IMGAMMI: ALGORITMO DE IMPUTAÇÃO PARA EXPERIMNTOS.

Pedro Marinho Amoedo[1]*(PQ), Clodoaldo Pires Araujo[2] (PQ), Manoel Carlos de Oliveira Junior[3] (PQ), Paulo Sergio Ribeiro da Silva[4] (PQ) e Jessica Caroline Correa da Silva[5] (IC).

* pamoedo@ufam.edu.br.

[1,5] Instituto de Ciências Sociais, Educação e Zootecnia-ICSEZ/UFAM
[3] Universidade Federal do Amazonas-UFAM
[2,4] Centro de estudos Superiores de Parintins-CESP/UEA.

Palavras Chave: AMMI generalizado, Estatística Tacc, Imputação.

Introdução

As análises mediante abordagem de modelo de efeitos principais aditivos e interação multiplicativa (AMMI) exigem que as matrizes de dados, provenientes de experimentos multiambientais, sejam completas, o que em geral não ocorre. Excelentes alternativas para contornar o problema das ausências para posterior análise são os métodos de imputações de dados. A imputação é uma das técnicas comumente empregada para resolver o problema das ausências, estima os dados ausentes por valores plausíveis e posteriormente as análises são realizadas sobre os dados completados[1,2,3]. Trabalhos bem aceitos para preencher as ausências em experimentos mutiambientais são os métodos que fazem uso da decomposição em valores singulares (DVS) de uma matriz, como o algoritmo EM-AMMI[4], o algoritmo EM+DVS[5], o método de imputação múltipla livre de distribuição (IMLD)[6], o método de imputação Biplot[2] e entre outros.

Assim, citado alguns aspectos sobre imputação de dados em experimentos multiambientais, tem-se como objetivo propor um algoritmo de imputação múltipla a partir de uma extensão do método EM-GAMMI e os resíduos simples do modelo de regressão linear, uma combinação do modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI) com o algoritmo EM e o resíduo simples da regressão linear.

Material e Métodos

Algoritmo de imputação múltipla IMGAMMI.

OS dados: foram realizadas simulações de retiradas aleatórias nas porcentagens de 10%, 20% e 30% sobre um conjunto de dados reais, delineamento em blocos ao acaso, estudo da resistência de soja à praga foliar, publicado por[7]. Este processo foi repetido 100 vezes para cada porcentagem de retirada, obtendo 300 matrizes diferentes com valores ausentes. Em seguida, para cada uma das matrizes com valores ausentes simulados, foram feitas as imputações. As imputações foram obtidas pelos algoritmos EM-GAMMI, IMGAMMI e EM+DVS. Na rotina do algoritmo IMGAMMI utilizou-se dos modelos GAMMI Poisson e GAMMI Gaussiano, com até dois termos multiplicativos, e respectivas funções de ligações logarítmica e identidade. As simulações de retiradas aleatórias de valores deram-se mediante uso da função SimIm do pacote multivariado ImputeR.

Para avaliação do método, tomou-se as estatísticas raiz do erro quadrático médio padronizado – NRMSE e a medida total (ou geral) de acurácia (Tacc).

O algoritmo: O método EM-GAMMI fornece no final, depois de atingir convergência, uma matriz X(2), que contém tanto imputações quanto estimativas dos valores observados. Então, como passo seguinte, calcula-se uma 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), t =1,...,M, e cada elemento de Ω_t é selecionado aleatoriamente com reposição dos elementos de $\hat{\varepsilon}$. A imputação múltipla é realizada ao se substituir os elementos ausentes da matriz X original pelos valores correspondentes de cada uma das t matrizes definidas por $X(2)+\Omega_t$. Utilizou-se t=5, número de imputações, pois de acordo com Van Buuren[8], t=5 imputações são suficientes para fazer inferências válidas. Dessa forma, obteve-se o processo de imputação múltipla com resíduos simples por meio de um modelo multiplicativo generalizado, o qual foi denominado de imputação múltipla GAMMI (IMGAMMI)..

Resultados e Discussão

A Tabela 3 apresenta as medianas da NRMSE para os métodos de imputações comparados: imputação múltipla IMGAMMI, EM+DVS e o EM-AMMI para cada nível de porcentagem de retirada. Nesta, pelo critério de menor valor da NRMSE, o método com melhor desempenho foi o

IMGAMMI, independentemente do nível de retirada. Assim, para o nível de ausência de 10%, o procedimento com melhor desempenho foi o IMGAMMI0 (mediana=0,199); para o nível de 20%, foi o IMGAMMI0 (mediana=0,2076) e para o nível de 30%, foi o IMGAMMI0 (mediana=0,1956). Observa-se ainda, que o procedimento EM+DVS obteve melhores resultados que o método clássico EM-AMMI com até dois termos multiplicativos para todos os níveis de retirada.

Tabela 1 – NRMSE mediana para os níveis de retirada aleatória (de 10%, 20% e 30%).

	10%	20%	30%
Método	Mediana	Mediana	Mediana
EM+DVS	0,971	0,9518	1,0006
EM-AMMI0	0,965	1,029	1,1922
EM-AMMI1	1,462	1,892	2,0158
EM-AMMI2	1,076	1,047	1,0323
IMGAMMI0	0,199	0,2076	0,1956
IMGAMMI1	0,23	0,2126	0,2024
IMGAMMI2	0,221	0,2341	0,2046

A Tabela 2 apresenta valores pequenos da estatística geral de acurácia - Tacc, indicam que o método é eficiente para predizer valores imputados próximos dos valores observados.

Tabela 2 - Mediana da distribuição da Tacc para os níveis de retirada aleatória (10%, 20% e 30%).

	Estatística Tacc		
Método	Me(10%)	Me(20%)	Me(30%)
IMGAMMI0	0,0001000	0,0001016	0,000102
IMGAMMI1	0,0001031	0,0001006	0,000098
IMGAMMI2	0,0000993	0,0001001	0,0000986

Me: mediana

Os resultados das estatísticas NMRSE e Tacc, pelos seus baixos valores obtidos, são evidências fortes da boa qualidade, eficácia, do algoritmo de imputação IMGAMMI. Também, pode ser observado uma maior eficiência do método IMGAMMI, em termos de NMRSE, quando comparado a métodos clássicos, como o procedimento EM+DVS e o EM-AMMI.

Conclusões

O procedimento IMGAMMI apresentou ser eficiente como técnica de imputação de dados, foi superior a métodos usuais com bastante destaque na literatura como o EM-AMMI e EM+DVS. Portanto, o procedimento de imputação múltipla IMGAMMI é uma excelente alternativa para fazer imputações de dados em experimentos multiambientais.

Agradecimentos

A Universidade Federal do Amazonas-ICSEZ/UFAM.
A Fundação Universitas de Estudos Amazônicos – F.UEA.
Big data analysis para monitoramento de dados- MULTI BDA/MULTILASER.

[1]ARCINIEGAS-ALARCÓN, S; GARCÍA-PEÑA, M; RODRIGUES, P. C. New multiple imputation methods for genotype-by-environment data that combine singular value decomposition and Jackknife resampling or weighting schemes.Computers and Electronics in Agriculture, v. 176, p. 105617, 2020.

[2]YAN, W. Biplot analysis of incomplete two-way data. Crop Science, v.53, p.48-57, 2013.

[3]RODRIGUES, P. C.; PEREIRA, D. G. S.; Mexia, J. T. A comparison between joint regression analysisand the additive main and multiplicative interaction model: the robustness with increasing amountsof missing data. Scientia Agricola, Piracicaba, v.68, p.679-686, 2011.

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

[5]PERRY, P. O. Cross-validation for unsupervised learning, 2009. 153p. Dissertation, Stanford University, 2009.

[6]BERGAMO, G. C.; Dias, C. T. d. S.; KRZANOWSKI, W. J. Distribution-free multiple imputation in an interaction matrix through singular value decomposition. \textit{Scientia Agricola}, Piracicaba, Braz., v.65, p.422-427, 2008.

[7]HADI, A. F.; MATTJIK, A.; SUMERTAJAYA, I. Generalized ammi models for assessing the endurance of soybean to leaf pest. Jurnal Ilmu Dasar, v.11, p.151-159, 2010.

[8]VAN BUUREN, S. Flexible imputation of missing data. 2.ed. Boca Raton: CRC Press, 2018. 416p..