

Single Imputation Techniques of Missing Values: an Experimental Application to Income Data

*Tecniche di Imputazione Singola di Valori Mancanti:
una Applicazione Sperimentale a Dati di Reddito*

Antonia Manzari, Orietta Luzi

ISTAT, Via C. Balbo 16, 00184 Roma, e-mail: manzari@istat.it, luzi@istat.it

Riassunto: La mancata risposta parziale è generalmente trattata dagli Istituti Nazionali di Statistica mediante tecniche di imputazione singola. Particolarmente critico è il caso dell'imputazione di variabili di reddito, in termini sia di analisi e modellazione del meccanismo di mancata risposta, sia di preservazione delle proprietà statistiche delle distribuzioni delle variabili oggetto di imputazione. In questo lavoro vengono presentati i primi risultati di uno studio sperimentale condotto in ISTAT al fine di valutare comparativamente l'accuratezza di diverse tecniche di imputazione singola di variabili di reddito.

Key words: Item nonresponse, income imputation, regression tree, regression.

1. Introduction

The problems arising from item nonresponse (some but not all of the responses are available for a survey unit) are mainly related to a loss in efficiency of estimates because of the reduced size of the complete data set, but also to the possible bias that may exist when nonrespondents are systematically different from respondents on the variable of interest. Furthermore, the “standard” software packages for complete data cannot be used to analyse the entire data set because the estimation of parameters with incomplete data requires more complex methods (e.g. the maximum likelihood method). Moreover, available-case analyses may produce inconsistent results depending on the set of variables used in the analyses.

National Statistical Institutes (NSI) are requested to provide complete data sets for public use. This task is generally achieved by compensating for item nonresponse by using single imputation methods. Single imputation consists in the assignment of a value to each missing value. It allows analysing the resulting complete data set by standard software packages producing results that are consistent with each other. The main drawback of single imputation is that standard variance formulas applied to the filled-in data systematically underestimate the variance of estimates, even if the model used to generate the imputations is correct (Little *et al.*, 2002). In fact, imputed data are generally considered as if they were actually observed thus neglecting the additional uncertainty due to nonresponse and imputation. Many different methods for single imputation have been proposed in literature. Kalton *et al.* (1982) describe the most important characteristics of the commonly used imputation methods and analyse advantages and drawbacks related to their use.

In this study the performance, in terms of accuracy, of five single imputation methods for numeric variables is comparatively evaluated. Three of them are tree-based

imputation techniques: the *random selection*, the *nearest neighbour* and the *mean imputation*. The fourth method is the *nearest neighbour* without tree structure, while the *stochastic regression imputation* is the fifth one.

Outlines of the imputation methods are given in section 2, while in section 3 the evaluation study on real data from the European Community Household Panel (ECHP in what follows) is presented together with summary results.

2. The imputation methods

The tree-based imputation methods considered in this paper are supported by the software *Weighted Automatic Interaction Detection (WAID)* (Chambers *et al*, 2001). The *WAID* software was developed as part of the AutImp project (<http://www.cbs.nl/en/services/autimp/autimp.htm>). It performs the imputation of missing values of categorical or numeric variables using a tree-based model. Given a variable (*response variable*) for which data are missing and a set of categorical explanatory variables, the software uses the complete cases to classify the data in terms of the explanatory variables values by successively splitting data into subsets that are increasingly more homogeneous with respect to the response variable. The terminal nodes of the (classification or regression) tree are then used as imputation classes for the missing values. *WAID* supports three imputation methods for numeric variables: the *random selection* (*Tree-RS* in the following), the *nearest neighbour* (*Tree-NN* in the following) and the *mean imputation* (*Tree-MEAN* in the following). In the first two methods, a complete record (*donor*) within the same terminal node of the missing record (*recipient*) is selected, and its value for the response variable is used to impute the corresponding missing value in the recipient. In the *Tree-RS* a donor is randomly selected. In the *Tree-NN*, the total distance between the recipient and each donor within the terminal node is calculated first. The total distance is the sum of distances for all the explanatory variables defining the tree. Then the donor closest to the recipient is selected. If there is more than one donor with the minimum total distance, one of them is randomly selected. In the *Tree-MEAN* the mean value computed over all donors belonging to the recipient's terminal node is used to impute the recipient.

The *nearest neighbour* method without tree structure (*NN* in what follows) has been applied by using a SAS program. The *NN* basically works as the *Tree-NN* but it uses the whole database as imputation class: in *NN* the explanatory variables are used only as *matching* variables to determine the distance between recipients and donors.

The *stochastic regression imputation* (*SRI* in the following) has been applied by using the software *IVEware* (<http://www.isr.umich.edu/src/smp/ive/>) which performs single or multiple imputation of missing values using a sequence of regression models (Raghunathan *et al*, 2001). Basically, each regression describes the relationship between the variable to impute and all the remaining variables and uses both the observed values and the values imputed in previous rounds.

3. The evaluation study

The accuracy of the imputation methods has been evaluated by a simulation approach. Each imputation method has been applied to a set of respondent data, artificially set to

missing by a MAR mechanism. Then some indicators have been used to evaluate the “distance” between the imputed values and the corresponding observed (true) values.

Real data from an Italian wave of the European Community Household Panel (ECHP) have been used. The *wage and salary earnings* variable has been selected as variable to impute (response variable, Y). In the ECHP, missing values of income components have been imputed by Eurostat as described in Eurostat (2002). In particular, for imputing the *wage and salary earnings* components, twelve auxiliary variables have been used (Eurostat, 2002, pp. 80-81). For the description of variables see Eurostat (2001).

Firstly, the 6,457 records for which Y should be reported have been considered. Out of them, 697 cases (about 11%) had a missing value for Y, while the remaining 5,760 records had an observed value for Y. We used a logistic model to regress the nonresponse indicator on the set of the twelve auxiliary variables used by Eurostat, and selected the best subset of predictors by a backward selection procedure (six predictors were retained¹). Then we estimated the probability of Y to be missing through a logistic function of the selected predictors. Out of the 5,760 complete records, we sampled 634 cases (11%) with a selection probability proportional to the estimated nonresponse probability. For each sampled unit, the observed value of Y was deleted. Finally, each imputation method was applied using all the twelve auxiliary variables.

Trees were generated measuring the within node heterogeneity by the ordinary within node sum of squares. Three different tree sizes (10, 30, 40) were considered.

The performance of each imputation method has been assessed according to three criteria: the preservation of the individual values, the preservation of the marginal distribution, and the preservation of the aggregates². The preservation of individual values has been assessed by three indicators: the distance d between the imputed values (Y'_i) and the corresponding true values (Y_i): $d = \sum_i^n |Y'_i - Y_i|/n$; the *slope* and the R^2 obtained by regressing Y on Y' . The preservation of the distribution has been assessed by using the Kolmogorov-Smirnov distance (d_{KS}) between the true values empirical distribution function and the corresponding imputed values empirical distribution function. The mean and the standard deviation have been considered for evaluating the preservation of aggregates: the absolute difference between these aggregates computed on imputed and true values have been used as indicators, thus obtaining the d_m and d_σ measures. All indicators have been computed on the set of the $n=634$ imputed values.

The overall process (missing values simulation, imputation, computation of indicators) has been replicated five times to take into account the variability due to the nonresponse mechanism. The indicators (particularly those related to the preservation of aggregates) showed a non negligible variation on the different nonrespondent samples. This variation is due to the variability of the nonresponse mechanism and to the randomness of imputations. A more complete comparative evaluation could be carried out taking into greater account the two mentioned sources of variability. Anyway, at this stage of the study, for each method and each indicator, the average of the five accuracy indices has been computed and reported in Table 1.

From Table 1, the SRI method seems to perform worse than the other methods with respect to all the considered criteria. As relating to the tree-based methods, results do

¹ Household's total net income per month categorised, Region, Occupation in current job, Main activity of the local unit of the business or organisation in current job, Status in employment, Job status.

² The preservation of associations between variables has not been considered because only one variable had to be imputed (univariate imputation) and all the imputation methods used all the auxiliary variables.

Table 1: Evaluation indicators of imputation methods for wage and salary earnings

Method	Tree size	Indicators					
		d	$slope$	R^2	d_{KS}	d_m	$d\sigma$
Tree-RS	10	9330.119	0.827	0.691	0.052	400.124	603.205
	30	8541.762	0.859	0.729	0.045	291.306	666.634
	40	8351.999	0.859	0.734	0.046	222.564	959.462
Tree-NN	10	7527.030	0.875	0.761	0.046	435.616	1269.516
	30	7409.695	0.881	0.771	0.044	390.729	868.541
	40	7397.979	0.876	0.774	0.040	528.126	933.071
Tree-MEAN	10	6708.599	1.000	0.813	0.199	377.710	5261.845
	30	6270.942	0.979	0.828	0.156	298.013	3865.014
	40	6196.396	0.971	0.831	0.151	233.706	3405.954
NN	-	7364.444	0.904	0.773	0.048	317.448	1305.832
SRI	-	10936.856	0.688	0.639	0.181	942.275	5315.717

not point out that one method is superior to others with respect to all the considered criteria: the *Tree-MEAN* is the best one in terms of preservation of individual values; the *Tree-NN* and the *Tree-RS* are the best in terms of preservation of distribution; in terms of preservation of the mean there is no remarkable difference between the *Tree-MEAN* and the *Tree-RS*; finally, in terms of preservation of data variability, the *Tree-RS* seems to perform better than the *Tree-NN* for lower tree sizes. It appears that increasing the tree size improves the imputation accuracy with respect to all the considered criteria only for the *Tree-MEAN*. Finally, the tree structure does not necessarily improve the imputation accuracy when the *nearest neighbour* approach is used: the *NN* provides better results than the *Tree-NN* with respect to all criteria but the preservation of the data variability.

In general, for the specific data and the nonresponse mechanism used in the application, the results show that the *Tree-MEAN* has to be preferred when individual values and mean are to be mainly preserved while donor based techniques better preserve the data distribution and variability. In particular, when the *Tree-MEAN* is used, the tree size helps in improving also the accuracy with respect to the data distribution and variability criteria.

References

- Chambers, R., Hoogland, J., Laaksonen, S., Mesa, D.M., Pannekoek, J., Piela, P., Tsai, P. and De Waal, T. (2001) *The AUTIMP Project: Evaluation of Imputation Software*. Research Paper 0122, Statistics Netherlands.
- Eurostat (2001) Doc.PAN 166/2001-12. *ECHP UDB Description of variables. Data Dictionnary, Codebook and Differences between Countries and Waves*.
- Eurostat (2002) Doc.PAN 164/2002-12. *Imputation of income in the ECHP*.
- Kalton, G. and Kasprzyk, D. (1982) Imputing for missing survey responses. *Proceedings of the section on Survey Research Methods, American Statistical Association*, 22-31.
- Little, R.J.A., Rubin, D.B.(2002) *Statistical Analysis with Missing Data*, second edition, Wiley, New York.
- Raughunathan T.E., Lepkowski J.M., Van Hoewyk J. and Solenberger P. (2001) A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, 27, No.1, 85-95.