Data imputation processes based on statistical analysis: the case of Kosovo census data

Marco Scarnò, Bekim Canolli, Servete Muriqi, Hisni Ferizi

Abstract: Data collected by statistical offices may contain errors, like “strange data”, aberrations, erroneous values occurred during the enumeration phase or during the data entry. These errors have to be corrected before reliable statistical information can be published. In the last two years the development group of the ADaMSoft software, in a strict collaboration with Istat, the Italian National Statistical Institute, enriched the software with a series of procedures aimed at localizing, in such context, the minimum number of fields that should be corrected and to attribute them more suitable values.

These imputation procedures have been successfully applied to a series of numerical questions during the last Kosovo population and housing census of 2011. The most frequent cause of error observed was due to missing values in the balancing totals.

1 Data Editing and Imputation

Errors in data can be detected by specifying several constraints that have to be satisfied by the values observed for each individual/respondent. These constraints are called edit rules (or edits for short); for continuous variables they are represented by sets of linear equalities or inequalities.

In a questionnaire these edits can derive directly from the structure of the questions (for example using the total of some fields to verify the coherence between different answers) or can be introduced by using some external information (like the minimum or the maximum of some variables).

When a record fails an edit, it is considered erroneous, otherwise it is considered correct. The main issue of the error localization problem lies in the fact that, if a record is found erroneous, this does not mean that all its values have to be considered invalid; in these cases the objective is to find the minimum number of fields to change in order for the record to satisfy all the edits at the same time. This problem is of great interest because it respects the need to keep as much as possible the information given by the respondents; the methodologies that are used in such context are called “data editing and imputation” and refer to Error Localization Problems (ELP).

The first description of a model for automatically identifying errors in a record was made by Fellegi and Holt [4]. There are some tools that can be used to solve ELP; however those that can be used for free cannot treat more complex cases. For this purpose the ADaMSoft development group added to its Open Source statistical software several procedures that can manage such kind of problems. These procedures were tested on some parts of the Kosovo census data and resulted able to identify and correct the errors found, mainly characterized by missing values in the total fields.

2 Editing and imputation steps in ADaMSoft

The mathematical solution to ELP is usually found by identifying the vertices of the polyhedron defined by a system of equalities and inequalities; an example of such vertex generation algorithm is the one proposed by Chernikova [1, 2]. For a complete review of all the methods see De Waal [3]. It should be noted that all the methods solve a system in which both the edits and the variables are the double of the original ones. This is because the evaluation of the positive and negative variation of each variable is supposed.

In our case we refer to a general purpose library, called *polco*, distributed under the Simplified BSD License and written in Java. This library receives as input a system of linear equalities and inequalities and produces, as output, a matrix with extreme rays as column vectors. For more details on how the library works see Terzer [7].

The details of the algorithm used in ADaMSoft were already presented in a previous work [5]; it can treat situations with many variables; in particular it identifies a hierarchy of these (for each erroneous record) and iteratively tries to solve a system of inequalities smaller than the original one. So, as first step, it considers only the corrections for the variables directly involved in the not satisfied edits. In the next step (if no solutions were found) to these are added all the other variables that are involved in those edits in which they appear. As stated before, this heuristic process will continue till valid solutions will be found. Obviously the algorithm will always converge; it will be much faster if the edits contain disjoint variables.

Last year other three procedures to perform the imputation steps were added in order to complete the package; in particular:

- **Deterministic imputation**: this procedure analyses each field previously identified as requiring imputation to determine if there is only one vector of possible values which would satisfy the original edits;
- **Donor imputation**: this procedure uses a nearest neighbour approach to find, for each record requiring imputation, a valid record that is most similar to it and that will allow the imputed recipient record to pass the user-specified post imputation edits (a set of edits that can be more relaxed in respecting to the original ones). The imputation is performed

if such a record is found. It should be noticed that donor imputation is the preferred method of imputation because all fields requiring imputation are taken from the same donor record and the relationships between the imputed variables are therefore retained;

- **Vertices imputation**; this procedure will use the solutions (vertices) found during the error localization step to find the correct values to impute; its results letting a record to satisfy all the edits but don't consider the relationship between the variables as they are observed in the data set.

For a detailed explanation of all the ADaMSoft procedures for E&I see [6].

3 The importance of correcting data in Kosovo census

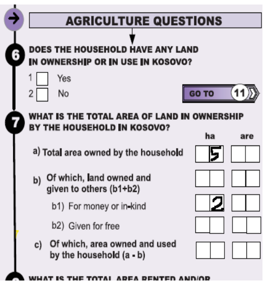
The 2011 population and housing census is an historic time in Kosovo because it is carried-out for the first time in 30 years. In this young country, such undertaking faces the specific challenge of getting confidence from a large public never exposed to census data, and therefore often dubious about its methodology and the reliability of its results. In such context, it is of crucial importance for the Kosovo Agency for Statistics (KAS) to analyse thoroughly the census data quality and simultaneously to correct any error found with the most proven and recognized statistical methods.

Thanks to a project of technical assistance financed by the European Union[2] KAS staff was trained to these methods and to the use of ADaMSoft software in the performance of data quality assessment and corrections. The exercise will above all result in coherent and consistent final census data, which is essential for the objectives of data relevance, accuracy and coherence, three of the main criteria for statistical data quality required by the European Statistics Code of Practice.

4 The details of the E&I in the Kosovo census

In the Kosovo Census questionnaire there is a part in which the household has to answer to 18 questions on the land owned and on its usage. Among the answers, some fields contain totals or subtotals. Figure 1 presents an example of such questions, with a sample of errors as it could have happened during the enumeration phase.

[2] "Technical Assistance to the Statistical Office of Kosovo for the preparation of a population census", an EU-funded project managed by the European Union Office in Kosovo; Implemented by Istat, CIRPS-Sapienza University of Rome, ICON-INSTITUT and Rrota



AGRICULTURE QUESTIONS

6 DOES THE HOUSEHOLD HAVE ANY LAND IN OWNERSHIP OR IN USE IN KOSOVO?

1 □ Yes
2 □ No GO TO 11

7 WHAT IS THE TOTAL AREA OF LAND IN OWNERSHIP BY THE HOUSEHOLD IN KOSOVO?

	ha	are
a) Total area owned by the household	5	
b) Of which, land owned and given to others (b1+b2)		
b1) For money or in-kind	2	
b2) Given for free		
c) Of which, area owned and used by the household (a - b)		

WHAT IS THE TOTAL AREA RENTED AND/OR

Figure 1: an example of the fields that were treated

It was possible to define nine edits that should be simultaneously satisfied by the values given by each respondent; among these 4 were inequalities while the others were equalities (balancing edits). A first analysis showed that 27% of cases (on a total of 297078 households) were incoherent. This implied the necessity to proceed with editing and imputation steps.

The ADaMSoft procedures took few hours to localize the minimum number of fields to impute and to substitute them with coherent values; in particular:

- The localization steps found that in 95.3% of records it was possible to consider a deterministic solution, while in the remaining records it was needed an imputation that uses the donors or the vertices methods;
- The donor imputation was able to correct all the records except 2, that were imputed using the vertices method.

Table 1 presents some results regarding the percentages of the changes in some variables by evaluating the ratio between their value after and before the editing and imputation steps.

Table 1: Some results of the editing and imputation steps in the Kosovo census data

Variable	*Percentages of changes after and before the E&I*
Land owned and really usable by the household	135.7%
Land owned and given to others for money	99.5%
Land owned and given to others	129.9%
Total area rented and used by others in Kosovo	114.7%

A further analysis verified that the most frequent cause of errors in the records were due to the enumeration phase, during which the totals were not reported. Only for 5% of the questionnaire the errors derived from not coherent values.

5 Conclusion

The results from the E&I steps showed that the data collected during the Kosovo census were corrected in the majority of the cases; the main source of incoherencies in these were due to a "relaxed" use of the questionnaire, i.e. to those situations in which the enumerator filled only the significant fields for each household.

But as important outcome of the training, the KAS capacity to autonomously deal with error localization problems and to proceed with editing and imputations on any type of data set represents the most valuable asset for further institutional development.

References

1. Chernikova, N.V., Algorithm for finding a general formula for the non-negative solutions of a system of linear equations. USSR Computational Mathematics and Mathematical Physics, 4, 151-158 (1964)
2. Chernikova, N.V.,. Algorithm for finding a general formula for the non-negative solutions of a system of linear inequalities. USSR Computational Mathematics and Mathematical Physics, 5, 228-233 (1965)
3. De Waal, T., Processing of erroneous and unsafe data. Ph.D. Thesis, Erasmus University Rotterdam (2003)
4. Fellegi, I.P., and Holt, D., A systematic approach to automatic edit and imputation, Journal of the American Statistical Association, 71, 17-35 (1976)
5. Scarnò M and Caramanna L., Two algorithms for Error Localization Problems, 45th Scientific Meeting of the Italian Statistical Society (2010)
6. http://adamsoft.caspur.it/English/Guide.html
7. Terzer, M. and Stelling, J., Large scale computation of elementary flux modes with bit pattern trees, Bioinformatics, July 28 (2008)