Research Article
Deterministic Imputation in Multienviroment Trials

Sergio Arciniegas-Alarcón,¹ Marisol García-Peña,¹ Wojtek Janusz Krzanowski,² and Carlos Tadeu dos Santos Dias¹

¹ Departamento de Ciências Exatas, Universidade de São Paulo/ESALQ, Cx. P.09, CEP 13418-900, Piracicaba, SP, Brazil
² College of Engineering, Mathematics and Physical Sciences Harrison Building, University of Exeter, North Park Road, Exeter, EX4 4QF, UK

Correspondence should be addressed to Sergio Arciniegas-Alarcón; sergio.arciniegas@gmail.com

Received 26 June 2013; Accepted 16 August 2013

Academic Editors: A. Escobar-Gutierrez and W. P. Williams

Copyright © 2013 Sergio Arciniegas-Alarcón et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper proposes five new imputation methods for unbalanced experiments with genotype-by-environment interaction (G×E). The methods use cross-validation by eigenvector, based on an iterative scheme with the singular value decomposition (SVD) of a matrix. To test the methods, we performed a simulation study using three complete matrices of real data, obtained from G×E interaction trials of peas, cotton, and beans, and introducing lack of balance by randomly deleting in turn 10%, 20%, and 40% of the values in each matrix. The quality of the imputations was evaluated with the additive main effects and multiplicative interaction model (AMMI), using the root mean squared predictive difference (RMSPD) between the genotypes and environmental parameters of the original data set and the set completed by imputation. The proposed methodology does not make any distributional or structural assumptions and does not have any restrictions regarding the pattern or mechanism of missing values.

1. Introduction

In plant breeding, multienviroment trials are important for testing the general and specific adaptations of cultivars. A cultivar developed in different environments will show significant fluctuations of performance in production relative to other cultivars. These changes are influenced by different environmental conditions and are referred to as genotype-by-environment interactions, or G×E. Often, G×E experiments are unbalanced because several genotypes are not tested in some environments. A common way of analyzing this type of study is by imputing the missing values and then applying established procedures on the completed data matrix (observed + imputed), for example, the additive main effects and multiplicative interaction model—AMMI—or factorial regression [1–5]. An alternative approximation is to work with the incomplete data using a mixed model with estimates based on maximum likelihood [6].

Several imputation methods have been suggested in the literature to solve the problem of missing values. One of the first was made by Freeman [7], who suggested imputing the missing values iteratively by minimizing the residual sum of squares and doing the G×E analysis on the completed table, reducing the degrees of freedom by the number of missing values. This work was developed by Gauch Jr. and Zobel [8], who made the imputations using the EM algorithm and the AMMI model or EM-AMMI. Some variants of this procedure using multivariate statistics (cluster analysis) were described in Godfrey et al. [9] and Godfrey [10]. Raju [11] proposed the EM-AMMI algorithm by treating the environments as random and suggested applying a robust statistic to the missing values in the stability analysis. Mandel [12] proposed the imputation to be made in incomplete two-way tables using linear functions of the rows (or columns). Other studies recommended by van Eeuwijk and Kroonenberg [13] as having good results in the case of missing values for G×E experiments were developed by Denis [14], Caliński et al. [15], and Denis and Baril [16]. They found that using imputations through alternating least squares with bilinear interaction models or AMMI estimates based on robust submodels could give results as good as those found with the EM algorithm. Additionally, Caliński et al. [17] introduced an algorithm
that combines the singular value decomposition (SVD) of a matrix with the EM algorithm, obtaining results very useful for experiments in which the alternating least squares have some problems, for instance, convergence failures [18]. Recently, Bergamo et al. [19] proposed a distribution-free multiple imputation method that was assessed by Arciniegas-Alarcón [20] and compared by Arciniegas-Alarcón and Dias [21] with algorithms that use fixed effects models in a simulation study with real data. Meanwhile, a deterministic imputation method without structural or distributional assumptions for multi-environment experiments was proposed by Arciniegas-Alarcón et al. [22]. The method uses a mixture of regression and lower-rank approximation. Finally, other studies to analyze multi-environment experiments with missing values can be found in the literature. For example, methodologies for stability analysis have been studied by Raju and Bhata [23] and Raju et al. [24, 25]. Recently, Pereira et al. [26], Rodrigues et al. [27], and Rodrigues [28] assessed the robustness of joint regression analysis and AMMI models without the use of data imputation.

2. Materials and Methods

2.1. Data Imputation Using the Cross-Validation by Eigenvector Method. The cross-validation method was presented by Bro et al. [29] to find the optimum number of principal components in any data set that can be arranged in a matrix form. In this approximation, principal component analysis (PCA) models are calculated with one or several samples left out and the model is used to predict these samples. The method used cross-validation “leave-one-out” and the same study showed it to be more efficient than other well-known methodologies used in multivariate statistics, such as those presented by Wold [30] and Eastment and Krzanowski [31]. Because of this finding, Arciniegas-Alarcón et al. [32] used the method to determine the best AMMI models in (G × E) experiments. This methodology is now presented.

Step 1. Consider the n × p matrix X with elements x_{ij}, (i = 1, ..., n; j = 1, ..., p). The matrix is divided into disjoint groups, each group is deleted in turn (leave-one-out), and a PCA model (T, P) is obtained from the remainder by solving

$$\min \|X^{(-g)} - TP^T\|^2_m$$

(1)

with m ≤ \min(n - 1, p - 1). Here X^{(-g)} represents the matrix after deleting the ith group (leave-one-out), \| \cdot \|_2 defines the squared Frobenius norm, P^T P = I, and T, P are scores and loadings matrices with dimensions (n - 1) × m and p × m respectively, where p is the number of columns and m is the number of components. Note that, in this method the deleted group corresponds to the ith row of X and according to Smilde et al. [33] the model (1) can be rewritten in terms of the singular value decomposition (SVD)

$$X^{(-g)} = UDV^T = \sum_{k=1}^m u_k d_k v_k^T,$$

(2)

where U = [u_1, u_2, ..., u_m], V = [v_1, v_2, ..., v_m], D = diag(d_1, d_2, ..., d_m), T = UD, and P = V.

Step 2. Estimate the score

$$t^{(-j)T} = x_j^{(-j)T} (P^{(-j)T} P^{(-j)})^{-1},$$

(3)

where P^{(-j)T} is the loading matrix found in Step 1 with the jth row excluded. x_j^{(-j)T} is a row vector containing the ith row of X except for the jth element.

Step 3. Estimate the element x_{ij} by

$$x_{ij}^{(m)} = t^{(-j)T} P_j,$$

(4)

p_j is the jth row of P.

Step 4. Find the prediction error of the (ij)th element, \hat{e}_{ij}^{(m)} = x_{ij} - x_{ij}^{(m)}.

Step 5. Obtain the criterion value

$$PRESS(m) = \sum_{i=1}^n \sum_{j=1}^p (e_{ij}^{(m)})^2.$$  

(5)

In order to make the imputation of missing values in the matrix from (G × E) experiments, a change is proposed in the method following the work of Krzanowski [34], Bergamo et al. [19], and Arciniegas-Alarcón et al. [22] using the singular value decomposition of a matrix [35].

Initially, suppose that n ≥ p and the matrix X has several missing values; in the case n < p, the matrix should first be transposed. The missing values are replaced by their respective column means \bar{X}_j, and after this has been done the matrix is standardized by columns, subtracting \bar{X}_j and dividing by \sigma_j (where \bar{X}_j and \sigma_j represent, resp., the mean and the standard deviation of the jth column). The eigenvector procedure using the SVD in expressions (2)–(4) is applied to the standardized matrix to find the imputation of the (i, j) element, denoted by \hat{x}_{ij}^{(m)}. After the imputation, the matrix must be returned to its original scale, x_{ij} = \bar{X}_j + \sigma_j \hat{x}_{ij}^{(m)}.

At this point the matrix does not have any missing values, but the imputations are rather basic and need to be refined. In the works that previously mentioned an iterative scheme is advocated, iterations continuing until the imputations achieve convergence (i.e., there is stability in successive imputed values), but Calinski et al. [17] showed that this convergence is not always necessary when using a method that combines the EM algorithm with SVD. Therefore, taking this into account, we will also consider fixing in advance the number of iterations between 0 and 3, as well as permitting
the process to run until convergence has been achieved. As regards to the computing effort, convergence can depend strongly on the size of matrix analyzed and also on the data structure (size of correlations, proportion of missing values, etc.), but in, for instance, the SVD method of Hastie et al. [36], convergence is achieved usually between 5 and 6 iterations, and in the Bergamo et al. [19] method it is achieved in between 20 and 50 iterations maximum.

On the other hand, the data imputation depends directly on (2) and (3). Equation (2) needs prior choice of the number of components (m) to extract from the SVD. Krzanowski [34] and Bergamo et al. [19] took m = min(n − 1, p − 1) with the objective of using the maximum amount of available information, but Hedderley and Wakeling [37] affirmed that if the estimation is based on the choice of a unique fixed number of dimensions, some of the lower dimensions may be essentially random. This can influence the imputation within an iterative scheme and can lead to the estimates becoming trapped in a cycle, hence preventing convergence. To solve this problem, they suggested including a test to check on the convergence rate, and in case a specific criterion is not being attained the number of dimensions should be reduced. Another option that has satisfactory results, suggested by Josse et al. [38], is through cross-validation based uniquely on the observed data. However, the computational cost of this option is likely to be high.

Taking into account all the above mentioned in the present study, for imputation of each missing value of the matrix X the value of m in (2) is allowed to be different in each SVD calculated and is chosen according to the criterion used by Arciniegas-Alarcón et al. [22]. Thus, m is chosen such that
\[
\left(\frac{\sum_{k=1}^{m} d_k^2}{\sum_{k=1}^{\min(n-1,p-1)} d_k^2}\right) \approx 0.75.
\]
Moreover, in (3), the Moore-Penrose generalized inverse can be used instead of the classic inverse as was studied in cross-validation by Dias and Krzanowski [39].

In this research, five imputation methods have been assessed. They are denoted Eigenvector0, Eigenvector1, Eigenvector2, Eigenvector3, and Eigenvector where the number indicates the number of iterations used while in the case of Eigenvector the process is iterated until convergence is achieved in the imputations.

These imputation methods are all deterministic imputations, and they have the advantage over other stochastic imputation methods (parametric multiple imputations) that the imputed values are uniquely determined and will always yield the same results when applied to a given data set. This is not necessarily true for the stochastic imputation methods [40].

2.2. The Data. To assess the imputation methods we used three data sets, published in Caliński et al. [41, page 227], Farias [42, page 115], and Flores et al. [43, page 274], respectively. In each case the data were obtained from a randomized complete block design with replication, and each reference offers an excellent description of the design if further details are required.

The first data set “Caliński” comprises an 18 × 9 matrix, for 18 pea varieties assessed in 9 different locations in Poland. The experiment was conducted by the Research Center for Cultivar Testing, Slupia Wielka, and the studied variable was mean yield (dt/ha).

The second data set “Farias” was obtained from Upland cotton variety trials (Ensia Estadual de Algodoeiro Herbáceo) in the agricultural year 2000/01, part of the cotton improvement program for the Cerrado conditions. The experiments assessed 15 cotton cultivars in 27 locations in the Brazilian states of Mato Grosso, Mato Grosso do Sul, Goiás, Minas Gerais, Rondônia, Maranhão, and Piauí. The studied variable was yield seed cotton (kg/ha).

The third data set “Flores” is in a 15 × 12 matrix, for 15 bean varieties assessed in 12 environments in Spain. The experiments were conducted by RAEA—Red Andaluza de Experimentación Agraria—where the studied variable was mean yield (kg/ha).

The three data matrices contained just the mean yield for each genotype in each environment, but the proposed methods work for any data set arranged in matrix form. For example, if information about the replications is available, an approach suggested by Bello [44] is to write the experiment in terms of a classic linear regression model in order to obtain the response vector and the design matrix, and then to join them into a single matrix and apply the proposed methods in this paper.

2.3. Simulation Study. Each original data matrix (“Caliński”, “Farias”, and “Flores”) was submitted to random deletion of values at the three rates 10%, 20%, and 40%. The process was repeated in each data set 1000 times for each percentage of missing values, giving a total of 3000 different matrices with missing values. Altogether, therefore, 9000 incomplete data sets were available, and for each one the missing values were imputed with the 5 Eigenvector algorithms described above using computational code in R [45].

The random deletion process for a matrix \(X\) \((n \times p)\) was conducted as follows. Random numbers between 0 and 1 were generated in R with the runif function. For a fixed \(r\) value \((0 < r < 1)\), if the \((pi + j)th\) random number was lower than \(r\), then the element in the \((i+1, j)\)th position of the matrix was deleted \((i = 0, 1, \ldots, n; j = 1, \ldots, p)\). The expected proportion of missing values in the matrix will be \(r\) [34]. This technique was used with \(r = 0.1, 0.2\) and 0.4 (i.e., 10%, 20%, and 40%).

2.4. Comparison Criteria. In general, the objective after imputation is to estimate model parameters from the complete table of information. One of the models frequently used in genotype-by-environmental trials is the AMMI model [46, 47], and for this reason the algorithms proposed in this paper will be compared through the genotypic and environmental parameters of the fitted AMMI models using the root mean squared predictive difference—RMSPD [39]. The AMMI model is first briefly presented.

The usual two-way ANOVA model to analyze data from genotype-by-environment trials is defined by
\[
y_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}
\]
\(i = 1, \ldots, n; j = 1, \ldots, p\) where \(\mu, a_i, b_j, (ab)_{ij},\) and \(e_{ij}\) are respectively, the overall mean, the genotypic and environmental main effects, the genotype-by-environment interaction, and an error term associated with the \(i\)th genotype and \(j\)th location. It is assumed that all effects except the error are fixed effects. The following reparametrization constraints are imposed: \(\sum_i (ab)_{ij} = \sum_j (ab)_{ij} = \sum a_i = \sum_j b_j = 0\). The AMMI model implies that interactions can be expressed by the sum of multiplicative terms. The model is given by

\[
y_{ij} = \mu + a_i + b_j + \theta_1 a_i\beta_{j1} + \theta_2 a_i\beta_{j2} + \cdots + e_{ij},
\]

(7)

where \(\theta_i, a_i, b_j, \) and \(\beta_{jl}(l = 1, 2, \ldots, \min(n-1, p-1))\) are estimated by the SVD of the matrix of residuals after fitting the additive part. \(\theta_i\) is estimated by the \(l\)th singular value of the SVD, \(a_i, b_j\) are estimated by the genotypic and environmental eigenvector values corresponding to \(\theta_i\).

Alternating regressions can be used in place of the SVD [48]; depending on the number of multiplicative terms, these models may be called AMMI0, AMMI1, and so forth.

An inherent requirement of the AMMI model is prior specification of the number of multiplicative components [49–51]. Rodrigues [28] made an exhaustive analysis of the related literature and concluded that usually two or three components can be used because, in general, one component is not enough to capture the entire pattern of response in the data, but with more than three components there are obvious visualization problems, and a huge quantity of noise is liable to be captured.

So, for the original matrices “Caliński”, “Farias”, and “Flores”, we fitted the AMMI2 and AMMI3 models. The same models were then fitted for each one of the 9000 sets of data that had been completed by imputation, and each set of parameters was compared with its corresponding set from the original data by using the RMSPD in the following way:

\[
\text{RMSPD (gen)} = \sqrt{\frac{\sum_{i=1}^{NG} (a_i - \hat{a}_i)^2}{NG}};
\]

\[
\text{RMSPD (env)} = \sqrt{\frac{\sum_{j=1}^{NE} (b_j - \hat{b}_j)^2}{NE}};
\]

\[
\text{RMSPD}_1 (genmult) = \sqrt{\frac{\sum_{i=1}^{NG} (\alpha_{ij} - \hat{\alpha}_{ij})^2}{NG}};
\]

\[
\text{RMSPD}_1 (envmult) = \sqrt{\frac{\sum_{j=1}^{NE} (\beta_{ij} - \hat{\beta}_{ij})^2}{NE}}.
\]

Here RMSPD (gen) represents the RMSPD among the estimated parameters for genotype main effects from the original data \(a_i\) and the corresponding parameters obtained from the completed data sets by imputation \(\hat{a}_i\). RMSPD (env) represents the RMSPD among the estimated parameters for environments main effects from the original data \(b_j\) and the corresponding parameters obtained from the completed data sets by imputation \(\hat{b}_j\). RMSPD1 (genmult) represents the equivalent RMSPD for the pairs of estimated parameters of genotype multiplicative components \(\alpha_{ij}, \beta_{ij}\), RMSPD1(envmult) represents the equivalent RMSPD for the pairs of estimated parameters of environments multiplicative components \(\beta_{ij}, \beta_{ij}\). In the statistics, NG represents the number of genotypes, NE the number of environments, and \(l = 2\) or 3 depending on the considered model AMMI2 or AMMI3.

The best imputation method is the one with the lowest values of RMSPD in each case. Summarizing, in each simulated data set with missing values, we applied the methods Eigenvector0, Eigenvector1, Eigenvector2, and Eigenvector3 and, then, in the completed data (observed and imputed) we fitted AMMI2, AMMI3 models for the calculation of the respectively RMSPD statistics. In order to visualize any differences more readily, the RMSPD values were standardized and the comparison was made directly. Note that because of the standardized scale, the values of the statistics can be either positive or negative.

### 3. Results

#### 3.1. Polish Pea Data

Figure 1 shows the RMSPD(gen) distribution on the standardized scale for the “Caliński” data set, showing each imputation method and each percentage. It can be seen that the Eigenvector distribution is left asymmetric and this asymmetry increases as the missing values percentage increases. In general, the Eigenvector distribution has values above zero and when the number of missing values increases, it is concentrated above one. This means that this method had the biggest differences among the additive genotypic parameters of the real and completed (by imputation) data.

The best method according to RMSPD(gen) is Eigenvector1, the method with just one iteration. This method has the smallest median for the 10% and 20% percentages. In the 40% percentage the medians of Eigenvector0 and Eigenvector1 are practically the same in the figure, but Eigenvector1 continues be preferable because it has the smallest dispersion. So, Eigenvector1 gave the smallest differences between the additive genotypic parameters of the real and completed data.

Figure 2 shows the RMSPD(env) on the standardized scale for the “Caliński” data set. It shows very similar behaviour to that of RMSPD(gen). Again the Eigenvector1 method presents the biggest differences among the additive environment parameters of the real and completed data because of the algorithm that maximizes the RMSPD(env). In this case, the RMSPD(env) is minimized with Eigenvector0 and Eigenvector1, and in all the percentages of missing values the two have nearly equal medians. However, Eigenvector1 has the smallest dispersion and that makes this again the method of choice.

The box plot analysis was useful in determining the best imputation method for the RMSPD(gen) and RMSPD(env) distributions, but in the case of RMSPD1(envmult), RMSPD2(genmult), RMSPD3(genmult) and RMSPD3(envmult), a more formal analysis can be used to compare the distributions; for instance the Friedman nonparametric test and, if this is significant, then the Wilcoxon test [52].
Table 1 shows the Friedman test statistics. It can be seen that a significant difference exists among the imputation methods for the 10% and 20% percentage of missing values, but with 40% the five methods have equivalent results. After the general test, it is necessary to make multiple pairwise comparisons for the two lower percentages.

Table 2 shows the Wilcoxon test to find the methods that are different. When $\text{RMSPD}_2^{2}(\text{genmul})$ for 10% was used, Eigenvector1 had significant differences with the other four methods. For 20%, Eigenvector1 was statistically different from Eigenvector, Eigenvector2, and Eigenvector3. For this percentage Eigenvector presents different results from Eigenvector0 and Eigenvector3. Joining the statistical differences found with the nonparametric test about $\text{RMSPD}_2^{2}(\text{genmul})$ and the correspond box plot in Figure 3, it can be said that for 10% and 20% the most efficient method is Eigenvector1,

**Figure 1**: Box plot of the $\text{RMSPD}(\text{gen})$ distribution in Calinski data set.
**Figure 2:** Box plot of the RMSPD\((env)\) distribution in Caliński dataset.

**Table 1:** Friedman test for the standardized RMSPD\((\cdot)\)—Caliński data set.

<table>
<thead>
<tr>
<th>Perc.</th>
<th>RMSPD(_2)(gen\textit{mult})</th>
<th></th>
<th>RMSPD(_2)(env\textit{mult})</th>
<th></th>
<th>RMSPD(_3)(gen\textit{mult})</th>
<th></th>
<th>RMSPD(_3)(env\textit{mult})</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Friedman (P) value</td>
<td>Friedman (P) value</td>
<td>Friedman (P) value</td>
<td>Friedman (P) value</td>
<td>Friedman (P) value</td>
<td>Friedman (P) value</td>
<td>Friedman (P) value</td>
<td>Friedman (P) value</td>
</tr>
<tr>
<td>10%</td>
<td>15.6256</td>
<td>0.0036</td>
<td>34.4896</td>
<td>0.0000</td>
<td>34.9368</td>
<td>0.0000</td>
<td>30.4928</td>
<td>0.0000</td>
</tr>
<tr>
<td>20%</td>
<td>10.7848</td>
<td>0.0291</td>
<td>11.3688</td>
<td>0.0227</td>
<td>16.7144</td>
<td>0.0022</td>
<td>11.1014</td>
<td>0.0254</td>
</tr>
<tr>
<td>40%</td>
<td>2.8416</td>
<td>0.5847</td>
<td>2.5568</td>
<td>0.6345</td>
<td>4.9496</td>
<td>0.2925</td>
<td>5.9448</td>
<td>0.2033</td>
</tr>
</tbody>
</table>
because it minimizes the median and presents the smallest dispersion compared with Eigenvector and Eigenvector0. The five methods all present similar results for the 40% deletion rate.

Table 2 shows the Wilcoxon test results for the 10% and 20% percentage of missing values using RMSPD\((genmult)\). There are significant differences among Eigenvector and Eigenvector2, Eigenvector3, and Eigenvector0 for the 10% deletion rate. Differences were found between Eigenvector1 and Eigenvector0, Eigenvector2 and Eigenvector3, respectively. For 20%, Eigenvector1 was different from Eigenvector2 and Eigenvector3; besides, there is a difference between Eigenvector0 and Eigenvector2.

However, Table 3 shows the Wilcoxon test results of the standardized RMSPD\((gen)\) and RMSPD\((env)\) values. In the 10% and 20% imputation percentages, there were significant differences between Eigenvector1 and Eigenvector, Eigenvector2 and Eigenvector3, respectively. Also, significant differences were detected between Eigenvector and the Eigenvector0 and Eigenvector3.

Finally, box plots were made for RMSPD\((env)\), RMSPD\((gen)\), and RMSPD\((mult)\), but are not presented here because they have similar behaviour to those in Figure 3, confirming that Eigenvector1 minimizes the median if it is compared with Eigenvector2 and Eigenvector3 and also has smaller dispersion than Eigenvector0. The method that always maximized all the statistics was Eigenvector, and for this reason it is the least recommended.

3.2. Brazilian Cotton Data. Figure 4 shows the RMSPD\((gen)\) distributions on the standardized scale for the “Farias” data set. The Eigenvector0 distribution is left asymmetric, and this asymmetry decreases as the missing values percentage increases. For the three percentages considered, the Eigenvector0 distribution is above one and very close to the other two, which means that this method had the biggest differences among the additive genotypic parameters of the real and completed (by imputation) data. With 10% imputation, the Eigenvector, Eigenvector2, and Eigenvector3 methods have very similar medians, but the smallest dispersion is achieved.
with Eigenvector2. Overall, when the missing values percentage increases Eigenvector achieves the best performance, because it minimizes RMSPD\(_{(\text{gen})}\). A similar behaviour is shown for RMSPD\(_{(\text{env})}\), as can be observed in Figure 5.

Table 4 shows the Friedman test statistics for RMSPD\(_{(\text{genmul})}\), RMSPD\(_{(\text{envmul})}\), RMSPD\(_{(\text{gen})}\), and RMSPD\(_{(\text{en})}\). There is a significant difference among the imputation methods for all the percentages of missing values, so multiple pairwise comparisons were made with the Wilcoxon test.

Table 5 shows the Wilcoxon tests for RMSPD\(_{(\text{genmul})}\) and RMSPD\(_{(\text{envmul})}\). They indicate that with 10% imputation, the majority of the compared pairs have a significant difference, but, for example, Eigenvector1 is not significantly different from Eigenvector, Eigenvector2 or Eigenvector3. For the other two percentages, 20% and 40%, Eigenvector is not statistically different from Eigenvector2, and Eigenvector3 which have similar performances.

Table 6 shows the Wilcoxon test for RMSPD\(_{(\text{en})}\). With 10% imputation, Eigenvector0 is different from all the others, while for RMSPD\(_{(\text{gen})}\) at the same percentage, Eigenvector1 was statistically different from Eigenvector2. With 20% and 40% of imputation, Eigenvector is not different from Eigenvector2 or Eigenvector3, and likewise Eigenvector3 is not different from Eigenvector2.

In order to make a definitive conclusion, box plots were made for RMSPD\(_{(\text{gen})}\), RMSPD\(_{(\text{en})}\), RMSPD\(_{(\text{genmul})}\), and RMSPD\(_{(\text{envmul})}\), but just one of them is presented because the distribution behaviour is similar for the others. From Figure 6, it can be concluded that the methods that minimize the median in all the percentages are Eigenvector, Eigenvector2, and Eigenvector3, and Tables 5 and 6 show that these methods are equivalent.

In summary, for the “Farias” data set, with the six standardized statistics, Eigenvector always showed good results and is therefore the recommended one.

3.3. Spanish Beans Data. Figure 7 shows the RMSPD\(_{(\text{gen})}\) distribution on the standardized scale for the “Flores” data set. Eigenvector has, in all the percentages, a left asymmetric distribution and maximizes the RMSPD\(_{(\text{gen})}\) median, therefore, it is the method that presents the biggest differences among the main genotypic parameters of the original and completed (by imputation) data. With 10% imputation, Eigenvector0 is the method which presents the best performance, while with 20% it is Eigenvector1 and
with 40% it is Eigenvector2, minimizing the median and
taking the RMSPD(gen) distribution to the bottom of the
standardized scale. Figure 8 presents a similar result, but
using RMSPD(env). From the figure it can be said that
with 20% imputation, Eigenvector0 and Eigenvector1 have
similar medians, but Eigenvector1 is preferred because it
has the smallest dispersion. With RMSPD(env), Eigenvector0
has right asymmetric distributions and Eigenvector1, Eigenvector2, and Eigenvector3 have approximately symmetric
distributions.

Table 7 shows the Friedman test for the statistics
RMSPD$_2$(genmult), RMSPD$_3$(enenv), RMSPD$_2$(genmult),
and RMSPD$_3$(enenv). It can be seen that significant
differences exist among the methods only for 10% imputation.
For this reason we restrict attention to this percentage.

Table 8 shows the 10 pairwise possible comparisons of
imputation methods considering just 10% imputation and
the statistics RMSPD$_2$(genmult), RMSPD$_3$(enenv), RMSPD$_2$(genmult), and RMSPD$_3$(enenv). Taken
across the statistics all the methods are different except
Eigenvector1 and Eigenvector0, but additionally, for
RMSPD$_2$(genmult) the pair Eigenvector1 and Eigenvector2
and for RMSPD$_3$(genmult) the pair Eigenvector2 and
Eigenvector0 are not significantly different.

Finally, to make a definitive conclusion about the four
analyzed statistics in Tables 7 and 8, the box plot for
RMSPD$_3$(genmult) is presented in Figure 9. Plots were made
of the other three statistics, but are not presented here because
the behaviour is similar. According to the box plot, the best
method is Eigenvector0 because it minimizes the median.

4. Discussion

We have presented five imputation methods and tested them
through a simulation study based on three multi-environment
trials and using six statistics derived from RMSPD. Overall,
for big trials (i.e., 450 observations in the data matrix)
Eigenvector should be used under convergence, while for
small trials (i.e., 162 or 180 observations in the data matrix)
two cycles of the process are enough in order to obtain good
results without convergence.

We used experiments with different species, in different
countries, and in different continents. Some of the results

<table>
<thead>
<tr>
<th>Percentage comparison</th>
<th>10%</th>
<th>20%</th>
<th>40%</th>
<th>10%</th>
<th>20%</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSPD$_2$(genmult)</td>
<td>Wilcoxon</td>
<td>Wilcoxon</td>
<td>Wilcoxon</td>
<td>Wilcoxon</td>
<td>Wilcoxon</td>
<td>Wilcoxon</td>
</tr>
<tr>
<td>Eigen-Eigen0</td>
<td>-12.7645*</td>
<td>-11.8392*</td>
<td>-5.5233*</td>
<td>-12.0995*</td>
<td>-11.3270*</td>
<td>-4.4629*</td>
</tr>
<tr>
<td>Eigen-Eigen1</td>
<td>-0.3505</td>
<td>-3.4137*</td>
<td>-3.8021*</td>
<td>-0.1890</td>
<td>-3.0716*</td>
<td>-2.4969*</td>
</tr>
<tr>
<td>Eigen-Eigen2</td>
<td>-2.6235*</td>
<td>-0.7163</td>
<td>-1.4214</td>
<td>-2.4094*</td>
<td>-0.4378</td>
<td>-0.3487</td>
</tr>
<tr>
<td>Eigen-Eigen3</td>
<td>-2.6720*</td>
<td>-0.8664</td>
<td>-0.1908</td>
<td>-2.7633*</td>
<td>-0.9311</td>
<td>-0.4897</td>
</tr>
<tr>
<td>Eigen2-Eigen0</td>
<td>-16.9878*</td>
<td>-13.0576*</td>
<td>-5.3009*</td>
<td>-16.2317*</td>
<td>-12.5528*</td>
<td>-4.9312*</td>
</tr>
<tr>
<td>Eigen3-Eigen0</td>
<td>-13.6133*</td>
<td>-12.6292*</td>
<td>-5.8550*</td>
<td>-12.8970*</td>
<td>-11.9226*</td>
<td>-5.2543*</td>
</tr>
<tr>
<td>Eigen2-Eigen1</td>
<td>-1.7703</td>
<td>-5.0465*</td>
<td>-3.6028*</td>
<td>-1.0721</td>
<td>-4.1600*</td>
<td>-2.8340*</td>
</tr>
<tr>
<td>Eigen3-Eigen1</td>
<td>-0.5797</td>
<td>-4.2466*</td>
<td>-4.3441*</td>
<td>-0.0083</td>
<td>-3.6872*</td>
<td>-3.5903*</td>
</tr>
<tr>
<td>Eigen3-Eigen2</td>
<td>-2.5865*</td>
<td>-1.6592</td>
<td>-1.3257</td>
<td>-2.3910*</td>
<td>-1.1999</td>
<td>-0.6422</td>
</tr>
</tbody>
</table>

* Significant difference 5%.
Figure 4: Box plot of the RMSPD(\(\text{gen}\)) distribution in Farias data set.

Table 7: Friedman test for the standardized RMSPD(\(\cdot\))—Flores data set.

<table>
<thead>
<tr>
<th>Perc.</th>
<th>RMSPD(_2(\text{gen_mult}))</th>
<th>RMSPD(_2(\text{env_mult}))</th>
<th>RMSPD(_3(\text{gen_mult}))</th>
<th>RMSPD(_3(\text{env_mult}))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Friedman</td>
<td>(P) value</td>
<td>Friedman</td>
<td>(P) value</td>
</tr>
<tr>
<td>10%</td>
<td>23.1512</td>
<td>0.0001</td>
<td>39.0136</td>
<td>0.0000</td>
</tr>
<tr>
<td>20%</td>
<td>5.0296</td>
<td>0.2843</td>
<td>2.2608</td>
<td>0.6879</td>
</tr>
<tr>
<td>40%</td>
<td>5.2256</td>
<td>0.2649</td>
<td>3.7480</td>
<td>0.4412</td>
</tr>
</tbody>
</table>
were as expected, but one important outcome is that the iterative aspect of the proposed algorithms should be obligatory when missing values are imputed in $G \times E$ experiments.

So there is a natural question for the applied researcher: how to choose the appropriate Eigenvector imputation method for experiments with different size to those illustrated in this paper? The answer depends on the imputation objective, because the imputation can be used in several ways: to establish one or more genotype-environment combinations that for some reason were not observed, or to follow the imputation with some further statistical modeling. The choice criteria can be extensive, but for the first objective it would be natural to find the imputation errors associated with each Eigenvector method. To find these errors, we can employ cross-validation, using the methodology proposed by Piepho [18] and studied in more detail via simulations in real data by Arciniegas-Alarcón et al. [32]. This methodology is now briefly presented.
Suppose a $G \times E$ experiment is arranged in a table with missing values. From the table of observed values, delete one cell at a time, impute all the missing values, and record the difference between estimated and actual data for the cell under consideration. Do this for all observed cells, and take the average of the squared differences. Denote this quantity by $D$. $D$ contains two components of variability: one due to predictive inaccuracy of the estimate, the other due to sampling error of the observed data. For this reason $D$ may be corrected by subtracting an estimate of the error of a
mean ($s^2$). The square root of $(D - s^2)$ may be taken as the imputation error. The Eigenvector method with smallest imputation error is the method to choose.

On the other hand, if the objective after imputation is inference from the parameter estimates of a statistical model [53, 54], the criterion for choosing the best Eigenvector method can be the standard error of the statistic of interest. The Eigenvector method that produces the smallest standard error will be the best. The modern treatment of missing values suggests multiple imputation as an alternative to find the standard error [55], but in the case of deterministic imputation a solution well known and tested with success can be applied. This is the proportional bootstrap method proposed by Bello [56], in which the proportion of present
and missing values that appear in each bootstrap sample is exactly equal to the proportion that appear in the original incomplete data.

Another aspect that can be of interest is the mechanism producing the missing data. Generally, in situations that involve the assessment of several genotypes in different environments, missing observations follow one of the definitions proposed by Little and Rubin [57], namely, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Values missing completely at random can occur, for example, when plants are damaged due to uncontrollable factors in the experiments,
or by incorrect data measurement or transcription. In this case the cause of the missing value is not correlated with the variable that has it. However, in the genotypes test program in which the cultivars are chosen during each year, using only the observed data without considering the missing values, the missing mechanism is clearly random MAR [58]. The last type of missing, MNAR, can be seen usually when the same subset of genotypes can be missing in some environments of the same subregion, because the plant breeder in the location does not like these genotypes. So, a genotype missing in one environment possibly will be missing too in other environments. In these cases, the mechanism that produces missing values is naturally not at random. The present study has focused exclusively on the MCAR mechanism, and further research is needed to study the remaining mechanisms.

Finally, the proposed methods in this paper have easy computational implementation, but one of the main advantages is that they do not make any distributional or structural assumptions and do not have any restrictions regarding the pattern or mechanism of missing data in $G \times E$ experiments.

**Acknowledgments**

Sergio Arciniegas-Alarcón thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brazil, (PEC-PG program) for the financial support. Marisol García-Peña thanks the National Council of Technological and Scientific Development, CNPq, Brazil, and the Academy of Sciences for the Developing World, TWAS, Italy, (CNPq-TWAS program) for the financial support. Carlos Tadeu dos Santos Dias thanks the CNPq for financial support.

**References**


Submit your manuscripts at http://www.hindawi.com