MULTIPLE IMPUTATION MIGAMMI ALGORITHM

Pedro Marinho AMOÊDO[1]
Sônia Maria De Stefano PIEDADE [2]
Carlos Tadeu Dos Santos DIAS [2]
Sergio ARCINIEGAS-ALARCÓN[3]

- ABSTRACT: Missing data are common in multi-environmental experiments however sophisticated they are. Thus, it is essential to use appropriate methods of analysis to reduce the impact generated by the loss of information. Data imputation consists in one of the most common techniques used to overcome the problem of missing values, it estimates missing data by plausible values; subsequently, the analyses are carried out on the complete data. This work aims to propose a new multiple imputation method for data from multi-environment trials, resulting from the proposal based on the simple residuals of a linear regression model. Alterations were made in the simple imputation algorithm EM-AMMI to accommodate the additive main effect and generalized multiplicative interaction GAMMI. The quality of the multiple imputations method was evaluated by using accurate general statistics distributions, which combines the variance among imputation and mean square deviation, and normalized root mean square error (NRMSE). For such, simulations of random values at levels of 10%, 20%, 30% and up to 40% were performed from two real data set and the obtained corresponding imputations. The overall mean accuracy and NRMSE results, given the low values obtained, considering the proposed method, demonstrate the high quality of the proposed multiple imputation algorithm MIGAMMI.

- KEYWORDS: Generalized AMMI; analysis of deviance; Tacc statistics; biplot imputation

[1]Universidade Federal do Amazonas, ICSEZ, Departamento de Zootecnia, CEP: 69152-240, Parintins, AM, Brasil. E-mail: *pamoedo@ufam.edu.br*
[2]Universidade de São Paulo - USP, ESALQ, Departamento de Ciências Exatas, Caixa Postal 9, CEP: 13418-900, Piracicaba, SP, Brasil. E-mail: *soniamsp@usp.br; ctsdias@usp.br*
[3]Universidad de La Sabana, Facultad de Ingeniería, Campus Puente del Común, Km 7, Autopista norte de Bogotá, Chía-Colombia. E-mail: *sergio.arciniegas@unisabana.edu.co*

1 Introduction

Though multi-environment trials are planned to be balanced, missing values may occur, whether due to control failure, human errors, or natural conditions, such as excessive rains, plague attacks, animal invasion, etc (ARCINIEGAS-ALARCON *et al.*, 2020; YAN, 2013; RODRIGUES *et al.*, 2011; BERGAMO, 2007). Missing data produce unbalanced trials that prevent data from being directly analyzed by effective traditional statistical methods. A typical example are the cultivars studied in different environments, in which the variable response is the mean of repetitions in each combination of level factors. In such trials, the additive main effect and multiplicative interaction are the best analysis approach, provided the variable distribution response is normal, independent, and identically distributed (RODRIGUES *et al.*, 2016; HADI *et al.*, 2010). However, under the existence of missing data, the applicability of models AMMI is unviable (GAUCH; ZOBEL, 1990; YAN, 2013).

Several strategies are applied to solve the problem of missing values, which commonly occur in many types of multi-environment trials, for instance, by deleting rows or columns that present missing values, to obtain a balanced subset; by filling in missing data through environmental means (input column) or through estimates obtained by any method, such as linear models or multiplicative mixed model. Each of these procedures can be used, but none of them are simple or entirely effective. The first one produces even more losses, since during the obtaining of the complete subset, it tends to eliminate other values, dramatically reducing the sample, which may result in deviation of patterns. The second one may not be adequate since missing values may occur. The third one is rather complex and involves several stages, whether in algebraic field or in computationally implementation (YAN, 2013).

Some of the works well accepted in filling in missing data in multi-environment trials are the methods of imputation that employ singular value decomposition (SVD) of a matrix, such as the algorithm EM-AMMI presented by Gauch and Zobel (1990), in which the authors introduce the additive main effect and multiplicative interaction (AMMI) in algorithm EM (Expectation-Maximization) to perform imputation. In this algorithm, the best results are achieved by including few multiplicative terms in AMMI model (PIEPHO, 1995; ARCINIEGAS-ALARCON; DIAS, 2009; ARCINIEGAS ALARCON *et al.*, 2014; PADEREWSKI *et al.*, 2014). Also, the algorithm EM+SVD presented by Perry (2009), the distribution-free multiple imputation (DFMI) by Bergamo *et al.* (2008), a method with no restriction on patterns or data missing mechanism and free of assumptions about distribution or data structure, the Biplot imputation method described by Yan (2013) and others of equal importance who use SVD.

Imputation is the process of filling in missing data with plausible values for subsequent analysis. In general, the methods of imputation are classified in single imputation and multiple imputation (MI). In single imputation, missing data are imputed only once and then the completed data are analyzed as if there were

no missing values. As it occurs a single time, it is not possible to quantify the uncertainty associated to imputations, which might be a constraint of the single imputation (ENDERS, 2010; BERGAMO, 2007). In multiple imputation, missing values are replaced by **m** values, creating **m** datasets with imputed values (RUBIN, 1978, 1987). Usually, multiple imputation (MI) consists of three stages: imputation of missing values, analysis of **m** datasets created, and combination of results created in **m** analysis (ZHANG *et al.*, 2003; SCHOMAKER; HEUMANN, 2018). In MI, imputation certainties are incorporated to the results, which makes MI more attractive and efficient for filling in missing data (BERGAMO *et al.*, 2008; VAN GINKEL *et al.*, 2019).

Despite the existence, for decades, of methods that deal with missing values, this issue is yet not fully addressed, leading many researchers not to utilize appropriate methods. Due to lack of knowledge, in most cases, they utilize simple approaches of deletion or replacement (PEUGH and ENDERS, 2004). Similar results to Peugh and Enders (2004), were presented by Rousseau *et al.* (2012), who observed that, in over one-third of the reviewed works, it was found no indication of missing values; yet, in half of the works in which missing data were reported, the adopted method was not understandable; and among the correct ones, the majority merely performed simple deletion of observations. According to the author, researchers only use these methods as they are standard statistical packages.

Suitable procedures of imputation are far more advantageous than a simple elimination of missing units, since the maintenance of the entire sample might help preventing the increase of errors caused by the reduction of sample size, completed data may be analyzed by efficient classic methods available in usual statistic programs. Furthermore, if data is to be analyzed by distinct individuals, one imputation before analysis will guarantee that the same dataset will be used by each one, which enables comparisons of results. On the other hand, the imputation might not be well implemented, some methods might present deficiencies, being disadvantageous (SCHAFER and GRAHAM, 2002).

Therefore, by presenting the literary aspects about data imputation in multi-environment trials, we aim to propose an algorithm of multiple imputation based on an extension of EM-AMMI method and the simple residual of linear regression model, a combination of the generalized additive main effects and multiplicative interaction model (GAMMI) with an algorithm EM and the simple residual of linear regression.

2 Materials and methods

2.1 Algorithm of single imputation EM-GAMMI

To carry out data imputations based on EM-GAMMI procedure, in trials with genotype-by-environment interaction or G×E, modifications were performed in the algorithm EM-AMMI: 1) the additive main effects and multiplicative interaction model (AMMI) was replaced by the generalized additive main effects and

multiplicative interaction model (GAMMI), enabling the algorithm to model other distributions, besides normal distribution, such as, Poisson, binomial, among others; 2) the singular value decomposition (SVD) was suppressed, the imputed values are then obtained directly from the adjustment of GAMMI model with **k** multiplicative terms; in this paper, consider $\mathbf{k} = 0, 1, 2$. The procedure of modification was viable due to the algorithm Van Eeuwijk (1995), which uses Nelder e Wedderburn (1972) approach of generalized linear models (GLM), as a basis for estimating the generalized AMMI model. According to Amoêdo (2021), the stages of functioning of the imputation algorithm, named EM-GAMMI, are described as follows:

Step 1 - Missing elements $[x_{ij}^m]$ of $\mathbf{X}$ are initially estimated by the observed overall mean values, plus the mean in row i (row main effect), plus mean in column j (column main effect), obtaining a full matrix $\mathbf{X}$. The initial filling is also possible by using an arbitrary value. In this article, means were employed to estimate the initial missing values.

Step 2 - A particular GLM with a specific link function is defined, then the parameters of the model GAMMI are estimated. Complete column entries of $\mathbf{X}$ are considered as an environment factor and entry rows as effect of genotype factor for the adjustment. The generalized AMMI model (GAMMI) for the mean response μ_{ij} in terms of linear prediction, as described in (1),

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \sum_{k=1}^{K} \sqrt{\lambda}_k \gamma_{ik} \delta_{jk} \tag{1}$$

where μ is a general mean, α_i e β_j represent rows and columns effects, γ_{ik} and δ_{jk} are row and column values for the k^{th} multiplicative component of the interaction terms, $\sqrt{\lambda_k}$ means the singular values of k-th component, and K is the number of multiplicative terms

Step 3 - The adjusted mean is calculated based on the model GAMMI with **k** multiplicative terms. Depending on the number of multiplicative terms used, the imputation method can be nominated EM-GAMMI-0, EM-GAMMI-1, EM-GAMMI-2, ..., EM-GAMMI-K.

Step 4 - Missing values (x_{ij}^m) in $\mathbf{X}$ are filled in (imputed) by appropriate EM-GAMMI estimates, adjusted means $(\widehat{\mu}_{ij})$. As the relation between $E(Y_{ij})$=μ_{ij} and the linear predictor η_{ij} does not occur in a direct way in the generalized linear model, they are united by the link function, the predicted values are returned to data scale utilizing $g^{-1}(\eta_{ij})$. Notice that $g(\cdot)$ is a function that links the mean $E(Y_{ij}) = \mu_{ij}$ to the linear η_{ij}. If the link function is the identity, the model in (1) is the model AMMI itself. The expression in (2) shows the obtaining of (μ_{ij}).

$$g(\mu_{ij}) = \eta_{ij} \Rightarrow \mu_{ij} = g^{-1}(\eta_{ij}) = E(Y_{ij}) \tag{2}$$

Step 5 - Convergence criteria: if Chebyshev's distance between the estimation of missing values, in two stages of successive iteration, is greater than the assumed accuracy (used pattern 0,01), the stages from 2 to 5 shall be repeated; otherwise, the algorithm converges and stops. The distance of Chebyshev, considering two vectors containing p imputed values $\mathbf{X}$ and $\mathbf{Y}$, is defined as:

$$d(\mathbf{X}, \mathbf{Y}) = \mathbf{max}(|\mathbf{X_1} - \mathbf{Y_1}|, |\mathbf{X_2} - \mathbf{Y_2}|, \ldots, |\mathbf{X_p} - \mathbf{Y_p}|)$$

2.2 Multiple Imputation (MIGAMMI) using single residual in linear regression model

This proposal of multiple imputation was developed on the method EM-GAMMI, complementary to the use of simple residual in linear regression model, presented by Srivastava e Dolatabadi (2009) and Arciniegas-Alarcón *et al.* (2014). Arciniegas-Alarcón *et al.* (2014) carried out multiple imputation based on the "biplot imputation" method, using simple residual in linear regression model $\mathbf{Y} = \mathbf{Q}\boldsymbol{\beta} + \mathbf{E}$, where $\mathbf{Y}$ $(n \times 1)$ is the vector that represents the variable response; $\mathbf{Q}$ $(n \times p)$ represents the design matrix; $\boldsymbol{\beta}$ $(p \times 1)$ is the vector of regression parameters and $\mathbf{E}$ $(n \times 1)$ is the random error vector. In line with the authors, missing data only occur in vector $\mathbf{Y}$, the explanatory variables that compose the model must be complete. This way, the linear regression model was rewrote as $(\mathbf{Y}_0/\mathbf{Y}_A) = (\mathbf{Q}_0/\mathbf{Q}_A)\boldsymbol{\beta} + \mathbf{E}$, where $\mathbf{Y}_0$ $(n_1 \times 1)$ corresponds to the subvector of n_1 observed data, $\mathbf{Y}_A$ $(n_0 \times 1)$ to the subvector that contains n_0 missing values, $\mathbf{Q}_0$ $(n_1 \times p)$ the submatrix of n_1 observed data and $\mathbf{Q}_A$ $(n_0 \times p)$ the submatrix of n_0 missing values, so that $n_0 + n_1 = n$. Then, the multiple imputation is obtained by: $\widehat{\mathbf{Y}}_{\mathbf{At}} = \mathbf{Q_A}(\mathbf{Q_0^T Q_0})^{-\mathbf{1}}\mathbf{Q_0^T Y_0} + \mathbf{E_t}$, where $\mathbf{t} = \mathbf{1}, \ldots, \mathbf{m}$, represent $\mathbf{m}$ imputations for each missing data; and $\mathbf{E_t}$ refers to $\mathbf{t}$-th random sample with replacement of size n_0 obtained from the residual vector $\mathbf{e} = \left(\frac{\mathbf{n_1}}{\mathbf{n_1 - p}}\right)^{\mathbf{0,5}} (\mathbf{Y_0} - \mathbf{Q_0 b_1})$, then $\mathbf{b_1} = (\mathbf{Q_0^T Q_0})^{-\mathbf{1}}\mathbf{Q_0^T Y_0}$ represents the least square estimate of $\boldsymbol{\beta}$, based on the observed data.

To perform MI with simple residual, the proposed modification to the EM-GAMMI algorithm is the following: the method EM-GAMMI provides at the end of its process a complete matrix $\mathbf{X^c}$, whose elements are the imputed values for the respective missing data and the estimate for the observed values. Then, the next step consists of obtaining the simple residual matrix via observed data, by the difference between the original matrix and the matrix that contains the estimate of the observed values, that is, $\widehat{\boldsymbol{\epsilon}} = \mathbf{X} - \mathbf{X}^c$. As the residual is only obtained for the observed values, the matrix $\widehat{\boldsymbol{\epsilon}}$ (n×p) is incomplete since we can only obtain residual for $(np - na)$ observed data. Subsequently, and based on the residual matrix $\widehat{\boldsymbol{\epsilon}}$, $\mathbf{t}$ different matrices Ω_t (n×p) are created, where $\mathbf{t} = \mathbf{1}, \ldots, \mathbf{m}$, as follows: each element that compose Ω_t is randomly chosen and the matrix $\widehat{\boldsymbol{\epsilon}}$ is replaced. The process of random selection with replacement is repeated on $\widehat{\boldsymbol{\epsilon}}$, $\mathbf{m}$ times, producing $\mathbf{m}$ matrices, $\Omega_1, \Omega_2, \ldots, \Omega_m$. Once obtained Ω_t, the following step consists of performing the multiple imputation, which is made when replacing the missing

elements $[x_{ij}^m]$ of matrix $\mathbf{X}$, by correspondent values of each $\mathbf{t}$ matrices that are construed by $\mathbf{X}^c+\Omega_t$, then the process of MI provides $\mathbf{X}^c+\Omega_1$, $\mathbf{X}^c+\Omega_2$, ..., $\mathbf{X}^c+\Omega_m$ complete matrices. After obtaining the imputations, $\mathbf{t}$ complete matrices (observed and imputed) are combined by the mean of $\mathbf{t}$ completed matrices, originating a single matrix, then, the missing elements in original $\mathbf{X}$ are imputed with the correspondent obtained means.

In this paper, it was used $t = 5$, number of multiple imputations, since, according to Rubin (1996), $t = 5$ imputations are enough to make valid inferences. To Van Buuren (2018), $t = 5$ provides good quality to the method, and it is unlikely that important conclusions are substantially altered if the limit t is higher than 5. Therefore, it was obtained a multiple imputation with simple residuals by means of a generalized multiplicative model, which was named multiple imputation GAMMI (MIGAMMI).

2.3 Description of the data used in the research

To evaluate MI procedure, two real datasets were considered, complete and derived from trials with genotype × environment interaction. The first dataset is a randomized block design, a study based on the resistance of soybean to foliar plague, published by Hadi *et al.* (2010). In this experiment, four genotypes of resultant hybrid soybean were used (Wilis, IAC-100, IAC-80 e W-80) and, 14 days after planting, the counting of foliar plagues in each plant was examined. After counting, five types of foliar plagues were classified in varieties (genotypes of soybean). Table 1 presents the population mean of the five foliar plagues in four genotypes of soybean. This dataset was particularly chosen, since the mean responses to the repetitions are expressed in interval scale and analyzed by GAMMI methodology (HADI *et al.*, 2010). In this way, it was possible to use the models AMMI and GAMMI for posterior use of imputation algorithms. This dataset was denominated foliar plague dataset for reference purposes.

Table 1 - Population mean of the five foliar plagues in four soybean genotypes

Genotypes	Types of foliar plague				
	Bemissia	Emproosca	Agronyza	Lamprosema	Longitarsaus
IAC-100	0.50	1.75	2.25	0.50	1.75
IAC-80	3.00	2.75	1.00	1.75	3.25
W-80	3.50	4.00	1.25	2.00	2.00
Wilis	4.00	3.00	1.00	1.75	4.00

Source: Hadi *et al.* (2010)

The second dataset utilized is part of a study performed in a randomized block design ceded by the researchers Spitti *et al.* (2019). In the study, 19 beans genotypes were observed in six different environments. Genotypes were evaluated on the grains tegument color according to the luminosity value (L), and also in relation to shelf

growing conditions. The variable response demonstrates the genotype tolerance (resistance) for pigment losses, that is, gradual change of grains color at 60 days. Table 2 illustrates the genotype mean values per environment, obtained from the six regions considered in the study. This dataset was denominated Acácia for reference purposes.

Table 2 - Beans genotype mean evaluated on the grains tegument color according to the luminosity

Genotypes	Regions					
	R1	R2	R3	R4	R5	R6
BRS Pérola	0.5041	0.4727	0.5036	0.4497	0.4840	0.4957
CHC 01-175-1	0.4987	0.4648	0.5105	0.4610	0.4747	0.5013
CNFC 11-948	0.5068	0.4703	0.5023	0.4618	0.5048	0.5110
CNFC 11-954	0.5013	0.4585	0.4867	0.4708	0.4992	0.4961
Gen 4-1F-19P	0.5263	0.5000	0.4909	0.4892	0.5241	0.5245
Gen 12-2F-67	0.5178	0.4681	0.5021	0.4790	0.5098	0.5184
Gen 20-4F-129	0.5122	0.4847	0.4844	0.4494	0.4987	0.5343
Gen 45-2F-293P	0.5244	0.4922	0.5083	0.4792	0.5326	0.5493
Gen 78-1A-59	0.5078	0.4907	0.4950	0.4717	0.5168	0.5291
Gen 86-12A-122	0.5055	0.4776	0.4907	0.4501	0.4878	0.5215
Gen 90-4A-160	0.5106	0.4692	0.4993	0.4588	0.5002	0.5228
Gen 104-1A-291	0.5314	0.4901	0.5109	0.4677	0.5197	0.5304
Gen 106-4A-317	0.5107	0.4882	0.4999	0.4497	0.5016	0.5346
Gen 106-6A-319	0.5195	0.4794	0.5014	0.4856	0.5143	0.5226
Gen 107-14A-336	0.5256	0.4777	0.5145	0.4563	0.5348	0.5552
Gen 125-10A-510	0.5123	0.4670	0.5103	0.4756	0.4987	0.5183
IAC Milênio	0.5219	0.4803	0.5063	0.4873	0.5017	0.5111
IAC Sintonia	0.5028	0.4682	0.4821	0.4588	0.4899	0.5276
LP 11-363	0.5394	0.4810	0.5201	0.4703	0.5018	0.5144

Source: Spitti *et al.* (2019)

2.4 Simulation procedure based on real data

Both datasets used in this study were submitted to randomized removal simulations at 10%, 20% and 30% for foliar plague data, and at 10%, 20%, 30% and 40% for Acácia data, since, according to Yan (2013), the number of missing values in experiments with genotype by environment interaction is lower than 40%. This process was repeated a hundred times for each percentage removed in both set of values, obtaining 300 distinct matrices for foliar plague dataset and 400 distinct matrices for Acácia dataset, totaling 700 matrices with simulated missing values. Subsequently, imputations were made for each one of the 700 matrices with simulated missing values. For the foliar plague dataset, three randomized removals were considered, since the matrix is small (size 4×5). Increases in removals at 40% would imply problems of convergence and loss of information, by complete deletion

of the row or column where they are located, among other issues.

The stages, simulations and predictions were conducted by computer routine developed and implemented to the programming language R (R Core Team 2020). It is worth mentioning, concerning the developed algorithm, the use of *gnm* function to adjust the GAMMI model up to two multiplicative terms. For foliar plague dataset, Poisson GAMMI and Gaussian GAMMI models were used, with their respective logarithmic function and identity. The model Poisson GAMMI was chosen due to the study of Hadi *et al.* (2010). For Acácia dataset it was used the model Binomial GAMMI, with the logit link function, since the data represent a proportion. The imputations were obtained by the algorithms EM-GAMMI, MIGAMMI, EM-AMMI (using the function EM-AMMI) and EM+DVS (using the function impute.svd). Simulations of random removal values or generation of missing data, assuming the missing at random mechanism - MAR (Missing at Random), were carried out by using the function *SimIm* from the multivariate *ImputeR package*.

GAMMI is one of the best models to analyze experiments with genotype $\times$ environment interaction, in cases in which occur violation of the suppositions of the ANOVA model, or when the response is a counting, a proportion, among others (HADI *et al.*, 2010). Hence, for each one of the matrices with missing values obtained by simulations from the junction of the algorithm EM with the GAMMI model up to $\mathbf{k}$ multiplicative terms ($\mathbf{k}$=0,1,2), using the simple residual linear regression model. Concerning the foliar plague set, it was assumed, for MIGAMMI and EM-GAMMI or (IM-AMMI) algorithms, Poisson and Gaussian logarithmic link function and identity, respectively. As for Acácia dataset, the binomial model with logit link function was employed by MIGAMMI, and EM-GAMMI.

2.5 Criteria used to evaluate the method

As evaluation criteria, it was used the statistics: normalized root mean square error - NRMSE, variance between imputation - V_E, average squared bias between the imputations mean and the original value deleted in the simulation study - ASB, general measure of performance (T_{acc}) and analysis of deviance - ANODEV. By NRMSE criterion, Ching *et al.* (2010), the algorithm is compared using the adjusted means, that is, the imputed values are compared to the correspondent observed values in the original dataset, in accordance with the equation (3). It is recognized as the best method in performance, the one which presents lower statistic value NRMSE.

$$\text{NRMSE} = \frac{\sqrt{\text{mean}(\mathbf{x}_{imp} - \mathbf{x}_{orig})^2}}{\mathbf{s}(\mathbf{x}_{orig})} \tag{3}$$

where $\mathbf{x}_{imp}$ and $\mathbf{x}_{orig}$ are vectors containing the respective mean imputed values and true values of the missing simulated observations and $\mathbf{s}(\mathbf{x}_{orig})$ represents the normalized standard deviation values contained in vector $\mathbf{x}_{orig}$.

2.5.1 General measure of performance (T_{acc})

According to Bergamo (2007), the general measure of performance T_{acc} is a measure of accuracy, used to evaluate a particular procedure of MI and, which can be decomposed in two components $T_{acc} = V_E + \text{ASB}$. The former, V_E, represents the variance between imputations, in general, small values of V_E indicate good accuracy of the method. ASB, in turn, represents the average squared bias between mean imputations ($\overline{Y}$) and the original value removed in the simulation study (VO). The method of multiple imputation will present good performance if the values of ASB are small. The statistics V_E and ASB are presented in (4).

$$V_E = \frac{1}{na}\sum_{l=1}^{na}\left[\frac{\sum_{m=1}^{M}(\widehat{y}_{ij(m)} - \overline{Y}_l)^2}{M-1}\right] \quad \text{e} \quad \text{ASB} = \frac{1}{na}\sum_{l=1}^{na} M\frac{(\overline{Y}_l - VO_l)^2}{M-1} \tag{4}$$

where, for each position (i, j) of random removals in the data matrix, $\mathbf{M}$ imputations are performed; VO_l original value removed at random; the index l represents the position of the removed value correspondent to the position (i, j) with $l = 1, \ldots, na$; na is the total number of removed values; $\hat{y}_{ij}$ is the value imputed to the respective value VO_i and $\overline{Y}_l$ represents the mean of imputation to the position l.

2.5.2 ANODEV

To analysis of deviance (ANODEV) it was used the statistics F_c, since GAMMI's model is easy to calculate, presents good performance and does not require special tables (ACORSI *et al.*,2016). Thus, the statistics F_c or test F_c is as follows:

$$Fc = \frac{\text{Dev. restricted}}{\text{D.F.sv restricted}} - \frac{\text{Dev. full/D.F. full}}{\hat{\phi}} \tag{5}$$

which approximates the $F_{(\text{D.F sources of variation; D.F. error})}$ distribution. Where Dev. - deviance, $\hat{\phi}$ - estimated dispersion parameter, D.F.sv - degrees of freedom from source of variation that is being tested, Dev. restricted - deviance from the current model (tested). For the calculation of $\hat{\phi}$, consult Acorsi *et al.* (2016).

In general, the statistics NRMSE and T_{acc}, used to compare and evaluate, offer an excellent insight into the performance of the method in analysis. Thus, in this paper, it was considered as a good data imputation method, the one that presented, smaller mean/median value for the NRMSE distribution, as well as small values for the distributions V_E, ASB and T_{acc}, since the imputation were obtained based on to 100 simulated matrices at different levels of randomized removals of values.

3 Results and discussion

3.1 Foliar plague dataset

Table 3 presents the means and medians of NRMSE for the foliar plague dataset, indicating the method of multiple imputation proposed (MIGAMMI), the method EM+SVD and the method EM-AMMI for each level of removal percentage. In this table, following NRMSE criterion, the method that presented the best performance was MIGAMMI, regardless of the removal level. Therefore, at 10% imputation, the best performance procedure was MIGAMMI0 (median=0.199). At 20% level, MIGAMMI0 was the best in performance (median=0.2076) and at 30%, MIGAMMI0, (median 0.1956). Also, the procedure EM+SVD obtained the best results in terms of NRMSE than the classic EM-AMMI, with up to two multiplicative terms for all the levels of removal.

Table 3 - Mean and median of the RMSE distribution, in which were made random removal (10%, 20% and 30%), from foliar plague dataset

Method	10%		20%		30%	
	Mean	Median	Mean	Median	Mean	Median
EM+SVD	0.925	0.971	1.1297	0.9518	1.0955	1.0006
EM-AMMI0	1.086	0.965	1.3770	1.0290	1.3025	1.1922
EM-AMMI1	2.168	1.462	2.3680	1.8920	2.1495	2.0158
EM-AMMI2	1.049	1.076	1.1930	1.0470	1.1218	1.0323
MIGAMMI0	0.216	0.199	0.2524	0.2076	0.2188	0.1956
MIGAMMI1	0.254	0.230	0.2454	0.2126	0.2230	0.2024
MIGAMMI2	0.227	0.221	0.2759	0.2341	0.2195	0.2046

Table 4 illustrates in terms of mean and median, the values of V_E and ASB. In this table, it was verified that the procedures of multiple imputation MI-AMMI0, MI-AMMI1 and MI-AMMI2 provided the major variance between imputations (V_E), regardless of the percentage of missing imputation, whereas the algorithm with minor variance between imputations was the MIGAMMI0, at 10% and 20% removal level, followed by MIGAMMI2, at 30% removal. However, as a complement of the analysis V_E and to take the best decision about which would be the procedure with the highest efficiency in prediction, it is necessary to analyze the average squared bias (ASB) as well as the general measure of accuracy T_{acc}.

In relation to ASB, the methods with the least deviance, according to the percentages of imputations adopted were: at 10% missing, MIGAMMI0, at 20% missing, MIGAMMI1 and at 30% missing, MIGAMMI2 (Table 4). In all cases, the procedures with major values of ASB were MI-AMMI0, MI-AMMI1, and MI-AMMI2. Otherwise, the algorithms MIGAMMI0, MIGAMMI1 and MIGAMMI2, considering their lower values of ASB, allowed to achieve major similarity between imputations and their respective original values, resulting in a most accurate

Table 4 - Mean and median of the combined variance between imputations (V_E) and average squared bias (ASB), corresponding to the random removal levels of data (10%, 20% and 30%) of foliar plague dataset

Method	10%		20%		30%	
	Mean	Median	Mean	Median	Mean	Median
			V_E			
MI-AMMI0	0.4618	0.4064	0.3986	0.3628	0.3412	0.3425
MI-AMMI1	0.4231	0.4130	0.3835	0.3530	0.3341	0.3159
MI-AMMI2	0.4282	0.4031	0.4169	0.4249	0.3398	0.3140
MIGAMMI0	0.3557	0.3362	0.3256	0.3046	0.2927	0.2867
MIGAMMI1	0.4083	0.3784	0.3238	0.3098	0.2836	0.2710
MIGAMMI2	0.3777	0.3486	0.3310	0.3296	0.2968	0.2961
			ASB			
MI-AMMI0	0.1103	0.0646	0.1024	0.0807	0.0838	0.0687
MI-AMMI1	0.1064	0.0590	0.1022	0.0764	0.0771	0.0597
MI-AMMI2	0.1237	0.0850	0.0901	0.0796	0.0862	0.0740
MIGAMMI0	0.0769	0.0495	0.0814	0.0676	0.0730	0.0565
MIGAMMI1	0.1047	0.0662	0.0739	0.0631	0.0748	0.0591
MIGAMMI2	0.0776	0.0608	0.0947	0.0719	0.0697	0.0582

method. Furthermore, as the percentage measure increased, it was expected an increase in ASB values for the procedures of imputations, which was not confirmed. It was in fact observed a small median increase in ASB for the procedure MI-AMMI1 in comparison to MIGAMMI0, the same was verified for the procedure MIGAMMI1 in comparison to MIGAMMI0. Such an event, when it occurs, might be justified as a decrease of values generated by the removal levels (sample decrease), because, in conformity with Arciniegas-Alarcón *et al.* (2014), the imputation error tends to increase, considering that the available information in the data matrix were decreased by the growth of removals percentage.

To decide the best method of imputation, the general statistics of accuracy T_{acc} must be considered. The statistics considers both the variance between imputations and the mean square deviation (Tabela 4). In figure 1, it is shown a T_{acc} distribution in terms of median for the MIGAMMI0 procedure (IMGA0), MIGAMMI1 (IMGA1) and MIGAMMI2 (IMGA2) at three levels of removal. The process of imputation MIGAMMI0 in this case presented lower median value for imputations at 10% and 20% random removals, followed by the MIGAMMI1 operation, for imputations at 30% of removals. Thus, at 10% of randomized removals, the medians of T_{acc} were: 0.4131 for MIGAMMI0; 0.4758 for MIGAMMI1; and 0.4270 for MIGAMMI2; against 0.4942 for MI-AMMI0; 0.4977 for MI-AMMI1 and 0.5305 for MI-AMMI2. For the percentage of 20%, the procedure MIGAMMI0 presented the best performance, median = 0.3847. In case of missing at 30%, the procedure MIGAMMI1 achieved the best performance, median = 0.3473. Therefore,

it is worth highlighting the good performance of MIGAMMI method, whether in comparison to the results of the presented methods and also due to the low values obtained by NRMSE statistics and general mean accuracy (T_{acc}).

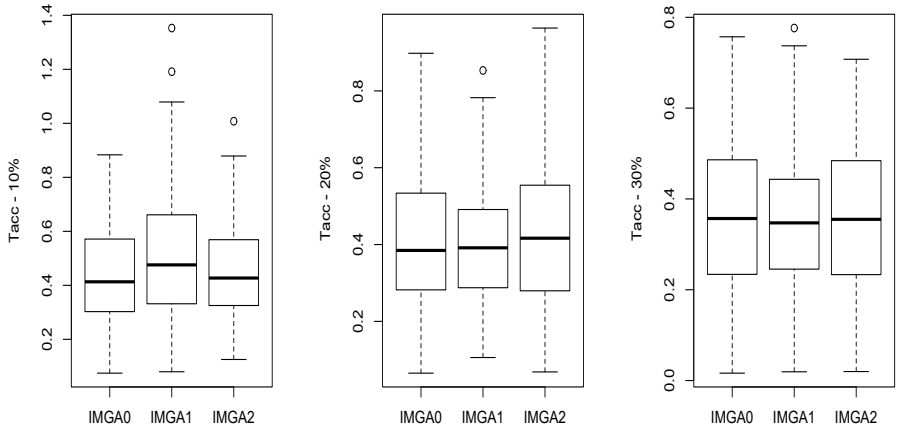

Figure 1 - Distribution of the general measure accuracy (T_{acc}), using the methods MIGAMMI0 (IMGA0), MIGAMMI1 (IMGA1) and MIGAMMI2 (IMGA2), for foliar plague dataset at levels of 10%, 20%, and 30% of removals.

3.2 Second dataset - Acácia

For the evaluated Acácia dataset, the method of imputation MIGAMMI invariably provided the lowest statistic value NRMSE, in means and medians terms, when compared to the method EM-GAMMI in all levels of removals (Table 5). The low values obtained from NRMSE suggest better predictions by the multiple imputation procedure MIGAMMI, that is, the imputed values come closer to the observed correspondents. For missing at 10%, the MIGAMMI0, MIGAMMI1 and MIGAMMI2 procedures evidenced similar performance, however, for computer facilities and economy of parameters MIGAMMI0 must be the chosen approach. For missing at 20%, the MIGAMMI1 presented better performance (mean = 0.1725), with 30% of missing, MIGAMMI2 presented better performance and with 40% of missing, MIGAMMI2 was the best method in performance. Also, it was verified the growth of NRMSE in the EM-GAMMI-0, EM-GAMMI-1 and EM-GAMMI-2 procedures, as the removal levels increase; the same can be observed for MIGAMMI0, MIGAMMI1 and MIGAMMI2 procedures, at moderate levels.

Table 5 - Mean and median of NRMSE distribution, in which were made random removal at 10%, 20%, 30% and 40%, of Acácia dataset

Method	10%		20%		30%		40%	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
EM-GAMMI-0	0.5128	0.5026	0.5232	0.5122	0.5699	0.5523	0.6111	0.5999
EM-GAMMI-1	0.5057	0.4960	0.5287	0.5179	0.5863	0.5737	0.6357	0.6310
EM-GAMMI-2	0.5140	0.5021	0.5404	0.5336	0.6022	0.5809	0.6550	0.6512
MIGAMMI0	0.1664	0.1612	0.1741	0.1724	0.1785	0.1744	0.1738	0.1725
MIGAMMI1	0.1683	0.1702	0.1725	0.1726	0.1757	0.1758	0.1737	0.1709
MIGAMMI2	0.1671	0.1528	0.1743	0.1707	0.1738	0.1733	0.1714	0.1690

Table 6 presents the variance between imputations (V_E) and the average squared bias (ASB), in terms of mean and median of 100 matrices submitted to the approaches of multiple imputations MIGAMMI0, MIGAMMI1 and MIGAMMI2. In this scenario, the methods presented small variance V_E with close values, in all levels of percentage imputation, since with up five decimal places means and medians results of V_E were approximately 0.00008. In relation to ASB, the procedures of imputations MIGAMMI0, MIGAMMI1 and MIGAMMI2, in four levels of missing, presented slight tendencies, that is, values of ASB close to zero, ranging from 0.0000194 to 0.0000207 in means terms, which indicates an excellent accuracy of the methods. It is worth emphasizing the small increase of ASB, which was expected, since by increasing the percentage of removals for the imputation, the size of the sample decreases.

Table 6 - Mean and median of combined variance between imputation (V_E) and average squared bias (ASB), corresponding to the random removal of data (10%, 20%, 30% and 40%) of Acácia dataset

Method	10%		20%		30%		40%	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
				V_E				
MIGA0	820	819	811	799	819	819	816	810
MIGA1	828	830	817	796	793	786	822	831
MIGA2	820	782	813	806	799	785	824	810
				ASB				
MIGA0	194	171	204	197	210	198	207	204
MIGA1	199	189	201	206	204	198	206	193
MIGA2	195	183	205	199	199	197	201	188

[1]MIGAMMI0 (MIGA0), MIGAMMI1 (MIGA1) e MIGAMMI2 (MIGA2)
[2]All values in the table are preceded by four decimal places, e.g., the first value is 0.0000820

Figure 2 shows the statistic distributions T_{acc} for the procedures MIGAMMI0 (IMGA0), MIGAMMI1 (IMGA1) and MIGAMMI2 (IMGA2). In this case, we

observed that the methods presented approximately symmetric distributions around the median for the imputations performed. The method with minor values for the parameter of median centrality, for missings at 10% was MIGAMMI2, for missings at 20% was MIGAMMI2, for missings at 30% was MIGAMMI1 and for missings at 40% was MIGAMMI2. On the other hand, it should be highlighted that all the approaches displayed small values of T_{acc}, close to zero, in all levels of performed imputations (Table 7). Therefore, the method MIGAMMI0 might be preferred, if considering the idea of parameter economy or computer facilities, since it exempts the inclusion of multiplicative terms.

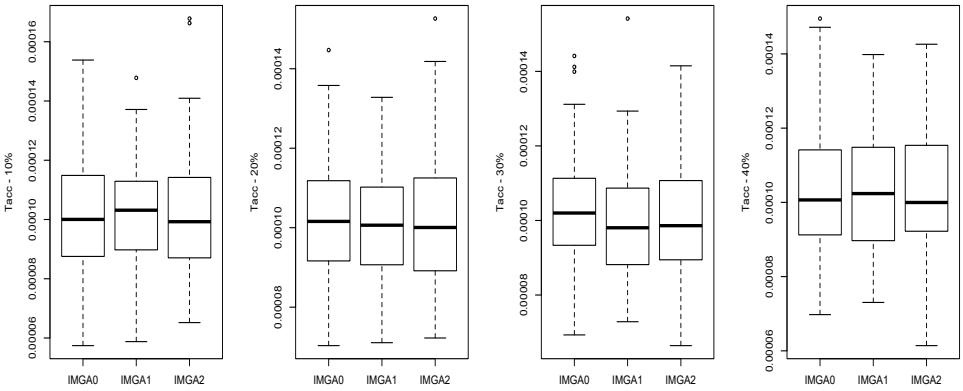

Figure 2 - Distribution of the general measure accuracy (T_{acc}), using the methods MIGAMMI0 (MIGA0), MIGAMMI1 (MIGA1) and MIGAMMI2 (MIGA2), for Acácia dataset at levels of 10%, 20%, 30% and 40% of removals.

Table 7 - Median of T_{acc} distribution to the random removal levels (10%, 20%, 30% and 40%) for Acácia dataset

Method	T_{acc}			
	Median (10%)	Median (20%)	Median (30%)	Median (40%)
MIGAMMI0	0.0001000	0.0001016	0.0001020	0.000101
MIGAMMI1	0.0001031	0.0001006	0.0000980	0.000102
MIGAMMI2	0.0000993	0.0001001	0.0000986	0.000100

Finally, when the results of procedures were analyzed for the both multi-environment datasets, it was found that the approaches of multiple imputation

MIGAMMI displayed better results than the EM+SVD, EM-AMMI in terms of NRMSE values, for foliar plague dataset, as well as, better results for Acácia dataset, both in terms of NRMSE, which compared the procedures EM-GAMMI0, EM-GAMMI1 and EM-GAMMI2 with the procedures MIGAMMI0, MIGAMMI1 and MIGAMMI2 and in terms of general measure of accuracy T_{acc}, due to their low values obtained. According to Rubin (1987), Carvalho *et al.* (2017), MI is more advantageous than single imputation, allowing increase in efficiency of the estimates, allowing valid inferences, reflecting the additional variability due to missing values and allows to compare the sensibility of the obtained inferences by different techniques of imputation, by simply using methods of complete data.

3.3 Application - foliar plague dataset

Once performed the stage of imputation, the following step is to carry out the analyses about the experiment, a convenient model is used for such proposal. Tables 8, 9 and 10 present the result of the analysis of deviance (ANODEV), in which we utilized the models GAMMIs, for removal levels at 10%, 20% and 30% respectively. True values removed at 10% level were 1.25 and 2.00 with imputed correspondents 1.40 and 2.00, at 20% level, the values removed were 2.75; 3.00; 2.25 and 1.25 with imputed correspondents 2.80; 3.00; 1.91 and 1.27 and at 30% level, the removed values by the process of simulation were 4.00; 3.00; 1.25; 1.00; 1.75 and 2.00 with imputed correspondents 4.0; 3.0; 1.29; 0.65; 2.0; 2.0. As expected, for well-balanced processes, besides the methods reproducing imputed values close to the observed correspondents, it was not found any substantial alterations in the inferences for the three levels of removal (Tables 8, 9 and 10) in comparison to the results presented by Hadi *et al.* (2010), relevant with up to two multiplicative terms in the model.

Table 8 - Deviance analysis of foliar plague dataset, after imputation by the method MIGAMMI0, with random removals at 10%

Source of variation	D.F.	Deviance	Deviance mean	F_c	p-value
Environment	4	4.1067	1.0267	76.05	0.0132
Genotype	3	2.8562	0.9521	70.52	0.0142
GAMMI1	6	3.6184	0.6031	44.67	0.0222
GAMMI2	4	0.9680	0.242	17.93	0.0542
Error	2	0.0270	0.0135		
Total	19	11.5763	0.6093		

Table 9 - Deviance analysis of foliar plague dataset, after imputation by the method MIGAMMI0, with random removals at 20%

Source of variation	D.F.	Deviance	Deviance mean	F_c	p-value
Environment	4	4.5929	1.1482	83.81	0.0120
Genotype	3	3.2505	1.0835	79.088	0.0127
GAMMI1	6	2.9942	0.4990	36.426	0.0272
GAMMI2	4	0.9159	0.2289	16.714	0.0579
Error	2	0.0274	0.0137		
Total	19	11.7809	0.620		

Table 10 - Deviance analysis of foliar plague dataset, after imputation by the method MIGAMMI0, with random removals at 30%

Source of variation	D.F.	Deviance	Deviance mean	F_c	p-value
Environment	4	4.3151	1.0787	101.29	0.0099
Genotype	3	2.8065	0.9355	87.84	0.0115
GAMMI1	6	4.0746	0.6791	63.77	0.0157
GAMMI2	4	1.0183	0.2545	23.90	0.0411
Error	2	0.0213	0.0107		
Total	19	12.2358	0.6439		

The results obtained provide some guide for future research related to missing data. For instance, to use new models GAMMIs for imputation in multi-environment data with overdispersion, other or new methods of imputation can be taken to make comparisons with the method MIGAMMI, new datasets can be taken for analyses. The methods EM-AMMI-0, EM-AMMI-1, EM+SVD, presented in literature, showed inferior performance to the MIGAMMI introduced in this paper. In Arciniegas-Alarcón and Dias (2009), the method EM-AMMI1 presented better performance than IMLD). In previous studies, Arciniegas *et al.* (2014), demonstrated the good performance of the methods EM+SVD and EM-AMMI when compared to other methods of imputation. Such indicators added to the results presented here, are solid findings of the good quality of MIGAMMI method.

Conclusions

In this paper, it was analyzed an approach of statistical multiple imputation of data in multi-environment trials and evaluated by statistical distribution NRMSE and general measure of accuracy T_{acc}. The procedure MIGAMMI exhibited the best results as a method of imputation, proving to be superior to the methods EM-GAMMI, EM-AMMI and EM+SVD, in both datasets used in the study.

Therefore, it was possible to conclude in favor of the procedure of multiple imputation MIGAMMI, the most efficient method to perform imputation, both in terms of NRMSE and in terms of overall accuracy statistic T_{acc}.

4 Acknowledgements

The authors thank the reviewers and editors for the comments and suggestions that helped improve the quality of this paper.

AMOÊDO, P. M; PIEDADE, S. M. P.; DIAS, C. T. S, ARCINIEGAS-ALARCÓN, S. Algoritmo de imputação múltipla MIGAMMI. *Braz. J. Biom.,* Lavras, v.40, n.1, p.1-20, 2022.

- *RESUMO: Dados ausentes são comuns em experimentos multiambientais por mais bem planejados que sejam, por isso, o uso de métodos de análises apropriados é essencial para reduzir o impacto gerado pela perda de informações. A imputação de dados é uma das técnicas comumente usada para contornar o problema das ausências, estima os dados ausentes por valores plausíveis e posteriormente as análises são realizadas sobre os dados completados. O presente trabalho tem por objetivo propor um novo método de imputação múltipla, para dados provenientes de experimentos multiambientais, resultante da proposta dos resíduos simples do modelo de regressão linear. Deste modo, modificações no algoritmo de imputação simples EM-AMMI foram realizadas, de forma a comportar o modelo de efeitos principais aditivos e interação multiplicativa generalizado GAMMI. A qualidade do método de imputação múltipla foi avaliada por meio das distribuições de uma estatística geral de acurácia que combina a variância entre imputações e o viés quadrático médio e da raiz normalizada do erro quadrático médio (NRMSE). Para tal, simulações de retiradas aleatória de valores nos níveis de 10%, 20%, 30% e até 40% foram geradas a partir de dois conjuntos de dados reais e as imputações correspondentes obtidas. Os resultados da medida geral acurácia e da NRMSE, pelos seus baixos valores obtidos em relação ao método proposto, servem de evidências da melhor qualidade do algoritmo de imputação múltipla IMGAMMI proposto.*

- *PALAVRAS-CHAVE:*
AMMI generalizado; análise de deviance; estatística Tacc; imputação biplot.

References

ACORSI, C. R. L; GUEDES, T. A; COAN, M.; PINTO, R. J. B; SCAPIM, C. A; PACHECO, C. A. P; GUIMARÃES, P. D. O.; CASELA, C. R. Applying the generalized additive main effects and multiplicative interaction model to analysis of maize genotypes resistant to grey leaf spot. *journal of Agricultural Science*, Cambridge, v.155, p.939-953, 2017.

AMOÊDO, P. M. *Modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI) para imputações de dados em experimentos multiambientais*, 2021. 45p. Thesis (Ph.D.) - L. S. E. Universidade de São Paulo, Piracicaba, 2021.

ARCINIEGAS-ALARCÓN, S.; DIAS, C. T. S. Data imputation in trials with genotype by environment interaction: an application on cotton data. *Revista Brasileira de Biometria*, São Paulo, v.27, p.125-138, 2009.

ARCINIEGAS-ALARCÓN, S.; DIAS, C. T. S.; GARCÍA-PÊNA, M. Imputação múltipla livre de distribuição em tabelas incompletas de dupla entrada. *Pesquisa Agropecuária Brasileira*, Brasília, v.49, p.683-691, 2014.

ARCINIEGAS-ALARCÓN, S; GARCÍA-PEÑA, M; RODRIGUES, P. C. New multiple imputation methods for genotype-by-environment data that combine singular value decomposition and Jackknife resampling or weighting schemes. *Computers and Electronics in Agriculture*, v.176, p.105617, 2020.

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; KRZANOWSKI, W.; DIAS, C. T. S. Imputing missing values in multi-environment trials using the singular value decomposition: An empirical comparison. *Communications in Biometry and Crop Science*, v.9, p.54-70, 2014

BERGAMO, G. C. *Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação*, 2007. 89p. Thesis (Ph.D.) - L. S. E. Universidade de São Paulo, Piracicaba, 2007.

BERGAMO, G. C.; Dias, C. T. d. S.; KRZANOWSKI, W. J. Distribution-free multiple imputation in an interaction matrix through singular value decomposition. *Scientia Agricola*, v.65, p.422-427, 2008.

CARVALHO, J. R. P. DE *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

CHING, W.; LI, L.; TSING, N.; TAI, C.; NG, T.; WONG, A.; CHENG, K. A weighted local least squares imputation method for missing value estimation in microarray gene expression data. *International journal of data mining and bioinformatics*, v.4, p.331-347, 2010.

ENDERS, C. K. *Applied missing data analysis*. Guilford: Guilford press, 2010. 382p.

GAUCH, H.; ZOBEL, R. W. Imputing missing yield trial data. *Theoretical and Applied Genetics*, v.79, p.753-761, 1990.

HADI, A. F.; MATTJIK, A.; SUMERTAJAYA, I. Generalized ammi models for assessing the endurance of soybean to leaf pest. *Jurnal Ilmu Dasar*, v.11, p.151-159, 2010.

PADEREWSKI, J.; RODRIGUES, P. C. The usefulness of em-ammi to study the influence of missing data pattern and application to polish post-registration winter wheat data. *Australian Journal of Crop Science*, v.8, p.640-645, 2014.

PERRY, P. O. *Cross-validation for unsupervised learning*, 2009. 153p. Dissertation, Stanford University, 2009.

PEUGH, J. L.; ENDERS, C. K. Missing data in educational research: A review of reporting practicesand suggestions for improvement. *Review of educational research*, v.74, p.525-556, 2004.

PIEPHO, H. P. Methods for estimating missing genotype-location combinations in multilocationtrials-an empirical comparison. *Informatik Biometrie und Epidemiologie in Medizin und Biologie*, v.26, p.335-349, 1995.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria., 2020.

RODRIGUES, P. C.; MONTEIRO, A.; LOURENÇO, V. M. A robust additive main effects and multiplicative interaction model for the analysis of genotype-by-environment data. *Bioinformatics*, v.32, p.58-66, 2016.

RODRIGUES, P. C.; PEREIRA, D. G. S.; Mexia, J. T. A comparison between joint regression analysisand the additive main and multiplicative interaction model: the robustness with increasing amountsof missing data. *Scientia Agricola*, v.68, p.679-686, 2011.

ROUSSEAU, M.; SIMON, M.; BERTRAND, R.; HACHEY, K. Reporting missing data: a study of selected articles published from 2003-2007. *Quality & Quantity*, v.46, p.1393-1406, 2012.

RUBIN, D. B. Multiple imputations in sample surveys-a phenomenological bayesian approachto nonresponse. In Proceedings of the survey research methods section of the American Statistical Association, 1978, Alexandria. *Proceedings. Alexandria: The American Statistical Association*, p.20-34, 1978.

RUBIN, D. B. *Multiple imputation for survey nonresponse*. New York: John Wiley & Sons, 1987. 320p.

RUBIN, D. B. Multiple imputation after 18+ years. *Journal of the American statistical Association*, v.91, p.473-489, 1996.

SCHAFER, J. L. ; GRAHAM, J. W. Missing data: our view of the state of the art. *Psychological methods*, v.7, p.147-177, 2002.

SCHOMAKER, M; HEUMANN, C. Bootstrap inference when using multiple imputation. *Statistics in medicine*, v.37, n.14, p.2252-2266, 2018.

SPITTI, A. M. D. S.; CARBONELL, S. A. M. ; DIAS, C. T. d. S.; SABINO, L. G.; CARVALHO, C. R. L.; CHIORATO, A. F. Genótipos de feijoeiro carioca para

tolerância ao escurecimento de grão pelos métodos natural e acelerado. *Ciência e Agrotecnologia*, v.43, 2019.

SRIVASTAVA, M. S.; DOLATABADI, M. Multiple imputation and other resampling schemes for imputing missing observations. *Journal of Multivariate Analysis*, v.100, p.1919-1937, 2009.

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

VAN EEUWIJK, F. A. Multiplicative interaction in generalized linear models. *Biometrics*, v.51. p.1017-1032, 1995.

VAN GINKEL, J. R.; LINTING, M.; RIPPE R. C. A. ; VAN DER VOORT, A. Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, v.102, p.297-308, 2019.

YAN, W. Biplot analysis of incomplete two-way data. *Crop Science*, v.53, p.48-57, 2013.

ZHANG, P. Multiple imputation: theory and method. *International Statistical Review*, v.71, p.581-592, 2003.

Recebido em 15.10.2020.

Aprovado após revisão em 02.07.2021.