ISSN: 2317-0840

# Introdução à metodologia AMMI

*Carlos T. S. Dias*[1], *Kuang Hongyu*[2,4], *Lúcio B. Araújo*[3†], *Maria Joseane C. Silva*[4], *Marisol García-Peña*[4], *Mirian F. C. Araújo*[3,4], *Priscila N. Faria*[3], *Sergio Arciniegas-Alarcón*[4]

[1] *Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ/USP).*
[2] *Universidade Federal do Mato Grosso, Departamento de Estatística (DES/UFMT).*
[3] *Universidade Federal de Uberlândia, Faculdade de Matemática (FAMAT/UFU).*
[4] *Programa de Pós-Graduação em Estatística e Experimentação Agronômica (USP/Esalq/PPGEEA)*

**Resumo:** *Este trabalho é baseado no minicurso "A Metodologia AMMI: Com Aplicação ao Melhoramento Genético"ministrado durante a 58ª RBRAS e 15º SEAGRO realizado em Campina Grande - PB e têm o objetivo de introduzir a metodologia AMMI para aqueles que têm e aqueles que não têm formação matemática. Não pretendemos apresentar um trabalho detalhado, mas a intenção é que sirva como uma luz para pesquisadores e estudantes ao nível de graduação e pós-graduação. Em outras palavras, é um trabalho para estimular a pesquisa e a busca por conhecimento em uma área de métodos estatísticos. Para isto faz- se um revisão sobre a interação genótipo × ambiente, define-se os modelos AMMI e alguns critérios de seleção e por fim gráfico biplot. Mais detalhes sobre o assunto pode ser consultado no material produzido para o Minicurso.*

**Palavras-chave:** Interação genótipo × ambiente; modelos AMMI.

**Abstract:** *This work is based on the short course "A Metodologia AMMI: Com Aplicação ao Melhoramento Genético"taught during the 58ª RBRAS and 15º SEAGRO held in Campina Grande - PB and aim to introduce the AMMI method for those that have and no have the mathematical training. We do not intend to submit a detailed work, but the intention is to serve as a light for researchers, graduate and postgraduate students. In other words, is a work to stimulate research and the quest for knowledge in an area of statistical methods. For this propose we make a review about the genotype × environment interaction, definition of the AMMI models and some selection criteria and biplot graphic. More details about it can be found in the material produced for the Short Course.*

**Keywords:** Genotype × environment interaction; AMMI models.

## Introdução

Em experimentos agrícolas, os ensaios são realizados em vários ambientes para, por meio de uma análise estatística adequada, isolar o componente de variabilidade devido à interação entre genótipos e ambientes (G ×E). Os programas de melhoramento têm como principal objetivo buscar técnicas capazes de detectar de maneira aprofundada essas interações para a seleção de genótipos consistentes e de elevada produtividade.

O método AMMI (*Additive Main effects and Multiplicative Interaction model*) surge, então, com a finalidade de estudar detalhadamente as interações (G × E) por meio da decomposição ortogonal da soma de quadrados dessas interações, fato que o torna vantajoso se comparado aos métodos tradicionais. Além disso, algumas razões pelas quais a metodologia AMMI vem crescendo em usuários e aplicações é o seu forte apelo para explicar o padrão de resposta da interação entre fatores em pesquisas experimentais, o pronto acesso a algum ambiente computacional para realizar a programação e, por conseguinte os cálculos e a aplicação prática no estudo da interação entre genótipos e ambientes no melhoramento genético vegetal ou animal, embora a metodologia seja universal, ou seja, pode ser aplicado à qualquer pesquisa nas diferentes áreas do conhecimento humano, que envolva o estudo de fatores e suas interações.

[†]Autor correspondente: lucio@famat.ufu.br.

## Interação Genótipo × Ambiente

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

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

Cockerham (1963) atribuiu o aparecimento de interações ($G \times E$) como sendo devido a respostas diferenciais do mesmo conjunto gênico em ambientes distintos ou pela expressão de diferentes conjuntos gênicos em diferentes ambientes. Quando um mesmo conjunto de genes se expressa em diferentes ambientes, as diferenças nas respostas podem ser explicadas pela heterogeneidade das variâncias genéticas e experimentais ou por ambas, e , quando diferentes conjuntos de genes se expressam em ambientes distintos, as diferenças nas respostas explicam-se pela inconsistência das correlações genéticas entre os valores de um mesmo caráter em dois ambientes (FALCONER, 1989). Segundo Cruz e Regazzi (1994), a interação ($G \times E$) também pode surgir em função de fatores fisiológicos e bioquímicos próprios de cada genótipo cultivado. Chaves et al. (1989) relatam ainda que a falta de ajuste do modelo estatístico adotado ao conjunto de dados pode ser uma das causas da interação ($G \times E$) significativa.

Várias metodologias têm sido propostas no sentido de entender melhor o efeito da interação ($G \times E$). Algumas dessas propostas são: zoneamento ecológico ou estratificação de ambientes, ou seja, identificar regiões ou sub-regiões onde o efeito da interação seja não significativo pode levar a identificação de genótipos que se adaptam a ambientes específicos e ainda identificar genótipos com uma ampla adaptação ou estabilidade (RAMALHO et al., 1993). Da importância dessa interação no campo experimental devem-se escolher os métodos estatísticos que melhor expliquem a informação contida nos dados, um daqueles métodos é o modelo de interação multiplicativa, também conhecido como o modelo de efeitos principais aditivos e interação multiplicativa - AMMI, que tem como objetivo selecionar modelos que expliquem o padrão de resposta da interação, deixando fora o ruído presente nos dados (ARCINIEGAS-ALARCÓN; DIAS, 2009).

Foram ilustrados por Allard e Bradshaw (1964) e ampliados aqui, alguns tipos de interações em que são considerados dois genótipos em dois ambientes. Nas Figuras 1, 2 e 3 são apresentadas três situações básicas que trazem diferentes consequências para o melhoramento. Em dois ambientes $A_1$ e $A_2$ são avaliados dois genótipos $G_1$ e $G_2$. Assume-se que o ambiente $A_1$ é mais favorável para a manifestação do caráter genérico Y, que pode ser qualquer variável métrica. Na Figura 1 há ausência de interação, ou seja, a mudança das condições ambientais afeta igualmente o comportamento dos genótipos e a diferença entre eles permanece constante nos dois ambientes. Observa-se nas Figuras 1c, 1d e 1e que os genótipos podem coincidir e portanto, terem o mesmo comportamento nos dois ambientes ou apresentarem diferenças

de um ambiente para outro, mas com ausência da interação. Nas Figuras 1 e 2, exceto Figura 1c, a mudança de ambiente afeta desigualmente a manifestação do caráter para os dois genótipos, ou seja, a diferença entre os genótipos varia entre ambientes. Na Figura 2, o efeito de cada ambiente não modifica a classificação dos genótipos, sendo o $G_1$ superior a $G_2$ em todas condições. Neste caso a interação é denominada simples ou quantitativa.

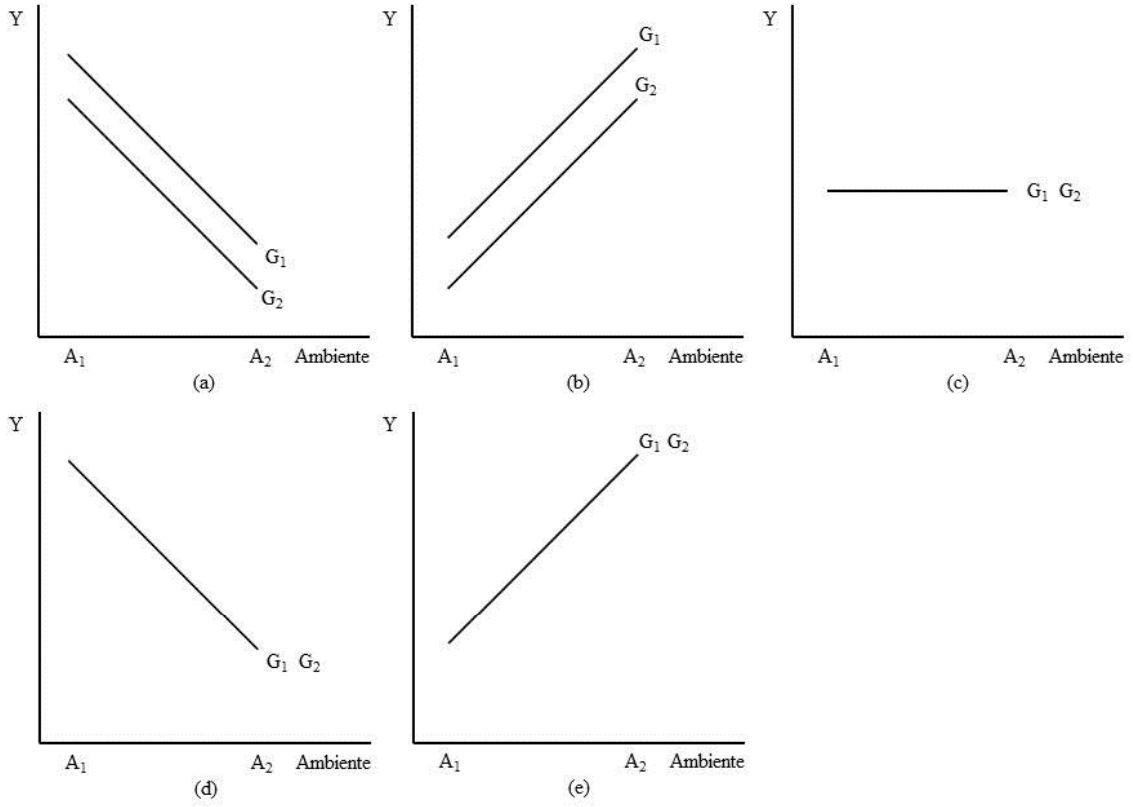


Figura 1: Comportamento de dois genótipos ($G_1$ e $G_2$) em duas condições ambientais ($A_1$ e $A_2$) com ausência de interação

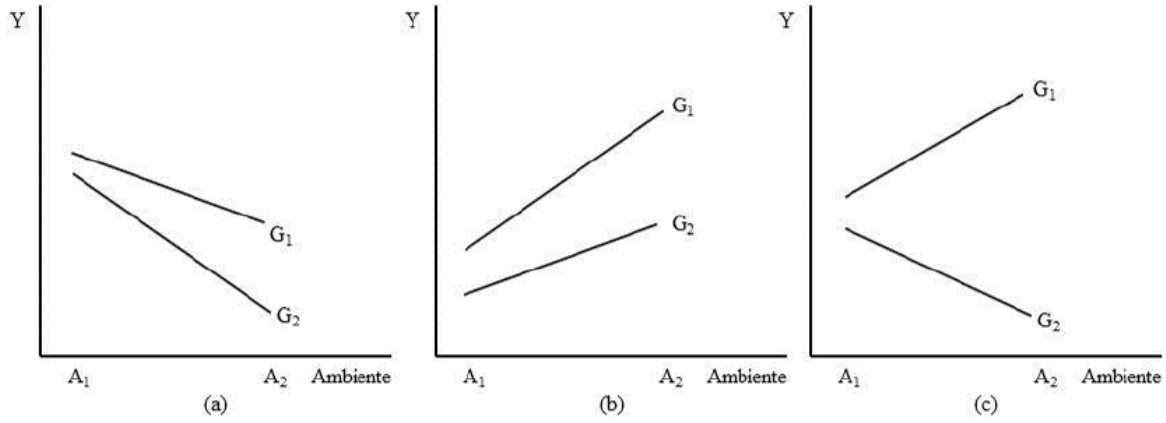


Figura 2: Comportamento de dois genótipos ($G_1$ e $G_2$) em duas condições ambientais ($A_1$ e $A_2$) com interação simples ou quantitativa

A Figura 3 apresenta uma mudança na classificação dos genótipos. Para este tipo, a interação é denominada cruzada ou qualitativa. Nas Figuras 1 e 2, as implicações para o melhoramento é que um mesmo genótipo $G_1$ é melhor adaptado às duas condições ambientais e uma seleção baseada na média dos ambientes beneficiará sempre o melhor genótipo. Na Figura 3, a seleção baseada na média dos ambientes não é capaz de satisfazer o conjunto dos ambientes podendo levar a uma seleção de genótipos mal adaptados a uma situação particular. Nas Figuras 1b e 2b ocorrem sinergismos e nas Figuras 1a e 2a antagonismo entre genótipos e ambientes, ao passar do ambiente $A_1$ para o ambiente $A_2$.

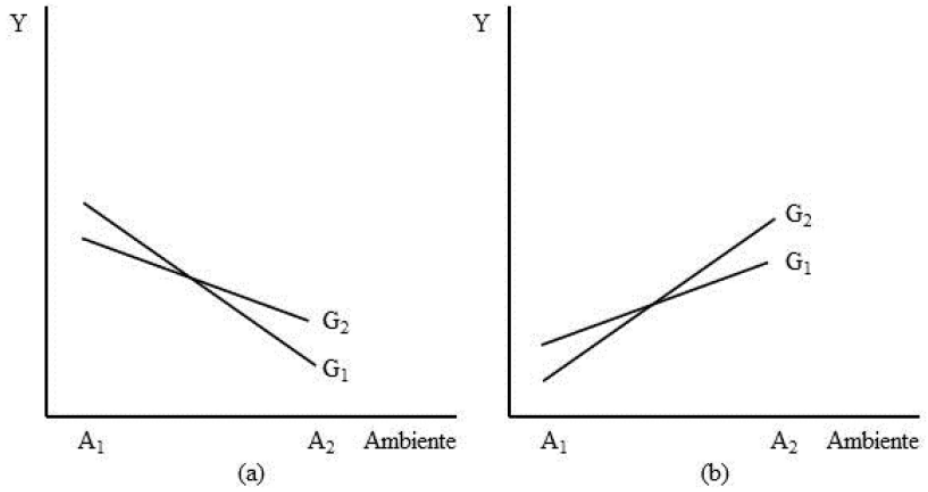


Figura 3: Comportamento de dois genótipos ($G_1$ e $G_2$) em duas condições ambientais ($A_1$ e $A_2$) com interação cruzada ou qualitativa

Para as três figuras anteriores o papel de $G_1$ e $G_2$ pode ser permutado com $A_1$ e $A_2$, obtendo-se interpretações semelhantes. Quando se consideram vários genótipos avaliados em vários ambientes, a combinação de situações como as das Figuras 1, 2 e 3 formam um emaranhado de situações, difícil de ser interpretado, exigindo métodos adequados de análise da interação genótipos ambientes (ARAÚJO, 2008).

As variáveis ambientais também podem ser classificadas em dois tipos: previsíveis e imprevisíveis (ALLARD; BRADSHAW, 1964). As variáveis previsíveis seriam as características gerais do clima e solos que ocorrem de maneira sistemática ou que estão sob controle do homem. Já as variáveis imprevisíveis correspondem às flutuações climáticas tais como quantidade e distribuição de chuvas, temperatura e outros fatores que não podem ser controlados pelo homem.

A interação genótipos $\times$ ambientes, é uma fonte de variação fenotípica, que na maioria dos casos é inseparável da variância ambiental (FALCONER, 1989). Na prática, para verificar a significância da interação de genótipos com ambientes, é necessário repetir o experimento várias vezes, pois se o experimento for realizado somente em um ambiente, poderá ocorrer uma superestimação dos ganhos genéticos (CROSSA et al., 1990).

Existe uma concordância geral entre melhoristas de plantas de que interação genótipos $\times$ ambientes tem um importante significado para a obtenção de variedades superiores. Entretanto, de acordo com Allard (1971) é muito difícil encontrar concordâncias sobre o que se deve conhecer em relação a interação genótipos $\times$ ambientes e como utilizá-la.

A natureza da interação genótipos $\times$ ambientes também deve ser considerada e não somente a verificação de sua existência (VENCOVSKY; BARRIGA, 1992). Assim a natureza pode ser simples e complexa. A interação de natureza simples indica a presença de genótipos adaptados em um grande número de ambientes, sendo possível fazer uma recomendação generalizada de cultivares. A interação com natureza complexa mostra que existem genótipos adaptados a apenas alguns ambientes, o que traz uma complicação ao pesquisador, quando da recomendação de cultivar.

A existência de interação genótipos × ambientes, produz uma barreira de dificuldades aos melhoristas na identificação de genótipos superiores, tanto no processo de seleção quanto no processo de recomendação de cultivares. Essa interação indica que o comportamento dos genótipos nos experimentos depende principalmente das condições ambientais a que são submetidos. Assim a resposta obtida de um genótipo, em comparação a outro, é variável, sendo que essas variações se apresentam devido a mudança de ambientes (OLIVEIRA; DUARTE; PINHEIRO, 2003; KANG, 1998).

Assim, a interação de genótipos × ambientes deve ser considerada, não como um problema ou um fator indesejável, cujo efeito deve ser minimizado, mas deve ser enfrentada como um fenômeno biológico natural, que deve ser bem conhecido para melhor aproveitá-lo no processo de seleção (CHAVES, 2001). Logo os genótipos que interagem positivamente com os ambientes podem fazer a diferença entre um bom e um ótimo cultivar (DUARTE; VENCOVSKY, 1999).

## A interação $G \times E$ e o enfoque estatístico

A existência de interação genótipos × ambientes tem sido reconhecida há longo tempo de acordo com Freeman (1973), sendo a referência mais antiga feita por Fisher e Mackenzie em 1923, a qual precede a análise de variância (ANOVA) conjunta. Desde então, muitos trabalhos tem sido feitos para análises estatísticas da interação genótipos × ambientes, seja por estatísticos, agrônomos, melhoristas e geneticistas.

Os métodos mais simples que utilizam componentes de variância para a interpretação dos resultados quando existem interações, incluem estudos de regressão, métodos baseados em análises modificadas e métodos envolvendo variáveis ambientais externas.

Estudos realizados sobre interação genótipos × ambientes mostraram que a regressão foi utilizada pela primeira vez por Yates e Cochran (1938), analisando grupos de experimentos. Neste caso, o grau de associação entre diferenças varietais pôde ser verificado pelo cálculo da regressão dos rendimentos das variedades isoladas sobre os rendimentos médios de todas as variedades, mostrando assim que a regressão explicava grande parte da interação em uma série de experimentos de cevada. O método de regressão foi também usado por Perkins e Jinks (1968) para estimação de parâmetros em um modelo genético aplicado à biometria.

Outro método usado é aquele no qual os dados são organizados em uma tabela de dupla entrada em que o processo de investigação da interação é feito através da ANOVA. Esta análise envolve vários experimentos, e por isso é possível determinar a magnitude da interação, através da razão do quadrado médio da interação ($QM_{G\times E}$) pelo quadrado médio do resíduo ($QMR_{\text{es}}$). A detecção de significância para a interação não esclarece, contudo, as implicações que estas possam ter sobre o melhoramento, de forma que, estudos de detalhamento deste componente da variação são em geral necessários.

Supondo-se que os genótipos foram avaliados nos diversos ambientes com $r$ repetições em delineamento experimental que permita a estimativa da variação residual em cada ambiente, o modelo mais simples e comum para a análise estatística de um conjunto de dados é dado por (ANOVA para grupos de experimentos):

$$Y_{ij} = \mu + g_i + e_j + (ge)_{ij} + \varepsilon_{ij} \quad (1)$$

sendo que: $Y_{ij}$ : é a resposta média do $i$-ésimo genótipo no $j$-ésimo ambiente; $\mu$: é uma constante comum às respostas (normalmente a média geral); $g_i$: é o efeito do $i$-ésimo genótipo (i = 1; 2; ...; g); $e_j$ : é o efeito do $j$-ésimo ambiente (j = 1; 2;...; e); $(ge)_{ij}$ : é o efeito da interação do $i$-ésimo genótipo com o $j$-ésimo ambiente; $\varepsilon_{ij}$ : é o erro experimental médio, assumido independente e $\varepsilon_{ij} \sim N(0; \sigma^2/r)$.

A forma básica de organização dos dados para análise da interação genótipos × ambientes está representada na Tabela 1, na qual estão representadas as médias dos $g$ tratamentos (genótipos) sobre $r$ repetições, derivadas de um experimento em delineamento apropriado. Quando todos os genótipos são avaliados em todos os ambientes, com o mesmo número de repetições por experimento, diz-se que os dados são balanceados.

A análise de variância feita sob o modelo (1) com base nos dados da Tabela 1 é bastante simples, sendo calculados os efeitos principais pelas marginais da Tabela 1 e a interação como desvios do modelo, em que $\widehat{ge}_{ij} = Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}$ são os correspondentes elementos da Tabela 1 para o estudo da interação. O erro experimental $\overline{Y}_{.j}$ é calculado para cada experimento, utilizando-se o resíduo na análise conjunta.

Tabela 1: Representação dos dados médios de $g$ genótipos avaliados em $e$ ambientes para um caráter genérico Y

| | Ambientes | | | | |
|---|---|---|---|---|---|
| Genótipo | 1 | 2 | ... | e | Média ($\overline{Y}_{i.}$) |
| 1 | $Y_{11}$ | $Y_{12}$ | ... | $Y_{1e}$ | ($\overline{Y}_{1.}$) |
| 2 | $Y_{21}$ | $Y_{22}$ | ... | $Y_{2e}$ | ($\overline{Y}_{2.}$) |
| | $\vdots$ | $\vdots$ | ... | $\vdots$ | $\vdots$ |
| g | $Y_{g1}$ | $Y_{g2}$ | ... | $Y_{ge}$ | ($\overline{Y}_{g.}$) |
| Médias ($\overline{Y}_{.j}$) | ($\overline{Y}_{.1}$) | ($\overline{Y}_{.2}$) | ... | ($\overline{Y}_{.e}$) | ($\overline{Y}_{..}$) |

Na solução do modelo (1) visando encontrar os estimadores dos parâmetros, pelo método de mínimos quadrados, admitindo-se as condições marginais:

$$\sum_{i=1}^{g} g_i = \sum_{j=1}^{e} e_j = \sum_{i=1}^{g} (ge)_{ij} = \sum_{j=1}^{e} (ge)_{ij} = \sum_{i=1}^{g}\sum_{j=1}^{e} (ge)_{ij} = 0.$$

Assim, a solução é:

$$\widehat{\mu} = \frac{\sum_{i,j} Y_{ij}}{ge} = \overline{Y}_{..}; \qquad \widehat{g}_i = \frac{\sum_{j} Y_{ij}}{e} - \widehat{\mu} = \overline{Y}_{i.} - \overline{Y}_{..}$$
$$\widehat{e}_j = \frac{\sum_{i} Y_{ij}}{g} - \widehat{\mu} = \overline{Y}_{.j} - \overline{Y}_{..}; \qquad \widehat{ge}_{ij} = Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}$$

Esses são estimadores não viesados dos parâmetros $\mu$, $g_i$, $e_j$ e $(ge)_{ij}$, sujeitos às condições marginais dadas acima. Se essas ou quaisquer outras condições não forem impostas sobre os parâmetros, então os estimadores dos parâmetros individuais são viesados.

A aproximação de mínimos quadrados $\hat{Y}_{ij}$ (ou valor predito) e seu respectivo resíduo, correspondente ao termo geral de interação $(\hat{ge})_{ij}$, são dados por: $\varepsilon_{ij} = Y_{ij} - \hat{Y}_{ij}$ em que $\hat{Y}_{ij} = \overline{Y}_{i.} + \overline{Y}_{.j} - \overline{Y}_{..}$ . Agora, pode-se construir a matriz de interações $GE_{(g\times e)} = \left[\widehat{ge}_{ij}\right]$:

$$GE_{(g\times e)} = \begin{bmatrix} \widehat{ge}_{11} & \widehat{ge}_{12} & \cdots & \widehat{ge}_{1e} \\ \widehat{ge}_{21} & \widehat{ge}_{22} & \cdots & \widehat{ge}_{2e} \\ \vdots & \vdots & \vdots & \vdots \\ \widehat{ge}_{g1} & \widehat{ge}_{g2} & \cdots & \widehat{ge}_{ge} \end{bmatrix}$$

As somas de quadrados devido aos efeitos principais e interação, referentes aos dados de médias como na Tabela 1, são obtidas por:

$$SQ_{Total} = \sum_{i=1}^{g}\sum_{j=1}^{e} Y_{ij}^2 - ge\overline{Y}_{..}^2, \text{ com } ge-1 \text{ graus de liberdade.}$$

$$SQ_G = e\sum_{i=1}^{g} \overline{Y}_{i.}^2 - ge\overline{Y}_{..}^2, \text{ com } g-1 \text{ graus de liberdade.}$$

$$SQ_E = g\sum_{j=1}^{e}\overline{Y}_{.j}^2 - ge\overline{Y}_{..}^2, \text{ com } e-1 \text{ graus de liberdade.}$$

$$SQ_{G\times E} = SQ_{Total} - SQ_G - SQ_E, \text{ com } (g-1)(e-1) \text{ graus de liberdade.}$$

Os quadrados médios correspondentes são obtidos dividindo-se cada soma de quadrados pelos respectivos graus de liberdade. O quadrado médio do resíduo ou do erro médio (*QMRes*) é calculado pela média ponderada dos $QM^{'}$s do resíduo (obtido nas ANOVA$^{'}$s individuais de experimentos), assumidos homogêneos, usando seus respectivos graus de liberdade como pesos, ou seja:

$$QM\text{Res} = \frac{\sum_j SQ_{\text{Res}_j}}{\sum_j GL_{\text{Res}_j}}$$

em que $j = 1,2,\ldots,e$ ambientes ou experimentos.

Uma consideração importante a ser feita para realização do teste $F$ e interpretação dos resultados da análise conjunta da variância, diz respeito à natureza fixa ou aleatória dos efeitos do modelo de análise (CHAVES, 2001). A natureza fixa ou aleatória da interação é determinada pelos efeitos principais. Se genótipos e ambientes são fixos, a interação será fixa. Se pelo menos um dos fatores for aleatório, a interação será aleatória. Na análise de ensaios multiambientais, consideram-se, em geral, os efeitos de genótipos como fixos e os efeitos de ambientes como aleatórios, de tal forma que o efeito da interação genótipos $\times$ ambientes é aleatório nesse caso.

O estudo da interação genótipos $\times$ ambientes possibilita a identificação de cultivares mais adaptados a determinadas regiões, onde as mesmas poderão expressar o seu potencial genético. Assim, estudos sobre a magnitude de tais interações podem ser úteis na regionalização de cultivares, objetivando indicar áreas onde as mesmas possam expressar o máximo que as condições ambientais particulares permitam, com respeito a respostas de genótipos, e possibilitar a exploração de efeitos específicos de adaptação para determinadas regiões.

No que diz respeito à adaptabilidade e estabilidade de cada genótipo, tais fenômenos não devem ser considerados iguais, apesar de estarem relacionados entre si. A adaptabilidade refere-se à capacidade de os genótipos aproveitarem vantajosamente o estímulo do ambiente e a estabilidade diz respeito à capacidade de os genótipos mostrarem comportamento altamente previsível em razão do estímulo do ambiente.

## O Modelo de efeitos principais aditivos e interação multiplicativa

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

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

- Quando toda a interação estiver em uma só linha (ou coluna). Isto pode ocorrer quando houver vários dados discrepantes na mesma linha (ou coluna).

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

O modelo AMMI combina dois procedimentos estatísticos: análise da variância e a decomposição por valores singulares. Em um único modelo têm-se componentes aditivos para os efeitos principais (linhas ou genótipos e colunas ou ambientes), e componentes multiplicativos para os efeitos da interação. Duarte e Vencovsky (1999) explicam que os efeitos principais, na parte aditiva (média, efeitos genotípicos e ambientais), são ajustados por uma análise de variância comum (univariada) aplicada à matriz de dados, resultando em um resíduo de não aditividade, isto é, na interação ($G \times E$), e essa interação, constituinte da parte multiplicativa do modelo, é, depois, analisada pela decomposição por valores singulares da matriz de resíduos ou interação.

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

$$y_{ij} = \mu + g_i + e_j + \lambda_1 \alpha_{1i} \gamma_{1j} + \lambda_2 \alpha_{2i} \gamma_{2j} + \ldots + \lambda_k \alpha_{ki} \gamma_{kj} \quad (2)$$

em que $\mu$ representa a média geral, $g_i$ é o efeito do i-ésimo genótipo (com $i = 1, 2, \ldots, g$), $e_j$ é o efeito do j-ésimo ambiente (com $j = 1, 2, \ldots, e$), $k$ = posto($\mathbf{GE}$) = $min(g-1, e-1)$, $\lambda_r$ (com $r = 1, \ldots, k$) é a raiz quadrada do r-ésimo autovalor das matrizes $(\mathbf{GE})(\mathbf{GE}^T)$ e $(\mathbf{GE}^T)(\mathbf{GE})$ de iguais autovalores não nulos, $\alpha_{ri}$ é o i-ésimo elemento (relacionado ao genótipo i) do r-ésimo autovetor de $(\mathbf{GE})(\mathbf{GE^T})$ associado a $\lambda_r^2$, $\gamma_{rj}$ é o j-ésimo elemento (relacionado ao ambiente j) do r-ésimo autovetor de $(\mathbf{GE})^T)(\mathbf{GE})$ associado a $\lambda_r^2$, com $\sum_i \alpha_{ri}^2 = \sum_j \gamma_{rj}^2 = 1$, para $r = 1, 2, \ldots, k$ e $\sum_i \alpha_{ri} \alpha_{r*i} = \sum_i \gamma_{rj} \gamma_{r*i} = 1$, para $r \neq r^* = 1, 2, \ldots, k$.

Dependendo do número de componentes multiplicativos o modelo (2) é notado por AMMI0, AMMI1 ou AMMI$k$ de forma genérica.

Além dos erros independentes, se eles tiverem distribuição normal, então os estimadores descritos são também os estimadores de máxima verossimilhança. Em experimentos $G \times E$, $y_{ij}$ representa a resposta do i-ésimo genótipo no j-ésimo ambiente, $\mu$ é a média geral,$\lambda_r^2$ fornece a proporção da variância devido à interação $G \times E$ no r-ésimo componente e $\alpha_{ri}$, $\gamma_{rj}$ representam os pesos para o genótipo $i$ e ambiente $j$ naquele componente de interação.

## Escolha do número apropriado de termos para descrever a interação

Na literatura existem dois tipos mais utilizados de procedimentos para determinar o número ótimo de termos no modelo AMMI (o valor de $k$ no modelo (2)). Um desses procedimentos consiste em fazer testes de significância dos termos multiplicativos e o outro procedimento consiste em fazer validação cruzada. Na validação cruzada os dados de repetições, para cada combinação de tratamentos são aleatoriamente divididos em dois subconjuntos, um subconjunto de dados para o ajuste do modelo e outro subconjunto para validação. As respostas preditas por um determinado modelo AMMI, são confrontadas com os respectivos dados de validação, calculando-se as diferenças entre esses valores. Obtém-se, em seguida a soma de quadrados dessas diferenças, dividindo-se o resultado pelo número delas. A raiz quadrada

desse resultado chama-se diferença preditiva média. Esse método foi estudado com mais detalhes em Dias (2005).

Os testes de hipóteses são aplicados usando os dados completos e os critérios adotados para a determinação do número de componentes multiplicativos tem sido objeto de várias pesquisas.Alguns resultados dos resulatdos obtidos são: Gollob (1968), Mandel (1971), Gauch (1988) Gauch e Zobel (1996), Milliken e Johnson (1989), Piepho (1995), Cornelius et al. (1996), Dias e Krzanowski (2003), Dias (2005) e Dias e Krzanowski (2006). A seguir são descritos alguns destes testes:

**1) Testes de razão de verossimilhança**

Se um pesquisador deseja usar o modelo dado em (2), então é preciso determinar o número de termos de interação multiplicativa necessário para que o modelo explique adequadamente os dados. Para tomar essa decisão, têm-se problemas muito parecidos aos problemas encontrados na construção dos modelos de regressão. O objetivo é encontrar um modelo parcimonioso (com poucos termos quanto seja possível), e ao mesmo tempo, obter um modelo adequado. Deve-se lembrar que em situações de regressão polinomial, sempre é possível ajustar um modelo de grau $(n-1)$ a $n$ observações, mas, tais modelos não são geralmente bons, pois funcionam bem na predição da resposta média dos valores observados, mas, podem funcionar muito mal na predição de respostas de valores não observados.

Uma situação similar se apresenta para os dados provenientes de uma estrutura de tratamentos em dupla entrada, na qual sempre é possível fazer $k = min(g-1; e-1)$ e ajustar exatamente esses dados com o modelo (2), mas, tal modelo não será provavelmente muito bom por causa do sobreajuste dos dados e pelo fato dele explicar além do padrão de resposta presente nos dados parte do erro de medida. Assim, é desejável um modelo com poucos componentes que ofereça um ajuste ótimo e que explique boa parte do padrão de resposta dos dados. Assumindo por enquanto que se conhece o procedimento necessário para testar as hipóteses, um procedimento razoável pode ser:

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

$$y_{ij} = \mu + \tau_i + \beta_j + \lambda_1 \alpha_{1i} \gamma_{1j} + \varepsilon_{ij} \tag{3}$$

(ii) Se não rejeitar $H_{01}$, conclui-se que os dados são aditivos e deve-se completar uma análise correspondente com o resultado. No entanto, se rejeita-se $H_{01}$, então se deve testar
$H_{02} : \lambda_2 = 0 \; ; \lambda_1 \neq 0$ vs: $H_{a2} : \lambda_2 \neq 0 \; ; \lambda_1 \neq 0$
no modelo $y_{ij} = \mu + \tau_i + \beta_j + \lambda_1 \alpha_{1i} \gamma_{1j} + \lambda_2 \alpha_{2i} \gamma_{2j} + \varepsilon_{ij}$

(iii) Continuar dessa forma sucessivamente até não rejeitar a hipótese $H_{01}$ .

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

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

$$U_1 = \frac{\lambda_1}{\sum\limits_{i,j} \widehat{g} e_{ij}^2} = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} > C_\alpha$$

em que $C_\alpha$ é o ponto crítico a $\alpha(100\%)$ obtido da tabela com os pontos críticos da distribuição de $U_1$ e apresentada em Milliken e Johnson (1989), com $p = min(g-1; e-1)$ e $n = max(g-1; e-1)$.

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

$$U_2 = \frac{\lambda_2}{\lambda_2 + \lambda_3 + \ldots + \lambda_p}.$$

Rejeita-se $H_{02}$ se $U_2$ for maior do que o valor crítico (ver tabela em Milliken e Johnson (1989)).

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

$$U_k = \frac{\lambda_k}{\lambda_k + \lambda_k + 1 + \ldots + \lambda_p}$$

Milliken e Johnson (1989) sugerem nestes casos usar os pontos críticos da distribuição de $U_1$ com $p = min(g,e) - k$ e $n = max(g,e) - k$, mas, para aqueles experimentos nos quais não existe na tabela o correspondente ponto crítico (grande número de genótipos ou ambientes) Cornelius et al. (1996) apresentam uma transformação da estatística para obter um teste F aproximado.

**2) Teste $F_R$**

Considere o modelo (2) aplicado em um experimento para avaliar genótipos e ambientes. Escrevendo a soma de quadrados da interação $(G \times E)$ tem-se que

$$SQ(G \times E) = \sum_i \sum_j \left(\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2 \text{ com } (g-1)(e-1) \text{graus de liberdade.}$$

Essa soma de quadrados pode ser escrita como:

$$SQ(G \times E) = \sum_{r=1}^{p} {\lambda_r}^2$$

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

$$SQ(G \times E) = \sum_{r=1}^{p} {\lambda_r}^2 = \sum_{r=1}^{k} {\lambda_r}^2 + \sum_{r=k+1}^{p} {\lambda_r}^2 = SQ(G \times E)_{PADRÃO} + SQ(G \times E)_{RUÍDO}.$$

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

$$F_{R,K} = \frac{\left(SQ(G \times E) - \sum_{r=1}^{k} \lambda_r^2\right)}{f_2 \times QM\,(Erro\ médio)}$$

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

Apresenta-se na tabela 2 a análise da variância a partir de médias segundo o sistema de Cornelius, em que n representa o número de repetições no experimento e $IPCA_k$ é a notação internacional para o i-ésimo componente da interação.

Tabela 2: Esquema da Análise de variância pelo sistema de Cornélius baseado em médias

| Fontes de variação | Graus de liberdade | Soma de quadrados |
|---|---|---|
| Genótipos (G) | $g-1$ | SQE |
| Ambientes (E) | $e-1$ | SQG |
| Interação (G × E) | $(g-1)(e-1)$ | SQ(G×E) |
| IPCA 1 | $(g-1-1)(e-1-1)$ | $\sum_{k=2}^{p}\lambda_k^2$ |
| IPCA 2 | $(g-1-2)(e-1-2)$ | $\sum_{k=3}^{p}\lambda_k^2$ |
| IPCA 3 | $(g-1-3)(e-1-3)$ | $\sum_{k=4}^{p}\lambda_k^2$ |
| ... | ... | ... |
| IPCA k | $(g-1-k)(e-1-k)$ | $\sum_{k=k+1}^{p}\lambda_k^2$ |
| ... | ... | ... |
| IPCA p | $(g-1-p)(e-1-p)$ | - |
| Erro médio | e(g-1)(r-1) | - |
| Total | $egr-1$ | - |

IPCA k: (Interaction Principal Components Analysis) modelo com k componentes k = 1,2,...p.

**3) Teste de Gollob**

O teste de Gollob (1968) distribui graus de liberdade às Somas de Quadrados $SQ_k$=$\lambda_k^2$ com $k = 1,2,\ldots,p$ e $r$ o número de repetições, contando o número de parâmetros no k-ésimo termo multiplicativo. Logo, o teste F é calculado como na análise de variância para modelos lineares e supõe sob a hipótese nula que, o numerador e o denominador da estatística F são distribuídos independentemente como uma variável qui-quadrado (CORNELIUS et al. 1996). Um resumo deste teste é apresentado na Tabela 3.

Tabela 3: Esquema da Análise de variância pelo sistema de Gollob baseado em médias

| Fontes de variação | Graus de liberdade | Soma de quadrados |
|---|---|---|
| Genótipos (G) | $g-1$ | SQE |
| Ambientes (E) | $e-1$ | SQG |
| Interação (G × E) | $(g-1)(e-1)$ | SQ(G×E) |
| IPCA 1 | $g+e-1-2\times 1$ | $\lambda_1^2$ |
| IPCA 2 | $g+e-1-2\times 2$ | $\lambda_2^2$ |
| IPCA 3 | $g+e-1-2\times 3$ | $\lambda_3^2$ |
| ... | ... | ... |
| IPCA k | $g+e-1-2\times k$ | $\lambda_k^2$ |
| ... | ... | ... |
| IPCA p | $g+e-1-2\times p$ | $\lambda_p^2$ |
| Erro médio | e(g-1)(r-1) | - |
| Total | $egr-1$ | - |

O teste F de Gollob não é válido porque os autovalores $\lambda_k^2$ são distribuídos como autovalores de uma matriz de Wishart e, portanto não tem distribuição qui-quadrado, além disso, o teste assume que $n\lambda_k^2/\sigma^2$ é distribuído como qui-quadrado e então obviamente não é válido (Dias, 2005).

No que se refere aos graus de liberdade, o método de Gollob é muito popular, pois o procedimento é fácil de aplicar, uma vez 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 (Dias, 2005).

**4) Avaliação preditiva por validação cruzada**

Em geral, é necessário o uso de procedimentos estatísticos computacionalmente intensivos para fazer predições, daí a importância que tem ultimamente os métodos livres de distribuições teóricas como os baseados em reamostragem jacknife, bootstrap e validação cruzada. O critério preditivo de avaliação pri-

oriza a capacidade de um modelo aproximar suas predições a dados não incluídos na análise (simulando respostas futuras ainda não mensuradas).

Um modelo que seletivamente recupera o padrão e relega ruídos a um resíduo desconsiderado na predição de respostas, pode resultar em melhor precisão do que os próprios dados. Esse é o princípio subjacente à proposta de Gauch (1988) para seleção do modelo AMMI introduzida por ele como *avaliação preditiva*.

Dessa forma, através da validação cruzada, os dados de repetições, para cada combinação de genótipos e ambientes, são divididos, por um critério aleatório, em dois subconjuntos: $(i)$ dados para o ajuste do modelo AMMI; e $(ii)$ dados de validação. As respostas preditas do modelo AMMI, são comparadas com os dados de validação, calculando-se as diferenças entre esses valores. Logo, é obtida a soma de quadrados dessas diferenças e o resultado dividido pelo número de respostas preditas. À raiz quadrada desse resultado é chamado de diferença preditiva média (RMSPD), Crossa et al. (1991) sugerem que o procedimento deve ser repetido 10 vezes, obtendo-se uma média dos resultados para cada membro da família de modelos.

Um pequeno valor de RMSPD indica sucesso preditivo do modelo, tal que o melhor modelo é aquele com o menor RMSPD. O modelo selecionado é então usado para para analisar os dados de todas as $m$ repetições, conjuntamente, em uma análise definitiva (DIAS, 2005).

Outros autores como Piepho (1994) sugere que o valor médio de RMSPD seja obtido a partir de 1000 randomizações diferentes e não 10 como propôs Crossa et al. (1991). O autor considera uma modificação da partição completamente aleatória dos dados (modelagem e validação) quando o ensaio é em blocos. Neste caso, ele recomenda sortear o bloco inteiro de um ensaio e não fazer componentes para cada combinação de genótipo e ambiente. Assim, a estrutura original de blocos é preservada. Contudo, apesar da coerência lógica desse tipo de proposta, estudos confirmando sua efetividade ainda não estão disponíveis. Gauch e Zobel (1996) sugerem que se faça o conjunto de dados de validação sempre com uma só observação para cada tratamento. Sendo assim, é mais provável para $m-1$ dados, encontrar um modelo que mais se aproxime do ideal para analisar o conjunto completo dos $m$ dados por tratamento.

Segundo Duarte e Vencovsky (1999), ao avaliar o modelo por validação cruzada, a análise AMMI deve partir das observações individuais propriamente ditas (dados de cada repetição dentro de experimentos). Por outro lado, se o modelo for avaliado por um teste $F$ a análise pode ser feita a partir das médias dos genótipos nos ambientes (experimentos), desde que se disponha dos quadrados médios residuais, obtidos nas análises de variâncias de cada experimento.

Dias e Krzanowski (2003) descrevem dois métodos que otimizam o processo de validação cruzada por validar o ajuste do modelo em cada um dos dados por vez e então combinar essa validação em uma medida simples e geral de ajuste.

**5) Método "leave-one-out"**

Dias e Krzanowski (2003) propuseram dois métodos baseados em um procedimento "leave-one-out"completo, que otimiza o processo de validação cruzada. No que segue, assume-se que se deseja predizer os elementos $x_{ij}$ $(ge_{ij})$ da matriz $\boldsymbol{X}$ $(\boldsymbol{GE})$por meio do modelo:

$$x_{ij} = \sum_{k=1}^{n} d_k u_{ik} v_{jk} + \varepsilon_{ij}$$

Os métodos são aqueles apresentados em Krzanowski (1987) e Gabriel (2002), no qual prediz-se o valor $\widehat{x}_{ij}^{n}$ de $x_{ij}$ $(i = 1,\ldots,g;\ j = 1,\ldots,e)$ para cada possível escolha de $n$ (o número de componentes), e a medida de discrepância entre o valor atual e predito como

$$PRESS(n) = \sum_{i=1}^{g}\sum_{j=1}^{e}(x_{ij}^{n} - x_{ij})^2$$

Contudo, para evitar viés, os dados $x_{ij}$ não devem ser usados nos cálculos de $x_{ij}^{n}$ para cada $i$ e $j$. Como conseqüência, apelo a alguma forma de validação cruzada é indicado, e os dois procedimentos

diferem na forma com que eles lidam com isso (DIAS, 2005). Ambos, entretanto, assumem que a DVS de $\boldsymbol{X}$ pode ser escrita como $\boldsymbol{X} = \boldsymbol{UDV^T}$.

O procedimento de validação cruzada padrão subdivide $\boldsymbol{X}$ em um certo número de grupos, deleta-se cada grupo por vez a partir dos dados, avalia-se os parâmetros do modelo ajustados a partir dos dados remanescentes, e prediz-se o valor deletado (WOLD, 1976, 1978). Krzanowski (1987) argumenta que a predição mais precisa resulta quando cada grupo deletado é tão pequeno quanto possível, que no presente caso é um simples elemento de $\boldsymbol{X}$. Denota-se por $\boldsymbol{X}^{(-i)}$ o resultado de deletar a $i$-ésima linha de $\boldsymbol{X}$ e centralizar em torno das médias das colunas. Denota-se por $\boldsymbol{X}_{(-j)}$ o resultado de deletar a $j$-ésima coluna de $\boldsymbol{X}$ e centralizar em torno das médias das colunas, seguindo o esquema dado por Eastment e Krzanowski (1982). Então pode-se escrever:

$$\boldsymbol{X}_{(-i)} = \overline{U}\,\overline{D}\overline{V}^T \text{ com } \overline{U} = (\overline{u}_{pt}), \overline{V} = (\overline{v}_{pt}) \text{ e } \overline{D} = diag(\overline{d}_1, \ldots, \overline{d}_l),$$
$$\boldsymbol{X}_{(-j)} = \widetilde{U}\widetilde{D}\widetilde{V}^T \text{ com } \widetilde{\boldsymbol{U}} = (\widetilde{u}_{pt}), \widetilde{\boldsymbol{V}} = (\widetilde{v}_{pt}) \text{ e } \widetilde{\boldsymbol{D}} = diag(\widetilde{d}_1, \ldots, \widetilde{d}_{l-1}).$$

Agora, considere-se o preditor

$$\widehat{x}_{ij}^n = \sum_{t=1}^{n} \left(\widetilde{u}_{it}\sqrt{\widetilde{d}_t}\right)\left(\overline{v}_{tj}\sqrt{\overline{d}_t}\right) \tag{4}$$

Cada elemento no lado direito da equação (4) é obtido da DVS de $\boldsymbol{X}$ centrada na média após omitir a $i$-ésima linha e a $j$-ésima coluna. Assim, o valor $x_{ij}$ não é usado no cálculo da predição, e o máximo uso dos dados é feito com os outros elementos de $\boldsymbol{X}$. Os cálculos aqui são exatos, assim não há problema com a convergência como nos procedimentos de maximização que têm sido aplicados ao modelo AMMI, mas que não garantem a convergência (DIAS; KRZANOWSKI, 2003).

Gabriel (2002), tomou uma mistura de regressão e aproximação de uma matriz de posto inferior como a base para sua predição. O algoritmo para validação cruzada de aproximações de posto inferior proposto pelo autor é como segue: Para a matrix $\boldsymbol{X}$, usa-se a partição

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \boldsymbol{X}_{1\cdot}^T \\ \boldsymbol{X}_{\cdot 1} & \boldsymbol{X}_{|11} \end{bmatrix}$$

e o ajuste aproximado da sub-matriz $\boldsymbol{X}_{|11}$ de posto $n$ usando a DVS é:

$$\boldsymbol{X}_{|11} = \sum_{k=1}^{n} \boldsymbol{u}_{(k)} d_k \boldsymbol{v}_{(k)}^T = \boldsymbol{UDV^T}.$$

Então prediz-se $x_{11}$ por $\widehat{x}_{11} = \boldsymbol{X}_{1\cdot}^T \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^T\boldsymbol{X}_{\cdot 1}$ e obtém-se o resíduo de validação cruzada $e_{11} = x_{11} - \widehat{x}_{11}$.

Similarmente, obtém-se o valor ajustado da validação cruzada $x_{ij}$ e os resíduos $e_{ij} = x_{ij} - \widehat{x}_{ij}$ para todos os outros elementos $x_{ij}$, $i = 1, \ldots, g$; $j = 1, \ldots, n$; $(i, j) \neq (1, 1)$. Cada um irá requerer uma partição diferente de $\boldsymbol{X}$.

Esses resíduos e valores ajustados podem ser sumarizados por

$$PRESS(n) = \frac{1}{ge}\sum_{i=1}^{g}\sum_{j=1}^{e} e_{ij}^2 \quad \text{e} \quad PRECORR(n) = Corr(x_{ij}, \widehat{x}_{ij} | \forall i, j),$$

respectivamente.

Com cada método, a escolha de $n$ pode ser baseada em alguma função apropriada de

$$PRESS(n) = \frac{1}{ge}\sum_{i=1}^{g}\sum_{j=1}^{e} (x_{ij}^n - x_{ij})^2$$

Contudo, as características dessa estatística diferem para os dois métodos. O procedimento de Gabriel produz valores que primeiro decresce e então (usualmente) cresce com $n$. Por essa razão ele sugere

que o valor ótimo de $n$ seja aquele que produz o mínimo da função PRESS. O procedimento de Eastment-Krzanowski produz, geralmente, um conjunto de valores que é monotonicamente não-crescente com $n$ (DIAS, 2005). Por isso, sugerem o uso de

$$W_n = \frac{\frac{PRESS(n-1)-PRESS(n)}{D_n}}{\frac{PRESS(n)}{D_r}}$$

em que $D_n$ é o número de graus de liberdade requeridos para ajustar o $n$-ésimo componente e $D_r$ é o número de graus de liberdade remanescentes após ajustar o $n$-ésimo componente. Considerações sobre o número de parâmetros a serem estimados juntos com todas as restrições nos autovetores em cada estágio, mostra que $D_n = g + e - 2n$. $D_r$ pode ser obtido por sucessivas subtrações, dando $(g-1)e$ graus de liberdade na matriz centrada na média $\boldsymbol{X}$, isto é, $D_1 = (g-1)e$ e $D_r = D_{r-1} - [g+e-(n-1)2]$, $r = 2,3,\ldots,(g-1)$, (Wold, 1978). $W_n$ representa o aumento na informação preditiva suprida pelo $n$-ésimo componente, dividido pela informação preditiva média em cada um dos componentes remanescentes. Assim, importantes componentes devem produzir valores de $W_n$ maiores que a unidade. Baseando-se a escolha de $n$ em $W_n$ pode ser vista como uma natural seleção de um melhor conjunto de variáveis regressoras ortogonais em análise de regressão múltipla (DIAS; KRAZANOWSKI, 2003).

Cornelius et al. (1993) compararam resultados de validação cruzada com aqueles obtidos após calcular a estatística PRESS nos modelos multiplicativos em dados MET (*MultiEnvironment Trials*) completos. A partição dos dados envolveu três repetições para modelagem e uma repetição para validação. Calcularam o RMSPD da estatística PRESS ajustando os valores de PRESS como $[PRESS/ge + 3s^2/4]^{\frac{1}{2}}$, em que $g$ e $e$ denotam o número de genótipos e ambientes no MET e $s^2$ é a variância residual conjunta dentro de ambientes. O termo em $s^2$ é um ajuste para a diferença em variância da validação dos dados nas médias de caselas, para tomar os resultados comparáveis ao RMSPD da divisão $3-1$ dos dados. Resultados em um MET com nove genótipos e vinte ambientes mostrou que PRESS é mais sensível a super ajuste do que os dados divididos (DIAS, 2005).

## Biplot

Muitos estudos observacionais ou experimentais produzem uma tabela de dupla entrada de dados a ser analisada. A origem mais comum de tais dados é de um experimento de dois fatores; se um fator tem $g$ níveis, o segundo tem $e$ níveis, e há $r$ observações repetidas em cada combinação de níveis de fator. O conceito de biplot foi desenvolvido por Gabriel (1971) como uma representação gráfica que apresenta ambas as entradas (por exemplo, cultivares) e os testadores (por exemplo, ambientes) de um conjunto de dados em uma tabela de dupla entrada. O biplot permite a visualização dos dados conforme as seguintes propriedades: a) inter-relação entre as entradas (por exemplo, genótipos); b) inter-relação entre os testadores (por exemplo, ambientes); c) inter-relação entre as entradas e os testadores. Trata-se de uma representação gráfica da informação em uma matriz $g \times e$. O "b" refere-se aos dois tipos de informações contidas em uma matriz de dados: as informações nas linhas pertencem a amostras ou unidades amostrais e aquelas nas colunas pertencem as variáveis. Esta representação gráfica permite a inspeção visual da posição de uma unidade amostral relativa à outra e a importância relativa de cada uma das variáveis à posição de qualquer unidade. Assim pode-se ver como as unidades amostrais se agrupam e quais variáveis contribuem para sua posição dentro dessa representação.

A análise Biplot é utilizada com quaisquer tipos de variáveis (contínuas ou discretas), quando a finalidade é aproximar os dados originais e realizar uma análise simultânea das relações entre indivíduos e/ou variáveis. A fundamentação teórica se baseia na aproximação da matriz de dados $\mathbf{X}$ , de ordem $(g \times e)$ de posto $p$ por uma matriz $Y$ de ordem $(g \times e)$ de posto $q$, em que $(q < p)$, por meio de sua DVS.

## Construção do Biplot - Gabriel (1971)

A construção de um biplot origina-se dos componentes principais amostrais. Seja $\mathbf{X}$ de ordem $(g \times e)$ e procura-se por uma aproximação $\mathbf{Y}$ de ordem $(g \times e)$ e posto 2 da matriz original $\mathbf{X}$. Essa aproximação de posto 2 de $\mathbf{X}$ é obtida pela decomposição singular de $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}' = \lambda_1\mathbf{u_1}\mathbf{v}'_1 + \lambda_2\mathbf{u_2}\mathbf{v}'_2 + \ldots + \lambda_p\mathbf{u_p}\mathbf{v}'_p = \sum_{i=1}^{p} \lambda_i\mathbf{u_i}\mathbf{v}'_i$$

com $\Lambda$=diag$(\lambda_1, \lambda_2, \ldots, \lambda_p)$ ; $\mathbf{U}$: matriz diagonal de autovetores de ${}_n\mathbf{X}_{pp}\mathbf{X}'_n$; $\mathbf{V}$: matriz ortogonal de autovetores de ${}_p\mathbf{X}'_{nn}\mathbf{X}_p$. sendo que $\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_p$

Então,

$$\mathbf{Y} = \lambda_1\mathbf{u_1}\mathbf{v}'_1 + \lambda_2\mathbf{u_2}\mathbf{v}'_2 = \begin{bmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \\ \vdots & \vdots \\ u_{1n} & u_{2n} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} & \ldots & v_{1p} \\ v_{21} & v_{22} & \ldots & v_{2p} \end{bmatrix}$$

Para obter o biplot, é necessário escrever $\mathbf{Y}$ como o produto de duas matrizes $\mathbf{GH}'$, em que $\mathbf{G}$ é uma matriz $(g \times 2)$ e $\mathbf{H}$ é uma matriz $(e \times 2)$. Isso pode ser feito de várias formas, mas (GABRIEL, 1971) sugere três fatorações bem simples:

$${}_n\mathbf{Y_p} = \begin{bmatrix} \mathbf{u}_{11}\sqrt{\lambda_1} & \mathbf{u}_{21}\sqrt{\lambda_2} \\ \mathbf{u}_{12}\sqrt{\lambda_1} & \mathbf{u}_{22}\sqrt{\lambda_2} \\ \vdots & \vdots \\ \mathbf{u}_{1n}\sqrt{\lambda_1} & \mathbf{u}_{2n}\sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} v_{11}\sqrt{\lambda_1} & v_{21}\sqrt{\lambda_2} & \ldots & v_{1p}\sqrt{\lambda_1} \\ v_{21}\sqrt{\lambda_2} & v_{22}\sqrt{\lambda_2} & \ldots & v_{2p}\sqrt{\lambda_2} \end{bmatrix} \tag{5}$$

$\mathbf{G}_1 \qquad\qquad\qquad\qquad \mathbf{H}'_1$

$${}_n\mathbf{Y_p} = \begin{bmatrix} \mathbf{u}_{11}\lambda_1 & \mathbf{u}_{21}\lambda_2 \\ \mathbf{u}_{12}\lambda_1 & \mathbf{u}_{22}\lambda_2 \\ \vdots & \vdots \\ \mathbf{u}_{1n}\lambda_1 & \mathbf{u}_{2n}\lambda_2 \end{bmatrix} \begin{bmatrix} v_{11} & v_{21} & \ldots & v_{1p} \\ v_{21} & v_{22} & \ldots & v_{2p} \end{bmatrix} \tag{6}$$

$\mathbf{G}_2 \qquad\qquad\qquad\qquad \mathbf{H}'_2$

$${}_n\mathbf{Y_p} = \begin{bmatrix} \mathbf{u}_{11} & \mathbf{u}_{21} \\ \mathbf{u}_{12} & \mathbf{u}_{22} \\ \vdots & \vdots \\ \mathbf{u}_{1n} & \mathbf{u}_{2n} \end{bmatrix} \begin{bmatrix} v_{11}\lambda_1 & v_{21}\lambda_2 & \ldots & v_{1p}\lambda_1 \\ v_{21}\lambda_2 & v_{22}\lambda_2 & \ldots & v_{2p}\lambda_2 \end{bmatrix} \tag{7}$$

$\mathbf{G}_3 \qquad\qquad\qquad\qquad \mathbf{H}'_3$

A fatoração (5), corresponde a uma fatoração geral onde nenhuma ênfase é dada a linha ou coluna de $\mathbf{Y}$. A fatoração (6), coloca ênfase nas linhas de $\mathbf{Y}$. A fatoração (7), coloca ênfase nas colunas de $\mathbf{Y}$.

O biplot consiste em plotar os $(g+e)$ vetores $\mathbf{g}'_i(i = 1, 2, \ldots, g)$ e $\mathbf{h}'_j(j = 1, 2, \ldots, e)$ em um plano. Os vetores $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_g$ são chamados "efeitos ou marcas de linhas" de $\mathbf{Y}$, enquanto os vetores $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_e$ são chamados "efeitos ou marcas de colunas" de $\mathbf{Y}$. Veja uma ilustração na Figura 4.

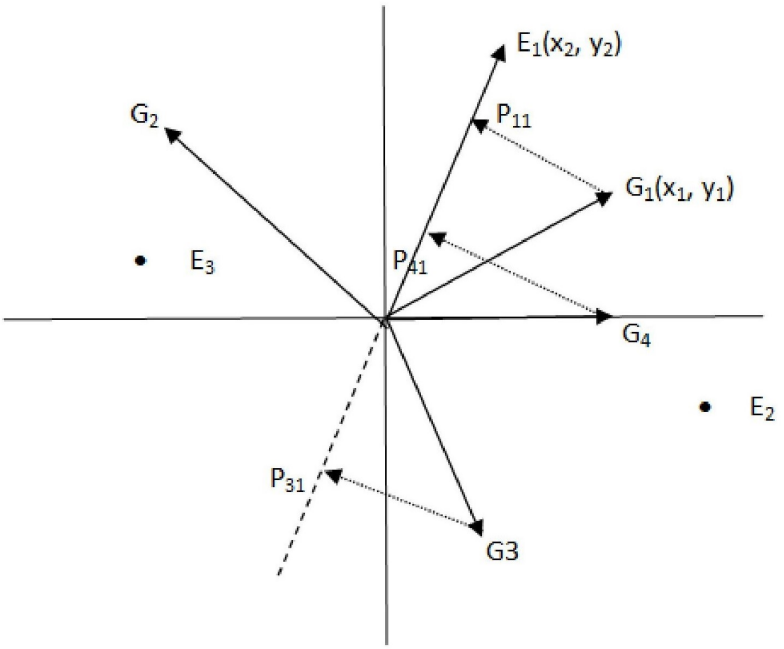


Figura 4: A geometria do biplot. Visualizando as projeções dos genótipos $G_1$, $G_2$, $G_3$ e $G_4$ sobre o ambiente $E_1$.

Cada elemento de $y_{ij}$ de $\mathbf{Y}$ é representado como produto interno $\mathbf{g}_i\mathbf{h}_j$ dos correspondentes vetores "efeito linha" e "efeito coluna". O comprimento da projeção de $\mathbf{g}_i$ sobre $\mathbf{h}_j$ é dado por:

$$\mathbf{L}_{P_{g_i/h_j}} = \frac{|\mathbf{g'}_i\mathbf{h}_j|}{\mathbf{L}_{\mathbf{h}_j}} \Rightarrow |\mathbf{g'}_i\mathbf{h}_j| = \begin{cases} \mathbf{L}_{\mathbf{h}_j}\mathbf{L}_{P_{g_i/h_j}} \\ \text{ou} = L_{\mathbf{g}_j}L_{\mathbf{h}_j}\cos(\theta) \\ \mathbf{L}_{\mathbf{g}_j}\mathbf{L}_{P_{h_i/g_j}} \end{cases}$$

O produto interno dos vetores $(\mathbf{g'}_i\mathbf{h}_j)$ pode ser visualizado como o produto do comprimento de um dos vetores vezes o comprimento da projeção do outro no primeiro. Isto é útil em permitir rápida avaliação visual da estrutura da matriz. Por exemplo, pode-se rapidamente ver quais linhas ou colunas são proporcionais a outras linhas ou colunas (mesma direção dos vetores correspondentes), e quais linhas ou colunas são zeros (ângulo reto entre efeitos linhas e colunas).

Se for de interesse que as relações entre linhas de $\mathbf{Y}$ sejam representadas pelas correspondentes relações dos vetores $\mathbf{g}$, as seguintes condições devem ser satisfeitas para quaisquer duas linhas $\mathbf{y'}_i$ e $\mathbf{y'}_j$ de $\mathbf{Y}$. Mas, $\mathbf{y'}_i\mathbf{y'}_j = \mathbf{g'}_i\mathbf{g'}_j$ equivale a $|\mathbf{y'}_i| = |\mathbf{g'}_j|$, isto é, $\mathbf{L}_{\mathbf{y}_i} = \mathbf{L}_{\mathbf{g}_i}$. Ainda, $|\mathbf{y}_i - \mathbf{y}_j| = |\mathbf{g}_i - \mathbf{g}_j|$ e $\cos(\mathbf{y}_i, \mathbf{y}_j) = \cos(\mathbf{g}_i, \mathbf{g}_j)$. A mesma ideia é válida para as colunas de $\mathbf{Y}$.

## Considerações Finais

Neste trabalho buscou-se de introduzir a metodologia AMMI de forma simplificada, mas a intenção é que sirva como uma luz para pesquisadores e estudantes ao nível de graduação e pós-graduação. Para isto faz- se um revisão sobre a interação genótipo $\times$ ambiente, define-se os modelos AMMI e alguns critérios de seleção e por fim gráfico biplot. Em Dias et al. (2013), é possivel obter mais detalhes sobre o assunto abordado neste trabalho e os seguintes tópicos: distribuição empírica dos autovalores da matriz de interação, utilizando reamostragem *bootstrap*; divergência genética utilizando reamostragem *bootstrap* e análise de agrupamento; Método de correção de autovalores viesados e a eficiência da correção; teste para confirmar a contribuição de genótipos e ambientes para a interação; imputação simples e múltipla para observações ausentes na matriz de interação; modelos AMMI bivariados utilizando a análise de

procrustes; modelo AMMI no estudo da interação entre QTL e ambiente; e generalização dos modelos AMMI para três fatores com uso dos modelos PARAFAC e TUCKER, gráficos *Joint plot* e *Triplot*.

## Agradecimentos

Os autores agradecem ao prof. Paulo Canas Rodrigues da Universidade Federal da Bahia e as agências finaciadoras de pesquisas: CAPES, CNPq e FAPEMIG, pelo suporte financeiro diversas pesquisas do grupo.

## Referências

ALLARD, R.W. *Princípios do melhoramento genético das plantas*. Rio de Janeiro: USAID/Edgard Blucher, 1971. 381p.

ALLARD, R.W. BRADSHAW, A.D. Implications of genotype-environmental interactions in applied plant breeding. *Crop Science*, Madison, v.4, n.5, p.503-508, 1964.

ARAÚJO, M.F.C.*Teste estatístico para contribuição de genótipos e ambientes na matriz de interação **GE***. 2008. 113p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2008.

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

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

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

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

CORNELIUS, P. L.; CROSSA J.; SEYEDSADR M. S. Tests and estimators of multiplicative models for variety trials. In: *Proceedings of Annual Kansas State University Conference on Applied Statistics in Agriculture*, 5th. Manhattan, KS. 25-27 April 1993. Dep. of Statistics, Kansas State Univ. Manhattan, KS. 1993. p.156-166.

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

CROSSA, J.; GAUCH, H. G.; ZOBEL, R. W. Additive main effects and multiplicative interaction analysis of two international maize cultivar trials, *Crop Science* 30(3), 493-500, 1990.

CROSSA, J.; FOX, P. N.; PFEIFER, W. H.; RAJARAM, S.; GAUCH, H. G. AMMI adjustment for statistical analysis of an international wheat yield trial, *Theoretical Applied of Genetics* 81, 27-37, 1991.

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

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

DIAS, C.T.S.; HONGYU, K.; ARAÚJO, L.B.; SILVA, M.J.C.; GARCÍA-PEÑA, M.; ARAÚJO, M.F.C.; RODRIGUES, P.C.; FARIA, P.N.; ARCINIEGAS-ALARCÓN, S. *Metodologia AMMI: Com Aplicação ao Melhoramento Genético*. In: 58ª Reunião da Região Brasileira da Sociedade Internacional de Biometria e 15º Simpósio de Estatística Aplicada à Experimentação Agronômica, 2013, Campina Grande-PB. Mini-Curso. 169p.

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

DIAS, C.T.S.; KRZANOWSKI, W.J. Choosing components in the additive main efect and multiplicative interaction (AMMI) models. *Scientia Agricola*, Piracicaba, v.63, n.2, p.169-175, 2006.

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

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

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

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

FISHER, R. A.; MACKENZIE, W. A. Studies in crop variation. II. The manurial response of different potato varieties. *Journal of Agricultural Science*, 13, 311-320, 1923.

FREEMAN, G.H. Statistical methods for the analysis of genotype-environment interactions. *Heredity*, 31, p.339-354, 1973.

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

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

GAUCH, H.G.: Model Selection and Validation for Yield Trials with Interaction. *Biometrics*, v. 44, p. 705–715, 1988.

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

GOLLOB, H.F. A statistical model which combines feature of factor analitic and analysis of variance techniques. *Psychometrika*, New York, v.33, p.73-115, 1968.

KANG, M.S. Using genotype-by-environment interaction for crop cultivar development. *Advances in Agronomy*, New York, v.62, p.199-252, 1998.

KANG, M.S.; MAGARI, R. New developments in selecting for phenotypic stability in crop breeding. In: KANG, M.S.; GAUCH, H.G. *Genotype-by-environment interaction*, Boca Raton: CRC Press, 1996. chap. 1. p. 1-14.

KRZANOWSKI, W. J. Cross-validation in principal component analysis, *Biometrics* 43, 575-584, 1987.

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

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

OLIVEIRA, A.B.; DUARTE, J.B.; PINHEIRO, J.B. Emprego da análise AMMI na avaliação da estabilidade produtiva em soja. *Pesquisa Agropecuária Brasileira*, Brasília, v.38, n.3, p.357-364, 2003.

PERKINS, J. M.; JINKS, J. L. Environmental and genotype-environmental components of variability. III. Multiple lines and crosses. *Heredity*, 23, p.339-356, 1968.

PIEPHO, H. P. Best linear unbiased prediction (BLUP) for regional yield trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis, *Theoretical and Applied Genetics*, New York, v.89, p.647-654,1994.

PIEPHO, H.P. Robustness of statistical test for multiplicative terms in the additive main efects and multiplicative interaction model for cultivar 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.

VENCOVSKY, R.; BARRIGA, P. *Genética biométrica no fitomelhoramento*. Ribeirão Preto: Sbg, 1992. 486p

YATES, F.; COCHRAN, W. G. The analysis of groups of experiments. *Journal of Agricultural Science*, 28, p.556-580, 1938.

WOLD, S. Pattern recognition by means of disjoint principal component models, *Pattern Recognition*, Great Britain, v.8, p.127-139, 1976.

WOLD, S. Cross-validatory estimation of the number of components in factor and principal component models, *Crop Science* 20, 397-405, 1978.