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

Alternativas de análise para experimentos G×E multiatributo

Marisol García Peña

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

Piracicaba
2015

Marisol García Peña
Estatística

Alternativas de análise para experimentos G×E multiatributo

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
2015

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

García Peña, Marisol
Alternativas de análise para experimentos G×E multiatributo / Marisol García Peña. - - Piracicaba, 2015.
72 p. : il.

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

1. Interação genótipos × ambientes 2. Dados de tripla entrada 3. Atributos 4. Modelos AMMI 5. Análise de procrustes generalizado 6. Dados faltantes 7. Imputação múltipla I. Título

CDD 575.10212
G216a

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

AGRADECIMENTOS

A minha família, cujo amor e força estiveram perto de mim durante todos os momentos apesar da distância. A vocês, um enorme abraço.

Ao Professor Carlos Tadeu dos Santos Dias, pela confiança depositada em mim, sugestões e apoio ao longo deste trabalho.

À Professora Kaye Basford da Universidade de Queensland - Australia e ao Professor Peter Kroonenberg da Universidade de Leiden - Holanda, pela disponibilidade, sugestões e as orientações compartilhadas na distância.

Aos professores e funcionários do programa de Pós-graduação em Estatística e Experimentação Agronômica do Departamento de Ciências Exatas da ESALQ/USP, pela atenção, auxílios permanentes e amizade.

Ao Sergio, meu bem! pelos momentos vividos, apoio e incentivo constantes ao longo desta jornada.

Aos amigos do doutorado, em especial ao Kuang e ao Nilton, pela amizade sincera, risadas nos momentos de descontração e horas de trabalho compartilhado.

Ao programa CNPq-TWAS pela concessão da bolsa de estudos.

SUMÁRIO

RESUMO

Alternativas de análise para experimentos G×E multiatributo

Geralmente, nos experimentos genótipo por ambiente (G×E) é comum observar o comportamento dos genótipos em relação a distintos atributos nos ambientes considerados. A análise deste tipo de experimentos tem sido abordada amplamente para o caso de um único atributo. Nesta tese são apresentadas algumas alternativas de análise considerando genótipos, ambientes e atributos simultaneamente. A primeira, é baseada no método de mistura de máxima verossimilhança de agrupamento - Mixclus e a análise de componentes principais de 3 modos - 3MPCA, que permitem a análise de tabelas de tripla entrada, estes dois métodos têm sido muito usados na área da psicologia e da química, mas pouco na agricultura. A segunda, é uma metodologia que combina, o modelo de efeitos aditivos com interação multiplicativa - AMMI, modelo eficiente para a análise de experimentos (G×E) com um atributo e a análise de procrustes generalizada, que permite comparar configurações de pontos e proporcionar uma medida numérica de quanto elas diferem. Finalmente, é apresentada uma alternativa para realizar imputação de dados nos experimentos (G×E), pois, uma situação muito freqüente nestes experimentos, é a presença de dados faltantes. Conclui-se que as metodologias propostas constituem ferramentas úteis para a análise de experimentos (G×E) multiatributo.

Palavras-chave: Interação genótipos × ambientes; Dados de tripla entrada; Atributos; Modelos AMMI; Análise de procrustes generalizado; Dados faltantes; Imputação múltipla

ABSTRACT

Alternatives of analysis of G×E trials multi-attribute

Usually, in the experiments genotype by environment (G×E) it is common to observe the behaviour of genotypes in relation to different attributes in the environments considered. The analysis of such experiments have been widely discussed for the case of a single attribute. This thesis presents some alternatives of analysis, considering genotypes, environments and attributes simultaneously. The first, is based on the mixture maximum likelihood method - Mixclus and the three-mode principal component analysis, these two methods have been very used in the psychology and chemistry, but little in agriculture. The second, is a methodology that combines the additive main effects and multiplicative interaction models - AMMI, efficient model for the analysis of experiments (G×E) with one attribute, and the generalised procrustes analysis, which allows compare configurations of points and provide a numerical measure of how much they differ. Finally, an alternative to perform data imputation in the experiments (G×E) is presented, because, a very frequent situation in these experiments, is the presence of missing values. It is concluded that the proposed methodologies are useful tools for the analysis of experiments (G×E) multi-attribute.

Keywords: Genotypes × environments interaction; Three-way data; Attributes; AMMI models; Generalised procrustes analysis; Missing values; Multiple imputation

1 INTRODUÇÃO

Nos experimentos genótipo por ambiente ($G \times E$) é freqüente que seja observado o desempenho dos genótipos em relação a vários atributos nos distintos ambientes, na literatura podem ser encontrados diversos mecanismos de análise para estes experimentos, mas considerando apenas um atributo por vez. Aqui são apresentadas algumas alternativas de análise envolvendo genótipos, ambientes e atributos simultaneamente.

As técnicas de análise de tripla entrada resumem toda a informação presente nos dados (todos os efeitos principais e todas as interações) de forma eficiente. Em concreto, resumem cada modo por meio de poucos componentes e adicionalmente, descrevem as relações entre eles (Kiers e Mechelen, 2001).

Na análise de dados de tripla entrada a informação é apresentada em arranjos em que cada dado é indexado com três índices: um que identifica ao indivíduo i ($i = 1, \ldots, I$), outro à condição j ($j = 1, \ldots, J$) e um terceiro que corresponde à variável (atributo) k ($k = 1, \ldots, K$); representando em um arranjo tridimensional: indivíduos, condições e variáveis (Frutos, 2014). No caso dos experimentos $G \times E$: genótipos, ambientes e atributos, (Figura 1.1).

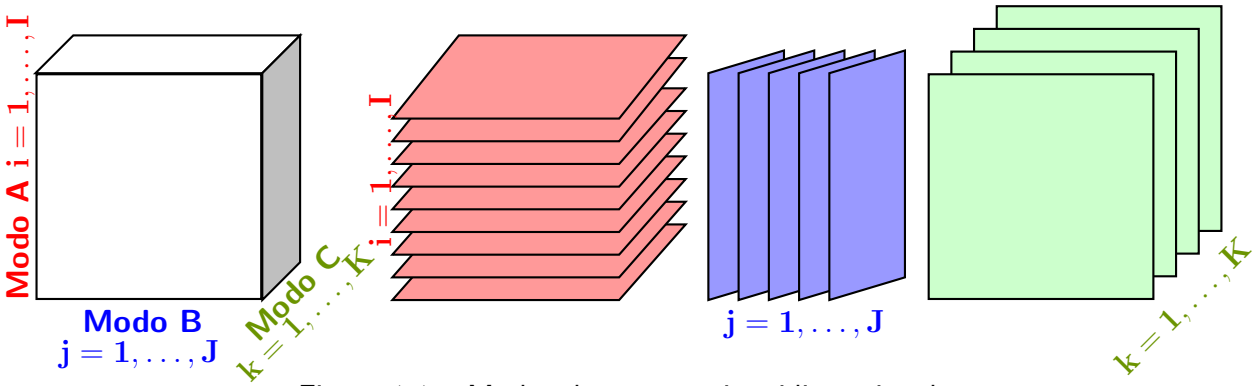


Figura 1.1 – Modos de um arranjo tridimensional

Os métodos utilizados para a análise de dados de tripla entrada são de caráter exploratório, porque identificam padrões e relações da estrutura interna presente entre os elementos das três entradas do arranjo, sem aplicar testes a estes padrões porque testar hipóteses assume uma pressuposição distribucional que em muitas ocasiões é desconhecida (Kroonenberg, 2008).

No capítulo 2, é feita uma revisão dos modelos Mixclus (método de mistura de máxima verossimilhança de agrupamento) proposto por Basford e McLachlan (1985) e 3MPCA (componentes principais de 3 modos) proposto por Kroonenberg (1983) para tabelas de tripla entrada com dados contínuos. Para cada um deles é apresentado o algoritmo correspondente, o método de seleção e uma aplicação destes modelos.

Muitas vezes se tem interesse em descrever as distintas interações que podem estar

presentes nas tabelas de múltiplas entradas. Isto é o que acontece nos dados $G \times E$, em que é medido o rendimento de uma série de genótipos em diferentes ambientes, podem ser utilizadas técnicas clássicas como a análise de variância para descrever os efeitos principais e determinar a existência da interação. No entanto, não permite chegar muito além de estabelecer se a interação é ou não significativa.

Uma das técnicas mais utilizadas para a análise deste tipo de dados é o modelo AMMI (Main effects and Multiplicative Interactions) (Gauch, 1988). Nos modelos AMMI são incorporados tantos termos multiplicativos quantos sejam necessários para explicar a variabilidade da interação de segunda ordem. É baseado na decomposição por valores singulares (SVD) da matriz de resíduos de interação do modelo linear associado. Este modelo é eficiente no caso de experimentos $G \times E$ com um único atributo, mas quando o número de atributos aumenta não há um procedimento claro a seguir, por isto, no capítulo 3, apresenta-se uma revisão de literatura sobre o modelo AMMI e a análise de procrustes generalizada, combinando estas duas metodologias para a análise dos experimentos $G \times E$ multiatributo.

Nos experimentos $G \times E$ é comum a presença de dados faltantes, produzidos por distintas razões, como fatores climáticos, doenças, presença de animais e erros na digitação ou tomada de dados. Levando em conta esta situação, no capítulo 4, é apresentada uma alternativa para realizar imputação de dados nos experimentos $G \times E$ com um atributo, como base para estudos futuros de imputação nos ensaios $G \times E$ multiatributo.

Finalmente, no capítulo 5 são considerados trabalhos futuros para continuar a linha de pesquisa.

Referências

BASFORD, K. E.; McLACHLAN, G. J. The mixture method of clustering applied to three-way data. **Journal of Classification**, New York, v.2, p.109-125, 1985.

FRUTOS, B. E. **Análisis de datos acoplados:** modelo T3-PCA. 371p. 2014. Tese (Doutorado em Estatística), Universidad de Salamanca, Salamanca. 2014.

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

KIERS, H. A. L.; MECHELEN, I. V. Three-way component analysis: Priciples and illustrative application. **Psychological Methods**, Washington, v.6, p.84-110, 2001.

KROONENBERG, P. M. **Applied Multiway Data Analysis**. New Jersey: Wiley-Interscience, 2008. 579p.

_____. **Three-mode principal component analysis:** Theory and applications. Leiden: DSWO Press, 1983. 398p.

2 ANÁLISE DE DADOS DE GIRASSOL PROVENIENTES DE UM ENSAIO GENÓTIPO×AMBIENTE MULTIATRIBUTO EM BRASIL[1]

Resumo

Nos experimentos multiambiente é comum coletar várias variáveis resposta ou atributos para determinar os genótipos com as melhores características agronômicas. Por essa razão é de importância contar com técnicas que permitam a análise de dados multiambientais multivariados. O objetivo principal é apresentar duas técnicas multivariadas: o método de mistura de máxima verossimilhança de agrupamento e a análise de componentes principais de 3 modos, para analisar conjuntamente genótipos, ambientes e atributos. Assim, podem ser feitas conclusões globais e detalhadas sobre o desempenho dos genótipos, ressaltar a utilidade do tratamento de dados de tripla entrada e fornecer uma alternativa de análise aos pesquisadores. Foram usados dados de girassol com vinte genótipos, oito ambientes e três atributos. Os procedimentos proporcionam um método analítico, relativamente fácil de aplicar e interpretar para descrever os padrões de desempenho e associações em experimentos multiambientais multivariados.

Palavras-chave: Dados de tripla entrada; Interação genótipo-ambiente; Agrupamento por mistura; Componentes principais

Abstract

In multienvironment trials it is common to measure several response variables or attributes to determine the genotypes with the best agronomic characteristics. Thus it is important to have techniques to analyse multivariate multi-environment data. The main objective is to present two multivariate techniques: the mixture maximum likelihood method of clustering and three-mode principal component analysis, to analyse jointly genotypes, environments and attributes. In this way, both global and detailed statements about the performance of the genotypes can be made, highlighting the utility of using three-way data in a direct way and providing an alternative analysis for researchers. We illustrate using sunflower data with twenty genotypes, eight environments and three attributes. The procedures provide an analytical procedure which is relatively easy to apply and interpret in order to describe the patterns of performance and associations in multivariate multi-environment trials.

Keywords: Three-way data; Genotype-by-environment interaction; Cluster via mixtures, Principal components

[1]Capítulo submetido para publicação na revista Communications in Biometry and Crop Science.

2.1 Introdução

Segundo Basford et al. (1991), a existência da interação genótipo por ambiente (G× E) significativa, complica a seleção e as estratégias de teste para os melhoristas. As interações representam diferenças na adaptação (ampla ou específica) e para tomar decisões de seleção objetivas é necessário entender a natureza dessas interações.

Como os melhoristas geralmente estão interessados em mais de um atributo que apresentam correlação, é de importância usar análises multivariadas de dados. No caso de experimentos G×E com um atributo só, têm-se metodologias que funcionam muito bem, como os modelos aditivos com interação multiplicativa - AMMI, mas para experimentos G×E multiatributo há muito pouco na literatura. Basford (1982); Basford e McLachlan (1985); Kroonenberg e Basford (1989); Basford et al. (1990, 1991); van Eeuwijk e Kroonenberg (1995, 1998) apresentam algumas técnicas de agrupamento e ordenação para a análise de dados de tripla entrada. Denis e Moro (1995) e Moro e Denis (1995) propuseram um modelo de tripla entrada para a análise multivariada da interação genótipo por ambiente, o qual tenta generalizar a situação multivariada, a regressão fatorial, agrupamento duplo (simultâneo de genótipos e ambientes) e os modelos biaditivos.

É difícil interpretar as interações complexas inerentes aos dados de tripla entrada (three-way). Se a avaliação dos genótipos é feita só usando um atributo, por exemplo rendimento, mesmo considerando-o como o atributo mais importante, muita informação contida nos dados é ignorada. Por outro lado, se são feitas análises separadas para cada atributo, há dificuldade para combinar com sucesso os resultados, além disso ignora as correlações entre os dados (Basford et al., 1990).

Neste capítulo, duas técnicas multivariadas, o método de mistura de máxima verossimilhança de agrupamento - MIXCLUS (Basford e McLachlan, 1985) e a análise de componentes principais de três modos - TMPCA, modelo de TUCKALS3 (Kroonenberg e De Leeuw, 1980; Kroonenberg, 1983), são usadas na análise de todos os genótipos, ambientes e atributos simultaneamente. São usados dados de tripla entrada de girassol, com vinte genótipos, oito ambientes e três atributos.

O objetivo principal é apresentar estas análises para mostrar que é possível manejar vários atributos em uma única análise, fazer conclusões globais e detalhadas sobre o desempenho dos genótipos, ressaltar a utilidade do tratamento de dados de tripla entrada desta forma e fornecer uma alternativa de análise aos pesquisadores.

As técnicas aqui apresentadas são usadas para analisar dados G×E multiatributo mas também servem na análise de qualquer conjunto de dados de tripla entrada.

2.2 Materiais e métodos

2.2.1 Dados Experimentais

Foram utilizados os resultados de um ano de experimentação correspondente aos Informes de avaliação de genótipos de girassol 2012/2013 da Embrapa. Foram avaliados 20 genótipos, 18 genótipos obtidos de cruzamento (híbridos) e 2 variedades em 8 ambientes das regiões sul e nordeste do Brasil. O delineamento experimental utilizado foi o aleatorizado em blocos com quatro repetições. Os experimentos dos 8 ambientes foram compostos de 1 parcela sendo, cada uma delas, constituída de quatro fileiras de seis metros de comprimento. Foram avaliados três atributos, rendimento de grãos de girassol por hectare (t/ha); teor de óleo (%) e altura da planta (cm).

2.2.2 Método de mistura de máxima verossimilhança de agrupamento - MIXCLUS

Se os genótipos podem ser agrupados de modo que dentro de um grupo tenham padrões de resposta semelhantes para cada um dos atributos nos ambientes, então, o melhorista pode examinar um conjunto de dados muito menor e portanto, integrar mais facilmente a informação inerente aos ensaios. O método de mistura de máxima verossimilhança de agrupamento pode ser aplicado para produzir um agrupamento de genótipos baseado no uso simultâneo de atributos e ambientes (Basford et al., 1990).

No uso do método de mistura de agrupamento, assume-se em primeiro lugar que há um número específico, NG, de grupos. Agrupamentos iniciais podem ser obtidos utilizando os resultados de outras técnicas de agrupamento, como k-médias ou agrupamento hierárquico, aplicadas aos dados genótipo por ambiente para um único atributo, informação *a priori* sobre os dados, ou simplesmente valores inicias aleatórios. A verossimilhança é formada sob a suposição de que os elementos são uma amostra de uma mistura em várias proporções destes grupos (π_{ng}, $ng = 1, \ldots, NG$), é por isso que é chamado de método de mistura, os detalhes estão em Basford e McLachlan (1985). A pressuposição mais comum e a utilizada aqui, é que a distribuição dos atributos de cada grupo é normal multivariada, isto é, a distribuição do vetor de atributos para o genótipo i ($i = 1, \ldots, I$) no ambiente j ($j = 1, \ldots, J$) é $f(\mathbf{x}_{ij}) = \sum_{ng=1}^{NG} \pi_{ng} f_{ng}(\underset{\sim}{x}_{ij})$, em que $f_{ng}(\underset{\sim}{x}_{ij}) \sim N(\underset{\sim}{\mu}_{ngj}, V_{ng})$. No modelo, os grupos têm diferentes vetores de médias e diferentes matrizes de correlação (Basford et al., 1990).

Conforme Basford et al. (1990), um dos objetivos da análise de mistura de máxima verossimilhança de agrupamento é estimar esses parâmetros desconhecidos no modelo (vetores de médias, matrizes de correlação e proporções de mistura). Isto é conseguido considerando a verossimilhança descrita anteriormente. A probabilidade de que cada elemento pertença a cada um dos grupos é calculada substituindo os parâmetros desconhecidos na expressão de probabilidade adequada com as suas estimativas de máxima verossimilhança, por isso é

chamado método de mistura de máxima verossimilhança. Cada elemento é atribuído ao grupo para o qual ele tem a maior probabilidade estimada.

Esta aproximação foi estendida a dados de tripla entrada por Basford e McLachlan (1985). O modelo assume que cada população tem o seu próprio vetor de médias, o qual pode ser diferente de um ambiente para outro; isto é, um grupo pode ter um bom rendimento em um ambiente, mas baixo no outro. No entanto, a estrutura de correlação entre os atributos em um grupo é a mesma em todos os ambientes. O modelo permite que as matrizes de correlação entre atributos sejam diferentes para grupos distintos. Isto é permitido também para a situação geral na qual existe interação entre genótipos e ambientes.

O método analisa os dados na forma original $x(i,j,k)$, isto é, sem centrar ou escalar. O método de mistura de máxima verossimilhança de agrupamento é descrito em detalhe em Basford e McLachlan (1985).

2.2.3 Análise de componentes principais de três modos - 3MPCA (modelo Tuckals3)

Um modelo para a análise de dados de tripla entrada foi proposto por Tucker (1966). Então, para dados classificados de acordo com três modos (genótipos, atributos e ambientes), tem-se.

$$Z_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}a_{ip}b_{jq}c_{kr} + e_{ijk}, \text{ com } i = 1,2,\ldots,I,\ j = 1,2,\ldots,J,\ k = 1,2,\ldots,K$$

em que, I, J e K representam o número de níveis do primeiro, segundo e terceiro modo respectivamente. $\mathbf{A}_{I\times P}$, $\mathbf{B}_{J\times Q}$ e $\mathbf{C}_{K\times R}$ são matrizes de marcadores associadas a cada modo, cujos elementos são a_{ip}, b_{jq} e c_{kr}. P, Q e R representam os números de componentes retidos em cada modo. $\mathbf{G}$ é o arranjo de tripla entrada e seus elementos indicam a relação entre os componentes de cada modo (Varela et al., 2009). O elemento g_{pqr} indica a magnitude da relação entre o componente p do primeiro modo, o componente q do segundo e o componente r do terceiro modo e seu valor ao quadrado indica a variação explicada por esta combinação de componentes. O arranjo $\mathbf{G}$ pode ser considerado como uma generalização da matriz diagonal de autovalores, associada à decomposição de dupla entrada (Varela et al., 2008; Araújo et al., 2011).

Tucker (1966) oferece uma solução para as matrizes $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{C}$ do modelo, porém, as soluções encontradas não são estimadores de mínimos quadrados. Para posto completo (tomando P, Q e R quantidade de componentes retidos no primeiro, segundo e terceiro modo respectivamente) é possível reproduzir o valor Z_{ijk}, mas com modelos inferiores, o ajuste gerado pode estar o suficientemente longe do verdadeiro valor de Z_{ijk}, como por exemplo, ao reter os primeiros componentes de cada modo (Varela e Torres, 2005).

Para tentar resolver esta situação, Kroonenberg e De Leeuw (1980) propuseram um algoritmo de mínimos quadrados alternantes (Tuckals3), que toma como solução inicial a de Tucker que se baseia em encontrar os estimadores para $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{C}$, tal que a soma de quadrados dos resíduos seja minimizada, equação 2.1.

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(Z_{ijk}-\widehat{Z}_{ijk}\right)^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(Z_{ijk}-\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R}a_{ip}b_{jq}c_{kr}g_{pqr}\right)^2 \qquad (2.1)$$

O objetivo básico do modelo no método é representar cada uma das entradas ou modos tão bem como seja possível em um espaço de dimensão menor formando combinações lineares (componentes) dos níveis dos modos. Em Kroonenberg (1983, 2008) podem ser encontrados detalhes teóricos e aplicações.

Seleção do número de componentes a ser retidos em cada modo

No algoritmo Tuckals3 é necessário especificar *a priori* o número de componentes que vão ser retidos, então é preciso ter um critério de seleção. Timmerman e Kiers (2000) propuseram um procedimento que indica os valores de P, Q e R, que podem ser considerados de modo que o algoritmo Tuckals3 leve a um ótimo global e não local.

O método consiste em calcular os valores de ajuste para todas as soluções possíveis, obtidas a partir do algoritmo de Tuckals3. Em cada solução é ajustado um modelo, com o objetivo de aproximar os valores de Z. Então, para cada solução, está associado um erro e um valor explicado. O ajuste coincide com a parte de Z explicada por cada solução (Varela e Torres, 2005).

As possíveis soluções devem satisfazer as seguintes condições, $P \leq QR$, $Q \leq PR$ e $R \leq PQ$, isto por que um modelo em que, por exemplo, $R > PQ$ fornece o mesmo ajuste que o modelo com $R = PQ$. Igualmente, modelos com $P > \max(I, JK)$, $Q > \max(J, IK)$ e/ou $R > \max(K, IJ)$ podem ser omitidos, porque estes modelos não ajustam melhor do que aqueles com P, Q e R igual ao $\max(I, JK)$, $\max(J, IK)$ e $\max(K, IJ)$, respectivamente. Como ilustração, se a solução $3 \times 1 \times 2$ coincide em ajuste com $2 \times 1 \times 2$, então a primeira pode ser eliminada por ter mais eixos (Timmerman e Kiers, 2000; Kiers e der Kinderen, 2003).

Após encontrar estes valores de ajuste nas soluções possíveis e para cada valor de $s = P + Q + R$, é selecionada aquela com melhor ajuste, quer dizer, o maior. Das soluções selecionadas (uma para cada valor de s), é feito o cálculo das diferenças de ajuste (DifFit) de uma solução com a anterior imediata ($dif_S = SQ_S - SQ_{S-1}$), ou seja, quanto é ganho no ajuste, ao aumentar o número de componentes total (s), em que SQ é a soma de quadrados do modelo s; dif_S é calculada para $s = 4, \ldots, S$, para $s = 3$, dif_S é igual ao ajuste do modelo com $P = Q = R = 1$, o que implica que dif_3 é comparado com o modelo com zero componentes (Timmerman e Kiers, 2000).

O próximo passo é eliminar todas aquelas soluções para as quais existe uma solução com mais componentes e uma maior diferença de ajuste associada. Neste passo busca-se obter uma similaridade com o PCA clássico, no qual ao aumentar o número de componentes, o ganho em ajuste é cada vez menor.

Para o conjunto de soluções selecionadas é calculado o quociente $b_S = \frac{dif_s}{dif_{s+1^*}}$, em que dif_{s+1^*} é o maior valor depois de dif_s. A solução ótima será aquela com quociente máximo e a DifFit associada seja maior do que o valor critico $\frac{\|Z\|^2}{(S_{\min}-3)}$, em que $S_{\min} = \min(I, JK) + \min(J, IK) + \min(K, IJ)$, sendo que modelos com valores abaixo deste limite não devem ser levados em conta.

Este método permite encontrar um equilíbrio ótimo entre o número de componentes retidos em cada modo e a variabilidade explicada pelo modelo. Timmerman e Kiers (2000) afirmam que se o número de eixos ou componentes têm sido selecionados adequadamente, raramente o algoritmo de Tuckals3 leva a um ótimo local.

Representação Biplot

Para representar graficamente três matrizes de componentes principais ou marcadores é utilizada uma generalização do Biplot clássico proposto por Gabriel (1971), veja-se também Gabriel (2002). Um biplot é uma representação simultânea de linhas e colunas de uma matriz de dados, na qual as colunas (aqui atributos) são representados por vetores e as linhas (aqui genótipos) por pontos. O valor de um atributo para um genótipo particular, pode ser estimado a partir da sua projeção sobre o vetor que representa o atributo considerado (Varela et al., 2006).

Como um biplot está construído para trabalhar com uma matriz, quando for introduzido um terceiro modo (várias matrizes), deve-se projetar sobre os componentes principais de um dos modos e realizar tantos biplots quantos componentes tenha o modo sobre o qual está se projetando. Isto é chamado de joint plot (Kroonenberg, 1983). Por exemplo, se for projetado sobre os componentes do terceiro modo, devem-se ajustar R biplots; para $R = 1$ um biplot à matriz $A * G1 * B$, para $R = 2$ um biplot $A * G2 * B$, e assim por diante até todos os componentes retidos no terceiro modo. Gr é a parte do arranjo de tripla entrada G, associada a $R = r$, isto é, ao componente r do terceiro modo (Varela e Torres, 2005).

As diferentes unidades de medida para os atributos fazem necessário ajustar suas escalas para que possam ser analisados em conjunto, pois, caso contrário, não é possível compará-los, isto é, os dados são centrados por subtração do efeito global de atributo (β_j) e do efeito ambiental (μ_k), $\widetilde{x}_{ijk} = x_{ijk} - \mu_k - \beta_j$. Os dados são escalonados dividindo pelo desvio padrão para cada atributo, calculado ao longo de todos os ambientes, $s_k = \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\widetilde{x}_{ijk}^2}$, estimadores de mínimos quadrados são usados para estimar μ_k e β_j.

2.3 Resultados

2.3.1 Método de mistura de máxima verossimilhança de agrupamento - MIXCLUS

Para selecionar o número de grupos, foram usados distintos valores iniciais para cada valor NG. O agrupamento inicial dos genótipos foi obtido após realizar a análise de agrupamentos para cada atributo (variável resposta) com o método de k-médias e agrupamento baseado em modelos. $NG = 2$ e 3 foram avaliados, no caso de $NG = 3$ foi encontrado que pode gerar autovalores muito pequenos ou negativos o que interfere no cálculo das inversas das matrizes de covariâncias dos grupos, o que indica que pode ser necessário diminuir o número de grupos. Levando em conta isto, decidiu-se trabalhar com $NG = 2$.

A Tabela 2.1 apresenta os valores estimados para o vetor de médias dos atributos nos ambientes para cada grupo com sua correspondente matriz de correlação e a respectiva composição. Os grupos estão formados por 7 e 13 genótipos respectivamente. Em geral o grupo 2 apresenta médias maiores para dois dos três atributos que o grupo 1. A correlação entre rendimento de grãos e teor de óleo é positiva em ambos os grupos (0,31 e 0,26). Por outro lado, a correlação entre rendimento de grãos e a altura da planta é positiva no primeiro grupo e negativa no segundo (0,27 e -0,10), o mesmo ocorre na correlação entre teor de óleo e altura da planta (0,24 e -0,06).

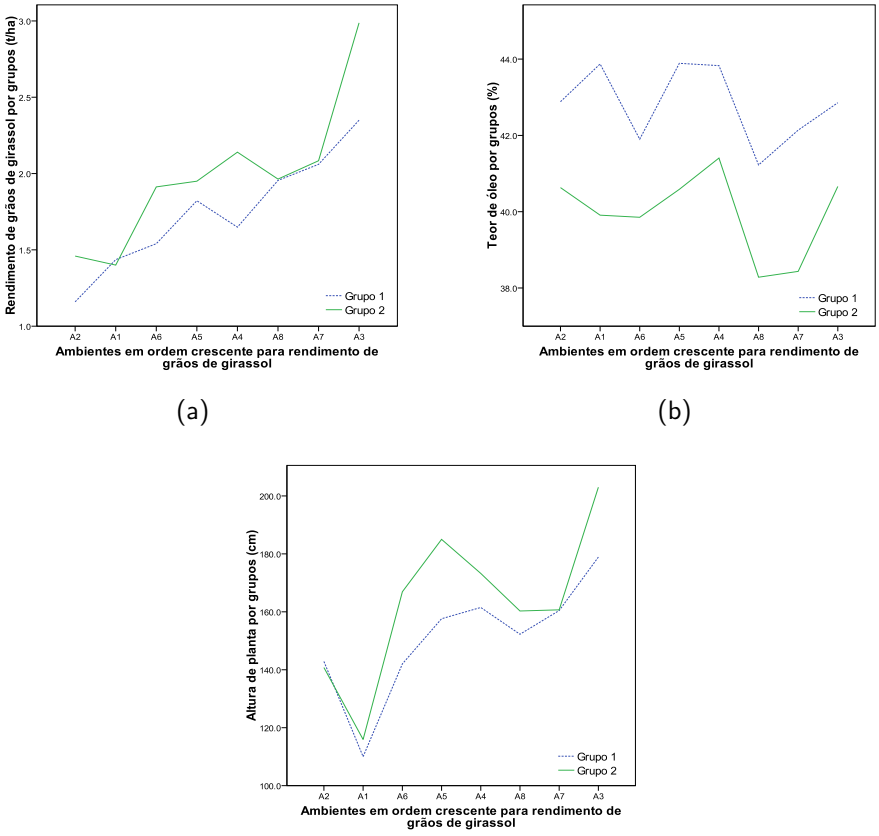
Tabela 2.1 – Médias estimadas com suas correspondentes matrizes de correlação e composição dos grupos a partir do método de mistura de agrupamento

Atributo	Grupo 1	Grupo2
Rendimento de grãos (t/ha)	1,746	1,987
Teor de óleo (%)	42,824	39,971
Altura da planta (cm)	150,698	163,225
Matriz de correlação	$\begin{bmatrix} 1 & & \\ 0,31 & 1 & \\ 0,27 & 0,24 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & & \\ 0,26 & 1 & \\ -0,10 & -0,06 & 1 \end{bmatrix}$
Composição	G3-BRSG35[1]	G1-BRSG30
	G6-BRSG38	G2-BRSG34
	G9-BRSG42	G4-BRSG36
	G10-Embrapa122[1]	G5-BRSG37
	G12-EXP26	G7-BRSG39
	G13-HELIO358	G8-BRSG40
	G15-MG341	G11-EXP24
		G14-M734
		G16-SRM767
		G17-SRM779CL
		G18-SRMCiro
		G19-SYN3950HO
		G20-V100964

[1] Variedade

A Figura 2.1, corresponde as médias dos ambientes por grupo estimadas para cada atributo. O grupo 2 tem maior desempenho no rendimento de grãos e na altura da planta que o grupo 1, o contrário ocorre no teor de óleo. Para o grupo 1, o ambiente com maior rendimento de grãos é o ambiente 3 - A3 e o menor o ambiente 2 - A2. Para o grupo 2, o maior também é o A3 mas o menor é o A1, os genótipos no grupo 2 no ambiente 3 podem ser uma boa opção para aumentar rendimento de grãos. A3 tem valores moderados para teor de óleo nos dois grupos. De outra forma, A1 tem teor de óleo alto (grupo 1) e moderado (grupo 2) e baixo rendimento de grãos em ambos os grupos. No caso da altura da planta, de novo os dois grupos apresentam comportamento similar, sendo o A3 o ambiente com maior rendimento de grãos e o A1 com menor. Não é surpresa, que o padrão para a altura da planta siga de perto o rendimento de grãos de girassol, mas com picos e depressões causadas pelo teor de óleo .

Nos dois grupos, o A4 e o A5 se caracterizam por ter valores moderados para rendimento de grãos e altura da planta e o maior valor para o teor de óleo. Da Figura 2.1, pode ser concluído que não há muita interação G×E, exceto para alguns ambientes no rendimento de grãos e altura da planta, pois o padrão de resposta por meio dos ambientes é similar.



(a) (b)

(c)

Figura 2.1 – Médias esperadas para os dois grupos, formados pelo método de mistura de agrupamento para rendimento de grãos, teor de óleo e altura da planta versus os ambientes

2.3.2 Análise de componentes principais de três modos - 3MPCA (modelo Tuckals3)

Inicialmente, o número de componentes principais para cada modo, deve ser determinado para ajustar o modelo Tuckals3 posteriormente. Neste caso $I = 20$ (número de genótipos), $J = 8$ (número de ambientes) e $K = 3$ (número de atributos), então existem $IJK = 480$ soluções possíveis, quanto ao número de componentes que devem ser retidos em cada modo, desde a solução $1 \times 1 \times 1$ ($P = 1, Q = 1, R = 1$) até a $20 \times 8 \times 3$ ($P = 20, Q = 8, R = 3$). Das 480 soluções, na primeira etapa foram selecionadas unicamente aquelas que satisfazem as condições ($P \leq QR, Q \leq PR$ e $R \leq PQ$), pois as demais são redundantes. Para as soluções selecionadas, foi calculada a porcentagem de variância explicada pelo ajuste do modelo Tuckals3, neste caso são obtidas diversas soluções para cada valor de $s = P + Q + R$. Na Tabela 2.2 é apresentada a melhor solução para cada valor de s, isto é, aquela com maior ajuste. Também aparece o valor da diferença de ajuste associado a cada solução, o quociente entre as diferenças e a porcentagem de variância explicada. Apenas são consideradas as diferenças seqüencialmente menores.

A solução ótima foi a $2 \times 1 \times 2$, pois apresenta um quociente maior entre as soluções com diferença de ajuste associada, superior ao valor critico ($1/[S_{min} - 3] = 1/28 = 0,035$). O fato de reter um componente só para os ambientes, reduz as diferenças entre os ambientes, isto também foi encontrado na análise de MIXCLUS.

Tabela 2.2 – Soluções de melhor ajuste para cada valor de $s = P + Q + R$

s=P+Q+R	Modelo (Solução PQR)	Diferença de ajuste	Quociente DifFit	% Ajuste
3	1×1×1	0,246	1,272	24,64
5	2×1×2	0,194	2,220	44,01
6	2×2×2	0,073	–	51,29
7	3×1×3	0,027	–	53,95
8	3×2×3	0,087	1,853	62,68
9	3×3×3	0,047	1,051	67,39
10	4×3×3	0,045	1,157	71,87
11	4×4×3	0,039	1,387	75,74
12	5×4×3	0,023	–	78,05
13	5×5×3	0,028	–	80,84

As Tabelas 2.3, 2.4 e 2.5, apresentam os componentes para os genótipos, ambientes e atributos para o melhor ajuste. Os dois componentes retidos para os genótipos e atributos dividem a variabilidade do modelo em 24% e 20%, respectivamente. Isto mostra que há variabilidade considerável entre os escores nos respectivos componentes para os genótipos e os atributos. Os dois componentes para genótipos e o primeiro componente para atributos são contrastes, enquanto que o segundo componente para atributos e aquele para ambientes são aproximadamente médias. Na Tabela 2.4 é observado que o primeiro componente dos

ambientes tem valores muito similares, com valores baixos para A2 e A8, isto indica que a variabilidade dos genótipos nos atributos é similar em todos os ambientes.

Tabela 2.3 – Componentes para genótipos

Genótipo	1	2
G1	-0,352	0,241
G2	-0,226	-0,049
G3	0,230	-0,133
G4	-0,310	-0,010
G5	0,025	0,259
G6	0,159	-0,093
G7	-0,156	-0,232
G8	0,001	0,040
G9	0,112	-0,350
G10	0,215	-0,574
G11	-0,171	0,139
G12	0,133	-0,041
G13	0,494	0,219
G14	-0,203	-0,175
G15	0,253	0,188
G16	0,129	0,161
G17	-0,348	-0,032
G18	-0,141	0,279
G19	0,080	0,201
G20	0,077	0,241
% Variação	24,0	20,0

Tabela 2.4 – Componentes para ambientes

Ambiente	1
A1	0,238
A2	0,229
A3	0,535
A4	0,377
A5	0,409
A6	0,456
A7	0,241
A8	0,169
% Variação	44,0

Tabela 2.5 – Componentes para atributos

Atributo	1	2
Rendimento de grãos	-0,178	0,714
Teor de óleo	0,806	0,500
Altura da planta	-0,564	0,490
% Variação	24,0	20,0

A relação entre os diferentes modos dos dados fornecidos pela análise de componentes principais de três modos são apresentados na Figura 2.2, a magnitude das relações pode ser quantificada pelos produtos internos entre os genótipos e atributos (Tabela 2.6). Produto interno alto indica bom desempenho com o atributo.

Os grupos encontrados com o método de mistura de máxima verossimilhança de agrupamento, estão claramente separados na Figura 2.2. A separação é baseada principalmente na altura da planta com os genótipos do grupo 2, apresentando um maior rendimento de grãos e uma maior altura que aqueles no grupo 1. Por outra parte, os genótipos G1 e G18 têm o maior rendimento de grãos e altura da planta (produtos internos 0,186, 0,175 e 0,260, 0,174), enquanto G10 e G9 têm os menores (produtos internos -0,350, -0,210 e -0,321, -0,187). Para teor de óleo o G13 e G15 do grupo 1 têm o melhor desempenho (produtos internos 0,423 e 0,246). É claro que o grupo 2 apresenta melhores características por meio do rendimento de grãos e altura da planta que o grupo 1.

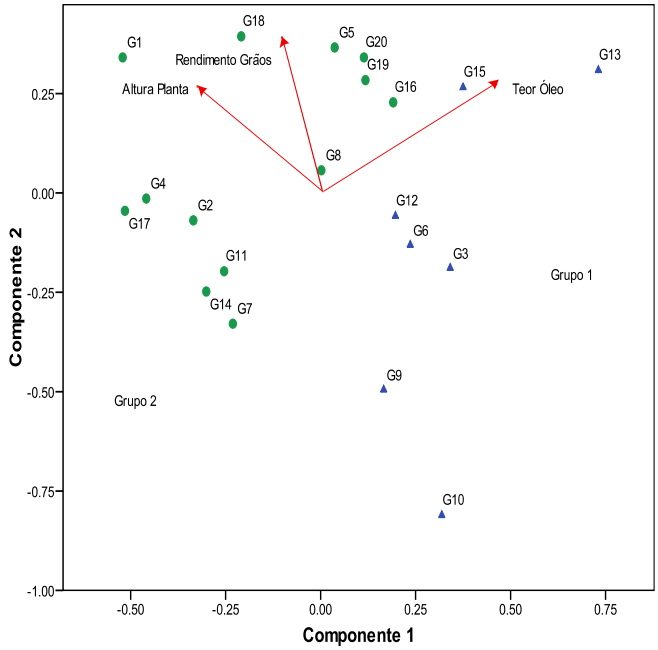


Figura 2.2 – Joint plot para genótipos e atributos para o primeiro componente de ambiente para a solução $2 \times 1 \times 2$

Tabela 2.6 – Produtos internos entre genótipos e atributos

Grupo	Genótipo	Rendimento de grãos	Teor de óleo	Altura da planta
1	G3	-0,108	0,106	-0,161
	G6	-0,075	0,073	-0,112
	G9	-0,210	-0,059	-0,187
	G10	-0,350	-0,075	-0,321
	G12	-0,043	0,075	-0,080
	G13	0,046	0,423	-0,154
	G15	0,066	0,246	-0,050
2	G1	0,186	-0,148	0,260
	G2	0,007	-0,174	0,090
	G4	0,041	-0,216	0,145
	G5	0,139	0,117	0,086
	G7	-0,105	-0,197	-0,013
	G8	0,022	0,016	0,015
	G11	-0,051	-0,171	0,029
	G14	-0,066	-0,207	0,031
	G16	0,070	0,151	0,000
	G17	0,035	-0,250	0,155
	G18	0,175	0,011	0,174
	G19	0,099	0,132	0,038
	G20	0,121	0,146	0,055

Como o modelo selecionado $2 \times 1 \times 2$, reduz as diferenças entre os ambientes e na prática é de interesse ter maiores detalhes sobre essas diferenças, os resultados correspondentes ao

modelo $2 \times 2 \times 2$ foram observados; neste caso os dois componentes para genótipos e atributos são muito similares aos encontrados com o modelo $2 \times 1 \times 2$, o mesmo acontece com o primeiro componente para ambientes. O segundo componente contém 7,9% da variação (Tabela 2.7) e é um contraste entre os ambientes A1 e A7 e os ambientes A3 e A6. Note que E3 e E7 têm potencial produtivo alto e os outros dois ambiente baixo (Figura 2.1).

Tabela 2.7 – Componentes para ambientes no modelo $2 \times 2 \times 2$

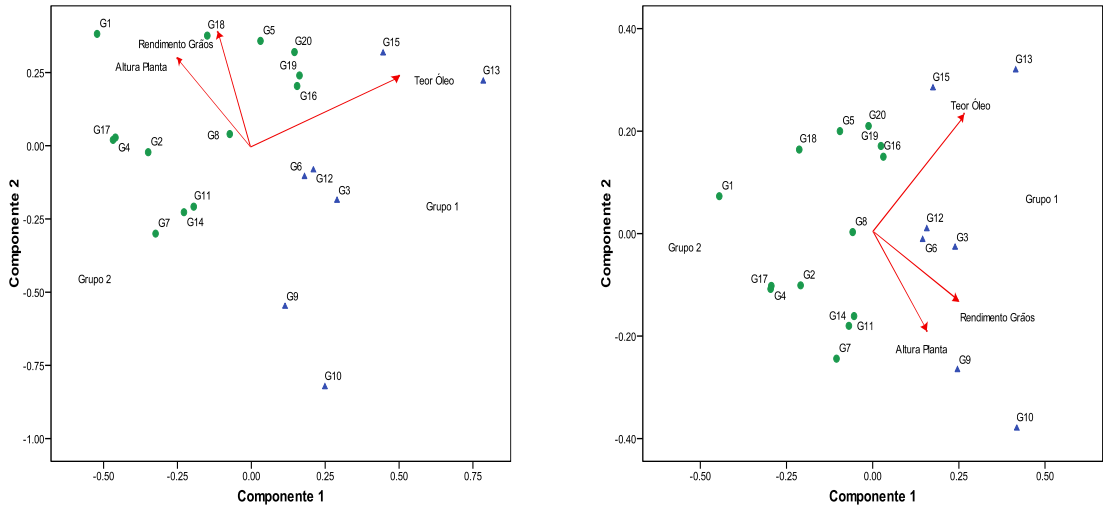
Ambiente	1	2
A1	0,244	0,540
A2	0,234	0,088
A3	0,534	-0,377
A4	0,385	-0,087
A5	0,397	0,140
A6	0,446	-0,373
A7	0,257	0,556
A8	0,174	0,289
% Variação	43,4	7.9

A Figura 2.3 apresenta o joint plot para genótipos e atributos para o primeiro e segundo componente dos ambientes. A Figura 2.3 (a) apresenta um comportamento muito similar à Figura 2.2, enquanto que a 2.3 (b) apresenta como a relação entre genótipos e atributos é diferente para os ambientes A1 e A7 em comparação com os ambientes A3 e A6. Comparado com o seu desempenho em todos os ambientes, G9 e G10 têm menores rendimentos e alturas em A3 e A6 que em A1 e A7, enquanto G1 e G18 têm rendimentos e alturas relativamente maiores em A3 e A6 do que em A1 e A7. Adicionalmente, G13 e G15 têm maior teor de óleo em A1 e A7 do que em A3 e A6, com o padrão contrário para G17, G4 e G1 (Tabela 2.8)

Tabela 2.8 – Produtos internos entre genótipos[1] e atributos para o segundo componente de ambiente no modelo $2 \times 2 \times 2$

Grupo	Genótipo	Rendimento de grãos	Teor de óleo	Altura da planta
1	G9	0,095	0,003	0,089
	G10	0,152	0,022	0,138
	G13	0,061	0,184	0,005
	G15	0,006	0,112	-0,026
2	G1	-0,119	-0,101	-0,084
	G2	-0,038	-0,079	-0,014
	G4	-0,059	-0,102	-0,027
	G7	0,006	-0,085	0,030
	G17	-0,059	-0,104	-0,026
	G18	-0,074	-0,018	-0,065

[1] Só genótipos com pelo menos um valor $\geq \mid 0,07 \mid$



(a) Primeiro componente de ambiente

(b) Segundo componente de ambiente

Figura 2.3 – Joint plot para genótipos e atributos para o primeiro e segundo componente dos ambientes para a solução $2 \times 2 \times 2$

2.4 Discussão e conclusão

A informação obtida a partir das análises conjuntas dos dados de girassol pode ser resumida da seguinte forma.

Os métodos integraram com sucesso os atributos aos dados da interação G×E. As análises ajudam na toma da decisão, estando a favor de produções e alturas moderadas e altas com teor de óleo moderado. Neste caso particular, selecionando os "melhores" genótipos do grupo 2 com alto rendimento, altura razoável e teor de óleo adequado e por outra parte, os melhores do grupo 1 com bom teor de óleo, rendimento e altura moderados, assim, os candidatos seriam G5, G18 (Grupo 1) e G13 (Grupo 2). Pode-se concluir que ambos os métodos permitiram uma interpretação simples da relação entre e dentro dos grupos dos genótipos. Isto ajuda no problema prático dos melhoristas.

Embora os grupos sejam mais fáceis de examinar que os genótipos individuais, deve-se lembrar que a seleção deve ser feita para genótipos. A maior vantagem destes métodos é que permitem trabalhar com dados de tripla entrada de maneira direta e é obtida uma visão global das respostas dos genótipos em distintos ambientes. Adicionalmente, a informação fornecida pelos métodos pode ser facilmente apresentada em figuras, o que torna mais simples a sua interpretação.

Referências

ARAÚJO, L. B.; VARELA, M.; ARAÚJO, M. F. C.; DIAS, C. T. S. Multiattribute response of Maize Genotypes tested in different coastal regions of Brazil. **International Journal of Agronomy**, Cairo, v.2011, p.1-6, 2011.

BASFORD, K. E. The use of multidimensional scaling in analysing multi-attribute genotype response across environments. **Australian Journal of Agricultural Research**, Sydney, v.33, p.473-480, 1982.

BASFORD, K. E.; McLACHLAN, G. J. The mixture method of clustering applied to three-way data. **Journal of Classification**, New York, v.2, p.109-125, 1985.

BASFORD, K. E.; KROONENBERG, P. M.; DeLACY, I. H.; LAWRENCE, P. K. Multiattribute evaluation of regional cotton variety trials. **Theoretical and Applied Genetics**, New York, v.79, p.225-234, 1990.

BASFORD, K. E.; KROONENBERG, P. M.; DeLACY, I. H. Three-way methods for multiattribute genotype $\times$ environment data: an illustrated partial survey. **Field Crops Research**, Amsterdam, v.27, p.131-157, 1991.

DENIS, J. B.; MORO. J. Multivariate generalizations for modeling two-way interaction. I. Defining and estimating models. **Biuletyn Oceany Odmian**, Poznan, v.26-27, p.43-56, 1995.

GABRIEL, K.R. The biplot graphic display of matrices with applications to principal components analysis. **Biometrika**, Cambridge, v.58, p.453-467, 1971

_____. Le biplot-outil d'exploration de données multidimensionelles. **Journal de la Societe Francaise de Statistique**, Paris, v.143, p.5-55, 2002.

KIERS, H. A. L.; der KINDEREN, A. A fast method for choosing the numbers of components in Tucker3 analysis. **British Journal of Mathematical and Statistical Psychology**, London, v.56, p.119-125, 2003.

KROONENBERG, P. M. **Three-mode principal component analysis:** Theory and applications. Leiden: DSWO Press, 1983. 398p.

_____. **Applied Multiway Data Analysis**. New Jersey: Wiley-Interscience, 2008. 579p.

KROONENBERG, P. M.; BASFORD, K. E. An investigation of multi-attribute genotype response across environments using three-mode principal component analysis. **Euphytica**, Wageningen, v.44, p.109-123, 1989.

KROONENBERG, P. M. e De LEEUW, J. Principal Component Analysis of Three-Mode Data by means of Alternating Least Squares Algorithms. **Psychometrika**, Colorado Springs, v.45, p.69-97, 1980.

MORO. J.; DENIS, J. B. Multivariate generalizations for modeling two-way interaction. II. Interpreting models and examples. **Biuletyn Oceany Odmian**, Poznan, v.26-27, p.57-72, 1995.

TIMMERMAN, M. E.; KIERS, H. A. L. Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local optima. **British Journal of Mathematical and Statistical Psychology**, London, v.53, n.1, p.1-16, 2000.

TUCKER, L. R. Some mathematical notes on three-mode factor analysis. **Psychometrika**, Colorado Springs, v.31, p.279-311 , 1966.

VAN EEUWIJK, F. A.; KROONENBERG, P. M. The simultaneous analysis of genotype by environment interaction for a number traits using three-way multiplicative modelling. **Biuletyn Oceany Odmian**, Poznan, v.26-27, p.83-96, 1995.

_____. Multiplicative models for interaction in three-way ANOVA, with applications to plant breeding. **Biometrics**, Washington, v.54, p.1315-1333, 1998.

VARELA, M.; TORRES, V. Aplicación del análisis de componentes prinicipales de tres modos en la caracterización multivariada de somaclones de king-grass. **Revista Cubana de Ciencia Agrícola**, La Habana, v.39, p.543-552, 2005.

VARELA, M.; CROSSA, J.; RANE, J.; JOSHI, A. K.; TRETHOWAN, R. Analysis of three-way interaction including multiattributes. **Australian Journal of Agricultural Research**, Sydney, v.57, p.1185-1193, 2006.

VARELA, M.; VICENTE, J. L.; GALINDO, P.; BLÁZQUEZ, A.; CASTILLO, J. G.; ESTÉVEZ, A. Una generalización de los modelos AMMI basada en el algoritmo de Tuckals3 para el análisis de componentes principales de tres modos. **Cultivos Tropicales**, La Habana, v.29, p.69-72, 2008.

VARELA, M.; CROSSA, J.; JOSHI, A. K; CORNELIUS, P. L.; MANES, Y. Generalizing the sites regression model to three-way interaction including multi-attributes. **Crop Science**, Madison, v.49, p.2043-2057 ,2009.

3 ALTERNATIVA DE ANÁLISE PARA OS MODELOS ADITIVOS COM INTERAÇÃO MULTIPLICATIVA (AMMI) MULTIATRIBUTO

Resumo

A interação genótipo-ambiente tem forte influência na seleção e recomendação de cultivares nos programas de melhoramento genético; os genótipos devem ter ótimos desempenhos em diferentes ambientes. Os modelos aditivos com interação multiplicativa - AMMI, são muito utilizados na análise da interação (G×E), pois permitem combinar componentes aditivos e multiplicativos em um mesmo modelo. Estes modelos têm demonstrado ser eficientes quando se tem apenas uma variável resposta ou atributo, mas quando o número de atributos aumenta não há um procedimento claro a seguir. O objetivo deste capítulo é propor uma metodologia de análise para os modelos AMMI multiatributo, realizando modelagens individuais dos atributos seguidas por procrustes generalizado.

Palavras-chave: Modelos AMMI; Interação genótipos × ambientes; Análise de procrustes generalizada; Atributos

Abstract

The genotype-environment interaction has an strong influence on the selection and recommendation of cultivars in the breeding programs; the genotypes should have optimum performance in different environments. The additive main effects and multiplicative interaction models - AMMI, are widely used in the analysis of the interaction (G×E), because permit combine additive and multiplicative components in the same model. These models have demonstrated be efficient with just one response variable or attribute, but when the number of attributes increases, there is not a clear procedure to follow. The aim of this chapter is to propose an analysis methodology for the multiattribute AMMI models, performing individual modeling of attributes followed by generalised procrustes.

Keywords: AMMI models; Genotypes × environments interaction; Generalised procrustes analysis; Attributes

3.1 Introdução

A análise da interação genótipo-ambiente tem sido um problema abordado pelos melhoristas durante muito tempo. Nas suas pesquisas realizam experimentos em vários locais e durante vários anos com o objetivo de selecionar variedades que sejam capazes de manter bons rendimentos em condições climáticas adversas; contribuindo a estender o ciclo médio do cultivo (Varela, 2002).

Os modelos aditivos com interação multiplicativa (AMMI) (propostos por Gauch (1988), baseado na idéia de Gollob (1968); Mandel (1969, 1971)), são muito utilizados na análise dos experimentos genótipo-ambiente (G×E), pois fornecem maior detalhe da soma de quadrados da interação e trazem vantagens na seleção de genótipos sobre metodologias tradicionais como a ANOVA. O uso do AMMI parece ser uma alternativa adequada para os programas de melhoramento, já que combina em um único modelo estatístico, componentes aditivos para os efeitos principais (genótipos e ambientes), e componentes multiplicativos para os efeitos da interação.

Aplicações do modelo AMMI têm aparecido freqüentemente durante as duas últimas décadas, e têm sido realizados vários artigos de revisão (Gauch, 2006; Yan et al., 2007; Gauch et al., 2008; Yang et al., 2009). A partir deles pode se concluir que os modelos AMMI são muito eficientes na análise dos dados no caso de uma variável resposta ou atributo.

Por outro lado, quando o número de atributos aumenta não há um procedimento claro a seguir, por isso García (2009) and García-Peña e Dias (2009) propuseram uma metodologia de análise para o caso bivariado, usando algumas técnicas multivariadas e testes paramétricos para a seleção do modelo.

O objetivo deste capítulo é propor uma alternativa para a análise dos Modelos Aditivos com Interação Multiplicativa (AMMI) com mais de duas variáveis resposta e dessa forma, fazer uma única recomendação de genótipos e ambientes para todos os atributos.

Neste capítulo, será apresentada a generalização da metodologia descrita em García (2009), usando modelagens AMMI individuais e em seguida comparadas com a análise de procrustes generalizada. Essa técnica permite fazer a comparação dos resultados obtidos para cada uma das variáveis (atributos) nas mesmas condições. São usados dados de feijão, com 13 genótipos, 9 ambientes e quatro atributos.

3.2 Material e métodos

3.2.1 Características dos dados

Os dados são relativos a experimentos com 13 genótipos de feijão que foram conduzidos em 9 ambientes distintos constituídos pelos anos agrícolas de 2000/2001, 2001/2002 e 2005/2006, nos municípios de Dourados e Aquidauana no estado de Mato Grosso do Sul, sendo que os experimentos foram instalados na época das águas (Dourados) e também na época da seca (Dourados e Aquidauana). Cada local é constituído de município e uma época de instalação. Tem-se ainda que em cada experimento foi utilizado um delineamento em blocos ao acaso, com 3 blocos em cada experimento. Os genótipos, em cada um dos ambientes, foram avaliados quanto aos seguintes caracteres; produtividade de grãos (t/ha), número médio de vagens por planta, número médio de sementes por vagem e massa de 100 sementes (gr).

3.2.2 Interação entre genótipos e ambientes (G×E).

As variações na resposta dos genótipos ou dos procedimentos agronômicos nos diferentes ambientes são conhecidas como a interação destes com o ambiente, ela pode ser devida a fatores físicos, adaptativos entre outros. Nos programas de melhoramento, a interação genótipos por ambientes (G×E) é de extrema importância, pois possibilita a seleção de genótipos, bem como, a determinação do número ideal de ambientes e genótipos a serem avaliados em cada fase da seleção (Fox et al., 1996).

Geralmente, os dados estão organizados em uma tabela de dupla entrada, com os genótipos nas linhas (g) e os ambientes nas colunas (e), as observações são representadas por Y_{ij}, em que g_i $(i = 1, 2, \ldots, g)$ e e_j $(j = 1, 2, \ldots, e)$. A análise da variância conjunta (ANOVA) permite determinar a magnitude da interação, usando a razão entre o Quadrado Médio da Interação $(QM_{G\times E})$ e o Quadrado Médio do Erro Médio (QM_{EM}). A interação significativa não é muito informativa por si só, então é necessário fazer um estudo mais detalhado sobre este componente (Ramalho et al., 1993).

Uma das metodologias muito usada atualmente é o modelo de efeitos aditivos com interação multiplicativa, AMMI; que tem como objetivo selecionar modelos que expliquem o padrão da interação.

3.2.3 Modelo de Efeitos Aditivos com Interação Multiplicativa - AMMI.

O modelo AMMI usa dois métodos na sua análise: análise de variância e a decomposição singular; no modelo se unem os termos aditivos dos efeitos principais e os termos multiplicativos para os efeitos da interação. Na primeira fase a análise de variância é aplicada à matriz

de médias $(Y_{(g\times e)})$ composta pelos efeitos principais na parte aditiva (média geral, efeitos genotípicos e ambientais), resultando em um resíduo de não aditividade, isto é, na interação (G×E), dada por $(\widehat{ge})_{ij}$. Essa interação constitui a parte multiplicativa do modelo, na segunda fase é analisada pela decomposição por valores singulares (DVS) da matriz de interações $(GE_{(g\times e)} = [(\widehat{ge}_{ij})])$ (Duarte e Vencovsky, 1999). O modelo AMMI para dois fatores (G e E) é apresentado como:

$$Y_{ij} = \mu + g_i + e_j + \sum_{k=1}^{p} \lambda_k \gamma_{ik} \alpha_{jk} + \varepsilon_{ij}; \tag{3.1}$$

em que Y_{ij} é a resposta média do i-ésimo genótipo no j-ésimo ambiente; μ a média geral; g_i o efeito do i-ésimo genótipo; e_j o efeito do j-ésimo ambiente e ε_{ij} erro experimental médio, $\varepsilon_{ij} \sim N(0, \frac{\sigma^2}{m})$ (m é o número de repetições).

O efeito de interação $(ge)_{ij}$ é modelado por $\sum_{k=1}^{n} \lambda_k \gamma_{ik} \alpha_{jk} + \rho_{ij}$, em que n é o número de componentes retidos no modelo $(n < p)$, com p o número de raízes características não nulas, $p = (1, 2, \ldots, \min\{g-1, e-1\})$; λ_k é o valor singular para o componente k, os valores λ_k estão ordenados $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$; 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 e ρ_{ij} são os ruídos presentes nos dados.

Os γ_{ik} e α_{jk} são obtidos sob as restrições $\sum_{i=1}^{g} \gamma_{ik}^2 = \sum_{j=1}^{e} \alpha_{jk}^2 = 1$ e $\sum_{i=1}^{g} \gamma_{ik}\gamma_{ik'} = \sum_{j=1}^{e} \alpha_{jk}\alpha_{jk'} = 0$ para $k \neq k'$. Assim como $\sum_{i=1}^{g} g_i = \sum_{j=1}^{e} e_j = \sum_{i=1}^{g} (ge)_{ij} = \sum_{j=1}^{e} (ge)_{ij} = 0$.

As estimativas da média geral (μ) e os efeitos principais (g_i e e_j) são obtidos no contexto simples de uma ANOVA de dupla entrada aplicada à matriz de médias $(Y_{g\times e})$. Os resíduos deste ajuste para os efeitos principais, equivalem ao termo das interações $GE_{g\times e} = [(g \times e)_{ij}]$ e os termos multiplicativos da interação são estimados por meio da decomposição singular - DVS.

Na abordagem AMMI não se busca recuperar toda a soma de quadrados de interação $SQ_{G\times E}$, mas apenas a parcela mais fortemente determinada por genótipos e ambientes (linhas e colunas da matriz GE), ou seja, o *padrão* (parte determinística ou sistemática). Assim, a interação do genótipo i com o ambiente j é descrita por: $\sum_{k=1}^{n} \lambda_k \gamma_{ik} \alpha_{jk}$, descartando-se o resíduo adicional ρ_{ij} dado por: $\sum_{k=n+1}^{p} \lambda_k \gamma_{ik} \alpha_{jk}$. Da mesma forma que em PCA esses eixos captam, sucessivamente, porções cada vez menores de variação presente na matriz GE $(\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_p^2)$. Por isso, o método AMMI é visto como um procedimento capaz de separar

padrão e ruído na análise da $SQ_{G\times E}$: $\sum_{k=1}^{n} \lambda_k \gamma_{ik} \alpha_{jk}$ e $\sum_{k=n+1}^{p} \lambda_k \gamma_{ik} \alpha_{jk}$, respectivamente (Weber et al., 1996).

3.2.4 Seleção do número de termos no modelo AMMI para descrever a interação.

Existem distintos métodos para selecionar o número de componentes a ser retidos no modelo, mas todos concordam em que o número deve ser o menor possível, para assim explicar a estrutura da interação (García-Peña e Dias, 2009). Aqui são apresentados alguns dos métodos mais populares.

Teste de Gollob.

O teste de Gollob (1968) distribui graus de liberdade às Somas de Quadrados $SQ_k = m\lambda_k^2$ com $k = 1, 2, \ldots, p$ e m o número de repetições, contando o número de parâmetros no k-ésimo termo multiplicativo. Logo, o teste F é calculado como na análise da variância para modelos lineares.

Na questão dos graus de liberdade, o método de Gollob (1968) é muito popular, este procedimento é muito fácil de aplicar, desde que o número de graus de liberdade para o k-ésimo componente da interação é simplesmente definido como $GL(IPCA_k) = g + e - 1 - 2k$, enquanto muitos outros procedimentos requerem simulações extensivas antes de serem usadas (Duarte e Vencovsky, 1999).

Teste F_R de Cornelius.

Estudos realizados por Piepho (1995) mostram que o teste F_R proposto por Cornelius, é bem mais robusto que o proposto por Gollob. A estatística do teste é definida como:

$$F_R = \frac{SQ_{G\times E} - \sum_{k=1}^{n} \lambda_k^2}{f_2 \; QM_{\text{erro médio}}}$$

em que, $f_2 = (g - 1 - n)(e - 1 - n)$ com n o número de termos multiplicativos incluídos no modelo. A estatística F_R, sob a hipótese nula de que não haja mais do que n termos determinando a interação, tem uma distribuição F aproximada com f_2 e $GL_{\text{erro médio}}$ graus de liberdade (Cornelius et al., 1996).

Um resultado significativo pelo teste sugere que pelo menos um termo multiplicativo a mais deve ser adicionado aos n já ajustados. Logo F_R pode ser visto como um teste para a significância dos $n + 1$ primeiros termos da interação. Observa-se que para $n = 0$, isto é,

quando nenhum termo multiplicativo é ajustado, o teste é equivalente ao F para a interação $G \times E$ global, na ANOVA conjunta, que é um teste exato. Nota-se também que os graus de liberdade do numerador de F_R é igual aos graus de liberdade para toda a interação menos os graus de liberdade atribuídos por Gollob (1968) aos n primeiros termos. Então, a aplicação de F_R é equivalente ao teste do resíduo AMMI para $G \times E$ (Duarte e Vencovsky, 1999).

Assim, pelo sistema de Gollob e Cornelius, a análise da variância conjunta completa (computadas a partir das médias) tem a estrutura como mostrada na Tabela 3.1.

Tabela 3.1 – Análise da variância conjunta completa computada a partir das médias usando os sistemas de Gollob e Cornelius

Fontes de variação	G.L.[1] Gollob	S.Q.[2] Gollob	G.L. Cornelius	S.Q. Cornelius
Genótipo (G)	$g-1$	SQ_G	-	-
Ambiente (E)	$e-1$	SQ_E	-	-
Interação (GE)	$(g-1)(e-1)$	SQ_{GE}	-	-
$IPCA_1$[3]	$g+e-1-(2\times 1)$	λ_1^2	$(g-1-1)(e-1-1)$	$\sum_{k=2}^{p}\lambda_k^2$
$IPCA_2$	$g+e-1-(2\times 2)$	λ_2^2	$(g-1-2)(e-1-2)$	$\sum_{k=3}^{p}\lambda_k^2$
$IPCA_3$	$g+e-1-(2\times 3)$	λ_3^2	$(g-1-3)(e-1-3)$	$\sum_{k=4}^{p}\lambda_k^2$
...	...	...	...	...
$IPCA_p$	$g+e-1-(2\times p)$	λ_p^2	-	-
Erro médio	$ge(m-1)$	$SQ_{\text{erro médio}}$	-	-
Total	$gem-1$	SQ_{TOTAL}	-	-

[1] G.L.: Graus de liberdade [2] S.Q.: Soma de quadrados

[3] IPCA$_k$: (interaction principal component analysis) modelo com k componentes, $k = 1, 2, \ldots, p$.

Representação gráfica - Biplot

A representação gráfica dos genótipos e ambientes mais utilizada para os modelos AMMI é o chamado *biplot*, Gabriel (2002). As coordenadas para os eixos singulares da interação, são obtidos a partir das matrizes $\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}$, resultantes da decomposição por valores singulares - DVS da matriz $\mathbf{GE}$. O biplot baseia-se na aproximação DVS de uma matriz, por outra de posto inferior. Para permitir a construção do gráfico o posto da matriz aproximada deverá, na realidade, ser igual a um, dois ou três, resultando num biplot em uma, duas ou três dimensões, respectivamente (Duarte e Vencovsky, 1999).

Deve ficar claro que o termo *biplot*, em alguns artigos tomado como sinônimo de AMMI (embora não seja exclusivo deste tipo de análise), não se refere a qualquer modelo estatístico particular, mas apenas a um tipo de gráfico contendo duas categorias de pontos ou marcadores. Aqui, uma delas referindo-se aos genótipos e a outra aos ambientes. Ademais, embora a maioria destes gráficos seja construída em duas dimensões, não é isto que determina o nome biplot, mas os dois tipos de marcadores, os quais podem ser representados, na prática, em uma, duas ou em até três dimensões (Gauch, 1992).

3.2.5 Análise de Procrustes Generalizada.

A análise de procrustes é um método que permite comparar duas o mais configurações de pontos, proporcionando uma medida numérica de quanto elas diferem. Quanto menor o valor da estatística maior similaridade entre as configurações.

No caso de duas configurações, sejam $\mathbf{X}$ e $\mathbf{Y}$ duas matrizes com dimensão $(n \times p)$ e $(n \times q)$ que representam as coordenadas dos n pontos em cada uma das configurações. Sem perda de generalidade, pode-se assumir que $p = q$ e que esta condição é atingida adicionando um número apropriado de colunas de zeros na menor matriz dentre as duas (Krzanowski, 2000).

A estatística de procrustes é composta por três passos, o primeiro é a translação que é um deslocamento de todos os pontos por meio de uma distância constante no mesmo sentido, quer dizer, uma translação fixa de toda a configuração; o segundo é a rotação que consiste em um deslocamento fixo de todos os pontos por meio de um ângulo constante, mantendo a distância de cada ponto ao centróide, significa, uma rotação fixa de toda a configuração; e finalmente a dilatação que é o estiramento ou encolhimento de todos os pontos por meio de uma constante em uma linha reta do ponto ao centróide da configuração, isto é, dilatação uniforme de toda a configuração (García-Peña e Dias, 2009).

O valor da estatística de procrustes no caso de duas configurações, será obtida por meio da realização dos 3 passos, na seqüência, de forma tal a tornar o valor final de M^2 tão pequeno quanto possível.

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

em que, $c = \frac{\text{traço}(\mathbf{\Sigma})}{\text{traço}(\mathbf{Y}\mathbf{Y}^\mathbf{T})}$ é o parâmetro de dilatação e $\mathbf{Q} = \mathbf{V}\mathbf{U}^\mathbf{T}$ a matriz de rotação, com $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathbf{T}$ a decomposição por valores singulares da matriz $\mathbf{X}^\mathbf{T}\mathbf{Y}$.

A idéia da análise de procrustes generalizada proposta por Gower (1975), é ajustar simultaneamente K configurações ou nuvens de pontos em um espaço geométrico por meio de translações, rotações e dilatações.

Sejam $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K$, K matrizes de dados, com dimensões $(n \times p_i)$, é possível assumir que todas têm dimensão $(n \times p)$, com $p = \max(p_1, p_2, \ldots, p_K)$, para conseguir isto, são adicionadas $(p - p_k)$ colunas de zeros a $\mathbf{X}_k$, com $k = 1, \ldots, K$. O objetivo é minimizar

$$M^2 = \sum_{k=1}^{K} \sum_{\substack{l=1 \\ k<l}}^{K} \text{traço}[(\widetilde{\mathbf{X}}_k - \widetilde{\mathbf{X}}_l)^T (\widetilde{\mathbf{X}}_k - \widetilde{\mathbf{X}}_l)]$$

em que, $\widetilde{\mathbf{X}}_k = \mathbf{s}_k \mathbf{X}_k \mathbf{T}_k + \mathbf{1}_n \mathbf{a}_k^T, \forall k \in [1, K]$; $\mathbf{s}_k$ é o vetor de dilatação; $\mathbf{T}_k$ matriz de rotação,

quando $\mathbf{T}$ é ortogonal é chamada de $\mathbf{Q}$; $\mathbf{a}_k$ é o vetor de translação e $\mathbf{1}_n$ vetor com todos os componentes iguais a 1 (Camiz e Denimal, 2011).

O critério deve ser minimizado de maneira iterativa, aqui é apresentado o critério proposto por Gower (1975) e melhorado por ten Berge (1977).

Como primeiro passo, as K translações $\mathbf{a}_k$ são encontradas; o critério é minimizado se todas as configurações resultantes têm o mesmo centróide, ou seja, $\forall k \in [1, K]$, $\widetilde{\mathbf{X}}_k^T \mathbf{1}_n = 0$. Isto é obtido simplesmente centrando as matrizes originais $\mathbf{X}_k$ com respeito ao seu correspondente centróide.

As rotações $\mathbf{T}_k$ e as dilatações $\mathbf{s}_k$ são encontradas iterativamente em duas fases, sendo o chute inicial $\mathbf{s}_k = 1, \forall k \in [1, K]$ e $\widetilde{\mathbf{X}}_k = \mathbf{X}_k$:

1. A rotação $\mathbf{T}_k$ é realizada para ajustar $\widetilde{\mathbf{X}}_k$ respeito a $G = \frac{1}{K-1} \sum_{l \neq k} \widetilde{\mathbf{X}}_l$ iterativamente para cada k até alcançar convergência. As configurações obtidas são de novo denotadas como $\widetilde{\mathbf{X}}_k, k \in [1, K]$, isto é, as $\widetilde{\mathbf{X}}_k$ são atualizadas.

2. Considerando agora, a matriz $\mathbf{B}$ com dimensão $\mathbf{K} \times \mathbf{K}$, cujos elementos são $b_{kl} =$ traço$(\widetilde{\mathbf{X}}_k^T \widetilde{\mathbf{X}}_l)$, o autovetor de $\mathbf{B}$ associado ao maior autovalor é denotado por $\mathbf{\Phi} = (\phi_k)_{k \in [1,K]}$. Assim, a dilatação que minimiza M^2 é $\mathbf{s}_k = \frac{\text{traço}(\widetilde{\mathbf{X}}_l^T \widetilde{\mathbf{X}}_l)}{\text{traço}(\widetilde{\mathbf{X}}_k^T \widetilde{\mathbf{X}}_k)} \phi_k$. De novo, as $\widetilde{\mathbf{X}}_k$ obtidas (dilatadas) são atualizadas e denotadas como $\widetilde{\mathbf{X}}_k$.

As fases 1 e 2 são então repetidas até alcançar convergência.

3.3 Resultados

Foi realizada a análise de variância conjunta para cada um dos atributos, considerando 13 genótipos e 9 ambientes, dando como resultado o ajuste de efeitos principais por ANOVA, isto corresponde à primeira etapa da análise AMMI.

Nas Tabelas 3.2 e 3.3 observa-se que tanto os genótipos como os ambientes são estatisticamente significativos. Um dos principais resultados de interesse é a soma de quadrados da interação, objeto da decomposição DVS, na segunda etapa da análise, $SQ_{G \times E} = 19,55$ que representou 29% da soma de quadrados total para a produtividade de grãos, para o número médio de vagens por planta, o número médio de sementes por vagem e a massa de 100 sementes foi $5824,72$ (43%), $18,50$ (13%) e $265,13$ (14%), respectivamente.

As estimativas das médias de genótipos e ambientes, assim como a média geral, ajustadas pelo modelo (sem interação), são apresentadas na Tabela 3.4. Nela, pode-se observar que os genótipos com maior produtividade média são G2, G8 e G1 e com menor os G12, G10 e G11, enquanto ao número médio de vagens por planta o genótipo G8 apresentou o maior número,

Tabela 3.2 – Análise da variância conjunta calculada a partir das médias para a produtividade de grãos e o número médio de vagens por planta

Fonte de		Produtividade de grãos				Número médio de vagens/planta			
variação	G.L.[1]	S.Q.[2]	Q.M.[3]	Teste F	Valor p	S.Q	Q.M.	Teste F	Valor p
Ambientes (E)	8	36,65	4,58	22,50	<2,2E-16	4974,13	621,77	10,25	3,06E-10
Genótipos (G)	12	10,54	0,88	4,31	1,9E-05	2644,50	220,38	3,63	0,0002
Resíduo (GE)	96	19,55	0,20	–	–	5824,72	60,67	–	–
Total	116	66,73	–	–	–	13443,35	–	–	–

[1] G.L.: Graus de liberdade [2] S.Q.: Soma de quadrados [3] Q.M.: Quadrado médio

Tabela 3.3 – Análise da variância conjunta calculada a partir das médias para o número médio de sementes por vagens e a Massa de 100 sementes

Fonte de		Número médio sementes/vagem				Massa de 100 sementes			
variação	G.L.[1]	S.Q.[2]	Q.M.[3]	Teste F	Valor p	S.Q	Q.M.	Teste F	Valor p
Ambientes (E)	8	108,10	13,51	70,11	<2,2E-16	709,37	88,67	32,11	<2,2E-16
Genótipos (G)	12	13,20	1,10	5,71	2,5E-07	969,68	80,81	29,26	<2,2E-16
Resíduo (GE)	96	18,50	0,19	–	–	265,13	2,76	–	–
Total	116	139,81	–	–	–	1944,18	–	–	–

[1] G.L.: Graus de liberdade [2] S.Q.: Soma de quadrados [3] Q.M.: Quadrado médio

seguido dos genótipos G3 e G7. Já os genótipos G10, G12 e G4 apresentaram os menores números de vagens em média.

Com relação ao número médio de sementes por vagem, encontra-se que os genótipos G3, G11 e G13 apresentaram os maiores valores e os G7, G4 e G9 os menores. Finalmente para a massa de 100 sementes, os genótipos com maior peso foram G4, G2 e G5; enquanto os mais leves foram G11, G3 e G12.

No caso dos ambientes, para a produtividade A9, A6 e A3 são considerados ambientes de alta produtividade, por outra parte, os ambientes com baixa produtividade são A7, A8 e A2. Para o número de vagens por planta os ambientes mais favoráveis são A6, A3 e A5, e os menos favoráveis A7, A2 e A1. Se é desejável um número alto de sementes por vagem os ambientes indicados são A9, A3 e A8 e os menos indicados A7, A6 e A5. Já os ambientes com maior desempenho em relação a massa de 100 sementes foram A8, A9 e A6 e os que apresentaram pouco peso foram A2, A1 e A5.

Destaca-se que o genótipo G2 apresenta valores altos nos atributos produtividade de grãos e massa de 100 sementes, e valores moderados para número de vagens por planta e número de sementes por vagem. Já o genótipo G8 apresenta excelente produtividade e número de vagens e para os outros dois atributos valores moderados. O ambiente A6 tem ótimo desempenho para produtividade, número de vagens e massa de 100 sementes, mas com número de sementes por vagem moderado; um outro ambiente com boas características é o A9, neste caso com valores moderados para número de vagens por planta e altos para os

outros atributos.

Tabela 3.4 – Médias dos genótipos e ambientes em relação a cada atributo

Genótipo	Prod[1]	Vag[2]	Sem[3]	Mcs[4]		Ambiente	Prod	Vag	Sem	Mcs
G1	2,11	37,45	4,81	22,87		A1	1,62	30,53	4,88	20,28
G2	2,26	36,11	4,62	27,12		A2	1,56	30,40	5,17	20,07
G3	1,95	42,33	5,18	18,46		A3	2,17	40,84	5,37	20,55
G4	1,65	29,56	4,15	27,28		A4	1,83	34,96	4,84	21,22
G5	1,91	35,97	4,40	24,80		A5	1,83	39,30	4,10	20,51
G6	1,61	35,50	4,41	23,66		A6	2,34	49,20	3,98	24,31
G7	1,64	38,69	3,95	23,35		A7	0,49	26,18	2,29	23,37
G8	2,25	43,13	4,77	22,36		A8	1,56	32,22	5,19	26,86
G9	1,60	37,53	4,32	23,51		A9	2,53	37,38	5,56	25,85
G10	1,34	27,06	4,52	22,27						
G11	1,59	36,30	5,00	17,26						
G12	1,29	27,59	4,71	19,90		Média Geral	1,77	35,67	4,60	22,56
G13	1,82	36,49	4,91	20,41						

[1] Produtividade de grãos
[2] Número médio de vagens por planta
[3] Número médio de sementes por vagem
[4] Massa de 100 sementes

A próxima etapa da análise corresponde ao ajuste da interação pela decomposição por valores singulares, aplicada a matriz $\mathbf{GE}$. Esta matriz terá posto $p = \min(12, 8) = 8$, conseqüentemente a $SQ_{G\times E}$ pode ser decomposta em até 8 componentes, a decomposição da matriz $\mathbf{GE}$ é apresentada na Tabela 3.5.

Tabela 3.5 – Proporção da $SQ_{G\times E}$ explicada por cada eixo

Eixo	Autovalor (λ_k^2)				Prop. $SQ_{G\times E}$/Eixo				Prop. Acumulada			
Singular	Prod[1]	Vag[2]	Sem[3]	Mcs[4]	Prod	Vag	Sem	Mcs	Prod	Vag	Sem	Mcs
1	9,480	2871,408	7,854	93,217	0,485	0,493	0,424	0,352	48,502	49,297	42,450	35,159
2	3,127	1090,026	6,061	61,883	0,160	0,187	0,328	0,233	64,499	68,011	75,208	58,499
3	2,523	943,279	1,785	49,052	0,129	0,162	0,096	0,185	77,407	84,205	84,855	77,000
4	2,202	495,618	1,106	32,662	0,113	0,085	0,060	0,123	88,673	92,714	90,833	89,319
5	1,120	288,671	0,838	12,116	0,057	0,050	0,045	0,046	94,403	97,670	95,362	93,888
6	0,675	104,090	0,550	7,484	0,035	0,018	0,030	0,028	97,854	99,457	98,334	96,711
7	0,359	31,358	0,207	6,456	0,018	0,005	0,011	0,024	99,691	99,995	99,451	99,146
8	0,060	0,271	0,102	2,264	0,003	0,000	0,005	0,009	100,000	100,000	100,000	100,000
Total	19,545	5824,722	18,502	265,133	1,000	1,000	1,000	1,000	–	–	–	–

[1] Produtividade de grãos
[2] Número médio de vagens por planta
[3] Número médio de sementes por vagem
[4] Massa de 100 sementes

Pela Tabela 3.5, observa-se que para cada um dos atributos considerados, o primeiro eixo singular da interação captura a maior parte da $SQ_{G\times E}$, 48,5%, 49,3%, 42,4% e 35,2% respectivamente. Enquanto que nos demais eixos há uma diminuição na porcentagem de padrão, por exemplo, o segundo eixo captura 16,0%, 18,7%, 32,8% e 23,3%, respectivamente para cada atributo.

Neste contexto, se fosse selecionado um modelo AMMI4 (quatro componentes multiplicativos) para o número de sementes por vagem, este explicaria 90,8% da soma de quadrados da interação entre genótipos e ambientes como resposta padrão e o 9,2% restante, seria considerado como ruído presente nos dados. Para selecionar adequadamente o número de componentes a ser retidos no modelo para cada atributo foi realizado o teste do resíduo AMMI, no qual é possível avaliar todos os modelos por meio dos métodos de seleção propostos por Gollob (1968) e por Cornelius et al. (1993).

Para a produtividade de grãos, tanto o critério de Gollob quanto o de Cornelius sugerem reter 7 dos 8 eixos de interação, isto é AMMI7 (valor p <0,002 e valor p=0,468). No caso do número de vagens por planta, o método de Gollob indica que 5 dos 8 eixos devem ser retidos (modelo AMMI5, valor p=0,048), enquanto pelo teste F_R de Cornelius, o modelo selecionado é o AMMI4, já que somente a partir de $IPCA_4$ o resíduo AMMI torna-se não significativo (valor p=0,574). Considerando o número de sementes por vagem, os dois métodos coincidem em indicar que só 3 eixos são necessários, quer dizer, um modelo AMMI3 (valor p=0,009 e valor p=0,284). De acordo com Gollob e Cornelius, para a massa de 100 sementes, os oito eixos de interação são requeridos, sugerindo assim, um modelo AMMI8 (valor p=0,049).

Diante da maior simplicidade representativa do modelo e das propriedades do teste F_R, sugere-se, o modelo AMMI7 para a produtividade, o AMMI4 para o número de vagens, o AMMI3 para o número de sementes e o AMMI8 para a massa de 100 sementes, como os melhores descritores do padrão de resposta diferencial dos genótipos aos ambientes. Detalhes adicionais da seleção, podem ser observados no Apêndice 5.

A última etapa da análise AMMI, consiste na representação gráfica dos genótipos e ambientes no denominado biplot, Gabriel (2002). Antes de realizar esta etapa, foi feita a análise de procrustes, que quantifica a diferença entre configurações de pontos, neste caso, comparando os marcadores resultantes após realizar as análises AMMI individuais, lembrando que quanto menor o valor da estatística, as configurações serão mais similares. M^2 foi calculada fazendo translações, rotações e dilatações simultâneas das matrizes de marcadores, para assim, avaliar a similaridade entre elas e estabelecer um único modelo para os quatro atributos considerados.

Na análise de procrustes é levado em conta cada um dos modelos escolhidos, isto é, os modelos AMMI3, AMMI4, AMMI7 e AMMI8. Os quatro modelos foram aplicados nos quatro atributos e obtidas as respectivas matrizes de marcadores para genótipos e ambientes, as quais, são o objeto de comparação nessa análise. Na Tabela 3.6, observa-se os valores da estatística M^2 para os marcadores de genótipos e ambientes segundo os modelos escolhidos pelo método de Cornelius. O menor valor da estatística foi encontrado quando os atributos foram modelados com o modelo AMMI3, indicando que as matrizes de marcadores são mais similares do que nos outros modelos. Assim o modelo AMMI3 explica 77,4%, 84,2%, 84,9% e 77,0% da soma de quadrados da interação entre genótipo e ambientes para cada atributo.

Tabela 3.6 – Análise de procrustes (M^2) para os marcadores de genótipos e ambientes

Modelo	M^2 Genótipos	M^2 Ambientes
AMMI3	55,636	0,570
AMMI4	66,257	1.535
AMMI7	78,821	2,286
AMMI8	70,012	3,108

Baseado na seleção do modelo pela análise de procrustes, é realizada a última etapa da análise AMMI, isto é, a representação gráfica de genótipos e ambientes no biplot, usando o modelo AMMI3 para todos os atributos. A partir das Figuras 3.1, 3.2, 3.3 e 3.4, são feitas, então, as devidas interpretações, procurando identificar genótipos e ambientes que menos contribuíram para a interação $G \times E$; combinações de genótipos e ambientes desejáveis em termos de adaptabilidade; relações entre os eixos de interação e características genotípicas e ambientais conhecidas.

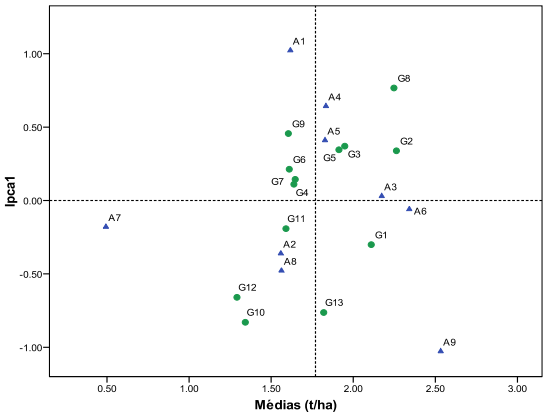
Levando em conta o anterior, são apresentados 4 biplots para cada atributo, o primeiro composto pelas médias e o $IPCA_1$ (primeiro eixo de interação), o segundo conformado pelo primeiro e segundo eixo de interação ($IPCA_1$ e $IPCA_2$), o seguinte pelo primeiro e terceiro eixo de interação ($IPCA_1$ e $IPCA_3$) e finalmente o biplot que apresenta o segundo e o terceiro eixo de interação ($IPCA_2$ e $IPCA_3$). Em resumo, o primeiro biplot, representa efeitos principais vs. interação $IPCA_1$ e os outros três, somente efeitos de interação.

Para a interpretação dos biplot, devem observar-se a magnitude e o sinal dos escores de genótipos e ambientes para os eixos de interação. Assim, escores baixos (próximos de zero) são próprios de genótipos e ambientes que contribuíram pouco ou quase nada para a interação, caracterizando-os como estáveis (Duarte e Vencovsky, 1999).

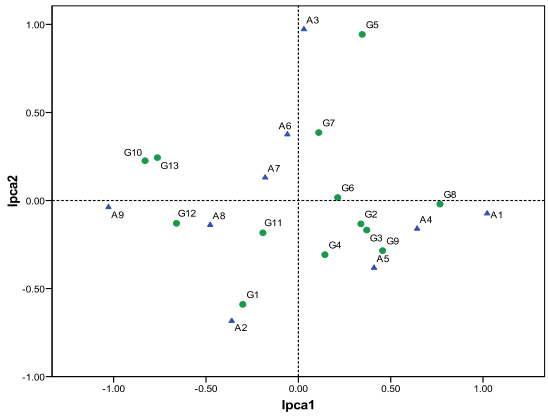
No primerio biplot, parte (a) das Figuras 3.1, 3.2, 3.3 e 3.4, a estabilidade é avaliada inspecionando $IPCA_1$; pontos situados em torno de zero em relação ao eixo $IPCA_1$, correspondem aos genótipos e ambientes mais estáveis. Quanto aos outros biplots, partes (b), (c) e (d) dessas Figuras, genótipos e ambientes estáveis, são aqueles que estão próximos à origem, quer dizer, escores praticamente nulos para os dois eixos de interação.

Na Figura 3.1, pode-se concluir que os genótipos mais estáveis e que menos contribuíram para a interação, foram os genótipos G1, G2, G3, G4, G6, G7 e G11; enquanto entre os ambientes nesse sentido foram A1, A3, A4, A6, A7, A8 e A9, mas entretanto para fins de recomendação de cultivares deseja-se uma alta performance na produtividade, que pode ser avaliada pelas médias. Assim, entre os genótipos estáveis destacam-se G1, G2 e G3 com uma produtividade média alta; G4, G6 e G7 produtividades moderadas em média e entre os ambientes A3, A4, A6 e A9.

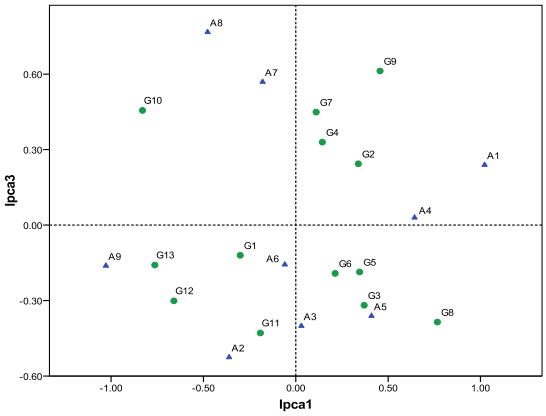
Alguns genótipos tiveram adaptações específicas a determinados ambientes, isto é, coordenadas dos genótipos estão próximas a coordenadas dos ambientes, G5 adaptou-se especificamente ao ambiente A3, G12 adaptou-se bem ao ambiente A9. Compete ao melhorista identificar tais características para assim discernir melhor os mecanismos determinantes da interação.



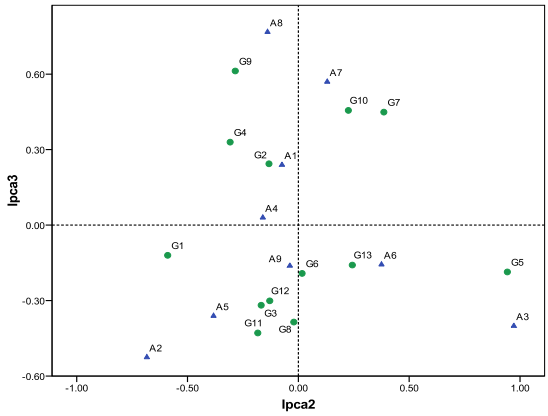
(a) A figura captura 48,5% de variabilidade



(b) A figura captura 64,5% de variabilidade



(c) A figura captura 61,4% de variabilidade

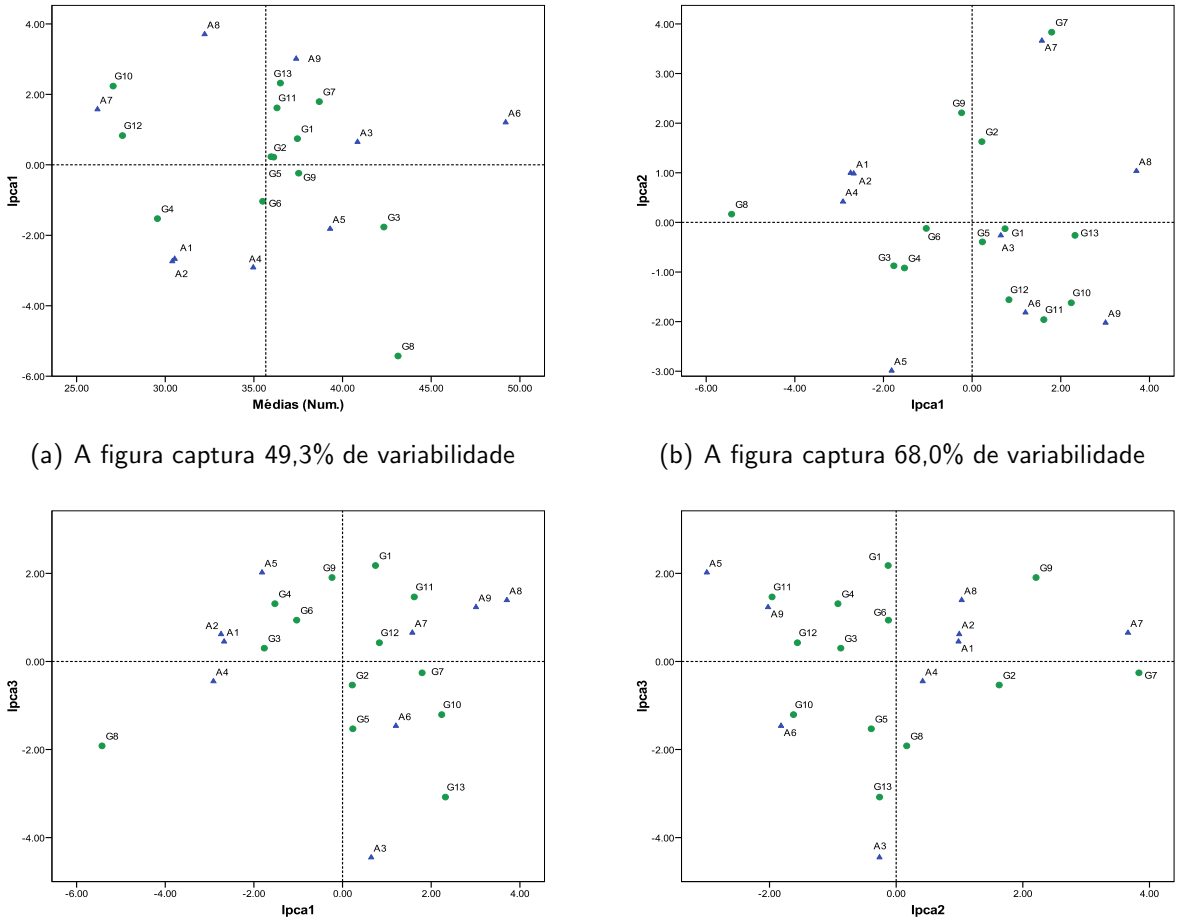


(d) A figura captura 28,9% de variabilidade

Figura 3.1 – Biplots para a produtividade de grãos (t/ha), em feijoeiro, com treze genótipos (G) e nove ambientes (A)

De acordo com a Figura 3.2, os genótipos que menos aportaram na interação foram G1, G2, G5, G6 e G9, no caso dos ambientes foram A3 e A4, todos eles estáveis. Sendo que os que apresentam melhor desempenho quanto ao número de vagens por planta são os genótipos G1 e G9 e nos ambientes destaca-se o A3.

Por outro lado, os genótipos que apresentaram adaptações específicas com os ambientes foram, o genótipo G7 com o ambiente A7, o genótipo G8 com o ambiente A4 e o genótipo G11 com o ambiente A9, este comportamento é observado nos distintos biplots do atributo número de vagens por planta.



(a) A figura captura 49,3% de variabilidade

(b) A figura captura 68,0% de variabilidade

(c) A figura captura 65,5% de variabilidade

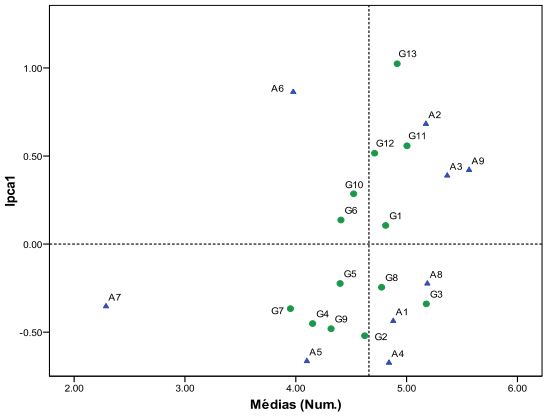
(d) A figura captura 34,9% de variabilidade

Figura 3.2 – Biplots para o número médio de vagens por planta, em feijoeiro, com treze genótipos (G) e nove ambientes (A)

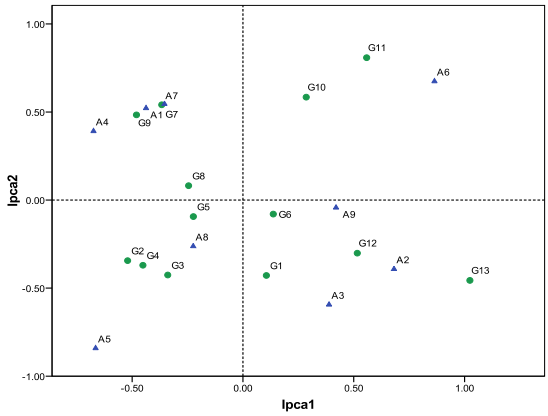
A partir da Figura 3.3, observa-se que os genótipos G1, G5, G6 e G8, se caracterizam por ser estáveis, o mesmo comportamento é apresentado pelo o ambiente A8, lembrando que os genótipos e ambientes estáveis devem ter um bom desempenho em relação ao número de sementes por vagem, destacam-se os genótipos G1 e G8 e o ambiente A8.

Ainda na Figura 3.3, é possível identificar as adaptações dos genótipos nos ambientes, por exemplo, o ponto do genótipo G3 ficou do mesmo lado que o ambiente A8, o que possibilita inferir que G3 é o melhor adaptado a A8, situações similares acontecem entre o genótipo G9 e o ambiente A1, o genótipo G11 e o ambiente A6 e o genótipo G12 e o ambiente A2.

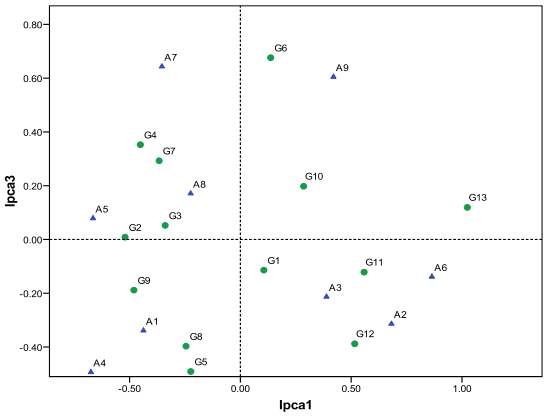
Segundo a Figura 3.4, os genótipos que contribuíram menos para a interação, foram G3, G4, G6, G9, G10, G11 e G13, além de serem estáveis, os genótipos indicam serem amplamente adaptados aos ambientes de teste. Já os ambientes detectados como estáveis para o atributo massa de 100 sementes foram A2, A4, A5, A7 e A8. Quanto ao desempenho desses genótipos e ambientes são ressaltados os genótipos G4, G6 e G9, e os ambientes A4, A7 e A8.



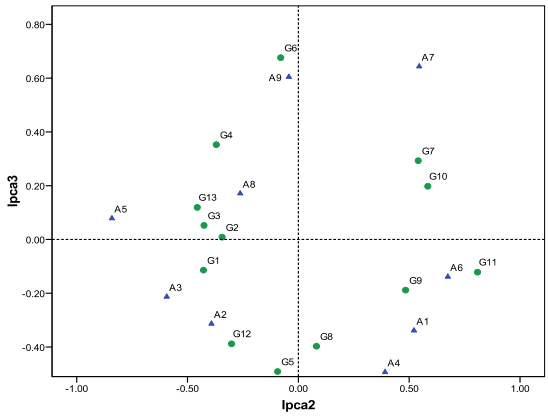
(a) A figura captura 42,4% de variabilidade



(b) A figura captura 75,2% de variabilidade



(c) A figura captura 52,1% de variabilidade



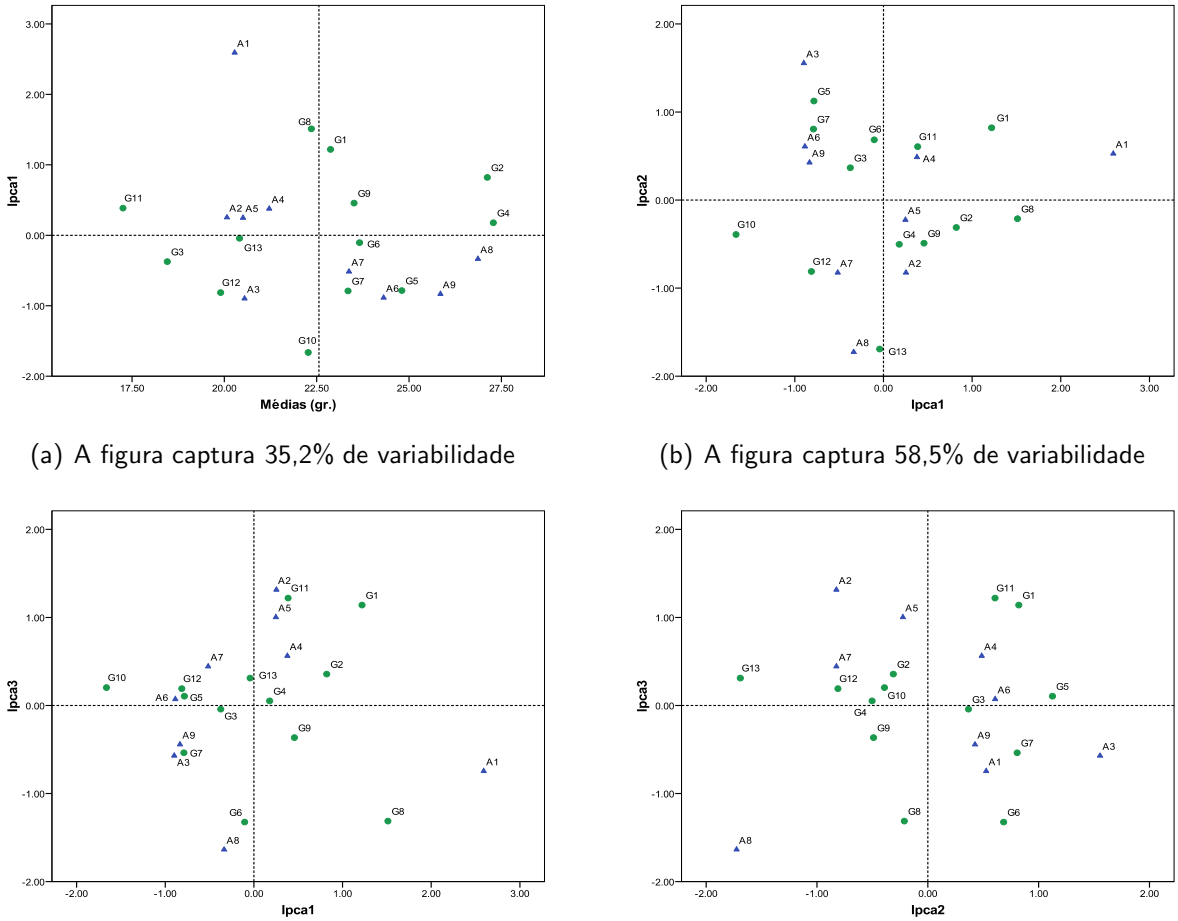
(d) A figura captura 42,4% de variabilidade

Figura 3.3 – Biplots para o número médio de sementes por vagem, em feijoeiro, com treze genótipos (G) e nove ambientes (A)

Adicionalmente, é possível identificar adaptabilidades dos genótipos aos ambientes no caso da massa de 100 sementes (Figura 3.4), nesse sentido, destaca-se a adaptabilidade do genótipo G5 com o ambiente A6, do genótipo G7 com o ambiente A9 e do genótipo G12 com o ambiente A7.

De acordo com os genótipos e ambientes selecionados como estáveis nos quatro atributos, pode ser recomendado o genótipo G1 como aquele que além de estável apresenta um bom desempenho para produtividade, número de vagens por planta e número de sementes por vagem, enquanto que para massa de 100 sementes seu desempenho é moderado. Também é ressaltado o comportamento dos genótipos G2, G3, G4, G6, G8 e G9.

Entre os ambientes, destacam-se, A3 e A4, o primeiro apresenta ótimo desempenho para produtividade e número de vagens por planta e o segundo com boa performance para produtividade e massa de 100 sementes. Outros ambientes com boas características são A6, A7, A8 e A9.



(a) A figura captura 35,2% de variabilidade

(b) A figura captura 58,5% de variabilidade

(c) A figura captura 53,7% de variabilidade

(d) A figura captura 41,8% de variabilidade

Figura 3.4 – Biplots para a massa de 100 sementes (gr.), em feijoeiro, com treze genótipos (G) e nove ambientes (A)

Com os resultados encontrados nas análises, compete ao melhorista identificar as características desejadas para assim discernir melhor os mecanismos determinantes da interação.

3.4 Discussão e conclusão

Pode-se concluir que a utilização conjunta das metodologias AMMI e procrustes torna mais eficaz a exploração da interação genótipo-ambiente, quando se tem vários atributos.

No uso da análise AMMI seguida da análise de procrustes, foi encontrado um modelo com alta similaridade entre as matrizes de marcadores nos quatro atributos considerados. Este modelo se caracteriza por ter poucos componentes, o que indica que está captando o padrão dos dados para os atributos e deixando fora o ruído.

Recomenda-se o uso das análise AMMI e procrustes conjuntamente no caso de dados multiatributo, pois apresentam bons resultados e facilitam a recomendação de genótipos e ambientes para os melhoristas.

Referências.

CAMIZ, S.; DENIMAL, J. J. Procrustes analysis and stock markets. **Case Studies in Business, Industry and Government Statistics**, Waltham, Massachusetts, v.4, n.2, p.93-100, 2011.

CORNELIUS, P. L.; CROSSA J.; SEYEDSADR M. S. Tests and estimators of multiplicative models for variety trials. In: ANNUAL KANSAS STATE UNIVERSITY CONFERENCE ON APPLIED STATISTICS IN AGRICULTURE, 1993, Manhattan. **Proceedings . . .** Manhattan: Statistics department, Kansas State University, 1993. p.156-166.

_____. Statistical tests and estimators of multiplicative models for genotype-by-environment interaction. In: KANG, M. S.; GAUCH, H. G. **Genotype-by-environment interaction**. Boca Raton: CRC Press, 1996. cap. 8. p.199-234.

DUARTE, J.B.; VENCOVSKY, R. **Interação genótipo $\times$ ambiente:** uma introdução à análise "AMMI". Ribeirão Preto: Sociedade Brasileira de Genética, 1999. p.60 (Série Monografias, 9).

FOX, P. N.; CROSSA, J.; ROMAGOSA, I. Multi-environment testing and genotype $\times$ environment interaction. In: KEMPTON, R.A.; FOX, P.N. (Ed.). **Statistical methods for plant variety evaluation**. New York: Chapman and Hall, 1996. cap. 8, p. 117-138.

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

GARCÍA, P. M. **Análise dos modelos AMMI bivariados**. 2009. 77p. Disssertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba. 2009.

GARCIA-PEÑA, M.; DIAS, C. T. S. Analysis of bivariate additive models with multiplicative interaction (AMMI). **Revista Brasileira de Biometria**, São Paulo, v.27, n.4, p.586-602, 2009.

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

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

_____. Statistical analysis of yield trials by AMMI and GGE. **Crop Science**, Madison, v.46, p.1488-1500, 2006.

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

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

GOWER, J. C. Generalized Procrustes Analysis. **Psychometrika**, Colorado Springs, v.40, p.33-1, 1975.

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

MANDEL, J. The partitioning of interactions in analysis of variance. **Journal of Research of the National Bureau of Standards, Series B**, Washington, v.73, p.309-328, 1969.

_____. A new analysis of variance model for non-adittive data. **Technometrics**, Alexandria, v.13, n.1, p.1-18, 1971.

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

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

TEN BERGE, J. M. F. Orthogonal Procrustes rotation for two or more matrices. **Psychometrika**, Colorado Springs, v.42, p.267-276, 1977.

VARELA, N. M. **Los métodos biplot como herramienta de análisis de interacción de orden superior en un modelo lineal/bilineal** 2002. 168p. Tese (Doutorado em Estatística), Universidad de Salamanca, Salamanca. 2002.

WEBER, W. E., WRICKE, G.; WESTERMANN, T. Selection of genotypes and prediction of performance by analysing genotype-by-environment interactions, 1996. In: KANG, M. S.; GAUCH, H. G. **Genotype-by-environment interaction**. Boca Raton: CRC Press, 1996. cap. 13. p.353-371.

YAN, W.; KANG, M. S.; MA, B.; WOODS, S.; CORNELIUS, P. L. GGE biplot vs. AMMI analysis of genotype-by-environment data. **Crop Science**, Madison, v.47, p.643-655, 2007.

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

4 PROCEDIMENTOS DE IMPUTAÇÃO MÚLTIPLA UTILIZANDO O ALGORITMO GABRIELEIGEN

Resumo

GabrielEigen é um sistema de imputação simples determinística sem pressuposições estruturais nem distribucionais, que utiliza uma mistura de regressão com aproximação de posto inferior de uma matriz por meio da decomposição por valores singulares. O objetivo desse capítulo é fornecer alternativas de imputação múltipla (IM) baseadas nesse sistema, adicionando quantidades aleatórias e gerando intervalos de confiança aproximados com diferente amplitude para as imputações usando validação cruzada (VC). As metodologias foram avaliadas por meio de um estudo de simulação em matrizes de dados reais considerando retiradas aleatórias em várias porcentagens e também foi exemplificada a situação quando as observações ausentes tinham um padrão sistemático. A qualidade das imputações foi avaliada por uma medida de acurácia geral (T_{acc}) que combina a variância entre imputações (V_b) e o viés quadrático médio delas em relação aos valores retirados (B). Concluiu-se que o melhor desempenho é obtido com intervalos cuja amplitude seja igual ao erro de imputação associado a GabrielEigen.

Palavras-chave: Imputação; Valores ausentes; Decomposição por valores singulares; Validação cruzada; Desbalanceamento

Abstract

GabrielEigen is a simple deterministic imputation system without structural or distributional assumptions, which uses a mixture of regression with lower-rank approximation of a matrix based on its singular value decomposition. The aim of this chapter is to provide multiple imputation alternatives (MI) based on this system by adding random quantities and generating approximate confidence intervals with different amplitude to the imputations using cross-validation (CV). These methods were assessed by a simulation study, using real data matrices in which values are deleted randomly at different rates and also was exemplified the situation when the missing observations had a systematic pattern. The quality of the imputations was evaluated by a measure of overall accuracy (T_{acc}) that combines the variance between imputations (V_b) and their mean square deviations in relation to the deleted values (B). It was concluded that the best performance is obtained at intervals whose amplitude is equal to the imputation error associated with GabrielEigen.

Keywords: Imputation; Missing values; Singular value decomposition; Cross-validation; Unbalanced

4.1 Introdução

Imputação é uma técnica na qual os elementos ausentes de uma matriz de dados são substituídos por valores plausíveis, dessa forma, é possível fazer análises válidas sobre a matriz de dados completada (observados + imputados). Recentemente, Arciniegas-Alarcón et al. (2010) propuseram um algoritmo de imputação livre de pressuposições distribucionais e estruturais que mistura regressão com aproximação de posto inferior de uma matriz.

O algoritmo foi chamado GabrielEigen e por ser determinístico, tem como vantagem sobre os métodos de imputação estocástica (imputação múltipla paramétrica) que os valores imputados são determinados de forma única e se o processo for repetido de qualquer forma no mesmo conjunto de dados, sempre fornecerá os mesmos resultados. Essa característica não é, necessariamente, verdadeira para os métodos de imputação estocástica (Bello, 1993; Arciniegas-Alarcón et al., 2013).

Como toda metodologia estatística, GabrielEigen tem limitações e uma delas é que fornece imputação simples, portanto, não leva em conta a incerteza produzida pelas imputações. Desse modo, se parâmetros de um modelo forem estimados a partir dos dados imputados, os erros-padrão serão subestimados, ou seja, os intervalos de confiança e os testes perderão a validade, mesmo que o modelo de imputação esteja correto (Josse et al, 2011; Josse e Husson, 2012a; Arciniegas-Alarcón et al., 2014a; Arciniegas-Alarcón, 2015).

Para resolver esse problema pode ser utilizada a imputação múltipla IM (Rubin, 1978, 1987). Descrições mais recentes da técnica são encontradas em Graham (2012), van Buuren (2012) e Rässler et al. (2013). Segundo van Ginkel e Kroonenberg (2014), a técnica consiste em quatro passos: (i) Os valores ausentes são estimados M vezes de acordo com um modelo estatístico especifico; (ii) Essas estimativas são substituídas no conjunto de dados, resultando em M versões completas plausíveis do conjunto de dados incompleto; (iii) Procedimentos estatísticos padrão são aplicados nesses M conjuntos de dados; (iv) Os resultados são combinados para obter as estimativas dos parâmetros e sua variabilidade.

A IM resolve de uma forma simples o problema de desbalanceamento que pode afetar os experimentos com interação genótipo por ambiente (G×E) e que causa dificuldades na aplicação dos modelos de efeitos aditivos com interação multiplicativa - AMMI e dos modelos de efeitos principais genotípicos com efeitos de interação genótipo por ambiente - GGE (Gauch, 2013; Paderewski, 2013; Forkman, 2015; Yan, 2015). Portanto, o objetivo desse capítulo é propor alternativas para o primeiro passo da IM utilizando GabrielEigen e avaliá-las por meio de um estudo de simulação baseado em matrizes reais provenientes de experimentos (G×E).

4.2 Material e métodos

4.2.1 Algoritmo de imputação GabrielEigen

O método, inicialmente, substitui as caselas vazias por valores arbitrários e subseqüentemente as imputações são refinadas por meio de um esquema iterativo que define uma partição diferente da matriz para cada valor ausente e utiliza a regressão linear das colunas (ou linhas) para obter a nova imputação. Nessa regressão, a matriz de delineamento é aproximada por uma matriz de posto inferior usando a decomposição por valores singulares (DVS) (Arciniegas-Alarcón et al., 2014b). O algoritmo é apresentado a seguir mais formalmente.

Considere uma matriz $\mathbf{X}$ de dimensão $(n \times p)$ com elementos x_{ij} $(i = 1, \ldots, n; j = 1, \ldots, p)$, em que alguns deles são ausentes. Note-se que esse processo requer $n \geq p$ e se não for o caso, então a matriz $\mathbf{X}$ deve ser primeiro transposta.

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

Passo 2. As colunas são padronizadas, subtraindo de cada elemento m_j e dividindo o resultado por s_j (em que m_j e s_j representam a média e desvio padrão da j-ésima coluna).

Passo 3. Usando a matriz padronizada, define-se a seguinte partição $\mathbf{X} = \begin{bmatrix} x_{ij} & \mathbf{x}_{1\cdot}^T \\ \mathbf{x}_{\cdot 1} & \mathbf{X}_{11} \end{bmatrix}$, em que o valor ausente na posição (i, j) está sempre na posição $(1, 1)$ da partição definida. Para cada valor faltante x_{ij}, os componentes da partição considerada serão diferentes e essa partição é obtida por meio de operações elementares nas linhas e colunas da matriz $\mathbf{X}$. Substitui-se a submatriz $\mathbf{X}_{11}$ pela aproximação de posto m utilizando a DVS: $\mathbf{X}_{11} = \sum_{k=1}^{m} \mathbf{u}_{(k)} d_k \mathbf{v}_{(k)}^T = \mathbf{U}\mathbf{D}\mathbf{V}^T$, em que $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$, $\mathbf{D} = diag(d_1, \ldots, d_m)$ e $m \leq \min\{n-1, p-1\}$. Então, usando a regressão $\mathbf{U}\mathbf{D}^{-1}\mathbf{V}^T\mathbf{x}_{\cdot 1}$ (ou $\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T\mathbf{x}_{\cdot 1}$) da primeira linha (ou da primeira coluna) omitindo a primeira coluna (ou linha) a imputação de x_{ij} está dada por $\widehat{x}_{ij}^{(m)} = \mathbf{x}_{1\cdot}^T \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T\mathbf{x}_{\cdot 1}$.

Passo 4. O processo de imputação depende do valor de m e sugere-se que m seja escolhido como o menor valor para o qual $\dfrac{\sum_{k=1}^{m} d_k^2}{\sum_{k=1}^{\min\{n-1,p-1\}} d_k^2} \approx 0,75$ (Arciniegas-Alarcón et al., 2010; Caliński et al., 1999).

Passo 5. Finalmente, os valores imputados devem ser retornados à sua escala original, $x_{ij} =$

$m_j + s_j \widehat{x}_{ij}^{(m)}$, substituindo-os na matriz $\mathbf{X}$. Então, o processo é iterado (voltando ao Passo 2) até alcançar estabilidade nas imputações.

4.2.2 Imputação múltipla - IM usando GabrielEigen

A literatura estatística oferece uma opção sem complexidade para produzir IM. Em análise multivariada se adverte que seja qual for o método de imputação utilizado, existe o perigo que as variâncias e as covariâncias sejam subestimadas a partir de uma matriz completada, uma vez que os valores imputados não permitem variação amostral natural. Uma forma de contornar esse problema é somar pequenos valores aleatórios a cada imputação (Krzanowski e Marriott, 1994). Utilizando essa ideia, é possível gerar IM a partir de um método de imputação simples somando M valores aleatórios a cada imputação.

Recentemente, na mesma linha, Srivastava e Dolatabadi (2009) propuseram fazer IM utilizando os resíduos simples de um modelo de regressão linear clássico, assumindo a matriz de delineamento completa e a variável independente incompleta. O processo consiste em ajustar um modelo de regressão aos dados observados e calcular os resíduos. Posteriormente, são obtidas M amostras aleatórias com reposição a partir dos resíduos com tamanho igual ao número de dados ausentes. O produto da matriz de delineamento dos dados faltantes pelo vetor de parâmetros da equação de regressão ajustada produz um vetor com as imputações. Por último, para produzir IM, ao vetor de imputações é somada, independentemente, cada uma das M amostras de resíduos simples. Uma discussão completa de IM com modelos lineares pode ser encontrada em Di Ciaccio (2011) e van Buuren (2012).

Levando em conta o que se expôs anteriormente, a primeira proposta apresentada aqui consiste em um procedimento com dois estágios aplicados na matriz $\mathbf{X}$ $(n \times p)$ com elementos x_{ij} $(i = 1, \ldots, n; j = 1, \ldots, p)$ que contém algumas observações ausentes. No primeiro estágio, é aplicado o algoritmo GabrielEigen para obter uma matriz $\mathbf{X}_G$ $(n \times p)$ completada (observados + imputados). No segundo estágio, são somados valores aleatórios às imputações, ou seja, $\mathbf{X}_G + (\mathbf{W} \circ \mathbf{E}_t)$, em que $\mathbf{W}$ $(n \times p)$ é uma matriz indicadora de dados faltantes, com zeros e uns, zero na posição (i, j) se a entrada corresponde a um dado observado em $\mathbf{X}$ e um no caso contrário. O símbolo "$\circ$" é o produto de Hadamard e $\mathbf{E}_t$ $(n \times p)$ é uma matriz de valores aleatórios com $t = 1, \ldots, M$.

As opções consideradas para construir a matriz $\mathbf{E}_t$ são as seguintes:

i) Gnorm: $\mathbf{E}_t$ está composta por valores aleatórios provenientes de uma distribuição $N(0, \widehat{\sigma}_j^2)$, em que $\widehat{\sigma}_j^2$ é a variância estimada da coluna j de $\mathbf{X}_G$. Essa forma de construir $\mathbf{E}_t$ foi inspirada pelo trabalho de Krzanowski (1988) que utilizou como imputações iniciais dentro de um esquema iterativo a média da coluna j mais uma quantidade aleatória com média

zero e variância igual à variância estimada a partir unicamente dos dados observados na coluna j.

ii) Gadd: $\mathbf{E}_t$ está composta por valores escolhidos aleatoriamente com reposição do conjunto de resíduos obtidos depois de ajustar um modelo aditivo $(x_{ij} = \mu + a_i + b_j + e_{ij})$ sobre a matriz $\mathbf{X}_G$, em que μ, a_i, b_j e e_{ij} são respectivamente, a média geral, os efeitos principais das linhas e colunas e o termo de erro associado à i-ésima linha e à j-ésima coluna. Essa forma de construir $\mathbf{E}_t$ foi inspirada pelos trabalhos de Denis e Baril (1992) e Arciniegas-Alarcón et al. (2014b) que mostraram o desempenho de um modelo aditivo para fazer imputação de dados.

iii) GLR: $\mathbf{E}_t$ está composta por valores escolhidos aleatoriamente com reposição do conjunto de resíduos obtidos de $\mathbf{X}_G - \mathbf{X}^{(m)}$, em que $\mathbf{X}^{(m)}$ corresponde a uma matriz de posto inferior, calculada por meio da DVS de $\mathbf{X}_G$ considerando apenas m componentes. Essa forma de construir $\mathbf{E}_t$ foi inspirada pelo trabalho de Arciniegas-Alarcón et al. (2014a) que generalizou para IM, o método de imputação simples baseado na DVS para análise biplot proposto por Yan (2013).

A segunda proposta consiste em gerar intervalos de confiança aproximados para cada valor imputado, calculando por meio de validação cruzada o erro de imputação associado ao algoritmo GabrielEigen (Piepho, 1995; Arciniegas-Alarcón et al., 2011, 2013). Uma vez estabelecidos os intervalos de confiança, são escolhidos aleatoriamente M valores dentro deles para produzir IM. A metodologia é apresentada a seguir.

Considere novamente a matriz incompleta $\mathbf{X}$ $(n \times p)$. Dos elementos observados na matriz, é deletado um por vez, imputando o dado deletado com GabrielEigen e guardando a diferença entre a estimativa e o dado atual para a casela sob consideração. Isto é feito para todos os elementos presentes e depois é calculada a média das diferenças ao quadrado. Denote essa quantidade por D, que contém dois componentes de variação: Um devido à inexatidão preditiva da imputação e o outro devido ao erro amostral dos dados presentes. Por essa razão D deve ser corrigida subtraindo uma estimativa do erro da média (s^2). A raiz quadrada de $(D - s^2)$ pode ser tomada como o erro de imputação (I_e) associado a GabrielEigen.

Assim, sendo $\widehat{x}_{ij}$ a imputação da posição (i, j) com GabrielEigen, então, um intervalo de imputação aproximado está dado por $\widehat{x}_{ij} \pm z_{1-\alpha} I_e$, em que $z_{1-\alpha}$ é o percentil apropriado da distribuição normal para um nível de confiança de $(1 - \alpha)\%$. Para produzir IM são escolhidos aleatoriamente M valores dentro do intervalo.

Aqui será considerado o intervalo clássico de 95%, com $z_{1-\alpha} = 1,96$ e que tem como amplitude aproximadamente $4I_e$. Decidiu-se também pesquisar o efeito que pode ter a diminuição da amplitude do intervalo de imputação, considerando intervalos de amplitude I_e e $2I_e$ ou de maneira equivalente com $z_{1-\alpha} = 0,5$ e $z_{1-\alpha} = 1$ que representam intervalos de 38% e

68%, respectivamente. A diminuição da amplitude permite a diminuição da variabilidade nas imputações, mas, pode aumentar o risco de baixa qualidade nas imputações múltiplas, pois com um intervalo pequeno as chances de conter o verdadeiro valor são baixas. A metodologia será chamada GCV1, GCV2 e GCV4 (GabrielEigen Cross-Validation) dependendo da amplitude considerada nos intervalos para produzir IM (I_e, $2I_e$ e $4I_e$).

A IM com GabrielEigen requer a especificação do número de componentes (m) retidos na DVS e do número de versões completadas (M) da matriz $\mathbf{X}$. Para estabelecer o m foi utilizada validação cruzada no lugar do critério descrito no Passo 3 do algoritmo original. Nesse caso, aplicou-se o processo explicado em García-Peña et al. (2014) que usa a função `cv.SVDImpute` do pacote `imputation` do software livre R (Wong, 2013; R Core Team, 2015).

Por outro lado, tipicamente um pequeno número de imputações $(3 \leq M \leq 10)$ é necessário para obter um bom desempenho da IM (Ounpraseuth et al., 2012) e se utilizará $M = 5$, uma vez que esse número permite atingir alta eficiência estatística em muitas aplicações práticas (van Buuren, 2012).

Finalmente, as propostas Gnorm, Gadd, GLR, GCV1, GCV2 e GCV4 são por construção, computacionalmente, menos intensivas que a proposta de IM feita por Arciniegas-Alarcón et al. (2014c) para GabrielEigen e que essencialmente consiste na inserção de pesos multiplicativos na equação de imputação. O custo computacional foi a principal razão para não considerá-la nesse trabalho, pois a seleção dos pesos está baseada em dupla validação cruzada, característica que não a faz eficiente para ser testada em estudos de simulação como esse.

4.2.3 Estudo de simulação

Para fazer o estudo de simulação mais realista, seguiu-se o protocolo proposto por Yan (2013) para avaliar, numericamente, novos métodos de imputação em matrizes (G×E). Os passos são os seguintes:

i) Escolher conjuntos reais de dados balanceados. Também é válido extrair subconjuntos balanceados a partir de experimentos incompletos (G×E).

ii) Em cada conjunto retirar valores, aleatoriamente, considerando diferentes porcentagens.

iii) Repetir o processo para cada porcentagem considerada. Por exemplo, 1000 vezes.

iv) Calcular, em cada repetição do processo, uma estatística para comparar as imputações com os valores reais retirados.

Três conjuntos de dados foram utilizados no estudo de simulação. O primeiro, é uma matriz de dimensão (20×7), com 20 progênies de *Eucalyptus grandis* avaliadas em sete locais das

regiões Sul e Sudeste do Brasil; a variável estudada foi altura média da árvore (m) (Lavoranti et al., 2007; Wright, 2012). O segundo, refere-se a uma tabela de dupla entrada (6×18), correspondente à avaliação de 6 genótipos de cevada em 18 ambientes em Alberta no Canadá. A variável de estudo foi rendimento (Mg há$^{-1}$) (Yang, 2007). Por último, o terceiro conjunto de dados está composto por uma matriz de dimensão (36×6), para 36 genótipos de trigo avaliados em 6 ambientes de diferente estresse hídrico na estação experimental da Universidade de Putra na Malásia; a variável de estudo foi rendimento médio dos grãos de planta (gr) (Rad et al., 2013).

A seleção dos conjuntos de dados foi baseada em um estudo prévio que determinou o número de componentes multiplicativos necessários para explicar a interação (G×E) por meio de um modelo AMMI (Gauch, 1992, 2013). Em cada conjunto foi aplicado o método de validação cruzada generalizada proposto por Josse e Husson (2012b) e disponível no pacote `FactoMineR` do software livre R (Husson et al., 2014; R Core Team, 2015). A Tabela 4.1 apresenta o erro quadrático médio da predição (EQMP) para escolher os componentes multiplicativos do modelo. EQMP é a soma das diferenças ao quadrado entre o valor verdadeiro e a sua predição, dividida pelo total de observações. O melhor modelo é aquele com menor EQMP, portanto, para os dados de eucalipto o melhor modelo é um AMMI1, para os dados de cevada é o AMMI2 e para os dados de trigo é um modelo AMMI3.

Tabela 4.1 – Valores do Erro Quadrático Médio da Predição (EQMP) usando validação cruzada generalizada na escolha do modelo AMMI para explicar a interação nas matrizes de dados originais (completos)

	EQMP		
Modelo	Eucalipto	Cevada	Trigo
AMMI1	0,5744	0,0502	5,08E-01
AMMI2	0,5834	0,0463	1,21E-01
AMMI3	0,6964	0,0584	6,85E-06
AMMI4	0,8123	0,0853	9,03E-06
AMMI5	1,1937	0,1565	1,50E-05
AMMI6	1,8987		

Os três conjuntos de dados têm diversos tamanhos e diferentes estruturas de interação, sendo amplamente representativos de experimentos (G×E). Por tal razão, as conclusões que forem obtidas a partir deles poderiam ser também relevantes para outras matrizes de dados multiambientais.

Em cada conjunto de dados foram retirados, aleatoriamente, 10%, 20% e 35% das observações, e o processo foi repetido 1000 vezes, obtendo no total 9000 conjuntos de dados incompletos sobre os quais foram aplicadas as seis metodologias propostas de IM por meio de programas computacionais implementados no software livre R (R Core Team, 2015).

As porcentagens escolhidas e o mecanismo de retirada dos dados estão plenamente justificados na literatura. Na prática em (G×E), o número de dados ausentes geralmente é menor que 40% (Yan, 2013) e por outro lado, com menos de 10% não teria muito ganho utilizar IM porque a imputação simples pode fornecer resultados razoavelmente bons (Schafer, 1999). Entretanto, o mecanismo aleatório da retirada de valores, representa situações comuns nos experimentos agrícolas, por exemplo, as plantas podem ser destruídas por animais, inundações ou durante a colheita e as medidas de rendimento podem ser inadequadamente inseridas nas bases de dados (Rodrigues et al., 2011). Uma discussão dos diferentes mecanismos que podem ser simulados em (G×E) encontra-se em Paderewski e Rodrigues (2014) e Arciniegas-Alarcón et al. (2014c).

No protocolo de Yan, a estatística escolhida para comparar as imputações com os valores retirados foi o erro de predição (P_e), que está definido como:

$$P_e = \left[\frac{1}{NM}\sum_{i,j}(MV_{ij} - PV_{ij})^2\right]^{\frac{1}{2}}$$

em que MV é o valor medido original, PV é o valor predito correspondente e NM é o número total de valores faltantes. O P_e é muito útil para avaliar métodos de imputação simples, mas, para avaliar a exatidão das estratégias de IM é preferível utilizar as estatísticas T_{acc}, V_b e B introduzidas por Penny e Jolliffe (1999) e usadas, recentemente, por Bergamo et al. (2008); Arciniegas-Alarcón et al. (2014a) e Arciniegas-Alarcón (2015).

T_{acc} é uma medida de acurácia geral composta pela soma da variância combinada entre imputações dentro de posições (Vb) e o viés quadrático médio entre a média das imputações e o valor original retirado no estudo de simulação (B). As estatísticas são apresentadas a seguir:

$$T_{acc} = V_b + B, \text{ em que } V_b = \frac{1}{na}\sum_{l=1}^{na}\left[\frac{\sum_{m=1}^{M}\left(\widehat{y}_{ij(m)} - \overline{Y}_l\right)^2}{M-1}\right] \text{ e } B = \frac{1}{na}\sum_{l=1}^{na} M\frac{\left(\overline{Y}_l - VO_l\right)^2}{M-1}.$$

"na" é o número total de valores retirados da matriz G×E. Cada valor retirado l tem sua correspondente posição (i, j) na matriz, isto é, na i-ésima linha e na j-ésima coluna. M é o número de imputações para o valor ausente l, $\widehat{y}_{ij(m)}$ é a m-ésima imputação para o dito valor, por meio de um dos métodos propostos. $\overline{Y}_l$ é a média das imputações produzidas para o valor ausente l e VO_l é o valor original l no conjunto de dados original completo.

Nesse estudo serão analisadas as três estatísticas, mas a decisão final para escolher o melhor sistema de IM será baseada na T_{acc}. Se V_b for muito grande, então o método pode não

ser muito confiável, mas um valor pequeno para essa variância não significa, necessariamente, que o método de imputação seja bom, pois as imputações podem ser viesadas. Um bom método de imputação será aquele com B pequeno, porque de outra forma indicará que as imputações diferem substancialmente do conjunto de dados observados. Idealmente, é requerido um método de imputação com valores pequenos tanto para V_b quanto para B, que implicam em valores baixos para T_{acc} (Penny e Jolliffe, 1999).

4.3 Resultados

4.3.1 Dados de eucalipto

A Tabela 4.2 mostra a média e mediana das V_b, B e T_{acc} nas diferentes porcentagens de valores retirados aleatoriamente (10, 20 e 35%) para o conjunto de dados de eucalipto. A menor variância em todas as porcentagens sempre foi obtida com o método GCV1, o que era esperado dentro dos esquemas que envolviam o cálculo de intervalos de confiança aproximados para as imputações. GCV1 também superou todas as metodologias que somavam erros aleatórios às imputações produzidas, inicialmente, por GabrielEigen, ou seja, Gnorm, Gadd e GLR. Por outra parte, o algoritmo que maximizou V_b em todos os casos foi Gnorm. Vale a pena ressaltar o desempenho do GLR, pois forneceu em todas as porcentagens, variâncias menores do que as produzidas com GCV4, ou seja, quando se consideraram intervalos de confiança de 95%.

No mesmo conjunto de dados, o menor viés (B) foi obtido pelo algoritmo GCV1 em todas as porcentagens e constitui um novo resultado, pois se esperava que a diminuição da amplitude dos intervalos de imputação tivesse como consequência um aumento de valores na estatística B. O fato indica que as imputações simples com GabrielEigen são de alta qualidade, pelo qual para incorporar variabilidade para produzir IM não se requer de uma amplitude muito grande nos intervalos de confiança. Os algoritmos com imputações mais viesadas foram Gnorm e GCV4, pois maximizaram B. Entretanto, o algoritmo GCV2 foi menos viesado do que GLR e Gadd com taxas baixas de dados faltantes (10 e 20%), mas quando se imputou 35% dos dados, a situação mudou e o algoritmo GLR teve um menor viés do que Gadd e GCV2 respectivamente (Tabela 4.2).

Por último, para os dados de eucalipto, foi analisada a estatística T_{acc} para determinar o melhor sistema de imputação (Tabela 4.2 e Figura 4.1). O melhor método em todos os casos foi claramente o GCV1, pois apresentou a distribuição com menor média e mediana dos valores da T_{acc}, seguido pelo GCV2. Em todas as porcentagens, os algoritmos com o desempenho mais baixo foram Gnorm e GCV4 porque maximizaram a T_{acc}. Entretanto, os métodos Gadd e GLR mostraram um desempenho que pode ser chamado de "intermediário", quer dizer, não superaram GCV1 nem GCV2, mas, tiveram melhores resultados que o Gnorm e GCV4.

Tabela 4.2 – Média e mediana da variância combinada entre imputações (V_b), do viés quadrático médio (B) e da medida geral de acurácia (T_{acc}) sob diferentes porcentagens de retirada aleatória em 1000 simulações a partir do conjunto de dados de eucalipto

Método	10%		20%		35%	
	Média	Mediana	Média	Mediana	Média	Mediana
			V_b			
Gnorm	1,2208	1,1908	1,1318	1,1162	0,9752	0,9668
Gadd	0,4107	0,4033	0,3665	0,3604	0,3025	0,3003
GLR	0,3966	0,3895	0,3543	0,3488	0,2929	0,2905
GCV4	0,8606	0,8627	0,8831	0,8808	0,9665	0,9647
GCV2	0,2241	0,2246	0,2299	0,2293	0,2516	0,2511
GCV1	0,0560	0,0561	0,0575	0,0573	0,0629	0,0628
			B			
Gnorm	1,2700	1,1868	1,2800	1,2464	1,3343	1,3169
Gadd	1,0640	1,0203	1,0908	1,0607	1,1612	1,1391
GLR	1,0591	1,0184	1,0892	1,0656	1,1594	1,1404
GCV4	1,1779	1,1115	1,2326	1,2068	1,3493	1,3172
GCV2	1,0240	0,9577	1,0665	1,0378	1,1732	1,1468
GCV1	0,9843	0,9237	1,0217	1,0022	1,1272	1,1080
			T_{acc}			
Gnorm	2,4908	2,3993	2,4119	2,3917	2,3095	2,2950
Gadd	1,4747	1,4331	1,4573	1,4236	1,4637	1,4370
GLR	1,4556	1,4081	1,4435	1,4161	1,4523	1,4267
GCV4	2,0385	1,9977	2,1157	2,0921	2,3158	2,2871
GCV2	1,2480	1,1843	1,2964	1,2787	1,4248	1,3942
GCV1	1,0403	0,9822	1,0792	1,0585	1,1901	1,1669

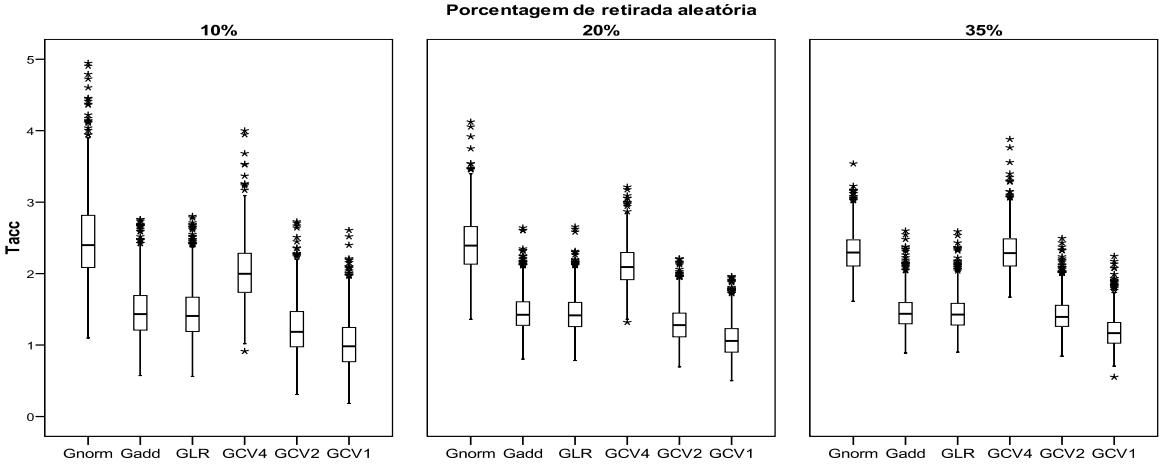


Figura 4.1 – Box plot da distribuição da medida de acurácia T_{acc} para os seis algoritmos no conjunto de dados de eucalipto

4.3.2 Dados de cevada

A Tabela 4.3 mostra a média e mediana das estatísticas V_b, B e T_{acc} nas diferentes porcentagens de valores retirados aleatoriamente (10, 20 e 35%) na matriz de cevada. Do mesmo modo que nas simulações com os dados de eucalipto, V_b foi minimizada sempre por GCV1 e maximizada com o uso de Gnorm e GCV4. A principal diferença em relação ao conjunto de dados de eucalipto se apresentou com os métodos GLR e GCV2 quando se imputou 35% dos dados, pois GLR teve uma média de variâncias igual a 0,0824 enquanto com GCV2 o valor foi 0,0975. Isto significa que em termos de variação entre imputações, GLR superou na porcentagem de ausência mais alta, o desempenho dos algoritmos que construíram intervalos de imputação com amplitude de $2I_e$ (ou de 68% de confiança) e $4I_e$ (ou de 95% de confiança).

Quanto à similaridade com os dados originais, o método com imputações menos viesadas (B) novamente foi GCV1, enquanto as imputações mais viesadas foram produzidas por Gnorm e GCV4. Vale a pena ressaltar que no caso da estatística B com 35% de imputação, os algoritmos Gadd e GLR tiveram melhor desempenho que os algoritmos GCV2 e GCV4.

Para tomar uma decisão definitiva sobre os algoritmos de IM considerados foi utilizada a medida T_{acc}. As distribuições mostradas na Figura 4.2 permitem identificar claramente os métodos de baixo desempenho, Gnorm e GCV4, mas o box plot não mostra de maneira eficiente as diferenças entre os restantes. Por essa razão, para escolher o melhor método se utilizaram as médias e medianas das distribuições (Tabela 4.3).

Essas estatísticas permitem estabelecer que T_{acc} foi minimizada em todas as porcentagens de imputação por GCV1, pelo qual, supera todos os outros sistemas de IM. Entretanto, GLR, Gadd e GCV2 apresentam diferentes resultados dependendo da porcentagem de imputação, isto é, para 10 e 20% GCV2 é melhor do que GLR e Gadd, mas quando aumentar para 35% de imputação, ocorre o contrário.

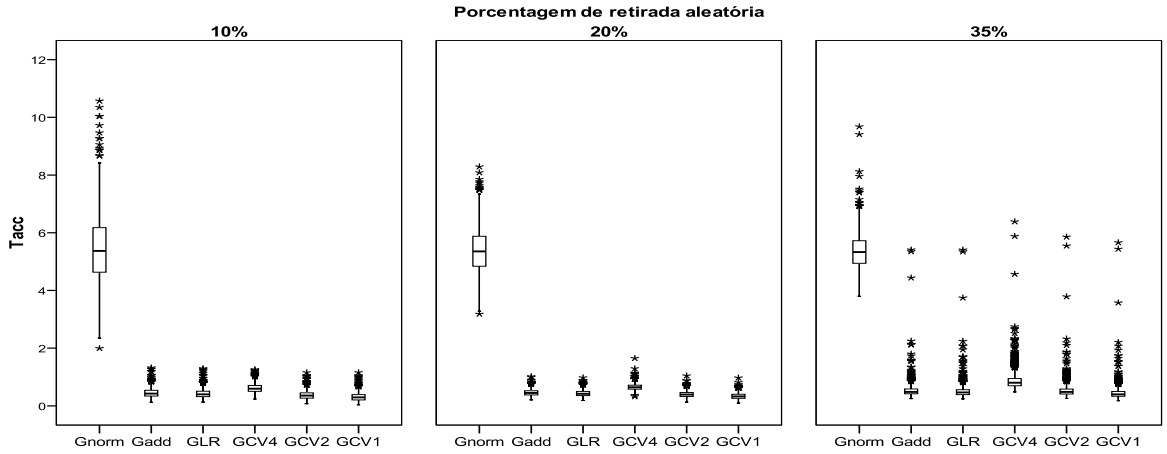


Figura 4.2 – Box plot da distribuição da medida de acurácia T_{acc} para os seis algoritmos no conjunto de dados de cevada

Tabela 4.3 – Média e mediana da variância combinada entre imputações (V_b), do viés quadrático médio (B) e da medida geral de acurácia (T_{acc}) sob diferentes porcentagens de retirada aleatória em 1000 simulações a partir do conjunto de dados de cevada

Método	10%		20%		35%	
	Média	Mediana	Média	Mediana	Média	Mediana
			V_b			
Gnorm	4,1148	4,0170	4,0525	4,0235	3,9476	3,9362
Gadd	0,1287	0,1248	0,1166	0,1140	0,1036	0,1005
GLR	0,1094	0,1072	0,0966	0,0944	0,0824	0,0799
GCV4	0,2528	0,2525	0,2717	0,2681	0,3747	0,3444
GCV2	0,0658	0,0657	0,0707	0,0698	0,0975	0,0897
GCV1	0,0165	0,0164	0,0177	0,0174	0,0244	0,0224
			B			
Gnorm	1,3578	1,2488	1,3401	1,2910	1,4156	1,3751
Gadd	0,3262	0,2952	0,3420	0,3304	0,4435	0,3906
GLR	0,3216	0,2909	0,3366	0,3240	0,4379	0,3849
GCV4	0,3585	0,3326	0,3838	0,3696	0,5241	0,4657
GCV2	0,3109	0,2881	0,3331	0,3220	0,4527	0,3947
GCV1	0,2982	0,2740	0,3195	0,3098	0,4332	0,3753
			T_{acc}			
Gnorm	5,4725	5,3720	5,3926	5,3543	5,3632	5,3310
Gadd	0,4550	0,4246	0,4586	0,4474	0,5471	0,4857
GLR	0,4310	0,3996	0,4332	0,4202	0,5204	0,4632
GCV4	0,6113	0,5960	0,6555	0,6458	0,8988	0,8037
GCV2	0,3767	0,3562	0,4038	0,3940	0,5503	0,4835
GCV1	0,3147	0,2926	0,3372	0,3276	0,4576	0,3988

4.3.3 Dados de trigo

A Tabela 4.4 mostra a média e mediana de V_b, B e T_{acc} nas diferentes porcentagens de valores retirados aleatoriamente (10, 20 e 35%) no conjunto de dados de trigo. A média e mediana das variâncias entre imputações foi maximizada em todos os casos por Gnorm e Gadd, enquanto foi minimizada por GCV1. Dos métodos restantes se destaca o GLR, porque apesar de que sempre apresentou médias de variâncias mais altas do que GCV1, teve menores valores de V_b quando comparado com GCV4 e GCV2 nas porcentagens 20 e 35%.

A maior similaridade entre as imputações e os dados retirados artificialmente (B) foi novamente obtida com GCV1, enquanto as imputações mais viesadas foram produzidas por Gnorm e Gadd. Quando foram imputados 10 e 20% dos valores da matriz, GCV2 foi menos viesado do que GLR e GCV4, mas quando se imputou 35% o resultado foi diferente e GLR teve um viés menor que GCV2 e GCV4 (Tabela 4.4).

Utilizou-se a T_{acc} para escolher o melhor método nos dados de trigo, portanto, apresentam-

Tabela 4.4 – Média e mediana da variância combinada entre imputações (V_b), do viés quadrático médio (B) e da medida geral de acurácia (T_{acc}) sob diferentes porcentagens de retirada aleatória em 1000 simulações a partir do conjunto de dados de trigo

Método	10%		20%		35%	
	Média	Mediana	Média	Mediana	Média	Mediana
			V_b			
Gnorm	3,4515	3,3803	3,3063	3,2894	2,9553	2,9420
Gadd	1,1222	1,1121	1,0705	1,0711	0,8622	0,8418
GLR	0,1569	0,1538	0,1775	0,1371	0,3830	0,4643
GCV4	0,5291	0,5262	0,7875	0,6495	1,9316	2,1086
GCV2	0,1396	0,1382	0,2092	0,1725	0,5044	0,5477
GCV1	0,0349	0,0345	0,0523	0,0431	0,1261	0,1369
			B			
Gnorm	1,4275	1,3556	1,6873	1,5333	2,6444	2,7064
Gadd	0,8454	0,8056	1,1299	0,9427	2,1137	2,2425
GLR	0,6149	0,5760	0,9051	0,7087	1,9978	2,1813
GCV4	0,6882	0,6658	0,9964	0,8255	2,3452	2,4887
GCV2	0,6015	0,5644	0,8773	0,7015	2,0075	2,1847
GCV1	0,5756	0,5361	0,8375	0,6676	1,9188	2,0821
			T_{acc}			
Gnorm	4,8789	4,8259	4,9936	4,9257	5,5997	5,5426
Gadd	1,9676	1,9389	2,2004	2,0546	2,9759	3,0199
GLR	0,7719	0,7333	1,0827	0,8406	2,3808	2,7326
GCV4	1,2173	1,1888	1,7839	1,4787	4,2767	4,7226
GCV2	0,7411	0,7068	1,0865	0,8880	2,5119	2,7580
GCV1	0,6105	0,5681	0,8898	0,7189	2,0449	2,2325

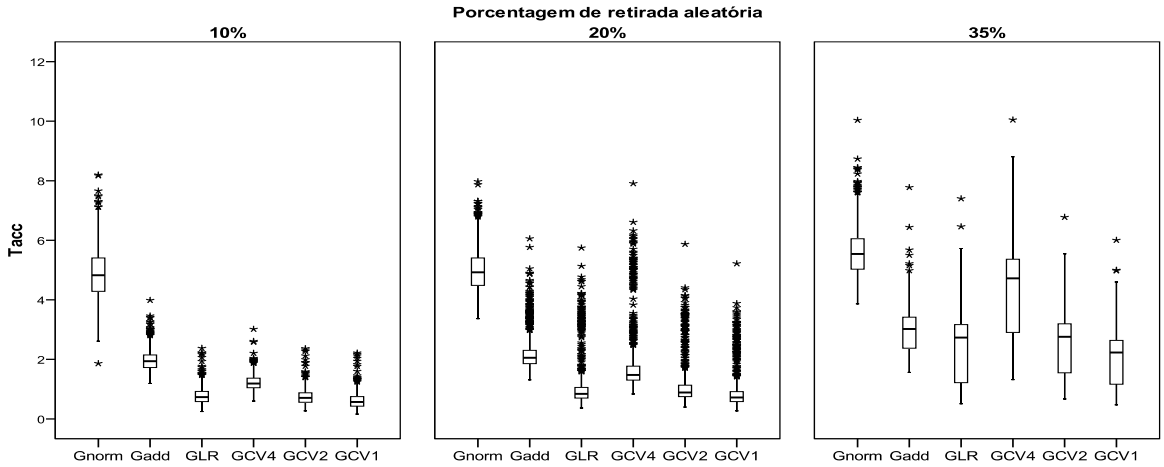


Figura 4.3 – Box plot da distribuição da medida de acurácia T_{acc} para os seis algoritmos no conjunto de dados de trigo

se as distribuições (Figura 4.3) e as correspondentes estimativas dos parâmetros de centralidade (média e mediana na Tabela 4.4). Pelo box plot é possível agrupar os algoritmos em baixo

e alto desempenho. O grupo de alto desempenho está composto por GCV1, GCV2 e GLR e no grupo de baixo desempenho estão GCV4, Gadd e Gnorm. Dentro do grupo de alto desempenho, GCV1 sempre teve os melhores resultados minimizando a T_{acc}. Por outra parte, apesar de que nunca superaram GCV1, os sistemas GCV2 e GLR tiveram um comportamento influenciado pela porcentagem de imputação. Por exemplo, GCV2 teve menores valores de T_{acc} do que GLR com 10% de retirada aleatória, mas GLR teve melhor desempenho do que GCV2 quando se imputou 20% e 35% dos dados.

4.4 Uma situação diferente: Valores ausentes não aleatórios

Os sistemas de IM propostos neste estudo não dependem de alguma pressuposição estrutural nos dados, mas, Bello (1993) adverte que essa falta de pressuposições não implica robustez e em alguns casos podem se produzir resultados distintos aos esperados. As diferentes estruturas que ocorrem em matrizes incompletas podem ser causadas por diferentes mecanismos de ausência dos dados (Little e Rubin, 2002; Paderewski e Rodrigues, 2014) e em experimentos (G×E) é possível encontrar situações com ausência de dados não aleatória (Missing Not At Random - MNAR). Por exemplo, podem ser encontradas matrizes incompletas com padrões sistemáticos porque ao longo dos anos novos genótipos de referência são incluídos e alguns outros são desconsiderados (Denis e Baril, 1992).

Para avaliar como afetam os dados faltantes não aleatórios às metodologias de IM propostas aqui, foram consideradas novamente as matrizes completas de eucalipto, cevada e trigo, mas, diferentemente do estudo de simulação, foram inseridos apenas uma vez dados ausentes de forma arbitrária.

Em cada uma das matrizes foram removidos dentro de cada ambiente, os genótipos pertencentes ao terço com menor altura no caso de eucalipto e ao terço com menor rendimento nos casos de cevada e trigo (personal communication of W. Yan 2014). Desse modo, nos dados de eucalipto se produziram 49 valores ausentes (35%), tendo como consequência a perda total de um genótipo. Entretanto, produziram-se 36 dados faltantes (33.33%) na matriz de cevada e 72 valores omissos na matriz de trigo (33.33%). No caso da matriz de trigo, a retirada arbitrária teve como consequência a perda total de cinco genótipos. Uma vez com as matrizes incompletas, aplicaram-se os métodos de IM e calculadas as correspondentes estatísticas V_b, B e T_{acc}. Os resultados apresentam-se na Tabela 4.5.

No conjunto de dados de eucalipto, as imputações menos viesadas (B) foram produzidas por GCV2, enquanto o sistema de imputação mais viesado foi Gnorm. Entretanto, a menor variância entre imputações (V_b) foi obtida com GCV1 e a maior obtida com Gnorm. A medida de acurácia geral T_{acc} indica que o melhor método foi GCV1 e se explica esse resultado porque embora GCV1 não obteve o melhor desempenho quanto à similaridade com os dados originais

Tabela 4.5 – Estatísticas V_b, B e T_{acc} depois de fazer uma retirada não aleatória nas matrizes de eucalipto, cevada e trigo

	Eucalipto		
Método	V_b	B	T_{acc}
Gnorm	0,3141	4,7464	5,0606
Gadd	0,1647	4,7218	4,8864
GLR	0,1628	4,7136	4,8764
GCV4	0,4269	4,7016	5,1284
GCV2	0,1172	4,6887	4,8059
GCV1	0,0271	4,7077	4,7348
	Cevada		
Método	V_b	B	T_{acc}
Gnorm	3,9782	1,5779	5,5561
Gadd	0,0879	0,4797	0,5676
GLR	0,0862	0,4302	0,5163
GCV4	0,2576	0,5354	0,7930
GCV2	0,0806	0,4437	0,5244
GCV1	0,0184	0,4454	0,4638
	Trigo		
Método	V_b	B	T_{acc}
Gnorm	0,6643	14,2154	14,8797
Gadd	0,5055	13,7940	14,2996
GLR	0,0021	14,0732	14,0753
GCV4	0,3224	13,9020	14,2244
GCV2	0,0845	13,9838	14,0683
GCV1	0,0187	14,0694	14,0881

retirados, compensou dita situação atingindo alta precisão por meio da minimização de V_b.

No conjunto de dados de cevada, novamente o melhor método de IM segundo a estatística T_{acc} foi GCV1. Como aconteceu no conjunto de dados de eucalipto, GCV1 teve um desempenho mais baixo que GLR e GCV2 quanto ao viés das imputações, mas nenhum deles o superou quanto à variabilidade em relação à média dos valores imputados. O método com mais baixo desempenho novamente foi Gnorm (Tabela 4.5).

Até este ponto, os resultados com valores faltantes não aleatórios não diferiram muito dos obtidos no estudo de simulação, mas, o conjunto de dados de trigo forneceu um novo resultado para levar em conta. Nesse caso, o GCV1 que sempre teve os melhores resultados porque minimizava o viés das imputações e/ou minimizava a variância V_b, foi superado por GCV2 e GLR respectivamente, utilizando T_{acc} como critério de avaliação. Aqui, GLR minimizou V_b e todos os sistemas, exceto Gnorm e GLR, tiveram maior similaridade (B) com os dados originais quando comparados com GCV1 (Tabela 4.5).

4.5 Discussão e conclusões

O objetivo foi cumprido, obtiveram-se metodologias para produzir imputação múltipla utilizando o algoritmo GabrielEigen. Foram consideradas matrizes (G×E) com estrutura de interação simples (eucalipto), moderada (cevada) e complexa (trigo), mas, de acordo com o estudo de simulação, GCV1 foi o algoritmo com melhor desempenho em todos os casos, portanto, o mais recomendado.

No segundo lugar, ficaram os algoritmos GCV2 e GLR, mas dependem fortemente da estrutura da interação e da porcentagem de imputação. Assim, GLR tem melhor desempenho do que GCV2 quando a porcentagem de dados faltantes for alta (∼35%) e a interação for moderada ou complexa. Nos outros casos pode ser preferível o GCV2.

Quanto aos resultados obtidos no exemplo, utilizando uma retirada arbitrária de valores, o GCV1 de novo foi o melhor com as matrizes de interação simples e moderada, mas GLR e GCV2 tiveram melhor desempenho quando se utilizou a matriz de interação complexa. Esse resultado pode ser a origem de estudos futuros e de recomendações práticas.

A retirada arbitrária foi realizada apenas uma vez na matriz de trigo, então, para confirmar a robustez do GCV1 às observações ausentes não aleatórias em estruturas de interação complexas, seriam necessárias pesquisas adicionais com simulações que podem utilizar o procedimento proposto aqui para gerar valores MNAR em experimentos (G×E). Enquanto essas pesquisas são desenvolvidas, a recomendação para essa situação específica é utilizar simultaneamente os algoritmos GCV1, GCV2 e GLR e avaliá-los com as estatísticas apresentadas aqui.

A estratégia de construir intervalos de confiança para as imputações utilizando validação cruzada teve sucesso, mas, um ponto para destacar é que nessa ocasião os intervalos de confiança de 95%, tradicionalmente utilizados em estatística, não forneceram os melhores resultados. Por último, o aspecto computacional não foi problema neste estudo, mas se as matrizes analisadas forem de maior tamanho, poderia ser considerada a validação cruzada "k-fold" no lugar da "leave-one-out" descritas em James et al. (2013).

Referências.

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

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

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; KRZANOWSKI, W. J.; DIAS, C. T. S. Deterministic imputation in multienvironment trials. **ISRN Agronomy**, Cairo, v.2013, p.1-17, 2013.

ARCINIEGAS-ALARCÓN, S.; DIAS, C. T. S.; GARCÍA-PEÑA, M. Distribution-free multiple imputation in incomplete two-way tables. **Pesquisa Agropecuária Brasileira**, Brasilia, v.49, p.683-691, 2014a.

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; KRZANOWSKI, W. J.; DIAS, C. T. S. Imputing missing values in multi-environment trials using the singular value decomposition: An empirical comparison. **Communications in Biometry and Crop Science**, Warsaw, v.9, p.54-70, 2014b.

_____. An alternative methodology for imputing missing data in trials with genotype-by-environment interaction: some new aspects. **Biometrical Letters**, Poznan, v.51, p.75-88, 2014c.

ARCINIEGAS-ALARCÓN, S. **Imputação de dados em experimentos multiambientais:** novos algoritmos utilizando a decomposição por valores singulares. 110p. 2015. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba. 2015.

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

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

CALIŃSKI, T.; CZAJKA, S.; DENIS, J. B.; KACZMAREK, Z. Further study on estimating missing values in series of variety trials. **Biuletyn Oceny Odmian**, Poznan, v.30, p.7-38, 1999.

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

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

FORKMAN, J. A resampling test for principal component analysis of genotype-by-environment interaction. **Acta et Commentationes Universitatis Tartuensis de Mathematica**, Tartu, v.19, p.27-33, 2015.

GARCÍA-PEÑA, M.; ARCINIEGAS-ALARCÓN, S.; BARBIN, D. Climate data imputation using the singular value decomposition: an empirical comparison. **Revista Brasileira de Meteorologia**, Rio de Janeiro, v.29, p.527-536, 2014.

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

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

GRAHAM, J. W. **Missing data. Analysis and Design**. New York, Springer, 2012. 339p.

HUSSON, F.; JOSSE, J.; MAZET, J. **FactoMineR:** Multivariate Exploratory Data Analysis and Data Mining with R. R package version 1.26. 2014. Disponível em: http://CRAN.R-project.org/package=FactoMineR. Acesso em: 21 ago. 2015.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning with applications in R**. New York, Springer, 2013. 441p.

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

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

_____. Selecting the number of components in principal component analysis using cross-validation approximations. **Computational Statistics and Data Analysis**, New york, v.56, p.1869-1879, 2012b.

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

KRZANOWSKI, W. J.; MARRIOTT, F. H. C. **Multivariate analysis, Part 1:** Distributions, Ordination, and inference. New York, John Wiley & Sons, 1994. 280p.

LAVORANTI, O. J.; DIAS, C. T. S.; KRZANOWSKI, W. J. Phenotypic stability and adaptability via AMMI model with bootstrap re-sampling. **Pesquisa Florestal Brasileira**, Colombo, v.54, p.45-52, 2007.

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

OUNPRASEUTH, S.; MOORE, P. C.; YOUNG, D. M. Imputation techniques for incomplete data in quadratic discriminant analysis. **Journal of Statistical Computation and Simulation**, New York, v.82, p.863-877, 2012.

PADEREWSKI, J. An R function for imputation of missing cells in two-way data sets by EM-AMMI algorithm. **Communications in Biometry and Crop Science**, Warsaw, v.8, p.60-69, 2013.

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

PENNY, K. I.; JOLLIFFE, I. T. Multivariate outlier detection applied to multiply imputed laboratory data. **Statistics in Medicine**, New York, v.18, p.1879-1895, 1999.

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

R CORE TEAM. **R:** A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2015. Disponível em: http://www.R-project.org/. Acesso em: 21 ago. 2015.

RAD, M. R. N.; KADIR, M. A.; RAFII, M. Y.; JAAFAR, H. Z. E.; NAGHAVI, M. R.; AHMADI, F. Genotype×environment interaction by AMMI and GGE biplot analysis in three consecutive generations of wheat (Triticum aestivum) under normal and drought stress conditions. **Australian Journal of Crop Science**, Sydney, v.7, p.956-961, 2013.

RÄSSLER, S.; RUBIN, D. B.; ZELL, E. R.Imputation. **WIREs Computational Statistics**, New York, v.5, p.20-29, 2013.

RODRIGUES, P.; PEREIRA, D. G. S.; MEXIA, J. T. A comparison between joint regression analysis and the additive main and multiplicative interaction model: the robustness with increasing amounts of missing data. **Scientia Agricola**, Piracicaba, v.68, p.679-686, 2011.

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

_____. **Multiple imputation for nonresponse in surveys**. New York: John Wiley & Sons, 1987. 258p.

SCHAFER, J. L. Multiple imputation: A primer. **Statistical Methods in Medical Research**, Thousand Oaks, v.8, p.3-15, 1999.

SRIVASTAVA, M. S.; DOLATABADI, M. Multiple imputation and other resampling scheme for imputing missing observations. **Journal of Multivariate Analysis**, New York, v.100, p.1919-1937, 2009.

VAN BUUREN, S. **Flexible imputation of missing data**. Boca Raton, Chapman & Hall/CRC, 2012. 343p.

VAN GINKEL, J. R.; KROONENBERG, P. M. Using generalized procrustes analysis for multiple imputation in principal component analysis. **Journal of Classification**, New York, v.31, p.242-269, 2014.

WONG, J. **Imputation**. R package version 2.0.1., 2013. Disponível em: https://github.com/jeffwong/imputation. Acesso em: 21 agos. 2015.

WRIGHT, K. **agridat:** Agricultural datasets. R package version 1.4., 2012. Disponível em: http://CRAN.R-project.org/package=agridat. Acesso: 21 ago. 2015.

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

_____. Mega-environment analysis and test location evaluation based on unbalanced multiyear data. **Crop Science**, Madison, v.55, p.113-122, 2015.

YANG, R. C. Mixed-model analysis of crossover genotype-environment interactions. **Crop Science**, Madison, v.47, p.1051-1062, 2007.

5 TRABALHOS FUTUROS

Com o objetivo de dar continuidade aos estudos apresentados aqui, é sugerido:

- Propor um modelo global para os modelos AMMI multiatributo, isto é, $Y(r;g,e)$ sendo Y a variável aleatória associada ao atributo r para o genótipo g e o ambiente e. $E[Y(r;g,e)] = \mu_{(r)} + \alpha_{(r;g)} + \beta_{(r;e)} + \sum_k \lambda_{(k:r)}\gamma_{(k:g)}\delta_{(k:e)}$, em que $Y(*;g,e)$ é independente e $V[Y(*;g,e)] = \Sigma$, matriz positiva com tamanho de acordo ao número de atributos. A dificuldade pode ser a estimação dos parâmetros, para o qual sugere-se usar uma aproximação bayesiana.

- Propor uma representação gráfica para os modelos AMMI multiatributo, robusta para dados discrepantes e na qual, possam ser avaliados conjuntamente genótipos, ambientes e atributos.

- Estender a aplicação dos mecanismos de imputação propostos para dados multiatributo.

APÊNDICES

Tabelas de seleção de número de componentes modelo AMMI

Tabela .1 – Análise da variância conjunta completa calculada a partir das médias usando os sistemas de Gollob e Cornelius para a produtividade de grãos

Fontes de variação	G.L.[1]	Q.M.[2] Gollob	F Gollob	Valor p	G.L. Cornelius	Q.M. Cornelius	F_R Cornelius	Valor p
Ambiente (E)	8	4,581	22,501	<2,2E-16	–	–	–	–
Genótipo (G)	12	0,878	4,312	1,9E-05	–	–	–	–
Interação (GE)	96	0,204	15,503	<2,0E-16	–	–	–	–
$IPCA_1$[3]	19	0,499	37,990	3,5E-58	77	0,131	9,953	7,9E-41
$IPCA_2$	17	0,184	14,005	2,0E-26	60	0,116	8,805	2,2E-33
$IPCA_3$	15	0,168	12,806	1,3E-22	45	0,098	7,472	4,5E-25
$IPCA_4$	13	0,169	12,897	8,0E-21	32	0,069	5,268	3,2E-14
$IPCA_5$	11	0,102	7,752	2,9E-11	21	0,052	3,967	9,8E-08
$IPCA_6$	9	0,075	5,707	4,3E-07	12	0,035	2,661	0,002
$IPCA_7$	7	0,051	3,905	5,0E-04	5	0,012	0,921	0,468
$IPCA_8$	5	0,012	0,921	0,468	0	0,000	0,000	–
Erro médio$_k$	216	0,013	–	–	–	–	–	–

[1] G.L.: Graus de liberdade [2] Q.M.: Quadrado médio

[3] IPCA$_k$: (interaction principal component analysis) modelo com k componentes, $k = 1, 2, \ldots, 17$.

Tabela .2 – Análise da variância conjunta completa calculada a partir das médias usando os sistemas de Gollob e Cornelius para o número médio de vagens por planta

Fontes de variação	G.L.[1]	Q.M.[2] Gollob	F Gollob	Valor p	G.L. Cornelius	Q.M. Cornelius	F_R Cornelius	Valor p
Ambiente (E)	8	621,770	10,248	3,0E-10	–	–	–	–
Genótipo (G)	12	220,380	3,632	0,0002	–	–	–	–
Interação (GE)	96	60,670	4,273	<2,0E-16	–	–	–	–
$IPCA_1$[3]	19	151,127	10,643	5,3E-22	77	38,355	2,701	7,8E-09
$IPCA_2$	17	64,119	4,515	5,0E-08	60	31,055	2,187	2,2E-05
$IPCA_3$	15	62,885	4,429	2,9E-07	45	20,445	1,440	0,046
$IPCA_4$	13	38,124	2,685	0,002	32	13,262	0,934	0,574
$IPCA_5$	11	26,243	1,848	0,048	21	6,463	0,455	0,982
$IPCA_6$	9	11,566	0,814	0,603	12	2,636	0,186	0,999
$IPCA_7$	7	4,480	0,315	0,946	5	0,054	0,004	1,000
$IPCA_8$	5	0,054	0,004	1,000	0	0,000	0,000	–
Erro médio$_k$	216	14,200	–	–	–	–	–	–

[1] G.L.: Graus de liberdade [2] Q.M.: Quadrado médio

[3] IPCA$_k$: (interaction principal component analysis) modelo com k componentes, $k = 1, 2, \ldots, 17$.

Tabela .3 – Análise da variância conjunta completa calculada a partir das médias usando os sistemas de Gollob e Cornelius para o número médio de sementes por vagem

Fontes de variação	G.L.[1]	Q.M.[2] Gollob	F Gollob	Valor p	G.L. Cornelius	Q.M. Cornelius	F_R Cornelius	Valor p
Ambiente (E)	8	13,513	70,111	<2,2E-16	–	–	–	–
Genótipo (G)	12	1,100	5,708	2,5E-07	–	–	–	–
Interação (GE)	96	0,193	3,483	1,7E-14	–	–	–	–
$IPCA_1$[3]	19	0,413	7,471	2,0E-15	77	0,1383	2,4992	9,8E-08
$IPCA_2$	17	0,357	6,443	3,1E-12	60	0,0765	1,3816	0,049
$IPCA_3$	15	0,119	2,150	0,009	45	0,0623	1,1254	0,284
$IPCA_4$	13	0,085	1,538	0,105	32	0,0530	0,9579	0,535
$IPCA_5$	11	0,076	1,377	0,185	21	0,0409	0,7385	0,789
$IPCA_6$	9	0,061	1,104	0,360	12	0,0257	0,4643	0,933
$IPCA_7$	7	0,030	0,.534	0,808	5	0,0203	0,3670	0,871
$IPCA_8$	5	0,020	0,.367	0,871	0	0,000	0,000	–
Erro médio$_k$	216	0,055	–	–	–	–	–	–

[1] G.L.: Graus de liberdade [2] Q.M.: Quadrado médio

[3] IPCA$_k$: (interaction principal component analysis) modelo com k componentes, $k = 1, 2, \ldots, 17$.

Tabela .4 – Análise da variância conjunta completa calculada a partir das médias usando os sistemas de Gollob e Cornelius para a massa de 100 sementes

Fontes de variação	G.L.[1]	Q.M.[2] Gollob	F Gollob	Valor p	G.L. Cornelius	Q.M. Cornelius	F_R Cornelius	Valor p
Ambiente (E)	8	88,672	32,106	<2,2E-16	–	–	–	–
Genótipo (G)	12	80,806	29,259	<2,2E-16	–	–	–	–
Interação (GE)	96	2,762	13,810	<2,0E-16	–	–	–	–
$IPCA_1$[3]	19	4,906	24,531	1,1E-43	77	2,233	11,163	1,0E-44
$IPCA_2$	17	3,640	18,201	6,6E-33	60	1,834	9,169	1,3E-34
$IPCA_3$	15	3,270	16,351	4,9E-28	45	1,355	6,776	1,0E-22
$IPCA_4$	13	2,512	12,562	2,5E-20	32	0,885	4,425	1,8E-11
$IPCA_5$	11	1,101	5,507	9,5E-08	21	0,772	3,858	1,8E-07
$IPCA_6$	9	0,832	4,158	5,8E-05	12	0,727	3,633	5,7E-05
$IPCA_7$	7	0,922	4,611	7,9E-05	5	0,453	2,264	0,049
$IPCA_8$	5	0,453	2,264	0,049	0	0,000	0,000	–
Erro médio$_k$	216	0,200	–	–	–	–	–	–

[1] G.L.: Graus de liberdade [2] Q.M.: Quadrado médio

[3] IPCA$_k$: (interaction principal component analysis) modelo com k componentes, $k = 1, 2, \ldots, 17$.