Missing Data Analysis Using Multiple Imputation in Relation to Parkinson's Disease

Conference Paper · November 2016
DOI: 10.1145/3010089.3010117

5 authors, including:

Rima Houari
Institut National des Sciences Appliquées de Toulouse
9 PUBLICATIONS **20** CITATIONS
SEE PROFILE

Ahcène Bounceur
Université de Bretagne Occidentale
134 PUBLICATIONS **500** CITATIONS
SEE PROFILE

Tahar Kechadi
University College Dublin
341 PUBLICATIONS **2,202** CITATIONS
SEE PROFILE

Abdelkamel A Kamel Tari
Université de Béjaïa
67 PUBLICATIONS **119** CITATIONS
SEE PROFILE

Missing Data Analysis Using Multiple Imputation in Relation to Parkinson's Disease

Rima Houari
LIMED Laboratory, University of Bejaia, Algeria
ri.houari@gmail.com

Ahcène Bounceur
Lab-STICC UMR CNRS 6285
University of Western Brittany, Brest, France
ahcene.bounceur@univ-brest.fr

Tahar Kechadi
UCD,University College Dublin
Belfield, Dublin 4, Ireland
tahar.kechadi@ucd.ie

A-Kamel Tari
LIMED Laboratory, University of Bejaia, Algeria
tarikamel59@gmail.com

Reinhardt Euler
Lab-STICC UMR CNRS 6285
University of Western Brittany, Brest, France
reinhardt.euler@univ-brest.fr

ABSTRACT

Missing data is an omnipresent problem in neurological control diseases, such as Parkinson's Disease. Statistical analyses on the level of Parkinson's Disease may be not accurate, if no adequate method for handling missing data is applied. In order to determine a useful way to treat missing data on Parkinson's stage, we propose a multiple imputation method based on the theory of Copulas in the data pre-processing phase of the data mining process. Our goal to use the theory of Copulas is to estimate the multivariate joint probability distribution without constraints of specific types of marginal distributions of random variables that represent the dimensions of our datasets. To evaluate the proposed approach, we have compared our algorithm with seven state-of-the-art imputation methods such as mean, regression, min, max, K-nearest neighbors, Markov Chain Monte Carlo, Expected Maximization methods, on the basis of six dataset cases containing 5%, 15%, 25% , 35%, 45% and 50% missing data percentages, respectively. The accuracy of each imputation method was evaluated using the Root Mean Square Error (RMSE) formula. Our results indicate that the proposed method outperforms significantly the existing algorithms.

Keywords

Data mining; Data pre-processing; Multi-dimensional Sampling; Copulas; Multiple Imputation; Missing data; Parkinson's Disease.

1. INTRODUCTION AND RELATED WORK

Missing data is unavoidable in health research, but their potential to undermine the validity of research results has often been neglected in the medical literature. This is due to the fact that statistical methods that can resolve problems arising from missing data have, until recently, not been easily available to medical researchers, especially in Parkinson's Disease, where any missing data can potentially cause difficulties in analyses and lead to bias and loss of information in epidemiological and clinical research.

Parkinson's Disease (PD) is the second most common adult-onset neurodegenerative disease after Alzheimer's [1], and it is evaluated to affect more than one million people in North America alone [2]. Moreover, these statistics are expected to grow because the population is getting older. PD is a degenerative disorder of the central nervous system and one of the most common movement disorders affecting people older than 60 years. The symptoms of PD include primary motor symptoms (e.g., resting tremor, rigidity, slow movement, speech problems, swallowing difficulty, etc.), and non-motor symptoms (e.g., pain, depression, etc.). The management of PD typically involves a complete dataset for the development of reliable and objective tools for assessing PD. However, the major challenge when studying PD is that the data is multi-dimensional, and may have a missing data in important parameters. There are several methods for handling these issues, described in a rich literature. According to [3][4], there are three possible strategies to deal with missing data. The first one is based on missing data ignoring techniques that simply omit the cases that contain missing data [5]. The second one represents missing data based modeling techniques, that define a model from available data and inferences based on the distribution of the data [6]. These methods assume a multi-normality of continuous outcome variables. The third one represents missing data imputation techniques [7][8] . These techniques are a strategy for completing a missing data in the dataset with a plausible value which is the estimation of the true value of the missing observation. These methods keep the full sample size, which can be advantageous for bias, precision, and accuracy. Different missing data imputation techniques are used, amongst which we can mention: mean [9], regression [10] , K-nearest neighbors [11], multiple imputation [5] [7],

ACM ISBN .
DOI:

etc.

In this paper, we will focus our attention on the use of missing data multiple imputation methods in the data preprocessing phase of the data mining process. Multiple imputation is a relatively flexible, general purpose approach to deal with missing data, but it needs to be applied carefully to avoid misleading conclusions.

The main goals of this paper are to handle Parkinson's Disease by analyzing the missing data, preserve essential characteristics of this dataset by preserving the relationships among the variables, and finally estimate missing values with the most accurate method and the smallest error.

The paper is organized as follows: the basic concepts are presented in Section 2, and Section 3 describes the proposed method. The experimental results are given in Section 4, and finally, Section 5 concludes the paper.

2. BASIC CONCEPTS

This section aims to introduce the basic concepts used in our approach. Our technique is based on probabilistic and sampling models, therefore, one needs to recall some fundamental concepts. These include the notions of a Probability Density Function (PDF), Cumulative Distribution Function (CDF), dependence and rank correlations of multivariate random variables to measure dependencies of the dimensions. In the following table we give the basic notations used throughout this paper.

Table 1: Primitives and their definitions

Primitive	Definition
X	$n \times m$ data matrix (random variable).
X^i	i^{th} row of the matrix X.
X_j	j^{th} column of the matrix X.
$F_j(.)$	CDF of the j^{th} column.
$f_j(.)$	PDF of the j^{th} column.
C	Gaussian Copula of the matrix X.
c	Density associated with C.
C_{ij}	Empirical Copula of the matrix X.
Σ	Correlation matrix of C.
X^t	Transposed matrix of X.
v_{ij}	Value of the i^{th} row and j^{th} column.

Let f be the Probability Density Function (PDF) of a random variable X. The probability distribution of X consists in calculating the probability $P(X_1 \leq x_1, X_2 \leq x_2, ..., X_m \leq x_m)$, $\forall (X_1, ..., X_m) \in R^m$. It is completely specified by the CDF F which is defined as follows [12]:

$$F(x_1, x_2, ..., x_m) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_m \leq x_m) \quad (1)$$

2.1 Modeling with Copulas

The first usage of Copulas is to provide a convenient way to generate correlated multivariate random variable distributions and to present a solution for the difficulties of transformation of the density estimation problem. Sklar's Theorem [13] showed that there exists a unique m-dimensional Copula C in $[0,1]^m$ with standard uniform marginal distributions $U_1, ..., U_m$. [13] states that every distribution function F with margins $F_1, ..., F_m$ can be written $\forall (X_1, ..., X_m) \in \mathbb{R}^m$ as:

$$F(X_1, ..., X_m) = C(F_1(X_1), ..., F_m(X_m)). \quad (2)$$

Figure 1 shows an overview of a general approach to use Copulas for the proposed approach, that requires the following steps.

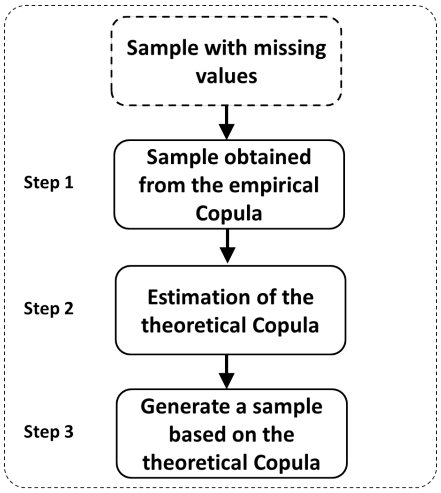

Figure 1: Main steps to use Copulas.

- Step 1: Empirical Copula

  To evaluate the suitability of a selected Copula with estimated parameter and to avoid the introduction of any assumptions on the distribution $F_i(X_i)$, one can utilize an empirical CDF of a marginal $F_i(X_i)$, to transform m samples of X into m samples of U. An empirical Copula is useful for examining the dependence structure of multivariate random vectors. Formally, an empirical Copula is given by the following equation [14]:

$$C_{ij} = \frac{1}{m}(\sum_{k=1}^{m} I_{(v_{kj} \leq v_{ij})}), i = 1, ..., n; \quad j = 1, ..., m. \quad (3)$$

  where the function $I_{(arg)}$ is the indicator function, which equals 1 if arg is true and 0 otherwise. Here, m is used to keep the empirical CDF less than 1, where m is the number of observations.

- Step 2: Theoretical Copula

  In the literature, various Copula families have been proposed. As the most frequently used we can cite the following: Gaussian, Student , and the Archimedean Copulas. In this paper, we will focus on the Copula that results from a standard multivariate Gaussian Copula. The difference between the Gaussian Copula and the joint normal CDF is that the Gaussian Copula allows to have different marginal CDF types from the joint distribution [15]. However, in probability theory and statistics, the multivariate normal distribution is a generalization of the one-dimensional normal distribution. The Gaussian Copula is defined as follows [12]:

$$C(\Phi(x_1),...,\Phi(x_m)) = \frac{1}{\mid \Sigma \mid^{\frac{1}{2}}} \exp(\frac{-1}{2} X^t(\Sigma^{-1} - I)X). \quad (4)$$

where $\Phi(x_i)$ is the CDF standard Gaussian distribution of $f_{i(x_i)}$, i.e., $X_i \sim N(0,1)$, and Σ is the correlation matrix. The resulting Copula $C(u_1,...,u_m)$ is called Gaussian Copula. The density associated with $C(u_1,...,u_m)$ is obtained with the following equation:

$$c(u_1,...,u_m) = \frac{1}{\mid \Sigma \mid^{\frac{1}{2}}} \exp\left[\frac{-1}{2}\xi^t(\Sigma^{-1} - I)\xi\right], \quad (5)$$

where $u_i = \Phi(x_i)$,
and $\xi = (\Phi^{-1}(u_1),....,\Phi^{-1}(u_m))^T$.

- Step 3: Generate a sample based on the theoretical Copula

  To illustrate the problem of invertible transformations of m-dimensional continuous random variables $X_1,...,X_m$ according to their CDF, into m independently uniformly-distributed variables $U_1 = F_1(X_1), U_2 = F_2(X_2),..., U_m = F_m(X_m)$ [12], let $f(x_1,x_2,...,x_m)$ be the probability density function of $X_1,...,X_m$, and let $c(u_1,u_2,...,u_m)$ be the joint probability density function of $U_1,U_2,...,U_m$. In general, the estimation of the probability density function $f(x_1,x_2,...,x_m)$ can provide a nonparametric form (unknown families of distributions).

  In this case, we estimate the probability density function $c(u_1,u_2,...,u_m)$ of $U_1,U_2,...,U_m$ instead of that of $X_1,...,X_m$ to simplify the density estimation problem, and then simulate it to achieve the random samples $X_1,...,X_m$ by using the inverse transformations $X_i = F_i^{-1}(U_i)$ [12].

2.2 Dependence and Rank Correlation

Since the Copula of a multivariate distribution describes its dependence structure, it might be appropriate to use measures of dependence which are Copula-based. The Pearson correlation measures the relationship Σ given by the equation 6.

$$\Sigma_{ij} = \frac{cov(X_i,X_j)}{\sigma_{X_i}\sigma_{X_j}} \quad (6)$$

where $cov(X_i,X_j)$ is the covariance of X_i, $i = 1,...,n$ and X_j, $j = 1,...,m$, and $\sigma_{X_i}, \sigma_{X_j}$ are the standard deviations of X_i and X_j.

Kendall's rank correlation (also known as Kendall's coefficient of concordance) is a non-parametric test that measures the strength of dependence between two random samples $X_p^i, X_{p'}^i$ of n observations with $i = 1,...,n$. The notion of concordance can be defined by the following equation:

$$\tau = P[(X_p^i - X_p^j)(X_{p'}^i - X_{p'}^j) > 0] - P[(X_p^i - X_p^j)(X_{p'}^i - X_{p'}^j) < 0]. \quad (7)$$

For the Gaussian Copula, Kendall's τ can be calculated as follows:

$$\tau = \frac{2}{\pi} \arcsin \Sigma_{X_i X_j}. \quad (8)$$

3. PROPOSED IMPUTATION APPROACH

In this section, we propose a new multiple imputation approach based on the theory of Copulas to address the problem of multivariate missing data. A Copula provides a suitable model of dependencies to compare with well-known multivariate data distributions in order to better distinguish the relationship between the data.

The main goals of this paper are to: (a) handle Parkinson's Disease by analyzing the missing data in the multivariate case, (b) preserve the essential characteristics of the data by preserving the relationships among the variables, (c) model multivariate random variables without imposing constraints to specific types of marginal distributions of data, (e) provide a valid statistical inference, (f) provide highest accuracy (g) estimate missing values with the most effective method.

The main steps of the proposed imputation approach can be described by the following algorithm:

Algorithm 1 Proposed algorithm

Input: Incomplete dataset $X_{(n_1,m)} = \{X_1, X_2, ..., X_m\}$
Output: Complete dataset $\tilde{X}$.
1: Generate a Sample $W_{(n_2,m)}$ based on a theoretical Copula.
2: Determine V which is the subset of rows that contain missing data v_j^k.
3: Determine $Index$ which contains the index of subset of no missing values set in the original data.
4: Determine R the subset of rows in the theoretical sample $W_{(n_2,m)}$ that verify $w_j^i = v_j^k$ by comparing (V,W,$Index$, ξ), where ξ is a relative error.
5: Complete the dataset X.

Here n_1 is the number of rows of the dataset X, n_2 is the number of rows of the dataset W. For more clarification, we consider an example taken from Parkinson's Disease dataset in the following table. We suppose that the dataset contains a missing data in the biomedical test (T_i) denoted by " ? ".

Table 2: Example of a Parkinson's Disease dataset containing missing data

Test	T1	T2	T3	T4	T5	T6
Patient 1	?	0.021	0.221	?	0.277	0.301
Patient 2	0.925	?	0.507	0.511	0.602	?
Patient 3	1.024	1.185	?	1.182	1.504	1.504
Patient 4	1.406	1.823	?	4.385	?	4.640

In this example (Table 2), we want to estimate the missing data $T3$ and $T5$ for patient 4. The entries of the row in question are: 1.406, 1.823, " ? ", 4.385, " ? ", 4.640.

To impute the missing data " ? ", we generate the appropriate Copula having the same parameters as the empirical sample.

By applying the step 4 of the algorithm 1, we obtain the Table 3 which is a collection of rows of a theoretical sample $W_{(7,6)}$ obtained by comparing the known values: 1.406, 1.823, 4.385, 4.640 of patient 4. The red boxes represent the estimated values. For each column of Table 3, we compute the mean, and we impute the results obtained in the PD dataset according to the index of each values. The complete

Table 3: Collection of rows of the theoretical sample obtained by comparing the known values of patient 4.

1	1.406	1.823	0.521	4.385	1.542	4.640
2	1.406	1.823	0.531	4.385	1.541	4.640
3	1.406	1.823	0.512	4.385	1.561	4.640
4	1.406	1.823	0.562	4.385	1.541	4.640
5	1.406	1.823	0.514	4.385	1.520	4.640
6	1.406	1.823	0.561	4.385	1.532	4.640
7	1.406	1.823	0.546	4.385	1.531	4.640

data for the patient number 4 are then: 1.406, 1.823, " 0.535 ", 4.385, " 1.538 ", 4.640.

4. EXPERIMENTAL RESULTS

Our experiments were performed on Parkinson's Disease real-world datasets taken from the machine learning repository [16], which are from the Healthcare database.

Parkinson's Disease.

The dataset was created by Athanasios Tsanas and Max Little of the University of Oxford, in collaboration with 10 medical centers in the US and Intel Corporation who developed the telemonitoring device to record the speech signals. This dataset was accumulated employing the Intel AHTD which is a telemonitoring system designed to facilitate remote and internet-enabled measurements of a variety of PD-related motor impairment symptoms. The dataset is collected at the 42 patient's home, with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring transmitted over the Internet, and processed appropriately in the clinic to predict the Unified Parkinson Disease Rating Scale score (UPDRS). Columns in this dataset represent 16 biomedical UPDRS scores. The information about these attributes includes: clinician's motor UPDRS score (Jitter(%), Jitter(Abs), Jitter:RAP, Jitter:PPQ5, Jitter:DDP); several measures of variation in fundamental frequency (Shimmer, Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11, Shimmer:DDA); several measures of variation in amplitude (NHR,HNR); two measures of ratio of noise to tonal components in the voice (RPDE); a nonlinear dynamical complexity measure (DFA); signal fractal scaling exponent (PPE). Each row corresponds to one of 5875 voice recordings from these individuals.

Simulation Study.

To evaluate our proposed approach, we have implemented and performed extensive simulation experiments with seven imputation methods: mean, regression, min, max, K-nearest neighbors, Markov Chain Monte Carlo, Expected Maximization methods, with six dataset cases containing 5%, 15%, 25%, 35%, 45% and 50% missing data percentages, respectively. The accuracy of each imputation method can be evaluated from its value of RMSE.

In this section, we first describe our simulator and then present our experimental results and discussions. The details of general simulation parameters are depicted in Table 4.

The accuracy is defined as the overall distance between estimated values $\tilde{X}$ and the true value X [17], as shown in the equation 9. RMSE tends to be dominated by outlying estimates far away from the true value. A good estimator should be accurate, so that its estimates are as close to the true value as possible.

Table 4: Simulation parameters

Method	**Parameters**
Copula	ξ= 0.01, Σ_{ij}
K-nearest neighbors	K = 10
MCMC	Iterations=1500
EM	Iterations=30 Stagnation tolerance=0.0001 Regression type=multiple regression

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \tilde{X}_i)^2}{n}} \tag{9}$$

where n = missing values percentage $\times$ the length of data.

By implementing different imputation techniques: mean, regression, min, max, K-nearest neighbors, Markov Chain Monte Carlo, Expected Maximization methods and the proposed approach, Table 5 shows the numerical results obtained for the missing values

Table 5: Comparison of Missing data accuracy

Methods	**5%**	**15%**	**25%**	**35%**	**45%**	**50%**
PA	0.019	0.021	0.221	0.231	0.277	0.301
EM	0.925	0.525	0.507	0.511	0.602	0.609
Regression	1.024	1.185	1.453	1.182	1.504	1.504
MCMC	1.406	1.823	1.949	4.385	4.749	4.640
K-nn	0.452	0.495	0.583	0.670	0.703	1.703
Mean	0.905	1.056	1.023	1.031	1.026	1.982
Max	0.578	7.634	7.777	7.133	7.601	7.353
Min	1.206	5.246	5.066	5.051	4.847	4.902

4.0.1 *Interpretation of the results*

According to Table 5 and Figure 2, we notice that the increase of missing values by 5 % to 50 % causes a decrease in accuracy for different methods.

Max imputation provides the lowest accuracy for most of the databases (15 % to 50 %), but it works well in the case of 5 % of missing data. The minimum error values achieve 0.019 for the Proposed Approach (PA), 0.507 for the EM method, 1.504 for regression imputation, 4.640 for the MCMC method, 1.703 for K-nearest neighbors method, 1.982 for the Mean imputation, 7.353 for the Max, and 4.902 for the Min method. On the other hand, the maximum error values achieve 0.301 for the Proposed Approach, 0.925 for the EM method, 1.024 for regression imputation, 1.406 for the MCMC method, 0.452 for K-nearest neighbors method, 0.905 for the Mean imputation, 0.578 for the Max, and 1.206 for the Min method. We can observe that EM and K-nearest neighbor approaches work well and have better results than Mean, MCMC, Min, and Max methods, respectively. The results obtained by the proposed approach are much better than EM, regression, MCMC, K-nearest neighbor, Mean,

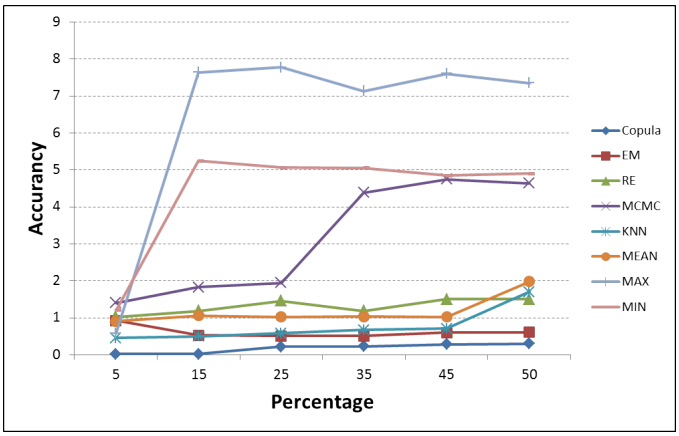

Figure 2: RMSE from eight imputation methods (mean, regression, min, max, K-nearest neighbors (K-nn), Markov Chain Monte Carlo, Expected Maximization (EM) methods and the proposed approach) on the basis of six dataset cases containing missing data from Parkinson's Disease dataset.

Max, and Min imputation for all datasets containing 5%, 15%, 25%, 35%, 45% and 50% missing data percentages, respectively.

5. CONCLUSION

In this paper, we have proposed a multiple imputation method based on sampling techniques to handle missing data from Parkinson's Disease. We have evaluated the proposed approach by comparing our algorithm with seven imputation methods such as, mean, regression, min, max, K-nearest neighbors, Markov Chain Monte Carlo, Expected Maximization methods, on the basis of six dataset cases containing 5%, 15%, 25%, 35%, 45% and 50% missing data percentages, respectively. The accuracy of each imputation method was evaluated using RMSE. The results obtained by the proposed approach are much better than EM, regression, MCMC, K-nearest neighbors, Mean, Max, and Min imputation for all dataset cases containing missing data percentages. In future work, we will compare the efficiency of the proposed method with respect to the use of classification methods.

References

[1] MC De Rijk. Prevalence of parkinson's disease in Europe: A collaborative study of population-based cohorts. *Neurology*, 54(5):s214323, 2000.

[2] Anthony E Lang and Andres M Lozano. Parkinson's disease. *New England Journal of Medicine*, 339(15):1044–1053, 1998.

[3] Julián Luengo, Salvador García, and Francisco Herrera. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and information systems*, 32(1):77–108, 2012.

[4] Rima Houari, Ahcéne Bounceur, Tahar Kechadi, Tari Abdelkamel, and Reinhardt Euler. A new method for estimation of missing data based on sampling methods for data mining. In *Advances in Computational Science, Engineering and Information Technology*, pages 89–100. Springer, 2013.

[5] Bhekisipho Twala. An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, 23(5):373–405, 2009.

[6] Pedro J García-Laencina, Pedro Henriques Abreu, Miguel Henriques Abreu, and Noémia Afonoso. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in biology and medicine*, 59:125–133, 2015.

[7] Wei Ding and Peter X-K Song. Em algorithm in gaussian copula with missing data. *Computational Statistics & Data Analysis*, 101:1–11, 2016.

[8] Cao Truong Tran, Mengjie Zhang, and Peter Andreae. A genetic programming-based imputation method for classification with missing data. In *European Conference on Genetic Programming*, pages 149–163. Springer, 2016.

[9] Ton J Cleophas and Aeilko H Zwinderman. Missing data imputation. In *Clinical Data Analysis on a Pocket Calculator*, pages 93–97. Springer, 2016.

[10] Soeun Kim, Catherine A Sugar, and Thomas R Belin. Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in medicine*, 34(11):1876–1888, 2015.

[11] Rima Houari, Ahcène Bounceur, A Kamel Tari, and M Tahar Kechadi. Handling missing data problems with sampling methods. In *Advanced Networking Distributed Systems and Applications (INDS), 2014 International Conference on*, pages 99–104. IEEE, 2014.

[12] Rima Houari, Ahcène Bounceur, M-Tahar Kechadi, A-Kamel Tari, and Reinhardt Euler. Dimensionality reduction in data mining: A copula approach. *Expert Systems with Applications*, 64:247–260, 2016.

[13] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, second edition, 2007.

[14] R. Houari, A. Bounceur, and T. Kechadi. A new approach for preprocessing of large multi-dimensional data using sampling methods. *Colloque sur l'Optimisation et les Systèmes d'Information COSI, Algeria*, 2013.

[15] Rima Houari, Ahcéne Bounceur, and M-Tahar Kechadi. A new method for dimensionality reduction of multi-dimensional data using copulas. In *Programming and Systems (ISPS), 2013 11th International Symposium on*, pages 40–46. IEEE, 2013.

[16] M. Lichman. Uci machine learning repository, university of california, irvine, school of information and computer sciences, http://archive.ics.uci.edu/ml, 2013.

[17] Bruno A Walther and Joslin L Moore. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, 28(6):815–829, 2005.