3-6 October 2022, (virtual)

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Editing

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

SIMONA ROSATI, MARIA TERESA BUGLIELLI, LAURA TOSCO
Istat | DCME

Outline

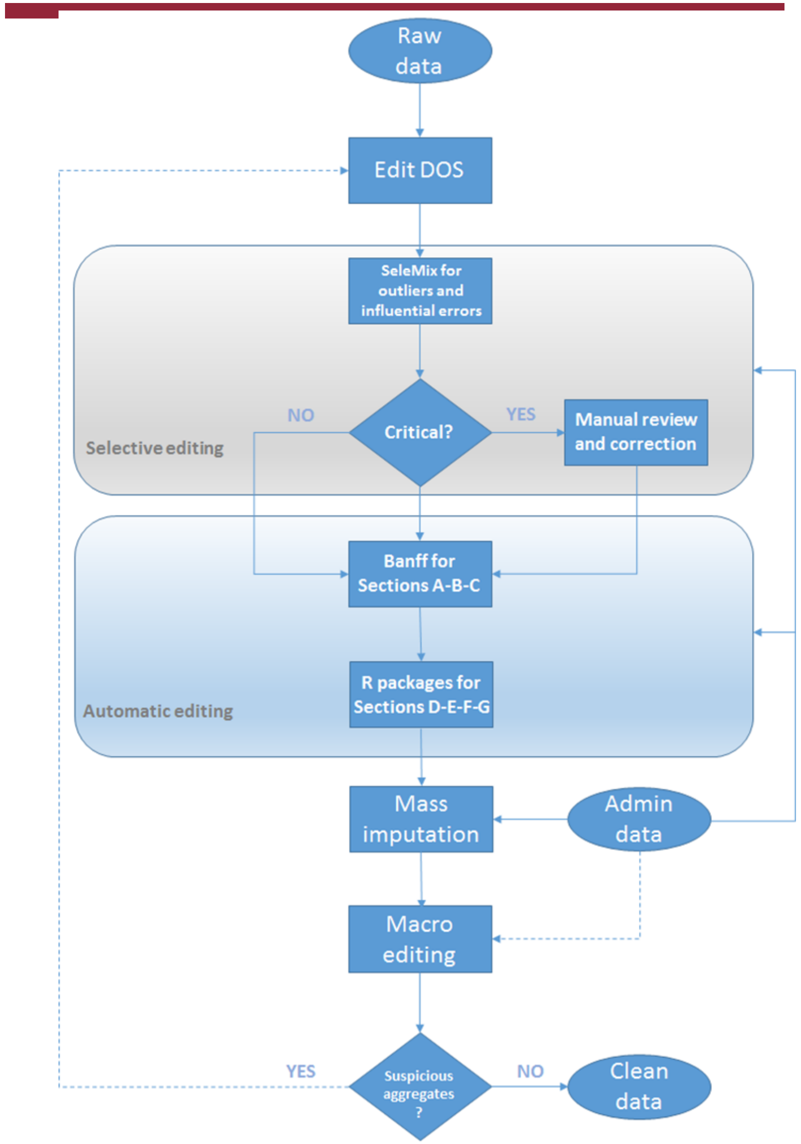
- Introduction
- The editing and imputation process
- Selective editing
- BANFF for quantitative variables
- R packages for mixed variables
- Mass imputation
- Conclusions

Introduction

- Questionnaire of the Agricultural Census 2020 divided into seven parts
  - Sections A-B-C: information on general characteristics of the agricultural holding, land use, size of livestock holdings, and manure management system; these sections included mainly **quantitative variables**
  - Sections D: information on the farm manager and on the other gainful activities (OGA) directly related to the farm
  - Section E: information on human resources employed by the agricultural holding
  - Section F: information, such as revenue, marketing, innovation, computerization, and others
  - Section G: information to evaluate the economic impact of the Covid-19 epidemic on the farms

- **Major aspects**
  - Very complex data set with a large number of qualitative and quantitative variables
  - Joint treatment of both qualitative and quantitative variables
  - integrated use of different methods and software tools

The edit and imputation (E&I) process



- E&I process completely redesigned: use of modern statistical E&I methodology consistent with internationally recognised principles and standards
- Use of auxiliary data sources as administrative data and registry data
  - at macro level for the validation of the census data
  - at micro level for comparisons between detected errors and auxiliary data when the census micro data were inconsistent, as well as to integrate missing data

Selective editing

- Detection of outliers and influential errors with the R package **SeleMix:** methodology based on contamination models

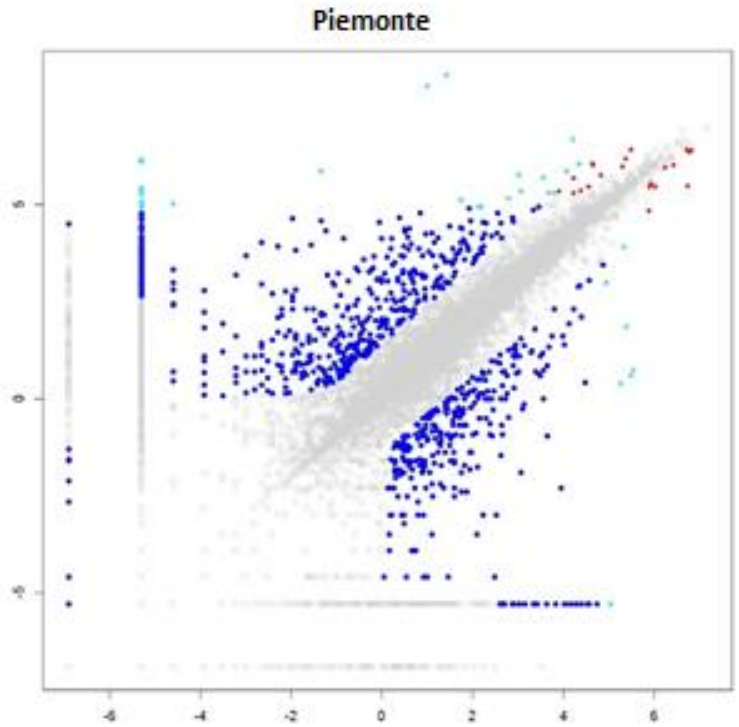
- Outliers and influential units (critical units) are manually reviewed by subject matter experts

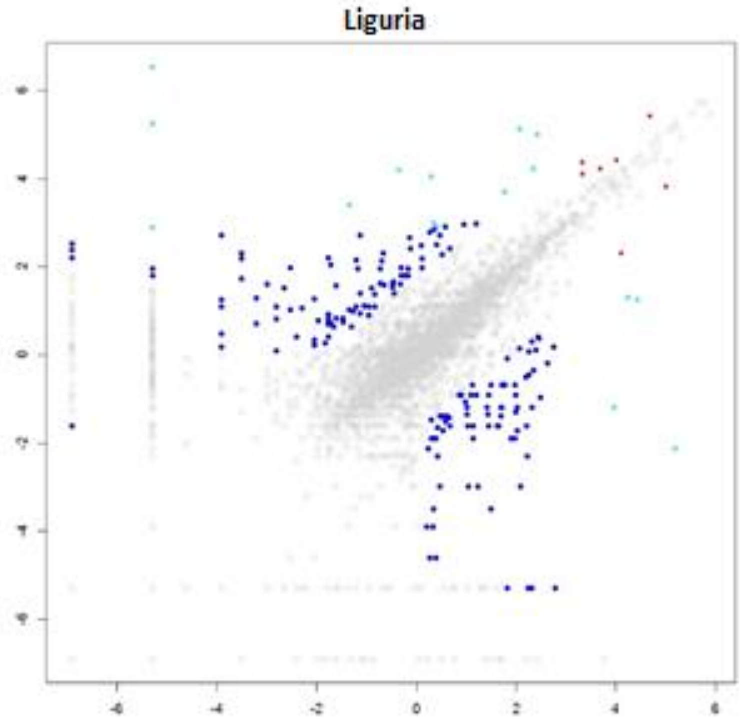
- To make this phase more efficient, selective editing started in advance during data collection phase (January 2021-July 2021).

    - Outliers and influential errors were detected at three different times

        1. in May
        2. in June
        3. at the end of data collection to identify residual outliers and influential errors

Outliers and influential errors of UAA for Piemonte and Liguria


Piemonte


Liguria

BANFF for quantitative variables 1/2

BANFF (SAS) procedures used

- **Proc Verifyedits:** allows the specification and analysis of the edit rules. A group of edits is checked for consistency, redundancy, determinacy and hidden equalities.
- **Proc Editstats:** applies a group of edits to a SAS dataset and determines if each observation passes, misses or fails each edit.
- **Proc Errorloc:** identifies the fields which must be changed in each individual record in error so that the record can be made to pass all the edits.
- **Proc Deterministic:** analyses each field previously identified as requiring imputation to determine if there is only one possible value which would satisfy the original edits
- **Proc DonorImputation:** uses a nearest neighbour approach to find for each record requiring imputation the valid record that is most similar to it.

BANFF for quantitative variables 2/2

Pros and cons of BANFF:

- ⊕ User-friendly and flexible
- ⊕ Minimizes the number of fields to change
- ⊕ Ensures that erroneous records are imputed to satisfy all the edits
- ⊖ Not for qualitative variables
- ⊖ Not for systematic errors
- ⊖ Edits must be linear equalities or inequalities
- ⊖ Imputation may be unsuccessful because no suitable donor is available
- ⊖ Possible post adjustments

R packages for mixed variables

Packages used:

- **validate**: provides functions to formulate validation rules written as positive logical formulas, to confront data and analyze or visualize the results
- **validatetools**: a set of functions for finding redundancies or contradictions between the rules formulated with validate
- **errorlocate**: implements functions to localize records violating defined validation rules; erroneous values are replacing with missing values to be imputed
- **VIM**: provides the functions to impute missing values as kNN and hotdeck

Pros and cons of R:

⊕ High flexibility in the design and realization of the E&I process

⊕ Simple handling of large and complex amounts of data

⊕ Joint treatment of quantitative and qualitative variables

⊖ Error localization can be time-consuming

⊖ Often it is necessary to guide the process to converge a ‘global’ solution

Mass imputation

- Imputing of records with a high number of missing values and unit nonresponses which was estimated eligible
- Carried out using Banff for sections A-B-C and R for the remaining sections of the questionnaire
- Units to be imputed were integrated with information from **administrative data sources** where possible
- Unit nonresponses without signals from administrative data were discarded
- Donor records were chosen among original (raw) records that passed all the edits

BANFF and R were both fast and simple

Conclusions

- Both BANFF and R ensure the internal consistency of records with respect to the set of the specified edit rules
- R for E&I was introduced for the first time in such a complex survey data
- R allows the joint treatment of qualitative and quantitative variables in a simple way
- Action: promote the use of R in surveys whose E&I process needs to be redesigned or modernised

thank you

SIMONA ROSATI sirosati@istat.it

MARIA TERESA BUGLIELLI bugliell@istat.it

LAURA TOSCO tosco@istat.it

Istat Istituto Nazionale di Statistica