Journal of Biometrics & Biostatistics

Schmitt et al., J Biomet Biostat 2015, 6:1
DOI: 10.472/2155-6180.1000224

Research Article Open Access

A Comparison of Six Methods for Missing Data Imputation

Peter Schmitt*, Jonas Mandel and Mickael Guedj

Department of Bioinformatics and Biostatistics, Pharnext, Paris, France

Abstract

Missing data are part of almost all research and introduce an element of ambiguity into data analysis. It follows that we need to consider them appropriately in order to provide an efficient and valid analysis. In the present study, we compare 6 different imputation methods: Mean, K-nearest neighbors (KNN), fuzzy K-means (FKM), singular value decomposition (SVD), bayesian principal component analysis (bPCA) and multiple imputations by chained equations (MICE). Comparison was performed on four real datasets of various sizes (from 4 to 65 variables), under a missing completely at random (MCAR) assumption, and based on four evaluation criteria: Root mean squared error (RMSE), unsupervised classification error (UCE), supervised classification error (SCE) and execution time. Our results suggest that bPCA and FKM are two imputation methods of interest which deserve further consideration in practice.

Keywords: Missing data; Imputation methods; Comparison study; Missing completely at random; bPCA

Introduction

Missing data are a common problem in most scientific research domains such as Biology [1], Medicine [2] or Climatic Science [3]. They can arise from different sources such as mishandling of samples, low signal-to-noise ratio, measurement error, non-response or deleted aberrant value. Rubin [4] defined missing data based on three missingness mechanisms [5]: data are missing completely at random (MCAR) when the probability of an instance (case) having a missing value for a variable does not depend on either the known values or the missing data; data are missing at random (MAR) when the probability of an instance having a missing value for a variable may depend on the known values but not on the value of the missing data itself; data are missing not at random (MNAR) when the probability of an instance having a missing value for a variable could depend on the value of that variable.

Missing data introduce an element of ambiguity into data analysis. They can affect properties of statistical estimators such as means, variances or percentages, resulting in a loss of power and misleading conclusions. A variety of techniques have been proposed for substituting missing values with statistical prediction, this process is generally referred to as 'missing data imputation' [5-7]. Most published articles in this field deal with the development of new imputation methods, however few studies report a global evaluation of existing methods in order to provide guidelines to make the more appropriate methodological choice in practice [8-10].

In the present study, we compare 6 different imputation methods: Mean, K-nearest neighbors (KNN) [1], fuzzy K-means (FKM) [11], singular value decomposition (SVD) [1], bayesian principal component analysis (bPCA) [12] and multiple imputations by chained equations (MICE) [6]. Comparison was performed on four real datasets of various sizes (small: variable numbers lower than 10 and large datasets: variable numbers greater than 10), under an MCAR assumption, and based on four evaluation criteria: Root mean squared error (RMSE), unsupervised classification error (UCE), supervised classification error (SCE) and execution time.

Methods

Imputation methods

Six imputation methods (described in supplementary methods) were selected in order to cover techniques broadly applied in the literature and representative of various statistical strategies.

Briefly, three of the six methods are based on imputation by the mean: Mean consists of replacing the missing data for a given variable by the mean of all known values of that variable; KNN defines for each sample or individual a set of K-nearest neighbors and then replaces the missing data for a given variable by averaging (non-missing) values of its neighbors; FKM is an extension of KNN based on fuzzy K-means clustering. SVD and bPCA are based on eigenvalues. Finally, MICE are an iterative algorithm based on chained equations that uses an imputation model specified separately for each variable and involving the other variables as predictors.

Datasets

Considering the possible variability of relative performances of methods across datasets, results were generated based on four reference datasets split in two groups of various size: small datasets (Iris and *E. coli*) and large datasets (Breast cancer 1 and 2), summarized in Table 1.

The Iris dataset is a very popular dataset introduced by Fisher [13] for an application of discriminant analysis. It provides for three species of iris flowers (setosa, versicolor, and virginica), four variables that are length and width of the sepal and the petal (in cm). For our study, we used the 100 flowers from the two most different species, versicolor and virginica.

The *E. coli* dataset was obtained from UCI machine learning repository [14]. The objective is to predict the cellular localization sites of 129 *E. coli* proteins [15] based on 5 variables.

The Breast cancer 1 dataset represents 80 tumor samples and 65 representative genes [16]. This set of tumor is organised into four molecular subtypes (termed basal, apocrine, luminal and normal-like) and according to the metastatic relapse at five years.

***Corresponding author:** Peter Schmitt, Department of Bioinformatics and Biostatistics, Pharnext, Paris, France, Tel: +33 1 69 47 70 00; E-mail: peter.schmitt@pharnext.com

Received April 24, 2015; **Accepted** May 14, 2015; **Published** May 22, 2015

Citation: Schmitt P, Mandel J, Guedj M (2015) A Comparison of Six Methods for Missing Data Imputation. J Biomet Biostat 6: 224. doi:10.4172/2155-6180.1000224

Datasets	Nb of samples	Nb of variables	Type of variables
Iris	100	4	flowers
E. coli	129	5	proteins
Breast cancer 1	80	65	gene expression
Breast cancer 2	89	60	gene expression

Table 1: Dataset used for imputation methods comparison.

The Breast cancer 2 dataset provides a 70 genes signature for prediction of metastasis-free survival, measured on 89 tumor samples [17]. These 70 genes highlight three grades of tumors: "poorly," "intermediate" and "well" with another risk factor: the metastatic relapse. For the needs of this study we only considered the "poorly" and "well" grades.

Evaluation criteria

Imputation methods were compared based on four measures of performance.

Root mean square error (RMSE) measures the difference between imputed and true values and is the figure of merit employed by most studies. Basically, it represents the sample standard deviation of that difference:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}\left(X_i^{obs}-X_i^{imputed}\right)^2}{n}}$$

Unsupervised classification error (UCE) assesses the preservation of internal structure by measuring how well the clustering of the complete dataset was preserved when clustering the imputed dataset. The approach used for unsupervised classification is Hierarchical Clustering with d=1-Pearson correlation as distance and Ward's aggregation. We defined the unsupervised classification error as:

UCE=% of misclassified samples

Supervised classification error (SCE) assesses the preservation of discriminative or predictive power by measuring the difference between subgroups predicted by supervised classification after missing data imputation and the actual subgroups (the metastatic relapse for Breast cancer 1 and 2). The approach used for supervised classification is linear discriminant analysis (LDA) on a set of variables selected a priori on each reference dataset without missing values. We defined the supervised classification error as:

SCE=1-AUC,

with AUC the area under the ROC curve of the predictive LDA model.

Finally the execution time was also assessed and compared between the six methods.

Principle of the Analysis

Figure 1 shows the general principle of the analysis. From the original datasets (without missing values), we introduced in the data a varying percentage of missing values (from 5% to 45%) generated under an MCAR assumption. These simulated missing values were imputed using the 6 methods and the 4 evaluation criteria (RMSE, UCE, SCE and execution time) were measured. Difference between the replaced values and the original true values was evaluated by RMSE criterion, the influence of the imputed values on the quality of clustering by UCE and SCE criteria (expressed in %), and finally the execution time in minutes. For the strength of this work, we performed 1000 simulations for each original dataset and for percentage of missing values i.e. 20,000 simulations. The results were averaged over the 1000 simulations.

Results

Six different imputation methods were selected in order to cover techniques broadly applied in the literature and representative of various statistical strategies. A simple Google search of each method associated with the term 'missing data' provides an idea of their respective popularity (Figure 2). As expected, Mean was produced the largest number of hits with more 21 000 results, followed by MICE,

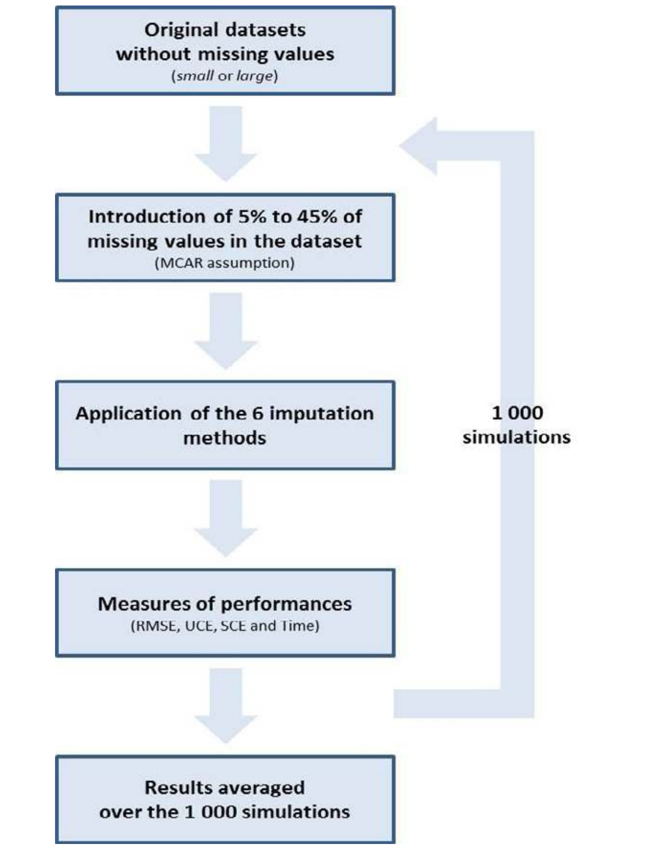
Original datasets
without missing values
(small or large)
Introduction of 5% to 45% of
missing values in the dataset
(MCAR assumption)
Application of the 6 imputation
methods
1 000
simulations
Measures of performances
(RMSE, UCE, SCE and Time)
Results averaged
over the 1 000 simulations

Figure 1: Principle of the analysis.

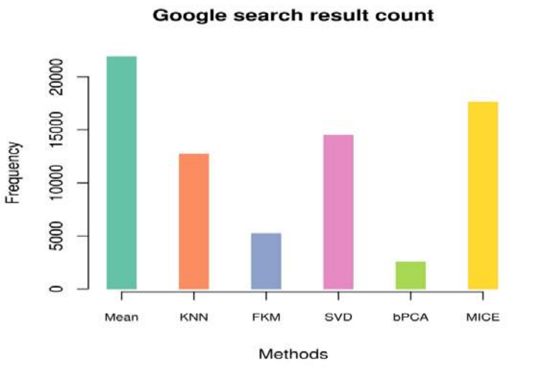

Figure 2: Google search result count. A simple Google search of each method associated to the term 'missing data' provides an idea of the respective popularity of the 6 methods.

SVD and KNN (17 600, 14 500 and 12 700 respectively). FKM and bPCA were found to be less popular with only 5 220 and 2 560 hits respectively. However, popular method doesn't necessarily mean the best method. So, a comparison of these methods was performed using four performance measures: RMSE, UCE, SCE and the execution time. Considering the possible variability of relative performances of methods across datasets, results were generated based on four reference datasets split in two groups of various sizes: small datasets (Iris and *E. coli*) and large datasets (Breast cancer 1 and 2), summarized in Table 1.

Figures 3 and 4 plots the average performances of each method as a function of the percentage of missing values (from 5% to 45% by 10%) for small and large datasets respectively where a low value involves a reliable imputation. As expected, the performances decreased with increasing percentage of missing values in all datasets. According to RMSE, UCE and SCE criteria and taking into account the reproducibility the 4 datasets, Mean was the less effective method when applied to the Breast cancer 1 dataset where the difference with other methods was more pronounced. The behaviors of SVD and MICE were not consistent from one dataset to another. In fact, MICE well performed with the small datasets whereas it was the second worst method (behind Mean) with the large ones. In contrast, the opposite is observed for SVD which performed well with the large datasets whereas its performances are deteriorated when applied on small datasets. KNN consistently stood between the best and the worst methods. Finally, bPCA and FKM consistently lie within the best methods across the different datasets and measures of performances. Specifically, FKM outperforms all other methods when applied to the small datasets based on the UCE and SCE criteria.

Execution time for each method is given in Figure 5. Mean, KNN, SVD and bPCA were all very fast with 0.5 to 10 sec duration following the missing value rate. FKM was slower but still shows a reasonable time of execution except when applied with the large dataset for 45% of missing values (around 25 min), ranging from 1 min to 15 min according to the size of the data and the rate of missing values. The execution time of MICE was related to the size of the dataset especially to the length of variables, very fast on the small dataset (around 5 to 10 sec), it reaches around 30 min on the largest dataset at the highest rate of missing values.

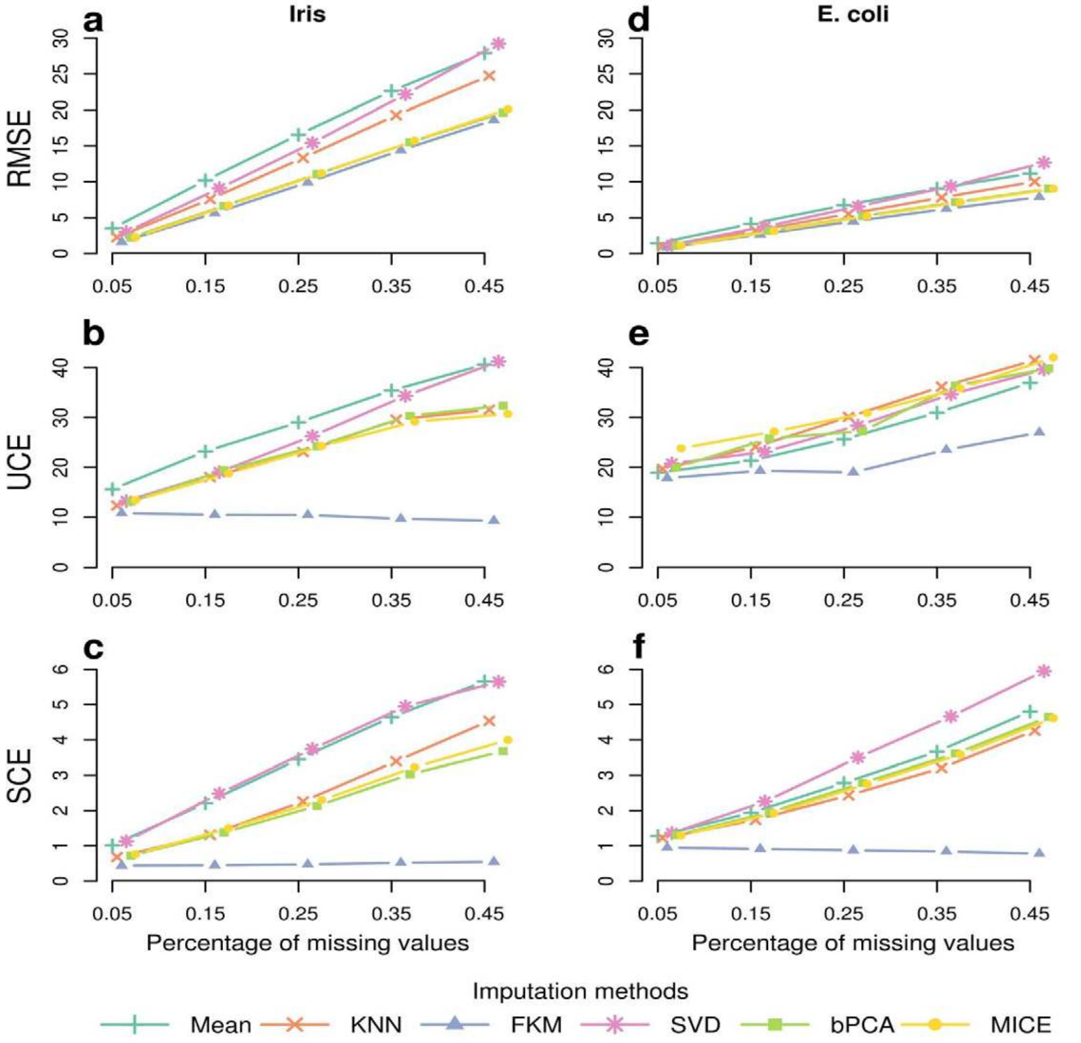

Figure 3: Evolution of RMSE, UCE and SCE with a varying percentage of missing values in the small datasets: Iris (a, b, c) and *E. coli* (d, e, f).

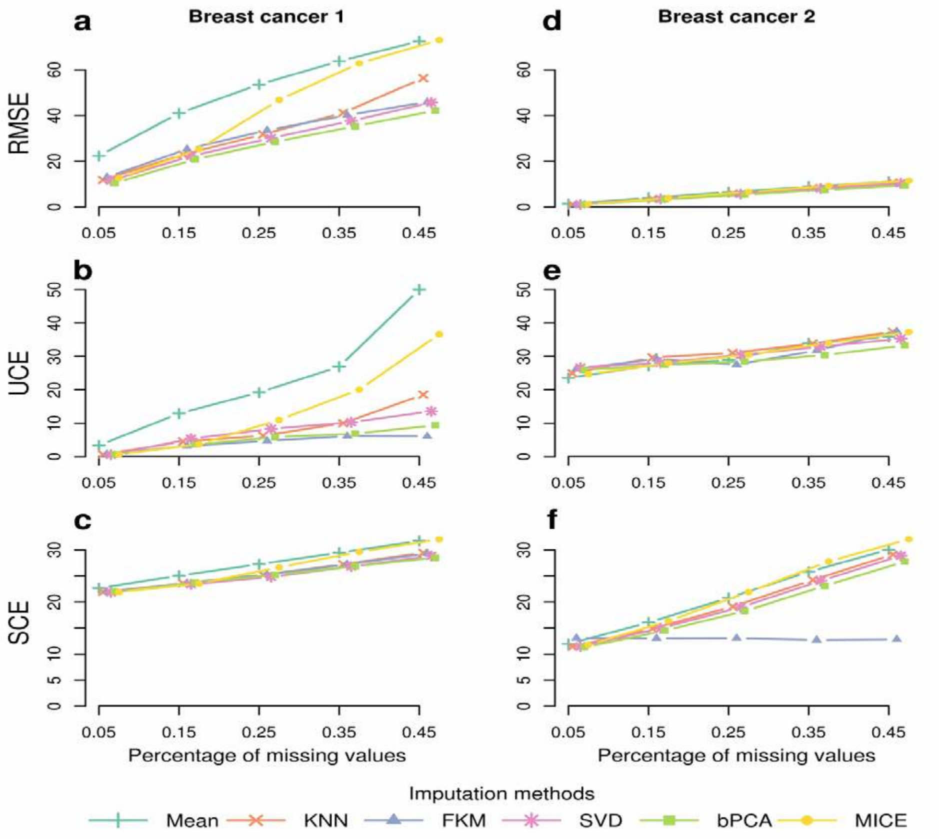

Figure 4: Evolution of RMSE, UCE and SCE with a varying percentage of missing values in the large datasets: Breast cancer 1 (a, b, c) and Breast cancer 2 (d, e, f).

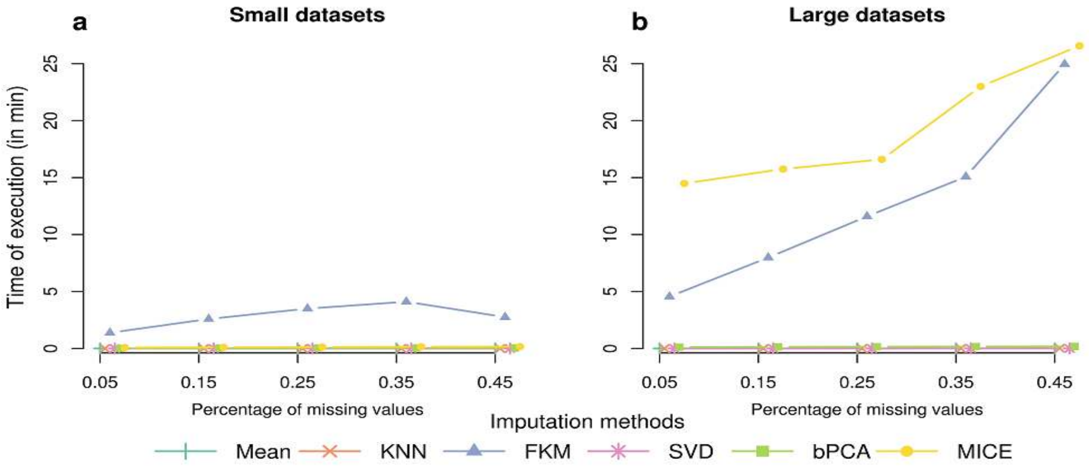

Figure 5: Evolution of execution time with a varying percentage of missing values in the small (a) and large (b) datasets.

Discussion and Conclusion

Missing data are a part of almost all research, and there are several alternative ways to overcome the drawbacks they produced. It was previous observed that neutral and well-designed comparison studies in computational sciences are necessary to ensure that previously proposed methods work as expected in various situations and to establish standards and guidelines [18]. However, only a few studies report an evaluation of existing imputation methods whose Brock et al. [8], Celton et al. [9] and Luengo et al. [10].

In the present study, we performed a neutral comparison of six imputation methods based on four real datasets of various sizes, under an MCAR assumption. Validation of imputation results is an important step and we consequently considered four evaluation criteria: Root mean squared error (RMSE), unsupervised classification error (UCE), supervised classification error (SCE) and execution time. While much attention has been paid to the imputation accuracy measured by RMSE, only a few studies have examined the effect of imputation on high-level analyses such as unsupervised and supervised classification [19, 20], or the time of execution [21].

Overall, results were consistent across the different situations and measures of performance are summarized in Table 2. They first suggest that the most popular methods (Mean, KNN, SVD and MICE) are not necessarily the most efficient, a conclusion also shared by Celton et al. (2010) [9]. It is not surprising for Mean in regards to the simplicity of the methodology: the method does not make use of the underlying correlation structure of the data and thus performs poorly. KNN represents a natural improvement of Mean that exploits the observed data structure. MICE are based on a much more complex algorithm and its behavior appears to be related to the size of the dataset: fast and efficient on the small datasets, its performance decreases and it becomes time-intensive when applied to the large datasets. A second main conclusion is that bPCA and FKM appeared to be the most robust imputation methods in the conditions tested here, with a significant advantage for FKM when applied to the small datasets.

The good results of bPCA were reported in two previous comparison studies by Sun et al. (2009) [22] and Celton et al. [9] where the approach confirmed better performances than Mean and KNN but they didn't compare with FKM. Actually, FKM is rarely used in this field; however, FKM outperformed all the methods considered in the comparison performed by Luengo et al. [10], including Mean, KNN, SVD and bPCA. However, they only considered the quality of imputation based on classification methods without worrying of the execution time that can be an exclude criterion. Consequently, FKM may represent the method of choice but its execution time can be a drag to its use and we consider bPCA as a more adapted solution to high-dimensional data.

Our study has several limitations. The treatment of missing data is a very widespread broad statistical problem and one should consider that there is no universal imputation method performing best in every situations. Our results are limited to data matrices of numerical values, and we did not consider the case of longitudinal or nominal data which would merit to be considered with careful attention [23]. In addition, our intention is also to provide general conclusions independent from the domain of application, and one could certainly further improve the accuracy of imputation methods by integrating specific domain knowledge into the imputation process [24]. Despite these limitations, this study provides a set of coherent observations across different settings.

	Small datasets				Large datasets				Overall
	RMSE	UCE	SCE	Execution time	RMSE	UCE	SCE	Execution time	
Mean	+	++	+	+++	+	+	+	+++	+
KNN	++	++	++	+++	++	+	++	+++	++
FKM	+++	+++	+++	++	++	++	+++	+	+++
SVD	+	++	+	+++	+++	++	+++	+++	++
bPCA	+++	++	++	+++	+++	+++	+++	+++	+++
MICE	+++	++	++	+++	+	+	+	+	++

Table 2: The results of our study based on four evaluation criteria. The number of "+" indicates the performance, from weak (+) to very good one (+++).

In conclusion, bPCA and FKM are two imputation methods of interest. They outperform more popular approaches such as Mean, KNN, SVD or MICE, and hence deserve further consideration in practice.

Author Contributions

PS and MG designed the comparison study. PS implemented the analysis. PS, JM and MG wrote the paper.

Acknowledgments

We thank L Belo, I Chumakov, L Heidsieck and D Cohen for supporting this work.

References

1. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al. (2001) Missing value estimation methods for dna microarrays. Bioinformatics 17: 520-525.
2. Lewis HD (2012) Missing data in clinical trials. New England Journal of Medicine 367: 2557-2558.
3. Schneider T (2001) Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values.
4. Rubin DB (1976) Inference and missing data. Biometrika 63: 581-592.
5. Little RJA, Rubin DB (2002) Statistical Analysis with Missing Data (2ndedn.) Wiley-Interscience.
6. Rubin DB (1987) Multiple Imputation for Nonresponse in Survey. John Wiley and Sons, Inc.
7. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH (2005) Missing-data methods for generalized linear models. Journal of the American Statistical Association 100: 332-346.
8. Brock G, Shaffer J, Blakesley R, Lotz M, Tseng G (2008) Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinformatics 9: 1-12.
9. Celton M, Malpertuy A, Lelandais G, Brevern A (2010) Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. BMC Genomics 11: 1-16.
10. Luengo J, Garca S, Herrera F (2012) On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowledge and Information Systems 32: 77-108.
11. Li D, Deogun J, Spaulding W, Shuart B (2004) Towards missing data imputation: A study of fuzzy k-means clustering method. In: Tsumoto S, Sowiski R, Komorowski J, Grzymaa-Busse J (eds.) Rough Sets and Current Trends in Computing. Springer Berlin Heidelberg 3066: 573-579.
12. Oba S, Sato Ma, Takemasa I, Monden M, Matsubara Ki, et al. (2003) A bayesian missing value estimation method for gene expression profile data. Bioinformatics 19: 2088-2096.
13. Horton P, Nakai K (1996) A probabilistic classification system for predicting the cellular localization sites of proteins. In: Proceeding of the Fourth International Conference on Intelligent Systems for Molecular Biology 4: 109-115.
14. Fisher RA (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics 7: 179- 188
15. Bache K, Lichman M (2003) UCI machine learning repository.
16. Guedj M, Marisa L, De Reynies A, Orsetti B, Schiappa R, et al. (2012). A refined molecular taxonomy of breast cancer. Oncogene 1 :1196-1206.
17. Van De Vijver MJ, He YD, vant Veer LJ, Dai H, Hart AA, et al (2002) A gene-expression signature as a predictor of survival in breast cancer. The New England journal of medicine 347: 1999-2009.
18. Boulesteix AL, Lauer S, Eugster MJA (2013) A plea for neutral comparison studies in computa- tional sciences. PLoS ONE 8: e61562.
19. de Brevern A, Hazout S, Malpertuy A (2004) Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. BMC Bioinformatics 5: 1-12.
20. Wang D, Lv Y, Guo Z, Li X, Li Y, et al. (2006) Effects of replacing the unreliable

cdna microarray measurements on the disease classification based on gene expression profiles and functional modules. Bioinformatics 22: 2883-2889.

21. Saunders JA, Morrow-Howell N, Spitznagel E, Dor P, Proctor EK, et al. (2006) Imputing missing data: A comparison of methods for social work researchers. Social Work Research 30: 19-31.

22. Sun Y, BragaNeto U, Dougherty ER (2009) Impact of missing value imputation on classification for dna microarray gene expression data: A model-based study EURASIP. J Bioinformatics Syst Biol 44:1.

23. Horton NJ, Kleinman KP (2007) Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. The American Statistician 61: 79-90.

24. Liew AWC, Law NF, Yan H (2011) Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in Bioinformatics 12: 498-513.