IMPUTAÇÃO DE DADOS EM EXPERIMENTOS COM INTERAÇÃO GENÓTIPO POR AMBIENTE: UMA APLICAÇÃO A DADOS DE ALGODÃO

Sergio ARCINIEGAS-ALARCÓN[1]
Carlos Tadeu dos Santos DIAS[2]

- RESUMO: Um problema comum em experimentos multiambientes são as combinações ausentes genótipo-ambiente. Recentemente, Bergamo propôs um método de imputação múltipla livre de distribuição na matriz de interação. O objetivo desse artigo é avaliar o novo desenvolvimento comparando-o com metodologias que já tiveram sucesso nos experimentos genótipo-ambiente com dados faltantes, como os mínimos quadrados alternados e as estimativas robustas, usando os modelos de efeitos aditivos com interação multiplicativa (AMMI). Foi feito um estudo de simulação baseado em dados reais, fazendo retiradas aleatórias de dados considerando diferentes porcentagens, imputando as observações e comparando as metodologias através de três critérios: a raiz quadrada da diferença preditiva média, a estatística de similaridade de Procrustes e o coeficiente de correlação de Spearman. Concluiu-se que a imputação múltipla não é melhor do que a imputação, baseada em um modelo aditivo sem interação e em termos de dispersão, os sub-modelos robustos oferecem os melhores resultados. Todos os métodos considerados apresentaram uma alta correlação entre os valores preditos e os valores verdadeiros.
- PALAVRAS-CHAVE: observações ausentes; imputação de dados; modelos AMMI; interação genótipo-ambiente.

1 Introdução

Muitas vezes os experimentos multiambientes são desbalanceados e vários genótipos não são testados em alguns ambientes. Para as recomendações de ambientes pode ser de interesse obter estimativas do desempenho de combinações que não foram testadas. Tais estimativas podem ser calculadas usando a informação daquelas combinações genótipo por ambiente (G×E) que foram atualmente observadas, além disso, é bem conhecido que uma das melhores opções na análise da interação (G×E) são os modelos de efeitos aditivos de interação multiplicativa (AMMI), pois exploram melhor as informações contidas nos dados do que a ANOVA tradicional (Duarte e Vencovsky, 1999), mas esses modelos têm alguns problemas na estimação dos parâmetros se existirem dados faltantes (Denis e Baril, 1992). Por exemplo, na estimação clássica dos modelos AMMI é preciso

1 Programa de Pós-Graduação em Estatística e Experimentação Agronômica, Escola Superior de Agricultura "Luiz de Queiroz" – ESALQ, Universidade de São Paulo – USP, Caixa Postal 9, CEP: 13418-900, Piracicaba, SP, Brasil. E-mail: *salarcon@esalq.usp.br*
2 Departamento de Ciências Exatas, Escola Superior de Agricultura "Luiz de Queiroz" – ESALQ, Universidade de São Paulo – USP, Caixa Postal 9, CEP: 13418-900, Piracicaba, SP, Brasil. E-mail: *ctsdias@esalq.usp.br*

encontrar a decomposição por valores singulares (DVS) da matriz de resíduos não aditivos, mas essa DVS não pode ser calculada na ocorrência de valores ausentes.

Uma abordagem para resolver o problema de dados faltantes em experimentos consiste em fazer modelos e considerações teóricas sobre as funções estimáveis, baseando-se unicamente nos dados presentes. Uma completa revisão dessa abordagem foi feita por Birkes et al. (1976), Dodge (1985) e Dodge e Zoppe (2004). Uma alternativa diferente no tratamento das observações ausentes foi apresentada por Rubin (1996), Little e Rubin (2002) e Schafer e Graham (2002), que recomendam usar métodos paramétricos de imputação múltipla ou métodos baseados em verossimilhança (por exemplo, com o algoritmo EM).

O problema também foi tratado especificamente para os experimentos G×E. 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, mas algumas alternativas desse procedimento usando estatística multivariada foram descritas em Goodfrey et al. (2002). Outros estudos que são recomendados por Van Eeuwijk e Kroonenberg (1998) no caso de dados faltantes para experimentos G×E com resultados razoavelmente bons, foram os desenvolvidos por Denis e Baril (1992) e Calinski et al. (1992). Eles encontraram que usando imputações através de modelos AMMI usando sub-modelos robustos ou baseados em mínimos quadrados alternados podem-se obter resultados tão bons como os encontrados com um algoritmo EM.

Recentemente, Bergamo (2007) e Bergamo et al. (2008) propuseram um método baseado em imputação múltipla livre de distribuição, o qual pode ser aplicado em matrizes de interação G×E com informação incompleta. Dada a informação histórica sobre imputação de dados em experimentos e especificamente em experimentos de dois fatores G×E decidiu-se fazer uma comparação das metodologias propostas por Bergamo (2007) e Bergamo et al. (2008), Calinski et al. (1992) e Denis e Baril (1992).

2 Material e métodos

2.1 Características dos dados

Os dados utilizados foram obtidos do Ensaio Estadual de Algodoeiro Herbáceo referente ao ano agrícola 2000/01, do programa de melhoramento do algodoeiro para as condições do Cerrado. Os experimentos foram avaliados em 27 localidades dos estados brasileiros de Mato Grosso, Mato Grosso do Sul, Goiás, Minas Gerais, Rondônia, Maranhão e Piauí. O delineamento experimental utilizado foi o aleatorizado em blocos completos, com 15 cultivares e quatro repetições. A parcela experimental foi constituída por quatro fileiras de 5 m de comprimento, com espaçamento de 0,80 m entre fileiras e uma densidade de sete plantas por metro linear. A área útil da parcela foi composta pelas duas fileiras centrais (Farias, 2005). A variável estudada foi produtividade de algodão em caroço (kg/ha) e para este trabalho só foram consideradas as médias das repetições de produtividade para cada genótipo em cada um dos locais.

2.2 Métodos de imputação

A imputação múltipla livre de distribuição (IMLD) proposta por Bergamo (2007) e Bergamo et al. (2008) consiste em um esquema iterativo usando a decomposição por valor

singular (DVS) de uma matriz para predizer as observações ausentes em uma matriz $\boldsymbol{Y}$ de dimensão (n×p), com p<n, se p<n a matriz deve ser transposta.

Para melhor entendimento, considere somente um valor perdido y_{ij} em $\boldsymbol{Y}$. Deve-se omitir a i-ésima linha de $\boldsymbol{Y}$ e calcular a decomposição por valor singular da matriz resultante de dimensão ((n-1)×p) denotada por $\boldsymbol{Y}^{(-i)}$, em que $\boldsymbol{Y}^{(-i)} = \bar{\boldsymbol{U}}\bar{\boldsymbol{D}}\bar{\boldsymbol{V}}^{T}$, $\bar{\boldsymbol{U}} = (\bar{u}_{sh})$, $\bar{\boldsymbol{V}} = (\bar{v}_{sh})$, $\bar{\boldsymbol{D}} = (\bar{d}_1, \ldots, \bar{d}_p)$. O passo seguinte consiste em omitir a j-ésima coluna de $\boldsymbol{Y}$ e obter a decomposição por valores singulares (DVS) da matriz resultante de dimensão (n×(p-1)) denotada por $\boldsymbol{Y}_{(-j)}$, em que $\boldsymbol{Y}_{(-j)} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{V}}^{T}$, $\tilde{\boldsymbol{U}} = (\tilde{u}_{sh})$, $\tilde{\boldsymbol{V}} = (\tilde{v}_{sh})$, $\tilde{\boldsymbol{D}} = (\tilde{d}_1, \ldots, \tilde{d}_{p-1})$. As matrizes $\bar{\boldsymbol{U}}$, $\bar{\boldsymbol{V}}$, $\tilde{\boldsymbol{U}}$ e $\tilde{\boldsymbol{V}}$ são matrizes ortonormais e $\tilde{\boldsymbol{D}}$ e $\bar{\boldsymbol{D}}$ são matrizes diagonais. Agora, combinando as duas DVS, $\boldsymbol{Y}^{(-i)}$ e $\boldsymbol{Y}_{(-j)}$, obtém-se o valor imputado através de

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

Bergamo et al. (2008) afirmam que com 5 imputações para cada observação ausente é suficiente para conhecer a variabilidade entre imputações, por essa razão sugerem usar $b = 20$, $\tilde{a} = 8,9,10,11,12$ e $\bar{a} = 12,11,10,9,8$ tal que $\bar{a} + \tilde{a} = b$. Cada combinação entre esse valores produzem uma imputação diferente.

Para mais de um valor perdido, um esquema iterativo é envolvido como segue: Os valores são imputados inicialmente pela média da respectiva coluna obtendo uma matriz $\boldsymbol{Y}$ completada e posteriormente padronizando as colunas, subtraindo de cada elemento m_j e dividindo o resultado por s_j (em que m_j e s_j representam a média e desvio padrão da j-ésima coluna calculados somente sobre os valores observados). Sobre a matriz padronizada é recalculada a imputação de cada valor ausente usando-se (1). Para os cálculos de cada estimativa são necessárias $\boldsymbol{Y}^{(-i)}$ e $\boldsymbol{Y}_{(-j)}$, as quais devem ser também padronizadas. Finalmente, a matriz $\boldsymbol{Y}$ deve ser retornada à sua escala original assim, $y_{ij} = m_j + s_j \hat{y}_{ij}$. Então, o processo é iterado até alcançar estabilidade nas imputações. Neste trabalho foi usada como estimativa dos dados faltantes a média das cinco imputações.

O segundo método de imputação envolvido na comparação foram os mínimos quadrados alternados (ALS) propostos por Calinski et al. (1992) e modificados por Piepho (1995). O método utiliza os modelos AMMI, muito usados na análise de dados G×E, cuja equação é dada por:

$$y_{ij} = \mu + \tau_i + \beta_j + \lambda_1 \alpha_{1i} \gamma_{1j} + \lambda_2 \alpha_{2i} \gamma_{2j} + \cdots + \varepsilon_{ij} \qquad (2)$$

(i=1,...,t; j=1,...,b) em que μ, τ_i, β_j representam a média geral e os efeitos genotípicos e ambientais, ε_{ij} é o termo do erro associado ao i-ésimo genótipo no j-ésimo ambiente, e λ_r, α_{1r} e $\gamma_{1r} (r = 1, \ldots)$ são estimados pela DVS da matriz de resíduos depois de ajustar a parte aditiva (Dias e Krzanowski, 2006). λ_r é estimado pelo r-ésimo valor singular da DVS, α_{1r} e γ_{1r} são estimados pelos correspondentes autovetores genotípicos e

ambientais correspondentes a λ_r. Dependendo do número de termos multiplicativos os modelos são chamados AMMI0, AMMI1 etc.

Assim, a idéia principal dos ALS consiste na estimação dos parâmetros do modelo (2) considerando alguns deles como conhecidos e estimar os parâmetros restantes por mínimos quadrados ordinários. O modelo (2) pode ser reescrito como

$$y_{ij} - \mu - \tau_i = \beta_j + \lambda_1 \alpha_{1i} \gamma_{1j} + \lambda_2 \alpha_{2i} \gamma_{2j} + \cdots + \varepsilon_{ij}.$$

Considerando como conhecidos todos os parâmetros, exceto os parâmetros ambientais β_j, γ_{1j} e γ_{2j} considerados como desconhecidos, o problema de estimação se reduz a um problema de regressão múltipla. Esse passo é conhecido como o **passo-linha**.

De outra maneira, o modelo (2) pode ser reescrito como

$$y_{ij} - \mu + \beta_j = \tau_i + \lambda_1 \alpha_{1i} \gamma_{1j} + \lambda_2 \alpha_{2i} \gamma_{2j} + \cdots + \varepsilon_{ij}$$

Considerando somente como desconhecidos os parâmetros genotípicos τ_i, α_{1i} e α_{2i}, o problema de estimação se reduz novamente a um problema de regressão múltipla. Esse passo é conhecido como o **passo-coluna.** A regressão deve ser feita somente sobre os dados observados. O algoritmo consiste em alternar o **passo-linha** e o **passo-coluna** até alcançar convergência. Entre os dois passos os dados são normalizados aplicando a estimação AMMI clássica. Dependendo do número de termos multiplicativos no modelo o algoritmo é denotado por ALS(0), ALS(1), ALS(2),... . Calinski et al. (1992) encontraram que o método ALS(0) pode ter melhor desempenho do que os métodos ALS(1) e ALS(2), além disso, na pesquisa desenvolvida por Piepho (1995) o método ALS(2) apresentou vários problemas de convergência e fica claro que usando os ALS, os componentes multiplicativos não deveriam ser maiores do que um e por essa razão no presente estudo foram considerados somente os algoritmos ALS(0) e ALS(1) descritos em Piepho (1995):

O método ALS(0) consiste em resolver o seguinte sistema de equações por mínimos quadrados alternados (ALS).

a) $y_{ij} - \hat{\mu} - \hat{\tau}_i = \beta_j$ (Resolver para β_j)

b) $y_{ij} - \hat{\mu} + \hat{\beta}_j = \tau_i$ (Resolver para τ_i)

Uma vez resolvido o sistema, normaliza-se calculando uma tabela de dupla entrada completa a partir dos parâmetros estimados atuais e se faz a análise AMMI0.

Para o método ALS(1), deve-se seguir os passos:

1) Preencher os dados ausentes com as imputações obtidas por ALS(0) e sobre a tabela de dados completada estimar os parâmetros do modelo AMMI1. As estimativas desses parâmetros serão os valores inicias para começar o algoritmo.

2) Resolver o seguinte sistema de equações por mínimos quadrados alternados (ALS)

$$y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j = \hat{\lambda}_1 \hat{\alpha}_{1i} \gamma_{1j} \quad \text{(Resolver para } \gamma_{1j} \text{)}$$

$$y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j = \hat{\lambda}_1 \alpha_{1i} \hat{\gamma}_{1j} \quad \text{(Resolver para } \alpha_{1i} \text{)}$$

Depois de resolver as equações, normaliza-se pelo cálculo de uma tabela completa de dupla entrada a partir dos parâmetros estimados atuais e se faz análise AMMI1.

3) Resolver o seguinte sistema de equações por mínimos quadrados alternados (ALS)

$$y_{ij} - \hat{\mu} - \hat{\tau}_i = \beta_j + \hat{\lambda}_1 \hat{\alpha}_{1i} \gamma_{1j} \qquad \text{(Resolver para } \beta_j \text{ e } \gamma_{1j}\text{)}$$

$$y_{ij} - \hat{\mu} + \hat{\beta}_j = \tau_i + \hat{\lambda}_1 \alpha_{1i} \hat{\gamma}_{1j} \qquad \text{(Resolver para } \tau_i \text{ e } \alpha_{1i}\text{)}$$

Uma vez resolvido o sistema, nomaliza-se calculando uma tabela de dupla entrada completa a partir dos parâmetros estimados atuais e se faz análise AMMI1.

Os mínimos quadrados alternados (ALS) são baseados unicamente nos dados observados. Denis e Baril (1992) argumentaram e subseqüentemente demonstraram através de exemplo, que em alguns experimentos os ALS podem levar a maus ajustes para os dados faltantes em tabelas de dupla entrada. Suponha-se que os valores ausentes fossem conhecidos e pudessem ser calculadas as regressões para os dados incompletos e para os dados completos. Se essas linhas de regressão diferirem consideravelmente, a extrapolação a partir da regressão obtida com os dados incompletos poderia ser altamente inexata. Para resolver esse problema, Denis e Baril (1992) sugerem fazer as análises sobre tabelas completas de dupla entrada, em que os dados faltantes são substituídos pelas estimativas de algum sub-modelo robusto. Resultados empíricos indicaram que uma ponderação igual para os valores ausentes e observados é aceitável (Denis e Baril, 1992 e Piepho, 1995). Com uma ponderação igual as análises são equivalentes à análise AMMI clássica para dados completos. Para a análise AMMI1 Denis e Baril (1992) propuseram AMMI0 como um sub-modelo robusto, da mesma maneira o AMMI1 pode ser usado como um sub-modelo robusto para uma análise AMMI2 etc. Neste trabalho foram considerados os métodos r-AMMI1 e r-AMMI2.

Concluindo esta seção, os métodos comparados serão: a média das imputações multiplas-IMLD, mínimos quadrados alternantes-ALS(0) e ALS(1) e estimativas AMMI com sub-modelos robustos, r-AMMI1 e r-AMMI2.

2.3 Estudo de simulação

Para comparar os métodos de imputação foi desenvolvido um estudo de simulação baseado no conjunto de dados reais. O conjunto de dados contém 405 observações. Esse conjunto foi submetido a retiradas aleatórias de diferentes porcentagens de dados. Foram consideradas as porcentagens de 10%, 20% e 30%, o processo foi repetido 1000 vezes para cada porcentagem, para um total de 3000 retiradas aleatórias, ou seja, 3000 conjuntos de dados diferentes foram gerados. No primeiro caso (10%) foram retiradas 41 interações, no segundo caso (20%) foram retiradas 81 interações e finalmente no 30% foram retiradas 122 interações. Para cada um dos 3000 conjuntos de dados com dados faltantes simulados, os algoritmos de imputação IMLD, ALS(0), ALS(1), r-AMMI1 e r-AMMI2 foram aplicados para predizer os valores ausentes através de um programa computacional implementado em SAS/IML (SAS INSTITUTE, 2004).

2.4 Critérios de comparação

Para comparar as estimativas dos diferentes métodos de imputação foram levados em conta três critérios de comparação. Cada matriz de dados completada (observados+imputados) foi comparada com a matriz original através da rotação de Procrustes (Krzanowski, 2000). Com a estatística de Procrustes se obtém uma medida da

diferença entre duas configurações de pontos e o método de imputação que minimize essa diferença indicará os melhores resultados.

Com cada um dos métodos de imputação de dados, foram calculadas as estimativas dos valores perdidos e a diferença preditiva média (RMSPD) dessas estimativas com os valores verdadeiros em cada um dos 3000 conjuntos de dados. Note-se que a estatística é construída com a soma de quadrados das diferenças entre os dados originais e as correspondentes imputações em cada retirada e dividindo pelo número dessas diferenças. A raiz quadrada desse resultado é conhecida como a estatística RMSPD (Dias e Krzanowski, 2003) e normalmente é usada em validação cruzada para escolher o melhor modelo AMMI, mas neste caso foi adaptada ao estudo como medida de qualidade das estimativas. A estatística é dada por:

$$RMSPD = \sqrt{\frac{\sum_{i,j}(y_{ij} - \hat{y}_{ij})^2}{NA}}$$

Em que y_{ij} representa a produção média do i-ésimo genótipo no j-ésimo ambiente no conjunto de dados original, $\hat{y}_{ij}$ representa a estimativa do i-ésimo genótipo no j-ésimo ambiente usando cada um dos métodos de imputação considerados e NA é número total de valores ausentes (i=1,...,15, j=1,...,27). Quanto menor seja a RMSPD, melhor será o método de imputação. Em cada uma das mil retiradas aleatórias feitas para cada porcentagem de retirada considerada (10%, 20% e 30% de dados) foram obtidas as RMSPD depois de imputar por IMLD, ALS(1), ALS(0), r-AMMI2 e r-AMMI1, mas, para conseguir ver as diferenças entre os diferentes métodos foram computados a média e o desvio padrão para calcular as estatísticas RMSPD padronizadas e sobre as quais foi feita diretamente a comparação.

O último critério de comparação considerado nesse estudo foi o coeficiente de correlação de Spearman (Sprent e Smeeton, 2001). Foi calculado este coeficiente de correlação não paramétrico entre cada valor ausente e seu correspondente dado verdadeiro. Quanto maior seja a correlação entre os valores imputados e os valores originais melhor será o método de imputação. Usou-se essa medida não paramétrica para evitar problemas de distribuição nos dados, uma vez que o coeficiente de correlação de Pearson é fortemente dependente da distribuição normal das variáveis.

3 Resultados e discussão

Na Tabela 1 são apresentadas as médias da RMSPD padronizada para cada porcentagem considerada no estudo de simulação. Observa-se que o método que minimiza as médias da RMSPD padronizada em todas as porcentagens de retirada é o ALS(0) com uma média de -0,57 para 10%, -0,68 para 20% e -0,73 para 30% de retirada dos dados, enquanto o método que maximiza esses valores em todos os casos é o IMLD. Assim, segundo essa comparação o melhor método poderia ser aquele que oferece um valor pequeno da média da RMSPD padronizada e corresponde aos mínimos quadrados alternados com estimativas aditivas. Note-se que resultados muito próximos daquele método são proporcionados pelo r-AMMI1.

Na Figura 1 se mostra o gráfico de caixas considerando as mil retiradas aleatórias de 10% de dados, observa-se que no caso dos mínimos quadrados alternados se obteve uma

distribuição simétrica quando foram imputados os dados com ALS(0) e assimétrica a direita quando a predição foi por ALS(1), ou seja, considerando um componente de interação multiplicativa. Todas as metodologias apresentam dados discrepantes, pois se têm muitos valores afastados do corpo principal dos dados, mas parece que a menor variabilidade é obtida com o, algoritmo r-AMMI1. Pode-se concluir que a maior diferença preditiva mediana é alcançada com a predição de interações faltantes através da média das imputações múltiplas (IMLD), enquanto a menor RMSPD mediana pode ser atingida usando estimativas aditivas ou estimativas baseadas em sub-modelos robustos, isto é ALS(0) e r-AMMI1.

Tabela 1 - Médias da RMSPD padronizada

Métodos	Porcentagem de retirada		
	10%	20%	30%
ALS(1)	0,0456	0,3107	0,5009
ALS(0)	-0,5785	-0,6848	-0,7308
IMLD	1,4992	1,4531	1,3850
r-AMMI1	-0,5631	-0,6369	-0,6740
r-AMMI2	-0,4032	-0,4420	-0,4810

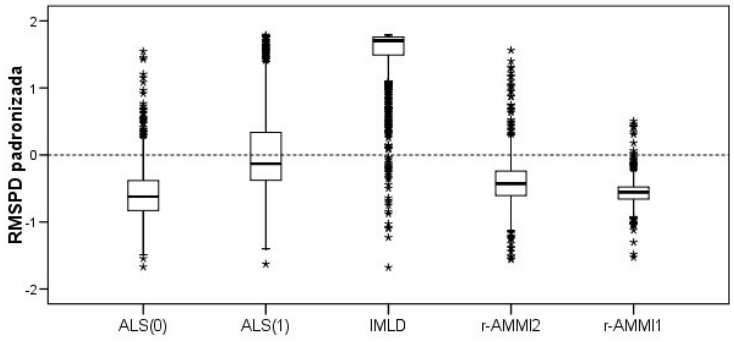

Figura 1 - Gráfico de caixas da distribuição da RMSPD padronizada com 10% de retirada dos dados.

Segundo a Tabela 2, a menor variância da RMSPD padronizada foi encontrada usando o método r-AMMI1 em todas as porcentagens de retirada. Para 10%, a menor variância da RMSPD foi 0,03, para 20% foi 0,02 e para 30% de retirada dos dados, a menor variância foi 0,01. Nesta análise, a variância foi maximizada nas diferentes porcentagens através dos mínimos quadrados alternados usando como modelo de imputação o AMMI1, ou seja, com o algoritmo ALS(1). A proposta baseada em imputação múltipla, IMLD, tem variâncias maiores do que os métodos ALS(0) e r-AMMI2. Até este ponto a nova metodologia IMLD não apresenta um desempenho melhor do que as predições baseadas em modelos AMMI.

Tabela 2 - Variâncias da RMSPD padronizada

Métodos	Porcentagem de retirada		
	10%	20%	30%
ALS(1)	0,3798	0,3780	0,3442
ALS(0)	0,1566	0,0604	0,0288
IMLD	0,2382	0,2153	0,2020
r-AMMI1	0,0316	0,0205	0,0102
r-AMMI2	0,1308	0,0484	0,0265

Para confirmar os resultados, foi feito o teste de Levene para homogeneidade de variâncias dos métodos. Para 10%, 20% e 30% de retirada dos dados, os valores da estatística de Levene foram: 67,51 (valor-p < 0,0001), 175,05 (valor-p < 0,0001) e 349,11 (valor-p < 0,0001) respectivamente.

Para concluir a comparação segundo o critério da RMSPD padronizada apresentam-se na Tabela 3 as principais estatísticas obtidas no estudo de simulação, sem considerar a porcentagem de retirada. Na tabela 3 Q3-Q1 representa a distância interquartil e conclui-se que o método que minimiza a distância interquartil é o método r-AMMI1 com um valor de 0,18. A mediana e a média da RMSPD apresentam os menores valores, -0,69 e -0,66 respectivamente, quando foram feitas as imputações com ALS(0). Entretanto, a menor variância, 0,02, foi alcançada imputando com as estimativas baseadas em um modelo sub-robusto, isto é o método r-AMMI1. De acordo com os objetivos deste trabalho, procurou-se avaliar a nova metodologia chamada IMLD e encontrou-se que segundo a RMSPD padronizada, a predição de interações através da média de imputações múltiplas não oferece um melhor desempenho do que aqueles métodos baseados em modelos de efeitos aditivos com interação multiplicativa. O método IMLD foi apenas superior ao método ALS(1) quando foram comparadas medidas de dispersão. Com os resultados obtidos, os melhores métodos segundo a RMSPD padronizada foram ALS(0) e r-AMMI1.

Tabela 3 - Estatísticas gerais da RMSPD padronizada

Estatísticas	Métodos de imputação				
	ALS(1)	ALS(0)	IMLD	r-AMMI2	r-AMMI1
Q3-Q1	0,9096	0,2963	0,4017	0,2557	0,1803
Mediana	0,1431	-0,6960	1,6478	-0,4636	-0,6288
Média	0,2857	-0,6647	1,4458	-0,4421	-0,6247
Variância	0,4020	0,0860	0,2206	0,0695	0,0229

Entretanto, na tabela 4 é feita a comparação dos métodos de imputação usando a estatística Procrustes. Nessa tabela aparece uma contagem do número de vezes que o valor da estatística Procrustes para o método IMLD ($IMLD_{Procrustes}$) foi maior ou menor do que as estatísticas $ALS(1)_{Procrustes}$, $ALS(0)_{Procrustes}$, $r-AMMI2_{Procrustes}$ e

$r - AMMI1_{Procrustes}$ em cada porcentagem de retirada considerada. Segundo a tabela 4, nos 1000 conjuntos de dados com 10% de observações ausentes simuladas o desempenho do IMLD não foi melhor do que os outros métodos, por exemplo, em 982 ocasiões a estatística $ALS(0)_{Procrustes}$ teve valores menores do que os valores da $IMLD_{Procrustes}$, o que significa que as matrizes de dados com as imputações por ALS(0) apresentam uma maior similaridade em relação à matriz original, do que podem apresentar as matrizes com imputações por IMLD. Mas, não somente com o ALS(0) acontece isso, com os métodos r-AMMI2 e r-AMMI1 mostrou-se um comportamento parecido, pois em 975 conjuntos de dados a estatística $r - AMMI2_{Procrustes}$ foi menor do que a estatística $IMLD_{Procrustes}$. Além disso, aumentando a porcentagem de retirada aleatória se obtém as mesmas conclusões. Com 30% de retirada aleatória, dos mil conjuntos de dados simulados com observações perdidas, em 999 deles a estatística $r - AMMI1_{Procrustes}$ foi menor do que a $IMLD_{Procrustes}$. Usando a rotação de Procrustes pode concluir-se que as matrizes de dados completadas com imputações feitas por mínimos quadrados alternados ou baseadas em sub-modelos robustos apresentam uma maior similaridade com a matriz de dados original considerada para o estudo de simulação.

Tabela 4 - Número modal em que a estatística Procrustes para IMLD foi maior ou menor do que as estatísticas Procrustes correspondentes aos outros métodos de imputação; calculado sobre 1000 conjuntos de dados para cada porcentagem de retirada

Comparação Procrustes	Porcentagem de retirada		
	10%	20%	30%
$IMLD_{Procrustes} < ALS(1)_{Procrustes}$	107	162	214
$IMLD_{Procrustes} > ALS(1)_{Procrustes}$	893	838	786
$IMLD_{Procrustes} < ALS(0)_{Procrustes}$	18	7	1
$IMLD_{Procrustes} > ALS(0)_{Procrustes}$	982	993	999
$IMLD_{Procrustes} < r - AMMI2_{Procrustes}$	25	6	3
$IMLD_{Procrustes} > r - AMMI2_{Procrustes}$	975	994	997
$IMLD_{Procrustes} < r - AMMI2_{Procrustes}$	14	5	1
$IMLD_{Procrustes} > r - AMMI2_{Procrustes}$	986	995	999

Em cada um dos conjuntos de informação com dados faltantes simulados, foram aplicados os métodos para imputar as observações ausentes e as imputações obtidas foram comparadas com as observações reais do experimento através do cálculo do coeficiente de correlação de Spearman. Os resultados foram muito parecidos nas diferentes porcentagens de retirada, mas por questão de espaço neste artigo só apresentam-se as estatísticas obtidas para a correlação de Spearman quando foram consideradas retiradas aleatórias de 30% e acompanhadas de um gráfico de caixas.

Observando a Tabela 5, em geral os cinco métodos estudados oferecem imputações altamente correlacionadas com os dados originais do experimento, pois a média e a mediana dos coeficientes de Spearman são superiores a 0,90. Entretanto, a menor correlação foi obtida através do ALS(1) com um coeficiente de 0,86 e a maior correlação resultou com a imputação por ALS(0) e r-AMMI1 com os valores de 0,9750 e 0,9747 respectivamente. O desvio padrão dos coeficientes é muito pequeno, bem como a distância interquartil. As diferenças entre as estatísticas são apenas por casas decimais, o que não permite escolher um único método como o melhor.

Tabela 5 - Estatísticas da correlação de Spearman imputando 30% de dados

Estatísticas	Métodos de Imputação				
	IMLD	ALS(1)	ALS(0)	r-AMMI2	r-AMMI1
Média	0,9387	0,9448	0,9568	0,9546	0,9563
Desvio padrão	0,0111	0,0134	0,0074	0,0079	0,0075
Mínimo	0,8861	0,8688	0,9296	0,9224	0,9288
Quartil 1 (Q1)	0,9323	0,9395	0,9520	0,9498	0,9516
Mediana	0,9400	0,9469	0,9570	0,9550	0,9564
Quartil 3 (Q3)	0,9465	0,9534	0,9620	0,9601	0,9616
Máximo	0,9648	0,9713	0,9750	0,9738	0,9747
Q3-Q1	0,0142	0,0139	0,0100	0,0103	0,0099

Na figura 2, o gráfico de caixas mostra o desempenho parecido dos algoritmos de imputação comparando-os através do coeficiente de Spearman. A caixa correspondente ao método IMLD se encontra um pouco abaixo das caixas dos outros métodos, mas isso não é suficiente para concluir que esse método é inferior aos demais levando em conta apenas esse critério.

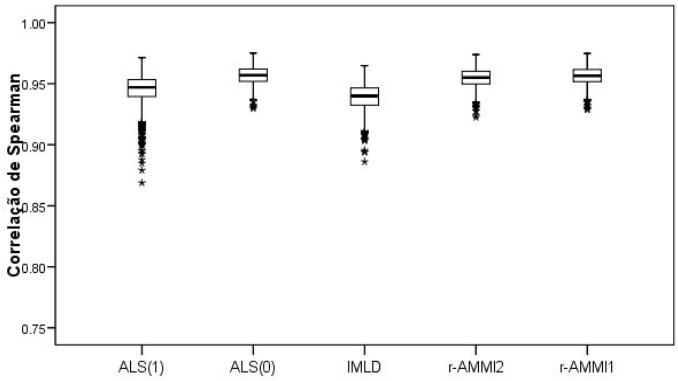

Figura 2 - Gráfico de caixas da distribuição do coeficiente de correlação de Spearman imputando 30% de dados.

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 os cultivares são escolhidos durante cada ano, baseados somente nos dados que foram observados sem considerar aqueles dados não observados, o mecanismo de ausência é claramente aleatório (Piepho e Möhring, 2005). 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 vários algoritmos de imputação (incluindo os mínimos quadrados alternados e estimativas robustas) considerando valores faltantes MNAR, pelo qual nesse estudo foi avaliada uma alternativa diferente, quer dizer o mecanismo MCAR com a inclusão de um novo método de imputação múltipla, mas os resultados obtidos sobre conjuntos de dados reais foram praticamente os mesmos, recomendando a imputação de combinações ausentes genótipo por ambientes através dos mínimos quadrados ordinários baseados em um modelo aditivo sem interação.

Os cinco métodos de imputação considerados nessa comparação não precisam estritamente da pressuposição de alguma distribuição sobre a variável resposta, por exemplo, a distribuição normal. O método de imputação múltipla foi proposto para evitar esse problema, enquanto a imputação com mínimos quadrados alternados e sub-modelos robustos provavelmente terá estimativas menos eficientes com distribuições muito viesadas ou com caldas longas (comunicação pessoal do Hans-Peter Piepho, 2008).

Finalmente, levando em conta a discussão final nesta seção e para dar continuidade desse estudo poderia ser interessante considerar o mecanismo de ausência MAR em experimentos genótipo por ambiente, estudar o efeito de diferentes distribuições de probabilidade sobre a precisão do método de imputação múltipla não paramétrica e compará-lo com a imputação que incorpore covariáveis genotípicas ou ambientais (temperatura e índices de chuva ou umidade).

Conclusões

A comparação dos algoritmos de imputação foi feita sobre um único conjunto de dados, pelo qual os resultados somente são validos para o exemplo atualmente investigado. Assim, segundo a RMSPD padronizada usada no estudo de simulação, os melhores métodos para imputar foram ALS(0) e r-AMMI1. Esses métodos apresentam os melhores resultados minimizando a média e a dispersão da distribuição. Baseado nesses critérios encontrou-se que a predição de interações em experimentos G×E com r-AMMI1 ou ALS(0) é mais recomendável do que a imputação com ALS(1), IMLD e r-AMMI2. Todos os métodos estudados neste trabalho apresentaram uma alta correlação entre o dado

imputado e o correspondente dado real no experimento. Segundo a estatística de Procrustes, a melhor similaridade entre as matrizes completadas por imputação e a matriz de dados original do experimento de algodão foi obtida com os métodos baseado nos modelos AMMI. Utilizando a RMSPD padronizada e a estatística de Procrustes, a metodologia proposta recentemente de predizer interações com a média dos valores obtidos por imputação múltipla livre de distribuição não mostrou melhores resultados do que as predições considerando apenas un modelo aditivo, ou seja, ALS(0).

Agradecimentos

Este trabalho é parte da dissertação do Mestrado em Estatística e Experimentação Agronômica do primeiro autor no Departamento de Ciências Exatas da ESALQ/USP, Piracicaba e foi realizado com o apoio do Conselho Nacional Cientifico e Tecnológico (CNPq), através do Programa PEC-PG.

ARCINIEGAS-ALARCÓN, S.; DIAS, C.T.S. Data imputation in trials with genotype by environment interaction: An application on cotton data. *Rev. Bras. Biom.*, São Paulo, v.27, n.1, p.125-138, 2009.

- *ABSTRACT: A common problem in multienvironment trials are the missing genotype-environmental combinations. Recently, Bergamo proposed a distribution-free multiple imputation method in the interaction matrix. The purpose of this paper is to evaluate the new development and compare it with methodologies that have success in the genotype-environmental trials with missing data, like the alternating least squares (ALS) and the robust estimates, using the Additive Main effects and Multiplicative Interaction Models (AMMI). Was made an simulation study based in real data, doing missed random considering different percentages, imputing the observations and comparing the methodologies through three criteria: the square root of the mean predictive difference, the Procrustes statistic and the Spearman's rank correlation coeficient. Was concluded that the multiple imputation is not better than the imputation based in a additive model without interaction, and the best results for the variance are obtained with robust sub-models. All the considerated methods in this study have a high correlation between the true and the imputed missing values.*
- *KEY WORDS: Missing data; data imputation; AMMI models; genotype-environment interaction.*

Referências

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

BERGAMO, G. C.; DIAS, C. T. S.; KRZANOWSKI, W. J. Distribution-free multiple imputation in an interaction matrix through singular value decomposition. *Sci. Agric*., Piracicaba, v.65, n.4, p.422-427, 2008.

BIRKES, D.; DODGE, Y.; SEELY, J. Spanning tests for estimable contrasts in classification models. *Ann. Stat.*, Corvallis, v.4, n.1, p.86-107, 1976.

CALINSKI, T. et al. EM and ALS algorithms applied to estimation of missing data in series of variety trials. *Biul. Oceny Odmian*, Poznan, v.24-25, p.7-31, 1992.

DENIS, J. B.; BARIL C. P. Sophisticated models with numerous missing values: the multiplicative interaction model as an example. *Biul. Oceny Odmian*, Poznan, v.24-25, p.33-45, 1992.

DIAS, C. T. S.; KRZANOWSKI, W. J. Model selection and cross validation in additive main effect and multiplicative interaction models. *Crop Sci*., Madison, v.43, p.865-873, 2003.

DIAS, C. T. S.; KRZANOWSKI, W. J. Choosing components in the additive main effect and multiplicative interaction (AMMI) models. *Sci. Agricola*, Piracicaba, v.63, n.2, p.169-175, 2006.

DODGE Y. *Analysis of experiments with missing data*. New York: John Wiley. 1985. 499p.

DODGE, Y.; ZOPPE, A. Adjusting the EM algorithm for design of experiments with missing data. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY INTERFACES, 26., 2004, Cavtat, Croatia. *Proceedings*…., p.9-12.

DUARTE, J. B.; VENCOVSKY, R. *Interação genótipo×ambiente*: uma introdução à análise "AMMI". Riberão Preto: Sociedade Brasileira de Genética, 1999. 60p. (Série Monografías).

FARIAS, F. J. C. *Índice de seleção em cultivares de algodoeiro herbáceo*. 2005. 121f. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2005.

GAUCH, H. G.; ZOBEL, R. W. Imputing missing yield trial data. *Theor. Appl. Genet.*, New York, v.79, p.753-761, 1990.

GODFREY, A. J. R. et al. Two-stage clustering in genotype-by-environment analyses with missing data. *J. Agric. Sci.*, Cambridge, v.139, p.67-77, 2002.

KRZANOWSKI, W. J. *Principles of multivariate analysis*: a user´s perspective. Oxford: University Press; 2000. 586p.

LITTLE, R. J.; RUBIN D. B. *Statistical analysis with missing data*. 2. ed. New York: John Wiley, 2002. 381p.

PIEPHO, H. P. Methods for estimating missing genotype-location combinations in multilocation trials - an empirical comparison. *Inform. Biom. Epidemiol. Med. Biol*. Stuttgart, vol.26, n.4, p.335-349, 1995.

PIEPHO, H. P; MÖHRING, J. Selection in cultivar trials-Is it ignorable?. *Crop Sci.* Madison, v.46, p.192-201, 2006.

RUBIN, D. B. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.*, Alexandria, v.91, n.434, p.473-489, 1996.

SAS INSTITUTE. *SAS/IML 9.1 user.s guide*. Carey, 2004. 1040p.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: our view of the state of the art. *Psychol. Methods*, Washington, v.7, n.2, p. 147-177, 2002.

SPRENT, P; SMEETON, N. C. *Applied nonparametric statistical methods*. 3th ed. Boca Raton: Chapman and Hall, 2001, 463p.

VAN EEUWIJK, F. A.; KROONENBERG, P. M. Multiplicative models for interaction in three-way ANOVA, with applications to plant breeding. *Biometrics*, Oxford, v.54, n.4, p.1315-1333, 1998.

Recebido em 11.02.2009.

Aprovado após revisão 14.05.2009.