

Evaluation of a New Approach for Edit and Imputation of Social and Demographical Data with Hierarchical Structure

Valutazione di un Nuovo Approccio per il Controllo e la Correzione di Dati Socio-Demografici con Struttura Gerarchica

Antonia Manzari, Anna Pezone, Alessandra Reale

ISTAT, Roma, Via C. Balbo 16 (00184)

manzari@istat.it, pezone@istat.it, reale@istat.it

Riassunto: In questo lavoro viene proposto un nuovo approccio, suggerito dalla Teoria dei Grafi, per ridurre, nel processo di correzione, la perdita di informazioni derivante da cancellazioni improprie dei dati, e per migliorare la qualità dell'imputazione, nel caso in cui vengano effettuati diversi passi di correzione per imputare dati con struttura gerarchica. Il nuovo approccio è stato confrontato con un approccio tradizionale usando i dati del Censimento della Popolazione del 1991; i risultati ottenuti hanno evidenziato un miglioramento consistente della qualità del processo di correzione. La metodologia proposta sarà, quindi, utilizzata per la correzione di alcune variabili rilevate con il Censimento della Popolazione del 2001.

Key words: hierarchical data, editing, imputation, graph theory.

1. Introduction

The Census Population (CP) data are characterised by a hierarchical structure: data are collected at the household (unit) level with information for each person (sub-unit) within the household. The editing and imputation (E&I) procedure, which has the aim to identify and to adjust errors in collected data by means of edit rules, has to take into account the constraints between the values of variables belonging to different persons within the household (*between persons* edit rules) together with the constraints between the values of variables belonging to a given person (*within person* edit rules). The large number of the questions (variables) and the complexity of their constraints don't always allow to perform one step the E&I process using only. The problem is, generally, tackled by dividing the variables into subsets that are handled in different E&I steps. The E&I steps are separately processed bearing in mind that if the variables subsets are unconnected (there are no edit rules defined on variables located in different subsets), the order of the processing runs is not of influence, while, if the subsets of variables are connected (there are edit rules defined on variables located in different subsets), during the performance of one of the processing runs it is necessary to maintain fixed all the variables imputed by run(s) processed previously. It is evident that the fewer the number of connections between the variables located in different subsets, the better will be the subdivision: the optimum result is obtained when the various subsets of variables are completely unconnected.

In case of CP data, if the choice is to handle some demographical variables as *Years of birth*, *Sex*, *Marital status*, *Years of marriage* and *Relation to Person 1* (*household variables* in the follow) in the first step and all the remaining variables (*individual variables* in the follow) in the subsequent steps, it is necessary to identify and, eventually to adjust the errors, taking into account the constraints between the *household variables* and the *individual variables* in order to preserve the maximum amount of collected information. The values of the *household variables* are not modifiable in the subsequent steps, therefore the inconsistencies between the *individual variables* and the *household variables* can be removed only changing the values of the *individual variables*. Therefore in some cases the values of the *individual variables* could be improperly deleted.

In this paper we propose an approach, suggested by the Graph Theory (Picard, 1980), to reduce the loss of information due to an improper deletion, and to improve the quality of the imputation of both the *household variables* and the *individual variables*. Section 2 describes the main characteristics of the method, while in Section 3 an evaluation study is illustrated on real data from the 1991 Population Census.

2. The methodology

A questionnaire is a sequence of questions. Each question can give rise to one or more variables. So the questionnaire can be considered as a sequence of variables. According with the graph theory, a questionnaire can be represented as a connected graph, where the vertices are the variables, while the answers defines 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 example, the *Marital status* is a *filter* variable for the *Years of marriage* variable. The *filter* variables are represented by vertices that give rise to more than one edge. In fact, the answer to a *filter* variable determines an edge entering into a subsequent vertex representing the *dependent* variable. Two vertices are adjacent if they are connected by an edge. A *path* is a sequence of distinct adjacent vertices. The *path* outlines a class of respondent.

A synthetic representation of the questionnaire can be obtained by means of variables subsets obtained by grouping adjacent vertices, where only the last one can be a *filter*. We name *groups* these subsets. As example, a *group* is composed by the questions whose answers are requested only by persons that go to a place of study or work: *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.

Representing the 1991 PC questionnaire, we determine 27 *groups*. We can observe that the *groups* of *household* variables are connected with some *groups* of *individual* variables mainly by means of the *Years of birth* variable. The *Years of birth* is one of the discriminant variables in outlining the path of a respondent. In fact, for 12 *groups*, it is a *filter* variable or it is connected to the variables located at least in a *group* by means of consistency edits (consistency edits check if a combination of values in a record is plausible).

Our propose is to handle the *household* variables in the first E&I step and to use the discriminant capability of the *Years of birth* variable, together with the information coming from the *groups*, to define a new auxiliary variable (the *Subset of Admissible Values (SAV) of Years of birth*) to control for the value of the *Years of birth* variable.

The *SAV of Years of birth* is defined for each individual within the household. The *SAV of Years of birth* is a composed variable obtained by the partition of the domain of *Years of birth* (1881-1991) in ten sub-domains. The breakpoints of each sub-domain are derived from the sub-domains of the *Years of birth* variable used in the *within person* edit rules. A dummy variable is assigned to each sub-domain (0 = not admissible sub-domain; 1 = admissible sub-domain). Many dummy variables can be equal to 1: when the sub-unit shows inconsistencies between the collected value of *Years of birth* variable and the collected values of the *individual* variables or when the value of *Years of birth* variable is missing and the information from the *groups* suggest different *SAVs of Years of birth*. We use the dummy variables derived from the *SAVs of Years of birth* as *key* variables to determine the records that can be used as donors. When *key* variables are specified, a failed edit record (record to be imputed) will be always corrected with a donor having at least an equal admissible sub-domain. In terms of *SAVs of Years of birth*, for each failed edit record, the selected donor will always have the values of the *Years of birth* variables inside of sub-domains consistent with the information provided by the *groups* of the failed edit record.

3. The evaluation study

We compared the performance, in terms of accuracy, of an E&I procedure with the new auxiliary variables, *SAVs of Years of birth* and without the new auxiliary variables.

Three different sets of data were compared: *original*, *perturbed* and *correct* data (Granquist, 1997). *Original* data correspond to the data observed under a perfect data production process, free of missing and inconsistencies according to the 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. We considered a subset of six-person household (about 20.000 units from a single region). *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 correspond to the edited and imputed values obtained by processing the *perturbed* data by means of the investigated E&I procedures. The *household variables* for the test were *Relation to Person 1*, *Sex*, *Marital Status*, *Year of birth*, and *Year of marriage*. The *individual variables* used for the test are related to: usual residence at 1986, educational level, professional or non professional status, time really worked, occupation, industry.

We strongly perturbed original data by replacing valid responses with *non-response* (the original value was randomly replaced by a missing value) or with *other valid response* (the original value was replaced by a wrong one randomly chosen in the admissible domain). The data were processed by a new E&I system called DIESIS (Bruni et al., 2001); the results refer to runs performed by the joint use of the “*first donors then fields*” and of the “*first fields then donors*” imputation algorithms with the *involved sub-units donor* and *involved sub-units imputation action* options. The distance functions used were the same for both comparing procedures.

For each *household variables*, we computed indicators evaluating the error detection performance of the editing process and evaluating how much the imputation procedure preserves the individual values (Manzari & Reale, 2001). The indicators used:

- the percentage of not modified data erroneously imputed (E_true);
- the percentage of modified data not imputed (E_mod) ;
- the percentage of imputed values for which imputation is a failure (I_imp).

For qualitative variables the imputation process is considered as a failure if the imputed value does not equal the original one. For numeric variables the imputation process is considered as a failure if the imputed value differs by more than $\pm 10\%$ from the original one. As Census data are the primary source of information about the *household typology*, which is a derived variable from all the considered *household* variables, we evaluated also the capability of the compared E&I procedures in preserving the original value and the distribution of this derived variable. We evaluated the preservation of the *household typology* variable by computing the percentage of off-diagonal entries considering the square tables obtained cross-classifying the original and corrected *household typology* categories.

Applying the *groups* and the *SAV of Year of birth*, we can observe a general better performance of the imputation process and a reduction of the failure indicators (E_true , E_mod and I_imp) for all variables; generally this reduction can be estimated more than 50%. Moreover the *SAV of Year of birth* preserves the maximum amount of collected information related to household and individual variables, and doesn't modify the marginal distribution of variables.

The obtained results support our opinion that this methodology can be successfully applied to the 2001 Population Census data.

4. Conclusions

In this paper is proposed a method to reduce the loss of information due to an improper deletion, and to improve the quality of the E&I of data with hierarchical structure.

The statistical performance of the proposed approach has been strictly evaluated and the results are very encouraging. The proposed approach, ensuring a better preservation of the original values, improves significantly the quality of the E&I process.

References

- R. Bruni, A. Reale, R. Torelli (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*.
- 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.
- 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", IASS Proceedings for ISI 2001 (in press).
- Picard C. F. (1980), *Graphs and questionnaires*, North-Holland, Netherlands.