CBCS
HTTP://AGROBIOL.SGGW.WAW.PL/CBCS

COMMUNICATIONS IN BIOMETRY AND CROP SCIENCE
VOL. 9, NO. 2, 2014, PP. 54–70

INTERNATIONAL JOURNAL OF THE FACULTY OF AGRICULTURE AND BIOLOGY,
WARSAW UNIVERSITY OF LIFE SCIENCES – SGGW, POLAND

REGULAR ARTICLE

Imputing missing values in multi-environment trials using the singular value decomposition: An empirical comparison

Sergio Arciniegas-Alarcón[1]*, Marisol García-Peña[1], Wojtek Krzanowski[2], Carlos Tadeu dos Santos Dias[1]

[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.
*Corresponding author: Sergio Arciniegas-Alarcón; E-mail: sergio.arciniegas@gmail.com

CITATION: Arciniegas-Alarcón, S., García-Peña, M., Krzanowski, W., Dias, C.T.S. (2014). Imputing missing values in multi-environment trials using the singular value decomposition: An empirical comparison. *Communications in Biometry and Crop Science* 9 (2), 54–70.

Received: 19 June 2014, Accepted: 7 October 2014, Published online: 31 October 2014

ABSTRACT

Missing values for some genotype-environment combinations are commonly encountered in multi-environment trials. The recommended methodology for analyzing such unbalanced data combines the Expectation-Maximization (EM) algorithm with the additive main effects and multiplicative interaction (AMMI) model. Recently, however, four imputation algorithms based on the Singular Value Decomposition of a matrix (SVD) have been reported in the literature (Biplot imputation, EM+SVD, GabrielEigen imputation, and distribution free multiple imputation - DFMI). These algorithms all fill in the missing values, thereby removing the lack of balance in the original data and permitting simpler standard analyses to be performed. The aim of this paper is to compare these four algorithms with the gold standard EM-AMMI. To do this, we report the results of a simulation study based on three complete sets of real data (eucalyptus, sugar cane and beans) for various imputation percentages. The methodologies were compared using the normalised root mean squared error, the Procrustes similarity statistic and the Spearman correlation coefficient. The conclusion is that imputation using the EM algorithm plus SVD provides competitive results to those obtained with the gold standard. It is also an excellent alternative to imputation with an additive model, which in practice ignores the genotype-by-environment interaction and therefore may not be appropriate in some cases.

Key Words: *AMMI; genotype × environment interaction; imputation; missing values; singular value decomposition.*

INTRODUCTION

In plant breeding, multi-environment trials are important for testing both 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 interaction, or G×E (Arciniegas-Alarcón et al. 2013).

Often, multi-environment experiments are unbalanced because several genotypes are not tested in some environments. Various methodologies have been proposed in order to solve this lack of balance caused by missing values; a useful list of references about this topic is available in Arciniegas-Alarcón et al. (2011, 2013). One of the first proposals was made by Freeman (1975), who suggested imputing the missing values in an iterative way by minimizing the residual sum of squares and then doing the G×E interaction analysis, reducing the degrees of freedom by the number of missing values. Gauch and Zobel (1990) developed this approach, doing the imputation by using the EM algorithm and incorporating the additive main effects and multiplicative interaction (AMMI) model; this is now known as the EM-AMMI approach. Alternative versions of this procedure using cluster analysis were described in Godfrey et al. (2002) and Godfrey (2004). Raju (2002) treated environments as random effects in the EM-AMMI algorithm, and suggested a robust statistic for the missing values in the stability analysis. Another option is to make the imputation in incomplete two-way tables using linear functions of rows (or columns) as proposed by Mandel (1993). Other methods for handling missing values in G×E experiments that showed good results were developed by Denis (1991), Caliński et al. (1992) and Denis and Baril (1992). They found that using imputations through alternating least squares with bilinear interaction models or AMMI estimates based on robust sub-models can give results as good as those found with the EM algorithm. A different approximation is to work with incomplete data under a mixed model structure with estimates based on maximum likelihood (Kang et al. 2004), but this approach can involve multiple steps and complicated procedures (Yan 2013). Other studies that consider lack of balance in multi-environment experiments are the stability analysis by Raju and Bathia (2003) and Raju et al. (2006, 2009). Finally, Pereira et al. (2007) and Rodrigues et al. (2011) assessed the robustness of joint regression analysis and AMMI models without the use of data imputation.

Recently, Bergamo et al. (2008), Perry (2009a), Arciniegas-Alarcón et al. (2010) and Yan (2013) described imputation systems that involve the Singular Value Decomposition (SVD) of a matrix, and therefore can be applied in any incomplete multi-environment experiments. So, the aim of this paper is to compare these four recent methods (henceforth denoted Biplot imputation, EM-SVD, GabrielEigen imputation and distribution free multiple imputation - DFMI) using as gold standard methodology, that is, the classic EM-AMMI algorithm proposed by Gauch and Zobel (1990).

MATERIALS AND METHODS

IMPUTATION METHODS

***EM-AMMI*:** We first briefly present the AMMI model (Gauch 1988, 1992) for complete experiments. The usual two-way ANOVA model for analysing data from genotype-by-environment trials is given by $y_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}$, $i = 1,...,n;\ j = 1,...,p$

where $\mu, a_i, b_j, (ab)_{ij}$ and e_{ij} are, respectively, the grand mean, the genotypic and environmental main effects, the genotype-by-environment interaction, and the error term associated with the *i*-th genotype and the *j*-th location. It is assumed that all effects except the error are fixed. The AMMI model further considers the interactions as a sum of multiplicative terms, so that the model is written as

$$y_{ij} = \mu + a_i + b_j + \sum_{i,j} \theta_l \alpha_{il} \beta_{jl} + e_{ij} .$$

The terms θ_l, α_{il} and β_{jl} $(l = 1,2,...,)$ can be estimated from the SVD of the interaction matrix: θ_l is estimated by l-th singular value of the SVD, while α_{il} and β_{jl} are estimated by the genotypic and environmental scores corresponding to θ_l. Depending on the number of multiplicative terms that have been included, these models can be called AMMI0, AMMI1, etc.

In incomplete trials, an iterative scheme built round the above procedure is used to obtain AMMI imputations from the EM algorithm. The additive parameters are initially set by computing the grand mean, genotype means and environment means obtained from the observed data. The residuals for the observed cells are initialized as the cell mean minus the genotype mean minus the environment mean plus the grand mean, and interactions for the missing positions are initially set to zero. The initial multiplicative parameters are obtained from the SVD of this matrix of residuals, and the missing values are filled by the appropriate AMMI estimates. In subsequent iterations, the usual AMMI procedure is applied to the completed matrix and the missing values are updated by the corresponding AMMI estimates. Iterations are stopped when changes in successive iterations become 'negligible' (see below).

Depending on the number of multiplicative terms employed, the imputation method may be referred to as EM-AMMI0, EM-AMMI1, etc. (Gauch and Zobel 1990). The studies of Caliński et al. (1992), Piepho (1995), Arciniegas-Alarcón and Dias (2009) and Paderewski and Rodrigues (2014) showed that the best results for imputation with AMMI models are given by including at most one multiplicative component, for which reason our study will consider only the EM-AMMI0 and EM-AMMI1 methods.

It is worth pointing out that the analysis model will not always be the same as the imputation model. The number of components selected for an AMMI-like model analysis must always depend on some tests, and it is common to use the RMSPD statistic for these. However, in this article AMMI will not be assessed as an analysis model but will be evaluated merely as an imputation model, and Arciniegas Alarcón et al. (2011) found that imputation errors associated with AMMI models increase as the number of multiplicative components increases.

***Biplot imputation*:** Recently, Yan (2013) described the following imputation method, using the fact that SVD forms the basis of biplot analysis (Gabriel, 1971; 2002). The method basically consists of substituting the missing values initially by arbitrary values in order to obtain a completed matrix, and then computing the SVD using only two components. This method is now presented more formally.

Biplot step 1. Consider a $n \times p$ matrix **X** with elements x_{ij} ($i=1,...,n$; $j=1,...,p$), where missing entries are denoted by x_{ij}^{aus}. Initially, these missing values are imputed by their respective columns means, thereby providing a completed matrix **X**.

Biplot step 2. The columns of completed matrix **X** are standardised, mean-centering by subtracting m_j and dividing the result by s_j (where m_j and s_j represent the mean and standard deviation of the j-th column). Denoting the standardised elements by p_{ij}, the matrix with elements p_{ij} will be denoted by **P**.

Biplot step 3: The SVD of the **P** matrix is calculated. Considering the first two principal components, we have

$$p_{ij} = \frac{\left(x_{ij} - m_j\right)}{s_j} = \sum_{k=1}^{2} \lambda_k \alpha_{ik} \gamma_{jk} + \varepsilon_{ij} ,$$

with singular values λ_k, eigenvectors for the rows α_{ik} and eigenvectors for the columns γ_{jk} for each of the k PC´s. ε_{ij} is the error for the row i in the column j. Removing this latter error term, the p_{ij} values may be updated, obtaining a new matrix called $\mathbf{P^{(2)}}$ with elements $p_{ij}^{(2)}$.

Biplot step 4: All the elements $p_{ij}^{(2)}$ in $\mathbf{P^{(2)}}$ are returned to their original scale, $\hat{x}_{ij}^{(2)} = m_j + s_j p_{ij}^{(2)}$, thus obtaining a new $\mathbf{X^{(2)}}_{(n \times p)}$ matrix. The missing elements x_{ij}^{aus} in the original matrix **X** are imputed by the corresponding values $\hat{x}_{ij}^{(2)}$ from $\mathbf{X^{(2)}}$.

Biplot step 5: The process is then iterated (back to **Biplot step 2**) until stability is achieved in the imputations. For example, iterations can be continued until the difference between predicted values in the current iteration and those in the previous one, across all missing values, is less than some pre-specified small value. Formally this can be expressed as continuing until $\frac{d}{\bar{y}} < 0.01$ (say), where

$$d = \left[\left(\frac{1}{na} \right) \sum_{i=1}^{na} \left(x_i - x_i^A \right)^2 \right]^{\frac{1}{2}} \text{ and } \bar{y} = \left[\left(\frac{1}{N} \right) \sum_{i=1}^{n} \sum_{j=1}^{p} y_{ij}^2 \right]^{\frac{1}{2}} .$$

Here na is the total number of missing values in the matrix **X**, x_i is the predicted value for the i-th missing cell in the current iteration, x_i^A is the predicted value for the i-th missing cell in the previous iteration, y_{ij} is the observed value (not missing) in the i-th row and j-th column, and N is the total number of observed values.

EM-SVD: Perry (2009a) presents the following imputation method that combines the EM algorithm with SVD. This method replaces the missing values of a G×E matrix initially by arbitrary values to obtain a completed matrix, and a SVD is then computed iteratively on this matrix. At the end of the process, when the iterations reach stability, a matrix containing the imputations for the missing values is obtained. We now present the method more formally.

Consider the $n \times p$ matrix **A** with elements A_{ij} ($i=1,...,n$; $j=1,...,p$), some of which are missing.

EM-SVD step 1: Let $I = \{(i, j) : A_{ij} \text{ isn't missing}\}$, the set of all the observed values.

EM-SVD step 2: For $1 \le j \le p$ let μ_j be the mean of the non-missing values in column j of **A;** set μ_j to 0 if all of the entries in column j are missing.

EM-SVD step 3: Define $\mathbf{A}^{(0)}$ by

$$A_{ij}^{(0)} = \begin{cases} A_{ij} \text{ if } (i,j) \in I \\ \mu_j \text{ otherwise} \end{cases}$$

EM-SVD step 4: Initialize the iteration count, $N \leftarrow 0$.

EM-SVD step 5: (Maximization) Compute the SVD: $\mathbf{A}^{(N)} = \sum_{i=1}^{p} d_i^{(N)} \mathbf{u}_i^{(N)} \mathbf{v}_i^{(N)T}$; let $\mathbf{A}_k^{(N)}$ denote the SVD truncated to *k* terms: $\mathbf{A}_k^{(N)} = \sum_{i=1}^{k} d_i^{(N)} \mathbf{u}_i^{(N)} \mathbf{v}_i^{(N)T}$.

EM-SVD step 6: (Expectation) Define the n×p matrix $\mathbf{A}^{(N+1)}$ as

$$A_{ij}^{(N+1)} = \begin{cases} A_{ij} \text{ if } (i,j) \in I \\ A_{k,ij}^{(N)} \text{ otherwise} \end{cases}$$

EM-SVD step 7: Set $RSS^{(N)} = \left\| \mathbf{A} - \mathbf{A}_k^{(N)} \right\|_{F,I}^2$. If $\left| RSS^{(N)} - RSS^{(N-1)} \right|$ is less than some pre-specified small value, then stop and output $\mathbf{A}_k^{(N)}$, which contains the imputed missing values. Otherwise, increment $N \leftarrow N+1$ and return to **EM-SVD step 5.**

***GabrielEigen imputation*:** Arciniegas-Alarcón et al. (2010) proposed the following imputation method that combines regression and lower-rank approximation using SVD. This method initially replaces the missing cells by arbitrary values, and subsequently the imputations are refined through an iterative scheme that defines a partition of the matrix for each missing value in turn and uses a linear regression of columns (or rows) to obtain the new imputation. In this regression the design matrix is approximated by a matrix of lower rank using the SVD. The algorithm is now presented more formally.

Consider the *n*×*p* matrix **X** with elements x_{ij} (*i=1,...,n; j=1,...,p*), some of which are missing. Note that this process requires $n \geq p$, and if this is not the case then the matrix **X** should first be transposed.

GabrielEigen step 1: The missing values are imputed initially by their respective column means, giving a completed matrix **X**.

GabrielEigen step 2: The columns are standardised, mean-centering by subtracting m_j and dividing the result by s_j (where m_j and s_j represent the mean and the standard deviation of the *j*-th column).

GabrielEigen step 3: Using the standardised matrix, define the following partition

$$\mathbf{X} = \begin{bmatrix} x_{ij} & \mathbf{x}_{\bullet 1}^T \\ \mathbf{x}_{\bullet 1} & \mathbf{X}_{11} \end{bmatrix},$$

where the missing value in position (i,j) is always in position (1,1) of the defined partition. For each missing value x_{ij} the components from the considered partition will be different, and this partition is obtained through elementary operations on the rows and columns of **X**.

Replace the submatrix $\mathbf{X}_{11}$ by its rank m approximation using the singular value decomposition (SVD): $\mathbf{X}_{11} = \sum_{k=1}^{m} \mathbf{u}_{(k)} d_k \mathbf{v}_{(k)}^T = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m]$, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m]$, $\mathbf{D} = diag(d_1, ..., d_m)$ and $m \le \min\{n-1, p-1\}$. The imputation of x_{ij} is given by $x_{ij}^{(m)} = \mathbf{x}_{1\bullet}^T \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T \mathbf{x}_{\bullet 1}$

GabrielEigen step 4: The imputation process is governed by the value of m, and it is suggested that m is chosen to be the smallest value for which

$$\frac{\sum_{k=1}^{m} d_k^2}{\sum_{k=1}^{\min\{n-1, p-1\}} d_k^2} \approx 0.75 .$$

GabrielEigen step 5: Finally, the imputed values must be returned to their original scale, $x_{ij} = m_j + s_j \hat{x}_{ij}^{(m)}$, replacing them in the matrix **X**. Then the process is iterated (back to **GabrielEigen step2**) until stability is achieved in the imputations.

Distribution free multiple imputation (DFMI): The distribution free multiple imputation (DFMI) method proposed by Bergamo et al. (2008) is an iterative scheme that uses the SVD of a matrix to predict missing values in a $n \times p$ matrix **Y**. As with the previous method, this method requires $n \ge p$ so if $n < p$ the matrix should first be transposed. Consider first just one missing value y_{ij} in **Y**. Then, the i-th row from **Y** is deleted and the SVD for the $(n-1) \times p$ resulting matrix $\mathbf{Y}^{(-i)}$ is calculated, where $\mathbf{Y}^{(-i)} = \overline{\mathbf{U}}\overline{\mathbf{D}}\overline{\mathbf{V}}^T$, $\overline{\mathbf{U}} = (\bar{u}_{sh})$, $\overline{\mathbf{V}} = (\bar{v}_{sh})$, $\overline{\mathbf{D}} = (\bar{d}_1, ..., \bar{d}_p)$. The next step is to delete the j-th column from **Y** and obtain the SVD for the $n \times (p-1)$ matrix $\mathbf{Y}_{(-j)}$, where $\mathbf{Y}_{(-j)} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^T$, $\tilde{\mathbf{U}} = (\tilde{u}_{sh})$, $\tilde{\mathbf{V}} = (\tilde{v}_{sh})$, $\tilde{\mathbf{D}} = (\tilde{d}_1, ..., \tilde{d}_{p-1})$. The matrices $\overline{\mathbf{U}}$, $\overline{\mathbf{V}}$, $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ are orthonormal, $\tilde{\mathbf{D}}$ and $\overline{\mathbf{D}}$ are diagonal matrices. Now, combining the two SVDs, $\mathbf{Y}^{(-i)}$ and $\mathbf{Y}_{(-j)}$, the imputed value is given by $\hat{y}_{ij} = \sum_{h=1}^{H} (\tilde{u}_{ih} \tilde{d}_h^{\,a})(\bar{v}_{jh} \bar{d}_h^{\,1-a})$, where $H = \min\{n-1, p-1\}$ and a is the weight in the interval [0,1] given to $\mathbf{Y}_{(-j)}$. Specification of a automatically determines the weight for $\mathbf{Y}^{(-i)}$. For example, a weight of 40% for $\mathbf{Y}_{(-j)}$ requires a=0.4 and the weight for $\mathbf{Y}^{(-i)}$ will be 60% or 1–a=0.6. Bergamo et al. (2008) affirm that 5 imputations for each missing value is sufficient to determine the variability among imputations, and for this reason suggest using weights of 40%, 45%, 50%, 55% and 60% for $\mathbf{Y}_{(-j)}$, namely, a=0.4, 0.45, 0.50, 0.55 and 0.60. Each value of a will provide a different imputation.

For more than one missing value, an iterative scheme is required as follows. Initially all missing values are replaced by their respective column means, giving a completed matrix **Y,** and then the columns are standardised by subtracting m_j from each element and dividing the result by s_j (where m_j and s_j represent the mean and the standard deviation of the j-th column calculated only from the observed values). Using the standardised matrix, the imputation for each missing value is recalculated using the expression for $\hat{y}_{ij}$. For the calculations of each estimate we need $\mathbf{Y}^{(-i)}$ and $\mathbf{Y}_{(-j)}$, which are also standardised. Finally,

the matrix **Y** is returned to its original scale, $y_{ij} = m_j + s_j \hat{y}_{ij}$. The process is iterated until stability is achieved in the imputations. In all our analyses the mean of the five imputations was used as the estimate of each missing value.

Arciniegas-Alarcón and Dias (2009) showed that in some cases imputation with models AMMI0, AMMI1 and AMMI2 can provide better results than imputation with DFMI. For this reason, some modifications can be made to the components of the expression for $\hat{y}_{ij}$, taking into account the work of Caliński et al. (1999) and Bro et al. (2008). First, the DFMI is based on the cross-validation method development by Eastment and Krzanowski (1982), and the estimates $\hat{y}_{ij}$ are biased because the matrices $\tilde{\mathbf{D}}$ and $\bar{\mathbf{D}}$ systematically underestimate $\mathbf{D}$. On average, this bias can be eliminated by correcting $\bar{\mathbf{D}}$ by a factor of $\sqrt{n/(n-1)}$ and $\tilde{\mathbf{D}}$ by a factor $\sqrt{p/(p-1)}$ (Bro et al. 2008). Second, in the DFMI method, $H = \min\{n-1, p-1\}$ in order to use the maximum amount of information available in the matrix, but Caliński et al. (1999) suggest that residual dispersion of the interaction measured by the eigenvalues is close to 75%. Taking this into account, the choice of *H* was modified to $H = \min\{w, k\}$ with *w* such that $\left(\sum_{h=1}^{w} \bar{d}_h^2 \Big/ \sum_{h=1}^{\min\{n-1,p-1\}} \bar{d}_h^2 \right) \approx 0.75$ and *k* such that $\left(\sum_{h=1}^{k} \tilde{d}_h^2 \Big/ \sum_{h=1}^{\min\{n-1,p-1\}} \tilde{d}_h^2 \right) \approx 0.75$.

THE DATA

In order to compare the imputation methods we considered three data sets, published in Lavoranti (2003, p. 91), Flores et al. (1998) and Santos (2008, p. 37). In each case the data were obtained from a randomized complete block design with replicates, and each reference provides full details of the corresponding design.

The data in the matrix "Lavoranti" came from experiments conducted in seven environments, in the south and southeast regions of Brazil, for 20 *Eucalyptus grandis* progenies from Australia. This was a randomized block design, with 6 plants per plot and 10 replicates in a space of dimension 3.0m by 2.0m. The studied variable was the mean tree height in meters (m). The data matrix has size 20×7.

The second data set "Santos" is a 15×13 matrix, with 15 sugar cane varieties assessed in 13 environments in Brazil. The experiment was conducted under the breeding program of RIDESA (Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético), where the studied variable was sugar cane yield (t/ha).

The third data set "Flores" is a 15×12 matrix, with 15 faba bean varieties assessed in 12 environments in Spain. The experiments were conducted by RAEA (Red Andaluza de Experimentación Agraria), and the studied variable was faba bean yield (kg/ha).

In Table 1, we present results from a preliminary study about the choice of the number of multiplicative components (to explain the G×E interaction) of the AMMI model that can be used for each selected data set, over which a simulation study is described below. The method of cross-validation "leave-one-out" by eigenvector (Bro et al. 2008, Gauch 2013) was used to select each model, the best model being the one that has the lowest PRESS statistic. It can be seen that an appropriate model for the "Lavoranti" data matrix is AMMI2, an AMMI1 model is appropriate for the data matrix "Santos" and an AMMI4 model for the "Flores" data. This preliminary study justifies the choice of data sets to evaluate imputation methods.

Table 1. Values of Predicted REsidual Sum of Squares (PRESS) using cross validation by eigenvector in choosing the AMMI model to explain the interaction in the original (complete) data matrices.

		PRESS	
Model	Lavoranti	Flores	Santos
AMMI1	75.1109	113.4852	**101.2898**
AMMI2	**73.8176**	125.1817	115.8897
AMMI3	100.3585	119.7403	129.3862
AMMI4	134.3914	**108.7088**	121.9672
AMMI5	575.0878	141.2513	159.6933
AMMI6	56146.3357	171.0122	146.1308
AMMI7	133.0000	289.9653	177.4717
AMMI8		718.3669	224.5685
AMMI9		1297.3257	234.6328
AMMI10		5204.1570	396.0781
AMMI11		20406.6026	406.0652
AMMI12		168.0000	119156.7004
AMMI13			182.0000

SIMULATION STUDY

Each original data matrix ("Lavoranti", "Flores", "Santos") was submitted to random deletion at three different percentages, namely 10%, 20%, and 40%. The process was repeated 1000 times for each percentage of missing values, giving a total of 3000 different matrices with missing values. Altogether, therefore, there were 9000 incomplete data sets, and for each one the missing values were imputed with the imputation algorithms described above using computational code in R (R Core Team 2013).

The random deletion process for a matrix $\mathbf{X}_{(n\times p)}$ was 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 $(i+1, j)$ position of the matrix was deleted $(i = 0,1,\ldots,n-1;\ j = 1,\ldots,p)$. The expected proportion of missing values in the matrix will be r (Krzanowski 1988). This technique was used with r = 0.1, 0.2 and 0.4.

For EM-SVD, the computational implementation "*bcv*" provided by Perry (2009b) in the R package was used, along with the function *impute.svd*, to make the imputations in a matrix with missing values. In the application of the EM-SVD algorithm, a prior choice is needed for each simulated incomplete matrix for the number of components k to use in the SVD. For this we used the function *cv.SVDImpute* from the "*imputation*" package by Wong (2013), which conducts cross-validation on the available information in the following way. From the matrix with observed values a random deletion is made of 30% of them, and then the EM-SVD imputation is conducted. Using the completed data, the root mean squared error (RMSE) is then computed for those data entries that were randomly deleted. This is repeated for a range of values of k, and the value of k with lowest RMSE is the one chosen to make the imputation. In this study, in each simulated incomplete matrix the cross-validation process was repeated 100 times and the selected k was the one that most often minimised the RMSE.

COMPARISON CRITERIA

Three criteria were used in order to compare the true values with the results obtained in the simulations: the Procrustes statistic M^2 (Krzanowski 2000), the normalised root mean squared error – NRMSE (Ching et al. 2010) and the Spearman correlation coefficient.

For the first criterion, each completed data matrix (observed+imputed) $\mathbf{Y}_{\text{imp}}$ was compared with the original matrix $\mathbf{X}_{\text{orig}}$ (before removing the data) using $M^2 = trace(\mathbf{X}^C_{orig}\mathbf{X}^{C\ T}_{orig} + \mathbf{Y}^C_{orig}\mathbf{Y}^{C\ T}_{orig} - 2\mathbf{X}^C_{orig}\mathbf{Q}\mathbf{Y}^{C\ T}_{orig})$, where $\mathbf{X}^C_{orig}$ and $\mathbf{Y}^C_{orig}$ are mean centered matrices, and $\mathbf{Q} = \mathbf{V}\mathbf{U}^T$ is the rotation matrix calculated from elements of the SVD of the matrix $\mathbf{X}^{C\ T}_{orig}\mathbf{Y}^C_{orig} = \mathbf{U}\Sigma\mathbf{V}^{\mathbf{T}}$. The M^2 statistic measures the difference between two configurations of points, so we look for the imputation method that minimises this difference.

The second criterion used was $NRMSE = \dfrac{\sqrt{mean\left(\mathbf{a}_{imp} - \mathbf{a}_{orig}\right)^2}}{sd\left(\mathbf{a}_{orig}\right)}$, where $\mathbf{a}_{imp}$ and $\mathbf{a}_{orig}$ are vectors containing the respective predicted and true values of the simulated missing observation and $sd\left(\mathbf{a}_{orig}\right)$ is the standard deviation of the values contained in the vector $\mathbf{a}_{orig}$. The best imputation method is the one with the lowest value of NRMSE.

The last comparison criterion considered was the Spearman correlation coefficient (Sprent and Smeeton 2001). This non-parametric correlation coefficient was calculated between each missing value and its corresponding true value. The imputation algorithm with the highest correlation provides the best performance. The non-parametric measure was used in order to avoid distribution problems in the data, since the Pearson correlation coefficient is strongly dependent on the normal distribution of variables.

RESULTS AND DISCUSSION

"LAVORANTI" MATRIX

Table 2 shows the NRMSE means and medians. The imputation methods maximising the criterion are Biplot and EM-AMMI1 for all the imputation percentages, so they are the poorest. The best imputation method is EM-SVD, with means of 0.2690, 0.2649 and 0.2774 for imputation rates of 10%, 20% and 40%, respectively. It can be seen that the GabrielEigen, DFMI and EM-AMMI0 algorithms obtain better results that the classical method that considers the AMMI1 model for imputation. So, according to NRMSE the most efficient method is EM-SVD, followed by EM-AMMI0, GabrielEigen, DFMI and last the EM-AMMI1 and Biplot methods.

Figure 1 shows the distributions of the Procrustes statistic M^2 for the different imputation percentages. The lower the statistic, the better is the imputation algorithm. So the poorest algorithms according to this criterion are the Biplot and EM-AMMI1 methods with a right asymmetric distribution. The algorithms DFMI, GabrielEigen, EM-SVD and EM-AMMI0 minimise the statistic and have comparable behaviour at 10% and 20% rates with approximately symmetric distributions. On the other hand, at 40% deletion the dispersions of DFMI and GabrielEigen increase, so the best algorithms at this percentage are EM-SVD and EM-AMMI0.

Table 2. NRMSE means and medians for the "Lavoranti" matrix

	Percentages of values deleted randomly					
	10%		20%		40%	
Method	Mean	Median	Mean	Median	Mean	Median
Biplot	0.3782	0.3454	0.4197	0.4024	0.5326	0.5217
DFMI	0.2870	0.2719	0.2774	0.2724	0.2881	0.2842
GabrielEigen	0.2780	0.2648	0.2757	0.2705	0.2920	0.2870
EM-SVD	0.2690	0.2552	0.2649	0.2617	0.2774	0.2756
EM-AMMI0	0.2705	0.2591	0.2661	0.2629	0.2776	0.2768
EM-AMMI1	0.4268	0.3993	0.4106	0.3918	0.4239	0.4103

Additionally, the Friedman non-parametric test was used to investigate differences among the M^2 values for the methods DFMI, GabrielEigen, EM-SVD and EM-AMMI0 at all the percentages. The tests were significant for all the cases (P<0.001), so the Wilcoxon test was subsequently used to make paired comparisons. Since the gold standard methodology was EM-AMMI, only the multiple comparisons that included the EM-AMMI0 method were considered. For 20% and 40% deletion the EM-AMMI0 method showed significant differences (P<0.001) with each of DFMI, GabrielEigen and EM-SVD. For 10% deletion no significant difference was found between EM-AMMI0 and EM-SVD. In summary, therefore, for M^2 the imputation methods in decreasing order of efficiency are EM-AMMI0, EM-SVD, GabrielEigen, DFMI, EM-AMMI1 and in last place Biplot.

Finally, the correlation coefficient distributions that were calculated in each simulated data set to compare the imputations with the real data are presented in Figure 2. The median of the correlations is high (> 0.85) for all the considered imputation systems, but the distributions with the highest dispersion are those of Biplot and EM-AMMI1. As the imputation percentage increases the lowest dispersion and highest median is obtained with the EM-SVD and EM-AMMI0 systems. According to the distributions, with 10% and 20% imputation any of the 4 systems DFMI, GabrielEigen, EM-SVD or EM-AMMI0 can be used, but if the missing values percentage is higher then EM-SVD and EM-AMMI0 give the best results.

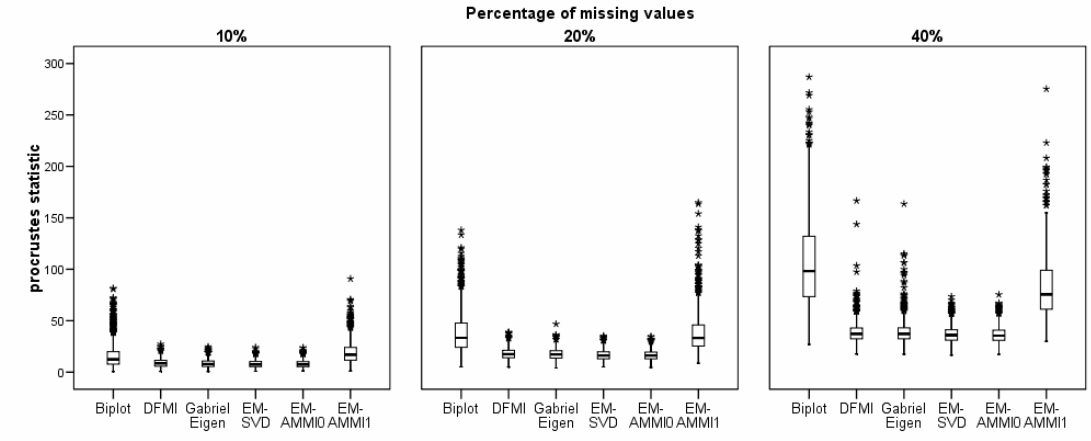

Figure 1. M^2 with different imputation percentages for the "Lavoranti" matrix

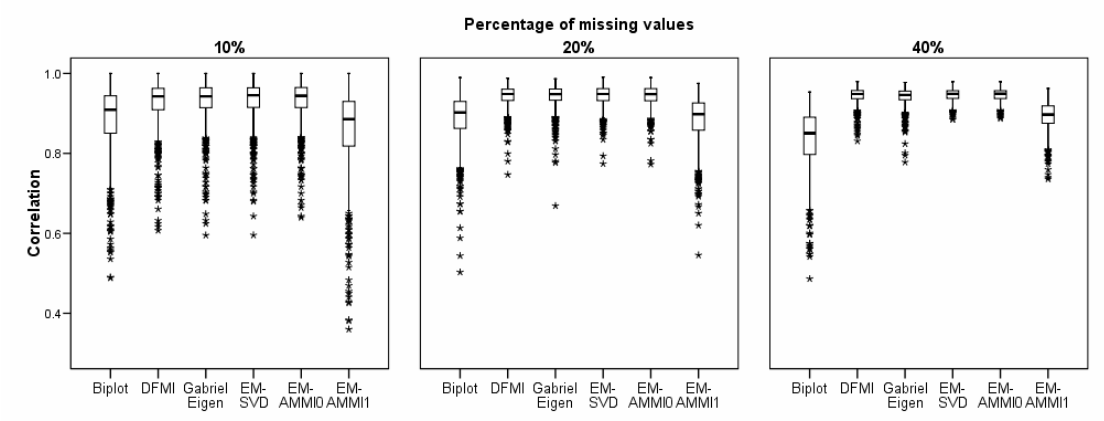

Figure 2. Box plot of the correlation distribution between real and imputed data for the "Lavoranti" matrix for 10%, 20% and 40% of missing values

"FLORES" MATRIX

Table 3 shows the NRMSE means and medians. EM-AMMI1 and Biplot maximize the criterion for all the imputation percentages, and therefore, as for the "Lavoranti" matrix, these are the poorest algorithms. The best imputation algorithm according to this criterion is EM-SVD, with means of 0.3947, 0.3899 and 0.4034 for imputation of 10%, 20% and 40% respectively; GabrielEigen, DFMI and EM-AMMI0 have better results than the classic algorithm EM-AMMI1 at all imputation percentages. Note that for 10% and 20% missing values, GabrielEigen have a better performance than DFMI, but when the percentage increase to 40% the opposite is the case. Thus, when NRMSE is considered, for the "Flores" matrix there are three groups. The first one contains the most efficient algorithms, namely EM-SVD and EM-AMMI0. The second one consists of the least efficient algorithms, EM-AMMI1 and Biplot, while the third one with the intermediate efficiency algorithms contains GabrielEigen and DFMI.

Table 3. NRMSE means and medians for the "Flores" matrix

	Percentages of values deleted randomly					
	10%		20%		40%	
Method	Mean	Median	Mean	Median	Mean	Median
Biplot	0.4837	0.4578	0.4781	0.4676	0.5695	0.5603
DFMI	0.4628	0.4367	0.4571	0.4458	0.4624	0.4548
GabrielEigen	0.4349	0.4090	0.4385	0.4251	0.4817	0.4734
EM-SVD	0.3947	0.3839	0.3899	0.3846	0.4034	0.4004
EM-AMMI0	0.3977	0.3810	0.3925	0.3858	0.4041	0.4036
EM-AMMI1	0.9553	0.5812	1.0818	0.6038	1.3877	0.8204

Using M^2, the EM-AMMI1 method had a very large dispersion so is the poorest method, and is therefore excluded from the comparisons that use the Procrustes statistic. Figure 3 presents the M^2 statistic distribution for the different imputation percentages. The method that maximises M^2 for all the considered percentages is Biplot while the methods that

minimise M^2 for all percentages are EM-SVD and EM-AMMI0. These latter methods have approximately symmetric distributions and the least dispersion among the considered methods. Once again, DFMI and GabrielEigen methods have intermediate efficiency. To determine whether EM-SVD or EM-AMMI0 is better, the Wilcoxon non-parametric test was used, giving a non-significant result (P=0.060) for 10% deletion and significance (P<0.001) for the other two percentages. At 20% and 40% deletion, the smaller medians show that EM-AMMI0 has the greatest similarity between the original matrix and the different matrices containing imputations.

Finally, Figure 4 shows the correlation coefficient distributions that were calculated in each simulated data set between imputations and the corresponding real data. For all the percentages, the biggest median and the smallest dispersion were obtained with EM-SVD and EM-AMMI0, so these two systems are the best according to the Spearman correlation coefficient. As with NRMSE and M^2, the poorest methods with the lowest correlations and high dispersion were EM-AMMI1 and Biplot while GabrielEigen and DFMI again exhibit intermediate efficiency.

"SANTOS" MATRIX

Table 4 shows the means and medians of NRMSE for the "Santos" matrix. In this case, the best imputation system is clearly EM-SVD because it minimises the statistic for all imputation percentages. Next comes EM-AMMI0, followed by GabrielEigen and DFMI. The poorest methods were Biplot and EM-AMMI1. As for both the "Flores" and "Lavoranti" matrices, all methods perform better than the system based on the AMMI1 model.

Table 4. NRMSE means and medians for the "Santos" matrix

	Percentages of values deleted randomly					
	10%		20%		40%	
Method	Mean	Median	Mean	Median	Mean	Median
Biplot	0.5858	0.5736	0.5724	0.5634	0.6065	0.6028
DFMI	0.5296	0.5209	0.5048	0.4982	0.4850	0.4788
GabrielEigen	0.4637	0.4520	0.4678	0.4602	0.4922	0.4843
EM-SVD	0.4147	0.4075	0.4132	0.4096	0.4233	0.4209
EM-AMMI0	0.4207	0.4100	0.4198	0.4184	0.4297	0.4261
EM-AMMI1	0.6731	0.5350	0.7946	0.5690	1.0414	0.7440

Turning next to the M^2 criterion, the EM-AMMI1 algorithm again had very large dispersions, so it was not compared with the other five methods for this criterion. The M^2 distributions are presented in Figure 5. The poorest performance is shown by Biplot, maximising M^2 as well as having the largest dispersion, while the methods with the best performance are EM-SVD and EM-AMMI0 with approximately symmetric distributions. GabrielEigen and DFMI methods again performed better than Biplot, but poorer than EM-SVD and EM-AMMI0.

As was done for the "Flores" data, a Wilcoxon test was conducted between EM-SVD and EM-AMMI0, resulting in a significant difference (P<0.001) for 10% and 20% deletion rates, but not for 40% deletion (P=0.074). Comparing the M^2 median values, the smallest ones were obtained with EM-SVD at all percentage rates.

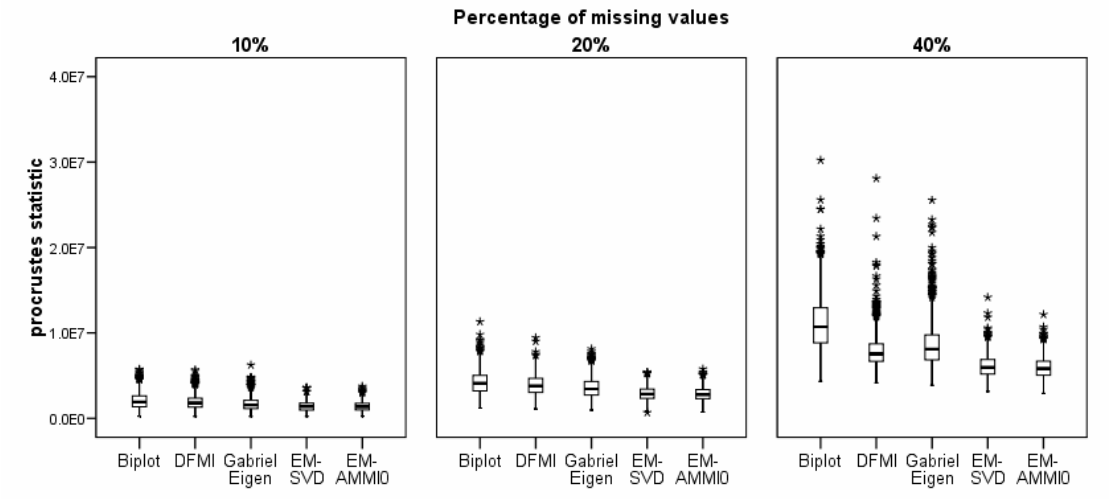

Figure 3. M^2 with different imputation percentages for the "Flores" matrix

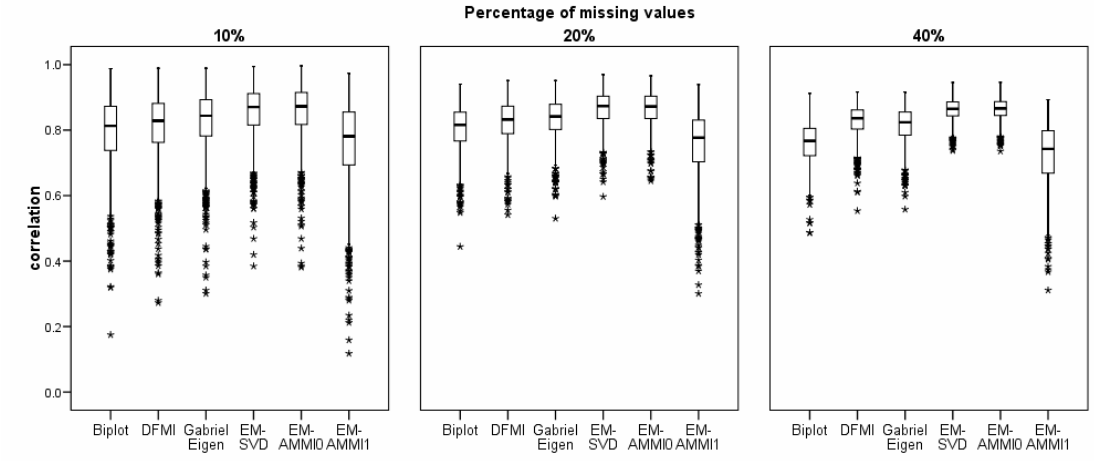

Figure 4. Box plot of the correlation distribution between real and imputed data for the "Flores" matrix for 10%, 20% and 40% of missing values

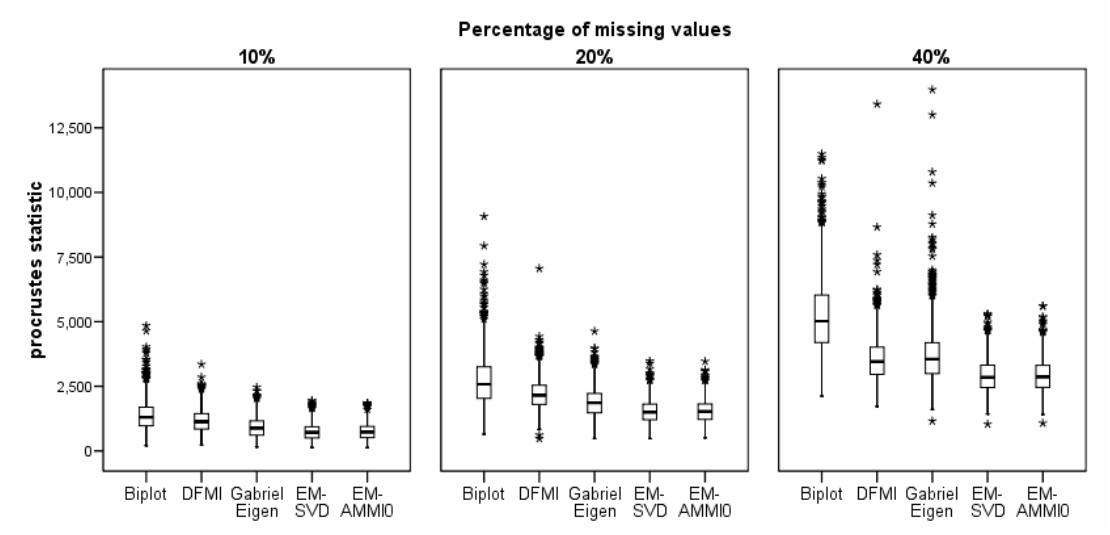

Figure 5. M^2 with different imputation percentages for the "Santos" matrix

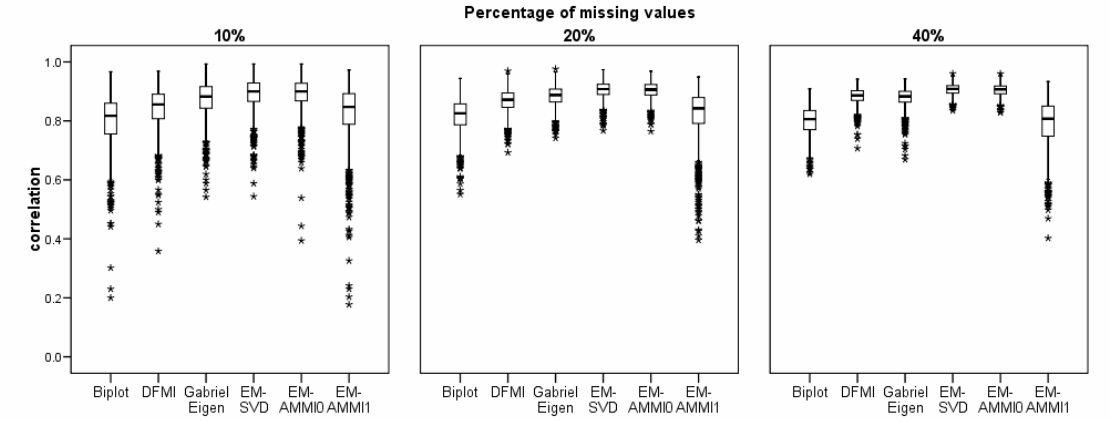

Figure 6. Box plot of the correlation distribution between real and imputed data for the "Santos" matrix for 10%, 20% and 40% of missing values

Figure 6 presents the correlation coefficient distributions. The best performers were EM-SVD and EM-AMMI0 which had the maximum correlation at all the imputation percentages, approximately symmetric distributions, and the lowest dispersions with high medians (>0.88). The algorithm with the highest dispersion was EM-AMMI1, while the GabrielEigen and DFMI algorithms were better than Biplot and EM-AMMI1 but poorer than EM-SVD and EM-AMMI0.

CONCLUSIONS

The results reported here from three multi-environmental data matrices provide some guidelines for future research and analysis about missing values in such trials. The methods DFMI, EM-SVD and GabrielEigen recently presented in the literature showed better performances than the algorithm EM-AMMI1, which belongs to the AMMI family of models proposed by Gauch and Zobel (1990). While in previous studies the AMMI1 model

outperformed DFMI (Arciniegas-Alarcón and Dias, 2009), the modifications to the latter proposed in this paper led to DFMI having better results than EM-AMMI1 in all cases.

According to the NRMSE statistic, the EM-SVD algorithm in general outperforms the other methods that use SVD, which is of interest if the principal aim of the researcher is to obtain estimates of missing G×E combinations. While some studies, such as those by Caliński (1992), Denis and Baril (1992), Piepho (1995) and Arciniegas-Alarcón et al. (2011), showed that the additive model was both simple and effective for solving the unbalance problem, the NRMSE results of the present study suggest that EM-SVD always outperforms EM-AMMI0, a method that is based on that model. However, when using the Procrustes criterion the situation changes and the greatest similarity between the original matrix and the matrices containing imputations is shown by EM-AMMI0, followed by EM-SVD.

Overall, therefore, considering jointly the three criteria, NRMSE, M^2 and Spearman correlation between original observations and the corresponding imputations, the most efficient methods are EM-SVD and EM-AMMI0 while the least efficient ones are Biplot and EM-AMMI1. The GabrielEigen and DFMI methods consistently lie intermediately between these two pairs. For 10% and 20% of missing values, GabrielEigen is better than DFMI, but when the percentage increases to 40% then DFMI is preferable. Moreover, DFMI has a characteristic exhibited by none of the other presented methods: it is the only method that provides a variance estimate among the imputations, enabling the uncertainty about the real values to be gauged.

It is important to remember that the analysis model does not have to be the same as the imputation model. For example, in the initial cross-validation study (Table 1) on the original "Santos" matrix the analysis model chosen to explain the interaction was AMMI1, so it might be tempting to think that in case of missing values the best imputation model would be EM-AMMI1. This does not have to be the case, because the missing observations change the structure of the interaction and, moreover, imputation errors associated with AMMI models increase as the number of multiplicative components increases. Thus, the results of this study suggest that the best approach is to impute with the EM-SVD system and then make an interaction analysis using AMMI models.

Further research about incomplete G×E experiments is needed and could involve, for example, other comparison criteria for the algorithms, missing value mechanisms other than missing completely at random (MCAR) as defined by Little and Rubin (2002) and Paderewski and Rodrigues (2014), or other gold standard methodologies, such as the mixed model (Piepho, 1998). Until such research provides contradictory evidence, it appears that EM-SVD is a very competitive alternative to AMMI models and especially so in relation to the additive model, which has the disadvantage of ignoring the interaction – a feature that in some situations can be inappropriate.

ACKNOWLEDGEMENTS

The first author thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brazil, (PEC-PG program) for the financial support. The second author 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.

REFERENCES

Arciniegas-Alarcón S., Dias C.T.S. (2009). Data imputation in trials with genotype by environment interaction: an application on cotton data. *Biometrican Brazilian Journal* 27, 125–138.

Arciniegas-Alarcón S., García-Peña M., Dias C.T.S., Krzanowski W.J.. (2010). An alternative methodology for imputing missing data in trials with genotype-by-environment interaction. *Biometrical Letters* 47, 1–14.

Arciniegas-Alarcón S., García-Peña M., Dias C.T.S. (2011). Data imputation in trials with genotype×environment interaction. *Interciencia* 36, 444–449.

Arciniegas-Alarcón S., García-Peña M., Krzanowski W.J., Dias C.T.S. (2013). *Deterministic imputation in multienvironment trials*. ISRN Agronomy vol. 2013, Article ID 978780, 17 pages. doi:10.1155/2013/978780.

Bergamo G.C., Dias C.T.S., Krzanowski W.J. (2008). Distribution-free multiple imputation in an interaction matrix through singular value decomposition. *Scientia Agricola* 65, 422–427.

Bro R., Kjeldahl K., Smilde A.K., Kiers H.A.L. (2008). Cross-validation of component models: a critical look at current methods. *Analytical and Bioanalytical Chemistry* 390, 1241–1251.

Caliński T., Czajka S., Denis J.B., Kaczmarek Z. (1992). EM and ALS algorithms applied to estimation of missing data in series of variety trials. *Biuletyn Oceny Odmian*, 24/25, 7–31.

Caliński T., Czajka S., Denis J.B., Kaczmarek Z. (1999). Further study on estimating missing values in series of variety trials. *Biuletyn Oceny Odmian* 30, 7–38.

Ching W., Li L., Tsing N., Tai C., Ng T. (2010). A weighted local least squares imputation method for missing value estimation in microarray gene expression data. *International Journal of Data Mining and Bioinformatics* 4, 331–347.

Denis J.B. (1991). Ajustements de modèles linéaires et bilinéaires sous contraintes linéaires avec données manquantes. *Revue de Statistique Appliquée* 39, 5–24.

Denis J.B., Baril C.P. (1992). Sophisticated models with numerous missing values: the multiplicative interaction model as an example. *Biuletyn Oceny Odmian* 24/25, 33–45.

Eastment H.T., Krzanowski W.J. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics* 24, 73–77.

Flores F., Moreno M.T., Cubero J.I. (1998). A comparison of univariate and multivariate methods to analyze G x E interaction. *Field Crops Research* 56, 271–286.

Freeman G.H. (1975). Analysis of interactions in incomplete two-ways tables. *Applied Statistics* 24, 46–55.

Gabriel K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58, 453–467.

Gabriel K.R. (2002). Le biplot - outil d´exploration de données multidimensionelles. *Journal de la Société Française de Statistique* 143, 5–55.

Gauch H.G. (1988). Model selection and validation for yield trials with interaction. *Biometrics* 44, 705–715.

Gauch H.G. (1992). *Statistical analysis of regional yield trials: AMMI analysis of factorial designs*. Elsevier, Amsterdam, The Netherlands.

Gauch H.G., Zobel R.W. (1990). Imputing missing yield trial data. *Theoretical and Applied Genetics* 79, 753–761.

Gauch H.G. (2013). A simple protocol for AMMI analysis of yield trials. *Crop Science* 53, 1860–1869.

Godfrey A.J.R., Wood G.R., Ganesalingam S., Nichols M.A., Qiao C.G. (2002). Two-stage clustering in genotype-by-environment analyses with missing data. *Journal of Agricultural Science* 139, 67–77.

Godfrey A.J.R. (2004). *Dealing with sparsity in genotype x environment analysis*. Dissertation, Massey University.

Kang M.S., M.G. Balzarini, Guerra J.L.L. (2004). *Genotype-by-environment interaction. In: Saxton AM (ed) Genetic analysis of complex traits using SAS*. SAS Institute Inc., Cary, NC, USA.

Krzanowski W.J. (1988). Missing value imputation in multivariate data using the singular value decomposition of a matrix. *Biometrical Letters* XXV, 31–39.

Krzanowski W.J. (2000). *Principles of multivariate analysis: A user's perspective*. Oxford: University Press, Oxford, England.

Lavoranti O.J. (2003). *Phenotypic stability and adaptability via AMMI model with bootstrap re-sampling*. Dissertation. University of São Paulo.

Little R.J.A., Rubin D.B. (2002). *Statistical analysis with missing data, 2nd edn*. Wiley, New York, USA.

Mandel J. (1993). The analysis of two-way tables with missing values. *Applied Statistics* 42, 85–93.

Paderewski J., Rodrigues P.C. (2014). 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* 8, 640–645.

Pereira D.G., Mexia J.T., Rodrigues P.C. (2007). Robustness of joint regression analysis. *Biometrical Letters* 44, 105–128.

Perry P.O. (2009a). *Cross-validation for unsupervised learning*. Dissertation, Stanford University.

Perry P.O. (2009b). *bcv: Cross-Validation for the SVD (Bi-Cross-Validation)*. R package version 1.0.

Piepho H.P. (1995). Methods for estimating missing genotype-location combinations in multilocation trials - an empirical comparison. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* 26, 335–349.

Piepho H.P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theoretical and Applied Genetics* 97, 195–201.

R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Raju B.M.K. (2002). A study on AMMI model and its biplots. *Journal of the Indian Society of Agricultural Statistics* 55, 297–322.

Raju B.M.K., Bhatia V.K. (2003). Bias in the estimates of sensitivity from incomplete GxE tables. *Journal of the Indian Society of Agricultural Statistics* 56, 177–189.

Raju B.M.K., Bhatia V.K., Kumar V.V.S. (2006). Assessment of sensitivity with incomplete data. *Journal of the Indian Society of Agricultural Statistics* 60, 118–125.

Raju B.M.K., Bhatia V.K., Bhar L.M. (2009). Assessing stability of crop varieties with incomplete data. *Journal of the Indian Society of Agricultural Statistics* 63, 139–149.

Rodrigues P.C., Pereira D.G.S., Mexia J.T. (2011). A comparison between joint regression analysis and the additive main and multiplicative interaction model: the robustness with increasing amounts of missing data. *Scientia Agricola* 68, 697–705.

Santos E.G.D. (2008). *Genotypes by locations interaction in sugarcane and perspectives of environmental stratification*. Dissertation. University of São Paulo.

Sprent P., Smeeton N.C. (2001). *Applied Nonparametric Statistical Methods*. Chapman and Hall, London, England.

Wong J. (2013). *Imputation: imputation. R package version 2.0.1*. http://CRAN.R-project.org/package=imputation.

Yan W. (2013). Biplot analysis of incomplete two-way data. *Crop Science* 53, 48–57.