Revista Ciência Agronômica, v. 51, n. 4, e20196817, 2020
Centro de Ciências Agrárias - Universidade Federal do Ceará, Fortaleza, CE
www.ccarevista.ufc.br

Scientific Article
ISSN 1806-6690

Multiple imputation to fill in missing data in soil physico-hydrical properties database[1]

Imputação múltipla para o preenchimento de dados faltantes em banco de dados de propriedades físico-hídricas do solo

Luciana Maria de Oliveira[2], Herdjania Veras de Lima[2], Sueli Rodrigues[3]*, Eduardo Jorge Maklouf Carvalho[4] and Lorena Chagas Torres[2]

ABSTRACT - Missing values in databases is a common issue and almost inevitable. Multiple imputation (MI) is an efficient statistical method for estimating missing values in an incomplete dataset. To test this approach for a soil database, we hypothesized that the imputation of missing data provides a statistically more accurate database than the complete case analysis (CCA). The overall goal of our study was to evaluate the efficiency of the MI using the MICE (Multivariate Imputation by Chained Equations) algorithm to fill in missing data in a database of soil physico-hydrical properties, and to show that it is more feasible to perform the imputation than the CCA. Preliminary analyses were performed to check the suitability of the proposed algorithm. Imputation of the missing data of each variable was adjusted using linear regression models. The variables with missing data comprise the model as the dependent variable and the other variables, which were correlated with the same, enter as covariates. The analysis was performed by comparing the values of the estimates, their standard errors and 95% confidence intervals. The pattern missing was multivariate and arbitrary and, organic matter was the variable with the largest amount of missing data. The significance of the covariates varied depending on the variable to be estimated. The results showed that the MICE presented better performance than CCA, since, although the statistical comparison of the two methods was similar, multiple imputation maintains the size of the database and preserves the general distribution.

Key words: Soil database. Incomplete data. Markov Chain Monte Carlo. Missing predictors.

RESUMO - Valores faltantes em banco de dados é um problema comum e quase inevitável. A imputação múltipla (IM) é um método estatístico eficiente para estimar valores ausentes em um conjunto de dados incompleto. Para testar essa abordagem em um banco de dados de solo, hipotetizamos que a imputação de dados ausentes fornece um banco de dados estatisticamente mais preciso do que a análise de casos completos (ACC). O objetivo geral do estudo foi avaliar a eficiência da IM usando o algoritmo MICE (Imputação Multivariada por Equações Encadeadas) para preencher dados ausentes em um banco de dados de propriedades físico-hídricas do solo e mostrar que é mais viável realizar a imputação do que a ACC. Análise preliminar do banco de dados foi realizada para verificar a adequação do algoritmo proposto. A imputação dos dados faltantes de cada variável foi ajustada usando modelos de regressão linear. Variáveis com dados faltantes entram no modelo como variável dependente e as outras como covariáveis. As análises foram realizadas comparando os valores das estimativas, seus erros padrão e intervalos de confiança de 95%. O padrão de faltas foi do tipo multivariado arbitrário e, a matéria orgânica foi a variável com a maior quantidade de dados faltantes. A significância das covariáveis variou de acordo com a variável a ser estimada. Os resultados mostraram que o MICE apresentou melhor desempenho que a ACC, pois, embora a comparação estatística dos dois métodos tenha sido semelhante, a imputação múltipla mantém o tamanho do banco de dados e preserva a distribuição geral.

Palavras-chave: Banco de dados de solo. Dados incompletos. Monte Carlo via Cadeias de Markov. Preditores de falta.

DOI: 10.5935/1806-6690.20200079
Editor do artigo: Professora Mirian Cristina Gomes Costa - mirian.costa@ufc.br
*Author for correspondence
Received for publication 15/07/2019; approved on 14/02/2020

[1] Parte da Tese do primeiro autor apresentada ao Programa de Pós-Graduação em Agronomia da Universidade Federal Rural da Amazônia/UFRA

[2] Programa de Pós-Graduação em Agronomia/PPGAGRO, Universidade Federal Rural da Amazônia/UFRA, Belém-PA, Brasil, luciana_oliveira658@hotmail.com (ORCID ID 0000-0003-1324-9318), herdjania.lima@ufra.edu.br (ORCID ID 0000-0001-8757-5090), loren_chagas@yahoo.com.br (ORCID ID 0000-0001-5548-9438)
[3] Universidade Federal do Piauí/UFPI, *Campus* Profa. Cinobelina Elvas/CPCE, Bom Jesus-PI, Brasil, rodsueli@gmail.com (ORCID ID 0000-0001-8249-2990)
[4] Empresa Brasileira de Pesquisa Agropecuária/EMBRAPA, Belém-PA, Brasil, eduardo.maklouf@embrapa.br (ORCID ID 0000-0003-3536-5598)

INTRODUCTION

Missing data in scientific studies are common and occur for a variety of reasons, resulting in incomplete databases, which may be a restriction for statistical analysis (AUDIGIER; HUSSON; JOSSE, 2015). However, problems related to missing data and the implemented solutions (when performed) are rarely mentioned in most publications. This may be due to the little importance given to the problem (e.g. reduction of the sample) or the lack of knowledge of the implemented solutions (often automatically) by statistical software (FIGUEREDO *et al*., 2000).

For example, for multiple regression analysis, the standard procedure in most statistical softwares, when data are missing, is the listwise deletion, which consists of removing all data for each case that has one or more missing values. This analysis is called complete case analysis. This proceeding can markedly reduce the available database and thus, induce to high predictive deviations in the parameter estimation, contesting the validity of the conclusions (PAES; POLETO, 2013).

The degree of the problem is even more significant when multivariate analyses are implemented since these analyses require complete data for all variables (FIGUEREDO *et al*., 2000). In soil science, an example is the estimation of pedotransfer functions that use easily determined soil properties such as soil texture and soil bulk density to predict more complex ones such as those related to soil water retention capacity (SILVA; ARMINDO, 2016). In general, the data used to determine pedotransfer functions, come from several locations, so missing data are common.

In some cases, an option to deal with an incomplete database is to fill in the missing values using simple methods such as mean, median, interpolation, and linear regression. These methods are named single imputation (RUBIN, 1976). However, the single imputation is limited because it does not take into account the uncertainty associated with predicting missing values based on the observed values (VAN BUUREN, 2018).

Currently, modern statistical procedures and software allow a more effective recourse to fill in these gaps. One of these methods is the Multiple Imputation (MI), which considers the variability among the imputations, generating complete data sets by filling the missing values through imputation models, generally more accurate than those provided by the single imputation methods (LITTLE; RUBIN, 2015).

Although the MI technique has been used in several areas (CARVALHO *et al*., 2017; PEDERSEN *et al*., 2017; POYATOS *et al*., 2018; SQUILLANTE JÚNIOR *et al*., 2018), in Soil Science, it is still little explored (CLIFFORD; DOBBIE; SEARLE, 2014; SHAO; MENG; SUN, 2017).

At selecting the MI method, it is recommended that different methodologies should be explored according to the characteristics of the data (KIM *et al*., 2015). The Multivariate Imputation by Chained Equations (MICE) is one of the many algorithms that perform MI based on the Monte Carlo Markov Chain (CARVALHO *et al*., 2017). MICE applications have been used in several areas, but in Soil Science this approach has not been used yet.

To test this approach for soil database, we hypothesized that the imputation of missing data provides a statistically more accurate database than the complete case analysis (CCA). The overall goal of our study was to evaluate the efficiency of the MI using the MICE algorithm to fill in missing data in a database of soil physico-hydrical properties, and to show that it is more feasible to perform the imputation than the CCA.

MATERIAL AND METHODS

Soil database

The soil database (SDB) used in the study stem from 24 municipalities of the state of Pará, northern Brazil. The SDB consists of 631 samples of two soil classes (*Latossolos* and *Argissolos* – Brazilian System Classification, (SANTOS *et al*., 2018) sampled at the depths of 0 to 60 cm between 1997 and 2014. The data were compiled from several sources (scientific papers, dissertations, thesis, Embrapa Research Bulletins and soil data surveys performed by Eastern Amazon Embrapa). Although the SDB includes quantitative and qualitative variables, only the following quantitative variables were considered for this study: sand, clay and silt contents, determined by sieving, sedimentation (pipette method) and difference, respectively; organic carbon content was estimated by the Walkley-Black method and the percent of soil organic matter (OM) was calculated by multiplying the organic carbon content by the factor 1.724 (WAXMAN; STEVENS, 1930); soil bulk density (Bd) by the core method; particle density (Pd) by the pycnometer method; soil total porosity (TP) by the saturation method; soil microporosity (Micro) as the water content at a water potential of -6kPa, corresponding to a 0.05 mm pore diameter in the soil water retention curve and taken as the limit between macro and microporosity (KIEHL, 1979); soil macroporosity (Macro) calculated as the difference between TP and Micro; soil water content at field capacity (FC) and permanent wilting point (PWP), considered as the soil moisture equilibrated at water potentials of -6kPa e -1500kPa, respectively. The latter two were determined on pressure plate extractor. All these methodologies are described in Claessen *et al*. (1997).

Preliminary analysis

Before the imputation process, three preliminary analyses of the missing data were performed to confirm the pattern, mechanism, and proportion of missing. The analyses were:

Pattern - the missing data pattern can be univariate (just one variable contains missing data) or multivariate (more than one variable contains missing data) (SONG; SHEPPERD, 2007). The multivariate pattern may occur as monotonous or arbitrary (RUBIN, 1987).

If the missing data pattern is univariate, the single imputation (SI) method is recommended, while the multiple imputation (MI) procedure is recommended for the multivariate pattern. In the latter, when the monotonous pattern occurs, the most indicated methods are Bayesian Linear Regression (BLR), and the Predictive Mean Matching (PMM), while for the arbitrary pattern the appropriate method is Monte Carlo Markov Chain (MCMC).

Mechanism - the missing data mechanisms represents the statistical relationship between the observations (variables) and the probability of missing data and, are classified into three categories (RUBIN, 1987): (i) *Missing completely at random* (MCAR), when the probability of the data missing depends on neither the observed nor the unobserved data; (ii) *Missing at random* (MAR) when the probability of missing data to some extent depends on the observed data; and (iii) *Not missing at random* (NMAR), when the probability of missing data depends on the missing data values themselves.

In practice, missing data are almost never MCAR, but instead somehow in between MAR and MNAR (GRAHAM, 2009). However, MAR and NMAR mechanisms are not identified by tests. The MCAR mechanism is tested by the Little test (1988) and, the lower the p value ($p<0.05$) the stronger is the evidence that the data is not MCAR.

Proportion - the proportion of missing data was checked through the frequency histograms. If the proportion is ≤ 5% the single imputation (SI) method can be used or the complete case analysis (CCA) can be considered. If the proportion is 5-15% it is still possible to use the SI method, however, the multiple imputation (MI) method is recommended. When the proportion of missing data is ≥ 15%, the appropriate procedure is the MI (HARRELL, 2016).

Multiple Imputation

Verified the conditions above, the chosen method was the multiple imputation by chain equations (MICE) since more than one variable has missing data, no defined pattern (multivariate and arbitrary pattern) was observed and the missing mechanism is MAR.

The MICE algorithm was performed for the set of variables (*x*) described above, some or all of which have missing values. The method consists of perform a series of regression models where each variable with missing data is modeled related to the other variables of the database (fully conditional specification – FCS). Linear regression models were carried out ($\hat{y} = \beta_0 + \beta_1 x + ... + \beta_n x$), where $\hat{y}$ is the variable to be imputed. The variables with missing data enter the imputation model as a dependent variable and the other variables that have a significant correlation ($p \leq 0.05$) with it, enter as covariates (independent variables).

The MICE procedure can be divided into three main steps: imputation, analysis, and combination, briefly described below:

Imputation - Generation of *m* complete data set. MICE perform a series of estimations where each variable takes its turn in being regressed on the other variables, that is, it loops through the variables predicting each variable dependent on the others. MICE runs through an iterative process: In the first iteration, the imputation model for the variable with the least missing values is estimated using only complete data. Next, the variable with the second least missing values is imputed using the complete data and the imputed values from the last iteration. After each variable has been through this process, the cycle is repeated using the data from the last iteration. Ten iterations were performed where the imputed values after the 10th and final iteration constitutes one imputed data set (STUART *et al*., 2009). Here, five versions of data sets ($m = 5$) were generated, since, according to Schafer and Olsen (1998), *m* for 3 to 5 is enough to obtain accurate estimates for most applications.

Analysis - Separately, the five versions of the data set were analyzed by traditional methods of statistical analysis (parameter estimates, standard errors and 95% confidence intervals).

Combination - The last step of the MICE was the combination of the results of the estimates of the *m* complete data sets, using the Rubin's method (1987). Five different sets of the point and variance estimates for a parameter *Q* were estimated. Let Q_j and U_j be the point and variance estimates from the *i*th imputed data set, i=1, 2, ..., *m*. Then the point estimate for *Q* from multiple imputations is the average of the *m* complete-data estimates:

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m} Q_j \tag{1}$$

Let Ū be the within-imputation variance, which is the average of the *m* complete-data estimates:

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m} U_j \tag{2}$$

and B be the between-imputation variance:

$$B = \frac{1}{m-1}\sum_{j=1}^{m} \left(Q_j - \bar{Q}_j\right)^2 \tag{3}$$

Then the variance estimate associated with Q̄ is the total variance:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B \tag{4}$$

Then, confidence intervals (95%) were built for the mean (Q̄) through t-Student approximation:

$$IC = \left[\bar{Q} \pm 1{,}96\sqrt{T}\,\right] \tag{5}$$

The relative efficiency (RE) of the MI of a point estimate based on *m* imputations was quantified through (6):

$$RE = 1 + \frac{FMI}{m} \tag{6}$$

where: FMI = B/B+Ō is the fraction of missing information (FMI) about *Q*, which ranges from 0 to 1 (SCHAFER; OLSEN, 1998). The FMI quantifies the accuracy of the estimate if there is no missing data.

Rubin (1987) introduced the missing information fraction (λ) (7) to indicate how much the estimates were influenced by the presence of missing data,

$$\lambda = \frac{B + \frac{B}{m}}{T} \tag{7}$$

The standard error (SE) of the parameter estimation is given by:

$$SE = \sqrt{\left(1 + \frac{\lambda}{m}\right)} \tag{8}$$

where λ is the missing information fraction and *m* is the number of complete dataset.

Imputation efficiency analysis

The efficiency of the imputation procedure was evaluated by means of the comparison of the estimated parameters (parameter estimates, standard errors and 95% confidence intervals), the determination coefficients of the imputation models and, graphical analyzes (probability density and box-plot).

All the analysis and multiple imputations were performed in the R program (R CORE TEAM, 2017), using the MICE (Multivariate Imputation by Chained Equations) package.

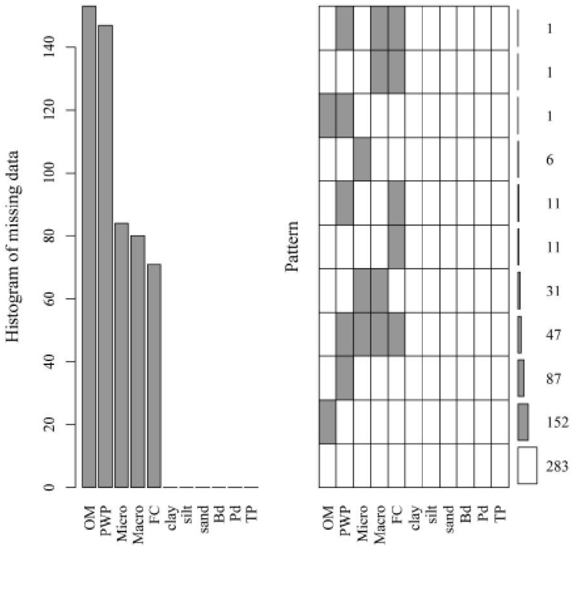
RESULT AND DISCUSSION

The preliminary analysis to identify the proportion and pattern of missing data is shown in Figure 1. The Figure 1a shows, in decreasing order, the number of missing information for each variable with missing data. From the eleven variables that make up the BDS, five have missing data, that is, 45%. The proportion of missing values ranged from 11.3 to 24.2%.

Figure 1b displays the pattern of missing values where the columns are the variables and the rows are the observations. There are 283 samples without gaps, and 348 cases with missing data, which corresponds to a percentage of 55.2. The pattern missing was multivariate and arbitrary. The organic matter is the variable with the largest amount of missing data and tends to miss in blocks of many observations.

From the Little's test (chi-square, χ^2, of 850.89, with 91 degrees of freedom and p-value = 0.000), it can

Figure 1 - Histogram in decreasing order of missing values (a) and missing data pattern (white correspond to observed values and gray missing values) (b) from a soil physical-hydric database. (OM - organic matter, PWP - permanent wilting point, Micro - Microporosity, Macro - macroporosity, FC - Field capacity, Bd - soil bulk density, Pd - particle density, TP - total porosity)



be stated, at a significance level of 5%, that the missing data is not MCAR. Therefore, in this study we assume the MAR (*missing at random*) mechanism data.

The results of the predictive models (multiple linear regressions) for the five soil properties with missing data (FC, PWP, Macro, Micro and MO) considering the complete case analysis (CCA) and multiple imputation (MI-MICE) are summarized in Table 1. The complete case results deviate notably from the imputed results.

Where the coefficients are zero mean that these variables have little significance to predict our variable of interest. The soil macroporosity was the only parameter to be imputed where none of the variables resulted in coefficient zero. In general, the soil bulk density was of most importance when predicting the outcomes variables.

The significance of the covariates (p-value) varied depending on the variable to be estimated (Table 1). For FC estimation, the covariates most related to soil structure (Micro, Macro and Ds) were significant in both methods. On the other hand, for the PWP model, those related to the texture (clay and sand content) were the most relevant. For the Macro estimation models, the significance varied according to the method used, only two covariates (Micro and FC) were more expressive for the estimation of this variable, when the CCA method was applied.

Table 1 - Parameters of estimates by complete case analysis (CCA) and multiple imputation by the MICE method (Multivariate Imputation by Chained Equations)

Covariates	Complete Case analysis (n=283)			Imputation – MICE (n=631)				
	β (standard errors)	IC [95%]	p-value	β (standard errors)	IC [95%]	p-value	FMI	λ
Intercept	63.4(39.3)	[-13.7; 140.5]	0.108	59.4(14.3)	[31.1; 87.6]	0.00	0.1	0.1
Microporosity	0.3(0.1)	[0.2; 0.4]	0.000**	0.2(0.0)	[0.1; 0.3]	0.00**	0.2	0.2
Macroporosity	-0.1(0.0)	[-0.2; 0.0]	0.004*	-0.2(0.0)	[-0.2; -0.1]	0.00**	0.3	0.3
Total porosity	-0.4(0.2)	[-0.7; 0.0]	0.051	-0.3(0.1)	[-0.5; -0.1]	0.017	0.2	0.2
Permanent wilting point	0.1(0.1)	[-0.2; 0.3]	0.608	0.1(0.1)	[-0.1; 0.2]	0.385	0.3	0.3
Soil bulk density	-21.4(6.7)	[-34.5; -8.3]	0.002*	-14.9(4.0)	[-22.8; -7.0]	0.000**	0.2	0.2
Clay	0.2(0.3)	[-0.5; 0.8]	0.585	0.1(0.1)	[0.0; 0.3]	0.091	0.0	0.0
Silt	0.1(0.3)	[-0.6; 0.7]	0.849	0.1(0.1)	[0.0; 0.3]	0.068	0.0	0.0
Sand	00(0.3)	[-0.7; 0.6]	0.945	-0.1(0.1)	[-0.3; 0.0]	0.145	0.0	0.0
Organic matter	00(0.4)	[-0.7; 0.8]	0.916	0.0(0.2)	[-0.3; 0.3]	0.963	0.3	0.2
Intercept	28.7(7.0)	[15.0; 42.5]	0.000	32.2(7.8)	[16.1; 48.3]	0.000	0.4	0.4
Microporosity	0.0(0.0)	[0.0; 0.1]	0.201	0.0(0.0)	[0.0; 0.1]	0.334	0.7	0.7
Macroporosity	0.0(0.0)	[-0.1; 0.0]	0.265	0.0(0.0)	[-0.1; 0.0]	0.083	0.6	0.5
Field capacity	0.0(0.0)	[0.0; 0.1]	0.135	0.0(0.0)	[0.0; 0.1]	0.398	0.2	0.2
Particle density	-2.2(1.7)	0.0;0.1]	0.196	-4.3(2.0)	[-8.5; -0.1]	0.045	0.6	0.5
Soil bulk density	2.1(0.9)	[-5.6; 1.1]	0.026*	2.2(0.9)	[0.3; 4.1]	0.024*	0.3	0.3
Clay	0.2(0.0)	[0.1; 0.3]	0.001**	0.2(0.1)	[0.1; 0.3]	0.000**	0.1	0.1
Silt	-0.1(0.0)	[-0.1; 0.0]	0.226	-0.1(0.0)	[-0.2; 0.0]	0.192	0.0	0.0
Sand	-0.3(0.0)	[-0.4; -0.2]	0.000**	-0.2(0.1)	[-0.3; -0.1]	0.000**	0.1	0.1
Intercept	27.0(23.2)	[-18.6;72.5]	0.247	17.2(20.0)	[-22.1; 56.5]	0.497	0.1	0.1
Microporosity	-0.4(0.1)	[-0.5; -0.3]	0.000**	-0.4(0.1)	[-0.5; -0.3]	0.000**	0.5	0.5
Total porosity	0.2(0.2)	[-0.1; 0.6]	0.241	0.4(0.2)	[0.1; 0.7]	0.005**	0.1	0.1
Permanent wilting point	-0.1(0.1)	[-0.4; 0.1]	0.245	-0.3(0.1)	[-0.6; 0.0]	0.050*	0.6	0.6
Field capacity	-0.2(0.1)	[-0.4; -0.1]	0.001**	-0.3(0.1)	[-0.5; -0.2]	0.000**	0.3	0.3
Soil bulk density	-12.5(6.7)	[-25.6;0.6]	0.063	-4.5(5.3)	[-15.0; 6.0]	0.496	0.1	0.1
Clay	0.2(0.1)	[0.0; 0.4]	0.058	0.2(0.1)	[0.0; 0.5]	0.032*	0.0	0.0
Silt	0.1(0.1)	[-0.1; 0.3]	0.387	0.1(0.1)	[-0.1; 0.3]	0.343	0.0	0.0
Sand	0.2(0.1)	[-0.1; 0.4]	0.128	0.1(0.1)	[-0.1; 0.3]	0.359	0.0	0.0
Intercept	16.4(30.0)	[-9.1; 41.9]	0.585	44.0(13.2)	[16.2; 71.7]	0.004	0.5	0.5
Macroporosity	-0.2(0.0)	[-0.2; -0.1]	0.000**	-0.3(0.0)	[-0.4; -0.2]	0.000**	0.4	0.4

Continuation Table 1

Permanent wilting point	0.2(0.1)	[0.0; 0.4]	0.074	0.0(0.1)	[-0.3; 0.4]	0.840	0.8	0.7
Field capacity	0.3(0.1)	[0.2; 0.4]	0.000**	0.3(0.0)	[0.2; 0.4]	0.000**	0.2	0.2
Total porosity	0.0(0.1)	[-0.1; 0.1]	0.454	0.1(0.1)	[0.0; 0.2]	0.118	0.6	0.5
Clay	0.0(0.3)	[-0.6; 0.6]	0.956	-0.2(0.1)	[-0.5; 0.0]	0.083	0.5	0.4
Silt	0.1(0.3)	[-0.5; 0.7]	0.690	-0.2(0.1)	[-0.5; 0.0]	0.079	0.5	0.4
Sand	0.0(0.3)	[-0.6; 0.5]	0.901	-0.3(0.1)	[-0.6; 0.0]	0.035*	0.5	0.5
Organic matter	1.5(0.3)	[0.8; 2.2]	0.000**	1.1(0.2)	[0.7; 1.5]	0.000**	0.4	0.3
Intercept	4.9(7.2)	[-9.2; 19.0]	0.495	-2.9(9.2)	[-22.6; 16.8]	0.756	0.6	0.5
Microporosity	0.0(0.0)	[0.0; 0.1]	0.000**	0.1(0.0)	[0.0; 0.1]	0.014**	0.9	0.8
Total porosity	-0.1(0.2)	[-0.4; 0.2]	0.521	0.2(0.2)	[-0.2; 0.6]	0.378	0.6	0.5
Field capacity	0.0(0.0)	[0.0; 0.02]	0.751	0.00(0.0)	[-0.1; 0.1]	0.968	0.8	0.8
Particle density	2.2(3.3)	[-4.3; 8.7]	0.501	-5.1(4.0)	[-13.8; 3.6]	0.225	0.6	0.5
Soil bulk density	-5.5(5.9)	[-17.1; 6.1]	0.356	4.8(7.3)	[-11.0; 20.7]	0.523	0.6	0.5
Silt	0.1(0.0)	[0.06; 0.1]	0.000**	0.0(0.0)	[0.0; 0.1]	0.000**	0.4	0.3
Sand	0.0(0.0)	[0.0; 0.03]	0.004**	0.0(0.0)	[0.0; 0.03]	0.047*	0.5	0.4

β = coefficients, IC (95%) = [lower confidence interval; upper confidence interval]; * and ** (significant at 5 and 1% probability, respectively), FMI = fraction of missing information, λ= proportion of the total variance that is attributable to the missing data

The largest fraction of missing information (FMI) observed in this study was in the covariate Micro to estimate OM (FMI = 0.9), i.e., less statistical certainty for estimating this variable. The OM was also superior in the efficiency of the estimates (RE = 85%), with the missing information fraction (λ = 0.8) obtained with 5 imputations.

The standard error of the parameter estimate was √(1 + 0,8/5 = 1,08 times greater than the standard error with an infinite number of imputations. It is worth mentioning that the largest proportions of the total variance were associated with the variables that presented missing data when they were inserted as covariates to estimate the others.

Figure 2 graphically compares the distributions of the observed (blue) and imputed (red) variables across imputation models. The distributions are very similar. Field capacity (FC) and permanent wilting point (PWP) seem to deviate a little more from the observed data.

The comparison of the data distribution, for each variable, taking into account the original data and submitted to the complete case analysis (CCA) and after the multiple imputation (MICE) is showed in Figure 3. It is noteworthy how the multiple imputations via MICE maintained the same behavior as the original data, and distributions changes were observed for the CCA, especially for the Micro, FC and OM variables (Figures 3b, 3c and 3e), where the central boxes, that represent 50% of the data, were reduced. The medians were larger in CCA for Micro, FC and PWP.

The initial examination of the SDB for the missing data pattern (Figure 1) is important for the selection of the imputation method to be used (HONAKER; KING; BLACKWELL, 2011). According to (FIGUEREDO *et al*., 2000) the problem of missing data in multivariate analysis has implications that threaten the conclusions validity.

Although the CCA is not a MI method, it is a reference to verify the estimates variability (AUDIGIER; HUSSON; JOSSE, 2015). The largest variance proportion attributed to missing data (FMI) was observed for OM (Table 1) once that this covariant had the highest estimated rate of lost information (λ).

The covariates inserted into the initial model (Table 1) were all those that showed a correlation with the estimated variable. Van Buuren and Oudshoorn (2000) suggest that the number of predictors used for imputation should be as broad as possible, since a large set of predictors tends to make the assumption of MAR more likely.

The convergence of MICE sampling was confirmed by the probability density function graphs (Figure 2), which presented approximately the same distribution and, curves similarity confirms that the Gibbs sampler algorithm converges.

Finally, the imputation efficiency of the missing data was evident in the box-plots graphs (Figure 3), since the MICE imputation method presented data distribution behavior similar to the observed data, both asymmetry and dispersion, in comparison to the submitted data to CCA. That is, the IM method preserved the original SDB characteristics. The differences observed in the CCA indicate that this approach does not allow generalizations for the entire population target of interest.

Although the results obtained with the MICE application did not stand out in relation to CCA (similar means and standard deviations), the preservation of the original variability of the data already demonstrates that the MI application is an appropriate alternative to complete a database with missing information, mainly for multivariate analysis.

In Soil Science, this situation can be exemplified by the estimation of pedotransfer functions, which, obtained from a multivariate approach, they are used to estimate soil properties that are either of onerous determination or are unavailable (MINASNY; HARTEMINK, 2011) and, often, the available databases have gaps, reducing the sample size considerably. Exemplifying for this purpose the SDB used in this study, failure to fill in the missing data would result in the reduction of the original SDB from 631 to 283 and consequently the change in data variability, as demonstrated (Figure 1). This significant modification of the database would possibly result in different models from those obtained for the complete bank, leading to the inaccuracy of the results.

When the number of cases available for multivariate analysis is decreased, the statistical power to detect significant effects is reduced, potentially

Figure 2 - Probability density functions of the observed data (blue line) and the five chains generated by MICE (red line) for the variables field capacity (FC), permanent wilting point (PWP), macroporosity (Macro), microporosity (Micro) and organic matter (OM)

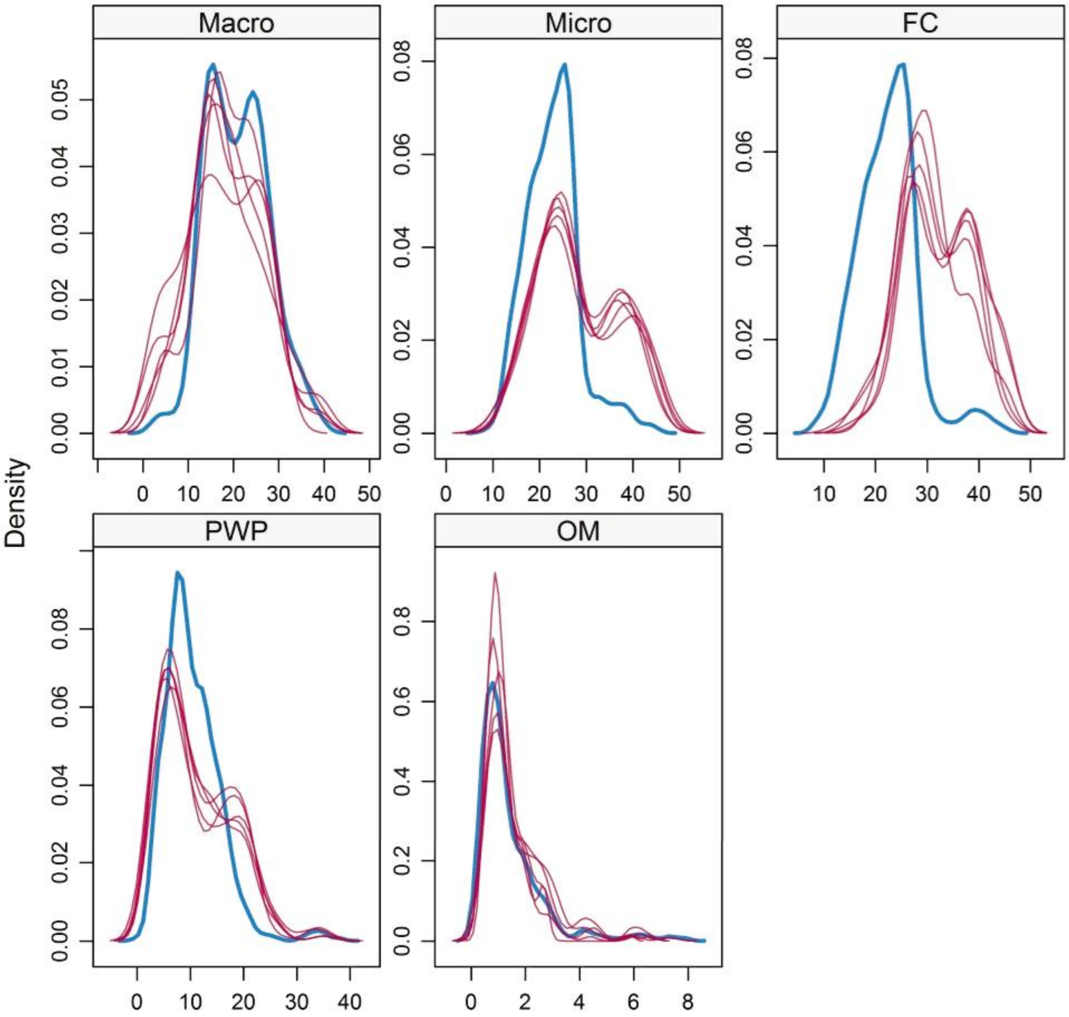
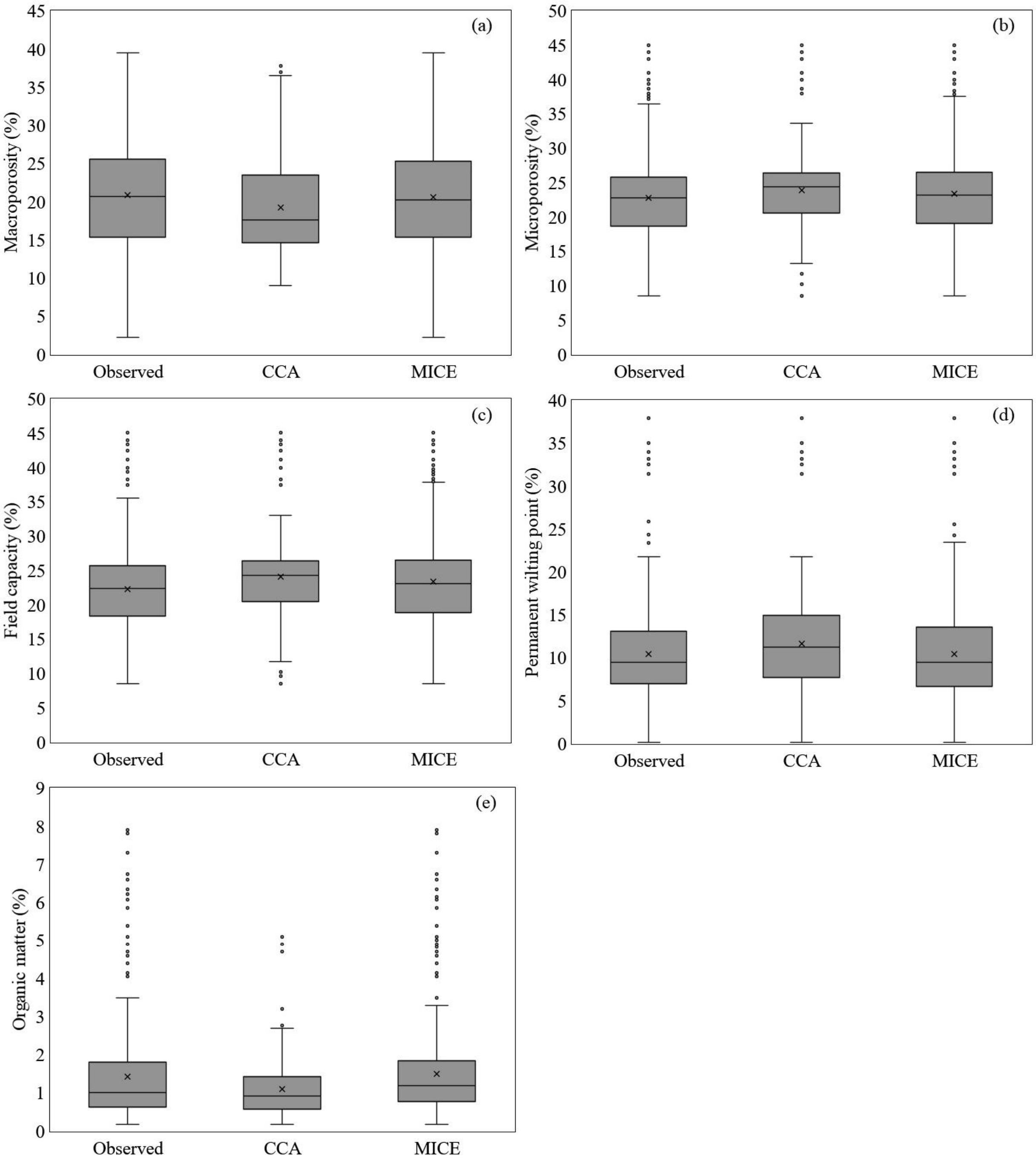


Figure 3 - Box-plot of the variables macroporosity (a), microporosity (b), field capacity (c), permanent wilting point (d) and organic matter (e) with original data, multiple imputation via MICE (Multivariate Imputation by Chained Equations)



leading to Type II error. The chances of Type II error increase when the original study sample is small, as it can occur in experimental studies assessing the treatment effectiveness. The main problem in listwise deletion (CCA) is whether the remaining sample size is sufficient to provide adequate statistical power, once missing data may cause the exclusion of much of the original data (FIGUEREDO *et al*., 2000).

Despite the methodological advances and demonstrations of the efficiency of MI in several areas (SQUILLANTE JÚNIOR *et al*., 2018) in Soil Science, this approach is still underutilized to deal with missing data. This work evidenced the advantages of this technique for the estimation of soil physico-hydrical properties data. Therefore, we understand that the results observed here can be used in studies with similar dataset. In this case, we recommend that the MI-MICE method be preferred over CCA. Since that analyzing just the complete cases, results in smaller sample sizes, that is, loss of information, with less statistical accuracy in the analyzes (NUNES; KLÜCK; FACHEL, 2009).

CONCLUSIONS

1. This paper has hypothesized that the imputation of missing data provides a statistically more accurate database than the complete case analysis. The results showed that the multiple imputation by chained equations presented better performance than the complete case analysis, since, although the statistical comparison of the two methods was similar, multiple imputation maintains the size of the database and preserves the general distribution;
2. Imputation data shows to be a fruitful approach for further studies in soil science, especially to deal with openly available soil database. Therefore more analysis needs to be carried out in order to validate the approach efficiency. With this study, we aim to help more soil researchers to get started with implementing multiple imputations techniques, such as Multivariate Imputation by Chained Equations, instead of inferior approaches in order to improve statistical analysis accuracy.

REFERENCES

AUDIGIER, V.; HUSSON, F.; JOSSE, J. Multiple imputation for continuous variables using a Bayesian principal component analysis. **Journal of Statistical Computation and Simulation,** v. 86, p. 2140-2156, 2015. DOI: http://dx.doi.org/10.1080/00949655.2015.1104683.

CARVALHO, J. R. P. *et al*. Modelo de imputação múltipla para estimar dados de precipitação diária e preenchimento de falhas. **Revista Brasileira de Meteorologia**, v. 32, p. 575-583, 2017. DOI: http://dx.doi.org/10.1590/0102-7786324006.

CLAESSEN, M. E. C. *et al*. **Manual de métodos de análise de solo**. Rio de Janeiro: Embrapa, 1997. 212 p.

CLIFFORD, D.; DOBBIE, M. J.; SEARLE, R. Non-parametric imputation of properties for soil profiles with sparse observations. **Geoderma**, v. 232/234, p. 10-18, 2014. DOI: https://doi.org/10.1016/j.geoderma.2014.04.026.

FIGUEREDO, A. J. *et al.* Multivariate modeling of missing data within and across assessment waves. **Addiction**, v. 95, p. 361-380, 2000. DOI: https://doi.org/10.1080/09652140020004287.

GRAHAM, J. W. Missing data analysis: making it work in the real world. **Annual Review of Psychology**, v. 60, p. 549-576, 2009.

HARRELL, J. F. E. **Regression modeling strategies**: with applications to linear models, logistic and ordinal regression, and survival analysis. 2. ed. New York: Springer International Publishing, 2016. 572 p.

HONAKER, J.; KING, G.; BLACKWELL, M. Amelia II: a program for missing data. **Journal of Statistical Software**, v. 45, p. 1-47, 2011. DOI: http://dx.doi.org/10.18637/jss.v045.i07.

KIM, M. *et al*. Comparative studies of different imputation methods for recovering streamflow observation. **Water Resource Research**, v. 7, p. 6847-6860, 2015. DOI: http://dx.doi.org/10.3390/w7126663.

KIEHL, E. J. **Manual de edafologia**: relações solo-planta. São Paulo: Ceres, 1979. 264 p.

LITTLE, R. J.; RUBIN, D. B. Missing data. **International Encyclopedia of the Social and Behavioral Sciences**, v. 15, p. 602-607, 2015. DOI: http://dx.doi.org/10.1016/B978-0-08-097086-8.42082-9.

LITTLE, R. J. A. A test of missing completely at random for multivariate data with missing values. **Journal of the American Statistical Association**, v. 83, p. 1198-1202, 1988. DOI: https://doi.org/10.2307/2290157.

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. **Caderno de Saúde Pública**, v. 25, n. 2, p. 268-278, 2009. DOI: http://dx.doi.org/10.1590/S0102-311X2009000200005.

MINASNY, B.; HARTEMINK, A. E. Predicting soil properties in the tropics. **Earth-Science Reviews**, v. 106, p. 52-62, 2011. DOI: http://dx.doi.org/10.1016/j.earscirev.2011.01.005.

PAES, Â. T.; POLETO, F. Z. Por dentro da estatística. **Educação Continuada em Saúde Einstein**, v. 11, p. 5-7, 2013.

PEDERSEN, A. B. *et al*. Missing data and multiple imputation in clinical epidemiological research. **Clinical Epidemiology**, v. 9, p. 157-166, 2017. DOI: https://doi.org/10.2147/CLEP.S129785.

POYATOS, R. *et al*. Gap-filling a spatially explicit plant trait database: comparing imputation methods and different levels of environmental information. **Biogeosciences**, v. 15, p. 2601-2617, 2018. DOI: https://doi.org/10.5194/bg-15-2601-2018.

R CORE TEAM. **R**: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2017. Disponível em: https://www.R-project.org/. Acesso em: 15 jan. 2018.

RUBIN, D. B. Inference and missing data. **Biometrika**, v. 63, p. 581-592, 1976. DOI: http://dx.doi.org/10.2307/2335739.

RUBIN, D. B. **Multiple imputation for nonresponse in surveys**. New York: John Wiley & Sons, 1987. 253 p.

SANTOS, H. G. *et al*. **Sistema brasileiro de classificação de solos**. 5. ed. rev. ampl. Brasília, DF: Embrapa Solos, 2018. 356 p.

SCHAFER, J. L.; OLSEN, M. K. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. **Multivariate Behavioral Research**, v. 33, p. 545-571, 1998. DOI: https://doi.org/10.1207/s15327906mbr3304_5.

SHAO, J.; MENG, W.; SUN, G. Evaluation of missing value imputation methods for wireless soil datasets. **Personal and Ubiquitous Computing**, v. 21, p. 113-123, 2017. DOI: https://doi.org/10.1007/s00779-016-0978-9.

SILVA, A. C.; ARMINDO, R. A. Importância das funções de pedotransferência no estudo das propriedades e funções hidráulicas dos solos do Brasil. **Multi-Science Journal**, v. 1, p. 31-37, 2016.

SONG, Q.; SHEPPERD, M. A new imputation method for small software project data sets. **Journal of Systems and Softwares**, v. 80, p. 51-62, 2007. DOI: https://doi.org/10.1016/j.jss.2006.05.003.

SQUILLANTE JÚNIOR, R. *et al*. Modeling accident scenarios from databases with missing data: a probabilistic approach for safety-related systems design. **Safety Science**, v. 104, p. 119-134, 2018. DOI: https://doi.org/10.1016/j.ssci.2018.01.001.

STUART, E. A. *et al*. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. **American Journal of Epidemiology**, v. 169, n. 9, p.1133-1139, 2009.

VAN BUUREN, S. **Flexible imputation of missing data**. 2. ed. Boca Raton: Chapman and Hall: CRC Press, 2018, 416 p.

VAN BUUREN, S.; OUDSHOORN, C. G. M. **Multivariate imputation by chained equations**: MICE V1.0 user´s manual. Leiden: TNO Preventie en Gezondheid, TNO/PG/VGZ/00.038, 2000.

WAXMAN, S. A.; STEVENS, K. R. A critical study of the methods for determining the nature and abundance of soil organic matter. **Soil Science**, v. 30, p. 97-116, 1930.