Genevile Carife Bergamo

Engenheiro Agrônomo

Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação

Orientador:

Prof. Dr. **CARLOS TADEU DOS SANTOS
DIAS**

Tese apresentada para obtenção do título de Doutor em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba

2007

ERRATA

BERGAMO, G.C. **Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação.** Piracicaba, 2007. 89p. Tese (Doutorado em Agronomia) Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

Página	Linha	Onde se lê	Leia-se
14	5	$y = X\theta + e$	$Y = X\theta + e$
15	Figura 1	R	R_2
16	20	$= f(y_i X_i, Z_i, \theta) f(r_i y_i, W_i, \psi),$	$= f(y_i X_i, Z_i, \theta) f(r_i y_i, W_i, \psi),$
16	27	$= f(r_i W_i, \psi)$	$= f(r_i W_i, \psi)$
17	2	$= f(y_i X_i, Z_i, \theta) f(r_i W_i, \psi),$	$= f(y_i X_i, Z_i, \theta) f(r_i W_i, \psi),$
17	4	$= f(y_i^{obs} X_i, Z_i, \theta) f(r_i W_i, \psi).$	$= f(y_i^{obs} X_i, Z_i, \theta) f(r_i W_i, \psi).$
17	7	$= f(r_i y_i^{obs}, W_i, \psi)$	$= f(r_i y_i^{obs}, W_i, \psi)$
17	8	Dividida em	Fatorada em
17	9	$= f(y_i X_i, Z_i, \theta) f(r_i y_i^{obs}, W_i, \psi),$	$= f(y_i X_i, Z_i, \theta) f(r_i y_i^{obs}, W_i, \psi),$
17	11	$= f(y_i^{obs} X_i, Z_i, \theta) f(r_i y_i^{obs}, W_i, \psi)$	$= f(y_i^{obs} X_i, Z_i, \theta) f(r_i y_i^{obs}, W_i, \psi)$
17	14	Dependente	Depende
17	16	$= \int f(y_i X_i, Z_i, \theta)$	$= \int f(y_i X_i, Z_i, \theta)$
		$f(r_i y_i, W_i, \psi) dy_i^{aus}.$	$f(r_i y_i, W_i, \psi) dy_i^{aus}.$
20	4	Padrão de ausência	Mecanismo de ausência
20	20	Mecanismo MAR	Mecanismo MCAR ou MAR
22	22	Número m de imputações	Número M de imputações
22	24	Valor p	Valor p
23	19	Segue numa	Segue uma
29	3	$\sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk}$ (n: termos da interação).	$\sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk}$
29	20	AMMI (12)	AMMI (13)
38	1	j -ésima coluna ($i; j$)	j -ésima coluna
39	3	Interação $G \times E$	Interação $G \times E$

Dados Internacionais de Catalogação na Publicação (CIP)
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Bergamo, Genevile Carife

Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação / Genevile Carife Bergamo. - - Piracicaba, 2007. 89 p. : il.

Tese (Doutorado) - - Escola Superior de Agricultura Luiz de Queiroz, 2007.
Bibliografia.

1. Análise de dados 2. Correlação genética e ambiental 3. Estatística aplicada 4. Inferência não-paramétrica I. Título

CDD 519.53

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte - O autor"

Dedicatória

A

Deus

Jamais teria conseguido realizar este trabalho sem a
Sua graça.

Ao meu grande amor,

Maria Inêz Barbosa Braga Bergamo,

pelo incentivo constante, por suportar a ausência e
as privações,

à minha mãe **Nair P. C. Bergamo** (in memoriam),

pela dedicação durante toda sua vida e

ao meu pai **Casério Bergamo** pelo estímulo e apoio.

Aos

“meninos”,

sempre felizes,

alegando minha vida.

AGRADECIMENTOS

Ao Prof. Dr. Carlos Tadeu dos Santos Dias, pelo conhecimento compartilhado e apoio, tornando possível a realização deste trabalho.

À Coordenadoria para o Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudo.

Ao Prof. Dr. Enrico Antônio Colosimo pela disponibilidade de material bibliográfico.

Ao Prof. Dr. Décio Barbin, pelas orientações, mesmo durante suas férias.

Aos professores e funcionários do Departamento de Ciências Exatas da ESALQ - USP, pela atenção e amizade.

À Universidade de Alfenas, na pessoa do Reitor Edson Antônio Velano, dos Gestores João Batista Magalhães, Fuad Haddad e Marlene Leite Godoy V. de Souza, por possibilitar o afastamento das minhas atividades de docência.

À minha grande amiga Ana Maria Souza de Araujo por todo o conhecimento transmitido, sem o qual este trabalho nem começaria, e pelo compartilhar diário sempre construtivo.

Aos amigos de turma, David José Miquelutti, Denise Nunes Viola, Elisabeth Strapasson, José Carlos Fogo e Pedro Ferreira Filho, pela amizade e companheirismo, nos momentos difíceis e prazerosos.

Aos colegas do doutorado e mestrado, em especial à Angela pelas traduções e ao Lúcio pelas “dicas” do Tex.

Ao amigo, da época do mestrado, Osmir pela disponibilidade em fornecer os dados.

Às funcionárias do “RUCAS” sempre prestativas, em especial à D.^a Expedita pelas “frutas”.

Às pessoas que compartilharam direta ou indiretamente para a realização deste trabalho.

SUMÁRIO

RESUMO	6
ABSTRACT	7
LISTA DE FIGURAS	8
LISTA DE TABELAS	9
1 INTRODUÇÃO	11
2 DESENVOLVIMENTO	13
2.1 Considerações gerais	13
2.1.1 Padrão de ausência dos dados	14
2.1.2 Mecanismo de ausência dos dados	16
2.1.3 Imputação simples	19
2.1.4 Imputação múltipla	20
2.1.4.1 Inferência na imputação múltipla	21
2.1.4.2 Eficiência na imputação múltipla	24
2.1.5 Interação Genótipos $\times$ Ambientes	26
2.2 Metodologia	32
2.2.1 Material	32
2.2.2 Método	32
2.3 Resultados e discussão	39
3 CONCLUSÕES	60
REFERÊNCIAS	61
APÊNDICES	65
ANEXOS	69

RESUMO

Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação

Algumas técnicas de análise estatística multivariada necessitam de uma matriz de dados completa, porém o processo de coleta dos dados frequentemente não leva a uma matriz com todos os dados. A imputação é uma técnica, na qual os dados ausentes são preenchidos com valores plausíveis, para uma posterior análise dos dados completados (*observados + imputados*). O objetivo deste trabalho é propor um método de imputação múltipla, resultante de uma mudança no procedimento, baseado na decomposição por valores singulares (DVS), desenvolvido por Krzanowski (1988). Assim, na matriz genótipos (20) $\times$ ambientes (7), proveniente de um ensaio com o delineamento aleatorizado em blocos em multiambientes com a cultura de *Eucalyptus grandis* (LAVORANTI, 2003), foram retirados valores aleatoriamente (5%, 10%, 30%), os quais foram imputados pelo método proposto. Os resultados obtidos por meio da medida geral de exatidão ou acurácia (T_{acc}), na matriz de interação $G \times E$ para os dados de altura de *E. grandis*, mostraram um viés pequeno, em relação aos valores originais, no entanto, seus valores foram maiores do que a variabilidade em relação à média dos valores imputados, indicando uma exatidão ou acurácia menor do método proposto em relação à sua alta precisão. A metodologia proposta utiliza o maior número de informação disponível, não possui qualquer restrição quanto ao padrão e mecanismo de ausência e é livre de suposição sobre a distribuição ou estrutura dos dados.

Palavras-chave: Imputação múltipla; Não-paramétrico; Decomposição por valor singular; Interação genótipo-ambiente

ABSTRACT

Multiple imputation with distribution-free using the singular value decomposition in interaction matrix

Some techniques of the multivariate statistical analysis need a complete data matrix, but the process of data collection usually does not supply a complete data matrix. The imputation is a technique, in which the missing data are replaced by plausible values, for a latter analysis of the complete data set (observed + imputed). This work aims to propose a multiple imputation method, as a product of a procedures change, based on the singular value decomposition (SVD) developed by Krzanowski (1988). Thus, in the genotype (20) $\times$ environment (7) matrix ($G \times E$), derived from a trial following the complete randomized blocks design considering the *Eucalyptus grandis* genotype in multienvironments (LAVORANTI, 2003), values were retrieved randomly (5%, 10%, 30%), which were imputed by the proposed method. The results obtained by means of the general measure of accuracy (T_{acc}), in the interaction $G \times E$ for the height of *E. grandis* data matrix, showed a small bias when compared to the original data, however, its values were greater than the variability in relation to the imputed data mean, indicating a smaller accuracy of the proposed method in relation to its precision. The proposed methodology uses the greater number of information available, it does not possess any restriction about the pattern and missing mechanism and it is free of suppositions about the data distribution or structure.

Keywords: Multiple imputation; Distribution-free; Singular value decomposition; Genotype-environment interaction

LISTA DE FIGURAS

Figura 1 - Representação esquemática de $\mathbf{Y}$ e $\mathbf{R}$ para um conjunto de dados bivariado	15
Figura 2 - Padrões arbitrário e monótono de ausência dos dados para um conjunto de dados multivariado	15
Figura 3 - Imputações e valores originais (VO) de alturas nas posições de retirada dos dados com 5% de ausência	40
Figura 4 - Média, erro padrão e desvio padrão de alturas (m) para as imputações em cada posição de retirada dos dados com 5% de ausência	41
Figura 5 - Média, erro padrão e desvio padrão de alturas (m) para as imputações com 10% de ausência e nas mesmas posições de retirada dos dados com 5% de ausência	46
Figura 6 - Imputações e valores originais (VO) de alturas com 10% de ausência e nas mesmas posições de retirada dos dados com 5% de ausência	47
Figura 7 - Média, erro padrão e desvio padrão de alturas (m) para as imputações com 30% de ausência e nas mesmas posições de retirada dos dados com 5% de ausência	53
Figura 8 - Imputações e valores originais (VO) de alturas com 30% de ausência e nas mesmas posições de retirada dos dados com 5% de ausência	54

LISTA DE TABELAS

Tabela 1 - Eficiências relativas da estimação de imputação múltipla pelo número de imputações M e fração de informação ausente λ	25
Tabela 2 - Matriz de dados para a interação genótipos (linhas) e ambientes (colunas) .	27
Tabela 3 - Localização do teste de progênes de <i>Eucalyptus grandis</i> em sete municípios brasileiros	32
Tabela 4 - Média de altura, em metros, dos genótipos de <i>E. grandis</i> aos 5 anos nos diferentes ambientes	33
Tabela 5 - Média de alturas, em metros, das imputações, segundo a posição (linha i , coluna j) de retirada aleatória (5%) da Tabela 4	39
Tabela 6 - Média e erro padrão das médias de alturas (m) dos ambientes completados pelas imputações, nos dados com 5% de ausência	41
Tabela 7 - Estimativa média ($\hat{\beta}^*$) das médias de alturas e medidas associadas a sua variabilidade, nos ambientes com valores imputados para dados com 5% de ausência. Teste t -Student para comparação com a média original dos ambientes	42
Tabela 8 - Autovalor (λ_k^2), porcentagem da soma de quadrados acumulada (PA) nos eixos singulares (ES), desdobramento da interação $G \times E$, teste F , segundo Gollob e F_R , segundo Cornelius para os dados de alturas originais (orig.) e a média (imp.) das imputações, com 5% de ausência	44
Tabela 9 - Média de alturas, em metros, das imputações, segundo a posição (linha i , coluna j) de retirada aleatória (10%) da Tabela 4	45
Tabela 10 - Média e erro padrão das médias de alturas (m) dos ambientes completados pelas imputações, nos dados com 10% de ausência	48
Tabela 11 - Estimativa média ($\hat{\beta}^*$) das médias de alturas e medidas associadas a sua variabilidade, nos ambientes com valores imputados para dados com 10% de ausência. Teste t -Student para comparação com as médias originais . . .	49

Tabela 12 -Autovalor (λ_k^2), porcentagem da soma de quadrados acumulada (PA) nos eixos singulares (ES), desdobramento da interação $G \times E$, teste F , segundo Gollob e F_R , segundo Cornelius para os dados de alturas originais (orig.) e a média (imp.) das imputações, com 10% de ausência	50
Tabela 13 -Média de alturas, em metros, das imputações, segundo a posição (linha i , coluna j) de retirada aleatória (30%) da Tabela 4	51
Tabela 14 -Média e erro padrão das médias de alturas (m) dos ambientes completados pelas imputações, nos dados com 30% de ausência	55
Tabela 15 -Estimativa média ($\hat{\beta}^*$) das médias de alturas e medidas associadas a sua variabilidade, nos ambientes com valores imputados para dados com 30% de ausência. Teste t -Student para comparação com as médias originais . . .	56
Tabela 16 -Autovalor (λ_k^2), porcentagem da soma de quadrados acumulada (PA) nos eixos singulares (ES), desdobramento da interação $G \times E$, teste F , segundo Gollob e F_R , segundo Cornelius para os dados de alturas originais (orig.) e a média (imp.) das imputações, com 30% de ausência	57
Tabela 17 -Medida geral da acurácia do método de imputação múltipla proposto, com 5%, 10% e 30% de ausência	59
Tabela 18 -Médias de alturas, em metros, dos genótipos nos ambientes, com 5% de ausência nos dados	66
Tabela 19 -Médias de alturas, em metros, dos genótipos nos ambientes, com 10% de ausência nos dados	67
Tabela 20 -Médias de alturas, em metros, dos genótipos nos ambientes, com 30% de ausência nos dados	68

1 INTRODUÇÃO

Algumas técnicas de análise estatística multivariada necessitam de uma matriz de dados completa, porém o processo de coleta dos dados frequentemente não leva a uma matriz com todos os dados, isto é, algumas variáveis não são registradas, ou mesmo alguns de seus valores estão ausentes, provocando falhas nos dados originais. Por exemplo, em um experimento agrícola, os dados não estão disponíveis porque alguns animais morreram, algumas plantas foram danificadas, houve falhas no seu desenvolvimento, não houve material suficiente para todos os tratamentos, ou porque os dados não foram transcritos, calculados ou digitados corretamente. No caso de ensaios com melhoramento genético de genótipos em vários ambientes, a matriz de interação genótipos por ambientes ($G \times E$) pode ser incompleta, pois os genótipos nem sempre estão alocados em todos os ambientes, devido principalmente à falta de material genético para todos os ambientes. Assim, quando ocorre a ausência de um ou mais valores é necessário utilizar-se alguma técnica, a qual pode eliminar as variáveis com valores ausentes, reduzindo a informação do ensaio, ou imputar os valores ausentes, permitindo analisar todas as variáveis.

A imputação é o preenchimento dos dados ausentes com valores plausíveis para uma posterior análise dos dados completos. Ela pode ser simples, quando somente um valor é colocado para cada dado ausente, ou múltipla, quando há mais de um valor em cada dado ausente. Do ponto de vista operacional, a imputação resolve o problema dos dados ausentes, permitindo ao analista prosseguir com suas análises sem qualquer problema. No entanto, do ponto de vista estatístico, um método de imputação sem critérios pode criar mais problemas do que resolvê-los, distorcendo estimativas, erros padrão e testes de hipóteses, como descrito por Little e Rubin (2002).

A imputação múltipla, tal como na simulação de parâmetros, é um método de simulação de Monte Carlo para a análise de dados incompletos. Descrita por Rubin (1987), no contexto da ausência de respostas nos estudos exploratórios não experimentais (*survey*), é uma técnica de uso geral, podendo também ser aplicada aos ensaios experimentais.

O objetivo deste trabalho consiste em propor um método para a primeira etapa de imputação múltipla, sem suposição sobre a distribuição ou estrutura dos dados, utilizando a decomposição por valor singular (DVS), em uma matriz de interação ($G \times E$), para modelos

cujas análises necessitam de uma matriz completa.

Uma aplicação será feita aos dados provenientes de ensaios com o delineamento aleatorizado em blocos em multiambientes com a cultura de *Eucalyptus grandis* utilizados por Lavoranti (2003).

2 DESENVOLVIMENTO

2.1 Considerações gerais

Métodos historicamente importantes, baseados nos estimadores de mínimos quadrados, foram propostos há mais de cinquenta anos.

Allan e Wishart (1930) desenvolveram expressões para obter estimativa de mínimos quadrados para um único valor ausente nos delineamentos aleatorizados em blocos e quadrado latino. Por exemplo, para um experimento aleatorizado em blocos com B blocos e T tratamentos, a estimativa de mínimos quadrados para um valor ausente no bloco b e tratamento t é dado por:

$$\frac{Ty_+^{(t)} + By_+^{(b)} - y_+}{(T-1)(B-1)},$$

em que $y_+^{(t)}$ e $y_+^{(b)}$ são as somas parciais dos valores observados no tratamento t e no bloco b , respectivamente, e y_+ é a soma de todos os valores observados. Expressões, para vários delineamentos experimentais, foram desenvolvidas como continuação desse trabalho por Wilkin-son, 1958.

Um previsor para várias observações, por meio da minimização da soma de quadrados de resíduos, foi proposto na área agrícola por Yates (1933). Por outro lado, Healy e Westmacott (1956) descreveram uma técnica iterativa muito conhecida, muitas vezes atribuída a Yates ou mesmo a Fisher, na qual os valores ausentes são, inicialmente, substituídos por quaisquer valores. Assim, feita a análise nos dados completos, obtêm-se valores preditos para cada valor ausente. Substituindo-se os valores ausentes por esses preditos, uma nova análise dos dados completos é realizada, e o processo iterativo continua até que os valores preditos não sofram grandes mudanças e a soma de quadrados do resíduo pare de decrescer.

Bartlett (1937) desenvolveu um método não iterativo no qual os valores ausentes também são, inicialmente, substituídos por quaisquer valores, geralmente zero ou a média geral dos dados disponíveis; uma covariável é definida para cada valor ausente, executando-se então uma análise de covariância. Esses métodos levam a uma perda de graus de liberdade, e, atualmente, com o processamento computacional bem mais rápido, são pouco utilizados.

Hartley e Hocking (1971) utilizam estimativas de máxima verossimilhança para dados incompletos com distribuição normal, ou seja, $\mathbf{Y}$ é o vetor de p -variáveis normalmente

e independentemente distribuídas $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, em que o vetor de médias ($\boldsymbol{\mu}$) com p -elementos e a matriz ($p \times p$) de variância e covariância devem ser estimados, quando elementos do vetor $\mathbf{Y}$ estão ausentes. Também utilizam estimativas de máxima verossimilhança, na análise de variância com unidades experimentais ausentes e na análise de regressão linear com valores de x ausentes. Nestes casos, têm-se um modelo linear $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}$, em que $\mathbf{e}$ são valores independentes de uma $N(0, \sigma^2)$ e $\mathbf{X}$ é a matriz do delineamento, com efeitos fixos, no caso da análise de variância.

Rubin (1976) descreve como o *processo de ausência dos dados* se relaciona com a inferência sobre o vetor de parâmetros dos dados ($\boldsymbol{\theta}$), procurando condições apropriadas, nas quais, o *processo de ausência dos dados* possa ser ignorado. Desenvolve uma classificação, complementada em Little e Rubin (1987), na qual, a análise de dados com valores ausentes leva em consideração o padrão e o mecanismo de ausência dos dados.

2.1.1 Padrão de ausência dos dados

Considere Y um conjunto de dados retangular ($n \times p$), em que as linhas são as unidades ($i = 1, 2, \dots, n$), representando uma amostra aleatória de alguma distribuição de probabilidade multivariada p -dimensional e as colunas são as variáveis ($j = 1, 2, \dots, p$), sendo que as variáveis respostas estão agrupadas em um vetor $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{ip})^T$. Admitindo-se $\mathbf{R} = (R_{ij})$ como uma matriz $n \times p$ indicadora da ausência dos dados tal que:

$$R_{ij} = \begin{cases} 1 & \text{se } Y_{ij} \text{ é observado} \\ 0 & \text{se } Y_{ij} \text{ é ausente} \end{cases}$$

os quais estão agrupados em um vetor $\mathbf{R}_i$, com dimensões iguais às de $\mathbf{Y}_i$.

Uma representação esquemática de $\mathbf{R}$ e $\mathbf{Y}$, considerando apenas duas variáveis Y_1 e Y_2 , em que Y_1 é observado de $1, 2, \dots, n$ e Y_2 é observado de $1, 2, \dots, n_1 < n$, está na Figura 1. Os pontos de interrogação representam as unidades com valores ausentes.

Obs.	Y_1	Y_2	R
1			1
2			1
⋮			⋮
⋮			⋮
n_1			1
⋮		?	0
⋮		⋮	⋮
⋮		⋮	⋮
n		?	0

Figura 1 - Representação esquemática de Y e R para um conjunto de dados bivariado

Os dois principais padrões de ausência de dados da Figura 2 representam a maneira como ocorrem os valores ausentes e os valores observados em um conjunto de dados.

Arbitrário						Monótono					
Obs.	Variáveis					Obs.	Variáveis				
	Y_1	Y_2	Y_3	...	Y_p		Y_1	Y_2	Y_3	...	Y_p
1				?		1					
2		?				2					?
⋮						⋮				?	?
⋮		?				⋮			?	?	?
⋮					?	⋮		?	?	?	?
⋮					?	⋮	?	?	?	?	?
n					?	n		?	?	?	?

Figura 2 - Padrões arbitrário e monótono de ausência dos dados para um conjunto de dados multivariado

Na ausência de dados com padrão arbitrário, também conhecido como geral, ou totalmente casual, os valores ausentes ocorrem de maneira intermitente. O padrão monótono, o qual pode surgir por uma permutação das linhas e colunas com valores ausentes, é cara-

característico de experimentos longitudinais, em que um indivíduo é medido p vezes ao longo do tempo. No entanto, quando um indivíduo sai do experimento (*dropout* ou *attrition*), a partir dessa ocasião, não haverá mais a sua resposta, ou seja, quando se observa um Y_j ausente, todos os seguintes $Y_{j+1}, \dots, Y_p$ também estarão ausentes, para todo $j = 1, \dots, p - 1$.

2.1.2 Mecanismo de ausência dos dados

Dividindo-se o vetor $\mathbf{Y}_i$ em dois subvetores $\mathbf{Y}_i^{obs}$, representando os valores de Y_{ij} para os quais $R_{ij} = 1$, ou seja, os valores observados e o subvetor $\mathbf{Y}_i^{aus}$, os valores ausentes ($R_{ij} = 0$). Um conjunto de *dados completos*, considerando os indicadores de ausência dos dados $\mathbf{R}_i$, é representado por $(\mathbf{Y}_i, \mathbf{R}_i)$.

Para considerar a classificação do mecanismo de ausência dos dados, proposta por Rubin (1976), Little e Rubin (1987) e seguindo a notação utilizada por Molenberghs e Verbeke (2005), deve-se considerar, primeiramente, a função de probabilidade ou densidade dos *dados completos*

$$f(\mathbf{y}_i, \mathbf{r}_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}), \quad (1)$$

em que, X_i é a covariável relacionada aos efeitos fixos e Z_i aos efeitos aleatórios, se aplicável, W_i a covariável relacionada ao *processo de ausência dos dados* ($\mathbf{R}_i$) e $\boldsymbol{\theta}$, $\boldsymbol{\psi}$ são os vetores associados, respectivamente, aos efeitos fixos, aleatórios (*processo de medição*) e ao *processo de ausência*, os quais parametrizam a distribuição conjunta.

Fatorando o modelo (1) obtém-se:

$$f(\mathbf{y}_i, \mathbf{r}_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i | X_i, Z_i, \boldsymbol{\theta}) f(\mathbf{r}_i | \mathbf{y}_i, W_i, \boldsymbol{\psi}), \quad (2)$$

em que o primeiro fator é a densidade marginal do *processo de medição* e o segundo a densidade do *processo de ausência*, condicionada às variáveis respostas ($\mathbf{Y}_i$). Assim, o mecanismo de ausência dos dados é classificado em três categorias, dependendo de $f(\mathbf{r}_i | \mathbf{y}_i, W_i, \boldsymbol{\psi}) = f(\mathbf{r}_i | \mathbf{y}_i^{obs}, \mathbf{y}_i^{aus}, W_i, \boldsymbol{\psi})$, segundo fator de (2),

- (1) Completamente Aleatório (MCAR - *Missing Completely At Random*) se a probabilidade de uma observação estar ausente é independente dos valores observados ($\mathbf{y}_i^{obs}$) e ausentes ($\mathbf{y}_i^{aus}$), ou seja, $f(\mathbf{r}_i | \mathbf{y}_i, W_i, \boldsymbol{\psi}) = f(\mathbf{r}_i, W_i, \boldsymbol{\psi})$

conseqüentemente (2) fica simplificada, com os dois fatores independentes

$$f(\mathbf{y}_i, \mathbf{r}_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i, | X_i, Z_i, \boldsymbol{\theta}) f(\mathbf{r}_i | W_i, \boldsymbol{\psi}),$$

e a distribuição conjunta de $\mathbf{y}_i^{obs}$ e $\mathbf{r}_i$ torna-se

$$f(\mathbf{y}_i^{obs}, \mathbf{r}_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i^{obs}, | X_i, Z_i, \boldsymbol{\theta}) f(\mathbf{r}_i | W_i, \boldsymbol{\psi}).$$

- (2) Aleatório (MAR - *Missing At Random*) se condicionada aos valores observados ($\mathbf{y}_i^{obs}$) a probabilidade de uma observação estar ausente é independente dos valores ausentes ($\mathbf{y}_i^{aus}$), ou seja, $f(\mathbf{r}_i | \mathbf{y}_i, W_i, \boldsymbol{\psi}) = f(\mathbf{r}_i, \mathbf{y}_i^{obs}, W_i, \boldsymbol{\psi})$

Novamente (2) fica dividida em

$$f(\mathbf{y}_i, \mathbf{r}_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i, | X_i, Z_i, \boldsymbol{\theta}) f(\mathbf{r}_i | \mathbf{y}_i^{obs}, W_i, \boldsymbol{\psi}),$$

com a distribuição conjunta de $\mathbf{y}_i^{obs}$ e $\mathbf{r}_i$ dada por:

$$f(\mathbf{y}_i^{obs}, \mathbf{r}_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i^{obs}, | X_i, Z_i, \boldsymbol{\theta}) f(\mathbf{r}_i | \mathbf{y}_i^{obs}, W_i, \boldsymbol{\psi}).$$

- (3) Não Aleatório (NMAR - *Not Missing At Random* ou MNAR *Missing Not At Random*) se condicionada aos valores observados ($\mathbf{y}_i^{obs}$) a probabilidade de uma medida estar ausente dependente dos valores ausentes ($\mathbf{y}_i^{aus}$). Nenhuma simplificação é possível, sendo que a distribuição conjunta de $\mathbf{y}_i^{obs}$ e $\mathbf{r}_i$ deve ser escrita como

$$f(\mathbf{y}_i^{obs}, \mathbf{r}_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\mathbf{y}_i, | X_i, Z_i, \boldsymbol{\theta}) f(\mathbf{r}_i | \mathbf{y}_i, W_i, \boldsymbol{\psi}) d\mathbf{y}_i^{aus}.$$

Como exemplo, numa pesquisa sobre o peso de pessoas, se ausência de resposta sobre o peso não está relacionada com o próprio peso do entrevistado e nem com qualquer outra variável, como por exemplo, a idade, o sexo desse entrevistado, então o mecanismo de ausência de valores para o peso é MCAR. Se as pessoas com sobrepeso tendem a não informar seu peso, a ausência de resposta sobre o peso depende do próprio peso, caracterizando o mecanismo de ausência MNAR. No entanto, se a ausência de resposta sobre o peso não depende do próprio peso, mas pode depender de outras variáveis (pessoas do sexo feminino tendem a não informar seu peso), diz-se que o mecanismo de ausência de valores para o peso é

MAR. Nesse exemplo e na maioria das situações reais, o mecanismo mais provável de ocorrer é o MNAR, pois a ausência de informação depende da própria variável.

Dempster, Laird e Rubin (1977) desenvolveram, sob um mecanismo MAR, o algoritmo *expectation-maximisation* (EM), o qual se tornou um método muito utilizado para otimização da verossimilhança em dados incompletos.

Little e Rubin (1987, 2002) reúnem os principais métodos voltados à análise de dados com valores ausentes. Um deles, a Análise de Caso Completo (*Complete-case analysis*), em que apenas as variáveis com todos os valores observados são analisadas, possui como vantagem a sua simplicidade, pois a análise é feita em um conjunto de dados completamente balanceado. No entanto, ocorre a perda de informação das variáveis excluídas da análise e um maior viés nos estimadores, quando o mecanismo de ausência é MAR, ao invés de MCAR.

Uma maneira de conferir a suposição MCAR, seria dividir as unidades observadas das variáveis Y_j em dois grupos: (1) aquelas unidades observadas em todas as variáveis, (2) aquelas com ausência numa das variáveis. Se ocorrer MCAR, os dois grupos deveriam ser amostras aleatórias da mesma população (VERBEKE; MOLENBERGHS, 1997).

Outro método é a Análise de Caso Disponível (*Available-case analysis*), no qual todos os valores observados são utilizados. Assim sendo, ele é mais eficiente que a análise de caso completo. As desvantagens estão nas mudanças que podem ocorrer entre as variáveis com os valores observados e aquelas com os valores ausentes e na sua utilização, somente válida, sob o mecanismo MCAR.

Estes dois métodos não utilizam as possíveis correlações existentes entre variáveis, ou seja, se na variável Y_j com valor ausente, representado, por exemplo, pela variável peso, há uma alta correlação com outra variável Y_p , por exemplo, a altura, é possível prever o valor ausente (y_{ij}) de Y_j por meio de Y_p , incluindo assim o valor imputado nas análises que envolvem Y_j . Os métodos utilizados no preenchimento dos valores ausentes são conhecidos como métodos de imputação, os quais podem ser aplicados para imputar um único valor para cada dado ausente (imputação simples) ou, em alguns casos, imputar mais de um valor (imputação múltipla), permitindo, assim, uma avaliação apropriada da incerteza da imputação.

2.1.3 Imputação simples

Vários métodos de imputação simples são utilizados, tanto em estudos exploratórios não experimentais (*Survey*), tais como a imputação *Hot deck*, *Cold deck*, ou a imputação por *Substituição* (DAVID et al., 1986; GROVES et al., 2002), como nos estudos experimentais. Nestes, Little e Rubin (1987, 2002) e Rubin (1987) mencionam vários métodos, dentre eles, a Imputação por meio da Última Observação (IUO) (*Last Observation Carried Forward - LOCF*) é um método de imputação simples, em que o valor ausente é substituído pelo último valor observado. Este método pode ser aplicado tanto a padrões monótonos como a não monótonos de ausência dos dados, embora seja típico das situações em que a ausência é monótona. Algumas suposições devem ser feitas para assegurar a validade deste método, ou seja, nas ausências monótonas, a partir do momento em que a resposta não é mais observada, todas as unidades devem ter um perfil constante, e no caso da ausência totalmente casual, durante as unidades ausentes.

Outro método de imputação simples é a Imputação por meio da Média, na qual todos os valores ausentes são substituídos pela média geral ($\bar{y}_{ij}$) dos valores observados. Nos experimentos longitudinais esta substituição pode ser feita, tanto pela média dos valores observados na i -ésima unidade em tempos distintos ($\bar{y}_{i*}$), como também pela média dos valores observados no j -ésimo tempo em unidades distintas ($\bar{y}_{*j}$). No caso do padrão de ausência geral, a substituição do valor ausente numa variável j é feita pela média dos valores observados nessa variável ($\bar{y}_{*j}$). Little e Rubin (1987, 2002) chamam essa imputação de *unconditional mean imputation*, pois não usa (*conditional*) os diferentes valores da unidade (i) para a qual a imputação é feita.

A Imputação por meio da Regressão foi proposta por Buck (1960) e revisada por Little e Rubin (1987, 2002), a qual chamaram de *conditional mean imputation*. Este é um método mais elaborado quando comparado à Imputação por meio da Média, utilizado para qualquer padrão de ausência dos dados e quando as variáveis possuem uma relação linear. Consiste em duas etapas, considerando as variáveis $Y_1, \dots, Y_p$ normais multivariadas, a primeira etapa calcula o vetor de médias $\boldsymbol{\mu}$ e a matriz de covariâncias Σ nos valores observados ($\mathbf{Y} \sim N(\boldsymbol{\mu}, \Sigma)$). Para as unidades com valores ausentes (Y_i^{aus}), utiliza as estimativas $(\boldsymbol{\mu}, \Sigma)$ numa regressão linear por quadrados mínimos das variáveis com unidades ausentes

em relação às variáveis em que essas mesmas unidades estão presentes (y_i^{obs}). Na segunda etapa, as unidades com valores ausentes são preditas pela substituição das unidades observadas na regressão apropriada. Os cálculos das diferentes regressões lineares podem ser feitos utilizando-se o operador *sweep* (LITTLE; RUBIN, 1987, 2002).

Uma limitação na maioria desses métodos de imputação simples é o padrão de ausência dos dados ser MCAR, mas uma outra importante limitação está nas inferências dos parâmetros, as quais baseadas nos dados completados (*observados + imputados*), não consideram a incerteza da imputação. Assim, os desvios padrões, calculados nos dados completados, não são estimados corretamente, ocasionando um valor p (probabilidade de significância) e intervalos de confiança incorretos. Uma maneira de contornar esses problemas é a utilização do método de imputação múltipla (IM).

2.1.4 Imputação múltipla

A IM foi proposta primeiramente por Rubin (1978), mas várias outras referências, tais como, Little e Rubin (1987, 2002); Rubin (1987); Rubin e Schenker (1986); Schafer (1997, 1999); Tanner e Wong (1987); Zhang (2003), fornecem excelentes descrições da técnica. A concepção fundamental do procedimento é substituir cada valor ausente por um conjunto de M valores imputados, ou seja, esses valores são “retirados” da distribuição dos próprios dados, representando a incerteza sobre o verdadeiro valor a ser imputado. Todos os M conjuntos de dados completados são analisados, individualmente, usando-se métodos padrões para análise de dados completos, cujos resultados são combinados em uma única análise. O mecanismo MAR, de ausência dos dados, é necessário para a aplicação da IM, embora Thijs et al (2002) tenham aplicado o método em condições MNAR.

Baracho (2003) resume o procedimento da IM em três etapas:

- (1) **Imputação:** Os valores ausentes são completados M vezes, gerando M conjuntos de dados completados.
- (2) **Análise:** Os M conjuntos de dados completados são analisados, usando procedimentos estatísticos de interesse.
- (3) **Combinação:** Os resultados das M análises são combinados em uma única inferência.

A imputação é a etapa mais crítica, pois nesta o mecanismo de ausência está sendo considerado. Uma suposição MAR permite gerar as imputações, partindo da distribuição dos dados ausentes condicionada aos dados observados. Quanto ao modelo utilizado na etapa de imputação, não é necessariamente o mesmo da etapa de análise, tornando o procedimento da IM mais atrativo, pois nem sempre o modelo utilizado para imputar é o mais adequado para analisar (BARACHO, 2003).

Ao combinar os M resultados das análises, a variância da estimativa combinada, consiste em variância dentro das imputações e em variância entre imputações, portanto, as incertezas dos dados imputados são incorporadas à inferência final.

2.1.4.1 Inferência na imputação múltipla

Molenberghs e Verbeke (2005) descrevem teoricamente as etapas da imputação múltipla, supondo uma amostra i.i.d. de vetores aleatórios $\mathbf{Y}_i$ ($n \times 1$), o interesse está na estimação de algum vetor $\boldsymbol{\theta}$ de parâmetros da distribuição de $\mathbf{Y}_i$. A imputação múltipla preenche os dados ausentes $\mathbf{Y}^{aus}$ várias vezes, usando os dados observados $\mathbf{Y}^{obs}$ e com os dados completados calcula $\boldsymbol{\theta}$. Se a distribuição de $\mathbf{Y}_i = (\mathbf{Y}_i^{obs}, \mathbf{Y}_i^{aus})$, com vetor de parâmetros $\boldsymbol{\theta}$, fosse conhecida, seria possível imputar $\mathbf{Y}_i^{aus}$ a partir da distribuição condicional $f(\mathbf{y}_i^{aus} | \mathbf{y}_i^{obs}, \boldsymbol{\theta})$. Como $\boldsymbol{\theta}$ não é conhecido, estima-se seu valor ($\hat{\boldsymbol{\theta}}$) por meio dos dados, sendo então, $f(\mathbf{y}_i^{aus} | \mathbf{y}_i^{obs}, \hat{\boldsymbol{\theta}})$ usada na imputação dos dados ausentes. Numa visão freqüentista, é possível incorporar a incerteza em $\boldsymbol{\theta}$, na etapa de imputação, utilizando-se métodos de reamostragem. Porém, em condições Bayesianas, $\boldsymbol{\theta}$ é uma variável aleatória, cuja distribuição é função dos dados. A aproximação Bayesiana depende de uma integração sobre $\hat{\boldsymbol{\theta}}$, a qual fornece uma informação da incerteza em $\boldsymbol{\theta}$. Da distribuição de $\boldsymbol{\theta}$, primeiramente gera-se um $\boldsymbol{\theta}^*$ aleatório e, então, colocando-se esse $\boldsymbol{\theta}^*$ em $f(\mathbf{y}_i^{aus} | \mathbf{y}_i^{obs}, \boldsymbol{\theta}^*)$, gera-se um $\mathbf{Y}_i^{aus}$ aleatório.

Após formular a distribuição de $\boldsymbol{\theta}$, o algoritmo de imputação é composto das seguintes etapas:

- (1) Gerar $\boldsymbol{\theta}^*$ da distribuição de $\boldsymbol{\theta}$.
- (2) Gerar $\mathbf{Y}_i^{aus*}$ de $f(\mathbf{y}_i^{aus} | \mathbf{y}_i^{obs}, \boldsymbol{\theta}^*)$.
- (3) Usando os dados completados $(\mathbf{Y}_i^{obs}, \mathbf{Y}_i^{aus*})$ e um método de estimação (máxima

verossimilhança, máxima verossimilhança restrita, método dos momentos, verossimilhança parcial) estimar um parâmetro de interesse, por exemplo β , por meio de $\hat{\beta} = \hat{\beta}(Y) = \hat{\beta}(Y_i^{obs}, Y_i^{aus*})$ e sua variância $U = \widehat{var}(\hat{\beta})$, chamada de variância dentro das imputações.

- (4) Independentemente, repetir as etapas 1, 2 e 3, M vezes. Os M conjuntos de dados completados produzirão $\hat{\beta}^m$ e U^m para $m = 1, \dots, M$.

As M estimativas imputadas para β devem ser combinadas em uma única, a fim de obter-se a estimativa da imputação múltipla. Para os dados completados e, supondo-se que a inferência sobre β possa ser feita por $(\beta - \hat{\beta}) \sim N(0, U)$, a estimativa da imputação múltipla é a média aritmética das M estimativas

$$\hat{\beta}^* = \frac{1}{M} \sum_{m=1}^M \hat{\beta}^m.$$

Além disso, as inferências para β serão baseadas na distribuição normal, $(\beta - \hat{\beta}^*) \sim N(0, V)$, em que

$$V = W + \left(\frac{M+1}{M}\right)B,$$

é a variabilidade total associada a $\hat{\beta}^*$, com

$$W = \frac{1}{M} \sum_{m=1}^M U^m,$$

a média das variâncias dentro das imputações e

$$B = \frac{1}{M-1} \sum_{m=1}^M (\hat{\beta}^m - \hat{\beta}^*)(\hat{\beta}^m - \hat{\beta}^*)^T,$$

a variância entre as imputações.

Os testes de hipóteses, associados a uma hipótese nula $\theta = \theta_0$, e os intervalos de confiança, não dependem apenas do tamanho do conjunto dos dados completados, mas também do número m de imputações. Li, Raghunathan e Rubin (1991) propõem o uso da distribuição F para calcular a probabilidade de significância (valor p), ou seja

$$\text{valor } p = P(F_{p,w} > F),$$

em que, $F_{p,w}$ é uma variável aleatória com distribuição F , p (comprimento do vetor de parâmetros θ) e w graus de liberdade, F é igual a

$$F = \frac{(\boldsymbol{\theta}^* - \boldsymbol{\theta}_o)^T \mathbf{W}^{-1} (\boldsymbol{\theta}^* - \boldsymbol{\theta}_o)}{p(1+r)},$$

$$w = 4 + (t-4) \left[1 + \frac{1}{r} \left(1 - \frac{2}{t} \right) \right]^2,$$

com

$$t = p(M-1),$$

e

$$r = \frac{1}{p} \left(1 + \frac{1}{M} \right) \text{tr}(\mathbf{B}\mathbf{W}^{-1})$$

é o aumento relativo na variância devido aos dados ausentes (RUBIN, 1987).

No caso univariado, se o tamanho do conjunto de dados completados é grande e o número de imputações (m) pequeno, os testes de hipóteses e os intervalos de confiança estão baseados na distribuição t-Student. Assim, assintoticamente $(\beta - \hat{\beta}^*)T^{(-1/2)} \sim t_\nu$, em que,

$$\hat{\beta}^* = \frac{1}{M} \sum_{m=1}^M \hat{\beta}^m \quad (3)$$

é a estimativa da imputação múltipla,

$$T = \bar{W} + \left(1 + \frac{1}{M} \right) B \quad (4)$$

é a estimativa da variância total,

$$\bar{W} = \frac{1}{M} \sum_{m=1}^M U^m \quad (5)$$

é a média das variâncias dentro das imputações,

$$B = \frac{1}{(M-1)} \sum_{m=1}^M (\hat{\beta}^m - \hat{\beta}^*)^2 \quad (6)$$

é a variância entre as imputações e

$$\nu = (M-1) \left[1 + \frac{\bar{W}}{(1 + M^{-1})B} \right]^2 \quad (7)$$

é o número de graus de liberdade, os quais estão baseados na suposição de que a inferência dos dados completados segue numa distribuição normal, isto é, o número de graus de liberdade dos dados completados, ν_{com} , é infinito e o número de imputações é finito. Quando ν_{com}

é pequeno e há somente uma pequena proporção de dados ausentes, o número de graus de liberdade definido por (7) pode ser muito maior do que ν_{com} , tornando (7) inapropriada para o cálculo do número de graus de liberdade. Barnard e Rubin (1999) recomendam o uso do número de graus de liberdade ajustados, ν^* , calculado por

$$\nu^* = \left(\frac{1}{\nu} + \frac{1}{\nu_{obs}} \right)^{-1}, \quad (8)$$

em que

$$\nu_{obs} = \left(\frac{\nu_{com} + 1}{\nu_{com} + 3} \right) \nu_{com} (1 - \gamma)$$

e

$$\gamma = \frac{1}{T} \left(1 + \frac{1}{M} \right) B.$$

2.1.4.2 Eficiência na imputação múltipla

Molenberghs e Verbeke (2005) comentam sobre a alta eficiência da imputação múltipla, até mesmo para valores pequenos de m , sendo que, em muitas aplicações, de 3 a 5 imputações são suficientes para obter excelentes resultados. Rubin (1987, p. 114) mostra a eficiência relativa (ER - na unidade da variância) de uma estimativa pontual baseada em m imputações, com relação a uma baseada em um número infinito de imputações, aproximadamente como,

$$ER = \left(1 + \frac{\lambda}{M} \right)^{-1}, \quad (9)$$

em que,

$$\lambda = \frac{1}{r + 1} \left(r + \frac{2}{\nu + 3} \right) \quad (10)$$

é a fração de informação ausente, a qual mede a precisão da estimativa (3), se nenhum dado está ausente e no caso univariado,

$$r = \frac{1}{\bar{W}} \left(1 + \frac{1}{M} \right) B \quad (11)$$

é o aumento relativo na variância devido aos dados ausentes (RUBIN, 1987). As eficiências relativas, para diferentes números de imputações (m) e frações de informação ausente (λ) da Tabela 1, mostram um ganho muito pequeno em eficiência após as primeiras imputações.

Tabela 1 - Eficiências relativas da estimação de imputação múltipla pelo número de imputações M e fração de informação ausente λ

m	λ				
	0,1	0,2	0,3	0,5	0,7
3	0,9677	0,9375	0,9091	0,8571	0,8108
5	0,9804	0,9615	0,9434	0,9091	0,8772
10	0,9901	0,9804	0,9709	0,9524	0,9346
20	0,9950	0,9901	0,9852	0,9756	0,9662

Schafer (1999) reafirma a necessidade de um número pequeno de imputações na IM, para um $\lambda = 0,5$ a estimativa baseada em $m = 5$ imputações tem um desvio padrão aproximadamente 5% ($\sqrt{1 + 0,5/5} = 1,049$) maior do que uma estimativa baseada em $m \rightarrow \infty$. Assim, a menos que as frações de informação ausente sejam muito grande, há um pequeno ou quase nenhum benefício prático ao utilizar entre cinco e dez imputações.

Vários modelos podem ser usados na primeira etapa da IM, dentre eles, quando o padrão de ausência dos dados é monótono e as variáveis são contínuas, há o método não paramétrico, do escore de propensão (*propensity score method*) introduzido inicialmente por Rosenbaum e Rubin (1983), sendo complementado por Rubin (1987) e Lavori; Dawson e Shera (1995). Assumindo uma normalidade multivariada é possível a utilização do método da regressão (*regression method*) descrito por Rubin (1987) ou do método da combinação preditiva da média (*predictive mean matching method*) desenvolvido por Heitjan e Little (1991) e Schenker e Taylor (1996). Já para as variáveis discretas binárias ou ordinais há o método da regressão logística (RUBIN, 1987) e para as binárias ou nominais o método da função discriminante. Se o o padrão de ausência dos dados é arbitrário e a variável contínua, pode ser utilizado o método (*MCMC method*) proposto por Schafer (1997), baseado no método Monte Carlo para cadeias de Markov.

Alguns softwares são utilizados para implementar e executar estes métodos, bem como as outras etapas da IM. Horton e Lipsitz (2001) descrevem e comparam os mais utilizados e citados: o SOLAS 3.0; o S-Plus com a biblioteca de ausência dos dados; o SAS 8.2 com a versão beta para o Proc MI, na primeira etapa da IM, e o Proc MIANALYZE na

terceira etapa, procedimentos estes incorporados à versão 9.1 com algumas mudanças; o MICE (*multiple imputation by chained equations*), cuja linguagem e interface é muito semelhante ao S-Plus, com uma vantagem em relação aos anteriores, é gratuito, pois faz parte do software R. Estes softwares podem analisar vários tipos de dados, inclusive a matriz **GE** proveniente da interação Genótipos $\times$ Ambientes, nos ensaios de melhoramento genético.

2.1.5 Interação Genótipos $\times$ Ambientes

Crossa (1990), revisando alguns métodos de análise estatística para ensaios de produtividade em multiambientes, coloca três principais objetivos agrícolas a serem atingidos nestes ensaios: (a) estimar com precisão e prever a produtividade, baseado em um número reduzido de dados experimentais, (b) determinar a estabilidade dos rendimentos e o padrão de resposta dos genótipos ou dos procedimentos agronômicos nos diferentes ambientes e (c) permitir uma orientação segura na seleção dos melhores genótipos ou procedimentos agronômicos.

As variações na resposta dos genótipos ou dos procedimentos agronômicos nos diferentes ambientes são conhecidas como a interação destes fatores com o ambiente. Nos programas de melhoramento, a interação genótipos por ambientes ($G \times E$) é de extrema importância, pois possibilita a seleção de genótipos, bem como, a determinação do número ideal de ambientes e genótipos a serem avaliados em cada fase da seleção (FOX et al., 1997).

Lavoranti (2003) representa a interação por meio de uma tabela de dupla entrada (Tabela 2), com os genótipos nas linhas, e os ambientes, onde os ensaios foram instalados, nas colunas. As n observações fenotípicas, ou seja, o resultado dos efeitos genéticos e ambientais, aos quais o genótipo foi exposto durante o seu desenvolvimento, são representadas pela variável Y_{ij} , associada a G_i ($i = 1, 2, \dots, g$) genótipos e E_j ($j = 1, 2, \dots, e$) ambientes.

Gauch (1992) considera a organização de um ensaio de produtividade, como sendo g genótipos, alocados em e ambientes com r repetições. Assim, Y_{ij} , na Tabela 2, pode representar a média das r repetições, seguindo o modelo matemático:

$$\bar{Y}_{ij} = \mu + \hat{G}_i + \hat{E}_j + \left(\widehat{GE} \right)_{ij} + \bar{\epsilon}_{ij} \quad (12)$$

em que:

$\bar{Y}_{ij}$: média (r repetições) da observação fenotípica do genótipo i no ambiente j ;

μ : média geral;

$\hat{G}_i = \bar{Y}_{i.} - \bar{Y}_{..}$ (efeito do genótipo i);

$\hat{E}_j = \bar{Y}_{.j} - \bar{Y}_{..}$ (efeito do ambiente j);

$(\widehat{GE})_{ij} = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$ (efeito da interação $G \times E$);

$\bar{\epsilon}_{ij}$: é o erro médio experimental assumindo ser normal e identicamente distribuído $\left(0, \frac{\sigma^2}{n}\right)$ com σ^2 , a variância do erro dentro de ambiente, assumida constante.

Tabela 2 - Matriz de dados para a interação genótipos (linhas) e ambientes (colunas)

Genótipos	Ambientes					Médias
	1	2	3	...	e	
1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1e}	$\bar{Y}_{1.}$
2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2e}	$\bar{Y}_{2.}$
3	Y_{31}	Y_{32}	Y_{33}	...	Y_{3e}	$\bar{Y}_{3.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
g	Y_{g1}	Y_{g2}	Y_{g3}	...	Y_{ge}	$\bar{Y}_{g.}$
Médias	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{.3}$	...	$\bar{Y}_{.e}$	$\bar{Y}_{..}$

Tradicionalmente, a estimativa da interação $G \times E$ é feita, considerando uma análise de variância com grupos de experimentos, no entanto, a interação significativa é extremamente complexa, necessitando de uma exploração mais detalhada, inclusive de uma análise de estabilidade e adaptabilidade fenotípica, tanto por métodos uni como multivariados. Uma dessas metodologias, muito utilizada ultimamente, é o modelo estatístico *AMMI* (*Additive Main effects and Multiplicative Interaction*), cujo objetivo principal é selecionar modelos que expliquem o padrão relacionado à interação ($SQ_{G \times E}$ Padrão), descartando os ruídos ($SQ_{G \times E}$ Ruídos), presentes nos dados e sem interesse agrônômico (LAVORANTI, 2003).

O modelo *AMMI*, proposto inicialmente por Mandel (1971), combina, num único modelo, efeitos aditivos de genótipos e de ambientes, de maneira tradicional, e efeitos multiplicativos para a interação $G \times E$, pela análise de componentes principais.

Zobel; Wright e Gauch (1988) descrevem o modelo *AMMI* como:

$$Y_{ij} : \mu + g_i + e_j + \sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk} + \rho_{ij} + \varepsilon_{ij} \quad (13)$$

em que:

Y_{ij} : resposta média do i -ésimo genótipo no j -ésimo ambiente;

μ : média geral;

g_i : efeito do i -ésimo genótipo, ($i = 1, 2, \dots, g$);

e_j : efeito do j -ésimo ambiente, ($j = 1, 2, \dots, e$);

λ_k : raiz quadrada do k -ésimo autovalor das matrizes $(\mathbf{GE})(\mathbf{GE})^T$ e $(\mathbf{GE})^T(\mathbf{GE})$ de iguais autovalores não nulos (λ_k^2 é o k -ésimo autovalor; $[GE_{G \times E} = (\hat{g}e_{ij})]$ matriz de interações obtida como resíduo do ajuste aos efeitos principais, por ANAVA, aplicada à matriz de médias;

γ_{ik} : i -ésimo elemento (relacionado ao genótipo i) do k -ésimo autovetor de $(\mathbf{GE})(\mathbf{GE})^T$ associado a λ_k^2 ;

α_{jk} : j -ésimo elemento (relacionado ao ambiente j) do k -ésimo autovetor de $(\mathbf{GE})^T(\mathbf{GE})$ associado a λ_k^2 ;

ρ_{ij} : ruídos presentes nos dados;

ε_{ij} : erro experimental médio;

i : variações de genótipos;

j : variações de ambientes;

p : raízes características não nulas, $p = (1, 2, \dots, \min(g-1, e-1))$.

Sob as restrições de identificabilidade $\sum_{i=1}^g g_i = \sum_{j=1}^e e_j = \sum_{i=1}^g (ge)_{ij} =$

$\sum_{j=1}^e (ge)_{ij} = 0$, uma DVS da matriz de interação $\mathbf{GE}$ dá origem ao termo $\sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk} + \rho_{ij}$.

Na metodologia *AMMI*, o termo **GE** é representado pela soma de p parcelas. Cada parcela é resultante da multiplicação de λ_k , expresso na mesma unidade de Y_{ij} , por um efeito genotípico (γ_{ik}) e um efeito ambiental (α_{jk}), ambos adimensionais, ou seja, $\sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk}$ (n: termos da interação). O termo λ_k traz uma informação relativa à interação $G \times E$, na k -ésima parcela e os efeitos γ_{ik} e α_{jk} representam os pesos do genótipo i e do ambiente j , naquela parcela da interação λ_k^2 (LAVORANTI, 2003).

A definição do número de eixos a serem retidos, para explicar a estrutura da interação, é o menor possível (dois ou três no máximo). Duarte e Vencovsky (1999) mencionam que um dos procedimentos usuais adotados para a definição do número de eixos a serem retidos consiste em determinar os graus de liberdade associados à parcela da $SQ_{G \times E}$ relacionada a cada membro da família *AMMI*. Obtém-se, então, o quadrado médio (QM) correspondente a cada parcela (ou modelo). Em seguida, é obtido um teste F avaliando-se a significância de cada componente em relação ao $QM_{erro\ médio}$. Dessa forma, o ponto de parada que determina a seleção do modelo ($AMMI_0, AMMI_1, \dots, ou AMMI_n$) baseia-se na significância do teste F para os sucessivos eixos da interação. O resíduo *AMMI*, reunindo os eixos descartados da interação, também pode ser testado de maneira a assegurar o seu caráter desprezível.

Um sistema, muito utilizado, para atribuir os graus de liberdade ao modelo *AMMI*, é o sistema de Gollob (1968). Segundo Gauch (1992), citado por Duarte e Vencovsky (1999), a proposta de Gollob é fundamentada na definição clássica de graus de liberdade, porque, de acordo com o modelo *AMMI* (12), cada eixo de interação tem “um” valor singular λ_k , “ g ” elementos para o vetor γ_k e “ e ” elementos para o vetor α_k , o que totaliza $(g + e + 1)$ parâmetros a serem estimados. Entretanto, para estimação dos parâmetros, os autovetores associados ao γ_k e α_k ficam sujeitos às restrições $\sum_i \gamma_{ik} = 0$, $\sum_j \alpha_{jk} = 0$, $\sum_i \gamma_{ik}^2 = 1$, $\sum_j \alpha_{jk}^2 = 1$ (comprimento unitário), e para a ortogonalidade dos eixos quando, $n > 1$ tem-se mais $2(n-1)$ restrições $\left(\sum_i \gamma_{in} \gamma_{ik} = 0, \sum_j \alpha_{jn} \alpha_{jk} = 0 \right)$ para $k = 1, 2, \dots, n-1$. Assim, chega-se à proposta de Gollob, para os graus de liberdade dos eixos (GL_{IPCA_k}) dada por:

$$GL_{IPCA_k} = g + e + 1 - 2 - 2 - 2(k-1) = g + e - 1 - 2k$$

$$k = 1, 2, \dots, n-1 \quad e \quad n = \text{mínimo}(g, e).$$

O ponto de parada para seleção do modelo é o n -ésimo componente principal da interação (CPI) significativo pelo teste F , segundo Gollob (1968).

Lavoranti (2003) comenta sobre a utilização de outros testes para a validação do modelo $AMMI_k$, dentre eles, o teste F_R , proposto por Cornelius; Seyedsadr e Crossa (1992). É um teste robusto para a escolha do modelo $AMMI$, segundo Piepho (1995), o qual apresenta a estatística F_R como:

$$F_R = \frac{SQ_{G \times E} - \sum_{k=1}^n \lambda_k^2}{f_2 \cdot QM_{erro \text{ médio}}}$$

em que, $f_2 = (g - 1 - n)(e - 1 - n)$ com n o número de termos multiplicativos incluídos no modelo. A estatística F_R , sob a hipótese nula de que não haja mais do que n termos determinando a interação, tem uma distribuição F aproximada com f_2 e $GL_{erro \text{ médio}}$ graus de liberdade.

O ajuste do modelo $AMMI$ só é possível para dados balanceados, no entanto, muitas vezes na Tabela 2 ocorrem Y_{ij} ausentes, os quais, segundo Gauch (1992), surgem basicamente de três maneiras:

- (1) acidentes causando a perda não intencional dos dados, ou seja, a perda de material por danos climáticos, físicos ou fisiológicos e mesmo a ocorrência de erros no manuseio dos resultados.
- (2) decisões intencionais e bem justificadas podem gerar perda de dados, um exemplo, é a falta de material genotípico para a locação em todos os ambientes ou a eliminação proposital de genótipos em diferentes anos de condução do ensaio.
- (3) retirada artificial de dados disponíveis, pelo menos temporariamente, para executar algum processo de validação ou método de imputação, nos quais a precisão dos valores imputados pode ser determinada, comparando-os com os dados originais.

O modelo $AMMI$ requer dados para todos os genótipos combinados com todos os ambientes, pois utiliza a técnica da decomposição por valores singulares (DVS). Assim, se ocorrer a ausência de algum Y_{ij} , é inviável o seu emprego. Gauch e Zobel (1990) desenvolveram

um modelo *AMMI* implementado com o algoritmo *EM* (*Expectation-Maximization*), representado por “*EM-AMMI*”, para suprir a ausência de dados, no entanto este modelo consiste em um método paramétrico e em uma imputação simples. Dear (1959) introduz um método baseado em componentes principais, o qual não requer suposições sobre a distribuição dos dados, como também Godfrey et al. (2002), com um método chamado agrupamento em dois estágios (*two-stage clustering*), o qual divide a distância Euclidiana quadrada em dois componentes independentes, a interação $G \times E$ e o efeito principal do genótipo, mas ambos são métodos de imputação simples.

Nos métodos de IM paramétricos há, normalmente, fortes suposições sobre a distribuição dos dados. Caso tais suposições não sejam atendidas, os M conjuntos de dados completados não são apropriados, produzindo estimadores inconsistentes, os quais levam a resultados enganosos. Alguns métodos semi e não-paramétricos para a IM foram desenvolvidos, entre eles, o proposto por Little e Rubin (1987, 2002), com o Bootstrap Bayesiano Aproximado (*Approximate Bootstrap Bayesian- ABB*); Lipsitz, Zhao e Molenberghs (1998) com uma imputação múltipla semiparamétrica; Aerts et al (2002) com um método de imputação múltipla local (semi e não-paramétrica) e Paddock, (2002) com um método completamente não-paramétrico, baseado nas árvores de Polya (*Polya's tree*).

No entanto, há uma carência quanto a um método de IM, voltado diretamente para a matriz de interação $G \times E$ (variável resposta) e, conseqüentemente, ser utilizado com a metodologia *AMMI*. Assim, partindo da DVS de uma matriz, Krzanowski (1988) desenvolveu um método de imputação simples baseado na maior quantidade possível de informação dos dados, porém, uma alteração neste método pode gerar vários valores para Y_{ij} ausente, transformando-o em um método de IM.

2.2 Metodologia

2.2.1 Material

Os dados utilizados neste trabalho foram obtidos de experimentos conduzidos em sete ambientes, nas regiões sul e sudeste do Brasil (Tabela 3), para 20 progênes de *Eucalyptus grandis* provenientes da Austrália (12 Km South of Ravenshoe-Mt Pandanus-QLD, lote 14.420). O delineamento utilizado foi aleatorizado em blocos, com 6 plantas por parcela e 10 repetições, no espaçamento 3,0 m por 2,0 m (LAVORANTI, 2003).

Tabela 3 - Localização do teste de progênes de *Eucalyptus grandis* em sete municípios brasileiros

Ambientes	Municípios/Estado	Regiões	Latitude (S)	Longitude (W)	Altitude (m)
1	Barra Ribeiro - RS	Sul	30°20'	51°14'	30
2	Telêmaco Borba - PR	Sul	24°15'	20°29'	850
3	Boa Esperança de Sul - SP	Sudeste	21°57'	48°32'	540
4	Guanhães - MG	Sudeste	18°40'	42°60'	900
5	Ipatinga - MG	Sudeste	19°15'	42°20'	250
6	Aracruz - ES	Sudeste	19°48'	40°17'	50
7	Caçapava - SP	Sudeste	23°03'	45°46'	650

Na Tabela 4, cada valor Y_{ij} representa a altura (m) média (10 blocos), das médias (6 plantas por parcela) de cada genótipo ($i = 1, 2, \dots, 20$) de *Eucalyptus grandis*, instalada nos diferentes ambientes ($j = 1, 2, \dots, 7$).

2.2.2 Método

Neste trabalho, a ausência de genótipos nos ambientes seguiu o padrão de ausência arbitrário, pois foram retirados, aleatoriamente, valores da matriz de interação (Tabela 4), numa proporção de 5% (Anexo A), 10% e 30%. O sistema estatístico SAS, por meio do SAS/IML (2004) e SAS/STAT (2004), foi utilizado no desenvolvimento de programas e na análise dos dados.

Tabela 4 - Média de altura, em metros, dos genótipos de *E. grandis* aos 5 anos nos diferentes ambientes

Genótipo	Ambientes						
	1	2	3	4	5	6	7
1	17,40	25,00	18,67	20,61	13,88	19,84	14,72
2	17,58	24,00	17,69	20,44	13,10	17,77	14,44
3	16,52	23,70	15,94	18,91	12,86	17,93	13,43
4	16,78	22,68	16,28	16,55	11,92	17,70	13,03
5	15,36	21,56	15,95	16,57	10,97	18,06	13,03
6	15,76	22,34	16,61	19,08	11,91	19,22	13,17
7	14,57	20,35	17,02	15,06	12,47	17,01	11,14
8	18,46	24,52	17,87	18,16	13,66	19,72	14,37
9	16,87	20,77	16,84	18,96	13,54	18,69	14,06
10	16,92	22,48	17,16	18,94	13,71	19,38	14,98
11	16,50	22,98	16,87	17,17	13,55	19,12	13,13
12	17,71	23,61	16,88	18,19	12,62	18,17	13,52
13	17,62	22,65	16,01	18,78	11,92	17,88	13,24
14	16,29	23,59	18,79	20,20	13,29	19,31	13,60
15	15,94	23,36	18,10	17,91	12,75	19,71	12,49
16	16,46	22,59	17,80	19,08	12,80	19,44	13,78
17	17,74	23,38	16,17	18,91	14,42	20,24	13,32
18	16,90	21,71	16,95	18,56	12,50	18,82	14,33
19	16,28	20,12	14,66	15,68	10,60	16,10	11,98
20	15,78	22,15	15,36	16,46	12,51	16,89	12,07
Média	16,67	22,68	16,88	18,21	12,75	18,55	13,39

Para a suposição livre de distribuição na variável resposta, os valores imputados foram obtidos por meio de uma mudança no procedimento de imputação simples desenvolvido por Krzanowski (1988). Este método parte, inicialmente, da afirmação feita por Good (1969),

na qual qualquer matriz $\mathbf{Y}_{(n,p)}$ pode ser decomposta por valor singular na forma

$$\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \quad (14)$$

em que $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_p$ e $\mathbf{D} = \text{diag}(d_1, \dots, d_p)$ com $d_1 \geq d_2 \geq \dots \geq d_p \geq 0$. As matrizes $\mathbf{Y}^T\mathbf{Y}$ e $\mathbf{Y}\mathbf{Y}^T$ têm os mesmos autovalores não nulos, e os elementos d_i são a raiz quadrada destes autovalores; a i -ésima coluna $\mathbf{v}_i = (v_{i1}, \dots, v_{ip})$ da matriz $\mathbf{V}_{p \times p}$ é o autovetor correspondente ao i -ésimo maior autovalor d_i^2 de $\mathbf{Y}^T\mathbf{Y}$; enquanto a j -ésima coluna $\mathbf{u}_j = (u_{1j}, \dots, u_{nj})^T$ da matriz $\mathbf{U}_{n \times p}$ é o autovetor correspondente ao j -ésimo maior autovalor d_j^2 de $\mathbf{Y}\mathbf{Y}^T$. A decomposição (14) tem sua representação elementar como

$$y_{ij} = \sum_{h=1}^p u_{ih} d_h v_{jh}. \quad (15)$$

Krzanowski (1987) usou esta representação como uma base para determinar a dimensionalidade de um conjunto de dados multivariados. Se a estrutura dos dados é essencialmente H -dimensional ($H < p$) então a variação na dimensão resultante ($p - H$) pode ser tratada como ruído aleatório. As características principais dos dados estarão supostamente no espaço dos H primeiros componentes principais. A correspondência entre as quantidades do lado direito de (15) e os eixos principais da configuração dos dados sugere o modelo de H -componentes

$$y_{ij} = \sum_{h=1}^H u_{ih} d_h v_{jh} + \epsilon_{ij}, \quad (16)$$

em que ϵ_{ij} é o ruído.

Supondo o modelo (16) para um valor específico de H , com uma única observação y_{ij} ausente na matriz de dados, tem-se y_{ij} estimado por

$$\hat{y}_{ij}^{(H)} = \sum_{h=1}^H u_{ih} d_h v_{jh}, \quad (17)$$

em que u_{ih}, d_h, v_{jh} , devem ser estimados com o restante dos dados. As melhores estimativas destes valores estão baseadas na maior quantidade possível de dados. Simbolizado, por $\mathbf{Y}^{(-i)}$ a matriz dos dados obtida, retirando-se a i -ésima linha de $\mathbf{Y}$, e por $\mathbf{Y}_{(-j)}$ a matriz dos dados obtida, retirando-se a j -ésima coluna de $\mathbf{Y}$, a decomposição de valor singular dessas matrizes fica

$$\mathbf{Y}^{(-i)} = \bar{\mathbf{U}} \bar{\mathbf{D}} \bar{\mathbf{V}}^T, \quad \bar{\mathbf{U}} = (\bar{u}_{sh}), \quad \bar{\mathbf{V}} = (\bar{v}_{sh}), \quad \bar{\mathbf{D}} = (\bar{d}_1, \dots, \bar{d}_p), \quad (18)$$

e

$$\mathbf{Y}_{(-j)} = \tilde{\mathbf{U}} \tilde{\mathbf{D}} \tilde{\mathbf{V}}^T, \quad \tilde{\mathbf{U}} = (\tilde{u}_{sh}), \quad \tilde{\mathbf{V}} = (\tilde{v}_{sh}), \quad \tilde{\mathbf{D}} = (\tilde{d}_1, \dots, \tilde{d}_{p-1}). \quad (19)$$

A estimativa de u_{ih} e v_{jh} em (17), obtida com o máximo dos dados de $\mathbf{Y}$, é $\tilde{u}_{ih}$ e $\tilde{v}_{jh}$, respectivamente, enquanto d_h pode ser estimado por $\bar{d}_h$, $\tilde{d}_h$ ou por alguma combinação dos dois. Uma forma adequada parece ser $\sqrt{\bar{d}_h} \sqrt{\tilde{d}_h}$, em que uma estimativa do valor ausente y_{ij} é dada por

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

Seguindo o preceito da máxima informação dos dados, usa-se o valor mais elevado disponível de H . De (19), este valor é, evidentemente, $p - 1$, então o valor imputado a y_{ij} será

$$\hat{y}_{ij} = \sum_{h=1}^{p-1} (\tilde{u}_{ih} \sqrt{\tilde{d}_h}) (\tilde{v}_{jh} \sqrt{\tilde{d}_h}). \quad (20)$$

As estimativas iniciais dos valores y_{ij} ausentes são feitas pela média $\bar{y}_j$ da j -ésima coluna. Para evitar qualquer influência de possíveis variações entre as colunas, por exemplo, a escala das variáveis, é recomendado aplicar uma padronização em $\mathbf{Y}$. Para os valores y_{ij} , inclusive os ausentes já substituídos pela média ($\bar{y}_j$), é calculada uma nova média ($\bar{y}'_j$) e um desvio padrão (dp_j) para cada coluna j , então y_{ij} é padronizado por $y'_{ij} = \frac{(y_{ij} - \bar{y}_j)}{dp_j}$. Padronização semelhante também é feita nas matrizes $\mathbf{Y}^{(-i)}$ e $\mathbf{Y}_{(-j)}$.

As estimativas de cada valor ausente são recalculadas usando-se (20) nas matrizes padronizadas. Para cada estimativa são necessárias duas decomposições de valores singulares, isto é, uma para cada i e j necessários. O processo iterativo continua até ser alcançada a estabilidade nos valores imputados. Finalmente, à matriz $\mathbf{Y}$ completada (*observados + imputados*) é aplicada uma operação para retorno à sua escala original, ou seja, se $y_{ij}^{(c)}$ representa cada valor da matriz $\mathbf{Y}$ completada, calcula-se novamente a média da coluna j ($\bar{y}_j^{(c)}$) e o seu desvio padrão ($s_j^{(c)}$). Cada valor da matriz $\mathbf{Y}$ completada, na escala original, é então obtido por, $y_{ij} = \bar{y}_j^{(c)} + s_j^{(c)} y_{ij}^{(c)}$.

A modificação proposta neste método, para gerar as imputações ($m = 1, \dots, M$) na primeira etapa da IM, consiste em uma mudança nos expoentes dos radicandos $\tilde{d}_h$ e $\tilde{d}_h$ em (20), ou seja, de uma maneira genérica, se $\sqrt[b]{d^a}$ for representada como uma potência fracionária $d^{\frac{a}{b}}$, o procedimento requer a mudança no numerador do expoente, tanto de $\tilde{d}_h^{\frac{a}{b}}$

como de $\bar{d}_h^{\frac{\bar{a}}{b}}$, de modo que a soma dos expoentes seja igual a 1 ($\frac{\bar{a}+\tilde{a}}{b} = 1$). Krzanowski (1988) sugere como estimativas para d_h em (17) uma combinação entre $\bar{d}_h$ de (18) e $\tilde{d}_h$ de (19), resultando na forma $\sqrt{\bar{d}_h}\sqrt{\tilde{d}_h}$, a qual admite influências iguais de (18) e (19). Assim, variando os expoentes de $\bar{d}_h$ e $\tilde{d}_h$, admite-se um peso maior para (18) ou (19) na estimativa final de y_{ij} em (20).

Cada mudança em $\tilde{a}$ e, conseqüentemente em $\bar{a}$, gera uma nova matriz $\mathbf{Y}$ completada, caracterizando, assim, um processo de geração dos M conjuntos de dados completados da primeira etapa da IM.

O número de imputações fica condicionado às mudanças nos expoentes e, segundo Molenberghs e Verbeke (2005); Rubin (1987); Schafer (1999), um número M de imputações entre 3 e 5 é suficiente para expressar a variabilidade entre imputações. Assim, com um número de 5 mudanças nos expoentes, além de estar em concordância com os referidos autores, há uma variação entre 40% e 60% nos pesos dados a (18) e (19), ou seja, partindo de um denominador fixo ($b = 20$, por exemplo), os valores assumidos por $\tilde{a}$ (8, 9, 10, 11 e 12) e respectivamente por $\bar{a}$ (12, 11, 10, 9 e 8) levam a uma variação (40%, 45%, 50%, 55% e 60%) nas proporções de (18) e (19) em

$$\hat{y}_{ij} = \sum_{h=1}^{p-1} (\tilde{u}_{ih} \bar{d}_h^{\frac{\tilde{a}}{b}}) (\bar{v}_{jh} \bar{d}_h^{\frac{\bar{a}}{b}}). \quad (21)$$

A metodologia acima descrita, e aqui proposta, utiliza a maior quantidade possível de dados de $\mathbf{Y}$ e independe de qualquer distribuição na variável resposta, podendo ser aplicada a qualquer matriz de dados numérica.

A implementação do método foi feita por meio de um programa desenvolvido no módulo IML do sistema estatístico SAS (Anexo B), o qual, após a sua execução, resultou em um arquivo de dados com os $M = 5$ conjuntos de dados completados e pronto para ser utilizado na segunda etapa da IM.

A análise dos $M = 5$ conjuntos de dados completados, individualmente, na segunda etapa da IM, utilizou o Proc UNIVARIATE do SAS (Anexo C), pois a finalidade foi obter a média de alturas de cada ambiente e seu erro padrão, as quais foram utilizadas na terceira etapa da IM, para uma comparação com as médias originais de alturas dos ambientes, apresentadas na última linha da Tabela 4.

O interesse na terceira etapa da IM foi combinar as $M = 5$ estimativas da média de alturas de cada ambiente ($\hat{\beta}^m$) em uma única ($\hat{\beta}^*$), por meio de uma média (3) e obter uma medida de sua variabilidade (6). Ambas foram utilizadas no cálculo da estatística *t*-Student, para testar as hipóteses de que a média ($\hat{\beta}^*$), resultante das $M = 5$ imputações, é igual à média original de alturas em cada ambiente. Esta etapa utilizou o Proc MIANALYZE do SAS (Anexo D).

Uma aplicação aos valores obtidos nas $M = 5$ imputações foi feita por meio do ajuste de um modelo *AMMI* para a decomposição da interação $G \times E$, ou seja, uma comparação na seleção do modelo *AMMI* ($AMMI_0, AMMI_1, \dots$, ou $AMMI_n$) nos dados sem a retirada aleatória de valores (dados originais) com a seleção do modelo *AMMI* proveniente de uma média das $M = 5$ imputações. Um programa desenvolvido no módulo IML do SAS (Anexo E), semelhante ao descrito por Duarte e Vencovsky (1999), foi utilizado na determinação do ponto de parada para a seleção do modelo de cada membro da família de modelos *AMMI* ($AMMI_0, AMMI_1, \dots$ ou $AMMI_n$), com base na significância do teste F , aplicado de acordo com o critério de Gollob (1968) e do teste F_R proposto por Cornelius; Seyedsadr e Crossa (1992) e Piepho (1995), para os sucessivos termos da interação $G \times E$.

Todos os procedimentos executados anteriormente também foram utilizados, considerando as ausências de 10% e 30% nos dados da Tabela 4. Aplicou-se a metodologia proposta (21) e a estabilidade nos valores imputados, ou seja, a convergência destes, para um único valor, foi alcançada com 20 iterações, mas, por segurança e rapidez na execução do programa, foram feitas 50 iterações. Para cada porcentagem de ausência, foram construídos gráficos de dispersão, considerando as mesmas posições de retirada dos dados com 5% de ausência, no i -ésimo genótipo do j -ésimo ambiente ($i;j$), uma vez que as retiradas aleatórias (Anexo A) possuíram um mesmo valor inicial como semente.

Uma medida de exatidão ou acurácia, para as diferentes porcentagens de ausência, levou em consideração a expressão

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

adaptada de Penny e Jolliffe (1999), em que M é o número de imputações, VO é o valor original retirado aleatoriamente na posição $l = 1, 2, \dots, na$, em que na representa o número

total de valores retirados correspondentes à i -ésima linha e j -ésima coluna $(i;j)$ da Tabela 4, e $\hat{y}_{ij}$ seu respectivo valor imputado (21) pelo método proposto. Esta expressão pode ser separada em duas partes,

$$acc_l = \frac{\sum_{m=1}^M (\hat{y}_{ij(m)} - \bar{Y}_l)^2}{M-1} + \frac{M(\bar{Y}_l - VO_l)^2}{M-1}, \quad (22)$$

em que $\bar{Y}_l$ é a média das imputações na posição l , a primeira parte representa uma variância e a segunda um viés, dos M valores imputados em cada posição.

Uma medida geral da acurácia T_{acc} pode ser calculada por meio da média das acc_l , ou seja,

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

em que, $na = g \times e \times porc$, com g o total de genótipos, e o total de ambientes e $porc$ a porcentagem de ausência dos dados (ANEXO F).

T_{acc} possui dois componentes,

$$T_{acc} = V_E + VQM, \quad (23)$$

em que

$$V_E = \frac{1}{na} \sum_{l=1}^{na} \left[\frac{\sum_{m=1}^M (\hat{y}_{ij(m)} - \bar{Y}_l)^2}{M-1} \right] \quad (24)$$

e

$$VQM = \frac{1}{na} \sum_{l=1}^{na} \frac{M(\bar{Y}_l - VO_l)^2}{M-1} \quad (25)$$

O primeiro representa a variância entre imputações (V_E), portanto, quanto maior o seu valor, menor é a precisão do método de imputação múltipla, mas um valor reduzido desta variância também não significa um bom método de imputação, pois o método pode ser tendencioso. O segundo componente representa o viés quadrático médio (VQM) entre os valores de $\bar{Y}$ e VO , quanto menor for o viés, mais as imputações assemelham-se aos valores originais. Assim, melhor será o método de imputação múltipla quanto menores forem V_E e VQM .

2.3 Resultados e discussão

Inicialmente foram retirados aleatoriamente 5% dos valores da matriz de interação $G \times E$, ou seja, 7 valores do conjunto de dados da Tabela 4. Neste sorteio, não ocorreram valores ausentes para os ambientes 1, 3 e 5 (Apêndice A), e o ambiente 4 concentrou aproximadamente 43% das ausências.

Tabela 5 - Média de alturas, em metros, das imputações, segundo a posição (linha i , coluna j) de retirada aleatória (5%) da Tabela 4

Posição		Imputação (m)					$\bar{Y}$	Var	CV	acc
retirada	VO	1	2	3	4	5				
(2;2)	24,00	22,37	22,35	22,35	22,35	22,35	22,35	0,0001	0,04	3,3802
(19;2)	20,12	21,37	20,93	20,93	20,93	20,93	21,02	0,0387	0,94	1,0511
(10;4)	18,94	21,06	21,05	21,03	21,02	21,01	21,03	0,0004	0,10	5,4690
(13;4)	18,78	16,90	17,08	17,09	17,10	17,11	17,06	0,0077	0,52	3,7243
(19;4)	15,68	14,58	14,97	14,98	15,00	15,01	14,91	0,0339	1,23	0,7763
(5;6)	18,06	17,50	17,58	17,57	17,55	17,54	17,55	0,0010	0,18	0,3272
(4;7)	13,03	12,82	12,80	12,80	12,80	12,81	12,81	0,0001	0,07	0,0635

VO : Valor original.

$\bar{Y}$: Média das imputações.

Var : Variância entre imputações.

CV : Coeficiente de variação.

acc : Acurácia.

Nas posição (10;4) e (19;2) as imputações ($m = 1, \dots, 5$) foram maiores do que o valor original, mas nas demais posições foram menores (Figura 3). A variabilidade dos valores imputados, em relação à média das imputações ($\bar{Y}$), em cada posição, foi pequena, expressa por variâncias próximas de zero, como pode ser constatado na Tabela 5 e Figura 4. Quanto à exatidão, ou seja, a diferença dos valores imputados em relação aos valores originais, foi maior do que a variabilidade, representada pela maior parte na composição da medida acc (22). Assim, o método proposto, com 5% de ausência dos dados, oscilou para mais ou para menos na estimativa dos valores ausentes, em relação aos valores originais (VO), no entanto a dispersão, em relação às médias das imputações ($\bar{Y}$), foi menor do que a diferença dos valores imputados em relação aos valores originais (VO).

Figura 3 - Imputações e valores originais (VO) de alturas nas posições de retirada dos dados com 5% de ausência

Figura 4 - Média, erro padrão e desvio padrão de alturas (m) para as imputações em cada posição de retirada dos dados com 5% de ausência

Na aplicação do Proc UNIVARIATE do SAS (Anexo C), segunda etapa da IM, como os ambientes 1, 3 e 5 não tiveram valores ausentes, suas médias e erros padrão foram iguais em cada conjunto de dados completados, não sendo incluídos no arquivo final do programa (*outuni*), com as médias de ambientes e os seus respectivos erros padrão para cada conjunto completado pelas imputações ($M = 5$), aqui representados na Tabela 6.

Tabela 6 - Média e erro padrão das médias de alturas (m) dos ambientes completados pelas imputações, nos dados com 5% de ausência

Imputação	Ambiente			
	2	4	6	7
1	22,6586 (0,2618)	18,1680 (0,3987)	18,5221 (0,2549)	13,3808 (0,2183)
2	22,6349 (0,2685)	18,1954 (0,3884)	18,5261 (0,2540)	13,3799 (0,2184)
3	22,6350 (0,2685)	18,1962 (0,3877)	18,5254 (0,2542)	13,3800 (0,2184)
4	22,6350 (0,2685)	18,1969 (0,3870)	18,5247 (0,2543)	13,3802 (0,2184)
5	22,6351 (0,2685)	18,1977 (0,3863)	18,5240 (0,2544)	13,3803 (0,2184)

Na terceira etapa da IM, por meio do Proc MIANALYZE do SAS (Anexo D), combinaram-se as estimativas das médias de alturas dos 4 ambientes com valores ausentes (2, 4, 6 e 7), provenientes dos 5 conjuntos de dados completados pelas imputações.

A Tabela 7 resume os resultados do programa, o qual possui o comando *edf*, necessário ao ajuste do número de graus de liberdade (7) por meio da expressão (8). Este comando corresponde ao número de graus de liberdade para os dados completados ν_{com} . Este valor foi utilizado no teste de hipóteses entre as médias de alturas dos ambientes provenientes dos dados completados e as médias dos dados originais, sem a ausência de dados, baseados na distribuição *t*-Student. Assim, $edf = 19$, pois as médias de alturas dos ambientes provêm de 20 genótipos.

Tabela 7 - Estimativa média ($\hat{\beta}^*$) das médias de alturas e medidas associadas a sua variabilidade, nos ambientes com valores imputados para dados com 5% de ausência. Teste *t*-Student para comparação com a média original dos ambientes

Estimativas	Ambientes			
	2	4	6	7
$\hat{\beta}^*$	22,63973	18,19082	18,52445	13,38023
B	0,00011	0,00016	$2,3 \times 10^{-6}$	$1,3 \times 10^{-7}$
$\bar{W}$	0,07138	0,15181	0,06470	0,04769
T	0,07151	0,15201	0,06471	0,04770
r	0,00188	0,00129	0,00004	$3,4 \times 10^{-6}$
λ	0,00188	0,00129	0,00004	$3,4 \times 10^{-6}$
ER	0,99963	0,99974	0,99999	0,99999
ν^*	17,2400	17,2500	17,2720	17,2730
MA	22,6770	18,2110	18,5500	13,3915
t	0,14 ^{ns}	0,05 ^{ns}	0,10 ^{ns}	0,05 ^{ns}
valor- p	0,8908	0,9593	0,9212	0,9594

ER : Eficiência relativa.

MA : Média original dos ambientes.

ns : não significativo, $p \geq 0,05$.

Outro comando disponível no Proc MIANALYZE é o *mu0*, o qual especifica o

valor de β_0 , sob a hipótese nula $\beta = \beta_0$ no teste t -Student. No presente caso, as médias de ambientes (MA) dos dados originais, Tabela 4, foram atribuídas aos valores de β_0 .

Não houve diferença entre as médias comparadas pelo teste t -Student com o menor valor- p igual a 0,8908, portanto cada estimativa média ($\hat{\beta}^*$) das médias de alturas dos ambientes nos dados completados, por meio do método proposto, foi semelhante à respectiva média de alturas nos dados originais (MA).

As estimativas das médias dos ambientes ($\hat{\beta}^*$) possuíram uma pequena variabilidade, sendo que a maior delas ocorreu no ambiente 4 ($T = 0,15201$). O aumento relativo na variância, devido à ausência de dados (r), bem como a precisão da estimativa se nenhum dado estivesse ausente (λ), foram valores próximos de zero, indicando uma pequena influência da ausência de dados nestas estimativas. Quanto à eficiência relativa (ER), o número de imputações utilizadas ($M = 5$) nas estimativas das médias dos ambientes foi 99,9% de uma estimativa feita com infinitas imputações.

Uma aplicação dos dados imputados pelo método proposto, considerando a média das 5 imputações, foi feita na determinação do efeito multiplicativo da interação $G \times E$, por meio da análise de componentes principais (ACP), pela decomposição da soma de quadrados da interação $G \times E$ em seis eixos ou componentes principais desta interação (IPCA). Na Tabela 8, os resultados dos testes F , segundo o critério de Gollob (1968), indicaram significância até o terceiro eixo, tanto nos dados originais (orig.) como naqueles provenientes da média das 5 imputações (imp.), o que caracterizou uma semelhança entre os dois conjuntos de dados, indicando a seleção do modelo $AMMI_3$. No entanto, com os dados provenientes das imputações, uma porção maior de $SQ_{G \times E}$ foi explicada pelos dois primeiros eixos (63,00%).

Um programa desenvolvido no módulo IML do SAS (Anexo E) foi utilizado tanto nesta análise, como no teste F_R , apresentado por Cornelius; Seyedsadr e Crossa (1992) e Piepho (1995) para a seleção do modelo $AMMI$. Na Tabela 8, os dados originais apresentaram os dois primeiros eixos ($IPCA_1$ e $IPCA_2$) significativos na análise da interação $G \times E$ pela ACP, ou seja, $p < 0,01$ para o primeiro eixo e $p < 0,05$ para o segundo eixo, indicando a seleção do modelo $AMMI_3$, uma vez que, somente a partir de $IPCA_3$, o resíduo $AMMI$ se tornou não significativo, enquanto nos dados imputados, apenas o primeiro eixo foi significativo ($p < 0,05$), indicando a seleção do modelo $AMMI_2$.

Tabela 8 - Autovalor (λ_k^2), porcentagem da soma de quadrados acumulada (PA) nos eixos singulares (ES), desdobramento da interação G×E, teste F , segundo Gollob e F_R , segundo Cornelius para os dados de alturas originais (orig.) e a média (imp.) das imputações, com 5% de ausência

ES	λ_k^2		PA (%)		FV	GL	Gollob <i>orig.</i>		Gollob <i>imp.</i>		GL	Cornelius <i>orig.</i>		Cornelius <i>imp.</i>	
	orig.	imp.	orig.	imp.			QM	F	QM	F		QM	F_R	QM	F_R
1	18,902	27,432	29,03	38,74	IPCA ₁	24	0,788	2,53**	1,143	3,38**	90	0,511	1,65**	0,482	1,42*
2	17,261	17,178	55,53	63,00	IPCA ₂	22	0,785	2,52**	0,781	2,31**	68	0,426	1,37*	0,385	1,14 ^{ns}
3	13,385	11,442	76,09	79,16	IPCA ₃	20	0,669	2,15**	0,572	1,69*	48	0,324	1,04 ^{ns}	0,307	0,91 ^{ns}
4	7,394	8,054	87,44	90,54	IPCA ₄	18	0,411	1,32 ^{ns}	0,447	1,32 ^{ns}	30	0,273	0,88 ^{ns}	0,223	0,66 ^{ns}
5	5,694	5,137	96,19	97,79	IPCA ₅	16	0,356	1,14 ^{ns}	0,321	0,95 ^{ns}	14	0,178	0,57 ^{ns}	0,112	0,33 ^{ns}
6	2,484	1,563	100,00	100,00	IPCA ₆	14	0,178	0,57 ^{ns}	0,112	0,33 ^{ns}					
					EM	216	0,311		0,338		216	0,311		0,338	

FV : Fontes de variação.

GL : Número de graus de liberdade.

QM : Quadrado médio.

EM : Erro médio.

** : significativo ao nível de probabilidade, $p < 0,01$.

* : significativo ao nível de probabilidade, $p < 0,05$.

ns : não significativo.

Na retirada aleatória de 14 valores do conjunto de dados da Tabela 4, ou seja, 10% dos valores (Apêndice B), somente o ambiente 1 não apresentou valor ausente. O ambiente 4 concentrou 50% das ausências, pois com o aumento nos dados ausentes, de 5% para 10%, as ausências neste ambiente também aumentaram.

Tabela 9 - Média de alturas, em metros, das imputações, segundo a posição (linha i , coluna j) de retirada aleatória (10%) da Tabela 4

Posição		Imputação (m)					$\bar{Y}$	Var	CV	acc
retirada	VO	1	2	3	4	5				
(2;2)	24,00	23,30	25,62	25,13	25,65	25,64	25,07	1,0263	4,04	2,4541
(19;2)	20,12	21,53	21,33	21,26	21,40	21,33	21,37	0,0101	0,48	1,9251
(6;3)	16,61	17,60	17,59	17,51	17,53	17,53	17,55	0,0017	0,23	1,1164
(1;4)	20,61	19,79	20,09	20,03	20,10	20,10	20,02	0,0176	0,66	0,4484
(3;4)	18,91	19,17	19,68	19,37	19,69	19,68	19,52	0,0566	1,22	0,5200
(6;4)	19,08	17,68	17,78	17,71	17,72	17,72	17,72	0,0013	0,21	2,3002
(10;4)	18,94	18,85	18,92	18,90	19,06	19,06	18,96	0,0097	0,51	0,0101
(13;4)	18,78	16,88	18,71	18,67	18,74	18,73	18,35	0,6723	4,47	0,9078
(19;4)	15,68	14,82	15,53	15,55	15,68	15,68	15,45	0,1306	2,33	0,1965
(20;4)	16,46	16,54	17,43	17,18	17,45	17,44	17,21	0,1530	2,27	0,8484
(6;5)	11,91	13,15	12,88	12,95	12,89	12,89	12,95	0,0124	0,88	1,3679
(5;6)	18,06	16,78	16,03	16,92	16,89	16,89	16,70	0,1425	2,27	2,4447
(4;7)	13,03	12,92	13,52	13,56	13,46	13,45	13,38	0,0703	1,96	0,2258
(7;7)	11,14	13,17	13,43	12,79	13,41	13,40	13,24	0,0750	2,06	5,5784

VO : Valor original.

$\bar{Y}$: Média das imputações.

Var : Variância entre imputações.

CV : Coeficiente de variação.

acc : Acurácia.

Os valores imputados continuaram oscilando para mais ou para menos em relação aos valores originais, com uma variabilidade pequena, ou seja, as variâncias continuaram próximas de zero (Tabela 9 e Figura 6), no entanto as posições (2;2), (13;4) e (4;7) apresentaram a primeira imputação menor, indicando que um peso maior para (18) em (21)

influenciou a primeira imputação. Algumas posições inverteram as estimativas dos dados ausentes, ou seja, as posições (2;2), (13;4) e (4;7) apresentaram valores estimados das $M = 5$ imputações maiores em relação às imputações com 5% de ausência e as posições (10;4) e (5;6) valores estimados menores, já as posições (19;2) e (19;4) mantiveram as estimativas semelhantes àquelas com 5% de ausência (Figuras 3 e 5).

Figura 5 - Média, erro padrão e desvio padrão de alturas (m) para as imputações com 10% de ausência e nas mesmas posições de retirada dos dados com 5% de ausência

O aumento ou a única diminuição na posição (19;2) dos coeficientes de variação (CV) nas mesmas posições de 5% de ausência, Tabelas 5 e 9, demonstraram um aumento na variabilidade relativa do método proposto, quando ocorreu o aumento de 5% para 10% na ausência dos dados. A acurácia apresentou, para as estimativas das imputações nas posições (10;4), (13;4) e (19;4), ou seja, no ambiente 4 e na posição (2;2), uma diminuição, quando comparada com os 5% de ausência; já nas posições (19;2), (5;6) e (4;7) ocorreu um aumento. No entanto, uma maior parte na composição da medida acc (22) continuou sendo devido à diferença entre os valores imputados e os valores originais (VO).

As médias de ambientes em cada conjunto completado pelas imputações ($M = 5$) e os seus respectivos erros padrão, mostrados na Tabela 10, não incluindo o ambiente 1,

Figura 6 - Imputações e valores originais (VO) de alturas com 10% de ausência e nas mesmas posições de retirada dos dados com 5% de ausência

Tabela 10 - Média e erro padrão das médias de alturas (m) dos ambientes completados pelas imputações, nos dados com 10% de ausência

Imputação	Ambientes					
	2	3	4	5	6	7
1	22,7125 (0,2613)	16,9306 (0,2407)	17,9740 (0,3498)	12,8107 (0,2114)	18,4860 (0,2648)	13,4871 (0,1836)
2	22,8184 (0,3007)	16,9302 (0,2406)	18,1950 (0,3319)	12,7975 (0,2107)	18,4487 (0,2797)	13,5305 (0,1802)
3	22,7905 (0,2902)	16,9262 (0,2400)	18,1586 (0,3288)	12,8009 (0,2108)	18,4931 (0,2625)	13,5004 (0,1840)
4	22,8207 (0,3014)	16,9271 (0,2402)	18,2102 (0,3301)	12,7980 (0,2107)	18,4914 (0,2631)	13,5264 (0,1803)
5	22,8196 (0,3011)	16,9270 (0,2401)	18,2088 (0,3210)	12,7982 (0,2107)	18,4915 (0,2630)	13,5259 (0,1803)

foram os resultados finais de um arquivo (*outuni*) gerado do Proc UNIVARIATE (Anexo C) para 10% de ausência, na segunda etapa da IM.

Tabela 11 - Estimativa média ($\hat{\beta}^*$) das médias de alturas e medidas associadas a sua variabilidade, nos ambientes com valores imputados para dados com 10% de ausência. Teste *t*-Student para comparação com as médias originais

Estimativas	Ambientes					
	2	3	4	5	6	7
$\hat{\beta}^*$	22,79233	16,92822	18,14929	12,80107	18,48214	13,51405
B	0,00215	$4,2 \times 10^{-6}$	0,01004	0,00003	0,00036	0,00037
$\bar{W}$	0,08489	0,05775	0,11169	0,04445	0,07112	0,03302
T	0,08747	0,05776	0,12374	0,04449	0,07155	0,03346
r	0,03042	0,00001	0,10781	0,00084	0,00601	0,01344
λ	0,02994	0,00001	0,10156	0,00084	0,00599	0,01334
ER	0,99405	0,99998	0,98009	0,99883	0,99880	0,99734
ν^*	17,7020	17,2710	15,0370	17,2580	17,1670	17,0310
MA	22,6770	17,8810	18,2110	12,7490	18,5500	13,3915
t	0,39 ^{ns}	0,20 ^{ns}	-0,18 ^{ns}	0,25 ^{ns}	-0,25 ^{ns}	0,67 ^{ns}
valor- p	0,7015	0,8465	0,8631	0,8079	0,8028	0,5119

ER : Eficiência relativa.

MA : Média original dos ambientes.

ns : não significativo, $p \geq 0,05$.

Um aumento nas estimativas de r e λ nos ambientes 2, 4, 6 e 7, com uma conseqüente diminuição nas respectivas eficiências relativas, ocorreu quando do aumento de 5% para 10% na ausência dos dados (Tabelas 7 e 11). As estimativas das médias dos ambientes ($\hat{\beta}^*$) continuaram com uma pequena variabilidade, sendo a maior delas proveniente do ambiente 4 ($T = 0,12374$). O teste *t*-Student, não significativo ($p \geq 0,05$), indicou a semelhança de cada média de alturas dos ambientes para os dados originais (MA) e sua respectiva estimativa média ($\hat{\beta}^*$) nos dados completados, por meio do método proposto (Tabela 11). Estes resultados foram obtidos do Proc MIANALYZE (Anexo D) para 10% de ausência, na terceira etapa da IM.

Tabela 12 - Autovalor (λ_k^2), porcentagem da soma de quadrados acumulada (PA) nos eixos singulares (ES), desdobramento da interação $G \times E$, teste F , segundo Gollob e F_R , segundo Cornelius para os dados de alturas originais (orig.) e a média (imp.) das imputações, com 10% de ausência

ES	λ_k^2		PA (%)		FV	GL		Gollob <i>orig.</i>		Gollob <i>imp.</i>		Cornelius <i>orig.</i>		Cornelius <i>imp.</i>	
	orig.	imp.	orig.	imp.		GL	QM	F	QM	F	QM	F _R	QM	F _R	F _R
1	18,902	28,716	29,03	52,26	IPCA ₁	24	0,788	2,53**	0,947	2,96**	90	0,511	1,65**	0,482	1,53**
2	17,261	15,985	55,53	57,78	IPCA ₂	22	0,785	2,52**	0,727	2,27**	68	0,426	1,37*	0,416	1,30*
3	13,385	12,515	76,09	76,47	IPCA ₃	20	0,669	2,15**	0,626	1,95*	48	0,324	1,04 ^{ns}	0,328	1,03 ^{ns}
4	7,394	10,172	87,44	91,65	IPCA ₄	18	0,411	1,32 ^{ns}	0,565	1,77*	30	0,273	0,88 ^{ns}	0,186	0,58 ^{ns}
5	5,694	4,109	96,19	97,79	IPCA ₅	16	0,356	1,14 ^{ns}	0,257	0,80 ^{ns}	14	0,178	0,57 ^{ns}	0,106	0,33 ^{ns}
6	2,484	1,480	100,00	100,00	IPCA ₆	14	0,178	0,57 ^{ns}	0,106	0,33 ^{ns}					
					EM	216	0,311		0,320		216	0,311		0,320	

FV : Fontes de variação.

GL : Número de graus de liberdade.

QM : Quadrado médio.

EM : Erro médio.

** : significativo ao nível de probabilidade, $p < 0,01$.

* : significativo ao nível de probabilidade, $p < 0,05$.

ns : não significativo.

Na aplicação do programa desenvolvido no módulo IML do SAS (Anexo E) para 10% de ausência, os dados provenientes da média das 5 imputações (imp.) foram significativos, pelo teste F , segundo o critério de Gollob (1968), até o quarto eixo, indicando a seleção do modelo $AMMI_4$. Este resultado divergiu dos dados originais (orig.), os quais foram significativos até o terceiro eixo. Na seleção do modelo $AMMI$, pelo teste de F_R , tanto os dados originais como os imputados foram significativos até os dois primeiros eixos ($IPCA_1$ e $IPCA_2$), sugerindo, portanto, a escolha do modelo $AMMI_3$ (Tabela 12).

Tabela 13 - Média de alturas, em metros, das imputações, segundo a posição (linha i , coluna j) de retirada aleatória (30%) da Tabela 4

(continua)

Posição		Imputação (m)					$\bar{Y}$	Var	CV	acc
retirada	VO	1	2	3	4	5				
(3;1)	16,52	17,27	16,93	17,01	16,74	17,55	17,10	0,0989	1,84	0,5214
(9;1)	16,87	14,92	15,43	15,33	15,24	15,84	15,35	0,1115	2,17	2,9910
(13;1)	17,62	16,46	16,44	16,38	16,23	16,53	16,41	0,0129	0,69	1,8407
(15;1)	15,94	17,19	17,01	17,31	17,44	17,02	17,19	0,0344	1,08	2,0008
(18;1)	16,90	15,70	16,11	16,02	15,76	16,03	15,92	0,0329	1,14	1,2231
(2;2)	24,00	23,78	23,94	23,14	23,22	22,94	23,40	0,1855	1,85	0,6302
(5;2)	21,56	20,96	21,53	21,36	21,39	21,39	21,33	0,0454	1,01	0,1135
(11;2)	22,98	22,57	22,89	22,09	23,01	22,83	22,68	0,1341	1,61	0,2480
(19;2)	20,12	21,60	21,62	21,88	21,88	21,90	21,78	0,0226	0,70	3,4414
(3;3)	15,94	16,75	17,35	16,65	16,50	17,03	16,86	0,1138	2,00	1,1643
(6;3)	16,61	17,45	17,58	17,58	17,15	17,27	17,41	0,0358	1,10	0,8300
(7;3)	17,02	14,79	15,34	15,71	15,66	14,48	15,20	0,2962	3,57	4,4403
(17;3)	16,17	16,91	18,25	18,58	18,68	17,38	17,96	0,6056	4,34	4,6096
(19;3)	14,66	15,48	15,10	15,25	14,86	15,73	15,28	0,1118	2,20	0,5959
(1;4)	20,61	21,15	21,27	20,00	20,05	19,76	20,45	0,4999	3,46	0,5341
(3;4)	18,91	17,82	18,65	17,26	16,76	17,86	17,67	0,4986	4,02	2,4195
(6;4)	19,08	18,85	19,12	18,55	17,96	18,57	18,61	0,1873	2,32	0,4629
(9;4)	18,96	18,48	18,14	17,99	18,26	19,14	18,40	0,2031	2,44	0,5918

Tabela 13- Média de alturas, em metros, das imputações, segundo a posição (linha i , coluna j) de retirada aleatória (30%) da Tabela 4

(conclusão)

Posição		Imputação (m)					$\bar{Y}$	Var	CV	acc
retirada	VO	1	2	3	4	5				
(10;4)	18,94	19,00	19,78	18,86	18,72	18,98	19,07	0,1725	2,17	0,1928
(12;4)	18,19	17,65	17,29	17,88	17,35	16,99	17,43	0,1193	1,97	0,8363
(13;4)	18,78	16,43	16,33	17,05	15,93	16,27	16,40	0,1666	2,49	7,2190
(15;4)	17,91	18,31	17,69	18,82	18,95	18,22	18,40	0,2551	2,75	0,5523
(17;4)	18,91	17,90	18,71	20,80	21,06	17,93	19,28	2,3801	8,01	2,5512
(19;4)	15,68	15,11	14,57	15,46	14,22	15,69	15,01	0,3703	4,07	0,9321
(20;4)	16,46	15,67	14,87	15,71	15,54	14,93	15,34	0,1665	2,68	1,7342
(6;5)	11,91	13,09	13,89	13,15	12,81	13,70	13,33	0,2017	3,38	2,7101
(8;5)	13,66	14,16	12,76	13,58	13,34	14,02	13,57	0,3486	4,13	0,3854
(12;5)	12,62	12,99	11,93	12,52	12,64	12,36	12,49	0,1516	3,11	0,1730
(16;5)	12,80	13,01	13,50	13,44	13,38	13,85	13,44	0,0919	2,23	0,5971
(2;6)	17,77	19,08	19,22	19,72	19,56	18,98	19,31	0,0997	1,64	3,0783
(5;6)	18,06	17,64	17,17	17,59	17,67	17,67	17,55	0,0458	1,22	0,3748
(15;6)	19,71	18,93	18,64	18,90	18,72	18,46	18,73	0,0368	1,03	1,2394
(16;6)	19,44	18,76	19,07	19,10	18,93	19,10	18,99	0,0214	0,78	0,2694
(17;6)	20,24	18,52	19,70	19,89	19,52	18,83	19,29	0,3452	3,05	1,4667
(19;6)	16,10	17,03	16,74	16,98	16,89	17,00	16,93	0,0143	0,69	0,8755
(4;7)	13,03	12,63	12,73	12,61	12,72	12,71	12,68	0,0030	0,44	0,1547
(6;7)	13,17	14,01	14,22	13,90	14,04	13,84	14,00	0,0219	1,05	0,8906
(7;7)	11,14	12,57	12,71	12,59	12,61	11,99	12,49	0,0809	2,30	2,3688
(8;7)	14,37	14,43	13,80	13,54	13,58	14,10	13,89	0,1400	2,70	0,4315
(11;7)	13,13	13,34	13,28	14,07	14,14	13,89	13,74	0,1661	2,96	0,6334
(12;7)	13,52	13,24	12,84	12,96	12,57	12,92	12,91	0,0565	1,87	0,5258
(13;7)	13,24	12,77	12,81	12,87	12,66	12,70	12,76	0,0070	0,66	0,2898

VO : Valor original.

Var : Variância entre imputações.

acc : Acurácia.

$\bar{Y}$: Média das imputações.

CV : Coeficiente de variação.

Considerando 30% de retirada aleatória do conjunto de dados da Tabela 4 (Apêndice C), ou seja, 42 valores, o sorteio possibilitou a distribuição destes valores ausentes em todos os ambientes (Tabela 13). O maior número relativo de ausências continuou no ambiente 4, com aproximadamente 26% das ausências, ou seja, 11 valores ausentes, um valor aproximadamente 1,5 vez maior do que os 7 valores ausentes do ambiente 7.

Nas mesmas posições de retirada dos dados com 5% de ausência, Figura 8, a oscilação nos valores imputados continuou para mais ou para menos em relação aos valores originais, no entanto, a variabilidade em relação às médias das imputações ($\bar{Y}$) aumentou para todas as posições, com exceção da posição (19;2), como constatado por meio do CV das Tabelas 5 e 13 e numa visualização das Figuras 4 e 7.

Figura 7 - Média, erro padrão e desvio padrão de alturas (m) para as imputações com 30% de ausência e nas mesmas posições de retirada dos dados com 5% de ausência

Esta variabilidade, em relação às 14 posições dos dados com 10% de ausência, também aumentou na maioria das posições, com exceção das posições (2;2), (13;4), (5;6) e (4;7), nas quais houve uma diminuição da variabilidade (Tabelas 9 e 13). A diferença entre os valores imputados e os originais continuou sendo a maior parte na composição das *acc* (22), sendo que, na posição (13;4), ocorreu o maior valor desta medida ($acc = 7,2190$), com

referência a todas as porcentagens de ausência.

Figura 8 - Imputações e valores originais (VO) de alturas com 30% de ausência e nas mesmas posições de retirada dos dados com 5% de ausência

Tabela 14 - Média e erro padrão das médias de alturas (m) dos ambientes completados pelas imputações, nos dados com 30% de ausência

Imputação	Ambientes						
	1	2	3	4	5	6	7
1	16,557 (0,226)	22,689 (0,273)	16,930 (0,240)	17,908 (0,377)	12,862 (0,218)	18,483 (0,202)	13,460 (0,196)
2	16,575 (0,211)	22,743 (0,267)	17,042 (0,245)	17,911 (0,419)	12,804 (0,221)	18,511 (0,225)	13,431 (0,192)
3	16,582 (0,215)	22,667 (0,261)	17,050 (0,241)	18,009 (0,372)	12,813 (0,210)	18,593 (0,223)	13,439 (0,192)
4	16,550 (0,220)	22,719 (0,260)	17,004 (0,251)	17,829 (0,419)	12,808 (0,210)	18,549 (0,216)	13,428 (0,198)
5	16,628 (0,211)	22,697 (0,258)	16,956 (0,244)	17,806 (0,361)	12,896 (0,227)	18,487 (0,203)	13,419 (0,203)

Na segunda etapa da IM, a aplicação do Proc UNIVARIATE (Anexo C), para 30% de ausência, gerou um arquivo (*outuni*) com as médias de ambientes em cada conjunto completado pelas imputações ($M = 5$) e os seus respectivos erros padrão Tabela 14.

A combinação das estimativas das médias de alturas dos ambientes, proveniente dos 5 conjuntos de dados completados pelas imputações, na terceira etapa da IM, foi feita pelo Proc MIANALYZE (Anexo D) para 30% de ausência (Tabela 15). O ambiente 4 apresentou a maior variabilidade ($T = 0,16004$) nas estimativas das médias dos ambientes ($\hat{\beta}^*$), as quais continuaram com uma pequena variabilidade, ou seja, os valores T próximos de zero. A ausência dos dados continuaram com pouca influência em $\hat{\beta}^*$ e T , pois, tanto λ quanto r continuaram com seus valores pequenos. A semelhança entre cada média de alturas dos ambientes para os dados originais (MA) e sua respectiva estimativa média ($\hat{\beta}^*$) nos dados completados, por meio do método proposto, foi verificada pelo teste t -Student não significativo ($p \geq 0,05$).

Tabela 15 - Estimativa média ($\hat{\beta}^*$) das médias de alturas e medidas associadas a sua variabilidade, nos ambientes com valores imputados para dados com 30% de ausência. Teste t -Student para comparação com as médias originais

Estimativas	Ambientes						
	1	2	3	4	5	6	7
$\hat{\beta}^*$	16,57864	22,70311	16,99626	17,89248	12,83649	18,52436	13,43542
B	0,00095	0,00085	0,00278	0,00639	0,00165	0,00218	0,00024
$\bar{W}$	0,04698	0,06942	0,05957	0,15237	0,04716	0,04577	0,03847
T	0,04812	0,07044	0,06291	0,16004	0,04914	0,04838	0,03876
r	0,02415	0,01465	0,05602	0,05035	0,04197	0,05707	0,00761
λ	0,02385	0,01454	0,05438	0,04903	0,04106	0,05537	0,00758
ER	0,99525	0,99710	0,98924	0,99029	0,99186	0,98905	0,99849
ν^*	16,8260	17,0080	16,1700	16,2910	16,4660	16,1480	17,1380
MA	16,6720	22,6770	17,8810	18,2110	12,7490	18,5500	13,3915
t	-0,43 ^{ns}	0,10 ^{ns}	0,46 ^{ns}	-0,80 ^{ns}	0,39 ^{ns}	-0,12 ^{ns}	0,22 ^{ns}
valor- p	0,6758	0,9228	0,6520	0,4374	0,6981	0,9086	0,8261

ER : Eficiência relativa. MA : Média original dos ambientes. ns : não significativo, $p \geq 0,05$.

Tabela 16 - Autovalor (λ_k^2), porcentagem da soma de quadrados acumulada (PA) nos eixos singulares (ES), desdobramento da interação G×E, teste F , segundo Gollob e F_R , segundo Cornelius para os dados de alturas originais (orig.) e a média (imp.) das imputações, com 30% de ausência

ES	λ_k^2		PA (%)		FV	GL	Gollob <i>orig.</i>		Gollob <i>imp.</i>		GL	Cornelius <i>orig.</i>		Cornelius <i>imp.</i>	
	orig.	imp.	orig.	imp.			QM	F	QM	F		QM	F_R	QM	F_R
1	18,902	28,520	29,03	52,26	IPCA ₁	24	0,788	2,53**	1,188	4,56**	90	0,511	1,65**	0,289	1,11 ^{ns}
2	17,261	15,250	55,53	80,21	IPCA ₂	22	0,785	2,52**	0,769	2,66**	68	0,426	1,37*	0,159	0,61 ^{ns}
3	13,385	5,705	76,09	90,66	IPCA ₃	20	0,669	2,15**	0,285	1,09 ^{ns}	48	0,324	1,04 ^{ns}	0,106	0,41 ^{ns}
4	7,394	2,402	87,44	95,07	IPCA ₄	18	0,411	1,32 ^{ns}	0,134	0,51 ^{ns}	30	0,273	0,88 ^{ns}	0,090	0,34 ^{ns}
5	5,694	2,067	96,19	98,85	IPCA ₅	16	0,356	1,14 ^{ns}	0,129	0,50 ^{ns}	14	0,178	0,57 ^{ns}	0,045	0,17 ^{ns}
6	2,484	0,625	100,00	100,00	IPCA ₆	14	0,178	0,57 ^{ns}	0,045	0,17 ^{ns}					
					EM	216	0,311		0,261		216	0,311		0,261	

FV : Fontes de variação.
GL : Número de graus de liberdade.
QM : Quadrado médio.
EM : Erro médio.
** : significativo ao nível de probabilidade, $p < 0,01$.
* : significativo ao nível de probabilidade, $p < 0,05$.
ns : não significativo.

Na seleção do modelo *AMMI* ($AMMI_0, AMMI_1, \dots$, ou $AMMI_n$), por meio do programa desenvolvido no módulo IML do SAS (Anexo E) para 30% de ausência, tanto o critério de Gollob (1968) como o critério de Cornelius; Seyedsadr e Crossa (1992) e Piepho (1995), Tabela 16, levaram a resultados diferentes na comparação entre os dados originais (orig.) e a média das 5 imputações (imp.), ou seja, nos dados originais a escolha recaiu no modelo $AMMI_3$, com $F = 2,15^{**}$ para o terceiro eixo e $F_R = 1,37^*$ para o segundo eixo, enquanto na média das 5 imputações, $F = 2,66^{**}$ para o segundo eixo e $F_R = 1,11^{ns}$ para o primeiro eixo, indicando a seleção do modelo $AMMI_2$.

Numa comparação final, entre os valores imputados aos dados com 5%, 10% e 30% de ausência nas posições (2;2), (19;2), (10;4), (13;4), (19;4), (5;6), (4;7), o maior número de dados ausentes ocorreu no ambiente 4. A variabilidade relativa foi pequena, aumentando à medida que a ausência dos dados foi de 5% para 10%, com exceção, na posição (19;2). No aumento de 10% para 30% de ausências, esta variabilidade tanto aumentou, nas posições (19;2), (10;4) e (19;4), como diminuiu, (2;2), (13;4), (5;6) e (4;7).

Na comparação entre a média de alturas dos ambientes para os dados originais (MA) e sua respectiva estimativa média ($\hat{\beta}^*$) nos dados completados, verificada pelo teste *t*-Student, foi não significativa ($p \geq 0,05$), indicando semelhança entre elas, para todas as porcentagens de ausência, sendo que a maior variabilidade (*T*) ocorreu no ambiente 4. Na composição desta variabilidade (4), a maior parte foi devido à média das variâncias dentro das imputações ($\bar{W}$) em relação à variância entre imputações (*B*), e a influência da ausência de dados nas estimativas $\hat{\beta}^*$ e *T* foi pequena, expressa por valores baixos da precisão da estimativa se nenhum dado estivesse ausente (λ) e do aumento relativo na variância devido à ausência de dados (*r*), o que levou a uma eficiência relativa (*ER*) próxima de 100,00%, em todas as porcentagens de ausência.

Na aplicação da média dos valores imputados na seleção de um modelo *AMMI*, em comparação com a seleção nos dados originais, tanto o critério de Gollob (1968), como o de Cornelius; Seyedsadr e Crossa (1992) e Piepho (1995), levaram a uma mesma seleção ou a seleções divergentes de modelos *AMMI*, nas diferentes porcentagens de ausência.

O maior componente da medida geral de exatidão ou acurácia (T_{acc}), em todas as porcentagens de ausência, foi o viés quadrático médio, Tabela 17, no entanto, com o

aumento na porcentagem de ausência dos dados, a variância entre imputações (24) aumentou, mas o viés quadrático médio (25) diminuiu, resultando em um aumento da acurácia, evidenciado pela diminuição na medida geral de exatidão ou acurácia (23).

Tabela 17 - Medida geral da acurácia do método de imputação múltipla proposto, com 5%, 10% e 30% de ausência

Ausência	Acurácia geral		
	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

V_E : Variância entre imputações.

VQM : Viés quadrático médio.

T_{acc} : Medida geral de acurácia.

3 CONCLUSÕES

De acordo com os objetivos deste estudo, pode-se concluir:

- o método proposto para a imputação múltipla utiliza o maior número de informação disponível da matriz de dados, não possui qualquer restrição quanto ao padrão e mecanismo de ausência e é livre de suposição sobre a distribuição ou estrutura dos dados;
- na matriz de interação $G \times E$ para os dados de altura de *E. grandis*, a variabilidade em relação à média dos valores imputados, gerados pelo método proposto, é pequena, indicando uma alta precisão ;
- o viés em relação aos valores originais também é pequeno, no entanto, seus valores são maiores do que a variabilidade em relação à média dos valores imputados, indicando uma exatidão ou acurácia menor do método proposto, na matriz de interação $G \times E$, para os dados de altura de *E. grandis*;

Como continuidade deste estudo, seria interessante corrigir o viés dos autovaleores na DVS, com a finalidade de melhorar a qualidade da estimação pelo método proposto; utilizar este método, considerando variáveis resposta com uma distribuição discreta (binária, binomial ou Poisson), uma vez que o método independe da distribuição dos dados; generalizar a metodologia para o caso de três (genótipo $\times$ ambiente $\times$ ano) ou mais fatores no arranjo dos dados.

REFERÊNCIAS

- AERTS, M.; CLAESKENS, G.; HENS, N.; MOLENBERGHS, G. Local multiple imputation. **Biometrika**, Oxford, v. 89, n. 2, p. 375-388, 2002.
- ALLAN, F.G.; WISHART, J. A method of estimating the yield of a missing plot in field experiments. **Journal of Agricultural Science**, Cambridge, v. 20, p. 399-406, 1930.
- BARACHO, S.M.L.N. **Tratamento de dados ausentes em estudos longitudinais**. 2003. 79 p. Dissertação (Mestrado em Estatística) - Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, 2003.
- BARNARD, J ; RUBIN, D.B. Small-sample degrees of freedom with multiple imputation. **Biometrika**, Oxford, v. 86, n. 4, p. 948-955, 1999.
- BARTLETT, M.S. Some examples of statistical methods of research in agricultural and applied biology. **Journal of Royal Statistical Society**, Oxford, v. 4, suppl., n. 2, p. 137-183, 1937.
- BUCK, S.F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. **Journal of Royal Statistical Society**, Oxford, Series B, v. 22, p. 302-306, 1960.
- CORNELIUS, P.L.; SEYEDSADR, M.; CROSSA, J. Using the shifted multiplicative model to search for "separability" in crop cultivar trials. **Theoretical and Applied Genetics**, New York, v. 84, n. 1-2, p. 161-172, 1992.
- CROSSA, J. Statistical analyses of multilocation trials. **Advances in Agronomy**, New York, v. 44, p. 55-85, 1990.
- DAVID, M.H.; LITTLE, R.J.A; SAMUHEL, M.E.; TRIEST, R.K. Alternative methods for CPS income imputation. **Journal of the American Statistical Association**, Alexandria, v. 81, p. 29-41, 1986.
- DEAR, R.E. A Principal Component Missing Data Method for Multiple Regression Models, **System Development Corporation**, Technical Report SP-86, 1959.
- DEMPSTER, A.P.; LAIRD, N.M.; RUBIN, D.B. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). **Journal of Royal Statistical Society**, Oxford, Series B, v. 39, p. 1-38, 1977.
- DUARTE, J.B.; VENCOSKY, R. **Interação genótipo × ambiente: uma introdução à análise "AMMI"**. Ribeirão Preto: Sociedade Brasileira de Genética, 1999. p. 60 (Série Monografias, 9).
- FOX. P.N.; CROSSA, J.; ROMAGOSA, I. Multi-environment testing and genotype × environment interaction. In: KEMPTON, R.A.; FOX, P.N. (Ed.). **Statistical methods for plant variety evaluation**. New York: Chapman and Hall, 1997. cap. 8, p. 117-138.
- GAUCH, H.G. **Statistical analysis of regional yield trials: AMMI analysis of factorial designs**. New York: Elsevier Science, 1992. 278 p.

GAUCH, H.G.; ZOBEL, R.W. Imputing missing yield trial data. **Theoretical and Applied Genetics**, New York, v. 79, p. 753-761, 1990.

GOLLOB, H.F. A statistical model which combines features of factor analytic and analysis of variance techniques. **Psychometrika**, Colorado Springs, v. 33, n. 1, p. 73-115, 1968.

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

GROVES, R.; DILLMAN, D.; ELTINGE, J.; LITTLE, R.J.A **Survey nonresponse**. New York: John Wiley & Sons, 2002. 500 p.

HARTLEY, H.O; HOCKING, R.R. The analysis of incomplete data. **Biometrics**, Washington, v. 27, n. 4, p. 783-823, 1971.

HEALY, M.J.R.; WESTMACOTT, M. Missing values in experiments analysed on automatic computers. **Applied Statistics**, London, v. 5, n. 3, p. 203-206, 1956.

HEITJAN, D.F.; LITTLE, R.J. Multiple imputation for the fatal accident reporting system. **Applied Statistics**, London, v. 40, n. 1, p. 13-29, 1991.

HORTON, N.J.; LIPSITZ, S.R. Multiple imputation in practice: comparison of software packages for regression models with missing variables. **Journal of the American Statistical Association**, Alexandria, v. 55, p. 244-254, 2001.

KRZANOWSKI, W.J. Cross-validation in principal component analysis. **Biometrics**, Washington 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**, [s.l.], v. 25, n. 1-2, p. 31-39, 1988.

LAVORANTI, O.J. **Estabilidade e adaptabilidade fenotípica através da reamostragem “bootstrap” no modelo AMMI**. 2003. 166 p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2003.

LAVORI, P.W. ; DAWSON, R. SHERA, D. A multiple imputation strategy for clinical trials with truncation patient data. **Statistics in Medicine**, Chichester, v. 14, p. 1913-1925, 1995.

LI, K.H.; RAGHUNATHAN, T.E.; RUBIN, D.B. Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distributions. **Journal of the American Statistical Association**, Alexandria, v. 86, p. 1065-1073, 1991.

LIPSITZ, S.R.; ZHAO, L.P.; MOLENBERGHS, G. A semi-parametric method of multiple imputation. **Journal of the Royal Statistical Society**, Oxford, Series B, v. 60, p. 127-144, 1998.

LITTLE, R.J.; RUBIN, D.B. **Statistical analysis with missing data**. New York, New Jersey: John Wiley & Sons, 1987. 278 p.

- LITTLE, R.J.; RUBIN, D.B. **Statistical analysis with missing data**. 2.ed. New York, New Jersey: John Wiley & Sons, 2002. 381 p.
- MANDEL, J. A new analysis of variance for non-additive data. **Technometrics**, Alexandria, v. 13, n. 1, p. 1-18, 1971.
- MOLENBERGHS, G.; VERBEKE, G. **Models for discrete longitudinal data**. New York: Springer Science+Business Media, 2005. 683 p.
- PADDOCK, S.M. Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse. **Biometrika**, Oxford, v. 89, n. 3, p. 529-538, 2002.
- PENNY, K.I.; JOLLIFFE, I.T. Multivariate outlier detection applied to multiply imputed laboratory data. **Statistics in Medicine**, Chichester, v. 18, p. 1879-1895, 1999.
- PIEPHO, H.P. Robustness of statistical test for multiplicative terms in the additive main effects and multiplicative interaction model for cultivar trial. **Theoretical and Applied Genetics**, New York, v. 90, p. 438-443, 1995.
- ROSENBAUM, P.R.; RUBIN, D.B. The central role of the propensity score in observational studies for causal effects. **Biometrika**, Oxford, v. 70, n. 1, p. 41-45, 1983.
- RUBIN, D.B. Inference and missing data. **Biometrika**, Oxford, v. 63, n. 3, p. 581-592, 1976.
- RUBIN, D.B. Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. **Proceedings of the Survey Research Methods Section of the American Statistical Association**, Alexandria, p. 20-34, 1978.
- RUBIN, D.B. **Multiple imputation for nonresponse in surveys**. New York: John Wiley & Sons, 1987. 258 p.
- RUBIN, D.B.; SCHENKER, N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. **Journal of the American Statistical Association**, Alexandria, v. 81, p. 366-374, 1986.
- SAS INSTITUTE **SAS/IML 9.1 User's guide**. Carey, NC: SAS Institute Inc., 2004. 1040 p.
- SAS INSTITUTE **SAS/STAT 9.1 User's guide**. Carey, NC: SAS Institute Inc., 2004. 5121 p.
- SCHAFER, J.L. **Analysis of incomplete multivariate data**. London: Chapman & Hall, 1997. 430 p.
- SCHAFER, J.L. Multiple Imputation: A Primer. **Statistical Methods in Medical Research**, London, v. 8, p. 3-15, 1999.
- SCHENKER, N.; TAYLOR, J.M.G. Partially parametric techniques for multiple imputation. **Computational statistics and data analysis**, Amsterdam, v. 22, n. 4, p. 425-446, 1996.

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

THIJS, H.; MOLENBERGHS, G.; MICHIELS, B.; VERBEKE, G.; CURRAN, D. Strategies to fit pattern-mixture models. **Biostatistics**, Oxford, v. 3, p. 245–265, 2002.

VERBEKE, G.; MOLENBERGHS, G. **Linear mixed models in practice : a SAS-oriented approach** . New York: Springer, 1997. 306 p.

WILKINSON, G.N. Estimation of missing values for the analysis of incomplete data. **Biometrics**, Washington, v. 14, n. 2, p. 257-286, 1958.

YATES, F. The analysis of replicated experiments when the field results are incomplete. **Empire Journal of Experimental Agriculture**, London, v. 1, p. 129-142, 1933.

ZHANG, P. Multiple imputation: theory and method. **International Statistical Review**, Voorburg, The Netherlands, v. 71, n. 3, p. 581-592, 2003.

ZOBEL, R.W.; WRIGHT, M.J.; GAUCH, H.G. Statistical analysis of a yield trial. **Agronomy Journal**, Madison, v. 80, p. 388-393, 1988.

APÊNDICES

APÊNDICE A - Conjunto com 5% de dados retirados aleatoriamente

Tabela 18 - Médias de alturas, em metros, dos genótipos nos ambientes, com 5% de ausência nos dados

Genótipo	Ambientes						
	1	2	3	4	5	6	7
1	17,40	25,00	18,67	20,61	13,88	19,84	14,72
2	17,58	...	17,69	20,44	13,10	17,77	14,44
3	16,52	23,70	15,94	18,91	12,86	17,93	13,43
4	16,78	22,68	16,28	16,55	11,92	17,70	...
5	15,36	21,56	15,95	16,57	10,97	...	13,03
6	15,76	22,34	16,61	19,08	11,91	19,22	13,17
7	14,57	20,35	17,02	15,06	12,47	17,01	11,14
8	18,46	24,52	17,87	18,16	13,66	19,72	14,37
9	16,87	20,77	16,84	18,96	13,54	18,69	14,06
10	16,92	22,48	17,16	...	13,71	19,38	14,98
11	16,50	22,98	16,87	17,17	13,55	19,12	13,13
12	17,71	23,61	16,88	18,19	12,62	18,17	13,52
13	17,62	22,65	16,01	...	11,92	17,88	13,24
14	16,29	23,59	18,79	20,20	13,29	19,31	13,60
15	15,94	23,36	18,10	17,91	12,75	19,71	12,49
16	16,46	22,59	17,80	19,08	12,80	19,44	13,78
17	17,74	23,38	16,17	18,91	14,42	20,24	13,32
18	16,90	21,71	16,95	18,56	12,50	18,82	14,33
19	16,28	...	14,66	...	10,60	16,10	11,98
20	15,78	22,15	15,36	16,46	12,51	16,89	12,07

Nota : Sinal convencional utilizado:

... Dado numérico retirado.

APÊNDICE B - Conjunto com 10% de dados retirados aleatoriamente

Tabela 19 - Médias de alturas, em metros, dos genótipos nos ambientes, com 10% de ausência nos dados

Genótipo	Ambientes						
	1	2	3	4	5	6	7
1	17,40	25,00	18,67	...	13,88	19,84	14,72
2	17,58	...	17,69	20,44	13,10	17,77	14,44
3	16,52	23,70	15,94	...	12,86	17,93	13,43
4	16,78	22,68	16,28	16,55	11,92	17,70	...
5	15,36	21,56	15,95	16,57	10,97	...	13,03
6	15,76	22,34	...	...	...	19,22	13,17
7	14,57	20,35	17,02	15,06	12,47	17,01	...
8	18,46	24,52	17,87	18,16	13,66	19,72	14,37
9	16,87	20,77	16,84	18,96	13,54	18,69	14,06
10	16,92	22,48	17,16	...	13,71	19,38	14,98
11	16,50	22,98	16,87	17,17	13,55	19,12	13,13
12	17,71	23,61	16,88	18,19	12,62	18,17	13,52
13	17,62	22,65	16,01	...	11,92	17,88	13,24
14	16,29	23,59	18,79	20,20	13,29	19,31	13,60
15	15,94	23,36	18,10	17,91	12,75	19,71	12,49
16	16,46	22,59	17,80	19,08	12,80	19,44	13,78
17	17,74	23,38	16,17	18,91	14,42	20,24	13,32
18	16,90	21,71	16,95	18,56	12,50	18,82	14,33
19	16,28	...	14,66	...	10,60	16,10	11,98
20	15,78	22,15	15,36	...	12,51	16,89	12,07

Nota : Sinal convencional utilizado:

... Dado numérico retirado.

APÊNDICE C - Conjunto com 30% de dados retirados aleatoriamente

Tabela 20 - Médias de alturas, em metros, dos genótipos nos ambientes, com 30% de ausência nos dados

Genótipo	Ambientes						
	1	2	3	4	5	6	7
1	17,40	25,00	18,67	...	13,88	19,84	14,72
2	17,58	...	17,69	20,44	13,10	...	14,44
3	...	23,70	...	...	12,86	17,93	13,43
4	16,78	22,68	16,28	16,55	11,92	17,70	...
5	15,36	...	15,95	16,57	10,97	...	13,03
6	15,76	22,34	...	...	...	19,22	...
7	14,57	20,35	...	15,06	12,47	17,01	...
8	18,46	24,52	17,87	18,16	...	19,72	...
9	...	20,77	16,84	...	13,54	18,69	14,06
10	16,92	22,48	17,16	...	13,71	19,38	14,98
11	16,50	...	16,87	17,17	13,55	19,12	...
12	17,71	23,61	16,88	...	...	18,17	...
13	...	22,65	16,01	...	11,92	17,88	...
14	16,29	23,59	18,79	20,20	13,29	19,31	13,60
15	...	23,36	18,10	...	12,75	...	12,49
16	16,46	22,59	17,80	19,08	...	...	13,78
17	17,74	23,38	...	...	14,42	...	13,32
18	...	21,71	16,95	18,56	12,50	18,82	14,33
19	16,28	...	...	...	10,60	...	11,98
20	15,78	22,15	15,36	...	12,51	16,89	12,07

Nota : Sinal convencional utilizado:

... Dado numérico retirado.

ANEXOS

ANEXO A - Programa para retirada aleatória (5%) de valores da matriz de interação (Tabela 4)

```

proc iml;
*reset print;
/*DEFINE O NÚMERO DE LINHAS, COLUNAS E A
PORCENTAGEM DE VALORES AUSENTES NA MATRIZ Y */
Y={17.40  25.00  18.67  20.61  13.88  19.84  14.72,
17.58  24.00  17.69  20.44  13.10  17.77  14.44,
16.52  23.70  15.94  18.91  12.86  17.93  13.43,
16.78  22.68  16.28  16.55  11.92  17.70  13.03,
15.36  21.56  15.95  16.57  10.97  18.06  13.03,
15.76  22.34  16.61  19.08  11.91  19.22  13.17,
14.57  20.35  17.02  15.06  12.47  17.01  11.14,
18.46  24.52  17.87  18.16  13.66  19.72  14.37,
16.87  20.77  16.84  18.96  13.54  18.69  14.06,
16.92  22.48  17.16  18.94  13.71  19.38  14.98,
16.50  22.98  16.87  17.17  13.55  19.12  13.13,
17.71  23.61  16.88  18.19  12.62  18.17  13.52,
17.62  22.65  16.01  18.78  11.92  17.88  13.24,
16.29  23.59  18.79  20.20  13.29  19.31  13.60,
15.94  23.36  18.10  17.91  12.75  19.71  12.49,
16.46  22.59  17.80  19.08  12.80  19.44  13.78,
17.74  23.38  16.17  18.91  14.42  20.24  13.32,
16.90  21.71  16.95  18.56  12.50  18.82  14.33,
16.28  20.12  14.66  15.68  10.60  16.10  11.98,
15.78  22.15  15.36  16.46  12.51  16.89  12.07};

NL=nrow(Y);
NC=ncol(Y); /* OBS. Para obter 10% e 30% de valores ausentes, substituir */
PORC=5;      /* PORC por 10 e 30 respectivamente */
create original from Y [colname={loc1 loc2 loc3 loc4 loc5 loc6 loc7}];

```

```

    append from Y;
close original;
/* CALCULA A PORCENTAGEM DE VALORES AUSENTES */
NOM=(NL*NC)*(PORC/100);
if NOM<1
    then
        NOM=ceil(NOM);
    else
        NOM=floor(NOM);
/* DEFINE DE FORMA ALEATÓRIA AS POSIÇÕES DOS VALORES AUSENTES NA LINHA
        E NA COLUNA */
do I=1 to NOM;
    add:  EL=int(ranuni(1061958)*NL)+1;
        EC=int(ranuni(23081957)*NC)+1;
        if (Y[EL,EC]^=.)
            then Y[EL,EC]=.;
            else goto add;
end;
count=0;
do i=1 to NL;
    do j=1 to NC;
        if Y[i,j]=. then;
            count=count+1;
        end;
    end;
end;
/* CRIA O CONJUNTO DE DADOS COM 5% DE VALORES AUSENTES */
create dados5 from Y [colname={loc1 loc2 loc3 loc4 loc5 loc6 loc7}];
    append from Y;
close dados5;
quit;

```


ANEXO B - Programa para estimativa dos valores ausentes pela expressão (21), com 5% de ausência

```

/* IMPUTAÇÃO MULTIPLA EM DADOS MULTIVARIADOS USANDO
   A DECOMPOSIÇÃO POR VALOR SINGULAR */
/*OBS: PADRONIZANDO  $Y^i$  E  $Y_j$  */

proc iml;
*reset print;
/* ENTRADA DOS DADOS */
use dados5;
read all into LOST;
Y=LOST;
/* NI É O NÚMERO DE ITERAÇÕES */
NI=50;
UI=shape(0,NI,1);
do I=1 to NI; UI[I]=I; end;
SOMA=SHAPE(0,NI,1);
/* CALCULA O NÚMERO DE LINHAS E COLUNAS DOS DADOS */
L=NROW(Y);
C=NCOL(Y);
/* DETERMINA O NÚMERO DE OBSERVAÇÕES PRESENTES (NÃO AUSENTES)
   E COLOCA NA MATRIZ NO */
YM=Y=.;
UM=j(L,L,1);
NO=UM*YM;
NO=L-NO;
do I=1 to C;
if NO[1,I]=1
then
do;
print 'VOCÊ TEM DESVIO PADRAO INFINITO';

```

```

        ABORT;
    end; /*if NO[1,I]=1*/
end; /*do I=1 to C*/
/* LOCALIZA AS LINHAS E COLUNAS DOS VALORES AUSENTES NA MATRIZ Y*/
    z=loc(nmiss(y));
    ya=int((z+C-1)/C);
    yb=mod(z,C);
    zcol=ncol(z);
    if (zcol > 0) then do;
        do i=1 to zcol;
            if yb[,i]=0 then yb[,i]=C;
        end;
    end;
    ymiss=t(ya)||t(yb);
L1=NROW(ymiss);
/* CRIA A MATRIZ COM TODAS AS IMPUTAÇÕES MÚLTIPLAS */
    NL=nrow(LOST);
    NC=ncol(LOST);
    IMPUTMULT=j(5*NL,NC+1,0);
    AUX1=-NL+1;
    AUX2=0;
    AUX3=1;
    do i=1 to 5;
        IMPUTMULT[NL*i-(NL-1):NL*i,1]=i;
    end;
/* OBTÉM AS IMPUTAÇÕES MÚLTIPLAS VARIANDO O NUMERADOR (numer) DA EXPRESSÃO
        YE_=UT*(DT##(numer/20)*DBU##(numer/20))*VBLU */
do numer=8 to 12;
print numer;
AUX1=AUX1+NL;

```

```

AUX2=AUX2+NL;
/* K É O ÍNDICE DAS ITERAÇÕES E P O ÍNDICE DA LINHA NAS POSIÇÕES DOS VALORES
      AUSENTES */

DO K=1 TO NI;
/* CALCULA A MÉDIA DAS COLUNAS DE Y */
YTOT=REPEAT(Y[+,],L,1);
YBAR=YTOT/NO;
/* SUBSTITUI OS VALORES AUSENTES PELAS RESPECTIVAS MÉDIAS, EM Y SE K=1 */
IF K=1 THEN
DO;
  DO I=1 TO L1;
    Y[ymiss[I,1],ymiss[I,2]]=YBAR[ymiss[I,1],ymiss[I,2]];
  END;
YM=Y;
END; /* IF K=1 */
SOMAP=0;
  DO P=1 TO L1;
/* PADRONIZA A MATRIZ Y, COM AS MÉDIAS DAS COLUNAS NAS POSIÇÕES(1a. VEZ) OU
      ESTIMATIVAS */

UM=j(L,L,1);
YBAR=(1/L)*UM*Y;
DIF=Y-YBAR;
DIF2=DIF##2;
STD=UM*DIF2;
STD=(STD/(L-1))##0.5;
YS=(Y-YBAR)/STD;
/* DEFINE O NÚMERO DA LINHA E DA COLUNA DO VALOR AUSENTE */
EL=ymiss[P,1];
EC=ymiss[P,2];
/* FAZ A DECOMPOSIÇÃO EM VALORES SINGULARES DA MATRIZ DOS DADOS Y */

```

```

CALL SVD(U1,D1,V1,YS);
D=DIAG(D1);
U=U1[,1:C];
VL=V1[,1:C]';
A=U*D*VL;
/* DEFINE A SUBMATRIZ Y_i, COM ELIMINAÇÃO DA i-ÉSIMA LINHA DE Y */
IF (EL=1) THEN Y_i=Y[2:L,];
ELSE IF (EL=L) THEN Y_i=Y[1:L-1,];
ELSE Y_i=Y[1:EL-1,]/Y[EL+1:L,];
/* PADRONIZA A MATRIZ Y_i */
UM_i=j(L-1,L-1,1);
YBAR_i=(1/(L-1))*UM_i*Y_i;
DIF_i=Y_i-YBAR_i;
DIF2_i=DIF_i##2;
STD_i=UM_i*DIF2_i;
STD_i=(STD_i/(L-2))##0.5;
Y_i=(Y_i-YBAR_i)/STD_i;
/* DEFINE A SUBMATRIZ Y_j, COM ELIMINAÇÃO DA j-ÉSIMA COLUNA DE Y */
IF (EC=1) THEN Y_j=Y[,2:C];
ELSE IF (EC=C) THEN Y_j=Y[,1:C-1];
ELSE Y_j=Y[,1:EC-1]||Y[,EC+1:C];
/* PADRONIZA A MATRIZ Y_j */
UM_j=j(L,L,1);
YBAR_j=(1/L)*UM_j*Y_j;
DIF_j=Y_j-YBAR_j;
DIF2_j=DIF_j##2;
STD_j=UM_j*DIF2_j;
STD_j=(STD_j/(L-1))##0.5;
Y_j=(Y_j-YBAR_j)/STD_j;
/* FAZ A DECOMPOSIÇÃO EM VALORES SINGULARES DA MATRIZ Y_i */

```

```

CALL SVD(UB,DB1,VB,Y_i);
DB=DIAG(DB1);
UB=UB[,1:C];
VBL=VB[,1:C]';
AY_i=UB*DB*VBL;
/* FAZ A DECOMPOSIÇÃO EM VALORES SINGULARES DA MATRIZ Y_j */
CALL SVD(UT,DT1,VT,Y_j);
DT=DIAG(DT1);
UT=UT[,1:C-1];
VTL=VT[,1:C-1]';
AY_j=UT*DT*VTL;
/* TROCA SINAIS DA MATRIZ UT PARA QUE FIQUEM IGUAIS AOS DA MATRIZ U */
SU=0; SUT=0;
DO I=1 TO L;
    DO J=1 TO (C-1);
        IF U[I,J]<0 THEN SU=-1;
            ELSE IF U[I,J]>0 THEN SU=1;
        IF UT[I,J]<0 THEN SUT=-1;
            ELSE IF UT[I,J]>0 THEN SUT=1;
        IF SU ^= SUT THEN UT[I,J]=-UT[I,J];
    END;
END;
/* TROCA SINAIS DA MATRIZ VBL PARA QUE FIQUEM IGUAIS AOS DA MATRIZ VL */
SU=0; SUT=0;
DO I=1 TO C;
    DO J=1 TO C;
        IF VL[I,J]<0 THEN SU=-1;
            ELSE IF VL[I,J]>0 THEN SU=1;
        IF VBL[I,J]<0 THEN SUT=-1;
            ELSE IF VBL[I,J]>0 THEN SUT=1;
    END;
END;

```

```

    IF SU ^= SUT THEN VBL[I,J]=-VBL[I,J];
END;
END;
/* ELIMINA ÚLTIMO ELEMENTO DE DB ÚLTIMA LINHA DE VBL */
DBU=DIAG(DB1[1:C-1,]);
VBLU=VBL[1:C-1,];
/* CALCULA AS IMPUTAÇÕES MÚLTIPLAS */
YE_=UT*(DT##(numer/20)*DBU##((20-numer)/20))*VBLU;
YE_=YBAR+STD#YE_;
/* COLOCA AS ESTIMATIVAS NAS RESPECTIVAS POSIÇÕES EM YM */
YM[ymiss[P,1],ymiss[P,2]]=YE_[ymiss[P,1],ymiss[P,2]];
END; /*P*/
/* COLOCA EM Y A MATRIZ YM COM OS VALORES AUSENTES ESTIMADOS E NA ESCALA
ORIGINAL */
Y=YM;
END; /*K*/
/* CRIA UMA MATRIZ COM TODAS AS IMPUTAÇÕES */
do j= 1 to NC;
    IMPUTMULT[AUX1:AUX2,j+1]=Y[1:20,j];
end;
end;
create miout5 from IMPUTMULT [colname={_Imputation_ amb1 amb2 amb3 amb4
    amb5 amb6 amb7}];
append from IMPUTMULT;
close miout5;
IMPUTMULT1=IMPUTMULT[1:20,2:8];
IMPUTMULT2=IMPUTMULT[21:40,2:8];
IMPUTMULT3=IMPUTMULT[41:60,2:8];
IMPUTMULT4=IMPUTMULT[61:80,2:8];
IMPUTMULT5=IMPUTMULT[81:100,2:8];

```

```
IMmedia=(INPUTMULT1+INPUTMULT2+INPUTMULT3+INPUTMULT4+INPUTMULT5)/5;  
create IMmedia from IMmedia;  
  append from IMmedia;  
  close IMmedia;  
quit;
```

ANEXO C - Programa para obter a média e o erro padrão dos $M = 5$ conjuntos de dados completados, com 5% de ausência

```
proc univariate data=miout5 noprint;
    var  amb1 amb2 amb3 amb4 amb5 amb6 amb7;
    output out=outuni mean=  amb1 amb2 amb3 amb4 amb5 amb6 amb7
           stderr=  Samb1 Samb2 Samb3 Samb4 Samb5 Samb6 Samb7;
    by _Imputation_;
run;
```


**ANEXO D - Programa para combinar as médias de alturas dos ambientes
dos $M = 5$ conjuntos de dados completados**

```
proc mianalyze data=outuni edf=19 mu0= 16.672 22.677 16.881  
18.211 12.749 18.55 13.3915;  
modeleffects amb1 amb2 amb3 amb4 amb5 amb6 amb7;  
stderr Samb1 Samb2 Samb3 Samb4 Samb5 Samb6 Samb7;  
run;
```

ANEXO E - Programa para a seleção do modelo AMMI, considerando os testes F , segundo Gollob, e F_R , segundo Cornelius

```

proc iml;
use original;
read all into original;
ORIG=original;
  ng=nrow(ORIG);
    na=ncol(ORIG);
    GM=ORIG[:,];
    AM=ORIG[:,];
    GG=ORIG[:,];
    CORR=(ng*na)*(GG##2);
    SQTR=ssq(ORIG)-CORR;
    SQGEN=na*ssq(GM)-CORR;
    SQAMB=ng*ssq(AM)-CORR;
    GLMOD=(ng-1)+(na-1)+(ng-1)*(na-1);
XG=GM*J(1,na,1);
XA=J(ng,1,1)*AM;
MG=J(ng,1,1)*GG*J(1,NA,1);
GE=ORIG-XG-XA+MG;
EIGV=eigval(GE*GE');
  nnul1=EIGV<-1E-9;
  nnul2=EIGV>1E-9;
  nnul=nnul1+nnul2;
  r=(nnul^=0)[+,];
SQGE=ssq(GE);
GLERR=((ng+na)*na)+ng+na;
QMERR=((1+(na/GLERR))*SQGE)/GLERR;
call SVD(eG,VS,eA,GE);
  S=diag(VS[1:r,]);

```

```

      U=eG[,1:r];
      VT=eA[,1:r]';
/* DESDOBRAMENTO DA SQ(GxE) POR AMMI (ACP ou DVS)*/
Li=VS##2;
AUTV=Li[1:r,];
SAUTV=sum(Li);
PRP=INV(SAUTV)*Li;
PROP=PRP[1:r,];
RESD1=SAUTV-Li[1:1,];
PACUM=CUSUM(PROP)*100;
LINH=(1:r)';
RESUMOrig=LINH||AUTV||PROP||PACUM;
colu={'CP/Eixo' 'Autovalor' 'Prop./CP' '%Acumulada'};
print
',
'-----',
'Autovalores da ACP <=> Desdobramento da SQ(QxE) por AMMI',
'Proporção da SQ(GxE) atribuída a cada CP (ou eixo singular de interação)',
'Percentagem da SQ(GxE) acumulada até o n-ésimo CP (ou eixo singular)',
',
'-----';
print RESUMOrig [colname=colu format=12.4];
/* ESCORES DE GENÓTIPOS E AMBIENTES E APROXIMAÇÃO SUCESSIVAS DE GE */
n=0;
  GXEI=J(ng,na,0);
  GLACU=0;
  SQGEI=ssq(GXEI);
  PROPI=INV(SAUTV)*SQGEI;
  PSQTR=INV(SQTR)*(SQGEN+SQAMB);
  XPO=XG+XA-MG;
do n=1 to r by 1;
  i=n;

```

```

VG=eG[,i];
ES=S[i,i];
VA=eA[,i];
    IN=VG*ES*VA';
    G=VG*(sqrt(ES));
    ESGE=G';
    H=sqrt(ES)*VA;
    ESAM=H';
if i=1 then ECP1=G//H;
if i=2 then ECP2=G//H;
if i=3 then ECP3=G//H;
    GEI=G*H';
    ZE=IN-GEI;
    SQGEI=SQGEI+ssq(GEI);
    PROPI=INV(SAUTV)*SQGEI;
GLAMI=ng+na-1-(2*i);
    QMAMI=inv(GLAMI)*(ssq(GEI));
GLACU=GLACU+GLAMI;
    GLRAMI=(ng-1)*(na-1)-GLACU;
if i<r then QMRAMI=inv(GLRAMI)*(SAUTV-SQGEI);
else QMRAMI=0;
FMODI=inv(QMERR)*QMAMI;
FRESI=INV(QMERR)*QMRAMI;
GXEI=GXEI+GEI;
    XPI=XG+XA-MG+GXEI;
PSQTR=inv(SQTR)*(SQGEN+SQAMB+SQGEI);
    prob1=1-probf(FMODI, GLAMI, GLERR);
if i<r then prob2=1-probf(FRESI, GLRAMI, GLERR);
else prob2={.};
FV1=GLAMI||QMAMI||FMODI||PROB1;

```

```

FV2=GLRAMI||QMRAMI||FRESI||PROB2;
FV3=GLERR||QMERR||{.}||{.};
FVorig=FV1//FV2//FV3;

lin={'AMMIn(GxE)' 'Res.GxE/AMMIn' 'Erro médio'};
col={'GL' 'QM' 'teste F' 'Pr>F'};

print 'Resultados da avaliação do modelo AMMI';
print FVorig [rowname=lin colname=col format=12.4];

end;

use IMmedia;
read all into IMmedia;
X=IMmedia;

ng=nrow(X);
na=ncol(X);
GM=X[, :];
AM=X[:, :];
GG=X[:, :];
CORR=(ng*na)*(GG##2);
SQTR=ssq(X)-CORR;
SQGEN=na*ssq(GM)-CORR;
SQAMB=ng*ssq(AM)-CORR;
GLMOD=(ng-1)+(na-1)+(ng-1)*(na-1);

XG=GM*J(1,na,1);
XA=J(ng,1,1)*AM;
MG=J(ng,1,1)*GG*J(1,NA,1);
GE=X-XG-XA+MG;
EIGV=eigval(GE*GE');

nnul1=EIGV<-1E-9;
nnul2=EIGV>1E-9;
nnul=nnul1+nnul2;

r=(nnul^=0)[+,];

```

```

SQGE=ssq(GE);
GLERR=((ng+na)*na)+ng+na;
QMERR=((1+(na/GLERR))*SQGE)/GLERR;
call SVD(eG,VS,eA,GE);
      S=diag(VS[1:r,]);
      U=eG[,1:r];
      VT=eA[,1:r]';
/* DESDOBRAMENTO DA SQ(GxE) POR AMMI (ACP ou DVS)*/
Li=VS##2;
AUTV=Li[1:r,];
SAUTV=sum(Li);
PRP=INV(SAUTV)*Li;
PROP=PRP[1:r,];
RES1=SAUTV-Li[1:1,];
PACUM=CUSUM(PROP)*100;
LINH=(1:r)';
RESUM0media=LINH||AUTV||PROP||PACUM;
colu={'CP/Eixo' 'Autovalor' 'Prop./CP' '%Acumulada'};
print
', -----',
',Autovalores da ACP <=> Desdobramento da SQ(QxE) por AMMI',
',Proporção da SQ(GxE) atribuída a cada CP (ou eixo singular de interação)',
',Percentagem da SQ(GxE) acumulada até o n-ésimo CP (ou eixo singular)',
', -----';
print RESUM0media [colname=colu format=12.4];
/* ESCORES DE GENÓTIPOS E AMBIENTES E APROXIMAÇÃO SUCESSIVAS DE GE */
n=0;
      GXEI=J(ng,na,0);
      GLACU=0;
      SQGEI=ssq(GXEI);

```

```

PROPI=INV(SAUTV)*SQGEI;
PSQTR=INV(SQTR)*(SQGEN+SQAMB);
XPO=XG+XA-MG;
do n=1 to r by 1;
    i=n;
    VG=eG[,i];
    ES=S[i,i];
    VA=eA[,i];
    IN=VG*ES*VA';
    G=VG*(sqrt(ES));
    ESGE=G';
    H=sqrt(ES)*VA;
    ESAM=H';
if i=1 then ECP1=G//H;
if i=2 then ECP2=G//H;
if i=3 then ECP3=G//H;
    GEI=G*H';
    ZE=IN-GEI;
    SQGEI=SQGEI+ssq(GEI);
    PROPI=INV(SAUTV)*SQGEI;
GLAMI=ng+na-1-(2*i);
    QMAMI=inv(GLAMI)*(ssq(GEI));
GLACU=GLACU+GLAMI;
    GLRAMI=(ng-1)*(na-1)-GLACU;
if i<r then QMRAMI=inv(GLRAMI)*(SAUTV-SQGEI);
else QMRAMI=0;
FMODI=inv(QMERR)*QMAMI;
FRESI=INV(QMERR)*QMRAMI;
GXEI=GXEI+GEI;
    XPI=XG+XA-MG+GXEI;

```

```

PSQTR=inv(SQTR)*(SQGEN+SQAMB+SQGEI);
        prob1=1-probf(FMODI, GLAMI, GLERR);
if i<r then prob2=1-probf(FRESI, GLRAMI, GLERR);
else prob2={.};
FV1=GLAMI||QMAMI||FMODI||PROB1;
FV2=GLRAMI||QMRAMI||FRESI||PROB2;
FV3=GLERR||QMERR||{.}||{.};
FVmedio=FV1//FV2//FV3;
        lin={'AMMIn(GxE)' 'Res.GxE/AMMIn' 'Erro médio'};
        col={'GL' 'QM' 'teste F' 'Pr>F'};
print 'Resultados da avaliação do modelo AMMI';
print FVmedio [rowname=lin colname=col format=12.4];
end;
quit;

```


ANEXO F - Programa para o cálculo da medida geral de exatidão T_{acc} , com 5% de ausência

```

proc iml;
use miout5;
read all into miout5;
IMPUT5=miout5;
M=5;
IMPUT5MULT1=IMPUT5[1:20,2:8];
IMPUT5MULT2=IMPUT5[21:40,2:8];
IMPUT5MULT3=IMPUT5[41:60,2:8];
IMPUT5MULT4=IMPUT5[61:80,2:8];
IMPUT5MULT5=IMPUT5[81:100,2:8]`;
IMmedia=(IMPUT5MULT1+IMPUT5MULT2+IMPUT5MULT3+IMPUT5MULT4+IMPUT5MULT5)/M;
use original;
read all into original;
ORIG=original;
/* Módulo para definição da função PRESS */
START PRESS(ARG1,ARG2);
    ARG3=SUM((ARG2-ARG1)##2);
RETURN(ARG3);
FINISH PRESS;
PRESS5MULT1=PRESS(ORIG,IMPUT5MULT1);
PRESS5MULT2=PRESS(ORIG,IMPUT5MULT2);
PRESS5MULT3=PRESS(ORIG,IMPUT5MULT3);
PRESS5MULT4=PRESS(ORIG,IMPUT5MULT4);
PRESS5MULT5=PRESS(ORIG,IMPUT5MULT5);
na=7;
Tacc5=(1/na)*(PRESS5MULT1+PRESS5MULT2+PRESS5MULT3+PRESS5MULT4+
PRESS5MULT5)/(M-1);
print Tacc5;

```

```

/* Módulo para definição da função PRESSTacc1 */
START PRESS1Tacc(ARG1,ARG2);
    ARG3=SUM((ARG2-ARG1)##2);
RETURN(ARG3);
FINISH PRESS1Tacc;
PRESS1Tacc5MULT1=PRESS1Tacc(IMmedia,IMPUT5MULT1);
PRESS1Tacc5MULT2=PRESS1Tacc(IMmedia,IMPUT5MULT2);
PRESS1Tacc5MULT3=PRESS1Tacc(IMmedia,IMPUT5MULT3);
PRESS1Tacc5MULT4=PRESS1Tacc(IMmedia,IMPUT5MULT4);
PRESS1Tacc5MULT5=PRESS1Tacc(IMmedia,IMPUT5MULT5);
parte1Tacc=(1/na)*(PRESS1Tacc5MULT1+PRESS1Tacc5MULT2+PRESS1Tacc5MULT3+
PRESS1Tacc5MULT4+PRESS1Tacc5MULT5)/(M-1);
print parte1Tacc;
/* Módulo para definição da função PRESSTacc2 */
START PRESS2Tacc(ARG1,ARG2);
    ARG3=SUM((ARG2-ARG1)##2);
RETURN(ARG3);
FINISH PRESS2Tacc;
PRESS2Tacc5MULT1=PRESS2Tacc(IMmedia,ORIG);
PRESS2Tacc5MULT2=PRESS2Tacc(IMmedia,ORIG);
PRESS2Tacc5MULT3=PRESS2Tacc(IMmedia,ORIG);
PRESS2Tacc5MULT4=PRESS2Tacc(IMmedia,ORIG);
PRESS2Tacc5MULT5=PRESS2Tacc(IMmedia,ORIG);
parte2Tacc=(1/na)*(PRESS2Tacc5MULT1+PRESS2Tacc5MULT2+PRESS2Tacc5MULT3+
PRESS2Tacc5MULT4+PRESS2Tacc5MULT5)/(M-1);
print parte2Tacc;
quit;

```