**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Imputação de dados em experimentos multiambientais:
novos algoritmos utilizando a decomposição por valores singulares**

Sergio Arciniegas Alarcón

Tese apresentada para obtenção do título de Doutor
em Ciências. Área de concentração: Estatística e
Experimentação Agronômica

**Piracicaba
2015**

Sergio Arciniegas Alarcón
Estatístico

**Imputação de dados em experimentos multiambientais:
novos algoritmos utilizando a decomposição por valores singulares**

Orientador:
Prof. Dr. **CARLOS TADEU DOS SANTOS DIAS**

Tese apresentada para obtenção do título de Doutor
em Ciências. Área de concentração: Estatística e
Experimentação Agronômica

**Piracicaba
2015**

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Arciniegas Alarcón, Sergio

Imputação de dados em experimentos multiambientais: novos algoritmos utilizando a decomposição por valores singulares / Sergio Arciniegas Alarcón. - - Piracicaba, 2015.
110 p. : il.

Tese (Doutorado) - - Escola Superior de Agricultura "Luiz de Queiroz".

1. Decomposição por valores singulares 2. Observações ausentes 3. Interação genótipo por ambiente 4. Melhoramento de plantas 5. Imputação 6. AMMI I. Título

CDD 519.50285
A674i

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

AGRADECIMIENTOS

Al Profesor Carlos Tadeu dos Santos Dias por la dirección y apoyo para desarrollar este trabajo.

Al Profesor Wojtek Krzanowski de la Universidad de Exeter - UK por las ideas y conocimientos que compartió conmigo a la distancia.

Al Profesor Décio Barbin, por transmitirme su conocimiento y colaborarme en aspectos de la investigación.

A todos los Profesores y funcionarios del Programa de Posgrado en Estadística y Experimentación Agronómica del Departamento de Ciencias Exactas de la ESALQ/USP, por la ayuda.

A Marisol García Peña, por estar a mi lado en todo este camino. Para ti toda mi gratitud y amor infinitos ¡Gracias Totales! Bjks

A mis amigos brasileños Kuang y José “Newton” por toda la ayuda dentro y fuera de los salones de clase.

Al programa PEC-PG de CAPES por el soporte económico para realizar este doctorado en Brasil.

A quienes olvido en este momento, pero directa o indirectamente contibuyeron a llegar al final feliz de este proyecto académico.

SUMÁRIO

RESUMO	9
ABSTRACT	11
1 INTRODUÇÃO	13
Referências	14
2 IMPUTAÇÃO DETERMINÍSTICA EM EXPERIMENTOS MULTIAMBIEN- TAIS	15
Resumo	15
Abstract	15
2.1 Introdução	16
2.2 Material e métodos	17
2.2.1 Imputação de dados utilizando o método de validação cruzada do autovetor	17
2.2.2 Os dados	20
2.2.3 Estudo de simulação	21
2.2.4 Critérios de comparação	21
2.3 Resultados	23
2.3.1 Dados poloneses de ervilha	23
2.3.2 Dados brasileiros de algodão	27
2.3.3 Dados espanhóis de feijão	30
2.4 Discussão	32
Referências	34
3 IMPUTAÇÃO MÚLTIPLA LIVRE DE DISTRIBUIÇÃO EM TABELAS IN- COMPLETAS DE DUPLA ENTRADA	41
Resumo	41
Abstract	41
3.1 Introdução	42
3.2 Material e métodos	43
3.3 Resultados e discussão	47
3.4 Conclusões	51
Referências	51
4 IMPUTAÇÃO DE VALORES AUSENTES EM EXPERIMENTOS MULTIAM- BIENTAIS USANDO A DECOMPOSIÇÃO POR VALORES SINGULARES: UMA COMPARAÇÃO EMPÍRICA	55
Resumo	55
Abstract	55
4.1 Introdução	56
4.2 Material e métodos	57
4.2.1 Métodos de imputação	57

4.2.2 Os dados	62
4.2.3 Estudo de simulação	63
4.2.4 Critérios de comparação	64
4.3 Resultados e discussão	65
4.3.1 Matriz “Lavoranti”	65
4.3.2 Matriz “Flores”	67
4.3.3 Matriz “Santos”	70
4.4 Conclusões	72
Referências	73
5 UMA METODOLOGIA ALTERNATIVA PARA IMPUTAR DADOS AUSEN- TES EM ENSAIOS COM INTERAÇÃO GENÓTIPO POR AMBIENTE: AL- GUNS ASPECTOS NOVOS	77
Resumo	77
Abstract	77
5.1 Introdução	78
5.2 Material e métodos	79
5.2.1 Imputação de dados usando GabrielEigen	79
5.2.2 Modificações propostas	80
5.2.3 Os dados	81
5.2.4 Estudo de simulação	81
5.2.5 Critérios de comparação	82
5.3 Resultados e discussão	83
5.3.1 Matriz “Denis-Baril”	83
5.3.2 Matriz “Calinski”	83
5.4 Conclusões	87
Referências	87
6 TRABALHOS FUTUROS	91
APÊNDICES	93
A Imputação de dados climáticos utilizando a decomposição por valores singulares: Uma comparação empírica	95
Resumo	95
Abstract	95
A.1 Introdução	96
A.2 Material e métodos	97
A.2.1 A decomposição por valores singulares de uma matriz (DVS)	97
A.2.2 Métodos de imputação	97
A.2.3 Características dos dados	100
A.2.4 Estudo de simulação	101
A.2.5 Critérios de comparação	102

A.3 Resultados e discussão	103
A.4 Conclusões	107
Referências	107

RESUMO

Imputação de dados em experimentos multiambientais: novos algoritmos utilizando a decomposição por valores singulares

As análises biplot que utilizam os modelos de efeitos principais aditivos com interação multiplicativa (AMMI) requerem matrizes de dados completas, mas, frequentemente os ensaios multiambientais apresentam dados faltantes. Nesta tese são propostas novas metodologias de imputação simples e múltipla que podem ser usadas para analisar dados desbalanceados em experimentos com interação genótipo por ambiente ($G \times E$). A primeira, é uma nova extensão do método de validação cruzada por autovetor (Bro et al, 2008). A segunda, corresponde a um novo algoritmo não-paramétrico obtido por meio de modificações no método de imputação simples desenvolvido por Yan (2013). Também é incluído um estudo que considera sistemas de imputação recentemente relatados na literatura e os compara com o procedimento clássico recomendado para imputação em ensaios ($G \times E$), ou seja, a combinação do algoritmo de Esperança-Maximização com os modelos AMMI ou EM-AMMI. Por último, são fornecidas generalizações da imputação simples descrita por Arciniegas-Alarcón et al. (2010) que mistura regressão com aproximação de posto inferior de uma matriz. Todas as metodologias têm como base a decomposição por valores singulares (DVS), portanto, são livres de pressuposições distribucionais ou estruturais. Para determinar o desempenho dos novos esquemas de imputação foram realizadas simulações baseadas em conjuntos de dados reais de diferentes espécies, com valores retirados aleatoriamente em diferentes porcentagens e a qualidade das imputações avaliada com distintas estatísticas. Concluiu-se que a DVS constitui uma ferramenta útil e flexível na construção de técnicas eficientes que contornem o problema de perda de informação em matrizes experimentais.

Palavras-chave: Decomposição por valores singulares; Observações ausentes; Interação genótipo por ambiente; Melhoramento de plantas; Imputação; AMMI

ABSTRACT

Data imputation in multi-environment trials: new algorithms using the singular value decomposition

The biplot analysis using the additive main effects and multiplicative interaction models (AMMI) require complete data matrix, but often multi-environments trials have missing values. This thesis proposed new methods of single and multiple imputation that can be used to analyze unbalanced data in experiments with genotype by environment interaction ($G \times E$). The first is a new extension of the cross-validation method by eigenvector (Bro et al., 2008). The second, corresponds to a new non-parametric algorithm obtained through modifications of the simple imputation method developed by Yan (2013). Also is included a study that considers imputation systems recently reported in the literature and compares them with the classic procedure recommended for imputation in trials ($G \times E$), it means, the combination of the Expectation-Maximization (EM) algorithm with the additive main effects and multiplicative interaction (AMMI) model or EM-AMMI. Finally, are supplied generalizations of simple imputation described by Arciniegas-Alarcón et al. (2010) that combines regression with lower-rank approximation of a matrix. All methodologies are based on singular value decomposition (SVD), so, are free of any distributional or structural assumptions. In order to determine the performance of the new imputation schemes were performed simulations based on real data set of different species, with values deleted randomly at different percentages and the quality of the imputations was evaluated using different statistics. It was concluded that SVD provides a useful and flexible tool for the construction of efficient techniques that circumvent the problem of missing data in experimental matrices.

Keywords: Singular value decomposition; Missing values; Genotype-by-environment interaction; Plant breeding; Imputation; AMMI

1 INTRODUÇÃO

Em vários estudos realizados nas diferentes áreas do conhecimento são planejados estatisticamente experimentos que envolvem dois fatores, cada fator pode apresentar um número diferente de níveis e geralmente o resultado é uma tabela de dupla entrada, onde cada casela da tabela contém a medição da variável de interesse, mas alguns problemas na coleta das observações ou no mesmo planejamento podem causar dificuldades nas análises posteriores. Nesse momento, o analista de dados pode encontrar dificuldades como dados discrepantes, falta de repetições se os custos foram considerados e dados faltantes por causa de questões climáticas, morte de animais, plantas danificadas, dados digitados ou mensurados erroneamente e muitas outras situações que acontecem quando se trabalha com dados reais.

No caso de dados ausentes, a perda de informação produz delineamentos desbalanceados que perdem a simetria e por exemplo, os testes de hipóteses de interesse como a diferença entre os tratamentos, precisam de um desenvolvimento teórico particular. Às vezes, quando se tem um grande número de observações faltantes, algumas funções paramétricas não são estimáveis e um cálculo errado de graus de liberdade para as somas de quadrados pode gerar inferências inadequadas e conclusões pouco verdadeiras do experimento. Uma possível solução ao problema consiste em repetir o experimento sob condições similares e dessa maneira obter novos valores para as observações perdidas. No entanto, esta solução, embora ideal, pode não ser viável em termos de tempo e dinheiro. Dodge (1985) e Little e Rubin (2002) apresentam duas das aproximações mais usadas para resolver esse problema. Dodge (1985) apresenta as considerações teóricas para fazer as análises baseadas unicamente nas observações presentes, enquanto Little e Rubin (2002) descrevem uma grande quantidade de métodos para a imputação de dados com o objetivo de preencher as caselas vazias.

É bem comum que nos experimentos bifatoriais se tenha apenas uma observação por casela e adicionalmente dados ausentes. Um exemplo desta situação pode apresentar-se nos experimentos multiambiente, em que cultivares são estudadas em diferentes localidades ou ambientes e cada casela na estrutura de dados experimentais pode corresponder à média das repetições de cada combinação dos níveis dos fatores. Esses tipos de experimentos têm muita aplicação no melhoramento genético de plantas e são conhecidos como experimentos genótipo por ambiente ($G \times E$).

Muitas vezes os experimentos multiambientes são desbalanceados e vários genótipos não são testados em alguns locais. 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 que foram atualmente observadas, além disso, é bem conhecido que uma

das melhores opções na análise da interação ($G \times E$) 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 fazer a decomposição por valores singulares (DVS) da matriz de resíduos não aditivos, mas essa DVS não pode ser calculada na ocorrência de valores ausentes.

Para resolver esses problemas pode ser utilizada a imputação de dados e o objetivo dessa tese é contribuir na área por meio de propostas não-paramétricas baseadas na DVS de uma matriz. Assim, a tese está composta de cinco capítulos e escrita de tal forma, que o leitor, se desejar, pode abordar cada capítulo independentemente, sem precisar de informações contidas em outras partes do manuscrito.

Os assuntos tratados são descritos a seguir. No capítulo 2 se estuda a imputação simples determinística em ensaios ($G \times E$), fornecendo novos sistemas de imputação utilizando o método de validação cruzada por autovetor. No capítulo 3 se propõe uma nova metodologia de imputação múltipla livre de distribuição para tabelas incompletas de dupla entrada. No capítulo 4 são estudados algoritmos de imputação recentemente relatados na literatura e comparados com a metodologia clássica para experimentos com interação ($G \times E$). No capítulo 5 se apresenta uma generalização da imputação Gabriel-Eigen em experimentos multiambientais, que utiliza uma mistura entre regressão e aproximação de posto inferior de uma matriz. Finalmente, no capítulo 6 são considerados trabalhos futuros para continuar a linha de pesquisa.

Referências

- DENIS, J.B.; BARIL C.P. Sophisticated models with numerous missing values: the multiplicative interaction model as an example. **Biuletyn Oceny Odmian**, Poznan, v.24-25, p.33-45, 1992.
- DODGE Y. **Analysis of experiments with missing data**. New York: John Wiley, 1985. 499p.
- DUARTE, J.B.; VENCOSKY, R. **Interação genótipo $\times$ ambiente: uma introdução à análise "AMMI"**. Riberão Preto: Sociedade Brasileira de Genética, 1999. 60p. (Série Monografias).
- LITTLE, R. J; RUBIN D.B. **Statistical analysis with missing data**. 2nd ed. New York: John Wiley, 2002. 381p.

2 IMPUTAÇÃO DETERMINÍSTICA EM EXPERIMENTOS MULTIAMBIENTAIS *

Resumo

Este capítulo propõe cinco novos métodos de imputação para experimentos desbalanceados com interação genótipo $\times$ ambiente ($G \times E$). Os métodos utilizam validação cruzada por autovetor e são baseados em um esquema iterativo com a decomposição por valores singulares (DVS) de uma matriz. Para testar os métodos, foi realizado um estudo de simulação usando três matrizes completas de dados reais, provenientes de ensaios $G \times E$ de ervilha, algodão e feijão respectivamente. A simulação de desbalanceamento foi feita por meio de retiradas aleatórias de 10%, 20% e 40% dos valores em cada matriz. A qualidade das imputações foi avaliada com o modelo de efeitos principais aditivos com interação multiplicativa (AMMI), utilizando a diferença preditiva média (RMSPD) entre os parâmetros genotípicos e ambientais dos dados originais e os dados completados por imputação. A metodologia proposta é livre de pressuposições distribucionais ou estruturais e não tem restrições em relação ao mecanismo ou padrão dos dados faltantes.

Palavras-chave: Modelos AMMI; Validação cruzada; Autovetor; Interação genótipo por ambiente; Imputação; Valores omissos; DVS

Abstract

This chapter proposes five new imputation methods for unbalanced experiments with genotype by-environment interaction ($G \times E$). The methods use cross-validation by eigenvector, based on an iterative scheme with the singular value decomposition (SVD) of a matrix. To test the methods, we performed a simulation study using three complete matrices of real data, obtained from ($G \times E$) interaction trials of peas, cotton, and beans, and introducing lack of balance by randomly deleting in turn 10%, 20%, and 40% of the values in each matrix. The quality of the imputations was evaluated with the additive main effects and multiplicative interaction model (AMMI), using the root mean squared predictive difference (RMSPD) between the genotypes and environmental parameters of the original data set and the set completed by imputation. The proposed methodology does not make any distributional or structural assumptions and does not have any restrictions regarding the pattern or mechanism of missing values.

Keywords: AMMI models; Cross-validation; Eigenvector; Genotype-by-environment interaction; Imputation; Missing values; SVD

*Capítulo aceito para publicação na revista ISRN Agronomy.

2.1 Introdução

No melhoramento genético de plantas, os ensaios multiambientais são importantes para testar a adaptação geral e específica das cultivares. Uma cultivar desenvolvendo-se em diferentes ambientes mostrará uma flutuação significativa do desempenho na produção relativa a outras cultivares. Essas mudanças são influenciadas por diferentes condições ambientais e são referidas como interação genótipo por ambiente ou $G \times E$. Muitas vezes, os experimentos $G \times E$ são desbalanceados porque vários genótipos não são testados em alguns ambientes.

Uma maneira muito comum de analisar dados provenientes desse tipo de estudos é imputar as observações ausentes e posteriormente, aplicar procedimentos já estabelecidos sobre a matriz de dados completada (observados + imputados), por exemplo, o modelo de efeitos aditivos com interação multiplicativa - AMMI ou a regressão fatorial (Gauch 2006; van Eeuwijk et al. 2005, 2007; Romagosa et al. 2008; Arciniegas-Alarcón e Dias 2009b). Uma aproximação alternativa é trabalhar com dados incompletos sob a estrutura do modelo misto com estimativas baseadas em máxima verossimilhança (Kang et al. 2004).

Vários métodos de imputação têm sido sugeridos na literatura para resolver o problema de dados faltantes. Um dos primeiros foi feito por Freeman (1975), que sugere imputar as observações ausentes de maneira iterativa minimizando a soma de quadrados do erro e sobre a tabela completa fazer as análises da interação $G \times E$, diminuindo os graus de liberdade pelo número de dados faltantes. Um trabalho bem aceito nestes experimentos foi o desenvolvido por Gauch e Zobel (1990), que fizeram a imputação usando o algoritmo EM e o modelo AMMI ou EM-AMMI.

Algumas alternativas desse procedimento usando estatística multivariada, especificamente análise de agrupamentos, foram descritas em Godfrey et al. (2002) e Godfrey (2004). Raju (2002) propôs o algoritmo EM-AMMI utilizando os ambientes como aleatórios e sugere aplicar uma estatística robusta aos dados omissos nas análises de estabilidade. Mandel (1993) propôs que a imputação seja feita nas tabelas incompletas de dupla entrada usando funções lineares das linhas ou colunas.

Outros estudos recomendados por van Eeuwijk e Kroonenberg (1998) que têm tido bons resultados no caso de observações faltantes para experimentos $G \times E$ foram desenvolvidos por Denis (1991), Calinski et al. (1992) e Denis e Baril (1992). Eles encontraram que usando imputações por meio dos mínimos quadrados alternados com modelos de interação bilinear ou estimativas AMMI baseadas em submodelos robustos podem-se obter resultados tão bons como os encontrados com o algoritmo EM. Adicionalmente, Calinski et al. (1999) introduziram um algoritmo que combina a decomposição por valores singulares (DVS) de uma matriz com o algoritmo EM, obtendo resultados muito úteis para

aqueles experimentos nos quais os mínimos quadrados alternados têm alguns problemas, como por exemplo, falhas de convergência (Piepho, 1995).

Recentemente, Bergamo et al. (2008) propuseram um método de imputação múltipla livre de distribuição que foi avaliado por Arciniegas-Alarcón (2008) e comparado por Arciniegas-Alarcón e Dias (2009a) com algoritmos que usam modelos de efeitos fixos em um estudo de simulação a partir de dados reais. Entretanto, um método de imputação determinística sem pressuposições estruturais nem distribucionais para os experimentos multiambientais foi proposto por Arciniegas-Alarcón et al. (2010). O método utiliza uma mistura de regressão com a aproximação de posto inferior de uma matriz.

Finalmente, outros estudos para fazer análises de experimentos multiambientais com dados incompletos podem ser encontrados na literatura. Por exemplo, metodologias para análises de estabilidade têm sido estudadas por Raju e Bathia (2003) e Raju et al. (2006). Nessa mesma linha, Pereira et al. (2007), Rodrigues et al. (2011) e Rodrigues (2012) avaliaram a robustez da análise de regressão conjunta e dos modelos AMMI sem utilizar imputação de dados.

Dada a informação histórica sobre imputação de dados em experimentos de dois fatores $G \times E$, o objetivo desse capítulo é propor um algoritmo de imputação determinística livre de pressuposições distribucionais e estruturais a partir de uma extensão do método de validação cruzada do autovetor apresentado por Bro et al. (2008).

2.2 Material e métodos

2.2.1 Imputação de dados utilizando o método de validação cruzada do autovetor

O método de validação cruzada foi apresentado por Bro et al. (2008) para encontrar o número ótimo de componentes principais em qualquer conjunto de dados que possa ser arranjado em forma matricial. Nesta aproximação, modelos para análises de componentes principais (ACP) são calculados com uma ou várias amostras deixadas fora “left-out” e o modelo é usado para prever tais amostras. O método usa validação cruzada “leave-one-out” e no mesmo estudo se mostrou mais eficiente do que metodologias bem reconhecidas e utilizadas na estatística multivariada, como as apresentadas por Wold (1978) e Eastment e Krzanowski (1982). Por essas características, Arciniegas-Alarcón et al. (2011) usaram o método para determinar o melhor modelo de efeitos principais aditivos e interação multiplicativa (AMMI) em experimentos genótipo por ambiente ($G \times E$). A metodologia é apresentada a seguir.

Passo 1. Considere a matrix $\mathbf{X}$ ($n \times p$) com elementos x_{ij} ($i = 1, \dots, n; j = 1, \dots, p$). A matriz é dividida em grupos disjuntos, cada grupo é deletado um por vez (leave-one-out)

e um modelo ACP ($\mathbf{T}, \mathbf{P}$) é obtido dos elementos remanescentes, resolvendo

$$\min \left\| \mathbf{X}^{(-i)} - \mathbf{TP}^T \right\|_m^2 \quad (2.1)$$

com $m \leq \min(n-1, p-1)$. Aqui $\mathbf{X}^{(-i)}$ representa a matriz depois de deletar o i -ésimo grupo (leave-one-out), $\|\bullet\|^2$ define a norma quadrada de Frobenius, $\mathbf{P}^T \mathbf{P} = \mathbf{I}$, e $\mathbf{T}, \mathbf{P}$ são as matrizes de escores e pesos com dimensões $(n-1) \times m$ e $p \times m$ respectivamente, em que p é o número de colunas e m é o número de componentes. Note-se que nesse método o grupo deletado corresponde à i -ésima linha de $\mathbf{X}$ e segundo Smilde et al. (2004), o modelo (2.1) pode ser reescrito em termos da decomposição em valores singulares (DVS)

$$\mathbf{X}^{(-i)} = \mathbf{UDV}^T = \sum_{k=1}^m \mathbf{u}_k d_k \mathbf{v}_k^T \quad (2.2)$$

em que $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m]$, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m]$, $\mathbf{D} = [d_1, d_2, \dots, d_m]$, $\mathbf{T} = \mathbf{UD}$ e $\mathbf{P} = \mathbf{V}$.

Passo 2. Estima-se o escore

$$\mathbf{t}^{(-j)T} = \mathbf{x}_i^{(-j)T} \mathbf{P}^{(-j)} \left(\mathbf{P}^{(-j)T} \mathbf{P}^{(-j)} \right)^{-1} \quad (2.3)$$

em que $\mathbf{P}^{(-j)T}$ é a matriz de pesos encontrada no passo 1 com a j -ésima linha excluída. $\mathbf{x}_i^{(-j)T}$ é o vetor linha que contém a i -ésima linha de $\mathbf{X}$ sem o j -ésimo elemento.

Passo 3. Estima-se o elemento x_{ij} por

$$\hat{x}_{ij}^{(m)} = \mathbf{t}^{(-j)T} \mathbf{p}_j^T \quad (2.4)$$

$\mathbf{p}_j^T$ é a j -ésima linha de $\mathbf{P}$

Passo 4. Encontra-se o erro de predição do elemento (i, j) , $e_{ij}^{(m)} = x_{ij} - \hat{x}_{ij}^{(m)}$.

Passo 5. Obtém-se o valor do critério

$$PRESS(m) = \sum_{i=1}^n \sum_{j=1}^p \left(e_{ij}^{(m)} \right)^2$$

Para fazer a imputação de dados faltantes na matriz dos experimentos ($\mathbf{G} \times \mathbf{E}$), uma mudança é proposta no método seguindo o trabalho de Krzanowski (1988), Bergamo et al. (2008) e Arciniegas-Alarcón et al. (2010) utilizando a decomposição em valores singulares de uma matriz (Good, 1969).

Inicialmente, suponha que $n > p$ e a matriz $\mathbf{X}$ tem vários valores ausentes; no caso que $n < p$ a matriz deve ser primeiro transposta. Os dados faltantes são substituídos pela média da correspondente coluna $\bar{x}_j$ e uma vez feito isso, a matriz é padronizada por

colunas, subtraindo $\bar{x}_j$ e dividindo por s_j (em que $\bar{x}_j$ e s_j representam a média e o desvio padrão da j -ésima coluna respectivamente). O procedimento do autovetor usando a DVS nas expressões (2.2) - (2.4) é aplicado à matriz padronizada para encontrar a imputação do elemento (i, j) , denotado por $\hat{x}_{ij}^{(m)}$. Após a imputação, a matriz deve ser retornada a sua escala original, $x_{ij} = \bar{x}_j + s_j \hat{x}_{ij}^{(m)}$.

Neste ponto, a matriz já não tem observações ausentes, mas, as imputações são bastante básicas e precisam ser refinadas. Nos trabalhos mencionados anteriormente, um esquema iterativo é requerido e as iterações continuam até que as imputações alcancem convergência (isto é, estabilidade em valores imputados sucessivos), mas, Calinski et al. (1999) mostraram que essa convergência nem sempre é necessária utilizando um método que combina o algoritmo EM com a DVS. Levando isso em conta, nesta pesquisa foi estabelecido de antemão o número de iterações entre 0 e 3, bem como o processo que alcança convergência. Com relação ao aspecto computacional, a convergência pode depender fortemente da dimensão da matriz analisada e também da estrutura dos dados (magnitude das correlações, proporção de dados faltantes, etc.), mas, por exemplo, a convergência no método DVS de Hastie et al. (1999) é alcançada usualmente entre 5 e 6 iterações e no método de Bergamo et al. (2008) a convergência é alcançada entre 20 e 50 iterações no máximo.

Por outro lado, a imputação de dados depende diretamente das equações (2.2) e (2.3). A equação (2.2) precisa da escolha prévia do número de componentes (m) para extrair da DVS. Krzanowski (1988) e Bergamo et al. (2008) fizeram $m = \min\{n - 1, p - 1\}$ com o objetivo de utilizar a máxima quantidade de informação disponível, mas, Hedderley and Wakeling (1995) afirmaram que se a estimação está baseada na escolha de um único número fixo de dimensões, algumas das menores dimensões podem ser essencialmente aleatórias. Isto pode influenciar a imputação dentro de um esquema iterativo e produzir que as estimações fiquem presas em um ciclo do processo evitando a convergência. Para resolver esse problema, eles sugerem envolver um processo que verifique a taxa de convergência e no caso de não atingir um determinado critério especificado, o número de dimensões deve ser reduzido. Outra opção com resultados satisfatórios, sugerida por Josse et al. (2011) para escolher o m ótimo pode ser por meio da validação cruzada baseada unicamente nos dados observados, mas, o custo computacional provavelmente será alto.

Levando em conta todas as situações acima descritas no presente estudo, para imputar cada observação ausente da matriz $\mathbf{X}$, o valor de m na equação (2.2) poderá ser diferente em cada DVS calculada e escolhido de acordo com o critério utilizado por Arciniegas-Alarcón et al. (2010). Assim, m é selecionado tal que $\left(\frac{\sum_{k=1}^m d_k^2}{\sum_{k=1}^{\min\{n-1, p-1\}} d_k^2} \right) \approx 0.75$.

Além disso, na equação (2.3) a inversa generalizada de Moore-Penrose pode ser utilizada no lugar da matriz inversa clássica como já foi estudado por Dias e Krzanowski (2003).

Nesta pesquisa cinco métodos de imputação têm sido avaliados, denotados por Eigenvector0, Eigenvector1, Eigenvector2, Eigenvector3 e Eigenvector, em que o número indica as iterações empregadas, enquanto no caso do Eigenvector o processo é iterado até atingir convergência nas imputações.

Os métodos de imputação descritos são todos determinísticos e têm como vantagem sobre outros métodos de imputação estocástica (imputação múltipla paramétrica) que os valores imputados são determinados de maneira única e sempre fornecerão os mesmos resultados quando sejam aplicados sobre um conjunto de dados específico. Isto não é necessariamente verdadeiro para os métodos de imputação estocástica (Bello, 1993).

2.2.2 Os dados

Para avaliar os métodos de imputação, foram considerados três conjuntos de dados completos publicados em Calinski et al. (2009), Farias (2005) e Flores et al. (1998) respectivamente. Em cada caso, os dados foram obtidos a partir de delineamentos experimentais aleatorizados em blocos com repetições e cada referência oferece uma excelente descrição do planejamento se detalhes específicos fossem requeridos.

O primeiro conjunto de dados “Calinski” está composto por uma matriz de dimensão 18×9 , com 18 variedades de ervilha avaliadas em 9 locais diferentes da Polônia. O experimento foi conduzido pelo Research Centre for Cultivar Testing, Słupia Wielka, em que a variável de interesse foi produção média medida em t/ha.

O segundo conjunto de dados “Farias” foi obtido 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 avaliaram 15 cultivares de algodão em 27 localidades dos estados brasileiros de Mato Grosso, Mato Grosso do Sul, Goiás, Minas Gerais, Rondônia, Maranhão e Piauí. A variável estudada foi produtividade de algodão em caroço (kg/ha).

O terceiro conjunto de dados “Flores” é uma matriz com dimensão 15×12 , com 15 variedades de feijão avaliadas em 12 ambientes da Espanha. O experimento foi conduzido pela RAEA - Red Andaluza de Experimentación Agraria-, em que a variável de interesse foi produção média (kg/ha).

As três matrizes de dados consideradas nesta pesquisa continham unicamente as médias de produção de cada genótipo em cada ambiente, mas, os métodos propostos aqui funcionam para qualquer conjunto de dados arranjado em forma matricial. Por exemplo,

se fosse disponibilizada a informação correspondente a repetições, uma aproximação sugerida por Bello (1995) é escrever o experimento em termos de um modelo de regressão linear clássico para obter o vetor de respostas, a matriz de delineamento e posteriormente juntar tudo em uma única matriz e sobre essa matriz aplicar os métodos propostos neste capítulo.

2.2.3 Estudo de simulação

Cada matriz de dados original (“Calinski”, “Farias” e “Flores”) foi submetida a retiradas aleatórias de três porcentagens, 10%, 20% e 40%. O processo foi repetido em cada conjunto de dados 1000 vezes para cada porcentagem de retirada, obtendo 3000 matrizes diferentes com valores omissos. No total, foram gerados 9000 conjuntos de dados incompletos e em cada um, os dados foram imputados com os 5 algoritmos *Eigenvec*_*tor* descritos acima por meio de um programa computacional implementado no R (R Development Core Team, 2012).

O processo de retirada aleatória para uma matriz $\mathbf{X}$ ($n \times p$) 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 foi menor do que r , então o elemento na posição $(i + 1, j)$ da matriz foi deletado ($i = 0, 1, \dots, n; j = 1, \dots, p$). A proporção esperada de dados ausentes na matriz será r (Krzanowski, 1988). Essa técnica foi utilizada com $r = 0.1, 0.2$ e 0.4 , ou seja, 10%, 20% e 40% respectivamente.

2.2.4 Critérios de comparação

Em geral, o objetivo após a imputação consiste em estimar parâmetros de um modelo a partir de tabelas com informação completa. Um dos modelos mais utilizados nos experimentos com interação genótipo $\times$ ambiente é o modelo AMMI (Gauch 1988,1992) e por essa razão, os algoritmos propostos neste capítulo serão comparados por meio dos parâmetros genotípicos e ambientais do modelo AMMI ajustado, utilizando a raiz quadrada da diferença preditiva média – RMSPD (Dias e Krzanowski, 2003). O modelo AMMI é apresentado brevemente à seguir.

O modelo ANOVA usual de dupla entrada para analisar dados de ensaios genótipo por ambiente está definido por

$$y_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}$$

($i = 1, \dots, n; j = 1, \dots, p$) em que μ , a_i , b_j , $(ab)_{ij}$ e e_{ij} são respectivamente, a média geral, os efeitos principais genotípicos e ambientais, a interação genótipo $\times$ ambiente e um termo de erro associado ao i -ésimo genótipo no j -ésimo local. Assume-se que todos os efeitos são

fixos, exceto o erro. As seguintes restrições de reparametrização são feitas: $\sum_i (ab)_{ij} = \sum_j (ab)_{ij} = \sum_i a_i = \sum_j b_j = 0$. O modelo AMMI implica que as interações podem ser expressas pela soma de termos multiplicativos. O modelo está dado por

$$y_{ij} = \mu + a_i + b_j + \theta_1 \alpha_{i1} \beta_{j1} + \theta_2 \alpha_{i2} \beta_{j2} + \cdots + e_{ij} \quad (2.5)$$

em que θ_l , α_{il} e β_{jl} ($l = 1, 2, \dots, \min(n-1, p-1)$) são estimados pela DVS da matriz de erros depois de ajustar a parte aditiva. θ_l é estimado pelo l -ésimo valor singular da DVS, α_{il} e β_{jl} são estimados pelos valores dos autovetores genotípicos e ambientais correspondentes a θ_l . Regressões alternantes podem ser utilizadas no lugar da DVS (Gabriel, 2002); dependendo do número de termos multiplicativos, estes modelos podem ser chamados AMMI0, AMMI1, etc.

Uma exigência inerente ao modelo AMMI é a escolha prévia do número de componentes multiplicativos (Dias e Krzanowski, 2006; García-Peña e Dias, 2009; Hongyu, 2012). Rodrigues (2012) fez uma análise exaustiva da literatura relacionada e concluiu que usualmente dois ou três componentes podem ser usados para a modelagem, porque em geral, um componente não é suficiente para capturar o padrão inteiro de resposta nos dados, mas, com mais de três componentes se captura uma grande quantidade de ruído, além das evidentes dificuldades para visualizá-los.

Assim, para as matrizes originais “Calinski”, “Farias” e “Flores” foram ajustados os modelos AMMI2 e AMMI3. Os mesmos modelos foram ajustados para cada um dos 9000 conjuntos de dados completados por imputação e cada conjunto de parâmetros foi comparado com o correspondente conjunto dos dados originais utilizando a RMSPD da seguinte maneira:

$$\begin{aligned} RMSPD(gen) &= \sqrt{\frac{\sum_{i=1}^{NG} (a_i - \hat{a}_i)^2}{NG}}; & RMSPD_l(genmult) &= \sqrt{\frac{\sum_{h=1}^l \sum_{i=1}^{NG} (\alpha_{ih} - \hat{\alpha}_{ih})^2}{(NG)l}}; \\ RMSPD(env) &= \sqrt{\frac{\sum_{j=1}^{NE} (b_j - \hat{b}_j)^2}{NE}}; & RMSPD_l(envmult) &= \sqrt{\frac{\sum_{h=1}^l \sum_{j=1}^{NE} (\beta_{jh} - \hat{\beta}_{jh})^2}{(NE)l}}. \end{aligned}$$

Aqui $RMSPD(gen)$ representa a RMSPD entre os parâmetros estimados para efeitos principais genotípicos dos dados originais a_i e os correspondentes parâmetros obtidos dos conjuntos completados por imputação $\hat{a}_i$. $RMSPD(env)$ representa a RMSPD entre os parâmetros estimados para efeitos principais ambientais dos dados originais b_j e os correspondentes parâmetros obtidos dos conjuntos completados por imputação

$\hat{b}_j$. $RMSPD_l(genmult)$ representa a RMSPD equivalente para os pares de parâmetros estimados correspondentes aos componentes multiplicativos genotípicos α_{ih} , $\hat{\alpha}_{ih}$ e $RMSPD_l(envmult)$ representa a RMSPD equivalente para os pares de parâmetros estimados correspondentes aos componentes multiplicativos ambientais β_{jh} , $\hat{\beta}_{jh}$. Nas estatísticas, NG representa o número de genótipos, NE o número de ambientes e $l = 2$ ou 3 , dependendo do modelo considerado AMMI2 ou AMMI3.

O melhor método de imputação será aquele com menores valores de RMSPD em cada caso. Resumindo, em cada conjunto de dados simulado com observações ausentes, foram aplicados os métodos Eigenvector, Eigenvector0, Eigenvector1, Eigenvector2 e Eigenvector3. Com os dados completados (observados + imputados) foram ajustados os modelos AMMI2, AMMI3 para o cálculo das respectivas estatísticas RMSPD. Para visualizar melhor qualquer diferença entre os algoritmos, os valores das RMSPD foram padronizados e sobre eles foi feita diretamente a comparação. Note-se que por causa da escala padronizada, os valores das estatísticas podem ser positivos ou negativos.

2.3 Resultados

2.3.1 Dados poloneses de ervilha

A Figura 2.1 apresenta a distribuição da $RMSPD(gen)$ na escala padronizada para o conjunto de dados “Calinski”, mostrando cada método de imputação e cada porcentagem. Observa-se que a distribuição para Eigenvector é assimétrica à esquerda e dita assimetria aumenta conforme aumenta a porcentagem de dados omissos. Em geral, a distribuição do Eigenvector está acima de zero e com o aumento de dados faltantes tende a se concentrar acima de 1, quer dizer que este método teve as maiores diferenças entre os parâmetros genotípicos aditivos dos dados reais e dos dados completados por imputação.

Figura 2.1 - Gráfico de caixas da distribuição da $RMSPD(gen)$ no conjunto de dados “Calinski”

O melhor método segundo a $RMSPD(gen)$ é o Eigenvector1, o método com apenas uma iteração. Este método apresenta a menor mediana nas porcentagens de 10% e 20%. Na porcentagem de 40% as medianas do Eigenvector0 e Eigenvector1 são praticamente iguais na figura, mas, o Eigenvector1 continua sendo preferível porque tem a menor dispersão. Assim, com Eigenvector1 se obteve as menores diferenças entre os parâmetros genotípicos aditivos dos dados reais e dos dados completados.

A Figura 2.2 mostra a $RMSPD(env)$ na escala padronizada para o conjunto de dados “Calinski”. Observa-se um comportamento muito similar ao encontrado com a $RMSPD(gen)$. Novamente o método Eigenvector apresenta as maiores diferenças entre os parâmetros ambientais aditivos dos dados reais e dos dados completados porque é o algoritmo que maximiza a $RMSPD(env)$. Neste caso, a $RMSPD(env)$ é minimizada com Eigenvector0 e Eigenvector1 e em todas as porcentagens de observações ausentes têm medianas aproximadamente iguais. No entanto, Eigenvector1 tem a menor dispersão e isso o torna o método mais recomendável.

Figura 2.2 - Gráfico de caixas da distribuição da $RMSPD(env)$ no conjunto de dados “Calinski”

A análise com gráfico de caixas foi útil para determinar o melhor método de imputação nas distribuições da $RMSPD(gen)$ e da $RMSPD(env)$, mas, no caso das $RMSPD_2(envmult)$, $RMSPD_2(genmult)$, $RMSPD_3(genmult)$ e $RMSPD_3(envmult)$, análises mais formais podem ser utilizadas para comparar as distribuições; por exemplo, o teste não-paramétrico de Friedman com posterior teste de Wilcoxon se o primeiro for significativo (Sprent e Smeeton, 2001).

A Tabela 2.1 apresenta as estatísticas do teste de Friedman. Observa-se que existe uma diferença significativa entre os métodos de imputação para as porcentagens de retirada de 10% e 20%, enquanto com o 40% os cinco métodos produzem resultados equivalentes. Depois do teste geral, é preciso fazer comparações para as duas porcentagens menores.

Tabela 2.1 - Teste de Friedman para a $RMSPD_l(\bullet)$ padronizada - Conjunto de dados “Calinski”

Porcentagem	Estatística							
	$RMSPD_2(genmult)$		$RMSPD_2(envmult)$		$RMSPD_3(genmult)$		$RMSPD_3(envmult)$	
	Friedman	valor-p	Friedman	valor-p	Friedman	valor-p	Friedman	valor-p
10%	15,6256	0,0036	34,4896	0,0000	34,9368	0,0000	30,4928	0,0000
20%	10,7848	0,0291	11,3688	0,0227	16,7144	0,0022	11,1104	0,0254
40%	2,8416	0,5847	2,5568	0,6345	4,9496	0,2925	5,9448	0,2033

Na Tabela 2.2 se encontra o teste de Wilcoxon para encontrar os métodos que diferem. Quando foi utilizada a $RMSPD_2(genmult)$ para 10%, Eigenvector1 teve diferenças significativas com os outros quatro métodos. Para 20%, Eigenvector1 foi estatisticamente diferente do Eigenvector, Eigenvector2 e Eigenvector3. Nessa mesma porcentagem, Eigenvector também apresentou resultados diferentes do Eigenvector0 e do Eigenvector3. Juntando as diferenças estatísticas encontradas com os testes não-paramétricos sobre a $RMSPD_2(genmult)$ e o gráfico de caixas correspondente na Figura 2.3, pode se afirmar que para 10% e 20% o método mais eficiente é o Eigenvector1, pois minimiza a mediana e apresenta a menor dispersão se for comparado com Eigenvector e Eigenvector0. Os cinco métodos apresentaram resultados similares para a porcentagem de 40%.

Tabela 2.2 - Teste de Wilcoxon para a $RMSPD_2(\bullet)$ padronizada - Conjunto de dados “Calinski”

Porcentagem	$RMSPD_2(genmult)$		$RMSPD_2(envmult)$	
	10%	20%	10%	20%
Comparação	Wilcoxon	Wilcoxon	Wilcoxon	Wilcoxon
Eigen - Eigen0	-0,2913	-2,4166 *	-0,8459	-1,7890
Eigen - Eigen1	-3,4322 *	-2,6145 *	-4,5972 *	-1,5540
Eigen - Eigen2	-1,0087	-1,0783	-2,0225 *	-0,1250
Eigen - Eigen3	-1,3178	-2,0335 *	-2,4155 *	-0,7970
Eigen1 - Eigen0	-2,0468 *	-0,1261	-2,8270 *	-0,5490
Eigen2 - Eigen0	-0,2997	-1,6703	-0,2598	-2,0420 *
Eigen3 - Eigen0	-0,3213	-1,3256	-0,0852	-1,6410
Eigen2 - Eigen1	-3,3075 *	-2,7537 *	-4,3006 *	-2,5030 *
Eigen3 - Eigen1	-3,5483 *	-2,2389 *	-5,0405 *	-2,3590 *
Eigen3 - Eigen2	-0,4955	-0,0203	-0,9271	-0,6170

*: Diferença significativa ao nível de significância de 5%

A Tabela 2.2 mostra os resultados do teste de Wilcoxon para as porcentagens de 10% e 20% de dados faltantes, utilizando a $RMSPD_2(envmult)$. Existem diferenças significativas entre o Eigenvector e os algoritmos Eigenvector2, Eigenvector3 e Eigenvector1

Figura 2.3 - Gráfico de caixas da distribuição da $RMSPD_2(genmult)$ no conjunto de dados “Calinski”

com 10% de retirada aleatória. Foram encontradas diferenças entre Eigenvector1 e Eigenvector0, Eigenvector2 e Eigenvector3 respectivamente. Para o 20% de retirada aleatória, Eigenvector1 foi diferente do Eigenvector2 e o Eigenvector3, além da diferença entre o Eigenvector0 e o Eigenvector2.

Entretanto, na Tabela 2.3 são apresentados os resultados do teste de Wilcoxon da $RMSPD_3(envmult)$ e da $RMSPD_3(genmult)$ padronizadas. Com 10% e 20% de imputação se encontram diferenças significativas entre Eigenvector1 e Eigenvector, Eigenvector2 e Eigenvector3 respectivamente. Também foram detectadas diferenças significativas entre Eigenvector e os algoritmos Eigenvector0 e Eigenvector3.

Tabela 2.3 - Teste de Wilcoxon para a $RMSPD_3(\bullet)$ padronizada - Conjunto de dados “Calinski”

	$RMSPD_3(genmult)$		$RMSPD_3(envmult)$	
Porcentagem	10%	20%	10%	20%
Comparação	Wilcoxon	Wilcoxon	Wilcoxon	Wilcoxon
Eigen - Eigen0	-1,7875	-2,6026 *	-2,1574 *	-1,9962 *
Eigen - Eigen1	-4,8856 *	-3,1579 *	-4,4210 *	-2,6059 *
Eigen - Eigen2	-1,8055	-1,9022	-1,7068	-1,2073
Eigen - Eigen3	-2,6978 *	-2,0075 *	-3,1627 *	-1,3278
Eigen1 - Eigen0	-1,8186	-0,2928	-1,1885	-0,4978
Eigen2 - Eigen0	-1,0934	-1,0855	-1,2860	-1,0310
Eigen3 - Eigen0	-0,9545	-1,2510	-1,1276	-1,1751
Eigen2 - Eigen1	-4,9417 *	-2,3846 *	-4,1071 *	-2,0129 *
Eigen3 - Eigen1	-4,5703 *	-2,5499 *	-4,1410 *	-2,3572 *
Eigen3 - Eigen2	-0,1905	-0,7254	-0,1788	-0,8727

* : Diferença significativa ao nível de significância de 5%

Finalmente, foram feitos os gráficos de caixas para $RMSPD_2(envmult)$, $RMSPD_3(genmult)$ e $RMSPD_3(envmult)$, mas, não se apresentam porque o comportamento foi muito similar ao apresentado na Figura 2.3, confirmando que Eigenvector1 minimiza a mediana se for comparado com Eigenvector2 e Eigenvector3 e também, tem menor dispersão do que Eigenvector0. O método que sempre maximizou todas as estatísticas foi o Eigenvector e por isso é o menos recomendável.

2.3.2 Dados brasileiros de algodão

A Figura 2.4 apresenta a distribuição da $RMSPD(gen)$ na escala padronizada para o conjunto de dados “Farias”. A distribuição para Eigenvector0 é assimétrica à esquerda e dita assimetria diminui conforme aumenta a porcentagem de dados omissos. Nas três porcentagens consideradas, a distribuição do Eigenvector0 está acima de 1 e muito próxima a 2, quer dizer que este método teve as maiores diferenças entre os parâmetros genotípicos aditivos dos dados reais e dos dados completados (por imputação). Com 10% de imputação, os métodos Eigenvector, Eigenvector2 e Eigenvector3 têm uma mediana aproximadamente igual, mas a menor dispersão é atingida com Eigenvector2. Em geral, pode se afirmar que conforme aumenta a porcentagem de valores retirados, o Eigenvector atinge um melhor desempenho, pois minimiza a $RMSPD(gen)$. Um comportamento similar ao descrito anteriormente é mostrado para a $RMSPD(env)$ e pode ser observado na Figura 2.5.

Figura 2.4 - Gráfico de caixas da distribuição da $RMSPD(gen)$ no conjunto de dados “Farias”

Na Tabela 2.4 são apresentadas as estatísticas dos testes de Friedman para $RMSPD_2(genmult)$, $RMSPD_2(envmult)$, $RMSPD_3(genmult)$ e $RMSPD_3(envmult)$. Existe diferença significativa entre os métodos de imputação para todas as porcentagens de observações ausentes, portanto, serão feitas comparações múltiplas não-paramétricas dois a dois por meio do teste de Wilcoxon.

Figura 2.5 - Gráfico de caixas da distribuição da $RMSPD(env)$ no conjunto de dados “Farias”

Tabela 2.4 - Teste de Friedman para a $RMSPD_l(\bullet)$ padronizada - Conjunto de dados “Farias”

Porcentagem	Estatística							
	$RMSPD_2(genmult)$		$RMSPD_2(envmult)$		$RMSPD_3(genmult)$		$RMSPD_3(envmult)$	
	Friedman	valor-p	Friedman	valor-p	Friedman	valor-p	Friedman	valor-p
10%	452,1168	0,0000	444,0952	0,0000	228,6352	0,0000	201,9368	0,0000
20%	313,0696	0,0000	295,0152	0,0000	193,6624	0,0000	173,3472	0,0000
40%	49,8712	0,0000	32,3296	0,0000	25,5240	0,0000	10,8736	0,0280

A Tabela 2.5 mostra os testes de Wilcoxon para $RMSPD_2(genmult)$ e $RMSPD_2(envmult)$. Eles indicam que com 10% de imputação a maioria dos pares comparados teve uma diferença significativa, mas, por exemplo, Eigenvector1 não teve um desempenho diferente dos algoritmos Eigenvector, Eigenvector2 e Eigenvector3. Nas outras duas porcentagens, 20% e 40%, Eigenvector não difere estatisticamente do Eigenvector2 nem do Eigenvector3 que apresentam desempenhos similares.

A Tabela 2.6 mostra os testes de Wilcoxon sobre a $RMSPD_3(envmult)$. Com 10% de imputação, Eigenvector0 é diferente de todos os demais algoritmos, enquanto para a $RMSPD_3(genmult)$ na mesma porcentagem, Eigenvector1 foi estatisticamente diferente do Eigenvector2. Entretanto, com 20% e 40% de imputação, Eigenvector não mostrou diferenças com Eigenvector2 nem com o Eigenvector3 e da mesma forma, Eigenvector3 não foi diferente do Eigenvector2.

Para tirar uma conclusão definitiva, foram realizados os gráficos de caixas para $RMSPD_3(envmult)$, $RMSPD_3(genmult)$, $RMSPD_2(genmult)$ e $RMSPD_2(envmult)$, mas, somente um deles é apresentado porque o comportamento das distribuições é semelhante. Da Figura 2.6 pode se concluir que os métodos que minimizam a mediana em

Tabela 2.5 - Teste de Wilcoxon para a $RMSPD_2(\bullet)$ padronizada - Conjunto de dados “Farias”

Porcentagem	$RMSPD_2(genmult)$			$RMSPD_2(envmult)$		
	10%	20%	40%	10%	20%	40%
Comparação	Wilcoxon	Wilcoxon	Wilcoxon	Wilcoxon	Wilcoxon	Wilcoxon
Eigen - Eigen0	-12,7645 *	-11,8392 *	-5,5233 *	-12,0995 *	-11,3270 *	-4,4629 *
Eigen - Eigen1	-0,3505	-3,4137 *	-3,8021 *	-0,1890	-3,0716 *	-2,4969 *
Eigen - Eigen2	-2,6235 *	-0,7163	-1,4214	-2,4094 *	-0,4378	-0,3487
Eigen - Eigen3	-2,6720 *	-0,8664	-0,1908	-2,7633 *	-0,9311	-0,4897
Eigen1 - Eigen0	-16,5991 *	-11,7349 *	-2,6190 *	-16,9885 *	-11,6590 *	-2,5653 *
Eigen2 - Eigen0	-16,9878 *	-13,0576 *	-5,3009 *	-16,2317 *	-12,5528 *	-4,9312 *
Eigen3 - Eigen0	-13,6133 *	-12,6292 *	-5,8550 *	-12,8970 *	-11,9226 *	-5,2543 *
Eigen2 - Eigen1	-1,7703	-5,0465 *	-3,6028 *	-1,0721	-4,1600 *	-2,8340 *
Eigen3 - Eigen1	-0,5797	-4,2466 *	-4,3441 *	-0,0083	-3,6872 *	-3,5903 *
Eigen3 - Eigen2	-2,5865 *	-1,6592	-1,3257	-2,3910 *	-1,1199	-0,6422

* : Diferença significativa ao nível de significância de 5%

Tabela 2.6 - Teste de Wilcoxon para a $RMSPD_3(\bullet)$ padronizada - Conjunto de dados “Farias”

Porcentagem	$RMSPD_3(genmult)$			$RMSPD_3(envmult)$		
	10%	20%	40%	10%	20%	40%
Comparação	Wilcoxon	Wilcoxon	Wilcoxon	Wilcoxon	Wilcoxon	Wilcoxon
Eigen - Eigen0	-9,2191 *	-9,2224 *	-4,1232 *	-8,2742 *	-8,9679 *	-2,1050 *
Eigen - Eigen1	-1,1084	-2,2832 *	-3,3175 *	-0,1224	-2,2990 *	-2,0120 *
Eigen - Eigen2	-0,6928	-0,1061	-0,8429	-1,1718	-0,2890	-0,3286
Eigen - Eigen3	-0,9784	-0,1836	-0,2097	-1,7433	-0,2468	-0,5434
Eigen1 - Eigen0	-11,1032 *	-8,6574 *	-1,0162	-10,6797 *	-8,5424 *	-0,2532
Eigen2 - Eigen0	-11,7189 *	-9,8996 *	-3,5702 *	-11,2791 *	-9,5638 *	-2,5008 *
Eigen3 - Eigen0	-9,8820 *	-9,3163 *	-4,3048 *	-8,8932 *	-9,1492 *	-2,7670 *
Eigen2 - Eigen1	-2,2248 *	-3,7149 *	-3,0406 *	-1,4067	-3,5119 *	-2,6124 *
Eigen3 - Eigen1	-1,5319	-2,3342 *	-3,5506 *	-0,5309	-2,4132 *	-2,8674 *
Eigen3 - Eigen2	-0,3787	-0,2394	-0,9666	-0,8871	-0,2512	-0,0848

* : Diferença significativa ao nível de significância de 5%

todas as porcentagens são Eigenvector, Eigenvector2 e Eigenvector3. Pelas Tabelas 2.5 e 2.6, esses métodos são equivalentes.

Figura 2.6 - Gráfico de caixas da distribuição da $RMSPD_2(genmult)$ no conjunto de dados “Farias”

Em resumo, para o conjunto de dados “Farias”, com as seis estatísticas padronizadas, o Eigenvector sempre mostrou bons resultados e é o algoritmo recomendado.

2.3.3 Dados espanhóis de feijão

Na Figura 2.7 é mostrada a distribuição da $RMSPD(gen)$ na escala padronizada para o conjunto de dados “Flores”. O Eigenvector tem uma distribuição assimétrica à esquerda em todas as porcentagens e maximiza a mediana da $RMSPD(gen)$, portanto, é o método que apresenta as maiores diferenças entre os parâmetros genotípicos principais dos dados originais e dos completados por imputação. Com 10% de imputação, o Eigenvector0 é o método que apresenta melhor desempenho, com 20% o Eigenvector1 e com 40% o Eigenvector2; minimizando a mediana e levando a distribuição da $RMSPD(gen)$ ao fundo da escala padronizada.

Na Figura 2.8 se encontra um resultado similar, mas, utilizando a $RMSPD(env)$. Dessa figura pode-se descrever que com 20% de imputação Eigenvector0 e Eigenvector1 têm medianas aproximadamente iguais, mas, é preferível o Eigenvector1 por causa da menor dispersão. Com a $RMSPD(env)$ o Eigenvector0 tem distribuições assimétricas à direita e Eigenvector1, Eigenvector2 e Eigenvector3 têm distribuições aproximadamente simétricas.

A Tabela 2.7 apresenta os testes de Friedman para as estatísticas $RMSPD_2(genmult)$, $RMSPD_2(envmult)$, $RMSPD_3(genmult)$ e $RMSPD_3(envmult)$. Pode-se observar que existem diferenças significativas entre os métodos somente com 10% de imputação. Por essa razão, a atenção será restrita para essa porcentagem.

Figura 2.7 - Gráfico de caixas da distribuição da $RMSPD(gen)$ no conjunto de dados “Flores”

Figura 2.8 - Gráfico de caixas da distribuição da $RMSPD(env)$ no conjunto de dados “Flores”

Tabela 2.7 - Teste de Friedman para a $RMSPD_l(\bullet)$ padronizada - Conjunto de dados “Flores”

Porcentagem	Estatística							
	$RMSPD_2(genmult)$		$RMSPD_2(envmult)$		$RMSPD_3(genmult)$		$RMSPD_3(envmult)$	
	Friedman	valor-p	Friedman	valor-p	Friedman	valor-p	Friedman	valor-p
10%	23,1512	0,0001	39,0136	0,0000	24,0736	0,0001	26,1888	0,0000
20%	5,0296	0,2843	2,2608	0,6879	1,8144	0,7698	1,0936	0,8953
40%	5,2256	0,2649	3,7480	0,4412	8,1944	0,0847	1,6856	0,7933

A Tabela 2.8 mostra as 10 possíveis comparações dois a dois dos métodos considerando unicamente 10% de imputação e as estatísticas $RMSPD_2(genmult)$, $RMSPD_2(envmult)$, $RMSPD_3(genmult)$ e $RMSPD_3(envmult)$. De acordo com elas, todos os métodos diferem, exceto Eigenvector1-Eigenvector0. Adicionalmente, não foi encontrada significância na diferença dos pares Eigenvector1-Eigenvector2 com a $RMSPD_2(genmult)$ e no par Eigenvector2-Eigenvector0 com a $RMSPD_3(genmult)$ respectivamente.

Tabela 2.8 - Teste de Wilcoxon para a $RMSPD_l(\bullet)$ padronizada (10% de imputação) - Conjunto de dados “Flores”

Comparação	Estatística			
	$RMSPD_2$ (<i>genmult</i>)	$RMSPD_2$ (<i>envmult</i>)	$RMSPD_3$ (<i>genmult</i>)	$RMSPD_3$ (<i>envmult</i>)
Eigen - Eigen0	-3,1132 *	-3,9729 *	-2,9800 *	-3,2533 *
Eigen - Eigen1	-2,7033 *	-3,4193 *	-2,6193 *	-2,6122 *
Eigen - Eigen2	-2,2662 *	-2,9110 *	-2,7950 *	-2,1989 *
Eigen - Eigen3	-2,2427 *	-3,0329 *	-2,5279 *	-2,8083 *
Eigen1 - Eigen0	-1,3053	-1,9408	-0,6860	-0,9421
Eigen2 - Eigen0	-2,4441 *	-3,2769 *	-1,6886	-2,1223 *
Eigen3 - Eigen0	-3,2117 *	-3,8341 *	-2,3675 *	-2,7069 *
Eigen2 - Eigen1	-1,8444	-2,8314 *	-2,0155 *	-2,5541 *
Eigen3 - Eigen1	-2,3102 *	-2,9518 *	-2,3568 *	-2,3958 *
Eigen3 - Eigen2	-2,2854 *	-2,4862 *	-2,4622 *	-2,0679 *

* : Diferença significativa ao nível de significância de 5%

Finalmente, para tirar uma conclusão definitiva das quatro estatísticas analisadas nas Tabelas 2.7 e 2.8 é apresentado na Figura 2.9 o gráfico de caixas para $RMSPD_2(genmult)$. Foram realizados os gráficos das outras três estatísticas, mas, não se apresentam porque o comportamento é similar. Segundo o gráfico de caixas, o melhor método é o Eigenvector0 porque minimiza a mediana.

2.4 Discussão

Apresentaram-se neste estudo cinco métodos de imputação que foram testados por meio de um estudo de simulação baseado em três experimentos multiambientais utilizando seis estatísticas derivadas da $RMSPD$. De maneira global, para experimentos grandes, neste caso com 405 observações, o Eigenvector_ deve ser utilizado sob convergência, enquanto para experimentos menores, de 162 ou 180 observações, dois ciclos do processo são suficientes para obter bons resultados sem precisar da convergência.

Foram utilizados experimentos com diferentes espécies, em distintos países e em diferentes continentes. Alguns dos resultados foram conforme o esperado, mas, um fato

Figura 2.9 - Gráfico de caixas da distribuição da $RMSPD_2(genmult)$ no conjunto de dados “Flores” com 10% de imputação

importante é que o aspecto iterativo dos algoritmos propostos deve ser obrigatório quando os valores omissos são imputados em matrizes provenientes de experimentos $G \times E$.

Assim, fica uma questão natural de interesse para o pesquisador aplicado: Como escolher o método de imputação `Eigenvector_` mais apropriado para experimentos com diferentes tamanhos aos apresentados neste capítulo? A resposta para essa questão depende do objetivo da imputação, pois a imputação pode ser utilizada com diferentes fins: Para estabelecer uma ou várias combinações genótipo-ambiente que por algum motivo não foram observadas ou fazer algum tipo de modelagem estatística posterior à imputação. Os critérios de escolha podem ser tão diversos, mas, para o primeiro objetivo seria natural encontrar os erros de imputação associados a cada um dos métodos `Eigenvector_`. Para encontrar esses erros pode ser utilizada a validação cruzada, usando a metodologia proposta por Piepho (1995) e estudada com mais detalhe sobre simulações em dados reais por Arciniegas-Alarcón et al. (2011). A metodologia é descrita brevemente a seguir.

Suponha um experimento $G \times E$ arranjado em uma tabela com dados omissos. Da tabela de valores observados, é deletado um por vez, imputando todos os dados faltantes e registrando a diferença entre a estimativa e o dado atual para a casela sob consideração. Isto é feito para todas as caselas observadas e depois é calculada a média das diferenças ao quadrado. Denote essa quantidade por D . D contém dois componentes de variação: Um devido à inexatidão preditiva da imputação e o outro devido ao erro amostral dos dados observados. Por essa razão D pode ser corrigida subtraindo uma estimativa do erro da média (s^2). A raiz quadrada de $(D - s^2)$ pode ser tomada como o erro de imputação. O método `Eigenvector` com menor erro de imputação seria o método recomendado.

Por outro lado, se o objetivo posterior à imputação é a inferência a partir da estimação de parâmetros de um modelo estatístico (Gauch et al., 2008; Rodrigues et al., 2009), o critério de escolha do melhor método `Eigenvector` pode ser o erro padrão da

estatística de interesse. O método Eigenvector que produza o menor erro padrão seria o melhor. O tratamento moderno de dados faltantes sugere a imputação múltipla como uma alternativa para encontrar erros padrão (Schafer e Graham, 2002), mas, no caso de imputação determinística pode ser aplicada uma solução já bem conhecida e testada com sucesso. Trata-se do método bootstrap proporcional proposto por Bello (1994), no qual a proporção de dados presentes e omissos que aparecem em cada amostra bootstrap é exatamente igual à proporção que aparece nos dados originais incompletos.

Outro aspecto que pode ser de interesse é o mecanismo de ausência dos dados. Geralmente, em situações que envolvem a avaliação de vários genótipos em diferentes ambientes, as observações ausentes seguem uma das definições propostas 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 contém a ausência. Entretanto, nos programas de teste de genótipos, nos quais as cultivares são escolhidas durante cada ano, usando somente os dados observados sem considerar os dados ausentes, o mecanismo de ausência é claramente aleatório MAR (Piepho e Möhring, 2006).

O último tipo de ausência, MNAR, pode ser visto usualmente quando o mesmo subconjunto de genótipos está ausente em um número de ambientes da mesma sub-região, porque o melhorista de plantas no local não aceita esses 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. Nesta pesquisa foi avaliado o mecanismo MCAR e pesquisa adicional é necessária para estudar os mecanismos restantes.

Finalmente, os métodos propostos neste capítulo têm fácil implementação computacional, mas, uma das principais vantagens é que são livres de pressuposições estruturais ou distribucionais e não dependem do mecanismo de ausência dos dados em experimentos $G \times E$.

Referências

ARCINIEGAS-ALARCÓN, S. **Imputação de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão**. 82p. 2008. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2008.

ARCINIEGAS-ALARCÓN, S.; DIAS, C.T.S. Imputação de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão. **Revista Brasileira de Biometria**, Jaboticabal, v.27, n.1, p.125-138, 2009a.

_____. Análise AMMI com dados imputados em experimentos de interação genótipo $\times$ ambiente de algodão. **Pesquisa Agropecuária Brasileira**, Brasília, v.44, n.11, p.1391-1397, 2009b.

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; DIAS, C.T.S.; KRZANOWSKI, W.J. An alternative methodology for imputing missing data in trials with genotype-by-environment interaction. **Biometrical Letters**, Poznan, v.47,n.1, p.1-14, 2010.

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; DIAS, C.T.S. Imputação de dados em experimentos com interação genótipo $\times$ ambiente. **Interciencia**, Caracas, v.36, n.6, p.444-449, 2011.

BELLO, A.L. Choosing among imputation techniques for incomplete multivariate data: a simulation study. **Communications in Statistics – Theory and Methods**, Oxon, v.22, p.853-877, 1993.

_____. A bootstrap method for using imputation techniques for data with missing values. **Biometrical Journal**, Weinheim, v.36, p.453-464, 1994.

_____. Imputation techniques in regression analysis: Looking closely at their implementation. **Computational Statistics & Data Analysis**, Amsterdam, v.22, p.853-877, 1995.

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

BRO, R.; KJELDAHL, K.; SMILDE, A.K.; KIERS, H.A.L. Cross-validation of component models: a critical look at current methods. **Analytical and Bioanalytical Chemistry**, Berlin, v.390, p.1241-1251, 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**, Poznan, v.24-25, p.7-31, 1992.

_____. Further study on estimating missing values in series of variety trials. **Biuletyn Oceny Odmian**, Poznan, v.30, p.7-38, 1999.

CALINSKI, T.; CZAJKA, S.; KACZMAREK, Z.; KRAJEWSKI, P.; PILARCZYK, W. Analyzing the Genotype-by-Environment Interactions Under a Randomization-Derived Mixed Model. **Journal of Agricultural, Biological and Environmental Statistics**, Berlin, v.14, p.224-241, 2009.

DENIS, J.B. Ajustements de modèles linéaires et bilinéaires sous contraintes linéaires avec données manquantes. **Revue de statistique appliquée**, Paris, v.39, p.5-24, 1991.

DENIS, J.B.; BARIL C.P. Sophisticated models with numerous missing values: the multiplicative interaction model as an example. **Biuletyn 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 Science**, Madison, v.43, p.865-873, 2003.

———. Choosing components in the additive main effect and multiplicative interaction (AMMI) models. **Scientia Agricola**, Piracicaba, v.63, n.2, p.169-175, 2006.

EASTMENT, H.T.; KRZANOWSKI, W.J. Cross-validators choice of the number of components from a principal component analysis. **Technometrics**, Alexandria, v.24, p.73-77, 1982.

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

FLORES, F.; MORENO, M.T.; CUBERO, J.I. A comparison of univariate and multivariate methods to analyze $G \times E$ interaction. **Field Crops Research**, New York, v.56, p.271-286, 1998.

FREEMAN, H.G. Analysis of interactions in incomplete two-ways tables. **Applied Statistics**, Chichester v.24, p.46-55, 1975.

GABRIEL, K.R. Le biplot – outil d'exploration de données multidimensionnelles. **Journal de la Societe Francaise de Statistique**, Paris, v.143, p. 5-55, 2002.

GARCÍA-PEÑA, M.; DIAS, C.T.S. Analysis of bivariate additive models with multiplicative interaction (AMMI). **Biometric Brazilian Journal**, Jaboticabal, v.27, p.586-602, 2009.

GAUCH, H.G. Model selection and validation for yield trials with interaction. **Biometrics**, Washington, v.44, p.705-715, 1988.

———. **Statistical analysis of regional yield trials: AMMI analysis of factorial designs**. Amsterdam: Elsevier, 1992. 278p.

———. Statistical Analysis of Yield Trials by AMMI and GGE. **Crop Science**, Madison, v.46, p.1488-1500, 2006.

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

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

GODFREY, A.J.R.; WOOD, G.R.; GANESALINGAM, S.; NICHOLS, M.A.; QIAO, C.G. Two-stage clustering in genotype-by-environment analyses with missing data. **Journal of Agricultural Science**, Cambridge, v.139, p.67-77, 2002.

GODFREY, A.J.R. **Dealing with sparsity in genotype $\times$ environment analysis**. 356p. Dissertation, Massey University, 2004.

GOOD, I. J. Applications of the singular decomposition of a matrix. **Technometrics**, Alexandria, v.11, p.823-831, 1969.

HASTIE, T.; TIBSHIRANI, R.; SHERLOCK, G.; EISEN, M.; BROWN, P.; BOTSTEIN, D. **Imputing missing data for gene expression arrays**. Technical Report, Division of Biostatistics, Standford University, 1999.

HONGYU, K. **Distribuição empírica dos autovalores associados à matriz de interação dos modelos AMMI pelo método bootstrap não-paramétrico**. 103p. 2012. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2012.

HEDDERLEY, D.; WAKELING, I. A comparison of imputation techniques for internal preference mapping, using Monte Carlo simulation. **Food Quality and Preference**, Paris, v.6, p.281-297, 1995.

JOSSE, J.; PAGÈS, J.; HUSSON, F. Multiple imputation in principal component analysis. **Advances in Data Analysis and Classification**, Berlin, v.5, p.231-246, 2011.

KANG, M.S.; BALZARINI, M.G.; GUERRA, J.L.L. Genotype-by-environment interaction. In: SAXTON, A.M. **Genetic analysis of complex traits using SAS**. SAS Institute Inc., Cary, NC, USA. 2004. chap. 5, p.69-96.

KRZANOWSKI, W.J. Missing value imputation in multivariate data using the singular value decomposition of a matrix. **Biometrical Letters**, Poznan, v.25, n.1-2, p.31-39, 1988.

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

MANDEL, J. The analysis of two-way tables with missing values. **Applied Statistics**, Chichester, v.42, p.85-93, 1993.

PEREIRA, D.G.; MEXIA, J.T.; RODRIGUES, P.C. Robustness of joint regression analysis. **Biometrical Letters**, Poznan, v.44, p.105-128, 2007.

PIEPHO, H.P. Methods for estimating missing genotype-location combinations in multilocation trials - an empirical comparison. **Informatik, Biometrie und Epidemiologie in Medizin und Biologie**, Stuttgart, v.26, n.4, p.335-349, 1995.

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

R DEVELOPMENT CORE TEAM. **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. Disponível em: <http://www.R-project.org/>. Acesso em: 15 dez. 2014.

RAJU, B.M.K. A study on AMMI model and its biplots. **Journal of the Indian Society of Agricultural Statistics**, New Delhi, v.55, p.297-322, 2002.

RAJU, B.M.K.; BHATIA, V.K. Bias in the estimates of sensitivity from incomplete $G \times E$ tables. **Journal of the Indian Society of Agricultural Statistics**, New Delhi, v.56, p.177-189, 2003.

RAJU, B.M.K.; BHATIA, V.K.; KUMAR, V.V.S. Assessment of sensitivity with incomplete data. **Journal of the Indian Society of Agricultural Statistics**, New Delhi, v.60, p.118-125, 2006.

RODRIGUES, P.J.C. **New strategies to detect and understand genotype-by-environment interactions and QTL-by-environment interactions**. 2012, 145p. Tese (Doutorado em Estatística e Gestão do Risco) - Universidade Nova de Lisboa, Lisboa, 2012.

RODRIGUES, P.J.C.; PEREIRA, D.G.; MEXIA, J.T. Structuring Genotype x Environment Interaction: an Overview. **Bulletin of Plant Breeding and Acclimatization Institute**, Blonie, v.250, p.225-236, 2009.

———. A comparison between joint regression analysis and the additive main and multiplicative interaction model: the robustness with increasing amounts of missing data. **Scientia Agricola**, Piracicaba, v.68, p.697-705, 2011.

ROMAGOSA, I.; VOLTAS, J.; MALOSETTI, M.; VAN EEUWIJK, F.A. Interacción Genotipo por Ambiente. In: ÁVILA, C.M.; ATIENZA, S.G.; MORENO, M.T.; CUBERO, J.I. **La adaptación al ambiente y los estreses abióticos en la mejora vegetal**. Sevilla: Instituto de Investigación y Formación Agraria y Pesquera, 2008. cap. 5, p.107-136.

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

SMILDE A.; BRO, R.; GELADI, P. **Multi-way analysis with applications in the chemical sciences**. England: John Wiley & Sons, 2004. 376 p.

SPRENT, P.; SMEETON, N.C. **Applied Nonparametric Statistical Methods**. London: Chapman and Hall, 2001. 463 p.

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.

VAN EEUWIJK, F.A.; MALOSETTI, M.; BOER, M.P. Modelling the genetic basis of response curves underlying genotype environment interaction. In: SPIERTZ, J.H.J.; STRUIK, P.C.; VAN LAAR, H.H. **Scale and Complexity in Plant Systems Research: Gene-Plant-Crop Relations**. New York: Springer. 2007. chap. 10, p.115-126.

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**, Australia, v.56, p.883-894, 2005.

WOLD, S. Cross-Validatory estimation of the number of components in factor and principal components models. **Technometrics**, Alexandria, v.20, p.397-405, 1978.

3 IMPUTAÇÃO MÚLTIPLA LIVRE DE DISTRIBUIÇÃO EM TABELAS INCOMPLETAS DE DUPLA ENTRADA *

Resumo

O objetivo deste capítulo 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.

Palavras-chave: Dados ausentes; Decomposição por valores singulares; Ensaios multi-ambiente; Experimentos desbalanceados; Interação genótipo×ambiente; Melhoramento de plantas

Abstract

The objective of this chapter 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.

Keywords: Missing data; Singular value decomposition; Multi-environment trials; Unbalanced experiments; Genotype×environment interaction; Plant breeding

*Capítulo aceito para publicação na revista Pesquisa Agropecuária Brasileira.

3.1 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 ($G \times E$) (Dias e Krzanowski, 2003).

Embora os experimentos com interação $G \times E$ 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 ou 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, 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 $G \times E$: 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 e Husson (2012a) e van 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 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 e 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 $G \times E$ 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 e Rubin, 2002). O objetivo deste capítulo 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.

3.2 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 capítulo para designar o algoritmo, descrito a seguir.

Considere a matriz $\mathbf{X}$, de dimensão $(n \times p)$ com elementos x_{ij} , em que alguns desses elementos estão ausentes (x_{ij}^{aus}) ($i = 1, \dots, n; j = 1, \dots, p$). Na imputação biplot, os dados faltantes são inicialmente imputados pela média dos valores observados em suas respectivas colunas, o que resulta em uma matriz $\mathbf{X}$ completada. As colunas da matriz $\mathbf{X}$ 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 e 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 $\mathbf{P}$. Em seguida, calcula-se a DVS da matriz $\mathbf{P}$ e os valores p_{ij} , são atualizados com uso de apenas dois CP da DVS, o que resulta em uma nova matriz, denominada $\mathbf{P}^{(2)}$, 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 matriz $\mathbf{X}^{(2)}$ de dimensão $(n \times p)$.

Os elementos ausentes x_{ij}^{aus} na matriz $\mathbf{X}$ original são imputados pelo correspondente valor $\hat{x}_{ij}^{(2)}$ de $\mathbf{X}^{(2)}$. 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, $\frac{d}{\bar{y}} < 0,01$, 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 $\mathbf{X}$ 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 e Dolatabadi (2009). Srivastava e Dolatabadi (2009) propuseram IM com uso dos resíduos simples do modelo de regressão linear clássico $\mathbf{Y} = \mathbf{Q}\beta + \mathbf{E}$, em que o vetor $\mathbf{Y}$ ($n \times 1$) representa a variável dependente; $\mathbf{Q}$ ($n \times p$) é a matriz de delineamento que contém as variáveis independentes; β ($p \times 1$) é o vetor desconhecido de parâmetros de regressão e $\mathbf{E}$ ($n \times 1$) é o vetor de erros aleatórios independentes e identicamente distribuídos.

Assume-se que os dados ausentes somente podem ocorrer no vetor $\mathbf{Y}$ e que todas as observações das variáveis independentes são disponíveis e completas. Portanto, o modelo pode ser reescrito como $\begin{pmatrix} \mathbf{Y}_O \\ \mathbf{Y}_A \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_O \\ \mathbf{Q}_A \end{pmatrix} \beta + \mathbf{E}$, em que $\mathbf{Y}_O$ ($n_1 \times 1$) corresponde ao subvetor dos n_1 dados observados e $\mathbf{Y}_A$ ($n_0 \times 1$) ao subvetor que contém n_0 valores ausentes, ao se levar em conta que $n_0 + n_1 = n$. A matriz $\mathbf{Q}$ é decomposta de forma semelhante. Assim, a imputação múltipla é obtida da seguinte maneira: $\hat{\mathbf{Y}}_{At} = \mathbf{Q}_A (\mathbf{Q}_O^T \mathbf{Q}_O)^{-1} \mathbf{Q}_O^T \mathbf{Y}_O + \mathbf{E}_t$ em que $t = 1, \dots, M$, sendo M o número de imputações para cada dado faltante e $\mathbf{E}_t$ é a t -ésima amostra aleatória com reposição de tamanho n_0 obtida do vetor de resíduos. Esse vetor é calculado pela expressão $\mathbf{e} = \left(\frac{n_1}{n_1 - p} \right)^{0,5} (\mathbf{Y}_O - \mathbf{Q}_O \mathbf{b}_1)$, em que $\mathbf{b}_1$ é a estimativa

de mínimos quadrados de β , baseada unicamente nos dados observados, ou seja, $\mathbf{b}_1 = (\mathbf{Q}_o^T \mathbf{Q}_o)^{-1} \mathbf{Q}_o^T \mathbf{Y}_o$.

A primeira modificação proposta no algoritmo de imputação biplot é a seguinte. O método fornece no final, depois de atingir convergência, uma matriz $\mathbf{X}^{(2)}$, 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, $\hat{\varepsilon} = \mathbf{X} - \mathbf{X}^{(2)}$. Naturalmente, a matriz $\hat{\varepsilon}$ tem dimensão $(n \times p)$ 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 ($n \times p$), em que $t = 1, \dots, M$, 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 $\mathbf{X}$ original pelos valores correspondentes de cada uma das t matrizes definidas por $\mathbf{X}^{(2)} + \Omega_t$.

No presente trabalho, adotou-se $M = 5$, uma vez que esse número permite atingir alta eficiência estatística em muitas aplicações práticas (van 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 capítulo, cinco grupos de avaliação. Assim, o método foi denominado IMBiplotGh, em que h representa o grupo e $h = 1, \dots, 5$. 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 calcular a DVS de $\mathbf{P}$, 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 $G \times E$, 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 $G \times E$ 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 $\mathbf{X}$ ($n \times p$) 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 ($i = 0, 1, \dots, n - 1$; $j = 1, \dots, p$). 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 $G \times E$ - 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 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.

3.3 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 3.1). 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 3.1). 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 porcentagem 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

Tabela 3.1 - 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

variância entre imputações quanto o viés quadrático médio (Figura 3.1). 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 entre imputações (V_E) foi o IMBiplotG2, para todas as percentagens consideradas, seguido por IMBiplotG1 e IMBiplotG3 (Tabela 3.2). Entretanto, a variância foi maximizada com uso de IMBiplotRes, também em todas as situações simuladas. Portanto, tanto nos dados de eucalipto quanto nos de cana-de-açúcar, a 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

Figura 3.1 - Distribuição da medida de acurácia geral (T_{acc}), com uso dos algoritmos IM-BiplotRes (Res) e IMBiplotGh (G1, G2, G3, G4 e G5), para o conjunto de dados de eucalipto (*Eucalyptus grandis*)

seja, IMBiplotG1, IMBiplotG2 e IMBiplotG3. Da mesma forma, nos dados de cana-de-açúcar, também constatou-se que o VQM aumentava com o crescimento da 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 3.2). Assim, o primeiro grupo, de alto desempenho, foi composto por IMBiplotG5, IMBiplotG4 e IMBiplotRes; 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 no conjunto de eucalipto, mas fez parte dos melhores métodos no conjunto de dados de cana-de-açúcar.

Tabela 3.2 - 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

Figura 3.2 - Distribuição da medida de acurácia geral (T_{acc}), com uso dos algoritmos IM-BiplotRes (Res) e IMBiplotGh (G1, G2, G3, G4 e G5), para o conjunto de dados de cana-de-açúcar (*Saccharum officinarum*)

Os algoritmos de IM apresentados neste capítulo têm como base a imputação biplot de Yan (2013), que utiliza uma aproximação de posto 2 para qualquer matriz $G \times E$. 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 são suficientes nem apropriados para determinar genótipos ga-

nhadores ou mega-ambientes. Por esse motivo, sugere-se o seguinte esquema simples de pré-processamento de dados, para uma matriz $G \times E$ incompleta.

Suponha a matriz $G \times E$ $\mathbf{X}$ ($n \times p$) com dados ausentes. Antes de aplicar os métodos propostos aqui, deve-se encontrar o posto de $\mathbf{X}$ 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 e 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 e Husson, 2012b). Esses métodos proporcionam o posto da matriz $\mathbf{X}$ 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 ao autor.

3.4 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 $\times$ ambiente ($G \times E$), 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.

Referências

- ALLISON, P.D. Handling Missing Data by Maximum Likelihood. **SAS Global Forum, Statistics and Data Analysis**. Orlando, n.312, p.1-21, 2012.
- ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; DIAS, C.T.S. Data imputation in trials with genotype $\times$ environment interaction. **Interciencia**, Caracas, v.36, p.444-449, 2011.
- ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; KRZANOWSKI, W.J.; DIAS, C.T.S. Deterministic imputation in multienvironment trials. **ISRN Agronomy**, Cairo, v.2013, p.1-17, 2013.

- 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.S.; KRZANOWSKI, W.J. Distribution-free multiple imputation in an interaction matrix through singular value decomposition. **Scientia Agricola**, Piracicaba, v.65, p.422-427, 2008.
- CECCARELLI, S.; GRANDO, S.; BAUM, M. Participatory plant breeding in water-limited environments. **Experimental Agriculture**, Cambridge, v.43, p.411-435, 2007.
- DIAS, C.T.S.; KRZANOWSKI, W.J. Model selection and cross validation in additive main effect and multiplicative interaction models. **Crop Science**, Madison, v.43, p.865-873, 2003.
- GABRIEL, K.R. The biplot graphic display of matrices with application to principal component analysis. **Biometrika**, Oxford, v.58, p.453-467, 1971.
- . Le biplot - outil d'exploration de données multidimensionnelles. **Journal de la Société Française de Statistique**, Paris, v.143, p.5-55, 2002.
- 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**, Rio de Janeiro, v.29, p.527-536, 2014.
- GAUCH JUNIOR, H. G. A simple protocol for AMMI analysis of yield trials. **Crop Science**, Madison, v.53, p.1860-1869, 2013.
- HAREL, O.; ZHOU, X.H. Multiple imputation: review of theory, implementation, and software. **Statistics in Medicine**, Berlin, v.26, p.3057-3077, 2007.
- HUSSON, F.; JOSSE, J. **missMDA**: Handling missing values with/in multivariate data analysis (principal component methods). R package version 1.7, 2013. Disponível em: <http://CRAN.R-project.org/package=missMDA>. Acesso em: 15 dez. 2014.
- JOSSE, J.; HUSSON, F. Handling missing values in exploratory multivariate data analysis methods. **Journal de la Société Française de Statistique**, Paris, v.153, p.79-99, 2012a.
- . Selecting the number of components in PCA using cross-validation approximations. **Computational Statistics and Data Analysis**, Amsterdam, v.56, p. 1869-1879, 2012b.

- KRZANOWSKI, W.J. Missing value imputation in multivariate data using the singular value decomposition of a matrix. **Biometrical Letters**, Poznan, 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 droughtprone ecosystem. **Field Crops Research**, New York, v.133, p.37–47. 2012.
- LAVORANTI, O.J. **Estabilidade e adaptabilidade fenotípica através da reamostragem “bootstrap” no modelo AMMI**. 2003. 166 p. Tese (Doutorado) - Universidade de São Paulo, Piracicaba.
- LITTLE, R.J.A.; RUBIN, D.B. **Statistical analysis with missing data**, 2nd edn. Wiley, New York, USA, 2002. 408p.
- PERRY, P. **Cross-validation for unsupervised learning**. Dissertation, Stanford University, 2009.
- R Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>. Acesso em: 15 dez. 2014.
- RÄSSLER, S.; RUBIN, D. B.; ZELL, E. R. Imputation. **WIREs Computational Statistics**, Oxford, v.5, p.20–29, 2013.
- 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**, Piracicaba, v.68, p.697-705, 2011.
- RUBIN, D.B. Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. **Proceedings of the Survey Research Methods Section of the American Statistical Association**, Alexandria, p. 20-34, 1978.
- _____. **Multiple imputation for nonresponse in surveys**. New York: John Wiley & Sons, 1987. 258 p.
- SANTOS, E.G.D. **Interação genótipos×locais em cana-de-açúcar e perspectivas de estratificação ambiental**. 2008. Dissertação (Mestrado) - Universidade de São Paulo, Piracicaba.
- SRIVASTAVA, M.S.; DOLATABADI, M. Multiple imputation and other resampling scheme for imputing missing observations. **Journal of Multivariate Analysis**, Amsterdam, v.100, p.1919–1937, 2009.

VAN BUUREN, S. **Flexible imputation of missing data**. Boca Raton: CRC press, 2012. 343p.

WONG, J. **Imputation**. R package version 2.0.1, 2013. Disponível em: <http://CRAN.R-project.org/package=imputation>. Acesso em: 15 dez. 2014.

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

YAN, W.; HOLLAND J.B. A heritability-adjusted GGE biplot for test environment evaluation. **Euphytica**, Amsterdam, v.171, p.355–369, 2010.

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

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

YANG, R.-C.; CROSSA, J.; CORNELIUS, P.L.; BURGUEÑO, J. Biplot analysis of genotype×environment interaction: Proceed with caution. **Crop Science**, Madison, v.49, p.1564–1576, 2009.

ZHANG, P. Multiple imputation: theory and method. **International Statistical Review**, Leidschenveen, v.71, p. 581-592, 2003.

4 IMPUTAÇÃO DE VALORES AUSENTES EM EXPERIMENTOS MULTI-AMBIENTAIS USANDO A DECOMPOSIÇÃO POR VALORES SINGULARES: UMA COMPARAÇÃO EMPÍRICA *

Resumo

Valores faltantes para algumas combinações genótipo-ambiente são comumente encontrados em experimentos multiambientais. A metodologia recomendada para analisar tais dados desbalanceados combina o algoritmo de Esperança-Maximização (EM) com o modelo de efeitos principais aditivos e interação multiplicativa (AMMI). Recentemente, no entanto, quatro algoritmos de imputação baseados na decomposição por valores singulares (DVS) têm sido relatados na literatura (imputação biplot, EM+DVS, imputação GabrielEigen e imputação múltipla livre de distribuição - IMLD). Esses algoritmos preenchem os valores ausentes, removendo a falta de balanceamento nos dados originais e permitem fazer análises padrão mais simples. O objetivo deste capítulo é comparar os quatro algoritmos com o padrão-ouro EM-AMMI. Para fazer isso, foram relatados os resultados de um estudo de simulação baseado em três conjuntos de dados reais completos (eucalipto, cana-de-açúcar e feijão) para várias porcentagens de imputação. As metodologias foram comparadas usando a raiz quadrada normalizada do erro quadrático médio, a estatística de similaridade de Procrustes e o coeficiente de correlação não-paramétrico de Spearman. Concluiu-se que a imputação utilizando o algoritmo EM+DVS proporciona resultados competitivos àqueles obtidos com o padrão-ouro. Também, é uma excelente alternativa à imputação com um modelo aditivo, o qual, na prática, ignora a interação genótipo por ambiente e portanto, pode ser inapropriado em alguns casos.

Palavras-chave: AMMI; Interação genótipo×ambiente; Imputação; Observações ausentes; Decomposição por valores singulares

Abstract

Missing values for some genotype-environment combinations are commonly encountered in multienvironment trials. The recommended methodology for analyzing such unbalanced data combines the Expectation-Maximization (EM) algorithm with the additive main effects and multiplicative interaction (AMMI) model. Recently, however, four imputation algorithms based on the Singular Value Decomposition of a matrix (SVD) have been reported in the literature (Biplot imputation, EM+SVD, GabrielEigen imputation, and distribution free multiple imputation - DFMI). These algorithms all fill in the missing values, thereby removing the lack of balance in the original data and permitting simpler standard analyses to be performed. The aim of this chapter is to compare these four algorithms with the gold standard EM-AMMI. To do this, we report the results of a simulation study based on three complete sets of real data (eucalyptus, sugar cane and beans) for various imputation percentages. The methodologies were compared using the normalised root mean squared error, the Procrustes similarity statistic and the Spearman correlation coefficient. The conclusion is that imputation using the EM algorithm plus SVD provides competitive results to those obtained with the gold standard. It is also an excellent alternative to imputation with an additive model, which in practice ignores the genotype-by-environment interaction and therefore may not be appropriate in some cases.

Keywords: AMMI; Genotype×environment interaction; Imputation; Missing values; Singular value decomposition

*Capítulo aceito para publicação na revista Communications in Biometry and Crop Science.

4.1 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 flutuações significativas do desempenho na produção relativa a outras cultivares. Essas mudanças são influenciadas por diferentes condições ambientais e são referidas como interação genótipo por ambiente ou $G \times E$ (Arciniegas-Alarcón et al., 2013).

Frequentemente, os experimentos multiambientais são desbalanceados porque vários genótipos não são testados em alguns ambientes. Várias metodologias têm sido sugeridas para resolver essa falta de balanceamento produzida pelos dados faltantes; uma lista útil de referências sobre esse assunto está disponível em Arciniegas-Alarcón et al. (2011, 2013). Uma das primeiras propostas foi feita por Freeman (1975), que sugere imputar as observações ausentes de maneira iterativa, minimizando a soma de quadrados do erro e então, fazer a análise da interação $G \times E$, diminuindo os graus de liberdade pelo número de dados faltantes.

Gauch e Zobel (1990) desenvolveram uma metodologia, fazendo a imputação por meio do algoritmo EM e incorporando o modelo de efeitos principais aditivos com interação multiplicativa (AMMI); este método é também conhecido como a aproximação EM-AMMI. Versões alternativas desse procedimento usando análise de agrupamentos foram descritas em Godfrey et al. (2002) e Godfrey (2004). Raju (2002) tratou os ambientes como efeitos aleatórios no algoritmo EM-AMMI e sugeriu uma estatística robusta para os dados omissos nas análises de estabilidade.

Outra opção, é fazer imputação em tabelas incompletas de dupla entrada usando funções lineares das linhas (ou colunas), como foi proposto por Mandel (1993). Outros métodos para manejar observações faltantes em experimentos $G \times E$ que mostraram resultados bons, foram desenvolvidos por Denis (1991), Calinski et al. (1992) e Denis e Baril (1992). Eles encontraram que usando imputações por meio dos mínimos quadrados alternados com modelos de interação bilinear ou estimativas AMMI baseadas em submodelos robustos, podem-se obter resultados tão eficientes quanto os encontrados com o algoritmo EM.

Uma aproximação distinta é trabalhar com dados incompletos sob a estrutura do modelo misto com estimativas baseadas em máxima verossimilhança (Kang et al. 2004), mas, esta aproximação pode envolver múltiplos passos e procedimentos complexos (Yan, 2013). Outros estudos que consideram desbalanceamento nos experimentos multiambientais, são as análises de estabilidade de Raju e Bathia (2003) e Raju et al. (2006, 2009). Finalmente, Pereira et al. (2007) e Rodrigues et al. (2011) avaliaram a robustez da análise de regressão conjunta e dos modelos AMMI sem utilizar imputação de dados.

Recentemente, Bergamo et al. (2008), Perry (2009a), Arciniegas-Alarcón et al. (2010) e Yan (2013) descreveram sistemas de imputação que envolvem a decomposição por valores singulares (DVS) de uma matriz, portanto, são aplicáveis em qualquer experimento multiambiental incompleto. Assim, o objetivo deste capítulo é comparar esses quatro recentes métodos (doravante denotados como imputação biplot, EM+DVS, imputação GabrielEigen e imputação múltipla livre de distribuição - IMLD) utilizando como metodologia padrão, o algoritmo EM-AMMI clássico proposto por Gauch e Zobel (1990).

4.2 Material e métodos

4.2.1 Métodos de imputação

EM-AMMI: Inicialmente, é apresentado o modelo AMMI (Gauch 1988, 1992) para experimentos completos. O modelo usual ANOVA de dupla entrada para analisar dados provenientes de ensaios genótipo por ambiente está dado por

$$y_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}, \quad i = 1, \dots, n; \quad j = 1, \dots, p$$

em que μ , a_i , b_j , $(ab)_{ij}$ e e_{ij} são, respectivamente, a média geral, os efeitos principais genotípicos e ambientais, a interação genótipo por ambiente e o termo de erro associado ao i -ésimo genótipo e ao j -ésimo local. Assume-se que todos os efeitos, exceto o erro, são fixos. O modelo AMMI considera ainda, as interações como uma soma de termos multiplicativos, de modo que o modelo é escrito como

$$y_{ij} = \mu + a_i + b_j + \sum_{l,j} \theta_l \alpha_{il} \beta_{jl} + e_{ij}.$$

Os termos θ_l , α_{il} e β_{jl} ($l = 1, 2, \dots$) podem ser estimados a partir da DVS da matriz de interação: θ_l é estimado pelo l -ésimo valor singular da DVS, enquanto α_{il} e β_{jl} são estimados pelos escores genotípicos e ambientais correspondentes a θ_l . Dependendo do número de termos multiplicativos que hajam sido incluídos, esse modelos são chamados AMMI0, AMMI1, etc.

Em ensaios incompletos, um esquema iterativo é construído em torno ao procedimento acima descrito, para obter imputações AMMI do algoritmo EM. Os parâmetros aditivos são inicialmente definidos pelo cálculo da média geral e das médias genotípicas e ambientais, obtidas a partir dos dados observados. Os resíduos das caselas observadas são iniciados como a média da casela menos a média genotípica menos a média ambiental mais a média geral; as interações para as posições faltantes são inicialmente definidas como zero. Os parâmetros multiplicativos iniciais são obtidos da DVS dessa matriz de resíduos e os valores ausentes são preenchidos pelas estimativas AMMI apropriadas. Em iterações subseqüentes, o procedimento AMMI usual é aplicado à matriz completada e

os valores faltantes são atualizados pelas correspondentes estimativas AMMI. O processo deve se realizar até obter alterações insignificantes em iterações sucessivas.

Dependendo do número de termos multiplicativos utilizado, o método de imputação pode ser nomeado EM-AMMI0, EM-AMMI1, etc. (Gauch e Zobel, 1990). Os estudos de Calinski et al. (1992), Piepho (1995), Arciniegas-Alarcón e Dias (2009) e Paderewski e Rodrigues (2014) mostraram que os melhores resultados para imputação com modelos AMMI se obtêm com a inclusão no máximo, de um componente multiplicativo, razão pela qual esse estudo considerará apenas os métodos EM-AMMI0 e EM-AMMI1.

Vale a pena ressaltar que o modelo de análise nem sempre será o mesmo modelo de imputação. O número de componentes selecionado para uma análise, utilizando o modelo AMMI, sempre dependerá de alguns testes e é comum o uso da estatística RMSPD para isso. Contudo, neste capítulo o AMMI não será avaliado como modelo para análise, mas, será avaliado apenas como modelo de imputação e Arciniegas-Alarcón et al. (2011) encontraram que os erros de imputação associados aos modelos AMMI aumentam conforme o número de componentes multiplicativos também aumenta.

Imputação biplot: Recentemente, Yan (2013) descreveu o seguinte método de imputação, usando o fato que a DVS constitui a base da análise biplot (Gabriel, 1971; 2002). O método basicamente consiste em substituir os valores ausentes, inicialmente, por valores arbitrários para obter uma matriz completada e então, calcular a DVS utilizando apenas dois componentes. Esse método é apresentado a seguir mais formalmente.

Biplot passo 1: Considere a matriz $\mathbf{X}$ de dimensão $(n \times p)$ com elementos x_{ij} ($i = 1, \dots, n; j = 1, \dots, p$), em que as entradas ausentes são denotadas por x_{ij}^{aus} . Inicialmente, os dados faltantes são imputados pela média dos valores observados na respectiva coluna, obtendo assim uma matriz $\mathbf{X}$ completada.

Biplot passo 2: As colunas da matriz $\mathbf{X}$ completada são padronizadas, 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). A matriz padronizada será denotada por $\mathbf{P}$ e os elementos correspondentes por p_{ij} .

Biplot passo 3: É calculada a DVS da matriz $\mathbf{P}$. Considerando os primeiros dois componentes principais (CP), tem-se

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

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's. ε_{ij} é o resíduo para a linha i na coluna j . Removendo esse último termo do erro, os valores p_{ij} podem ser atualizados, obtendo uma nova matriz chamada $\mathbf{P}^{(2)}$ com elementos $p_{ij}^{(2)}$.

Biplot passo 4: Todos os elementos $p_{ij}^{(2)}$ em $\mathbf{P}^{(2)}$ são retornados à sua escala original, $\hat{x}_{ij}^{(2)} = m_j + s_j p_{ij}^{(2)}$, assim obtém-se uma nova matriz $\mathbf{X}^{(2)}$ de dimensão $(n \times p)$. Os elementos ausentes x_{ij}^{aus} na matriz $\mathbf{X}$ original são imputados pelo correspondentes valores $\hat{x}_{ij}^{(2)}$ de $\mathbf{X}^{(2)}$.

Biplot passo 5: O processo é então iterado (voltando ao **Biplot passo 2**) até alcançar estabilidade nas imputações. Por exemplo, as iterações podem continuar até que a diferença entre os valores preditos na iteração atual e na iteração anterior, de todos os dados ausentes, seja menor que algum pequeno valor especificado. Formalmente, isto pode ser expresso como continuar até que $\frac{d}{\bar{y}} < 0,01$, em que

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

Aqui, na é o número total de valores ausentes na matriz $\mathbf{X}$, x_i é o valor predito para o i -ésimo dado faltante na iteração atual, x_i^A é o valor predito para o i -ésimo dado faltante na iteração anterior, 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.

EM+DVS: Perry (2009a) apresenta o seguinte método de imputação que combina o algoritmo EM com a DVS. Esse método substitui os valores ausentes de uma matriz $G \times E$, inicialmente, por valores arbitrários para obter uma matriz completada e uma DVS é então calculada iterativamente sobre essa matriz. No final do processo, quando as iterações alcançarem estabilidade, uma matriz contendo as imputações para as observações ausentes é obtida. A seguir, apresenta-se mais formalmente o método.

Considere a matriz $\mathbf{A}$ de dimensão $(n \times p)$ com elementos A_{ij} ($i = 1, \dots, n$; $j = 1, \dots, p$) e com alguns deles faltantes.

EM+DVS passo 1: Seja $I = \{(i, j) : A_{ij} \text{ não é omissa}\}$, o conjunto de todos os valores observados.

EM+DVS passo 2: Para $1 \leq j \leq p$ seja μ_j a média dos valores não faltantes na coluna j de $\mathbf{A}$; μ_j será 0 se todas as entradas na coluna j são faltantes.

EM+DVS passo 3: Defina $A^{(0)}$ por

$$A_{ij}^{(0)} = \begin{cases} A_{ij} & \text{se } (i, j) \in I \\ \mu_j & \text{em caso contrário} \end{cases}$$

EM+DVS passo 4: Iniciar a contagem das iterações, $N \leftarrow 0$.

EM+DVS passo 5: (Maximização) Calcular a DVS: $\mathbf{A}^{(N)} = \sum_{i=1}^p d_i^{(N)} \mathbf{u}_i^{(N)} \mathbf{v}_i^{(N)T}$;

seja $\mathbf{A}_k^{(N)}$ que denota a DVS truncada com k termos: $\mathbf{A}_k^{(N)} = \sum_{i=1}^k d_i^{(N)} \mathbf{u}_i^{(N)} \mathbf{v}_i^{(N)T}$.

EM+DVS passo 6: (Esperança) Defina a matriz $\mathbf{A}^{(N+1)}$ de dimensão $(n \times p)$ como

$$A_{ij}^{(N+1)} = \begin{cases} A_{ij} & \text{se } (i, j) \in I \\ A_{k,ij}^{(N)} & \text{em caso contrário} \end{cases}$$

EM+DVS passo 7: Seja $RSS^{(N)} = \left\| \mathbf{A} - \mathbf{A}_k^{(N)} \right\|_{F,I}^2$. Se $|RSS^{(N)} - RSS^{(N-1)}|$ é menor que algum valor especificado, declare convergência e obtenha a matriz $\mathbf{A}_k^{(N)}$, que conterá as imputações dos valores ausentes. Caso contrário, aumente $N \leftarrow N + 1$ e volte para **EM+DVS passo 5**.

Imputação GabrielEigen: Arciniegas-Alarcón et al. (2010) propuseram o seguinte método de imputação que mistura a regressão com aproximação de posto inferior utilizando a DVS. Esse método inicialmente substitui as caselas faltantes por valores arbitrários e subsequentemente, as imputações são refinadas por meio de um esquema iterativo que define uma partição diferente da matriz para cada valor ausente e utiliza uma regressão linear das colunas (ou linhas) para obter a nova imputação. Nessa regressão, a matriz de delineamento é aproximada por uma matriz de posto inferior usando a DVS. O algoritmo é apresentado mais formalmente a seguir.

Considere a matriz $\mathbf{X}$ de dimensão $(n \times p)$ com elementos x_{ij} ($i = 1, \dots, n; j = 1, \dots, p$), alguns dos quais são ausentes. Note-se que esse processo requer $n \geq p$, no caso contrário, então a matriz $\mathbf{X}$ deve primeiro ser transposta.

GabrielEigen passo 1: Os valores ausentes são imputados inicialmente pela média da respectiva coluna, obtendo uma matriz $\mathbf{X}$ completada.

GabrielEigen passo 2: As colunas são padronizadas, 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).

GabrielEigen passo 3: Usando a matriz padronizada, define-se a seguinte partição

$$\mathbf{X} = \begin{bmatrix} x_{ij} & \mathbf{x}_{1\bullet}^T \\ \mathbf{x}_{\bullet 1} & \mathbf{X}_{11} \end{bmatrix}$$

em que o valor ausente na posição (i, j) está sempre na posição $(1, 1)$ da partição definida. Para cada valor ausente x_{ij} , os componentes da partição considerada serão diferentes e essa partição é obtida por meio de operações elementares nas linhas e colunas de $\mathbf{X}$. Substitui-se a submatriz $\mathbf{X}_{11}$ pela sua aproximação de posto m utilizando a DVS:

$$\mathbf{X}_{11} = \sum_{k=1}^m \mathbf{u}_k d_k \mathbf{v}_k^T = \mathbf{U} \mathbf{D} \mathbf{V}^T$$

em que $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m]$, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m]$, $\mathbf{D} = \text{diag}(d_1, \dots, d_m)$

e $m \leq \{n - 1, p - 1\}$. A imputação de x_{ij} está dada por $\hat{x}_{ij}^{(m)} = \mathbf{x}_{1\bullet}^T \mathbf{V} \mathbf{D}^{-1} \mathbf{U}^T \mathbf{x}_{\bullet 1}$.

GabrielEigen passo 4: O processo de imputação depende do valor de m e sugere-se

que seja o menor valor para o qual $\frac{\sum_{k=1}^m d_k^2}{\min\{n-1, p-1\}} \approx 0,75$.

GabrielEigen passo 5: Finalmente, os valores imputados devem ser retornados à sua escala original, $x_{ij} = m_j + s_j \hat{x}_{ij}^{(m)}$, substituindo-os na matriz $\mathbf{X}$. Então, o processo passa por iteração (voltando ao **GabrielEigen passo 2**) até alcançar estabilidade nas imputações.

Imputação múltipla livre de distribuição (IMLD): O método de imputação múltipla livre de distribuição (IMLD) proposto por Bergamo et al. (2008) é um esquema iterativo que usa a DVS de uma matriz para prever valores ausentes em uma matriz $\mathbf{Y}$ ($n \times p$). Como com GabrielEigen, esse método requer $n \geq p$ e se $n < p$, a matriz deve primeiro ser transposta.

Considere inicialmente apenas um valor ausente y_{ij} em $\mathbf{Y}$. Então, a i -ésima linha de $\mathbf{Y}$ é eliminada e a DVS da matriz resultante $\mathbf{Y}^{(-i)}$, de dimensão $(n - 1) \times p$, é calculada; em que $\mathbf{Y}^{(-i)} = \overline{\mathbf{U}} \overline{\mathbf{D}} \overline{\mathbf{V}}^T$, $\overline{\mathbf{U}} = (\overline{u}_{sh})$, $\overline{\mathbf{V}} = (\overline{v}_{sh})$, $\overline{\mathbf{D}} = (\overline{d}_1, \dots, \overline{d}_p)$. O seguinte passo consiste em eliminar a j -ésima coluna de $\mathbf{Y}$ e obter a DVS da matriz $\mathbf{Y}_{(-j)}$, de dimensão $n \times (p - 1)$, em que $\mathbf{Y}_{(-j)} = \tilde{\mathbf{U}} \tilde{\mathbf{D}} \tilde{\mathbf{V}}^T$, $\tilde{\mathbf{U}} = (\tilde{u}_{sh})$, $\tilde{\mathbf{V}} = (\tilde{v}_{sh})$, $\tilde{\mathbf{D}} = (\tilde{d}_1, \dots, \tilde{d}_{p-1})$. As matrizes $\overline{\mathbf{U}}$, $\overline{\mathbf{V}}$, $\tilde{\mathbf{U}}$ e $\tilde{\mathbf{V}}$ são ortonormais, $\tilde{\mathbf{D}}$ e $\overline{\mathbf{D}}$ são matrizes diagonais. Combinando as duas DVS, $\mathbf{Y}^{(-i)}$ e $\mathbf{Y}_{(-j)}$, o valor imputado está dado por

$$\hat{y}_{ij} = \sum_{h=1}^H \left(\tilde{u}_{ih} \tilde{d}_h^{(a)} \right) \left(\overline{v}_{jh} \overline{d}_h^{(1-a)} \right) \quad (4.1)$$

em que $H = \min\{n - 1, p - 1\}$ e a é o peso dado a $\mathbf{Y}_{(-j)}$ no intervalo $[0, 1]$.

A especificação de a automaticamente determina o peso para $\mathbf{Y}^{(-i)}$. Por exemplo, um peso de 40% para $\mathbf{Y}_{(-j)}$ requer $a = 0,4$ e o peso para $\mathbf{Y}^{(-i)}$ será de 60% ou $1 - a = 0,6$. Bergamo et al. (2008) afirmam que 5 imputações para cada valor ausente são suficientes para determinar a variabilidade entre imputações e por essa razão sugerem utilizar pesos de 40%, 45%, 50%, 55% e 60% para $\mathbf{Y}_{(-j)}$, isto é, $a = 0,4, 0,45, 0,50, 0,55$ e $0,60$. Cada valor de a fornecerá uma imputação diferente.

Para mais de um valor faltante, um esquema iterativo é necessário. Inicialmente, todos os dados ausentes são substituídos pelas médias das respectivas colunas, obtendo uma matriz $\mathbf{Y}$ completada, então as colunas são padronizadas subtraindo m_j de cada elemento 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 apenas sobre os valores observados. Usando a matriz padronizada, a imputação para cada valor omissa é recalculada utilizando a equação (4.1). Para os cálculos de cada estimativa são necessárias $\mathbf{Y}^{(-i)}$ e $\mathbf{Y}_{(-j)}$, as quais também devem ser padronizadas. Finalmente, a matriz $\mathbf{Y}$ é retornada à sua escala original, $y_{ij} = m_j + s_j \hat{y}_{ij}$. O processo passa por iterações até alcançar estabilidade nas imputações. Em todas as análises deste capítulo a média das cinco imputações foi usada como estimativa de cada valor ausente.

Arciniegas-Alarcón e Dias (2009) mostraram que em alguns casos, a imputação com modelos AMMI0, AMMI1 e AMMI2 podem proporcionar melhores resultados do que a imputação com IMLD. Por essa razão, algumas modificações podem ser realizadas nos componentes da expressão (4.1), levando em conta o trabalho de Calinski et al. (1999) e Bro et al. (2008).

Primeiro, a IMLD está baseada no método de validação cruzada desenvolvido por Eastment e Krzanowski (1982) e as estimativas $\hat{y}_{ij}$ são viesadas porque as matrizes $\tilde{\mathbf{D}}$ e $\overline{\mathbf{D}}$ sistematicamente subestimam $\mathbf{D}$. Em média, esse viés pode ser eliminado corrigindo $\overline{\mathbf{D}}$ por um fator de $\sqrt{n/(n-1)}$ e $\tilde{\mathbf{D}}$ pelo fator $\sqrt{p/(p-1)}$ (Bro et al., 2008).

Segundo, no método IMLD, $H = \min\{n-1, p-1\}$ para utilizar a máxima quantidade de informação disponível na matriz, mas, Calinski et al. (1999) sugerem que a dispersão residual da interação medida pelos autovalores seja próxima de 75%. Levando isso em conta, a escolha de H foi modificada para $H = \min\{w, k\}$, com w tal que $\left(\sum_{h=1}^w \overline{d}_h^2 / \sum_{h=1}^{\min\{n-1, p-1\}} \overline{d}_h^2\right) \approx 0,75$ e k tal que $\left(\sum_{h=1}^k \tilde{d}_h^2 / \sum_{h=1}^{\min\{n-1, p-1\}} \tilde{d}_h^2\right) \approx 0,75$.

4.2.2 Os dados

Para comparar os métodos de imputação foram considerados três conjuntos de dados, publicados em Lavoranti (2003, p.91), Flores et al. (1998) e Santos (2008, p.37). Em cada caso, os dados foram obtidos de um delineamento experimental aleatorizado em blocos com repetições e cada referência oferece os detalhes completos do correspondente planejamento.

Os dados na matriz “Lavoranti” obtiveram-se de experimentos conduzidos em sete ambientes, nas regiões sul e sudeste do Brasil, para 20 progênies de *Eucalyptus grandis* provenientes da Austrália. O delineamento utilizado foi aleatorizado em blocos, com 6 plantas por parcela e 10 repetições, no espaçamento 3.0m por 2.0m. A variável estudada foi altura média da árvore em metros (m). A dimensão da matriz é 20×7 .

O segundo conjunto de dados “Santos” é uma matriz 15×13 , com 15 variedades de cana-de-açúcar avaliadas em 13 ambientes do Brasil. O experimento foi conduzido

no programa de melhoramento genético da RIDESA (Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético), em que a variável estudada foi produção de cana-de-açúcar (t/ha).

O terceiro conjunto de dados “Flores” é uma matriz de dimensão 15×12 , com 15 variedades de feijão avaliadas em 12 ambientes da Espanha. Os experimentos foram conduzidos pela RAEA (Red Andaluza de Experimentación Agraria) e a variável estudada foi produção de feijão (kg/ha).

Na Tabela 4.1, apresentam-se os resultados de um estudo preliminar sobre a escolha do número de componentes multiplicativos, para explicar a interação $G \times E$, do modelo AMMI que pode ser utilizado nos conjuntos de dados selecionados e sobre os quais, um estudo de simulação foi realizado. O método de validação cruzada “leave-one-out” por autovetor (Bro et al., 2008; Gauch, 2013) foi usado para escolher cada modelo e o melhor modelo será aquele que tiver o menor valor da estatística PRESS. Pode-se observar que um modelo apropriado para a matriz de dados “Lavoranti” é o AMMI2, um modelo AMMI1 é apropriado para a matriz “Santos” e um modelo AMMI4 para os dados “Flores”. Esse estudo preliminar justifica a seleção dos conjuntos de dados para avaliar os métodos de imputação.

Tabela 4.1 - Valores da soma de quadrados de predição (PRESS) usando validação cruzada por autovetor na escolha do modelo AMMI para explicar a interação nas matrizes de dados originais (completos)

Modelo		PRESS	
		Flores	Santos
AMMI1	75.1109	113.4852	101.2898
AMMI2	73.8176	125.1817	115.8897
AMMI3	100.3585	119.7403	129.3862
AMMI4	134.3914	108.7088	121.9672
AMMI5	575.0878	141.2513	159.6933
AMMI6	56146.3357	171.0122	146.1308
AMMI7		289.9653	177.4717
AMMI8		718.3669	224.5685
AMMI9		1297.3257	234.6328
AMMI10		5204.1570	396.0781
AMMI11		20406.6026	406.0652
AMMI12			119156.7004

4.2.3 Estudo de simulação

Cada matriz de dados original (“Lavoranti”, “Flores”, “Santos”) foi submetida a retiradas aleatórias em três diferentes porcentagens, isto é, 10%, 20% e 40%. O processo foi

repetido 1000 vezes para cada porcentagem de retirada, obtendo 3000 matrizes diferentes com valores omissos. No total, por conseguinte, foram gerados 9000 conjuntos de dados incompletos e em cada um, os dados faltantes foram imputados com os algoritmos descritos anteriormente por meio de um programa computacional implementado no R (R Core Team, 2013).

O processo de retirada aleatória para uma matriz $\mathbf{X}$ ($n \times p$) 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 foi menor do que r , então o elemento na posição $(i + 1, j)$ da matriz foi deletado ($i = 0, 1, \dots, n - 1; j = 1, \dots, p$). A proporção esperada de dados ausentes na matriz será r (Krzanowski, 1988). Essa técnica foi utilizada com $r = 0, 1; 0, 2$ e $0, 4$.

Para EM+SVD, foi utilizada a implementação computacional fornecida por Perry (2009b) no pacote de R, chamado “*bcv*”. Esse pacote, contém a função *impute.svd* para encontrar as imputações a partir de uma matriz com dados omissos. Na aplicação do algoritmo EM+SVD, devia ser escolhido de antemão para cada matriz incompleta simulada, o número de componentes k a ser utilizado na DVS. Para isso, foi utilizada a função *cv.SVDImpute* do pacote “*imputation*” (Wong, 2013), o qual faz validação cruzada sobre a informação disponível em uma matriz da seguinte forma: Dos dados observados na correspondente matriz, é feita uma retirada aleatória de 30% deles e então, a imputação por EM+SVD é realizada. Usando os dados completados, é calculada a raiz do erro quadrático médio (RMSE) para as entradas que foram aleatoriamente deletadas. Isto é repetido para uma série de valores de k , e o valor de k com menor RMSE é o escolhido para fazer a imputação. Neste estudo, sobre cada matriz incompleta simulada foi repetido o processo de validação cruzada 100 vezes e o k selecionado foi aquele que mais frequentemente minimizasse a RMSE.

4.2.4 Critérios de comparação

Três critérios foram utilizados para comparar os valores verdadeiros com os resultados obtidos nas simulações: a estatística M^2 de Procrustes (Krzanowski, 2000), a raiz normalizada do erro quadrático médio - NRMSE (Ching et al., 2010) e o coeficiente de correlação não-paramétrico de Spearman.

Para o primeiro critério, cada matriz de dados completada (observados+imputados) $\mathbf{Y}_{imp}$ foi comparada com a matriz original $\mathbf{X}_{orig}$, antes de retirar os dados, utilizando

$$M^2 = \text{traço} (\mathbf{X}_{orig}^C \mathbf{X}_{orig}^{C^T} + \mathbf{Y}_{imp}^C \mathbf{Y}_{imp}^{C^T} - 2\mathbf{X}_{orig}^C \mathbf{Q} \mathbf{Y}_{imp}^{C^T})$$

em que $\mathbf{X}_{orig}^C$ e $\mathbf{Y}_{imp}^C$ são matrizes centradas por coluna e $\mathbf{Q} = \mathbf{V}\mathbf{U}^T$ é a matriz de rotação calculada com elementos da DVS da matriz $\mathbf{X}_{orig}^C \mathbf{Y}_{imp}^{C^T} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. A estatística M^2 mede

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

O segundo critério utilizado foi

$$NRMSE = \frac{\sqrt{\text{média}(\mathbf{a}_{imp} - \mathbf{a}_{orig})^2}}{dp(\mathbf{a}_{orig})}$$

em que $\mathbf{a}_{imp}$ e $\mathbf{a}_{orig}$ são vetores contendo os respectivos valores preditos e os valores verdadeiros das observações ausentes simuladas e $dp(\mathbf{a}_{orig})$ é o desvio padrão dos valores contidos no vetor $\mathbf{a}_{orig}$. Quanto menor seja a NRMSE melhor será o método de imputação.

O último critério de comparação foi o coeficiente de correlação de Spearman (Sprent e Smeeton, 2001). Esse coeficiente de correlação não-paramétrico foi calculado entre cada valor ausente e seu correspondente dado verdadeiro. O algoritmo de imputação com a maior correlação fornecerá o melhor desempenho. 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.

4.3 Resultados e discussão

4.3.1 Matriz “Lavoranti”

A Tabela 4.2 apresenta as médias e medianas da NRMSE. Os métodos de imputação que maximizaram o critério foram o Biplot e EM-AMMI1 em todas as porcentagens, portanto, eles são os de mais baixo desempenho. O melhor método de imputação foi EM+DVS, com médias de 0,2690, 0,2649 e 0,2774 para taxas de imputação de 10%, 20% e 40%, respectivamente. Pode-se observar que os algoritmos GabrielEigen, IMLD e EM-AMMI0 obtiveram melhores resultados que o método clássico que considera o modelo AMMI1 para imputação. Então, de acordo com a NRMSE, o método mais eficiente é EM+DVS, seguido por EM-AMMI0, GabrielEigen, IMLD e por último, os métodos EM-AMMI1 e Biplot.

A Figura 4.1 mostra a distribuição da estatística M^2 de Procrustes para diferentes porcentagens de imputação. Quanto menor seja o valor da estatística, melhor será o algoritmo de imputação. Então, de acordo com esse critério, os algoritmos de baixo desempenho são os métodos Biplot e EM-AMMI1 com uma distribuição assimétrica à direita. Os algoritmos IMLD, GabrielEigen, EM+DVS e EM-AMMI0 minimizaram a estatística e têm comportamento comparável nas taxas de 10% e 20% com distribuições aproximadamente simétricas. Por outro lado, com 40% de imputação as dispersões de IMLD e GabrielEigen aumentaram, portanto, os melhores algoritmos nessa porcentagem foram EM+DVS e EM-AMMI0.

Tabela 4.2 - Médias e medianas da NRMSE na matriz “Lavoranti”

Método	Porcentagens de valores retirados aleatoriamente					
	10%		20%		40%	
	Média	Mediana	Média	Mediana	Média	Mediana
Biplot	0,3782	0,3454	0,4197	0,4024	0,5326	0,5217
IMLD	0,2870	0,2719	0,2774	0,2724	0,2881	0,2842
GabrielEigen	0,2780	0,2648	0,2757	0,2705	0,2920	0,2870
EM+DVS	0,2690	0,2552	0,2649	0,2617	0,2774	0,2756
EM-AMMI0	0,2705	0,2591	0,2661	0,2629	0,2776	0,2768
EM-AMMI1	0,4268	0,3993	0,4106	0,3918	0,4239	0,4103

Figura 4.1 - M^2 com diferentes porcentagens de imputação na matriz “Lavoranti”

Adicionalmente, o teste não-paramétrico de Friedman foi utilizado para investigar as diferenças entre os valores da M^2 para os métodos IMLD, GabrielEigen, EM+DVS e EM-AMMI0 em todas as porcentagens. Os testes foram significativos em todos os casos ($p < 0.001$), pelo qual, foi utilizado subsequentemente o teste de Wilcoxon para fazer comparações dois a dois. Uma vez que a metodologia padrão-ouro foi EM-AMMI, apenas foram consideradas as comparações múltiplas que incluíam o método EM-AMMI0. Nas retiradas de 20% e 40% o método EM-AMMI0 teve diferenças significativas ($p < 0.001$) com IMLD, GabrielEigen e EM+DVS. Na taxa de 10%, não se encontraram diferenças significativas entre EM-AMMI0 e EM+DVS. Resumindo, segundo a M^2 , os métodos de imputação em ordem decrescente de eficiência são EM-AMMI0, EM+DVS, GabrielEigen, IMLD, EM-AMMI1 e em último lugar o Biplot.

Finalmente, na Figura 4.2, apresentam-se as distribuições dos coeficientes de correlação que foram calculados em cada conjunto de dados simulado para comparar as imputações com os dados reais. A mediana das correlações é alta ($> 0,85$) para todos os

sistemas de imputação considerados, mas as distribuições com maior dispersão são as correspondentes a Biplot e EM-AMMI1. Conforme aumenta a porcentagem de imputação, as menores dispersões e as maiores medianas são obtidas com os sistemas EM+DVS e EM-AMMI0. De acordo com as distribuições de 10% e 20% de imputação, pode ser utilizado qualquer um dos 4 sistemas IMLD, GabrielEigen, EM+DVS ou EM-AMMI0, mas, se a porcentagem de dados omissos for maior, então EM+DVS e EM-AMMI0 terão os melhores resultados.

Figura 4.2 - Box plot da distribuição de correlação entre dados reais e dados imputados na matriz “Lavoranti” para 10%, 20% e 40% de observações ausentes

4.3.2 Matriz “Flores”

A Tabela 4.3 apresenta as médias e medianas da NRMSE. EM-AMMI1 e Biplot maximizaram o critério em todas as porcentagens de imputação, portanto, como na matriz “Lavoranti”, esses algoritmos foram os de baixo desempenho. O melhor algoritmo de imputação de acordo com esse critério foi o EM+DVS, com médias de 0,3947, 0,3899 e 0,4034 para imputação de 10%, 20% e 40%, respectivamente; GabrielEigen, IMLD e EM-AMMI0 tiveram melhores resultados que o algoritmo clássico EM-AMMI1 em todas as porcentagens de imputação.

Note-se que para 10% e 20% de dados ausentes, GabrielEigen teve melhor desempenho que IMLD, mas quando a porcentagem aumentou para 40% ocorreu o caso contrário. Assim, quando for considerada a NRMSE na matriz “Flores”, três grupos podem ser determinados. O primeiro contém os algoritmos mais eficientes, isto é, EM+DVS e EM-AMMI0. O segundo grupo está composto pelos algoritmos menos eficientes, EM-AMMI1 e Biplot, enquanto o terceiro grupo contém os algoritmos de eficiência intermediária, ou seja, GabrielEigen e IMLD.

Tabela 4.3 - Médias e medianas da NRMSE na matriz “Flores”

Método	Porcentagens de valores retirados aleatoriamente					
	10%		20%		40%	
	Média	Mediana	Média	Mediana	Média	Mediana
Biplot	0,4837	0,4578	0,4781	0,4676	0,5695	0,5603
IMLD	0,4628	0,4367	0,4571	0,4458	0,4624	0,4548
GabrielEigen	0,4349	0,4090	0,4385	0,4251	0,4817	0,4734
EM+DVS	0,3947	0,3839	0,3899	0,3846	0,4034	0,4004
EM-AMMI0	0,3977	0,3810	0,3925	0,3858	0,4041	0,4036
EM-AMMI1	0,9553	0,5812	1,0818	0,6038	1,3877	0,8204

Utilizando a M^2 , o método EM-AMMI1 teve a maior dispersão, por essa razão torna-se o método de mais baixo desempenho e excluído das comparações que envolveram a estatística de Procrustes. A Figura 4.3 apresenta a distribuição da estatística M^2 nas diferentes porcentagens de imputação.

O método que maximizou M^2 em todas as porcentagens consideradas foi o Biplot, enquanto os métodos que minimizaram M^2 em todas as porcentagens foram EM+DVS e EM-AMMI0. Esses dois métodos, têm distribuições aproximadamente simétricas com a menor dispersão entre todos os algoritmos considerados. Novamente, os sistemas IMLD e GabrielEigen tiveram eficiência intermediária.

Para determinar qual é melhor, entre EM+DVS e EM-AMMI0, o teste não-paramétrico de Wilcoxon foi usado. Obteve-se um resultado não significativo ($p = 0,060$) para 10% de imputação, enquanto se encontraram resultados significativos nas outras porcentagens ($p < 0,001$). Nas taxas de 20% e 40%, as menores medianas mostram que EM-AMMI0 têm a maior similaridade entre a matriz original e as diferentes matrizes que continham imputações.

Finalmente, a Figura 4.4 mostra as distribuições dos coeficientes de correlação que foram calculados em cada conjunto de dados simulados, entre as imputações e os correspondentes dados reais. Em todas as porcentagens, a maior mediana e a menor dispersão foram obtidas com EM+DVS e EM-AMMI0, o que significa que de acordo com a correlação de Spearman são os dois melhores sistemas.

Da mesma forma que ocorreu com NRMSE e M^2 , os métodos de baixo desempenho com menores correlações e alta dispersão foram EM-AMMI1 e Biplot, enquanto GabrielEigen e IMLD novamente exibiram eficiência intermediária.

Figura 4.3 - M^2 com diferentes porcentagens de imputação na matriz “Flores”

Figura 4.4 - Box plot da distribuição de correlação entre dados reais e dados imputados na matriz “Flores” para 10%, 20% e 40% de observações ausentes

4.3.3 Matriz “Santos”

Na Tabela 4.4 se apresentam as médias e medianas da NRMSE para a matriz “Santos”. Nesse caso, o melhor sistema de imputação foi claramente EM+DVS porque minimizou a estatística em todas as porcentagens de imputação. Em seguida, encontram-se EM-AMMI0, GabrielEigen e IMLD. Os métodos de baixo desempenho foram Biplot e EM-AMMI1. Igual que nas matrizes “Flores” e “Lavoranti”, todos os métodos resultaram melhores que o sistema baseado no modelo AMMI1.

Tabela 4.4 - Médias e medianas da NRMSE na matriz “Santos”

Método	Porcentagens de valores retirados aleatoriamente					
	10%		20%		40%	
	Média	Mediana	Média	Mediana	Média	Mediana
Biplot	0,5858	0,5736	0,5724	0,5634	0,6065	0,6028
IMLD	0,5296	0,5209	0,5048	0,4982	0,4850	0,4788
GabrielEigen	0,4637	0,4520	0,4678	0,4602	0,4922	0,4843
EM+DVS	0,4147	0,4075	0,4132	0,4096	0,4233	0,4209
EM-AMMI0	0,4207	0,4100	0,4198	0,4184	0,4297	0,4261
EM-AMMI1	0,6731	0,5350	0,7946	0,5690	1,0414	0,7440

O segundo critério avaliado foi M^2 , novamente o algoritmo EM-AMMI1 teve as maiores dispersões e por essa razão, não foi comparado com os outros cinco métodos. As distribuições de M^2 são apresentadas na Figura 4.5. O desempenho mais baixo foi mostrado pelo Biplot, pois maximizou M^2 além de ter uma grande dispersão. Entretanto, os métodos com melhores resultados foram EM+DVS e EM-AMMI0 com distribuições aproximadamente simétricas. GabrielEigen e IMLD, novamente, foram melhores que o Biplot, mas não conseguiram superar EM+DVS e EM-AMMI0.

Como se fez no conjunto de dados “Flores”, um teste de Wilcoxon foi realizado entre EM+DVS e EM-AMMI0, resultando em uma diferença significativa ($p < 0,001$) para as taxas de retirada de 10% e 20%, mas não significativa para a taxa de 40% ($p = 0,074$). Comparando as medianas de M^2 , as menores se obtiveram com EM+DVS em todas as porcentagens.

Na Figura 4.6 se apresentam as distribuições do coeficiente de correlação. Os métodos de melhor desempenho foram EM+DVS e EM-AMMI0, que maximizaram a correlação em todas as porcentagens de imputação e tiveram distribuições aproximadamente simétricas com altas medianas ($> 0,88$) e as menores dispersões. O algoritmo que teve a maior dispersão foi EM-AMMI1, enquanto os sistemas GabrielEigen e IMLD foram melhores que Biplot e EM-AMMI1, mas superados por EM+DVS e EM-AMMI0.

Figura 4.5 - M^2 com diferentes porcentagens de imputação na matriz “Santos”

Figura 4.6 - Box plot da distribuição de correlação entre dados reais e dados imputados na matriz “Santos” para 10%, 20% e 40% de observações ausentes

4.4 Conclusões

Os resultados obtidos aqui, a partir de três matrizes multiambientais, proporcionam alguns guias para pesquisa futura e análises relacionadas com dados omissos nesses ensaios. Os métodos IMLD, EM+DVS e GabrielEigen recentemente apresentados na literatura, mostraram melhores desempenhos do que o algoritmo EM-AMMI1, o qual pertence à família de modelos AMMI proposta por Gauch e Zobel (1990). Enquanto em estudos anteriores, o modelo AMMI1 superou a IMLD (Arciniegas-Alarcón e Dias, 2009), as modificações propostas neste capítulo levaram a que IMLD tivesse melhores resultados do que EM-AMMI1 em todos os casos.

De acordo com a estatística NRMSE, o algoritmo EM+DVS em geral supera os outros métodos que utilizam a DVS, o que é de interesse se o objetivo principal do pesquisador for obter estimativas das combinações $G \times E$ ausentes. Enquanto alguns estudos, como os desenvolvidos por Calinski et al. (1992), Denis e Baril (1992), Piepho (1995) e Arciniegas-Alarcón et al. (2011), mostraram que o modelo aditivo era tanto simples quanto efetivo para resolver o problema de desbalanceamento, os resultados da NRMSE no presente estudo sugerem que EM+DVS sempre supera ao EM-AMMI0, método que está baseado nesse modelo. No entanto, quando se utilizou o critério de Procrustes a situação mudou e a maior similaridade entre a matriz original e as matrizes contendo imputações foi mostrada por EM-AMMI0, seguido pelo EM+DVS.

Em geral, portanto, considerando conjuntamente os três critérios, NRMSE, M^2 e a correlação de Spearman entre as observações originais as correspondentes imputações, os métodos mais eficientes foram EM+DVS e EM-AMMI0 enquanto os menos eficientes foram Biplot e EM-AMMI1. Os métodos GabrielEigen e IMLD se encontram consistentemente no intermédio dos pares citados. Para 10% e 20% de dados ausentes GabrielEigen foi melhor do que IMLD, mas quando a porcentagem aumentar para 40% então é preferível IMLD. Além disso, IMLD exibe uma característica que não apresenta algum dos outros métodos apresentados: é o único sistema que fornece uma estimativa da variância entre imputações que representa a incerteza sobre os valores reais a serem preenchidos.

É importante lembrar que nem sempre o modelo de análise deve ser o mesmo modelo de imputação. Por exemplo, no estudo inicial de validação cruzada (Tabela 4.1) realizado na matriz original “Santos”, o modelo de análise escolhido para explicar a interação foi AMMI1, então pode ser tentador pensar que no caso de dados faltantes o melhor modelo de imputação seria EM-AMMI1. Esse não é o caso, porque as observações ausentes mudam a estrutura da interação e além disso, os erros de imputação associados aos modelos AMMI aumentam conforme aumenta o número de componentes multiplicativos. Assim, os resultados deste estudo sugerem que a melhor aproximação é imputar com o sistema EM+DVS e posteriormente, fazer uma análise da interação usando os modelo AMMI.

Mais pesquisa sobre experimentos $G \times E$ incompletos é necessária e pode envolver, por exemplo, outros critérios de comparação dos algoritmos, mecanismos de ausência dos dados diferentes do completamente aleatório (MCAR) e definidos por Little e Rubin (2002) e Paderewski e Rodrigues (2014) ou considerar outras metodologias padrão-ouro como o modelo misto (Piepho, 1998). Até que tal pesquisa forneça evidência contraditória, o EM+DVS aparece como uma alternativa muito competitiva aos modelos AMMI e especialmente em relação ao modelo aditivo, que tem a desvantagem de ignorar a interação, característica que em algumas situações pode ser inadequada.

Referências

- ARCINIEGAS-ALARCÓN, S.; DIAS, C.T.S. Data imputation in trials with genotype by environment interaction: an application on cotton data. **Biometric Brazilian Journal**, Jaboticabal, v.27, p.125-138, 2009.
- ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; DIAS, C.T.S.; KRZANOWSKI, W.J. An alternative methodology for imputing missing data in trials with genotype-by-environment interaction. **Biometrical Letters**, Poznan, v.47, p.1-14, 2010.
- ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; DIAS, C.T.S. Data imputation in trials with genotype $\times$ environment interaction. **Interciencia**, Caracas, v.36, p.444-449, 2011.
- ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; KRZANOWSKI, W.J.; DIAS, C.T.S. Deterministic imputation in multienvironment trials. **ISRN Agronomy**, Cairo, v.2013, p.1-17, 2013.
- BERGAMO, G.C.; DIAS, C.T.S.; KRZANOWSKI, W.J. Distribution-free multiple imputation in an interaction matrix through singular value decomposition. **Scientia Agricola**, Piracicaba, v.65, p.422-427, 2008.
- BRO, R.; KJELDAHL, K.; SMILDE, A.K.; KIERS H.A.L. Cross-validation of component models: a critical look at current methods. **Analytical and Bioanalytical Chemistry**, Berlin, v.390, p.1241-1251, 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 Ocen Odmian**, Poznan, v.24-25, p.7-31, 1992.
- _____. Further study on estimating missing values in series of variety trials. **Biuletyn Ocen Odmian**, Poznan, v.30, p.7-38, 1999.

- CHING, W.; LI, L.; TSING, N.; TAI, C.; NG, T. A weighted local least squares imputation method for missing value estimation in microarray gene expression data. **International Journal of Data Mining and Bioinformatics**, Olney, v.4, p.331-347, 2010.
- DENIS, J.B. Ajustements de modèles linéaires et bilinéaires sous contraintes linéaires avec données manquantes. **Revue de Statistique Appliquée**, Paris, v.39, p.5-24, 1991.
- DENIS, J.B.; BARIL, C.P. Sophisticated models with numerous missing values: the multiplicative interaction model as an example. **Biuletyn Oceny Odmian**, Poznan, v.24-25, p.33-45, 1992.
- EASTMENT, H.T.; KRZANOWSKI, W.J. Cross-validatory choice of the number of components from a principal component analysis. **Technometrics**, Alexandria, v.24, p.73-77, 1982.
- FLORES, F.; MORENO, M.T.; CUBERO, J.I. A comparison of univariate and multivariate methods to analyze G×E interaction. **Field Crops Research**, New York, v.56, p.271-286, 1998.
- FREEMAN, G.H. Analysis of interactions in incomplete two-ways tables. **Applied Statistics**, Chichester, v.24, p.46-55, 1975.
- GABRIEL, K.R. The biplot graphic display of matrices with application to principal component analysis. **Biometrika**, Oxford, v.58, p.453-467, 1971.
- . Le biplot - outil d'exploration de données multidimensionnelles. **Journal de la Société Française de Statistique**, Paris, v.143, p.5–55, 2002.
- GAUCH, H.G. Model selection and validation for yield trials with interaction. **Biometrics**, Oxford, v.44, p.705-715, 1988.
- . **Statistical analysis of regional yield trials: AMMI analysis of factorial designs**. Elsevier, Amsterdam, The Netherlands, 1992. 278p.
- GAUCH, H.G.; ZOBEL, R.W. Imputing missing yield trial data. **Theoretical and Applied Genetics**, New York, v.79, p.753-761, 1990.
- GAUCH, H.G. A simple protocol for AMMI analysis of yield trials. **Crop Science**, Madison, v.53, p.1860-1869, 2013.
- GODFREY, A.J.R.; WOOD, G.R.; GANESALINGAM, S.; NICHOLS, M.A.; QIAO, C.G. Two-stage clustering in genotype-by-environment analyses with missing data. **Journal of Agricultural Science**, Cambridge, v.139, p.67-77, 2002.

GODFREY, A.J.R. **Dealing with sparsity in genotype×environment analysis**. Dissertation, Massey University, 2004.

KANG, M.S.; BALZARINI, M.G.; GUERRA, J.L.L. Genotype-by-environment interaction. In: SAXTON AM. **Genetic analysis of complex traits using SAS**. SAS Institute Inc., Cary, NC, USA. 2004. p.69–96.

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

———. **Principles of multivariate analysis: A user's perspective**. Oxford: University Press, Oxford, England, 2000. 608p.

LAVORANTI, O.J. **Phenotypic stability and adaptability via AMMI model with bootstrap re-sampling**. Dissertation. University of São Paulo, 2003.

LITTLE, R.J.A.; RUBIN, D.B. **Statistical analysis with missing data**, 2nd edn. Wiley, New York, USA, 2002. 408p.

MANDEL, J. The analysis of two-way tables with missing values. **Applied Statistics**, Chichester, v.42, p.85-93, 1993.

PADEREWSKI, J.; RODRIGUES, P.C. The usefulness of EM-AMMI to study the influence of missing data pattern and application to Polish post-registration winter wheat data. **Australian Journal of Crop Science**, Australia, v.8, p.640–645, 2014.

PEREIRA, D.G.; MEXIA, J.T.; RODRIGUES, P.C. Robustness of joint regression analysis. **Biometrical Letters**, Poznan, v.44, p.105-128, 2007.

PERRY, P.O. **Cross-validation for unsupervised learning**. Dissertation, Stanford University. 2009a.

———. **bcv: Cross-Validation for the SVD (Bi-Cross-Validation)**. R package version 1.0. 2009b. Disponível em: <http://ptrckpry.com/code/>. Acesso em: 15 dez. 2014.

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

———. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. **Theoretical and Applied Genetics**, New York, v.97, p.195-201, 1998.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>. Acesso em: 15 dez. 2014.

RAJU, B.M.K. A study on AMMI model and its biplots. **Journal of the Indian Society of Agricultural Statistics**, New Delhi, v.55, p.297-322, 2002.

RAJU, B.M.K.; BHATIA, V.K. Bias in the estimates of sensitivity from incomplete $G \times E$ tables. **Journal of the Indian Society of Agricultural Statistics**, New Delhi, v.56, p.177-189, 2003.

RAJU, B.M.K.; BHATIA, V.K.; KUMAR, V.V.S. Assessment of sensitivity with incomplete data. **Journal of the Indian Society of Agricultural Statistics**, New Delhi, v.60, p.118-125, 2006.

RAJU, B.M.K.; BHATIA, V.K.; BHAR, L.M. Assessing stability of crop varieties with incomplete data. **Journal of the Indian Society of Agricultural Statistics**, New Delhi, v.63, p.139-149, 2009.

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**, Piracicaba, v.68, p.697-705, 2011.

SANTOS, E.G.D. **Genotypes by locations interaction in sugarcane and perspectives of environmental stratification**. 64p. 2008. Dissertação (Mestrado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, 2008.

SPRENT, P.; SMEETON, N.C. **Applied Nonparametric Statistical Methods**. Chapman and Hall, London, 2001. 544p.

WONG, J. **Imputation**: imputation. R package version 2.0.1, 2013. Disponível em: <http://CRAN.R-project.org/package=imputation>. Acesso em: 15 dez. 2014.

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

5 UMA METODOLOGIA ALTERNATIVA PARA IMPUTAR DADOS AUSENTES EM ENSAIOS COM INTERAÇÃO GENÓTIPO POR AMBIENTE: ALGUNS ASPECTOS NOVOS *

Resumo

Um problema comum em ensaios multiambientais surge quando algumas combinações genótipo por ambiente são faltantes. Em Arciniegas-Alarcón et al. (2010) foi descrito um método de imputação para estimar os valores omissos; o algoritmo computacional consiste em uma mistura de regressão com aproximação de posto inferior de uma matriz, baseada na decomposição por valores singulares (DVS). No presente capítulo, fornecem-se duas extensões a essa metodologia por meio da inclusão de pesos escolhidos por validação cruzada, permitindo imputação tanto simples quanto múltipla. Os três métodos foram avaliados e comparados em um estudo de simulação, utilizando um conjunto de dados reais completo, em que valores foram retirados aleatoriamente considerando diferentes taxas. A qualidade das imputações foi avaliada usando três medidas: a estatística de Procrustes, a correlação quadrada entre matrizes e a raiz quadrada normalizada do erro quadrático médio entre as estimativas e os valores observados. Os métodos não fazem nenhuma pressuposição estrutural nem distribucional e podem ser utilizados para qualquer padrão ou mecanismo de ausência dos dados.

Palavras-chave: Validação cruzada; Decomposição por valores singulares; Imputação; Interação genótipo por ambiente; Pesos; Dados faltantes

Abstract

A common problem in multi-environment trials arises when some genotype-by-environment combinations are missing. In Arciniegas-Alarcón et al. (2010) we outlined a method of data imputation to estimate the missing values, the computational algorithm for which was a mixture of regression and lower-rank approximation of a matrix based on its singular value decomposition (SVD). In the present chapter we provide two extensions to this methodology, by including weights chosen by cross-validation and allowing multiple as well as simple imputation. The three methods are assessed and compared in a simulation study, using a complete set of real data in which values are deleted randomly at different rates. The quality of the imputations is evaluated using three measures: the Procrustes statistic, the squared correlation between matrices and the normalised root mean squared error between these estimates and the true observed values. None of the methods makes any distributional or structural assumptions, and all of them can be used for any pattern or mechanism of the missing values.

Keywords: Cross-validation; Singular value decomposition; Imputation; Genotype-by-environment interaction; Weights; Missing values

*Capítulo aceito para publicação na revista Biometrical Letters

5.1 Introdução

No melhoramento genético de plantas, os ensaios multiambientais são importantes para testar a adaptação geral e específica das cultivares. Uma cultivar desenvolvendo-se em diferentes ambientes mostrará uma flutuação significativa do desempenho na produção relativa a outras cultivares. Essas mudanças são influenciadas por diferentes condições ambientais e são referidas como interação genótipo por ambiente ou $G \times E$ (Dias e Krzanowski, 2003; Gauch, 2013).

Usualmente, os ensaios multiambientais dão origem a conjuntos de dados incompletos (Rodrigues et al., 2011; Bergamo et al., 2008). Possíveis formas de analisar tais ensaios são: (i) extrair um subconjunto de dados balanceado eliminando aqueles genótipos ou ambientes que contêm valores faltantes (Yan et al., 2011); (ii) preencher as parcelas vazias com médias ambientais; ou (iii) preencher as parcelas vazias com estimativas obtidas a partir de modelos multiplicativos ou modelos lineares mistos (Kumar et al., 2012; Arciniegas-Alarcón et al., 2011). Essas estratégias podem contornar o desbalanceamento nos dados, mas nenhuma delas é simples e efetiva (Yan, 2013). A primeira estratégia não utiliza toda a informação disponível, a segunda pode ter problemas quando houverem muitos dados faltantes e a terceira, envolve múltiplos passos e procedimentos complexos (Yan, 2013).

Segundo Little e Rubin (2002) e Di Ciaccio (2011) podem se distinguir três mecanismos de ausência em ensaios $G \times E$. Os dados têm "ausência aleatória" (MAR) se o mecanismo que causa as omissões é independente dos valores não observados. Se as omissões também são independentes dos valores observados, então os dados têm "ausência completamente aleatória" (MCAR). Finalmente, se o mecanismo que causa as omissões depende dos valores não observados, os dados têm "ausência não aleatória" (MNAR).

Piepho e Möhring (2006) observam que em um programa de teste de genótipos, em que a seleção de cultivares de cada ano está baseada apenas nos dados coletados sem considerar os não observados, o mecanismo de ausência dos dados é claramente MAR. Por outro lado, Rodrigues et al. (2011) afirmam que MCAR ocorre quando as plantas forem destruídas por animais, inundações ou durante a colheita e também, quando as medições de rendimento forem erroneamente realizadas e inadequadamente inseridas nas bases de dados.

O terceiro mecanismo, MNAR, é considerado por alguns pesquisadores como o mais comum, porque nos ensaios pode ser encontrado um claro padrão na forma dos valores ausentes (Paderewski e Rodrigues, 2014). Conforme Piepho (1995), MNAR ocorre quando o mesmo subconjunto de genótipos está ausente em um número de ambientes da mesma subregião, porque o melhorista não aceita esses 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.

Um novo método de imputação livre de distribuição para ensaios $G \times E$ foi recentemente proposto por Arciniegas-Alarcón et al. (2010). O sistema foi chamado GabrielEigen e utiliza uma mistura de regressão e aproximação de posto inferior de uma matriz. No presente capítulo, descreve-se uma modificação desse método que usa pesos escolhidos por validação cruzada e então, fornece-se uma extensão ao caso de imputação múltipla.

5.2 Material e métodos

5.2.1 Imputação de dados usando GabrielEigen

Suponha que a matriz $\mathbf{X}$ ($n \times p$) contém elementos x_{ij} ($i = 1, \dots, n; j = 1, \dots, p$), alguns deles ausentes. As linhas representam os genótipos e as colunas os ambientes.

Passo 1. Inicialmente são inseridas as médias por coluna nas respectivas parcelas vazias, obtendo dessa forma uma matriz $\mathbf{X}$ completada.

Passo 2. Padronizar as colunas de $\mathbf{X}$ subtraindo m_j de cada elemento e dividir o resultado por s_j (em que m_j e s_j são respectivamente a média e o desvio padrão da j -ésima coluna).

Passo 3. Utilizando a matriz padronizada, substitui-se cada entrada ausente x_{ij} por $\hat{x}_{ij}^{(m)} = \mathbf{x}_{1\bullet}^T \mathbf{V} \mathbf{D}^+ \mathbf{U}^T \mathbf{x}_{\bullet 1}$. Aqui, os vetores $\mathbf{x}_{1\bullet}^T$, $\mathbf{x}_{\bullet 1}$ e as matrizes $\mathbf{V}$, $\mathbf{D}$ e $\mathbf{U}$ são obtidos da partição $\mathbf{X} = \begin{bmatrix} x_{ij} & \mathbf{x}_{1\bullet}^T \\ \mathbf{x}_{\bullet 1} & \mathbf{X}_{11} \end{bmatrix}$ com $\mathbf{X}_{11} = \sum_{k=1}^m \mathbf{u}_k d_k \mathbf{v}_k^T = \mathbf{U} \mathbf{D} \mathbf{V}^T$, em que $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m]$, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m]$, $\mathbf{D} = \text{diag}(d_1, \dots, d_m)$ e $m \leq \{n-1, p-1\}$ é escolhido como o menor valor que satisfaça $(\sum_{k=1}^m d_k^2) / \left(\sum_{k=1}^{\min\{n-1, p-1\}} d_k^2 \right) \approx 0,75$. Assim, $\mathbf{U} \mathbf{D} \mathbf{V}^T$ é a decomposição por valores singulares (DVS) de $\mathbf{X}_{11}$ e $\mathbf{D}^+$ é a inversa generalizada Moore-Penrose de $\mathbf{D}$. Também, para cada observação ausente, os componentes da partição considerada serão diferentes e essa partição é obtida por meio de operações elementares nas linhas e colunas de $\mathbf{X}$.

Passo 4. Finalmente, os valores imputados são retornados à sua escala original, $x_{ij} = m_j + s_j \hat{x}_{ij}^{(m)}$, substituindo-os na matriz $\mathbf{X}$.

Os passos 2 a 4 passam por iteração até alcançar estabilidade nas imputações. Esse processo assume que $n > p$, no caso contrário a matriz deve ser transposta antes de conduzir as iterações.

5.2.2 Modificações propostas

A primeira proposta consiste em incluir pesos no passo 3, substituindo $\hat{x}_{ij}^{(m)}$ por $\hat{x}_{ij}^{(w,m)} = w (\mathbf{x}_{1\bullet}^T \mathbf{V} \mathbf{D}^+ \mathbf{U}^T \mathbf{x}_{\bullet 1})$. Aqui, o peso w é obtido por validação cruzada, utilizando como critério preditivo a raiz quadrada da diferença preditiva média (RMSPD) dos dados observados na matriz incompleta (Gauch e Zobel, 1990; Sabaghnia et al., 2012). O algoritmo modificado, com o novo passo 3, mas mantendo iguais os outros passos de GabrielEigen, será chamado WGabriel.

Para ver como o peso w é escolhido, considere um ensaio $G \times E$ arranjado em uma tabela com valores ausentes. Dos valores observados e para qualquer valor específico de w , é eliminado um por vez, imputando o valor eliminado com WGabriel e registrando a diferença entre a estimativa e o dado atual para a parcela sob consideração. Isto é feito para todas as parcelas observadas e depois, é calculada a média das diferenças ao quadrado. Denote essa quantidade por D . A raiz quadrada de D é a RMSPD baseada nos valores observados, chamada RMSPD(obs) para o valor de w .

Como as colunas da matriz $\mathbf{X}$ são padronizadas no passo 3 de WGabriel, faz sentido permitir que o valor de w seja positivo, negativo ou zero. w negativo representa uma mudança na magnitude e direção da imputação, quer dizer uma mudança de um valor positivo para negativo, ou vice-versa, na escala padronizada com o objetivo de minimizar a RMSPD(obs). No entanto, w deve ser escolhido de tal forma que os valores imputados retornados à escala original $x_{ij} = m_j + s_j \hat{x}_{ij}^{(w,m)}$, não fiquem fora do intervalo de valores existentes. Para atingir isso, sugere-se que os valores de w no intervalo $[-2, 2]$ sejam testados com passos de 0,005 ou 0,01, isto é, testar no total 801 ou 401 pesos respectivamente. Intervalos maiores poderiam ser considerados, mas, aumentará o risco de falhas de convergência do algoritmo. A RMSPD(obs) é obtida para cada um dos pesos e o w que forneça o valor mínimo da estatística será o escolhido para a análise. Esse peso será denotado como w^* .

Josse e Husson (2012) e van Buuren (2012) advertem que os sistemas de imputação simples, como o WGabriel, não levam em conta a incerteza produzida pelas imputações e se posteriormente fossem estimados parâmetros a partir dos dados completados que incluem imputações, o erro padrão será subestimado. É bem conhecido que esse problema se resolve com a utilização da imputação múltipla (IM) (Rubin, 1978; Josse et al., 2011).

A IM envolve três passos distintos (Bergamo et al., 2008): **1. Imputação:** Os dados faltantes são estimados M vezes, gerando M conjuntos de dados completados; **2. Análise:** Os M conjuntos de dados completados são analisados, usando procedimentos estatísticos apropriados para o problema estudado; **3. Combinação:** Os M conjuntos de resultados são combinados em uma única inferência. Em várias aplicações práticas se encontrou uma alta eficiência estatística usando $M = 20$ (Schafer e Graham, 2002).

O método WGabriel pode ser estendido da seguinte forma para obter imputação múltipla livre de distribuição. w^* é o peso que fornece a melhor diferença preditiva ou a menor RMSPD(obs), utilizando toda a informação disponível em uma matriz incompleta. Para produzir 20 ou mais conjuntos diferentes de dados completados é possível usar 20 ou mais pesos distintos em WGabriel. Esses pesos devem ser próximos de w^* , pois no caso contrário a RMSPD pode estar afastada do mínimo para alguns deles. Assim, por exemplo, se $w^* = 0,7$ e o passo escolhido anteriormente foi 0,01, sugere-se um intervalo para w entre 0,6 e 0,8, obtendo no total 21 conjuntos de dados completados. Esse método de imputação será chamado IMWG(0,01).

A seguir, são comparados os três métodos de imputação: GabrielEigen, WGabriel e IMWG(0,01). Como IMWG(0,01) produz múltiplos conjuntos completados, a média dos valores imputados será usada (Kroonenberg, 2008), de modo que seja factível compará-lo com os métodos de imputação determinística simples.

5.2.3 Os dados

Consideraram-se dois conjuntos de dados, a matriz “Denis-Baril” e a matriz “Calinski”. A primeira corresponde a um ensaio $G \times E$ com 26 genótipos de trigo avaliados em 5 ambientes franceses, que foi submetida a uma retirada arbitrária de 37% das entradas da matriz, isto é, 48 valores omissos (Denis e Baril, 1992). Esse conjunto de dados, disponibilizado no programa estatístico de acesso livre R (Wright, 2012), foi usado apenas ilustrativamente para mostrar os valores da RMSPD para dados observados e para dados faltantes com diferentes pesos em WGabriel.

O segundo conjunto de dados é balanceado e foi utilizado para comparar os algoritmos de imputação. É uma matriz de dimensão (18×9) , para 18 variedades de ervilha avaliadas em 9 diferentes locais da Polônia. O experimento foi conduzido pelo Research Centre for Cultivar Testing, Slupia Wielka, e a variável de interesse foi rendimento médio em t/ha (Calinski et al., 2009).

5.2.4 Estudo de simulação

A matriz “Calinski” foi submetida a retiradas aleatórias de valores em diferentes porcentagens, isto é, 10%, 20% e 30%. O processo se repetiu 100 vezes para cada porcentagem, obtendo no total, 300 conjuntos de dados incompletos e em cada um deles os dados omissos foram imputados com os três algoritmos descritos anteriormente por meio de um código no R (R Development Core Team, 2013).

O processo de retirada aleatória para uma matriz $\mathbf{X}$ ($n \times p$) foi conduzido da seguinte forma. 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 foi menor do que r , então o elemento na posição $(i + 1, j)$ da matriz foi deletado ($i = 0, 1, \dots, n - 1; j = 1, \dots, p$). A proporção esperada de dados ausentes na matriz será r (Krzanowski, 1988). Essa técnica foi utilizada com $r = 0, 1; 0, 2$ e $0, 3$.

5.2.5 Critérios de comparação

Três critérios se usaram para comparar os dados atuais com os resultados da simulação: a estatística M^2 de Procrustes (Krzanowski, 2000); a correlação quadrada entre matrizes, $corr^2$ (Gabriel, 2002); e a raiz quadrada normalizada do erro quadrático médio, NRMSE (Ching et al., 2010). A seguir se apresentam os detalhes computacionais de cada um dos critérios.

Primeiro, cada matriz de dados completada contendo valores observados+imputados, $\mathbf{Y}_{imp}$, foi comparada com a matriz original $\mathbf{X}_{orig}$ usando

$$M^2 = \text{traço} (\mathbf{X}_{orig} \mathbf{X}_{orig}^T + \mathbf{Y}_{imp} \mathbf{Y}_{imp}^T - 2\mathbf{X}_{orig} \mathbf{Q} \mathbf{Y}_{imp}^T)$$

em que $\mathbf{Q} = \mathbf{V} \mathbf{U}^T$ é a matriz de rotação calculada com elementos da DVS da matriz $\mathbf{X}_{orig}^T \mathbf{Y}_{imp} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$. A estatística M^2 mede a diferença entre duas configurações de pontos, assim, o método de imputação que minimize essa diferença indicará que o método produz a correspondência mais próxima entre os dados originais e a respectiva matriz com imputações.

Semelhantemente, M^2 foi utilizada para comparar as matrizes de interação $\mathbf{GE}_{orig}$ e $\mathbf{GE}_{imp}$, em que $\mathbf{GE}_{orig}$ e $\mathbf{GE}_{imp}$ são as matrizes de resíduos depois de ajustar os efeitos principais por ANOVA das matrizes $\mathbf{X}_{orig}$ e $\mathbf{Y}_{imp}$ respectivamente (García-Peña e Dias, 2009). Entretanto, as matrizes $\mathbf{GE}_{orig}$ e $\mathbf{GE}_{imp}$ também foram comparadas usando o coeficiente

$$corr^2 (\mathbf{B}, \hat{\mathbf{B}}) = \frac{\text{traço}^2 \{ \mathbf{B}^T \hat{\mathbf{B}} \}}{\text{traço} \{ \mathbf{B}^T \mathbf{B} \} \text{traço} \{ \hat{\mathbf{B}}^T \hat{\mathbf{B}} \}}$$

em que $\mathbf{B}$ e $\hat{\mathbf{B}}$ representam respectivamente, as matrizes $\mathbf{GE}_{orig}$ e $\mathbf{GE}_{imp}$ centradas por coluna. O melhor algoritmo de imputação com esse critério será aquele com maior $corr^2$.

O terceiro critério utilizado foi a NRMSE definida como:

$$NRMSE = \frac{\sqrt{\text{média} (\mathbf{a}_{imp} - \mathbf{a}_{orig})^2}}{dp (\mathbf{a}_{orig})}$$

em que $\mathbf{a}_{imp}$ e $\mathbf{a}_{orig}$ são vetores contendo os respectivos valores preditos e os valores verdadeiros das observações ausentes simuladas. $dp (\mathbf{a}_{orig})$ é o desvio padrão dos valores

contidos no vetor $\mathbf{a}_{orig}$. O melhor método de imputação com esse critério será aquele com a menor NRMSE.

5.3 Resultados e discussão

5.3.1 Matriz “Denis-Baril”

A matriz completa foi submetida a uma retirada arbitrária de 48 valores, a partir dos quais se obteve a RMSPD tanto para dados observados quanto para dados omissos. A Figura 5.1 mostra as duas curvas da RMSPD para diferentes pesos. Consideraram-se todos os pesos no intervalo $[-2, 2]$, mas, apenas os pesos no intervalo $[0, 5; 1, 2]$ são mostrados aqui porque nessa região as curvas foram minimizadas.

Figura 5.1 - Raiz quadrada da diferença preditiva média (RMSPD) na matriz “Denis-Baril”

A distância entre as duas curvas pode ser interpretada como a discrepância entre a imputação utilizando a RMSPD(obs) e a RMSPD real dos dados ausentes - RMSPD(aus). Existe também, um ponto de interseção das curvas, mas nesse ponto não se minimizou a RMSPD(aus) nem a RMSPD(obs); RMSPD(aus) e RMSPD(obs) foram minimizadas em $w = 1,08$ e $w = 0,935$ respectivamente. Esses resultados levam a uma conclusão importante e é que as curvas se minimizaram em $w \neq 1$, portanto, o método WGabriel é apropriado.

5.3.2 Matriz “Calinski”

Na matriz anterior foi possível calcular a RMSPD(aus) para a comparação, mas em aplicações práticas isso é impossível. Por essa razão, o estudo de simulação na matriz “Calinski” leva em conta unicamente a RMSPD(obs) como critério de seleção dos pesos. A Figura 5.2 apresenta as distribuições de M^2 quando cada matriz completada, com valores

observados+imputados, comparou-se com a matriz original. Lembre-se que o melhor método de imputação será aquele que minimize M^2 . Para 10% de retirada aleatória os três métodos têm resultados semelhantes, mas quando a porcentagem de imputação aumentar, M^2 para GabrielEigen também aumenta. Portanto, IMWG(0,01) e WGabriel são os melhores métodos.

Figura 5.2 - Distribuição da estatística de Procrustes (M^2) entre as matrizes com valores imputados e a matriz “Calinski”

A Figura 5.3 mostra os valores de M^2 quando comparadas as matrizes de resíduos $\mathbf{GE}_{imp}$, depois de fazer ANOVA nas matrizes completadas, com a matriz original $\mathbf{GE}_{orig}$. Um comportamento muito semelhante ao da Figura 5.2 é evidente: quando aumentar a porcentagem das retiradas aleatórias, os melhores métodos de imputação são IMWG(0,01) e WGabriel. Uma distribuição aproximadamente simétrica se observou nas porcentagens de 10% e 20% para todos os métodos. Com retiradas de 30%, a simetria continua nas distribuições de M^2 para WGabriel e IMWG(0,01), mas, no caso de GabrielEigen se tem uma distribuição assimétrica à direita.

As distribuições da correlação (quadrada) entre as diferentes matrizes $\mathbf{GE}_{imp}$ e $\mathbf{GE}_{orig}$ são apresentadas na Figura 5.4. Nas retiradas aleatórias de 10%, correlações altas com uma mediana aproximada de 0,87 foram obtidas para os três algoritmos de imputação. Quando a porcentagem de retirada aumentou, a correlação diminuiu mais rápido para GabrielEigen do que para os outros sistemas. Na porcentagem de 20%, a mediana da $corr^2$ para GabrielEigen foi 0,7174, enquanto para WGabriel foi 0,7632 e para IMWG(0,01) foi 0,7649.

Em 30%, a mediana da $corr^2$ do algoritmo GabrielEigen diminuiu para 0,5528, enquanto para WGabriel e IMWG(0,01) a mediana teve valores de 0,6292 e 0,6297 respectivamente. Em geral, todos os métodos apresentaram correlações positivas moderadas

Figura 5.3 - Distribuição da estatística de Procrustes (M^2) entre as matrizes $\mathbf{GE}_{imp}$ e $\mathbf{GE}_{orig}$

Figura 5.4 - Distribuição da correlação quadrada ($corr^2$) entre as matrizes $\mathbf{GE}_{imp}$ e $\mathbf{GE}_{orig}$

e altas. Com retiradas de 10%, os métodos mostraram distribuições simétricas, mas com 20% WGabriel e IMWG(0,01) tiveram distribuições assimétricas à esquerda. De acordo com a $corr^2$, portanto, o melhor método foi IMWG(0,01), seguido pelo WGabriel e por último, GabrielEigen.

O terceiro critério de comparação foi a NRMSE e as médias e medianas da estatística, para os três métodos de imputação, são mostradas na Tabela 5.1. O melhor método será aquele que minimize a estatística. Em todas as porcentagens de imputação, os dois métodos propostos foram melhores do que GabrielEigen. Na porcentagem de 10%, WGabriel foi o melhor método, seguido pelo IMWG(0,01). Para 20% de imputação, a ordem foi a mesma, mas com médias muito próximas. Com 30% de imputação, a média

de IMWG(0,01) foi a menor, mas novamente, muito próxima de WGabriel.

Tabela 5.1 - Médias e medianas da raiz quadrada normalizada do erro quadrático médio na matriz “Calinski”

Método	Porcentagens de valores retirados aleatoriamente					
	10%		20%		30%	
	Média	Mediana	Média	Mediana	Média	Mediana
GabrielEigen	0,3512	0,3382	0,3538	0,3510	0,3709	0,3660
IMWG(0.01)	0,3426	0,3375	0,3293	0,3215	0,3345	0,3277
WGabriel	0,3419	0,3376	0,3292	0,3252	0,3346	0,3289

Finalmente, um aspecto importante é o relacionado com os pesos utilizados na matriz “Calinski”, que foram escolhidos por validação cruzada e dos quais depende o sistema de imputação múltipla livre de distribuição IMWG(0,01).

Figura 5.5 - Distribuição dos pesos para WGabriel na matriz “Calinski”

A Figura 5.5 mostra as distribuições dos pesos para WGabriel. Quando a porcentagem de valores ausentes aumentou, a distribuição dos pesos foi fortemente assimétrica à esquerda com mediana próxima a 0,4. Nesse caso, as estatísticas básicas dos pesos podem ser de interesse para avaliar os parâmetros de centralidade e variabilidade, portanto, apresentam-se na Tabela 5.2.

A sugestão de fazer validação cruzada no intervalo $[-2, 2]$, foi certamente justificada nesse conjunto de dados porque os pesos caíram dentro dele. Mais especificamente, do box plot para 10% de imputação, observa-se que os pesos caíram no intervalo $[-1, 3; 0, 9]$, enquanto para 20% e 30%, caíram em $[-1, 3; 0, 97]$ e $[-1, 27; 0, 96]$ respectivamente.

Tabela 5.2 - Estatísticas dos pesos para WGabriel

Estatísticas	Porcentagens de valores retirados aleatoriamente		
	10%	20%	30%
Média	0,3649	0,2734	0,1152
Mediana	0,4200	0,4000	0,3600
Desvio padrão	0,3103	0,4837	0,6350
Q3-Q1(*)	0,3100	0,3900	1,1000

(*)Q3-Q1=Distância interquartílica

5.4 Conclusões

Os procedimentos de imputação propostos neste capítulo, WGabriel e IMWG(0,01), forneceram os melhores resultados na matriz utilizada no estudo de simulação. Esses métodos minimizaram M^2 , NRMSE e maximizaram $corr^2$ em todas as porcentagens de imputação consideradas.

Para situações com alta porcentagem de valores omissos ($> 10\%$), o método mais favorável é IMWG(0,01), pois com esse método de imputação múltipla livre de distribuição é também possível obter uma estimativa da variância entre imputações que quantifica a incerteza sobre os valores reais a serem imputados. Entretanto, essa estimativa não pode se obter com os outros dois métodos, razão pela qual seria necessário o uso de técnicas adicionais de reamostragem como o bootstrap proporcional.

Neste estudo, o processo utilizado para produzir artificialmente valores omissos na matriz foi MCAR. Contudo, isso não quer dizer que os métodos não possam ser usados quando os valores faltantes em aplicações práticas forem MAR ou MNAR; por exemplo, quando os dados tiverem um padrão claro. O único requisito para aplicação dos métodos consiste em que o conjunto de dados possa ser arranjado de forma matricial.

Os três métodos apresentados não fazem alguma pressuposição distribucional ou estrutural e não têm restrições quanto ao padrão ou mecanismo de ausência dos dados faltantes. No entanto, uma avaliação mais extensa dos métodos será necessária antes de tirar conclusões definitivas. Uma boa alternativa para fazê-lo é por meio da construção de mais simulações em outras matrizes completas de dados reais.

Referências

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

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; DIAS, C.T.S.; KRZANOWSKI, W.J. An alternative methodology for imputing missing data in trials with genotype-by-environment interaction. **Biometrical Letters**, Poznan, v.47, p.1-14, 2010.

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

CALINSKI, T.; CZAJKA, S.; KACZMAREK, Z.; KRAJEWSKI, P.; PILARCZYK, W. Analyzing the Genotype-by-Environment Interactions Under a Randomization-Derived Mixed Model. **Journal of Agricultural, Biological and Environmental Statistics**, Berlin, v.14, p.224-241, 2009.

CHING, W.; LI, L.; TSING, N.; TAI, C.; NG, T. A weighted local least squares imputation method for missing value estimation in microarray gene expression data. **International Journal of Data Mining and Bioinformatics**, Olney, v.4, p.331-347, 2010.

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

DI CIACCIO, A. Bootstrap and nonparametric predictors to impute missing data. In: FICHET, B.; PICCOLO, D.; VERDE, R.; VICHI, M. **Classification and Multivariate Analysis for Complex Data Structures, Studies in Classification, Data Analysis, and Knowledge Organization**. Berlin: Springer. 2011. Part IV, p.203-210.

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

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

GARCÍA-PEÑA, M.; DIAS, C.T.S. Analysis of bivariate additive models with multiplicative interaction (AMMI). **Biometric Brazilian Journal**, Jaboticabal, v.27, p. 586-602, 2009.

GAUCH, H.G. A simple protocol for AMMI analysis of yield trials. **Crop Science**, Madison, v.53, p.1860-1869, 2013.

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

JOSSE, J.; PAGÈS, J.; HUSSON, F. Multiple imputation in PCA. **Advances in data analysis and classification**, Berlin, v.5, p. 231-246, 2011.

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

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

———. **Principles of multivariate analysis: A user's perspective**. Oxford: University Press, 2000. 586p.

KROONENBERG, P.M. **Applied multiway data analysis**. John Wiley & Sons, New York, NY, 2008. 579p.

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 droughtprone ecosystem. **Field Crops Research**, New York, v.133, p. 37–47, 2012.

LITTLE, R.; RUBIN, D. **Statistical analysis with missing data**. 2nd ed. John Wiley & Sons, New York, NY, 2002. 408p.

PADEREWSKI, J.; RODRIGUES, P.C. The usefulness of EM-AMMI to study the influence of missing data pattern and application to Polish post-registration winter wheat data. **Australian Journal of Crop Science**, Australia, v.8, p. 640–645, 2014.

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

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

R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>. Acesso em: 15 dez. 2014.

RODRIGUES, P.; 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**, Piracicaba, v.68, p. 679-686, 2011.

RUBIN, D.B. Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. In: SURVEY RESEARCH METHODS SECTION OF THE AMERICAN STATISTICAL ASSOCIATION, 1978, Alexandria. **Proceedings ...** Alexandria: The American Statistical Association, 1978, p.20-34.

SABAGHNIA, N.; KARIMIZADEH, R.; MOHAMMADI M. Model selection in additive main effect and multiplicative interaction model in durum wheat. **Genetika**, Zemun, v.44, p. 325 – 339, 2012.

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

VAN BUUREN, S. **Flexible imputation of missing data**. Boca Raton, CRC press, 2012. 342p.

WRIGHT, K. **agridat**: Agricultural datasets. R package version 1.4, 2012. Disponível em: <http://CRAN.R-project.org/package=agridat>>. Acesso em: 15 dez. 2014.

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

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

6 TRABALHOS FUTUROS

Como continuidade dos estudos apresentados nessa tese, seria interessante:

- Comparar os algoritmos aqui propostos com a imputação múltipla paramétrica utilizando um modelo linear clássico.
- Considerar os mecanismos de ausência MNAR e MAR nos estudos de simulação para experimentos ($G \times E$).
- Estudar a imputação de dados climáticos utilizando a DVS (ver Apêndice).
- Estender a aplicação dos sistemas de imputação aqui propostos para dados categorizados.
- Propor uma extensão dos métodos de imputação simples determinística à imputação múltipla por meio de técnicas de reamostragem como: bootstrap, validação cruzada e jackknife.
- Corrigir o viés dos autovalores na DVS para melhorar a qualidade das imputações.

APÊNDICES

A Imputação de dados climáticos utilizando a decomposição por valores singulares: Uma comparação empírica *

Resumo

Um problema comum em dados climáticos é a informação ausente. Recentemente, foram desenvolvidos quatro métodos de imputação que têm como base a decomposição por valores singulares de uma matriz (DVS). O objetivo deste capítulo é avaliar os novos desenvolvimentos fazendo uma comparação por meio de um estudo de simulação baseado em duas matrizes completas de dados reais. Uma matriz corresponde à precipitação histórica de Piracicaba/SP – Brasil, enquanto a outra matriz corresponde às características meteorológicas multivariadas na mesma cidade desde o ano 1997 até 2012. No estudo foram feitas retiradas aleatórias de diferentes porcentagens com posterior imputação, comparando as metodologias por meio de três critérios: a raiz quadrada normalizada do erro quadrático médio, a estatística de similaridade de Procrustes e o coeficiente de correlação não paramétrico de Spearman. Concluiu-se que a DVS deve ser utilizada unicamente quando sejam analisadas matrizes multivariadas e no caso de matrizes de precipitação, a imputação pela média mensal supera o desempenho de métodos baseados na DVS.

Palavras-chave: Imputação; DVS; Observações ausentes

Abstract

A common problem in climate data is missing information. Recently, four methods have been developed which are based in the singular value decomposition of a matrix (SVD). The aim of this chapter is to evaluate the new developments making a comparison by means of a simulation study based on two complete matrices of real data. One corresponds to the historical precipitation of Piracicaba / SP - Brazil and the other matrix corresponds to multivariate meteorological characteristics in the same city from year 1997 to 2012. In the study, values were deleted randomly at different percentages with subsequent imputation, comparing the methodologies by three criteria: the normalised root mean squared error, the similarity statistic of Procrustes and the Spearman correlation coefficient. It was concluded that the SVD should be used only when multivariate matrices are analysed and in matrices of precipitation, the monthly mean outperforms methods based on the SVD.

Keywords: Imputation; SVD; Missing values

*Apêndice aceito para publicação na Revista Brasileira de Meteorologia.

A.1 Introdução

Frequentemente, nos estudos de climatologia são necessárias observações completas (sem falta de informação) para realizar as análises apropriadas dos dados que são coletados durante um determinado período de tempo, em diferentes estações de uma ou várias regiões de interesse. É comum encontrar séries de dados climatológicos com dados omissos devido a várias razões, como falhas dos instrumentos de medição, condições climáticas extremas e erros na digitação. Uma maneira muito comum de analisar dados provenientes de estudos com informação faltante é imputar as observações ausentes e posteriormente, aplicar procedimentos clássicos sobre os dados completados (observados + imputados). Um método amplamente usado na literatura é utilizar a média como imputação.

Existem muitos métodos estatísticos para resolver o problema de dados omissos como Little e Rubin (2002), Schafer (1997), McLachlan e Krishnan (1997), Tanner e Wong (1987). Outras metodologias com desenvolvimentos recentes são os métodos de imputação simples e múltipla. Eles são uma alternativa às aproximações baseadas em verossimilhanças e são recomendados por alguns autores (Rubin, 1987; Allison 2001; Schafer, 1997; Schafer e Graham, 2002). Em meteorologia, os primeiros estudos usavam estimações de regressão linear simples ou simplesmente a média (Paulhus e Kohler, 1952). Outros autores compararam métodos de imputação como a análise discriminante múltipla, regressão linear múltipla, razão normal e os algoritmos de Esperança-Maximização ou EM (Young, 1992; Makhuvha et al., 1997; Xia et al. 1999a, b). Estudos recentes usaram aproximações para completar séries de temperatura (mínima e máxima) e precipitação total, como descrito em Eischeid et al. (2000). Entretanto, Teegavarapu e Chandramouli (2005), usaram o método modificado da distância entre as estações com dados faltantes e as estações de referência para estimação; também são usados conceitos de redes neurais, interpolação estocástica e kriging. Schneider (2001) trabalhou com o algoritmo EM regularizado e sugere seu uso quando o número de variáveis seja maior do que o tamanho da amostra. Cano e Andreu (2010), estudaram a imputação múltipla definindo uma estratégia de construção da base de dados (usando defasagens como variáveis de suporte e desta maneira criar uma aproximação para manejar o efeito da distância) para evitar ruído nas simulações.

Outros autores usaram métodos alternativos para a imputação de dados como o mapa de auto-ordenamento, perceptron multicamada, método do vizinho mais próximo (Junninen et al., 2004; Ramos-Calzado et al., 2008; Coulibaly e Evora, 2007; Lucio et al., 2007; Kalteh e Berndtsson, 2007; Kalteh e Hjorth, 2009). Smith e Aretxabaleta (2007) apresentaram um método alternativo à análise da função ortogonal empírica, baseado no algoritmo EM para estimar os parâmetros de um modelo de mistura Gaussiano, indicando, também, que este método permite uma interpretação clara da variabilidade associada com cada regime. Aly et al. (2009) compararam quatro métodos de interpolação estocástica

e determinística para imputar dados faltantes em precipitação diária, por outro lado, Lo Presti et al. (2010), propuseram um método em duas etapas, a primeira, consiste em identificar as estações vizinhas e similares à estação com dados faltantes (coeficiente de similaridade) e a segunda, usar um método de regressão para realizar as estimações. Existem outros estudos no âmbito de imputação de dados como os descritos em Yozgatligil et al. (2013).

Os métodos de imputação utilizados neste estudo serão a Média, EM-SVD, imputação biplot, imputação por meio de uma aproximação de posto inferior ponderada (ou EMSJ) e imputação GabrielEigen. A comparação será feita por meio da raiz quadrada normalizada do erro quadrático médio, a estatística de similaridade de Procrustes e o coeficiente de correlação de Spearman. Neste trabalho, serão criados três cenários de dados omissos, 10%, 20% e 40% para os dados descritos posteriormente na seção de material e métodos. Depois, esses valores são estimados usando os métodos já mencionados e para os dados completados (observados + imputados) são calculados e comparados os valores dos critérios de comparação.

A.2 Material e métodos

A.2.1 A decomposição por valores singulares de uma matriz (DVS)

Esta ferramenta é a base que utiliza a maioria dos métodos de imputação considerados neste estudo, daí, a importância para apresentá-la inicialmente. Qualquer matriz $\mathbf{S}$ ($n \times p$) pode ser decomposta por valor singular na forma $\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, em que $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_p$ e $\mathbf{D} = \text{diag}(d_1, \dots, d_p)$ com $d_1 \geq d_2 \geq \dots \geq d_p \geq 0$. As matrizes $\mathbf{S}^T\mathbf{S}$ e $\mathbf{S}\mathbf{S}^T$ têm os mesmos autovalores e os elementos d_i são a raiz quadrada destes autovalores; a i -ésima coluna $\mathbf{v}_i = (v_{i1}, \dots, v_{ip})^T$ da matriz $\mathbf{V}$ ($p \times p$) é o autovetor correspondente ao i -ésimo maior autovalor d_i^2 de $\mathbf{S}^T\mathbf{S}$; enquanto a j -ésima coluna $\mathbf{u}_j = (u_{1j}, \dots, u_{nj})^T$ da matriz $\mathbf{U}$ ($n \times p$) é o autovetor correspondente ao j -ésimo maior autovalor d_j^2 de $\mathbf{S}\mathbf{S}^T$. A decomposição tem sua representação elementar como $s_{ij} = \sum_{h=1}^p u_{ih}d_hv_{jh}$ (Krzanowski, 1988).

A.2.2 Métodos de imputação

Média: É um método de imputação simples muito usado nas ciências sociais e consiste em imputar cada valor ausente pela média da correspondente variável ou coluna se os dados estão arranjados em uma matriz $\mathbf{X}$ de dimensão ($n \times p$) com n linhas e p colunas. É um método rápido, mas dependendo do conjunto de dados e da porcentagem de dados faltantes pode apresentar desvantagens tais como a compressão da distribuição das variáveis e a distorção das relações entre elas (Durrant, 2005).

EM-SVD: Perry (2009) apresenta um método de imputação que mistura o algoritmo EM (Esperança-Maximização) com a decomposição em valores singulares de uma matriz (DVS). A metodologia é apresentada a seguir. Considere a matriz $\mathbf{A}$ de dimensão $(n \times p)$ com elementos A_{ij} ($i = 1, \dots, n$; $j = 1, \dots, p$) e com alguns deles faltantes.

EM-SVD passo 1: Seja $I = \{(i, j) : A_{ij} \neq *\}$, ou seja, o conjunto I representa todos os valores observados.

EM-SVD passo 2: Para $1 \leq j \leq p$ seja μ_j a média dos valores não faltantes na coluna j de $\mathbf{A}$ ou 0 se todas as caselas na coluna j são faltantes.

EM-SVD passo 3: Defina $A^{(0)}$ por

$$A_{ij}^{(0)} = \begin{cases} A_{ij} & \text{se } (i, j) \in I \\ \mu_j & \text{em caso contrário} \end{cases}$$

EM-SVD passo 4: Iniciar a contagem das iterações em zero, $N \leftarrow 0$.

EM-SVD passo 5: É feita a maximização calculando a DVS de $\mathbf{A}^{(N)}$, assim, $\mathbf{A}^{(N)} = \sum_{i=1}^p d_i^{(N)} \mathbf{u}_i^{(N)} \mathbf{v}_i^{(N)T}$ e calculando a DVS de $\mathbf{A}_k^{(N)}$, ou seja, a DVS truncada com k termos, tal que $\mathbf{A}_k^{(N)} = \sum_{i=1}^k d_i^{(N)} \mathbf{u}_i^{(N)} \mathbf{v}_i^{(N)T}$.

EM-SVD passo 6: É calculada a esperança definindo a matriz $\mathbf{A}^{(N+1)}$ de dimensão $(n \times p)$ como

$$A_{ij}^{(N+1)} = \begin{cases} A_{ij} & \text{se } (i, j) \in I \\ A_{k,ij}^{(N)} & \text{em caso contrário} \end{cases}$$

EM-SVD passo 7: Calcular $RSS^{(N)} = \left\| \mathbf{A} - \mathbf{A}_k^{(N)} \right\|_{F,I}^2$, se $|RSS^{(N)} - RSS^{(N-1)}|$ é pequena, declare convergência e obtenha a matriz $\mathbf{A}_k^{(N)}$ que conterá as imputações dos valores ausentes. Caso contrário, aumente $N \leftarrow N + 1$ e volte para **EM-SVD passo 5**.

Imputação biplot: Recentemente, Yan (2013) descreveu um método de imputação utilizando a DVS e a qual é a técnica básica para análise biplot (Gabriel, 1971; 2002). O método é apresentado a seguir.

Biplot passo 1: Considere a matriz $\mathbf{X}$ de dimensão $(n \times p)$ com elementos x_{ij} ($i = 1, \dots, n$; $j = 1, \dots, p$), em que alguns deles são ausentes x_{ij}^{aus} . Inicialmente, os dados faltantes são imputados pela média dos valores observados na respectiva coluna, obtendo assim uma matriz $\mathbf{X}$ completada.

Biplot passo 2: As colunas da matriz $\mathbf{X}$ completada são padronizadas 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). Os elementos padronizados serão notados por p_{ij} e modelados por meio de um biplot bidimensional, ou seja, $p_{ij} = \frac{(x_{ij} - m_j)}{s_j} =$

$\sum_{k=1}^2 \lambda_k \alpha_{ik} \gamma_{jk} + \varepsilon_{ij}$. 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's. ε_{ij} é o residuo para a linha i na coluna j . A matriz com elementos padronizados p_{ij} será denotada por $\mathbf{P}$.

Biplot passo 3: É calculada a DVS da matriz $\mathbf{P}$ e os valores p_{ij} são atualizados utilizando apenas dois CP's da DVS obtendo uma nova matriz chamada $\mathbf{P}^{(2)}$ com elementos $p_{ij}^{(2)}$.

Biplot passo 4: Todos os elementos $p_{ij}^{(2)}$ em $\mathbf{P}^{(2)}$ devem ser retornados à sua escala original assim $\hat{x}_{ij}^{(2)} = m_j + s_j p_{ij}^{(2)}$. Desta maneira é obtida uma nova matriz $\mathbf{X}^{(2)}$ de dimensão $(n \times p)$. Os elementos ausentes x_{ij}^{aus} na matriz $\mathbf{X}$ original são imputados pelo correspondente valor $\hat{x}_{ij}^{(2)}$ de $\mathbf{X}^{(2)}$.

Biplot passo 5: O processo é iterado (voltando ao **Biplot passo 2**) até alcançar estabilidade nas imputações. Por exemplo, as iterações devem ser realizadas até que $\frac{d}{\bar{y}} < 0.01$. Definindo $d = \left[\left(\frac{1}{na} \right) \sum_{i=1}^{na} (x_i - x_i^A)^2 \right]^{\frac{1}{2}}$ e $\bar{y} = \left[\left(\frac{1}{N} \right) \sum_{i=1}^g \sum_{j=1}^e y_{ij}^2 \right]^{\frac{1}{2}}$. Em que d representa a diferença entre os valores preditos para todos os valores ausentes na iteração atual e na iteração anterior. Nessa estatística “ na ” é o número total de valores ausentes na matriz $\mathbf{X}$, x_i é o valor predito para o i -ésimo dado faltante na iteração atual e x_i^A na iteração anterior. Entretanto, 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, sendo N o número total de valores observados.

Imputação por meio de uma aproximação de posto inferior ponderada

(ou EMSJ): Srebro e Jaakkola (2003) apresentam um algoritmo EM simples e eficiente para calcular a aproximação de posto inferior ponderada de uma matriz. No caso de dados faltantes as ponderações podem ser 0 se for ausente e 1 se for observado. Este método, recentemente, foi considerado por Canas (2012) para propor os modelos WAMMI (Weighted Additive Main effects and Multiplicative Interaction) na análise da interação genótipo×ambiente com heterogeneidade de variâncias. A metodologia é apresentada a seguir.

Considere a matriz $\mathbf{Y}$ de dimensão $(n \times p)$ com elementos y_{ij} ($i = 1, \dots, n; j = 1, \dots, p$) sendo alguns deles faltantes. Construa uma matriz $\mathbf{W}$ ($n \times p$) com valores $w_{ij} = 0$ se y_{ij} for faltante e $w_{ij} = 1$ em caso contrário. Construa a matriz $\mathbf{1}$ ($n \times p$) com uns em todas as posições. De maneira iterativa calcule

$$\mathbf{X}^{(t+1)} = DVS (\mathbf{W} \odot \mathbf{Y} + (\mathbf{1} - \mathbf{W}) \odot \mathbf{X}^{(t)})$$

em que t faz referência ao número da iteração no processo, $\odot$ representa o produto de

Hadamard. Para $t = 0$, $\mathbf{X}$ deve ser iniciada como $\mathbf{X}^{(0)} = \mathbf{0}$. O processo é iterado enquanto a soma de quadrados entre duas iterações consecutivas $\mathbf{X}^{(t+1)}$ e $\mathbf{X}^{(t)}$ seja maior do que um valor especificado (por exemplo, 10^{-9}). A saída deste procedimento são as matrizes $\mathbf{U}_k$, $\mathbf{D}_k$ e $\mathbf{V}_k$ tal que $\tilde{\mathbf{Y}} = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T$, em que k é o posto da aproximação. Da matriz $\tilde{\mathbf{Y}}$ são obtidas as imputações para os valores ausentes da matriz $\mathbf{Y}$ original.

Imputação GabrielEigen: Arciniegas-Alarcón et al. (2010) propuseram um método de imputação que mistura a regressão com aproximação de posto inferior utilizando a DVS. O método é apresentado a seguir. Suponha uma matriz $\mathbf{X}$ de dimensão $(n \times p)$ com elementos x_{ij} ($i = 1, \dots, n; j = 1, \dots, p$), em que alguns deles são ausentes.

GabrielEigen passo 1: Os valores ausentes são imputados, inicialmente, pela média da respectiva coluna obtendo uma matriz $\mathbf{X}$ completada.

GabrielEigen passo 2: São padronizadas as colunas da matriz $\mathbf{X}$, 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).

GabrielEigen passo 3: Sobre a matriz padronizada é recalculada a imputação de cada valor x_{ij} ausente usando-se $\hat{x}_{ij}^{(m)} = \mathbf{x}_{1\bullet}^T \mathbf{V} \mathbf{D}^{-1} \mathbf{U}^T \mathbf{x}_{\bullet 1}$, em que os vetores $\mathbf{x}_{1\bullet}^T$, $\mathbf{x}_{\bullet 1}$ e as matrizes $\mathbf{V}$, $\mathbf{D}$ e $\mathbf{U}$, são obtidos da partição $\mathbf{X} = \begin{bmatrix} x_{ij} & \mathbf{x}_{1\bullet}^T \\ \mathbf{x}_{\bullet 1} & \mathbf{X}_{11} \end{bmatrix}$, com $\mathbf{X}_{11} = \sum_{k=1}^m \mathbf{u}_{(k)} d_k \mathbf{v}_{(k)}^T = \mathbf{U} \mathbf{D} \mathbf{V}^T$, sendo $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m]$, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m]$, $\mathbf{D} = \text{diag}(d_1, \dots, d_m)$ e $m \leq \{n-1, p-1\}$. Para cada observação faltante os componentes da partição considerada serão diferentes e dita partição é obtida por meio de operações elementares nas linhas e colunas da matriz $\mathbf{X}$.

GabrielEigen passo 4: O processo de imputação depende da escolha do valor para

$$m \text{ e o critério será o seguinte: } m \text{ tal que, } \frac{\sum_{k=1}^m d_k^2}{\min\{n-1, p-1\}} \approx 0,75.$$

GabrielEigen passo 5: Finalmente, os valores imputados devem ser retornados à sua escala original assim, $x_{ij} = m_j + s_j \hat{x}_{ij}^{(m)}$, substituindo-os na matriz $\mathbf{X}$. Então, o processo é iterado (voltando ao **GabrielEigen passo2**) até alcançar estabilidade nas imputações. Todo o processo deve ser feito sobre a matriz $\mathbf{X}$, tal que $n > p$, caso contrario, a matriz deve ser transposta.

A.2.3 Características dos dados

Para avaliar os métodos de imputação foram considerados dois conjuntos de dados completos, obtidos da série de dados climatológicos do campus Luiz de Queiroz na cidade de Piracicaba – SP, coletada pela estação meteorológica (localizada juntamente com a

pluviométrica) do Departamento de Engenharia de Biosistemas da Universidade de São Paulo. Os dados podem ser acessados em <http://www.leb.esalq.usp.br/posto.html>.

O primeiro conjunto de dados corresponde à matriz histórica de totais mensais de precipitação em milímetros desde o ano 1917 até o ano 2012. Assim, a matriz “univariada” tem dimensão (96×12) , em que as linhas representam os anos e as colunas os meses.

O segundo conjunto de dados corresponde à matriz “multivariada” de dados diários (coletados entre as 0h e as 24h) a partir do mês de janeiro de 1997 até dezembro de 2012. Em 1997 começou a coleta das seguintes 4 variáveis: radiação global (cal/cm), umidade relativa média (%), amplitude térmica (graus Celsius) e chuva (mm). Para obter um conjunto completo foram considerados somente os dias com a informação coletada em todas as variáveis, obtendo uma matriz de dimensão (5818×4) , em que as linhas representam os dias e as colunas representam as variáveis.

Desde o ponto de vista prático, considerar uma matriz “multivariada” é muito importante, pois, a maioria dos problemas na natureza não são univariados e qualquer variável das consideradas pode ser afetada pela ocorrência de dados faltantes. Desde o ponto de vista estatístico, a presença conjunta de várias variáveis é justificada pela matriz de correlação entre elas (Tabela A.1), sendo todas as correlações significativas ($valor - p < 0.001$). Assim, por exemplo, a umidade relativa (UR) média tem uma correlação moderada positiva com a chuva ($r = 0,40$) e moderada negativa com a radiação global ($r = -0,49$). Também pode se observar uma alta correlação negativa com a amplitude térmica ($-0,73$).

Tabela A.1 - Matriz de correlação

Correlações	UR média	Chuva	Rad. Global	Ampl. Term.
UR média	1,00	0,40	-0,49	-0,73
Chuva	0,40	1,00	-0,27	-0,37
Rad.Global	-0,49	-0,27	1,00	0,49
Ampl. Term.	-0,73	-0,37	0,49	1,00

A.2.4 Estudo de simulação

Cada matriz de dados (univariada e multivariada) foi submetida a retiradas aleatórias de diferentes porcentagens. Foram consideradas as porcentagens de 10%, 20%, e 40%. O processo foi repetido em cada conjunto de dados 1000 vezes para cada porcentagem de retirada, obtendo 3000 matrizes diferentes com valores omissos. No total, foram gerados 6000 conjuntos de dados incompletos e em cada um, os dados foram imputados com os 5 algoritmos já descritos por meio de um programa computacional implementado no R (R Core Team, 2013).

O processo de retirada aleatória para uma matriz $\mathbf{X}$ ($n \times p$) 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 foi menor do que r , então o elemento na posição $(i + 1, j)$ da matriz foi deletado ($i = 0, 1, \dots, n - 1; j = 1, \dots, p$). A proporção esperada de dados ausentes na matriz será r (Krzanowski, 1988). Essa técnica foi utilizada com $r = 0, 1, 0, 2$ e $0, 4$.

Para os métodos de imputação pela Média, EM-SVD e EMSJ foi utilizada a implementação computacional fornecida por Wong (2013). Na aplicação dos algoritmos EM-SVD e EMSJ em cada uma das matrizes incompletas simuladas devia ser escolhido de antemão o número de componentes k a ser utilizado na DVS e para isso foi utilizada a função *cv.SVDImpute* e que faz validação cruzada sobre a informação disponível em uma matriz da seguinte maneira: Dos dados observados na correspondente matriz é eliminado, aleatoriamente, 30% deles e, posteriormente, é feita a imputação por EM-SVD. Usando os dados completados, é, finalmente calculada a raiz do erro quadrático médio (RMSE) sobre a porção de dados que foi artificialmente removida. O valor de k com menor RMSE é o escolhido para fazer a imputação. Neste estudo, sobre cada matriz incompleta simulada foi repetido o processo de validação cruzada 100 vezes e o k selecionado foi aquele que mais frequentemente minimizasse a RMSE.

A.2.5 Critérios de comparação

No conjunto de dados “univariado” foram adotados três critérios para comparar os resultados obtidos nas simulações: a estatística M^2 de Procrustes (Krzanowski, 2000), a raiz normalizada do erro quadrático médio - NRMSE (Ching et al., 2010) e o coeficiente de correlação não-paramétrico de Spearman (Arciniegas-Alarcón e Dias, 2009).

Assim, cada matriz de dados completada (observados+imputados) $\mathbf{Y}_{imp}$ foi comparada com a matriz original $\mathbf{X}_{orig}$ por meio de

$$M^2 = \text{traço} (\mathbf{X}_{orig} \mathbf{X}_{orig}^T + \mathbf{Y}_{imp} \mathbf{Y}_{imp}^T - 2\mathbf{X}_{orig} \mathbf{Q} \mathbf{Y}_{imp}^T)$$

em que $\mathbf{Q} = \mathbf{V} \mathbf{U}^T$ é a matriz de rotação e pode ser calculada com elementos da decomposição por valores singulares da matriz $\mathbf{X}_{orig}^T \mathbf{Y}_{imp} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$. Com a estatística M^2 se obtém uma medida da diferença entre duas configurações de pontos e o método de imputação que minimize essa diferença indicará os melhores resultados.

O segundo critério utilizado foi a NRMSE definida como:

$$NRMSE = \frac{\sqrt{\text{média} (\mathbf{a}_{imp} - \mathbf{a}_{orig})^2}}{dp(\mathbf{a}_{orig})}$$

em que $\mathbf{a}_{imp}$ e $\mathbf{a}_{orig}$ são vetores contendo os respectivos valores preditos e os valores verdadeiros das observações ausentes simuladas. $dp(\mathbf{a}_{orig})$ é o desvio padrão dos valores contidos no vetor $\mathbf{a}_{orig}$. Quanto menor seja a NRMSE melhor será o método de imputação.

O terceiro critério de comparação no conjunto de dados “univariado” foi o coeficiente de correlação de Spearman. 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.

Para o conjunto de dados “multivariado” foram utilizados dois critérios de comparação: a estatística M^2 apresentada anteriormente e por causa das diferentes escalas, uma versão modificada da NRMSE segundo Audigier et al. (2013) chamada NRMSE2. Para isto, deve ser construída para cada matriz incompleta simulada uma matriz indicadora de observações ausentes $\mathbf{W}$ com elementos w_{ij} (como foi descrito no método de imputação EMSJ)

$$NRMSE2 = \sqrt{\frac{\sum_{i=1}^n \sum_{j=1}^p w_{ij} \left(\frac{x_{ij} - \hat{x}_{ij}}{s_{\mathbf{x}_j}} \right)^2}{\sum_{i=1}^n \sum_{j=1}^p w_{ij}}}$$

Na NRMSE2 x_{ij} representa o valor original, $\hat{x}_{ij}$ representa o valor imputado e $s_{\mathbf{x}_j}$ representa o desvio padrão verdadeiro da variável j na qual foi simulada a observação ausente. Quanto menor seja a NRMSE2 melhor será o método de imputação.

A.3 Resultados e discussão

Matriz univariada de precipitação

Na Tabela A.2 apresentam-se a média e a mediana da NRMSE. Observa-se que em todas as porcentagens o método menos recomendável é o EMSJ, pois maximizou o critério (com valores médios de 2,0299, 1,4459 e 2,4480) e o mais recomendável seria a imputação por Média mensal porque o minimizou com um valor aproximadamente de 0,68 em todas as porcentagens. Estabelecendo uma ordem do mais eficiente ao menos eficiente segundo a NRMSE, os métodos seriam Média, GabrielEigen ou EM-SVD, Biplot e finalmente o EMSJ.

Estudando com mais detalhe o método EMSJ e procurando uma explicação do baixo desempenho, encontrou-se que existiram múltiplos problemas de convergência, especialmente quando a porcentagem de retirada foi 40% (em geral, nem 20000 iterações foram

Tabela A.2 - Média e mediana da NRMSE na matriz de precipitação

Método	Porcentagem de retirada aleatória					
	10%		20%		40%	
	Média	Mediana	Média	Mediana	Média	Mediana
Média	0,6856	0,6828	0,6862	0,6831	0,6882	0,6875
GabrielEigen	0,7156	0,7127	0,7328	0,7309	0,8096	0,8082
EM-SVD	0,7409	0,7364	0,7507	0,7492	0,8031	0,7916
Biplot	0,8171	0,8107	0,8453	0,8391	0,8919	0,8868
EMSJ	2,0299	2,0130	1,4459	1,4407	2,4480	1,0894

suficientes para convergir). Os efeitos de tais problemas de convergência podem ser vistos adicionalmente quando foi estudada a correlação entre os dados imputados com os correspondentes dados verdadeiros na Figura A.1.

Figura A.1 - Correlação com diferentes porcentagens de imputação na matriz de precipitação

Quando se imputaram 10% e 20% os métodos apresentaram distribuições de correlações aproximadamente simétricas, sendo o Biplot o método com correlação mais baixa (mediana próxima de 0,6) e a Média com correlação mais alta (mediana próxima de 0,8). No entanto, nessas duas porcentagens, independente do método, as correlações sempre foram moderadas ou altas. Na porcentagem de 40 % a imputação pela Média manteve seu bom desempenho e o EMSJ confirmou que não deve ser utilizado quando a quantidade de dados faltantes seja grande, pois a variação de resultados pode ser alta. De acordo com os resultados da NRMSE e da correlação de Spearman decidiu-se não considerar mais neste conjunto de dados o EMSJ por se tornar um método não competitivo e pouco prático por causa de sua convergência lenta.

Na Figura A.2 se apresenta a distribuição da estatística M^2 de Procrustes utilizando diferentes porcentagens de imputação. Lembre-se que quanto menor seja dita estatística, melhor será o método de imputação. Assim, com 10% de imputação poderiam ser utilizados os métodos GabrielEigen ou Média, mas, conforme aumenta a porcentagem a Média torna-se novamente o método mais recomendado. Segundo a M^2 , a ordem do método mais eficiente ao menos eficiente seria: Média, EM-SVD, GabrielEigen e por último o Biplot.

Figura A.2 - M^2 com diferentes porcentagens de imputação na matriz de precipitação

Todos os anteriores resultados foram verificados por meio de testes não-paramétricos de Friedman na comparação de todos os métodos, com posteriores testes não paramétricos de Wilcoxon para comparações dois a dois (Sprenst e Smeeton, 2001). Em todos os casos houve significância; não se apresentam, por questão de espaço e simplicidade para o leitor.

Matriz multivariada de precipitação

Na Tabela A.3 apresentam-se a média e a mediana da NRMSE2. Observa-se que, para todas as porcentagens de retirada aleatória o método GabrielEigen foi o melhor porque sempre minimizou o critério. A diferença dos resultados encontrados na matriz univariada de precipitação o GabrielEigen apresentou melhor desempenho do que a imputação utilizando a Média da variável. Por exemplo, com 10% de imputação a média da NRMSE2 para GabrielEigen foi 0,8057, enquanto para imputação por Média foi de 0,9987. Pode-se observar que em todas as porcentagens o algoritmo EMSJ teve o mais baixo desempenho.

Pode ser estabelecida uma ordem para os métodos de acordo com sua porcentagem de imputação. Assim, com 10% e 20% de imputação, a ordem do algoritmo mais eficiente ao menos eficiente será: GabrielEigen, Média, Biplot, EM-SVD e EMSJ. Com 40% de

Tabela A.3 - Média e mediana da NRMSE2 na matriz multivariada

Método	Porcentagem de retirada aleatória					
	10%		20%		40%	
	Média	Mediana	Média	Mediana	Média	Mediana
Média	0,9987	0,9964	0,9993	0,9972	1,0002	0,9997
Biplot	1,0483	1,0472	1,0149	1,0140	0,9729	0,9736
GabrielEigen	0,8057	0,8013	0,8232	0,8213	0,8703	0,8703
EM-SVD	1,2285	1,2280	1,2040	1,2040	1,1490	1,1488
EMSJ	2,1279	2,1229	2,5231	2,5220	7,2241	9,1977

imputação a ordem já descrita é alterada porque o Biplot supera a Média. Vale a pena comentar que problemas de convergência do EMSJ similares aos encontrados na matriz univariada se mantiveram na matriz multivariada.

Na Figura A.3 se apresenta a distribuição da estatística M^2 de Procrustes utilizando diferentes porcentagens de imputação na matriz multivariada. Com este critério, o método EMSJ teve uma dispersão muito grande, razão pela qual não permitiu fazer a comparação simultânea com os outros quatro métodos. Por essa característica, é o método menos recomendado e excluído da análise quando foi considerada a M^2 . O método que minimizou M^2 em todas as porcentagens foi o GabrielEigen, quer dizer, com este método se obteve a maior similaridade entre as matrizes completadas por imputação e a matriz original a partir da qual foi feito o estudo de simulação.

Figura A.3 - M^2 com diferentes porcentagens de imputação na matriz multivariada

Por outro lado, o método com mais baixo desempenho foi o EM-SVD, porque em todas as porcentagens consideradas maximizou o critério. Os métodos, Média e Biplot,

foram melhores que o EM-SVD, mas, foram superados pelo GabrielEigen.

A.4 Conclusões

Os resultados obtidos neste estudo a partir de duas matrizes reais de dados climáticos oferecem alguns guias para análises e pesquisas futuras sobre observações ausentes. Primeiro, se for considerada uma matriz univariada com dados faltantes, a imputação pela Média oferece melhores resultados do que outro algoritmo que envolva a DVS. Segundo, se a matriz de análise for multivariada e exista alguma razão para assumir correlação não nula entre as variáveis, o método GabrielEigen deveria ser considerado. Em geral, pode-se concluir também, que o algoritmo EMSJ deve ser utilizado unicamente com porcentagens baixas de observações ausentes, ou seja, menos de 20%.

Referências

- ALLISON, P.D. **Missing data**. Sage university papers series on quantitative applications in the social sciences. Sage, Thousand Oaks, pp 07–136, 2001.
- ALY, A.; PATHAK, C.; TEEGAVARAPU, R.S.V.; AHLQUIST, J.; FUELBERG, H. Evaluation of improvised spatial interpolation methods for infilling missing precipitation records. In: WORLD ENVIRONMENT WATER RESOURCES CONGRESS, 2009, Kansas City. **Proceedings...** Kansas City: American Society of Civil Engineers, 2009. p.5914-5923.
- ARCINIEGAS-ALARCÓN, S.; DIAS, C.T.S. Data imputation in trials with genotype by environment interaction: an application on cotton data. **Biometric Brazilian Journal**, Jaboticabal, v.27, p.125-138, 2009.
- ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; DIAS, C.T.S.; KRZANOWSKI, W.J. An alternative methodology for imputing missing data in trials with genotype-by-environment interaction. **Biometrical Letters**, Poznan, v.47, p.1-14, 2010.
- AUDIGIER, V.; HUSSON, F.; JOSSE, J. **A principal components methods to impute missing values for mixed data**, 2013. Disponível em: <http://arxiv.org/abs/1301.4797>. Acesso em: 15 dez. 2014.
- CANAS, P.J. **New strategies to detect and understand genotype-by-environment interactions and QTL-by-environment interactions**. Dissertation, Universidade Nova de Lisboa, Lisboa, 2012.
- CANO, S.; ANDREU, J. Using multiple imputation to simulate time series: a proposal to solve the distance effect. **WSEAS Transactions Computers**, New York, v.9, n.7, p.768–777, 2010.

CHING, W.; LI, L.; TSING, N.; TAI, C.; NG, T. A weighted local least squares imputation method for missing value estimation in microarray gene expression data. **International Journal of Data Mining and Bioinformatics**, Onley, v.4, n.3, p.331-347, 2010.

COULIBALY, P.; EVORA, N.D. Comparison of neural network methods for in filling missing daily weather records. **Journal of Hydrology**, Amsterdam, v.341, p.27-41, 2007.

DURRANT, G.B. Imputation methods for handling item-nonresponse in the social sciences: a methodological review, **NCRM (NCRM Working Paper Series, (NCRM-002))**, Southampton, 2005.

EISCHEID, J.K.; PASTERIS, P.A.; DIAZ, H.F.; PLANTICO, M.S.; LOTT, N.J. Creating a serially complete, national daily time series of temperature and precipitation for the western United States. **Journal of Applied Meteorology**, Washington, v.39, n.9, p.1580-1591, 2000.

GABRIEL, K.R. The biplot graphic display of matrices with application to principal component analysis. **Biometrika**, Oxford, v.58, n.3, p.453-467, 1971.

———. Le biplot - outil d'exploration de données multidimensionnelles. **Journal de la Société Française de Statistique**, Paris, v.143, n.3-4, p.5-55, 2002.

JUNNINEN, H.; NISKA, H.; TUPPURAINEN, K.; RUUSKANEN, J.; KOLEHMAINEN, M. Methods for imputation of missing values in air quality data sets. **Atmospheric Environment**, Amsterdam, v.38, n.18, p.2895-2907, 2004.

KALTEH, A.M.; BERNDTSSON, R. Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). **Hydrological Sciences Journal**, Vienna, v.52, n.2, p.305-317, 2007.

KALTEH, A.M.; HJORTH, P. Imputation of missing values in a precipitation-runoff process database. **Hydrology Research**, London, v.40, n.4, p.420-432, 2009.

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

———. **Principles of multivariate analysis: A user's perspective**. Oxford: University Press, 2000. 586 p.

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

LO PRESTI, R.; BARCA, E.; PASSARELLA, G. A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). **Environmental Monitoring and Assessment**, New York, v.160, n.1–4, p.1–22, 2010.

LUCIO, P.S.; CONDE, F.C.; CAVALCANTI, I.F.A.; SERRANO, A.I.; RAMOS, A.M.; CARDOSO, A.O. Spatiotemporal monthly rainfall reconstruction via artificial neural network—case study: south of Brazil. **Advances in Geosciences**, Göttingen, v.10, p.67–76, 2007.

MAKHUVHA, T.; PEGRAM, G.; SPARKS, R.; ZUCCHINI, W. Patching rainfall data using regression methods. 2. Comparisons of accuracy, bias and efficiency. **Journal of Hydrology**, Amsterdam, v.198, n.1–4, p.308–318, 1997.

McLACHLAN, G.; KRISHNAN, T. **The EM algorithm and extensions**. Wiley, New York, 1997.

PAULHUS, J.L.H.; KOHLER, M.A. Interpolation of missing precipitation records. **Monthly Weather Review**, Washington, v.80, p.129–133, 1952.

PERRY, P. **Cross-validation for unsupervised learning**. Dissertation, Stanford University, 2009.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>. Acesso em: 15 dez. 2014.

RAMOS-CALZADO, P.; GÓMEZ-CAMACHO, J.; PÉREZ-BERNAL, F.; PITA-LÓPEZ, M.F. A novel approach to precipitation series completion in climatological datasets: application to Andalusia. **International Journal of Climatology**, Chichester, v.28, n.11, p.1525–1534, Rev A 45, 3403, 2008.

RUBIN, D.B. **Multiple imputation for nonresponse in surveys**. New York: Wiley, 1987. 320 p.

SCHAFER, J.L. **Analysis of incomplete multivariate data**. London: Chapman and Hall/CRC, 1997. 444 p.

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

SCHNEIDER, T. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. **Journal of Climate**, Washington, v.14, p.853–871, 2001.

- SMITH, K.W.; ARETXABALETA, A.L. Expectation–maximization analysis of spatial time series. **Nonlinear Processes in Geophysics**, Göttingen, v.14, n.1, p.73–77, 2007.
- SPRENT, P.; SMEETON, N.C. **Applied nonparametric statistical methods**. 3th ed. Boca Raton: Chapman and Hall, 2001. 463p.
- SREBRO, N.; JAAKKOLA, T. Weighted low-rank approximations. **Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)**, Washington, D.C., 2003.
- TANNER, M.A.; WONG, W.H. The calculation of posterior distributions by data augmentation. **Journal of the American Statistical Association**, Washington, v.82, n.398, p.528-540, 1987.
- TEEGAVARAPU, R.S.V.; CHANDRAMOULI, V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. **Journal of Hydrology**, Amsterdam, v.312, n.1–4, p.191–206, 2005.
- WONG, J. **Imputation**: imputation. R package version 2.0.1, 2013. Disponível em: <http://CRAN.R-project.org/package=imputation>. Acesso em: 15 dez. 2014.
- XIA, Y.; FABIAN, P.; STOHL, A.; WINTERHALTER, M. Forest climatology: reconstruction of mean climatological data for Bavaria, Germany. **Agricultural and Forest Meteorology**, Amsterdam, v.96, n.1-3, p.117-129, 1999a.
- . Forest climatology: estimation of missing values for Bavaria Germany. **Agricultural and Forest Meteorology**, Amsterdam, v.96, n.1-3, p.131-144, 1999b.
- YAN, W. Biplot analysis of incomplete two-way data. **Crop Science**, Madison, v.53, n.1, p.48-57, 2013.
- YOUNG, K.C. A three-way model for interpolating for monthly precipitation values. **Monthly Weather Review**, Washington, v.120, p.2562-2569, 1992.
- YOZGATLIGIL, C.; ASLAN, S.; IYIGUN, C.; BATMAZ, I. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. **Theoretical and Applied Climatology**, Berlin, v.112, p.143-167, 2013.