Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Imputação de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão

Sergio Arciniegas Alarcón

Dissertação apresentada para obtenção do título de Mestre em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2008

Sergio Arciniegas Alarcón
Estatístico

Imputação de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão

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

Dissertação apresentada para obtenção do título de Mestre em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2008

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Arciniegas Alarcón, Sergio
Imputação de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão / Sergio Arciniegas Alarcón. - - Piracicaba, 2008.
82 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura Luiz de Queiroz, 2008.
Bibliografia.

1. Algodão 2. Análise de dados 3. Análise de variância 4. Correlação genética e ambienta 5. Delineamento experimental 6. Genética estatística I. Título

CDD 519.5
A674i

DEDICATÓRIA

A Deus pela força e pela saúde que me deu.

À minha mãe Sonia Alarcón e aos meus irmãos Natalie e Camilo por suportar a ausência.

À família Alarcón-Castro pelo apoio constante para alcançar esse objetivo (sem esquecer ao Edgar).

AGRADECIMENTOS

Ao Prof. Dr. Carlos Tadeu dos Santos Dias pela orientação e apoio para desenvolver este trabalho.

Ao Prof. Dr. Roland Vencovsky do Departamento de Genética pelo conhecimento compartilhado e por fornecer os dados para a pesquisa.

A todos os Professores e funcionários do Programa de Pós-graduação em Estatística e Experimentação Agronômica do Departamento de Ciências Exatas da ESALQ/USP pela atenção.

Ao Lúcio, colega de doutorado, pela grande ajuda na minha chegada no Brasil e pela amizade durante os dois anos do mestrado.

Ao programa PEC-PG do CNPq pelo apoio financeiro para meus estudos no Brasil.

SUMÁRIO

RESUMO

Imputação de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão

Os experimentos multiambientes são um tipo especial dos experimentos bifatoriais, muito usados em melhoramento genético de plantas, nos quais algumas cultivares são avaliadas em diferentes locais. Geralmente nesses estudos se encontra uma resposta diferencial das cultivares em cada local que é chamada de interação genótipo × ambiente ou G×E, que é bem explicada por modelos de efeitos aditivos e interação multiplicativa (AMMI). Frequentemente os experimentos G×E podem ser desbalanceados e um ou vários genótipos não serem testados em alguns locais. Às vezes para o pesquisador recomendar os ambientes pode ser de interesse obter estimativas daquelas combinações genótipo × ambiente que não foram testadas e tais estimativas podem ser calculadas explorando a informação inerente a aquelas combinações que foram atualmente obtidas. Além do interesse do pesquisador por essas estimativas, os dados ausentes podem causar alguma modificação na estimação tradicional dos parâmetros nos modelos AMMI, pois para estimar os parâmetros é necessário um processo sequencial fazendo uma análise de variância com uma posterior decomposição por valor singular da matriz de residuais, a qual não pode ser calculada se existir uma matriz de interação com dados faltantes. Para resolver esses problemas Bergamo (2007) e Bergamo et al. (2008) propuseram uma nova técnica através do uso de imputação múltipla livre de distribuição (IMLD) e é por essa razão que se decidiu avaliar o recente desenvolvimento comparando-o com algumas metodologias de imputação que têm sido usadas com sucesso nos experimentos G×E com dados ausentes como os mínimos quadrados alternados ALS(0), ALS(1) (CALINSKI et al., 1992) e estimativas robustas r-AMMI1 e r-AMMI2 (DENIS; BARIL, 1992). Assim, foi desenvolvido um estudo de simulação baseado em uma matriz de dados reais genótipos (15) × ambientes (27) do ensaio estadual de algodoeiro herbáceo 2000/01 (FARIAS, 2005), fazendo retiradas aleatórias de 10%, 20% e 30%, imputando os dados e comparando os métodos através da raiz quadrada da diferença preditiva média (RMSPD), a estatística de similaridade de Procrustes e o coeficiente de correlação não paramétrico de Spearman. Também foi feita uma análise sobre a escolha de componentes multiplicativos de um modelo AMMI quando se têm matrizes completadas (observados + imputados). Os resultados do estudo de simulação mostraram que segundo a distribuição da RMSPD padronizada, o método r-AMMI1 é o melhor, superando o IMLD. Entretanto, utilizando a estatística de Procrustes se encontrou que completando matrizes com ALS(0) se obtém a maior similaridade com relação à matriz de dados originais, também foi mostrado que os cinco métodos considerados têm uma alta correlação entre as imputações e os correspondentes dados reais. Finalmente, recomenda-se utilizar a imputação de dados para a estimação dos parâmetros de um modelo AMMI sob ocorrência de dados ausentes, mas para determinar o número de componentes multiplicativos é preferível tomar a decisão somente sobre a informação observada.

Palavras-chave: Imputação de dados; Modelos AMMI; Interação genótipo-ambiente

ABSTRACT

Data imputation in trials with genotype by environment interaction: an application on cotton data

The multienvironment trials are a special type of the two-factor experiments, widely used in genetic improvement of plants, where some cultivars are assessed in different locations. Generally, in these studies there is a differential response of cultivars in each location that is called genotype × environment interaction, or G × E, which is well explained by the additive main effects and multiplicative interaction models (AMMI). Often the experiments G×E may be unbalanced and one or several genotypes were not tested in some locations. Sometimes for the environments recommendations, the researcher may be interested in obtain estimates of those combinations G × E that were not tested and such estimates can be calculated using the information of those combinations that were actually obtained. Additionally to the interest of the researchers in these estimates, the missing data may cause some problems in the classical estimation of parameters in the AMMI models, because the parameter estimation need of a sequential process doing an analysis of variance followed by a singular value decomposition, which can not be calculated if there is a matrix of interaction with missing data. To solve these problems Bergamo (2007) and Bergamo et al. (2008) proposed a new technique using the distribution free multiple imputation (IMLD), and for this reason was decided to evaluate the recent development through the comparison with some methods of imputation that have been used successfully in experiments G×E with missing data like the AMMI estimates based on alternating least squares ALS(0), ALS(1) (CALINSKI et al. 1992) and AMMI estimates with robust sub-model r-AMMI1 and r-AMMI2 (DENIS; BARIL, 1992). Thus, was developed a simulation study based on a matrix of true data genotypes (15) × environments (27) of the upland cotton variety trials (ensaio estadual de algodoeiro herbáceo) 2000/01 (FARIAS, 2005), doing missed random (10%, 20%, 30%), imputing the data and comparing the methods through the root mean square predictive difference (RMSPD) of the true value, the Procrustes statistic and the Spearman´s ranks correlation coefficient. Also was made an analysis on the choice of the multiplicative components of an AMMI model after imputation on the complete data sets (observed + imputed). The results of the simulation study has shown that according to the distribution of RMSPD standardized, the r-AMMI1 method is better than the IMLD. However, using the Procrustes statistic was found that imputing data matrix with ALS(0), is obtained the greatest similarity related to the true data matrix. The five methods considered show high correlation between the true and the imputed missing values. Finally, is recommended using the imputation data for the estimation of the parameters of an AMMI model under the presence of missing data, but for choosing the number of multiplicative terms is preferable take the decision only on the observed information.

Keywords: Data imputation; AMMI models; Genotype-environment interaction

LISTA DE FIGURAS

LISTA DE TABELAS

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 erradamente 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 do 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.

Levando em conta as situações acima descritas, neste trabalho foi desenvolvido um estudo baseado em experimentos bifatoriais, onde se tem apenas uma observação por casela e adicionalmente dados ausentes. Um exemplo desta situação pode apresentar-se nos experimentos multiambiente, onde cultivares são estudados 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 tem muita aplicação no melhoramento genético de plantas e são conhecidos como experimentos genótipo por ambiente (G×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×E) são os modelos de efeitos aditivos de interação multiplicativa (AMMI), pois exploram melhor as informações contidas nos dados do que a ANOVA tradicional (CROSSA, 1990 apud DUARTE; VENCOSVSKY, 1999), mas esses modelos têm alguns problemas na estimação dos parâmetros se existirem dados faltantes (DENIS; BARIL, 1992). Por exemplo, na estimação clássica dos modelos AMMI é preciso encontrar a decomposição por valor singular(DVS) da matriz de resíduos não aditivos, mas essa DVS não pode ser calculada na ocorrência de valores ausentes.

Para resolver esses problemas Bergamo (2007) e Bergamo et al. (2008) propuseram uma nova técnica através do uso de imputação múltipla, usando a média das imputações para estimar combinações G×E. Assim, a proposta do presente trabalho é avaliar o recente desenvolvimento comparando-o com algumas metodologias de imputação que têm sido usadas com sucesso nos experimentos genótipo × ambiente com dados ausentes, como os mínimos quadrados alternados-ALS (CALINSKI et al., 1992; PIEPHO, 1995b) e estimativas robustas como r-AMMI (DENIS; BARIL, 1992; PIEPHO, 1995b).

2 DESENVOLVIMENTO

2.1 Interação entre genótipos e ambientes (G×E)

A interação (G×E) é definida como o comportamento diferencial de genótipos em função da diversidade ambiental. Neste sentido, na presença da interação, os resultados das avaliações podem mudar de um ambiente para outro, ocasionando mudanças na posição relativa dos genótipos ou mesmo na magnitude das suas diferenças (FALCONER; MACKAY, 1996). Para Chaves (2001), a interação (G×E) deve ser encarada como um fenômeno biológico em suas implicações no melhoramento de plantas e não como um simples efeito estatístico, cumprindo buscar a explicação evolutiva do evento se se quiser tirar proveito de seus efeitos benéficos indesejáveis sobre a avaliação de genótipos e recomendação de cultivares. Diferenças em adaptação de genótipos em populações resultam, evidentemente, de diferenças de constituição gênica para os caracteres importantes nesta adaptação. A reação diferencial às mudanças ambientais pode-se dar desde os mecanismos de regulação gênica até caracteres morfológicos finais.

Segundo Duarte e Vencovsky (1999) a interação (G×E) representa uma das principais dificuldades encontradas pelo melhorista durante sua atividade seletiva. Nas etapas preliminares desse processo (com avaliações normalmente em uma só localidade), a interação (G×E) pode inflacionar as estimativas da variância genética, resultando em superestimativas dos ganhos genéticos esperados com a seleção (ganhos reais inferiores aos previstos). Nas fases finais, via de regra, os ensaios são conduzidos em vários ambientes (locais, anos e/ou épocas), o que possibilita o isolamento daquele componente de variabilidade; muito embora, neste momento, a intensidade de seleção seja baixa, o que já minimizaria seus efeitos sobre previsões de ganho genético. Por outro lado, a presença dessa interação, na maioria das vezes, faz com que os melhores genótipos em um determinado local não o sejam em outros. Isso dificulta a recomendação de genótipos (cultivares) para toda a população de ambientes amostrada pelos testes. Estatisticamente isso decorre da impossibilidade de interpretar, de forma aditiva, os efeitos principais de genótipos e de ambientes (KANG; MAGARI, 1996).

Cockerham (1963) atribuiu o aparecimento de interações (G×E) como sendo devido a respostas diferenciais do mesmo conjunto gênico em ambientes distintos ou pela expressão de diferentes conjuntos gênicos em diferentes ambientes. Quando um mesmo conjunto de genes se expressa em diferentes ambientes, as diferenças nas respostas podem ser explicadas pela heterogeneidade das variâncias genéticas e experimentais ou por ambas, e , quando diferentes

conjuntos de genes se expressam em ambientes distintos, as diferenças nas respostas explicam-se pela inconsistência das correlações genéticas entre os valores de um mesmo caráter em dois ambientes (FALCONER, 1989). Segundo Cruz e Regazzi (1994), a interação (G×E) também pode surgir em função de fatores fisiológicos e bioquímicos próprios de cada genótipo cultivado. Chaves et al. (1989) relatam ainda que a falta de ajuste do modelo estatístico adotado ao conjunto de dados pode ser uma das causas da interação (G×E) significativa.

Várias metodologias têm sido propostas no sentido de entender melhor o efeito da interação (G×E). Algumas dessas propostas são: zoneamento ecológico ou estratificação de ambientes, ou seja, identificar regiões ou sub-regiões onde o efeito da interação seja não significativo, pode levar a identificação de genótipos que se adaptam a ambientes específicos e ainda identificar genótipos com uma ampla adaptação ou estabilidade (RAMALHO et al., 1993). Da importância dessa interação no campo experimental devem-se escolher os métodos estatísticos que melhor expliquem a informação contida nos dados, um daqueles métodos é o modelo de interação multiplicativa, também conhecido como o modelo de efeitos principais aditivos e interação multiplicativa - AMMI.

2.2 O Modelo de interação multiplicativa (ou Modelo AMMI)

Em geral, um modelo de efeitos principais aditivos e interação multiplicativa (AMMI) pode ser útil para qualquer conjunto de dados provenientes de experimentos com dois fatores de classificação cruzada e é muito apropriado em certas situações descritas por Milliken e Johnson (1989), como por exemplo:

- Quando a interação estiver presente no modelo, mas não existirem diferenças nos tratamentos das linhas, nem nos tratamentos das colunas.
- Quando a interação estiver presente em uma só casela. Esse pode ser o caso no qual a observação seja um dado discrepante, que também pode ocorrer se uma combinação particular de tratamentos dá resultados muito raros quando for aplicada na unidade experimental (tratamentos de controle). A combinação de dois tratamentos de controle pode causar a interação nos dados e um simples modelo aditivo não pode ser ajustado.
- Quando toda a interação estiver em uma só linha (ou coluna). Isto pode ocorrer quando houverem vários dados discrepantes na mesma linha (ou coluna).

O modelo AMMI é uma boa alternativa de análise, pois esses modelos ajudam a interpretação e melhor compreensão do fenômeno da interação de fatores, problema que se

encontra presente no melhoramento genético de plantas, especificamente no estudo da interação genótipo por ambiente (G×E). Vários autores afirmam que esta metodologia é melhor do que os métodos baseados em regressão. Crossa (1990 apud DUARTE; VENCOSVSKY, 1999) argumenta que a análise de regressão linear não é informativa se a linearidade falhar e depende do grupo de genótipos e ambientes incluídos e tende a simplificar modelos de resposta, explicando a variação devida à interação em uma única dimensão, quando na realidade ela pode ser bastante complexa. Esses procedimentos em geral, não informam sobre interações específicas de genótipos com ambientes (se positivas ou negativas), dificultando explorar vantajosamente os efeitos da interação. É por isso, que Crossa (1990 apud DUARTE; VENCOSVSKY, 1999) sugere a aplicação de métodos multivariados como a análise de componentes principais (ACP), a análise de agrupamentos e o procedimento AMMI.

O modelo AMMI combina dois procedimentos estatísticos: análise da variância e a decomposição por valores singulares. Em um único modelo tem-se componentes aditivos para os efeitos principais (linhas ou genótipos e colunas ou ambientes), e componentes multiplicativos para os efeitos da interação. Duarte e Vencovsky (1999) explicam que os efeitos principais, na parte aditiva (média, efeitos genotípicos e ambientais), são ajustados por uma análise de variância comum (univariada) aplicada à matriz de dados, resultando em um resíduo de não aditividade, isto é, na interação (G×E), e essa interação, constituinte da parte multiplicativa do modelo, é, depois, analisada pela decomposição por valores singulares da matriz de resíduos ou interação. Em seguida se apresenta o modelo AMMI de uma forma geral para dois fatores (T e B) de acordo com Milliken e Johnson (1989), nesse caso um fator pode corresponder a genótipos, por exemplo o fator T, e o outro fator B pode corresponder aos ambientes.

Seja μ_{ij} a resposta esperada quando os níveis de tratamentos T_i e B_j são aplicados em uma unidade experimental dada, em que $i = 1, 2, ..., t$ e $j = 1, 2, ..., b$. Usando os resultados da decomposição de matrizes, é possível mostrar que qualquer matriz de dimensão $t \times b$ das μ_{ij} sempre pode ser decomposta da seguinte maneira:

$$\begin{aligned} \mu_{ij} &= \mu + \tau_i + \beta_j + \lambda_1\alpha_{1i}\gamma_{1j} + \lambda_2\alpha_{2i}\gamma_{2j} + \cdots + \lambda_k\alpha_{ki}\gamma_{kj} \\ & i = 1, 2, ..., t. \quad j = 1, 2, ..., b. \end{aligned} \tag{1}$$

em que μ representa a média geral, τ_i é o efeito do i-ésimo genótipo, β_j é o efeito do j-ésimo ambiente, $k = posto(\mathbf{\Omega})$, $\mathbf{\Omega} = (w_{ij})$ em que $w_{ij} = \mu_{ij} - \overline{\mu}_{i\bullet} - \overline{\mu}_{\bullet j} + \overline{\mu}_{\bullet\bullet}$, λ_r (com $r = 1, ..., k$) é a raiz quadrada do r-ésimo autovalor das matrizes $\mathbf{\Omega}\mathbf{\Omega}^T$ e $\mathbf{\Omega}^T\mathbf{\Omega}$ de iguais autovalores não

nulos, α_{ri} é o i-ésimo elemento (relacionado ao genótipo i) do r-ésimo autovetor de $\mathbf{\Omega\Omega}^T$ associado a λ_r^2, γ_{rj} é o j-ésimo elemento (relacionado ao ambiente j) do r-ésimo autovetor de $\mathbf{\Omega}^T\mathbf{\Omega}$ associado a λ_r^2.

Para reduzir o número de possíveis valores para os parâmetros em (1) e sintetizar a apresentação do modelo de efeitos principais aditivos e interação multiplicativa, é assumido sem perda de generalidade que:

$$\begin{aligned} \sum_i \tau_i = \sum_j \beta_j = 0 \\ |\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_k| \\ \sum_i \alpha_{ri} = \sum_j \gamma_{rj} = 0 \text{ para } r = 1, 2, ..., k \\ \sum_i \alpha_{ri}^2 = \sum_j \gamma_{rj}^2 = 1 \text{ para } r = 1, 2, ..., k \\ \sum_i \alpha_{ri}\alpha_{r^*i} = \sum_i \gamma_{ri}\gamma_{r^*i} = 0 \text{ para } r \neq r^* = 1, 2, ..., k \end{aligned} \tag{2}$$

Nesse modelo, qualquer contraste nas médias μ_{ij}, o qual mede a interação, pode ser escrito como uma combinação linear dos w_{ij}. Também se tem que $k \leq \min(b-1, t-1)$.

Agora, em uma situação real, dado um conjunto de dados experimentais em uma tabela de dupla entrada com uma observação por casela y_{ij} (essa observação pode ser a média das repetições de cada tratamento em um delineamento balanceado), podem-se considerar modelos da seguinte maneira:

$$\begin{aligned} y_{ij} = \mu + \tau_i + \beta_j + \lambda_1\alpha_{1i}\gamma_{1j} + \lambda_2\alpha_{2i}\gamma_{2j} + \cdots + \lambda_k\alpha_{ki}\gamma_{kj} + \epsilon_{ij} \\ i = 1, 2, ..., t. \quad j = 1, 2, ..., b. \end{aligned} \tag{3}$$

em que é assumido que os parâmetros satisfazem as restrições dadas em (2) e que os erros ϵ_{ij} são independentemente distribuídos com média zero e variância comum σ^2. Para obter os resultados dos testes de hipóteses é necessário também assumir normalidade nos erros. Dependendo do número de componentes multiplicativos o modelo (3) é notado por AMMI0, AMMI1 ou AMMIk de forma genérica.

Um conjunto de estimativas de mínimos quadrados dos parâmetros nesse modelo é

$$\begin{aligned}
\widehat{\mu} &= \frac{\sum_{i,j} y_{ij}}{(t \times b)} = \overline{y}_{\bullet\bullet} \\
\widehat{\tau}_i &= \frac{\sum_{j} y_{ij}}{b} - \widehat{\mu} = \overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet} \qquad i = 1, 2, ..., t. \\
\widehat{\beta}_j &= \frac{\sum_{i} y_{ij}}{t} - \widehat{\mu} = \overline{y}_{\bullet j} - \overline{y}_{\bullet\bullet} \qquad j = 1, 2, ..., b. \\
\widehat{\lambda}_r^2 &= l_r \qquad\qquad\qquad r = 1, 2, ..., k.
\end{aligned} \tag{4}$$

em que $l_1 > l_2 > \cdots > l_k > l_{k+1} > \cdots > l_p$ são os autovalores não nulos de $\mathbf{Z}^T\mathbf{Z}$ ou $(\mathbf{Z}\mathbf{Z}^T)$,

$$\begin{aligned}
&\mathbf{Z} = (z_{ij}). \\
&z_{ij} = y_{ij} - \overline{y}_{i\bullet} - \overline{y}_{\bullet j} + \overline{y}_{\bullet\bullet}. \\
&p = \min(b-1, t-1). \\
&\widehat{\boldsymbol{\alpha}}_r : \text{autovetor normalizado de } \mathbf{Z}\mathbf{Z}^T \text{ correspondente ao autovalor não nulo, } l_r,\ r = 1, 2, ..., k. \\
&\widehat{\boldsymbol{\gamma}}_r : \text{autovetor normalizado de } \mathbf{Z}^T\mathbf{Z} \text{ correspondente ao autovalor não nulo, } l_r,\ r = 1, 2, ..., k.
\end{aligned} \tag{5}$$

O sinal adequado para $\widehat{\lambda}_r$ pode ser obtido assim,

$$\widehat{\lambda}_r = sinal\left(\widehat{\boldsymbol{\alpha}}_r^T \mathbf{Z} \widehat{\boldsymbol{\gamma}}_r\right), r = 1, 2, ..., k.$$

Além dos erros independentes, se eles tiverem distribuição normal, então os estimadores acima descritos são também os estimadores de máxima verossimilhança. Em experimentos G×E, y_{ij} representa a resposta do i-ésimo genótipo no j-ésimo ambiente, μ é a média geral, τ_i e β_j os efeitos genotípicos e ambientais, λ_r^2 fornece a proporção da variância devido à interação G×E no r-ésimo componente e α_{ri}, γ_{rj} representam os pesos para o genótipo i e ambiente j naquele componente de interação.

2.2.1 Escolha do número apropriado de termos para descrever a interação

Na literatura existem dois tipos de procedimentos para determinar o número ótimo de termos no modelo AMMI (o k no modelo (3)). Um desses procedimentos consiste em fazer testes de significância dos termos multiplicativos e o outro procedimento consiste em fazer

validação cruzada. Na validação cruzada os dados de repetições, para cada combinação de tratamentos são aleatoriamente divididos em dois subconjuntos, um subconjunto de dados para o ajuste do modelo e outro subconjunto para validação. As respostas preditas por um determinado modelo AMMI, são confrontadas com os respectivos dados de validação, calculando-se as diferenças entre esses valores. Obtém-se, em seguida a soma de quadrados dessas diferenças, dividindo-se o resultado pelo número delas. A raiz quadrada desse resultado chama-se diferença preditiva média. Esse método foi estudado com mais detalhes em Duarte e Vencovsky (1999) e Dias (2005). Neste trabalho por causa da estrutura dos dados experimentais, só serão considerados testes estatísticos sobre os componentes multiplicativos, pois não se dispõe das repetições das observações e apenas está disponível a matriz de médias.

Os testes de hipóteses se fazem usando os dados completos, e os critérios adotados para a determinação do número de componentes multiplicativos tem sido objeto de várias pesquisas como as desenvolvidas por Gollob (1968), Mandel (1971), Gauch (1988) e Gauch e Zobel (1988 apud KANG; GAUCH, 1996) entre outros. Mas, levando em conta os resultados e recomendações dos estudos feitos por Milliken e Johnson (1989), Piepho (1995a), Cornelius et al. (1996), Dias e Krzanowski (2003), Dias (2005) e Dias e Krzanowski (2006), os métodos propostos para usar neste trabalho serão os testes de razão de verossimilhança e o teste F_R, os quais serão descritos a seguir:

Testes de razão de verossimilhança. Se um pesquisador deseja usar o modelo dado em (1), então é preciso determinar o número de termos de interação multiplicativa necessário para que o modelo explique adequadamente os dados. Para tomar essa decisão, têm-se problemas muito parecidos aos problemas encontrados na construção dos modelos de regressão. O objetivo é encontrar um modelo parcimonioso (com poucos termos quanto seja possível), e ao mesmo tempo, obter um modelo adequado. Deve-se lembrar que em situações de regressão polinomial, sempre é possível ajustar um modelo de grau $(n-1)$ a n observações, mas, tais modelos não são geralmente bons, pois funcionam bem na predição da resposta média dos valores observados, mas, podem funcionar muito mal na predição de respostas de valores não observados.

Uma situação similar se apresenta para os dados provenientes de uma estrutura de tratamentos em dupla entrada, na qual sempre é possível fazer $k = \min(b-1, t-1)$ e ajustar exatamente esses dados com o modelo (1), mas, tal modelo não será provavelmente muito bom por causa do sobreajuste dos dados e pelo fato dele explicar além do padrão de resposta presente nos dados parte do erro de medida. Assim, é desejável um modelo com poucos

componentes que ofereça um ajuste ótimo e que explique boa parte do padrão de resposta dos dados.

Assumindo por enquanto que se conhece o procedimento necessário para testar as hipóteses, um procedimento razoável pode ser:

1. Testar $H_{01} : \lambda_1 = 0$ vs. $H_{a1} : \lambda_1 \neq 0$ no modelo

$$y_{ij} = \mu + \tau_i + \beta_j + \lambda_1 \alpha_{1i} \gamma_{1j} + \epsilon_{ij} \tag{6}$$

2. Se não rejeitar H_{01}, conclui-se que os dados são aditivos e deve-se completar uma análise correspondente com o resultado. No entanto, se rejeita-se H_{01}, então se deve testar

$$H_{02} : \lambda_2 = 0 \ , \ \lambda_1 \neq 0 \ \ vs. \ \ H_{a2} : \lambda_2 \neq 0 \ , \ \lambda_1 \neq 0$$

no modelo

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

3. Se não rejeitar H_{02}, conclui-se que o modelo apropriado é o modelo dado em (6) e deve-se completar a análise de acordo com o resultado encontrado. Se rejeitar H_{02}, então se deve testar

$$H_{03} : \lambda_3 = 0 \ , \ \lambda_2 \neq 0, \ \lambda_1 \neq 0 \ \ vs. \ \ H_{a3} : \lambda_3 \neq 0 \ , \ \lambda_2 \neq 0, \ \lambda_1 \neq 0$$

no modelo

$$y_{ij} = \mu + \tau_i + \beta_j + \lambda_1 \alpha_{1i} \gamma_{1j} + \lambda_2 \alpha_{2i} \gamma_{2j} + \lambda_3 \alpha_{3i} \gamma_{3j} + \epsilon_{ij}$$

4. Continuar dessa forma sucessivamente até não rejeitar a hipótese.

Uma desvantagem do procedimento acima descrito é parecida à desvantagem encontrada no procedimento de escolha sequencial desenvolvido para problemas de regressão múltipla. Por exemplo, um termo da interação pode explicar apenas uma pouca proporção da variação da interação e por tal razão esse termo pode não ser significativo. Assim, é preferível usar um procedimento parecido ao explicado acima, mas com uma pequena diferença. Podem-se testar hipóteses sucessivas, depois das quais é possível concluir que o valor certo de k é o valor para o qual foi feita a última rejeição. Segundo Milliken e Johnson (1989), em geral o modelo de interação multiplicativa em aplicacões sobre dados reais precisa de um máximo de dois componentes e em muitas ocasiões apenas um termo da interação é necessário.

Um teste de razão de verossimilhança para H_{01} versus H_{a1}, pode fazer-se rejeitando H_{01} se

$$U_1 = \frac{l_1}{\sum_{ij} z_{ij}^2} > C_\alpha$$

em que C_α é o ponto crítico a $\alpha(100\%)$ obtido da tabela com os pontos críticos da distribução de U_1 e apresentada em Milliken e Johnson (1989) (p.171) com $p = \min(t-1, b-1)$ e $n = \max(t-1, b-1)$. l_1, z_{ij} foram definidos na página 17 e t, b definidos na página 3. Note-se também que U_1 é igual a

$$U_1 = \frac{l_1}{l_1 + l_2 + \cdots + l_p}.$$

Agora, uma estatística de razão de verossimilhança para testar H_{02} versus H_{a2} é

$$U_2 = \frac{l_2}{l_2 + l_3 + \cdots + l_p}.$$

Rejeita-se H_{02} se U_2 for maior do que o valor crítico da tabela apresentada em Milliken e Johnson (1989) (p.173). Nessa tabela encontram-se os pontos críticos aproximados da distribuição de U_2.

Em geral, as estatísticas de razão de verossimilhança para a hipótese H_{0k} versus H_{ak}, $k = 3, 4, ..., p-1$, são dadas por:

$$U_k = \frac{l_k}{l_k + l_{k+1} + \cdots + l_p}.$$

Milliken e Johnson (1989) sugerem nestes casos usar os pontos críticos da distribuição de U_1 com $p = \min(t, b) - k$ e $n = \max(t, b) - k$, mas, para aqueles experimentos nos quais não existe na tabela o correspondente ponto crítico (grande número de genótipos ou ambientes) Cornelius et al. (1996) apresentam uma transformação da estatística para obter um teste F aproximado. Para testar o $k - ésimo$ termo, o teste F_{teste} aproximado pode ser encontrado como segue:

$$
\begin{aligned}
p &= \min(t-1, b-1); \;\; n = \max(t-1, b-1), \\
Q_k &= [(p-k+1)U_k - 1]/(p-k); \;\; c_1^* = \frac{u_{1k} - (n-k+1)}{(n-k+1)\,(p-k)}; \\
c_2^* &= \frac{(p-k+1)\,(n-k+1)\,u_{2k}^2 - 2u_{1k}^2}{(n-k+1)^2\,[(p-k+1)\,(n-k+1)+2]\,(p-k)^2}; \\
d^* &= c_1^*\,(1-c_1^*) - c_2^*; \;\; a^* = dc_1^*/c_2^*; \quad b^* = d^*\,(1-c_1^*)\,/c_2^*; \\
F_{teste} &= b^* U_k / a^*\,(1-U_k)\,.
\end{aligned}
$$

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

$$
\begin{aligned}
u_{1k} = E_{r,c}\left(l_1/\sigma^2\right) = &-0,64679880 + 1,0068336\,(r+c) - 7,1495083 \times 10^{-9} r^2 c^2 \\
&+ 0,082395238\left[(\ln r)^2 + (\ln c)^2\right] + 0,53767438\,(\ln r \ln c)\,(\ln r + \ln c) \\
&- 0,091580971\,(\ln r \ln c)\left[(\ln r)^2 + (\ln c)^2\right] \\
&+ 0,021644307\,(\ln r \ln c)\left[(\ln r)^3 + (\ln c)^3\right] \\
&+ 5,3529799 \times 10^{-3}\,(\ln r)^3\,(\ln c)^3 - 0,76227733\left(e^{-r} + e^{-c}\right) \\
&- 0,020829655\,(r/c + c/r) + 1,7482806 \times 10^{-3}\,(rc - |r-c|)
\end{aligned}
$$

$$
\begin{aligned}
u_{2k} = DP_{r,c}\left(l_1/\sigma^2\right) = &-0,015802857\,(r+c) + 2,3780161 \times 10^{-9} rc\left(r^2 + c^2\right) \\
&+ 1,7371131\,(\ln r + \ln c) - 0,33301620\left[(\ln r)^2 + (\ln c)^2\right] \\
&+ 0,11442045\left[(\ln r)^3 + (\ln c)^3\right] - 0,035296928\,(\ln r \ln c)\,[\ln r + \ln c] \\
&+ 0,033246016\,(r/c + c/r) - 5,7298685 \times 10^{-9}\left[(r/c)^4 + (c/r)^4\right] \\
&+ 1,8747692\left(e^{-r} + e^{-c}\right) - 1,7476731 \times 10^{-13} r^2 c^2\left(r^2 + c^2\right) \\
&+ 6,9946263 \times 10^{-16} r^3 c^3\,(r+c) + 2,3238523 \times 10^{-5}\,|r-c|^2
\end{aligned}
$$

em que $r = \max(t, b) - k$ e $c = \min(t, b) - k$.

Teste $\mathbf{F_R}$. Considere-se o modelo (7) aplicado em um experimento para avaliar genótipos e ambientes, τ_i representa os genótipos e β_j representa os ambientes, onde os componentes multiplicativos representam a interação desses fatores (G×E).

$$y_{ij} = \mu + \tau_i + \beta_j + \lambda_1\alpha_{1i}\gamma_{1j} + \lambda_2\alpha_{2i}\gamma_{2j} + \cdots + \lambda_k\alpha_{ki}\gamma_{kj} + \epsilon_{ij} \quad (7)$$
$$i = 1, 2, ..., t. \quad j = 1, 2, ..., b.$$

Escrevendo a soma de quadrados da interação (G×E) tem-se que

$$SQ\left(G \times E\right) = \sum_i \sum_j \left(y_{ij} - \overline{y}_{i\bullet} - \overline{y}_{\bullet j} + \overline{y}_{\bullet\bullet}\right)^2 \text{ com } \left(t-1\right)\left(b-1\right) \text{ graus de liberdade.}$$

Além do anterior, essa soma de quadrados pode ser escrita como,

$$SQ(G \times E) = \sum_{r=1}^{p} l_r$$

em que $l_1 > l_2 > \cdots > l_k > l_{k+1} > \cdots > l_p$ são os autovalores de $\mathbf{Z}^T\mathbf{Z}$ ou $(\mathbf{Z}\mathbf{Z}^T)$, $p = \min(t-1, b-1)$ e a matriz $\mathbf{Z}$ está definida pelos componentes $z_{ij} = y_{ij} - \overline{y}_{i\bullet} - \overline{y}_{\bullet j} + \overline{y}_{\bullet\bullet}$.

Então, a idéia é escolher o melhor k, de tal maneira que a soma de quadrados da interação possa ser separada em uma parte determinística (padrão) e outra parte que conterá ruído, assim,

$$SQ\left(G\times E\right) = \sum_{r=1}^{p} l_r = \sum_{r=1}^{k} l_r + \sum_{r=k+1}^{p} l_r = SQ(G \times E)_{PADRÃO} + SQ(G \times E)_{RUÍDO}.$$

A estatística teste de Cornelius reescrita por Dias e Krzanowski (2003) com k termos multiplicativos no modelo é dada por,

$$F_{R,k} = \frac{\left(SQ(G \times E) - \sum_{r=1}^{k} l_r\right)}{\left(f_2 * QM\left(Erro\ médio\right)\right)}$$

com $f_2 = \left(t-1-k\right)\left(b-1-k\right)$. O erro médio, segundo Duarte e Vencovsky (1999) é originário das análises individuais de variância dos b experimentos. Este é o teste F_R de Cornelius et al. (1996) que sob a hipótese nula de que não mais que k termos determinam a interação, de tal forma que o teste estatístico tem uma distribuição F com f_2 gl e os graus de liberdade do quadrado médio do resíduo. Um resultado significativo para o teste sugere que no mínimo um ou mais termos multiplicativos devem ser adicionados aos k já incluídos.

Apresenta-se na tabela 1 a análise da variância a partir de médias segundo o sistema de

Cornelius, em que n representa o número de repetições no experimento e IPCAi é a notação internacional para o i-ésimo componente da interação.

Tabela 1 - Esquema da análise de variância pelo sistema de Cornelius baseado em médias

Fonte de variação	**Graus de liberdade**	**Somas de quadrados**
Genótipos (G)	$t-1$	SQ(G)
Ambientes (E)	$b-1$	SQ(E)
Interação (G×E)	$(t-1)(b-1)$	SQ(G×E)
IPCA1	$(t-1-1)(b-1-1)$	$\sum_{r=2}^{p} l_r$
IPCA2	$(t-1-2)(b-1-2)$	$\sum_{r=3}^{p} l_r$
IPCA3	$(t-1-3)(b-1-3)$	$\sum_{r=4}^{p} l_r$
IPCAp	$-$	$-$
Erro médio	$b(t-1)(n-1)$	
Total	$tbn-1$	

2.3 Métodos para a imputação de dados faltantes

O problema de dados ausentes não é novo na literatura para experimentos estatisticamente planejados, e duas abordagens diferentes são usadas para resolver essa situação. A primeira abordagem e talvez a mais comum consiste em preencher as caselas vazias utilizando métodos de imputação para depois fazer a respectiva análise. Uns dos primeiros autores em considerar esta opção foram Allan e Wishart (1930 apud DODGE, 1985), que desenvolveram uma fórmula para imputar dados faltantes em um delineamento aleatorizado em blocos e em um delineamento em quadrado latino. Depois, Yates (1933 apud DODGE, 1985) mostrou que os valores certos para preencher as caselas em um delineamento por blocos deveriam ser escolhidos com o critério da minimização da soma de quadrados do erro, estabelecendo-se como um dos critérios mais referenciados nos livros clássicos de experimentos, como o de Cochran e Cox (1957). O método de Yates (1933 apud DODGE, 1985) foi levado para outros delineamentos por Cornish (1940, 1941 apud DODGE, 1985) considerando várias observações ausentes. Anos depois, Healy e Westmacott (1956 apud LITTLE; RUBIN, 2002) propuseram um método iterativo para imputar os dados faltantes, com esse método, valores iniciais são inseridos, a análise é feita e então, de cada casela com dado ausente é calculado o resíduo e é substraído dos valores iniciais. Uma vez que os valores perdidos são imputados, a análise completa é feita incluíndo as imputações e os graus de liberdade da soma de quadrados do erro são diminuídos pelo número de dados que foram imputados inicialmente. Logo depois, Hartley (1956) apresentou uma alternativa para quando se tem uma só observação faltante,

mas recomendou usar um método iterativo quando se tem mais de uma observação faltante. Outro método muito conhecido foi o proposto por Bartlett (1937 apud DODGE, 1985), este método é não iterativo, teóricamente obtém os mesmos resultados do método de Yates (1933 apud DODGE, 1985), mas leva em conta uma técnica que às vezes ajuda na imputação de dados; a análise de covariância. Outros autores que historicamente fizeram seu aporte neste assunto foram Rubin (1972,1976) e Li (1982 apud DODGE, 1985).

Outra abordagem para resolver o problema de dados faltantes, diferente da imputação, consiste em fazer modelos e considerações teóricas sobre as funções estimáveis baseado unicamente nos dados presentes do experimento. Uma completa revisão dessa abordagem foi feita por Birkes et al. (1976), Dodge (1985) e Dodge e Zoppe (2004). Mais recentemente no tratamento para os dados faltantes Rubin (1996), Little e Rubin (2002) e Schafer (2002) recomendam usar métodos parâmétricos de imputação múltipla ou métodos baseados em verossimilhança (como por exemplo o algoritmo EM) ao invés dos métodos que usavam modelos lineares, como os procedimentos citados acima.

O problema dos dados faltantes também foi tratado especificamente para os experimentos que envolvem este estudo, quer dizer os experimentos de genótipos e ambientes. Um trabalho aceito nestes experimentos foi o desenvolvido por Gauch e Zobel (1990), que fizeram a imputação através do uso do algoritmo EM e o modelo AMMI, mas algumas alternativas desse procedimento usando estatística multivariada foram descritas em Goodfrey et al. (2002). Outros estudos que são recomendados por Van Eeuwijk e Kroonenberg (1998) no caso de observações faltantes para experimentos (G×E) com resultados razoavelmente bons, foram os desenvolvidos por Denis e Baril (1992) e Calinski et al. (1992). Eles encontraram que usando imputações através dos mínimos quadrados alternados (usando modelos robustos e modelos com estrutura de interação bilinear) podem-se obter resultados tão bons como os encontrados com o algoritmo EM. Recentemente, Bergamo (2007) propôs um excelente método baseado em imputação múltipla livre de distribuição, o qual pode ser aplicado en matrizes de interação G×E com observações ausentes para resolver o problema de estimação de parâmetros de um modelo AMMI. Dada a informação histórica sobre imputação de dados em experimentos e especificamente em experimentos de dois fatores multiplicativos (G×E) decidiu-se fazer uma comparação das metodologias propostas por Bergamo (2007), Bergamo et al. (2008), Piepho (1995b), Calinski et al. (1992) e Denis e Baril (1992), as quais são apresentadas a seguir.

2.3.1 Imputação Múltipla Livre de Distribuição (IMLD)

Qualquer matriz $\mathbf{Y}$ de dimensão $(n \times p)$ pode ser decomposta por valor singular na forma

$$\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{8}$$

em que $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_p$ e $\mathbf{D} = diag(d_1, d_2, ..., d_p)$ com $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$. As matrizes $\mathbf{Y}^T\mathbf{Y}$ e $\mathbf{Y}\mathbf{Y}^T$ têm os mesmos autovalores não nulos, e os elementos d_i são a raiz quadrada destes autovalores; a i-ésima coluna $\mathbf{v}_i = (v_{i1}, v_{i2}, ..., v_{ip})^T$ da matriz $\mathbf{V}_{(p\times p)}$ é o autovetor correspondente ao i-ésimo maior autovalor d_i^2 de $\mathbf{Y}^T\mathbf{Y}$; enquanto a j-ésima coluna $\mathbf{u} = (u_{1j}, u_{2j}, ..., u_{nj})^T$ da matriz $\mathbf{U}_{(n\times p)}$ é o autovetor correspondente ao j-ésimo maior autovalor d_j^2 de $\mathbf{Y}\mathbf{Y}^T$. A decomposição (8) tem uma representação elementar

$$y_{ij} = \sum_{h=1}^{p} u_{ih} d_h v_{jh}. \tag{9}$$

Segundo Krzanowski (1988), essa representação pode ser usada para determinar a dimensionalidade de um conjunto de dados multivariados. Se a estrutura dos dados é esencialmente H-dimensional $(H < p)$ então a variação na dimensão resultante $(p - H)$ pode ser tratada como um ruído aleatório. As características principais dos dados estarão supostamente no espaço dos H primeiros componentes principais. A correspondência entre as quantidades do lado direito de (9) e os eixos principais da configuração dos dados sugere o modelo de H componentes

$$y_{ij} = \sum_{h=1}^{H} u_{ih} d_h v_{jh} + \epsilon_{ij}, \tag{10}$$

em que ϵ_{ij} é o ruído.

Supondo o modelo (10) para um valor específico de H, com uma única observação y_{ij} ausente na matriz de dados, tem-se y_{ij} predito por

$$\widehat{y}_{ij}^{(H)} = \sum_{h=1}^{H} u_{ih} d_h v_{jh}, \tag{11}$$

em que u_{ih}, d_h, v_{jh}, devem ser preditos com o restante dos dados. As melhores imputações destes valores estão baseadas na maior quantidade possível de dados. Simbolizando, por $\mathbf{Y}^{(-i)}$ a matriz de dados obtida, retirando-se a i-ésima linha de $\mathbf{Y}$, e por $\mathbf{Y}_{(-j)}$ a matriz dos dados obtida, retirando-se a j-ésima coluna de $\mathbf{Y}$, a decomposição por valor singular dessas

matrizes fica

$$\mathbf{Y}^{(-i)} = \overline{\mathbf{U}}\,\overline{\mathbf{D}}\,\overline{\mathbf{V}}^T,\ \overline{\mathbf{U}} = (\overline{u}_{sh}),\ \overline{\mathbf{V}} = (\overline{v}_{sh}),\ \overline{\mathbf{D}} = (\overline{d}_1, ..., \overline{d}_p) \tag{12}$$

e

$$\mathbf{Y}_{(-j)} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}^T,\ \widetilde{\mathbf{U}} = (\widetilde{u}_{sh}),\ \widetilde{\mathbf{V}} = (\widetilde{v}_{sh}),\ \widetilde{\mathbf{D}} = (\widetilde{d}_1, ..., \widetilde{d}_{p-1}) \tag{13}$$

A estimativa de u_{ih} e v_{jh} em (11), obtida com o máximo dos dados de $\mathbf{Y}$ é $\widetilde{u}_{ih}$ e $\overline{v}_{jh}$, respectivamente, enquanto d_h pode ser estimado por $\overline{d}_h$, $\widetilde{d}_h$ ou por alguma combinação dos dois. Uma forma adequada pode ser $\sqrt{\overline{d}_h}\sqrt{\widetilde{d}_h}$ em que uma imputação do valor ausente y_{ij} é dada por:

$$\widehat{y}_{ij}^{(H)} = \sum_{h=1}^{H} \left(\widetilde{u}_{ih}\sqrt{\widetilde{d}_h}\right)\left(\overline{v}_{jh}\sqrt{\overline{d}_h}\right).$$

Usando o valor mais elevado de H, de (13), este valor é $p-1$, então o valor imputado a y_{ij} será

$$\widehat{y}_{ij} = \sum_{h=1}^{p-1} \left(\widetilde{u}_{ih}\sqrt{\widetilde{d}_h}\right)\left(\overline{v}_{jh}\sqrt{\overline{d}_h}\right). \tag{14}$$

As imputações iniciais dos valores y_{ij} ausentes são feitas pela média $\overline{y}_j$ da j-ésima coluna. Para evitar qualquer influência de possíveis variações entre as colunas, por exemplo, a escala das variáveis, é recomendado aplicar uma padronização em $\mathbf{Y}$. Para os valores y_{ij}, inclusive os ausentes já substituídos pela média $(\overline{y}_j)$, é calculada uma nova média $(\overline{y}_j^*)$ e um desvio padrão (s_j) para cada coluna j, então y_{ij} é padronizado por $y_{ij}^* = \frac{\left(y_{ij}-\overline{y}_j\right)}{s_j}$. Padronização semelhante é feita nas matrizes $\mathbf{Y}^{(-i)}$ e $\mathbf{Y}_{(-j)}$.

As imputações de cada valor ausente são recalculadas usando-se (14) nas matrizes padronizadas. Para cada estimativa são necessárias duas decomposições por valores singulares, isto é, uma para cada i e j necessários. O processo iterativo continua até ser alcançada a estabilidade nos valores imputados. Finalmente à matriz $\mathbf{Y}$ completada (observados + imputados) é aplicada uma operação para retorno à sua escala original, ou seja, se $y_{ij}^{(c)}$ representa cada valor da matriz $\mathbf{Y}$ completada, calcula-se novamente a média da coluna j $(\overline{y}_j^{(c)})$ e seu desvio padrão $(s_j^{(c)})$. Cada valor da matriz $\mathbf{Y}$ completada, na escala original, é então obtido por $y_{ij} = \overline{y}_j^{(c)} + s_j^{(c)} y_{ij}^{(c)}$.

Bergamo (2007) e Bergamo et al. (2008) apresentam uma proposta para gerar m imputações $(m = 1, 2, ..., S)$. A proposta consiste em mudar os expoentes dos radicandos $\overline{d}_h$

e $\widetilde{d}_h$ em (14), ou seja, de uma maneira genérica, se $\sqrt[b]{a}$ for representada como uma potência fracionária $d^{\frac{a}{b}}$, o procedimento requer a mudança no numerador do expoente, tanto de $\widetilde{d}_h^{\frac{\widetilde{a}}{b}}$ como de $\overline{d}_h^{\frac{\overline{a}}{b}}$, de modo que a soma dos expoentes seja igual a 1 $\left(\frac{\overline{a}+\widetilde{a}}{b}=1\right)$. Uma sugestão para as estimativas de d_h em (11) consiste em fazer uma combinação entre $\overline{d}_h$ de (12) e $\widetilde{d}_h$ de (13), resultando na forma $\sqrt{\overline{d}_h}\sqrt{\widetilde{d}_h}$, a qual admite influências iguais de (12) e (13) na estimativa final de y_{ij} em (14).

Cada mudança em $\widetilde{a}$ e, consequentemente em $\overline{a}$, gera uma nova matriz $\mathbf{Y}$ completada, caracterizando, assim, um processo de geração dos S conjuntos de dados completados. Segundo Bergamo et al. (2008) um número de $S=5$ é suficiente para expressar variabilidade entre imputações. Assim o autor propõe fazer as imputações através de

$$\widehat{y}_{ij}=\sum_{h=1}^{p-1}\left(\widetilde{u}_{ih}\widetilde{d}_h^{\frac{\widetilde{a}}{b}}\right)\left(\overline{v}_{jh}\overline{d}_h^{\frac{\overline{a}}{b}}\right), \tag{15}$$

então, por exemplo, para um denominador $b=20$ os valores assumidos por $\widetilde{a}$ serão $(8,9,10,11,12)$ e os valores assumidos por $\overline{a}$ $(12,11,10,9,8)$ lembrando que $\frac{\widetilde{a}+\overline{a}}{b}=1$.

Bergamo (2007) apresenta uma ampla variedade de análises multivariadas para testar a eficácia do método de imputação, mas neste trabalho somente considerar-se-a metodologia proposta para o modelo AMMI. Nesse modelo aconselha-se fazer as 5 imputações de acordo com (15) para cada valor ausente e logo obter um único valor para as análises posteriores proveniente da média dessas $S=5$ imputações.

2.3.2 Mínimos Quadrados Alternados (ALS) e estimativas com um sub-modelo robusto (r-AMMI)

Considere o modelo (16) para a análise da variância de dois fatores, que pode ser usado em dados de um experimento genótipo por ambiente (G×E)

$$y_{ij}=\mu+\tau_i+\beta_j+(\tau\beta)_{ij}+e_{ij} \tag{16}$$

$(i=1,...,t;\ j=1,...,b)$ em que μ representa a média geral, τ_i e β_j são os efeitos principais genotípicos e ambientais, $(\tau\beta)_{ij}$ representa a interação (G×E) e e_{ij} é o termo do erro associado ao i-ésimo genótipo e ao j-ésimo ambiente. Tem-se assumido que todos os efeitos são fixos, exceto o erro, que justifica impor as seguintes restrições de reparametrização

$$\sum_{i=1}^{t}(\tau\beta)_{ij} = \sum_{j=1}^{b}(\tau\beta)_{ij} = \sum_{i=1}^{t}\tau_i = \sum_{j=1}^{b}\beta_j = 0$$

Além disso, também assume-se que as interações têm uma estrutura bilinear,

$$(\tau\beta)_{ij} = \sum_{r=1}^{R}\theta_r\gamma_{ri}\alpha_{rj} \qquad R \leq \min(t-1, b-1)$$

em que θ_r, γ_{ri}, α_{rj} são estimados pela decomposição por valor singular (DVS) da matriz de resíduos depois de ajustar a parte aditiva. θ_r é estimado pelo r-ésimo valor singular da DVS, γ_{ri} e α_{rj} são estimados pelos valores dos autovetores genotípicos e ambientais correspondentes a θ_r. Os parâmetros γ_{ri} e α_{rj} satisfazem as seguintes condições

$$\begin{aligned}
\sum_{i=1}^{t}\gamma_{ri} &= \sum_{j=1}^{b}\alpha_{rj} = 0 \\
\sum_{i=1}^{t}(\gamma_{ri})^2 &= \sum_{j=1}^{b}(\alpha_{rj})^2 = 1, \text{ para qualquer } r \\
\sum_{i=1}^{t}\gamma_{ri}\gamma_{r^*i} &= \sum_{j=1}^{b}\alpha_{rj}\alpha_{r^*j} = 0 \quad \text{para qualquer } r \neq r^*.
\end{aligned}$$

Então o modelo pode ser escrito como:

$$\begin{aligned}
y_{ij} &= \mu + \tau_i + \beta_j + \sum_{r=1}^{R}\theta_r\gamma_{ri}\alpha_{rj} + e_{ij} \\
E(e_{ij}) &= 0,\ var(e_{ij}) = \sigma^2,\ cov(e_{ij}, e_{i^*j^*}) = 0
\end{aligned} \tag{17}$$

para cada (i, j) e $(i^*, j^*) \neq (i, j)$.

No caso dos dados incompletos os parâmetros do modelo (17) não podem ser estimados simultaneamente através de um simples procedimento de mínimos quadrados, mas, para resolver o problema pode ser aplicado um algoritmo iterativo. Em cada iteração os parâmetros de um tipo determinado são estimados, por exemplo, aqueles relacionados com os genótipos (subscrição i), e os outros parâmetros (correspondentes aos ambientes j) considerados como conhecidos. No algoritmo, de uma iteração para outra, o papel dos parâmetros é intercambiado e o procedimento cessa quando um nível de exatidão for alcançado.

No entanto, existem algumas limitações na aplicação deste algoritmo no modelo (17), as quais dependem da estrutura dos dados faltantes. Se ocorrem muitos dados faltantes e as caselas perdidas na estrutura de dupla entrada apresentam alguma forma particular, então a identificabilidade do modelo pode ser violada. Nesses casos segundo Calinski et al. (1992) a identificabilidade do modelo pode ser atingida através da diminuição do número R. Na literatura ainda não é muito comum encontrar condições necessárias e suficientes para este problema, mas é conhecido que cada linha e cada coluna devem ter pelo menos $(R+1)$ dados observados ou não faltantes. No presente trabalho R não será muito grande, levando em conta as conclusões dos trabalhos feitos por Milliken e Johnson (1989), Calinski et al. (1992), Denis e Baril (1992) e Piepho (1995b).

1) O CASO DE $R = 0$ ou ALS(0).

Quando $R = 0$, o modelo (17) se reduz ao modelo aditivo dos parâmetros μ, τ_i e β_j, o qual é um modelo linear e o procedimento de mínimos quadrados pode ser aplicado diretamente usando as matrizes de delineamento dos dados ausentes e dos dados presentes no experimento. Não obstante, nesta seção e nas seguintes apresenta-se como fazer as imputações dos dados faltantes através desse método iterativo.

a) Estimação inicial.

Apenas $\mu^{(0)}$ e $\beta_j^{(0)}$ são necessários. $\mu^{(0)}$ é igual a zero, enquanto $\beta_j^{(0)}$ é igual à média correspondente às observações presentes em cada ambiente $(j = 1, ..., b)$.

b) A h-ésima iteração.

i) A etapa linha.

Valores τ_i^* intermediários são calculados a partir dos valores $\mu^{(h-1)}$ e $\beta_j^{(h-1)}$. Para qualquer $i = 1, 2, ..., t$, α_i^* é obtido através da minimização da soma de quadrados

$$\sum_j \left[y_{ij} - \left(\mu^{(h-1)} + \tau_i^* + \beta_j^{(h-1)} \right) \right]^2 , \tag{18}$$

o somatório é feito somente sobre os valores y_{ij} presentes. Neste caso o procedimento corresponde ao cálculo de uma média.

ii) A etapa coluna.

Valores β_j^* são calculados a partir de $\mu^{(h-1)}$ e τ_i^* minimizando

$$\sum_i \left[y_{ij} - \left(\mu^{(h-1)} + \tau_i^* + \beta_j^*\right)\right]^2 \ , \tag{19}$$

o somatório é feito somente sobre os valores y_{ij} presentes, para qualquer $j = 1, 2, ..., b$.

iii) A etapa de normalização.

Finalmente, os valores dos parâmetros para a h-ésima iteração são obtidos impondo as restrições descritas acima na apresentação do modelo:

$$\begin{aligned}
\mu^{(h)} &= \mu^{(h-1)} + (t)^{-1}\sum_{i=1}^{t}\tau_i^* + (b)^{-1}\sum_{j=1}^{b}\beta_j^* \\
\tau_i^{(h)} &= \tau_i^* - (t)^{-1}\sum_{i=1}^{t}\tau_i^* \\
\beta_j^{(h)} &= \beta_j^* - (b)^{-1}\sum_{j=1}^{b}\beta_j^*
\end{aligned} \tag{20}$$

2) O CASO DE $R = 1$ ou ALS(1).

Neste caso ($R = 1$), apenas se tem um termo bilinear, θ_1, γ_{1i} e α_{1j}, os quais serão denotados como θ, γ_i, e α_j respectivamente.

a) A estimação inicial.

Somente $\mu^{(0)}$, $\beta_j^{(0)}$, $\theta^{(0)}$ e $\alpha_j^{(0)}$ são necessários. Os valores $\mu^{(0)}$ e $\beta_j^{(0)}$ são obtidos aplicando o modelo com $R = 0$, $\theta^{(0)}$ é igual a um e $\alpha_j^{(0)}$ é gerado pseudo-aletoriamente.

b) A h-ésima iteração.

Como no caso $R = 0$, cada iteração é decomposta em três etapas:

i) A etapa linha.

Valores de τ_i^* e γ_i^* são calculados a partir dos valores dos outros parâmetros obtidos na iteração $(h-1)$. Mais exatamente, τ_i^* e γ_i^* são dados para qualquer $i = 1, 2, ..., t$ através da minimização

$$\sum_j \left[y_{ij} - \left(\mu^{(h-1)} + \tau_i^* + \beta_j^{(h-1)} + \theta^{(h-1)} \gamma_i^* \alpha_j^{(h-1)} \right) \right]^2,$$

a soma é feita somente para os y_{ij} presentes.

ii) A etapa coluna.

Os valores intermediários β_j^* e α_j^* são calculados apartir dos valores calculados anteriormente. Mais exatamente, β_j^* e α_j^* são obtidos para qualquer $j = 1, 2, ..., b$ minimizando

$$\sum_j \left[y_{ij} - \left(\mu^{(h-1)} + \tau_i^* + \beta_j^* + \theta^{(h-1)} \gamma_i^* \alpha_j^* \right) \right]^2,$$

o somatório é feito sobre os valores y_{ij} presentes.

iii) A etapa de normalização.

Finalmente os valores dos parâmetros para a h-ésima iteração são obtidos impondo as restrições, mas é muito melhor usar o procedimento geral apresentado por Calinski et al. (1992), o qual é válido também quando $R > 1$.

Uma matriz completa $(t \times b)$ denotada por $\mathbf{M}$ é calculada de acordo com o modelo

$$M_{ij} = \mu^{(h-1)} + \tau_i^* + \beta_j^* + \theta^{(h-1)} \gamma_i^* \alpha_j^*. \tag{21}$$

Então $\mu^{(h)} = (tb)^{-1} \sum_{i=1}^{t} \sum_{j=1}^{b} M_{ij}$, $\tau_i^{(h)} = (b)^{-1} \sum_{j=1}^{b} M_{ij} - \mu^{(h)}$ para qualquer $i = 1, 2, ..., t$ e $\beta_j^{(h)} = (t)^{-1} \sum_{i=1}^{t} M_{ij} - \mu^{(h)}$ para qualquer $j = 1, 2, ..., b$.

Seja a matriz $\mathbf{U}$ de tamanho $(t \times b)$, tal que os elementos desta matriz são $U_{ij} = M_{ij} - \left(\mu^{(h)} + \tau_i^{(h)} + \beta_j^{(h)} \right)$. Isto é, uma parte da matriz $\mathbf{M}$, então $\boldsymbol{\gamma}^{(h)} = \left[\gamma_1^{(h)}, \gamma_2^{(h)}, ..., \gamma_t^{(h)} \right]^T$ é o autovetor normalizado associado ao único autovalor diferente de zero de $\mathbf{U}\mathbf{U}^T$, $\theta^{(h)}$ é a raiz quadrada desse autovalor não nulo, e $\boldsymbol{\alpha}^{(h)} = \left[\alpha_1^{(h)}, \alpha_2^{(h)}, ..., \alpha_b^{(h)} \right]^T$ está dado por $\left[\theta^{(h)} \right]^{-1} \mathbf{U}^T \boldsymbol{\gamma}^{(h)}$.

3) O CASO DE $R = 2$ OU ALS(2)

a) A estimação inicial.

Somente $\mu^{(0)}$, $\beta_j^{(0)}$, $\theta_1^{(0)}$, $\theta_2^{(0)}$, $\alpha_{1j}^{(0)}$ e $\alpha_{2j}^{(0)}$ são necessários. Os valores de $\mu^{(0)}$, $\beta_j^{(0)}$, $\theta_1^{(0)}$ e $\alpha_{1j}^{(0)}$ são obtidos aplicando o modelo com $R = 1$, $\theta_2^{(0)}$ é igual a um e $\alpha_{2j}^{(0)}$ é gerado pseudo-aleatoriamente.

b) A h-ésima iteração.

Como no caso de $R = 0$ e $R = 1$, cada iteração é decomposta em três etapas:

i) A etapa linha.

τ_i^*, γ_{1i}^* e γ_{2i}^* são calculados a partir dos valores dos parâmetros obtidos na iteração $(h-1)$. Mais exatamente, τ_i^*, γ_{1i}^* e γ_{2i}^* são dados para qualquer $i = 1, 2, ..., t$ através da minimização de

$$\sum_j \left[y_{ij} - \left(\mu^{(h-1)} + \tau_i^* + \beta_j^{(h-1)} + \theta_1^{(h-1)} \gamma_{1i}^* \alpha_{1j}^{(h-1)} + \theta_2^{(h-1)} \gamma_{2i}^* \alpha_{2j}^{(h-1)} \right) \right]^2,$$

o somatório é feito somente para os y_{ij} presentes.

ii) A etapa de normalização.

Os valores dos parâmetros para a h-ésima iteração são obtidos impondo as restrições adicionais. O procedimento descrito no caso $R = 1$ pode ser usado, exceto para o cálculo dos dois termos multiplicativos.

Seja a matriz $\mathbf{U}$ de tamanho $(t \times b)$, tal que os elementos são $U_{ij} = M_{ij} - (\mu^{(h)} + \tau_i^{(h)} + \beta_j^{(h)})$. Então $\boldsymbol{\gamma}_1^{(h)}$ e $\boldsymbol{\gamma}_2^{(h)}$ são os autovetores normalizados associados aos dois autovalores diferentes de zero de $\mathbf{U}\mathbf{U}^T$; $\theta_1^{(h)}$ e $\theta_2^{(h)}$ são as raízes quadradas dos autovalores correspondentes e os vetores $\boldsymbol{\alpha}_r^{(h)}$ são dados por $\left[\theta_r^{(h)}\right]^{-1} \mathbf{U}^{-1} \boldsymbol{\gamma}_r^{(h)}$ para $r = 1, 2$.

iii) A etapa coluna.

Valores intermediários β_j^*, α_{1j}^*, α_{2j}^* são calculados a partir dos valores calculados anteriormente. Mais exatamente β_j^*, α_{1j}^*, α_{2j}^* são obtidos para qualquer $j = 1, 2, ..., b$ minimizando

$$\sum_i \left[y_{ij} - \left(\mu^{(h-1)} + \tau_i^* + \beta_j^* + \theta_1^{(h-1)} \gamma_{1i}^* \alpha_{1j}^* + \theta_2^{(h-1)} \gamma_{2i}^* \alpha_{2j}^* \right) \right]^2,$$

a soma é feita somente para os y_{ij} presentes.

iv) A etapa de normalização em ii) é feita novamente.

Calinski et al. (1992) encontraram que o método ALS(0) pode ter melhor desempenho do que os métodos ALS(1) e ALS(2), além disso, na pesquisa desenvolvida por Piepho (1995b) o método ALS(2) apresentou vários problemas de convergência e fica claro que usando os mínimos quadrados alternados com o objetivo de imputar interações, os componentes multiplicativos não deveriam ser maiores do que um. Baseado nas conclusões dessas pesquisas

neste trabalho serão considerados os algoritmos ALS(0) e ALS(1), mas implementando as modificações que fez Piepho (1995b) no algoritmo anteriormente apresentado para ALS(1).

No algoritmo dos mínimos quadrados alternados (ALS) originalmente apresentado por Calinski et al. (1992) os termos α_{rj} no modelo (17) são inicializados pseudo-aleatoriamente de uma distribuição normal padrão, mas Piepho (1995b) encontrou que esses termos podem ser inicializados de um sub-modelo robusto. Ele afirma que em alguns casos a inicialização aleatória produz resultados divergentes para diferentes sementes iniciais do gerador dos números aleatórios, e é por essa razão que é preferível inicializar apartir de um sub-modelo. Assim, o algoritmo de Calinski et al. (1992) modificado por Piepho (1995b) para ALS(1) é o seguinte:

1. Preencher as caselas vazias com as estimativas ALS(0) encontradas em (18), (19) e (20). Obter as estimativas de um modelo com um componente multiplicativo (AMMI1) da tabela completada e inicializar todos os parâmetros com essas estimativas.

2. Resolver o seguinte sistema de equações por mínimos quadrados alternados:

(a)
$$y_{ij} - \left(\widehat{\mu} + \widehat{\tau}_i + \widehat{\beta}_j\right) = \widehat{\theta}_1 \widehat{\gamma}_{1i} \alpha_{1j} \qquad \text{(Resolver para } \alpha_{1j})$$

(b)
$$y_{ij} - \left(\widehat{\mu} + \widehat{\tau}_i + \widehat{\beta}_j\right) = \widehat{\theta}_1 \gamma_{1i} \widehat{\alpha}_{1j} \qquad \text{(Resolver para } \gamma_{1i})$$

Depois de resolver as duas equações, deve-se normalizar como foi explicado em (21), isto é calcular uma tabela de dupla entrada dos parâmetros atuais e fazer uma análise AMMI1.

3. Resolver o seguinte sistema de equações por mínimos quadrados alternados:

(a)
$$y_{ij} - (\widehat{\mu} + \widehat{\tau}_i) = \beta_j + \widehat{\theta}_1 \widehat{\gamma}_{1i} \alpha_{1j} \qquad \text{(Resolver para } \beta_j \text{ e } \alpha_{1j})$$

(b)
$$y_{ij} - \left(\widehat{\mu} + \widehat{\beta}_j\right) = \tau_i + \widehat{\theta}_1 \gamma_{1i} \widehat{\alpha}_{1j} \qquad \text{(Resolver para } \tau_i \text{ e } \gamma_{1i})$$

Depois de resolver as duas equações, deve-se normalizar como foi explicado em (21), isto é calcular uma tabela de dupla entrada dos parâmetros atuais e fazer uma análise AMMI1.

Os mínimos quadrados alternados (ALS) são baseados unicamente nos dados observados. Denis e Baril (1992) argumentaram e subsequentemente demonstraram através de exemplo, que em alguns experimentos os ALS podem levar a maus ajustes para os dados faltantes em tabelas de dupla entrada. Imagine que os valores ausentes fossem conhecidos e pudessem ser calculadas as regressões para os dados incompletos e para os dados completos. Se essas linhas de regressão diferirem consideravelmente, a extrapolação apartir da regressão obtida com os dados incompletos poderia ser altamente inexata. Para resolver esse problema, Denis e Baril (1992) sugerem fazer as análises sobre tabelas completas de dupla entrada, em que os dados faltantes são substituídos pelas estimativas de algum sub-modelo robusto. Resultados empíricos indicaram que uma ponderação igual para os valores ausentes e observados é aceitável (DENIS; BARIL, 1992; PIEPHO, 1995b). Com uma ponderação igual as análises são equivalentes à análise AMMI clássica para dados completos. Para a análise AMMI1 Denis e Baril (1992) propuseram AMMI0 como um sub-modelo robusto, da mesma maneira o AMMI1 pode ser usado como um sub-modelo robusto para uma análise AMMI2 etc. Neste trabalho serão considerados os métodos r-AMMI1 e r-AMMI2. Note-se que o método r-AMMI0 é equivalente com o algoritmo ALS(0).

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

2.4 Rotação Procrustes (Medida de qualidade de ajuste)

Uma ferramenta da análise multivariada apresentada por Mardia et al. (1979) pode ser usada para fazer comparações de configurações de pontos em duas matrizes. Essa ferramenta é conhecida como rotação Procrustes e será descrita na sequência. Seja $\mathbf{X}$ a matriz $(n \times p)$ das coordenadas de n pontos obtidas a partir de $\mathbf{D}$ baseada em uma técnica determinada. Suponha que $\mathbf{Y}$ é a matriz $(n \times q)$ de coordenadas de outro conjunto de pontos obtidos por uma técnica diferente. Seja $q \leq p$. Adicionando colunas de zeros na matriz $\mathbf{Y}$, é possível assumir que $\mathbf{Y}$ tenha a dimensão $(n \times p)$.

A medida de qualidade de ajuste é obtida pelo movimento dos pontos $\mathbf{y}_r$ relativo aos pontos $\mathbf{x}_r$ até que a soma de quadrados do erro seja mínima.

$$\sum_{r=1}^{n} (\mathbf{x}_r - \mathbf{y}_r)^T (\mathbf{x}_r - \mathbf{y}_r)$$

O movimento dos pontos pode fazer-se através da rotação, reflexão e translação, isto é,

$$\mathbf{A}^T\mathbf{y}_r - \mathbf{b}, \quad r = 1, 2, ..., n,$$

em que $\mathbf{A}^T$ é uma matriz $(p \times p)$ ortogonal. Então deve-se resolver

$$R^2_{procrustes} = \min_{\mathbf{A},\mathbf{b}} \sum_{r=1}^{n} \left(\mathbf{x}_r - \mathbf{A}^T\mathbf{y}_r - \mathbf{b}\right)^T \left(\mathbf{x}_r - \mathbf{A}^T\mathbf{y}_r - \mathbf{b}\right) \tag{22}$$

para $\mathbf{A}$ e para $\mathbf{b}$ que são encontrados por mínimos quadrados. Os valores são dados por Mardia et al. (1979) no seguinte teorema.

Teorema 1 *Seja* $\mathbf{X}(n \times p)$ *e* $\mathbf{Y}(n \times p)$ *dois conjuntos de dados e por conveniência centrados na origem, tal que* $\overline{\mathbf{x}}$=$\overline{\mathbf{y}}$= $\mathbf{0}$. *Seja* $\mathbf{Z} = \mathbf{Y}^T\mathbf{X}$ *e usando a decomposição por valores singulares, pode-se escrever*

$$\mathbf{Z} = \mathbf{V\Gamma U}^T,$$

em que $\mathbf{V}$ *e* $\mathbf{U}$ *são matrizes* $(p \times p)$ *ortogonais e* Γ *é uma matriz diagonal de elementos não negativos. Então os valores que minimizam (22) são dados por,*

$$\widehat{\mathbf{b}}= \mathbf{0}, \quad \widehat{\mathbf{A}}= \mathbf{VU}^T.$$

e também

$$R^2_{procrustes} = traço(\mathbf{XX}^T) + traço(\mathbf{YY}^T) - 2traço(\mathbf{\Gamma}). \tag{23}$$

$\widehat{\mathbf{A}}$ *é chamada a rotação Procrustes de* $\mathbf{Y}$ *em relação a* $\mathbf{X}$.

Nesse trabalho $\mathbf{X}$ será a matriz de dados original completa do experimento genótipo por ambiente e $\mathbf{Y}$ será a matriz de dados depois de fazer imputação de dados. Então, por exemplo, $\mathbf{Y}_1$ pode representar a matriz de dados com observações imputadas pelo método IMLD e $\mathbf{Y}_2$ a matriz de dados com imputações por ALS. Serão feitas duas rotações Procrustes, uma rotação de $\mathbf{Y}_1$ em relação a $\mathbf{X}$ e outra rotação de $\mathbf{Y}_2$ em relação a $\mathbf{X}$. Em cada rotação pode-se calcular um R^2, notados por R^2_1 e R^2_2, então se $R^2_1 < R^2_2$ pode-se concluir que o método mais recomendável para imputar será IMLD, caso contrário será o método ALS.

2.5 Material e métodos

2.5.1 Características dos dados

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

2.5.2 Metodologia

Análise dos dados originais Dado que se tinha um experimento completo, isto é sem observações ausentes, foi feita uma análise AMMI baseada nos dados originais de Farias (2005) (ver Apêndice A). Neste ponto foi encontrado o número de componentes necessário para um modelo de efeitos aditivos e interação multiplicativa. Os critérios para escolher o número de componentes multiplicativos foi através do teste de Cornelius et al. (1996) e o teste de Milliken e Johnson (1989), os dois explicados anteriormente. O teste de Cornelius F_R leva em conta uma estimativa da variância, enquanto o método por máxima verossimilhança (MILLIKEN; JOHNSON, 1989) define o número de componentes de acordo com uma estatística baseada nos autovalores da matriz de dados. O teste F_R já mostrou robustez quanto à falta de homocedasticidade e normalidade nos erros (PIEPHO, 1995a), enquanto o teste por máxima verossimilhança pode ser aplicado também naqueles experimentos nos quais não se tem uma estimativa da variância e não se têm mais do que uma observação por combinação de tratamentos. O objetivo de fazer esta análise foi comparar o número de componentes multiplicativos nos dados originais com o número de componentes escolhido baseado em dados que contêm imputações e assim determinar se pode ser correto testar interação sobre dados imputados.

Estudo de simulação Para avaliar o desempenho da metodologia recentemente proposta por Bergamo et al. (2008) e Bergamo (2007) para imputar interações em experimentos G×E decidiu fazer-se um estudo comparativo com outras metodologias, assim, a recente proposta de predizer interações através da média das imputações múltiplas livres de distribuição foi comparada com a imputação usando mínimos quadrados alternados (ALS) e com a imputação utilizando estimativas AMMI de sub-modelos robustos (r-AMMI). Mas, para fazer essa comparação foi desenvolvido um estudo de simulação baseado em um conjunto recente de dados reais (FARIAS, 2005). As simulações foram implementadas no SAS/IML (SAS INSTITUTE, 2004).

O conjunto de dados contém a informação correspondente à média de produtividade de algodão (kg/ha) em 15 genótipos e 27 ambientes, corresponde a 405 observações. Esse conjunto foi submetido a retiradas aleatórias de diferentes porcentagens de dados. Foram consideradas as porcentagens de 10%, 20% e 30%, mas para dar consistência aos resultados obtidos, o processo foi repetido 1000 vezes para cada porcentagem, para um total de 3000 perdas aleatórias, ou seja 3000 conjuntos de dados diferentes foram gerados. No primeiro caso (10%) foram retiradas 41 interações, no segundo caso (20%) foram retiradas 81 interações e finalmente no terceiro caso 30% foram retiradas 122 interações. Para cada um dos 3000 conjuntos de dados com dados faltantes simulados, os algoritmos IMLD, ALS(0), ALS(1), r-AMMI1 e r-AMMI2 foram aplicados para predizer os valores ausentes através de um programa computacional.

Critérios de comparação Para comparar as estimativas dos diferentes métodos de imputação foram levados em conta quatro critérios de comparação. Dois dos quatro critérios fazem uma comparação baseado no conjunto de dados completados (observados+imputados), isto é quando em cada conjunto de dados simulado, os valores ausentes foram substituídos pelas correspondentes predições de cada método. Assim, sobre as matrizes de dados completadas foram encontrados os números de componentes multiplicativos de um modelo AMMI de acordo com os métodos F_R, máxima verossimilhança e os resultados obtidos foram comparados com os resultados das análises dos dados originais.

Da mesma maneira, cada matriz de dados completada foi comparada com a matriz original através da rotação de Procrustes. Com a estatística de Procrustes se obtem uma medida da diferença entre duas configurações de pontos e o método de imputação que minimize essa diferença indicará os melhores métodos.

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

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

em que y_{ij} representa a produção média do i-ésimo genótipo no j-ésimo ambiente no conjunto de dados original, $\widehat{y}_{ij}$ representa a estimativa do i-ésimo genótipo no j-ésimo usando cada um dos métodos de imputação considerados e NA o número total de valores ausentes ($i = 1, ..., 15$, $j = 1, ..., 27$). Quanto menor seja a $RMSPD$, melhor será o método de imputação.

O último critério de comparação considerado nesse estudo 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 metodo 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.

2.6 Resultados e discussão

2.6.1 Análise do conjunto original de dados

Inicialmente foi feita uma análise para encontrar o número de componentes multiplicativos do modelo AMMI que deveria ser usado no conjunto de dados original. Para encontrar esse número foi aplicado o teste F_R de Cornelius e o teste de máxima verossimilhança denotado como F_{teste}. Os resultados se apresentam a seguir:

Tabela 2 - Análise da variância pelo sistema de Cornelius baseada em dados originais

F.V.	GL	SQ	QM	F_R	Valor-p
Genótipos (G)	14	2175438,3000	155388,4500	1,76	0,0423
Ambientes (E)	26	684073204,0000	26310508,0000	298,73	0,0000
Interação (G×E)	364	32059271,0000	88074,9204	1,97	0,0000
IPCA1	325	23948838,4850	73688,7338	1,64	0,0000
IPCA2	288	18060443,5680	62709,8735	1,40	0,0001
IPCA3	253	14003995,6848	55351,7616	1,24	0,0132
IPCA4	220	10863885,1200	49381,2960	1,10	0,1664
IPCA5	189	8219276,5473	43488,2357	0,97	0,5946
IPCA6	160	6309347,8560	39433,4241	0,88	0,8462
IPCA7	133	4950038,8609	37218,3373	0,83	0,9132
IPCA8	108	3614043,9924	33463,3703	0,75	0,9730
IPCA9	85	2423746,3210	28514,6626	0,64	0,9956
IPCA10	64	1475275,6864	23051,1826	0,51	0,9995
IPCA11	45	829241,1630	18427,5814	0,41	0,9998
IPCA12	28	376314,3048	13439,7966	0,30	0,9999
IPCA13	13	114550,6921	8811,5917	0,20	0,9991
IPCA14	–	–	–	–	–
Erro médio/n	1134		44797,4597		

Na tabela 2 segundo o sistema de Cornelius, observa-se que com 5% de probabilidade, existe uma diferença significativa da produção de algodão entre os genótipos (G), entre os ambientes (E) e um efeito significativo da interação (G×E). Segundo esse sistema o melhor modelo para explicar a interação multiplicativa seria o AMMI4, já que somente apartir do IPCA4 o resíduo AMMI torna-se não significativo. Os dados originais só tinham as médias das repetições, mas para encontrar o quadrado médio do erro médio/n se calculou uma estimativa da variância do experimento usando os quadrados médios das análises individuais apresentados em Farias (2005) (ver Apêndice B). A estimativa da variância foi $\widehat{QME} = 179189{,}8389$ com 1134 graus de liberdade associados e então o QM(Erro médio/n) $= 179189{,}8389/4 = 44797{,}4597$.

Entretanto, usando o método de máxima verossimilhança denotado por F_{teste}, o resultado é diferente, pois é recomendado outro modelo com um número diferente de componentes para a interação.

Tabela 3 - Teste de máxima verossimilhança

Componente	d_1	d_2	F_{Teste}	Valor-p
IPCA1	78,2087	546,4341	1,6981	0,0004
IPCA2	71,4969	473,4628	1,4836	0,0095
IPCA3	64,8902	405,8620	1,1395	0,2277
IPCA4	58,3937	343,5691	1,0111	0,4591
IPCA5	52,0135	286,5234	1,0443	0,4000
IPCA6	45,7569	234,6659	0,8101	0,8012
IPCA7	39,6325	187,9386	0,5467	0,9869
IPCA8	33,6493	146,2803	0,7564	0,8270
IPCA9	27,8134	109,6184	0,9561	0,5353
IPCA10	22,1200	77,8520	1,1063	0,3591
IPCA11	16,5347	50,8327	1,0277	0,4458
IPCA12	10,9692	28,3996	1,2144	0,3224
IPCA13	5,3720	10,7817	1,2896	0,3377

Na tabela 3 d_1 e d_2 representam os graus de liberdade da estatística de verossimilhança transformada (F_{teste}) e se observa que o último componente significativo ao 5% de probabilidade é o IPCA2, assim um modelo adequado poderia ser o AMMI2. Os resultados dessas análises dos componentes multiplicativos serão utilizados posteriormente para avaliar os métodos de imputação, mas uma interpretação do obtido nesta análise é que o modelo mais adequado para os dados se encontra entre o modelo AMMI2 e o modelo AMMI4, o que significa que AMMI3 deveria ser levado em conta para análises posteriores. No começo, a análise pode ser feita com dois componentes multiplicativos por ser a estrutura mais simples, mas no caso de encontrar problemas nas análises dos resíduos como heterogeneidade de variância ou falta de normalidade, aumentar componentes pode ser uma alternativa.

A seguir se apresentam os resultados do estudo de simulação segundo os critérios de comparação escolhidos para os algoritmos de imputação.

2.6.2 Comparação dos métodos de imputação através da $RMSPD$ padronizada

Em cada uma das mil retiradas aleatórias feitas para cada porcentagem de perda considerada (10%, 20% e 30% de informação) foram obtidas as $RMSPD$ depois de imputar

por IMLD, ALS(1), ALS(0), r-AMMI2 e r-AMMI1, mas, para conseguir ver as diferenças entre os diferentes métodos foram computados a média e o desvio padrão para calcular as estatísticas $RMSPD$ padronizadas e sobre as quais foi feita diretamente a comparação. Por exemplo, na primeira retirada aleatória de 10% após a imputação, as estatísticas $RMSPD$ foram: 536,6643 para IMLD, 411,6015 para ALS(1), 387,6654 para ALS(0), 411,4858 para r-AMMI2 e 388,9791 para r-AMMI1. A média desses valores é 427,28 e o desvio padrão 62,24, então, substraindo essa média de cada $RMSPD$ encontrada e dividindo o resultado pelo desvio padrão, obtém-se os valores da estatística padronizada. A $RMSPD$ padronizada para IMLD foi 1,76, para ALS(1) -0,25, para ALS(0) -0,64, para r-AMMI2 -0,25 e para r-AMMI1 -0,62. Sobre essas estatísticas $RMSPD$ padronizadas foram feitas as comparações e conclusões deste trabalho, pelo qual daqui em adiante $RMSPD$ fará referência aos valores padronizados. Os melhores métodos de imputação serão aqueles que minimizem a $RMSPD$. A seguir se mostra o gráfico de caixas considerando as mil retiradas aleatórias do 10% de dados:

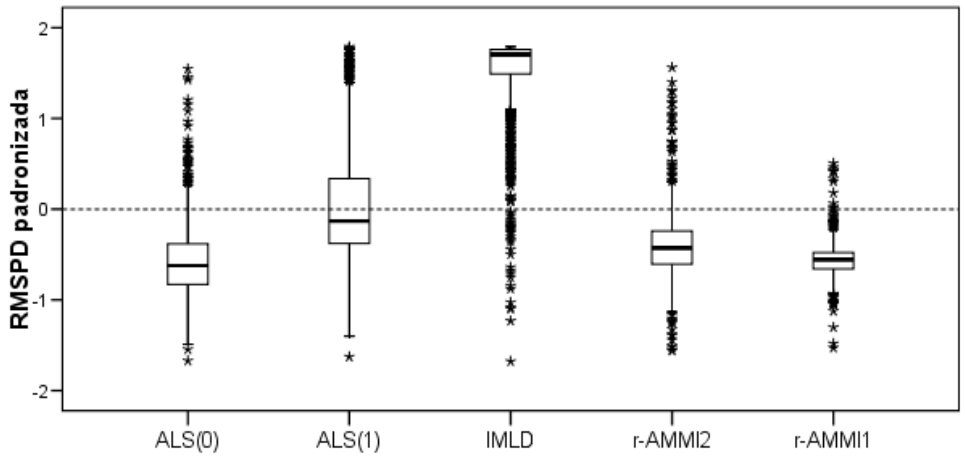

Figura 1 - Box plot da distribuição da RMSPD padronizada com 10% de retirada dos dados

O gráfico de caixas dá uma idéia da posição, dispersão, assimetria, caudas e possíveis dados discrepantes na diferença preditiva ($RMSPD$). Observa-se que a distribuição da diferença preditiva padronizada é assimétrica à esquerda usando a metodologia de imputação múltipla, IMLD. No caso dos mínimos quadrados alternados se obteve uma distribuição simétrica quando foram imputados os dados com ALS(0) e assimétrica a direita quando a predição foi por ALS(1), ou seja considerando uma componente de interação multiplicativa. Todas as metodologias apresentam dados discrepantes, pois têm-se muitos valores afastados do corpo principal dos dados, mas parece que a menor variabilidade é obtida com o, algoritmo r-AMMI1. Pode-se concluir que a maior diferença preditiva mediana é alcançada com a

predição de interações faltantes através da média das imputações múltiplas (IMLD), enquanto a menor $RMSPD$ mediana pode ser atingida usando estimativas aditivas ou estimativas baseadas em sub-modelos robustos, isto é ALS(0) e r-AMMI1. Agora é apresentada uma tabela com as médias da $RMSPD$ padronizada para cada porcentagem considerada no estudo de simulação

Tabela 4 - Médias da RMSPD padronizada

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

Na tabela 4 observa-se que o método que minimiza as médias da diferença preditiva padronizada $RMSPD$ em todas as porcentagens de perda é o ALS(0) com uma média de -0,57 para 10%, -0,68 para 20% e -0,73 para 30% de retirada dos dados, enquanto o método que maximiza esses valores em todos os casos é o IMLD. Assim, segundo essa comparação o melhor método poderia ser aquele que oferece um valor pequeno da média da $RMSPD$ padronizada e corresponde aos mínimos quadrados alternados com estimativas aditivas. Note-se que resultados muito próximos daquele método são proporcionados pelo r-AMMI1.

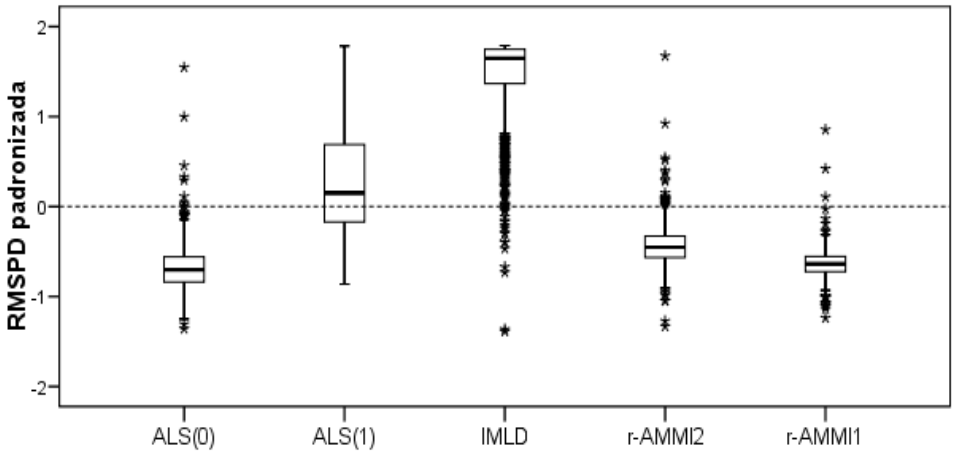

Figura 2 - Box plot da distribuição da RMSPD padronizada com 20% de retirada dos dados

Continuando com a comparação, na figura 2 apresenta-se um gráfico de caixas para a distribuição da estatística padronizada ($RMSPD$) quando foi considerado 20% de infor-

mação perdida. Observa-se que o comportamento das metodologias de imputação continua aproximadamente igual ao encontrado na figura 1 com o 10% de informação faltante, pois a assimetria da distribuição da $RMSPD$ padronizada com IMLD e ALS(1) se mantém, além disso, a menor mediana da $RMSPD$ continua sendo obtida com o algoritmo ALS(0) e com as predições de r-AMMI1. Entretanto, com o aumento na porcentagem de dados perdidos os pontos discrepantes desapareceram para o ALS(1) e o tamanho das caixas aumentou para esse método e também para IMLD, enquanto diminuiu ou se manteve para as metodologias restantes. Entretanto, para estudar a dispersão da estatística de comparação serão analizadas as variâncias da $RMSPD$ na tabela 5.

Tabela 5 - Variâncias da RMSPD padronizada

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

A menor variância da $RMSPD$ padronizada foi encontrada usando o método r-AMMI1 em todas as porcentagens de perda. Para 10%, a menor variância da $RMSPD$ foi 0,03, para 20% foi 0,02 e para 30% de retirada dos dados, a menor variância foi 0,01. Nesta análise, a variância foi maximizada nas diferentes porcentagens através dos mínimos quadrados alternados usando como modelo de imputação o AMMI1, ou seja, com o algoritmo ALS(1). A proposta baseada em imputação múltipla, IMLD, têm variâncias maiores do que os métodos ALS(0) e r-AMMI2. Até este ponto a nova metodologia IMLD não apresenta um desempenho melhor do que as predições baseadas em modelos AMMI. Para confirmar os resultados, foi feito o teste de Levene para homogeneidade de variâncias dos métodos. Para 10%, 20% e 30% de retirada dos dados, os valores da estatística de Levene foram: 67,51 (valor-p < 0,0001), 175,05 (valor-p < 0,0001) e 349,11 (valor-p < 0,0001) respectivamente, indicando a rejeição da hipótese de homogeneidade de variâncias dos métodos.

Possivelmente a média da $RMSPD$ padronizada não seja o melhor critério para comparar, levando em conta o grande número de casos extremos que mostraram os gráficos de caixas para todos os métodos de imputação e é por isso que a mediana deve ser analizada, mas a diferença do apresentado anteriormente junto com os gráficos de caixas, será considera-

da uma tabela contendo a mediana para cada um dos métodos de predicão nas diferentes porcentagens de perda (tabela 6).

Tabela 6 - Medianas da RMSPD padronizada

	Porcentagem de retirada		
Métodos	10%	20%	30%
ALS(1)	-0,1305	0,1519	0,3826
ALS(0)	-0,6219	-0,7022	-0,7262
IMLD	1,7064	1,6491	1,5723
r-AMMI1	-0,5562	-0,6387	-0,6759
r-AMMI2	-0,4270	-0,4517	-0,4973

ALS(0) tem as menores medianas das $RMSPD$ padronizadas, seguido pelo metodo r-AMMI1. Comparando o ALS(0) com o IMLD, o desempenho do primeiro método é melhor, pois com ele se minimiza a mediana da $RMSPD$. Embora, o IMLD utiliza toda a informação disponível na matriz de dados para as imputações, isso não parece ser suficiente pois as predições aditivas com ALS(0) e as predições com r-AMMI1 parecem ter uma melhor qualidade segundo a estatística considerada nesta seção.

Novamente, levando em conta os dados afastados das caixas da $RMSPD$ padronizada para todos os métodos de imputação, foi utilizada uma medida de dispersão alternativa à variância e que envolve duas medidas de localização resistentes das distribuições, trata-se da diferença entre o quartil 3 e o quartil 1 conhecida como a distância interquartil e que é exibida na tabela 7.

Tabela 7 - Distância interquartil Q3-Q1 da RMSPD padronizada

	Porcentagem de retirada		
Métodos	10%	20%	30%
ALS(1)	0,7146	0,8598	0,9287
ALS(0)	0,4450	0,2788	0,2134
IMLD	0,2690	0,3800	0,5029
r-AMMI2	0,3644	0,2385	0,1865
r-AMMI1	0,1814	0,1682	0,1402

Segundo esse critério, as maiores distâncias interquartis em todas as porcentagens de perda foram produzidas com a imputação por mínimos quadrados alternados com um componente multiplicativo, ou seja ALS(1). Entretanto, a menor distância interquartil se obteve

em todos os casos com o método baseado em estimativas de sub-modelos robustos r-AMMI1. Note-se que nesse caso o método IMLD obteve uma menor distância entre os quartis do que o ALS(0) quando foram feitas as retiradas aleatórias do 10%, o valor para ALS(0) foi 0,44, enquanto para IMLD foi 0,26. A seguir mostra-se o gráfico de caixas obtido quando foram simulados conjuntos de dados com o 30% de observações ausentes.

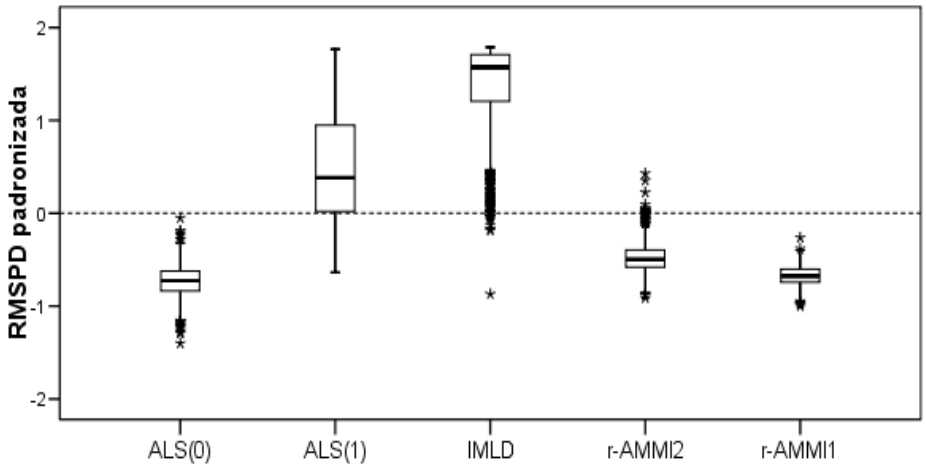

Figura 3 - Box plot da distribuição da RMSPD padronizada com 30% de retirada dos dados

Comparando o gráfico 3 com os gráficos apresentados para as outras porcentagens de perda consideradas, pode-se dizer que a dispersão da $RMSPD$ dos métodos IMLD e ALS(1) aumenta conforme aumenta o número de dados faltantes, pois as caixas aumentam de tamanho, caso contrário acontece com ALS(0), r-AMMI1 e r-AMMI2 porque as respectivas caixas diminuiram de tamanho. Também se obervou que com um maior número de dados perdidos os dados discrepantes da $RMSPD$ tenderam a diminuir. Nos três gráficos apresentados da distribuição da $RMSPD$ padronizada o método IMLD não apresentou um melhor comportamento do que aqueles que utilizaram modelos AMMI para predizer as interações, em geral o método IMLD apresentou maiores medianas, médias e medidas de dispersão consideradas do que os métodos ALS(0) e r-AMMI1.

Para verificar a diferença da $RMSPD$ padronizada entre os cinco métodos de imputação considerados, foi feito o teste não paramétrico de Friedman em cada porcentagem de retirada, os valores da estatística $\chi^2_{Friedman}$ foram: 2308,13 (valor-p < 0,0001) para 10%, 2987,538 (valor-p < 0,0001) para 20 % e 3307,82 (valor-p < 0,0001) para 30% de retirada dos dados. Uma vez confirmada a diferença média entre os métodos, fizeram-se comparações múltiplas (Apêndice C) confrontando os algoritmos de imputação dois a dois, encontrando que existem diferenças significativas entre todos os métodos, exceto entre ALS(0) e r-AMMI1.

Finalmente, na tabela 8 apresentam-se as principais estatísticas obtidas no estudo de simulação para a $RMSPD$ padronizada nos três mil conjuntos de dados gerados a partir do conjunto de dados real com dados faltantes simulados, sem considerar a porcentagem de perda. Na tabela 8 Q3-Q1 representa a distância interquartil.

Tabela 8 - Estatísticas gerais da RMSPD padronizada

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

Conclui-se pelo estudo em geral da estatística $RMSPD$ padronizada, que o método que minimiza a distância interquartil é o método r-AMMI1 com um valor de 0,18. A mediana e a media da $RMSPD$ apresentam os menores valores, -0,69 e -0,66, quando foram feitas as imputações com ALS(0). Entretanto, a menor variância, 0,02, foi alcançada imputando com as estimativas baseadas em um modelo sub-robusto, isto é o método r-AMMI1. De acordo com os objetivos deste trabalho, procurou-se avaliar a nova metodologia chamada IMLD e encontrou-se que segundo a $RMSPD$ padronizada, a predição de interações através da média de imputações múltiplas não oferece um melhor desempenho do que aqueles métodos baseados em modelos de efeitos aditivos com interação multiplicativa. O método IMLD foi apenas superior ao método ALS(1) quando foram comparadas medidas de dispersão. Com os resultados obtidos, os melhores métodos segundo a $RMSPD$ padronizada foram ALS(0) e r-AMMI1.

2.6.3 Comparação dos métodos de imputação através da estatística de Procrustes

A estatística de Procrustes $R^2_{procrustes}$ definida em (23) mede a diferença de duas configurações de pontos, quanto menor o valor da estatística, as configurações serão mais similares. Neste trabalho foi usada essa estatística para comparar os resultados após a imputação pelos cinco métodos considerados, ALS(1), ALS(0), IMLD, r-AMMI2 e r-AMMI1. A comparação foi feita sobre as matrizes completadas (observados+imputados) e a matriz original de dados. Assim, por exemplo, a matriz de dados original chamada $\mathbf{X}_{(15\times 27)}$ é submetida a uma retirada aleatória de 10% dos seus elementos, os valores ausentes são imputados por

IMLD e as imputações obtidas são inseridas na matriz de dados incompleta, essa nova matriz completada é chamada de Y_{IMLD}. Então, a estatística $R^2_{procrustes}$ foi calculada fazendo a rotação da matriz Y_{IMLD} em relação à matriz $\mathbf{X}_{(15\times27)}$, o valor da estatística foi denotado por $\text{IMLD}_{procrustes}$. Além disso, sobre a mesma retirada aleatória do 10% foram encontradas as matrizes $Y_{ALS(1)}$, $Y_{ALS(0)}$, $Y_{r-AMMI2}$, $Y_{r-AMMI1}$ e calculadas as correspondentes estatísticas de Procrustes relativas à matriz original $\mathbf{X}_{(15\times27)}$, os valores dessas estatísticas foram denotados por $\text{ALS(1)}_{procrustes}$, $\text{ALS(0)}_{procrustes}$, $\text{r-AMMI2}_{procrustes}$ e $\text{r-AMMI1}_{procrustes}$. Sobre essas estatísticas são comparadas as imputações, e quanto menor seja o valor da $R^2_{procrustes}$ melhor será o método de imputação. Todo o exposto anteriormente foi envolvido em um estudo de simulação considerando três porcentagens de perda aleatória (10%, 20%, 30%) e fazendo mil repetições para cada uma delas, os resultados se mostram a seguir.

De acordo com um dos objetivos deste trabalho, procura-se avaliar a recente proposta baseada em IMLD (BERGAMO, 2007) para imputar interações. É por essa razão que se apresenta inicialmente uma comparação dos resultados obtidos com os valores das estatísticas de Procrustes para IMLD. Na tabela 9 aparece uma contagem do número de vezes que o valor da estatística $\text{IMLD}_{procrustes}$ foi maior ou menor do que as estatísticas $\text{ALS(1)}_{procrustes}$, $\text{ALS(0)}_{procrustes}$, $\text{r-AMMI2}_{procrustes}$ e $\text{r-AMMI1}_{procrustes}$ em cada porcentagem de perda considerada.

Tabela 9 - Número de vezes que a estatística de Procrustes para IMLD foi maior ou menor do que as estatísticas de Procrustes correspondentes aos outros métodos de imputação; 1000 conjuntos de dados simulados para cada porcentagem de retirada

	Porcentagem de retirada		
Comparação Procrustes	10%	20%	30%
$\text{IMLD}_{procrustes} < \text{ALS(1)}_{procrustes}$	107	162	214
$\text{IMLD}_{procrustes} > \text{ALS(1)}_{procrustes}$	893	838	786
$\text{IMLD}_{procrustes} < \text{ALS(0)}_{procrustes}$	18	7	1
$\text{IMLD}_{procrustes} > \text{ALS(0)}_{procrustes}$	982	993	999
$\text{IMLD}_{procrustes} < \text{r-AMMI2}_{procrustes}$	25	6	3
$\text{IMLD}_{procrustes} > \text{r-AMMI2}_{procrustes}$	975	994	997
$\text{IMLD}_{procrustes} < \text{r-AMMI1}_{procrustes}$	14	5	1
$\text{IMLD}_{procrustes} > \text{r-AMMI1}_{procrustes}$	986	995	999

Segundo a tabela 9, nos 1000 conjuntos de dados com o 10% de observações ausentes simuladas o desempenho do IMLD não foi melhor do que os outros métodos, por exemplo, em 982 ocasiões a estatística $\text{ALS(0)}_{procrustes}$ teve valores menores do que os valores

da $\text{IMLD}_{procrustes}$, o que significa que as matrizes de dados com as imputações por ALS(0) apresentam uma maior similaridade em relação à matriz original, do que podem apresentar as matrizes com imputações por IMLD. Mas, não somente com o ALS(0) acontece isso, com os métodos r-AMMI2 e r-AMMI1 mostrou-se um comportamento parecido, pois em 975 conjuntos de dados a estatística $\text{r-AMMI2}_{procrustes}$ foi menor do que a estatística $\text{IMLD}_{procrustes}$. Além disso, aumentando a porcentagem de perda aleatória se obtêm as mesmas conclusões. Com 30% de perda aleatória, dos mil conjuntos de dados simulados com observações perdidas, em 999 deles a estatística $\text{r-AMMI1}_{procrustes}$ foi menor do que a $\text{IMLD}_{procrustes}$. Usando a rotação de Procrustes pode concluir-se que as matrizes de dados completadas com imputações feitas por mínimos quadrados alternados ou baseadas em sub-modelos robustos apresentam uma maior similaridade com a matriz de dados original considerada para o estudo de simulação.

Já que foi encontrado que segundo a estatística de Procrustes o método IMLD não apresenta um bom desempenho comparado com as outras opções de imputação, apresenta-se na tabela 10 a comparação baseada no algoritmo ALS(1).

Tabela 10 - Número de vezes que a estatística de Procrustes para ALS(1) foi maior ou menor do que as estatísticas de Procrustes correspondentes aos métodos ALS(0) e r-AMMI; 1000 conjuntos de dados simulados para cada porcentagem de retirada

Comparação Procrustes	Porcentagem de retirada		
	10%	20%	30%
$\text{ALS(1)}_{procrustes} < \text{ALS(0)}_{procrustes}$	220	76	19
$\text{ALS(1)}_{procrustes} > \text{ALS(0)}_{procrustes}$	780	924	981
$\text{ALS(1)}_{procrustes} < \text{r-AMMI2}_{procrustes}$	253	103	36
$\text{ALS(1)}_{procrustes} > \text{r-AMMI2}_{procrustes}$	747	897	964
$\text{ALS(1)}_{procrustes} < \text{r-AMMI1}_{procrustes}$	113	28	8
$\text{ALS(1)}_{procrustes} > \text{r-AMMI1}_{procrustes}$	887	972	992

Considerando 10% de perda aleatória, o método que apresentou mais vezes uma estatística de Procrustes inferior à $\text{ALS(1)}_{procrustes}$, foi o método r-AMMI1 com 887 vezes, de um total de 1000 conjuntos considerados com observações ausentes simuladas. Isto significa que em 887 vezes as matrizes completadas com imputações através de r-AMMI1 mostraram uma maior similaridade com a matriz de dados do experimento original. Observando apenas o 10% de perda aleatória, os melhores métodos seriam em ordem, r-AMMI1, ALS(0) e r-AMMI2. Aumentando a porcentagem de perda os resultados são confirmados, pois por exemplo, nos 1000 conjuntos com o 20% de perda aleatória, apenas em 76 conjuntos o ALS(1) apresentou

matrizes completadas mais similares do que as matrizes completadas utilizando o ALS(0).

Com as análises feitas até esse momento, dos cinco métodos de imputação escolhidos inicialmente, três mostraram um desempemho bem diferente e é por isso que a ultima tabela desta seção faz uma comparação entre eles usando a estatística de Procrustes como critério.

Tabela 11 - Número de vezes que a estatística de Procrustes para ALS(0) foi maior ou menor do que as estatísticas de Procrustes correspondentes aos métodos r-AMMI; 1000 conjuntos de dados simulados para cada porcentagem de retirada

	Porcentagem de retirada		
	10%	20%	30%
$\text{ALS(0)}_{procrustes} < \text{r-AMMI2}_{procrustes}$	668	788	868
$\text{ALS(0)}_{procrustes} > \text{r-AMMI2}_{procrustes}$	332	212	132
$\text{ALS(0)}_{procrustes} < \text{r-AMMI1}_{procrustes}$	562	633	669
$\text{ALS(0)}_{procrustes} > \text{r-AMMI1}_{procrustes}$	438	367	331
$\text{r-AMMI2}_{procrustes} < \text{r-AMMI1}_{procrustes}$	302	193	104
$\text{r-AMMI2}_{procrustes} > \text{r-AMMI1}_{procrustes}$	698	807	896

Pela tabela 11 observa-se que com o método ALS(0) se produziram as matrizes com a melhor similaridade da matriz original de dados, pois em um maior número de ocasiões obteve-se uma estatística $\text{ALS(0)}_{procrustes}$ menor do que as estatísticas produzidas após a imputação com r-AMMI1 e r-AMMI2. Dos mil conjuntos de dados gerados com 10% de informação perdida, em 668 deles, a $\text{ALS(0)}_{procrustes}$ foi menor do que a $\text{r-AMMI2}_{procrustes}$ e em 562 foi menor do que a $\text{r-AMMI1}_{procrustes}$. Aumentando o número de porcentagem de perda aleatória a situação se mantém, isto é, nos 1000 conjuntos simulados com o 30% de perda, em 669 obteve-se melhor similaridade usando ALS(0) do que o obtido com r-AMMI1.

Finalmente, conclui-se que comparando através da rotação de procrustes a matriz de dados original com matrizes de dados que contêm dados imputados em diferentes porcentagens de perda, o melhor método é o ALS(0) e nenhum dos outros métodos (IMLD, r-AMMI1, r-AMMI2 e ALS(1)) apresentam um resultado aproximado.

2.6.4 Comparação dos métodos de imputação através da correlação de Spearman

Neste estudo foram gerados três mil conjuntos de informação com dados faltantes simulados baseado em um experimento completo, do total de conjuntos de dados, mil deles tiveram 10% de perda aleatória, outros mil conjuntos tiveram 20% de perda aleatória e os 1000 restantes tiveram 30% de perda. Em cada um dos conjuntos foram aplicados os métodos para imputar as observações ausentes por IMLD, ALS(1), ALS(0), r-AMMI1 e r-AMMI2. As imputações por cada um dos métodos foi comparada com as observações reais do experimento e calculado o coeficiente de correlação de Spearman, assim, o critério para escolher o melhor método de imputação foi aquele que apresentasse as maiores correlações. A seguir apresentam-se as estatísticas obtidas para a correlação de Spearman quando foram consideradas retiradas aleatórias do 10% e acompanhadas de um gráfico de caixas.

Tabela 12 - Estatísticas da correlação de Spearman imputando 10% de dados

	Métodos de Imputação				
Estatísticas	IMLD	ALS(1)	ALS(0)	r-AMMI2	r-AMMI1
Média	0,9314	0,9464	0,9514	0,9498	0,9514
Desvio padrão	0,0238	0,0218	0,0177	0,0184	0,0174
Mínimo	0,8059	0,7502	0,8559	0,8441	0,8577
Quartil 1 (Q1)	0,9186	0,9358	0,9416	0,9399	0,9418
Mediana	0,9348	0,9502	0,9537	0,9524	0,9545
Quartil 3 (Q3)	0,9485	0,9606	0,9638	0,9629	0,9636
Máximo	0,9794	0,9864	0,9880	0,9883	0,9883
Q3-Q1	0,0300	0,0248	0,0221	0,0230	0,0218

Observando a tabela 12, em geral os cinco métodos estudados oferecem imputações altamente correlacionadas com os dados originais do experimento, pois a média e a mediana dos coeficientes de Spearman são superiores a 0,90. Entretanto, a menor correlação foi obtida através do ALS(1) com um coeficiente de 0,75 e a maior correlação resultou com a imputação por r-AMMI1 e r-AMMI2 com um valor de 0,9883. O desvio padrão dos coeficientes é muito pequeno, bem como a distância interquartil. As diferenças entre as estatísticas são apenas por casas decimais, o que não permite escolher um único método como o melhor.

Na figura 4, o gráfico de caixas mostra o desempenho parecido dos algoritmos de imputação comparando-os através do coeficiente de Spearman. A caixa correspondente ao método IMLD se encontra um pouco abaixo das caixas dos outros métodos, mas isso não

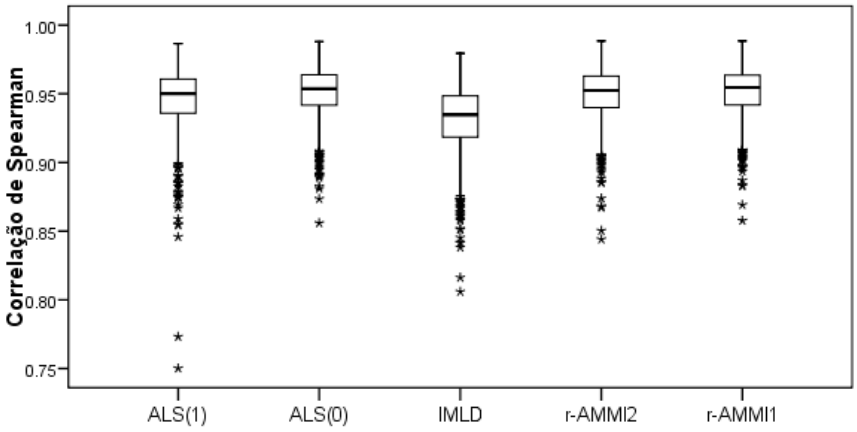

Figura 4 - Distribuição do coeficiente de correlação de Spearman imputando 10% de dados

é suficiente para concluir que esse método é inferior aos demais. A seguir apresenta-se o resultado quando foi aumentada a porcentagem das retiradas aleatórias a 20%.

Na tabela 13 e na figura 5 as conclusões com 20% de informação perdida praticamente são iguais às obtidas anteriormente, todos os métodos apresentam um ótimo desempenho segundo a correlação de Spearman, pois o mínimo valor encontrado foi 0,84 com o método ALS(1), isto indica uma alta correlação positiva entre as imputações e os valores originais do experimento. Novamente o desvio padrão é bem pequeno para todos os métodos e a média e a mediana da correlação de Spearman é bem alta. Finalmente se apresentam os resultados com a maior porcentagem de perda, ou seja 30% de observações.

Tabela 13 - Estatísticas da correlação de Spearman imputando 20% de dados

	Métodos de Imputação				
Estatísticas	IMLD	ALS(1)	ALS(0)	r-AMMI2	r-AMMI1
Média	0,9377	0,9474	0,9555	0,9536	0,9551
Desvio padrão	0,0144	0,0146	0,0103	0,0106	0,0103
Mínimo	0,8715	0,8459	0,8983	0,9035	0,8991
Quartil 1 (Q1)	0,9290	0,9405	0,9495	0,9473	0,9492
Mediana	0,9387	0,9492	0,9564	0,9545	0,9561
Quartil 3 (Q3)	0,9480	0,9568	0,9628	0,9609	0,9623
Máximo	0,9734	0,9797	0,9780	0,9792	0,9797
Q3-Q1	0,0190	0,0163	0,0133	0,0135	0,0131

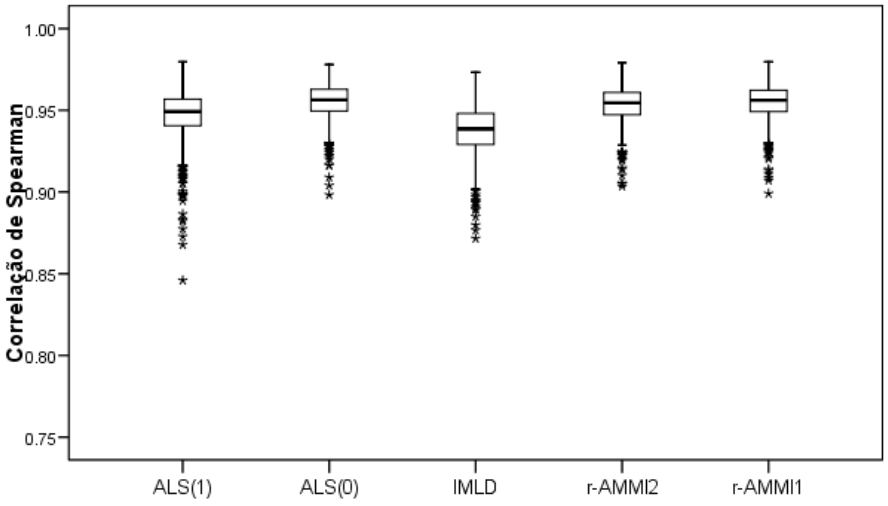

Figura 5 - Distribuição do coeficiente de correlação de Spearman imputando 20% de dados

Conclui-se através da figura 6 e da tabela 14 que o aumento na porcentagem da informação perdida não está associada com a correlação entre imputações e valores reais do experimento considerado para o estudo de simulação. Todos os métodos em todas as porcentagens de perda apresentam uma alta correlação positiva entre o valor imputado e o valor original. De acordo com os resultados apresentados qualquer método pode oferecer um bom desempenho segundo a estatística de Spearman, mas se além disso se leva em conta a estrutura do algoritmo, o método escolhido deveria ser o ALS(0), pois apresenta a estrutura mais simples.

Tabela 14 - Estatísticas da correlação de Spearman imputando 30% de dados

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

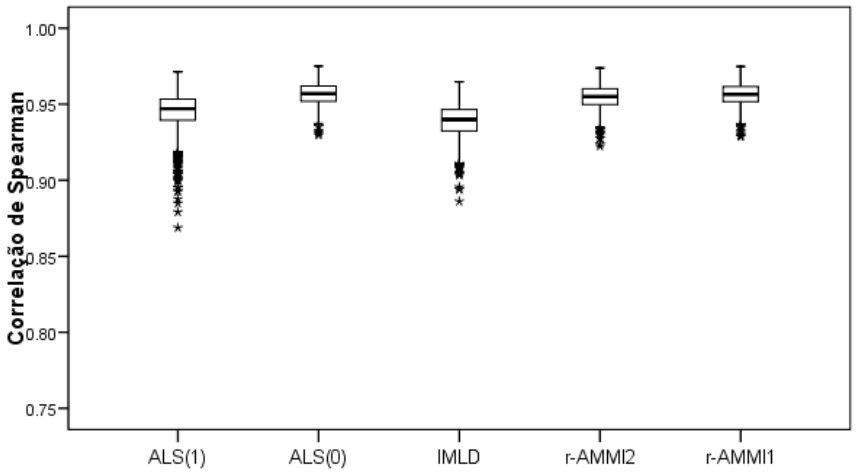

Figura 6 - Distribuição do coeficiente de correlação de Spearman imputando 30% de dados

2.6.5 Escolha dos modelos AMMI nos conjuntos de dados completados (observados + imputados)

Anteriormente foi encontrado na análise do conjunto de dados original que para explicar a estrutura da interação era necessário um modelo AMMI4 segundo o critério de Cornelius e um modelo AMMI2 segundo o método de máxima verossimilhança. Nesta seção se comparou os diferentes métodos de imputação levando em conta o número de componentes multiplicativos de um modelo AMMI para as matrizes de dados completadas (observados+imputados). Assim, em cada um dos três mil conjuntos de dados gerados com as diferentes porcentagens de perda aleatória foram aplicados os métodos de imputação e posteriormente feitos os testes respectivos para determinar o número de componentes multiplicativos. O objetivo desta análise, além de comparar os algoritmos de predição, é estabelecer se é recomendável fazer a escolha do número de componentes AMMI sobre as matrizes completadas.

Na tabela 15 se apresenta o resultado do número de componentes encontrado sobre as matrizes completadas, depois da imputação do 10% de informação. Observa-se que nos 1000 conjuntos de dados com 10% de imputação através de ALS(0), ALS(1), r-AMMI1 e r-AMMI2 o modelo que mais número de vezes foi escolhido foi o AMMI2, seguido do modelo AMMI1. Quando foram imputados os dados ausentes com ALS(0), em 533 vezes o AMMI2 foi recomendado. Esse mesmo modelo foi recomendado em 529 ocasiões imputando por ALS(1), em 556 conjuntos imputando por r-AMMI1 e em 688 imputando por r-AMMI2. Caso diferente aconteceu com o método de imputação IMLD, pois o modelo que foi escolhido o maior número de vezes quando se imputou por esse método foi o modelo AMMI3 (208

Tabela 15 - Modelos AMMI escolhidos através do método de máxima verossimilhança nos 1000 conjuntos de dados com imputação do 10%

	Métodos de imputação				
Modelo	ALS(0)	IMLD	ALS(1)	r-AMMI1	r-AMMI2
AMMI0	17	0	0	1	0
AMMI1	189	25	244	213	68
AMMI2	533	194	529	556	688
AMMI3	132	208	107	104	118
AMMI4	6	114	9	7	12
AMMI5	7	107	8	7	11
AMMI6	3	43	2	2	3
AMMI7	0	14	0	0	2
AMMI8	0	19	0	0	0
AMMI9	3	17	7	5	5
AMMI10	10	40	10	11	9
AMMI11	12	51	12	12	13
AMMI12	38	66	33	36	34
AMMI13	50	102	39	46	37

vezes), o segundo modelo mais escolhido em conjuntos de dados com predições por IMLD foi AMMI2 em 194 vezes. Note-se que o método de máxima verossimilhança foi bastante liberal, pois por exemplo em 102 conjuntos de dados com predições IMLD o modelo AMMI13 foi selecionado.

Continuando com a análise no estudo de simulação, apresentam-se os resultados para a porcentagem de perda 20%. Na tabela 16 se observa que aumentando a porcentagem de retirada a 20%, o número de vezes que é recomendado o modelo AMMI2 diminui em matrizes com elementos imputados pelos métodos ALS(0), ALS(1) e r-AMMI1. Por exemplo com o 10% de imputação por ALS(0) o modelo AMMI2 foi recomendado 533 vezes, agora com 20% de informação imputada pelo mesmo método o AMMI2 é recomendado em 364 conjuntos de dados. Caso contrário acontece com o método r-AMMI2, pois com 10% de dados imputados o modelo AMMI2 foi selecionado em 688 ocasiões e agora, com o 20% de imputação o número de ocasiões aumenta para 696.

Esses resultados fazem concluir que para tomar alguma decisão sobre a estrutura da interação em experimentos desbalanceados é preferível escolher o número de componentes multiplicativos sobre dados que não contenham nenhum tipo de imputação. Aqui o número de componentes multiplicativos está relacionado diretamente com o modelo de imputação usando mínimos quadrados alternados ou sub-modelo robustos. Um resultado totalmente diferente

Tabela 16 - Modelos AMMI escolhidos através do método de máxima verossimilhança nos 1000 conjuntos de dados com imputação do 20%

	Métodos de imputação				
Modelos AMMI	ALS(0)	IMLD	ALS(1)	r-AMMI1	r-AMMI2
AMMI0	56	0	0	0	0
AMMI1	232	2	348	321	21
AMMI2	364	12	370	412	696
AMMI3	204	63	139	139	137
AMMI4	35	99	31	33	32
AMMI5	20	162	14	19	21
AMMI6	3	136	8	4	7
AMMI7	5	90	5	4	1
AMMI8	2	70	4	3	5
AMMI9	4	71	4	3	9
AMMI10	8	79	9	7	7
AMMI11	14	73	19	13	19
AMMI12	28	61	26	21	18
AMMI13	25	82	23	21	27

foi encontrado quando se fizeram as imputações com o IMLD, porque esse método não leva em conta modelos AMMI para a imputação, por isso com 20% de retirada aleatória os modelos mais vezes escolhidos foram AMMI5 e AMMI6. Para confirmar o descrito anteriormente se apresentam os resultados com 30% de dados imputados.

Na tabela 17, observa-se que com 30% de imputação de dados com ALS(0), ALS(1) e r-AMMI1, os modelos mais vezes escolhidos foram AMMI1, AMMI2 e AMMI3. Nos conjuntos de dados completados através de imputações com r-AMMI2 o modelo mais escolhido foi o AMMI2 (682 vezes). Quando foram imputadas as observações com IMLD os modelos mais selecionados foram AMMI6, AMMI7, AMMI8 e AMMI9. Com esta tabela se confirma que não é recomendável tomar alguma decisão sobre a interação baseado em dados imputados, porque essa decisão dependerá muito do método de predição de observações ausentes.

Levando em conta todo o exposto anteriormente, se se tem um conjunto de dados com dados faltantes e se deseja analizar a interação através de um modelo AMMI, o procedimento sugerido é: escolher o modelo baseado só na informação observada e depois de conhecer o número de componentes multiplicativos para explicar a interação, imputar os dados ausentes e sobre a matriz completada (observados+imputados) fazer a estimação dos parâmetros.

Inicialmente a escolha do modelo AMMI nos conjuntos de dados completados se fez por máxima verossimilhança, esse método utiliza somente os autovalores da matriz de interação.

Tabela 17 - Modelos AMMI escolhidos através do método de máxima verossimilhança nos 1000 conjuntos de dados com imputação do 30%

	Métodos de imputação				
Modelos AMMI	ALS(0)	IMLD	ALS(1)	r-AMMI1	r-AMMI2
AMMI0	51	0	0	0	0
AMMI1	261	0	393	363	2
AMMI2	304	4	308	326	682
AMMI3	203	6	141	144	154
AMMI4	67	33	43	49	35
AMMI5	22	76	23	29	29
AMMI6	7	131	10	10	12
AMMI7	8	155	9	6	8
AMMI8	8	149	10	9	6
AMMI9	4	116	7	3	7
AMMI10	14	90	6	11	8
AMMI11	13	89	13	5	7
AMMI12	18	79	16	16	21
AMMI13	20	72	21	29	29

O método foi aplicado diretamente sem fazer nenhuma correção pelos dados faltantes e além disso, esse método não considerou a estimativa da variância quando esta está disponível. Por essa razão foi feita novamente uma análise para determinar o melhor modelo AMMI, mas usando o sistema de Cornelius e corrigindo o quadrado médio do erro médio e seus respectivos graus de liberdade.

No conjunto de dados original com 15 genótipos, 27 ambientes e apenas uma observação por cada combinação genótipo×ambiente que correspondia à média de quatro repetições, também se tinha uma estimativa da variância $\widehat{QME} = 179189{,}8389$ com 1134 graus de liberdade associados, que foi calculada através da média dos quadrados médios do erro obtidos nas análises individuias, ou seja em cada ambiente. Como se tinham as médias por combinação de tratamentos e a informação do delineamento experimental, foi possível calcular a média geral do experimento resultando em $\overline{X}_{geral} = 3505{,}0866$. Com a média geral do experimento e com a estimativa da variância se calculou o coeficiente de variação para o experimento $CV = \sqrt{\widehat{QME}}/\ \overline{X}_{geral} = 0{,}1207$ (12,07%). Uma vez desenvolvido o anterior, explicar-se-a como se fez a análise corrigida usando o sistema de Cornelius sobre dados que continham imputações.

Por exemplo, considere-se a primeira retirada aleatória do 30% no conjunto de dados originais, ou seja, foram consideradas 122 observações (de um total de 405) como ausentes.

Após a retirada foram imputados o dados pelos cinco método já expostos: ALS(0), ALS(1), IMLD, r-AMMI1 e r-AMMI2, obtendo cinco conjuntos de dados completados (observados + imputados) e sobre cada um desses foi aplicado o sistema de Cornelius apresentado na tabela 1, mas para testar as hipóteses se calculou um novo número de graus de liberdade do Erro médio/n,

$$\text{G.L.(Erro médio}/n) = \sum_{j=1}^{b} GL_{Erroj}$$

$$GL_{Erroj} = [t - 1 - \text{Faltantes}_j]\,(n-1)$$

em que j representa os ambientes, $t = 15$ genótipos, Faltantes_j é o número de médias de tratamentos ausentes no ambiente j e $n = 4$ é o número de repetições.

Além disso, foi feita uma correção no valor do QM(Erro médio/n), encontrando em cada conjunto de dados completado uma nova média geral do experimento e multiplicando essa média pelo CV dos dados originais, o resultado foi a nova estimativa do QM(Erro médio/n) para testar as hipóteses. Isto faz sentido porque se assume que os efeitos de tratamentos e os efeitos do erro são independentes. Todo o processo anteriormente exposto foi aplicado nas outras retiradas do 30% e da mesma maneira nas outras porcentagens de perda aleatória (20% e 30%). Em seguida se apresentam os resultados das análises corrigidas no estudo de simulação usando o sistema Cornelius para os conjuntos completados (observados+imputados) na escolha do melhor modelo AMMI.

Tabela 18 - Modelos AMMI escolhidos através do método de Cornelius nos 1000 conjuntos de dados com imputação do 10%

	Métodos de imputação				
Modelos AMMI	ALS(0)	IMLD	ALS(1)	r-AMMI1	r-AMMI2
AMMI0	0	0	0	0	0
AMMI1	0	0	0	0	0
AMMI2	46	18	144	104	225
AMMI3	879	826	830	869	770
AMMI4	75	156	26	27	5

Observa-se na tabela 18 que o sistema de Cornelius e bem menos liberal do que o método de máxima verossimilhança, pois o número máximo de componentes selecionadas sobre os conjuntos de dados com o 10% de informação imputada foi quatro. Nos 1000 conjuntos de

dados considerados para essa porcentagem de imputação, o modelo mais vezes escolhido foi o AMMI3. Quando se imputou por ALS(0) o modelo AMMI3 foi escolhido em 879 vezes, seguido pelo método r-AMMI1 em 869 ocasiões.

Tabela 19 - Modelos AMMI escolhidos através do método de Cornelius nos 1000 conjuntos de dados com imputação do 20%

	Métodos de imputação				
Modelo	ALS(0)	IMLD	ALS(1)	r-AMMI1	r-AMMI2
AMMI0	0	0	0	0	0
AMMI1	44	1	307	175	142
AMMI2	787	534	684	795	858
AMMI3	169	464	9	30	0
AMMI4	0	1	0	0	0

Entretanto, nos 1000 conjuntos com imputação do 20% (tabela 19) o modelo mais escolhido foi o AMMI2. Com o método r-AMMI2 esse modelo foi escolhido 858 vezes e 534 com IMLD. Os modelos AMMI0 e AMMI4 foram os menos selecionados para essa porcentagem. Nos conjuntos com dados imputados por IMLD se apresentou a maior frequencia para o modelo AMMI3 (464 vezes) comparando-o com os outros métodos de predição de observações ausentes.

Tabela 20 - Modelos AMMI escolhidos através do método de Cornelius nos 1000 conjuntos de dados com imputação do 30%

	Métodos de imputação				
Modelos AMMI	ALS(0)	IMLD	ALS(1)	r-AMMI1	r-AMMI2
AMMI0	58	0	0	27	5
AMMI1	713	57	995	933	946
AMMI2	229	840	5	40	49
AMMI3	0	103	0	0	0
AMMI4	0	0	0	0	0

Finalmente se apresenta na tabela 20 o resultado quando se imputou 30% dos dados. Observa-se que nessa porcentagem o modelo mais escolhido foi o AMMI1 nos conjuntos com dados imputados por ALS(0), ALS(1), r-AMMI1 e r-AMMI2. Enquanto com IMLD o resultado foi diferente, pois o modelo mais escolhido foi AMMI2 em 840 vezes. Nesta porcentagem se esperava que nos conjuntos com imputações através de r-AMMI2 o modelo mais selecionado fosse o AMMI2, mas em apenas 49 conjuntos isso aconteceu.

3 CONCLUSÕES

De acordo com os objetivos deste estudo, pode-se concluir:

- segundo a RMSPD padronizada usada no estudo de simulação, os melhores métodos para imputar foram ALS(0) e r-AMMI1. Esses métodos apresentam os melhores resultados considerando a média e a mediana da distribuição, além de minimizar a dispersão. Baseado nesses critérios encontrou-se que a predição de interações em experimentos G×E com r-AMMI1 ou ALS(0) é mais recomendável do que a imputação com ALS(1), IMLD e r-AMMI2.

- segundo a estatística de Procrustes, a melhor similaridade entre as matrizes completadas por imputação e a matriz de dados original do experimento de algodão foi obtida através de ALS(0).

- todos os métodos estudados neste trabalho apresentaram uma alta correlação entre o dado imputado e o correspondente dado real no experimento.

- sobre um conjunto de dados real com observações ausentes recomenda-se determinar o número de componentes multiplicativos em um modelo AMMI a partir da informação observada e estimar os parâmetros desse modelo da maneira usual completando a informação através de imputação.

- a metodologia proposta recentemente de predizer interações com a média dos valores obtidos por imputação múltipla livre de distribuição não mostrou melhores resultados do que as predições considerando apenas un modelo aditivo, ou seja ALS(0).

Imputar com métodos que apenas consideram estimativas aditivas parece insatisfatório, pois ignorar a interação para as predições nesse tipo de experimentos pode ser visto como inapropiado, por isso para dar uma continuidade deste estudo seria interessante fazer imputação de dados incorporando covariáveis genotípicas ou ambientais, como por exemplo a temperatura e os índices de chuva ou umidade. A regressão feita sobre covariáveis pode explicar uma parte importante da interação e isso pode ser usado para imputar combinações G×E ausentes.

REFERÊNCIAS

ALLAN, F.E.; WISHART, J. A method of estimating the yield of a missing plot in field experimental work. In: DODGE Y. **Analysis of experiments with missing data**. New York: John Wiley. 1985. chap. 5, p. 93-162.

BARTLETT, M.S. Some examples of statistical methods of research in agriculture and applied biology. In: DODGE Y. **Analysis of experiments with missing data**. New York: John Wiley. 1985. chap. 5, p. 93-162.

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

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

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

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.

CHAVES, J.L. Interação de cultivares com ambientes. In: NASS, L.L.; VALOIS, A.C.C.; MELO, I.S.; VALADARES, M.C. **Recursos genéticos e melhoramento de plantas**. Rondonópolis: Fundação MT, 2001. p.673-713

CHAVES, L.J.; VENCOVSKY, R.; GERALDI, I.O. Modelo não linear aplicado ao estudo da interação de genótipos × ambientes em milho. **Pesquisa Agropecuária Brasileira**, v.24, n.2, p. 259-269, 1989.

COCHRAN, W.G.; COX, G. **Experimental designs.** New York: John Wiley, 1957. 611p.

COCKERHAM, C.C. Estimation of genetics variance. In: HANSON, W.D.; ROBINSON, H.F. Ed. **Statistical genetics and plant breeding**. Madison: National Academy of Sciences, 1963. chap. 2, p.53-94.

CORNELIUS, P.L.; CROSSA J.; SEYEDSADR M.S. Statistical tests and estimators of multiplicative models for genotype-by-environment interaction. In: KANG, M.S.; GAUCH, H.G. **Genotype-by–environment interaction**. Boca Raton: CRC Press, 1996. chap. 8, p.199-234.

CORNISH, E.A. The estimation of missing values in quasi-factorial designs. In: DODGE Y. **Analysis of experiments with missing data**. New York: John Wiley, 1985. chap. 5, p. 93-162.

_____. The analysis of quasi factorial designs with incomplete data: lattice squares, 1941. In: DODGE Y. **Analysis of experiments with missing data**. New York: John Wiley, 1985. chap. 5, p. 93-162.

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

CRUZ, C.D.; REGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético.** Viçosa: UFV, 1994. 390p.

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. **Métodos para a escolha de componentes em modelo de efeito principal aditivo e interação multiplicativa**. 73 p. 2005. Tese (Livre-Docencia no Departamento de Ciências Exatas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2005.

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.

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

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

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

FALCONER, D.S. **Introduction to quantitative genetics**. Harlow: Longman, 1989, 438p.

FALCONER, D.S; MACKAY, T.F.C. **Introduction to quantitative genetics.** Harlow: Longman, 1996, 446p.

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.

GAUCH, H.G. Model selection and validation for yield trials with interactions. **Biometrics**, Washington, v.44, n.3, p. 705-715, 1988.

GAUCH, H.G.; ZOBEL, R.W. Predictive ans postdictive sucess of statistical analysis of yield trials. In: KANG, M.S.; GAUCH, H.G. **Genotype-by-environment interaction**, Boca Raton: CRC Press, 1996. chap. 8. p. 199-234.

_____. 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.

GOLLOB, H.F. A statistical model which combines features of factor analitic and analysis of variance techniques. **Psychometrika**, Colorado Springs, v.33, n.1, p.73-115, 1968.

HARTLEY, H.O. A plan for programming analysis of variance for general purpose computers. **Biometrics**, Washington, v.12, n.2, p.110-122, 1956.

HEALY, M; WESTMACOTT, M. Missing values in experiments analyzed on automatic computers. In: LITTLE, R. J; RUBIN D.B. **Statistical analysis with missing data**. 2nd ed. New York: John Wiley, 2002. chap. 2, p.24-40.

KANG, M.S.; MAGARI, R. New developments in selecting for phenotypic stability in crop breeding. In: DUARTE, J.B.; VENCOVSKY, R. **Interação genótipo × ambiente:** uma introdução à análise "AMMI". Riberão Preto: Sociedade Brasileira de Genética, 1999. 60p. (Série Monografías).

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.

LI, C.C. Analysis of unbalanced data. A pre-program introduction. In: DODGE Y. **Analysis of experiments with missing data**. New York: John Wiley. 1985. chap. 5, p. 93-162.

LIU, G.; CORNELIUS, P. L. Simulations and derived approximations for the means and standard deviations of the characteristic roots of a wishart matrix. **Communications in Statistics - Simulation and Computation**, London, v.30, n.4, p.963-989. 2001.

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

MANDEL, J. A new analysis of variance for non-additive data. **Technometrics**, Alexandria, v.13, n.1, p.1-18, 1971.

MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. **Multivariate analysis**. Amsterdam: Academic Press. 1979. 521p.

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

PIEPHO, H.P. Robustness of statistical test for multiplicative terms in the additive main effects and multiplicative interaction model for cultivar trial. **Theoretical and Applied Genetics**, New York, v.90, p.438-443, 1995a.

_____. 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, 1995b.

RAMALHO, M.A.P.; SANTOS, J.B.; ZIMMERMANN, M.J.O. **Genética quantitativa em plantas autógamas**: aplicações ao melhoramento do feijoeiro. Goiânia: UFG, 1993. 271p.

RUBIN, D.B. A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. **Applied Statistics**, London, v.21, n.2, p.136-141, 1972.

_____. Inference and missing data. **Biometrika**, Oxford, v.63, n. 3, p.581-592, 1976.

_____. Multiple imputation after 18+ years. **Journal of the American Statistical Association**, Alexandria, v.91, n.434, p.473-489, 1996.

SAS INSTITUTE **SAS/IML 9.1 User's guide**. Carey: SAS Institute Inc., 2004. 1040p.

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.

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.

YATES, F. The analysis of replicated experiments when the field results are incomplete. In: DODGE Y. **Analysis of experiments with missing data**. New York: John Wiley, 1985. chap. 5, p. 93-162.

APÊNDICES

APÊNDICE A - Conjunto de dados original

Tabela 21 - Produtividade média de algodão em caroço (kg/ha) obtidas no ensaio estadual de algodoeiro herbáceo (FARIAS, 2005)

	Genótipos														
Ambientes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	2810,75	3048,20	2701,87	2456,45	3301,00	2857,10	3407,08	3159,21	2814,97	3073,83	2441,75	2882,77	3257,51	3052,28	3318,45
2	2946,10	3732,32	3387,70	3470,29	3651,41	3173,62	3396,99	3066,83	2912,82	3117,24	2465,53	3163,18	3339,86	3417,05	3464,04
3	4096,82	4641,21	4799,52	4775,91	5148,10	4968,95	4275,96	4128,76	4789,80	4989,78	3403,00	5031,45	4817,58	4331,52	4753,70
4	6017,46	5645,27	6643,78	6467,41	6411,86	6231,32	5675,83	6993,75	5974,41	5678,60	6129,95	6381,31	6107,73	6238,27	6354,93
5	5916,08	5621,66	6838,21	6518,80	6474,40	6335,48	5779,98	6236,88	6282,71	6297,98	6834,04	6027,18	6063,28	5896,64	6793,77
6	4838,41	5085,60	5391,13	4720,36	5659,16	4206,52	5246,70	5836,92	5371,67	5611,94	5423,07	5250,87	5110,60	5416,13	5636,93
7	2490,03	2187,29	2484,48	2406,71	2733,06	2576,13	2238,67	2174,79	2281,72	2608,07	2215,06	2035,91	2351,16	2562,25	2669,17
8	5088,38	4655,09	4132,92	5331,41	4842,57	4800,91	4681,48	4948,12	4803,68	4766,19	5310,58	4432,89	4503,72	4960,61	4893,96
9	2184,50	1852,59	2080,35	1762,32	2470,59	2208,11	2449,20	2073,40	2653,90	2006,74	2091,45	2371,98	2415,04	2512,25	2399,76
10	2791,39	2585,85	2996,92	2427,53	2913,60	3003,86	2456,70	2963,59	3062,19	2680,29	3098,30	2328,93	2763,61	2574,74	3019,14
11	3287,72	3796,56	3038,86	3399,38	3728,23	3550,20	3066,63	3739,07	3728,23	3754,62	3763,51	3741,57	3934,60	3764,62	3588,25
12	897,13	980,46	873,94	856,86	905,46	937,54	740,76	617,16	890,05	929,63	962,40	673,54	747,56	872,97	851,16
13	2444,20	2930,26	2280,32	3049,69	2619,18	1308,20	2535,85	2477,53	2583,07	2208,11	2069,23	2249,77	2074,79	2835,82	1833,15
14	3963,84	4225,96	2737,55	3908,38	4014,04	3494,50	3429,29	3855,30	3862,83	4174,58	4065,89	3154,95	3323,41	3981,48	4220,15
15	4678,02	4455,50	4746,80	4343,27	4612,80	4199,55	4764,42	5136,75	4795,30	4800,75	4290,37	4350,17	4538,36	4444,02	4861,52
16	3372,71	3507,02	3436,02	3860,57	3776,46	3944,35	3410,49	3884,27	3454,02	3942,41	3198,77	3447,94	3667,99	3517,85	3369,13
17	4452,96	3800,32	5456,99	4164,26	4602,24	4914,96	4226,99	4467,27	3983,51	4402,76	4303,26	4269,68	4540,65	4135,77	4518,88
18	3886,25	4892,50	4486,25	4562,50	4976,25	4250,50	4938,75	4746,25	4726,25	4526,25	4537,50	4110,00	4501,25	4811,25	4918,75
19	3202,50	3626,25	3601,25	3500,00	3306,25	3636,25	3682,50	3467,50	3676,25	3746,25	3501,25	3652,50	3585,00	3522,50	3222,50
20	2660,75	2583,00	2457,00	2738,50	2674,50	2783,00	2874,50	2691,50	2944,25	2932,75	2802,50	2397,00	2474,75	2541,25	2488,50
21	2085,90	2654,18	2147,00	2159,22	2149,51	2207,28	2401,14	2058,68	2162,28	2435,86	2241,99	2168,95	2231,16	2223,38	2445,31
22	2425,42	2160,92	2697,06	1842,09	2347,57	1990,24	2353,57	2223,27	2309,96	2200,02	2106,40	1935,80	2070,95	2277,74	2542,58
23	4207,91	4543,99	3241,34	3980,15	3666,30	3649,63	3999,60	3621,86	4080,14	3588,53	3510,76	4232,91	3944,05	3755,18	3955,16
24	2686,00	2616,25	2843,50	2626,25	2170,00	1943,75	2202,50	2372,50	2382,00	2417,50	2190,00	2317,50	2476,25	2455,00	2682,50
25	2842,50	2786,25	2820,00	2780,00	2104,50	2205,00	2887,50	2895,00	2910,00	2581,25	2895,00	2716,25	2932,50	2643,75	2805,00
26	1756,25	1771,25	1841,25	1795,00	2057,50	2006,25	2010,00	1800,00	2030,00	1787,50	1857,50	1965,00	1937,50	1891,25	1901,25
27	4531,49	3969,04	4899,51	4241,52	4728,97	4272,07	4489,55	4644,53	4561,21	4350,95	4797,29	4487,05	4508,15	4770,91	4779,24

APÊNDICE B - Quadrados médios do erro das análises individuais

Tabela 22 - Quadrados médios do erro (QME) por ambiente (FARIAS, 2005)

Ambiente	QME	Ambiente	QME	Ambiente	QME
1	253587,47	10	127382,04	19	118231,11
2	148478,15	11	145679,20	20	117460,84
3	234382,30	12	41095,06	21	68923,84
4	373925,48	13	274872,21	22	63855,08
5	476324,06	14	144252,00	23	433052,14
6	300225,11	15	186198,44	24	17855,75
7	108692,29	16	299868,06	25	31589,08
8	212577,78	17	179556,69	26	15213,90
9	208189,42	18	167641,22	27	89016,93

APÊNDICE C - Comparações múltiplas não paramétricas para a RMSPD padronizada entre os métodos de imputação

Tabela 23 - Diferenças das R_i para 10% de retirada dos dados

Métodos	ALS(1)	IMLD	r-AMMI2	r-AMMI1
ALS(0)	1477	2807	578	52
ALS(1)	-	1330	899	1529
IMLD	-	-	2229	2859
r-AMMI2	-	-	-	630

Diferença mínima significativa ($\alpha = 0{,}05$) = 192,9
R_i é a soma das ordens atribuídas ao método i nos 1000 conjuntos de dados simulados.
i = ALS(0), ALS(1), IMLD, r-AMMI2, r-AMMI1.

Tabela 24 - Diferenças das R_i para 20% de retirada dos dados

Métodos	ALS(1)	IMLD	r-AMMI2	r-AMMI1
ALS(0)	2268	3131	985	121
ALS(1)	-	863	1283	2147
IMLD	-	-	2146	3010
r-AMMI2	-	-	-	864

Diferença mínima significativa ($\alpha = 0{,}05$) = 192,9
R_i é a soma das ordens atribuídas ao método i nos 1000 conjuntos de dados simulados.
i = ALS(0), ALS(1), IMLD, r-AMMI2, r-AMMI1.

Tabela 25 - Diferenças das R_i para 30% de retirada dos dados

Métodos	ALS(1)	IMLD	r-AMMI2	r-AMMI1
ALS(0)	2577	3247	1210	181
ALS(1)	-	670	1367	2396
IMLD	-	-	2037	3066
r-AMMI2	-	-	-	1029

Diferença mínima significativa ($\alpha = 0{,}05$) = 192,9
R_i é a soma das ordens atribuídas ao método i nos 1000 conjuntos de dados simulados.
i = ALS(0), ALS(1), IMLD, r-AMMI2, r-AMMI1.

APÊNDICE D - Programa para gerar os resultados da simulação fazendo retiradas aleatórias de 10 % e imputando com IMLD

```
dm 'log;clear;output;clear';
PROC IML;
PORCENTP=10;
REPETICOES=1000;
SIMULACAO_IMMULTIPLA=J(REPETICOES,7,0);
ORIGINAL={
2810.75 3048.2 2701.87, ...,
,..., 4508.15 4770.91 4779.24
};
KR=ORIGINAL;
NUMFIL=NROW(KR);
NUMCOL=NCOL(KR);
NUMPERD1=(NUMFIL*NUMCOL)*PORCENTP/100;
NUMPERD=CEIL(NUMPERD1);
DO INDREP=1 TO REPETICOES;
MISSING=KR;
DO INDPERD=1 TO NUMPERD;
PERDER: FILAP=INT(RANUNI(2147483638-INDREP)*NUMFIL)+1;
COLUMP=INT(RANUNI(4411600+INDREP)*NUMCOL)+1;
IF (MISSING[FILAP,COLUMP]^=.) THEN MISSING[FILAP,COLUMP]=.;
ELSE GOTO PERDER;
END;/*DO INDPERD=1 TO NUMPERD;*/
INDQUALIDADE=SSQ(MISSING);
/*PRINT MISSING;*/
LOST=MISSING;
Y=LOST;
/* NI É O NÚMERO DE ITERACÕES */
NI=50;
UI=shape(0,NI,1);
do I=1 to NI; UI[I]=I; end;
SOMA=SHAPE(0,NI,1);
/*PRINT SOMA, UI;*/
/* CALCULA O NÚMERO DE LINHAS E COLUNAS DOS DADOS */
L=NROW(Y);
```

```
C=NCOL(Y);
/* DETERMINA O NÚMERO DE OBSERVACOES PRESENTES (NAO AUSENTES)
E COLOCA NA MATRIZ NO */
YM=Y=.;
UM=j(L,L,1);
NO=UM*YM;
NO=L-NO;
do I=1 to C;
if NO[1,I]=1
then
do;
print "VOCE TEM DESVIO PADRAO INFINITO";
ABORT;
end; /*if NO[1,I]=1*/
end; /*do I=1 to C*/
/*PRINT YM ,UM, NO;*/
/* LOCALIZA AS LINHAS E COLUNAS DOS VALORES AUSENTES NA MATRIZ Y*/
z=loc(nmiss(y));
ya=int((z+C-1)/C);
yb=mod(z,C);
zcol=ncol(z);
if (zcol > 0) then do;
do i=1 to zcol;
if yb[,i]=0 then yb[,i]=C;
end;
end;
ymiss=t(ya)||t(yb);
L1=NROW(ymiss);
/*PRINT YMISS , L1;*/
/* CRIA A MATRIZ COM TODAS AS IMPUTACOES MÚLTIPLAS */
NL=nrow(LOST);
NC=ncol(LOST);
IMPUTMULT=j(5*NL,NC+1,0);
AUX1=-NL+1;
AUX2=0;
AUX3=1;
```

```
do i=1 to 5;
IMPUTMULT[NL*i-(NL-1):NL*i,1]=i;
end;
/*PRINT IMPUTMULT;*/
/* OBTEM AS IMPUTACOES MULTIPLAS VARIANDO O NUMERADOR (numer) DA EXPRESSAO
YE_=UT*(DT##(numer/20)*DBU##(numer/20))*VBLU */
do numer=8 to 12;
/*print numer;*/
AUX1=AUX1+NL;
AUX2=AUX2+NL;
/* K É O ÍNDICE DAS ITERACOES E
P O ÍNDICE DA LINHA NAS POSICOES DOS VALORES
AUSENTES */
DO K=1 TO NI;
/* CALCULA A MÉDIA DAS COLUNAS DE Y */
YTOT=REPEAT(Y[+,],L,1);
YBAR=YTOT/NO;
/* SUBSTITUI OS VALORES AUSENTES PELAS RESPECTIVAS MÉDIAS, EM Y SE K=1 */
IF K=1 THEN
DO;
DO I=1 TO L1;
Y[ymiss[I,1],ymiss[I,2]]=YBAR[ymiss[I,1],ymiss[I,2]];
END;
YM=Y;
END;/*IF K=1*/
SOMAP=0;
DO P=1 TO L1;
/* PADRONIZA A MATRIZ Y, COM AS MÉDIAS DAS COLUNAS NAS POSICOES(1a. VEZ) OU
ESTIMATIVAS */
UM=j(L,L,1);
YBAR=(1/L)*UM*Y;
DIF=Y-YBAR;
DIF2=DIF##2;
STD=UM*DIF2;
STD=(STD/(L-1))##0.5;
YS=(Y-YBAR)/STD;
```

```
/* DEFINE O NÚMERO DA LINHA E DA COLUNA DO VALOR AUSENTE */
EL=ymiss[P,1];
EC=ymiss[P,2];
/* FAZ A DECOMPOSICAO EM VALORES SINGULARES DA MATRIZ DOS DADOS Y */
CALL SVD(U1,D1,V1,YS);
D=DIAG(D1);
U=U1[,1:C];
VL=T(V1[,1:C]);
A=U*D*VL;
/* DEFINE A SUBMATRIZ Y_i, COM ELIMINACAO DA i-ÉSIMA LINHA DE Y */
IF (EL=1) THEN Y_i=Y[2:L,];
ELSE IF (EL=L) THEN Y_i=Y[1:L-1,];
ELSE Y_i=Y[1:EL-1,]//Y[EL+1:L,];
/* PADRONIZA A MATRIZ Y_i */
UM_i=j(L-1,L-1,1);
YBAR_i=(1/(L-1))*UM_i*Y_i;
DIF_i=Y_i-YBAR_i;
DIF2_i=DIF_i##2;
STD_i=UM_i*DIF2_i;
STD_i=(STD_i/(L-2))##0.5;
Y_i=(Y_i-YBAR_i)/STD_i;
/* DEFINE A SUBMATRIZ Y_j, COM ELIMINACAO DA j-ÉSIMA COLUNA DE Y */
IF (EC=1) THEN Y_j=Y[,2:C];
ELSE IF (EC=C) THEN Y_j=Y[,1:C-1];
ELSE Y_j=Y[,1:EC-1]||Y[,EC+1:C];
/* PADRONIZA A MATRIZ Y_j */
UM_j=j(L,L,1);
YBAR_j=(1/L)*UM_j*Y_j;
DIF_j=Y_j-YBAR_j;
DIF2_j=DIF_j##2;
STD_j=UM_j*DIF2_j;
STD_j=(STD_j/(L-1))##0.5;
Y_j=(Y_j-YBAR_j)/STD_j;
/* FAZ A DECOMPOSICAO EM VALORES SINGULARES DA MATRIZ Y_i */
CALL SVD(UB,DB1,VB,Y_i);
DB=DIAG(DB1);
```

```
UB=UB[,1:C];
VBL=T(VB[,1:C]);
AY_i=UB*DB*VBL;
/* FAZ A DECOMPOSICAO EM VALORES SINGULARES DA MATRIZ Y_j */
CALL SVD(UT,DT1,VT,Y_j);
DT=DIAG(DT1);
UT=UT[,1:C-1];
VTL=T(VT[,1:C-1]);
AY_j=UT*DT*VTL;
/* TROCA SINAIS DA MATRIZ UT PARA QUE FIQUEM IGUAIS AOS DA MATRIZ U */
SU=0; SUT=0;
DO I=1 TO L;
DO J=1 TO (C-1);
IF U[I,J]<0 THEN SU=-1;
ELSE IF U[I,J]>0 THEN SU=1;
IF UT[I,J]<0 THEN SUT=-1;
ELSE IF UT[I,J]>0 THEN SUT=1;
IF SU ^= SUT THEN UT[I,J]=-UT[I,J];
END;
END;
/* TROCA SINAIS DA MATRIZ VBL PARA QUE FIQUEM IGUAIS AOS DA MATRIZ VL */
SU=0; SUT=0;
DO I=1 TO C;
DO J=1 TO C;
IF VL[I,J]<0 THEN SU=-1;
ELSE IF VL[I,J]>0 THEN SU=1;
IF VBL[I,J]<0 THEN SUT=-1;
ELSE IF VBL[I,J]>0 THEN SUT=1;
IF SU ^= SUT THEN VBL[I,J]=-VBL[I,J];
END;
END;
/* ELIMINA ÚLTIMO ELEMENTO DE DB ÚLTIMA LINHA DE VBL */
DBU=DIAG(DB1[1:C-1,]);
VBLU=VBL[1:C-1,];
/* CALCULA AS IMPUTACOES MÚLTIPLAS */
YE_=UT*(DT##(numer/20)*DBU##((20-numer)/20))*VBLU;
```

```
YE_=YBAR+STD#YE_;
/* COLOCA AS ESTIMATIVAS NAS RESPECTIVAS POSICOES EM YM */
YM[ymiss[P,1],ymiss[P,2]]=YE_[ymiss[P,1],ymiss[P,2]];
END;/*P*/
/*PRINT K YE_;*/
/* COLOCA EM Y A MATRIZ YM COM OS VALORES AUSENTES ESTIMADOS E NA ESCALA
ORIGINAL */
Y=YM;
END;/*K*/
/* CRIA UMA MATRIZ COM TODAS AS IMPUTACOES */
do j= 1 to NC;
IMPUTMULT[AUX1:AUX2,j+1]=Y[1:L,j];
end;
end;
/*PRINT IMPUTMULT;*/
NCOLIMPUMULT=NCOL(IMPUTMULT);
IMPUTMULT1=IMPUTMULT[1:L,2:NCOLIMPUMULT];
IMPUTMULT2=IMPUTMULT[1+L:L*2,2:NCOLIMPUMULT];
IMPUTMULT3=IMPUTMULT[1+(2*L):L*3,2:NCOLIMPUMULT];
IMPUTMULT4=IMPUTMULT[1+(3*L):L*4,2:NCOLIMPUMULT];
IMPUTMULT5=IMPUTMULT[1+(4*L):L*5,2:NCOLIMPUMULT];
IMMEDIA=(IMPUTMULT1+IMPUTMULT2+IMPUTMULT3+IMPUTMULT4+IMPUTMULT5)/5;
/***********************************************
*    ROTACAO PROCRUSTES          *
***********************************************/
IMPUTADOS=IMMEDIA;
NLINORIGINAL=NROW(ORIGINAL);
ORIGINALTOT=REPEAT(ORIGINAL[+,],NLINORIGINAL,1);
ORIGINALBAR=ORIGINALTOT/NLINORIGINAL;
ORIGINALCENTRADA=ORIGINAL-ORIGINALBAR;
NLINIMPUTADOS=NROW(IMPUTADOS);
IMPUTADOSTOT=REPEAT(IMPUTADOS[+,],NLINIMPUTADOS,1);
IMPUTADOSBAR=IMPUTADOSTOT/NLINIMPUTADOS;
IMPUTADOSCENTRADA=IMPUTADOS-IMPUTADOSBAR;
Z_PROCRUSTES=T(ORIGINALCENTRADA)*IMPUTADOSCENTRADA;
CALL SVD(V_PROCRU,GAMMA_PROCRU,U_PROCRU,Z_PROCRUSTES);
```

```
GAMMA_PROCRUSTES=DIAG(GAMMA_PROCRU);
ROTACAO_PROCRU=V_PROCRU*T(U_PROCRU);
R2_PROCRUSTES=  TRACE(ORIGINALCENTRADA*T(ORIGINALCENTRADA))
+TRACE(IMPUTADOSCENTRADA*T(IMPUTADOSCENTRADA))
    -(2*TRACE(GAMMA_PROCRUSTES));
/*****************************************************************
*            RMSPD (DIFERENCA PREDITIVA MÉDIA)                   *
*****************************************************************/
VETORIMPUTADOS=J(L1,1,0);
VETORORIGINAL=J(L1,1,0);
DO i_RMSPD=1 TO L1;
VETORIMPUTADOS[i_RMSPD,1]=IMMEDIA[YMISS[i_RMSPD,1],YMISS[i_RMSPD,2]];
VETORORIGINAL[i_RMSPD,1]=ORIGINAL[YMISS[i_RMSPD,1],YMISS[i_RMSPD,2]];
END;/*DO i_RMSPD=1 TO L1;*/
DIF_RMSPD=VETORIMPUTADOS-VETORORIGINAL;
NL_DIF=NROW(DIF_RMSPD);
SQ_DIF_RMSPD=SSQ(DIF_RMSPD);
RMSPD=SQRT(SQ_DIF_RMSPD/NL_DIF);
/*******************************************************
* CORRELACAO DE SPEARMAN       *
*******************************************************/
ORDENSORIGINAL=RANKTIE(VETORORIGINAL);
ORDENSIMPUTADOS=RANKTIE(VETORIMPUTADOS);
DIF_ORDENS=ORDENSORIGINAL-ORDENSIMPUTADOS;
SQ_DIF_ORDENS=SSQ(DIF_ORDENS);
N_SPEARMAN=NROW(DIF_ORDENS);
R_SPEARMAN=1-((6*SQ_DIF_ORDENS)/((N_SPEARMAN##(3))-N_SPEARMAN));
/*****************************************************************
* MILLIKEN-JOHNSON                          *
*****************************************************************/
MVIMPUTADOS=IMMEDIA;
**********************MATRIZ GXE_MV E
SOMAS DE QUADRADOS*****************************;
MEDLINHAS_MV=MVIMPUTADOS[,:];
MEDCOLUN_MV=MVIMPUTADOS[:,];
NGEN_MV=NROW(MVIMPUTADOS);
```

```
NAMB_MV=NCOL(MVIMPUTADOS);
MEDIA_MV=MVIMPUTADOS[:,:];
L_MV=MEDLINHAS_MV*J(NCOL(MEDLINHAS_MV),NAMB_MV,1);
C_MV=J(NGEN_MV,NROW(MEDCOLUN_MV),1)*MEDCOLUN_MV;
M_MV=J(NGEN_MV,NAMB_MV,1)*MEDIA_MV;
GXE_MV=MVIMPUTADOS-L_MV-C_MV+M_MV;
SQGXE_MV_MV=SSQ(GXE_MV);
SQGEN_MV=NAMB_MV*(SSQ(MEDLINHAS_MV-MEDIA_MV));
SQAMB_MV=NGEN_MV*(SSQ(MEDCOLUN_MV-MEDIA_MV));
GLGEN_MV=NGEN_MV-1;
GLAMB_MV=NAMB_MV-1;
GLGXE_MV=GLGEN_MV*GLAMB_MV;
QMGEN_MV=SQGEN_MV/GLGEN_MV;
QMAMB_MV=SQAMB_MV/GLAMB_MV;
QMGXE_MV=SQGXE_MV_MV/GLGXE_MV;
FGEN_MV=QMGEN_MV/QMGXE_MV;
VALORP_FGEN_MV=1-PROBF(FGEN_MV,GLGEN_MV,GLGXE_MV);
FAMB_MV=QMAMB_MV/QMGXE_MV;
VALORP_FAMB_MV=1-PROBF(FAMB_MV,GLAMB_MV,GLGXE_MV);
/*PRINT GXE_MV, MEDLINHAS_MV MEDIA_MV, MEDCOLUN_MV,QMGEN_MV QMAMB_MV
QMGXE_MV, FGEN_MV VALORP_FGEN_MV [FORMAT=12.4],
FAMB_MV VALORP_FAMB_MV [FORMAT=12.4];*/
******************DVS DA MATRIZ GXE_MV*******************************;
AUTOVALORES_MV=EIGVAL(GXE_MV*T(GXE_MV));
NNUL1_MV=AUTOVALORES_MV<-1E-6;
NNUL2_MV=AUTOVALORES_MV>1E-6;
NNUL_MV=NNUL1_MV+NNUL2_MV;
R_MV=(NNUL_MV^=0)[+,];
CALL SVD(EG_MV,VS_MV,EA_MV,GXE_MV);
U_MV=EG_MV[,1:R_MV];
S_MV=DIAG(VS_MV[1:R_MV,]);
VT_MV=EA_MV[,1:R_MV];
/*PRINT U_MV,AUTOVALORES_MV S_MV,VT_MV;*/
**************DESDOBRAMENTO DA SQ(GXE_MV)POR DVS*****************;
**************PROPORCOES*********************;
VS_MV2=VS_MV##2;
```

```
AUTV_MV=VS_MV2[1:R_MV];
SOMAAUTOVAL_MV=SUM(VS_MV2);
PROPOR_MV=INV(SOMAAUTOVAL_MV)*VS_MV2;
PROPORCAO_MV=PROPOR_MV[1:R_MV];
RESD1_MV=SOMAAUTOVAL_MV-VS_MV2[1:1,];
PACUM_MV=CUSUM(PROPORCAO_MV)*100;
LINH_MV=T(1:R_MV);
RESUMO_MV=LINH_MV||AUTV_MV||PROPORCAO_MV||PACUM_MV;
/*PRINT RESUMO_MV [FORMAT=12.4];*/
*************************************************************************
TESTE DE RAZÃO DE MAXIMA VEROSSIMILHANÇA (Milliken-Johnson, 1989)
com a aproximação F proposta por Cornelius et al.(1996)
*************************************************************************;
RESULTADOSMV=J(R_MV-1,5,0);
VETORTEMPO=J(2,1,0);
VETORTEMPO[1,1]=NAMB_MV;
VETORTEMPO[2,1]=NGEN_MV;
PMV1=MIN(VETORTEMPO);
QMV2=MAX(VETORTEMPO);
VETORGL=J(2,1,0);
VETORGL[1,1]=GLGEN_MV;
VETORGL[2,1]=GLAMB_MV;
PMINMV=MIN(VETORGL);
QMAXMV=MAX(VETORGL);
DO j=1 TO R_MV-1 BY 1;
PMV=PMV1-j;
QMV=QMV2-j;
USUB1j= -(0.64679880)
+(1.0068336*(PMV+QMV))
-(7.1495083*(10##(-9))*(PMV**2)*(QMV**2))
+(0.082395238*(((LOG(PMV))##2)+((LOG(QMV))##2)))
+(0.53767438*(LOG(PMV)*LOG(QMV))*(LOG(PMV)+LOG(QMV)))
-(0.091580971*(LOG(PMV)*LOG(QMV))*((LOG(PMV))##2+(LOG(QMV))##2))
+(0.021644307*(LOG(PMV)*LOG(QMV))*(LOG(PMV)##3+LOG(QMV)##3))
+(5.3529799*(10##(-3))*((LOG(PMV))##3)*((LOG(QMV))##3))
-(0.76227733*(EXP(-PMV)+EXP(-QMV)))
```

```
-(0.020829655*((PMV/QMV)+(QMV/PMV)))
+(1.7482806*10##(-3)*((PMV*QMV)-ABS(PMV-QMV)));
USUB2j= -(0.015802857*(PMV+QMV))
+(2.3780161*(10##(-9))*PMV*QMV*((PMV##2)+(QMV##2)))
+(1.7371131*(LOG(PMV)+LOG(QMV)))
-(0.33301620*((LOG(PMV)##2)+(LOG(QMV)##2)))
+(0.11442045*((LOG(PMV)##3)+(LOG(QMV)##3)))
-(0.035296928*LOG(PMV)*LOG(QMV)*(LOG(PMV)+LOG(QMV)))
+(0.033246016*((PMV/QMV)+(QMV/PMV)))
-(5.7298685*(10##(-9))*(((PMV/QMV)##4)+((QMV/PMV)##4)))
+(1.8747692*(EXP(-PMV)+EXP(-QMV)))
-(1.7476731*(10##(-13))*(PMV##2)*(QMV##2)*((PMV##2)+(QMV##2)))
+(6.9946263*(10##(-16))*(PMV##3)*(QMV##3)*(PMV+QMV))
+(2.3238523*(10##(-5))*((ABS(PMV-QMV))##2));
RSSj_1=SUM(AUTV_MV[j:R_MV]);
LAMBDAj=AUTV_MV[j,]/RSSj_1;
QSUB_j=(((PMINMV-j+1)*LAMBDAj)-1)/(PMINMV-j);
C1MV=(USUB1j-(QMAXMV-j+1))/((QMAXMV-j+1)*(PMINMV-j));
C2MV= (((PMINMV-j+1)*(QMAXMV-J+1)*(USUB2j##2))-(2*(USUB1j##2))) /
(((QMAXMV-j+1)##2)*(((PMINMV-j+1)*(QMAXMV-j+1))+2)*((PMINMV-j)##2));
DMV=(C1MV*(1-C1MV))-C2MV;
AMV=DMV*C1MV/C2MV;
BMV=DMV*(1-C1MV)/C2MV;
F_MILLIKEN=(BMV*QSUB_j)/(AMV*(1-QSUB_j));
GL1F_MILLIKEN=2*AMV;
GL2F_MILLIKEN=2*BMV;
VALORP_FMILLIKEN=1-PROBF(F_MILLIKEN,GL1F_MILLIKEN,GL2F_MILLIKEN);
/*PRINT j F_MILLIKEN "  " GL1F_MILLIKEN "  "
GL2F_MILLIKEN " " VALORP_FMILLIKEN [FORMAT=12.4];*/
RESULTADOSMV[j,1]=j;
RESULTADOSMV[j,2]=GL1F_MILLIKEN;
RESULTADOSMV[j,3]=GL2F_MILLIKEN;
RESULTADOSMV[j,4]=F_MILLIKEN;
RESULTADOSMV[j,5]=VALORP_FMILLIKEN;
END;
ALPHA=0.05;
```

```
NUMCOMPVEROSS=0;
MINRESMV=MIN(RESULTADOSMV[,5]);
IF MINRESMV>ALPHA THEN VALORPMV=0;
ELSE VALORPMV=1;
DO INDIMV=1 TO R_MV-1 BY 1 WHILE(VALORPMV>ALPHA);
POS=R_MV-INDIMV;
NUMCOMPVEROSS=RESULTADOSMV[POS,1];
VALORPMV=RESULTADOSMV[POS,5];
END;
/*PRINT "RESULTADOS TESTE DE MILLIKEN" ,"  ",
RESULTADOSMV [FORMAT=12.4] "    " NUMCOMPVEROSS;*/
*************************************************************************
*       TESTE DE CORNELIUS                      *
*************************************************************************;
CORNEIMPUTADOS=IMMEDIA;
**********MATRIZ GXE E SOMAS DE QUADRADOS******************;
MEDLINHAS_CORNE=CORNEIMPUTADOS[,:];
MEDCOLUN_CORNE=CORNEIMPUTADOS[:,];
NGEN_CORNE=NROW(CORNEIMPUTADOS);
NAMB_CORNE=NCOL(CORNEIMPUTADOS);
MEDIA_CORNE=CORNEIMPUTADOS[:,:];
L_CORNE=MEDLINHAS_CORNE*J(NCOL(MEDLINHAS_CORNE),NAMB_CORNE,1);
C_CORNE=J(NGEN_CORNE,NROW(MEDCOLUN_CORNE),1)*MEDCOLUN_CORNE;
M_CORNE=J(NGEN_CORNE,NAMB_CORNE,1)*MEDIA_CORNE;
GXE_CORNE=CORNEIMPUTADOS-L_CORNE-C_CORNE+M_CORNE;
SQGXE_CORNE=SSQ(GXE_CORNE);
SQGEN_CORNE=NAMB_CORNE*(SSQ(MEDLINHAS_CORNE-MEDIA_CORNE));
SQAMB_CORNE=NGEN_CORNE*(SSQ(MEDCOLUN_CORNE-MEDIA_CORNE));
GLGEN_CORNE=NGEN_CORNE-1;
GLAMB_CORNE=NAMB_CORNE-1;
GLGXE_CORNE=GLGEN_CORNE*GLAMB_CORNE;
QMGEN_CORNE=SQGEN_CORNE/GLGEN_CORNE;
QMAMB_CORNE=SQAMB_CORNE/GLAMB_CORNE;
QMGXE_CORNE=SQGXE_CORNE/GLGXE_CORNE;
FGEN_CORNE=QMGEN_CORNE/QMGXE_CORNE;
VALORP_FGEN_CORNE=1-PROBF(FGEN_CORNE,GLGEN_CORNE,GLGXE_CORNE);
```

```
FAMB_CORNE=QMAMB_CORNE/QMGXE_CORNE;
VALORP_FAMB_CORNE=1-PROBF(FAMB_CORNE,GLAMB_CORNE,GLGXE_CORNE);
/*PRINT GXE_CORNE, MEDLINHAS_CORNE MEDIA_CORNE, MEDCOLUN_CORNE,
QMGEN_CORNE QMAMB_CORNE QMGXE_CORNE,
FGEN_CORNE VALORP_FGEN_CORNE [FORMAT=12.4],
FAMB_CORNE VALORP_FAMB_CORNE [FORMAT=12.4];*/
******************DVS DA MATRIZ GXE_CORNE*******************************;
AUTOVALORES_CORNE=EIGVAL(GXE_CORNE*T(GXE_CORNE));
NNUL1_CORNE=AUTOVALORES_CORNE<-1E-6;
NNUL2_CORNE=AUTOVALORES_CORNE>1E-6;
NNUL_CORNE=NNUL1_CORNE+NNUL2_CORNE;
R_CORNE=(NNUL_CORNE^=0)[+,];
CALL SVD(EG_CORNE,VS_CORNE,EA_CORNE,GXE_CORNE);
U_CORNE=EG_CORNE[,1:R_CORNE];
S_CORNE=DIAG(VS_CORNE[1:R_CORNE,]);
VT_CORNE=EA_CORNE[,1:R_CORNE];
/*PRINT U_CORNE,AUTOVALORES_CORNE S_CORNE,VT_CORNE;*/
**************DESDOBRAMENTO DA SQ(GXE_CORNE)POR DVS*****************;
**************PROPORCOES********************;
VS2_CORNE=VS_CORNE##2;
AUTV_CORNE=VS2_CORNE[1:R_CORNE];
SOMAAUTOVAL_CORNE=SUM(VS2_CORNE);
PROPOR_CORNE=INV(SOMAAUTOVAL_CORNE)*VS2_CORNE;
PROPORCAO_CORNE=PROPOR_CORNE[1:R_CORNE];
RESD1_CORNE=SOMAAUTOVAL_CORNE-VS2_CORNE[1:1,];
PACUM_CORNE=CUSUM(PROPORCAO_CORNE)*100;
LINH_CORNE=T(1:R_CORNE);
RESUMO_CORNE=LINH_CORNE||AUTV_CORNE||PROPORCAO_CORNE||PACUM_CORNE;
/*PRINT RESUMO_CORNE [FORMAT=12.4];*/
***************ESCOLHA DOS COMPONENTES
MULTIPLICATIVOS DO MODELO AMMI***************;
ALPHA_CORNE=0.05;
AUX_QMEMCORNE=IMMEDIA;
CV_ORIG=0.120769685;
TOTALAUX_QMEMCORNE=SUM(AUX_QMEMCORNE*4);
MEDIAAUX_QMEMCORNE=TOTALAUX_QMEMCORNE/(15*27*4);
```

```
QMEMEDIO_CORNE1=(CV_ORIG*MEDIAAUX_QMEMCORNE)##2;
QMERRMEDIO_CORNE=QMEMEDIO_CORNE1/4;
/*
GLERRMEDIO_CORNE=90;
QMERRMEDIO_CORNE=9679.597;
*/
RESULTADOS_CORNE=J(R_CORNE,5,0);
/*CALCULO CORRIGIDO DOS GRAUS DE LIBERDADE DO ERRO MÉDIO PARA
TESTAR SIGNIFICANCIA SOB PRESENCA DE DADOS FALTANTES*/
INDFALTANTES_CORNE=NMISS(MISSING);
NUMFALTANTES_PORAMBIENTE=INDFALTANTES_CORNE[,:]*NUMCOL;
GLERROMEDIO_PORAMBIENTES=J(NUMFIL,1,42);
DO i=1 TO NUMFIL;
IF NUMFALTANTES_PORAMBIENTE[i,1]^=0
THEN GLERROMEDIO_PORAMBIENTES[i,1]=(14-NUMFALTANTES_PORAMBIENTE[i,1])*3;
END;
GLERRMEDIO_CORNE=SUM(GLERROMEDIO_PORAMBIENTES);
DO i=0 TO R_CORNE-1 BY 1;
GLCOMPONENTEm=(NGEN_CORNE-1-i)*(NAMB_CORNE-1-i);
CORNELIUS1=SUM(VS2_CORNE[i+1:R_CORNE]);
CORNELIUS2=GLCOMPONENTEm*QMERRMEDIO_CORNE;
QMCORNELIUS=CORNELIUS1/GLCOMPONENTEm;
CORNELIUS=CORNELIUS1/CORNELIUS2;
VALORP_CORNE=1-PROBF(CORNELIUS, GLCOMPONENTEm, GLERRMEDIO_CORNE);
RESULTADOS_CORNE[i+1,1]=i;
RESULTADOS_CORNE[i+1,2]=GLCOMPONENTEm;
RESULTADOS_CORNE[i+1,3]=QMCORNELIUS;
RESULTADOS_CORNE[i+1,4]=CORNELIUS;
RESULTADOS_CORNE[i+1,5]=VALORP_CORNE;

/*PRINT i GLCOMPONENTEm QMCORNELIUS
CORNELIUS [FORMAT=12.4] VALORP_CORNE [FORMAT=12.4];*/
END;
VALORESP_CORNE=RESULTADOS_CORNE[,5]<ALPHA_CORNE;
NUMCOMPCORNE=SUM(VALORESP_CORNE);
/*PRINT "RESULTADOS TESTE DE CORNELIUS","    ",
```

```
RESULTADOS_CORNE [FORMAT=12.4];*/
/*PRINT VALORESP_CORNE [FORMAT=12.4] NUMCOMPCORNE [FORMAT=12.4];*/
/*PRINT INDREP R2_PROCRUSTES;*/
PRINT INDQUALIDADE INDREP R2_PROCRUSTES RMSPD R_SPEARMAN
NUMCOMPVEROSS NUMCOMPCORNE QMERRMEDIO_CORNE GLERRMEDIO_CORNE PORCENTP;
SIMULACAO_IMMULTIPLA[INDREP,1]=INDQUALIDADE;
SIMULACAO_IMMULTIPLA[INDREP,2]=INDREP;
SIMULACAO_IMMULTIPLA[INDREP,3]=R2_PROCRUSTES;
SIMULACAO_IMMULTIPLA[INDREP,4]=RMSPD;
SIMULACAO_IMMULTIPLA[INDREP,5]=R_SPEARMAN;
SIMULACAO_IMMULTIPLA[INDREP,6]=NUMCOMPVEROSS;
SIMULACAO_IMMULTIPLA[INDREP,7]=NUMCOMPCORNE;
END;/*DO INDREP=1 TO REPETICOES;*/
PRINT SIMULACAO_IMMULTIPLA;
create IMP_MULTIPLA from
SIMULACAO_IMMULTIPLA
[colname={ID_IM INDREP_IM PROCRUSTES_IM RMSPD_IM
SPEARMAN_IM NUMCOMPVEROSS_IM NUMCOMPCORNE_IM}];
append from SIMULACAO_IMMULTIPLA;
QUIT;
PROC PRINT DATA=IMP_MULTIPLA; RUN;
**Para detalhes entrar em contato com: sergio.arciniegas@gmail.com**
```