AJAS / SPAS

African Journal of Applied Statistics
Vol. 5 (1), 2018, pages 323–336.
DOI: http://dx.doi.org/10.16929/ajas/323.218

ISSN 2316-0861

Influence of Missing Value Imputations on the Performance of Canonical Correspondence Analysis: Ecological Applications

Matthews Lazaro[1,2], **Micheline Gbeha**[3] **and Romain Glèlè Kakaï**[1*]

[1]Laboratoire de Biomathématiques et d'Estimations Forestières (LABEF), University of Abomey-Calavi, 04 B.P. 1525 Cotonou, Benin, [2]Faculty of Health Sciences, Daeyang University, P.O. Box 30330, Lilongwe, Malawi, [3] Département de Mathématiques, Faculté des Sciences et Techniques, Université d'Abomey-Calavi, Bénin

Received March 3, 2018; April 23, 2018; Published Online Month June, 2018

Abstract. This paper assessed the influence of four imputation methods of missing values on the performance of canonical correspondence analysis (CCA). Missingness was introduced in complete multivariate normal data sets under three missing mechanisms : MCAR, MAR and NMAR. Results showed that mean imputation recorded the best performance under MCAR and MAR while for NMAR, median imputation was the best.

Key words: Complete Data; Missingness; Simulation, Accuracy Criteria; Imputation Methods.
AMS 2010 Mathematics Subject Classification : 62-07; 62P12

Presented by Dr. Lotsi Anani, University of Ghana, Accra Legon.
Corresponding Member of the Editorial Board.

*Corresponding author Romain Glèlè Kakaï: glele.romain@gmail.com
Matthews Lazaro : lazaromatthews@yahoo.com
Micheline Gbeha : micgbeha@gmail.com

Full Abstract (ENGLISH) The main objective of this study was to assess the influence of four imputation methods of missing values (mean, median, random forest and zero) on the performance of canonical correspondence analysis (CCA). Firstly, complete multivariate normal environmental data sets were simulated by taking into account sample size, number of variables, proportion of noise and correlation between variables. Thereafter, missingness in the complete data sets was artificially introduced at 0.1, 0.3 and 0.5 under three missing mechanisms: MCAR, MAR and NMAR. For each combination of factors, CCA was applied and constrained inertia was assessed between the complete data set and imputed data set. Results obtained showed that mean imputation recorded the best performance when data was MCAR and MAR. However, under NMAR, median imputation was the best preferred method. The study showed that beyond a missing value proportion of 30 % the performance of imputation methods significantly reduced.

Résumé (FRENCH) L'objectif principal de cette étude est d'évaluer l'influence de quatre méthodes d'imputation de valeurs manquantes (imputation par moyenne, médiane, forêt aléatoire et zero) sur la performance de l'analyse des correspondances canoniques (ACC). Tout d'abord, des données complètes de distribution Normale multivariée ont été générées en prenant en compte la taille des échantillons, le nombre de variables, la proportion de bruit et la correlation entre les variables. Ensuite, des valeurs manquantes ont été artificiellement introduites dans les données environnementales (10, 30 et 50 %) suivant trois mécanismes: MCAR, MAR et NMAR. Pour chaque combinaison des facteurs, l'ACC a été appliquée et l'inertie sous contrainte des données environnementales complètes et imputées a été calculée. Les résultats obtenus montrent que l'imputation par moyenne présentait la meilleure performance dans le cas de MCAR et MAR. Toutefois, sous un NMAR, l'imputation par médiane était la meilleure. L'étude a montré qu'à partir d'une proportion de valeurs manquantes de 30 %, la performance des méthodes d'imputation décroit significativement.

1. Introduction

Canonical Correspondence Analysis (CCA) is a multivariate method to relate species communities to known variation in the environment. CCA is an extension of correspondence analysis (CA) with predictor variables. If CA is applied to the (n x m) matrix $\mathbf{Y}(y_{ik} \geq 0)$, CCA considers this matrix as a matrix of multivariate responses and requires a second (n x p) matrix **Z** with predictor variables (columns of **Z**). In ecology, **Y** normally constitutes the species data with y_{ik} the presence or absence (1/0) or abundance of species k in site i and **Z** contains environmental variables with z_{ij} the measurement of environmental variable j in site i. Because CCA uses data on environment to structure the community analysis, CCA has been called a method for direct gradient analysis (Ter Braak, 1986).

Currently CCA is one of the most popular ordination techniques in community ecology. Based on the findings of Okland (1996), many ecologists use CCA as if it is yet another ordination technique, when in fact they differ in objectives. Indeed

CCA is easily misused because it is a relatively complex method. Besides, the performance of the method has not been intensively explored and documented in the literature (Mccune, 1997).

A fundamental but poorly understood characteristic of CCA is how it responds to missing values imputation methods in environmental data. CCA like any other standard multivariate methods is based on the eigen decomposition of a cross product matrix (e.g., covariance matrix) and thus requires complete data sets. Whatever precaution one takes, both species and environmental matrices can contain missing values and then require a particular attention during the statistical analysis. Actually, they may pose a great threat to the validity and generalizability of study results due to selection bias and loss of statistical power and precision when the sample size is reduced (Jan and Petr, 2003).

Rubin (1976) distinguished three mechanisms generating missing data namely missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR means that the probability that an observation is missing is not related to its value or to any other values in the data set. MAR means that the probability that an observation is missing is related to the values for some other observed variables. Finally, NMAR means that the probability that an observation is missing is related to its value.

There are many solutions to the missing values such as the zero, mean or median of the variables, and more recently random forest (RF) (Piotr *et al.*, 2014). The present study therefore assess the influence of four different missing value imputation approaches on CCA using Monte Carlo methods. A comparison of the missing data imputation methods has provided evidence of the most appropriate method to be used to fill up missing data in environmental data for CCA.

2. Methods

2.1. Principle of Canonical Correspondence Analysis

Let us suppose a survey of *n* sites (samples) lists the abundances or occurrences (presence scored as 1, absence as 0) of *m* species and the values of *q* environmental variables $(q < n)$. Let y_{ik} be the abundance or presence/absence (1/0) of species k $(y_{ik} > 0)$, and z_{ij} the value of environmental variable j at site i . The first step in indirect gradient analysis is to summarize the main variation in the species data by ordination (Ter Braak, 1986). This results into the response model for the species being bell-shaped function and is denoted as:

$$E(y_{ik}) = c_k exp[\frac{1}{2t_k^2}(x_i - \mu_k)^2] \quad (1)$$

where $E(y_{ik})$ represents the expected (average) value of y_{ik} at site i that has score of x_i on the ordination axis; the parameters for the k are c_k, the maximum of that species' response curve; μ_k, the mode or optimum and t_k , tolerance, a measure of ecological amplitude. Then the second step in indirect gradient analysis is to estimate site scores which mathematically is given by:

$$x_i = b_0 + \sum_{j=1}^{q} b_j z_{ij} \tag{2}$$

where b_0 is the intercept and b_j , the regression coefficient for environmental variable j . The species data are thus indirectly related to the environmental variables, via the ordination axis.

Canonical correspondence analysis simultaneously estimates the species optima, the regression coefficients and, hence, the site scores by using the model described by (1), in conjunction with (2) (see Ter Braak (1986) for more information)

2.2. Methods of handling missing data

Zero imputation: Zero imputation is used to replace all missing value with zeros. This method although common in ecology, produces downward-biased standard errors since the zeros are treated as knowns rather than probabilistic estimates (Lall , 2016). This means that essentially this does not solve any problem in terms of quality statistical results.

Mean Imputation: Mean Imputation consists of replacing the missing data for a given feature (attribute) by the mean (quantitative attribute) of all known values of that attribute in the class where the instance with missing attribute belongs (Kantardzic, 2003) . According to Little and Rubin (2002), among the weaknesses of mean imputation are (i) sample size is overestimated, (ii) variance is underestimated, (iii) correlation is negatively biased, and (iv) the distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean. Replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical tests based on it. The function *meanimp*( ) which is available in the package *ForImp* of R software was used to apply the method (Barbiero *et al.*, 2015).

Median imputation (MDI): Since the mean is affected by the presence of outliers, it seems natural to use the median instead just to ensure robustness. In this case the missing data for a given feature is replaced by the median of all known values of that attribute in the class where the instance with the missing feature belongs (Acuña and Rodriguez, 2004). This method is also a recommended choice when the distribution of the values of a given feature is skewed. The function *medianimp*( ) of the package *ForImp* was considered in the R software to implement the method.

Random Forest (RF): The method treats the variable of the missing value as the response variable and borrows information from other variables by the resamplingbased classification and regression trees to grow a random forest for the final prediction. The method is repeated until the imputed values reach convergence. RF is a classification and regression technique that can deal with both parametric and non-parametric data sets of complex linear and non-linear problems (Breiman, 2001). The approach is based on the estimation of imputation error which is calculated for the bootstrapped samples and therefore no separate cross-validation is required. Package *missForest* of R software (Stekhoven and Buehlmann, 2012) is used to implement this method and number of trees was fixed to 100.

2.3. Simulation design for assessing the influence of missing data imputes

Factors considered: Normally distributed complete data set with a varying correlation structure was generated. We varied correlation between variables (0, 0.4, 0.8), sample sizes (20, 50, 100), number of environmental variables (4, 7, 10), proportion of noise (0, 0.2, 0.4), and then introduced different proportions of missing values (0.1, 0.3, 0.5), in all the variables with three different generating missing value mechanisms: MAR (Missing at Random), MCAR (Missing Completely at Random), NMAR (Not Missing at Random).

Complete data sets: The complete data sets used in the simulation design is a combination of 3 levels of correlation of variables, 3 sample sizes, 3 numbers of environmental variables and 3 levels of noise. This yielded a total of 81 combinations of the complete data sets. The value of 5 was fixed as the variance in the covariance variance matrix ($\sum$). The following mean vector was specified for the environmental matrix: $\mu = [1000, 25, 8, 120, 10.5, 14, 7.5, 12.5, 9, 7]$. The mean values were randomly selected since the mean of the variable has no effect on the performance of CCA so is the unit of the variable (Palmer, 1993). These values were greater than 5 (variance) to ensure positive values in the environmental data set. The *MASS* package of R software was used to generate this multivariate normal distributed data using the *mvrnorm*( ) function.

Incomplete data sets: using the generated complete data sets, different proportions of missing values were artificially introduced thus at 10 %, 30 % and 50 % with three different missing data mechanisms (MAR, MCAR, NMAR). Under this simulation design, there were 81 conditions of complete datasets x 3 different proportions of missing value x 3 different missing data mechanisms. This gave a total of 729 combinations of the factors considered in case of incomplete data sets. To get meaningful results from the study, each combination was replicated 500 times to have a total of 40,500 complete datasets and 364,500 incomplete data sets.

Community (Species) data set: the community (species) responses data matrix containing 100 species was generated with varying sample sizes depending on environmental data matrix. The simulated species data followed a Poison distribution (counts of species) with parameter lambda equals 1. This was achieved by using the function *rpois*() in R. To simulate a typical ecological data, 50 % of values in the data set was replaced by zeros. The four missing value imputations (zero, mean, median, and Random Forest) were then applied on each of 364,500 incomplete data sets of environmental variables thereby yielding a total of 1,458,000 records (364,500 x 4).

2.4. Performance evaluation criteria

(a) Constrained inertia.

The total inertia in the species data is the sum of eigenvalues of the constrained and the unconstrained axes, and is equivalent to the sum of eigenvalues or total inertia of CA. Moreover, explained inertia, compared to total inertia was used as a measure of goodness of fit for the performance of CCA. Thus, this was used as a measure of how well species composition is explained by the environmental variables. However, goodness of fit for CCA is elusive, because the arch effect itself has some inertia associated with it (Bakus, 2007). Nevertheless, we used this approach since there exist no alternative option.

Using the inertia from all 729 data conditions with missing data and 81 conditions of complete data sets, different imputations methods were assessed. Accordingly the actual inertia was obtained from the complete (control) data sets (with 81 conditions) generated and a comparison was made with inertia obtained from the similar data condition but with a particular imputation method and any departure was assumed to be attributed to the imputation method. Based on this, different error statistics such as relative error (E_r) and relative bias (B_r) used by Glèlè Kakaï and Palm (2009) were computed for each imputation under various conditions.

$$B_r = \frac{x_i - x_t}{x_t} \qquad E_r = \frac{\mid x_i - x_t \mid}{x_t}$$

where x_i is the measured inertia value from imputed data, x_t is the inertia value from complete data set (true value) and B_r and E_r are relative bias and relative error respectively. MRE indicates how close the observed data points are to the model's predicted values while MRB indicates how close, generally in one direction, the observed data points are to the model's predicted values. The lower the MRE absolute value the better the fit and the same is with MRB. Furthermore, mean, variance and Coefficient of variation (CV) of mean relative error (MRE) were also computed to give the overall performance of each substitution method under different combinations of the variables. The lower the value of CV the less the dispersion within a variable and *vise-versa*.

(b) Number of constrained axes.

We performed CCA and evaluated the total explained variance (TEV) captured in the canonical components (CCAs). With the aid of plots, we illustrated how each of the different substitutions can influence variance across the analysis. We recorded overall mean number of constrained components across each imputation method. In this study, 70 % cut off point was selected as criterion in estimation of the number of canonical components (constrained axes) to be used.

(c) Agreement of samples scores (RV coefficients).

To further assess the performance of the imputation methods, we computed the RV coefficient (Escoufier, 1973) to evaluate the agreement between the scores for the individuals (respectively, the variables) obtained by the different imputation methods and those obtained from CCA of the complete data set. In this way, a matrix of sample score for imputed data and that of complete data under a particular condition were used. This was achieved by using a function *coeffRV( )* of *FactoMineR* package in R. A value of 1 indicates a perfect agreement between configurations while a value of 0 shows lack of agreement. If we denote two positive semi-definite matrices of same dimensions by **S** and **T**, then RV coefficient between them is defined as (Escoufier, 1973):

$$RV = \frac{trace\mathbf{S}^T\mathbf{T}}{\sqrt{(trace\mathbf{S}^T\mathbf{S}) * (trace\mathbf{T}^T\mathbf{T})}} = \frac{\sum_i^I \sum_j^I s_{i,j} t_{i,j}}{\sqrt{(\sum_i^I \sum_j^I s_{i,j}^2)}\sqrt{(\sum_i^I \sum_j^I t_{i,j}^2)}} \qquad (3)$$

where S, S, s and t are positive semi-definite matrices with a value of i and j row and column respectively. The RV coefficient allowed us to see performance of each imputation method. The closer to 1 the RV is, the more similar the two matrices.

3. Results

3.1. Performance of missing data imputation methods based on constrained Inertia

Table 1 shows that under MCAR and MAR missing mechanisms, mean imputation has in general best performance. However, under NMAR, mean and median imputations have the best performance followed by Random forest. We also notice that under MAR, both Random forest and median imputations have the same overall performance. In general, under the three missing processes, zero imputation was overall the worst performer.

It is observed from Table 1 that the performance of median imputation under MCAR decreases with higher proportion of noise (0.4) while concurrently, the performance of RF increases. For NMAR, median imputation is the best performer when proportion of the missing value is lowest and when proportion of noise is moderate (0.2). Interestingly, with a missing proportion of 0.5, the performance of zero imputation becomes the best. However, under MAR, the performance of the random forest increases with increase in sample size and number of environmental variables. The same trend is observed with increase in proportion of noise and correlation between variables. In almost all considered cases of combinations of factors, mean imputation gives relatively the best results.

(a) - Mean Relative Error (MRE)

Since the median rank approach only indicates the best or the least imputation method without necessarily giving quantitative performance, Table 2 presents the

Table 1. Median rank of different imputation methods under different combinations of the factors.

	MCAR				NMAR				MAR			
	M	Med	R	Z	M	Med	R	Z	M	Med	R	Z
Overall	1	2	3	4	2	2	3	4	1	3	3	4
$n = 20$ with $p = 4$	1	2	3	4	2	2	3	4	1	2	3	4
$n = 20$ with $p = 7$	1	2	3	4	2	2	3	4	1	3	3	4
$n = 20$ with $p = 10$	1	2	3	4	2	2	3	4	1	3	3	4
$n = 50$ with $p = 4$	1	2	3	4	2	2	3	4	1	3	2	4
$n = 50$ with $p = 7$	1	2	3	4	2	2	3	4	1	3	2	4
$n = 50$ with $p = 10$	1	2	3	4	2	2	3	4	1	3	2	4
$n = 100$ with $p = 4$	1	2	3	4	2	2	3	4	1	3	2	4
$n = 100$ with $p = 7$	1	2	3	4	2	2	3	4	1	3	2	4
$n = 100$ with $p = 10$	1	2	3	4	2	2	3	4	1	3	3	4
$prop.m = 0.1$	1	2	3	4	2	1	3	4	1	3	2	4
$prop.m = 0.3$	1	2	3	4	1	2	3	4	1	3	2	4
$prop.m = 0.5$	1	2	3	4	4	2	3	1	1	3	3	4
$prop.n = 0$	1	2	3	4	1	2	3	4	2	1	3	4
$prop.n = 0.2$	1	2	3	4	2	1	3	4	1	3	2	4
$prop.n = 0.4$	1	3	2	4	2	2	3	4	1	3	2	4
$coR = 0$	1	2	3	4	2	2	3	4	1	3	3	4
$coR = 0.4$	1	2	3	4	2	2	3	4	1	3	2	4
$coR = 0.8$	1	2	3	4	2	2	3	4	1	3	2	4

prop.m stands for missing proportion, prop.n stands for proportion of noise, coR stands for correlation, M stands for Mean, Med stands for Median, R stands for Random Forest and Z stands for Zero; n = sample size ; p = number of variables.

mean, Coefficient of variation (CV) and variance of mean relative error for each imputation method under three different missing mechanisms.

The same trend as observed in Table 1 is noticed here. In general, across MCAR and MAR, mean imputation which seems to be the best method according (Table 1) has the lowest mean of MRE. Furthermore, it is noticed that under MCAR, the median and RF imputations have the same value of mean of MRE then followed by the zero while with MAR. The trend however, changes under NMAR where both the mean and median imputations have the same mean value of error. It is evident that zero imputation has highest values of variances of MRE under all the three missing mechanisms thereby implying lack of stable performance. The CV of

MRE for each imputation method displays a very important trend across the three missing mechanisms. By each imputation method, MCAR and MAR have smaller values of CV than NMAR.

Table 2. Mean, Coefficient of variation (CV) and Variance (Var) of Mean Relative Error (MRE)

Imputation	Missing mechanisms								
	MCAR			NMAR			MAR		
	Mean	CV(%)	Var	Mean	CV(%)	Var	M Mean	CV(%)	Var
Mean	0.0015	98.55	$2.3 * 10^{-6}$	0.0013	143.13	$3.6 * 10^{-6}$	0.0014	98.49	$3.6 * 10^{-6}$
Median	0.0017	101.68	$2.9 * 10^{-6}$	0.0013	151.05	$3.7 * 10^{-6}$	0.0015	98.16	$3.8 * 10^{-6}$
RF	0.0017	93.75	$2.4 * 10^{-6}$	0.0014	133.60	$3.7 * 10^{-6}$	0.0016	92.32	$4.0 * 10^{-6}$
Zero	0.0027	81.86	$5.2 * 10^{-6}$	0.0023	96.67	$5.0 * 10^{-6}$	0.0026	87.22	$8.1 * 10^{-6}$

(b) - Mean Relative Bias (MRB).

The results of MRB in general indicate that mean, median and random forest imputations have lower absolute values of MRB (Figure 1). Figure 1 shows the best behavior of the mean, median and random forest imputation across all the three missing mechanisms. Nevertheless, the mean and median imputations obtain the lowest median values of bias under NMAR although they have a lot of extreme values of relative bias. In addition, the dispersion around the median values of mean relative bias is less pronounced in the case of mean, median and random forest across all missing mechanisms. The zero imputation has a lot of extreme values of relative bias (Figure 1). The same trend of results is noticed under MCAR and MAR where the incidence of extreme values among mean, median and random forest is less noticeable.

(c) - Number of constrained axes.

To have an idea on how each imputation method affects the number of constrained components retained, we recorded the number of components within each combination of factors for the number of replications (500) (Figure 2). Based on the results obtained, there is no significant difference in the number of retained components between imputed data and complete data set. It is noticed that in general the mean number of components in complete data set is 4.64. A drop in mean number of components is noticed when one moves from MCAR to MAR although not so significant. As it can be observed in Figure 2, under MCAR, when zero is applied as a substitution method, a mean of 4.60 components is required to cover over 70 % of explained variance. This result is followed by RF method where a mean of 4.62 is needed to cover similar level of explained variance. For median and mean replacement, a mean of 4.63 is required to reach comparable level of variance.

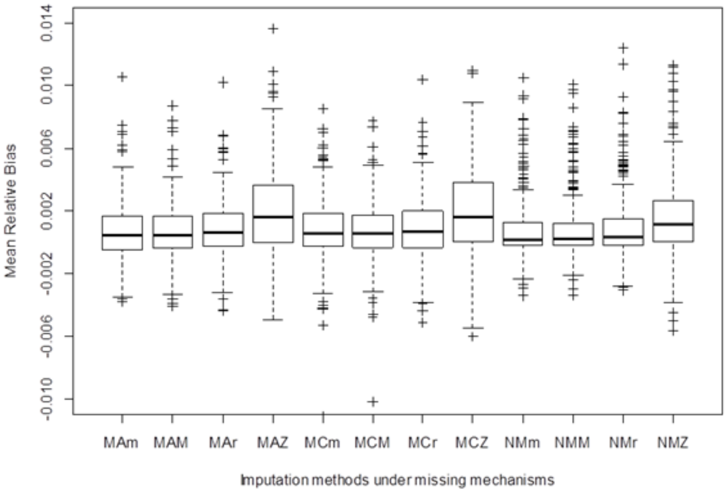

Fig. 1. Boxplot for the mean relative bias of imputation methods under three missing mechanisms. **Legend**: On the x-axis, the first two letters are initials for the missing mechanisms (MA=Missing at Random; MC=Missing Completely at Random; NM=Not Missing at Random) and the last letter is for the imputation method (M=mean; m=Median; r=Random forest; Z=zero). For instance MAm corresponds to combination of Missing at Random and Median imputation.

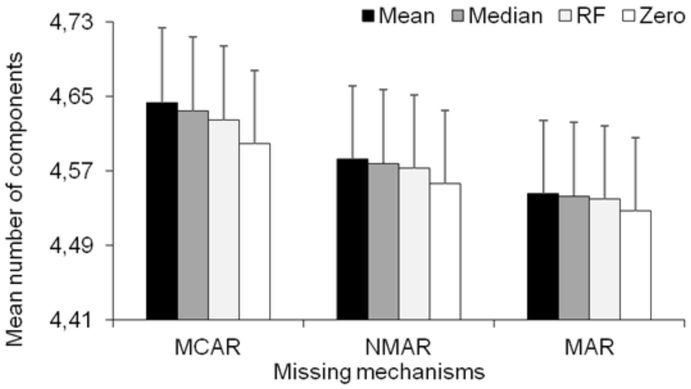

Fig. 2. Mean number of constrained components of four imputation methods under the three missing mechanisms.**Legend**: On the x-axis, the first two letters are initials for the missing mechanisms (MA=Missing at Random; MC=Missing Completely at Random; NM=Not Missing at Random) and the last letter is for the imputation method (M=mean; m=Median; r=Random forest; Z=zero). For instance MAm corresponds to combination of Missing at Random and Median imputation

3.2. Performance of missing data imputation methods based on agreement of samples scores (RV coefficients)

According to results presented in Table 3, missing mechanisms and proportion of missing values greatly affect the performance of imputation methods. Furthermore, the RV coefficients for the imputation methods also increase with decrease in the proportion of missing values. Under MCAR, when the missing proportion is 0.5, the imputation methods give unsatisfactory results, zero imputation being the worst affected. The same trend is noticed under NMAR and MAR. However, under NMAR when missing proportion is 0.5, mean, median and RF record worse performance than zero imputation (Table 1). The performance of imputations however remains relatively constant with increase in proportion of noise, correlation between variables, number of variables and sample size. Taking all factors into account and according to RV coefficients, mean imputation performs the best under MCAR and MAR, while median imputation is frequently the best under NMAR. With each combination of factors by each imputation method (mean, median and random forest), RV coefficients under NMAR are the lowest among the three missing mechanisms. However, the zero imputation has the highest values of RV coefficients under this mechanism.

Table 3. RV coefficient of samples scores between imputed data set and complete data set according to the combinations of some factors.

	MCAR				NMAR				MAR			
	M	Med	R	Z	M	Med	R	Z	M	Med	R	Z
Overall	0.7505	0.7310	0.7121	0.3479	0.6700	0.6937	0.6549	0.5486	0.7146	0.6820	0.6804	0.4319
$n = 20$ with $p = 4$	0.7415	0.7158	0.6928	0.2920	0.6486	0.6787	0.6280	0.4519	0.7008	0.6671	0.6552	0.3542
$n = 20$ with $p = 7$	0.7702	0.7524	0.7427	0.3835	0.6572	0.6987	0.6463	0.5944	0.7370	0.7126	0.7050	0.4718
$n = 20$ with $p = 10$	0.8000	0.7840	0.7748	0.4471	0.6669	0.7166	0.6584	0.6791	0.7661	0.7428	0.7314	0.5350
$n = 50$ with $p = 4$	0.7349	0.7099	0.6799	0.2752	0.6610	0.6716	0.6390	0.4576	0.6935	0.6506	0.6499	0.3614
$n = 50$ with $p = 7$	0.7435	0.7252	0.7080	0.3623	0.6716	0.6901	0.6598	0.5603	0.7102	0.6768	0.6814	0.4468
$n = 50$ with $p = 10$	0.7532	0.7385	0.7227	0.4135	0.6842	0.7090	0.6746	0.6155	0.7240	0.6960	0.6989	0.4987
$n = 100$ with $p = 4$	0.7330	0.7077	0.6783	0.2599	0.6650	0.6827	0.6434	0.4463	0.6902	0.6465	0.6471	0.3439
$n = 100$ with $p = 7$	0.7372	0.7191	0.7005	0.3293	0.6809	0.6928	0.6651	0.5423	0.7008	0.6662	0.6716	0.4178
$n = 100$ with $p = 10$	0.7414	0.7267	0.7092	0.3682	0.6941	0.7028	0.6800	0.5902	0.7089	0.6794	0.6832	0.4571
$prop.m = 0.1$	0.9022	0.8917	0.8878	0.5214	0.9854	0.9944	0.9707	0.4959	0.8924	0.8717	0.8789	0.5596
$prop.m = 0.3$	0.8081	0.7893	0.7796	0.3712	0.9205	0.8861	0.8888	0.4613	0.7869	0.7535	0.7606	0.4493
$prop.m = 0.5$	0.5412	0.5121	0.4689	0.1511	0.1041	0.2006	0.1054	0.6887	0.4645	0.4208	0.4018	0.2866
$prop.n = 0$	0.7588	0.7557	0.7154	0.1746	0.7298	0.7782	0.7108	0.4327	0.7463	0.7493	0.7040	0.2719
$prop.n = 0.2$	0.7470	0.7345	0.7100	0.3870	0.6694	0.6964	0.6530	0.5135	0.6705	0.6365	0.6409	0.4595
$prop.n = 0.4$	0.7458	0.7029	0.7110	0.4821	0.6107	0.6064	0.6011	0.6997	0.7270	0.6602	0.6964	0.5642
$coR = 0$	0.7524	0.7329	0.7129	0.3477	0.6708	0.6949	0.6537	0.5500	0.7170	0.6842	0.6809	0.4315
$coR = 0.4$	0.7514	0.7317	0.7127	0.3477	0.6696	0.6934	0.6548	0.5489	0.7153	0.6827	0.6805	0.4328
$coR = 0.8$	0.7478	0.7285	0.7108	0.3483	0.6695	0.6928	0.6564	0.5470	0.7116	0.6791	0.6799	0.4312

n stands for sample size, p stands for number of environmental variables, Prop.m stands for missing proportion, prop.n stands for proportion of noise, coR stands for correlation, M stands for Mean, Med stands for Median, R stands for random forest and Z stands for Zero.

4. Discussion

Missing data can lead to problems that affect the interpretation and inferences of research results, the understanding and explanation of conclusions made, the strength of the study design, the validity of conclusions about relationships between variables and may limit the representativeness of the sample (Morais, 2013). So far several studies have been carried out to compare the performance of different missing imputations (Anani *et al.* , 2017; Niass *et al.*, 2015; Dray and Josse, 2014). Despite these numerous studies, multivariate analysis of incomplete data sets has received little attention in ecology (Dray and Josse, 2014). The uniqueness of this study therefore is that four imputation methods have been assessed under different conditions in environmental data on performance of CCA. The study revealed that the performance of imputation methods on CCA is affected by proportion of missing value and missing value mechanisms.

When missing values are MCAR and MAR, mean imputation yielded a better performance in accuracy. Hening (2009) pointed out that mean and median gave satisfactory output comparing different missing data imputation methods in Ohio University Student Retention Database. This study has observed that under NMAR, mean and median imputations gave similar results thereby implying loss of power by mean imputation. Furthermore, when data is normally distributed then both mean and median imputations will provide very similar results (Hrydziuszko and Viant, 2011). This is being supported by the lowest values of RV coefficients under NMAR that have been reported for each combination of some factors among the three missing value mechanisms in this study. It is also noticed that in case of NMAR, median imputation has the best performance among the four methods. Mean and median imputation seemed to have low MRE because the values imputed for the missing data do not yield large differences although the standard deviations and variances for mean imputation are underestimated due to centralization of the distribution. In addition, the median is less vulnerable to outliers than the arithmetic mean and is therefore considered to be more robust (Steuer *et al.*, 2007). This tells that in case the two methods have the same performance, median imputation is preferred to mean imputation. Our findings also pointed out that Random forest gave acceptable results across the three missing mechanisms. We noticed that in general random forest was coming on third position in terms of performance. Since random forest imputation can handle both parametric and non-parametric data sets, it is expected that it has the best performance (Piotr *et al.*, 2014). Generally, although there is no established cutoff from the literature regarding an acceptable percentage of missing data in a data set for valid statistical inferences (Dong and Peng, 2013), this study has demonstrated that when the missing proportion increases up to 50 %, the performance of mean, median, random forest and zero imputations on CCA is greatly affected, the worst results being observed under NMAR. The results further demonstrated that although zero imputation was negatively affected with increase in missing proportion, it gave surprising results under NMAR as it emerged the best imputation when missing proportion was 0.5. Although zero imputation seemed the best, the obtained RV coefficients were not appealing. The results are in agreement to Bennett (2001) who observed that statistical analysis is likely to be biased when more than 10 % of data are missing. However, Dray and Josse (2014) reported that proportion of missing values does not affect the performance of mean imputation in PCA. We also noticed that different imputation methods do not significantly affect the number of retained constrained axes. Based on the results it is evident that the number of retained components from the imputed data sets was not quite different from the number of components in complete data sets.

5. Conclusion

This study investigated the performance of four missing value imputation in environmental data on the canonical correspondence analysis under three missing mechanisms. It is therefore concluded based on the results that the performance of these methods indeed depends on missing mechanisms and proportion of missing data. The study has shown that when data is missing completely at random or missing at random and normally distributed, then among the tested four methods, mean imputation is favoured. However, if the data is not missing at random, median imputation is the best. The study has also concluded that random forest has stable performance although not the best performer under this study. It has been also recommended not to use zero imputation in handling missing data especially when the data has no or few zeros. The study has also shown that indeed handling missing data when data is NMAR is complex issue and need several approaches. Based on this study, it is concluded that before doing a missing value imputation the distribution of the data, the missing mechanisms and the missing proportion be examined in order to prescribe the best imputation method.

Acknowledgment. We would like to thank CSAA for the financial support through European Union. We also wish to thank the presenter of the article for the insightful comments and recommendations that helped to improve the quality and the editors for a high professional handling of the publication process.

References

Acuña, E., and Rodriguez, C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. *Classification, Clustering, and Data Mining Applications*, 639-647.

Anani, L., Asiedu, L., Katsekpor, J. (2017). Comparison of Imputation Methods for Missing Values in Longitudinal Data Under Missing Completely at Random (mcar) mechanism. *African Journal of Applied Statistics*, 4(1), 241-258.

Bakus, G.J. (2007). *Quantitative Analysis of Marine Biological Communities Field Biology and Environment*. A John Wiley & Sons, Inc. Publication.

Barbiero, A., Ferrari, P.A., and Manzi, G. (2015). *Imputation of Missing Values Through a Forward Imputation Algorithm, R Package "ForImp"*.

Bennett D.A. (2001). How can I deal with missing data in my study? *Aust. N Z J Public Health*, 25(5), 464-469.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45, 5-32.

Dray, S., and Josse, J. (2014). Principal component analysis with missing values: a comparative survey of methods. *Plant. Ecol.*, 216, 657-667.

Dong, Y., and Peng, C-Y.J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2(1), 222.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29, 751-760.

Glèlè Kakaï, R. and Palm, R. (2009). Empirical comparison of error rate estimators in logistic discriminant analysis. *Journal of Statistical Computation and Simulation*, 79(2), 111-120.

Hening, A.D. (2009). *Missing Data Imputation Method Comparison in Ohio University Student Retention*. Master thesis, the faculty of the Russ College of Engineering and Technology of Ohio University.

Hrydziuszko, O., and Viant, M.R. (2011). Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline. *Metabolomics*,161-174.

Jan, L., and Petr, S. (2003). *Multivariate Analysis of Ecological Data using CANOCO, United States of America*. Cambridge University Press, New York.

Kantardzic, M. (2003). *Data Mining – Concepts, Models, Methods, and Algorithms*. IEEE.

Lall, R. (2016). How Multiple Imputation Makes a Difference. *Political Analysis*, 24, 414-433.

Little, R.J.A. and Rubin, D.B. (2002). Bayes and multiple imputation. *Statistical Analysis with Missing Data, Second Edition*, 200-220.

Mccune, B. (1997). Influence of noisy environmental data on canonical correspondence analysis. *Ecology*, 78(8), 2617-2623.

Morais, S.F. (2013). *Dealing with Missing data: An Application in the Study of Family History of Hypertension*. A Master Dissertation, Faculty of Medicine of the University of Porto.

Niass, O., Diongue, A.K., and Toure, A. (2015). Analysis of missing data in sereoepidemiologic studies. *African Journal of Applied Statistics*, 2(1), 29-37.

Okland, R.H. (1996). Are Ordination and Constrained Ordination Alternative or Complementary Strategies in General Ecological Studies? *Journal of Vegetation Science*, 7(2), 289-292.

Palmer, M.W. (1993). Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology*, 74, 2215-2230.

Piotr, S.G., Yun, X., Helen, L.K., Elon, C., David, I.E., Emily, G.A., Michael, L.T., and Royston, G. (2014). Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data. *Metabolites*, 4, 433-452.

Rubin, D. (1976). Inference and missing data. *Biometrika*, 69(3), 581-592.

Stekhoven, D.J., and Buehlmann, P. (2012). MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28, 112-118.

Steuer, R., Morgenthal, K., Weckwerth, W., and Selbig, J. (2007). A gentle guide to the analysis of metabolomic data. *Methods Mol. Biol.*, 358, 105-126.

Ter Braak, C.J.F. (1986). Canonical Correspondence Analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67, 1167-1179.