DISTRIBUTION-FREE MULTIPLE IMPUTATION IN AN INTERACTION MATRIX THROUGH SINGULAR VALUE DECOMPOSITION

Genevile Carife Bergamo[1,2]; Carlos Tadeu dos Santos Dias[3]*; Wojtek Janusz Krzanowski[4]

[1] *UNIFENAS - C.P. 23 - 37130-000 - Alfenas, MG - Brasil.*
[2] *USP/ESALQ - Programa de Pós-Graduação em Estatística e Experimentação Agronômica.*
[3] *USP/ESALQ - Depto. de Ciências Exatas, C.P. 09 - 13418-900 - Piracicaba, SP- Brasil.*
[4] *School of Engineering, Computer Science & Mathematics - University of Exeter -North Park Road - Exeter, EX4 4QE - UK.*
**Corresponding author <ctsdias@esalq.usp.br>*

ABSTRACT: Some techniques of multivariate statistical analysis can only be conducted on a complete data matrix, but the process of data collection often misses some elements. Imputation is a technique by which the missing elements are replaced by plausible values, so that a valid analysis can be performed on the completed data set. A multiple imputation method is proposed based on a modification to the singular value decomposition (SVD) method for single imputation, developed by Krzanowski. The method was evaluated on a genotype × environment (G × E) interaction matrix obtained from a randomized blocks experiment on *Eucalyptus grandis* grown in multienvironments. Values of *E. grandis* heights in the G × E complete interaction matrix were deleted randomly at three different rates (5%, 10%, 30%) and were then imputed by the proposed methodology. The results were assessed by means of a general measure of performance (T_{acc}), and showed a small bias when compared to the original data. However, bias values were greater than the variability of imputations relative to their mean, indicating a smaller accuracy of the proposed method in relation to its precision. The proposed methodology uses the maximum amount of available information, does not have any restrictions regarding the pattern or mechanism of the missing values, and is free of assumptions on the data distribution or structure.
Key words: missing data, nonparametric, eigenvalue, eigenvector, genotype-environment

IMPUTAÇÃO MÚLTIPLA LIVRE DE DISTRIBUIÇÃO EM MATRIZ DE INTERAÇÃO POR MEIO DE DECOMPOSIÇÃO POR VALOR SINGULAR

RESUMO: Algumas técnicas de análise estatística multivariada só podem ser realizadas com uma matriz de dados completa, porém o processo de coleta dos dados freqüentemente leva a uma matriz com dados ausentes. A imputação é uma técnica, na qual os dados ausentes são preenchidos com valores plausíveis, para uma posterior análise na matriz completa. Neste trabalho, nós propomos um método de imputação múltipla, baseado no método da decomposição por valores singulares (DVS) para imputação simples, desenvolvido por Krzanowski, e avaliado numa matriz de interação genótipos × ambientes (G × E), proveniente de um ensaio com o delineamento aleatorizado em blocos em multiambientes com genótipos de *Eucalyptus grandis*. Valores da altura de *E. grandis* da matriz completa de interação G × E foram retirados aleatoriamente em três diferentes proporções (5%, 10%, 30%), os quais foram imputados valores dados pelo método proposto. Os resultados obtidos por meio da medida geral de exatidão ou acurácia (T_{acc}) mostraram um viés pequeno, em relação aos valores originais. No entanto, seus valores foram maiores do que a variabilidade dos valores imputados em relação à sua média, indicando uma exatidão ou acurácia menor do método proposto em relação à sua precisão. A metodologia proposta utiliza o maior número de informação disponível, não possui qualquer restrição quanto ao padrão e mecanismo de ausência de dados e é livre de suposição sobre a distribuição ou estrutura dos dados.
Palavras-chave: dados ausentes, não-paramétrico, autovalor, autovetor, genótipo-ambiente

INTRODUCTION

Imputation is a technique in which missing elements of a data matrix are replaced by plausible values, so that a valid analysis can be made on the complete data set (observed + imputed). Various imputation methods have been proposed over the years, but the current interest is now focussed mostly on multiple imputation (MI). MI was first proposed by Rubin (1978), and several other references including Little & Rubin (1987, 2002); Rubin (1987); Rubin & Schenker (1986); Schafer (1997, 1999); Tanner & Wong (1987) and Zhang (2003) provide excellent descriptions of the technique. The basic idea of the procedure is to replace each missing value by a set of *M* imputed values that are drawn from the data distribution, with the variation in these values representing the uncertainty about the true value to be imputed.

The MI procedure involves three distinct steps (i) **Imputation**: The missing values are estimated *M* times, generating *M* completed data sets; (ii) **Analysis**: The *M* completed data sets are analyzed, using appropriate statistical procedures for the problem at hand, and (iii) **Combination**: The *M* separate sets of results are combined into one single inference.

The imputation is the most critical step, and the model used in this step is not necessarily the same as the one used in the analysis step. This makes the MI procedure more attractive, as the model used to impute data is not always the best one suited for the analysis.

On combining the results of the *M* analyses, the combined estimative variance consists of the variance within the imputations and, therefore, the uncertainty in the imputed data is incorporated into the final inference.

A method is here proposed for the first step of the multiple imputation, without making any assumptions about the data distribution or structure, by using the singular value decomposition (SVD) of a genotype × environment (G × E) interaction matrix for those models whose analysis needs a complete matrix. The performance of the method is then investigated on randomly deleted entries in the G × E interaction matrix obtained from an experiment on *Eucalyptus grandis* progenies.

MATERIAL AND METHODS

The data used in this study were obtained from experiments conducted in seven environments, at the south and southeast regions of Brazil, on 20 *Eucalyptus grandis* progenies obtained from Australia (km 12, South of Ravenshoe-Mt Pandanus-QLD, 43°41'10" W and 22°45'30" S at 33 m above sea level, lot 14,420). A randomized block design with six plants per plot and ten replicates was used, the whole experiment taking up a space of dimension 3.0 m by 2.0 m (Lavoranti, 2003).

The value y_{ij} represents the mean tree height (m) over 10 blocks of the plot means for the i^{th} genotype (i = 1,2,...,20) in the j^{th} environment (j = 1,2,...7) of *Eucalyptus grandis*. The resulting 20 × 7 matrix of y_{ij} values constitutes the G × E interaction matrix. Values from this matrix were then deleted randomly to form three different incomplete matrices differing in the proportions of deleted values. The first matrix had 5% deleted, the second 10%, and the third 30%. The same random number seed was used in generating each set of deletions (i^{th} row; j^{th} column), so that the deleted entries of the second matrix: (6;3), (1;4), (3;4), (6;4), (20;4), (6;5), (7;7) included those deleted in the first: (2;2), (19;2), (10;4), (13;4), (19;4), (5;6), (4;7), and the deleted entries of the third matrix: (3;1), (9;1), (13;1), (15;1), (18;1), (5;2), (11;2), (3;3), (7;3), (17;3), (19;3), (9;4), (12;4), (15;4), (17;4), (8;5), (12;5), (16;5), (2;6), (15;6), (16;6), (17;6), (19;6), (6;7), (8;7), (11;7), (12;7), (13;7), included those deleted in each of the other two. The SAS statistical system, by means of the SAS (2004a) and SAS (2004b), was used in the development of the programs and data analyses described below.

For a distribution free approach, the imputed values were obtained by means of a modification to the simple imputation system developed by Krzanowski (1988). This method starts from the observation made by Good (1969), that any matrix $\boldsymbol{Y}_{(n,p)}$ can be decomposed by the singular value decomposition into the form

$$\boldsymbol{Y} = \boldsymbol{UDV}^T, \qquad (1)$$

in which $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}_p$, $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}_n$ and $\boldsymbol{D} = diag(d_1, ..., d_p)$ with $d_1 \geq d_2 \geq, ..., \geq d_p \geq 0$. The matrices $\boldsymbol{Y^TY}$ and $\boldsymbol{YY^T}$ have the same eigenvalues, and the elements d_i are the square roots of these eigenvalues; the i^{th} column $\mathbf{v}_i = (v_{i1},..., v_{ip})^T$ of the $\boldsymbol{V}_{p\times p}$ matrix is the eigenvector corresponding to the i^{th} largest eigenvalue d_i^2 of $\boldsymbol{Y^TY}$; while the j^{th} column $\boldsymbol{u}_j = (u_{1j},..., u_{nj})^T$ of the $\boldsymbol{U}_{n\times p}$ matrix is the eigenvector corresponding to the j^{th} largest eigenvalue d_j^2 of $\boldsymbol{YY^T}$. The decomposition (1) has its elementwise representation as follows:

$$y_{ij} = \sum_{h=1}^{p} u_{ih} d_h v_{jh}. \qquad (2)$$

Krzanowski (1987) used this representation as a basis for determining the dimensionality of a multivariate data set. If the data structure is essentially H-dimensional then the variation in the remaining (p – H) dimensions can be treated as random noise. The main features of the data can thus be supposed to lie in the space of the first H principal components. The correspondence between the quantities on the right-hand side of (2) and the principal axes of the data configuration suggests, therefore, the H-component model

$$y_{ij} = \sum_{h=1}^{H} u_{ih} d_h v_{jh} + \varepsilon_{ij}\,, \qquad (3)$$

in which ε_{ij} is a residual term.

Supposing that the model (3) holds for a specified value of H, but the single observation y_{ij} is missing from the data matrix, then y_{ij} can be estimated by

$$\hat{y}_{ij}^{(H)} = \sum_{h=1}^{H} u_{ih} d_h v_{jh}\,, \qquad (4)$$

in which the u_{ih}, d_h, v_{jh}, must be estimated from the remaining data. The best estimates of these latter quantities will be those that use the maximal amount of data. Denote by $\boldsymbol{Y}^{(-i)}$ the data matrix obtained on deleting the i^{th} row from $\boldsymbol{Y}$, and by $\boldsymbol{Y}_{(-j)}$ the data matrix obtained on deleting the j^{th} column from $\boldsymbol{Y}$, and suppose that the singular value decompositions of these matrices are given by:

$$\boldsymbol{Y}^{(-i)} = \bar{\boldsymbol{U}}\ \bar{\boldsymbol{D}}\bar{\boldsymbol{V}}^T, \quad \bar{\boldsymbol{U}} = (\bar{u}_{sh}), \quad \bar{\boldsymbol{V}} = (\bar{v}_{sh}), \quad \bar{\boldsymbol{D}} = (\bar{d}_1,\ldots,\bar{d}_p), \qquad (5)$$

and

$$\boldsymbol{Y}_{(-j)} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{V}}^T, \quad \tilde{\boldsymbol{U}} = (\tilde{u}_{sh}), \quad \tilde{\boldsymbol{V}} = (\tilde{v}_{sh}), \quad \tilde{\boldsymbol{D}} = (\tilde{d}_1,\ldots,\tilde{d}_{p-1}). \qquad (6)$$

The maximum-data estimates of u_{ih} and v_{jh} in (4) are clearly $\tilde{u}_{ih}$ and $\bar{v}_{jh}$, respectively, while d_h can be estimated either by $\bar{d}_h$, $\tilde{d}_h$ or by some combination of the two. Krzanowski (1988) suggested using $\sqrt{\bar{d}_h}\sqrt{\tilde{d}_h}$ as a suitable compromise, so that an estimate of the missing value y_{ij} is given by:

$$\hat{y}_{ij}^{(H)} = \sum_{h=1}^{H} (\tilde{u}_{ih}\sqrt{\tilde{d}_h})(\bar{v}_{jh}\sqrt{\bar{d}_h})$$

Following the maximum-data precept, the highest possible value of H should be used. From (6) this is evidently $p - 1$, so that the value to be imputed for y_{ij} will be:

$$\hat{y}_{ij} = \sum_{h=1}^{p-1} (\tilde{u}_{ih}\sqrt{\tilde{d}_h})(\bar{v}_{jh}\sqrt{\bar{d}_h}). \qquad (7)$$

An iterative numerical process is needed to find the appropriate quantities in (7). Starting with initial "guesses" for the missing values, each iteration requires singular value decompositions for the reduced matrices $\boldsymbol{Y}^{(-i)}$, $\boldsymbol{Y}_{(-j)}$ for every (i, j) where there is a missing value. The use of (7) provides the updated imputation for that position, and the process continues until convergence (i.e. stability in the successive imputed values). The initial "guesses" for the missing values y_{ij} are most easily provided by the mean $\bar{y}_j$ of the j^{th} column of the existing values. To avoid any possible variation in the influence among columns, for example caused by different measurement scales, it is recommended to first standardise $\boldsymbol{Y}$ (completed using the initial "guesses"). The $\boldsymbol{Y}^{(-i)}$ and $\boldsymbol{Y}_{(-j)}$ matrices should also be standardised at each step, so that all operations are conducted on standardised quantities. At the end, the completed matrix $\boldsymbol{Y}$ (i.e. observed + imputed values) should be returned to its original scale. Thus if $y_{ij}^{(c)}$ represents each value of the completed matrix, the j^{th} column mean ($\bar{y}_j^{(c)}$) and standard deviation ($s_j^{(c)}$) are calculated and each value of the completed matrix $\boldsymbol{Y}$ is obtained in its original scale as $y_{ij} = \bar{y}_j^{(c)} + s_j^{(c)} y_{ij}^{(c)}$.

The modification to this method that is here proposed is a generalisation to the exponents of $\tilde{d}_h$ and $\bar{d}_h$ in (7) when generating the imputations (m = 1,..., M) in the first step of the MI. For full generality, if d $\sqrt[b]{d^a}$ is represented as a fractional power $d^{\frac{a}{b}}$, we propose changing the exponents in (7) to $\tilde{d}_h^{\frac{\tilde{a}}{b}}$ and $\bar{d}_h^{\frac{\bar{a}}{b}}$ subject to the exponents summing to 1 (i.e. $\frac{\tilde{a}+\bar{a}}{b}=1$). In this way, different weights can be assigned to (5) and (6) in the final estimate of y_{ij} in (7), by varying the exponents of $\bar{d}_h$ and $\tilde{d}_h$, whereas the current form forces them to have equal weights.

Each different value of $\tilde{a}$ and consequently of $\bar{a}$ generates a new completed matrix $\boldsymbol{Y}$, thus providing a mechanism for generating the M different completed data sets at the first step of the multiple imputation process.

The number of imputations is governed by the number of different exponents used. According to Molenberghs & Verbeke (2005); Rubin (1987) and Schafer (1999), between 3 and 5 imputations should be enough to characterise the variability between imputations. Thus, if it is decided on 5 changes in the exponents, between 40% and 60% variation can be produced in the weight given to (5) and (6) by starting with a fixed denominator (b = 20, for instance) and taking values (8, 9, 10, 11 and 12) for $\tilde{a}$ and (12, 11, 10, 9 and 8) respectively for $\bar{a}$. These choices lead to a variation of (40%, 45%, 50%, 55% and 60%) respectively in the proportions of (5) and (6) in

$$\hat{y}_{ij} = \sum_{h=1}^{p-1} (\tilde{u}_{ih}\tilde{d}_h^{\frac{\tilde{a}}{b}})(\bar{v}_{jh}\bar{d}_h^{\frac{\bar{a}}{b}}). \qquad (8)$$

The methodology described above and proposed here uses the greatest quantity of values in $\boldsymbol{Y}$ and does not depend on any distribution in the response variable, being applicable to any matrix of numeric data.

Implementation of the method was made by means of a program developed in the IML module of the statistical system SAS, which, after its execution, resulted in a data file with the M = 5 completed data sets, ready to be used in the second step of the MI.

The procedures were applied to the three reduced data matrices, containing 5%, 10% and 30% of missing values. Stability of the imputed values, or their convergence to a single value, was generally achieved within 20 iterations, but, for reasons of security and generality in the program execution, 50 iterations was the default setting. For each percentage of missing values, dispersion graphics were produced for those positions in the data matrix that were randomly deleted at the 5% deletion level. This could be done because the random deletions had the same initial value as seed for all the different percentage deletion levels.

As a measure of performance of the method at missing value position l (falling in row i and column j), the following expression was used:

$$acc_l = \frac{\sum_{m=1}^{M}(\hat{y}_{ij_{(m)}} - VO_l)^2}{M-1},$$

adapted from Penny & Jolliffe (1999), in which M is the number of imputations at that missing value position, VO_l is the original randomly deleted value at that position, and $\hat{y}_{ij(m)}$ is the m^{th} imputation at that position using (8) according to the proposed method. This expression is computed for l = 1,2,...,na, where na is the total number of missing values. The expression can be separated into two terms,

$$acc_l = \frac{\sum_{m=1}^{M}(\hat{y}_{ij_{(m)}} - \bar{Y}_l)^2}{M-1} + \frac{M(\bar{Y}_l - VO_l)^2}{M-1} \qquad (9)$$

where $\bar{Y}_l$ is the actual value imputed at position l, so that the first term represents a variance over the M values in each position and the second a bias in the final imputation. Thus the first term is a measure of *precision* and the second term a measure of *accuracy* at position l.

An overall measure of performance T_{acc} may be computed by averaging the acc_l measures as follows:

$$T_{acc} = \frac{\sum_{l=1}^{na} acc_l}{na}.$$

in which $na = g \times e \times porc$, with g representing the total number of genotypes, e representing the total number of environments and *porc* representing the percentage of missing data.

T_{acc} may similarly be broken into two components,

$$T_{acc} = V_E + VQM \qquad (10)$$

in which

$$V_E = \frac{1}{na}\sum_{l=1}^{na}\left[\frac{\sum_{m=1}^{M}(\hat{y}_{ij_{(m)}} - \bar{Y}_l)^2}{M-1}\right] \qquad (11)$$

and

$$VQM = \frac{1}{na}\sum_{l=1}^{na}\frac{M(\bar{Y}_l - VO_l)^2}{M-1} \qquad (12)$$

The first component (V_E) represents the pooled variance between imputations within positions, therefore the greater its value is, the smaller is the precision of the multiple imputation method. However, a small value for this component does not necessarily mean that the imputation method is good, for the method may be biased. The second component (*VQM*) represents the average squared bias between the values of $\bar{Y}$ and VO, so the smaller the bias is, the bigger is the number of imputations that are similar to the original values and the greater is their accuracy. Therefore, the smaller are the values of V_E and *VQM*, the larger is the multiple imputation method.

RESULTS AND DISCUSSION

Table 1 gives the results for 5%, 10% and 30% of missing values of a random deletion in the matrix. The variability between imputations within each missing value is relatively small, as shown by the small coefficients of variation. Furthermore, this variability increases as the percentage of missing data rises from 5% to 10%, except at the position (19;2) (Figures 1 and 2). When the percentage of missing values increases from 10% to 30%, the variability increases at positions (19;2), (10;4) and (19;4), and it decreases at positions (2;2), (13;4), (5;6) and (4;7) (Figures 2 and 3). The overall precision of multiple imputation is good for these data (Tables 1-2). As for the accuracy, which is the difference between the imputed values and the original values, this shows less consistency than the precision. Values of *acc* are nearly all larger than their counterparts *Var* in Table 1, but the pattern is more variable in the other table. So there is no clear relationship between precision and accuracy over the individual missing value positions.

On aggregating over missing value positions, the larger component of the general measure (T_{acc}) in (10) at all the missing percentages is the average squared bias, see Table 2. However, as the percent-

Table 1 - Imputations for mean heights (m) at each position (i^{th} row; j^{th} column) of a random deletion (5%, 10% and 30% missing).

Deletion position	VO	Imputation								
		1	2	3	4	5	$\bar{Y}$	Var	CV	acc
5% missing										
(2;2)	24.00	22.37	22.35	22.35	22.35	22.35	22.35	0.0001	0.04	3.380
(19;2)	20.12	21.37	20.93	20.93	20.93	20.93	21.02	0.0387	0.94	1.051
(10;4)	18.94	21.06	21.05	21.03	21.02	21.01	21.03	0.0004	0.10	5.469
(13;4)	18.78	16.90	17.08	17.09	17.10	17.11	17.06	0.0077	0.52	3.724
(19;4)	15.68	14.58	14.97	14.98	15.00	15.01	14.91	0.0339	1.23	0.776
(5;6)	18.06	17.50	17.58	17.57	17.55	17.54	17.55	0.0010	0.18	0.327
(4;7)	13.03	12.82	12.80	12.80	12.80	12.81	12.81	0.0001	0.07	0.064
10% missing										
(2;2)	24.00	23.30	25.62	25.13	25.65	25.64	25.07	1.0263	4.04	2.454
(19;2)	20.12	21.53	21.33	21.26	21.40	21.33	21.37	0.0101	0.48	1.925
(10;4)	18.94	18.85	18.92	18.90	19.06	19.06	18.96	0.0097	0.51	0.010
(13;4)	18.78	16.88	18.71	18.67	18.74	18.73	18.35	0.6723	4.47	0.908
(19;4)	15.68	14.82	15.53	15.55	15.68	15.68	15.45	0.1306	2.33	0.197
(5;6)	18.06	16.78	16.03	16.92	16.89	16.89	16.70	0.1425	2.27	2.445
(4;7)	13.03	12.92	13.52	13.56	13.46	13.45	13.38	0.0703	1.96	0.226
30% missing										
(2;2)	24.00	23.78	23.94	23.14	23.22	22.94	23.40	0.1855	1.85	0.630
(19;2)	20.12	21.60	21.62	21.88	21.88	21.90	21.78	0.0226	0.70	3.441
(10;4)	18.94	19.00	19.78	18.86	18.72	18.98	19.07	0.1725	2.17	0.193
(13;4)	18.78	16.43	16.33	17.05	15.93	16.27	16.40	0.1666	2.49	7.219
(19;4)	15.68	15.11	14.57	15.46	14.22	15.69	15.01	0.3703	4.07	0.932
(5;6)	18.06	17.64	17.17	17.59	17.67	17.67	17.55	0.0458	1.22	0.375
(4;7)	13.03	12.63	12.73	12.61	12.72	12.71	12.68	0.0030	0.44	0.155

VO: Original value.   : Imputations means. *Var*: Variance between imputations. *CV* : Coefficient of variation. *acc* : Accuracy.

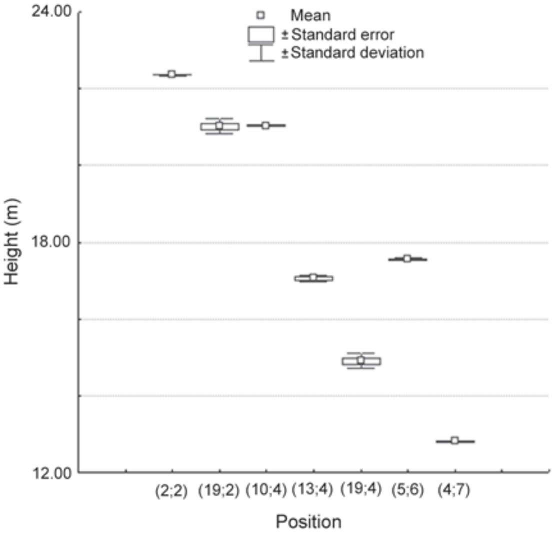

Figure 1 - Mean imputation height, standard error and standard deviation with 5% missing values.

Table 2 - General performance measure of the proposed multiple imputation method, with 5%, 10% and 30% missing values.

Missing	General accuracy		
	V_E	VQM	T_{acc}
5%	0.0119	2.1012	2.1131
10%	0.1700	1.2831	1.4531
30%	0.2070	1.2011	1.4081

age of missing values rises, the variation between imputations defined by expression (11) increases while the average squared bias defined by expression (12) decreases. This results in an improvement of the overall performance, which can be seen through the decrease in the general measure as defined by expression (10).

The proposed method for multiple imputation uses the maximum amount of information available from the original data matrix and is free from assump-

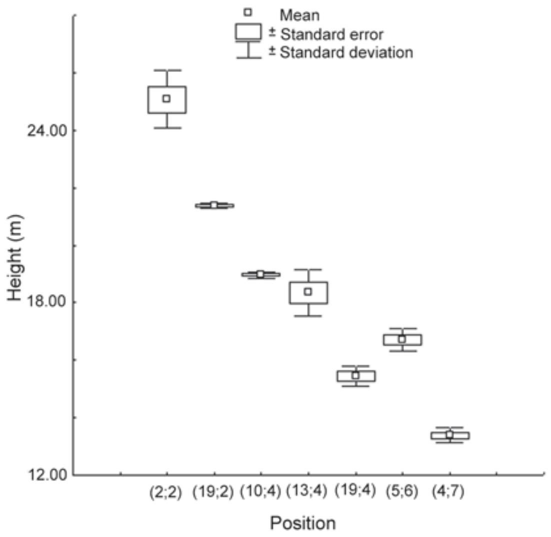

Figure 2 - Mean imputation height, standard error and standard deviation with 10% missing values, for the same positions as the data with 5% missing values.

tions about the distribution or structure of the data. In the G × E interaction matrix for the *E. grandis* height data, the variability in relation to the mean of imputed values is small, indicating a high precision, and the bias in relation to the original values is also small, indicating good accuracy. However, its values are greater than the variability in relation to the imputed values mean, indicating less accuracy in the proposed model.

ACKNOWLEDGMENTS

To CAPES and CNPq for the award of a scholarship to G.C. Bergamo and C.T.S. Dias, respectively.

REFERENCES

GOOD, I.J. Some applications of the singular value decomposition of a matrix **Technometrics**, v.11, p.823-831, 1969.

KRZANOWSKI, W.J. Cross-validation in principal component analysis. **Biometrics**, v.43, p.575-584, 1987.

KRZANOWSKI, W.J. Missing value imputation in multivariate data using the singular value decomposition of a matrix. **Biometrical Letters**, v.25, p.31-39, 1988.

LAVORANTI, O.J. Estabilidade e adaptabilidade fenotípica através da reamostragem "bootstrap" no modelo AMMI. Piracicaba: USP/ESALQ, 2003. 166p. Tese (Doutorado).

LITTLE, R.J.; RUBIN, D.B. **Statistical analysis with missing data**. 2nd ed. New York: John Wiley, 1987. 278p.

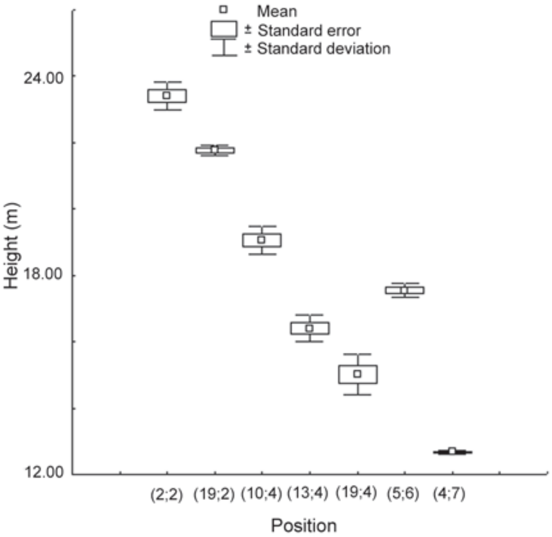

Figure 3 - Mean imputation height, standard error and standard deviation with 30% missing values, for the same positions as the data with 5% missing values.

LITTLE, R.J.; RUBIN, D.B. **Statistical analysis with missing data**. New York: John Wiley, 2002. 381p.

MOLENBERGHS, G.; VERBEKE, G. **Models for discrete longitudinal data**. New York: Springer Science, 2005. 683p.

PENNY, K.I.; JOLLIFFE, I.T. Multivariate outlier detection applied to multiply imputed laboratory data. **Statistics in Medicine**, v.18, p.1879-1895, 1999.

RUBIN, D.B. Multiple imputation in sample surveys: a phenomenological Bayesian approach to nonresponce. In: SURVEY RESEARCH METHODS SECTION OF THE AMERICAN STATISTICAL ASSOCIATION, 1978. **Proceedings**. p.20-34. Available at: http://www.amstat.org/sections/srms/Proceedings/. Accessed 20 Nov. 2006.

RUBIN, D.B. **Multiple imputation for nonresponce in surveys**. New York: John Wiley, 1987. 258p.

RUBIN, D.B.; SCHENKER, N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. **Journal of the American Statistical Association**, v.81, p.366-374, 1986.

SAS INSTITUTE. **SAS/IML 9.1:** user's guide. Carey: SAS Institute, 2004a. 1040p.

SAS INSTITUTE. **SAS/STAT 9.1:** user's guide. Carey: SAS Institute, 2004b. 5121p.

SCHAFER, J.L. **Analysis of incomplete multivariate data**. London: Chapman & Hall, 1997. 430p.

SCHAFER, J.L. Multiple imputation: a primer. **Statistical Methods in Medical Research**, v.8, p.3-15, 1999.

TANNER, M.A.; WONG, W.H. The calculation of posterior distributions by data augmentation (with discussion). **Journal of the American Statistical Association**, v.82, p.528-550, 1987.

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

Received March 29, 2007
Accepted January 07, 2008