WP. 37
ENGLISH ONLY

**UNITED NATIONS**
**ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Oslo, Norway, 24-26 September 2012)

Topic (vi): New and emerging methods

# USE OF MACHINE LEARNING METHODS TO IMPUTE CATEGORICAL DATA

Pilar Rey del Castillo, EUROSTAT

## I. Introduction

1. Non-response in statistical surveys is a well-known problem, and literature on the issue is widely available. Missing information in datasets is also a common scenario in the field of machine learning where different solutions continue to be proposed. These two domains – statistical surveys and machine learning – address the problem with different approaches and use different evaluation criteria to compare the results.

2. The strategies to deal with non-response follow two main approaches: prevention before it occurs and correction if – no matter how carefully the tasks have been performed – data are still missing. One way of dealing with non-response once it has occurred is to use imputation procedures. As a first step these can be classified into two large (non-disjoint) groups: model-based procedures and algorithmic procedures. The latter are usually classification and/or prediction methods from the machine learning domain. Within statistical agencies both types of method are typically used.

3. This paper deals with the case of imputation of categorical variables because it displays significant features which are not always taken into account. We will show that sometimes the practical recommendations made from the statistical approach are not suitable enough but just reuse procedures originally designed for numeric variables. This led us to test other methods from the field of machine learning.

4. Among these two procedures were chosen: a new neural networks classifier that has recently been extended to deal with mixed numeric and categorical data as inputs, and a Bayesian networks classifier. In this paper the two methods are used to impute categorical data in a microdata file from an opinion poll and the results are compared with those from another traditional imputation method by applying a practical cross-validation evaluation criterion.

5. The remainder of this paper is organised as follows. The next section contains a brief review of the state of the art with imputation in statistical surveys and the criteria for evaluating the procedures, putting particular emphasis on practical advice and recommendations for the case of categorical data. Section III presents the context of the imputation problem to be solved, and introduces the above-mentioned new neural networks and Bayesian networks classifiers. Next, section IV, reports the results of the experiment – comparing the machine learning classifiers with another traditional method. Finally, a number of remarks and conclusions are presented in section V.

# II. State of the art of categorical data imputation in statistical surveys

## A. A general introduction to ways of treating non-response

6. Before starting with imputation, let's introduce the general approach to dealing with non-response once it has occurred in surveys, independently of whether the data are numeric or categorical. Countless procedures to deal with missing values in datasets have been developed since the problem emerged. There is not even agreement between the statistical and computational intelligence communities about how to classify the methods. Trying to cover the different approaches, a possible compromise taxonomy broadly groups them into three categories which are not mutually exclusive: deletion procedures, tolerance procedures and imputation procedures (Little and Rubin, 2002; Song and Shepperd, 2007).

7. *Deletion procedures* are based on units containing complete records. The strategy consists of discarding incomplete units and using only the units with complete data for further analysis. This is easy to carry out but is satisfactory only if the amount of data missing is small. Although these methods are not usually recommended, they are by far the most widely used because this is the default option in many statistical software packages.

8. *Tolerance procedures* are the strategies to deal with missing values that are internal to the data mining and statistical analysis procedures, that is, that work directly with datasets from which data are missing without removing missing value records or completing them. The significant feature of these methods is that they are not general solutions to the problem, but special techniques appropriated only to each specific procedure. *Weighting procedures* are included in this group. Among these, *calibration* (Särndal, 2007) provides a systematic way to integrate auxiliary information in the estimation process. The method is being used increasingly in small and medium-sized countries where much auxiliary information is available from registers.

9. The last group, *imputation procedures*, are the techniques that solve the non-response problem by replacing each missing value by an estimate or imputation. One advantage of this strategy is that it splits up non-response adjustment and data analysis into two separate stages. Datasets including imputed data may be analysed using different standard techniques and possible discrepancies would not be affected by the treatment of the missing data.

10. Among these imputation techniques two large (non-disjoint) categories can be distinguished: algorithmic procedures and model-based procedures. The former – usually classification and/or prediction methods from the machine learning domain – use an algorithm to produce the results, with an underlying model implied. On the other hand, in model-based procedures the predictive distributions have a formal statistical model, and the assumptions are explicit.

11. The *algorithmic methods* include some of the procedures used to tackle non-response from the beginnings of statistical surveys: *deck imputation, nearest-neighbour imputation, mean imputation* and *general* and *logistic regression imputation*. The group also includes some of the best-known computational intelligence classification and prediction techniques: *neural networks, genetic algorithms, Bayesian networks, classification trees, fuzzy sets, hybrid procedures*, etc. Generally speaking, prediction methods are used for imputation of continuous variables and classification methods for discrete or categorical. In this last case every category or value of the variable to be imputed is associated with a classifier class, and the estimation for a missing data input consists of the classification category.

12. All the above-mentioned methods to impute missing data are sometimes criticised because the imputed data are used later as if they were known, without taking into account the uncertainty due to non-response. These may in turn cause the sampling variance of estimators to be underestimated. To solve this problem, model-based imputation methods have been developed. These provide estimates of variance that take into account the incompleteness of the data. *Model-based methods* in the wide sense include all the imputation methods making assumptions about the probability distribution of the variables and/or about the missingness mechanism. They constitute the state of the art in imputation of missing data for the statistical inference community.

13. *Multiple Imputation* (*MI*) methods appear to be the most model-based procedures used for handling missing data in multivariate analysis at this time. They consist of replacing each missing value by a set of imputations drawn from the assumed model and combining these later in a particular way. One of its suggested advantages is that only a small number of imputations (between three and five) are needed in order to obtain relatively efficient estimators. Many *MI* methods have been developed using different models assumptions for continuous, categorical and mixed continuous and categorical data. Statistical software packages often now include *MI* as one of the options and this is making a big contribution to its dissemination. Nonetheless, there is controversy about its actual efficiency and the unverifiable assumptions on the models. Another important point to emphasize is that in many applications the issue of non-response bias is much more crucial than that of variance thus downing the main argument supporting *MI*.

14. Given the variety of procedures and points of view when it comes to addressing the missing data problem, the criteria for evaluating and comparing performance become crucial. A general discussion on these criteria would go beyond the scope of this paper. Instead, the next section summarises the principles for measuring the results of the imputation procedures as an introduction to the method subsequently proposed for categorical data.

## B. The criteria for evaluating the results of imputation methods

15. Although the problem of missing data in datasets is essentially the same, the methods to deal with it are judged significantly differently from the statistical inference and machine learning perspectives.

16. Basic criteria for evaluating statistical procedures had been set by Neyman and Pearson (Neyman and Pearson, 1933; Neyman, 1937). From this Rubin (1976, 1996) developed an extended framework for statistical surveys with non-response that is still in use today. The framework sets the goal that a statistical procedure should make valid and efficient inferences about a population of interest, considering the method for handling missing values as part of the overall procedure. That is, the method should yield unbiased parameter estimates, with the possibility of assessing the degree of uncertainty about its estimation even though non-response has happened. The meaning of this paradigm for evaluation of the missing data treatments can be summarised in the following paragraphs:

17. *"… I believe that statistically valid must be a frequency concept, averaging over randomization distributions generated by known sampling mechanisms (used to collect data) and posited distributions for the response mechanisms (the processes underlying non-response)". (Rubin,1996).*

18. *"…Judging the quality of missing data procedures by their ability to recreate the individual missing values (according to hit-rate, mean square error, etc.) does not lead to choosing procedures that result in valid inference, which is our objective". (Rubin,1996).*

19. *"… We can essentially never be sure that the data base constructor's model is appropriate, but assuming it is, and assuming that the ultimate user is performing an analysis that would be valid if there were no missing data, we can expect that the ultimate user will obtain a valid inference". (Rubin,1996).*

20. These ideas must not be taken as mathematical results, but more as practical guidelines. The important point to stress is that the goodness of a missing data treatment – following the statistical inference perspective – must be assessed for each of the features (bias, mean square error, etc.) of the estimator obtained. And also that these features are linked to the assumed data model, whose assumptions are usually untestable. In other words, the correctness of a missing data treatment is predicated on the correctness of the model specifications.

21. The machine learning perspective has not an evaluation criterion for non-response imputation methods on its own but looks at the problem from the general artificial intelligence framework. It does not try to build a theoretical model able to describe the behaviour of the variables, but has the more modest aim of fitting as close as possible to this behaviour, not involving modelling assumptions.

Performance is typically evaluated by its empirical results through simulating missing data and then measuring the closeness between the real and the imputed data.

22. The next section gives details of the methods and recommendations specifically for imputation of missing categorical data from the statistical inference perspective.

## C. The case of imputation of missing categorical data

23. In statistical surveys *MI* has become the most widely used method of imputation. Like other model-based procedures, it requires specification of an imputation model. Usually the choice depends on the type of variables in the input dataset. For continuous variables, the most widely accepted is the multivariate normal model. Schafer (1997) showed that small deviations from the normality assumption can produce rather robust inferences.

24. In the case of categorical variables, the set of cells in the contingency table has a multinomial distribution, with the log-linear model being the most appropriated for imputation. These models prove flexible to specify variable dependence but can cause practical problems because of non-unique or invalid estimates (Molenberghs et al., 1999) and can be applied only when the number of variables is not very big, that is, when the full multi-way cross-tabulation required for the log-linear analysis can be computationally processed.

25. When there are mixed continuous and categorical data, a combination of a log-linear and multivariate normal model would be an appropriate option. Another possibility is to use logistic regression models, drawing the *MI* imputations from the estimated posterior probability distribution. But, sometimes, practical problems are found here too, at the regression estimation step (Santner and Duffy, 1986). And most of these procedures are also able to deal with only a small number of variables. Because of all the above-mentioned difficulties, different solutions have been studied and proposed.

26. Some studies recommend approximating the imputation model by a continuous variables model. For example, for the binary case, Rubin and Schenker (1986) proposed using a *Gaussian* distribution to obtain the imputations, and then converting them later to dichotomous by using a cut-off point based on the maximum likelihood estimate of the *Bernouilli* probability, whereas Schafer (1997) proposed using a normal distribution, rounding to the closest integer.

27. Their simplicity together with the non-existence of any alternatives have led to the above-mentioned methods also being used when dealing with non-binary categorical variables. Yucel and Zaslavsky (2003) made practical recommendations on rounding and Van Ginkel et al. (2007) obtained reliable results using *MI* with the multivariate normal model and ordinal non-dichotomous data, rounding the non-integer imputations to the nearest feasible integer.

28. There have been some criticisms of this way of proceeding. For example, Horton et al. (2003) warned about the possible bias when rounding and Ake (2005) estimated the bias using normal distribution and rounding non-binary categorical data. On the other hand, Allison (2006) and Demirtas (2008) compared, by simulation studies, the results of *MI* with ordinal non-binary data using multivariate normal approximation and rounding on the one hand, with the results of using other options such as logistic regressions and the log-linear model on the other, and concluded that the former should be used only in exceptional cases.

29. Apart from the criticisms expressed from the practical perspective, a contradiction seems to exist between the theoretical framework underlying the statistical approach and the recommendations to use a normal distribution to approximate categorical data models. That is, as the assessment of the missing data treatment is based on its features when making statistical inferences, a model for approximating data that is clearly inadequate cannot be evaluated positively. This tendency to apply to categorical data procedures originally designed for continuous data is a common practice in statistical inference, for example, in sampling finite populations, where the categorical is frequently the trivial and easiest case. But this is not necessarily the same for probability models and estimation, due to the different topological features of the continuous and categorical spaces.

30. The difficulties become even more relevant when dealing with non-binary and non-ordinal categorical data, where the distributions are frequently oddly shaped or far from symmetry. This led us to seek other imputation methods, namely from the area of machine learning.

# III. The imputation problem to be solved and the solutions proposed

## A. Context

31. Surveys consisting of questions about individual opinions or attitudes are known as opinion polls. These have proven to be an especially fast and easy-to-use means of gathering information about a population because they normally simplify the most technical phases of the survey process. As in most surveys, there is usually total and/or partial non-response in polls. Total non-response is commonly addressed at the sampling design stage. This paper focuses on partial non-response.

32. The typical way of dealing with partial non-response in polls is to add "*Don't know/Not applicable*" and treat it like any other category. But this is not highly recommendable because it can produce other problems at the analysis of results stage. Nevertheless, it is widely applied in polls due to its straightforwardness.

33. In election polls, though, there is one variable – *which political party do you intend to vote for in the next general elections? (*referred to as *voting intention* from now on*)* – for which the procedure outlined above is not good enough, and missing values are imputed using other methods.

34. To evaluate the imputation procedures, the microdata file from a 2008 Spanish general election poll (number 2750 in the Sociological Research Centre's survey catalogue[1]) was chosen. Apart from the *voting intention* question, the poll contains others giving rise to different types of variables:

- Quantitative variables: questions answered by entering a numerical value. They include questions on *ideological self-location* (asking respondents to place themselves ideologically on a scale from *1* to *10*, with *1* being the extreme left and *10* the extreme right). Others focus on the *rating of three specific political figures*, *likelihood to vote*, and *likelihood to vote for three specific political parties*, all of which are rated on a scale from *0* to *10*.
- Ordered categorical variables: questions answered by entering categories that are so well ordered that they are easy and straightforward to convert into quantitative variables. They refer to government and opposition party ratings. The answer categories are "*very good*", "*good*", "*fair*", "*bad*" and "*very bad*", which we convert into the values, *1, 0.75, 0.5, 0.25* and *0*, respectively, assuming that they are ordered equidistantly.
- Categorical variables with non-ordered categories: questions including *voting intention* and similar points, such as *voting memory* (party the respondent voted for at the last general elections); the *autonomous community*; *which of the likely candidates the respondent would prefer to see as president of the government*; *how sure/definite the respondents' voting intention is*; *the political party the respondent tips to win*, and *the political party the respondent would prefer to see win*.

35. Although non-response is found in all these variables, this paper focuses on imputation of the *voting intention* variable, using for the tests the *13.280* interviews in the file with an answer to this question and also no missing values for the rest of the variables. The *voting intention* is a qualitative variable in which *11* categories were considered, including the names of each of the biggest political parties, "*blank vote*", "*abstention*" and a category called "*others*".

## B. New neural networks classifier for imputing categorical data

36. Neural networks seem a promising approach that has previously been used for imputation. Duliba (1991), Gupta and Lam (1996), Abdella and Marwala (2005), Nelwamondo et al. (2007), and Ssali and Marwala (2007) have all used neural networks to impute continuous data. Nordbotten (1998) presented an interesting study of neural networks imputation of some population census variables for individuals

[1] http://www.cis.es/cis/opencms/EN/1_encuestas/catalogoencuestas.html

not included in the sample, while Amer (2006) also used neural networks for imputing in surveys with complex designs, integrating the sampling weights.

37. Most of the previous studies used the multilayer perceptron with back-propagation learning algorithm, although this neural networks architecture is considered as sub-optimal and non-efficient (Vellido et al. 1999). Another neural networks structure – fuzzy min-max neural networks – is tested in this paper. Neuro-fuzzy computing is one of the most popular hybridisations in the artificial intelligence literature because it offers the generic benefits of neural networks – like massive parallelism and robustness – and, at the same time, uses fuzzy logic to model vague or qualitative knowledge and convey uncertainty.

38. The fuzzy min-max neural network classifier is a supervised learning method. The original model was developed by Simpson (1992). It is a classification method that separates the joint input variables space into classes of any size and shape with nonlinear boundaries. Its structure is a three-layer feed-forward network with one hidden node, where the activation functions are fuzzy sets membership functions and the learning consists of incrementally adjusting the number and volume of the fuzzy sets in a neural network framework. All input variables in the network are required to correspond to continuously valued variables, which can be a significant constraint in many real-world situations.

39. A new procedure that extends the input to categorical variables by introducing new fuzzy sets, a new operation and a new architecture has been developed recently (Rey del Castillo and Cardeñosa, 2011). This method may be used to impute the *voting intention* from the rest of the variables in the dataset proposed. For the purposes of imputation, each class or classification category is matched with one of the different values which the variable to be imputed takes. Hence, the imputed value is the category of classification, corresponding in this case to the class with the highest degree of membership.

### C. Bayesian networks classifier for imputing categorical data

40. Different Bayesian networks structures have been proposed before to impute missing data in classification problems (García and Hruschka, 2005; Hruschka et al. 2007) and also in the field of official statistics (Thibaudeau and Winkler, 2002; Di Zio et al., 2004).

41. For the purpose of imputing the *voting intention* variable, a Bayesian network using the *17* available variables is constructed. To improve the results, the network structure, that is, the structure of dependences between the nodes or variables, is learned in two steps. The first can be considered an optimisation problem, where a quality measure of a network structure given the training data must be maximised. As several search algorithms and several metrics are available, we use different combinations of search algorithms and metrics and, for each of these combinations, select the structure with the best score. From the selected structures, in the second step the one which provides the best imputation results is finally chosen. The network parameters are then estimated (or "learned", in machine learning jargon) by maximum likelihood.

42. The Bayesian network learned is later used as a classifier for imputation, matching each class or classification category with one of the *voting intention* categories.

## IV. Comparison of the imputation results

43. The aim now is to compare the results of the imputations when using the two proposed classifiers with the results of a classical method of imputation from the statistical inference perspective. The same dataset with the same imputation variable, *voting intention*, and the same *10* numeric and *6* categorical variables will be used in all the experiments.

44. The performance of the methods will be assessed following a practical evaluation criterion frequently used in supervised classification procedures: the correctly imputed rate, that is, the percentage of imputed values that exactly match the original data over the inputs where *voting intention* is not missing. A ten-fold cross-validation, randomly partitioning the test data (*13.280* interviews) into *10* parts (folds) is performed. A single fold is retained as the validation data for testing the model, whereas the

remaining *9* are used as training data. The cross-validation process is then repeated 10 times with each of the *10* folds, and the results are averaged to produce a single estimation. The advantage of this method is that all the observations are used for both training and validation and each observation is used for validation exactly once, while providing non-biased estimations of the correctly imputed rate.

45. The classical imputation method taken as the baseline for comparison must be appropriate to deal with numeric and categorical variables as inputs. Use of the multivariate normal as a proxy distribution lacks statistical underpinning in this case, where the number of categories of the *voting intention* non-ordered variable is really big (*11*). Consequently, *MI* using logistic regression on the *16* remaining variables – producing *5* imputations for each missing value – was selected. This procedure constitutes the state of the art in imputation of opinion polls. As the correctly imputed rate is not defined for *MI*, we calculated it in the logical way, as the average of the correctly imputed rate for each of the *5* complete datasets, where each individual imputation is obtained from the remaining *9* folds. The result is a *correctly imputed rate* of *66.0%*, generated using SAS/STATS software, version 9.2 of the SAS System for Windows, Copyright 2002-2008 by SAS Institute Inc., Cary, NC, USA.

46. The new fuzzy min-max neural networks classifier extended to accept categorical variables as inputs has a number of parameters to be tuned at the learning stage, selecting the values which provide the best performance. The result of the cross-validation performed over the same *10* folds and for the set of parameters with the best score is a *correctly imputed rate* of *86.1%*, obtained using a Fortran program with an IBM XL Fortran Enterprise Edition v10.1 for Aix compiler, specially adapted to parallel computing. One weakness of the kind of learning used is that the learning set order could have an impact on the results. The process was therefore repeated several times with a number of different randomisations of the input dataset. The resulting rates were similar, thereby confirming the robustness of the method.

47. The figures for the Bayesian networks cross-validation over the same *10* folds were generated using the Weka data mining software (Hall et al., 2009). This is an open source project that contains a collection of machine learning algorithms and tools that make it easy to test and compare different Bayesian networks learning procedures. The result for the network with the best performance – obtained using a Hill-Climber search algorithm and an AIC (Akaike Information Criterion) metric – is a *correctly imputed rate* of *87.4%*.

# V. Final remarks

From the discussion and the results set out above, a number of general comments may be made:

- The results reported here apply to one specific opinion poll. However, it must be emphasised that differences of a similar order of magnitude – about *20* to *22* percentage points – between the machine learning classifiers on one side, and the logistic regression *MI* on the other, have been found whenever the three methods have been compared using datasets from other opinion polls. That is, assuming that the correctly imputed rate is a valid evaluation criterion for comparing the methods for imputation of categorical data, the machine learning classifiers clearly outperform classical statistical inference procedures such as logistic regression *MI*.
- This paper presents the simplest case where there are missing data exclusively on one of the variables. Further studies are needed to extend the results to more real situations with non-response on most of the variables.
- Machine learning procedures seem easier to automate. Apart from non-dependence on model assumptions, other plausible reasons should be further tested:
    - The step of selection of the auxiliary variables – something that needs to be done with care in statistical inference – is not required because the procedures fit appropriately to the existing variables, not breaking down when the number is large.
    - The presence of outliers, especially dangerous with numeric variables, has less impact on the results, that is, the procedures are more robust.

- If the above assertions are true, the machine learning imputation methods may be used for massive imputation tasks. This could be a solution to problems such as integration of data from different surveys, small area estimation and others that must be addressed in the domain of statistical surveys and which are difficult to solve using *deletion* or *tolerance procedures* for dealing with non-response.
- The previous comment follows from the study of imputation of categorical variables, but it is possible that similar conclusions may be drawn on numeric variables.
- An interesting task for the near future could be to test the hypotheses set out above.

# VI. References


Abdella, M. and Marwala, T. (2005), *Treatment of Missing Data Using Neural Networks and Genetic Algorithms*, in Proceedings of the International Joint conference on Neural Networks, Montreal, Canada, July 31- August 4 2005.

Ake, C. F. (2005), *Rounding After Multiple Imputation with Non-Binary Categorical Covariates*, SAS Conference Proceedings: SAS User Group International 30, Philadelphia, PA, April 2005.

Allison, P. (2006), *Multiple Imputation of Categorical Variables under the Multivariate Normal Model*, paper presented at the Annual Meeting of the American Sociological Association, Montreal Convention Center, Montreal, Quebec, Canada, August 2006.

Amer, R. F. (2006), *Neural Network Imputation in Complex Survey Design*, in Proceedings of World Academy of Science, Engineering and Technology, vol. 12, pp. 76-81, March 2006.

Demirtas, H. (2008), *On Imputing Continuous Data When the Eventual Interest Pertains to Ordinalized Outcomes Via Threshold Concept*, Computational Statistics and Data Analysis, vol. 52, pp. 2261-2271.

Di Zio, M., Scanu, M., Coppola, L., Luzi, O. and Ponti, A. (2004), *Bayesian Networks for Imputation*, Journal of the Royal Statistical Society, Section A, vol. 167, part 2, pp. 309-322.

Duliba, K. A. (1991), *Contrasting Neural Nets with Regression in Predicting Performance in the Transportation Industry*, in Proceedings of the Twenty- Fourth Annual International Conference on System Sciences, vol. 4, pp. 163-170, January 1991.

García, J. T. A. and Hruschka, E. R. (2005), *Naïve Bayes as an Imputation Tool for Classification Problems*, in Proceedings of the Fifth International Conference on Hybrid Intelligent Systems, Rio de Janeiro, vol. 1, pp. 497-499, Los Alamitos, CA, USA: IEEE Computer Society.

Gupta, A., and Lam, M. S. (1996), *Estimating Missing Values Using Neural Networks*, Journal of the Operational Research Society, vol. 47, pp. 229-238.

Hall, M., Frank, E., Holmes, G., Pfahringger, B., Reutemann, P. and Witten, H. (2009), *The Weka Data Mining Software: An Update*, SIGKDD Explorations, vol. 11, no. 1.

Horton, N. J., Lipsitz, S. R. and Parzen, M. (2003), *A Potential for Bias when Rounding in Multiple Imputation*, The American Statistician, vol. 57, no. 4, pp. 229-232, November 2003.

Hruschka, E. R. Jr., Hruschka, E. R. and Ebecken, N. F. (2007), *Bayesian Networks for Imputation in Classification Problems*, Journal of Intelligent Information Systems, vol. 29, Issue 3, pp. 231-252, December 2007.

Little, R. J. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd edition, John Wiley and Sons, New York.

Molenberghs, G., Goetghebeuer, E. J. T., Lipsitz, R., and Kenward, M. G. (1999), *Nonrandom Missingness in Categorical Data: Strengths and Limitations*, The American Statistician, vol. 53, no. 2, pp. 110-118, May 1999.

Nelwamondo, F. V., Mohamed, S. and Marwala, T. (2007), *Missing Data: A Comparison of Neural Network and Expectation Maximization Techniques*, Current Science, vol. 93, no. 11, pp. 1514-1521, December 2007.

Neyman, J. and Pearson, E. S. (1933), *On the Problem of Most Efficient Tests of Statistical Hypotheses*, Philosophical Transactions of the Royal Society of London, Series A, no. 231, pp. 289-337.

Neyman, J. (1937), *Outline of a Theory of Statistical Estimation Based on the Classic Theory of Probability*, Philosophical Transactions of the Royal Society of London, Series A, no. 236, pp. 330-380.

Nordbotten, S. (1998), *Estimating Population Proportions from Imputed Data*, Computational Statistics and Data Analysis, vol. 27, pp. 291-309.

Rey-del-Castillo, P., and Cardeñosa, J. (2012), *Fuzzy Min–Max Neural Networks for Categorical Data: Application to Missing Data Imputation*, Neural Computing and Applications, vol. 21, no. 6 (2012), pp. 1349-1362, DOI 10.1007/s00521- 011-0574-x, Springer-Verlag London.

Rubin, D. B. (1976), *Inference and Missing Data*, Biometrika, no. 63, pp. 581-592.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.

Rubin, D. B. (1996), *Multiple Imputation After 18+ Years*, Journal of the American Statistical Association, vol. 91, no. 434, Applications and Case Studies, June 1996.

Rubin, D. B. and Schenker, N. (1986), *Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse*, Journal of the American Statistical Association, vol. 81, no. 394, Survey Research Methods, June 1986.

Särndal, C. E. (2007), *The Calibration Approach in Survey Theory and Practice*, Survey Methodology, vol. 33, No. 2, pp. 99-119, December 2007.

Santner, T. J. and Duffy, D. E. (1986), *A Note on A. Albert and J. A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models*, Biometrika, vol. 73, pp. 755-758.

Schafer, J. L. and Graham, J. W. (2002), *Missing Data: Our View of the State of the Art*, Psychological Methods, vol. 7, no. 2, pp. 147-177.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London, 1997.

Simpson P. K. (1992), *Fuzzy Min-Max Neural Networks- Part 1: Classification*, IEEE Transactions on Neural Networks, vol. 3, pp. 776-786.

Song, Q. and Shepperd, M. (2007), *Missing Data Imputation Techniques*, International Journal of Business Intelligence and Data Mining, vol. 2, no. 3, pp. 261-291.

Ssali, G., and Marwala, T. (2007), *Estimation of Missing Data Using Computational Intelligence and Decision Trees*, arXiv: 0709.1640v1, URL http://arxiv.org/abs/0709.1640 .

Thibaudeau Y., Winkler W.E. (2002). *Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints*. Research Report Series, Statistics, n. 2002-09, U.S. Bureau of the Census.

Van Ginkel, J. R., Van der Ark, L. A. and Sijtsma, K. (2007), *Multiple Imputation of Item Scores when Test Data are Factorially Complex*, British Journal of Mathematics and Statistical Psychology, vol. 60, pp. 315-337.

Vellido, A., Lisboa, P.J.G., and Vaughan, J. (1999), *Neural Networks in Business: a Survey of Applications (1992-1998)*, Expert Systems with Applications, vol. 17, pp. 51-70.

Yucel, R. M. and Zaslavsky, A. M. (2003), *Practical Suggestions on Rounding in Multiple Imputation*, Proceedings of the Joint American Statistical Association Meeting, Section on Survey Research Methods, Toronto, Canada, August 2003.