WP. 2
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (i): Editing of administrative and Census data

The Main Innovations of Data Editing and Imputation for the 2010 Italian Agricultural Census

Key Invited Paper

Prepared by Gianpiero Bianchi, Rosa Maria Lipsi, Giuseppina Ruocco, and Michele Antonio Salvatore
Istituto Nazionale di Statistica, Italy[1]

I. Introduction

1. In planning the ongoing 6th Agriculture Census, a new Editing and Imputation System (E&IS) has been implemented, in order to reduce the total census error. The E&I System is a combination of different methodologies chosen and tuned on the basis of several simulations and validation tests, to verify the effectiveness of statistical methods adopted. In brief, the E&I strategy allows to identify the most appropriate method to solve problems of missing, invalid or inconsistent values in collected data, preserving the largest amount of available information. The following paragraphs describe the main features of the E&I System implemented. In particular, after deepening the E&I strategy, the fitting of the editing process in data collection stage is considered. As scheduled in the survey organization plan, the follow-up to the most relevant units, having anomalous values, has contributed to gather accurate information for the clerical review of outliers. Finally, the main editing and imputation steps and the adopted methodologies are highlighted.

II. Overview of the E&I System

A. The E&I strategy

2. According to the EU recommended practices for editing and imputation in Cross-sectional Business Surveys (Luzi et al., 2007), the overall E&I strategy is based on the following main guidelines:

(a) Adopting a quality oriented approach by performing the E&I process from data collection to the final figures;
(b) Review of the outliers and influential errors (selective editing) by enumerators, before the end of data capturing, in order to reduce the errors that may have substantial impact on data dissemination. After data census field operations, detected outliers and influential errors will be manually reviewed, by experienced Istat staff;
(c) Scheduling of two main correction phases, centrally managed by Istat, once the data capture stage is over;
(d) Use of techniques that minimize the number of changes especially for the treatment of not influential random errors;
(e) Computing quality indicators to monitor the main steps of E&IS;

[1] gianbia@istat.it, lipsi@istat.it, giruocco@istat.it, salvatore@istat.it

(f) Ad-hoc documentation to evaluate the outcome of the procedures, paying particular attention to changes due to the E&IS process.

3. The E&I process is performed in two main stages. While data editing (for the error localization) will be repeated at the beginning and at the end of each E&I step on the whole set of data, in the first stage, only the variables for the dissemination of provisional figures (primary variables, describing the farm structure) will be corrected.
In the second stage, all the remaining variables (secondary variables) will be corrected, considering the constraints between primary and secondary variables and preserving the values imputed in the previous runs of procedures.

4. This approach adopted has resulted in many innovations introduced in the new E&I System. Some of them are briefly described below:

(a) Inclusion of a subset of edit rules in the data capture stage;
(b) Use of Forward Search methods for the outliers detection;
(c) Use of administrative sources (for example, AGEA Archive, Viticultural Land Register, Bovine animals Register) for micro and macro data checks;
(d) Use of score functions to prioritize records to be manually reviewed in order to identify and treat potentially influential errors;
(e) Use of minimum change based model or nearest neighbour approach for localizing residual random erroneous values, to preserve as much respondent data as possible;
(f) Mix of different imputation methods as nearest neighbour approach or model based imputation.

The scheduling and the monitoring of all procedures and the interactive corrections will be managed by CONCERT, a Java web application.

5. An Oracle database was implemented to test the methodologies and to manage collected data in a more efficient way to run the procedures.

B. Data editing and data capturing

6. In order to prevent and correct fatal errors and missing values, a subset of editing rules has been integrated in the data capture system. Due to the large number of items collected (approximately 650), in this stage, only the subset of variables for provisional figures will be checked more accurately (about 230 variables). Data collection is based on a multi-channel data capture technique, a mix of the traditional face-to-face interview (paper or computer based) and the option to fill in the questionnaire via web.

7. To limit the respondent burden during data capturing, in the data collection system, a subset of 220 checking rules has been implemented. These edits refer to the expected relations between variables, or aim at reducing missing or invalid responses, especially for the items concerning farms identification and localization.

8. In addition to the questionnaire editing, another automatic check has been introduced before the final release of data, to identify potential errors slipped during data gathering. For this purpose, two kind of editing rules were considered:

(a) Fatal edit rules, which point out the presence of a fatal error and force to restore the situation of correctness of data;
(b) Query edit-rules, which underline the need for an assessment of the information, but not the obligation to modify the data.

The type and the number of rules to be applied and their characterization, fatal or query, depends on the type of data capture and on the availability to carry out further controls in the field. In most cases, the choice of such rules reflects the need to simplify data release, in order to reduce potential obstacles in filling or recording the questionnaires.

9. The E&I activities at the data collection stage have been differentiated, according to:

(a) the data collection technique;
(b) the involvement of the regional Statistical and Agricultural Offices in census operations;

(c) the entity responsible for data entry (agriculture holdings, Regions, companies in charge of paper questionnaires recording).

10. Before the end of field enumeration operations, and while data collection network is still in force, two distinct procedures will be launched to detect influential errors and outlier values. The first procedure is a micro-editing check, which underlines inconsistent data by analysing at unit level the coherence between the answers referring to related topics. Anomalous units are manually reviewed by data collection staff. Particularly, each enumerator will receive a list of units having anomalous or inconsistent data. The need of further assessment will depend on the value of the score function computed to prioritize micro data review in selective editing. In order to avoid to contact again the respondents, this step will be supported by auxiliary available information (time series, statistics or administrative sources).

11. Special procedures must be adopted for detecting outlier values for each of the main surfaces (UAA, total area, vineyard and olive plantations). These procedures are based on robust statistical methods, such as the robust technique of Forward Search (Riani and Atkinson, 2000; Riani and Atkinson, 2001). The analysis is carried out considering the farm size of the units. By assuming the hypothesis of perfect match between census and administrative data, the regression line would have vanishing intercept and unitary slope. The points being distant from this regression line are supposed to be outliers values.

12. At the beginning, data are divided in two groups. The parameters are estimated from the subset of observations supposed to have a distribution model not affected by outliers values. Then, these estimates are used to test the presence of outliers units in the remaining group. In the following steps, the observations near the fitted model are added to the first subset of units without outliers. The analysis of the differences between the parameters computed before and after the increase of observations in the considered subset will enhance anomalous units. The graphic approach implemented in this method enhances both, model inadequacy and model adjustment. A test of hypothesis is performed to verify the null hypothesis of absence of outliers values. Once the interactive correction is concluded, the remaining stages of data E&I will be carried out by performing automated methods, as described in the next paragraph.

III. The E&I System: general framework and main steps

A. General framework

13. The aim of an E&IS is to identify and treat the non sampling errors, minimizing the loss of collected information. The detection of errors is based on a set of consistent and not redundant edit rules, corresponding to logical and/or mathematical constraints. The whole set of edits contains more than 1000 checking rules. These edit rules concern single variables, or refer to the expected relations among them. The constraints must be satisfied simultaneously by the values of the statistical units. The checking rules consider both, the values of variables of the individual unit (Within-unit edit rules), and those of the same variables resulting from the aggregation of different units belonging to the same subset of analysis (Between-unit edit rules). As an instance, for each farm (unit), both the total vineyard area and the surface of each type of grape variety (subunit) are collected. So, in addition to the edits concerning the relation between the total vineyard area and the Utilised Agricultural Area (Within-unit edit rules), some checking rules refer to the surface of the single wine variety (Between subunit edit rules).

14. The large number of variables and the complexity of their constraints don’t always allow to perform one step of E&I process only. The problem is tackled by dividing the variables into subsets that are treated in different E&I steps. The E&I steps will be separately launched, considering that, if the subsets of variables are unconnected (there are no edit rules defined on variables belonging to different subsets), the order of the processing runs is not of influence. If the subsets of variables are connected (there are edit rules referring to variables of different subsets), during the performance of one of the processing runs, it is necessary to maintain fixed all the variables imputed by previous runs. It is evident that the best result is obtained when the various subsets of variables are completely unconnected.
In the case of the Agricultural census, if the choice is to handle the primary variables in the first steps and all the remaining variables in the subsequent steps, it is necessary to identify and, eventually to adjust the errors, taking into account the constraints between the primary and the secondary variables, in order to

preserve the maximum amount of collected information. The values of the primary variables are not modifiable in the subsequent steps. The inconsistencies between primary and secondary variables can be removed only by changing the values of the secondary variables. An approach based on the Graph Theory will be used to reduce the loss of information due to an improper deletion, and to improve the quality of the imputation of both primary and secondary variables. To make the description of this process easier to deal with, the main features of the E&I steps are described separately.

B. The main E&I steps

15. The following steps will be repeated in each E&I stage.

(a) Automatic error detection:

(i) Micro-editing: use of edit rules for detecting errors in collected data at individual level, according to the whole set of checking rules.

(ii) Macro-editing: the aim is to validate the information, according to the different aggregates, describing the structure of agriculture system. (e.g. the total number of farms collected, the Utilised Agricultural Area, area invested in major crops, etc..). A set of indicators computed on the main figures will enhance the need of further assessment of the E&I process. Preliminary results will be compared with estimates from the available statistical and administrative sources.

(b) Treatment of errors:

(i) Detected outliers and influential errors will be manually reviewed by experienced staff. In particular, only a subset of the records containing errors that have relevant impact on data dissemination will be clerically edited, by using score functions to prioritize micro data review in selective editing. The available auxiliary information from statistical or administrative registers will be used in these steps too.

(ii) Random errors, that are not influent, will be treated by automatic methods.

16. In the E&I System, the localization of random errors is based on two editing approaches implemented in DIESIS system (Data Imputation Editing System - Italian Software), the data driven and the theoretical minimum change, respectively named ‘first donors then fields’ and ‘first fields then donors’ algorithms.

17. The ‘first donors then fields’ algorithm first identifies a subset of potential donors and then determines the minimum number of variables to impute on these donors. The potential donors are the passed edit units (farms) which are as similar as possible to the failed edit unit. The similarity between each failed and each passed edit unit is calculated by a function defined as the weighted sum of the distances (for quantitative variables) or similarities (for qualitative variables).
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 unit, the algorithm determines the values to impute considering the donor unit being as similar as possible to the farms recipient. If possible, this algorithm imputes the variables simultaneously.

18. DIESIS system (Bruni et al., 2001; Bianchi et al., 2005) was developed in C + + language, and was used for the first time to treat the demographic variables in 2001 Population and Housing Census. In DIESIS, the detection of errors is based on an Integer Linear Programming solved using Branch and Cut methods (Manzari and Reale, 2002; Bianchi et al., 2005). The DIESIS system allows to deal both with qualitative and quantitative variables. A rigorous statistical evaluation of its performance (Bianchi et al., 2009b) has confirmed the suitability of this software for localizing errors in agricultural census data.
According to the principle of the minimum absolute change, the correctness of each wrong unit is restored by modifying the minimum number of values. The localization of errors is also made conditionally, according to the questionnaire flow (relative minimum change), or analysing the answers of those units belonging to the subset of the closest donors to wrong units. The localization stage, jointly for the qualitative and quantitative variables, will be performed by methods of Operational Research and, in particular Whole Linear Programming with efficient computational resolutions. These methods are implemented in the software DIESIS.

19. To impute erroneous values in addition to probabilistic methods, the deterministic approach is foreseen. For this purpose, a study is being conducted for identifying the most appropriate correction approach for each set of related questions. As a whole, the imputation process will be a combination of the following methodologies:

(a) deductive methods, if the values to impute are uniquely determined by the values assumed by other variables;
(b) methods based on rules like "if-then" (deterministic);
(c) methods of the nearest neighbour donor;
(d) model based techniques.

20. Another simulation study actually ongoing, has the aim to test the imputation of missing non linearly dependent data through conditional Copula functions in continuous variables. This test is being conducted by Istat, in cooperation with researchers from the University of Bologna (Bianchi et al.; 2009a). This new approach allows to preserve the variables distribution, as the missing values to impute are randomly chosen from the conditional distribution of missing values, given the observed values. This method can also be used for variables characterised by non linear dependency. In the treatment of continuous variables, for the detection of the erroneous values (localization of the variables most likely responsible for edit violations), the algorithms implemented in DIESIS will be used. Model based techniques will be preferred for the imputation of the detected continuous variables.

21. The whole process of E&I will be monitored by a set of quality indicators both, on the data distributions and on the performances of the scheduled editing steps. Standard analysis of detected errors and imputed values will contribute to measure the quality of final results.

IV. Conclusions

22. The new E&I System is one of the innovative projects launched to improve data accuracy and consistency for the ongoing census. The strategy adopted for data editing and imputation is the outcome of different simulation studies, carried out to mix the most suitable methodologies. One of the long-term goals of the E&I strategy is to set up a mix of standardised tools, procedures and methodologies, reusable for the next farm structure surveys planned in the inter-census years. Achieving this goal improves standardisation of processes between distinct surveys collecting similar information. The first results of the procedures already implemented are very encouraging and allow to trust in reducing the efforts of coping with timeliness constraints.

V. References

Bianchi G., Di Lascio F. M. L., Giannerini S., Manzari A., Reale A., Ruocco G. (2009a) *Exploring copulas for the imputation of missing nonlinearly dependent data*, Seventh Scientific Meeting of the CLAssification and Data Analysis Group of the Italian Statistical Society Università di Catania (Italy). September 9-11, 2009.

Bianchi G., Manzari A., Pezone A., Reale A., Saporito G., (2005) *New procedures for editing and imputation of demographic variables*. United Nations Statistical Commission and Economic Commission for Europe UNECE. Conference of European Statisticians .Work Session on Statistical Data Editing. 16-18 May 2005, Ottawa Canada. (Supporting paper).

Bianchi G., Manzari A., Reale A., Salvi S. (2009b) Valutazione dell'idoneità del software DIESIS all'individuazione dei valori errati in variabili quantitative. Istat - *Collana Contributi Istat* – n. 1 – 2009.

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

Luzi et al. (2007). EDIMBUS. Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys, August 2007.

Manzari A., Reale A., (2002) *Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation methodology*, Proceedings, International Association of Survey Statisticians, 634-655.

Riani M., Atkinson A. C. (2000). *Robust Diagnostic Data Analysis: Trasformations in Regression*. TECHNOMETRICS. vol. 42, pp. 384-394 ISSN: 0040-1706. With discussion.

Riani M., Atkinson A. C. (2001). A Unified Approach to Outliers, Influence, and Transformations in Discriminant Analysis. *Journal Of Computational And Graphical Statistics. vol. 10, pp. 513-544 ISSN: 1061-8600l.*