DE GRUYTER
OPEN

DOI: 10.2478/bile-2014-0006

Biometrical Letters
Vol. 51 (2014), No. 2, 75-88

An alternative methodology for imputing missing data in trials with genotype-by-environment interaction: some new aspects

Sergio Arciniegas-Alarcón[1], Marisol García-Peña[1],
Wojtek Janusz 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 - Brasil, e-mail: sergio.arciniegas@gmail.com

[2]College of Engineering, Mathematics and Physical Sciences, Harrison Building,
University of Exeter, North Park Road, Exeter, EX4 4QF, United Kingdom

Summary

A common problem in multi-environment trials arises when some genotype-by-environment combinations are missing. In Arciniegas-Alarcón et al. (2010) we outlined a method of data imputation to estimate the missing values, the computational algorithm for which was a mixture of regression and lower-rank approximation of a matrix based on its singular value decomposition (SVD). In the present paper we provide two extensions to this methodology, by including weights chosen by cross-validation and allowing multiple as well as simple imputation. The three methods are assessed and compared in a simulation study, using a complete set of real data in which values are deleted randomly at different rates. The quality of the imputations is evaluated using three measures: the Procrustes statistic, the squared correlation between matrices and the normalised root mean squared error between these estimates and the true observed values. None of the methods makes any distributional or structural assumptions, and all of them can be used for any pattern or mechanism of the missing values.

Key words: cross-validation, singular value decomposition, imputation, genotype-by-environment interaction, weights, missing values

1. Introduction

In plant breeding, multi-environment trials are important for testing general and specific cultivar adaptation. A cultivar grown in different environments will show significant fluctuations in yield performance relative to other cultivars. These changes are influenced by the different environmental conditions and are referred to as genotype-by-environment interaction or G×E (Dias and Krzanowski, 2003; Gauch, 2013).

Multi-environment trials usually give rise to incomplete data sets (Rodrigues et al., 2011; Bergamo et al., 2008). Possible ways of analysing such trials are: (i) extracting a balanced subset of data by deleting those genotypes or environments that contain missing values (Yan et al., 2011); (ii) filling the missing cells with environmental means; or (iii) filling the missing cells with estimated values obtained from fitted multiplicative or mixed linear models (Kumar et al., 2012; Arciniegas-Alarcón et al., 2011). These strategies may overcome the lack of balance in the data, but none of them is both simple and effective (Yan, 2013). The first strategy does not make use of all available information, the second one may have problems when too many values are missing, and the third one involves multiple steps and complicated procedures (Yan, 2013).

Following Little and Rubin (2002) and Di Ciaccio (2011) we can distinguish between three missing data mechanisms in G×E trials. Data are said to be "missing at random" (MAR) if the mechanism causing the omissions is independent of the unobserved data values. If the omissions are also independent of the observed data values, then the data are said to be "missing completely at random" (MCAR). Finally, if the mechanism causing the omissions depends on the unobserved values, the data are said to be "missing not at random" (MNAR).

Piepho and Möhring (2006) note that in a cultivar testing program, where cultivars are selected each year on the basis of the data thus far collected but not on unobserved data, the missing-data mechanism is clearly MAR. On the other hand, Rodrigues et al. (2011) state that MCAR occurs when the plants may be destroyed by animals, floods or during harvesting, and the yield measurements may be erroneously performed and inadequately introduced into the data base. The third mechanism, MNAR, is considered by some researchers to be the most common one, because in the trials a clear pattern in the shape of the missing values can be found (Paderewski and Rodrigues, 2014). According to Piepho (1995), MNAR occurs when the same subset of genotypes may be missing in a number of environments of the same subregion, because of local growers dislike of those genotypes. Similarly, a genotype missing in one place is likely to be missing in other places as well. In these cases, the mechanism that leads to the missing data is clearly not a random one.

A new distribution-free imputation method was recently proposed for G×E trials using a mixture of regression and lower-rank approximation of a matrix by Arciniegas-Alarcón et al. (2010), who called this method

GabrielEigen. In the present paper we first describe a modification of this method that uses weights chosen by cross-validation, and then go on to provide an extension to the case of multiple imputation.

2. Material and methods

2.1. Data imputation using GabrielEigen

Suppose that the $(n \times p)$ matrix $\mathbf{X}$ contains elements x_{ij} $(i = 1, \dots, n; j = 1, \dots, p)$, some of which are missing. The rows represent genotypes and the columns the environments.

Step 1. Start by inserting into each missing entry the mean of its column, thereby obtaining a completed matrix $\mathbf{X}$.

Step 2. Standardise the columns of $\mathbf{X}$ by subtracting m_j from each element and dividing the result by s_j (where m_j and s_j are respectively the mean and the standard deviation of the jth column).

Step 3. Using the standardised matrix, replace each original missing entry x_{ij} by $\widehat{x}_{ij}^{(m)} = \mathbf{x}_{1\cdot}^T \mathbf{V}\mathbf{D}^+\mathbf{U}^T \mathbf{x}_{\cdot 1}$. Here the vectors $\mathbf{x}_{1\cdot}^T$, $\mathbf{x}_{\cdot 1}$ and the matrices $\mathbf{V}$, $\mathbf{D}$ and $\mathbf{U}$ are obtained from the partition $\mathbf{X} = \begin{bmatrix} x_{ij} & \mathbf{x}_{1\cdot}^T \\ \mathbf{x}_{\cdot 1} & \mathbf{X}_{11} \end{bmatrix}$ with $\mathbf{X}_{11} = \sum_{k=1}^{m} \mathbf{u}_{(k)} d_k \mathbf{v}_{(k)}^T = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_m]$, $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_m]$, $\mathbf{D} = diag(d_1, \dots, d_m)$ and $m \leqslant \min\{n-1, p-1\}$ is chosen to be the smallest value satisfying $(\sum_{k=1}^{m} d_k^2)/(\sum_{k=1}^{\min\{n-1,p-1\}} d_k^2) \approx 0.75$. Thus $\mathbf{U}\mathbf{D}\mathbf{V}^T$ is the singular value decomposition (SVD) of $\mathbf{X}_{11}$ and $\mathbf{D}^+$ is the Moore-Penrose generalised inverse of $\mathbf{D}$. Also, for each missing observation the components of the considered partition will be different, and this partition is obtained through elementary operations on the rows and columns of $\mathbf{X}$.

Step 4. Finally, return the imputed values to their original scale, $x_{ij} = m_j + s_j \widehat{x}_{ij}^{(m)}$, replacing them in the matrix $\mathbf{X}$.

Steps 2 to 4 are iterated until the imputations achieve stability. This process assumes that $n > p$, so if this is not the case the matrix should first be transposed before conducting the iterations.

2.2. Proposed modifications

Our first proposal is to include weights in Step 3, replacing $\widehat{x}_{ij}^{(m)}$ by $\widehat{x}_{ij}^{(w,m)} = w\left(\mathbf{x}_{1\boldsymbol{.}}^T\mathbf{V}\mathbf{D}^+\mathbf{U}^T\mathbf{x}_{\boldsymbol{.}1}\right)$. Here the weight w is obtained by cross-validation, using as the predictive criterion the root mean squared predictive difference (RMSPD) from the observed data in the incomplete matrix (Gauch and Zobel, 1990; Sabaghnia et al., 2012). The modified algorithm, with the new Step 3 but keeping the other steps of GabrielEigen the same, will be called WGabriel.

To see how the weight w is chosen, consider a G×E trial arranged in a table with missing values. From the observed values, and for any specific value of w, delete one cell at a time, impute the deleted value with WGabriel, and record the difference between the 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. The square root of D is the RMSPD based on the observed values, namely RMSPD(obs) for that value of w.

Because the matrix $\mathbf{X}$ is standardised by columns in step 3 of WGabriel, it makes sense to allow the value of w to be positive, negative or zero. Negative w represents a change in the magnitude and direction of the imputation, which means a change from a positive to a negative value, or vice versa, in the standardized scale with the objective of minimising RMSPD(obs). However, the chosen value of w should be such that on returning to the original scale (Step 4), the imputed values $x_{ij} = m_j + s_j\widehat{x}_{ij}^{(w,m)}$ do not lie outside the range of existing values. To achieve this, we suggest that all w values in the interval [-2,2] should be tested at steps of 0.005 or 0.01, i.e. testing a total of either 801 or 401 weights respectively (larger intervals could be considered, but the risk of convergence failures in the algorithm will be increased). The value of RMSPD(obs) is obtained for each of these weights, and the value of w which gives the minimum of this statistic is chosen for analysis. This weight will be denoted w^*.

Josse and Husson (2012) and van Buuren (2012) have warned that simple imputation systems such as WGabriel do not take into account the uncertainty produced by the imputations, and if later parameters are estimated from augmented data that includes the imputed values, the standard error will be underestimated. It is well known that this problem can be solved by the use of multiple imputation (MI) (Rubin, 1978; Josse et al., 2011). MI involves three distinct steps (Bergamo et al., 2008): **1. Imputation**:

The missing values are estimated M times, generating M completed data sets; **2. Analysis**: The M completed data sets are analysed, using appropriate statistical procedures for the problem under study; **3. Combination**: The M separate sets of results are combined into one single inference. In many practical applications it has been found that a high statistical efficiency can be achieved by using M=20 (Schafer and Graham, 2002).

The WGabriel method can be extended to allow distribution-free multiple imputation in the following way. w^* is the weight that provides the best predictive difference, i.e. minimum RMSPD(obs), using all the available information in an incomplete matrix. So to produce 20 or more different completed data sets we could use 20 or more different weights in WGabriel. These weights should all be close to w^*, as otherwise RMSPD may be far from minimum for some of them. Thus, for example, if w^*=0.7 and the previously chosen step was 0.01, we suggest a range for w between 0.6 and 0.8, giving a total of 21 completed data sets. This imputation method will be called MIWG(0.01).

In the following we compare the three imputation methods: GabrielEigen, WGabriel and MIWG(0.01). As MIWG(0.01) is a method that produces multiple completed sets, the mean of the imputed values will be used (Kroonenberg, 2008) so that this method can be compared with the simple deterministic imputation algorithms.

2.3. The data

We consider two data sets, the “Denis-Baril Matrix” and the “Caliński Matrix”. The former is a complete G×E trial with 26 wheat genotypes evaluated in 5 French environments, which was subjected to the arbitrary deletion of 37% of the entries, giving 48 missing values (Denis and Baril, 1992). This data set, available in the free statistical software R (Wright, 2012), is used here only illustratively, to show values of RMSPD for the observed data and for missing values with different weights in WGabriel. The second data set is complete, and will be used to compare the imputation algorithms. It is a matrix of size 18×9, for 18 pea varieties evaluated in 9 different locations in Poland. The experiment was conducted by the Research Centre for Cultivar Testing, Słupia Wielka, and the variable of interest was mean yield in dt/ha (Caliński et al., 2009).

2.4. Simulation study

The "Caliński Matrix" was submitted to random deletion of values at different percentages, namely 10%, 20% and 30%. The process was repeated 100 times for each percentage of missing values, giving a total of 300 incomplete data sets, and in each set the missing values were imputed in turn with the three algorithms described above using code in R (R Development Core Team, 2013). The R code is available from the authors on request.

The random deletion process for a matrix $\mathbf{X}_{(n\times p)}$ was conducted as follows. Random numbers between 0 and 1 were generated in R with the *runif* function. For a fixed value of r $(0 < r < 1)$, if the $(pi + j)$-th random number was lower than r, then the element in position $(i+1, j)$ in position of the matrix was deleted $(i = 0, 1, \dots, n-1; j = 1, \dots, p)$. The expected proportion of missing values in the matrix is thus r (Krzanowski, 1988). This technique was used with r =0.1, 0.2 and 0.3.

2.5. Comparison criteria

Three criteria were used to compare the actual data with the simulation results: the M^2 Procrustes statistic (Krzanowski, 2000); the squared correlation between matrices, $corr^2$ (Gabriel, 2002); and the normalised root mean squared error, NRMSE (Ching et al., 2010). The computational details for each of these criteria now follow.

First, each completed data matrix containing observed+imputed values, $\mathbf{Y}_{imp}$, was compared with the original matrix $\mathbf{X}_{orig}$ using $M^2 = trace(\mathbf{X}_{orig}\mathbf{X}_{orig}^T + \mathbf{Y}_{imp}\mathbf{Y}_{imp}^T - 2\mathbf{X}_{orig}\mathbf{Q}\mathbf{Y}_{imp}^T)$ where $\mathbf{Q} = \mathbf{V}\mathbf{U}^T$ is the rotation matrix calculated from elements of the SVD of the matrix $\mathbf{X}_{orig}^T\mathbf{Y}_{imp} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. The M^2 statistic measures the difference between two configurations of points, so the imputation method that minimises this difference indicates the method that yields the closest match between the original data values and the corresponding imputed ones after deletion.

Similarly, M^2 was used to compare the matrices of interactions $\mathbf{GE}_{orig}$ and $\mathbf{GE}_{imp}$ where $\mathbf{GE}_{orig}$ and $\mathbf{GE}_{imp}$ are the residual matrices after fitting the main effects by ANOVA in the matrices $\mathbf{X}_{orig}$ and $\mathbf{Y}_{imp}$ respectively (García-Peña and Dias, 2009). However, the matrices $\mathbf{GE}_{orig}$ and $\mathbf{GE}_{imp}$ were also compared using the coefficient $corr^2(\mathbf{B}, \widehat{\mathbf{B}}) = \frac{tr^2\left\{\mathbf{B}^T\widehat{\mathbf{B}}\right\}}{tr\{\mathbf{B}^T\mathbf{B}\}tr\left\{\widehat{\mathbf{B}}^T\widehat{\mathbf{B}}\right\}}$, where $\mathbf{B}$ and $\widehat{\mathbf{B}}$ represent respectively the matrices $\mathbf{GE}_{orig}$ and $\mathbf{GE}_{imp}$ centered by columns. The best imputation algorithm with this criterion is the one with highest $corr^2$.

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

3. Results and discussion

3.1. Denis-Baril matrix

The complete matrix was subjected to arbitrary deletion of 48 values, from which we can obtain the RMSPD for the observed and the missing values. Figure 1 shows the two RMSPD curves over different weights. All weights in the interval [-2,2] were considered, but only the weights in the interval [0.5, 1.2] are shown here, because this is where the curves were minimised.

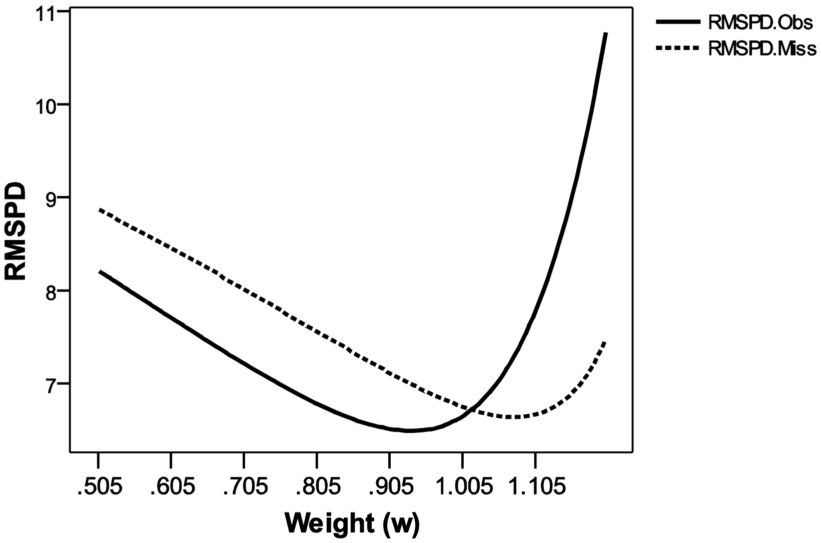

Figure 1. Root mean squared predictive difference (RMSPD) in the Denis-Baril Matrix

The distance between the two curves can be interpreted as the discrepancy between the imputation using RMSPD(obs) and the real RMSPD of the missing values - RMSPD(miss). There is a point of intersection of the two curves, but at this point neither RMSPD(miss) nor RMSPD(obs) is minimised; RMSPD(miss) and RMSPD(obs) are minimised at $w = 1.08$ and $w = 0.935$ respectively. An important conclusion, therefore, is that since the curves are minimised when $w \neq 1$, the WGabriel method is appropriate.

3.2. Caliński matrix

In the previous matrix it was possible to calculate RMSPD(miss) for comparison, but in practical applications this is impossible. For this reason, the simulation study in the "Caliński matrix" takes into account only RMSPD(obs) as the criterion of choice for the weights. Figure 2 presents the M^2 distributions when each completed matrix (i.e. containing observed+imputed values) was compared with the original matrix. Recall that the best imputation method is the one minimising M^2. For 10% deletion the three methods have similar results, but when the imputation percentage increases, M^2 for the GabrielEigen method also increases so IMWG(0.01) and WGabriel are the better methods.

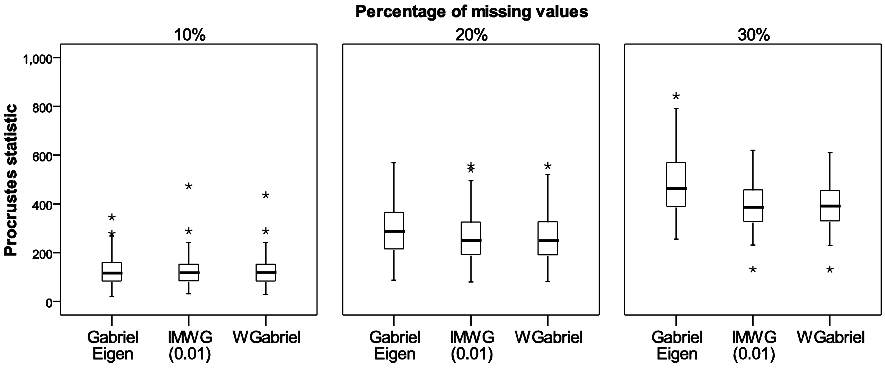

Figure 2. Procrustes statistic (M^2) distribution between imputed values matrices and the Caliński matrix

Figure 3 shows the M^2 values when comparing the $\mathbf{GE}_{imp}$ ANOVA residual matrices obtained from the completed matrices with the $\mathbf{GE}_{orig}$ original matrix. Very similar behaviour to that of Figure 2 is evident: when the percentage of random deletions increases the best imputation methods are MIWG(0.01) and WGabriel. An approximately symmetric distribution is observed at 10% and 20% deletion for all methods. With 30% deletion, the symmetry remains for the M^2 distributions in WGabriel and MIWG(0.01), but the GabrielEigen method has a right-asymmetric distribution.

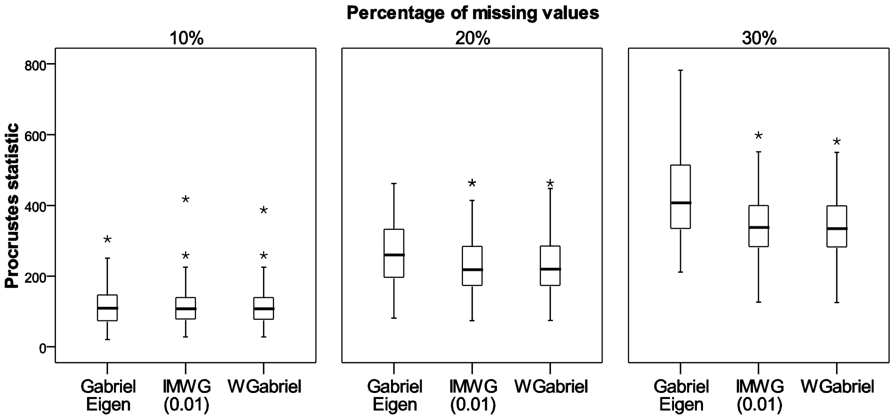

Figure 3. Procrustes statistic (M^2) distribution between the matrices $\mathbf{GE}_{imp}$ and $\mathbf{GE}_{orig}$

The (squared) correlation distributions between the different $\mathbf{GE}_{imp}$ matrices and $\mathbf{GE}_{orig}$ are presented in Figure 4. High correlations with a median of approximately 0.87 were obtained for all three imputation algorithms at 10% deletion. When the deletion percentage increases, the correlation decreases faster for the GabrielEigen method than for the others. At 20% deletion, the *corr*^{2} median for GabrielEigen was 0.7174, while for WGabriel it was 0.7632 and for IMWG(0.01) it was 0.7649.

At 30%, the median *corr*^{2} in the GabrielEigen algorithm decreases to 0.5528, while for WGabriel and IMWG(0.01) the median has values of 0.6292 and 0.6297 respectively. In general, all methods show moderate and high positive correlations. At 10% deletion all methods show symmetric distributions, but at 20% WGabriel and IMWG(0.01) have a left-asymmetric distribution. According to *corr*^{2}, therefore, the best method is IMWG(0.01), with WGabriel second and GabrielEigen last.

The third comparison criterion was NRMSE, and the means and medians of this statistic for the three imputation methods are shown in Table 1. The best method is the one that minimises the value of the statistic. For all deletion percentages the two proposed methods were better than

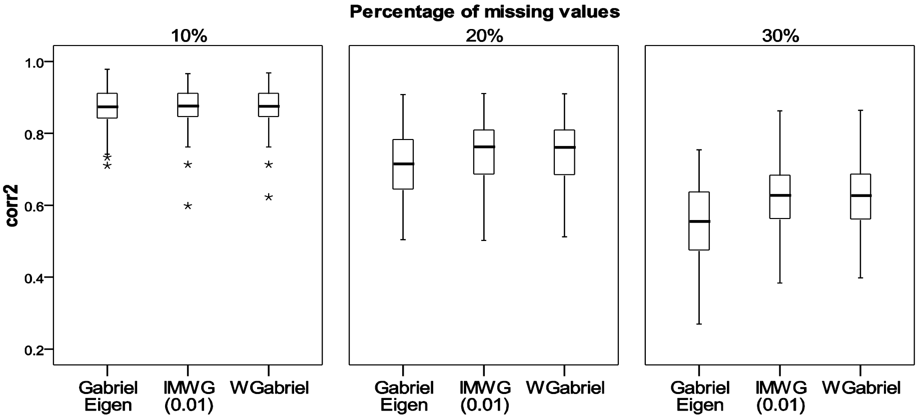

Figure 4. Squared correlation ($corr^2$) distribution between the matrices $\mathbf{GE}_{imp}$ and $\mathbf{GE}_{orig}$

GabrielEigen. At 10% deletion WGabriel was the best method followed by IMWG(0.01). For 20% deletion the ranking was the same but the two mean values were very close. At 30% deletion the mean of IMWG(0.01) is the lowest but the mean of WGabriel is again very close.

Table 1. Mean and median for normalised root mean squared error in the Caliński matrix

	Percentages of values deleted randomly					
	10%		20%		30%	
Method	Mean	Median	Mean	Median	Mean	Median
GabrielEigen	0.3512	0.3382	0.3538	0.3510	0.3709	0.3660
IMWG(0.01)	0.3426	0.3375	0.3293	0.3215	0.3345	0.3277
WGabriel	0.3419	0.3376	0.3292	0.3252	0.3346	0.3289

Finally, an important aspect is that of the weights used in the “Caliński matrix”, which were found by cross-validation for WGabriel, because the distribution-free multiple imputation by MIWG(0.01) method depends on these weights.

Figure 5 shows the weight distributions for WGabriel. When the missing values percentage increases, the weight distribution is more strongly left-asymmetric with a median close to 0.4. In this case the basic statistics of the weight could be of interest in order to assess the centrality parameters and the variability, so these are presented in Table 2.

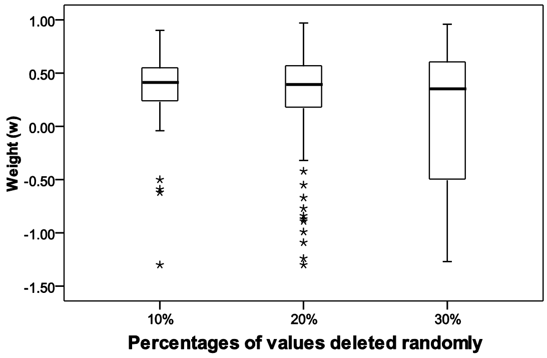

Figure 5. Weight distribution for WGabriel in the Caliński matrix

Table 2. Weight statistics for WGabriel

	Percentages of values deleted randomly		
	10%	20%	30%
Mean	0.3649	0.2734	0.1152
Median	0.4200	0.4000	0.3600
Standard Dev.	0.3103	0.4837	0.6350
Q3-Q1(*)	0.3100	0.3900	1.1000

(*)Q3-Q1=interquartile distance

The suggestion to apply cross-validation in the interval [-2,2] was certainly justified in this data set, because the values of the weights fell inside it. More specifically, from the box plot for 10% imputation we see that the weights lay in the interval [-1.3, 0.9], while for 20% and 30% they were in [-1.3, 0.97] and [-1.27, 0.96] respectively.

4. Conclusions

The WGabriel and MIWG(0.01) imputation procedures proposed here give the best results for the data matrix in the simulation study. These methods minimise both M^2 and NRMSE, and maximise $corr^2$ for all percentages of deleted values values. For situations with a high missing value percentage (>10%) the most favourable method is MIWG(0.01), because with such a distribution-free multiple imputation method it is also possible to obtain a variance estimate among the imputations which quantifies the uncertainty about the real values to be imputed. This possibility does not arise with the other two methods, so for these methods it would be necessary to use extra resampling techniques such as proportional bootstrap in order to obtain variance estimates.

In this study the process used to artificially introduce missing values into the matrix was MCAR. However, that is not to say of course that the methods cannot be used when the missing values in a practical application are MAR or MNAR; for instance, when the data have a clear pattern. The only requirement for application of the methods is that the data set can be arranged in matrix form.

The three methods presented do not make any distributional or structural assumptions, and do not have any restrictions regarding the pattern or mechanism of missing data. However, more extensive evaluation of all the methods will be necessary before definitive conclusions can be reached. A good practical way forward would be to build further simulations around other complete sets of real data.

Acknowledgements

The first author thanks the Coordenação de Aperfeicoamento de Pessoal de Nivel 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., García-Peña M., Dias C.T.S. (2011): Data imputation in trials with genotype×environment interaction. Interciencia 36(6): 444–449.

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): 1–14.

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(4): 422–427.

Caliński T., Czajka S., Kaczmarek Z., Krajewski P., Pilarczyk W. (2009): Analyzing the Genotype-by-Environment Interactions Under a Randomization-Derived Mixed Model. Journal of Agricultural, Biological and Environmental Statistics 14(2): 224–241.

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(3): 331–347.

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.

Di Ciaccio A. (2011): Bootstrap and nonparametric predictors to impute missing data. In: B. Fichet et al. (eds.), Classification and Multivariate Analysis for Complex Data Structures, Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag Berlin Heidelberg.

Dias C.T.S., Krzanowski W.J. (2003): Model selection and cross validation in additive main effect and multiplicative interaction models. Crop Science 43: 865–873.

Gabriel K.R. (2002): Le biplot - outil d'exploration de données multidimensionelles. Journal de la Société Française de Statistique 143(3-4): 5–55.

García-Peña M., Dias C.T.S. (2009): Analysis of bivariate additive models with multiplicative interaction (AMMI). Biometric Brazilian Journal 27(4): 586–602.

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

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

Josse J., Pagès J., Husson F. (2011): Multiple imputation in PCA. Advances in data analysis and classification 5(3): 231–246.

Josse J., Husson F. (2012): Handling missing values in exploratory multivariate data analysis methods. Journal de la Société Française de Statistique 153(2): 79–99.

Krzanowski W.J. (1988): Missing value imputation in multivariate data using the singular value decomposition of a matrix. Biometrical Letters XXV(1-2): 31–39.

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

Kroonenberg P.M. (2008): Applied multiway data analysis. John Wiley & Sons.

Kumar A., Verulkar S.B., Mandal N.P., Variar M., Shukla V.D., Dwivedi J.L., Singh B.N., Singh O.N., Swain P., Mall A.K., Robin S., Chandrababu R., Jain A., Haefele S.M., Piepho H.P., Raman A. (2012): High-yielding, drought-tolerant, stable rice genotypes for the shallow rainfed lowland droughtprone ecosystem. Field Crops Research 133: 37–47.

Little R., Rubin D. (2002): Statistical analysis with missing data. 2nd ed. John Wiley & Sons, New York, NY.

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.

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., Möhring J. (2006): Selection in cultivar trials - Is it ignorable? Crop Science 46: 192–201.

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

Rodrigues P., 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(6): 679–686.

Rubin D.B. (1978): Multiple imputation in sample surveys: a phenomenological Bayesian approach to nonresponse. In: Survey Research Methods Section Of The American Statistical Association. Proceedings: 20–34.

Sabaghnia N., Karimizadeh R., Mohammadi M. (2012): Model selection in additive main effect and multiplicative interaction model in durum wheat. Genetika 44(2): 325–339.

Schafer J.L., Graham J.W. (2002): Missing data: our view of the state of the art. Psychological Methods 7(2): 147–177.

van Buuren S. (2012): Flexible imputation of missing data. CRC press.

Wright K. (2012): agridat: Agricultural datasets. R package version 1.4. http://CRAN.R-project.org/package=agridat>

Yan W., Pageau D., Frégeau-Reid J., Durand J. (2011): Assessing the representativeness and repeatability of test locations for genotype evaluation. Crop Science 51: 1603–1610.

Yan W. (2013): Biplot analysis of incomplete two-way data. Crop Science 53(1): 48–57.