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

Imputação AMMI *Bootstrap* Não-paramétrico em dados
multiambientais

Maria Joseane Cruz da Silva

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

Piracicaba
2017

Maria Joseane Cruz da Silva
Bacharel em Estatística

**Imputação AMMI *Bootstrap* Não-paramétrico em dados
multiambientais**
versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

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

**Piracicaba
2017**

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

Silva, Maria Joseane Cruz da

Imputação AMMI *Bootstrap* Não-paramétrico em dados multiambientais/
Maria Joseane Cruz da Silva. - - versão revisada de acordo com a resolução
CoPGr 6018 de 2011 - -Piracicaba, 2017.

69 p.

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

1. Imputação de dados 2. Interação genótipo-ambiente 3. Modelos AMMI 4.
Método Bootstrap I. Título

Dedicatória

Dedico...

*Aos amores da minha vida, **meus pais**, que fazem da
minha vida um caminho simples de se viver.*

AGRADECIMENTOS

Primeiramente a Deus por mais um objetivo alcançado.

Aos meus irmãos em especial a Rosineide, Rosicyde, José Wilson e José, pelo incentivo e apoio em mais um passo da minha vida.

Ao Prof. Dr. Carlos Tadeu dos Santos Dias pelo apoio, carinho, amizade e orientação durante o desenvolvimento deste trabalho.

À CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, pela bolsa de mestrado concedida.

A todos os professores em que convivi durante o curso de doutorado em Estatística e Experimentação Agronômica, me dando a oportunidade de mais uma vez participar desta família ESALQ, fornecendo conhecimentos e possibilidades de alto nível, com as quais tive a oportunidade de trabalhar em minhas pesquisas: Dercio, Roseli A. Leandro, Clarice G. B. Demétrio, Sílvio S. Zocchi, César G. de Lima, Sônia M. D. S. Piedade, Edwin M. M. Ortega, Cristian Villegas, Taciana e Renata Alcarde.

As secretárias (Solange Assis e Luciane Brajão) e funcionários do LCE/ESALQ/USP pelo apoio, carinho e dedicação no decorrer do curso.

Aos colegas do curso de Pós-Graduação em Estatística e Experimentação Agronômica da ESALQ/USP, os quais fizeram parte desta fase.

Um agradecimento especial eu reservo aos meus amigos Rick Anderson, Erasnilson, Marisol Miranda e Ana Júlia pelo enorme apoio, os quais não mediram esforço quando precisei.

Com carinho agradeço ao meu companheiro Lucas Nunes pelo apoio, compreensão e paciência, meu muito obrigada!

De forma geral, a todos os amigos que ajudaram a escrever mais uma página da minha história, aqueles que passaram e muito deixaram, e os que permanecem e me presenteiam com algo único todos os dias.

Muito obrigado!

*"A vontade de Deus nunca irá levá-lo
aonde a Graça de Deus não irá protegê-lo".*

Chico Xavier

SUMÁRIO

RESUMO	7
ABSTRACT	8
LISTA DE FIGURAS	9
LISTA DE TABELAS	10
1 INTRODUÇÃO	11
2 OBJETIVOS ESPECÍFICOS	13
3 REVISÃO BIBLIOGRÁFICA	15
3.1 Interação genótipos $\times$ ambiente	15
3.2 Métodos utilizados para estudar a interação G X E	17
3.3 Análise AMMI	18
3.4 Análise biplot e a metodologia AMMI	21
3.5 Método <i>Bootstrap</i>	24
3.6 Métodos de imputação de dados	25
3.6.1 Falta de dados	25
3.6.2 Métodos utilizados para estimar observações ausentes	26
3.6.3 Padrão e mecanismos de observações ausentes	28
3.6.4 Fases da imputação de dados	29
4 MATERIAL E MÉTODOS	32
4.1 Descrição dos dados	32
4.2 Métodos Estatísticos	33
4.2.1 Método de imputação via metodologia AMMI com reamostragem <i>bootstrap</i> Não-paramétrico (IAMMI-BNP)	33
4.2.2 Métodos de comparação	35
4.3 Descrição do processo de imputação	37
4.4 Simulação 1	37
4.5 Simulação 2	37
5 RESULTADOS E DISCUSSÃO	39
5.1 Resultado 1: estudo de simulação 1	39
5.2 Resultado 2: estudo de simulação 2	43
6 CONCLUSÃO	49
REFERÊNCIAS	51
APÊNDICES	55

RESUMO

Imputação AMMI *Bootstrap* Não-paramétrico em dados multiambientais

Em estudos multiambientais, o processo de recomendação de genótipos com maior produção e a determinação de genótipos estáveis são de suma importância para os melhoristas. Porém, quando ocorre falta de genótipo em um ou mais ambientes este processo passa a ter dificuldades. Pois, este procedimento depende de métodos estatísticos que necessitam de uma matriz de dados sem dados em falta. Desde 1976 diversos matemáticos e estatísticos estudam, continuamente, uma forma de lidar com dados em falta em dados multiambientais buscando obter um método que estime, de forma precisa, as unidades ausentes sem perda de informação. Desta forma, esta pesquisa propõe um novo método de imputação baseado na metodologia AMMI fazendo reamostragens *Bootstrap* Não-paramétrico na matriz de médias de interação genótipos e ambientes ($G \times E$), o modelo de imputação AMMI *Bootstrap* Não-paramétrico (IAMMI-BNP). Para estudo de simulação foi considerado o conjunto de dados referente a procedência S. of Ravenshoe - Mt Pandanus - QLD (14.420) de *Eucalyptus grandis* coletada na Austrália em 1983. Com a finalidade de obter estimativas precisas dos valores em falta, foi considerado dois estudos de simulação. O primeiro considerou 2000 reamostragens no sentido linha da matriz de interação $G \times E$ considerando duas porcentagens de perda de dados (10% e 20 %). O segundo estudo de simulação, considerou 200 reamostragens na matriz de falta (10%) e três diferentes modelos de IAMMI-BNP: $IAMMI_0$ -BNP, que considera apenas os efeitos principais do modelo AMMI; $IAMMI_1$ -BNP e $IAMMI_2$ -BNP que considera um e dois eixos multiplicados do modelo AMMI, respectivamente. De forma geral, de acordo com os métodos de comparação o método de imputação proposto nos dois estudos de simulação forneceu valores imputados próximos dos originais. Considerando os estudos de simulação com 10% de perda, a eficiência do método de imputação proposto foi melhor quando se utilizou o modelo $IAMMI_2$ -BNP (com dois eixos multiplicativos). O teste das ordens assinaladas de Wilcoxon mostrou que os valores imputados não influenciaram na estimativa da média, indicando que valores médios dos dados imputados de cada ambiente foram estatisticamente semelhantes aos valores médios originais.

Palavras-chave: Imputação de dados; Interação genótipo-ambiente; Modelos AMMI; *Bootstrap* Não-paramétrico

ABSTRACT

AMMI imputation Non-parametric bootstrap in multienvironmental data

In multienvironment studies, the process of recommendation of genotypes with higher production and the determination of stable environments are of utmost importance for plant breeders. However, when there is missing of genotype in one or more environments this process show difficulties. Therefore, this procedure depends on statistical methods that complete data matrix required. Since 1976 various mathematical and statistical study, continually, one way of dealing with the loss of information on data multienvironments, seeking to obtain a method that estimate, precisely, the missing units without loss of information. In this way, the purpose of this study is develop a new method of apportionment based on the methodology AMMI doing reamostragens bootstrap nonparametric in the array of means of genotype x environment interaction (GE). For the study of simulation was considered the data set concerning the origin of S. Mexico City - Mt Pandanus - QLD (14,420) of *Eucalyptus grandis* collected in Australia in 1983. It was performed two studies of simulation. The first performed 2000 resampling on the lines of the interaction matrix G X E, for two percentages of missing data (10% and 20%). The second simulation study considered 200 replicates in the missing data set (10 %) and three different models of IMAMMI-BNP: $AMAMMI_0$ -BNP, which considers only the main effects of the AMMI model; $IAMMI_1$ -BNP and $IAMMI_2$ -BNP which considers one and two axes multiplied by the AMMI model, respectively. In general, according to the comparison methods, the imputation method proposed in the two simulation studies provided imputed values similar to the originals. Considering the simulation studies with 10 % loss, the efficiency of the proposed imputation method was better when using the $IAMMI_2$ -BNP model (with two multiplicative axes). The Wilcoxon test of the orders showed that the values imputed had no influence on the mean estimate, indicating that mean values of the data imputed from each environment were statistically similar to the original mean values.

Keywords: Multiple imputation; Genotype-environment interaction; AMMI models; Non-parametric bootstrap

LISTA DE FIGURAS

Figura 1 - Ausência de interação entre Genótipos e Ambientes	16
Figura 2 - Interação simples entre Genótipos e Ambientes	16
Figura 3 - Interação complexa entre Genótipos e Ambientes	17
Figura 4 - Gráfico de caixas para os valores das 2000 estimativas obtidas pelo método IAMMI-BNP em cada posição de uma matriz com 10% de perda	40
Figura 5 - Gráfico de caixas para os valores das 2000 estimativas obtidas pelo método IAMMI-BNP em cada posição de uma matriz com 20% de perda	41
Figura 6 - Gráfico de dispersão da diferença entre o valor estimado e o valor original a 10% nas posições estimadas	42
Figura 7 - Gráfico de dispersão da diferença entre o valor estimado e o valor original a 20% nas posições estimadas	42
Figura 8 - Gráfico de caixas para os valores estimados $IAMMI_0$ -BNP dos dados com 10% de perda	44
Figura 9 - Gráfico de caixas para os valores estimados pela $IAMMI_1$ -BNP com 10% de perda	44
Figura 10 - Gráfico de caixas para os valores estimados pela $IAMMI_2$ -BNP Bootstrap Não-paramétrico dos dados com 10% de perda	44
Figura 11 - Gráfico de dispersão da diferença entre os valores estimados e o valor original a 10% de perda nas posições estimadas pela $IAMMI_0$ bootstratp Não-paramétrico	45
Figura 12 - Gráfico de dispersão da diferença entre os valores estimados e o valor original a 10% de perda nas posições estimadas pela $IAMMI_1$ bootstratp Não-paramétrico	46
Figura 13 - Gráfico de dispersão da diferença entre os valores estimados e o valor original a 10% de perda nas posições estimadas pela $IAMMI_2$ bootstratp Não-paramétrico	46

LISTA DE TABELAS

Tabela 1 - Dados para a análise da interação genótipos e ambientes	19
Tabela 2 - Esquema da análise de variância utilizando o modelo AMMI pelo método de Gollob	21
Tabela 3 - Marcadores de genótipos e ambientes a serem incluídos no gráfico biplot $AMMI_2$	23
Tabela 4 - Localização do teste de progênies de <i>Eucalyptus grandis</i>	32
Tabela 5 - Médias de alturas, em metros, das progênies (Prog.) de <i>E. grandis</i> aos 5 anos da Procedência, 12 km S. of Ravenshoe-Mt Pandanus-QLD	32
Tabela 6 - Matriz de dados para a análise da interação genótipos $\times$ ambientes . . .	33
Tabela 7 - Conjunto de dados com 10% de falta referente as médias de alturas, em metros, das progênies de <i>E. grandis</i> aos 5 anos (Procedência, 12 km S. of Ravenshoe-Mt Pandanus-QLD (14.420))	39
Tabela 8 - Conjunto de dados com 20% de falta referente as médias de alturas, em metros, das progênies (Prog.) de <i>E. grandis</i> aos 5 anos (Procedência, 12 km S. of Ravenshoe-Mt Pandanus-QLD (14.420))	39
Tabela 9 - Estimativas Bootstrap para cada posição (i, j) em falta da matriz com 10% de perda	40
Tabela 10 - Estimativas Bootstrap para cada posição (i, j) em falta da matriz com 20% de perda	41
Tabela 11 - Valores do teste das ordens assinaladas de Wilcoxon, do teste da raiz quadrada da diferença preditiva média (RMSPD) e da correlação de Spearman nas diferentes porcentagens de perda	42
Tabela 12 - Estimativas agrupadas ($\hat{Q}, \bar{U}, B, T$), valores do incremento relativo da variância (r), da taxa de unidades ausentes (λ) e a eficiência relativa (ER) em cada ambiente com valores imputados (matriz de dados continha 10% de perda) . .	43
Tabela 13 - Estimativas agrupadas ($\hat{Q}, \bar{U}, B, T$), valores do incremento relativo da variância (r), da taxa de unidades ausentes (λ) e a eficiência relativa (ER) em cada ambiente com valores imputados (matriz de dados continha 20% de perda) . .	43
Tabela 14 - Valores originais das alturas (coluna 2), estimativa da média de cada ambiente (coluna 3), valores estimados pelo método $IAMMI_0$ -BNP, $IAMMI_1$ -BNP e $IAMMI_2$ -BNP para cada casela em falta	45
Tabela 15 - Estimativas agrupadas ($\hat{Q}, \bar{U}, B, T$), valores do incremento relativo da variância (r), da taxa de unidades ausentes (λ) e a eficiência relativa (ER) em cada ambiente com valores imputados (dados 1 com 10% de perda)	46
Tabela 16 - Estimativas agrupadas ($\hat{Q}, \bar{U}, B, T$), valores do incremento relativo da variância (r), da taxa de unidades ausentes (λ) e a eficiência relativa (ER) em cada ambiente com valores imputados com o modelo $IAMMI_1$ -BNP (dados 1 com 10% de perda)	47
Tabela 17 - Estimativas agrupadas ($\hat{Q}, \bar{U}, B, T$), valores do incremento relativo da variância (r), da taxa de unidades ausentes (λ) e a eficiência relativa (ER) em cada ambiente com valores imputados com o modelo $IAMMI_2$ -BNP (dados 1 com 10% de perda)	47
Tabela 18 - Valores do teste das ordens assinaladas de Wilcoxon, do teste da raiz quadrada da diferença preditiva média (RMSPD) e da correlação de Spearman aos diferentes modelos de imputação	47

1 INTRODUÇÃO

Em programa de melhoramento genético é comum o interesse em obter respostas estáveis de um ou mais genótipos em diferentes ambientes, fornecendo ao melhorista estabilidade biológica dos mesmos. Desta forma, experimentos são realizados em vários ambientes com o intuito de estudar qualquer mudança no comportamento dos genótipos, a qual é caracterizada pela variação das suas respostas conforme o material estudado. Este comportamento é conhecido como interação genótipo $\times$ ambiente ($G \times E$) e tem sido alvo de estudo nos programas de melhoramento, pois a presença de diferenças entre os ambientes de seleção e de produção podem fornecer desempenho diferenciado, causando resposta inferior ao esperado e baixa eficiência na condução dos programas de melhoramento.

Os efeitos da interação dificilmente é nula e sua presença dificulta o processo de identificação de diferenças significativas entre os genótipos e de recomendação do genótipo mais promissor. Para um estudo acurado desses efeitos, métodos estatísticos são utilizados com o intuito de interpretar esta variação e obter informações precisas sobre a adaptabilidade e estabilidade dos genótipos. Primeiro, os melhoristas verificam a existência da interação $G \times E$, que pode ser detectada usando a análise conjunta de experimentos (análise de variância conjunta). Esta metodologia mede o efeito da interação observando a divisão do quadrado médio da interação $G \times E$ ($QM_{G \times E}$) pelo quadrado médio residual (QMR).

Uma vez que foi detectado o efeito da interação, outros métodos estatísticos são utilizados para interpretar a variação fornecida pelos genótipos em seus respectivos ambientes. Dentre esses métodos destaca-se a metodologia AMMI (*Additive Main effects and Multiplicative Interaction model*), a qual combina em um único modelo os componentes aditivos para os efeitos principais (genótipo, ambiente) e componentes multiplicativos para o efeito da interação (genótipo $\times$ ambiente), e estuda a adaptabilidade e estabilidade dos genótipos aos respectivos ambientes. Porém, o seu uso apresenta dificuldade na estimação dos parâmetros quando existe dados faltantes, impossibilitando o cálculo do efeito de interação presente nos modelos AMMI, representado pela decomposição por valores singulares (DVS), cujo método depende de uma matriz completa para ser utilizado (ARCINIEGAS-ALARCÓN; DIAS, 2009).

Visando solucionar este problema foram desenvolvidas diversas técnicas para lidar com a perda de dados. Os primeiros métodos desenvolvidos são conhecidos como métodos de deleção, os quais incluem a análise de casos completos e descarte de casos (LOBO et al., 2006). Em seguida foram desenvolvidos os métodos de imputação simples e múltiplas, que consistem na substituição dos valores em falta por valores estimados conforme o mecanismo de dados em falta: ausência totalmente aleatória (*Missing completely at Random - MCAR*), ausência de forma aleatória (*Missing at Random - MAR*) e ausência não aleatória (*Missing Not at Random - MNAR*) (ENDERS, 2010; RUBIN, 1976). Em estudo de melhoramento genético BERGAMO et al. (2008), Arciniegas-Alarcón e Dias (2009) e Arciniegas-Alarcón et al. (2013) descrevem o método de imputação livres da distribuição dos dados perdidos, o qual forneceu resultados precisos.

Constantemente, diversos pesquisadores estudam métodos estatísticos que forneçam estimativas mais próxima do valor real e com menor viés em dados de melhoramento genético, porém nem sempre estas estimativas são eficientes. Diante destas considerações este trabalho propõe um método de imputação baseado na metodologia AMMI com reamostragem *Bootstrap* Não-paramétrico na matriz de médias de interação genótipo por ambiente, cuja a finalidade é obter estimativas precisas dos valores em falta em dados provenientes de programa de melhoramento genético.

2 OBJETIVOS ESPECÍFICOS

Este trabalho de pesquisa tem por objetivo:

- i. Retirar porcentagens aleatórias (10% e 20%) dos valores em cada matriz de dados considerada nesta pesquisa (matriz de médias da interação genótipo e ambiente ($G \times E$)), gerando um padrão arbitrário de dados faltantes;
- ii. Estimar os valores em falta via Modelo AMMI fazendo reamostragem *Bootstrap* Não-paramétrico na matriz de médias de interação $G \times E$;
- iii. Estimar os valores em falta considerando três tipos de Modelo AMMI *Bootstrap* Não-paramétrico na matriz de médias de interação $G \times E$ ($IAMMI_0$ -BNP, $IAMMI_1$ -BNP e $IAMMI_2$ -BNP);
- iv. Verificar a eficiência do método proposto em cada caso estudado observando as médias, desvios padrão de cada conjunto de dados imputado em suas respectivas posições (i, j) com dados faltantes;
- v. Verificar a medida da qualidade das estimativas por meio da raiz quadrada da diferença preditiva média (RMSPD); do coeficiente de correlação de Spearman Não-paramétrico e do teste de Wilcoxon.

3 REVISÃO BIBLIOGRÁFICA

3.1 Interação genótipos $\times$ ambiente

Em programas de melhoramento genético, o processo de seleção de genótipos é embasado nas informações fenotípicas dos indivíduos, porém este procedimento se torna complexo devido a grande maioria dos caracteres de importância agrônoma serem controlados por um grande número de genes que são influenciados pelos efeitos ambientais na expressão do fenótipo (TOPPA & JADOSKI, 2011). Estes efeitos dão origem ao grupo de parâmetros genéticos definidos como interação genótipos $\times$ ambientes ($G \times E$). Neste sentido, pode-se dizer que o efeito da interação $G \times E$ é caracterizado pela variação sofrida pelos genótipos ao serem testados em diferentes ambientes.

Esta mudança dificulta a identificação dos genótipos adaptáveis e com alta produtividade, pois um grupo de genótipos podem se desenvolver com melhor qualidade em um ambiente e produzir resultados ruins em outro. Diante destes inconvenientes causados pela interação, recomenda-se a estratificação da região de adaptação da cultura em subregiões mais homogêneas. Porém, este processo não elimina o efeito da interação por absoluto, em razão da ocorrência de fatores incontrolláveis dos ambientes contra os quais a estratificação não oferece êxito. Se não existisse interação, um genótipo poderia se adaptar à maioria dos ambientes, apenas um experimento em um local seria necessário para fornecer resultados gerais e a fase de avaliação dos genótipos deixaria de ser a parte mais complexa.

Uma outra característica de interesse dos melhoristas é encontrar genótipos estáveis. Esta estabilidade pode ser classificada em estática e dinâmica. Na estabilidade estática os genótipos são estáveis em diferentes ambientes e fornecem pequena variação entre os ambientes. Porém, na estabilidade dinâmica os genótipos se caracterizam como estável em cada ambiente particular.

Em genética, genótipo é a constituição genética total de um organismo, ou seja, um gene que é transmitido dos genitores à sua progênie e os fenótipos correspondem à forma alternativa de expressão de uma característica, sendo dependente do genótipo, do ambiente e da expressão de um genótipo em um ambiente (RAMALHO et al., 1989 apud SQUILASSI, 2003). De forma geral, a manifestação de um fenótipo é o resultado da ação do genótipo sob influência do ambiente, assim para um dado genótipo podem ocorrer vários fenótipos, dependendo do efeito do ambiente. Ele é observável em níveis físicos, morfológicos, anatômico e bioquímicos.

Cada fenótipo é representado por $F = G + E + GE$, em que F é o efeito do fenótipo, G o efeito de genótipo, E o efeito de ambiente e GE representa o efeito da interação $G \times E$. Uma vez que não existe mudança na característica do genótipo de um ambiente para outro, a expressão é reescrita excluindo o termo referente a interação $G \times E$, ou seja, $F = G + E$.

Mori et al. (1986) relata que os parâmetros genéticos estimados não devem conter nenhum efeito dos ambientes, no entanto, isto é verídico quando não ocorre interação com ambiente. Entretanto, na maioria das vezes os efeitos da interação dificilmente é nula e testar os genótipos em uma série de ambientes diversificados, de tal maneira que se possa separar com segurança as variâncias genéticas das variâncias ocasionadas pela interação $G \times E$, passa a ser alvo de interesse dos pesquisadores.

De acordo com Faria (2009) existem duas condições que influenciam no efeito da interação $G \times E$. Primeiro são as variações ambientais que ocorrem de ambiente para ambiente dentro da área de distribuição da cultura como as características climáticas, solo e técnicas agrônomicas. Segundo são as frequências e distribuições de chuvas, temperatura

do ar e do solo, ocorrência de geadas etc.

Segundo Squilassi (2003) a resposta dos genótipos aos ambientes podem ser classificada em três tipos. A seguir, esta classificação é apresentada de forma gráfica considerando dois genótipos (Gen_1 e Gen_2) avaliados em dois ambientes (Amb_1 e Amb_2).

1. O primeiro tipo é conhecido como ausência de interação, em que não ocorre interação entre genótipos e ambientes, indicando que a mudança nas condições ambientais não influenciam no comportamento dos genótipos.

Figura 1 - Ausência de interação entre Genótipos e Ambientes

2. O segundo, é caracterizado como interação simples ou quantitativa, em que existe mudança na magnitude de performance dos genótipos, porém a ordem de seleção dos genótipos aos ambientes são as mesmas.

Figura 2 - Interação simples entre Genótipos e Ambientes

3. A interação cruzada ou qualitativa ocorre quando os genótipos respondem de forma diferenciada aos ambientes e a ordenação na classificação dos genótipos aos diferentes ambientes é alterada, indicando que o genótipo que respondeu positivamente em um ambiente pode não apresentar em outro.

Figura 3 - Interação complexa entre Genótipos e Ambientes

Neste contexto, os melhoristas buscam estudar os efeitos causados pela interação $G \times E$ sobre o material de pesquisa, explorar as informações contidas nos dados usando técnicas estatísticas com a finalidade de obter respostas precisas sobre a adaptabilidade e estabilidade dos genótipos aos diferentes ambientes, visando orientar os programas de melhoramento genético no que concerne ao desenvolvimento das cultivares superiores.

3.2 Métodos utilizados para estudar a interação $G \times E$

Uma vez que foi detectado o efeito da interação é de suma importância para o melhorista interpretar a variação fornecida pelos genótipos em seus respectivos ambientes. Desta forma, utiliza-se métodos estatísticos capazes de detectar esta variação. Dentre esses métodos encontra-se o método de Wricke que tem como base a estimação de um parâmetro de estabilidade obtido da decomposição da soma de quadrados da interação $G \times E$, a qual se divide em partes referentes a cada genótipo isoladamente, quantificando a contribuição de cada genótipo, separadamente, na interação $G \times E$ (WRICKE, 1962). Este processo apresenta falta de precisão na identificação da estabilidade dos genótipos e falta de informação sobre os ambientes.

SHUKLA (1972) propôs outro método que consiste em estimar a variância da estabilidade por meio da soma da variância dentro e entre ambientes, assim um genótipo é considerado estável quando a variância entre ambientes é igual a zero. Outro método é conhecido como regressão linear, o qual considera a variação da variável em estudo em relação a um índice ambiental, permitindo identificar genótipos estáveis recomendados de forma generalizada (SQUILASSI, 2003). No entanto, este procedimento apresenta falha na suposição de independência, o que não é permitido pelo método de regressão.

Com a tentativa de desenvolver um método que não dependa de suposições sobre o comportamento dos dados, foi proposto os métodos Não-paramétricos baseados no estudo da variação dos postos dentro de cada ambiente (SQUILASSI, 2003). Assim, os genótipos com menor variância dos postos é considerado mais estável. Os métodos supracitados, não permitem identificar a possibilidade do melhor genótipo em um particular ambiente não o ser em outro.

Existem diversos métodos estatísticos utilizados para esta finalidade, porém alguns fornecem respostas mais precisas que outros. Na pesquisa genética, a metodologia AMMI (*additive main effects and multiplicative interaction model*) tem ganhado destaque, devido a sua eficiência no processo de seleção de genótipos. A metodologia permite ao melhorista, identificar os genótipos que se mostraram mais adaptáveis e estáveis aos ambientes, o que pode ocorrer a um ou mais ambientes. Diversas generalizações tem sido

desenvolvidas e são continuamente estudadas com a finalidade de fornecer um método estatístico mais preciso e eficaz no estudo destes efeitos.

Lavoranti (2003), em sua pesquisa, desenvolveu um método baseado na amostragem *Bootstrap*, no modelo AMMI, que permitiu as análises gráficas e numéricas, da estabilidade e adaptabilidades fenotípicas de progênies de *Eucalyptus grandis* procedentes de três localidades australianas e implantadas em sete testes de procedências e progênies nas regiões Sul e Sudeste do Brasil. Seu trabalho fornecem resultados satisfatórios, interpretando melhor a estabilidade e adaptabilidade fenotípica dos genótipos.

Silva Filho et al. (2008) em seu estudo sobre “Comparação de métodos para avaliar a adaptabilidade e estabilidade produtiva em algodoeiro” verificou que a identificação de adaptabilidades específicas obtida pela análise AMMI, mostrou-se importante no estudo do comportamento dos genótipos.

Meotti et al. (2012) avaliaram o efeito causado pelas épocas de semeadura no desempenho agrônomo de cultivares de soja em São Domingos, SC, cuja finalidade foi indicar as cultivares mais estáveis e adaptadas a cada época. O autor utilizou a metodologia AMMI para estudar o desempenho produtivo das cultivares e a avaliar a adaptabilidade e a estabilidade das cultivares nas diferentes épocas de semeadura. Diante dos resultados, o autor observou que em ambos os anos agrícolas, as semeaduras em 15/10 e 15/11 maximizaram o número de vagens por planta, o número de grãos por vagem, a estatura das plantas, o número de ramos, a massa de mil sementes e a produtividade de grãos.

3.3 Análise AMMI

A metodologia AMMI (*additive main effects and multiplicative interaction model*) foi desenvolvida visando entender os efeitos da interação genótipos $\times$ ambientes, presentes, em experimentos multiambientais. Tal metodologia foi, à princípio, proposta por Mandel (1971) com a finalidade de selecionar modelos que expliquem o efeito causado pela variabilidade da interação por meio do estudo da variabilidade padrão, descartando os ruídos presentes nos dados.

A metodologia AMMI foi aplicada na agricultura a partir de 1984, ganhando destaque após 1994 (DUARTE; VENCOSKY, 1999). Dentre as aplicações na agricultura, a motivação realizada por Gauch (1988) contribuiu fortemente para o sucesso de tal metodologia. O autor analisou dados experimentais com a finalidade de garantir maior precisão e eficiência à análise da interação.

De forma geral, a metodologia AMMI é utilizada em dados que apresentam efeitos principais e interação (GAUCH JÚNIOR, 1992). Segundo Faria (2012) tal metodologia é aplicada em dados que fornecem uma variável resposta, não permite o uso de uma matriz de várias linhas ou colunas que contenham variáveis.

A análise AMMI combina em um único modelo duas metodologias importantes da estatística experimental para explicação da interação $G \times E$: os componentes aditivos referentes aos efeitos principais (genótipo, ambiente), os quais são estudados por meio da análise de variância usual (ANOVA) e os componentes multiplicativos que se referem ao efeito da interação (genótipo e ambiente), cuja parte é modelada pela Análise de Componentes Principais (ACP). O modelo AMMI pode conter ou não o efeito de bloco na parte aditiva, dependendo de como tenha sido planejado o experimento.

Considerando o conjunto de dados dispostos na tabela de dupla entrada de dimensão $g \times e$ (Tabela 1), em que as linhas representam o efeito dos genótipos, as colunas os efeitos de ambientes e Y_{ij} representa a resposta média de cada genótipo em seu respectivo ambiente, obtidas de r repetições.

Tabela 1 - Dados para a análise da interação genótipos e ambientes

Genótipos	Ambientes					Médias
	1	2	3	...	e	
1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1e}	$\bar{Y}_{1.}$
2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2e}	$\bar{Y}_{2.}$
3	Y_{31}	Y_{32}	Y_{33}	...	Y_{3e}	$\bar{Y}_{3.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
g	Y_{g1}	Y_{g2}	Y_{g3}	...	Y_{ge}	$\bar{Y}_{g.}$
Médias	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{.3}$	...	$\bar{Y}_{.e}$	$\bar{Y}_{..}$

Considerando a tabela de dupla entrada acima, assume-se que as respostas Y_{ij} seguem o modelo matemático:

$$Y_{ij} = \underbrace{\mu + g_i + e_j}_{\text{Parte Aditiva}} + \underbrace{\sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk}}_{\text{Parte Multiplicativa}} + \bar{\varepsilon}_{ij} \quad (1)$$

em que Y_{ij} é a resposta média do genótipo i ($i=1, 2, \dots, g$) no ambiente j ($j=1, 2, \dots, e$), μ é a média geral, g_i é o efeito do genótipo i , e_j é o efeito do ambiente j (parte aditiva do modelo), o termo $\sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk}$ é o efeito de interação (parte multiplicativa do efeito da interação $((ge)_{ij})$ a ser modelada), sendo λ_k é a raiz quadrada do k -ésimo autovalor das matrizes $(GE)(GE)^T$ e $(GE)^T(GE)$ (ambas com iguais autovalores não nulos), γ_{ik} é o i -ésimo elemento relacionado ao genótipo i do k -ésimo autovetor de $(GE)(GE)^T$ associado a λ_k^2 , α_{jk} é o j -ésimo elemento relacionado ao ambiente j do k -ésimo autovetor de $(GE)^T(GE)$ associado a λ_k^2 , $\bar{\varepsilon}_{ij}$ é o erro experimental médio associado ao i -ésimo genótipo no j -ésimo ambiente ($\bar{\varepsilon}_{ij} \sim N(0, \frac{\sigma^2}{r})$), $\frac{\sigma^2}{r}$ é a variância média constante do erro dentro do ambiente, r é o número de repetições ou blocos, i é a variação de genótipos, j é a variação de ambientes e p é o termo que relaciona a interação com a quantidade de matrizes obtidas da decomposição da matriz $G \times E$.

A matriz de efeitos da interação GE é obtida com base nas estimativas dos efeitos principais, a partir da solução de mínimos quadrados ordinários para o sistema de equações normais, do modelo não aditivo $Y_{ij} = \mu + g_i + e_j + \varepsilon_{ij}$, sob as restrições de identificabilidade $\sum_i g_i = \sum_j e_j = 0$ (DUARTE; VENCOSKY, 1999). Assim,

$$\hat{g}e_{ij} = \bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..} \quad (2)$$

em que, $\hat{\mu} = \bar{y}_{..}$, $\hat{g}_i = \bar{y}_{i.} - \bar{y}_{..}$, $\hat{e}_j = \bar{y}_{.j} - \bar{y}_{..}$, $\bar{y}_{..} = \frac{\sum_{i,j} y_{ij}}{ge}$, $\bar{y}_{i.} = \frac{\sum_j y_{ij}}{e}$ e $\bar{y}_{.j} = \frac{\sum_i y_{ij}}{g}$. Seja a matriz de médias da interação $G \times E$ com a estrutura de dados dada na tabela 1, em que Y_{ij} é a resposta média do genótipo i no ambiente j ,

utilizando a expressão 2 teremos a seguinte matriz de efeitos da interação GE.

$$\hat{GE} = \begin{bmatrix} \hat{g}e_{11} & \hat{g}e_{12} & \hat{g}e_{13} & \dots & \hat{g}e_{1e} \\ \hat{g}e_{21} & \hat{g}e_{22} & \hat{g}e_{23} & \dots & \hat{g}e_{2e} \\ \hat{g}e_{31} & \hat{g}e_{32} & \hat{g}e_{33} & \dots & \hat{g}e_{3e} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{g}e_{(g1)} & \hat{g}e_{(g2)} & \hat{g}e_{(g3)} & \dots & \hat{g}e_{(ge)} \end{bmatrix}$$

Nesta matriz $\hat{\mathbf{G}}\mathbf{E}$, é feito o ajuste dos efeitos multiplicativos, procurando obter um modelo (expressão 1) com um número de eixos suficientes para explicar o comportamento dos genótipos aos diferentes ambientes.

Diante do exposto anteriormente, o modelo pode conter vários termos multiplicativos ao se realizar a decomposição da matriz $\mathbf{G} \times \mathbf{E}$ pela decomposição em valores singulares (DVS). Visto que, um modelo com muitos termos pode resultar em dificuldades na interpretação dos resultados, assim é utilizado a metodologia AMMI, que tem por objetivo geral obter um modelo que resuma parte da interação $\mathbf{G} \times \mathbf{E}$ em apenas alguns eixos, ou seja, selecione um modelo AMMI com poucos termos multiplicativos (obtidos pela DVS) que despreze um resíduo adicional considerado como ruído (DUARTE; VENCOSKY, 1999). Desta forma, se observa no modelo 1 que a parte multiplicativa pode ser decomposta como,

$$\sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk} = \underbrace{\sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk}}_{\text{variabilidade padrao}} + \underbrace{\sum_{k=n+1}^p \lambda_k \gamma_{ik} \alpha_{jk}}_{\text{ruído}} \quad (3)$$

sendo $n < p$, $\rho_{ij} = \sum_{k=n+1}^p \lambda_k \gamma_{ik} \alpha_{jk}$ e p são as raízes características não nulas ($p = (1, 2, \dots, \min(g-1, e-1))$).

Desta forma, a matriz de interação $\mathbf{G} \times \mathbf{E}$ é modelada pela expressão 4 sob as restrições de identificabilidade $\sum_{i=1}^g g_i = \sum_{j=1}^e e_j = \sum_{i=1}^g (ge)_i = \sum_{j=1}^e (ge)_j = 0$. Sendo, $\mathbf{G} \times \mathbf{E}$ representado pelos seguintes termos: soma de p parcelas, cada uma resultante da multiplicação λ_k , a qual fornece uma informação sobre a k -ésima parcela da interação $\mathbf{G} \times \mathbf{E}$ e é expressa na mesma unidade de Y_{ij} por um efeito genotípico (γ_{ik}) e um efeito ambiental (α_{jk}), ambos adimensionais ($\sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk}$) e representam os pesos do genótipo i e do ambiente j na parcela da interação. Logo o modelo 1 é reescrito como,

$$Y_{ij} = \underbrace{\mu + g_i + e_j}_{\text{Parte Aditiva}} + \underbrace{\sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk}}_{\text{Parte Multiplicativa}} + \rho_{ijr} + \bar{\varepsilon}_{ij} \quad (4)$$

A estrutura da interação $\mathbf{G} \times \mathbf{E}$ é explicada, decompondo a soma de quadrado do efeito multiplicativo por meio da Análise de Componentes Principais (ACP), e assim, formando uma família de modelos $AMMI_0, AMMI_1, \dots, AMMI_k$, em que k representa o número de eixos ou componentes da interação $\mathbf{G} \times \mathbf{E}$, a serem considerados no modelo (GAUCH JÚNIOR; ZOBEL, 1996).

Para selecionar o número de eixos, um dos métodos utilizados na literatura é o método de Gollob (DUARTE; VENCOSKY, 1999), o qual atribui graus de liberdade às somas de quadrados ($SQ_k = \lambda_k^2$) e conta o número de parâmetros no k -ésimo termo multiplicativo. Em seguida, o teste F é calculado para cada k -ésimo eixo que pode ser considerado no modelo assim como na análise de variância usual. Desta forma, o esquema da análise de variância para análise AMMI, considerando um delineamento aleatorizado

em blocos, é descrita na Tabela 2.

Tabela 2 - Esquema da análise de variância utilizando o modelo AMMI pelo método de Gollob

Fonte de variação	GL	SQ	QM	F
Blocos.d.amb	$e(k-1)$	SQB.d.amb	-	-
Genótipos(G)	$g-1$	SQG	QMG	QMG/ QM(em)
Ambientes(E)	$e-1$	SQE	QME	QME/ QM(em)
Interação (G×E)	$(g-1)(e-1)$	SQ(G×E)	QMGE	$F_{g \times e}$
IPCA1	$g+e-1-(2 \times 1)$	λ_1^2	QM_{IPCA1}	$QM_{IPCA1}/QM(em)$
IPCA2	$g+e-1-(2 \times 2)$	λ_2^2	QM_{IPCA2}	$QM_{IPCA2}/QM(em)$
IPCA3	$g+e-1-(2 \times 3)$	λ_3^2	QM_{IPCA3}	$QM_{IPCA3}/QM(em)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
IPCAp	$g+e-1-(2 \times p)$	λ_p^2	QM_{IPCAp}	$QM_{IPCAp}/QM(em)$
Erro médio(em)	$ge(k-1)$	SQ(em)	QM(em)	
Total	$gek-14$	SQTotal		

em que g é o número de genótipos, e é o número de ambientes, λ é o valor singular da matriz de efeitos da interação GE, $IPCA_k$ é o eixo de interação da análise de componentes principais ($k=1, 2, \dots, p$)

As somas de quadrados para as fontes de variação são escritas da seguinte maneira,

$$SQB.d.amb = SQB.d.E_1 + SQB.d.E_2 + SQB.d.E_3 + \dots + SQB.d.E_k$$

$$SQG = a \sum_i \hat{g}_i^2, \quad SQE = g \sum_j \hat{e}_j^2$$

$$SQ(G \times E) = \sum_{k=1}^p \lambda_k^2 = \sum_{k=1}^n \lambda_k^2 + \sum_{k=n+1}^p \lambda_k^2$$

$$SQem = \sum_{i,j} \hat{\epsilon}_{ij}^2$$

sendo $\hat{g}_i = \bar{y}_{i.} - \bar{y}_{..}$, $\hat{e}_j = \bar{y}_{.j} - \bar{y}_{..}$, $\hat{\epsilon}_{ij} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$, $SQB.d.amb$ é a soma de quadrados de blocos dentro de ambientes, SQG é a soma de quadrados do efeito de genótipos, SQE é a soma de quadrados do efeito de ambientes, $SQ(G \times E)$ é a soma de quadrados do efeito da interação $G \times E$ e $SQem$ é a soma de quadrados no erro médio. O quadrado médio (QM) para cada fonte de variação é determinado pela razão entre sua respectiva SQ e graus de liberdade. A significância do k -ésimo eixo é observado por meio do teste F de Gollob à um determinado nível de significância. Os resultados da análise são apresentados graficamente em um biplot, o qual observa a magnitude dos escores de genótipos e ambientes para os eixos determinado pelo teste comentado anteriormente (DUARTE; VENCOSKY, 1999).

3.4 Análise biplot e a metodologia AMMI

A análise biplot foi desenvolvida por Gabriel (1971) com o objetivo de representar graficamente as informações de uma matriz de dados $n \times p$, utilizando resultados

obtido da análise de componentes principais. As informações contidas nas linhas se referem as amostras ou unidades experimentais e as informações contidas nas colunas se referem as variáveis.

Em experimentos multiambientais, as linhas da matriz são representadas pelos genótipos e as colunas pelos ambientes. Cada combinação de linha e coluna representa a média, das r repetições, das informações fornecido pelo i -ésimo genótipo no j -ésimo ambiente para alguma variável resposta.

Nesta matriz aplica-se a análise AMMI, a qual realiza a análise de variância (ANOVA) para a parte aditiva do modelo (genótipo, ambiente) e interação genótipo $\times$ ambiente dada pela equação 1, sendo que o ajuste da parte multiplicativa é realizado pela metodologia dos componentes principais (DVS).

Segundo Duarte e Vencovsky (1999), na análise AMMI o gráfico biplot é construído realizando a decomposição da matriz de efeitos da interação GE por meio da DVS, a qual é dada pela por:

$$\hat{\mathbf{G}}\mathbf{E} = \mathbf{U}\mathbf{S}\mathbf{V}' = \mathbf{U}\mathbf{S}^{1/2}\mathbf{S}^{1/2}\mathbf{V}' = \mathbf{G}\mathbf{H}'$$

Tendo selecionado os n eixos à serem incluídos no gráfico biplot, os quais são importantes para explicar o comportamento dos genótipos aos diferentes ambientes, teremos uma decomposição aproximada da matriz de efeitos da interação GE dada por,

$$\hat{\mathbf{G}}\mathbf{E} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}' = \sum_{k=1}^n \lambda_k \gamma_k \alpha'_k$$

em que as colunas de $\tilde{\mathbf{U}}$ representam os n primeiros vetores $\gamma_{p \times 1}$, as linhas de $\tilde{\mathbf{V}}'$ representam os n primeiros vetores de $\alpha'_{1 \times e}$ e $\tilde{\mathbf{S}}$ é uma matriz diagonal com os primeiros valores singulares ($\lambda_1, \lambda_2, \dots, \lambda_n$ com $n < p$). A expressão 5 pode ser reescrita como segue.

$$\hat{\mathbf{G}}\mathbf{E} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}' = \tilde{\mathbf{U}}\tilde{\mathbf{S}}^{1/2}\tilde{\mathbf{S}}^{1/2}\tilde{\mathbf{V}}' = \tilde{\mathbf{G}}\tilde{\mathbf{H}}' = \sum_{k=1}^n (\lambda_k^{1/2}\gamma_k)(\lambda_k^{1/2}\alpha'_k) \quad (5)$$

Desta forma, $\tilde{\mathbf{G}}$ terá os g vetores conhecidos como marcadores das linhas, cada um com n componentes representando as coordenadas de genótipos para cada eixo selecionado e $\tilde{\mathbf{H}}'$ terá os e marcadores de colunas, cada um com n componentes representando as coordenadas de ambientes para os mesmos n eixos selecionados anteriormente.

Deste modo, tendo dados com a mesma estrutura da tabela 1, se obtém a matriz de efeitos da interação $GE_{(g \times e)} = (ge_{ij})$ por meio da expressão 2.

$$\hat{\mathbf{G}}\mathbf{E} = \begin{bmatrix} \hat{ge}_{11} & \hat{ge}_{12} & \hat{ge}_{13} & \dots & \hat{ge}_{1e} \\ \hat{ge}_{21} & \hat{ge}_{22} & \hat{ge}_{23} & \dots & \hat{ge}_{2e} \\ \hat{ge}_{31} & \hat{ge}_{32} & \hat{ge}_{33} & \dots & \hat{ge}_{3e} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{ge}_{g1} & \hat{ge}_{g2} & \hat{ge}_{g3} & \dots & \hat{ge}_{ge} \end{bmatrix}$$

A matriz $\hat{\mathbf{G}}\mathbf{E}$ é composta em valores singulares (expressão 5) obtendo as informações das coordenadas de genótipos e ambientes para os n eixos escolhidos. De forma geral, teremos a seguinte decomposição exata.

$$\hat{\mathbf{G}}\mathbf{E} = \mathbf{U}\mathbf{S}\mathbf{V}' = \mathbf{U}\mathbf{S}^{1/2}\mathbf{S}^{1/2}\mathbf{V}' = \mathbf{G}\mathbf{H}'$$

em que,

$$\mathbf{G} = \begin{bmatrix} g_{11} & g_{12} & \dots & g_{1n} \\ g_{21} & g_{22} & \dots & g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{g1} & g_{g2} & \dots & g_{gn} \end{bmatrix} \cdot \begin{bmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_n} \end{bmatrix} =$$

$$\begin{bmatrix} g_{11}\sqrt{\lambda_1} & g_{12}\sqrt{\lambda_2} & \dots & g_{1n}\sqrt{\lambda_n} \\ g_{21}\sqrt{\lambda_1} & g_{22}\sqrt{\lambda_2} & \dots & g_{2n}\sqrt{\lambda_n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{g1}\sqrt{\lambda_1} & g_{g2}\sqrt{\lambda_2} & \dots & g_{gn}\sqrt{\lambda_n} \end{bmatrix}$$

e

$$\mathbf{H}' = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_n} \end{bmatrix} \cdot \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{e1} & h_{e2} & \dots & h_{en} \end{bmatrix} =$$

$$\begin{bmatrix} \sqrt{\lambda_1}h_{11} & \sqrt{\lambda_1}h_{12} & \dots & \sqrt{\lambda_1}h_{1n} \\ \sqrt{\lambda_2}h_{21} & \sqrt{\lambda_2}h_{22} & \dots & \sqrt{\lambda_2}h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\lambda_e}h_{e1} & \sqrt{\lambda_e}h_{e2} & \dots & \sqrt{\lambda_e}h_{en} \end{bmatrix}$$

Observa-se que, as linhas de $\mathbf{G}$ são os marcadores de genótipos e as colunas de $\mathbf{H}'$ são os marcadores de ambientes. Tendo as matrizes $\mathbf{G}$ e $\mathbf{H}'$ pode-se obter o gráfico biplot. Para plotar o gráfico biplot é importante selecionarmos o número de componentes a serem incluídos no modelo. Segundo Crossa (1990) os dois primeiros eixos, geralmente, representam a maior parte do padrão de resposta técnica relacionada com a interação ($G \times E$). Por isto, frequentemente, é usado os dois primeiros eixos no gráfico biplot caracterizando o modelo *AMMI2*. Assim, pode-se plotar os g genótipos e os e ambientes num sistema cartesiano de n eixos ($n \leq 3$). Considerando $n = 2$, para construir o gráfico biplot, basta que se tome apenas as duas primeiras coordenadas para os marcadores de genótipos e de ambientes das matrizes $\mathbf{G}$ e $\mathbf{H}'$, obtendo as matrizes aproximadas $\tilde{\mathbf{G}}$ e $\tilde{\mathbf{H}}'$. E, conseqüentemente, as colunas de $\tilde{\mathbf{G}}$ e linhas de $\tilde{\mathbf{H}}'$ vão compor as coordenadas (abscissa X e ordenadas Y) no gráfico biplot (Tabela 3).

Tabela 3 - Marcadores de genótipos e ambientes a serem incluídos no gráfico biplot *AMMI2*

Pontos	Genótipos e ambientes	biplot <i>AMMI2</i>	
		X	Y
		<i>IPCA1</i>	<i>IPCA2</i>
1	G_1	$g_{11}\sqrt{\lambda_1}$	$g_{12}\sqrt{\lambda_2}$
2	G_2	$g_{21}\sqrt{\lambda_1}$	$g_{22}\sqrt{\lambda_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
g	G_g	$g_{g1}\sqrt{\lambda_1}$	$g_{g2}\sqrt{\lambda_2}$
1	E_1	$\sqrt{\lambda_1}h_{11}$	$\sqrt{\lambda_2}h_{21}$
2	E_2	$\sqrt{\lambda_1}h_{12}$	$\sqrt{\lambda_2}h_{22}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
e	E_e	$\sqrt{\lambda_1}h_{1e}$	$\sqrt{\lambda_2}h_{2e}$

Com as informações dos eixos das abscissas X e das ordenadas Y teremos o gráfico biplot num sistema cartesiano de 2 eixos. Em geral, o gráfico biplot fornece a visualização do melhor desempenho dos genótipos a determinados ambientes como também mostra a adaptabilidade e a estabilidade das cultivares aos diferentes ambientes.

Observando-se os escores de genótipos e ambientes para os eixos da interação pode-se obter informações importantes sobre o comportamento dos genótipos nos diferentes ambientes. Assim, escores próximos de zeros indicam os genótipos e ambientes que contribuíram pouco para a interação, sendo assim classificados como estáveis. Se formos considerar uma análise AMMI com duas componentes ($IPCA1$ e $IPCA2$) suficientes para explicar a variação dos genótipos aos ambientes, ou seja, $AMMI_2$, então o biplot pode ser interpretado observando os escores próximos de zeros (que situam-se próximos da origem do gráfico) para os dois eixos da interação, caracterizando os genótipos e ambientes como estáveis. Outra forma de interpretação é observar os genótipos que têm uma performance desejável, ou seja, com alta produtividade, sendo avaliados por meio das médias dos efeitos principais. Mais detalhes podem ser encontradas em Duarte e Vencovsky (1999).

3.5 Método *Bootstrap*

Nas pesquisas científicas a fase de análise de dados é de suma importância, pois a partir dela são obtidos resultados que permitem obter conclusões sobre uma determinada característica em estudo (produção de uma variedade, altura de planta, número de danos causados por determinada praga etc). Para esta finalidade se utiliza métodos estatísticos que fornecem resultados numéricos à um determinado nível de confiança fornecendo, ao pesquisador, conclusões com base na característica em análise.

Existem diversos métodos estatísticos que podem ser utilizados conforme o objetivo da pesquisa, desde os mais simples como uma estatística descritiva até os mais complexos como, por exemplo, os métodos bayesianos. Dentre os métodos existentes na literatura estatística se encontram os métodos de reamostragens. Ambos tem por finalidade estimar com precisão características populacionais e fazer inferência sobre a população. Desta forma, são realizadas reamostragens sobre a amostra representativa da população de interesse da pesquisa, a qual é considerada como uma população finita. Dentre as diversas técnicas de reamostragem se destaca o método de reamostragem *Bootstrap* proposto por Efron (1979), o qual é composto por métodos paramétricos e Não-paramétricos.

Faria (2012) menciona que, se a amostra original representa a população da qual foi retirada, assim, por meio da reamostragem *Bootstrap* considerando a amostra como uma "população", e realizando reamostragens com reposição, é possível estimar parâmetros da população (média, variância, percentis etc.), gerando estimativas consistentes do parâmetro de interesse.

A ideia base deste método é estimar um parâmetro θ com base na reamostragem da amostra de interesse, a fim de construir uma distribuição empírica para θ . Assim, seja uma amostra aleatória original $Y = (y_1, y_2, \dots, y_n)^T$, a b -ésima amostra de tamanho n ($b = 1, 2, \dots, B$) obtida da amostra original Y por meio de uma amostragem simples com reposição, e θ^b a b -ésima estimativa de θ obtida aplicando um método de estimação padrão para $Y^{(b)}$, então, a estimativa média ($\hat{\theta}_{boot}$) e a estimativa da variância ($\hat{V}_{boot}$) das B estimativas *Bootstrap* ($\hat{\theta}_{boot}^b$) são dadas por:

$$\hat{\theta}_{boot} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_{boot}^b \quad (6)$$

$$\hat{V}_{boot} = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_{boot}^b - \hat{\theta}_{boot})^2 \quad (7)$$

Segundo He (2006), a estimativa da variância para θ é uma estimativa consistente quando n e B tendem ao infinito e consequentemente tem distribuição aproximadamente normal, assim o intervalo de confiança para θ é expresso por:

$$CI_{boot}(\theta) = \hat{\theta} \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_{boot}} \quad (8)$$

em que $z_{1-\frac{\alpha}{2}}$ representa o percentil da distribuição normal. Se a distribuição *Bootstrap* não for normal, existe um intervalo de confiança alternativo calculado empiricamente usando os percentis da distribuição empírica de θ . Este intervalo é expresso por:

$$CI_{emp}(\theta) = (\hat{\theta}^{(b,l)}, \hat{\theta}^{(b,u)}) \quad (9)$$

em que $\hat{\theta}^{(b,l)}$ e $\hat{\theta}^{(b,u)}$ são os percentis $\alpha/2$ e $1-\alpha/2$ da distribuição *Bootstrap* empírica. Efron (1979) relata que, o método fornece estimativas confiáveis para distribuições *Bootstrap* normal se as reamostragens forem feitas com $B = 200$ vezes. Entretanto, para distribuições *Bootstrap* não normal são necessárias amostras grandes a partir de $B = 2000$.

3.6 Métodos de imputação de dados

3.6.1 Falta de dados

Nas diversas pesquisas científicas (sejam elas qualitativas ou quantitativas, realizadas em campo, em domicílio, em laboratório e etc.) a falta de informação para um ou mais indivíduos é um problema comum. Este fato surge, na maioria dos casos, devido aos fatores que não foram controlados no decorrer da pesquisa, ou simplesmente fatores que não podem ser controlados pelo pesquisador.

Uma das causas da ocorrência de dados em falta é o erro de digitação cometido na fase da coleta de dados ao se fazer o registro das informações. Por isto, esta fase deve ser realizada com muito cuidado para evitar falhas deste tipo e, consequentemente, influenciar na realização das análises e fornecer resultados falseados.

Outro fator que gera falta de dados é o caso do indivíduo omitir as informações. Este caso ocorre geralmente nas pesquisas domiciliares, em que o indivíduo deixa de fornecer informações particulares como salário, idade etc. Outro fator similar é quando o indivíduo deixa a pesquisa, por algum motivo particular, seja devido a uma reação ao medicamento ou por não resistir à pesquisa, este é um fator bem característico das pesquisas médicas, em que os pacientes são submetidos a aplicação de algum novo medicamento. Em problemas que envolvem séries temporais, a ausência de valores de precipitação de séries históricas das estações meteorológicas geram perda de informações no conjunto de dados (SILVA, 2012).

Em programa de melhoramento genético a falta pode ocorrer de várias formas. Arciniegas-Alarcón e Dias (2009) comentam algumas situações que podem gerar falta de dado em experimentos multiambientais:

- i. A falta de informação pode surgir em programa de teste de genótipos, pois as culti-

vares são escolhidas durante cada ano, levando em consideração os dados que foram observados, sem considerar dados não observados;

- ii. A falta de dados emerge quando o mesmo subconjunto de genótipos está ausente em um número de ambientes da mesma sub-região, porque o melhorista no local não aceita esses genótipos, por apresentarem algumas características agronômicas indesejáveis.

A falta de atenção, durante a realização da pesquisa ou experimento, pode causar diversos problemas que podem influenciar na resposta do indivíduo ao material em estudo. Dentre eles tem-se: a existência de dados em falta complicando o uso dos métodos multivariados; o surgimento de observações discrepantes; os resultados falseados quando não se tem conhecimento da área experimental.

Diante destes problemas percebe-se que, para realizar uma pesquisa é importante ter bastante atenção, principalmente na fase de planejamento dos experimentos, pois um erro cometido pode comprometer o restante do processo da pesquisa e consequentemente, afeta os resultados. De forma geral, quando o processo de realização dos experimentos não é executado com cautela, ou melhor, da forma apropriada, os resultados simplesmente fornecem uma montanha de informações que podem ser inúteis, muitas vezes difíceis de se realizar se analisar por meio de métodos estatísticos.

3.6.2 Métodos utilizados para estimar observações ausentes

O problema de observações ausentes é comum nas diversas áreas do conhecimento científico e tem sido alvo de preocupação dos melhoristas de plantas, pois a sua existência complica o processo de seleção de genótipos. Para realizar o processo de seleção de genótipos é de suma importância, o pesquisador tomar conhecimento de algumas informações sobre a adaptabilidade e estabilidade dos genótipos aos ambientes em estudo.

Tais informações são geralmente obtidas por meio de métodos estatísticos como a metodologia multivariada. Entretanto, para que se possa utilizar estes métodos é necessário que a matriz de dados esteja completa, o que na verdade nem sempre ocorre em ensaios multiambientais. Então, existem duas formas de lidar com este problema: desconsiderar as informações ausentes da análise de dados (métodos de remoção) ou utilizar métodos que possam estimar estas informações ausentes com uma precisão considerável (Métodos de imputação).

Os métodos de remoção de dados são classificados em análise de casos completos e análise de descarte de casos (variáveis observadas) (LOBO et al., 2006; SILVA, 2012). A priori estes métodos se mostram eficientes quando o número de dados em falta é relativamente pequeno. Entretanto, com o aumento de dados em falta perde-se informações sobre a variável em estudo, superestimando os parâmetros, fornecendo resultados tendenciosos e apresenta menor poder devido ao menor tamanho amostral quando os dados em falta são removidos.

Rubin (1976) e Rubin (1978) sugeriu métodos embasados na substituição dos valores em falta por estimativas obtidas de algum método estatístico. Estes métodos são conhecidos na literatura como métodos de imputação de dados faltantes ou métodos de substituição de dados faltantes, os quais são classificados em simples e múltiplos. Os métodos de imputação simples consistem em estimar, apenas uma vez, cada valor em falta por algum método estatístico utilizando as informações presentes no conjunto de dados, e assim se obtém uma matriz de dados completa.

Dentre os métodos simples existentes, os que mais se destacaram foram o método de imputação pela média dos dados disponível em cada variável em estudo;

a imputação pelo método do vizinho mais próximo; a imputação *hot deck*; imputação pela regressão linear e imputação pela máxima verossimilhança. Mesmo diante do seu grande uso, Rubin (1976) relata que os métodos simples apresentam aumento no viés das estimativas, fornecendo valores estimados distantes dos verdadeiros.

Os métodos de imputação múltipla surgiram na literatura como tentativa de controlar o viés fornecido pelas estimativas ao se utilizar os métodos estatísticos com o restante dos dados. Estes métodos refletem a incerteza sobre as estimativas dos dados em falta ao substituir cada valor em falta por um conjunto de valores razoáveis sobre o valor a ser imputado (SILVA, 2012). Desta forma, a variabilidade é introduzida nos conjuntos de dados imputados, fornecendo resultados das análises mais generalizáveis que os métodos de imputação simples (JELICIC et al., 2009).

Ayuyev et al. (2009) propuseram, em sua pesquisa, um método de imputação baseado em agrupamento dinâmico em vários tipos de dados. Estes dados podem conter variáveis categóricas, contínuas ou simplesmente uma mistura delas, geralmente estão presente em pesquisa de opinião pública. A falta ocorre quando o indivíduo deixa de responder a um ou mais itens e responde os demais, gerando uma falta característico de um mecanismo de falta aleatória (MAR). O método, definido como imputação baseado em clusterização dinâmica (*Dynamic Clustering Imputation - DCI*), consiste em considerar a dissimilaridade das medidas entre os indivíduos no conjunto de dados, em seguida estas distâncias são utilizadas no algoritmo de agrupamento para identificar os casos similares e executar um específico grupo dinâmico de imputação de valores em falta.

Em programa de melhoramento genético, como solução dos dados em falta, foram desenvolvidos diversos métodos com o intuito de fornecer estimativas próximas do valor real, e assim obter uma matriz de dados completa que possa fornecer informações precisas sobre o comportamento dos genótipos nos diferentes ambientes.

Arciniegas-Alarcón e Dias (2009) citam alguns métodos importante como o uso do algoritmo EM (*Expectation Maximization*) juntamente com a metodologia AMMI (*additive main effects and multiplicative interaction model*) sugeridos por Gauch e Zobel (1988). Este método considera no processo de estimação dos dados em falta tanto os efeitos aditivos como os efeitos da interação genótipo por ambiente, proporcionando estimativas dos valores em falta com o máximo de informação contida nos dados disponíveis. Outra alternativa é baseada nos modelos multivariados, utilizando sub-modelos robustos ou métodos de mínimos quadrados alternados nos modelos AMMI sugeridos (DENIS; BARIL, 1992; CALISKI et al., 1992).

Um método que forneceu estimativas precisas dos valores ausentes foi desenvolvido por Bergamo (2007) e Bergamo et al. (2008) conhecido como imputação de dados livre de distribuição (IMDVS). Este método é uma extensão múltipla do método sugerido por Kranowski (1988), e tem como vantagem o fato de não exigir distribuição dos dados, por exemplos, a distribuição normal.

Oliveira (2012) realizou um estudo comparando diversos métodos de imputação em experimentos implantados em vários ambientes sobre a cultura de café conilon. Estes métodos consistiram da imputação simplesmente pela média, outros modelos com base em modelos de regressão linear simples, considerando como variáveis explicativas índices ambientais, média dos genótipos que foram observados em todos os ambientes, respostas dos genótipos correlacionados com o genótipo que se deseja estimar a falta e por fim, um modelo de imputação que conserva a soma de quadrados da interação $G \times E$ inalterada. Dentre os métodos comparados, o autor observou que os modelos com base em modelos de regressão forneceram médias próximas dos valores médios reais, exceto o modelo considerando como variável explicativa as respostas médias dos genótipos de maior

correlação com a que se pretende estimar o valor ausente.

Recentemente Arciniegas-Alarcón et al. (2013) propuseram um método de imputação com base em modelos determinísticos. Este método utiliza a técnica de validação cruzada por vetor baseando-se em um esquema iterativo com decomposição em valores singulares (DVS) da matriz. Diante do desenvolvimento dos métodos de imputação, observa-se que diversos métodos de imputação são continuamente propostos e várias generalizações são freqüentemente geradas com o intuito de obter um método de imputação que forneça estimativas com menos viés.

3.6.3 Padrão e mecanismos de observações ausentes

Os métodos de imputação desenvolvidos por Rubin (1976) e Rubin (1978) estimam os valores ausentes de acordo com o mecanismo que gerou a falta e do padrão de comportamento das observações em falta. Os mecanismos de dados faltantes consistem em descrever as relações entre os dados em falta e a probabilidade de ausência, fornecendo informações sobre o que gerou a falta de dados.

Estes mecanismos são divididos em três situações de falta. A ausência totalmente aleatória (*Missing completely at Random - MCAR*) que ocorre quando a falta de dados não tem relação com alguma outra variável medida e com a própria variável em falta. Um exemplo deste mecanismo pode ser observado quando ocorre plantas danificadas devido a fatores que o pesquisador não tem controle. Assim, a perda dos dados não está relacionada que não foi medida e nem com variáveis que possam ser medidas no decorrer do experimento, ocorreu simplesmente por conta de fatores inesperados.

O segundo mecanismo é conhecido como ausência aleatória (*Missing at Random - MAR*), o qual ocorre de forma aleatória indicando que a perda da observação ocorreu devido à alguma informação contida no conjunto de dados. Em experimentos multiambiais, este mecanismo é característico da falta ocorrida nos programas que realizam testes de genótipos, em que as cultivares são escolhidas durante cada ano conforme os dados observados sem considerar os dados não observados.

O terceiro mecanismo é a ausência não aleatória (*Missing Not at Random - MNAR*), em que a falta depende das informações da variável que contém a falta e/ou de outras informações não mensuradas. Este tipo de mecanismo é considerado em dados multiambiais, quando um subconjunto de genótipos não são considerados em um número de ambientes da mesma sub-região, porque simplesmente o melhorista decidiu não usar. Assim, a falta de genótipos em um ambiente ocorrerá nos demais.

Os padrões de dados consistem em fornecer informações sobre o comportamento dos dados ausentes, descrevendo a localização dos valores em falta. A literatura classifica os padrões em algumas situações. A primeira é conhecida como padrão univariado (*univariate pattern*) que ocorre quando a falta surge apenas na variável em estudo, sendo o caso de falta ocorrida em experimentos que envolvem delineamentos experimentais.

Temos também o padrão de não resposta (*unit nonresponse pattern*, o qual surge quando o indivíduo deixa de responder em uma pesquisa por algum motivo particular, sendo o caso de pesquisas realizadas por meio de questionários como o censo, ou seja, os dados em falta podem ocorrer quando um indivíduo não responde a um item do questionário porque está entediado, cansado ou não entendem determinado item (JELICIC et al., 2009)

O padrão monótono (*monotone pattern*) que ocorre quando o indivíduo deixa a pesquisa ou simplesmente não pode continuar, pode-se observar este tipo de padrão em pesquisas médicas quando o indivíduo deixa a pesquisa porque ocorreu alguma reação ao medicamento testado. Tal padrão é característico de experimentos que observam os

indivíduos ao longo do tempo. E por fim, o padrão geral (*general pattern*) que pode ser chamado de arbitrário, o qual apresenta uma dispersão dos valores faltantes na matriz de dados. Uma descrição completa pode ser vista em Enders (2010) e Silva (2012).

3.6.4 Fases da imputação de dados

De forma geral, para se realizar a imputação múltipla é importante seguir algumas fases como:

1. **Fase de imputação:** neste primeiro momento os dados em falta são estimados, em m tempos, por algum método de imputação formando m conjuntos de dados completos (m pode variar de 3 à 10 imputações, segundo Rubin (1978) com $m = 5$ se tem uma boa estimativa do valor em falta);
2. **Fase de análise:** nesta fase, para cada conjunto de dados completos se realiza uma análise de dados conforme a estatística de interesse como a média, variância, desvio padrão, parâmetros de regressão etc;
3. **Fase de agrupamento:** uma vez realizada a fase de análise, se faz a análise de agrupamento para produzir inferência dos resultados a serem imputados. Para esta finalidade, se agrupa as estimativas dos parâmetros obtidos na fase de análise dos m conjuntos de dados completos por meio da média aritmética das estimativas, dada por,

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^m \hat{Q}_i$$

em que $\hat{Q}_i$ é uma estimativa do i -ésimo parâmetro considerado correspondente aos m -ésimo conjunto de dados imputado, o qual é considerado como uma variável aleatória. Nesta fase, também é importante combinar os erros padrões das estimativas. Porém, o erro padrão é composto de duas variações, devido a duas fontes envolvidas na variação das estimativas. Uma delas é a variação dentro das imputações expresso por,

$$\bar{U} = \frac{1}{m} \sum_{i=1}^m \hat{U}_i$$

sendo $\hat{U}_i$ a variância do m -ésimo conjunto de dados imputados. A outra variação quantifica a variabilidade de uma estimativa em todas as demais. Esta variação é caracterizada como a variação entre as imputações, ou seja,

$$B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2.$$

Obtida as duas fontes de variação das estimativas pode-se encontrar a variância combinada e em seguida o desvio padrão combinado, ou seja,

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B$$

$$DP = \sqrt{T}$$

sendo $(1 + \frac{1}{m})$ a correção para números infinitos de imputações.

4 MATERIAL E MÉTODOS

4.1 Descrição dos dados

Para o desenvolvimento desta pesquisa, cujo principal objetivo é fornecer um método que estime as observações em falta próximas dos valores reais, foi realizado um estudo de simulação em um conjunto de dados sobre a altura de procedências de *Eucalyptus grandis*. Estes dados, foram obtidos em programas de melhoramento genético em diferentes ambientes e são os mesmos usados por Lavoranti (2003). A estrutura de dados refere-se a procedência 12 km S.of Ravenshoe-Mt Pandanus-QLD-(14.420), a qual foi coletada na Austrália em 1983 por meio de um convênio estabelecido com a CSIRO (Commonwealth Scientific and Industrial Research Organization) e sob o patrocínio do Banco Mundial BIRD com a Embrapa Florestas. Desta forma, foi possível realizar, por meio de parcerias com empresas florestais privadas, sete testes combinados de procedência e progênie nas regiões sul e sudeste do Brasil (Tabela 4), formando assim 7 ambientes.

Tabela 4 - Localização do teste de progênie de *Eucalyptus grandis*

Ambientes	Município/Estado	Regiões	Latitude	Longitude	Atitude
A ₁	Barra Ribeiro-RS	Sul	30°20'	51°14'	30
A ₂	Telêmico Borba-PR	Sul	24°150'	20°29'	850
A ₃	Boa Esperança do Sul-SP	Sudeste	21°57'	48°32'	540
A ₄	Guanhães-MG	Sudeste	18°40'	42°60'	900
A ₅	Ipatinga-MG	Sudeste	19°15'	42°20'	250
A ₆	Aracruz-ES	Sudeste	19°48'	40°17'	50
A ₇	Caçapava-SP	Sudeste	23°03'	45°46'	650

Para a realização destes testes foi considerado o delineamento em blocos casualizados, com 6 plantas por parcelas e 10 repetições, no espaçamento de 3,0 m por 2,0 m. Destes experimentos, foi disponibilizada apenas a média de alturas para cada genótipo de *Eucalyptus grandis* nos diferentes ambientes (Tabela 5). Estes dados são expressos em uma tabela de dupla entrada em que as linhas se referem aos genótipos e as colunas aos ambientes.

Tabela 5 - Médias de alturas, em metros, das progênie (Prog.) de *E. grandis* aos 5 anos da Proce-
dência, 12 km S. of Ravenshoe-Mt Pandanus-QLD

	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇
G ₁	17,40	25,00	18,67	20,61	13,88	19,84	14,72
G ₂	17,58	24,00	17,69	20,44	13,10	17,77	14,44
G ₃	16,52	23,70	15,94	18,91	12,86	17,93	13,43
G ₄	16,78	22,68	16,28	16,55	11,92	17,70	13,03
G ₅	15,36	21,56	15,95	16,57	10,97	18,06	13,03
G ₆	15,76	22,34	16,61	19,08	11,91	19,22	13,17
G ₇	14,57	20,35	17,02	15,06	12,47	17,01	11,14
G ₈	18,46	24,52	17,87	18,16	13,66	19,72	14,37
G ₉	16,87	20,77	16,84	18,96	13,54	18,69	14,06
G ₁₀	16,92	22,48	17,16	18,94	13,71	19,38	14,98
G ₁₁	16,50	22,98	16,87	17,17	13,55	19,12	13,13
G ₁₂	17,71	23,61	16,88	18,19	12,62	18,17	13,52
G ₁₃	17,62	22,65	16,01	18,78	11,92	17,88	13,24
G ₁₄	16,29	23,59	18,79	20,20	13,29	19,31	13,60
G ₁₅	15,94	23,36	18,10	17,91	12,75	19,71	12,49
G ₁₆	16,46	22,59	17,80	19,08	12,80	19,44	13,78
G ₁₇	17,74	23,38	16,17	18,91	14,42	20,24	13,32
G ₁₈	16,90	21,71	16,95	18,56	12,50	18,82	14,33
G ₁₉	16,28	20,12	14,66	15,68	10,60	16,10	11,98
G ₂₀	15,78	22,15	15,36	16,46	12,51	16,89	12,07

Destes dados, foi realizado um estudo de simulação retirando aleatoriamente 10% e 20% por cento dos dados, os quais foram considerados como valores ausentes e, assim, gerando um padrão arbitrário de dados em falta para cada matriz. Em seguida, o método de imputação proposto é utilizado para obter as estimativas dos valores ausentes, o qual é descrito no item 4.2.1. Depois, os resultados obtidos foram comparados utilizando os métodos de comparação descritos no item 4.2.2, cuja finalidade foi obter conclusões sobre a eficiência do método na estimação dos valores em falta.

4.2 Métodos Estatísticos

4.2.1 Método de imputação via metodologia AMMI com reamostragem *bootstrap* Não-paramétrico (IAMMI-BNP)

O método *Bootstrap* Não-paramétrico fornece algumas vantagens que chama atenção dos pesquisadores para sua aplicação em dados faltantes. Este método, na sua versão não-paramétrica, não requer conhecimento de qualquer distribuição de probabilidade da variável de interesse, não exigindo, também, o conhecimento do mecanismo dos dados faltantes.

Desta forma, para utilizar o método de reamostragem *Bootstrap* em dados multiambientais em programa de melhoramento genético, é necessário que os dados estejam dispostos em uma tabela de dupla entrada, em que as linhas representam os genótipos testados, as colunas são representadas pelos ambientes considerados no experimento e o resultado dos efeitos genéticos e ambientais são representados pela variável Y_{ij} associada ao i -ésimo genótipo e o j -ésimo ambiente (LAVORANTI, 2003). Considere Y_{ij} o valor médio das r repetições ao se estudar o efeito do genótipo g alocado ao ambiente e , o qual pode ser expresso conforme a Tabela 6.

Tabela 6 - Matriz de dados para a análise da interação genótipos $\times$ ambientes

Genótipos	Ambientes					Médias
	1	2	3	...	e	
1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1e}	$\bar{Y}_{1.}$
2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2e}	$\bar{Y}_{2.}$
3	Y_{31}	Y_{32}	Y_{33}	...	Y_{3e}	$\bar{Y}_{3.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
g	Y_{g1}	Y_{g2}	Y_{g3}	...	Y_{ge}	$\bar{Y}_{g.}$
Médias	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{.3}$	...	$\bar{Y}_{.e}$	$\bar{Y}_{..}$

Em experimentos multiambientais é comum que, pesquisadores queiram identificar o genótipo ou grupo de genótipos que são superiores em um ou mais ambientes, conforme o interesse econômico. Entretanto, em experimentos que envolvem o estudo de diferentes genótipos em vários ambientes podem ocorrer valores em falta, devido a alguns fatores não controláveis no decorrer do experimento ou escassês de material para alguns genótipos.

Existem vários métodos que estimam as informações perdidas por meio de alguma técnica estatística, porém a maioria dos métodos ainda apresenta viés nas estimativas dos valores ausentes. Com a finalidade de obter um método que estime os valores ausentes com menos viés, esta pesquisa propõe um novo método de imputação que estima

os valores ausentes via metodologia AMMI fazendo reamostragens *Bootstrap* na matriz de médias da interação $G \times E$.

Considere um conjunto de dados, assim como apresentado na Tabela 6, como uma amostra aleatória e considere cada linha da matriz, uma unidade amostral multivariada, mas que pode conter unidades ausentes. A princípio, os valores ausentes são substituídos por uma medida de posição (média), formando um conjunto de dados completo $\mathbf{Y}_{cm}$. Neste novo conjunto de dados, utiliza-se a metodologia *Bootstrap* Não-paramétrico no sentido linha da matriz de médias da interação $G \times E$, a fim de obter estimativas *Bootstrap* ($\hat{y}_{ij}^1, \hat{y}_{ij}^2, \dots, \hat{y}_{ij}^B$) para cada posição (i, j) ausente. Sendo $b = 1, 2, \dots, B$, o processo é realizado seguindo três passos.

- **Passo 1:** A partir da matriz de dados completada e padronizada $\mathbf{Y}_{cm}$, gerar uma amostra *Bootstrap* $\mathbf{Y}^b$ considerando a unidade amostral multivariada e sorteando no sentido “linha” da matriz;
- **Passo 2:** Neste passo, ajusta-se a interação $G \times E$ fazendo uma decomposição em valores singulares (DVS). Em seguida, calcula-se os efeitos da média geral ($\hat{\mu}$), os efeitos de genótipos ($\hat{g}_i$) e os efeitos de ambientes ($\hat{e}_j$). Após cada cálculo, o algoritmo utiliza o modelo de predição 10 da metodologia AMMI para estimar os valores ausentes, gerando o b -ésimo valor estimado referente a posição (i, j) da matriz de interação $G \times E$ em falta.

$$\hat{y}_{ij}^b = \hat{\mu} + \hat{g}_i + \hat{e}_j + \sum_{k=1}^n \hat{\lambda}_k \hat{\gamma}_{ik} \hat{\alpha}_{jk} \quad (10)$$

em que $\hat{\mu} = \bar{Y}_{..}$, $\hat{g}_i = \bar{Y}_{i.} - \bar{Y}_{..}$, $\hat{e}_j = \bar{Y}_{.j} - \bar{Y}_{..}$, n é o número de componentes principais retidos no modelo AMMI ($n \leq p$), λ_k é o k -ésimo valor singular da matriz de interação GE obtida no passo b (GE^b), α_{jk} é o elemento correspondente ao j -ésimo ambiente no k -ésimo vetor singular da matriz de interação GE^b e γ_{ik} é o vetor singular do i -ésimo genótipo no k -ésimo componente principal da interação GE^b .

- **Passo 3:** Uma vez repetido o passo 1 e 2, B vezes, se obtém B estimativas para cada posição (i, j) em falta, calcula-se $\hat{y}_{boot(ij)}$ para a posição (i, j) em falta, o qual será substituído na matriz de médias $G \times E$ fornecendo uma matriz final de dados imputada Y_{imp} .

$$\hat{y}_{boot(ij)} = \frac{1}{B} \sum_{b=1}^B \hat{y}_{ij}^b \quad (11)$$

$$\hat{V}_{boot(ij)} = \frac{1}{B-1} \sum_{b=1}^B (\hat{y}_{ij}^b - \hat{y}_{boot(ij)})^2 \quad (12)$$

Após gerados as B estimativas de cada valor ausente, pode-se calcular intervalos de confiança *Bootstrap* para o valor imputado. He (2006) comenta que este método tem grande vantagem por realizar o processo B vezes (uma vez para cada amostra *Bootstrap*), incorporando assim, a incerteza dos valores imputados.

Observa-se que, como o valor em falta é estimado via família AMMI ($AMMI_0, AMMI_1, \dots, AMMI_k$), se em uma das B etapas o modelo AMMI selecionado for o completo, com todos os termos multiplicativos a resposta estimada equivale a média da casela relativa ao genótipo i e o ambiente j .

4.2.2 Métodos de comparação

Para verificar a eficiência do método proposto na estimação dos valores em falta, foi obtida a medida da qualidade das estimativas dos valores em falta por meio da raiz quadrada da diferença preditiva média (RMSPD), cuja estatística foi desenvolvida com base na soma de quadrado das diferenças entre os valores originais em cada posição (i, j) escolhida a ser ausente e as suas respectivas imputações, dividindo pelo número de dados ausentes. A estatística RMSPD pode ser expressa da seguinte maneira,

$$RMSPD = \sqrt{\frac{\sum_{i,j} (y_{ij} - y_{boot(ij)})^2}{N}} \quad (13)$$

em que y_{ij} é o verdadeiro valor da casela na matriz de dados antes da retirada dos dados do i -ésimo genótipo no j -ésimo ambiente, $y_{boot(ij)}$ é a estimativa *Bootstrap* no i -ésimo genótipo no j -ésimo ambiente e N é o número de valores ausentes no conjunto de dados simulado. Segundo Arciniegas-Alarcón e Dias (2009), um menor valor da estatística RMSPD indica eficiência no método de imputação. Outro método utilizado para o estudo da qualidade das estimativas, é o coeficiente de correlação Não-paramétrico de Spearman entre valor imputado e o valor original, cujo método fornece valores entre -1 e 1. Uma vez que seu valor esteja próximo de -1 ou 1, isto indica que os valores originais e os valores estimados apresentam relação forte entre si negativa ou positiva, indicando eficiência do método.

Ainda neste passo, também foi utilizado o teste das ordens assinadas de Wilcoxon (*Wilcoxon Signed Ranks Test*). Este teste foi proposto por Wilcoxon em 1945 (CAMPOS, 1979) e como o próprio nome diz, se baseia nos postos das diferenças entre cada par de valores. É utilizado quando se pretende verificar a significância entre os resultados de duas amostras pareadas, ou seja, amostras que são avaliadas na mesma unidade amostral conforme a variável em estudo. Geralmente, é utilizado em pesquisas que pretendem avaliar alguma variável antes e depois de alguma situação.

Nesta pesquisa, o teste será utilizado com a finalidade de verificar se existe diferença significativa entre os valores originais e os valores imputados pelo método proposto, ou seja, se os valores nas duas situações (antes do valor ser retirado do conjunto de dados e depois de ser retirado) são estatisticamente semelhantes a um determinado nível de significância. Assim, será considerado como pares cada valor original e imputado ($Y_{o(ij)}, Y_{boot(ij)}$) referente a cada casela selecionada aleatoriamente de acordo com a porcentagem de falta (10%). Portanto, as diferenças entre o valor original e o valor imputado de cada casela serão dadas por,

$$D_i = Y_{boot(ij)} - Y_{o(ij)}$$

em seguida estas diferenças serão transformadas em valor absoluto e as ordens desses

valores absolutos serão classificados em ordem crescente (O_i) e a estatística de teste será,

$$T = \sum_{i=1}^n O_i A_i$$

sendo A_i uma variável indicadora em que o valor de D_i será 1 se $D_i > 0$ e 0 se $D_i < 0$. A estatística de teste representa a soma das ordens das diferenças positivas $D_i > 0$. E por fim, estimativas agrupadas propostas por Rubin (1976) adaptadas ao método de imputação utilizado, foram obtidas. Nesta fase do processo de pesquisa, será considerada as B estimativas *Bootstrap* e, em seguida, a média aritmética ($\bar{Q}_j$) das estimativas médias de todos os conjuntos de dados completados nas b 's reamostragens *bootstraps* ($b = 1, 2, \dots, B$) é obtida, ou seja,

$$\bar{Q}_j = \frac{1}{B} \sum_{b=1}^B Q\hat{boot}_b \quad (14)$$

em que $Q\hat{boot}_b$ é a b -ésima média do b -ésimo conjunto de dados imputado referente ao ambiente j , o qual é considerado como uma variável aleatória. A variação dentro das imputações é dada por,

$$\bar{U}_j = \frac{1}{B} \sum_{b=1}^B U\hat{boot}_b \quad (15)$$

sendo $U\hat{boot}_i$ a variância da b -ésima variância obtida do b -ésimo conjunto de dados imputado referente ao ambiente j . A variação entre as imputações tem a seguinte forma,

$$V_j = \frac{1}{B-1} \sum_{b=1}^B (Q\hat{boot}_i - \bar{Q})_j. \quad (16)$$

Obtida as duas fontes de variação das estimativas pode-se encontrar a variância combinada dado por,

$$T_j = \bar{U}_j + (1 + \frac{1}{B})V_j \quad (17)$$

sendo $(1 + \frac{1}{B})$ a correção de números infinitos de imputações.

Após obter todas as informações acima foi calculado a eficiência relativa (ER) da imputação de uma estimativa pontual baseada em B imputações dada por,

$$ER_j = (1 + (\lambda/b))^{-1} \quad (18)$$

em que,

$$\lambda = \left(\frac{1}{r_j + 1} \left(r_j + \frac{2}{v_{bj} + 3} \right) \right)^{-1} \quad r_j = \frac{1}{\bar{U}_j (1 + \frac{1}{b}) \hat{V}_j}$$

$$v_{bj} = (b-1) \left[1 + \frac{\bar{U}_j}{(1 + b^{-1}) \hat{V}_j} \right]^2$$

sendo λ_j a fração de informação ausente, a qual mede a precisão da estimativa se nenhum

valor ausente ocorrer, r_j é o aumento relativo na variância devido aos dados ausentes, $\bar{U}_j$ é a variância dentro das imputações, $\hat{V}_j$ é a estimativa da variância das B_j estimativas *Bootstrap*, v_{bj} são os graus de liberdade da distribuição t-Student referente ao j -ésimo ambiente.

4.3 Descrição do processo de imputação

O conjunto de dados referente as médias de alturas, em metros, das progênies de *E. grandis*, 12 km S. of Ravenshoe-Mt Pandanus-QLD, foi submetido a retiradas aleatórias de 10% e 20% da matriz de dados. Os valores selecionados, em cada porcentagem, foram retirados da matriz de dados formando duas novas matrizes com caselas vazias, ou seja, matrizes de dados em falta. Para cada uma delas, foram obtidas as médias de cada coluna (ambientes), as quais foram consideradas como estimativas iniciais para os dados em falta, e assim, duas matrizes completas foram geradas. Nesta pesquisa, foi considerado dois estudos de simulação (simulação 1 e simulação 2), os quais estão descritos a seguir. Todas as análises e gráficos foram implementadas por meio do *software* R 3.3.1 (R Core Team, 2016).

4.4 Simulação 1

Neste estudo de simulação, foram consideradas as matrizes com 10% e 20% de perda. Para cada matriz completada foi utilizado o método proposto para estimar os dados faltantes (Imputação AMMI *Bootstrap* Não-paramétrico). Este método é consiste na reamostragem da matriz em falta das médias das r repetições do i -ésimo genótipo ao j -ésimo ambiente, a qual foi completada pela estimativa da média de cada ambiente. Uma vez realizada a reamostragem na matriz, o próximo passo consiste em analisar o número de componentes a serem incluídos no modelo AMMI (expressão 10) para estimação dos dados em falta.

Desta forma, para selecionar o número de eixos a serem incluídos no modelo de imputação, em cada reamostragem, foi feito a decomposição em valores singulares (DVS) da matriz reamostrada (matriz *Bootstrap*). De acordo com os resultados da DVS, para determinar o número de eixo a ser incluso no modelo de imputação, foram considerados os eixos que explicam 95% da variabilidade total dos dados. Estes eixos são incluídos no modelo AMMI (equação 10) e as estimativas são obtidas, depois substituídas na matriz em suas respectivas caselas em falta. Esta nova matriz é guardada em um arquivo com formato de lista e se retorna ao primeiro passo de estimação, que é da reamostragem da matriz completada. Realizando este procedimento $B = 2000$ vezes, 2000 novas matrizes completadas com estimativas *Bootstrap* são obtidas.

Com as 2000 matrizes calcula-se a estimativa final para cada casela em falta utilizando a equação 11, as quais são substituídas na matriz em falta fornecendo uma matriz imputada. Tendo a matriz imputada, o próximo passo é fazer um estudo de comparação utilizando o cálculo da *RMSPD*, da Eficiência Relativa (*ER*) e do teste das ordens assinaladas de Wilcoxon com a finalidade de verificar a eficiência do método de imputação.

4.5 Simulação 2

Com a finalidade de estudar, de forma mais precisa, o comportamento dos valores imputados nas diversas reamostragens, verificando se o número de eixos inseridos no modelo de imputação e o número reamostragens influenciam na estimativa final dos

dados faltantes, foi considerado apenas o conjunto de dados com 10% de perda e três modelo de imputação AMMI *Bootstrap* Não-paramétrico: com apenas os eixos principais (IAMMI0-BNP), com apenas um eixo (IAMMI1-BNP) e com dois eixos (IAMMI2-BNP). Segundo Duarte e Vencovsky (1999), geralmente dois ou três eixos são necessários para explicar a variabilidade da estrutura dos dados.

Inicialmente, os valores em falta no conjunto de dados, foi substituído pela média de cada ambiente, gerando uma matriz completada. Em seguida, padronizou-se a matriz completada o sentido coluna, para evitar qualquer influência sobre a estimação das unidades ausentes. Na matriz padronizada foi realizado o processo de imputação. Após o processo ter sido realizado 200 vezes, para cada modelo de imputação, 200 novas matrizes estimadas são guardadas e suas estimativas *bootstraps* são obtidas.

Com as 200 matrizes, calcula-se a estimativa final para cada casela em falta utilizando a equação 11, as quais são substituídas na matriz em falta fornecendo uma matriz imputada para cada modelo de imputação. Tendo a matriz imputada, o próximo passo é fazer um estudo de comparação utilizando o cálculo da *RMSPD*, a Eficiência Relativa (*ER*) e do teste das ordens assinaladas de Wilcoxon com a finalidade de verificar a eficiência do método.

5 RESULTADOS E DISCUSSÃO

5.1 Resultado 1: estudo de simulação 1

Os conjuntos de dados com 10% e 20% falta de observações, encontram-se nas Tabelas 7 e 8. Observa-se que, para um conjunto de dados com 140 caselas à 10% de perda foram retiradas 14 caselas e à 20% de perda foram retiradas 28 caselas.

Tabela 7 - Conjunto de dados com 10% de falta referente as médias de alturas, em metros, das progênies de *E. grandis* aos 5 anos (Procedência, 12 km S. of Ravenshoe-Mt Pandanus-QLD (14.420))

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
G_1	17,40	25,00	18,67	20,61	13,88	19,84	y_{17}
G_2	17,58	24,00	17,69	20,44	13,10	y_{26}	y_{27}
G_3	16,52	23,70	15,94	18,91	12,86	17,93	13,43
G_4	16,78	22,68	y_{43}	16,55	11,92	17,70	13,03
G_5	y_{51}	21,56	15,95	16,57	10,97	18,06	13,03
G_6	15,76	22,34	16,61	19,08	11,91	19,22	13,17
G_7	y_{71}	20,35	17,02	15,06	y_{75}	y_{76}	11,14
G_8	18,46	24,52	17,87	18,16	13,66	19,72	y_{87}
G_9	16,87	20,77	16,84	18,96	13,54	18,69	14,06
G_{10}	16,92	22,48	17,16	18,94	13,71	19,38	14,98
G_{11}	$y_{11.1}$	$y_{11.2}$	16,87	17,17	13,55	19,12	13,13
G_{12}	17,71	23,61	16,88	18,19	12,62	18,17	13,52
G_{13}	17,62	22,65	16,01	18,78	11,92	17,88	13,24
G_{14}	16,29	23,59	18,79	20,20	13,29	19,31	$y_{14.7}$
G_{15}	15,94	23,36	18,10	17,91	$y_{15.5}$	19,71	12,49
G_{16}	16,46	22,59	17,80	19,08	12,80	19,44	13,78
G_{17}	17,74	23,38	16,17	18,91	14,42	20,24	13,32
G_{18}	16,90	21,71	16,95	18,56	12,50	$y_{18.6}$	14,33
G_{19}	16,28	20,12	14,66	15,68	10,60	16,10	11,98
G_{20}	15,78	22,15	15,36	16,46	12,51	16,89	12,07

Tabela 8 - Conjunto de dados com 20% de falta referente as médias de alturas, em metros, das progênies (Prog.) de *E. grandis* aos 5 anos (Procedência, 12 km S. of Ravenshoe-Mt Pandanus-QLD (14.420))

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
G_1	17,40	25,00	18,67	y_{14}	13,88	19,84	y_{17}
G_2	17,58	24,00	17,69	20,44	13,10	y_{26}	y_{27}
G_3	16,52	23,70	15,94	18,91	12,86	17,93	13,43
G_4	16,78	22,68	y_{43}	16,55	11,92	17,70	13,03
G_5	y_{51}	21,56	15,95	16,57	10,97	18,06	13,03
G_6	15,76	22,34	16,61	19,08	y_{65}	y_{66}	13,17
G_7	y_{71}	20,35	17,02	y_{74}	y_{75}	y_{76}	11,14
G_8	18,46	24,52	17,87	18,16	13,66	19,72	y_{87}
G_9	16,87	20,77	y_{93}	18,96	y_{95}	18,69	14,06
G_{10}	16,92	22,48	17,16	$y_{10.4}$	$y_{10.5}$	19,38	14,98
G_{11}	$y_{11.1}$	$y_{11.2}$	16,87	17,17	13,55	19,12	13,13
G_{12}	17,71	23,61	16,88	18,19	12,62	18,17	13,52
G_{13}	17,62	22,65	16,01	$y_{13.4}$	$y_{13.5}$	17,88	13,24
G_{14}	16,29	23,59	18,79	20,20	13,29	19,31	$y_{14.7}$
G_{15}	15,94	23,36	18,10	$y_{15.4}$	$y_{15.5}$	19,71	12,49
G_{16}	16,46	22,59	17,80	19,08	12,80	19,44	13,78
G_{17}	17,74	23,38	16,17	18,91	14,42	20,24	13,32
G_{18}	16,90	21,71	16,95	$y_{18.4}$	12,50	$y_{18.6}$	$y_{18.7}$
G_{19}	16,28	20,12	14,66	15,68	10,60	16,10	11,98
G_{20}	15,78	$y_{20.2}$	15,36	16,46	12,51	16,89	12,07

Nesta análise, foi considerado 2000 reamostragens com reposição no sentido linha da matriz de dados completada, fornecendo

As estimativas finais ($\hat{y}_{bootij}$), obtidas pelo método IAMMI-BNP utilizando a equação 11, encontram-se nas Tabelas 9 e 10 juntamente com os valores da variabilidade das 2000 estimativas *Bootstrap* correspondente as respectivas caselas em falta.

Diante destes resultados, observa-se que a 10% de perda dos dados, de acordo com a variabilidade das 2000 estimativas obtida pelo método proposto, apenas a posição (11,2) forneceu com maior variabilidade comparada com as demais posições (Tabela 9). Porém, com 20% de perda 5 caselas ((7,4), (10,4), (13,4), (15,4) e (18,4))apresentaram maior variabilidade (17, 85 % das caselas forneceram desvios padrão ($\hat{\sigma}_{boot}$) acima de 1) (Tabela 10).

Tabela 9 - Estimativas Bootstrap para cada posição (i, j) em falta da matriz com 10% de perda

Posições	VO	$\hat{y}_{boot}$	$\hat{V}_{boot}$	$\hat{\sigma}_{boot}$
1=(5,1)	15,36	16,894	0,4115	0,641
2=(7,1)	14,57	16,894	0,419	0,647
3=(11,1)	16,50	16,903	0,401	0,633
4=(11,2)	22,98	22,649	1,044	1,022
5=(4,3)	16,28	16,894	0,694	0,833
6=(7,5)	12,47	12,729	0,521	0,722
7=(15,5)	12,75	12,726	0,531	0,7284
8=(2,6)	17,77	18,659	0,621	0,788
9=(7,6)	17,01	18,677	0,609	0,7808
10=(18,6)	18,82	18,697	0,565	0,752
11=(1,7)	14,72	13,192	0,618	0,786
12=(2,7)	14,44	13,184	0,612	0,782
13=(8,7)	14,37	13,182	0,586	0,765
14=(14,7)	13,60	13,199	0,589	0,768

VO: valor original

$\hat{y}_{boot}$: média das imputações

$\hat{V}_{boot}$: variância das B estimativas *bootstraps* ($B=2000$)

$\hat{\sigma}_{boot}$: desvios padrão em cada conjunto de valores imputados

As Figuras 4 e 5 mostram o comportamento das 2000 estimativas, em cada casela em falta, conforme a porcentagem de perda (Gráfico de Caixas). Os pontos verdes, representam os valores estimados pelo método proposto e os pontos vermelhos, os valores originais do conjunto de dados. Desta forma, pode-se visualmente observar que, à 10% de perda (Figura 4) metade das caselas apresentaram valores estimados próximos dos reais e a variabilidade das estimativas em cada casela parecem ser similares, exceto para casela (11,2) que se refere ao 4º gráfico, o qual mostra uma variabilidade maior, assim como foi discutido anteriormente.

Figura 4 - Gráfico de caixas para os valores das 2000 estimativas obtidas pelo método IAMMI-BNP em cada posição de uma matriz com 10% de perda

A 20% de perda, nota-se que a maioria das caselas apresentaram estimativas próximas aos valores reais. Quanto a variabilidade das estimativas, o gráfico mostra o que foi discutido com base nas informações da Tabela 10.

Figura 5 - Gráfico de caixas para os valores das 2000 estimativas obtidas pelo método IAMMI-BNP em cada posição de uma matriz com 20% de perda

Tabela 10 - Estimativas *Bootstrap* para cada posição (i, j) em falta da matriz com 20% de perda

Posições	VO	$\hat{y}_{boot}$	$\hat{V}_{boot}$	$\hat{\sigma}_{boot}$
1=(5,1)	15,36	16,89551	0,4150305	0,6442286
2=(7,1)	14,57	16,89710	0,4378385	0,6616937
3=(11,1)	16,50	16,90580	0,4110385	0,6411228
4=(11,2)	22,98	22,65583	0,9353329	0,9671261
5=(20,2)	22,15	22,65864	0,8730717	0,9343831
6=(4,3)	16,28	16,90005	0,7934651	0,8907666
7=(9,3)	16,84	16,88051	0,7734980	0,8794873
8=(1,4)	20,61	18,22060	0,9889029	0,9944360
9=(7,4)	15,06	18,21182	1,049115	1,024263
10=(10,4)	18,94	18,19846	1,047932	1,023686
11=(13,4)	18,78	18,19649	1,026785	1,013304
12=(15,4)	17,91	18,19467	1,041907	1,020739
13=(18,4)	18,76	18,23584	1,021581	1,010733
14=(6,5)	11,91	12,71159	0,5677636	0,7535009
15=(7,5)	12,47	12,73654	0,5632928	0,7505283
16=(9,5)	13,54	12,71368	0,5517936	0,7428281
17=(10,5)	13,71	12,72698	0,5423296	0,7364303
18=(13,5)	11,92	12,73397	0,5527634	0,7434806
19=(15,5)	12,75	12,7250	0,5782813	0,7604481
20=(2,6)	17,77	18,61058	0,6612663	0,8131828
21=(6,6)	19,22	18,62481	0,6589161	0,8117364
22=(7,6)	17,01	18,64110	0,6461593	0,8038404
23=(18,6)	18,82	18,66891	0,6020267	0,7759038
24=(1,7)	14,72	13,11724	0,4551039	0,6746139
25=(2,7)	14,44	13,10694	0,4527096	0,6728370
26=(8,7)	14,37	13,10205	0,4525568	0,6727234
27=(14,7)	13,60	13,12377	0,4398993	0,6632491
28=(18,7)	14,33	13,11341	0,4482645	0,6695256

VO: valor original

$\hat{y}_{boot}$: média das imputações

$\hat{V}_{boot}$: variância das B estimativas *bootstraps* ($B=2000$)

$\hat{\sigma}_{boot}$: desvios padrão em cada conjunto de valores imputados

Na Figuras 6 e 7, encontram-se os gráficos de dispersão das diferenças entre os valores originais e os valores estimados pelo método de imputação nas diferentes porcentagens de perda. Graficamente tem-se que, em ambas as porcentagens, as estimativas das observações faltantes aparentemente foram próximas dos valores reais, com exceção da posição $(2=7,1)$ a 10% e 20%, as posições $(8=(1,4), 9=(7,4))$ a 20% de perda.

Figura 6 - Gráfico de dispersão da diferença entre o valor estimado e o valor original a 10% nas posições estimadas

Figura 7 - Gráfico de dispersão da diferença entre o valor estimado e o valor original a 20% nas posições estimadas

A Tabela 11 apresenta valores de algumas estatísticas que foram utilizadas para verificar a eficiência do método de imputação, ao se estimar os valores em falta. Nela se encontra, o valor- p do teste das ordens assinaladas de Wilcoxon (Teste de Wilcoxon), o valor do teste da raiz quadrada da diferença preditiva média (RMSPD) e o valor da correlação de Spearman para ambas porcentagens de perda. Desta forma pode-se observar que, por meio do teste de Wilcoxon, em ambas as porcentagens de falta, foi verificado que não há evidências para acreditar que os valores estimados sejam diferentes dos valores originais, ao nível de 5% de significância. A estatística da RMSPD mostrou menor valor a 20% de perda, indicando que o método de imputação foi mais eficiente na estimação quando o conjunto de dados apresentou 20% de perda. Observando os valores do coeficiente de correlação de Spearman, para ambas as porcentagens de perda, foram próximos de 1, sendo menor quando um conjunto de dados tem 20% de falta, indicando eficiência nas estimativas em cada porcentagem de falta.

Tabela 11 - Valores do teste das ordens assinaladas de Wilcoxon, do teste da raiz quadrada da diferença preditiva média (RMSPD) e da correlação de Spearman nas diferentes porcentagens de perda

Testes	10%	20%
Teste Wilcoxon (valor-p)	0,5416	0,8314
RMSPB	0,2992	0,2254
Correlação Spearman	0,9297	0,9146

As Tabelas 12 e 13, apresentam os valores das estimativas agrupadas dos ambientes em cada porcentagem de perda. Assim, independente da porcentagem de perda, o aumento na variância devido a falta de dados (r) e a precisão da estimativa se nenhum dado estivesse ausente (λ) apresentaram valores pequenos, indicando pouca influência dos dados ausentes nas estimativas das médias em cada ambiente. A eficiência relativa (ER) em todos os casos de falta se mostrou com valor 1, indicando alta eficiência nas estimativas obtida pelo método de imputação. Isto pode ter ocorrido devido ao número grande de estimativas obtidas (2000 estimativas).

O teste das ordens assinaladas de Wilcoxon foram não significativos ao nível de 5% de significância em todos os ambientes nos diferentes casos de falta de dados (10% e 20%) ($p > 0,05$). Assim, as médias em cada ambiente, após a imputação de dados, foram consideradas estatisticamente semelhantes as médias originais.

Tabela 12 - Estimativas agrupadas ($\hat{Q}$, $\bar{U}$, B , T), valores do incremento relativo da variância (r), da taxa de unidades ausentes (λ) e a eficiência relativa (ER) em cada ambiente com valores imputados (matriz de dados continha 10% de perda)

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
U	0,56715	1,76063	1,15202	2,35448	0,97437	1,17028	0,817856
Q	16,8851	22,66049	16,91169	18,21100	12,76076	18,67163	13,17282
VO	16,6720	22,6770	16,8810	18,2110	12,7490	18,5500	13,3915
B	0,00307	0,00261	0,00174	0,00000	0,00280	0,00493	0,00599
r	0,00541	0,00148	0,00151	0,00000	0,00287	0,00421	0,00732
λ	2,899e-08	2,192e-09	2,265e-09	0	8,227e-09	1,759e-08	5,270e-08
ER	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
T	0,57022	1,76324	1,15376	2,35448	0,97717	1,17522	0,82385
TW	94,5	118	112	101,5	128	103	140,5
p	0,7089 ^{ns}	0,1634 ^{ns}	0,8082 ^{ns}	0,9108 ^{ns}	0,1908 ^{ns}	0,9553 ^{ns}	0,1913 ^{ns}

TW : Estatística de Wilcoxon; p : valor p do teste de Wilcoxon; VO : valor original

Tabela 13 - Estimativas agrupadas ($\hat{Q}$, $\bar{U}$, B , T), valores do incremento relativo da variância (r), da taxa de unidades ausentes (λ) e a eficiência relativa (ER) em cada ambiente com valores imputados (matriz de dados continha 20% de perda)

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
U	0,56886	1,78420	1,19533	1,77786	0,93458	1,19171	0,73419
Q	16,88542	22,68622	16,91403	18,18089	12,75139	18,63627	13,09667
VO	16,6720	22,6770	16,8810	18,2110	12,7490	18,5500	13,3915
B	0,00313	0,00473	0,00397	0,01804	0,00907	0,00719	0,00575
r	0,00550	0,00265	0,00332	0,01014	0,00970	0,00603	0,00782
λ	2,992e-08	7,003e-09	1,098e-08	1,005e-07	9,209e-08	3,589e-08	6,020e-08
ER	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000
T	0,57199	1,78894	1,19931	1,79591	0,94365	1,1989	0,73994
TW	94,5	114	102	124,5	118	117	156,5
p	0,7089 ^{ns}	0,2226 ^{ns}	0,9256 ^{ns}	0,478 ^{ns}	0,3652 ^{ns}	0,6677 ^{ns}	0,05687 ^{ns}

TW : Estatística de Wilcoxon; p : valor p do teste de Wilcoxon; VO : valor original

5.2 Resultado 2: estudo de simulação 2

As 200 estimativas *Bootstraps* para cada casela vazia foi obtida, para cada modelo de imputação considerado ($IAMMI_0$ -BNP, $IAMMI_1$ -BNP e $IAMMI_2$ -BNP), as quais estão apresentadas nas Figuras 8, 9 e 10. Assim, observa-se que as estimativas das posições apresentam variabilidade similar em ambos os modelos. Porém, o modelo $IAMMI_2$ -BNP forneceu uma variabilidade maior das 200 estimativas das posições. E por fim, estimativas finais ($\hat{y}_{bootij}$) estão na Tabelas 8.

Figura 8 - Gráfico de caixas para os valores estimados $IAMMI_0$ -BNP dos dados com 10% de perda

Figura 9 - Gráfico de caixas para os valores estimados pela $IAMMI_1$ -BNP com 10% de perda

Figura 10 - Gráfico de caixas para os valores estimados pela $IAMMI_2$ -BNP Bootstrap Não-paramétrico dos dados com 10% de perda

Tabela 14 - Valores originais das alturas (coluna 2), estimativa da média de cada ambiente (coluna 3), valores estimados pelo método $IAMMI_0$ -BNP, $IAMMI_1$ -BNP e $IAMMI_2$ -BNP para cada casela em falta

Posição	Original	Média	$IAMMI_0$ -BNP	$IAMMI_1$ -BNP	$IAMMI_2$ -BNP
(1,7)	14,42	13,169	13,172	13,155	13,149
(2,6)	17,77	18,670	18,682	18,659	18,681
(2,7)	14,44	13,169	13,179	13,210	13,206
(4,3)	16,28	16,913	16,882	16,858	16,808
(5,1)	15,36	16,883	16,879	16,906	16,864
(7,1)	14,57	16,883	16,839	16,847	16,822
(7,5)	12,97	12,764	12,688	12,699	12,618
(7,6)	17,00	18,671	18,653	18,648	18,616
(8,7)	14,37	13,169	11,140	13,175	13,232
(11,1)	16,50	16,883	16,899	16,846	16,920
(11,2)	22,98	22,661	22,670	22,642	22,639
(14,7)	13,60	13,169	13,1215	15,117	13,060
(15,5)	12,75	12,764	12,727	12,705	12,656
(18,6)	18,82	18,671	18,702	18,719	18,775

Na Figuras 11, 12, 13, os pontos representam a diferença entre o valor estimado ($\hat{y}_{bootij}$) e o original (VO). Desta forma, pode-se visualmente observar que, algumas caselas apresentaram distância acima de 1 ((5,1), (7,1) e (7,6)) e abaixo de -1 ((1,7), (2,7) e (8,7)) para ambos os modelos de imputação

Figura 11 - Gráfico de dispersão da diferença entre os valores estimados e o valor original a 10% de perda nas posições estimadas pela $IAMMI_0$ bootstratp Não-paramétrico

As Tabelas 15, 16 e 17 apresentam os valores das estimativas agrupadas dos ambientes ao se utilizar cada modelo de imputação. Assim, observa-se que, independentemente do modelo de imputação considerado o aumento na variância devido a falta de dados (r) e a precisão da estimativa se nenhum dado estivesse ausente (λ) apresenta valores pequenos, indicando pouca influência dos dados ausentes nas estimativas das médias em cada ambiente. A eficiência relativa (ER) em todos os casos de falta se mostrou com valor 1 indicando alta eficiência nas estimativas obtida pelo método de imputação. Isto ocorreu devido ao número grande de estimativas obtidas (200 estimativas).

Figura 12 - Gráfico de dispersão da diferença entre os valores estimados e o valor original a 10% de perda nas posições estimadas pela $IAMMI_1$ bootstratp Não-paramétrico

Figura 13 - Gráfico de dispersão da diferença entre os valores estimados e o valor original a 10% de perda nas posições estimadas pela $IAMMI_2$ bootstratp Não-paramétrico

Tabela 15 - Estimativas agrupadas ($\hat{Q}$, $\bar{U}$, B , T), valores do incremento relativo da variância (r), da taxa de unidades ausentes (λ) e a eficiência relativa (ER) em cada ambiente com valores imputados (dados 1 com 10% de perda)

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
U	0,589	1,736	1,145	2,354	0,984	1,174	0,821
Q	16,881	22,661	16,912	18,211	12,759	18,671	13,169
VO	16,672	22,677	16,881	18,211	12,749	18,550	13,392
B	0,004	0,003	0,003	0,000	0,003	0,005	0,006
r	0,006	0,001	0,002	0,000	0,003	0,004	0,007
λ	3,72e-07	6,25e-09	2,05e-08	0,000	9,62e-08	1,99e-07	5,92e-07
ER	1,000	1,000	1,000	1,000	1,000	1,000	1,000
T	0,593	1,737	1,152	2,354	0,977	1,179	0,828
TW	173,5	201,5	195,5	200	197	189	236
p	0,482 ^{ns}	0,978 ^{ns}	0,914 ^{ns}	1 ^{ns}	0,946 ^{ns}	0,787 ^{ns}	0,336 ^{ns}

TW : Estatística de Wilcoxon; p : valor p do teste de Wilcoxon; p : valor original

Tabela 16 - Estimativas agrupadas ($\hat{Q}$, $\bar{U}$, B , T), valores do incremento relativo da variância (r), da taxa de unidades ausentes (λ) e a eficiência relativa (ER) em cada ambiente com valores imputados com o modelo $IAMMI_1$ -BNP (dados 1 com 10% de perda)

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
U	0,566	1,758	1,158	2,355	0,977	1,172	0,821
Q	16,882	22,661	16,910	18,211	12,759	18,672	13,169
VO	16,672	22,677	16,881	18,211	12,749	18,550	13,392
B	0,003	0,003	0,002	0,000	0,003	0,005	0,006
r	0,005	0,001	0,002	0,000	0,003	0,004	0,007
λ	2,7e-07	1,9e-08	3,2e-08	0,000	8,1e-08	2,01e-07	5,92e-07
ER	1,000	1,000	1,000	1,000	1,000	1,000	1,000
T	0,568	1,759	1,160	2,377	0,979	1,178	0,827
TW	174,5	201,5	197,5	200	197	189,5	236
p	0,498 ^{ns}	0,978 ^{ns}	0,957 ^{ns}	1 ^{ns}	0,946 ^{ns}	0,787 ^{ns}	0,336 ^{ns}

TW : Estatística de Wilcoxon; p : valor p do teste de Wilcoxon; p : valor original

Tabela 17 - Estimativas agrupadas ($\hat{Q}$, $\bar{U}$, B , T), valores do incremento relativo da variância (r), da taxa de unidades ausentes (λ) e a eficiência relativa (ER) em cada ambiente com valores imputados com o modelo $IAMMI_2$ -BNP (dados 1 com 10% de perda)

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
U	0,660	1,874	1,263	2,354	1,142	1,466	1,027
Q	16,877	22,659	16,907	18,211	12,752	18,673	13,167
VO	16,672	22,677	16,881	18,211	12,749	18,550	13,392
B	0,008	0,008	0,007	0,000	0,011	0,022	0,018
r	0,017	0,004	0,006	0,000	0,009	0,015	0,027
λ	1,4e-06	1,9e-07	3,3e-07	0,000	9,23e-07	2,10e-06	2,89e-06
ER	1,000	1,000	1,000	1,000	1,000	1,000	1,000
T	0,668	1,882	1,270	2,354	1,153	1,488	1,045
TW	177,5	202,5	198,5	200	198	189,5	236
p	0,552 ^{ns}	0,957 ^{ns}	0,978 ^{ns}	1 ^{ns}	0,968 ^{ns}	0,787 ^{ns}	0,336 ^{ns}

TW : Estatística de Wilcoxon; p : valor p do teste de Wilcoxon; p : valor original

A Tabela 18 apresenta valores de algumas estatísticas que foram utilizadas para verificar a eficiência dos métodos de imputação considerados a 10% de falta. Nela, encontra-se o valor p do teste das ordens assinaladas de Wilcoxon (Teste de Wilcoxon), o valor do teste da raiz quadrada da diferença preditiva média (RMSPD) e o valor da correlação de Spearman para ambos modelos. Desta forma pode-se observar que, por meio do teste de Wilcoxon, em ambos os modelos, foi verificado que não há evidências para acreditar que os valores estimados sejam diferentes dos valores originais ao nível de 5% de significância. A estatística da RMSPD mostrou valores similares em ambos os modelos, porém o menor valor foi obtido com o modelo $IAMMI_2$ -BNP, indicando que o método de imputação com dois eixos foi mais eficiente na estimação das unidades ausentes. Observando os valores do coeficiente de correlação de Spearman, para ambos os modelos, foram próximos de 1, sendo menor quando o conjunto de dados foi estimado pelo modelo $IAMMI_2$ -BNP, indicando uma eficiência nas estimativas com 10% de perda.

Tabela 18 - Valores do teste das ordens assinaladas de Wilcoxon, do teste da raiz quadrada da diferença preditiva média (RMSPD) e da correlação de Spearman aos diferentes modelos de imputação

Testes	$IAMMI_0$ -BNP	$IAMMI_1$ -BNP	$IAMMI_2$ -BNP
Teste Wilcoxon (valor-p)	0,626	0,626	0,715
RMSPD	0,296	0,297	0,293
Correlação Spearman	0,945	0,969	0,943

6 CONCLUSÃO

Diante dos resultados obtidos pode-se concluir, que nas diferentes porcentagens de perda (10% e 20%), em ambos os estudos de simulação, a variabilidade das estimativas em cada casela em falta foram aparentemente semelhantes. De forma geral, de acordo com os métodos de comparação o método de imputação proposto nos dois estudos de simulação forneceu valores imputados próximos dos originais. Porém, no estudo de simulação 1, a eficiência do método foi melhor com 20% de perda. Entretanto, considerando os estudos de simulação com 10% de perda, a eficiência do método de imputação proposto foi melhor quando se utilizou o modelo $IAMMI_2$ -BNP (com dois eixos multiplicativos). O teste das ordens assinaladas de Wilcoxon mostrou que os valores imputados não influenciaram na estimativa da média, indicando que valores médios (considerando dados imputados em cada ambiente) foram estatisticamente semelhantes aos valores médios originais, em cada ambiente.

REFERÊNCIAS

- ARCINIEGAS-ALARCÓN, S.P; DIAS, C. T. dos S. Imputação de dados em experimentos com interação genótipo por ambiente: Uma aplicação a dados de algodão. **Revista Brasileira de Biometria**, São Paulo, v.27, n.1, p.125-138, 2009.
- ARCINIEGAS-ALARCÓN, S.P; GARCÍA-PENÁ, M.; KRZANOWSKI, W. J.; DIAS, C. T. dos S. Deterministic Imputation in Multienvironment Trials. **ISRN Agronomy**, v. 2013, 17p., 2013. Disponível em <http://dx.doi.org/10.1155/2013/978780>.
- AYUYEV, V. V.; JUPIN, J.; HARRIS, P. W.; OBRADOVIC, Z. Dynamic Clustering-Based Estimation of Missing Values in Mixed Type Data. **Data warehousing and knowledge discovery**, Lecture notes in computer science, V. 5691, p.366-377, Springer, 2009.
- BERGAMO, G. C. **Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação**, 2007. 89 p. 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. dos 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.
- 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.
- CROSSA, J. Statistical analysis of multilocation trials. **Advance in Agronomy**, New York, v.44, p.55-85, 1990.
- 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.
- DUARTE, J. B.; VENCOVSKY, R. **Interação genótipo $\times$ ambiente**: Uma introdução à análise "AMMI". Ribeirão Preto, S.P, 60p., 1999.
- EFRON, B. Bootstrap methods: another look at jakknife. **Annals of Statistics**, Madison, V.7, n.1, p.1-26, 1979.
- ENDERS, C.K. **Applied Missing Data Analysis**. Guilford Press, Inc.72 Spring Street, New York, 2010.
- FARIA, A. P.; MODA-CIRINO, V.; BURATTO, J. S.; SILVA, F. B. da; DESTRO, D. Interação genótipo $\times$ ambiente na produtividade de grãos de linhagens e cultivares de feijão. **Acta Scientiarum. Agronomy**, Maringá, v.31, n.4, p.579-585, 2009.
- FARIA, P.N. **Utilização de técnicas multivariadas na análise da divergência genética via modelo AMMI com reamostragem *Bootstrap***. 2012. 128 p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2012.
- GABRIEL, K. R. The biplot graphic display of matrices with application to principal component analysis. **Biometrika**, Oxford, v.58, n.3, p.453-467, 1971.
- GAUCH, H. G. Model selection and validation for yield trials with interaction. **Biometrics**, Washington, v.44, p.705-715, 1988.
- GAUCH JÚNIOR, H. G. **Statistical analysis of regional yields trials: AMMI analysis of factorial designs**. Amsterdam. Elsevier, 278p., 1992.

GAUCH JÚNIOR, H. G.; ZOBEL, R. W. AMMI analysis of yields trials. In: Kang, M. S.; GAUCH JÚNIOR, H.G. **Genotype-by-environment interaction**. boca Raton: CRC Press, cap.4, p.85-122, 1996.

HE, Y. Missing Data Imputation for Tree-Based Models. 2006. 81p. Dissertation (Doctor of Philosophy in Statistics). University of California, Los Angeles, 2006.

JELICIC, H.; PHELPS, E.; LERNER, R. M. Use of Missing Data Methods in Longitudinal Studies: The Persistence of Bad Practices in Developmental Psychology. **Developmental Psychology**, v.45, No.4, p. 1195–1199, 2009.

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

LAVORANTI, O. J. **Estabilidade e adaptabilidade fenotípica através da reamostragem “bootstrap” no modelo AMMI**. 2003. 166 p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2003.

LOBO, R. S; KALIL, R. C. E. **Um estudo de caso do impacto da imputação de dados na mineração de dados**. 2006. 100 p. Trabalho de Conclusão de Curso (Bacharel em Ciência da Computação)- Universidade Federal do Pará, Belém, 2006.

MANDEL, J. A new analysis of variance model for nonadditive data. **Technometrics**, Washington, v.13, n.1, p.1-18, 1971.

MEOTTI, G. V.; BENIN, G.; SILVA, R. R.; BECHE, E.; MUNARO, L. B. Épocas de semeadura e desempenho agrônômico de cultivares de soja. **Pesq. agropec. bras.**, Brasília, v.47, n.1, p.14-21, jan. 2012.

MORI, E. S.; LELLO, L. R. B. de; KAGEVAMA, P. Y. Efeitos de interação genótipos X ambientes em profênes de *Eucalyptus saligna* SMITH. **Instituto de Pesquisas e Estudos Florestais - IPEF**, n.33, p.19-25, Agos. 1986.

OLIVEIRA, R. L. R. de. **Imputação de médias para a análise de estabilidade e adaptabilidade em experimentos conjuntos incompletos: uma aplicação em café conilon**, 2012. 52 p. Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Viçosa, Viçosa-MG, 2012.

R Core Team (2016). R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. URL <https://www.R-project.org/>.

RUBIN, D. B. Inference and Missing Data. **Biometrika**, V.63, Issue 3 , p. 581-592, 1976.

RUBIN, D. B. Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. **Proceedings of the Survey Research Methods Section**, pages 20-28, 1978.

SILVA, M. J. C. da. **Imputação múltipla: uma comparação e eficiência em experimentos multiambientais**, 2012. 122 p. Dissertação (Mestre em Ciências com ênfase em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2012.

SILVA FILHO, J. L. da; MORELLO, C. de L.; FARIAS, F. J. C.; LAMAS, F. M.; PEDROSA, M. B.; RIBEIRO, J. L. Comparação de métodos para avaliar a adaptabilidade e estabilidade produtiva em algodoeiro. **Pesq. agropec. bras.**, Brasília, v.43, n.3, p.349-355, mar. 2008.

SHUKLA, G. K. Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity*, v.29, n.2, p.237-245, 1972.

SQUILASSI, M. G. **Interação de genótipos com ambientes**. 2003, 47p. Disponível em <http://www.cpatc.embrapa.br>.

TOPPA, E. V. B.; JADOSKI, C. J. O uso dos marcadores moleculares no melhoramento genético de plantas. **Scientia Agraria Paranaensis – SAP**, Mal. Cdo. Rondon, v 12, n 1, jan-mar, p.1-5, 2013.

WRICKE, G. Zur berechnung der okovalenz bei sommerweizen und hafer. **Pflanzenzuchtung**, v.52, p.127-138, 1962.

APÊNDICES

APÊNDICE A - Processo de imputação utilizado no estudo de simulação 1

A.1 Programação para gerar a falta na matriz de dados originais e depois substituí-la pela média

```
#>GERANDO ALEATORIAMENTE AS POSIÇÕES EM FALTA>>>>
matriz_falta<-function(matriz,p,seed){
  p=(p)*100
  f<-round((p*length(matriz))/100)
  set.seed(seed)
  na<-runif(length(matriz))
  uni<-order(na) [1:f]
  matriz[uni]<-NA
  matriz
}
mf<-matriz_falta(matriz=dados,p=0.10,seed=6861944) #10 % de falta

#-----FUNÇÃO IMPUTE MÉDIA-----

#Substitui a posição em falta com a média da sua respectiva coluna

impute.media<-function(x){
  nc=ncol(x)
  nl=nrow(x)
  xc=matrix(c(rep(0,nc*nl)),nl,nc)
  xm<-apply(x,2,mean,na=T)
  for(j in 1:nc){
    coluna=x[,j]
    peso_faltante1=is.na(coluna) #qual o valor que ? NA?
    coluna[peso_faltante1]=xm[j]
    xc[,j]=coluna
  }
  return(xc)
}

yc=impute.media(x=mf) #MATRIZ COMPLETADA PELAS MÉDIAS
```

A.2 Programação para estimar os dados em falta usando o método de imputação AMMI Bootstrap Não-paramétrico

#Fornece uma matriz boot com o valor em falta estimado

```
imp_boot_AMMIb<-function(matriz,mf){
  ng=nrow(matriz) #n?mero de linhas
  nl=nrow(matriz)
  nc=ncol(matriz)

  res<-matrix(0,nl,nc)
  for(j in 1:nl){
    lin<-round(runif(1,1,nl))
    res[j,]<-matriz[lin,]
  }
  boot<-res
```

```

#Ajuste dos efeitos principais
yb..<-(sum(boot))/(nl*nc)
yb..2<-matrix(c(rep(yb..,nc*nl)),nl,nc)
ybi.<-apply(boot,1,mean)
ybi._2<-matrix(rep(ybi., nc),nl,nc)
yb.j<-apply(boot,2,mean,na=T)
yb.j_2<-matrix(rep(yb.j, each=nl),nl,nc)

gi<-yb.-yb.. #efeito dos genotipos
ej<-yb.j-yb.. #efeito de ambientes
geij=boot-ybi._2-yb.j_2+yb..

#>>>>>>Ajuste da interação G X E POR DVS>>>>>>>>>
dvs.ge<-svd(geij)
Lambda<-round(dvs.ge$d,digits=2)
lamb<-diag(Lambda)
lambda2<-Lambda^2
total<-sum(lambda2)
#proporção da soma de quadrado por eixo
prop<-lambda2/total
acum.p<-cumsum(prop)
eixo<-seq(1:nc)
fim<-cbind(eixo,lambda2,prop,acum.p)
#Escolhendo o número de eixos
p_expl<-acum.p[acum.p<0.95]
#determinei pergatar os eixos com porcentagem
# explicada abaixo de 0.95
neixo<-length(p_expl)

lambda<-dvs.ge$d
gama<-dvs.ge$u
alpha<-dvs.ge$v

imp<-matrix(rep(0,nc*nl),nl,nc)
for(l in 1:nl){
  for(j in 1:nc){

    if(is.na(mf[l,j])){
      mais<-0
      neixo=1
      for(k in 1:neixo) {
        yeij<-yb..+gi[l]+ej[j]+(mais+(lambda[k]*gama[l,k]*alpha[j,k]))
        mais<-yeij
      }
      imp[l,j]<-yeij #y11
    }
    else {
      imp[l,j]<-mf[l,j]
    }
    #saidas[[nb]]<-imp
  }
}
return(imp)

```

```

}

#Fornece as b matriz Bootstrap com os valores em falta

boots<-function(matriz,mf,nb){
  set.seed(6861944)
  bi<-list()
  for(i in 1:nb){
    bi[[i]]=imp_boot_AMMIb(matriz=matriz,mf=mf)
  }
  mais<-matrix(0,nrow(mf),ncol(mf))
  for(b in 1:nb){
    boot_imp<-(mais+bi[[b]])
    mais<-boot_imp
  }
  #return(bi)
  boot_imp<-(1/nb)*(boot_imp)
  bi$boot_imp<-boot_imp
  list(bi=bi)
}

matrizes<-boots(matriz=yc,mf=mf,nb=2000)

posicao.est<-function(matrix_boots,mf,nb){
  n<-length(mf)
  #separa amostra
  posicao_falta<-is.na(mf)
  pf<-mf[posicao_falta]
  nf<-length(pf) #n?mero de posi??es em falta
  resul<-matrix(0,nb,n)
  Bboots<-c()
  for(i in 1:n){
    if(is.na(mf[i])){
      for(b in 1:nb){
        Bboots[b]<-matrix_boots[[b]][i]
      }
      resul[,i]<-Bboots
    }
  }
  k=0
  est.boots<-matrix(0,nb,1)
  zero<-matrix(0,nb,1)
  for(p in n:1){
    if(is.na(mf[p])){
      est.boots[,k+1]<-resul[,p]
      est.boots<-cbind(zero,est.boots)
    }
  }
  est.boots<-est.boots[,2:ncol(est.boots)]
  list(est.boots= est.boots)
}

```

```

posicao<-posicao.est(matrix_boots=matrizes$bi,mf=mf,nb=2000)
boxplot(posicao$est.boots,xlab="Posições estimadas",ylab="Estimativas bootstrap AMMI")
#>>>ESTIMATIVA FINAL BOOT DE CADA POSICAO>>>>>>

yboot_est<-function(matrizes_boot,mf,nb){
  nc<-ncol(mf)
  nl<-nrow(mf)
  for(j in 1:nc){
    for(i in 1:nl){

      if(is.na(mf[i,j])){
        mais<-0
        mais2<-0
        for(b in 1:nb){
          ybootij<-mais+matrizes_boot[[b]][i,j]
          mais<-ybootij
        }
        ybootij=ybootij/nb
        for(b in 1:nb){
          somabootij<-mais2+(matrizes_boot[[b]][i,j]-ybootij)^2
          mais2<-somabootij
        }
        vbootij<-somabootij/(nb-1)
        dpbootij<-sqrt(vbootij)
        #cat("ybootij",i,j,ybootij,vbootij,"\n")
        cat("ybootij",i,j,ybootij,vbootij,dpbootij,"\n")
      }
    }
  }
}
est_final<-yboot_est(matrizes_boot=matrizes$bi,mf=mf,nb=2000)

```

A.3 Programação para realizar os cálculos dos métodos de comparação

#RMSPD E CORRELAÇÃO DE SPEARMAN

```

medidas.relacao<-function(mo,mi,mf){
  posicao_falta<-is.na(mf)
  pf<-mf[posicao_falta]
  nf<-length(pf)
  n<-length(mf)
  soma<-0
  for(i in 1:n){
    if(is.na(mf[i])){
      rmspd<-soma+(mo[i]-mi[i])^2
      soma<-rmspd
    }
  }
  RMSPD=sqrt(rmspd)/nf

  p1<-c(rep(0,n))
  p2<-c(rep(0,n))

```



```

nf<-length(posicao.falta)
k=0
est.boots<-matrix(0,nb,1)
zero<-matrix(0,nb,1)
for(p in n:1){
  if(is.na(mf[p])){
    est.boots[,k+1]<-resul[,p]
    est.boots<-cbind(zero,est.boots)
  }
}
est.boots<-est.boots[,2:ncol(est.boots)]

list(est.boots= est.boots)
}

#nb estimativas para cada posição estimada
posicao_est=posicao.est(matrix_boots=matrix_boot,mf=mf,nb=200)
estimativas<-as.matrix(posicao_est$est.boots)

boxplot(estimativas,xlab="Posições estimadas",
ylab="Estimativas AMMI0 bootstrap-NP(nb=200)")
medias<-apply(estimativas,2,mean)
points(medias, col="red", add=T,pch=16)

#boxplot(estimativas,xlab="Posições estimadas",ylab="Estimativas bootstrap AMMI")

#>>>ESTIMATIVA FINAL BOOT DE CADA POSICAO>>>>

yboot_est<-function(matrizes_boot,mf,nb){
  nc<-ncol(mf)
  nl<-nrow(mf)
  for(j in 1:nc){
    for(i in 1:nl){

      if(is.na(mf[i,j])){
        mais<-0
        mais2<-0
        for(b in 1:nb){
          ybootij<-mais+matrizes_boot[[b]][i,j]
          mais<-ybootij
        }
        ybootij=ybootij/nb
        for(b in 1:nb){
          somabootij<-mais2+(matrizes_boot[[b]][i,j]-ybootij)^2
          mais2<-somabootij
        }
        vbootij<-somabootij/(nb-1)
        dpbootij<-sqrt(vbootij)
        #cat("ybootij",i,j,ybootij,vbootij,"\n")
        cat("ybootij",i,j,ybootij,vbootij,dpbootij,"\n")
      }
    }
  }
}

```

```

    }
  }
  #yboot_est(matrizes_boot=matrix_boot,mf=mf,nb=200)

  #Gera a matriz final com as estimativas
  # boots calculadas e #substituidas<

  matrix_final<-function(matrizes_boot,mf,yp,nb){
    nc=-ncol(mf)
    nl=-nrow(mf)
    matriz_final<-matrix(rep(0,nc*nl),nl,nc)

    for(j in 1:nc){
      for(i in 1:nl){
        if(is.na(mf[i,j])){
          mais<-0
          for(b in 1:nb){
            ybootij<-mais+matrizes_boot[[b]][i,j]
            mais<-ybootij
            ybootij=ybootij/nb
            matriz_final[i,j]=ybootij
          }
          #ybootij=ybootij/nb
          #matriz_final[i,j]=ybootij
        }
        else if(!is.na(mf[i,j])) matriz_final[i,j]<-yp[i,j]
      }
    }
    list(matriz_final=matriz_final)
  }

  Yfp<-matrix_final(matrizes_boot=matrix_boot,mf=mf,yp=yp,nb=200)

  #RETORNA AO VALOR ORIGINAL
  mi=as.matrix(Yfp$matriz_final)

  matriz.ori<-function(mi,yc){
    nl=nrow(mi)
    nc=ncol(mi)
    matriz.imp=matrix(rep(0,nc*nl),nl,nc)

    for(z in 1:nc){
      xp=matrix(mi[,z])
      mb=mean(yc[,z])
      media=matrix(rep(mb,nl)) #Vetor
      variancia=var(yc[,z])
      dp=sqrt(variancia)
      xo=(dp*xp)+media
      matriz.imp[,z]=xo
    }
    return(matriz.imp)
  }
}

```

```
original=matriz.ori(mi=mi,yc=yc)
```

B.2 Programação para realizar a Imputação AMMI Bootstrap Não-paramétrico com um eixo ou dois eixo

```
#Matriz padronizada da matriz obtida yc no Anexo A
yp=scale(yc)
```

```
#FORNECE UMA MATRIZ COM O VALOR EM FALTA ESTIMADO
```

```
#Função de imputação quando se utiliza mais de um eixo no modelo de imputação
```

```
funcao.imp<-function(matriz, mf, neixo,lambda, gama, alpha, gi, ej,yb..){
  imp<-matrix(rep(0,ncol(matriz)*nrow(matriz)),
    nrow(matriz),ncol(matriz))
  for(j in 1:ncol(matriz)){
    for(i in 1:nrow(matriz)){
      #
      if(is.na(mf[i,j])){
        mais<-0
        #ADAPTAR ESTA SOMA PARA QUANDO NENHUM EIXO FOR SELECIONADO
        for(k in 1:neixo) {
          yeij<-yb..+gi[i]+ej[j]+(mais+(lambda[k]*gama[i,k]*alpha[j,k]))
          mais<-yeij
        }
        imp[i,j]<-yeij #y11
      }
      else {
        imp[i,j]<-matriz[i,j]
      }
    }
  }
  return(imp)
}

#dados_imp=funcao.imp(matriz=yp,neixo, mf=mf, lambda=lambda,
gama=gama, alpha=alpha, gi=gi, ej=ej, yb..=yb..)
```

```
result<-imp_boot_AMMIb(matriz=yc,mf=mf) #A mesma função utilizada
anteriormente quando não se tinha eixo algum no modelo
```

```
#NB MATRIZES BOOTS
```

```
matrix_boot=boots(matriz=yc,mf=mf,nb=200)
```

```
#POSIÇÕES COM SUAS ESTIMATIVAS BOOTS
```

```
posicao_est=posicao.est(matrix_boots=matrix_boot,mf=mf,nb=200)
```

```
estimativas<-as.matrix(posicao_est$est.boots)
```

```
#>>>ESTIMATIVA FINAL BOOT DE CADA POSICAO>>>>
```

```
Yfp<-matrix_final(matrizes_boot=matrix_boot,mf=mf,yp=yp,nb=200)
```

```
original=matriz.ori(mi=mi,yc=yc)
```

B.3 Programa para realizar o cálculos dos Métodos de comparação em amos os modelos

```
medidas.relacao<-function(mo,mi,mf){
```