



**Economic and Social
Council**

Distr.
GENERAL

ECE/CES/AC.6/2008/9
4 March 2008

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Meeting on Population and Housing Censuses

Eleventh Meeting
Geneva, 13-15 May 2008
Item 4 of the provisional agenda

DATA EDITING AND VALIDATION

**An overview of Editing and Imputation methods for the
next Italian Censuses**

Note by National Institute of Statistics of Italy

Summary

The National Institute of Statistics of Italy is preparing the next Population and Housing Censuses that will be held on 2011. In this paper, a short description of the main editing procedures for the 2011 Census is given. The impact of some likely census innovations on the editing and validation is also reviewed.

I. INTRODUCTION

1. Census data characteristics make the editing and validation phase a very complex task. Some characteristics are shared with data from other household surveys, e.g. the hierarchical structure of the collection and analysis units (households and individuals, buildings and dwellings) and the mixture of many qualitative and quantitative variables. Other characteristics are peculiar to census data, e.g. the huge amount of data that requires high efficient computational solutions and the presence of some small but important groups in the population difficult to enumerate (the elderly, babies, frontier workers, foreign persons, etc.), that demand for imputation actions able to preserve their actual distributions. A well-organized editing and validation strategy needs to be arranged in order to take into account the census data characteristics in an adequate way.
2. The editing and validation strategy adopted for handling the 2001 Italian census data was characterized by methodological and technical innovations aiming to tackle specific editing problems. In particular, new techniques and software tools were developed, implemented and successfully used so as to improve the quality of final results.
3. At present it is not still possible to give full details of the 2011 editing and validation strategy because some innovations affecting the editing and validation phases are being discussed for a possible introduction in the next Census.
4. As a matter of fact, it is stated that the 2011 editing and validation strategy will be built upon the experiences of the 2001 Census: a composite editing and validation process will be arranged by a hierarchical combination of several procedures addressing to specific problems. In particular, some editing procedures used in 2001 will be adapted to the 2011 questionnaire and re-used, some others will be improved and/or modified for taking account of the innovations that are going to be adopted. The use of generalized software will enable an easy data fitting and a prompt adjustment of the procedures to the new strategy's design.
5. The 2011 editing and validation strategy will also aim to reduce the processing time for contributing to improve the timeliness in disseminating results (EUROSTAT, 2008). With regard to this matter, a key role will be played by technologically up-to-dated hardware, highly efficient algorithms and proper planning, implementation and managing of the editing and validation procedures.
6. This document brings out the essential components of the 2011 editing and validation process pointing out some features to be taken into account in the preparing phase. Particularly, Section II describes the main editing procedures that will be used for the 2011 Census, while Section III points out the impact on the editing and validation strategy of the under discussion innovations in the survey design. Finally, in Section IV some considerations about issues are now being analyzed and future work are provided.

II. MAIN EDITING PROCEDURES FOR THE 2011 CENSUS

7. The editing strategy for the 2001 Census aimed at performing plausible imputations preserving the maximum amount of collected information. The strategy adopted to accomplish

this task consists in singling out the editing sub-problems and finding an appropriate solution for each of them. In consequence, an editing process composed of several (connected) procedures addressing to specific sub-problems and implementing suitable methods was put in place (Bianchi *et al.*, 2004 and 2005).

8. For computational and operational considerations, the editing of the variables from the Private Household Sheet¹ of the 2001 Population Census was performed in two sequential steps. The editing of the demographic variables (Year of birth, Sex, Relationship to the household reference person, Marital Status and marriage) was done before the editing of the other variables, named individual variables (Citizenship, Presence and accommodation, Educational degree and training, Professional and non-professional status, Work activity, Place of study or work). As a result, a lot of individual responses could be incorrectly erased (or improperly imputed) depending on the imputed values for the demographic variables. This problem was solved by using an ad hoc auxiliary variable (see sub-section A) based on the identification of the respondent path.

9. The optimization techniques played an important role in the treatment of the demographic variables. A new software based on optimization techniques was in fact developed for implementing the data driven and the (theoretical) minimum change approaches used for processing the 2001 demographic variables (see sub-sections B and C).

10. The identification of the respondent path was also an essential step in the editing strategy adopted for the individual variables. The paths identify classes of respondents having common characteristics (see sub-sections A and D). These classes were used as strata for error localization and imputation. The deterministic approach was essentially used for the error localization of the individual variables, while the new values were assigned by means of model based imputation (see sub-section D).

11. Particular attention was reserved to the editing of some small groups in the population, such as the centenarian, babies, frontier workers. Specific procedures were arranged for them using also the record linkage with the 1991 Census data (see sub-section E).

12. The editing strategy for the 2011 Census will be based on that one adopted in 2001. In particular, the same processing order (first the demographic variables and then the individual ones) will be applied and the main procedures used for treating the demographic and individual variables will be adapted and re-used. In the following sub-sections, a short description of the methodological aspects of these procedures is given.

A. Identification of the respondent path

13. According with the graph theory (Picard, 1980), a questionnaire can be represented as a connected graph, where the vertices are the variables and the answers define the edges. When the answer to a variable is required only for some values given to a previous collected variable, the

¹ Census form collecting information on persons who usually reside in the dwelling.

foregoing collected variable is named filter and the succeeding one is named dependent. As example, the Marital status is a filter variable for the dependent variable Years of marriage. 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 path is a sequence of distinct adjacent vertices. A path identifies a class of respondents having common characteristics. A path can therefore be considered a summary variable useful for identifying respondent sub-populations.

14. Missing responses and/or erroneous values can make uncertain the identification of the right path for some respondents and so the identification of the proper respondent classes. In fact, a failed edit record can have an erroneous path (not admissible path), that is a path not consistent with the questionnaire's compilation rules. In these cases it is necessary to find the most likely (admissible) path/s.

15. For 2001 Census an automatic procedure for the identification of the most likely path of each respondent was implemented. The procedure located the most likely path from the set of the admissible ones on the basis of the analysis of the responses given to filter and dependent questions. If a failed edit record admitted two or more admissible paths (multiple solutions), one was randomly selected on the basis of the observed frequency distribution of the respondent's most likely paths.

16. The respondent path was used to:

- (a) Compute the new auxiliary variable Subset of Admissible Value (SAV) of the Year of birth variable (Manzari et al., 2002). The SAV of Year of birth (in the following simply the SAV) was computed for each person within the household and defined a sub-domain of the Year of birth consistent with the highest number of provided responses. The SAV was used to identify donor strata when handling the demographic variables: a person in a passed edit household was a suitable donor for a person in a failed edit household (recipient person) if and only if his Year of birth was inside the SAV of the recipient person. Since the individual variables were handled in the second editing step, their values could be conditioned by the imputed value of the Year of birth variable and improper deletions could occur. The use of the SAV allowed to impute a Year of birth consistent with the highest number of individual variables so that, the loss of information due to improper deletions was strongly reduced.
- (b) Impute nonresponse or inconsistencies of Professional status, Working activity and Place of study or work variables. In particular, for each failed edit record a most likely path was first selected and a value to impute was then chosen by taking into account the respondent path, Place of dwelling, Class of age, Sex and other strata variables.

B. Joint use of data driven and minimum change approach by DIESIS

17. Population census data are collected at the household level with information for each person within the household. The demographic variables are related among different persons within the household by the so called between-persons edits and also inside the person by the

within-person edits. Therefore, the preservation of the relationships between the demographic variable can be guaranteed if the variables are processed at the same time for all the persons belonging to the same household that is at household level. To accomplish this task an editing system able to use both the between-person edits and the within-person edits should be used. As some demographic edits are expressed by linear inequalities, the editing system should also be able to deal with qualitative and quantitative variables simultaneously.

18. Preparing for the 2001 Census a new software, the Data Imputation and Edit System - Italian Software (DIESIS), was joint developed by National Institute of Statistics of Italy (ISTAT) and academic researchers (Department of Computer and Systems Science of the University of Roma "La Sapienza") (Bruni *et al.*, 2001). The DIESIS system allows to deal with qualitative and quantitative variables simultaneously, at household and individual level. After a rigorous statistical evaluation of its performance (Manzari and Reale, 2001), the DIESIS system was successfully used for imputing nonresponse and resolve inconsistent responses for the 2001 demographic variables.

19. Two editing approaches are implemented in the DIESIS system, the data driven and the (theoretical) minimum change, through the first donors then fields and the first fields then donors algorithms.

20. 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 and each passed edit household is calculated by a function 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 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). By using this algorithm the imputed values for a household come from a single donor household.

21. The first fields then donors algorithm first determines the minimum (weighted) number of variables to impute and identifies the potential donors (as previously described). Then, for each recipient person, the algorithm takes the values to impute from the donor person as similar as possible to the recipient one. This algorithm imputes the variables of one person in turn. If possible, the variables inside the person are imputed simultaneously. Note that the imputed values for a household may come from two or more donor households.

22. The two algorithms were jointly used for the treatment of the demographic variables, in order to balance the plausibility of the imputation actions with the preservation of the collected information. The first donors then fields algorithm was 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 was exceedingly high in comparison with the number of changes proposed by the second algorithm (the extent was set on the basis of the household size).

23. The first donors then fields algorithm was mainly used to process the households having common structure that are usually those having smaller household size. For these households it was generally possible to find enough potential donors. Otherwise, in the treatment of households having uncommon structure, usually those with largest size, few donors were generally available, and often they were not very similar to the failed edit household. In these cases the data driven imputation action would have required many changes to obtain an adjusted household passing the edits, therefore the minimum change approach was preferred.

C. Identification of the household reference person and potential couples

24. Two important procedures were run before the processing of the demographic variables. One aimed at validating the household reference person in the household, the other was designed for the identification of potential couples in the household.

25. One of the most important demographic variables is the Relationship to the household reference person. It is the basis variable for specifying all the between-person edits and most of the within-person edits. Moreover, it is necessary to define the family nucleus and hence the family structure. For each household, the reference person needs to be located (prior to editing the demographic variables) in order to allow all the remaining persons to define their relationship to it. Three possible erroneous situations can occur about the reference person:

- (a) one person has declared to be the reference person but his/her SAV (17 years old or younger) is not consistent with such a role;
- (b) more than one person has declared to be the reference person;
- (c) no one in the household has declared to be the reference person.

26. The procedure used to validate the household reference person for the 2001 Census data was based on optimization techniques and was carried out adapting the first fields then donors algorithm implemented in the DIESIS system to the specific problem. The procedure was composed of two main steps. In the first step the potential reference persons were identified. In particular:

- (i) in cases a) and c) the potential reference persons were all the persons in the household having a SAV consistent with the reference person role, that is the ones having a SAV consistent with an age of 18 years or older;
- (ii) in case b) the potential reference persons were the persons that had declared to be the reference person if their SAV was consistent with the reference person role, otherwise the potential reference persons were all the persons in the household having a SAV consistent with the reference person role.

27. If no person in the household had a consistent SAV, the potential reference persons were all the persons regardless of the consistent SAV requirement.

28. In the second step, the procedure selected the potential reference person which allowed to restore the household edit consistency changing the minimum (weighted) number of values of the demographic variables.

29. The definition of the family nucleus, and hence the family structure, is based on the analysis of the couples in the household. Some couples have unique relationship to the reference person (reference person, wife/husband; father, mother; father-in-law, mother-in-law). Other couples have non-unique relationship to the reference person (e.g. son/daughter, son/daughter-in-law). The persons forming unique couples are well identified in the household, if they are present. In the 2001, their editing was done by a set of ad-hoc between-person edits (couple edits). A more difficult problem is the editing of the non-unique couples because the persons forming them need to be identified first.

30. In order to preserve as much as possible the observed non-unique couples, for the 2001 Census a two-step strategy, drawn by the Canadian strategy (Bankier *et al.*, 1997; Bankier, 1999), was performed. In the first step the non-unique couples were identified prior to editing (potential couples). In particular, the identification of the potential couples was done by assigning a score to each possible pair of persons in the household and selecting the pairs with largest score. The score was based on the provided responses to the demographic variables and reflected the likelihood of the pair being couple.

31. In the second step, the couple edits were applied only to the persons forming the potential couples (components). At this aim an ad-hoc derived auxiliary variable was used. A potential couple could be either retained or eliminated by imputation. In the first case the components had appropriate values for the Relationship to the household reference person, Sex, Age, Marital status and Year of marriage variables, like the components of the unique couples. In the second case there were no couple constraints for the components.

D. Editing of individual variables

32. The individual variables are related only inside the person by the within-person edits, therefore they were processed at individual level (with the exception of the Citizenship variable). The editing process was composed of different procedures customized to the collected unit (private household, institutional household, person temporarily resident) and the specific measured topic (Place of birth, Citizenship, Presence and accommodation, Educational degree and training, Professional and non-professional status, Work activity, Place of study or work).

33. An analysis of the erroneous situations was first performed in order to classify the kind of errors (e.g. systematic coding or compilation errors, random errors). Then, the imputation of nonresponse and inconsistencies was carried out by using different methods. In particular, the systematic coding errors were treated by deterministic imputation while the standard rejection method was used for imputing non-response and inconsistencies due to random errors. The imputed values were drawn from the distribution functions calculated on the passed edit records grouped into strata defined also by the respondent path (see sub-section A).

E. Validation of centenarians

34. Centenarians represent a relatively small but important proportion of the total population. The strategy used for validating the centenarians enumerated in the 2001 Population Census is a combination of several methodologies directly applied to the raw micro-data captured by optical reading (Nuccitelli *et al.*, 2006). The main key stages of the strategy are as follows:

- 1) automatically match the records which refer to individuals born in the period 1888-1901 and enumerated in the 1991 Census with the records which refer to individuals born in the same period and enumerated in the 2001 Census, using exact (i.e. deterministic) matching;
- 2) automatically check for internal consistency of records unlinked at stage 1;
- 3) manually check for consistency with the respective questionnaire images which are stored during the process of optical reading of some ambiguous cases from stage 2.

35. With regard to the matching stage, two records were labelled as linked if the following conditions were simultaneously satisfied:

- (a) the records perfectly agreed on fields common to both data sets (matching variables);
- (b) the used matching key identified the links uniquely and did not contain missing values.

36. The individuals (enumerated in the 2001) corresponding to the linked records were considered validated. The unlinked individual presenting consistent items at the automatic check of stage 2 were considered validated.

III. EDITING AND VALIDATION PROCESS UNDER INNOVATIONS OF THE SURVEY DESIGN

37. Some innovations to the 2011 census survey design are currently being discussed at ISTAT (Crescenzi and Fortini, 2007).

38. Among the innovations most likely to be introduced, the ones having the major impact on the editing and validation phases are:

- (a) enumeration by short- and long-form questionnaires;
- (b) availability of registers:
 - (i) local registers on individuals, and their organization in private and institutional households, having their usual residence in the municipality – obtained by local population registers;
 - (ii) integrative registers on individuals, and their organization in private and institutional households, not usually residing but temporarily present in the municipality – obtained with data from local or central auxiliary sources;
 - (iii) local residential address lists;
- (c) use of multi-mode data collection: questionnaires collected by enumerators, transmit by mail or web, or collected by computer-assisted telephone interviewing (CATI).

39. For the 2011 Population Census a limited number of questions (mainly on demographic characteristics) could be asked to every person and housing unit while additional questions (on social-economic characteristics) could be asked to a sample of persons and housing units. In this case two census forms are necessary. The short-form questionnaire could contain information on the demographic variables (Sex, Year of birth, Marital status, Relationship to the household

reference person, Citizenship, Presence and accommodation, Educational level). Differently, the long-form questionnaire could contain information on the variables in the short-form and also School enrolment, Degree and professional training, Professional status, Working activity, Place of study or work and journey to study or work. Each household should receive either a short-form or a long-form.

40. The choice of treating the demographic variables (all contained into the short-form) before the individual variables is well suited to the new strategy of enumeration. However, the need of determining the *SAV* (see Section II A) with two different procedures could arise: one for the short-form, the other for the long-form. It must be noted that, for the individuals receiving the short-form, the *SAV* could be less effective, because the available information used for the *SAV* computation is strongly reduced.

41. The delivery of the long-form on sample basis will be able to have a negative impact on the availability (and therefore on the choice) of donors for the small groups in the population. This could reduce the accuracy of the imputation methods (donor-based and distribution-based) as it strongly depends on the availability of a large number of potential donors. High rates of item nonresponse joined to scarce donor availability could heavily cut down the quality of the final estimates. This negative issue could be prevented by managing the data collection and the donor pool selection phases with especial care.

42. Furthermore, the validation phase of the individual variables will be made more complex by the collection on sample basis. The raising weights should be promptly available for allowing both the test and tuning of the editing procedures and the comparison between the final data and the macro information coming from external sources or from administrative files. In this regard, a more effective validation phase could be performed if a detailed set of indicators (control system) was available to support the in-depth analysis of the inconsistencies between census data and other available sources. Therefore, a control system should be accurately planned and implemented.

43. A considerable contribution for the enhancement of the quality of the final data could be brought by the availability of some information (e.g. Sex and Date of birth) contained into local registers provided by the municipalities (points b1-b3). The information from the registers are usually correct and therefore could be used to:

- a) improve the quantitative control of the forms, that is the verification of the correspondence between the number of returned questionnaires and the "expected" number of questionnaires, with the goal of reducing under-enumeration;
- b) complete the census file by imputation of missing or inconsistent census values.

44. In this case, some studies aiming at determining effective strategies for the use of the register's macro- and micro-data have to be performed.

45. In particular, the record linkage between census data and register data could allow accurate imputation of missing or inconsistent census values. The difficulties related to the linkage between census and register data could mainly be due to the absence of reliable identifying information (in previous Censuses, the fiscal code of each individual was not

collected and the full name was not available to ISTAT) and to the required processing time. If the fiscal code was collected by the 2011 form, it could be used as unique identifier (matching variable) for the link with the register data. In this case, since the questionnaires will be mostly self-completed, also the full name should be made available in order to check the correctness of the matching variable. Otherwise, mistakes could occur in the matching phase resulting in useless or harmful use of the linked data for imputation purpose.

46. As alternative strategy, the register data could be added to the donor records and the donor-based imputation could be adopted. This strategy aiming at improving the similarity between recipient and donor records (donor records could be quite dissimilar from some recipient records) and hence the imputation performance. The accuracy and efficiency characteristics of the two alternative uses of the register data (linked-based imputation and enlarged donor-based imputation) should be assessed by an ad hoc study.

47. Finally, the impact of the multi-mode data collection is not negligible on the quantitative control of the forms. Procedures in aim to verify duplicate questionnaires, due to the use of different data collection modes, need to be arranged in order to avoid over-enumeration.

IV. CONCLUSIONS

48. The editing and validation processes for the next population Census will be mainly built upon the experience of the 2001 Census, but the strategies need to be adequately arranged in order to take into account the innovations likely to be introduced in the survey design. Special attention must be devoted to:

- a) the impact of the long-short form technique on the identification of the respondent path process, on the donor selection and on the validation phase;
- b) the opportunity to use information from local population registers, or other integrative sources, at macro and micro level;
- c) the availability of local residential address lists;
- d) the use of multi-mode data collection.

49. In particular, opportunities to the above points b)-d) allow to perform editing activities during the data collection process. Therefore the implementation of suitable supporting systems is necessary.

50. These innovations will have a relevant impact on the whole editing and validation process. Currently, some issues have already been analysed and solutions have been suggested, but others still need to be addressed. In addition, some studies are being carried out in order to reduce the computational time of the quantitative control process and to improve the detection of systematic errors. Since the timeliness in disseminating results must be improved, the editing and validation task is doubtless ambitious.

REFERENCES

- Bankier M., Houle A. and Luc M. (1997) 1996 Canadian Census Demographic Variables Imputation, *Proceedings of the UN/ECE Work Session on Statistical Data Editing*, Czech Republic (Prague).
- Bankier M. (1999) Experience with the New Imputation Methodology used in the 1996 Canadian Census with Extension for future Censuses, *Proceedings of the UN/ECE Work Session on Statistical Data Editing*, Italy (Rome).
- 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.
- Bianchi G., Manzari A., Pezone A., Reale A., Saporito G. (2005) New procedures for editing and imputation of demographic variables, *Proceedings of the UN/ECE Work Session on Statistical Data Editing*, Canada (Ottawa).
- 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*.
- Crescenzi F. and Fortini M. (2007) Due strategie per l'uso censuario di dati anagrafici, *Internal document (in italian)*, ISTAT.
- EUROSTAT (2008) Proposal for a Regulation of the European Parliament of the Council on population and housing censuses. Brussels, 11 January 2008.
- Manzari A. and Reale A. (2001) 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 *IASS Proceedings 53rd Session of The International Statistical Institute, August 22-29, 2001, Seoul*, pp. 634-655.
- Manzari A., Pezone A., Reale A. (2002) 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.
- Nuccitelli A., Pezone A., Reale A., (2006) The Validation of the Census Micro-Data on the Oldest Old Living in Italy, *Proceedings of Q2006 European Conference on Quality in Survey Statistics*, United Kingdom (Cardiff).
- Picard C. F. (1980) *Graphs and questionnaires*, North-Holland, Netherlands.
