UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing

3-6 October 2022, (virtual)

Multiple software systems for the editing and imputation process of the 7th General Census of Agriculture

Simona Rosati, Maria Teresa Buglielli, and Laura Tosco (Istat – Italian National Institute of Statistics, Italy)[1]

sirosati@istat.it, bugliell@istat.it, tosco@istat.it

I. Introduction

1. The 7th General Census of Agriculture was characterized by a very complex data set with a large number of qualitative and quantitative variables: hundreds of variables and more than 1 million of records were processed. When handling with both qualitative and quantitative variables, different methods and techniques are required for the treatment of item nonresponses and inconsistencies. This often leads to variables being treated separately according to their nature, which entails an equally complex data editing and imputation process.

2. The statistical data editing process of the Agricultural Census 2020 was designed and developed as follows:

(a) outliers and influential errors were identified by means the R package SeleMix;
(b) Banff procedures were used for data editing and imputation of sections A-B-C of the questionnaire, with a large number of quantitative variables;
(c) the remaining sections of the questionnaire, from D to G, were entirely treated with R packages for editing and imputation.

3. The aim of this paper is twofold. Firstly, a description of the editing and imputation process is provided, showing the strategy, and the use of the related software tools in detail. Secondly, the different software tools are compared in terms of their advantages and disadvantages. In addition, a section is devoted to mass imputation based on administrative and census data combined, performed with both R and Banff.

II. The editing and imputation process

4. The editing and imputation (E&I) process of the 7th General Census of Agriculture was completely redesigned in order to use modern statistical E&I methodology consistent with internationally recognised principles and standards (GSDEM, 2019). As already mentioned, the major difficulty was to deal jointly with common relationships between qualitative and quantitative variables. This entailed the integrated use of several methods and software tools depending on the nature of the variables involved in the E&I process.

5. For a better understanding of the E&I process, a brief description of the data involved is given below.

6. The questionnaire of the Agricultural Census 2020 was divided into seven parts on the basis of quite broad thematic areas as follows:

- Sections A-B-C collected information on general characteristics of the agricultural holding, utilised agricultural area (UAA), size of livestock holdings, and manure management system; these sections included mainly quantitative variables.

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the Italian National Institute of Statistics. The authors thank Diego Moretti (Istat, dimorett@istat.it) for contributing to the computer elaborations.

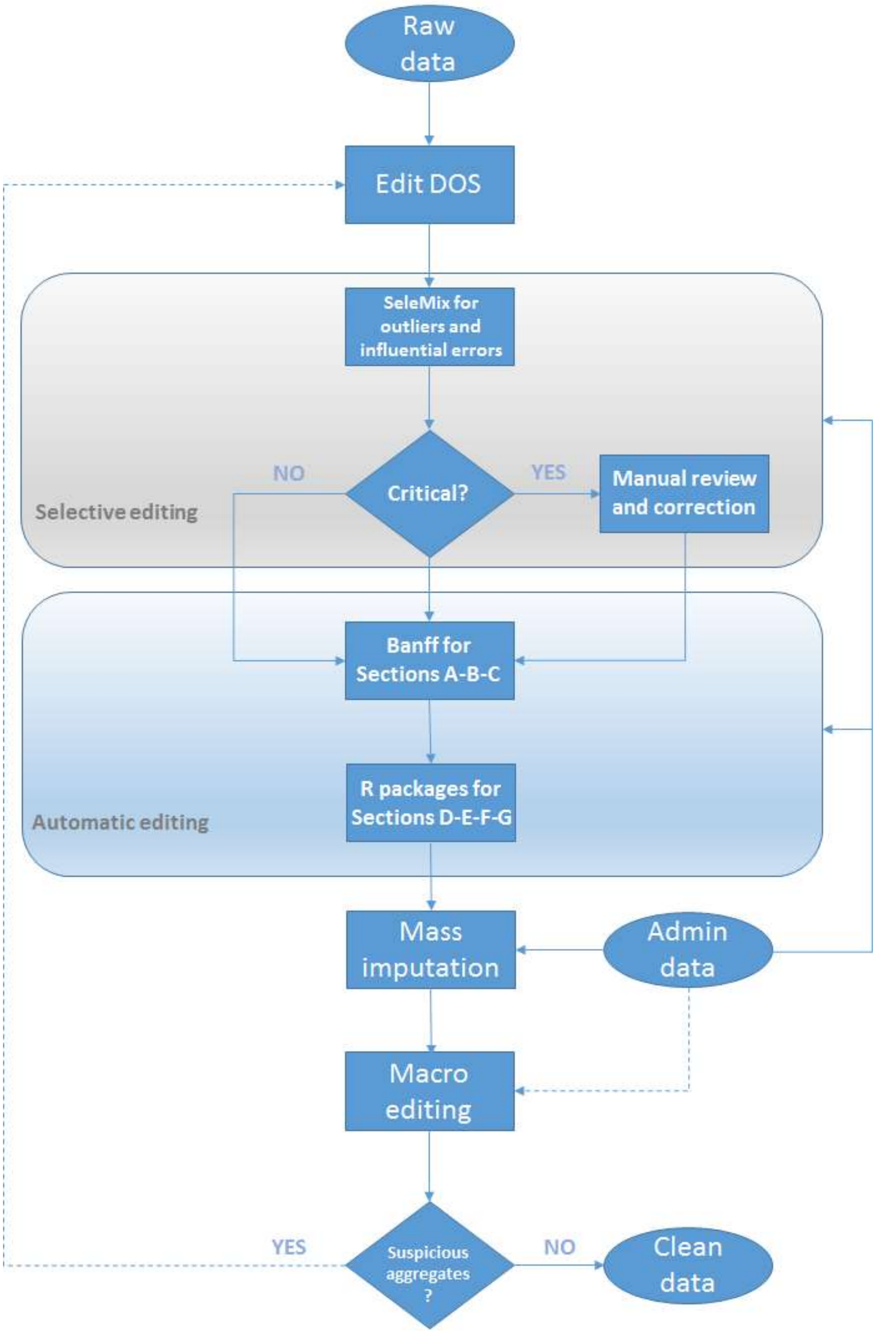
- Section D collected information on the manager, if not the same as the holder, and other gainful activities (OGA) directly related to the farm.
- Section E collected information on the holder and the human resources employed by the farm.

7. Additional information, such as revenue, marketing, innovation, computerization, and others, was collected in Section F. A specific part of the questionnaire was designed to evaluate the economic impact of the Covid-19 epidemic on the farms, in terms of changes in production volume, lost workdays, and the possible decline in sales on the national and international market (Section G).

8. Auxiliary data sources, such as administrative data and registry data, were used both at the macro level, for the validation of the census data, and at the micro level for comparisons between detected errors and auxiliary data, when the census micro data were inconsistent.

9. The E&I process flow can be depicted in Figure 1; details on methods and software tools will be provided in the next sections.

Figure 1. The E&I process flow - Agricultural Census 2020



III. Selective editing

10. After correcting any systematic or other obvious errors (‘edit DOS’ in Figure 1), outliers and influential errors were detected. The R package **SeleMix** (Guarnera and Buglielli, 2020) was used for this purpose. The underlying methodology is based on particular latent class models known in the literature as *contamination models*. More specifically, true data are modelled through a normal or log-normal distribution, and the “intermittent nature” of the error mechanism is captured through a Bernoullian random variable associated with the error occurrence. Given the assumptions, the resulting distribution for the observed data is a mixture of two Gaussian distributions with the same mean vector but proportional covariance matrices, where the “largest” one corresponds to contaminated data (Di Zio and Guarnera, 2013). For each unit, a score is calculated in terms of the difference between the predicted value and the observed value of variables of interest. Then all units are sorted (in descending order) according to their score. Influential units (critical units) are those with the highest score.

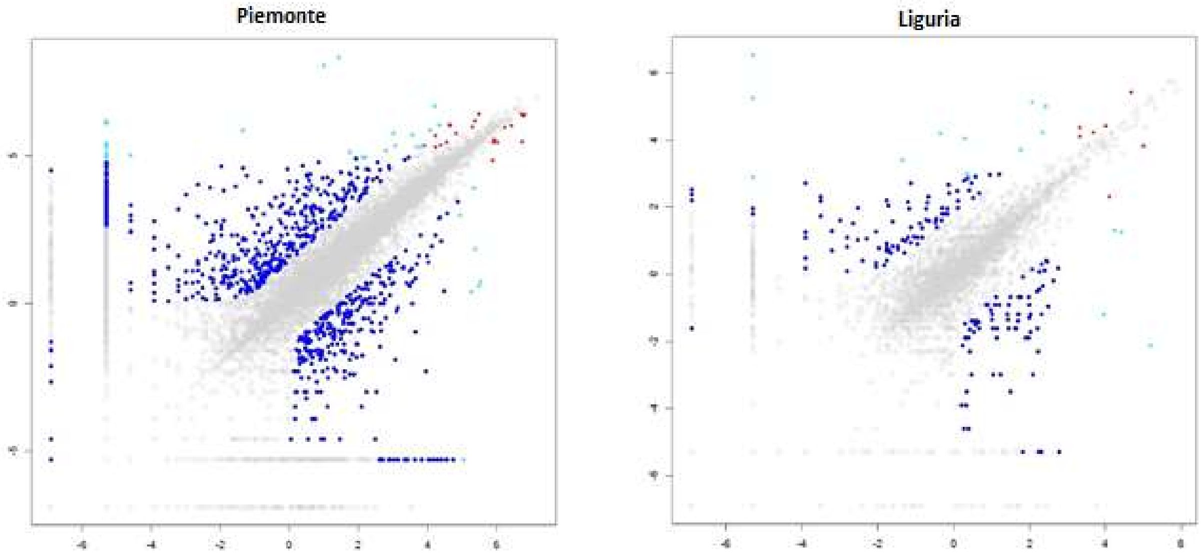
11. Outliers and influential units were manually reviewed by subject matter experts, as they could be affected by errors.

12. Outliers and potential influential errors were detected for the main variables, such as total area of the agricultural holding, utilised agricultural area (UAA), and livestock units (LSU) by geographical region. Moreover, for each outlier and influential unit any inconsistencies relating to the variables involved were flagged in order to facilitate the interactive editing.

13. To make this operation more efficient, selective editing started in advance, i.e. during data collection phase, which began in January 2021 and ended in July 2021. During this period, outliers and potential influential errors were detected at two different times when enough data were available: the first was in May (time 1) and the second in June (time 2). An additional step was necessary, at the end of data collection, to identify any residual outliers and influential units. At time 1, 1,694 units containing outliers or potential influential errors were identified, while at time 2, 4,383 units were identified. All these units, with outliers or potential influential errors, were distributed to the respective regional institutions for manual review.

14. Figure 2 shows the outliers and influential errors of UAA for two Italian regions, Piemonte and Liguria. Outliers are highlighted in blue, while influential errors are in red; if they are both outlier and influential they are highlighted in blue cyan.

Figure 2. Outliers and influential errors of UAA for Piemonte and Liguria



Piemonte

Liguria

IV. Banff for quantitative variables

A. The procedures for E&I

15. Quantitative variables (sections A-B-C of the questionnaire) affected by random errors were treated with Banff. The Banff system for E&I is a collection of specialized SAS procedures developed at Statistics Canada. It is a generalized system to process numeric and continuous variables, using consistency rules (edits) that must be expressed in linear form (Banff, 2014).

16. The independence of the SAS procedures in Banff gives the user a great deal of freedom and flexibility. However, this independence also entails more responsibility for the user in ensuring that the inputs are of good quality, and that the outputs are interpreted and applied correctly (Banff, 2014). The procedures used to perform the E&I process at this stage are listed below:

(a) **Proc Verifyedits**. Basically, this procedure allows the specification and analysis of the edit rules. In this procedure, a group of edits is checked for consistency, redundancy, determinacy and hidden equalities. The procedure also generates the extreme points, or vertices, of the feasible region or convex region bounded by the edits. These points represent the most extreme acceptable data records and give the user a better understanding of the shape of the feasible region which is being specified. The extremal points are determined through the use of Chernikova's Algorithm (Chernikova, 1964, 1965). A complete description of its use in Banff may be found in Schiopu-Kratina and Kovar (1989). For the Agricultural Census 2020 more than a thousand edits were specified. This was one of the most difficult tasks of the overall E&I process; it took several weeks to obtain a complete set of coherent and non-redundant edits.

(b) **Proc Editstats**. This procedure applies a group of edits to a SAS dataset and determines if each observation passes, misses or fails each edit. Five tables summarizing the status codes are produced and may be used to fine-tune the group of edits or to evaluate the effects of imputation. The summary statistical tables that are offered as output include:

- Number of records that pass, miss or fail each edit.
- Number of records that pass, miss or fail *k* edits.
- Overall counts of records that pass, miss or fail.
- Number of times each variable is involved in an edit that passes, misses or fails.
- Number of times each variable contributes to the overall record status.

(c) **Proc Errorloc**. The purpose of the error localization procedure is to identify the fields which must be changed in each individual record in error so that the record can be made to pass all the edits. The original data are not changed in this procedure. The fields which require imputation are identified during the execution of error localization, but no imputation actually takes place. The strategy used by Banff is to minimize the number of fields requiring imputation. In other words, it would be impossible to make a record pass the edits by changing fewer fields than the number of fields identified in the solution provided by error localization. This is an application of the *Rule of Minimum Change* as proposed by Fellegi and Holt (1976) and developed by Sande (1979).

(d) **Proc Deterministic** This procedure analyzes each field previously identified as requiring imputation to determine if there is only one possible value which would satisfy the original edits. If such a value is found, it is imputed during execution of this procedure.

(e) **Proc DonorImputation**. The DonorImputation procedure uses a nearest neighbour approach to find, for each record requiring imputation, the valid record that is most similar to it and that will allow the imputed recipient record to pass the user-specified post imputation edits. The imputation is performed if such a record is found.

B. Advantages and disadvantages

17. We observed some advantages and disadvantages of Banff that can be summarized as follows.

(a) Advantages:

   a. Simple to use and flexible. As already mentioned, the independence of the SAS procedures in Banff gives the user a great deal of freedom and flexibility.
   b. Minimizes the number of fields to change, hence it preserves as much of the original data as possible.
   c. Ensures that erroneous records are imputed in such a fashion to satisfy all the edits.

(b) Disadvantages:

   a. Not suitable for handling qualitative variables.
   b. Not suitable for systematic errors.
   c. Edits must be linear equalities or inequalities.
   d. Imputation may be unsuccessful because no suitable donor is available. Parameters need to be changed or some edit constraints need to be relaxed.
   e. Possible post adjustments.

V. Treating mixed type of variables

A. R packages for E&I

18. To implement the E&I process on data of the remaining sections of the questionnaire, from D to G, we designed and implemented R scripts using the following main packages:

(a) *validate*, which provides functions to formulate edit rules written as positive logical formulas, to confront data and analyze or visualize the results. Rules can be per-field, in-record, cross-record or cross-dataset (Van der Loo *et al.*, 2022).

(b) *validatetools*, a set of functions for finding redundancies or contradictions between the rules formulated with validate (De Jonge *et al*., 2020).

(c) *errorlocate*, which implements functions to localize records violating defined edit rules (De Jonge and Van der Loo, 2022).

(d) VIM, which provides the functions to impute missing values or erroneous values replacing with missing values to be imputed. We used the *hotdeck* function that implements the sequential hot deck algorithm, or alternatively, a random hot-deck algorithm, and the *kNN* function that implements the K-Nearest Neighbor Imputation method based on a variation of the Gower distance for numerical, categorical, ordered and semi-continuous variables (Templ *et al*. 2022; Kowarik and Templ, 2016). However, we encountered some computational resource limits with R that were solved by decomposing a problem into sub-problems.

19. The R scripts to perform the E&I process at this phase, have a common skeleton process structure that we briefly summarize in the following numbered list:

(1) Edit rules are written in a separated file in the form of a set of logical positive formulas.

(2) The edit rules file is the input of the *validator* function, implemented in the *validate* package, that creates a *validator object*, an internal data structure of the package.

(3) Edit rules are evaluated applying the *confront* function in *validate* package. This function evaluates each rule, one by one, on the input dataset and outputs a so-called *confrontation object* containing the logical results of each evaluation.

(4) The initial dataset is divided into the set of records violating the edit rules and the set of records satisfying the rules, obtained applying the *violating* and the *satisfying* functions in *validate* package, respectively.

(5) Localization of errors (fields in records violating the edit rules) using the *errorlocate* function.

(6) Applying of the *replace_errors* function (in the *errorlocate* package). This function sets the values of the erroneous variables to NA.

(7) Imputation of NA values set in step 6. Two different functions implementing two different methods were used to impute the Census data. Specifically, we applied the *KNN* function, in VIM package, to impute item nonresponses and erroneous values, while we applied the *hotdeck* function, in VIM package, to impute values for unit nonresponses. After imputation, all data should satisfy the edit rules.

B. Pros and cons

20. Testing the R packages to realize the E&I process of the Agricultural Census 2020 we observed some positives and some negatives that are listed below.

(a) **Pros**:

- R language and its implemented packages give experts great flexibility in design and realization of the E&I process.
- Easy management of large and complex amounts of data.
- Joint treatment of quantitative and qualitative variables.

(b) **Cons**:

- Error localization can be time-consuming and sometimes requires specific strategies to reduce the complexity of the algorithm.
- Despite the iterative application of steps 5, 6 and 7 (*errorlocate* and *replace_errors* functions and imputation, respectively), the process did not converge to a global solution, that is, a solution that satisfied the set of edit rules. To achieve the global solution it was necessary to guide the process by choosing which variables to impute from those involved in the failed edit rules, in addition to those already localized.

VI. Mass imputation

21. As shown in Figure 1 the last step of the E&I process of the 7th General Census of Agriculture was the mass imputation, which consisted of imputing on the one hand records with a high number of missing values, and on the other hand units nonresponse that were estimated eligible.

22. Mass imputation was carried out using both Banff and R. Banff was applied to the sections of the questionnaire including mainly quantitative variables (sections A-B-C), while R packages for E&I were used for the remaining sections of the questionnaire with mixed types of variables.

23. Before performing the mass imputation, the units to be imputed were integrated with information from administrative data sources where possible (e.g. for a given farm, the total area of the farm was imputed with the corresponding value from administrative data). Unit nonresponses without administrative data were discarded. Subsequently, for each erroneous record (with missing values) a suitable donor record was chosen among original (raw) records that passed all the edits (exact records).

24. Mass imputation was relatively fast and simple with both R and Banff. In Banff, *Proc Massimputation* was applied, while in R, VIM package was used.

25. It is worth noting that administrative data were used throughout the E&I process, both to review outliers and potential influential errors, and to resolve inconsistencies on the core variables.

VII. Conclusions

26. One of the main objectives of the E&I process of the Agricultural Census 2020 was to ensure the internal consistency of records with respect to the set of the specified edit rules. To achieve this aim two main software tools were used: Banff for quantitative variables and R for mixed types of variables.

27. R for E&I was initially introduced as an experiment, as it was used for the first time in such a complex survey data. Then, based on the good results, we realised that it could be used regularly also in this context. Moreover, it makes possible to deal with mixed types of variables as well as to easily manage large and complex data sets. For these reasons, we will promote the use of R and its packages in those surveys whose E&I process needs to be redesigned or modernised.

References

Banff. *Functional description of the Banff system for edit and imputation*, Version 2.06, Banff Support Team, Statistics Canada, 2014.

N.V. Chernikova. Algorithm for finding a general formula for the nonnegative solutions of a system of linear equations. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 4: 151-158, 1964.

N.V. Chernikova. Algorithm for finding a general formula for the nonnegative solution of a system of linear inequalities. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 5: 228-233, 1965.

M.P.J. van der Loo, E. de Jonge and P. Hsieh. *validate: Data Validation Infrastructure.* R package version 1.1.1, URL https://CRAN.R-project.org/package=validate, 2022.

E. de Jonge, M.P.J. van der Loo, and J. Daalmans. *validatetools: Checking and Simplifying Validation Rule Sets.* R package version 0.5.0, URL https://CRAN.R-project.org/package=validatetools, 2020.

E. de Jonge and M.P.J. van der Loo. *errorlocate: Locate Errors with Validation Rules*. R package version 1.1, URL https://CRAN.R-project.org/package=errorlocate, 2022.

M. Di Zio and U. Guarnera. Contamination Model for Selective Editing. *Journal of Official Statistics*, 29(4): 539-555, 2013.

GSDEM. *Generic Statistical Data Editing Model.* Version 2.0, Unece, URL https://statswiki.unece.org/display/sde/GSDEM, 2019.

U. Guarnera and M.T. Buglielli. *SeleMix: Selective Editing via Mixture Models*. R package version 1.02, URL https://CRAN.R-project.org/package=SeleMix , 2020.

I.P. Fellegi and D. Holt. A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, 71(353):17–35, 1976.

A. Kowarik and M. Templ. Data Validation Infrastructure for R. *Journal of Statistical Software*, 74(7), 1–16, 2016.

G. Sande. Numerical Edit and Imputation, *42nd International Statistical Institute Meeting*, Manila, Philippines, 1979.

I. Schiopu-Kratina and J.G. Kovar. Use of Chernikova's algorithm in the Generalized Edit and Imputation System. *Methodology Branch Working Paper*, No. BSMD-89-001E, Statistics Canada, 1989.

M. Templ, A. Kowarik, A. Alfons, G. de Cillia, B. Prantner, and W. Rannetbauer. *VIM: Visualization and Imputation of Missing Values*. R package version 6.2.2, URL https://CRAN.R-project.org/package=VIM, 2022.