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

COMMUNICATIONS IN BIOMETRY AND CROP SCIENCE
Vol. **11, NO. 2, 2016, PP. 149–163**

INTERNATIONAL JOURNAL OF THE FACULTY OF AGRICULTURE AND BIOLOGY,
Warsaw University of Life Sciences **– SGGW, POLAND**

REGULAR ARTICLE

Multiple imputation procedures using the GabrielEigen algorithm

Marisol García-Peña[1]*, Sergio Arciniegas-Alarcón[2], Wojtek Krzanowski[3], Décio Barbin[4]

[1]Pontificia Universidad Javeriana, Departamento de Matemáticas, Bogotá - Colombia, Carrera 7 43-82.
[2]Fundación Universitaria Konrad Lorenz, Escuela de Negocios, Bogotá - Colombia, Carrera 9 bis 62-43.
[3]University of Exeter, College of Engineering, Mathematics and Physical Sciences, Harrison Building,North Park Road, Exeter, EX4 4QF, UK.
[4]Universidade de São Paulo/ESALQ, Departamento de Ciências Exatas, Cx.P.09, CEP. 13418-900, Piracicaba, SP, Brasil.

*Corresponding author: Marisol García-Peña; E-mail: luzmara@gmail.com

CITATION: García-Peña M., Arciniegas-Alarcón S., Krzanowski W., Barbin D. (2016). Multiple imputation procedures using the GabrielEigen algorithm. *Communications in Biometry and Crop Science* 11, 149–163.

Received: 23 March 2016, Accepted: 7 September 2016, Published online: 23 September 2016

ABSTRACT

GabrielEigen is a simple deterministic imputation system without structural or distributional assumptions, which uses a mixture of regression and lower-rank approximation of a matrix based on its singular value decomposition. We provide multiple imputation alternatives (MI) based on this system, by adding random quantities and generating approximate confidence intervals with different widths to the imputations using cross-validation (CV). These methods are assessed by a simulation study using real data matrices in which values are deleted randomly at different rates, and also in a case where the missing observations have a systematic pattern. The quality of the imputations is evaluated by combining the variance between imputations (V_b) and their mean squared deviations from the deleted values (B) into an overall measure (T_{acc}). It is shown that the best performance occurs when the interval width matches the imputation error associated with GabrielEigen.

Key Words: *imputation; missing values; singular value decomposition; cross-validation; unbalanced.*

INTRODUCTION

Imputation is a technique in which the missing elements of a matrix are replaced by plausible values, thereby making possible a valid analysis of the completed data matrix (observed + imputed). Recently, Arciniegas-Alarcón et al. (2010) proposed an imputation

algorithm without distributional or structural assumptions that uses a mixture of regression and lower-rank approximation of a matrix.

The algorithm was called GabrielEigen and it is deterministic, so has the advantage over stochastic imputation methods (parametric multiple imputation) that the imputed values are uniquely determined, and if the process is repeated on the same data set it will always provide the same results. This characteristic is not necessarily true for stochastic imputation methods (Bello 1993, Arciniegas-Alarcón et al. 2013).

As with any statistical methodology, GabrielEigen has limitations and one of them is that it provides simple imputation, therefore does not take into account the uncertainty produced by the imputations. Thus, if the parameters of a model are estimated from the imputed values, the standard errors will be underestimated, so that confidence intervals and tests may lose validity even if the imputation model is correct (Josse et al. 2011, Josse and Husson 2012a, Arciniegas-Alarcón et al. 2014a).

Multiple imputation (MI, Rubin 1978, 1987) solves this problem. More recent descriptions of the technique can be found in Graham (2012), van Buuren (2012) and Räsler et al. (2013). According to van Ginkel and Kroonenberg (2014), the technique involves four steps: (i) The missing values are estimated *M* times according to a specified statistical model; (ii) These estimates are placed in turn in the data set, resulting in *M* plausible complete versions of the incomplete data set; (iii) Standard statistical procedures are applied to these *M* data sets; (iv) the results are combined to obtain parameter estimates and their variability.

MI solves in a simple way the lack of balance that can affect experiments with genotype-by-environment interaction (G×E), causing difficulties in the application of either additive main effects and multiplicative interaction models (AMMI) or genotype main effects and genotype-by-environment interaction models (GGE, Gauch 2013, Paderewski 2013, Forkman 2015, Yan 2015). Therefore, the aim of this paper is to propose alternatives to the first step of MI using GabrielEigen and to evaluate them by a simulation study based on real matrices from G×E experiments.

MATERIALS AND METHODS

GABRIELEIGEN IMPUTATION ALGORITHM

The algorithm initially replaces the missing cells by arbitrary values and subsequently the imputations are refined through an iterative scheme that defines a different partition of the matrix for each missing value and uses linear regression of columns (or rows) to obtain the new imputation. In this regression, the design matrix is approximated by a low-rank matrix using singular value decomposition (SVD) (Arciniegas-Alarcón et al. 2014b). The algorithm is now presented more formally.

Consider the $n \times p$ matrix $\boldsymbol{X}$ 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 $\boldsymbol{X}$ should first be transposed.

Step 1: The missing values are imputed initially by their respective column means, giving a completed matrix $\boldsymbol{X}$.

Step 2: The columns are standardised by subtracting $\bar{x}_j$ from each element and dividing the result by s_j, where $\bar{x}_j$ and s_j represent respectively the mean and the standard deviation of the *j*-th column.

Step 3: Using the standardised matrix, define the next partition

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

where the missing value in the *(i,j)* position is always in the *(1,1)* position of the defined partition. For each missing value x_{ij}, the components of the considered partition will be

different and that partition is obtained through elementary operations to the rows and columns of matrix **X**. Replace the submatrix $\mathbf{X}_{11}$ by its rank *m* approximation using the 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 $m \leq \min\{n-1, p-1\}$. The imputation of x_{ij} is defined by $\hat{x}_{ij}^{(m)} = \boldsymbol{x}_{i\bullet}^T \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^T \boldsymbol{x}_{\bullet 1}$.

Step 4: The imputation process depends on the value of *m*, and it is suggested that *m* should be chosen as the lowest value for which $\frac{\sum_{k=1}^{m} d_k^2}{\sum_{k=1}^{\min\{n-1, p-1\}} d_k^2} \approx 0.75$.

Step 5: Finally, the imputed values must be returned to the original scale, $x_{ij} = m_j + s_j \hat{x}_{ij}^{(m)}$, replacing them in the matrix **X**. This process is then iterated from **Step 2** until the imputations achieve stability.

MULTIPLE IMPUTATION - MI USING GABRIELEIGEN

It is known that, whatever imputation method is used, there is a risk of underestimating variances and covariances from a completed matrix, since the imputed values do not allow natural sample variation. One way to circumvent this problem is add small random values to each imputation (Krzanowski and Marriott 1994). Using this idea, it is possible to generate an MI scheme from a simple imputation method, just by adding *M* random values to each imputation.

Recently, following the same line, Srivastava and Dolatabadi (2009) proposed an MI scheme by using the simple residuals from a classic linear regression model, assuming the design matrix to be complete and the independent variable to be incomplete. The process consists of fitting a regression model to the observed data and calculating the residuals. *M* random samples, of size equal to the number of missing values, are then obtained with replacement from these residuals. The product of the design matrix of missing values by the vector of parameters from the fitted regression equation produces a vector containing the imputations. Finally, to produce MI, each of the *M* samples of simple residuals is added to the vector of imputations independently. A complete discussion of MI with linear models can be found in Di Ciaccio (2011) and van Buuren (2012).

Taking all the above into account, we came up with two proposals for MI using GabrielEigen. Our first proposal is a two-stage procedure applied to the matrix **X** *(n×p)* with elements x_{ij} *(i=1,...,n; j=1,...,p)* that contain some missing values. In the first stage, the GabrielEigen algorithm is applied to obtain a completed matrix $\mathbf{X}_\mathbf{G}$ *(n×p)* (observed+imputed). In the second stage, random values are added to the imputations, i.e. $\mathbf{X}_\mathbf{G} + (\mathbf{W} \circ \mathbf{E}_t)$, where **W** *(n×p)* is a indicator matrix of zeros and ones, with zero at the *(i,j)* position if that position corresponds to an observed value in **X** and one if the value is missing, "$\circ$" represent the Hadamard product and $\mathbf{E}_t$ *(n×p)* is a matrix of random values with $t = 1, ..., M$.

The options we considered for the matrix $\mathbf{E}_t$ are as follows:

i) Gnorm: $\mathbf{E}_t$ is composed of random values from a $N(0, \hat{\sigma}_j^2)$ distribution, where $\hat{\sigma}_j^2$ is the estimated variance of column *j* of $\mathbf{X}_\mathbf{G}$. This way of building $\mathbf{E}_t$ was inspired by the work of Krzanowski (1988) who used, as initial imputations within an iterative scheme, the mean of the *j*-th column plus a random quantity having zero mean and variance equal to the estimated variance from only the observed values in *j*-th column.

ii) Gadd: $\mathbf{E}_t$ is composed of randomly chosen values with replacement from the set of residuals obtained after fitting an additive model $(x_{ij} = \mu + a_i + b_j + e_{ij})$ to the matrix $\mathbf{X}_\mathbf{G}$.

This way of building $\mathbf{E}_t$ was inspired by the works of Denis and Baril (1992) and Arciniegas-Alarcón et al. (2014b) who discussed the performance of imputed values from an additive model.

iii) GLR: $\mathbf{E}_t$ is composed of values chosen randomly with replacement from the set of residuals obtained from $\boldsymbol{X}_G - \boldsymbol{X}^{(m)}$, where $\boldsymbol{X}^{(m)}$ corresponds to a lower-rank matrix, calculated by the SVD of $\mathbf{X}_G$ using m components. This way of building $\mathbf{E}_t$ was inspired by the work of Arciniegas-Alarcón et al. (2014a) who generalised for MI the simple imputation method based on the SVD for biplot analysis proposed by Yan (2013).

Our second proposal consists of generating approximate confidence intervals for each missing value, calculating through cross-validation the associated imputation error using the GabrielEigen algorithm (Piepho 1995, Arciniegas-Alarcón et al. 2011, 2013). Once the confidence intervals have been obtained, M values within them are chosen randomly to produce MI. A formal statement of the method is as follows.

Consider first the incomplete matrix $\mathbf{X}$ $(n \times p)$. One cell is deleted at a time from the observed elements in the matrix. The deleted value is imputed using GabrielEigen and the difference between the estimated and the actual value for the relevant cell is recorded. This is done for all the observed values and the average of the squared differences is denoted by D. D contains two components of variability, one due to predictive inaccuracy of the imputation and the other due to sampling error of the observed values. For this reason, D should be corrected by subtracting an estimate of the error of the mean (s^2). The square root of ($D - s^2$) may be taken as the imputation error (I_e) associated with GabrielEigen.

So, if the imputation in the (i,j) position with GabrielEigen is denoted by $\hat{x}_{ij}$, an approximate imputation interval is given by $\hat{x}_{ij} \pm z_{1-\alpha} I_e$, where $z_{1-\alpha}$ is the appropriate point of the normal distribution for a confidence level of $(1-\alpha)\%$. In order to produce MI, M values are chosen randomly within the interval.

For a 95% interval $z_{1-\alpha} = 1.96$, and the width of the interval is approximately $4I_e$. We also wished to study the effect of decreasing the interval width to I_e and $2I_e$, or equivalently with $z_{1-\alpha} = 0.5$ and $z_{1-\alpha} = 1$ representing 38% and 68% intervals respectively. The decrease in width reduces variability in the imputations, but may increase the risk of low quality in the multiple imputations. The methodology will be denoted GCV1, GCV2 and GCV4 depending on the interval width producing MI.

MI with GabrielEigen also requires specification of the number (m) of components to be retained in the SVD, and the number (M) of completed versions of the matrix $\mathbf{X}$. Cross-validation, rather than the criterion described in Step 3 of the original algorithm, was used to determine m by applying the process explained in Garcia-Peña et al. (2014) based on the *cv.SVDImpute* function from the *imputation* package of software R (Wong 2013, R Core Team 2015). Typically, a small number of imputations ($3 \leq M \leq 10$) is necessary to obtain a good performance of MI (Ounpraseuth et al. 2012), so we decided to fix $M=5$ since this number achieves high statistical efficiency in many practical applications (van Buuren 2012).

Note that the proposals Gnorm, Gadd, GLR, GCV1, GCV2 and GCV4 are, by construction, computationally less intensive than the MI proposal for GabrielEigen by Arciniegas-Alarcón et al. (2014c), which essentially consists of inserting multiplicative weights in the imputation equation. The selection of these weights requires double cross-validation, so we did not include the method in our simulations on grounds of computational cost and efficiency.

SIMULATION STUDY

In order to develop a realistic simulation study we followed the protocol proposed by Yan (2013) to assess new imputation methods for (G×E) matrices. The steps are:

i) Choose real balanced data sets, or extract balanced subsets from incomplete G×E experiments.
ii) In each set delete values randomly at different percentages.
iii) Repeat the process for each percentage a large number of times (e.g. 1000).
iv) Calculate in each repetition of the process a statistic to compare the imputations with the real values deleted.

Three data sets were used in our simulation study. The first data set (Lavoranti et al. 2007, Wright 2012), is a 20×7 matrix giving the mean tree heights (m) of 20 *Eucalyptus grandis* progenies assessed in 7 locations in the south and southeast regions of Brazil. The second data set (Yang 2007), is a 6×18 matrix giving the yield (Mg há$^{-1}$) of 6 barley genotypes assessed in 18 environments in Alberta, Canada. The third data set (Rad et al. 2013) is a 36×6 matrix giving the mean plant grain yield (gr) of 36 wheat genotypes assessed in 6 environments under normal and drought stress conditions, in Experiments Farm of University Putra, Malaysia.

The choice of data sets was based on a previous study that determined the number of multiplicative components necessary to explain the G×E interaction through an AMMI model (Gauch 1992, 2013). In each set the generalised cross-validation method proposed by Josse and Husson (2012b) and available in the *FactoMineR* package of the software R (Husson et al. 2014, R Core Team 2015) was applied. Table 1 presents the mean square error of prediction (MSEP) to choose the multiplicative components of the model. The best model is the one with lowest MSEP, so the best models for the eucalyptus data, for the barley data, and for the wheat data are AMMI1, AMMI2 and AMMI3 respectively.

Table 1. Values of Mean Square Error of Prediction (MSEP) using generalised cross-validation in choosing the AMMI model to explain the interaction in the original (complete) data matrices.

	MSEP		
Model	Eucalyptus	Barley	Wheat
AMMI1	**0.5744**	0.0502	5.08E-01
AMMI2	0.5834	**0.0463**	1.21E-01
AMMI3	0.6964	0.0584	**6.85E-06**
AMMI4	0.8123	0.0853	9.03E-06
AMMI5	1.1937	0.1565	1.50E-05
AMMI6	1.8987		

The three data sets have different sizes and interaction structures, and are broadly representative of G×E experiments. For this reason, the conclusions that are derived from them should also be relevant to most other matrices of multienvironmental data.

In each data set we deleted randomly 10%, 20% and 35% of values; the process was repeated 1000 times, giving a total of 9000 different incomplete data sets; for each one, the 6 MI proposed methodologies were applied using code in the software R (R Core Team 2015).

The chosen percentages and the value deletion mechanism have been fully justified in the literature. In (G×E) practice, generally the number of missing values is lower than 40% (Yan 2013) but anything under 10% would not benefit much from MI because simple imputation can provide fairly good results (Schafer 1999). Moreover, the value deletion mechanism represents common situations in agricultural experiments as, for example, the plants can be destroyed by animals, floods or during the harvest, and the yield measurements may be erroneously performed and introduced in the data base (Rodrigues et al. 2011). A discussion about the different mechanisms that can be simulated in (G×E) can be found in Paderewski and Rodrigues (2014) and Arciniegas-Alarcón et al. (2014c).

In Yan's protocol, the chosen statistic to compare the imputations with the deleted values was the prediction error (P_e), defined as:

$$P_e = \left[\frac{1}{NM} \sum_{i,j} \left(MV_{ij} - PV_{ij} \right)^2 \right]^{\frac{1}{2}}$$

where MV is the true value, PV is the predicted value and NM is the total number of missing values. P_e is very useful in assessing simple imputation methods, but to assess the accuracy of MI strategies it is preferable to use the statistics T_{acc}, V_b and B introduced by Penny and Jolliffe (1999) and recently used by Bergamo et al. (2008) and Arciniegas-Alarcón et al. (2014a).

T_{acc} is a measure of overall accuracy formed from the sum of the pooled variance between imputations within positions (V_b) and the mean squared deviation between the mean of the imputations and the corresponding original value deleted in the simulation study (B). These statistics are given by:

$$T_{acc} = V_b + B \text{, where}$$

$$V_b = \frac{1}{na} \sum_{l=1}^{na} \left[\frac{\sum_{q=1}^{M} \left(\hat{x}_{ij(q)} - \overline{X}_l \right)^2}{M-1} \right] \text{ and } B = \frac{1}{na} \sum_{l=1}^{na} M \frac{\left(\overline{X}_l - VO_l \right)^2}{M-1}.$$

Here "na" is the total number of deleted values from the G×E matrix, and deleted value l has position (i, j) in the matrix, i.e. in the i-th row and the j-th column. M is the number of imputations for the missing value l, $\hat{x}_{ij(q)}$ is the q-th imputation for that value according to the proposed methods, $\overline{X}_l$ is the mean of the imputations produced for the missing value l and VO_l is the original value l in the complete original data set.

In this study, all three statistics will be analysed, but the final decision to choose the best MI system will be based on T_{acc}. If V_b is too large, then the method may not be very reliable, but a small value for this variance does not necessarily mean that the imputation method is good, because the imputation may be biased. A good imputation method will be one with small B, as otherwise the imputations differ substantially from the observed data set. Ideally, an imputation method is required with small values for both V_b and B, which together imply a low value for T_{acc} (Penny and Jolliffe 1999).

RESULTS

EUCALYPTUS DATA

Table 2 shows the mean and median of V_b, B and T_{acc} for the different percentage of values deleted randomly (10, 20 and 35%) for the eucalyptus data set. The smallest variance at all percentages was always obtained with GCV1, which was as expected within the schemes that involved the calculation of approximate confidence intervals for the imputations. GCV1 was also better than all methodologies which added random error to the imputation initially produced by GabrielEigen, namely Gnorm, Gadd and GLR. On the other hand, the algorithm that maximised V_b in all the cases was Gnorm. It is worth noting the performance of GLR, because it provided, at all percentages, smaller variances than GCV4, i.e. using 95% confidence intervals.

In the same data set, the lowest bias (B) was obtained with GCV1 at all the percentages. This is an interesting result, because it was expected that decreasing the width of imputation intervals would increase the value of the B statistic. Thus simple imputations with GabrielEigen are of high quality, as incorporating variability to produce MI does not require

a very large interval width. The algorithms with more biased imputations, maximizing B, were Gnorm and GCV4. The GCV2 algorithm was less biased than GLR and Gadd at low rates of missing values (10 and 20%), but when imputing 35% of the data, the situation changed and the GLR algorithm had a lower bias than Gadd and GCV2 respectively (Table 2).

Table 2. Means and medians of pooled variance between imputations (V_b), mean square deviation (B) and measure of overall accuracy (T_{acc}) at different percentages of values deleted randomly in 1000 simulations from eucalyptus data set.

Method	10%		20%		35%	
	Mean	Median	Mean	Median	Mean	Median
			V_b			
Gnorm	1.2208	1.1908	1.1318	1.1162	0.9752	0.9668
Gadd	0.4107	0.4033	0.3665	0.3604	0.3025	0.3003
GLR	0.3966	0.3895	0.3543	0.3488	0.2929	0.2905
GCV4	0.8606	0.8627	0.8831	0.8808	0.9665	0.9647
GCV2	0.2241	0.2246	0.2299	0.2293	0.2516	0.2511
GCV1	0.0560	0.0561	0.0575	0.0573	0.0629	0.0628
			B			
Gnorm	1.2700	1.1868	1.2800	1.2464	1.3343	1.3169
Gadd	1.0640	1.0203	1.0908	1.0607	1.1612	1.1391
GLR	1.0591	1.0184	1.0892	1.0656	1.1594	1.1404
GCV4	1.1779	1.1115	1.2326	1.2068	1.3493	1.3172
GCV2	1.0240	0.9577	1.0665	1.0378	1.1732	1.1468
GCV1	0.9843	0.9237	1.0217	1.0022	1.1272	1.1080
			T_{acc}			
Gnorm	2.4908	2.3993	2.4119	2.3917	2.3095	2.2950
Gadd	1.4747	1.4331	1.4573	1.4236	1.4637	1.4370
GLR	1.4556	1.4081	1.4435	1.4161	1.4523	1.4267
GCV4	2.0385	1.9977	2.1157	2.0921	2.3158	2.2871
GCV2	1.2480	1.1843	1.2964	1.2787	1.4248	1.3942
GCV1	1.0403	0.9822	1.0792	1.0585	1.1901	1.1669

Finally, turning to the statistic T_{acc} (Table 2 and Figure 1), the best method in all the cases was clearly GCV1, followed by GCV2. At all percentages, the algorithms with lowest performance (maximizing T_{acc}) were Gnorm and GCV4, while the Gadd and GLR methods were poorer than GCV1 or GCV2 but better than Gnorm and GCV4.

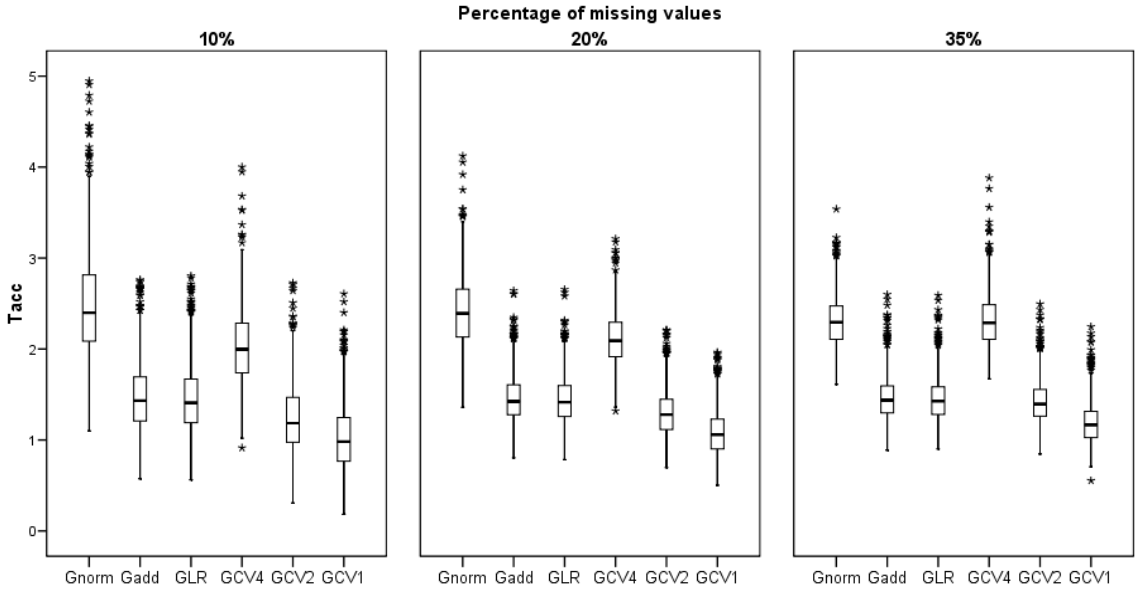

Figure 1. Box plot of the measure of overall accuracy *Tacc* distribution for the six algorithms in eucalyptus data set.

BARLEY DATA

Table 3 shows the mean and median of V_b, B and T_{acc} for the different percentages of values deleted randomly (10, 20 and 35%) for the barley matrix. In the same way as for the eucalyptus data simulations, V_b was always minimised by GCV1 and maximised with Gnorm and GCV4. The main difference from the eucalyptus data was shown by the GLR and GCV2 methods at 35% imputation, where GLR had an average variance equal to 0.0824 while with GCV2 the value was 0.0975. This means that in terms of variation between imputations, GLR was better at the higher missing percentages than the algorithms using intervals of width $2I_e$ (68% confidence) and $4I_e$ (95% confidence).

As regards similarity with the original data, the least biased method was again GCV1, while the most biased imputations were produced by Gnorm and GCV4. It is worth noting that in the case of the B statistic at 35% imputation, Gadd and GLR had better performance than GCV2 and GCV4.

To make a definite decision about the MI algorithms we used the measure T_{acc}. The distributions shown in Figure 2, clearly identify the poorest methods, Gnorm and GCV4, but the box plot does not show efficiently the differences between the others. Therefore, to choose the best method, we used the means and medians of the distributions (Table 3).

These statistics establish that T_{acc} was minimised at all the imputation percentages by GCV1 but GLR, Gadd and GCV2 give different results depending on the imputation percentage: for 10 and 20%, GCV2 is better than GLR and Gadd, but when the imputation increases to 35%, the opposite occurs.

Table 3. Means and medians of pooled variance between imputations (V_b), mean square deviation (B) and measure of overall accuracy (T_{acc}) at different percentages of values deleted randomly in 1000 simulations from barley data set.

Method	10%		20%		35%	
	Mean	Median	Mean	Median	Mean	Median
			V_b			
Gnorm	4.1148	4.0170	4.0525	4.0235	3.9476	3.9362
Gadd	0.1287	0.1248	0.1166	0.1140	0.1036	0.1005
GLR	0.1094	0.1072	0.0966	0.0944	0.0824	0.0799
GCV4	0.2528	0.2525	0.2717	0.2681	0.3747	0.3444
GCV2	0.0658	0.0657	0.0707	0.0698	0.0975	0.0897
GCV1	0.0165	0.0164	0.0177	0.0174	0.0244	0.0224
			B			
Gnorm	1.3578	1.2488	1.3401	1.2910	1.4156	1.3751
Gadd	0.3262	0.2952	0.3420	0.3304	0.4435	0.3906
GLR	0.3216	0.2909	0.3366	0.3240	0.4379	0.3849
GCV4	0.3585	0.3326	0.3838	0.3696	0.5241	0.4657
GCV2	0.3109	0.2881	0.3331	0.3220	0.4527	0.3947
GCV1	0.2982	0.2740	0.3195	0.3098	0.4332	0.3753
			T_{acc}			
Gnorm	5.4725	5.3720	5.3926	5.3543	5.3632	5.3310
Gadd	0.4550	0.4246	0.4586	0.4474	0.5471	0.4857
GLR	0.4310	0.3996	0.4332	0.4202	0.5204	0.4632
GCV4	0.6113	0.5960	0.6555	0.6458	0.8988	0.8037
GCV2	0.3767	0.3562	0.4038	0.3940	0.5503	0.4835
GCV1	0.3147	0.2926	0.3372	0.3276	0.4576	0.3988

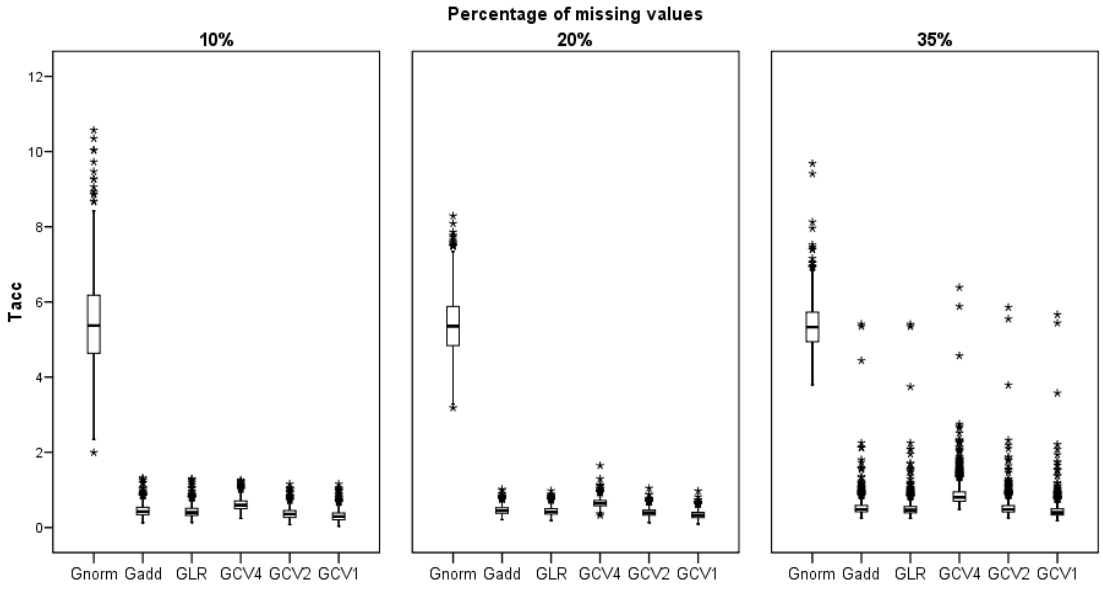

Figure 2. Box plot of the measure of overall accuracy *Tacc* distribution for the six algorithms in barley data set.

WHEAT DATA

Table 4 shows the mean and median of V_b, B and T_{acc} for the different percentages of values deleted randomly (10, 20 and 35%) for the wheat data. The mean and median of variances between imputations was maximised in all the cases by Gnorm and Gadd, and minimised by GCV1. Of the remaining methods we can highlight GLR, because although it always had average variances higher than GCV1, it had smaller values of V_b when compared with GCV4 and GCV2 at 20 and 35% imputation.

The greatest similarity between imputations and the artificially deleted data (B) was again obtained with GCV1, while the most biased imputations were produced by Gnorm and Gadd. At 10 and 20% imputation GCV2 was less biased than GLR and GCV4, but at 35% GLR had a smaller bias than GCV2 and GCV4 (Table 4).

Table 4. Means and medians of pooled variance between imputations (V_b), mean square deviation (B) and measure of overall accuracy (T_{acc}) at different percentages of values deleted randomly in 1000 simulations from wheat data set.

Method	10%		20%		35%	
	Mean	Median	Mean	Median	Mean	Median
			V_b			
Gnorm	3.4515	3.3803	3.3063	3.2894	2.9553	2.9420
Gadd	1.1222	1.1121	1.0705	1.0711	0.8622	0.8418
GLR	0.1569	0.1538	0.1775	0.1371	0.3830	0.4643
GCV4	0.5291	0.5262	0.7875	0.6495	1.9316	2.1086
GCV2	0.1396	0.1382	0.2092	0.1725	0.5044	0.5477
GCV1	0.0349	0.0345	0.0523	0.0431	0.1261	0.1369
			B			
Gnorm	1.4275	1.3556	1.6873	1.5333	2.6444	2.7064
Gadd	0.8454	0.8056	1.1299	0.9427	2.1137	2.2425
GLR	0.6149	0.5760	0.9051	0.7087	1.9978	2.1813
GCV4	0.6882	0.6658	0.9964	0.8255	2.3452	2.4887
GCV2	0.6015	0.5644	0.8773	0.7015	2.0075	2.1847
GCV1	0.5756	0.5361	0.8375	0.6676	1.9188	2.0821
			T_{acc}			
Gnorm	4.8789	4.8259	4.9936	4.9257	5.5997	5.5426
Gadd	1.9676	1.9389	2.2004	2.0546	2.9759	3.0199
GLR	0.7719	0.7333	1.0827	0.8406	2.3808	2.7326
GCV4	1.2173	1.1888	1.7839	1.4787	4.2767	4.7226
GCV2	0.7411	0.7068	1.0865	0.8880	2.5119	2.7580
GCV1	0.6105	0.5681	0.8898	0.7189	2.0449	2.2325

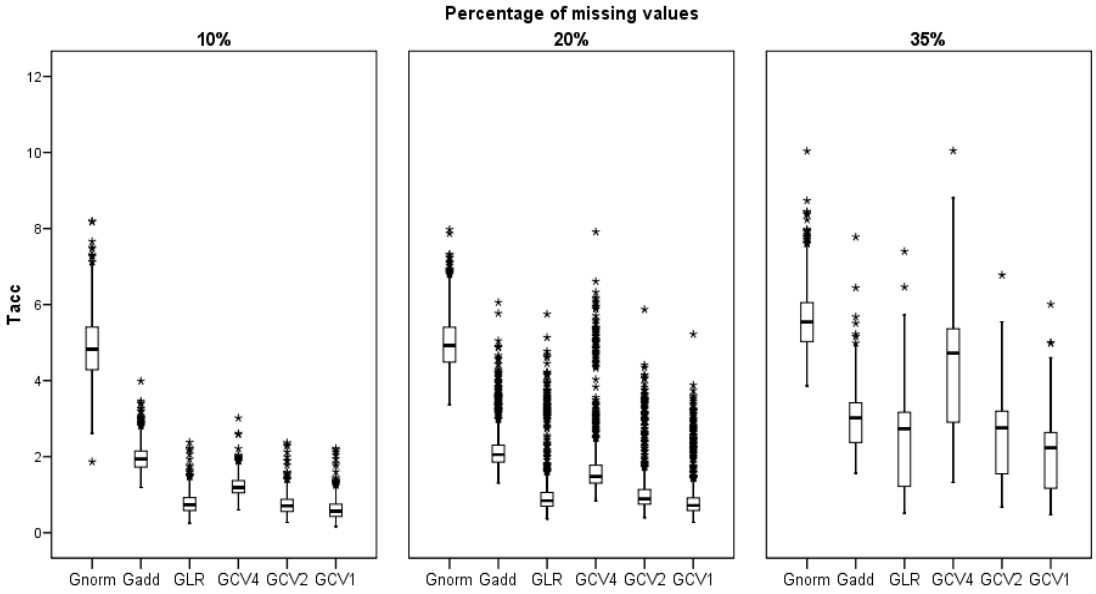

Figure 3. Box plot of the measure of overall accuracy *Tacc* distribution for the six algorithms in wheat data set.

As before we present summaries for T_{acc} in Figure 3 and Table 4. Based on the box plot it is possible to place the algorithms into low and high performance groups. The high performance group contains GCV1, GCV2 and GLR and in the low performance group are GCV4, Gadd and Gnorm. In the high performance group, GCV1 always had the best results minimising T_{acc}. On the other hand, although never better than GCV1, the performances of GCV2 and GLR depended on the imputation percentage. For instance, GCV2 had lower T_{acc} values than GLR with 10% randomly deleted values, but GLR had better performance than GCV2 at 20% and 35% imputation.

A DIFFERENT SITUATION: MISSING VALUES NOT AT RANDOM

The MI systems proposed in this study do not depend on any structural assumptions but Bello (1993) warns that this lack of assumptions does not imply robustness and in some cases may produce unexpected results. The different structures can be caused by different mechanisms of missing data (Little and Rubin 2002, Paderewski and Rodrigues 2014) and in G×E experiments it is possible to find situations with missing values not at random (MNAR). For example, incomplete matrices with systematic patterns can arise because over the years new reference genotypes are included and some others are disregarded (Denis and Baril 1992).

To assess how missing values not at random affect the MI methodologies here proposed, we considered again the complete matrices of eucalyptus, barley and wheat, but differently from the previous simulation study as missing values were created systematically once only.

For each of the matrices, a third of the genotypes were deleted in each environment. The genotypes belonging to the third with least height were deleted in the eucalyptus data and the third with lowest yield were deleted in the barley and wheat data sets (personal communication of W. Yan 2014). Therefore, in the eucalyptus data we had 49 missing values (35%), resulting in a total loss of a genotype, while in the barley matrix there were 36 missing values (33.33%) and 72 missing values in the wheat matrix (33.33%). In the case of the wheat matrix, the arbitrary deletion resulted in the total loss of five genotypes. Once we had the incomplete matrices, the MI methods were applied and the corresponding statistics V_b, B and T_{acc} were calculated. The results are shown in Table 5.

Table 5. V_b, B and T_{acc} statistics after one deletion not at random in the eucalyptus, barley and wheat matrices.

	Eucalyptus		
Method	V_b	B	T_{acc}
Gnorm	0.3141	4.7464	5.0606
Gadd	0.1647	4.7218	4.8864
GLR	0.1628	4.7136	4.8764
GCV4	0.4269	4.7016	5.1284
GCV2	0.1172	4.6887	4.8059
GCV1	0.0271	4.7077	4.7348
		Barley	
Method	V_b	B	T_{acc}
Gnorm	3.9782	1.5779	5.5561
Gadd	0.0879	0.4797	0.5676
GLR	0.0862	0.4302	0.5163
GCV4	0.2576	0.5354	0.7930
GCV2	0.0806	0.4437	0.5244
GCV1	0.0184	0.4454	0.4638
		Wheat	
Method	V_b	B	T_{acc}
Gnorm	0.6643	14.2154	14.8797
Gadd	0.5055	13.7940	14.2996
GLR	0.0021	14.0732	14.0753
GCV4	0.3224	13.9020	14.2244
GCV2	0.0845	13.9838	14.0683
GCV1	0.0187	14.0694	14.0881

In the eucalyptus data set, the least biased imputations (B) were produced by GCV2, while the most biased imputation system was Gnorm. However, the smallest variance between imputations (V_b) was obtained with GCV1 and the biggest with Gnorm. The measure of overall accuracy T_{acc} indicates that the best method was GCV1 and to explain this result note that although GCV1 did not have the best performance in terms of similarity with the original values deleted, it offset this situation by having high accuracy (minimising V_b).

In the barley data, again the best MI method according to T_{acc} was GCV1. As happened in the eucalyptus data set, GCV1 had a poorer performance than GLR and GCV2 in terms of imputation bias, but none of them outperform it in terms of the variability of mean imputed values. The method with poorest performance was again Gnorm (Table 5).

Up to this point, the results with missing values not at random did not differ much from those obtained in the simulation study, but the wheat data set proved to be the exception. In this case GCV1, which previously was best because it minimised the imputation bias and/or minimised the variance V_b, was outperformed by GCV2 and GLR respectively, using T_{acc} as evaluation criterion. Here, GLR minimised V_b, and all the systems except Gnorm and GLR had greater similarity with the original values (B) than GCV1 (Table 5).

DISCUSSION AND CONCLUSIONS

We have fulfilled our objective of producing multiple imputation systems using the GabrielEigen algorithm, and have shown GCV1 to have the best performance on (G×E) matrices that had simple (eucalyptus), moderate (barley) and complex (wheat) interaction structure.

The second best is either GCV2 or GLR, but their ordering depends strongly on the interaction structure and the imputation percentage. GLR has the better performance when the missing values percentage is high (~35%) and the interaction structure is moderate or complex, but in other cases it may be preferable to use GCV2.

In the case of systematic occurrence of missing values, our results showed that GCV1 was again the best for the simple and moderate interaction matrices, but GLR and GCV2 performed better when used on a complex interaction matrix. However, this is an area for more detailed future studies and practical recommendations. The systematic deletion was carried out only once in the wheat matrix, but to confirm the robustness of GCV1 to missing values not at random in complex interaction structures would require additional research, with simulations that can use the procedure proposed here to generate MNAR in G×E experiments. Our recommendation in this case is to use all three algorithms GCV1, GCV2 and GLR and to assess them on any particular data set using the statistics here presented.

The simulation study presented in this work was done from a complete data set, but the applied researcher may wish to conduct a similar study on an incomplete matrix. In order to do this, we suggest the methodology described by Arciniegas-Alarcón et al. (2016): delete randomly some of the entries of the matrix (for example, between 10% and 30%), apply the imputation algorithms, and calculate the statistic of comparison (for instance, Tacc). Repeat the deletion process (for example, 100 times), calculate the mean or median of all the values, and the method with the lowest mean or median is the recommended method.

The strategy of constructing confidence intervals for the imputations using cross-validation was successful, but a point to highlight is that the 95% confidence intervals traditionally used in statistics have not provided the best results. Finally, the computational aspect in this study was not a problem, but if the analysed matrices are larger then "k-fold" cross-validation could be considered instead of "leave-one-out" as described in James et al. (2013).

REFERENCES

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*, Article ID 978780, 17 pages.

Arciniegas-Alarcón S., Dias C.T.S., García-Peña M. (2014a). Distribution-free multiple imputation in incomplete two-way tables. *Pesquisa Agropecuária Brasileira* 49, 683–691.

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

Arciniegas-Alarcón S., García-Peña M., Krzanowski W.J., Dias C.T.S. (2014c). An alternative methodology for imputing missing data in trials with genotype-by-environment interaction: some new aspects. *Biometrical Letters* 51, 75–88.

Arciniegas-Alarcón S., García-Peña M., Krzanowski W.J. (2016). Missing value imputation in multi-environment trials: Reconsidering the Krzanowski method. *Crop Breeding and Applied Biotechnology* 16, 77–85.

Bello A.L. (1993). Choosing among imputation techniques for incomplete multivariate data: a simulation study. *Communications in Statistics – Theory and Methods* 22, 853–877.

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.

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: Fichet B. et al. (Eds.). *Classification and Multivariate Analysis for Complex Data Structures, Studies in Classification, Data Analysis, and Knowledge Organization*. Springer-Verlag Berlin Heidelberg.

Forkman J. (2015). A resampling test for principal component analysis of genotype-by-environment interaction. *Acta et Commentationes Universitatis Tartuensis de Mathematica* 19, 27–33.

García-Peña M., Arciniegas-Alarcón S., Barbin D. (2014). Climate data imputation using the singular value decomposition: an empirical comparison. *Revista Brasileira de Meteorologia* 29, 527–536.

Gauch H.G. (1992). *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*. Elsevier, Amsterdam.

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

Graham JW (2012). *Missing data. Analysis and Design*. Springer.

Husson F., Josse J., Le S., Mazet J. (2014). *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R*. R package version 1.26. http://CRAN.R-project.org/package=FactoMineR.

James G., Witten D., Hastie T., Tibshirani R. (2013). *An introduction to statistical learning with applications in R*. Springer.

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

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

Josse J., Husson F. (2012b) Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics and Data Analysis* 56, 1869–1879.

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., Marriott F.H.C. (1994). *Multivariate analysis, Part 1: Distributions, Ordination, and Inference*. Edward Arnold.

Lavoranti O.J., Dias C.T.S., Krzanowski W.J. (2007). Phenotypic stability and adaptability via AMMI model with bootstrap re-sampling. *Pesquisa Florestal Brasileira* 54, 45–52.

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

Ounpraseuth S., Moore P.C., Young D.M. (2012). Imputation techniques for incomplete data in quadratic discriminant analysis. *Journal of Statistical Computation and Simulation* 82, 863–877.

Paderewski J. (2013). An R function for imputation of missing cells in two-way data sets by EM-AMMI algorithm. *Communications in Biometry and Crop Science* 8, 60–69.

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.

Penny K.I., Jolliffe I.T. (1999). Multivariate outlier detection applied to multiply imputed laboratory data. *Statistics in Medicine* 18, 1879–1895.

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.

R Core Team (2015). *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/.

Rad M.R.N., Kadir M.A., Rafii M.Y., Jaafar H.Z.E., Naghavi M.R., Ahmadi F. (2013). Genotype × environment interaction by AMMI and GGE biplot analysis in three consecutive generations of wheat (*Triticum aestivum*) under normal and drought stress conditions. *Australian Journal of Crop Science* 7, 956–961.

Rässler S., Rubin D.B., Zell E.R. (2013). Imputation. *WIREs Computational Statistics* 5, 20–29.

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, 679–686.

Rubin D.B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, Alexandria, 20–34.

Rubin D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Schafer J.L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research* 8, 3–15.

Srivastava M.S., Dolatabadi M. (2009). Multiple imputation and other resampling scheme for imputing missing observations. *Journal of Multivariate Analysis* 100, 1919–1937.

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

van Ginkel J.R., Kroonenberg P.M. (2014). Using generalized procrustes analysis for multiple imputation in principal component analysis. *Journal of Classification* 31, 242–269.

Wong J. (2013). *Imputation. R package version 2.0.1*. Available at: https://github.com/jeffwong/imputation. Accessed on: 21 Aug. 2015.

Wright K. (2012). *agridat: Agricultural datasets. R package version 1.4*. Available at: http://CRAN.R-project.org/package=agridat. Accessed on: 21 Aug. 2015

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

Yan W. (2015). Mega-environment analysis and test location evaluation based on unbalanced multiyear data. *Crop Science* 55, 113–122.

Yang R.C. (2007). Mixed-model analysis of crossover genotype–environment interactions. *Crop Science* 47, 1051–1062.