Imputação múltipla livre de distribuição em tabelas incompletas de dupla entrada

Sergio Arciniegas-Alarcón[1], Carlos Tadeu dos Santos Dias[1] e Marisol García-Peña[1]

[1]Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Departamento de Ciências Exatas, Caixa Postal 9, CEP 1341-8900 Piracicaba, SP, Brasil. E-mail: sergio.arciniegas@gmail.com, ctsdias@usp.br, luzmara@gmail.com

Resumo – O objetivo deste trabalho foi propor um novo algoritmo de imputação múltipla livre de distribuição, por meio de modificações no método de imputação simples recentemente desenvolvido por Yan para contornar o problema de desbalanceamento de experimentos. O método utiliza a decomposição por valores singulares de uma matriz e foi testado por meio de simulações baseadas em duas matrizes de dados reais completos, provenientes de ensaios com eucalipto e cana-de-açúcar, com retiradas aleatórias de valores em diferentes percentagens. A qualidade das imputações foi avaliada por uma medida de acurácia geral que combina a variância entre imputações e o viés quadrático médio delas em relação aos valores retirados. A melhor alternativa para imputação múltipla é um modelo multiplicativo que inclui pesos próximos a 1 para os autovalores calculados com a decomposição. A metodologia proposta não depende de pressuposições distribucionais ou estruturais e não tem restrições quanto ao padrão ou ao mecanismo de ausência dos dados.

Termos para indexação: dados ausentes, decomposição por valores singulares, ensaios multiambiente, experimentos desbalanceados, interação genótipo x ambiente, melhoramento de plantas.

Distribution-free multiple imputation in incomplete two-way tables

Abstract – The objective of this work was to propose a new distribution-free multiple imputation algorithm, through modifications of the simple imputation method recently developed by Yan in order to circumvent the problem of unbalanced experiments. The method uses the singular value decomposition of a matrix and was tested using simulations based on two complete matrices of real data, obtained from eucalyptus and sugarcane trials, with values deleted randomly at different percentages. The quality of the imputations was evaluated by a measure of overall accuracy that combines the variance between imputations and their mean square deviations in relation to the deleted values. The best alternative for multiple imputation is a multiplicative model that includes weights near to 1 for the eigenvalues calculated with the decomposition. The proposed methodology does not depend on distributional or structural assumptions and does not have any restriction regarding the pattern or the mechanism of the missing data.

Index terms: missing data, singular value decomposition, multi-environment trials, unbalanced experiments, genotype x environment interaction, plant breeding.

Introdução

No melhoramento genético de plantas, ensaios multiambientais são importantes para testar a adaptação geral e específica das cultivares. O cultivo em diferentes ambientes geralmente mostra flutuação significativa no desempenho relativo das cultivares. Essa flutuação é influenciada por condições ambientais e é conhecida como interação genótipo por ambiente (GxE) (Dias & Krzanowski, 2003).

Embora os experimentos com interação GxE sejam planejados para serem balanceados, é comum a ocorrência de valores ausentes por diversos motivos, como a retirada de genótipos de baixo desempenho, a consideração de novos genótipos, erros humanos e causas naturais (Rodrigues et al., 2011). Assim, experimentos desbalanceados são usualmente obtidos e não podem ser analisados diretamente por metodologias clássicas eficientes, como a do modelo de efeitos principais aditivos e interação multiplicativa (AMMI) ou da análise biplot GGE (Yan et al., 2007; Yang et al., 2009; Gauch Junior, 2013). A principal dificuldade nesse sentido é que essas metodologias envolvem a decomposição por valores singulares (DVS) das matrizes, a qual não existe para matrizes com dados ausentes (Gabriel, 2002).

As seguintes alternativas possibilitariam a análise de experimentos incompletos sobre a interação GxE: extração de um subconjunto balanceado que elimine os genótipos ou os ambientes com dados faltantes (Ceccarelli et al., 2007; Yan et al., 2011); preenchimento das parcelas vazias com médias ambientais; e preenchimento dos dados faltantes com estimativas obtidas por métodos que envolvam, por exemplo, o uso de modelos multiplicativos ou de modelos lineares mistos (Arciniegas-Alarcón et al., 2011; Kumar et al., 2012). Essas estratégias podem resolver o problema de desbalanceamento, mas nenhuma delas é simples e efetiva (Yan, 2013). A primeira não utiliza toda a informação disponível; a segunda pode resultar em problemas no caso de grande quantidade de observações ausentes, além de superestimar ou subestimar o valor real; e a terceira demanda múltiplos passos e procedimentos complexos (Yan, 2013).

Recentemente, Yan (2013) propôs um procedimento iterativo, baseado na DVS, para imputar dados faltantes em uma tabela de dupla entrada. O procedimento fornece imputação simples, mas, conforme Josse & Husson (2012a) e Buuren (2012) advertem, não leva em conta a incerteza produzida pelas imputações. Desse modo, se os parâmetros forem estimados a partir dos dados imputados, os erros-padrão serão subestimados, ou seja, os intervalos de confiança e os testes perderão a validade, mesmo que o modelo de imputação esteja correto.

A imputação múltipla (IM) pode resolver esse tipo de problema (Rubin, 1978, 1987). Descrições mais recentes da técnica são encontradas em Zhang (2003), Harel & Zhou (2007), Allison (2012) e Rässler et al. (2013). Segundo Bergamo (2007), a IM envolve três passos distintos: imputação, em que os valores ausentes são estimados M vezes e geram M conjuntos de dados completados (observados+imputados); análise, em que os M conjuntos de dados completados são analisados com procedimentos estatísticos apropriados para o problema em estudo; e combinação, em que os M conjuntos separados de resultados são combinados em uma única inferência.

A etapa mais crítica é a imputação, e o modelo utilizado nesse passo não precisa ser o mesmo que o usado na etapa de análise, o que torna a IM mais atrativa, pois nem sempre o modelo mais adequado para imputar é o mais adequado para analisar. Ao combinar os resultados das M análises, a variância da estimativa combinada consiste na variância dentro das imputações e na variância entre imputações; portanto, as incertezas dos dados imputados são incorporadas à inferência final.

Na literatura sobre experimentos GxE incompletos, há vários sistemas de imputação (Arciniegas-Alarcón et al., 2013), mas a maioria deles não quantifica a incerteza sobre os valores reais a serem imputados. Nos casos em que é possível estimar essa incerteza, como com o uso da IM paramétrica, os sistemas dependem fortemente das distribuições de probabilidade e do mecanismo de ausência dos dados (Little & Rubin, 2002).

O objetivo deste trabalho foi propor um novo algoritmo de imputação múltipla livre de distribuição, por meio de modificações no método de imputação simples recentemente desenvolvido por Yan para contornar o problema de desbalanceamento de experimentos.

Material e Métodos

Yan (2013) descreveu um método de imputação que usa a DVS para realizar a análise biplot (Gabriel, 1971, 2002), a partir de dados incompletos. Por essa razão, García-Peña et al. (2014) chamaram o método de "imputação biplot", notação que também será utilizada no presente trabalho para designar o algoritmo, descrito a seguir.

Considere a matriz X, de dimensão ($n \times p$) com elementos x_{ij} ($i=1,...,n$; $j=1,...,p$), em que alguns desses elementos estão ausentes (x_{ij}^{aus}). Na imputação biplot, os dados faltantes são inicialmente imputados pela média dos valores observados em suas respectivas colunas, o que resulta numa matriz X completada. As colunas da matriz X completada são, então, padronizadas ao se subtrair m_j de cada elemento e dividir o resultado por s_j; em que m_j representa a média da j-ésima coluna e s_j, o desvio-padrão. Os elementos padronizados são notados por p_{ij} e modelados por meio de um biplot bidimensional (Yan & Holland, 2010):

$$p_{ij} = \frac{(x_{ij} - m_j)}{s_j} = \sum_{k=1}^{2} \lambda_k \alpha_{ik} \gamma_{jk} + \varepsilon_{ij}.$$

Os valores p_{ij} são decompostos em dois componentes principais (CP), com valores singulares λ_k, autovetores para as linhas α_{ik} e autovetores para as colunas γ_{jk}, para cada um dos k-ésimos CP; em que ε_{ij} é o resíduo

para a linha i na coluna j. A matriz com elementos padronizados p_{ij} é denotada por P. Em seguida, calcula-se a DVS da matriz P, e os valores p_{ij} são atualizados com uso de apenas dois CP da DVS, o que resulta numa nova matriz, denominada $P^{(2)}$, com elementos $p_{ij}^{(2)}$. Todos os elementos $p_{ij}^{(2)}$ são, então, retornados à sua escala original por meio da expressão $\hat{x}_{ij}^{(2)} = m_j + s_j p_{ij}^{(2)}$. Assim, obtém-se uma nova matrix $X^{(2)}$ de dimensão (n×p). Os elementos ausentes x_{ij}^{aus} na matriz X original são imputados pelo correspondente valor $\hat{x}_{ij}^{(2)}$ de $X^{(2)}$. Por último, o processo da imputação biplot passa por iteração até que se alcance estabilidade nas imputações. Por exemplo, as iterações são realizadas até que, $d / \bar{y} < 0{,}01$, que define,

$$d = \left[\left(\frac{1}{na}\right)\sum_{i=1}^{na}(x_i - x_i^A)^2\right]^{\frac{1}{2}} \text{ e } \bar{y} = \left[\left(\frac{1}{N}\right)\sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij}^2\right]^{\frac{1}{2}}$$

em que: d representa a diferença entre os valores preditos, para todos os valores ausentes na iteração atual (x_i) e na iteração anterior (x_i^A); na é o número total de valores ausentes na matriz X; e x_i e x_i^A são os valores preditos para o i-ésimo dado faltante na iteração atual e na iteração anterior, respectivamente. Além disso, uma grande média pode ser calculada como $\bar{y}$, em que: y_{ij} é o valor observado (não ausente) na i-ésima linha e na j-ésima coluna, e N é o número total de valores observados.

Para realizar a imputação múltipla a partir do algoritmo descrito, sugerem-se duas aproximações que estão de acordo com os trabalhos de Bergamo et al. (2008) e Srivastava & Dolatabadi (2009).

Srivastava & Dolatabadi (2009) propuseram IM com uso dos resíduos simples do modelo de regressão linear clássico $Y=Q\beta+E$, em que o vetor Y (n×1) representa a variável dependente; Q (n×p) é a matriz de delineamento que contém as variáveis independentes; β (p×1) é o vetor desconhecido de parâmetros de regressão; e E (n×1) é o vetor de erros aleatórios independentes e identicamente distribuídos. Assume-se que os dados ausentes somente podem ocorrer no vetor Y e que todas as observações das variáveis independentes são disponíveis e completas. Portanto, o modelo pode ser reescrito como $(Y_0/Y_A)=(Q_0/Q_A)\beta+E$, em que Y_0 (n_1×1) corresponde ao subvetor dos n_1 dados observados e Y_A (n_0×1) ao subvetor que contém n_0 valores ausentes, ao se levar em conta que $n_0+n_1=n$. A matriz Q é decomposta de forma semelhante. Assim, a imputação múltipla é obtida da seguinte maneira: $\hat{Y}_{At}=Q_A(Q^t{}_0Q_0)^{-1}Q^T{}_0Y_0+E_t$, em que t=1,...,M, sendo M o número de imputações para cada dado faltante; e E_t é a t-ésima amostra aleatória com reposição de tamanho n_0 obtida do vetor de resíduos. Esse vetor é calculado pela expressão $e = (n_1 / n_1 - p)^{0,5} (Y_0 - Q_0 b_1)$, em que b_1 é a estimativa de mínimos quadrados de β, baseada unicamente nos dados observados, ou seja, $b_1=(Q^t{}_0Q_0)^{-1}Q^T{}_0Y_0$.

A primeira modificação proposta no algoritmo de imputação biplot é a seguinte. O método fornece no final, depois de atingir convergência, uma matriz $X^{(2)}$, que contém tanto imputações para valores ausentes quanto estimativas dos valores observados. Por essa razão, como passo intermediário para produzir imputação múltipla, pode-se calcular a matriz de resíduos simples para os dados observados por meio da diferença entre a matriz original e a matriz que contém as imputações, isto é, $\hat{\varepsilon} = X - X^{(2)}$. Naturalmente, a matriz $\hat{\varepsilon}$ tem dimensão (n×p) e é incompleta, porque somente podem ser obtidos os resíduos para (np-na) dados. A partir dos resíduos que podem ser efetivamente calculados em $\hat{\varepsilon}$, são construídas t matrizes diferentes, denotadas por Ω_t (n×p), em que t =1, ..., M, e cada elemento de Ω_t é escolhido aleatoriamente com reposição dos elementos de $\hat{\varepsilon}$. Em seguida, a imputação múltipla é realizada ao se substituir os elementos ausentes x_{ij}^{aus} da matriz X original pelos valores correspondentes de cada uma das t matrizes definidas por $X^{(2)}+\Omega_t$.

No presente trabalho, adotou-se M=5, uma vez que esse número permite atingir alta eficiência estatística em muitas aplicações práticas (Buuren, 2012). Dessa forma, obteve-se a imputação múltipla com resíduos simples por meio de um modelo multiplicativo. Esse método foi denominado IMBiplotRes.

Bergamo et al. (2008), no entanto, propuseram fazer imputação múltipla por meio de um esquema que utiliza a combinação de duas DVS de uma matriz para imputar cada dado faltante, com mudanças nos expoentes das matrizes de autovalores obtidas a partir das decomposições. Os autores sugeriram que os expoentes fossem escolhidos do intervalo entre 0,4 e 0,6; em percentagem, a escolha seria 40, 45, 50, 55 e 60%, para M=5. Desse modo, a segunda proposta para produzir imputação múltipla, com uso da imputação biplot como base, consiste em substituir o modelo biplot bidimensional pelo modelo:

$$p_{ij} = \frac{(x_{ij} - m_j)}{s_j} = \sum_{k=1}^{2} \lambda_k^w \alpha_{ik} \gamma_{jk} + \varepsilon_{ij},$$

em que w pode ser considerado como o peso para o autovalor λ_k, e a inclusão de diferentes pesos produzirá diferentes imputações para cada valor ausente.

Diferentemente do estudo de Bergamo et al. (2008), essa segunda proposta realiza apenas uma DVS, na etapa de imputação, para todos os dados ausentes, e w pode assumir qualquer valor no intervalo entre 0 e 1. Por essa razão, foram considerados, no presente trabalho, cinco grupos de avaliação. Assim, o método foi denominado IMBiplotGh, em que h representa o grupo e h=1, ..., 5. Os algoritmos são, respectivamente: IMBiplotG1, com w=0, 0,05, 0,1, 0,15 e 0,2; IMBiplotG2, com w=0,25, 0,30, 0,35, 0,40 e 0,45; IMBiplotG3, com w=0,5, 0,55, 0,60, 0,65 e 0,7; IMBiplotG4, com w=0,75, 0,80, 0,85, 0,90 e 0,95; e IMBiplotG5, com w=0,96, 0,97, 0,98, 0,99 e 1.

Pesos menores do que 0 e maiores do que 1 foram avaliados previamente, mas, em ambos os casos, o algoritmo apresentou problemas de convergência. Além disso, os pesos também devem ser incluídos no algoritmo de imputação biplot depois de se calcular a DVS de P, ou seja, na atualização dos valores p_{ij}.

Para avaliar os métodos de imputação propostos, foram usados dois conjuntos de dados reais balanceados provenientes de experimentos GxE, publicados em Lavoranti (2003) e Santos (2008). Em cada caso, os dados foram obtidos a partir de delineamentos experimentais aleatorizados em blocos com repetições; porém, cada um desses trabalhos oferece uma excelente descrição do planejamento, se detalhes específicos fossem requeridos. O primeiro conjunto de dados (Lavoranti, 2003) é composto por uma matriz de dimensão 20×7, isto é, 20 progênies de *Eucalyptus grandis*, avaliadas em sete locais das regiões Sul e Sudeste do Brasil, tendo-se estudado a variável altura média (m). O segundo conjunto de dados (Santos, 2008) refere-se a uma matriz de dimensão 15×13, proveniente de um experimento com 15 variedades de cana-de-açúcar (*Saccharum officinarum* L.), em 13 locais do Brasil. A variável coletada foi teor de açúcar médio (pol de cana, %).

Cada matriz de dados originais foi submetida a retiradas aleatórias, em diferentes percentagens. Foram retirados 10, 20 e 35% dos dados, uma vez que, segundo Yan (2013), o número de dados ausentes em GxE geralmente é menor que 40%. O processo foi repetido 1.000 vezes, em cada conjunto de dados, para cada percentagem de retirada, tendo-se obtido 3.000 matrizes diferentes com dados ausentes. No total, foram gerados 6.000 conjuntos de dados incompletos, e, em cada um deles, os dados foram imputados com os seis algoritmos de IM descritos, por meio de um programa computacional implementado no R (R Development Core Team, 2014).

O processo de retirada aleatória para uma matriz X (n×p) foi o seguinte: números aleatórios entre 0 e 1 foram gerados no R com a função "runif"; para um valor fixo de r (0<r<1), se o (pi+j)-ésimo número aleatório for menor do que r, então o elemento na posição (i+1, j) da matriz foi deletado (i=0, 1, ..., n-1; j=1, ..., p). A proporção esperada de dados ausentes na matriz será r (Krzanowski, 1988). Essa técnica foi utilizada com r=0,1, 0,2 e 0,35.

Para medir a exatidão das imputações, foram adotadas as estatísticas T_{acc}, V_E e VQM, utilizadas por Bergamo et al. (2008). A estatística T_{acc} é uma medida de acurácia geral composta pela soma da variância combinada entre imputações dentro de posições (V_E) e o viés quadrático médio entre a média das imputações e o valor original retirado no estudo de simulação (VQM). As estatísticas são apresentadas pela equação $T_{acc}=V_E+VQM$, com

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

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

em que na é o número total de valores retirados da matriz GxE – cada valor retirado (l) tem sua correspondente posição (i, j) na matriz, isto é, na i-ésima linha e na j-ésima coluna; M é o número de imputações para o valor ausente l; $\hat{y}_{ij(m)}$ é a m-ésima imputação para o dito valor, obtida por meio de um dos métodos propostos; $\bar{Y}_l$ é a média das imputações produzidas para o valor ausente l; e VO_l é o valor original l no conjunto de dados original completo.

Considerou-se um bom método de imputação aquele que apresentou, conjuntamente, os menores valores de V_E e VQM. Ressalta-se que apenas o valor reduzido de

V_E não significaria uma boa qualidade da imputação, uma vez que o método pode ser tendencioso.

Resultados e Discussão

O método de imputação IMBiplotRes, para o conjunto de dados de eucalipto, forneceu sempre a maior variância entre imputações (V_E), independentemente da percentagem de imputação, enquanto o algoritmo com menor variância foi o IMBiplotG2, seguido pelo IMBiplotG1 e pelo IMBiplotG3 (Tabela 1). No entanto, para que se possa tomar uma decisão definitiva sobre qual seria o melhor algoritmo, é necessário que se analise ainda o VQM e a medida geral T_{acc}.

No mesmo conjunto de dados, o método com o menor viés para as percentagens de imputação foi o IMBiplotG5, seguido pelo IMBiplotG4 e pelo IMBiplotRes (Tabela 1). Em todos os casos, os sistemas de imputação mais viesados, ou seja, com maiores valores de VQM, foram IMBiplotG1 e IMBiplotG2. Assim, o algoritmo IMBiplotG5, por ter apresentado menor VQM, permitiu que se atingisse a maior similaridade entre as imputações e seus valores originais, o que resulta em maior precisão. Destaca-se que os métodos com menor V_E acabaram sendo os mais tendenciosos. Além disso, observou-se que o VQM aumentou à medida que a percentagem de imputação também aumentava, para todos os sistemas, o que é esperado, pois, se a informação disponível na matriz diminuir, o erro no modelo de imputação deve aumentar.

Contudo, para decidir qual seria o melhor método de imputação, a estatística de acurácia geral T_{acc} também deve ser considerada. Essa estatística leva em conta tanto a variância entre imputações quanto o viés quadrático médio (Figura 1). Todos os sistemas de imputação apresentam distribuições aproximadamente simétricas. O algoritmo com menor parâmetro de centralidade (distribuição mais próxima de zero) foi o IMBiplotG5, seguido do IMBiplotG4, em todas as percentagens de imputação consideradas. Assim, para 35% de retiradas aleatórias, as medianas da T_{acc} foram: 1,41 para IMBiplotG5; 1,46, para IMBiplotG4; 1,58, para IMBiplotG3; 1,67, para IMBiplotG2; 1,70, para IMBiplotRes; e 1,71, para IMBiplotG1. O algoritmo IMBiplotG5 também proporcionou o melhor desempenho nas outras percentagens; ou seja, apresentou as menores medianas dos valores de T_{acc}. Por último, ressalta-se que o algoritmo IMBiplotRes superou o IMBiplotG2 com as menores medianas de T_{acc}, nas taxas de 10 e 20% de imputação.

Para verificar a consistência dos resultados no estudo de simulação com dados de eucalipto, utilizou-se o conjunto de dados de cana-de-açúcar. Nesse conjunto de dados, o método com a menor média de variância

Tabela 1. Média e mediana da variância combinada entre imputações (V_E) e do viés quadrático médio (VQM), sob diferentes percentagens (10, 20 e 35%) de retirada aleatória de dados, para o conjunto de dados de eucalipto (*Eucalyptus grandis*).

Método	10%		20%		35%	
	Média	Mediana	Média	Mediana	Média	Mediana
			V_E			
IMBiplotRes	0,2782	0,2716	0,2590	0,2551	0,2294	0,2258
IMBiplotG1	0,0018	0,0017	0,0016	0,0016	0,0014	0,0014
IMBiplotG2	0,0008	0,0005	0,0006	0,0004	0,0004	0,0003
IMBiplotG3	0,0032	0,0031	0,0030	0,0029	0,0028	0,0028
IMBiplotG4	0,0192	0,0181	0,0154	0,0159	0,0106	0,0074
IMBiplotG5	0,0037	0,0019	0,0032	0,0018	0,0030	0,0015
			VQM			
IMBiplotRes	1,2428	1,1724	1,3575	1,3187	1,5009	1,4824
IMBiplotG1	1,6509	1,5719	1,6820	1,6665	1,7210	1,7158
IMBiplotG2	1,5829	1,5078	1,6257	1,6167	1,6786	1,6758
IMBiplotG3	1,4454	1,3579	1,5067	1,5054	1,5897	1,5863
IMBiplotG4	1,2361	1,1540	1,3323	1,3011	1,4638	1,4544
IMBiplotG5	1,1748	1,0942	1,2814	1,2417	1,4347	1,4137

entre imputações (V_E) foi o IMBiplotG2, para todas as percentagens consideradas, seguido por IMBiplotG1 e IMBiplotG3 (Tabela 2). Entretanto, a variância foi maximizada com uso do IMBiplotRes, também em todas as situações simuladas. Portanto, tanto nos dados de eucalipto quanto nos de cana-de-açúcar, a

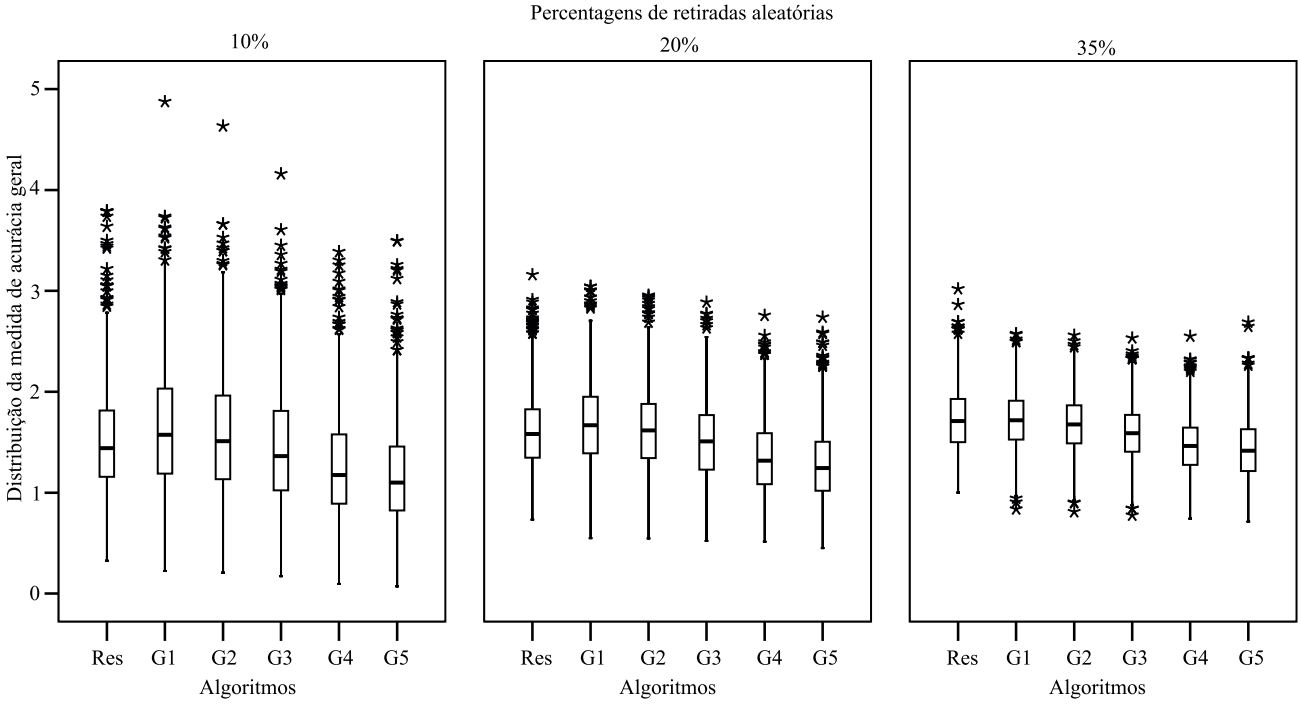

Figura 1. Distribuição da medida de acurácia geral (T_{acc}), com uso dos algoritmos IMBiplotRes (Res) e IMBiplotGh (G1, G2, G3, G4 e G5), para o conjunto de dados de eucalipto (*Eucalyptus grandis*).

Tabela 2. Média e mediana da variância combinada entre imputações (V_E) e do viés quadrático médio (VQM), sob diferentes percentagens (10, 20 e 35%) de retirada aleatória de dados, para o conjunto de dados de cana-de-açúcar (*Saccharum officinarum*).

Método	10%		20%		35%	
	Média	Mediana	Média	Mediana	Média	Mediana
			V_E			
IMBiplotRes	0,0904	0,0892	0,0919	0,0903	0,0935	0,0928
IMBiplotG1	0,0006	0,0006	0,0005	0,0005	0,0004	0,0004
IMBiplotG2	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001
IMBiplotG3	0,0020	0,0022	0,0016	0,0017	0,0008	0,0004
IMBiplotG4	0,0080	0,0063	0,0054	0,0050	0,0042	0,0041
IMBiplotG5	0,0022	0,0012	0,0019	0,0005	0,0009	0,0003
			VQM			
IMBiplotRes	0,3598	0,3376	0,4270	0,4142	0,5141	0,5095
IMBiplotG1	0,7330	0,7164	0,7359	0,7296	0,7561	0,7521
IMBiplotG2	0,6937	0,6764	0,7036	0,6993	0,7301	0,7274
IMBiplotG3	0,6058	0,5862	0,6351	0,6321	0,6808	0,6785
IMBiplotG4	0,4382	0,4090	0,4950	0,4861	0,5652	0,5653
IMBiplotG5	0,3476	0,3209	0,4163	0,4025	0,5015	0,4998

V_E teve o mesmo comportamento. Quanto ao VQM, o método com menor viés foi o IMBiplotG5, seguido pelo IMBiplotRes e pelo IMBiplotG4, em todas as percentagens consideradas. Os métodos mais tendenciosos foram, novamente, os com menor V_E, ou seja, IMBiplotG1, IMBiplotG2 e IMBiplotG3. Da mesma forma, nos dados de cana-de-açúcar, também constatou-se que o VQM aumentava com o aumento na percentagem de retirada aleatória, para todos os sistemas considerados.

Quanto à distribuição da T_{acc} dos dados de cana-de-açúcar, os seis algoritmos foram agrupados conforme seu desempenho (Figura 2). Assim, o primeiro grupo, de alto desempenho, foi composto por IMBiplotG5, IMBiplotG4 e IMBiplotRes, e o segundo grupo, por IMBiplotG3, IMBiplotG2 e IMBiplotG1. Também, com os dados de cana-de-açúcar, o melhor método de imputação foi o IMBiplotG5, que minimizou a estatística de acurácia geral em todas as percentagens.

Em ambos os estudos de simulação, o melhor desempenho foi atingido pelo IMBiplotG5, seguido do IMBiplotG4; portanto, os pesos w a serem considerados no modelo de imputação multiplicativo devem ser maiores do que 0,75. Esse resultado, no entanto, foi diferente do obtido no sistema de imputação descrito por Bergamo et al. (2008), que propuseram o intervalo de 0,4 a 0,6.

Para confirmar os pesos encontrados no presente trabalho, procedeu-se a simulações com o conjunto de dados de produtividade média (kg ha^{-1}) publicados por Yan et al. (2007) para 18 cultivares de trigo (*Triticum aestivum* L.), avaliadas em nove ambientes, em Ontário, no Canadá. Novamente, o IMBiplotG5 foi o método mais eficiente. Dessa forma, os pesos (ou expoentes) sugeridos para os autovalores de uma IM a partir da imputação biplot seriam: 0,96, 0,97, 0,98, 0,99 e 1. Destaca-se que o IMBiplotRes apresentou resultados inconsistentes, pois apresentou baixo desempenho

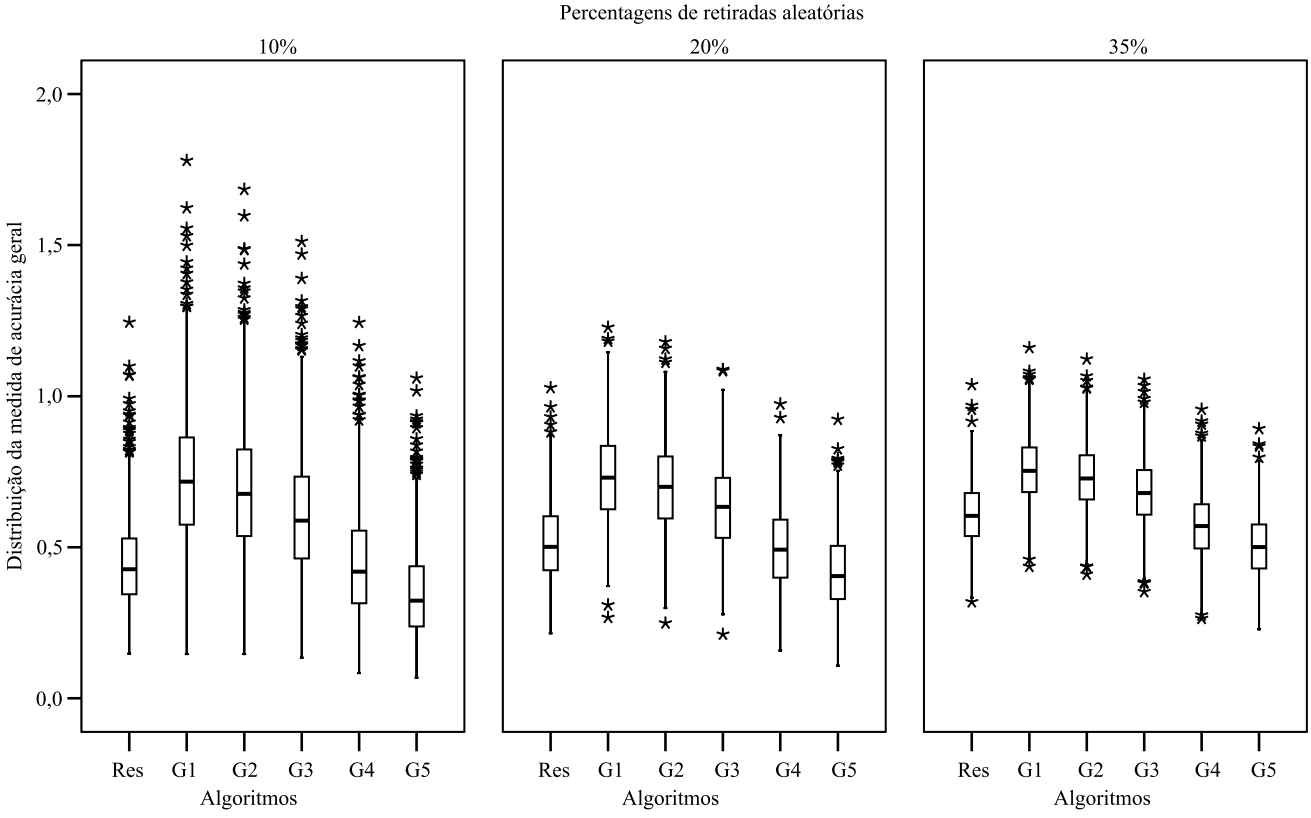

Figura 2. Distribuição da medida de acurácia geral (T_{acc}), com uso dos algoritmos IMBiplotRes (Res) e IMBiplotGh (G1, G2, G3, G4 e G5), para o conjunto de dados de cana-de-açúcar (*Saccharum officinarum*).

no conjunto de eucalipto, mas fez parte dos melhores métodos no conjunto de dados de cana-de-açúcar.

Os algoritmos de IM apresentados no presente trabalho têm como base a imputação biplot de Yan (2013), que utiliza uma aproximação de posto 2 para qualquer matriz GxE. Recentemente, Yang et al. (2009) afirmaram que todas as matrizes não podem ser analisadas com essa aproximação, pois, para algumas matrizes experimentais, ela, e especificamente o biplot, não é suficiente nem apropriada para determinar genótipos ganhadores ou mega-ambientes. Por esse motivo, sugere-se o seguinte esquema simples de pré-processamento de dados, para uma matriz GxE incompleta.

Suponha a matriz GxE X (n×p) com dados ausentes. Antes de aplicar os métodos propostos aqui, deve-se encontrar o posto de X, e uma maneira rápida de fazê-lo é por meio de validação cruzada (VC). Na literatura, vários esquemas de VC sobre dados incompletos podem ser encontrados, mas os recomendados no presente trabalho são os fornecidos por Husson & Josse (2013) e Wong (2013), que implementaram, nos pacotes imputation e missMDA do R (R Development Core Team, 2014), métodos que combinam a regularização com a análise de componentes principais (ACP), e a DVS com o algoritmo EM (Perry 2009; Josse & Husson, 2012b). Esses métodos proporcionam o posto da matriz X incompleta, que pode ser utilizado para IM com os sistemas sugeridos. Se o posto fosse diferente de 2, somente seria necessário inseri-lo no modelo multiplicativo de imputação. Finalmente, os algoritmos apresentados são de fácil implementação computacional e o código pode ser solicitado aos autores.

Conclusões

1. Os métodos de imputação múltipla (IM) propostos não dependem de pressuposições distribucionais ou estruturais e não têm restrições quanto ao padrão ou ao mecanismo de ausência de dados em experimentos genótipo x ambiente (GxE), ou em qualquer conjunto de dados que possa ser arranjado de forma matricial.

2. Com os sistemas de IM descritos, é possível obter uma estimativa da variância entre as imputações que represente a incerteza sobre os valores verdadeiros a serem imputados.

Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), ao Programa de Estudantes-Convênio de Pós-graduação (PEC-PG), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Academia de Ciências para os Países em Desenvolvimento (CNPq-TWAS), pelo apoio financeiro.

Referências

ALLISON, P.D. **Handling missing data by maximum likelihood**. 2012. Available at: <http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf>. Accessed on: 14 Aug. 2014.

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; DIAS, C.T. dos S. Data imputation in trials with genotype x environment interaction. **Interciencia**, v.36, p.444-449, 2011.

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; KRZANOWSKI, W.J.; DIAS, C.T. dos S. Deterministic imputation in multienvironment trials. **ISRN Agronomy**, v.2013, 2013. DOI: 10.1155/2013/978780.

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. DOI: 10.1590/S0103-90162008000400015.

BUUREN, S. van. **Flexible imputation of missing data**. Boca Raton: CRC Press, 2012. 343p. DOI: 10.1201/b11826.

CECCARELLI, S.; GRANDO, S.; BAUM, M. Participatory plant breeding in water-limited environments. **Experimental Agriculture**, v.43, p.411-435, 2007. DOI: 10.1017/S0014479707005327.

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. DOI: 10.2135/cropsci2003.0865.

GABRIEL, K.R. Le biplot – outil d'exploration de données multidimensionelles. **Journal de la Société Française de Statistique**, v.143, p.5-55, 2002.

GABRIEL, K.R. The biplot graphic display of matrices with application to principal component analysis. **Biometrika**, v.58, p.453-467, 1971. DOI: 10.1093/biomet/58.3.453.

GARCÍA-PEÑA, M.; ARCINIEGAS-ALARCÓN, S.; BARBIN, D. Imputação de dados climáticos utilizando a decomposição por valores singulares: uma comparação empírica. **Revista Brasileira de Meteorologia**, v.29, 2014. DOI: 10.1590/0102-778620130005.

GAUCH JUNIOR, H.G. A simple protocol for AMMI analysis of yield trials. **Crop Science**, v.53, p.1860-1869, 2013. DOI: 10.2135/cropsci2013.04.0241.

HAREL, O.; ZHOU, X.-H. Multiple imputation: review of theory, implementation, and software. **Statistics in Medicine**, v.26, p.3057-3077, 2007. DOI: 10.1002/sim.2787.

HUSSON, F.; JOSSE, J. **missMDA**: handling missing values with/in multivariate data analysis (principal component methods). Version 1.7. Available at: <http://CRAN.R-project.org/package=missMDA>. Accessed on: 15 out. 2013.

JOSSE, J.; HUSSON, F. Handling missing values in exploratory multivariate data analysis methods. **Journal de la Société Française de Statistique**, v.153, p.79-99, 2012a.

JOSSE, J.; HUSSON, F. Selecting the number of components in principal component analysis using cross-validation approximations. **Computational Statistics and Data Analysis**, v.56, p.1869-1879, 2012b. DOI: 10.1016/j.csda.2011.11.012.

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

KUMAR, A.; VERULKAR, S.B.; MANDAL, N.P.; VARIAR, M.; SHUKLA, V.D.; DWIVEDI, J.L.; SINGH, B.N.; SINGH, O.N.; SWAIN, P.; MALL, A.K.; ROBIN, S.; CHANDRABABU, R.; JAIN, A.; HAEFELE, S.M.; PIEPHO, H.P.; RAMAN, A. High-yielding, drought-tolerant, stable rice genotypes for the shallow rainfed lowland drought-prone ecosystem. **Field Crops Research**, v.133, p.37-47. 2012. DOI: 10.1016/j.fcr.2012.03.007.

LAVORANTI, O.J. **Estabilidade e adaptabilidade fenotípica através da reamostragem "Bootstrap" no modelo AMMI**. 2003. 166 p. Tese (Doutorado) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

LITTLE, R.J.A.; RUBIN, D.B. **Statistical analysis with missing data**. 2nd ed. Hoboken: Wiley, 2002. 408p. DOI: 10.1002/9781119013563.

PERRY, P.O. **Cross-validation for unsupervised learning**. 2009. 165p. Thesis (PhD) – Stanford University, Stanford.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2014.

RÄSSLER, S.; RUBIN, D.B.; ZELL, E.R. Imputation. **WIREs Computational Statistics**, v.5, p.20-29, 2013. DOI: 10.1002/wics.1240.

RODRIGUES, P.C.; PEREIRA, D.G.S.; MEXIA, J.T. A comparison between joint regression analysis and the additive main and multiplicative interaction model: the robustness with increasing amounts of missing data. **Scientia Agricola**, v.68, p.697-705, 2011. DOI: 10.1590/S0103-90162011000600012.

RUBIN, D.B. **Multiple imputation for nonresponse in surveys**. New York: John Wiley and Sons, 1987. 258p. DOI: 10.1002/9780470316696.

RUBIN, D.B. Multiple imputation in sample surveys – a phenomenological Bayesian approach to nonresponse. **Proceedings of the Survey Research Methods Section, American Statistical Association**, p.20-34, 1978. Available at: <https://www.amstat.org/sections/srms/proceedings/papers/1978_004.pdf>. Accessed on: 14 Aug. 2014.

SANTOS, É.G.D. dos. **Interação genótipos x locais em cana-de-açúcar e perspectivas de estratificação ambiental**. 2008. 63p. Dissertação (Mestrado) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

SRIVASTAVA, M.S.; DOLATABADI, M. Multiple imputation and other resampling scheme for imputing missing observations. **Journal of Multivariate Analysis**, v.100, p.1919-1937, 2009. DOI: 10.1016/j.jmva.2009.06.003.

WONG, J. **Imputation**. Version 2.0.1. Available at: <http://CRAN.Rproject.org/package=imputation>. Accessed on: 15 out. 2013.

YAN, W. Biplot analysis of incomplete two-way data. **Crop Science**, v.53, p.48-57, 2013. DOI: 10.2135/cropsci2012.05.0301.

YAN, W.; HOLLAND, J.B. A heritability-adjusted GGE biplot for test environment evaluation. **Euphytica**, v.171, p.355-369, 2010. DOI: 10.1007/s10681-009-0030-5.

YAN, W.; KANG, M.S.; MA, B.; WOODS, S.; CORNELIUS, P.L. GGE biplot vs. AMMI analysis of genotype-by-environment data. **Crop Science**, v.47, p.641-653, 2007. DOI: 10.2135/cropsci2006.06.0374.

YAN, W.; PAGEAU, D.; FRÉGEAU-REID, J.; DURAND, J. Assessing the representativeness and repeatability of test locations for genotype evaluation. **Crop Science**, v.51, p.1603-1610, 2011. DOI: 10.2135/cropsci2011.01.0016.

YANG, R.-C.; CROSSA, J.; CORNELIUS, P.L.; BURGUEÑO, J. Biplot analysis of genotype x environment interaction: proceed with caution. **Crop Science**, v.49, p.1564-1576, 2009. DOI: 10.2135/cropsci2008.11.0665.

ZHANG, P. Multiple imputation: theory and method. **International Statistical Review**, v.71, p.581-592, 2003. DOI: 10.1111/j.1751-5823.2003.tb00213.x.

Recebido em 7 de abril de 2014 e aprovado em 27 de agosto de 2014