4Th International Conference
New Challenges for Statistical Software - The Use of R in Official Statistics
Bucharest, 7-8 April 2016

DATA EDITING FOR COMPLEX SURVEYS IN PRESENCE OF ADMINISTRATIVE DATA: AN APPLICATION TO FSS 2013 LIVESTOCK SURVEY DATA BASED ON THE JOINT SEQUENTIAL USE OF DIFFERENT PACKAGES

Elena Catanese *(catanese@istat.it)

*Italian National Institute of Statistics

Introduction

- Data editing and imputation (E&I) in complex sample business surveys is a task which is usually split into two steps:

1. **selective editing** techniques applied to the primary target estimates

->to identify a potential set of influential errors (usually interactive editing)

2. **automatic identification and imputation** of inconsistencies and missing values.

- Within this framework, the present paper reviews the Italian data editing strategy adopted applied to 2013 Farm Structure Survey livestock data.

- In this edition this process has been entirely carried out in the R environment

Selective Editing: General Overview

- Selective Editing is the art of finding «*potential*» **influential errors** in sample survey data.

- It is often adopted in business surveys

- Influential errors: errors having the highest impact on target estimates

- The definition of outlier is widely abused in statistics, though, it may in some cases overlap with the concept of influential error

- In practice units are prioritized according to a *score function* based both on a :
  1. *risk component* which evaluates the probability of being an error
  2. *influence component* which evaluates the impact on target estimates

The R package *Selemix : the model*

- Basic assumptions:

  Observed data is a mixture of two Gaussians distributions

  1. One representing «true data»
  2. One representing the error mechanism, which is assumed to be a bernoullian process with parameter π

- True data are modelled through a normal or log-normal distribution, resulting from a standard multivariate regression model

- The error follows an additive mechanism represented by a Gaussian r.v. with mean 0 and covariance proportional to the one of the regression model

- This distribution can be estimated by maximizing the likelihood based on *n* sample units via an ECM algorithm

The R package *Selemix 2: the functions*

It consists of three main functions:

1. Ml.est : model estimation

where all the parameters have to be estimated(the regressione man and variance, while the probability of the bernoullian processs and the proportional variance of the error may be fixed or not).
It also computes for each observation the probability of being an outlier

2. Pred.y : model prediction, which essentially provides you

the linear model predicted values.
Useful to proceed with automatic corrections

3. Sel.edit : where fixed the accuracy level and given a set of weights the units are ranked and selected so that the user can proceed with *interactive editing.*

It has two terms the risk component(probability of being an outlier), influence component (weights).

Automatic imputations methods for compositional data

- The encountered errors dealing with Livestock data editing in this phase were:

  Inconsistencies of compositional data, to be split in:

1. inconsistencies in presence of missing items and 0s
2. minor inconsistencies in presence of non0s (ie. Deltadifference usually lower)

- Two different packages were tested and used:

  rspa and *RobComposition*

- No donor techniques were applied because of the presence of administrative data

The R package *RobComposition*

- Robust Estimation for Compositional Data, offers many robust methods for imputation and for analysis of **compositional data** (sum costraint)
- Particularly suitable when dealing with missing values and 0s
- Methods based on the k-nearest neighbour by means of **Aitchinson distance**, that takes into account ratios (of compositions) and exploits *ln (log-ratios)* like distribution similarity (dissimilarity) measures
- It should preserve distributions
- Main functions (at least for the needs of this work)

1. *impKNNa* → only K.nearest neighbour and missing values
2. *impCoda* → other methods and also 0s are allowed

The R package *rspa*

- The rspa (record successive projection algorithm) package applies a minimal adjusting to numerical variables such that the end result obeys a predefined set of linear equations (or inequalities), i.e. it minimizes the distance between the initial and final vector
- In this case the norm is induced by a diagonal positive matrix (the higher the weights, the stronger is the inertia)
- The set of costraints are defined through the **editrules** package, which is automatically inherited in *rspa*

- funtions:

editmatrix-> to define constraints

adjustRecords -> to impute with minimal adjustment

Data Editing of livestock in 2013 survey: Selective Editing 1

- **Selective editing model prediction part** needs **auxiliary variables**

- For these purposes 2010 Census data, and administrative data available from 2013 Livestock Register (LR) were used
- A previous work of record linkage had to be carried out to use Livestock Register

SeleMix output applied using as auxiliary variable: ***A livestock Register, B Census 2010 values***

Respondents	Influential Obeservation _A	Influential Obeservation_ B	% Incidence_ A	% Incidence_ B
Cattle				
9111	**8**	**494**	**0.1%**	**5.4%**
Sheep				
4693	**191**	**488**	**4.1%**	**10.4%**
Goats				
1883	**128**	**314**	**6.8%**	**16.7%**
Pigs				
2698	**249**	**349**	**9.2%**	**12.9%**

- When information is outdated, ie. auxiliary information not so correlated, SeleMix doesn't apply properly
- The variability of results for livestock categories with LR has to be linked with the *quality of the registers*, which are differently regulated

Data Editing of livestock in 2013 survey: non influential errors

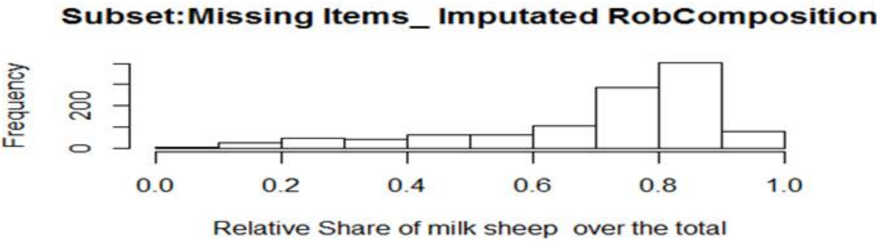
- Almost no action has been performed for cattle
- Analysis performed on the output of the selective editing and on the patterns of the "outliers" (not equivalent, the sel.editing uses *weights*) a predominant presence of influential errors at Nuts-level gave evidence of:

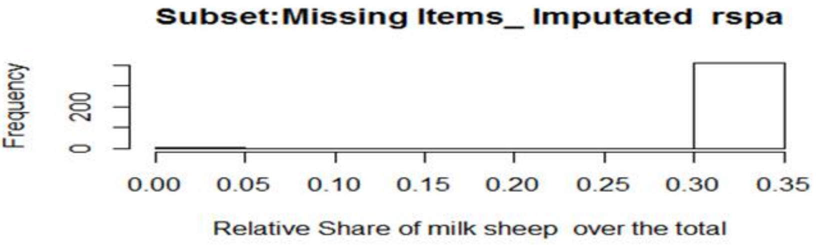
1. Record linkage problems for Pigs -> total imputation rate turned to be 3.3% instead of 9.2%
2. Patterns of missing items for sheep and goats

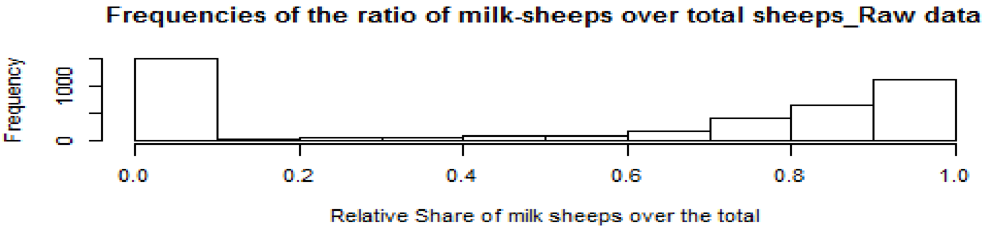
(probably **enumerator effect**, note cattle and sheep are present in different NUTS2 regions)

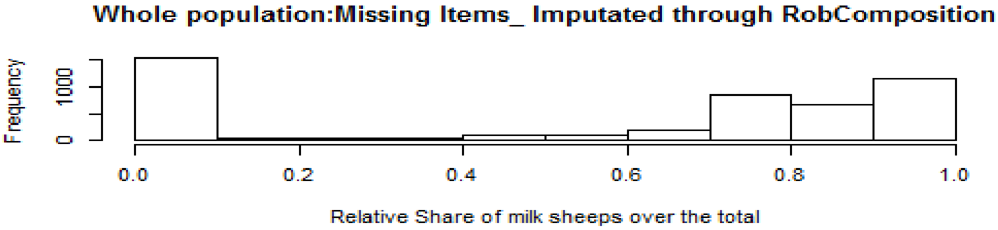
- Thus since the Register records only total animals and not their subcategories -> this yield to the problem of imputation of compositional data in presence of many missing items and or 0s.
- rspa vs RobCompostion has been checked

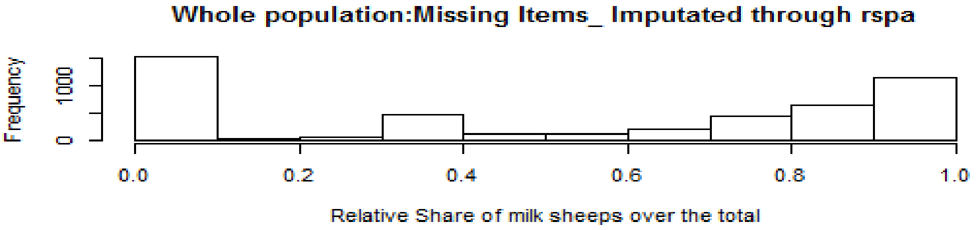
Data Editing of livestock in 2013 survey: non influential errors


Subset:Missing Items_ Imputated RobComposition


Subset:Missing Items_ Imputated rspa


Frequencies of the ratio of milk-sheeps over total sheeps_Raw data


Whole population:Missing Items_ Imputated through RobComposition


Whole population:Missing Items_ Imputated through rspa

- It has been decided to split the imputation part into two steps:

  first RobComposition

  then Rspa

Conclusions

- The present work reviews only a small proportion of the E&I phase of the 2013 FSS survey concerning livestock where for the first time administrative Register were used for two purposes: check and impute data; assess the quality of the information of the Registers
- Selective editing techniques were fruitful to :

1 verify the quality of the Registers

2 reduce the impact of manual checking and re-contact techniques,

3 detect systematic errors, such as record linkage problems for pigs, and to recognize patterns of missing items for sheep (9.9) and goats (10.3%) (probably linkable to enumerator work) thus introducing under-estimation.

- According to the overall pattern of errors, some R packages may be more suitable than others.
- FSS 2013 livestock experience should be useful for those that must follow a E&I phase to achieve a good performance in terms of efficiency , as well as a unified massive treatment of primary (totals) and secondary (linked to the primary by constraints) variables without the use of deterministic rules for imputation methods

References

Memobust, Handbook on Methodology of Modern Business Statistics 2014, www.cros-portal.eu

Aitchison J(1982) The statistical analysis of compositional data with discussion. J. Royal.Stat. Soc., Series B Statistical Methodology) 44 2): 139±17

De Waal T.,Pannekoek J, Scholtus S.(2011) Handbook of Statistical data Editing J.Wiley&Sons

Hron, K. and Templ, M. and Filzmoser, P. (2010) Imputation of missing values for compositional data using classical and robust methods Computational Statistics and Data Analysis, vol 54 (12),

pages 3095-3107

Cran Package: https://cran.r-project.org/web/packages/robCompositions/robCompositions.pdf

Buglielli, M.T., Di Zio, M., Guarnera, U. (2010), *Use of Contamination Models for Selective Editing*, European Conference on Quality in Survey Statistics Q2010, Helsinki, 4-6 May 2010

Kozak R.(2005) The BANFF system for automated editing and imputation. Proceedings of SSC Annual meeting. June 2005 proceeding of the survey methods Section

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

Cran Package: http://cran.r-project.org/web/packages/SeleMix/index.html

http://cran.r-project.org/web/packages/editrules/index.html

https://cran.r-project.org/web/packages/rspa/index.html

VIM package: https://cran.r-project.org/web/packages/VIM/index.html

M. Templ, A. Alfons, P. Filzmoser (2012) Exploring incomplete data using visualization tools.

Journal of Advances in Data Analysis and Classification**H**ildreth,C.(1957) A quadratic programming procedure Naval Research Logistics Quarterly V. 4, Issue 1, pages 79–85

StatMatch package: https://cran.r-project.org/web/packages/StatMatch/index.html

D’Orazio M., Di Zio M., Scanu M. (2006) Statistical Matching, Theory and Practice. Wiley, Chichester

Latouche M., Berthelot J.M. (1992), Use of Score Functions to Prioritise and Limit Recontacts in

Editing Business Surveys, *Journal of Official Statistics*, 8, 3, Part II.

Lawrence, D., and McKenzie, R. (2000), The General Application of Significance Editing. *Journalof Official Statistics*, **16**, 243-253.

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

Recommended practices for Editing and Imputation in cross-sectional Business surveys (O.Luzi et al 2007)

Kovar, et al., 1988, Overview and Strategy for the Generalized Edit and Imputation System. Statistics Canada, Methodology Branch

Bankier, M. (2011), “Imputing Numeric and Qualitative Variables Simultaneously”, A Technical Report Detailing the Methodology of CANCEIS, Internal report,Statistics Canada

Bankier, (2000). Canadian Census Minimum change Donor imputation methodology. Technical Report.

M.A. Hidiroglou and J.M. Berthelot. Statistical editing and imputation for periodic business surveys. Survey Methodology, 12(1):73–83, June 1986. Statistics Canada

Andridge, R.R. and Little, R.J.A. (2010) A Review of Hot Deck Imputation for Survey Non-response.International Statistical Review.78, 40–64.

R package: https://cran.r-project.org/web/packages/HotDeckImputation/index.html

Bankhofer, U. and Joenssen, D.W. (2014) On Limiting Donor Usage for Imputation of Missing Datavia Hot Deck Methods. In: M. Spiliopoulou, L. Schmidt-Thieme, and R. Jannings (Eds.):

Data Analysis, Machine Learning and Knowledge Discovery. Studies in Classification, Data Analysisand Knowledge Organization, 3–11. Berlin/Heidelberg: Springer.