IMPUTAÇÃO DE DADOS EM EXPERIMENTOS COM INTERAÇÃO GENÓTIPO×AMBIENTE

Sergio Arciniegas-Alarcón, Marisol García-Peña e Carlos Tadeu dos Santos Dias

RESUMO

O objetivo deste trabalho foi estudar os erros de predição associados a quatro métodos de imputação de dados aplicados para resolver o problema de desbalanceamento em experimentos com interação genótipo×ambiente (G×A). Um estudo de simulação foi realizado com base em quatro matrizes completas de dados reais, obtidas em ensaios de interação G×A de ervilha, algodão, feijão e eucalipto respectivamente. A simulação de desbalanceamento foi feita com retiradas aleatórias de 10, 20 e 40% dos dados em cada matriz. Os erros de predição foram encontrados utilizando-se validação cruzada, e foram testados intervalos de confiança de 95% para as observações ausentes. Para imputação de dados foram considerados algoritmos usando modelos de efeitos aditivos sem interação e estimativas de modelos de efeitos aditivos com interação multiplicativa baseadas em submodelos robustos. Em geral, os melhores erros de predição foram apresentados após a imputação por meio de um modelo aditivo sem interação.

Introdução

No melhoramento genético de plantas, os ensaios multiambientais são importantes para testar a adaptação geral e específica das cultivares. Uma cultivar desenvolvendo-se em diferentes ambientes mostrará uma flutuação significativa do desempenho na produção relativa a outras cultivares. Essas mudanças são influenciadas por diferentes condições ambientais e são referidas como interação genótipo por ambiente ou G×A.

Muitas vezes os experimentos G×A são desbalanceados e vários genótipos não são testados em alguns ambientes. Uma maneira muito comum de analisar dados provenientes desse tipo de estudos é imputar as observações ausentes e posteriormente, aplicar procedimentos clássicos, como por exemplo, o modelo de efeitos aditivos com interação multiplicativa (AMMI) ou a regressão fatorial (van Eeuwijk *et al*., 2005, 2007; Gauch, 2006; Romagosa *et al*., 2008; Arciniegas-Alarcón e Dias, 2009b) sobre a matriz de dados completada (observados + imputados). Uma aproximação alternativa é trabalhar com dados incompletos sob a estrutura do modelo misto com estimativas baseadas em máxima verossimilhança (Kang *et al*., 2004).

Na literatura têm sido sugeridos vários métodos de imputação para resolver o problema de dados faltantes. Um dos primeiros estudos foi o feito por Freeman (1975), que sugere imputar as observações ausentes de maneira iterativa minimizando a soma de quadrados do erro e sobre a tabela completa fazer as análises da interação G×A diminuindo os graus de liberdade pelo número de dados faltantes. Um trabalho aceito nestes experimentos foi o desenvolvido por Gauch e Zobel (1990), que fizeram a imputação através do uso do algoritmo EM e o modelo AMMI. Algumas alternativas desse procedimento usando estatística multivariada (análise de agrupamentos) foram descritas em Godfrey *et al*. (2002). Mandel (1993) propôs fazer a imputação em tabelas incompletas de dupla entrada usando funções lineares das linhas (ou colunas). Outros estudos que são recomendados por van Eeuwijk e Kroonenberg (1998) no caso de observações faltantes para experimentos G×A com resultados razoavelmente bons, foram os desenvolvidos por Denis (1991), Calinski *et al*. (1992) e Denis e Baril (1992). Eles encontraram que usando imputações através dos mínimos quadrados alternados com modelos de interação bilinear ou estimativas AMMI baseadas em submodelos robustos pode-se obter resultados tão bons como os encontrados com o algoritmo EM. Adicionalmente, Calinski *et al*. (1999) apresentaram um algoritmo que combina a decomposição por valores singulares (DVS) de uma matriz com o algoritmo EM, obtendo resultados muito úteis para aqueles experimentos em que os mínimos quadrados alternados apresentam alguns problemas, como por exemplo, falhas de convergência. Mais recentemente, Bergamo *et al*. (2008) propuseram um método de imputação múltipla livre de distribuição que foi avaliado por Arciniegas-Alarcón (2008) comparando-o com outros algoritmos em um estudo de simulação a partir de dados reais. Finalmente, um método de imputação determinística sem pressuposições estruturais nem distribucionais para os experimentos multiambientais foi proposto por Arciniegas-Alarcón *et al*. (2010). O método utiliza uma mistura de regressão com a aproximação de posto inferior de uma matriz.

Geralmente, os estudos que envolvem observações ausen-

PALAVRAS-CHAVE / Modelos AMMI / Observações Ausentes / Validação Cruzada /

Recebido: 26/01/2010. Modificado: 24/04/2011. Aceito: 29/04/2011.

Sergio Arciniegas-Alarcón. Mestre em Estatística e Experimentação Agronômica, Universidade de São Paulo (USP), Brasil. Docente Investigador, Departamento de Estadística, Universidad Nacional de Colombia. e-mail: sergio.arciniegas@gmail.com

Marisol García-Peña. Mestre em Estatística e Experimentação Agronômica, Universidade de São Paulo (USP), Brasil. Docente Investigador, Departamento de Estadística, Universidad Nacional de Colombia. e-mail: mgarciap@unal.edu.co

Carlos Tadeu dos Santos Dias. Ph.D. em Estatística e Experimentação Agronômica, USP, Brasil. Docente, ESALQ-USP, Brasil. Endereço: Departamento de Ciências Exatas, ESALQ-USP. Caixa Postal 9, CEP 13418-900 Piracicaba, SP, Brasil. e-mail: ctsdias@esalq.usp.br

DATA IMPUTATION IN TRIALS WITH GENOTYPE×ENVIRONMENT INTERACTION

Sergio Arciniegas-Alarcón, Marisol García-Peña and Carlos Tadeu dos Santos Dias

SUMMARY

The aim of this work was the study of prediction errors associated with four imputation methods applied to solve the problem of unbalance in experiments with genotype×environment (G×E) interaction. A simulation study was carried out based on four complete matrices of real data obtained in trials of interaction G×E of pea, cotton, beans and eucalyptus, respectively. The simulation of unbalance was done with random withdrawal of 10, 20 and 40% in each matrix. The prediction errors were found using cross-validation and were tested in classic intervals of 95% for missing data. For data imputation, algorithms were considered using models of additive effects without interaction and model estimates of additive effects with multiplicative interaction based on robust submodels. In general, the best prediction errors were obtained after imputation through an additive model without interaction.

IMPUTACIÓN DE DATOS EN EXPERIMENTOS CON INTERACCIÓN GENOTIPO×AMBIENTE

Sergio Arciniegas-Alarcón, Marisol García-Peña y Carlos Tadeu dos Santos Dias

RESUMEN

El objetivo de este trabajo fue estudiar los errores de predicción asociados a cuatro métodos de imputación de datos aplicados para resolver el problema de desbalanceamiento en experimentos con interacción genotipo×ambiente (G×A). Se realizó un estudio de simulación con base en cuatro matrices completas de datos reales, obtenidas en ensayos de interacción G×A de arveja, algodón, fríjol e eucalipto, respectivamente. La simulación de desbalanceamiento fue realizada por medio de pérdidas aleatorias de 10, 20 y 40% de los datos en cada matriz. Los errores de predicción fueron encontrados utilizando validación cruzada, y fueron evaluados intervalos de confianza de 95% para las observaciones ausentes. Para la imputación de datos fueron considerados algoritmos usando modelos de efectos aditivos sin interacción y estimaciones de modelos de efectos aditivos con interacción multiplicativa basadas en submodelos robustos. En general, los mejores errores de predicción se presentaron después de la imputación por medio de un modelo aditivo sin interacción.

---

tes, como os descritos acima, têm principalmente como foco de pesquisa a imputação do dado e por isso, a diferença deles, o objetivo deste trabalho consiste em estudar os erros de predição (ou imputação) associados aos métodos de imputação usados em experimentos de interação G×A. Para delimitar a pesquisa foram considerados quatro algoritmos de imputação e para encontrar os respectivos erros foi utilizada a técnica conhecida como validação cruzada.

Material e Métodos

Os dados utilizados correspondem a quatro matrizes de experimentos reais G×A publicados em Calinski *et al*. (2009) com genótipos de ervilha, de algodão em Farias (2005), de feijão em Flores *et al*. (1998) e de eucalipto em Lavoranti (2003). As matrizes têm dimensões 18×9, 15×27, 15×12 e 20×7 (cultivares×locais) respectivamente, e contêm as produções médias ($kg \cdot ha^{-1}$), exceto, nos dados de eucalipto, nos quais foi coletada a média de altura em metros. Cada referência traz um completo detalhe dos delineamentos experimentais usados, mas por questão de espaço não são apresentados aqui.

Para imputar observações, foram considerados os resultados descritos em Arciniegas-Alarcón e Dias (2009a). Nesse trabalho se encontrou que as imputações realizadas por mínimos quadrados alternados com um modelo aditivo sem interação (ALS(0)) e as estimativas AMMI baseadas em submodelos robustos (r-AMMI) podem ser mais eficientes do que a imputação múltipla livre de distribuição baseada na DVS de uma matriz. Na prática, o método ALS(0) é equivalente a fazer estimação mínimos quadrados utilizando um modelo AMMI0 ou $y_{ij} = \mu + a_i + b_j + e_{ij}$ sem envolver alguma iteração, em que y_{ij} representa a variável estudada, a_i, b_j representam os efeitos principais genotípicos e ambientais, e e_{ij} representa o termo do erro associado ao i-ésimo genótipo e ao j-ésimo ambiente.

Entretanto, o método r-AMMI proposto por Denis e Baril (1992) utiliza os modelos AMMI, tais como os AMMIk ou $y_{ij} = \mu + a_i + b_j + \lambda_1\alpha_{1i}\gamma_{1j} + \lambda_2\alpha_{2i}\gamma_{2j} + ... + \lambda_k\alpha_{ki}\gamma_{kj} + e_{ij}$, em que y_{ij}, a_i, b_j e e_{ij} já foram definidos anteriormente e os k componentes multiplicativos representam a interação G×A, em que λ_k, α_{1k} e γ_{1k} (k=1,...,) são estimados pela DVS da matriz de resíduos depois de ajustar a parte aditiva. λ_k é estimado pelo k-ésimo valor singular da DVS, e α_{1k} e γ_{1k} são estimados pelos correspondentes autovetores genotípicos e ambientais correspondentes a λ_k. O método sugere que no caso de observações ausentes em experimentos G×A, devem-se fazer análises sobre tabelas completas, em que os dados faltantes são substituídos pelas estimativas de um submodelo robusto. Resultados empíricos indicaram que uma ponderação igual para os valores ausentes e observados é aceitável. Com uma ponderação igual, as análises são equivalentes às análises AMMI clássicas para dados completos. Para a análise AMMI1, Denis e Baril (1992) propuseram AMMI0 como um submodelo robusto; da mesma maneira o AMMI1 pode ser usado como um submodelo robusto para uma análise AMMI2, e assim sucessivamente. Em geral, para uma análise AMMIk+1 um submodelo robusto pode ser AMMIk, e então o método é

chamado r-AMMIk. Levando em conta o exposto acima, neste trabalho foram considerados os métodos de imputação: modelo aditivo sem interação, r-AMMI1, r-AMMI2 e r-AMMI3.

Uma vez estabelecidos os métodos de imputação, pode ser muito útil a comparação do desempenho de cada um deles com a estrutura da interação que apresenta cada matriz de dados original completa. Para isso, encontraram-se o número de componentes multiplicativos dos modelos AMMI que deveriam ser usados. Para encontrar esse número, foi utilizada a validação cruzada '*leave-one-out*' por autovetor. O método sugerido por Bro *et al.* (2008) se descreve a seguir: Seja **X** a matriz de interação de dimensão (I×J), a qual pode ser padronizada (Dias e Krzanowski, 2003):

1) Subdivide-se **X** em um certo número de grupos, deleta-se cada grupo por vez a partir dos dados e se calcula um modelo de análise de componentes principais (o qual pode ser expresso em termos de duas matrizes **T** e **P**) sobre os elementos remanescentes.

2) Para cada componente principal e k= 1,...,min(I-1,J-1) e para cada grupo i=1,..., I, então

a) Para a coluna (ou colunas deletadas) j=1,..., J:

i) Estimar $t^{(-j)T} = X_i^{(-j)T} P^{(-j)} \left(P^{(-j)T} P^{(-j)} \right)^{-1}$, em que $P^{(-j)}$ é a matriz de ponderações **P** encontrada no passo 1 com a j-ésima linha excluída. $X_i^{(-j)T}$ é um vetor linha contendo a i-ésima linha de **X** sem o j-ésimo elemento.

ii) Estimar o elemento $x^{(ij)}$ através de $\hat{x}_{ij}^{(k)} = \mathbf{t}^{(-j)T} \mathbf{p}_j^T$, em que $\mathbf{p}_j$ é a j-ésima linha de **P**.

iii) Encontrar o erro de predição para o (i,j)-ésimo elemento $e_{ij}^{(k)} = x_{ij} - \hat{x}_{ij}^{(k)}$

b) Estimar a medida de discrepância entre o valor atual (da matriz original) e o valor predito como $\text{PRESS}(k) = \sum_{I} \sum_{J} \left(e_{ij}^{(k)} \right)^2$

Outro aspecto que deve ser considerado é o mecanismo de ausência dos dados. Geralmente, em situações que envolvem a avaliação de vários genótipos em diferentes ambientes podem ser encontradas observações ausentes seguindo a definição proposta por Little e Rubin (2002), isto é, completamente aleatórias (MCAR), aleatórias (MAR) e não aleatórias (MNAR). Valores faltantes completamente aleatórios podem ocorrer quando se têm, por exemplo, plantas danificadas devido a fatores não controláveis nos experimentos ou porque os dados foram digitados e mensurados erradamente. Nesse caso a causa da perda não está correlacionada com a variável que contêm a ausência. Entretanto, naqueles programas de teste de genótipos, nos quais as cultivares são escolhidas durante cada ano, baseadas somente nos dados que foram observados sem considerar aqueles dados não observados, o mecanismo de ausência é claramente aleatório (Piepho e Möhring, 2006). O último tipo de ausência, MNAR, pode ser visto usualmente quando o mesmo subconjunto de genótipos pode estar ausente em um número de ambientes da mesma sub-região, porque o melhorista de plantas no local não gosta desses genótipos. Assim, um genótipo ausente em um ambiente, possivelmente será também ausente em outros ambientes. Nesses casos o mecanismo que produz valores faltantes é naturalmente não aleatório.

Piepho (1995) estudou o comportamento de várias técnicas de imputação considerando valores faltantes MNAR, pelo qual nessa pesquisa foi avaliada uma alternativa diferente, quer dizer, o mecanismo MCAR. Para estudar os erros de predição associados aos algoritmos de imputação escolhidos em experimentos de interação G×A, foi desenvolvido um estudo de simulação baseado nos quatro conjuntos de dados reais por meio de um programa computacional em SAS/IML (SAS, 2004).

Cada matriz de dados foi submetida a retiradas aleatórias em diferentes porcentagens. Foram consideradas as porcentagens de 10, 20 e 40%, o processo foi repetido 1000 vezes para cada porcentagem em cada matriz, gerando 12000 conjuntos de dados incompletos. Em cada um desses conjuntos gerados foram encontrados os erros de predição utilizando a técnica de validação cruzada descrita a seguir. Da tabela de dados presentes, foi deletado um por vez, imputando o dado deletado e guardando a diferença entre a estimativa e o dado atual para a casela sob consideração. Isto foi feito para todas as caselas presentes e depois foi calculada a média das diferenças ao quadrado. Denote essa quantidade por D. D contém dois componentes de variação: um devido à inexatidão preditiva da estimativa ou imputação e o outro devido ao erro amostral dos dados presentes. Por essa razão, D deve ser corrigida subtraindo uma estimativa do erro da média (s^2). A raiz quadrada de ($D-s^2$) pode ser tomada como o erro de predição para um valor imputado. Em cada conjunto de dados simulado foram encontrados os erros de predição associados a cada método de imputação, mas para conseguir ver facilmente as diferenças entre os diferentes métodos foi computada a média e o desvio padrão para calcular os erros de predição padronizados. Além disso, no estudo de simulação foram avaliados os erros de predição por meio do cálculo de intervalos de imputação de 95% para as observações ausentes da seguinte maneira: $\hat{\theta} \pm z_{1-\alpha} \sqrt{(D-s^2)}$, em que $\hat{\theta}$ representa a imputação e $z_{1-\alpha}$, com α=0,05. Dado que se tinham as matrizes completas dos experimentos, foi calculado o número modal de intervalos que continham o respectivo dado real. Por exemplo, no conjunto de dados de eucalipto, uma retirada de 20% é equivalente a retirar 28 dados, para cada dado retirado foi encontrado um intervalo de imputação e depois de obter os 28 intervalos, foram verificados quantos deles continham o dado real. O número modal foi calculado sobre mil repetições.

Resultados e Discussão

Na tabela I se encontra a informação sobre a escolha do número de componentes multiplicativos (para explicar a interação G×A) do modelo AMMI que pode ser utilizado nos conjuntos de dados originais, sobre os quais foi feito posteriormente o estudo de simulação. Para escolher o modelo foi utilizada avaliação preditiva por validação cruzada e o melhor modelo será aquele que apresente a menor estatística PRESS. Pode se observar que nos dados de eucalipto um modelo apropriado pode ser o AMMI2, nos dados de ervilha e algodão um modelo AMMI1 e finalmente, nos dados de feijão um modelo AMMI4 pode ser o mais apropriado.

Com relação ao estudo de simulação, nas figuras são apresentados os gráficos de caixas dos erros de predição padronizados associados aos métodos de imputação em cada conjunto de dados considerado. Na Figura 1 se mostram os resultados do estudo de simulação nos dados de ervilha e observa-se que em todas as porcentagens de retirada aleatória se tem uma distribuição assimétrica à direita quando se imputaram os dados com um modelo aditivo sem interação, uma distribuição aproximadamente simétrica quando se imputaram com r-AMMI1, e uma distribuição assimétrica à esquerda quando se imputaram com r-AMMI2 e r-AMMI3. Segundo essa figura, a menor mediana dos erros preditivos padronizados foi encontrada com a imputação por meio de um modelo aditivo. Entretanto, na Figura 2 se têm os resultados obtidos nos dados de algodão e observa-se que em todas as porcentagens de retirada aleatória se tem uma distribuição aproxi-

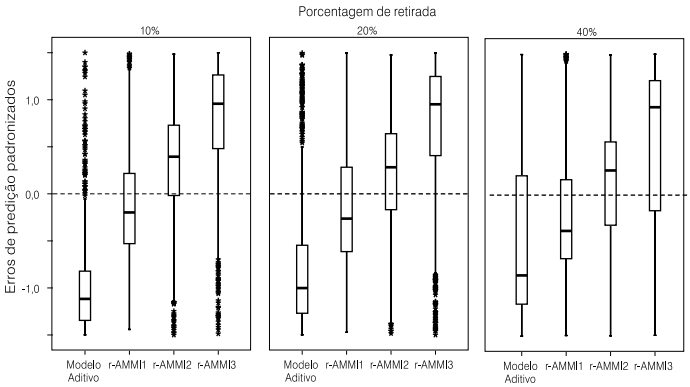

Figura 1. Distribuição dos erros de predição padronizados associados aos métodos de imputação nos dados de ervilha.

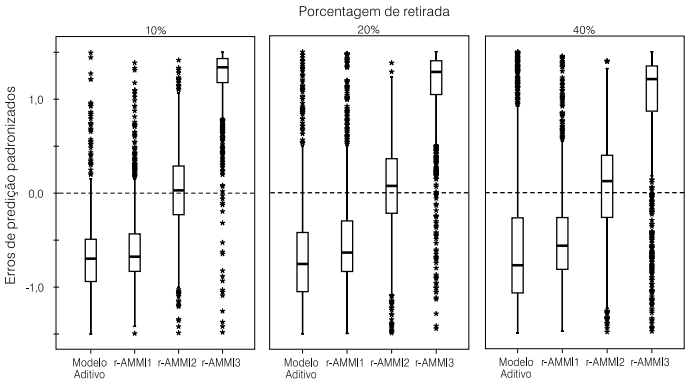

Figura 2. Distribuição dos erros de predição padronizados associados aos métodos de imputação nos dados de algodão.

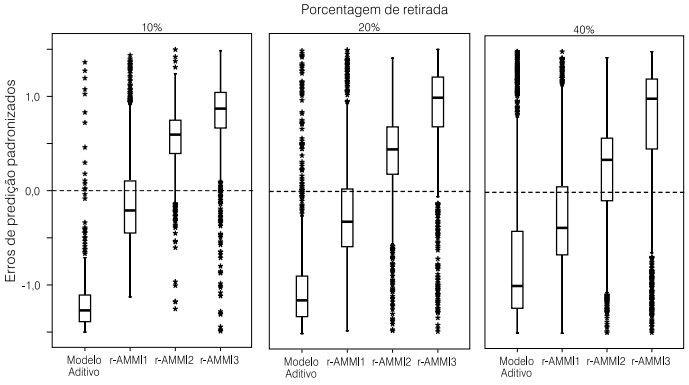

Figura 3. Distribuição dos erros de predição padronizados associados aos métodos de imputação nos dados de eucalipto.

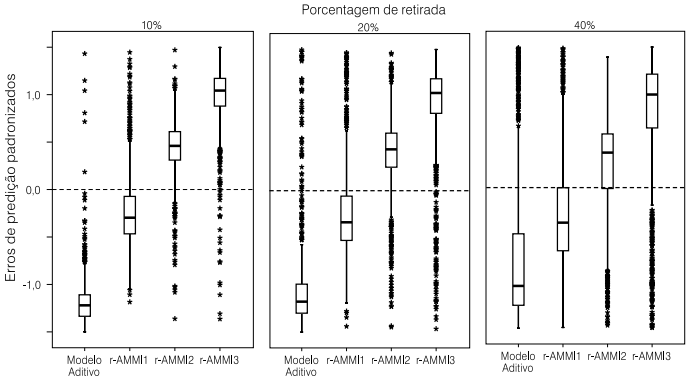

Figura 4. Distribuição dos erros de predição padronizados associados aos métodos de imputação nos dados de feijão.

TABELA I
VALOR DA SOMA DE QUADRADOS DA PREDIÇÃO (PRESS) UTILIZANDO VALIDAÇÃO CRUZADA POR AUTOVETOR NA ESCOLHA DO MODELO AMMI PARA EXPLICAR A INTERAÇÃO NOS CONJUNTOS DE DADOS ORIGINAIS COMPLETOS

	PRESS			
Modelo	Eucalipto	Ervilha	Feijão	Algodão
AMMI1	75,11086	**156,0466**	113,4852	**19,80334**
AMMI2	**73,81755**	161,0563	125,1817	20,66733
AMMI3	100,35854	177,8808	119,7403	22,939
AMMI4	134,39137	242,7536	**108,7088**	24,3314
AMMI5	575,08776	316,3643	141,2513	26,04652
AMMI6	56146,33573	317,2348	171,0122	29,48473
AMMI7		875,7188	289,9653	36,90663
AMMI8		12304,1671	718,3669	54,182
AMMI9			1297,3257	64,25948
AMMI10			5204,157	80,85339
AMMI11			20406,6026	73,46169
AMMI12				132,6654
AMMI13				381,61369
AMMI14				373582,7168

madamente simétrica quando se usou o método r-AMMI2. Nessa figura é possível concluir que a menor mediana dos erros de predição foi obtida com um modelo aditivo e com r-AMMI1. Nas Figuras 3 e 4 são apresentados os resultados correspondentes aos dados de eucalipto e de feijão. Nesses dados a menor mediana dos erros de predição foi detectada com o método do modelo aditivo, e para as porcentagens de 10 e 20% todos os algoritmos de imputação tiveram uma distribuição aproximadamente simétrica.

Na Tabela II são apresentadas as principais estatísticas, tais como a média, mediana, variância e distância interquartil (Q3-Q1) dos erros de predição padronizados associados a cada método de imputação. Os melhores erros de predição são aqueles que tenham os menores valores das estatísticas na escala padronizada e conjuntamente possuam uma dispersão pequena. Em geral, para os quatro conjuntos de dados considerados os melhores erros foram obtidos por meio dos métodos do modelo aditivo e r-AMMI1. Por exemplo, nos dados de ervilha com uma retirada aleatória de 10% as médias para esses métodos foram -0,9863 e -0,1034, enquanto que para os métodos r-AMMI2 e r-AMMI3 foram 0,3152 e 0,7744. Um comportamento parecido ocorreu com os demais conjuntos de dados nas diferentes porcentagens de retirada aleatória.

Para verificar as diferenças dos erros de predição entre os métodos de imputação considerados, foi feito o teste não paramétrico de Friedman em cada porcentagem de retirada de cada conjunto de dados. Os valores da estatística $T_{Friedman}$ (Sprent e Smeeton, 2001) no conjunto de dados de ervilha foram: 1096,88 ($p<0,0001$) para 10%, 528,25 ($p<0,0001$) para 20% e 217,22 ($p<0,0001$) para 40% de retirada dos dados. No conjunto de dados de algodão os valores da $T_{Friedman}$ foram: 2141,62 ($p<0,0001$) para 10%, 1309,72 ($p<0,0001$) para 20% e 740,39 ($p<0,0001$) para 40% de retirada dos dados. Entretanto, no conjunto de dados de feijão os valores correspondentes à $T_{Friedman}$ foram: 5880,63 ($p<0,0001$) para 10%, 2105,67 ($p<0,0001$) para 20% e 613,27 ($p<0,0001$) para 40%. Por ultimo, no conjunto de dados de eucalipto os valores de $T_{Friedman}$ foram 2840,67 ($p<0,0001$), 1467,33 ($p<0,0001$) e 476,69 ($p<0,0001$) para as porcentagens de retirada 10, 20 e 40% respectivamente.

TABELA II
ESTATÍSTICAS DOS ERROS DE PREDIÇÃO PADRONIZADOS ASSOCIADOS AOS MÉTODOS DE IMPUTAÇÃO

Conjunto de dados	% de retirada	Estatísticas	Modelo aditivo	r-AMMI1	r-AMMI2	r-AMMI3
Ervilha	10%	Média	-0,9863	-0,1034	0,3152	0,7744
		Variância	0,2580	0,3267	0,3267	0,4075
		Mediana	-1,1165	-0,1979	0,3950	0,9580
		Q3-Q1	0,5236	0,7462	0,7467	0,7825
	20%	Média	-0,7606	-0,1330	0,1969	0,6967
		Variância	0,4957	0,4289	0,3918	0,5651
		Mediana	-1,0015	-0,2634	0,2825	0,9519
		Q3-Q1	0,7226	0,8969	0,8076	0,8409
	40%	Média	-0,4416	-0,2192	0,1334	0,5274
		Variância	0,8373	0,4635	0,3875	0,7753
		Mediana	-0,8532	-0,3805	0,2622	0,9346
		Q3-Q1	1,3642	0,8393	0,8849	1,3824
Algodão	10%	Média	-0,6813	-0,5807	0,0363	1,2257
		Variância	0,1875	0,1775	0,1884	0,1424
		Mediana	-0,6970	-0,6762	0,0296	1,3399
		Q3-Q1	0,4495	0,3991	0,5181	0,2557
	20%	Média	-0,6577	-0,4971	0,0458	1,1090
		Variância	0,3295	0,2925	0,2280	0,2394
		Mediana	-0,7558	-0,6351	0,0729	1,2832
		Q3-Q1	0,6282	0,5366	0,5791	0,3583
	40%	Média	-0,5191	-0,4700	0,0392	0,9498
		Variância	0,6095	0,2764	0,2935	0,4283
		Mediana	-0,7725	-0,5646	0,1229	1,2090
		Q3-Q1	0,7992	0,5515	0,6617	0,4807
Eucalipto	10%	Média	-1,2051	-0,1300	0,5425	0,7926
		Variância	0,0934	0,2242	0,1056	0,1859
		Mediana	-1,2700	-0,2101	0,5944	0,8706
		Q3-Q1	0,2825	0,5524	0,3535	0,3777
	20%	Média	-0,9954	-0,2215	0,3829	0,8340
		Variância	0,2883	0,2882	0,2142	0,3282
		Mediana	-1,1489	-0,3195	0,4438	0,9889
		Q3-Q1	0,4275	0,6087	0,4995	0,5251
	40%	Média	-0,6510	-0,2491	0,2053	0,6948
		Variância	0,7011	0,3942	0,3294	0,5666
		Mediana	-0,9968	-0,3785	0,3461	0,9966
		Q3-Q1	0,8186	0,7242	0,6657	0,7441
Feijão	10%	Média	-1,1844	-0,2282	0,4360	0,9766
		Variância	0,0666	0,1388	0,0856	0,1104
		Mediana	-1,2205	-0,2966	0,4600	1,0431
		Q3-Q1	0,2267	0,3948	0,3001	0,2918
	20%	Média	-1,0388	-0,2379	0,3923	0,8844
		Variância	0,2526	0,2325	0,1666	0,2775
		Mediana	-1,1766	-0,3350	0,4388	1,0364
		Q3-Q1	0,3074	0,4696	0,3609	0,3663
	40%	Média	-0,6834	-0,2855	0,2256	0,7433
		Variância	0,7078	0,3323	0,2993	0,5104
		Mediana	-1,0480	-0,3735	0,3729	0,9924
		Q3-Q1	0,7623	0,6716	0,5819	0,5734

Uma vez confirmado que pelo menos um algoritmo de imputação tem um parâmetro de centralidade diferente dos outros três algoritmos, fizeram-se comparações múltiplas confrontando os métodos de imputação dois a dois, encontrando que no conjunto de dados de ervilha existem diferenças significativas dos erros de predição padronizados entre os quatro métodos de imputação. No conjunto de dados de algodão não se encontraram diferenças significativas dos erros de predição entre os métodos que usaram um modelo aditivo e o r-AMMI1 para as porcentagens de retirada de 10 e 40%. Entretanto, nos dados de feijão e de eucalipto todos os métodos considerados tiveram diferenças estatísticas significativas. Similarmente, foi testada a hipótese de homogeneidade de variâncias nos erros de predição por meio da estatística de Levene, encontrando que somente se têm variâncias homogêneas no conjunto de dados de algodão quando se imputou 10%, o valor da estatística foi 2,24 (p= 0,0819).

Os erros de predição por meio de validação cruzada foram utilizados no cálculo de intervalos de confiança de 95% para as imputações e os resultados são apresentados na Tabela III. Segundo essa tabela os quatro métodos de predição apresentam um desempenho muito parecido e por isso uma escolha lógica do método de imputação seria o modelo aditivo por causa de simplicidade. Observa-se, por exemplo, no conjunto de dados de feijão que uma retirada aleatória de 20% é equivalente a 36 dados e no estudo de simulação sobre 1000 repetições do processo, o número modal de intervalos que continham o verdadeiro dado foi 34. Assim, ~94,4% dos intervalos contém o dado real. No conjunto de dados de eucalipto, uma retirada aleatória de 10% é equivalente a 14 dados e depois das 1000 repetições o número modal de intervalos contendo o dado real foi também 14, quer dizer que nesse caso ~100% dos intervalos contém o dado verdadeiro. Note-se que uma interpretação similar pode ser feita para os outros conjuntos de dados nas outras porcentagens de perda aleatória.

Conclusões

Neste estudo dos erros de predição, tiveram-se diferentes estruturas segundo os métodos utilizados para resolver o problema de desbalanceamento em experimentos com interação, mas sempre foram minimizados quando se imputou por meio de um modelo aditivo sem interação e naturalmente, com erros pequenos a distância entre os limites dos intervalos de imputação será também pequena e a qual, é uma característica desejável nessa classe de estudos estatísticos.

REFERÊNCIAS

Arciniegas-Alarcón S (2008) *Imputação de Dados em Experimentos com Interação Genótipo por Ambiente: Uma Aplicação a Dados de Algodão*. Tese. Universidade de São Paulo. Brasil. 82 pp.

Arciniegas-Alarcón S, Dias CTS (2009a) Imputaçaõ de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão. *Rev. Bras. Biometr. 27*: 125-138.

Arciniegas-Alarcón S, Dias CTS (2009b) Análise AMMI com dados imputados em experimentos de interação genótipo x ambiente de algodão. *Pesq. Agropec. Brás. 44*: 1391-1397.

Arciniegas-Alarcón S, García-Peña M, Dias CTS, Krzanowski WJ (2010) An alternative methodology for imputing missing data in trials with genotype-by-environment interaction. *Biometr. Lett.* 47: 1-14.

Bergamo GC, Dias CTS, Krzanowski WJ (2008) Distribution-free multiple imputation in an interaction matrix through singular value decomposition. *Sci. Agric. 65*: 422-427.

Bro R, Kjeldahl K, Smilde AK, Kiers HAL (2008) Cross-validation of component models: a critical look at current methods. *Anal. Bioanal. Chem. 390*: 1241-1251.

Calinski T, Czajka S, Denis JB, Kaczmarek Z (1992) EM and ALS algorithms applied to estimation of missing data in

TABELA III
NÚMERO MODAL DE INTERVALOS DE PREDIÇÃO DE 95% PARA OBSERVAÇÕES AUSENTES QUE CONTÉM O RESPECTIVO DADO REAL RETIRADO NO ESTUDO DE SIMULAÇÃO SOBRE 1000 REPETIÇÕES

Conjunto de dados	% de retirada	Número de dados retirados	Modelo aditivo	r-AMMI1	r-AMMI2	r-AMMI3
Ervilha	10%	17	17	16	16	17
	20%	33	31	31	31	30
	40%	65	61	62	61	61
Algodão	10%	41	39	39	39	40
	20%	81	76	77	77	77
	40%	162	151	151	151	154
Eucalipto	10%	14	14	14	14	14
	20%	28	27	27	27	27
	40%	56	53	53	53	54
Feijão	10%	18	17	17	17	17
	20%	36	34	34	34	34
	40%	72	69	69	69	69

series of variety trials. *Biuletyn Oceny Odmian 24-25*: 7-31.

Calinski T, Czajka S, Denis JB, Kaczmarek Z (1999) Further study on estimating missing values in series of variety trials. *Biuletyn Oceny Odmian 30*: 7-38.

Calinski T, Czajka S, Kaczmarek Z, Krajewski P, Pilarczyk W (2009) Analyzing the genotype-by-environment interactions under a randomization-derived mixed model. *J. Agric. Biol. Env. Stat. 14*: 224-241.

Denis JB (1991) Ajustements de modèles linéaires ei bilinéaires sous contraintes linéaires avec données manquantes. *Rev. Stat. Appl.* 39: 5-24.

Denis JB, Baril CP (1992) Sophisticated models with numerous missing values: the multiplicative interaction model as an example. *Biuletyn Oceny Odmian 24-25*: 33-45.

Dias CTS, Krzanowski WJ (2003) Model selection and cross validation in additive main effect and multiplicative interaction models. *Crop Sci.* 43: 865-873.

Farias FJC (2005) Índice de Seleção em Cultivares de Algodoeiro Herbáceo. Tese. Universidade de São Paulo. Brasil. 121 pp.

Flores F, Moreno MT, Cubero JI (1998) A comparison of univariate and multivariate methods to analyze G x E interaction. Field Crops Res. 56: 271-286.

Freeman HG (1975) Analysis of interactions in incomplete two-ways tables. *Appl. Stat.* 24: 46-55.

Gauch HG (2006) Statistical Analysis of Yield Trials by AMMI and GGE. *Crop Sci*. 46: 1488-1500.

Gauch HG, Zobel RW (1990) Imputing missing yield trial data. *Theor. Appl. Genet. 79*: 753-761.

Godfrey AJR, Wood GR, Ganesalingam S, Nichols MA, Qiao CG (2002) Two-stage clustering in genotype-by-environment analyses with missing data. *J. Agric. Sci.* 139: 67-77.

Kang MS, Balzarini MG, Guerra JLL (2004) Genotype-by-environment interaction. Em Saxton AM (Ed.). *Genetic Analysis of Complex Traits Using SAS*. SAS Institute Inc., Cary, NC, EEUU. pp. 69-96.

Lavoranti OJ (2003) *Estabilidade E Adaptabilidade Fenotípica Através da Reamostragem "Bootstrap" no Modelo AMMI*. Tese. Universidade de São Paulo. Brasil. 184 pp.

Little RJ, Rubin DB (2002) *Statistical Analysis with Missing Data*. 2ª. ed. Wiley. Nova Iorque, EEUU. 381 pp.

Mandel J (1993) The analysis of two-way tables with missing values. *Appl. Stat. 42*: 85-93.

Piepho HP (1995) Methods for estimating missing genotype-location combinations in multilocation trials - an empirical comparison. *Inf. Biometr. Epidemiol. Med. Biol.* 26: 335-349.

Piepho HP, Möhring J (2006) Selection in cultivar trials-Is it ignorable? *Crop Sci. 46*: 192-201.

Romagosa I, Voltas J, Malosetti M, van Eeuwijk FA (2008) Interacción genotipo por ambiente. Em Ávila CM, Atienza SG, Moreno MT, Cubero JI (Eds.) *La Adaptación al Ambiente y los Estreses Abióticos en la Mejora Vegetal*. Instituto de Investigación y Formación Agraria y Pesquera. Junta de Andalucía. España. pp. 107-136.

SAS (2004) *SAS/IML 9.1 User.s guide*. SAS Institute Inc. Cary, NC, EEUU. 1040 pp.

Sprent P, Smeeton NC (2001) *Applied Nonparametric Statistical Methods*. 3ª ed. Chapman and Hall. Boca Raton, FL, EEUU. 463 pp.

van Eeuwijk FA, Kroonenberg PM (1998) Multiplicative models for interaction in three-way ANOVA, with applications to plant breeding. *Biometrics 54*: 1315-1333.

van Eeuwijk FA, Malosetti M, Yin X, Struik PC, Stam P (2005) Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. *Austr. J. Agric. Res. 56*: 883-894.

van Eeuwijk FA, Malosetti M, Boer MP (2007) Modelling the genetic basis of response curves underlying genotype×environment interaction. In: Spiertz JHJ, Struik PC, van Laar HH (Eds.) *Scale and Complexity in Plant Systems Research: Gene-Plant-Crop Relations*. Springer, Wageningen UR Frontier Series. Nova Iorque, EEUU. pp. 115-126.