Data Editing and Imputation in Business Surveys Using "R"

Elena ROMASCANU (elena.romascanu@insse.ro)
National Institute of Statistics, Romania

ABSTRACT

Purpose – *Missing data are a recurring problem that can cause bias or lead to inefficient analyses. The objective of this paper is a direct comparison between the two statistical software features R and SPSS, in order to take full advantage of the existing automated methods for data editing process and imputation in business surveys (with a proper design of consistency rules) as a partial alternative to the manual editing of data.*

Approach – *The comparison of different methods on editing surveys data, in* R *with the '*editrules' *and 'survey' packages because inside those, exist commonly used transformations in official statistics, as visualization of missing values pattern using 'Amelia' and 'VIM' packages, imputation approaches for longitudinal data using 'VIMGUI' and a comparison of another statistical software performance on the same features, such as SPSS.*

Findings – *Data on business statistics received by NIS's (National Institute of Statistics) are not ready to be used for direct analysis due to in-record inconsistencies, errors and missing values from the collected data sets. The appropriate automatic methods from R packages, offers the ability to set the erroneous fields in edit-violating records, to verify the results after the imputation of missing values providing for users a flexible, less time consuming approach and easy to perform automation in R than in SPSS Macros syntax situations, when macros are very handy.*

Keywords: *Business Surveys, Automated Edit Rules, Missing Values, Pattern of Missing, Random vs. Systematic Errors, Multiple Imputation, Non-Response Weights, Statistical software R, SPSS, SQL*

INTRODUCTION

This paper is concerned only with an essential aspect of business surveys post data capturing stage, the treatment of numerical data under linear constrains done by computationally-intensive techniques. In order to build the editing strategy and to provide high quality statistical information, the methods discussed in this paper could be considered appropriate for identifying random errors. Therefore, the treatment of errors should be done accordingly to their origin: random or not-random (systematic errors) and to treat those non-random firstly, applying any automatic method. Improving editing techniques

for business surveys, means to make them less costly than traditionally has been done, while maintaining the accuracy.

Business surveys can be classified in two broad categories: those producing short-term statistics and those focusing on structural statistics. Therefore, the term "survey" will refer to "a sample survey". The sampling frame is create from the Romanian Business Register (REGIS), which contains all enterprises, authorities and organizations as well as their local units that carried out any economic activity, their size or if they belong to the private or public sector. Registers containing detailed legal unit data records on a business population are used, but cannot always deliver, even after maintenance process or updating all specific information required. Conducting surveys is usually designed to obtain information directly from businesses and is widely used by Statistical Institutes due to its flexibility to ask specific questions.

The data on business statistics received by NIS's are not ready to be used for direct analysis due to in-record inconsistencies, errors and missing values from the collected data sets. To produce statistical output these problems have to be treated using: error detection, correction and imputation. Edit and imputation (E&I) are known as one of the most important aspect of business surveys but a very time consuming process for NISs. The process of dealing with data cleaning methods, become a strength in finding the best practices and having micro or macro data base ready to be analysed by different data user.

METHODOLOGY

The intention is to find, explore and use the proper and easy way to deal with method thus reducing the time for validation at the expense of other important phases of surveys that in turn require error checking. Since data from surveys often contain errors, it is desirable to detect these errors. To exemplify this, two files were used from a business survey sample in order to demonstrate some of the R statistical software tool functionalities with simple examples. The main advantage are increasing the efficiency of the editing processes and make use of existing automated methods (with a proper design of consistency rules) as a partial alternative to manual editing of data. In statistical offices those methods and tools are used at the post data receiving stage, for indentifying and elimination of errors that could otherwise affect the collected data. Modern goals of editing statistical data, especially for business surveys, can reduce the potential for bias arising from influential or not-influential errors. Editing has a major role in the data cleaning process but its most useful role derives from its ability to provide information about the survey process, quality measures and improvements for future surveys.

Another consideration worth taking into account is the resources or time-consuming features and it has been estimated that National Statistics Institutes spend around 40% of their resources on editing and imputing data (De Waal, et al.2011). For efficiency reasons, it may be desirable to edit at least part of a data file by automatic methods (see "MEMOBUST Handbook, Statistical Data Editing – Main Module"). It is recognized that the fatal errors (e.g., invalid or inconsistent entries) should be removed from the data sets in order to maintain accuracy and to facilitate further automated data processing and analysis.

The goal of automatic editing is to accurately detect and treat errors and missing values in a fully automated manner, without human intervention. A recent development in NSIs is represented by the increasing use of administrative data sources, as opposed to the more traditional data collection by sample surveys, approaches and constraints for reducing response burden.

It is known that not all data need to be corrected i.e. not all data containing errors need to be corrected to the smallest detail (over-editing). Studies prove (Granquist and Kovar, 1997) that there is no need to eliminate all errors in the data set to obtain reliable publication figure. The main products of statistical offices are tables containing aggregated data, so, small errors in the individual records are acceptable for proper and tend to cancel out when aggregated. But on the other hand, the influential errors and other relevant errors like unit measure errors and other systematic errors generally produce high impact on published figures.

The traditional goal of edit, detect and correct errors in the collected data is very labour-intensive and time-consuming with a degree of inefficiency because the measurement error is not the only source of error in statistical output. Generally, there are major differences in choosing the proper technique depending on the kind of data: numerical or categorical. Many national statistical institutes (NSIs) use nowadays automatic editing. Almost automatic editing methods treat a record of data in two steps: first, an attempt is made to identify the variables with erroneous or missing values (the error localisation problem) and second, new values are imputed to obtain a valid record.

RESULTS

This material aims to explain that the containing packages 'editrules', 'VIMGUI', 'survey' and 'Amelia' inside the R Project for statistical computing, with the exclusive use in editing and imputation process are performing well in localisation problems.

R Project has a lot of packages implementing various functions to handle missing values and missing value imputations (note: this is only a partial

list): ‘Amelia’, ‘arrayImpute’, ‘bcv’, ‘cat’, ‘crank’, ‘CVThresh’, ‘crank’, ‘compositions’, ‘Design’, ‘dprep’, ‘eigenmodel’, ‘EMV’, ‘FAwR’, ‘Hmisc impute’, ‘imputeMDR’, ‘MADAM’, ‘mclust’, ‘Mfuzz’, ‘mi’, ‘mitools’, ‘mice’, ‘missMDA’, ‘mimR’, ‘mix’, ‘MImix’, ‘MIfuns’, ‘monomvn’, ‘mvnmle’, ‘norm’, ‘nnc’, ‘optmatch’, ‘pan’, ‘pcaMethods’, ‘prabclus’, ‘rama’, ‘randomForest’, ‘rconfifers’, ‘relaimpo’, ‘robCompositions’, ‘rrp’, ‘scrime’, ‘SDisc’, ‘simsalabim’, ‘VIM’, ‘vmv’, ‘yaImpute’.

Is useful and may solve specific and clearly identified situations in business surveys using the content of these items in contrast with much-used SPSS through which, one could test and determine the pattern of missing data but even the SPSS 22 (latest version), provides poor option for handling missing data, even though offers the Little’s MCAR test as a measurement tool regarding data missing but do not offer parallel box plots or scatter plot matrix with information about missing/imputed values as one could make use of, in VIM. Although SPSS performs better and unravel some good imputation methods, including stochastic regression and EM imputation, there are voices considering that SPSS missing value analysis has been biased and limited in the types of imputations. This situation was improved with reference to the last five versions. Still, to explain all relations between them, that latest version of SPSS traverses the space to R, using the R Integration Package for SPSS Statistics, which provides the ability to use R programming features within SPSS Statistics. This feature requires the SPSS - integration plug-in for R, installed with SPSS Statistics - Essentials for R (see SPSS 22, tutorials are available by choosing Help-Working with R). With these tools one has everything he needs to create custom procedures in R. In addition, the quality of imputation can’t be visually explored using various univariate, bivariate, multiple and multivariate plot methods as ‘VIM’ or ‘mice’ R packages can do, but returning on the general descriptive statistics submenu and then to check plots of the means and standard deviations by iteration and imputation for each scale dependent variable for which values are imputed.

Other example is the identification of null values and missing values using relational databases which is easy for experienced NISs staff, but sparingly time consuming. T-SQL creates an object called a rule that specifies the acceptable values that can be inserted into that column. A rule can be any expression valid in a WHERE clause and can include elements such as arithmetic operators, relational operators, and predicates (for example, IN, LIKE, BETWEEN) but thinking on imputations using SQL, is preferred not continue the process of imputation of missing values. Even if there is the opportunity to lead this process in a DBMS without considering replacing missing by mean or median values, but by building a k-means clustering

model based on, it may be desirable or even necessary to perform a statistical analysis in a statistical package rather than in the database. On the other hand, R package 'sqldf' or 'DBI' package speaks well the language of relational databases helping to achieve the same goals that we achieve using SQL so we can find again a path of overlap the lineament among R environment and other often used software packages.

Using R, a record of data can be represented as a vector of fields or variables called its domain. Examples of variables and domains are the size class with domain (small, medium, large), the number of employees with domain (0,1,2…n) , and profit with domain (0, ∞). Edit rules are derived from conditions that should be satisfied by the values of single variables or combinations of variables in records. For the purpose of automatic editing, all edit rules must be checkable per record, and may therefore not depend on values in fields of other records.

Examples of edit rules are given below:

Annual turnover ≥ 0, should be non-negative; profit = turnover – total costs; IF (size class = "medium") THEN (10 ≤ Number of employees < 49) for mixed data containing both character and numerical field; NACE code check for validity, WHERE IN (select code from Nomenclature).

'EDITRULES': A PACKAGE FOR PARSING, APPLYING, AND MANIPULATING DATA CLEANING

A good way to see and test errors in an automatic way is installing R package 'editrules', a useful tool in detecting errors that can be expressed by checking after constructing rules based on: linear equations, restriction in well know if-else form and conditional restrictions on numerical or character type data. Automatic editing means data are checked and adjusted by computer. Those rules can be written and defined in a file .txt format.

>rules1 <- editfile("edit.txt")

>ver1 <- violatedEdits(rules1, *file*) # indicating which record violates the rules defined by yourself from the original file used

>plot(ver1)

>summary(ver1)

Edit violations, 650 observations, 0 completely missing (0%):

- Returns NA when edits cannot be checked because of missing values in the data.

- rules1- character vector with: 'constraintsm', 'editset', 'editmatrix' or 'editarray'.

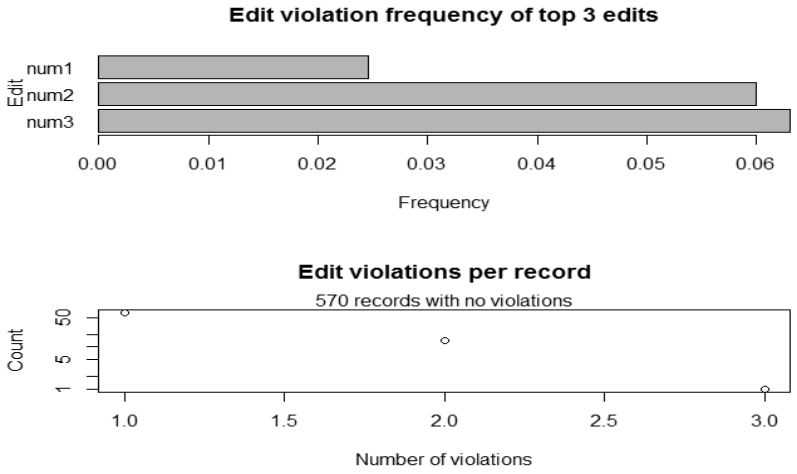
editname	freq	rel
num3	41	6.3%
num2	39	6%
num1	16	2.5%

Edit violations per record:

errors	freq	rel
0	570	87.7%
1	65	10%
2	14	2.2%
3	1	0.2%

Graphic for edit violation rules

Figure 1



As for further localization, as stated by authors, (Van der Loo, 2011) 'searchBest' method gives the lowest-weight solution to the error localization problem. Apart from 'searchBest', there are other solvers in the backtracking object returned by 'errorLocalizer', namely

-'searchNext' - search the next solution in the binary tree;

-'searchAll' - return all solutions encountered while traversing the binary tree in a branch-and-bound manner;

-'searchBest'- returns a random lowest-weight solution if multiple are found;

- ‘reset’ - reset backtracker object to initial state;

In practice, automatic editing introduced by this package is based on Fellegi-Holt paradigm (see Fellegi and Holt, 1976) which consider that the smallest (weighted) number of field is settled, which allow the record to be imputed consistently. In fact, the method Fellegi-Holt only provides a list of variables ready for imputation process in order to have clean data based on edit rules, but it does not provide the final value to impute. Is needed another level strictly based on accurate imputation rules.

The list of edit violations seen above produces a list of fields and observation of violated imposed rules of editing. Edit rules (also called validity rules) impose conditions that should be satisfied by the values of single variables or combinations of variables in a record.

Besides systematic errors, data also contain non-systematic, random errors that are not caused by a systematic reason, but randomly. An example is typing too many digits. To identify such non-systematic errors, Fellegi-Holt paradigm is suitable and recommended (De Waal, 2003) because that the data in a record should be made to satisfy the specified edits by changing the fewest possible (weighted) number of fields. To each variable a non-negative weight, the so-called reliability weight, is assigned that indicates the reliability of the values of this variable. The higher the weight of a variable, the more reliable the corresponding values are considered to be. If all weights are equal, the generalized Fellegi-Holt paradigm reduces to the original Fellegi-Holt paradigm.

This method works well for a record that contains fewer errors. Given such a minimal index set, De Wall, (De Wall, 2003) construct the implied edit, given by:

IF $v_r \in D_r$, $v_i \in \bigcap_{j \in S} F_i^j$ for i=1,…,r-1,r+1,…,m

THEN $(x_1,\ldots,x_n) \in \varnothing$

Some records are not suitable for automatic editing and when the records contain many errors based on a set of predefined maximum number of errors then, the records will not be introduced in the process of automated editing but will be considered for other method such as, reweighting (see ‘survey’ R package).

Rules can be defined with common R syntax and parsed to an internal (matrix-like format) and can be manipulated with variable elimination and value substitution methods, allowing for feasibility checks. Data can be tested against the rules and erroneous fields can be found based on Fellegi and Holt’s generalized principle.

The discussion that emphasizes specific preference or the need of a specific software power to use in enterprise statistical surveys conducted in the offices of statistics exists because our objective is getting inferential analysis accuracy, rigor and build as well as possible avoiding the generalization errors. Diagnostic procedures yield information about the nature of missing data and potential biases due to missing data. We need this evaluation since both numerical and graphical diagnostics procedures provide information by which to better handle, diagnose and interpret missing data and their impact on study results.

'AMELIA II': A PACKAGE FOR MISSING DATA

'Amelia II'- "multiply imputes" missing data in a single cross-section (such as a survey), from a time series (like variables collected for each year in a country) or from a time-series-cross-sectional data set (such as data collected by years for each of several countries).
'Amelia II' implements our bootstrapping-based algorithm that gives essentially the same answers as the standard IP or EM approaches, is usually considerably faster than existing approaches and can handle many more variables.

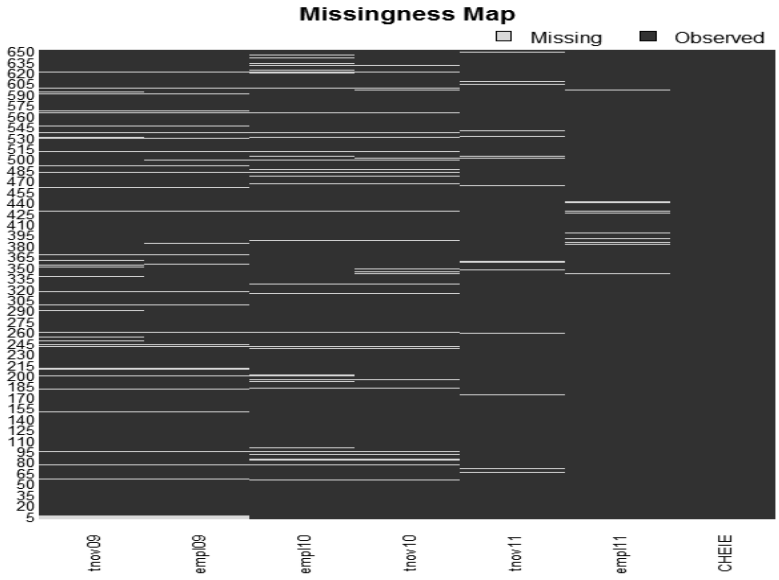
```
>install.packages('Amelia', repos="http://r.iq.harvard.edu", type = "source")
>library(Amelia)
>AmeliaView()
```

The imputation model in 'Amelia II' assumes that the complete data (that is, both observed and unobserved) are multivariate normal. Amelia requires both the multivariate normality and the MAR assumption (or the simpler special case of MCAR).

When using multiple imputations, the main idea is to identify the variables to be included in the imputation model. Is needed to include at least as much information as will be used in the analysis model. This means that any variable present in the analysis model should also be in the imputation model including, of course, any transformations or interactions of variables that will appear in the analysis model.

Missingness map

Figure 2

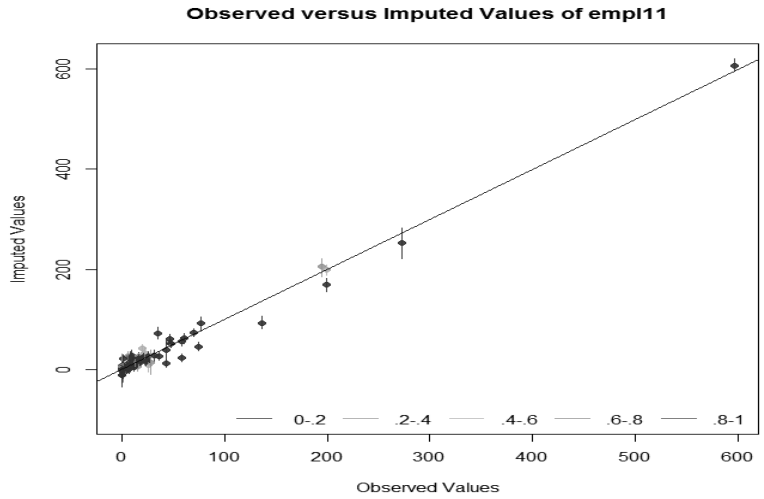


Missing values are in tan and observed values are in red.

The missing values map is an important tool for understanding the patterns of missingness in the data and indicate different ways to improve the imputation model. Variables considered are number of employees (as auxiliary variable) and annual turnover of enterprises. The correlation between those variables is important and could be a feasible solution also for small area estimation as well, in structural surveys.

Example of diagnostic graph

Figure 3



The color of the line (as coded in the legend) represents the fraction of missing observations in the pattern of missing values for that observation.

For each observation, Amelia also plots 90% confidence intervals that allow the user to visually inspect the behavior of the imputation model. By checking how many of the confidence intervals cover the y = x line, one can tell how often the imputation model can confidently predict the true value of the observation.

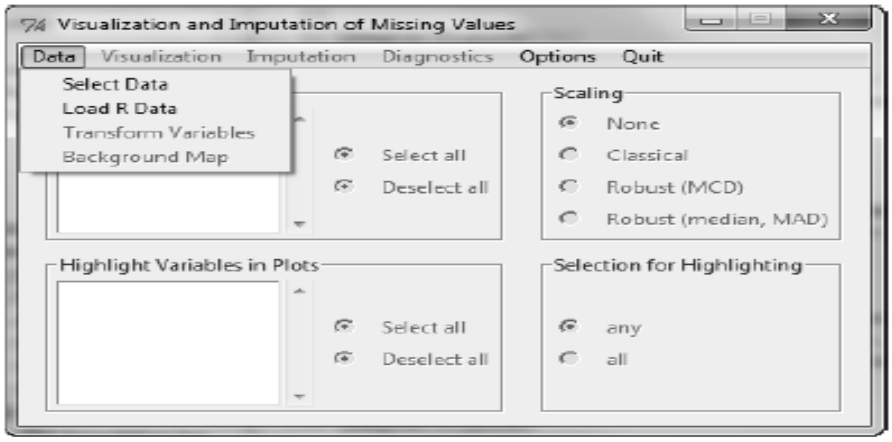
A typical scenario for a business survey is that data for relatively small businesses with simple structures are taken or derived from tax returns, whereas surveys are used to collect data from the key units (usually those that are largest and/or have the most complex structures)

'VIM' : a package for visualization and imputation of missing values

'VIMGUI()' This package introduces new tools for the visualization of missing and/or imputed values, which can be used for exploring the data and the structure of the missing and/or imputed values. Depending on this structure of the missing values, the corresponding methods may help to identify the mechanism generating the missings and allows to explore the data including missing values. In addition, the quality of imputation can be visually explored using various multiple plot methods. A graphical user interface allows an easy handling of the implemented plot methods (cran.r-project.org).

The VIM GUI and it's menu for importing data

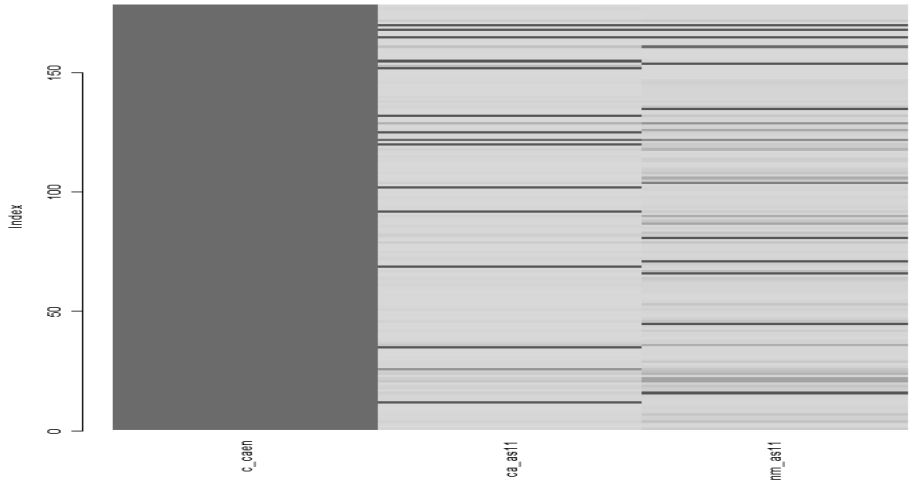
Figure 4




```
>activedataset <-spss.get("C:/FILE.SAV",use.value.labels=FALSE,lowernames=TRUE,
force.single=TRUE,charfactor=TRUE,to.data.frame = TRUE)
 >originaldataset <- activedataset
 >matrixplot(activedataset, sortby = "c_caen")
```

Matrix plot of missing values

Figure 5



Using the function ‘matrixplot’ one can create a color matrix plot in which the data cells are represented by a colored rectangle. The data matrix plot can also be sorted by clicking inside the plot space on the variable’s column which one wants to sort by. This is an example of pattern of missing at random (MAR).

From Imputation menu one can choose between many methods of missing values as: KNN, hotdeck, IRMI. The GUI has two menus for graphical methods: “Visualization” created for analysis of missing value before imputation and “Diagnostics” is designed to see the outcome after imputation process.

Variables sorted by number of missings:

Variable	Count
ca_as11	13 (annual turnover)
nm_as11	10 (n.employee)
c_caen	0

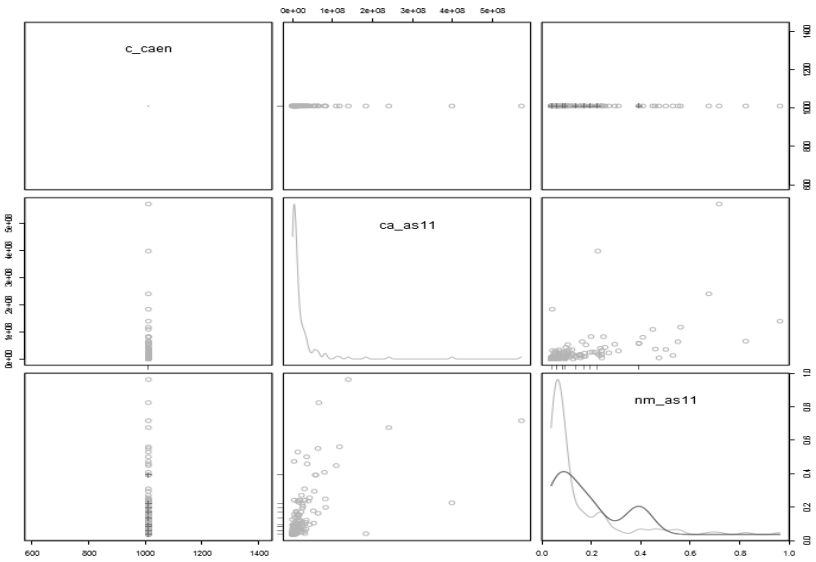
For exemplification and due to incomplete item values we make use of hot deck method - nearest neighbor imputation, used to compensate for non-response in sample surveys.

The k- Nearest Neighbor Imputation based on a variation of the Gower Distance for numerical, categorical, ordered and semi-continuous variables has the following usage:

```
>activedataset <- kNN(activedataset, variable= c( “nm_as11” ) , k= 5 ,  dist_
var= c( “ca_as11” ) ,  weights= NULL ,  numFun= median ,  catFun= maxCat
,  impNA= TRUE ,  addRandom= FALSE ,  mixed= NULL ,  mixed.constant=
NULL )
```

Scatterplot Matrix of the variables, imputed values in *nm_as11* are highlighted

Figure 6



Statistical surveys tend to suffer from varying degrees of non-response, which affects the efficiency of the sampling process, and the quality of the resulting statistics. Non-response typically takes one of two forms, "unit non-response", in which no data are supplied for the unit concerned, or "item non-response", in which a partial return is provided, but some data items are blank.

A more convenient alternative may be to decide that if data not provided by a particular date, of a single units, are or not vital to the survey results (e.g. smaller businesses in a business survey), they are instead taken from administrative sources.

Relying on a strong correlation between the administrative data and the survey data, the survey data can be either replaced directly with administrative data or indirectly through the production of modeled values based on the relationship between the two sets of data. As such, more and more business survey estimates are being based on a combination of survey and administrative information.

NSIs are increasingly turning to the use of administrative data to reduce the cost of surveys and to reduce the burden on respondents.

The basic way for estimating is summing weighted variable values for the units that happened to be in the sample. Suppose having 100% response rate, this situation gives an unbiased estimator. This is the Horvitz-Thompson (abbreviated as H-T) estimator for the population total. A more advanced point estimator is the generalized regression (GREG) estimator.

For surveys, it is important to know if, and how, non-respondents differ from respondents. This is imperative knowing if we are making correct inference from sample data. There are some main elements that characterize business survey data (Granquist, 1995). Firstly, responses to items of interest often present highly skewed distributions, in other words, a small number of units substantially contribute to the total estimate. Furthermore, information on the surveyed businesses is often available from a previous survey or can be drawn from administrative sources.

As collected, micro-data often include implausible or impossible values, for example arising from multiple forms of survey error (Groves 1989), such as reporting and measurement error. NSIs prefer not to release such faulty values and so undertake a process usually referred to as "edit and imputation" (De Waal et al. 2011).

Editing and imputation is a set of activities detecting erroneous and missing data and treating these data accordingly. When there are incidences of missing values in quantitative data items, such as sales and fixed investment, the current practice is to compile the survey result by imputing the missing value with the "previous fiscal year's value obtained from the non-responding enterprise." Like household surveys, business surveys often use one of the following methods to account for non-response: follow-up, imputation, or weighting adjustments. Imputation is done at the unit or item level and is the process of creating non-missing by inferring from other data what a missing value "should" be. (Singh and Petroni, 1988)

Unit non-response is usually treated by weighting the responding cases accordingly. In some applications even unit non-response is treated with unit imputation, meaning that one unit missing is replaced by another unit close to the first one in a metric way, using nearest neighbor technique, but reweighting is perhaps the most common method because is an approach which can be use to tackle bias resulting from non-response. The main intention with reweighting the data is to adjust the original inclusion probabilities by the response probabilities. When a stratified sampling survey is conducted with imperfect response it is desirable to rescale the sampling weights to observe the non-response.

Edit and imputation (E&I) are known as one of the most important aspect of business surveys but a very time consuming process for NISs. The

process of dealing with data cleaning methods to strength and consider the best practices for having micro or macro data base ready to be analyzed by different data user. Some forms of imputation are known as logical or deductive imputation but mostly when dealing with item non-response as opposing the earlier discussed, unit non-response.

All business surveys suffer from effect of non-response. Are well known the reasons why this happens but the important fact is that in those surveys the non-response is rarely Missing Completely at Random (MCAR). Systematic non-response patterns (MNAR) are responsible for biases in survey estimates and is imposed the use of weighting methods. Dealing with those aspects is not easy because on one hand, agreed methods refer to weighting and imputation but on the other hand considering the presence of systematic missing pattern is not always appropriate for common imputation. The assumption about randomness (MCAR, MAR, MNAR) must always be evaluated. Methods such as ratio, regression, nearest neighbor imputation are appropriate for business surveys where one can use other sources, preferred those longitudinal, in order to reduce non-response bias. Perhaps the most valuable R Package under assumption of MNAR is 'mice' (Multivariate Imputation by Chained Equations); it can handle both MAR and Missing Not at Random (MNAR). Multiple imputations under MNAR, requires additional modeling assumptions. The default methods given by 'mice' package are in partially presented in package 'MissingDataGUI', a GUI for Missing Data Exploration.

'SURVEY': A PACKAGE FOR ANALYSIS OF COMPLEX SURVEY SAMPLES

Also for non responses one can also take into account the R package 'survey'. 'Nonresponse()' combines stratified tables of population size, sample size, and sample weight. 'SparseCells' identifies cells that need combining.

```
>sparseCells(nr)
Cells: 3 5 7 11
Indices:
strVar1   strtVar2   strtVar3
3  "No"     "Yes"    "E"
5  "No"     "No"     "H"
7  "No"     "Yes"    "H"
11 "No"     "Yes"    "M"
```

Summary:
NR wt wt n
3 Inf Inf 0
5 3.2 108 3
7 Inf Inf 0
11 Inf Inf 0

'Neighbors' # describes the cells adjacent to one specified cell
>neighbours(3,nr) # look at neighbours
>nonresponse(object, nbour.index) # create a nonresponse object

Cells: 4 7 1
Indices:
strVar1 strtVar2 strtVar3
4 "Yes" "Yes" "E"
7 "No" "Yes" "H"
1 "No" "No" "E"

Summary:
NR wt wt n
4 0.92 31.1 112
7 Inf Inf 0
1 1.04 35.2 12

Function 'joinCells' :
>joinCells(nr1,3,11,8) # collapse some contiguous cells
>nonresponse(sample.weight, sample.count, table)
12 original cells, 8 distinct cells remaining
Joins:
3 5 7
3 5 7 8 11
counts NRweights totalwts
Min. : 3.00 Min. :0.6840 Min. :23.15
1st Qu.: 7.00 1st Qu.:0.8956 1st Qu.:30.31
Median : 11.00 Median :0.9793 Median :33.15
Mean : 22.88 Mean :1.1461 Mean :38.79
3rd Qu.: 15.50 3rd Qu.:1.3142 3rd Qu.:44.48
Max. :112.00 Max. :2.0977 Max. :71.00

When the collapsing is complete, use 'weights()' to extract the non-response weights.

CONCLUSIONS

In this paper, several recent approaches in missing data methods for identifying missing values in data sets were tested. R has data structures for example on edit rules that allow thinking more about the statistics performed than about the internal representation of data and, on the other hand, the automation is easier to perform in R than in SPSS, with concern on validity, reliability, power of the study.

The techniques and functionality discussed in this article represent a very small percentage of the available methods for identifying, displaying, and imputing missing values.

REFERENCES:

1. De Waal, T., 2000, *An Optimality Proof of Statistics Netherlands' New Algorithm for Automatic Editing of Mixed Data*. Report, Statistics Netherlands, Voorburg.
2. De Waal, T. and R. Quere (2003), *A Fast and Simple Algorithm for Automatic Editing of Mixed Data*. Submitted to Journal of Official Statistics
3. De Waal, T., Pannekoek, J., and Scholtus, S. (2011*), Handbook of Statistical Data Editing and Imputation*, Wiley
4. Fellegi, I.P. and D. Holt (1976), *A Systematic Approach to Automatic Edit and Imputation.* Journal of the American Statistical Association 71, pp. 17-35.
5. Granquist, L. and Kovar, J.G. (1997). *Editing of survey data: how much is enough? In Survey Measurement and Process Quality*, Lyberg et al (eds.). Wiley, New York, 415-435
6. Groves, R. M. (1989), *Survey Errors and Survey Costs*, New York: Wiley.
7. Honaker, James, and Gary King. 2010. *What to do About Missing Values in Time Series Cross-Section Data*. American Journal of Political Science 54, no. 3: 561-581
8. M. Templ, A. Alfons, A. Kowarik, and B. Prantner. VIM: *Visualization and Imputation of Missing Values*, 2011a. URL http://CRAN.R-project.org/package=VIM.
9. Memobust, *Handbook on Methodology of Modern Business Statistics*, 2014, www.cros-portal.eu
10. Scholtus, S. (2009). *Automatic correction of simple typing errors in numerical data with balance edits*. Technical Report 09046, Statistics Netherlands, Den Haag
11. Singh, R., & Petroni, R. (1988). *Nonresponse Adjustment Methods for Demographic Surveys at the U.S. Bureau of the Census*.
12. Van der Loo, M., E. De Jonge, and S. Scholtus (2011). *Correction of rounding, typing and sign errors with the deducorrect package*. Technical

Report 201119, Statistics Netherlands, The Hague.
13. R package version 3.0.0.
14. 'Amelia' package-Honaker, J., King, G., & Blackwell, M. (2010). Available at: http://cran.r-project.org/web/packages/Amelia/Amelia.pdf
15. 'editrules' package - Edwin de Jonge, Mark van der Loo (2013). Available at: http://cran.r-project.org/web/packages/editrules/index.html
16. 'survey' package – Thomas Lumley,(2013). Available at: http://cran.r-project.org/web/packages/survey/index.html
17. 'VIM' package - Templ, M., Alfons, A., & Kowarik, A. (2010). Available at: http://cran.r-project.org/web/packages/VIM/vignettes/VIM-EU-SILC.pdf