Análise AMMI com dados imputados em experimentos de interação genótipo x ambiente de algodão

Sergio Arciniegas-Alarcón[1] e Carlos Tadeu dos Santos Dias[1]

[1]Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Departamento de Ciências Exatas, Caixa Postal 9, CEP 13418-900 Piracicaba, SP. E-mail: sergio.arciniegas@gmail.com, ctsdias@esalq.usp.br

Resumo – O objetivo deste trabalho foi avaliar a conveniência de definir o número de componentes multiplicativos dos modelos de efeitos principais aditivos com interação multiplicativa (AMMI) em experimentos de interações genótipo x ambiente de algodão com dados imputados ou desbalanceados. Um estudo de simulação foi realizado com base em uma matriz de dados reais de produtividade de algodão em caroço, obtidos em ensaios de interação genótipo x ambiente, conduzidos com 15 cultivares em 27 locais no Brasil. A simulação foi feita com retiradas aleatórias de 10, 20 e 30% dos dados. O número ótimo de componentes multiplicativos para o modelo AMMI foi determinado usando o teste de Cornelius e o teste de razão de verossimilhança sobre as matrizes completadas por imputação. Para testar as hipóteses, quando a análise é feita a partir de médias e não são disponibilizadas as repetições, foi proposta uma correção com base nas observações ausentes no teste de Cornelius. Para a imputação de dados, foram considerados métodos usando submodelos robustos, mínimos quadrados alternados e imputação múltipla. Na análise de experimentos desbalanceados, é recomendável escolher o número de componentes multiplicativos do modelo AMMI somente a partir da informação observada e fazer a estimação clássica dos parâmetros com base nas matrizes completadas por imputação.

Termos para indexação: *Gossypium hirsutum*, desbalanceamento, imputação de dados, modelos AMMI.

AMMI analysis with imputed data in genotype x environment interaction experiments in cotton

Abstract – The objective of this work was to evaluate the convenience of defining the number of multiplicative components of additive main effect and multiplicative interaction models (AMMI) in genotype x enviroment interaction experiments in cotton with imputed or unbalanced data. A simulation study was carried out based on a matrix of real seed-cotton productivity data obtained in trials with genotype x environment interaction carried out with 15 genotypes at 27 locations in Brazil. The simulation was made with random withdrawals of 10, 20 and 30% of the data. The optimal number of multiplicative components for the AMMI model was determined using the Cornelius test and the likelihood ratio test onto the matrix completed by imputation. A correction based on the data missing in the Cornelius procedure was proposed for testing the hypothesis when the analysis is made from averages and the repetitions are not available. For data imputation, the methods considered used robust submodels, alternating least squares and multiple imputation. For analysis of unbalanced experiments, it is advisable to choose the number of multiplicative components of the AMMI model only from the observed information and to make the classical estimation of parameters based on the matrices completed by imputation.

Index terms: *Gossypium hirsutum*, unbalanced data, data imputation, AMMI models.

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 na produtividade em relação a outras cultivares. Essas mudanças são influenciadas por diferentes condições ambientais e são referidas como interação genótipo x ambiente (G x E).

Muitas vezes os experimentos de interação genótipo x ambiente são desbalanceados e vários genótipos não são testados em alguns ambientes. Uma maneira muito comum de analisar esse tipo de estudos é imputar as observações ausentes e, posteriormente, na matriz de dados completada (observados + imputados), encontrar o modelo de efeitos aditivos com interação multiplicativa (AMMI) (Gauch Junior, 2006; Gauch Junior et al., 2008) para explicar a interação. Além do modelo AMMI para a análise, após a imputação, outras opções, como a regressão

fatorial, poderiam ser consideradas (Van Eeuwijk et al., 2005, 2007; Romagosa et al., 2008).

Nos experimentos de interação genótipo x ambiente balanceados, a escolha do melhor modelo AMMI usando testes estatísticos foi estudada por Milliken & Johnson (1989), Cornelius et al. (1996) e Dias & Krzanowski (2003, 2006). Eles detectaram que o teste de Cornelius denotado por F_R, proposto para a seleção do número ótimo de componentes multiplicativos, é bem recomendável quando se dispõe de uma estimativa independente da variância. Quando essa estimativa não é confiável ou não existe, pode ser usado o teste de razão de verossimilhança, baseado nos autovalores da matriz de interação.

Nos experimentos de interação genótipo x ambiente com observações ausentes, existem várias alternativas de análise. Calinski et al. (1992) detectaram que o uso de estimativas AMMI baseadas em mínimos quadrados alternados pode ser uma boa solução na cultura do trigo em matrizes de dimensão 10x28 e 15x12 (cultivares x locais). Denis & Baril (1992) também sugerem usar as estimativas AMMI, baseadas na aplicação de submodelos robustos na cultura do trigo em uma matriz de dimensão 7x82 (cultivares x locais). Mais recentemente, Bergamo (2007) e Bergamo et al. (2008) propuseram um método de imputação múltipla livre de distribuição (IMLD) aplicado especificamente à matriz de interação G x E com genótipos de *Eucalyptus grandis* em uma matriz de dimensão 20x7 (cultivares x locais) com base na decomposição por valores singulares, sem usar o modelo AMMI e sem pressuposições estruturais ou distribucionais sobre os conjuntos de dados.

O objetivo deste trabalho foi avaliar a conveniência de definir o número de componentes multiplicativos do modelo AMMI em experimentos de interação genótipo x ambiente de algodão com matrizes de dados que contêm imputações ou apenas a informação observada (matriz incompleta).

Material e Métodos

Os dados utilizados foram obtidos do Ensaio Estadual de Algodoeiro Herbáceo referente ao ano agrícola de 2000/2001, 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 de blocos completos ao acaso, com 15 cultivares e quatro repetições (Farias, 2005). A variável estudada foi produtividade de algodão em caroço (kg ha^{-1}) e, para este trabalho, foram disponibilizadas as médias das repetições de produtividade de cada genótipo em cada um dos locais, isto é, foi utilizada uma matriz de dados de dimensão 15x27.

Para este estudo, foram considerados dois testes de significância para determinar o número ótimo de termos multiplicativos nos modelos AMMI: o teste de Cornelius denotado por F_R e o teste de razão de verossimilhança. Considere-se o seguinte modelo AMMI para avaliar genótipos e ambientes:

$y_{ij} = \mu + g_i + e_j + \lambda_1\alpha_{1i}\gamma_{1j} + \lambda_2\alpha_{2i}\gamma_{2j} + ... + \lambda_k\alpha_{ki}\gamma_{kj} + \varepsilon_{ij}$ (1),

em que y_{ij} é a produtividade média de algodão em caroço, g_i (i = 1, ..., 15) representa os genótipos, e_j (j = 1, ..., 27) representa os ambientes e os componentes multiplicativos representam a interação genótipo x ambiente. A estatística do teste de Cornelius reescrita por Dias & Krzanowski (2006) para o modelo AMMI com k termos multiplicativos é dada por:

$$F_{R,k} = \frac{\left(SQ(G \times E) - \sum_{r=1}^{k} l_r\right)}{f_2 * QM\ (Erro\ médio)}$$; em que f_2 = (15-1-k) (27-1-k), SQ(G×E) representa a soma de quadrados da interação G×E, l_r é o r-ésimo maior autovalor da matriz Z^TZ, e a matriz Z está definida pelos elementos $z_{ij} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$. O erro médio, segundo Duarte & Vencovsky (1999), é originário das análises de variância individuais dos 27 ambientes. Este é o teste F_R de Cornelius et al. (1996) com a hipótese nula de que somente k termos determinam a interação, de tal forma que a estatística $F_{R,k}$ tem uma distribuição F aproximada com graus de liberdade (f_2,$GL_{Erro\ médio}$), em que f_2 já foi definida anteriormente e $GL_{Erro\ médio}$ representa os graus de liberdade do quadrado médio do resíduo. Um resultado significativo para o teste sugere que no mínimo um ou mais termos multiplicativos devem ser adicionados aos k já incluídos.

Entretanto, o teste de razão de verossimilhança para o k-ésimo componente multiplicativo do modelo AMMI está baseado na estatística: $U_k = l_k/(l_k + l_{k+1} + ... + l_p)$, em que p = min (15-1, 27-1) e l_k é o k-ésimo maior autovalor da matriz Z^TZ que já foi definida anteriormente para o teste F_R. A distribuição exata da estatística U_k é apresentada por Milliken & Johnson (1989) para um número pequeno de genótipos e ambientes, mas, para aqueles casos em que se tem um número grande tanto de

genótipos quanto de ambientes, Cornelius et al. (1996) afirmam que pode ser utilizada uma transformação da estatística para obter um teste F aproximado (F_{teste}). Assim, para testar o k-ésimo termo, a estatística F_{teste} pode ser encontrada como segue, em que G e E representam o número de genótipos e de ambientes, respectivamente: $p = \min(G-1, E-1)$; $n = \max(G-1, E-1)$; $Q_k = [(p - k + 1)U_k - 1]/(p - k)$;

$$c_1^* = \frac{u_{1k} - (n-k+1)}{(n-k+1)(p-k)};$$

$$c_2^* = \frac{(p-k+1)(n-k+1)u_{2k}^2 - 2u_{1k}^2}{(n-k+1)^2[(p-k+1)(n-k+1)+2](p-k)^2};$$

$d^* = c_1^*(1 - c_1^*) - c_2^*$; $a^* = dc_1^*/c_2^*$; $b^* = d^*(1 - c_1^*)/c_2^*$; $F_{teste} = b^*U_k/a^*(1 - U_k)$.

A estatística F_{teste} tem uma distribuição F aproximada com graus de liberdade ($2a^*$, $2b^*$) e os valores u_{1k} e u_{2k} correspondem à esperança e ao desvio-padrão de (l_1/σ^2). Liu & Cornelius (2001) encontraram as funções polinomiais para esses valores por meio de simulação.

Para imputar observações, foram considerados os algoritmos ALS(0), ALS(1), IMLD, r-AMMI2 e r-AMMI1 estudados por Arciniegas-Alarcón & Dias (2009). Os métodos ALS(0) e ALS(1) foram propostos por Calinski et al. (1992), modificados por Piepho (1995) e correspondem às estimativas AMMI baseadas em mínimos quadrados alternados (ALS) usando como modelo de imputação um modelo aditivo sem interação (AMMI0) e um modelo com um componente multiplicativo (AMMI1), respectivamente. A ideia principal dos ALS consiste na estimação iterativa dos parâmetros do modelo (1) considerando alguns deles como conhecidos e na estimação dos parâmetros restantes por mínimos quadrados ordinários. A regressão deve ser feita somente sobre os valores observados.

O método ALS(0) consiste em resolver o seguinte sistema de equações por mínimos quadrados alternados (ALS): passo A – $y_{ij} - \hat{\mu} - \hat{g}_i = e_j$ (resolver para e_j); passo B – $y_{ij} - \hat{\mu} - \hat{e}_j = g_i$ (resolver para g_i). Uma vez resolvido o sistema, normaliza-se calculando-se uma tabela de dupla entrada preenchida a partir dos parâmetros estimados atuais e fazendo-se a análise AMMI0. O processo de imputação com ALS(0) volta para o passo A até alcançar convergência.

Para o método ALS(1), devem-se seguir os passos. No primeiro, devem-se preencher os dados ausentes com as imputações obtidas por ALS(0) e, sobre a tabela de dados preenchida, estimar os parâmetros do modelo AMMI1. As estimativas desses parâmetros serão os valores iniciais para começar o algoritmo. No segundo, deve-se resolver o seguinte sistema de equações por mínimos quadrados alternados (ALS): $y_{ij} - \hat{\mu} - \hat{g}_i - \hat{e}_j = \hat{\lambda}_1\hat{\alpha}_{1i}\gamma_{1j}$ (resolver para γ_{1j}); $y_{ij} - \hat{\mu} - \hat{g}_i - \hat{e}_j = \lambda_1\alpha_{1i}\hat{\gamma}_{1j}$ (resolver para α_{1i}). Depois de resolver as equações, é feita a normalização, pelo cálculo de uma tabela completa de dupla entrada a partir dos parâmetros estimados atuais, e a análise AMMI1. No terceiro passo, deve-se resolver o seguinte sistema de equações por mínimos quadrados alternados (ALS): $y_{ij} - \hat{\mu} - \hat{g}_i = e_j + \hat{\lambda}_1\hat{\alpha}_{1i}\gamma_{ij}$, (resolver para e_j e γ_{1j}); $y_{ij} - \hat{\mu} - \hat{e}_j = g_i + \hat{\lambda}_1\alpha_{1i}\hat{\gamma}_{1j}$, (resolver para g_i e α_{1i}).Uma vez resolvido o sistema, são feitas a normalização, calculando-se uma tabela de dupla entrada completa a partir dos parâmetros estimados atuais, e a análise AMMI1. O processo de imputação com ALS(1) volta para o passo 2 até alcançar convergência.

O método r-AMMI proposto por Denis & Baril (1992) sugere que, no caso de observações ausentes em experimentos de interação genótipo x ambiente, devem-se fazer análises sobre tabelas completas nas quais 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 (Denis & Baril, 1992; Piepho, 1995). Com uma ponderação igual, as análises são equivalentes à análise AMMI clássica para dados completos. Para a análise AMMI1, Denis & 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. Neste trabalho, foram considerados os métodos r-AMMI1 e r-AMMI2.

A imputação múltipla livre de distribuição (IMLD) proposta por 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 Y de dimensão (n x p), com n>p, ou seja, o número de linhas é maior do que o número de colunas. Se n<p, a matriz deve ser transposta. Para melhor entendimento, considere-se somente um valor perdido Y_{ij} em Y. Deve-se fazer a restrição: omissão da i-ésima linha de Y e calcular a decomposição por valor singular da

matriz resultante de dimensão [(n - 1) x p] denotada por $Y^{(-i)}$, em que $Y^{(-i)} = \overline{U}\overline{D}\overline{V}^{T}$, $\overline{U} = (\overline{u}_{sh})$ $\overline{V} = (\overline{v}_{sh})$, $\overline{D} = (\overline{d}_1, ..., \overline{d}_p)$ O passo seguinte consiste na omissão da j-ésima coluna de Y e na obtenção da decomposição por valores singulares (DVS) da matriz resultante de dimensão (n x (p-1)) denotada por $Y_{(-j)}$, em que $Y_{(-j)} = \widetilde{U}\widetilde{D}\widetilde{V}^{T}$, $\widetilde{U} = (\widetilde{u}_{sh})$, $\widetilde{V} = (\widetilde{v}_{sh})$, $\widetilde{D} = (\widetilde{d}_1, ..., \widetilde{d}_{p-1})$. As matrizes $\overline{U}$, $\overline{V}$, $\widetilde{U}$ e $\widetilde{V}$ são ortonormais, e $\widetilde{D}$ e $\overline{D}$ são diagonais. Agora, combinando as duas DVS, $Y^{(-i)}$ e $Y_{(-j)}$, obtém-se o valor imputado por meio de:

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

Bergamo et al. (2008) afirmam que cinco imputações para cada observação ausente são suficientes 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 esses valores produz uma imputação diferente. Para mais de um valor perdido, um esquema iterativo deve ser envolvido.

Para avaliar a conveniência de escolher um modelo AMMI sobre experimentos com dados imputados, foi desenvolvido um estudo de simulação baseado no conjunto de dados reais de ensaios da cultura do algodão por meio do programa computacional em SAS/IML (SAS Institute, 2004). O conjunto de dados original contém 405 observações. Esse conjunto foi submetido a retiradas aleatórias de diferentes percentagens de dados. Foram consideradas as percentagens de 10, 20 e 30%, o processo foi repetido mil vezes para cada percentagem, para um total de três mil perdas aleatórias, ou seja, três mil conjuntos de dados diferentes foram gerados. No primeiro caso (10%), foram retirados 41 elementos da matriz de interações; no segundo caso (20%), foram retirados 81 elementos e, finalmente, no terceiro caso (30%), foram retirados 122 elementos. Em cada um dos 3 mil conjuntos de dados gerados foram aplicados os métodos de imputação e, posteriormente, sobre a matriz de dados completada (observados + imputados), foram feitos os testes de significância (F_R e F_{teste}) para determinar o número de componentes multiplicativos. Para comparar os resultados, o F_{teste} foi aplicado diretamente sem fazer nenhuma correção pelos dados faltantes, mas, usando o teste F_R, foram corrigidos os graus de liberdade (GL) do erro médio para testar as hipóteses de significância. Dado que só se tinha a matriz de médias do experimento, os graus de liberdade foram calculados assim: $GL(\text{Erro médio}) = \sum_{j=1}^{b} GL_{Erroj}$ com $GL_{Erroj} = \left[t\text{-}1\text{-}\text{Faltantes}_j \right]$; em que j representa os ambientes; t = 15 representa os genótipos; Faltantes_j representa o número de médias de tratamentos ausentes no ambiente j e n = 4 é o número de repetições. Além disso, foi feita uma correção no valor do QM (Erro médio), encontrando em cada conjunto de dados completado (observados + imputados) uma nova média geral do experimento e multiplicando essa média pelo CV dos dados originais. Isso faz sentido porque se assume que os efeitos dos tratamentos e os efeitos do erro são independentes. O resultado dividido pelo número de repetições foi a nova estimativa do QM (Erro médio), que foi usada para testar as hipóteses. Para os dados originais (só médias), foi possível calcular uma estimativa da variância do experimento, pois em Farias (2005) são apresentados os quadrados médios das análises individuais.

Resultados e Discussão

Segundo o sistema de Cornelius (F_R), observou-se que, com 5% de probabilidade, o melhor modelo para explicar a interação no conjunto de dados original e balanceado seria o AMMI4, pois somente a partir do IPCA4 – eixo de interação da análise de componentes principais – o resíduo torna-se não significativo (p = 0,1664). Entretanto, usando o método F_{teste}, detectou-se que o último componente significativo é o IPCA2. Assim, um modelo adequado seria o AMMI2.

Nos 1.000 conjuntos de dados com 10% de imputação por meio de ALS(0), ALS(1), r-AMMI1 e r-AMMI2, observou-se que o modelo mais escolhido corresponde ao AMMI2, seguido do modelo AMMI1 pelo estudo de simulação usando o F_{teste} (Tabela 1). Quando foram imputados os dados ausentes com ALS(0), em 533 vezes o AMMI2 foi recomendado. Esse mesmo modelo foi recomendado em 529 ocasiões imputado com ALS(1), em 556 ocasiões imputado com r-AMMI1, e em 688 ocasiões imputado com r-AMMI2. Caso diferente aconteceu com o método IMLD, pois o modelo escolhido, o maior número de vezes, quando se imputou foi o modelo AMMI3 (208 vezes), e o segundo modelo mais escolhido em conjuntos de dados com predições por IMLD foi o AMMI2 (194 vezes). Note-se que o F_{teste} foi bastante liberal, pois, por

exemplo, o modelo AMMI13 foi selecionado em 102 conjuntos de dados com predições IMLD.

Na percentagem de retirada de 20%, o número de vezes em que é recomendado o modelo AMMI2 diminuiu em matrizes com elementos imputados pelos métodos ALS(0), ALS(1) e r-AMMI1 (Tabela 1). Por exemplo, com 10% de imputação por ALS(0), o modelo AMMI2 foi recomendado 533 vezes; com 20% de informação imputada pelo mesmo método, o AMMI2 foi recomendado em 364 conjuntos de dados. O contrário aconteceu com o método r-AMMI2, pois, com 10% de dados imputados, o modelo AMMI2 foi selecionado em 688 ocasiões e com 20% de imputação, aumentou para 696.

Tabela 1. Modelos AMMI escolhidos por meio do método F_{teste} nos 1.000 conjuntos de dados com imputação em diferentes percentagens.

Modelos AMMI	Métodos de imputação				
	ALS(0)	IMLD	ALS(1)	r-AMMI1	r-AMMI2
			Imputação de 10%		
AMMI0	17	0	0	1	0
AMMI1	189	25	244	213	68
AMMI2	533	194	529	556	688
AMMI3	132	208	107	104	118
AMMI4	6	114	9	7	12
AMMI5	7	107	8	7	11
AMMI6	3	43	2	2	3
AMMI7	0	14	0	0	2
AMMI8	0	19	0	0	0
AMMI9	3	17	7	5	5
AMMI10	10	40	10	11	9
AMMI11	12	51	12	12	13
AMMI12	38	66	33	36	34
AMMI13	50	102	39	46	37
			Imputação de 20%		
AMMI0	56	0	0	0	0
AMMI1	232	2	348	321	21
AMMI2	364	12	370	412	696
AMMI3	204	63	139	139	137
AMMI4	35	99	31	33	32
AMMI5	20	162	14	19	21
AMMI6	3	136	8	4	7
AMMI7	5	90	5	4	1
AMMI8	2	70	4	3	5
AMMI9	4	71	4	3	9
AMMI10	8	79	9	7	7
AMMI11	14	73	19	13	19
AMMI12	28	61	26	21	18
AMMI13	25	82	23	21	27
			Imputação de 30%		
AMMI0	51	0	0	0	0
AMMI1	261	0	393	363	2
AMMI2	304	4	308	326	682
AMMI3	203	6	141	144	154
AMMI4	67	33	43	49	35
AMMI5	22	76	23	29	29
AMMI6	7	131	10	10	12
AMMI7	8	155	9	6	8
AMMI8	8	149	10	9	6
AMMI9	4	116	7	3	7
AMMI10	14	90	6	11	8
AMMI11	13	89	13	5	7
AMMI12	18	79	16	16	21
AMMI13	20	72	21	29	29

Esses primeiros resultados sugerem que, para tomar alguma decisão sobre a estrutura da interação em experimentos desbalanceados, é preferível escolher o número de componentes multiplicativos sobre dados que não contenham nenhum tipo de imputação, pois o número de componentes multiplicativos está relacionado diretamente com o modelo de imputação, usando mínimos quadrados alternados ou submodelos robustos. Um resultado totalmente diferente foi encontrado quando se fizeram as imputações com IMLD, porque esse método não leva em conta modelos AMMI para a predição. Por isso, com 20% de retirada aleatória, os modelos mais escolhidos pelo IMLD foram AMMI5 e AMMI6 (Tabela 1).

Com 30% de imputação de dados com ALS(0), ALS(1) e r-AMMI1, os modelos escolhidos mais vezes por cada método foram AMMI1, AMMI2 e AMMI3 (Tabela 1). Nos conjuntos de dados completados por meio de imputações com r-AMMI2, o modelo mais escolhido foi o AMMI2 (682 vezes). Quando foram imputadas as observações com IMLD, os modelos mais selecionados foram AMMI6, AMMI7, AMMI8 e AMMI9. Os dados da Tabela 1 confirmam que não é recomendável tomar uma decisão sobre a escolha do número de componentes escolhidos usando o modelo baseado em dados imputados, porque essa decisão dependerá muito do método de imputação de observações ausentes. Com relação aos métodos de imputação, segundo Arciniegas-Alarcón (2008), à medida que aumenta o número de perdas, diminui a variância da estatística de qualidade do método, conhecida como raiz quadrada da diferença preditiva média (RMSPD), o que indica maior precisão dos algoritmos.

Entretanto, verificou-se que o sistema de Cornelius é bem menos liberal do que o método F_{teste}, pois o número máximo de componentes selecionados sobre os conjuntos de dados com 10% de informação imputada foi quatro (Tabela 2). Nos 1.000 conjuntos de dados considerados para essa percentagem de imputação, o modelo escolhido mais vezes foi o AMMI3. Quando se imputou por ALS(0), o modelo AMMI3 foi escolhido 879 vezes, seguido pelo método r-AMMI1, escolhido em 869 ocasiões.

Nos 1.000 conjuntos com imputação de 20%, o modelo mais escolhido foi o AMMI2 (Tabela 2). Com o método r-AMMI2, esse modelo foi escolhido 858 vezes e com o IMLD, 534 vezes. Os modelos AMMI0 e AMMI4 foram os menos selecionados para essa percentagem. A maior frequência foi apresentada para o modelo AMMI3 (464 vezes) nos conjuntos com dados imputados por IMLD, comparando-o aos outros métodos de predição de observações ausentes.

Finalmente, quando se imputou 30% dos dados, o modelo mais escolhido foi o AMMI1 nos conjuntos com dados imputados por ALS(0), ALS(1), r-AMMI1 e r-AMMI2 (Tabela 2). Com o IMLD, o resultado foi diferente, pois o modelo mais escolhido foi o AMMI2 (840 vezes). Nesta percentagem, esperava-se que, nos conjuntos com imputações pelo r-AMMI2, o modelo mais selecionado fosse o AMMI2, pois a base para imputação foi justamente esse modelo, mas em apenas 49 conjuntos isso aconteceu.

Os resultados obtidos foram diferentes nas diversas percentagens de imputação consideradas e dependeram muito do método de imputação, bem como do teste para escolher o melhor modelo AMMI. Por essa razão, em experimentos com desbalanceamento, a principal recomendação consiste em definir o número de componentes multiplicativos usando somente as observações presentes. Não existe um teste exato para essa situação, mas sim uma aproximação, que consiste em usar inicialmente as tabelas de análise de variância para dados desbalanceados, oferecidas pelos diferentes pacotes estatísticos que usam um modelo aditivo com interação, nas quais se têm as somas de quadrados e graus de liberdade para a interação total e para o efeito dos genótipos ajustado pelo efeito dos ambientes (ou o contrário). Nessas tabelas, podem ser inseridas as somas de quadrados para cada componente multiplicativo, baseadas na informação observada, calculando-as por mínimos quadrados alternados ou por meio de algoritmos EM e Newton-Raphson propostos por Cornelius & Seyedsadr (1997). Sobre essa tabela de análise de variância aumentada, podem ser calculadas as significâncias para os modelos AMMI1, AMMI2 e AMMI3, pois, segundo Ebdon & Gauch Junior (2002), o melhor modelo para explicar o padrão de resposta dos dados geralmente não tem mais do que três componentes multiplicativos.

Além do modelo escolhido, deve-se verificar que os valores preditos pelo modelo para todas as observações (ausentes e presentes) façam sentido, ou seja, estejam no intervalo determinado pelo valor mínimo e pelo valor máximo dos dados observados (Denis & Baril, 1992). Se isso não ocorre, a melhor opção é fazer a estimação de parâmetros sobre a matriz de dados completada (usando imputação) e, neste artigo, foram apresentadas cinco alternativas para fazê-lo.

Tabela 2. Número de modelos AMMI escolhidos por meio do método de Cornelius em 1.000 conjuntos de dados para cada percentagem de retirada aleatória.

Modelos AMMI	Métodos de imputação				
	ALS(0)	IMLD	ALS(1)	r-AMMI1	r-AMMI2
	Percentagem de retirada: 10%				
AMMI0	0	0	0	0	0
AMMI1	0	0	0	0	0
AMMI2	46	18	144	104	225
AMMI3	879	826	830	869	770
AMMI4	75	156	26	27	5
	Percentagem de retirada: 20%				
AMMI0	0	0	0	0	0
AMMI1	44	1	307	175	142
AMMI2	787	534	684	795	858
AMMI3	169	464	9	30	0
AMMI4	0	1	0	0	0
	Percentagem de retirada: 30%				
AMMI0	58	0	0	27	5
AMMI1	713	57	995	933	946
AMMI2	229	840	5	40	49
AMMI3	0	103	0	0	0
AMMI4	0	0	0	0	0

Conclusões

1. Em experimentos desbalanceados de interação genótipo x ambiente, o número de componentes multiplicativos de um modelo AMMI deve ser determinado em uma matriz de dados que não contenha nenhum tipo de imputação.

2. As matrizes experimentais completadas com dados imputados são úteis para estimar os parâmetros AMMI de uma maneira tradicional quando são analisados experimentos incompletos ou desbalanceados.

3. O método de imputação de observações ausentes deve levar em conta a possível existência de relação direta entre o número de componentes multiplicativos dos modelos AMMI e os algoritmos de imputação que utilizam esses modelos.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico, pelo apoio financeiro.

Referências

ARCINIEGAS-ALARCÓN, S. **Imputação de dados em experimentos com interação genótipo por ambiente**: uma aplicação a dados de algodão. 2008. 82p. Dissertação (Mestrado) - Universidade de São Paulo, Piracicaba.

ARCINIEGAS-ALARCÓN, S.; DIAS, C.T. dos S. Imputação de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão. **Revista Brasileira de Biometria**, v.27, p.125-138, 2009.

BERGAMO, G.C. **Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação**. 2007. 89p. Tese (Doutorado) - Universidade de São Paulo, Piracicaba.

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

CALINSKI, T.; CZAJKA, S.; DENIS, J.B.; KACZMAREK, Z. EM and ALS algorithms applied to estimation of missing data in series of variety trials. **Biuletyn Oceny Odmian**, v.24-25, p.7-31, 1992.

CORNELIUS, P.L.; CROSSA, J.; SEYEDSADR, M.S. Statistical tests and estimators of multiplicative models for genotype-by-environment interaction. In: KANG, M.S.; GAUCH, H.G. **Genotype-by-environment interaction**. Boca Raton: CRC Press, 1996. p.199-234.

CORNELIUS, P.L.; SEYEDSADR, M.S. Estimation of general linear-bilinear models for two-way tables. **Journal of Statistical Computation and Simulation**, v.58, p.287-322, 1997.

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

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

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

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

EBDON, J.S.; GAUCH JUNIOR, H.G. Additive main effect and multiplicative interaction analysis of national turfgrass perfomance trials: I. interpretation of genotype x environment interaction. **Crop Science**, v.42, p.489-496, 2002.

FARIAS, F.J.C. **Índice de seleção em cultivares de algodoeiro herbáceo**. 2005. 121p. Tese (Doutorado) - Universidade de São Paulo, Piracicaba.

GAUCH JUNIOR, H.G. Statistical analysis of yield trials by AMMI and GGE. **Crop Science**, v.46, p.1488-1500, 2006.

GAUCH JUNIOR, H.G.; PIEPHO, H.-P.; ANNICCHIARCO, P. Statistical analysis of yield trials by AMMI and GGE: further considerations. **Crop Science**, v.48, p.866-889, 2008.

LIU, G.Z.; CORNELIUS, P.L. Simulations and derived approximations for the means and standard deviations of the characteristic roots of a Wishart matrix. **Communications in Statistics - Simulation and Computation**, v.30, p.963-989, 2001.

MILLIKEN, G.A.; JOHNSON, D.E. **Analysis of messy data**. New York: Chapman & Hall, 1989. v.2, 199p.

PIEPHO, H.P. Methods for estimating missing genotype-location combinations in multilocation trials – an empirical comparison. **Informatik, Biometrie und Epidemiologie in Medizin und Biologie**, v.26, p.335-349, 1995.

ROMAGOSA, I.; VOLTAS, J.; MALOSETTI, M; VAN EEUWIJK, F.A. Interacción genótipo por ambiente. In: ÁVILA, C.M.; ATIENZA, S.G.; MORENO, M.T.; CUBERO, J.I. (Ed.). **La adaptación al ambiente y los estreses abióticos en la mejora vegetal**. Sevilla: Consejería de Agricultura y Pesca, 2008. p.107-136.

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

VAN EEUWIJK, F.A.; MALOSETTI, M.; BOER, M.P. Modelling the genetic basis of response curves underlying genotype x environment interaction. In: SPIERTZ, J.H.J.; STRUIK, P.C.; LAAR, H.H. van (Ed.). **Scale and complexity in plant systems research**: gene-plant-crop relations. New York: Springer, 2007. p.113-124. (Wageningen UR frontis series).

VAN EEUWIJK, F.A.; MALOSETTI, M.; YIN, X.; STRUIK, P.C.; STAM, P. Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. **Australian Journal of Agricultural Research**, v.56, p.883-894, 2005.

Recebido em 21 de fevereiro de 2009 e aprovado em 23 de outubro de 2009