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 R Packages

Elena CATANESE (catanese@istat.it),
Italian National Institute of Statistics (Istat)

ABSTRACT

Data editing and imputation (E&I) in complex sample business surveys is a task which is usually split into two steps to gain efficiency in terms of time and human resources: first selective editing techniques are applied to the primary target estimates variables in order to identify a potential set of influential errors that require usually manual editing and a second part of automatic identification and imputation of inconsistencies and missing values.
Within this framework, the present paper reviews the Italian top-down data editing strategy adopted and automated imputation showing the experience applied to 2013 Farm Structure Survey livestock data.

In this edition this process has been entirely carried out in the R environment by means of different R packages.

Keywords: *Selective editing; data editing; Business Surveys, Automated Edit Rules, Imputation of Missing Values, Compositional Data, Random vs. Systematic Errors, Influential non Influential Errors, Statistical software R*

JEL Classification: *C10, C88*

1. INTRODUCTION

Data editing and imputation (E&I) is still one of the most resource-consuming activities of National Statistical Institutes; for business surveys it is generally reckoned to account for up to 40% of survey cost (De Waal

et al 2011). This may account even much more especially in presence of complex surveys where many numeric variables are collected and where many coherence rules between variables must be fulfilled in order to provide internally consistent records.

In particular data may be affected by various typologies of errors such as missing items, logical inconsistencies, mistakes of unit of measure. Moreover in case of data obtained from a survey based on a probabilistic sample, even small errors for units with a high survey weight may strongly affect the quality of final survey estimates. For this reason the terminology of outlier is often inflated.

The general framework for E&I adopted by the official statistics (Luzi at el 2007, Memobust 2014) for continuous variables in business surveys consists in applying first a "selective editing " procedure in which only "outliers", here meaning a small selection of records with potentially the most influential errors are manually checked and eventually edited by domain experts, then the residual or uninfluential errors are usually corrected in an automated way, even if in some cases this step may be ignored without any impact on the quality of final estimates.

The advantage of this combined approach is on the one hand to minimize time and resources available and on the other hand to reduce the various inconsistencies which may arise from a non-parsimonious outliers imputation.

There are many methods and methodologies available in literature which may be used to define the data editing strategy for both the selective editing approach and the automated one.

Generally speaking to detect potentially influential errors , all selective editing techniques, require both a "*prediction*" of the true value (or an anticipated value) and a *score function* (Lawrence, D., and McKenzie, R. 2000). Each score function is built by taking into account two important components: *influence* and *risk*. A cut off value is then set to decide when a record should be treated. That is, a given record is suspicious if the value of a score function exceeds a certain value, then the imputation part may be done by many alternative methods such as interactive treatment, but also model-based. Within this selective editing framework Istat has developed an R-package, SeleMix (Di Zio, M., Guarnera, U. 2013), which also offers an imputation step by computing a predictive value. It is worth to notice that an alternative to selective editing may be constituted by robust estimation techniques.

The second part of the E&I consists in the treatment of the remaining non influential errors, such as missing items or logical inconsistencies. A variety of automated methods and software exist both for the localization and

imputation part. The localization may be deterministic, or rule-based on the Fellegi-Holt (F-H) paradigm (Fellegi, I. P. and Holt, D. 1976), or data-driven. Among the software that use the FH method there is BANFF (Kozak R.2005) Sas based developed by Statistics Canada and Edit-rules (de Jonge, Van der Loo 2013) which is an R package, and CANCEIS (Bankier 2000,2011) also developed from Statistics Canada, which follows the data driven approach. The methods for automatic data imputation, which are also implemented in the above mentioned software and package, may be based on explicit models such as mean value, regression, etc., or on implicit models such as cold-deck, hot-deck, or on combined methods such as predictive mean matching. There are many available R packages to perform automated editing according to the different approaches such as MICE (multivariate imputation by chained equations), for predictive mean matching, rspa for minimal numeric record adjustment under edit restriction (Van der Loo 2014), among the most popular for Hot-Deck imputation (Andridge, Little 2010) there are StatMatch (M.D'Orazio 2006) which implements Nearest Neighbor imputation with different metrics, VIM (Templ, A. et al. 2012), HotDeckImputation (Joenssen et al. 2014) , or RobComposition (Hron, K at al 2010) which specifically focuses on missing values for compositional data but also many other R packages for either deterministic corrections as well as hot-deck imputations or based on regression models, that for the sake of simplicity are not cited nor described here.

The present work focuses on the editing and imputation strategy adopted for the 2013 Italian Farm Structure Survey (FSS) which is a complex survey on agriculture holdings, foreseen by the EU Regulations, where agricultural utilized area and crops, livestock, labor-forces, irrigation, other gainful activities are surveyed. Indeed this paper discusses the data editing flow applied to livestock, tackling both the selective editing phase which was carried on using different types of auxiliary data (Census and administrative data) and the automatic treatment and localization for the remaining errors such as missing items and logical inconsistencies. It will be shown how some of the above mentioned R packages can be combined into a production system that improves data quality in several consecutive editing steps, bearing in mind which have been the best tools and their pros and cons under the specific considered real editing phase.

The rest of the paper is structured as follows: Section 2 provides an overview of the functionalities of the R-packages that have been used in the specific E&I phase; Section 3 focuses to the case study, i.e. to the editing of livestock data of the 2013 Farm Structure Survey , moreover the general architecture of the data editing flow is provided as well as main the sources of errors which were encountered and how it has been decided to treat them

after some preliminary analysis, results are discussed and some quality indicators that synthetize the impact of the FSS2013 livestock editing phase are provided; Section 4 summarizes results and discuss how these methods could be usefully applied to improve quality and timeliness of the editing phases of continuous variables in complex surveys.

2. AN OVERVIEW OF THE R-BASED PACKAGES UTILIZED

In this section a brief overview of the R-packages which have been utilized in the editing phase of 2013 FSS livestock data, is provided. Some details of some functions implemented within these packages are given, in order to let the reader understand the terminology and the results discussed in the next section. For further details refer to the CRAN site where the complete documentation is available.

SELEMIX

Sele = Selective Editing, Mix = by aim of a mixture model, is a package for the detection of outliers and influential errors using a latent variable model

There are three main functions provided with the package:

Ml.est : model estimation, where the parameters of the regression model are estimated and that performs outlier detection by means of posterior probabilities

Pred.y : model prediction, which essentially provides you the linear model predicted values. This may be useful if you do not intend to proceed with interactive corrections, but automatic ones

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

It is important to underline that this package is useful in both detecting influent observations as well as for imputation purposes, but it needs an auxiliary variable that must be linked at micro level for the model estimation function, while for the selective editing part the use of auxiliary information in terms of macro aggregate is allowed.

- **Ml.est**

The basic assumption of the underlying parametric model is that observed data is a mixture of two Gaussians distributions, where one represents «true data» (aka error free) and the other one represents the error mechanism, which is assumed to be a bernoullian process

True data are modelled through a normal or log-normal distribution

$$y_i^* \sim N(\mu_i, \Sigma) \quad u_i = N(0, \Sigma) \tag{1}$$

resulting from a standard multivariate regression model:

$$Y^* = XB + U \tag{2}$$

Where X represents the covariates, B is coefficients, U are the normal residuals

Observed data is error free or erroneous according to *n* independent realizations of a *bernoulli* r.v. with parameter π.

The error follows an additive mechanism represented by a Gaussian r.v. with mean 0 and covariance proportional $\Sigma_\epsilon \propto \Sigma$ i.e. $\Sigma_\epsilon = (\alpha - 1)\Sigma$ where $\alpha > 1$ α is the variance inflation factor,

Thus:

$$y_i \sim (1-\pi) N(y_i, \mu_i, \Sigma) + \pi N(y_i, \mu_i, \alpha\Sigma) \tag{3}$$

(where $\mu_i = \boldsymbol{B}^T x_i$)

Observed data represent the mixture of the 2 regression models, having the same coefficient matrix B but different and proportional residual variance-covariance matrix

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

#Log_linear case, the covariance matrix of error data is 3 (α) times the other one, $\pi = w$ #erroneous data represent 5% of observed ones. The two basic parameters lambda, w may be fixed #or not.
sel<-ml.est(FSS$var1,x=CENSUS$var1,model="LN",lambda=3,w=0.05,lambda.fix=FALSE, w.fix=FALSE,graph=TRUE)

 # linear case, now after few trials by letting w vary it seemed 10% was an acceptable range value, #so it was decided to let it fix. No use of a priori auxiliary information
sel2<-ml.est(FSS[,2:3],x=NULL,model="N",lambda=3,w=0.10,lambda.fix=FALSE, w.fix=TRUE,graph=TRUE)

Chosen the covariates X the model always estimates B and ∑, while the user may decide whether or not to estimate π, the mixing proportion of the contamination model (a priori probability of being erroneous) whose default value is 0.05 and α whose is 3.

It is important to check final estimates of both α, π, some values may be a hint that the linear model has been wrongly defined i.e. $\pi > 10\%$ is usually unacceptable).

Outlier analysis can be performed based on the vector of posterior probabilities (p.p.): units with p.p> t (a *user defined* parameter) are flagged as outlier. Default value for t is 0.5

 # this attribute provides the user the set of outliers either 1 or 0.
sel2$outlier

- **Pred.y**

Predictions are obtained from the fitted distribution of the true data conditional on the observed data

$$f(y_i^* | y_i) = \tau(y_i)\delta(y_i^* - y_i) + (1 - \tau(y_i))N(y_i^*, \mu_i^*, \Sigma_i^*) \quad (4)$$

where $= \mu_i^* = (y_i + (\alpha - 1)\mu_i) / \alpha$ and $\Sigma_i^* = (1 - 1/\alpha)\Sigma$

The conditional expected value $E(y_i^* | y_i)$ is $= \tau(y_i)\, y_i + (1 - \tau(y_i))\mu_i^*$

 # posterior probability for identifying outliers t.outl=0.6
ypred<pred.y(y=FSS[,2:3],x=NULL,B=sel2$B,sigma=sel2$sigma,lambda=sel2$lambda,w=0.10,model="N",t.outl=0.6)

- **Sel.edit**

The expected error is the score function and is defined as:

$$y_i - E(y_i^* | y_i) = (1 - \tau(y_i))(y_i - \mu_i^*) \quad (5)$$

The first term refers to the risk component $(1 - \tau(y_i))$, while the second one $(y_i - \mu_i^*)$ to the influence component. As usual as in selective editing this is a mix of both components.

In practice in SeleMix the user must define a threshold value η.

Suppose the target aggregate to estimate is the total of the variable, standard HT estimator, then the relative individual error for the unit *i* with respect to the variable Y_j is defined as the ratio between the weighted expected error and estimate of the target parameter $\hat{T_j}$.

$$r_{ij} = \frac{\omega_i (y_{ij} - E(y_{ij}^* | y_{ij}))}{\hat{T_j}} \quad (6)$$

$|r_{ij}|$ defines the *local score function*, and the global score function per unit *i* is

$$GS(i) = \max j\, |r_{ij}| \quad (7)$$

In practice the user defines a threshold / accuracy level η so that, observations are ordered decreasingly according to *GS*, then the first k units are selected so that all the left ones have a residual error below η

 # here t.*sel=0.01is the accuracy threshold values, I am using as total one provided from an external #macro source. I use the predicted values as true values*

sel1<-sel.edit(y=FSS$var1,ypred=ypred$FSS_var1,wgt=FSS$wgt,t.sel=0.01,tot=26600392)

 # I assume the true values are represented by an auxiliary variable

sel1<-sel.edit(y=FSS$var1,ypred= CENSUS$var1,wgt=FSS$wgt, t.sel=0.01)

ROB-COMPOSITION

This package, Robust Estimation for Compositional Data, offers many robust methods for imputation and for analysis of compositional data, i.e. data that are characterized by the sum constraint, a common situation in official statistics as well as in other sciences.

Among the robust analysis functions PCA, Factor Analysis, Discriminant analysis (Fischer rule), Anderson-Darling normality tests are provided.

The imputation functions deal with missing values and rounded zeros for compositional data using both classical and robust methods. To impute missing values there are two functions available, *impKNNa* and *impCoda*. These are both based on the log-ratio approach to the statistical analysis of compositional data proposed by Aitchison (1982), who introduced a metric, the Aitchinson distance, that takes into account ratios (of compositions) and exploits *ln* like distribution similarity (dissimilarity) distances

$$d_A = \sqrt{\frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D}\left(ln\frac{x_i}{x_j} - ln\frac{y_i}{y_j}\right)^2} \quad (8)$$

Where x=(x_1,…, x_D)represents a composition.

The first function *impKNNa* imputes missing values using a k-nearest neighbor procedure based on the Aitchinson distance. This method allows not to introduce any bias in the relative proportion of partials and to preserve their multivariate data structure.

The second function *impCoda* one initializes missing values with proper values then different iterative methods for the imputation of missing values such as least squares regression or least trimmed squares regression can be chosen in order to find better estimates for the former initial estimates.

Finally imputed values are perturbed in the direction of the predictor by values drawn from a normal distribution related to the corresponding residuals. By default values are initialized by the k-nearest neighbor method. Also 0's may be imputed by using this function.

```
# imputed values are in the object $xImp, only NA are imputed
Dat1 <- impKNNa(dat)$xImp
# This also imputes 0s. In some cases for quantitative values these should not
be imputed
Dat1 <- impCoda(dat)
Dat1$xImp
```

RSPA AND EDITRULES

The rspa (record successive projection algorithm) package implements methodology a to apply a minimal adjusting to numerical variables such that the end result obeys a predefined set of linear equations (or inequalities), which again may useful in the case of compositional data.

The package uses the successive projection algorithm of Hildreth (1957) so that give a vector x_0 where $Ax \leq B$ the substitution value is found by minimizing the usual norm induced by the definite positive symmetric matrix *W*, which in the package can only be diagonal (positive) :

$$(x_0 - x_1)^T W (x_0 - x_1) \tag{9}$$

The meaning of W is that it allows to choose among the components which are more likely to be changed (high weight implies that the variable will be changed less than variables with low weight).

The main function is *adjust*, a generic function solving a quadratic optimization problem implemented in C. It allows several definitions of the constraint for vectors *AdjustRecords* is an equivalent extension defined for data frames. It is worth to notice that adjust does not perform any consistency check. Those must be defined by using the editmatrix function which is contained in the editrules package. This package is specifically implemented to define rules and consistency checks as well as feasibility checks, such as for instance to remove redundant consistency rules (isFeasible function) . The error localization is based on the Fellegi-Holt paradigm, thus providing the minimal set of variables to be adjusted in order to let the entire record to obey all the rules.

```
#allows equality and for istance if a value must be positive
E1<editrules::editmatrix(expression(OVI1+OVI2+OVI3==TOTOVI,
OVI1>=0,
OVI2>=0,
OVI3>=0))
#set which values are allowed to change TRUE and which are fixed FALSE
aa<-array(dim=c(nrow(dat1),ncol(dat1)))
aa1[,1:3]<-TRUE
aa1[,4]<-FALSE
dat1<- data.frame(dat1)
cap<- rspa::adjustRecords(E1,dat1,adjust=aa1)
ovi1<-cap$adjusted
#check if it works
dat1[,5]<-dat1[,4]-dat1[,3]-dat1[,2]-dat1[,1]
```

3. DATA EDITING OF LIVESTOCK OBSERVED IN 2013 FSS SURVEY

The Farm structure surveys (FSS) are EU regulated and provide harmonized data on agricultural holdings in the EU, including: number of holdings, land use and area (main crops), livestock, farm labor force (including age, gender and relationship to the holder), type of activity, economic size of the holdings, other gainful activities on the farm, system of farming, machinery, organic farming. They are run at regular intervals, every two or three years. Every 10 years they are carried out in the form of Agricultural Census, while in intermediate periods they are sample based. FSS data is also used in other policy areas such as environment, regional development and climate change. Italian 2013 FSS edition sampled 44168 holdings from the 1620081 holdings of the reference frame list, which was the 2010 Census list of active farms. The survey was carried by enumerators supposed to make face to face interviews to holders and using CAWI techniques to report and send questionnaires to Istat.

All EU-National Institutes of Statistics(NIS) must provide complete records of FSS to Eurostat. This means that therefore in-record inconsistencies must be removed and that validated data can be used for direct analysis, on the other hand this implies that much of the work must be devoted to check data. For this reason it must be mentioned that a wide variety of methods for automatic imputations (R and SAS based) have been adopted according to the thematic area. Broadly speaking the questionnaire could be divided into three main thematic areas, that have internally many consistency checks: crops, livestock and labor.

In practice in the 2013 edition the selecting editing phase has been carried out through SeleMix, that has been applied to the major continuous variables such as utilized agricultural area, total area, irrigated and irrigable area, totals of the major livestock characteristics such as cattle, pigs, sheep and goats.

Most of the variables under investigation in 2013 could all be linked at record level with the previous 2010 Census values, except in some cases in presence of new holdings arising from splits or merging. For this reason initially all the main aggregates in terms of crops and livestock categories were checked through SeleMix using Census data as auxiliary variable. In addition during the collection phase an additional administrative source became available to Istat, the national livestock registers, available for the following species: 1) cattle and buffalos, 2) sheep and goats, 3) pigs. Cattle and buffalos register was the first one to be set up in 2005, it was recognized as official and obligatory by the European Commission in 13/2/2006 and it is strictly regulated. Each animal needs to be electronically identified within 20 days (birth or bought) by the holder, and owns an ID. For this reason the Register is very updated and it provides for each holder very detailed information on the animals, grouped by sex, age and typology. Since 1/1/2010 also sheep and goats animals need to be electronically and individually identified, but a major delay in individual registration is allowed. The Sheep and Goat register provides for each holder the total amount of animals, without any additional information. Concerning Pigs Register the holder is not obliged to electronically identify all individuals, and there all less strict law requirements to be fulfilled.

A preliminary work of record linkage had been carried out to link the holder identification of the Register with the Statistical codes of the holding used at Istat, in order to apply the SeleMix package.

Applying SeleMix using Census data as auxiliary variable for the regression model and by letting it be the "true" value in the selective part, turned out to give satisfactory results for total utilized areas, and agricultural utilized area. Unfortunately this did not hold true when applied to livestock. In the former case the incidence of influential observations remained below 5%, while for livestock they ranged from 5,4% for cattle to 16,7% for goats (see Table 1), thus in contrast with the estimation model where the probability of the error was hypothesized to be 5% . In this case 2010 Census information concerning livestock turned out to be too outdated, thus giving poor results when Selective Editing techniques were applied, moreover the incidence of influential information remained always over 10% for all species except cattle (see Table 1).

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

Table 1

Respondents	Influential Obeservations_A	Influential Obeservations_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%**

For this reason it has been decided to carry out the same analysis by using the 2013 data available from the livestock registers which were more updated and were showing a higher correlation with surveyed data. Indeed by applying SeleMix under the same model hypothesis (i.e. parameters such as lambda, inflation factor, posterior probabilities of being outlier and accuracy level, assuming in both cases that Census 2010 and Livestock Register represent true data), the method provided an effective gain in efficiency for all livestock species when FSS data was compared to Livestock Register (see table 1). The comparison with the cattle register gave impressive results, only 8 influential observations out of 9111, the incidence for goats and sheep lowered of more than a half, from 10.4% to 4.1% for sheep, and from 16.7% to 6.8% for goats, while the gain for pigs was not so significant as it lowered from 12.7% to 9.2%. This less satisfactory gain for pigs may be suggesting this Register may be considered as the worse accurate, and that probably either the record linkage procedure was misclassifying holders and owners, or that during the collection phase the enumerator attributed wrongly confusing holders with owners and vice versa. That indeed turned out to be the problem and indeed by checking a smaller subset of influential observations affected it was necessary to impute only 3.3% (see table 3) of the records instead of 9.2% which were predicted as influential by the model. Of course when applying a general methodology to real data, there are many underlying hypothesis and assumptions that may compromise results whenever they don't hold true.

After this step the patterns associated to the influential observations of sheep and goats were further analyzed and some re-contacts at NUTS2 level gave evidence of the correctness of the Register values and helped in the identification of localized item non-response associated to livestock(sheep and goats) that emerged through the record linkage procedure. For this reason

in has been decided to consider as true the value of the Register and subsequentially use it for automatic imputations.

Given for true the total amounts of sheep and goats, it was necessary to attribute the partials to each of the three possible subcategories, because the Register did not contain this information. Moreover the set to be imputed was constituted by more than 80% of item non responses, while the remaining 20% could have been corrected also through deterministic rules such as rescaling partials in order to achieve the correct total.

Composition of sheep in terms of their subcategories

Table 2

	Milk Sheep	Other Females	Males
raw_data_not_weighted	68.2%	21.1%	10.7%
raw_data_weighted	64.2%	24.3%	11.5%
validated_data_weighted	63.8%	24.8%	11.4%
validated_data_not_weighted	68.0%	21.3%	10.7%

It is well known that regression and parametric models are strongly affected by the presence of 0s and missing values, and a popular alternative approach is multiple imputation, but in this case the use of donors wouldn't have allowed to keep constant the total edited value of livestock, and a subsequent work would have been necessary to restore totals. For this reason different approaches were tested by using two packages that specifically deal with compositional data.

The idea was not introduce any bias in the relative proportion of partials and to preserve their distribution, since as expected weighted and not weighted composition were not differing (see table 2). It turned out that rspa in the imputation step of missing values attributes on average to each subcategory 1/3 as shown in figure 1, while milk sheep represent on average 68.2% of the total. It would have been possible to adjust weights in order to try achieve the average values for each subcategory, that would have in any case remained close to a flat step. In any case this would have led to a bias in the overall distribution as shown in figure 2 and 3, where the rspa imputation produces a step around 0.3-0.4 both for milk sheep and other females. The change of weights would have simply moved the step.

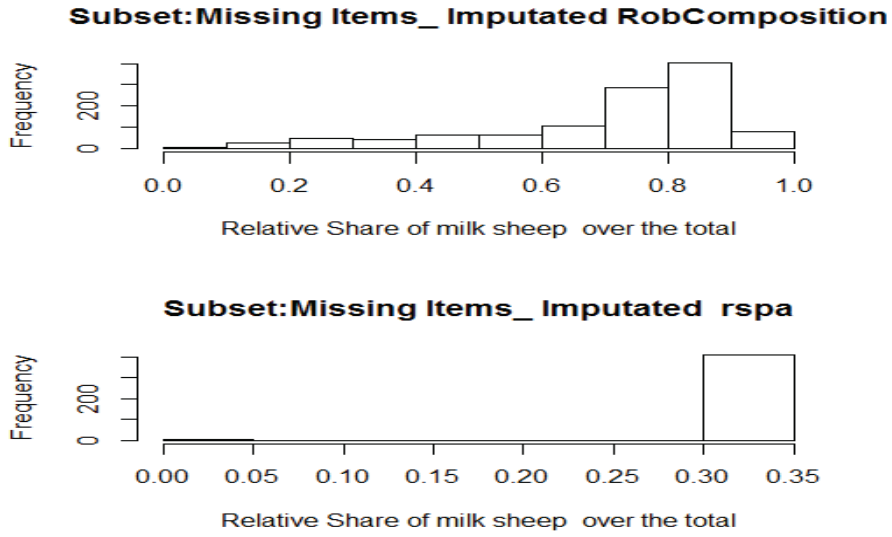
On the contrary the package RobComposition preserves well the relative proportions as shown in figure 1. By analyzing the distribution of the relative shares of milk sheep and other females after the imputation step, it seems that RobComposition smoothen a bit the distribution towards the mean value, but preserves much more the distribution shape of the frequencies

observed in the raw data , as shown in figure 2 and 3 where are reported the figures for the two subcategories referred to the whole population, before and after imputation.

For this reason it has been judged impossible to apply only rspa as initially thought. Nonetheless rspa has been applied as a subsequent step of RobCompsition, to impute the remaining 20% inconsistencies and the small ones arising from the rounding of RobComposition.

Comparison of the imputation of missing items of the two different R packages rspa and RobComposition. Frequency of the ratio of milk sheep over the total, relative to to the imputed subset

Figure 1



Imputation rates for livestock categories

Table 3

Missing items Imputation rate	Overall imputation rate	Species
0.1%	0.1%	**Cattle**
9.9%	11.1%	**Sheep**
10.3%	11.5%	**Goats**
2.1%	3.3%	**Pigs**
3.2%	5.7%	**Livestock (excluding poultry and rabbits)**

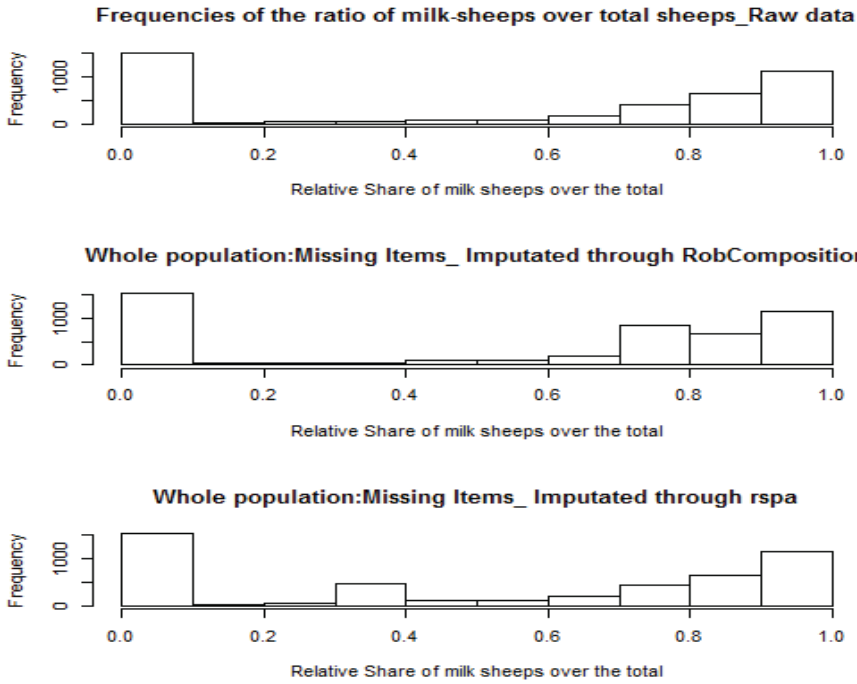
Summarizing first, the set of missing items of sheep and goats were imputed through RobComposition, then rspa was applied to the remaining residual inconsistencies for all livestock categories.

The overall procedure for livestock gave rise on the whole to an imputation rate of 5.7%, including modification rate, net imputation rate and cancellation rate, as shown in table 3. The treatment of pigs reduced to a small subset of influential errors that were due a problem of attribution of the animals, holder vs owner. Cattle information resulted to be very precise (0.1% imputation rate), while that did not hold true for sheep and goats, which, it must be stressed, have a very different geographic distribution compared to cattle where the highest imputation rates 11.1% and 11.5% were recorded, as reported in table 3.

On the whole the approach with R gave significant spare of time when compared with other parallel approaches that were more focusing in the development of consistency checks at record level.

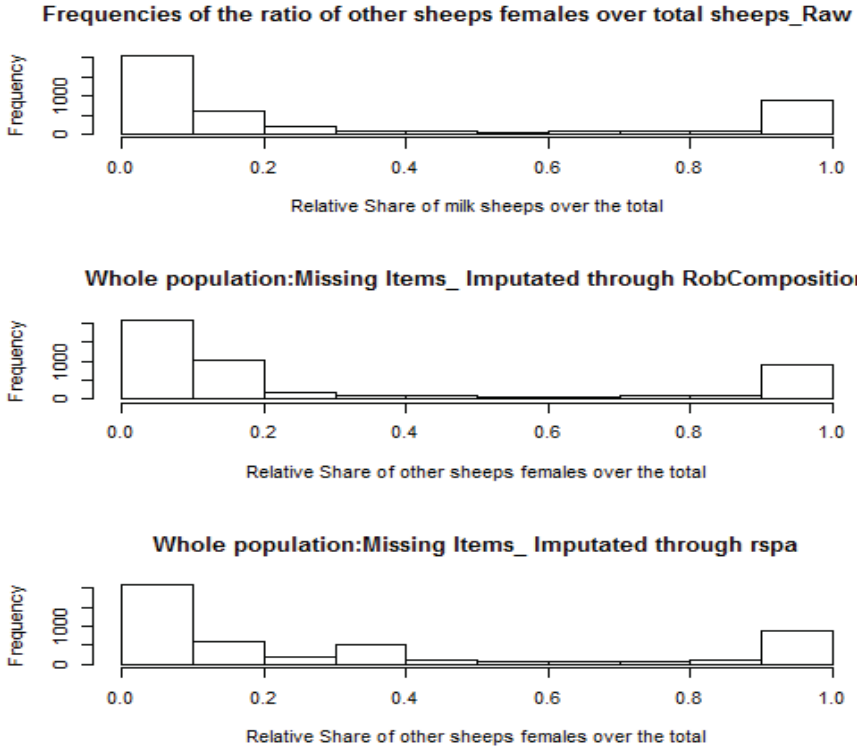
Comparison of the frequencies of the ratio of milk sheep over the total, relative to the whole population for: a) raw data b) data edited thorough RobComposition c) data edited thorough rspa

Figure 2



Comparison of the frequencies of the ratio of other sheep females over the total, relative to the whole population for: a) raw data b) data edited thorough RobComposition c) data edited thorough rspa

Figure 3



4. 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.

These analysis and the relative E&Is' adopted choices were massively carried out by means of R packages, that were chosen among the variety offered by the R environment, thus giving rise to a mixed and combined strategy.

Selective editing techniques were fruitful to verify both the quality of the Register, confirming experts' opinion, and to reduce the impact of manual checking and re-contact techniques, but also helped in the detection of

systematic errors, such as record linkage problems for pigs, and to recognize patterns of missing items for sheep (9.9% see table 3) and goats (10.3%) that where significantly high in some NUTS2 regions, this probably linkable to enumerator work, thus introducing under-estimation.

According to the overall pattern of errors, some packages may be more suitable than others.

This work reviews the combination of different R packages that were applied from raw to validated data.

FSS 2013 livestock experience should be useful for those that in practice must follow a E&I phase and want to achieve a good performance in terms of efficiency and quality, 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.

Overall imputation rates in the present work (see table 3) remained acceptable, confirming that non over editing has been performed.

References

1. **Aitchison J**, 1982, *The statistical analysis of compositional data with discussion*. J. Royal.Stat. Soc., Series B Statistical Methodology) 44 2): 139±17
2. **Andridge, R.R., Little, R.J.A.,** 2010, *A Review of Hot Deck Imputation for Survey Non-response*.International Statistical Review.78, 40–64.
3. **Bankier, M.**, 2011, *"Imputing Numeric and Qualitative Variables Simultaneously"*, A Technical
4. **Bankier**, 2000, *Canadian Census Minimum change Donor imputation methodology*. Technical Report.
5. **Bankhofer, U.**, **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.):
6. **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
7. **D'Orazio M., Di Zio M., Scanu M.**, 2006, *Statistical Matching, Theory and Practice*. Wiley, Chichester
8. **De Waal T.,Pannekoek J, Scholtus S.,** 2011, *Handbook of Statistical data***,** Editing J.Wiley&Sons
9. **Di Zio, M., Guarnera, U.**, 2013, *A Contamination Model for Selective Editing*, Journal of Official Statistics. Volume 29, Issue 4, Pages 539-555
10. **Fellegi, I. P.** and **Holt, D.,** 1976, *"A Systematic Approach to Automatic Edit and Imputation",* Journal of the American Statistical Association, 71, 17-35.
11. **Hidiroglou, M.A.**, **Berthelot, J.M.,** 1986, *Statistical editing and imputation for periodic business surveys.* Survey Methodology, 12(1):73–83, June 1986. Statistics Canada
12. **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
13. **Kovar, et al.**, 1988, *Overview and Strategy for the Generalized Edit and Imputation System*. Statistics Canada, Methodology Branch

14. **Kozak R**., 2005, *The BANFF system for automated editing and imputation*. Proceedings of SSC Annual meeting. June 2005 proceeding of the survey methods Section
15. **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.
16. **Lawrence, D., and McKenzie, R.** 2000, *The General Application of Significance Editing*. Journal of Official Statistics, **16**, 243-253.
17. **Templ, M., Alfons, A., Filzmoser**, **P.,** 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

Packages and Methodologies

18. Memobust, Handbook on Methodology of Modern Business Statistics 2014, ***www.cros-portal.eu***
19. Cran Package: ***https://cran.r-project.org/web/packages/robCompositions/robCompositions.pdf***
20. CranPackage:***http://www.istat.it/it/strumenti/metodi-e-strumenti-it/strumenti-di-elaborazione/SeleMix*** (only in italian)
21. Cran Package: ***http://cran.r-project.org/web/packages/SeleMix/index.html***
22. 'editrules' package - **Edwin de Jonge, Mark van der Loo**, 2013. Available at: ***http://cran.r-project.org/web/packages/editrules/index.html***
23. 'rspa' package - **Van der Loo**, *Adapt numerical records to fit (in)equality records with successive projection Algorithm.* ***https://cran.r-project.org/web/packages/rspa/index.html***
24. VIM package: ***https://cran.r-project.org/web/packages/VIM/index.html***
25. StatMatch package: ***https://cran.r-project.org/web/packages/StatMatch/index.html***
26. Recommended practices for Editing and Imputation in cross-sectional Business surveys (O.Luzi et al 2007)
27. Report Detailing the Methodology of CANCEIS, Internal report, Statistics Canada
28. R package: ***https://cran.r-project.org/web/packages/HotDeckImputation/index.html***
29. Data Analysis, Machine Learning and Knowledge Discovery. Studies in Classification, Data Analysisand Knowledge Organization, 3–11. Berlin/Heidelberg: Springer.