Acta Scientiarum

http://www.periodicos.uem.br/ojs/
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v46i1.66185

Data variability in the imputation quality of missing data

Elisandra Lúcia Moro Stochero[1], Alessandro Dal'Col Lúcio[2] and Luciane Flores Jacobi[3]*

[1]Secretaria de Educação, Prefeitura Municipal de Santa Maria, Rua Alameda Montevideo, 313, Edifício Sobral Pinto, 1° e 2° andares, 97010-004, Santa Maria, Rio Grande do Sul, Brazil. [2]Departamento de Fitotecnia, Universidade Federal de Santa Maria, Santa Maria, Rio Grande do Sul, Brazil. [3]Departamento de Estatística, Universidade Federal de Santa Maria, Santa Maria, Rio Grande do Sul, Brazil. *Author for correspondence. E-mail: luciane.jacobi@ufsm.br

ABSTRACT. Imputation methods were developed to define estimates for missing data and hence solve possible problems generated by the loss of this information. This study aims to assess whether data variability influences the results obtained after applying an imputation method. Incomplete databases were generated from complete real databases of experiments of tomato plants conducted using the randomized block design with three replications and 12 treatments by removing different amounts of data. The evaluated variables consisted of fruit weight per plant, number of fruits per plant, and average fruit length and width, forming eight balanced databases. Subsequently, the distribution-free multiple imputation method was applied, generating complete databases from imputation. The number of missing information influenced the accuracy measures for the data in this study. Data imputation was inadequate when there was high variability but more precise and accurate in cases of low variability. It confirmed the importance of assessing data variability before choosing to apply the imputation method.

Keywords: missing data; data imputation; randomized block design; distribution-free multiple imputation.

Received on December 6, 2022.
Accepted on May 10, 2023.

Introduction

Loss of information is common during the data collection process, which can interfere with data analysis results or even prevent the application of statistical methods that require a complete database. Missing data is a common problem faced by researchers and can occur for many reasons (Eze & Chukwunenye, 2019). According to Austin, White, Lee, and Buuren (2021), some possible situations for missing data include non-response of units, which occurs when the data collection procedure fails, and longitudinal studies in which participants may be present in some data collections and absent in others. The absence of certain data can lead to an incorrect analysis if it is not considered.

The main advances in research involving missing data emerged in the 1970s with the maximum likelihood estimation and multiple imputations (Enders, 2010). Moreover, according to the author, Rubin was responsible for establishing a classification system for missing data problems.

In order to find the conditions and support researchers in the decision to consider or not the process that causes the absence of data, Rubin (1976) presented examples of these processes and the conditional distribution that corresponds to the process that generated the data absence in each case. Little and Rubin (1987) presented the multiple imputation method. According to Pedersen et al. (2017), studies related to this topic have been carried out since then, and new methods have been developed and applied in various fields.

According to Banzatto and Kronka (2013), there is the possibility of reaching the end of an experiment in Agrarian Sciences and facing the loss of value of one or more plots, called missing plots. Among the possible causes, the authors mentioned plant death or disease, failure of the experimenter in collecting the data, loss of notes about the plot, a value very discrepant from the others and must be discarded, and a plot presenting a very doubtful value.

Caution is needed to define whether it actually is missing data. A study aiming at harvesting the fruits of each plot, for example, may present three situations: the plant died; the plant did not die but there is no fruit; and the plant did not die but the fruit is not ready to be harvested. These situations cannot be classified as missing data, given that the information in the empty cell is not unknown but null; there is true information, but it is simply not possible to identify it (Schafer & Graham, 2002).

Importantly, some statistical procedures were developed to work with complete data sets, such as the area of data mining and machine learning (Yu, Zhou, Chen, & Lai, 2020), which generate challenges for the researcher when empty cells are found. Initially, according to Enders (2010), the most common procedure was to address these situations considering ad hoc techniques, "manipulating" the data set even before performing the analysis.

According to Salgado, Azevedo, Proença, and Vieira (2016), dealing with missing data is necessary, either by deleting incomplete observations or by replacing any missing values with an estimated value based on the other information available, a process called imputation. Both methods can significantly affect the conclusions that can be drawn from the data. One of the ad hoc techniques applied is known as listwise exclusion, in which the general removal of the cases with missing data is performed (Lall, 2016). Depending on the sample size and the number of variables, it may result in a significant reduction in the sample size available for the data analysis. According to Kang (2013), missing data can reduce the statistical power of a study and produce biased estimates, leading to invalid conclusions.

Banzatto and Kronka (2013) mentioned possible situations in which missing plots occur, as previously commented, and indicated a way to work with these losses, i.e., the procedure that may be adopted to calculate an estimation for the missing plot. The analysis is straightforward with the complete database, and using complex methods in the presence of incomplete, unbalanced blocks is unnecessary.

Other methods can be found in the literature, some simple and others complex, whose application is currently accessible due to computational advancement. Although the number of research efforts in this context has been growing, especially in health, there are not many studies considering imputation methods in Agrarian Sciences, in which statistical methods of analysis and experimental design are distinct, with the number of publications in this context directed at unbalanced experiments being even less significant. Deepening and disseminating the study and application of imputation methods, together with the concern to obtain accurate and quality results, is interesting and of utmost importance (Jinubala & Jeyakumar, 2021). In fact, data loss can also occur in agricultural experimentation due to erasures, failure in filling in, or even actual loss of experimental units.

For instance, Peng, Lei, and Junyi (2022) and Boomgard-Zagrodnik and Brown (2022) evaluated specific imputation methods in the Agrarian Sciences within the context of missing data, that is, k-nearest neighbor algorithm and machine learning imputation, respectively. Peng et al. (2022) concluded that the k-nearest neighbor algorithm shows a good and stable performance when the missing data rate is lower than 10% and can meet the usage requirements. However, these studies do not have as their main focus the influence that the characteristics of the observed data may present on the results of the imputation methods, with the values of the database being used to determine the estimates for the missing data.

Therefore, this study aimed to verify whether the variability of data stemming from one experiment with a randomized block design (RBD) influences the imputation quality of missing data.

Material and methods

A method proposed by Bergamo, Dias, and Krzanowski (2008), which has as a starting point the singular value decomposition (SVD) for simple imputation, developed by Krzanowski in 1988, was applied to determine the values to be imputed. SVD was applied without making any assumption regarding the data distribution or structure, serving as a basis to determine the dimensionality of a multivariate data set, where a matrix Y_{nxp} is factored as $Y = UDV^T$, U is the matrix formed by the eigenvectors, D is the diagonal matrix formed by the eigenvalues of Y^TY, and V^T is the transposed matrix of the that formed by the eigenvectors of YY^T.

Matrices Y^TY and YY^T have the same eigenvalues and the elements d_i are the square roots of the eigenvalues. The i-th column $v_i = (v_{i1},\cdots, v_{ip})^T$ of the matrix is the eigenvector corresponding to the i-th largest eigenvalue d_i^2 of Y^TY. In the matrix U_{nxp}, the j-th column $u_i = (u_{i1},\cdots, u_{ip})^T$ is the eigenvector corresponding to the j-th largest eigenvalue d_j^2 of YY^T. Therefore, the decomposition of Y may be given as:

$$y_{ij} = \sum_{h=1}^{p} u_{ih} d_h v_{jh}$$

Bergamo et al. (2008) proposed a generalization for the exponents of the largest eigenvalue for the distribution-free multiple imputation method. The following expression was used as a performance measure of the method at the position of the missing value (absent in the row and column), where vo_l is the original value randomly eliminated at this position and $\hat{y}_{ij_{(m)}}$ is the m-th imputation at this position.

$$acc_l = \frac{\sum_{m=1}^{M}(\hat{y}_{ij_{(m)}} - vo_l)^2}{M-1}$$

This expression is calculated for $l = 1,2,\cdots, na$, where na is the total number of missing values. The expression may be separated into two terms, where $\underline{Y_l}$ is the actual value imputed at position l, the first term represents a variance over the values M at each position, and the second term represents a bias in the final imputation. Thus, the first term is a measure of precision that refers to the random errors and the second is a measure that refers to the systematic error at position l.

$$acc_l = \frac{\sum_{m=1}^{M}(\hat{y}_{ij_{(m)}} - \underline{Y_l})^2}{M-1} + \frac{M(\underline{Y_l} - vo_l)^2}{M-1}$$

A general performance measure T_{acc} may be calculated through the mean of the measures acc_l, where $na = g \times e \times porc$, with g representing the total number of genotypes, e representing the total number of environments, and $porc$ representing the percentage of missing data.

$$T_{acc} = \frac{\sum_{l=1}^{na} acc_l}{na}$$

T_{acc} can be divided into two components:

$$T_{acc} = V_E + VQM$$

where:

$$V_E = \frac{1}{na}\sum_{l=1}^{na}\left[\frac{\sum_{m=1}^{M}(\hat{y}_{ij_{(m)}} - \underline{Y_l})^2}{M-1}\right]$$

and

$$VQM = \frac{1}{na}\sum_{l=1}^{na}\frac{M(\underline{Y_l} - vo_l)^2}{M-1}$$

The first component V_E represents the variation grouped among imputations within positions; therefore, the higher the value is, the lower the precision of the multiple imputation method. However, a small value for this component does not necessarily mean that the imputation method is good since the method may be polarized. The second component VQM represents the mean squared bias between the values of $\underline{Y}$ and vo; therefore, the smaller the bias is, the higher the number of imputations that are similar to the original values and the better the precision. Thus, the smaller the values of V_E and VQM are, the better the multiple imputation method.

The distribution-free multiple imputation method was applied to the actual datasets of a balanced experiment belonging to the Sector of Plant Experimentation of the Federal University of Santa Maria. The databases have information on two experiments performed to verify the assumptions of the mathematical model and assess the effect of the application of a potato bioproduct on the productivity, fruit quality, and leaf color of tomato plants. One of the experiments was conducted in a plastic tunnel and the other in the field, using a random block design with three replicates and 12 treatments.

Two experiments using the salad tomato, hybrid Grandeur, were conducted at the Department of Plant Science at the Federal University of Santa Maria (latitude 29°43′ S, longitude 53°43′ W, and 95 m altitude), in Santa Maria, Rio Grande do Sul State, Brazil. The regional climate, according to the Köppen classification (Moreno, 1961), is Cfa (humid subtropical with no defined dry season and hot summers), and the soil is classified as an arenic dystrophic Red Argisol (Santos et al., 2006).

The experiments were conducted simultaneously during the spring-summer season (P-V), from August 16, 2010, to January 27, 2011. One experiment was conducted in a plastic tunnel 3.5 m high in the central part, 25 m long, and 4 m wide, with a useful area of 19.2 m long and 3.6 m wide, covered with a 100-micron low-density polyethylene (LDPE) film, with anti-UV additive, and north-south orientation. The other experiment was conducted in the field.

Seedlings were transplanted with four true leaves and arranged in three rows (beds without mulching) with 0.15 m high and 0.40 m wide and drip irrigation. The spacing was 0.8 m between plants and 1.2 m between rows, with a total of 24 plants per row.

Fertilization at transplanting was performed with 65 kg N ha^{-1}, 230 kg P_2O_5 ha^{-1}, and 65 kg K_2O ha^{-1}. Top dressing fertilization was carried out 21 days after transplanting, applying 35 kg N ha^{-1} and 35 kg K_2O ha^{-1}. Subsequently, additional top dressing fertilization was applied at 15-day intervals, totaling seven applications during the crop cycle, each consisting of 30 kg N ha^{-1} and 30 kg K_2O ha^{-1}. All fertilizations, as well as liming, were performed according to the results of soil analysis per row (Sociedade Brasileira de Ciência do Solo, 2004). The plants were grown on a single stem and all other cultural treatments were performed according to the crop recommendations (Filgueira, 2008).

The plot consisted of two plants in the direction of the planting row. A randomized block design with three replications and 12 treatments was used. The treatments consisted of the combination of two application intervals of Acrescent Solus® (after all fruit harvests and in alternate harvests) with the recommended mineral top dressing fertilization plus the doses of 1, 2, 3, and 4 L ha^{-1} of Acrescent Solus® applied to the soil, and (T1) mineral top dressing fertilization (without Acrescent Solus®), (T2) mineral top dressing fertilization plus 50 L ha^{-1} of Acrescent Solus® applied at 30 and 60 days after transplanting, (T3) replacement of mineral top dressing fertilization with the 100 L ha^{-1} of Acrescent Solus® applied every 15 days, and (T4) mineral top dressing fertilization plus 0.5 L ha^{-1} of Acrescent Solus® applied after all harvests.

Ten harvests were performed (from 11/11/2010 to 01/27/2011) and the following traits were evaluated at each harvest: fruit weight per plant, using a digital scale with a 1-g precision; number of fruits per plant; and average fruit length and width, measured with a caliper with a 1-mm precision. Among the ten harvests, the third met all the assumptions of the mathematical model, both in the field experiment and in the plastic tunnel. Thus, it was chosen for this approach.

Only the complete databases obtained upon the third harvest conducted on the field and the third harvest in the plastic tunnel were considered in the present study. Fruit weight per plant (g), number of fruits per plant, and average fruit length and width (cm) were assessed at the harvests. Thus, eight balanced databases were formed with actual data, with the adopted notations to differentiate them shown in Table 1.

Table 1. Variables assessed in each database and the notation adopted as a reference for each variable.

Experiment with tomato	Analyzed variables	Notation
In the field	Fruit weight per plant (g)	D1
	Number of fruits per plant (g)	D2
	Fruit length (mm)	D3
	Average fruit width (mm)	D4
In the tunnel	Fruit weight per plant (g)	D5
	Number of fruits per plant (g)	D6
	Fruit length (mm)	D7
	Average fruit width (mm)	D8

The positions of observations to be excluded from all balanced databases were randomly determined after organizing the databases with columns being blocks and rows being treatments. Thus, three new unbalanced databases were generated from each initial complete database, with the exclusion of 5, 15, and 30% of the observations, which are percentages adopted by Bergamo et al. (2008). After rounding, it resulted in the removal of one, five, and ten observations, respectively, in the present study.

The positions of the observations taken were the same for all variables, and the values for the exponents were the same used in the study by Bergamo et al. (2008). Subsequently, the distribution-free multiple imputation method was applied, and the "new" database, now incomplete, allowed determining the estimates for each missing data, being compared with the respective actual values that were taken at the first point of this process, following the same steps used by Bergamo et al. (2008).

The analysis of data variability and precision of the obtained results started with the initial complete databases, the unbalanced databases, and the databases completed with the imputation method. The variability was determined considering the coefficient of variation of the data and the mean of all observations of each database and each column (blocks), given that the mean of each column is the starting point to develop the imputation method.

The experimental coefficient of variation, which relates to the standard deviation in terms of the percentage of the arithmetic mean, was also verified. At first, the coefficient of variation of the values present in the databases is shown, and then the experimental coefficient of variation.

According to Banzatto and Kronka (2013), this coefficient is used to compare the variability of one's results with that obtained by researchers who work with similar materials, and it refers to an idea of the precision of the experiment: the lower the coefficient is, the better the precision. This coefficient is given as follows when a data set is assessed:

$$CV = \frac{s}{\hat{m}} \cdot 100$$

where $s = \sqrt{Q.M.Res.}$ is the standard deviation, $\hat{m} = \frac{G}{IJ}$ is the estimate of the mean, $Q.M.Res.$ is the mean square of residuals, $G = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}$, x_{ij} is the value observed in treatment i and block j, I is the number of treatments, and J is the number of blocks.

The coefficient of variation may be classified according to Table 2.

Table 2. Classification of the experimental coefficient of variation.

CV	Assessment	Precision
< 10%	Low	High
10% to 20%	Intermediate	Intermediate
20% to 30%	High	Low
> 30%	Very high	Very low

Source: Pimentel Gomes (1985).

Lastly, the results were compared to verify whether there was a difference between the results from the data with more or less variability and whether the imputation quality was better in some of the databases than others. These procedures were carried out using the software R Core Team (2017) and RStudio (2009-2017).

Results and discussion

The means, standard deviations, and coefficients of variation of each column (blocks) of each experiment could be obtained by generating incomplete and complete databases with 5, 15, and 30% of missing data and organizing them. These results were considered important in the analysis process because the mean of the respective column is entered to initiate the process of generating the databases imputed into the empty cells.

The results of the precision assessment measures V_E, VQM, and T_{acc} was obtained after the first steps. Table 3 shows these measures and compares them with the result of the coefficient of variation of the set of observations of the initial databases.

Table 3. Coefficients of variation of the set of observations of the initial database and performance measures of the imputed values.

Database****	Coefficient of variation	Precision measures								
		V_E*			VQM**			T_{acc}***		
	Database Initial Data	5%	15%	30%	5%	15%	30%	5%	15%	30%
D1	51%	18	12	7	519930	170525	243185	519948	170538	243192
D2	49%	0.00007	0.00004	0.00003	0.00001	2	6	0.00008	2	6
D3	11%	0.00354	0.00003	0.00004	354	23	33	354	23	33
D4	11%	0.01709	0.00056	0.00028	778	76	45	778	76	45
D5	49%	15	6	3	16802	713611	496391	16617	713616	496395
D6	43%	0.00055	0.00005	0.00002	50	6	3	50	6	3
D7	15%	0.00342	0.00236	0.00135	65	23	60	65	23	60
D8	16%	0.0004	0.00003	0.0001	1634	286	274	1634	286	274

*V_E: grouped variation among imputations within positions; **VQM: mean square bias between imputed and observed values; ***T_{acc}: sum of VE and VQM; ****Abbreviations of the database column are described in Box 1.

Among the eight actual complete (balanced) databases, D3, D4, D7, and D8 presented low coefficients of variation, while D1, D2, D5, and D6 presented higher coefficients of variation. Databases D1 and D5 presented a higher grouped variation among imputations, indicated by the VE measure and, therefore, the method precision was low. Moreover, the more observations were removed, the lower the VE value, resulting in better precision. According to Little and Rubin (2002), loss of precision depends not only on the fraction of complete cases and pattern of missing data but also on the extent to which complete and incomplete cases differ and the parameters of interest.

This measure has low values in D2, D3, D4, D6, D7, and D8, indicating that the variation among imputations is low and hence the precision is good. The results of the comparison of VE values with the

coefficients of variation show that the precision among imputations was good in all databases that presented low coefficients of variation. However, some databases had good precision among imputations and others showed low precision when the coefficient of variation was higher.

We assessed the databases D1, which presented low precision and a high coefficient of variation, D2, which presented good precision and a high coefficient of variation, and D3, which presented good precision and a low coefficient of variation, when verifying the causes of these differences. We removed 5% of the observations from different positions, i.e., the same column from which the observation was removed in the first execution of the imputation. The results (Table 4) were assessed after the removal of different observations. The results by Bleidorn, Pinto, Schmidt, Mendonça, and Reis (2022) indicate that any imputation methodology can be considered for 5% missing data.

Table 4. Precision values of results after removing different values from the same database.

Database	Position of the removed value	Removed value	Estimated value	V_E	VQM	T_{acc}
D1	(7;1)	1587	574.8	16.6204	1280678	1280694
	(8;1)	1216	917.76	5.1197	11186	111191
	(5;1)	600.5	1529.5	4.4209	1078938	1078942
	(2;1)	2054	1002	0.9085	1383265	1383266
D2	(7;1)	5.5	2.12	0.0008	14.2718	14.2726
	(8;1)	3.5	6.69	0.0006	12.7408	12.7415
	(5;1)	3	6.95	0.0007	19.5025	19.5033
	(2;1)	8	4.68	0.0001	13.0543	13.8055
D3	(7;1)	61	66.16	0.0005	33.2769	33.2774
	(8;1)	69.29	62.52	0.0001	57.3525	57.3527
	(5;1)	62	68.57	0.0022	487.1505	487.153
	(2;1)	58.88	75.4	0.0045	341.159	341.164

*V_E: grouped variation among imputations within positions; **VQM: mean square bias between imputed and observed values; ***T_{acc}: sum of VE and VQM; ****Abbreviations of the database column are described in Box 1.

The data have a high variability because the coefficients of variation (Table 3) of D1 and D2 are higher than 30% (Pimentel Gomes, 1985), showing some values with a considerable difference from the others. Therefore, we chose to remove higher and lower values to assess if they would influence the results. Table 4 shows that the lowest VQM values are for the data set D2, followed by D3, both with good accuracy in imputing the data, but one with a high and the other with a low coefficient of variation. According to Ni, Leonard, Guin, and Feng (2005), multiple imputations produce unbiased estimates for missing values and preserve the natural variability of the observed data.

As mentioned before, the fact that V_E is low does not guarantee that the method is good (Bergamo et al., 2008). High values were found when considering the results referring to VQM, which represents the mean squared bias between the values of $\underline{Y}$ and vo. Only three out of the 12 estimated values were similar to the original values (D3 at positions (7.1), (8.1), and (5.1), as shown in Table 4). It indicates that the number of imputations similar to the original values is small.

The experimental coefficients of variation of the original databases, unbalanced databases, and databases balanced after entering the estimates of the removed values were estimated to verify if it would influence the results of the experimental analysis (Table 5).

Table 5. Experimental coefficient of variation for the complete database, the incomplete database, and the database completed with imputation for the study data.

Database Data*	Coefficient of variation Database						
	Complete	Incomplete			With imputation		
		5%	15%	30%	5%	15%	30%
D1	49.86%	50.24%	52.17%	47.98%	49.60%	47.20%	42.65%
D2	49.99%	51.03%	52.03%	45.10%	49.99%	47.44%	39.00%
D3	10.78%	10.71%	11.61%	13.50%	10.70%	10.45%	10.55%
D4	11.62%	11.33%	11.85%	12.06%	11.20%	11.22%	11.90%
D5	48.08%	47.45%	50.00%	51.94%	47.69%	46.61%	40.46%
D6	39.63%	39.60%	42.02%	47.07%	43.76%	39.28%	36.34%
D7	15.29%	12.67%	12.78%	10.93%	14.19%	14.44%	12.82%
D8	15.35%	12.65%	12.88%	10.87%	12.42%	12.67%	10.55%

*Abbreviations of the database column are described in Box 1.

Based on the classification of the experimental coefficient of variation shown in Box 2, we have that D1, D2, D5, and D6 had higher coefficients of variation, indicating low precision. These coefficients underwent an increase in the unbalanced data and remained close to the values referring to the initial data when the estimates for the empty cells were imputed (Table 5). This difference may be caused by the sample size of the database with incomplete data being smaller than the database with complete and imputed data. According to Santos and Dias (2021), the upper limit of the coefficient of variation depends on the sample size.

In turn, the experimental coefficients found were low for D3, D4, D7, and D8, thus showing a high precision. The analysis of D3 and D4 shows that results remained very close to the original ones in the unbalanced data and the databases completed with the imputation.

The results show that good approximations are not always obtained for the actual values of the data that were removed using the applied method. Stochero, Jacobi, and Lúcio (2020) applied the imputation method and observed no better results compared to the analysis with the unbalanced database. In this case, the observed value and the imputed value were closer for the data with considerably low variability (CV lower than 20%). However, the results regarding precision were not always so satisfactory when considering the data with higher variability (CV higher than or equal to 20%).

The value to be imputed tended to be lower than the actual value when averaging the column at the initial starting point when a higher value was removed. In turn, the value to be imputed increased when the one removed was smaller. Thus, different results for the precision measures can be obtained depending on the value removed.

In the present study, initial, actual, and complete data were available, which allowed for comparison and verification that data variability influences the outcome of imputation. The application of a certain data imputation method must be carefully analyzed to verify if it will bring results with less bias or if it is more reliable to work with an unbalanced dataset when the analysis can be performed even under these conditions. We suggest that future studies verify if other imputation methods and experimental designs are influenced by data variability.

Conclusion

Data variability negatively influenced the results by applying the imputation method. Considering the variability of the assessed data is of utmost importance. Databases with low variability ensure imputed results closest to the actual ones. Resorting to imputation does not always provide appropriate precision when the number of pieces of missing information is relatively low.

References

Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2021). Missing data in clinical research: A tutorial on multiple imputation. *Canadian Journal of Cardiology*, *37*(9), 1322-1331 DOI: https://doi.org/10.1016/j.cjca.2020.11.010

Banzatto, D. A., & Kronka, S. N. (2013). *Experimentação agrícola* (4. ed.). Jaboticabal, SP: Funep.

Bergamo, G. C., Dias, C. T. S., & Krzanowski, W. J. (2008). Distribuition-free multiple imputation in an interaction matrix through singular value decomposition. *Scientia Agricola*, *65*(4), 422-427. DOI: https://doi.org/10.1590/S0103-90162008000400015

Bleidorn, M. T., Pinto, W. P., Schmidt, I. M., Mendonça, A. S. F., & Reis, J. A. T. (2022). Methodological approaches for imputing missing data into monthly flows series. *Revista Ambiente & Água*, *17*(2), 1-27. DOI: https://doi.org/10.4136/ambi-agua.2795

Boomgard-Zagrodnik, J. P., & Brown, D. J. (2022). Machine learning imputation of missing Mesonet temperature observations. *Computers and Electronics in Agriculture*, *192*, 106580. DOI: https://doi.org/10.1016/j.compag.2021.106580

Enders, C. K. (2010). *Applied missing data analysis* (2. ed). New York, NY: The Guilford Press. Retrieved on July 12, 2021 from http://hsta559s12.pbworks.com/w/file/fetch/52112520/enders.applied

Eze, F. C., & Chukwunenye, V. G. (2019). Comparing methods of estimating missing values in one-way analysis of variance. *International Journal of Trend in Scientific Research and Development*, *3*(2), 994-1000. DOI: https://doi.org/10.31142/ijtsrd18599

Filgueira, F. A. R. (2008). Novo manual de olericultura: *agrotecnologia moderna na produção e comercialização de hortaliças*. Viçosa, MG: UFV.

Jinubala, V., & Jeyakumar, P. (2021). Methodologies for imputation of missing values in rice pest data. *Current Journal of Applied Science and Technology*, *40*(5), 64-73. DOI: https://doi.org/10.9734/cjast/2021/v40i531304

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal Anesthesiology*, *64*(5), 402-406. DOI: https://doi.org/10.4097/kjae.2013.64.5.402

Lall, R. (2016). How multiple imputation makes a difference. *Political Analysis*, *24*(4), 414-433. DOI: https://doi.org/10.1093/pan/mpw020

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. *Journal of Educational Statistics*, *16*(2), 150-155. DOI: https://doi.org/10.2307/1165119

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New Jersey, NY: John Wiley & Sons Inc.

Moreno, J. A. (1961). *Clima no Rio Grande do Sul*. Porto Alegre, RS: Secretaria da Agricultura.

Ni, D., Leonard, J. D., Guin, A., & Feng, C. (2005). Multiple imputation scheme for overcoming the missing values and variability issues in ITS data. *Journal of Transportation Engineering*, *131*(12), 931-938. DOI: https://doi.org/10.1061/(asce)0733-947x(2005)131:12(931)

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, *9*, 157-166. DOI: https://doi.org/10.2147/CLEP.S129785

Peng, W., Lei, Y., & Junyi, Z. (2022). Research on missing data filling method of wind power generation based on k-nearest neighbor algorithm. In *5th International Conference on Data Science and Information Technology (DSIT)*. Shanghai, CH, IEEE. DOI: https://doi.org/10.1109/DSIT55514.2022.9943846

Pimentel Gomes, F. (1985). *Curso de estatística experimental*. São Paulo, SP: Nobel.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, AT: R Foundation for Statistical Computing. Retrieved on July 12, 2021 from https://www.R-project.org/

RStudio Team (2009-2017). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc. Retrieved on July 12, 2021 from http://www.rstudio.com/

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. DOI: https://doi.org/10.1093/biomet/63.3.581

Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2016). Missing data. In *Secondary analysis of electronic health records*. Cambridge, US: Springer. DOI: https://doi.org/10.1007/978-3-319-43742-2_13.

Santos, C., & Dias, C. (2021). Note on the coefficient of variation properties. *Brazilian Electronic Journal of Mathematics*, *2*(4), 101-111. DOI: https://doi.org/10.14393/BEJOM-v2-n4-2021-58062

Santos, H. G., Jacomine, P. K. T., Anjos, L. H. C., Oliveira, V. A., Oliveira, J. B., Coelho, M. R., ... Cunha, T. J. F. (2006). *Sistema brasileiro de classificação de solos* (2. ed.). Rio de Janeiro, RJ: Embrapa Solos.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147-177. DOI: https://doi.org/10.1037/1082-989X.7.2.147

Sociedade Brasileira de Ciência do Solo. (2004). *Manual de adubação e de calagem para os Estados do Rio Grande do Sul e de Santa Catarina*. Núcleo Regional Sul. Porto Alegre, RS: Comissão de Química e Fertilidade do Solo - RS/SC.

Stochero, E. L. M., Jacobi, L. F., & Lúcio, A. D. (2020). Imputação de dados na análise de variância em experimentos no delineamento inteiramente casualizado. *Ciência e Natura*, *42*, 1-13. DOI: https://doi.org/10.5902/2179460X40446

Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2020). Missing data preprocessing in credit classification: One-hot encoding or imputation? *Emerging Markets Finance and Trade*, *58*(2), 472-482. DOI: https://doi.org/10.1080/1540496X.2020.1825935