UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS AGRONÔMICAS
CÂMPUS DE BOTUCATU

UTILIZAÇÃO DE TÉCNICAS MULTIVARIADAS NA AVALIAÇÃO DA DIVERGÊNCIA GENÉTICA DE POPULAÇÕES DE GIRASSOL (*Helianthus annuus L.*)

ANA VERGÍNIA LIBOS MESSETTI

Tese apresentada à Faculdade de Ciências Agronômicas da Universidade Estadual Paulista - Câmpus de Botucatu, para obtenção do título de Doutor em Agronomia - Área de Concentração em Energia na Agricultura.

BOTUCATU-SP
Junho - 2007

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS AGRONÔMICAS
CÂMPUS DE BOTUCATU

UTILIZAÇÃO DE TÉCNICAS MULTIVARIADAS NA AVALIAÇÃO DA DIVERGÊNCIA GENÉTICA DE POPULAÇÕES DE GIRASSOL (*Helianthus annuus L.*)

ANA VERGÍNIA LIBOS MESSETTI

Orientador: Prof. Dr. Carlos Roberto Padovani

Tese apresentada à Faculdade de Ciências Agronômicas da Universidade Estadual Paulista - Câmpus de Botucatu, para obtenção do título de Doutor em Agronomia - Área de Concentração em Energia na Agricultura.

BOTUCATU- SP
Junho - 2007

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"

FACULDADE DE CIÊNCIAS AGRONÔMICAS

CAMPUS DE BOTUCATU

CERTIFICADO DE APROVAÇÃO

TÍTULO: **UTILIZAÇÃO DE TÉCNICAS MULTIVARIADAS NA AVALIAÇÃO DA DIVERGÊNCIA GENÉTICA DE POPULAÇÕES DE GIRASSOL (*Helianthus annuus* L.)**

ALUNA: ANA VERGÍNIA LIBOS MESSETTI

ORIENTADOR: PROF. DR. CARLOS ROBERTO PADOVANI

Aprovado pela Comissão Examinadora

PROF. DR. CARLOS ROBERTO PADOVANI

PROF. DR. ADRIANO WAGNER BALLARIN

PROF. DR. JOSÉ CARLOS MARTINEZ

PROFª DRª JACINTA LUDOVICO ZAMBOTI

PROFª DRª MARIE OSHIIWA

Data da Realização: 01 de junho de 2007.

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉCNICA DE AQUISIÇÃO E TRATAMENTO DA INFORMAÇÃO - SERVIÇO TÉCNICO DE BIBLIOTECA E DOCUMENTAÇÃO UNESP - FCA - LAGEADO - BOTUCATU (SP)

Messetti, Ana Verginia, 1964-
M584u Utilização de técnicas multivariadas na avaliação da divergência genética de populações de girassol / Ana Verginia Messetti. - Botucatu : [s.n.], 2007.
ix, 87 f. : il. color., gráfs, tabs.

Tese (Doutorado)-Universidade Estadual Paulista, Faculdade de Ciências Agronômicas, Botucatu, 2007
Orientador: Carlos Roberto Padovani
Inclui bibliografia

1. Girassol. 2. Análise multivariada. 4. Divergência genética. 4. Análise de variância. I. Padovani, Carlos Roberto. II. Universidade Estadual Paulista "Júlio de Mesquita Filho" (Campus de Botucatu). Faculdade de Ciências Agronômicas. III. Título.

Para ser sábio, é preciso primeiro temer ao Deus Eterno.
Ele dá compreensão aos que obedecem aos seus mandamentos.
Que Deus seja louvado para sempre! **Salmo 111: 10**

À Deus, meu eterno protetor.
Ao meu esposo Valter e meus queridos
filhos, Giulia (*in memorian*), Valter Luiz e Ana Clara
DEDICO

AGRADECIMENTOS

A Deus por ter me concedido força e coragem para enfrentar todas as dificuldades e concluir este trabalho.

Ao Prof. Dr. Carlos Roberto Padovani pela orientação, ensinamentos, incentivo e amizade.

Ao meu esposo Valter Luiz Messetti, pelo apoio e incentivo no decorrer do curso e aos meus filhos Giulia Libos Messetti (In Memorian), Valter Luiz Libos Messetti e Ana Clara Libos Messetti.

A toda minha família, que sempre me apoiou, incentivou nos momentos difíceis.

Aos amigos do Departamento de Matemática Aplicada e Estatística, pelo constante incentivo, amizade e apoio.

Aos amigos Jacinta, Rogério, Simone e Vanderli, pela agradável convivência no curso de pós-graduação.

Ao Eng. Agronômo Marcelo de Oliveira da EMBRAPA/Soja - Londrina por ter fornecido os dados de seus experimentos com girassol.

A Profª Elenice Pimentel pela revisão gramatical.

À Profª. Corina Maria Tedeschi Busnardo pela versão do resumo na língua inglesa.

A Ilza Almeida de Andrade e a Maria Aparecida Letrari pelos ajustes finais e correções das referências bibliográficas.

A todos aqueles que direta ou indiretamente colaboraram para a realização deste trabalho.

SUMÁRIO

Página

LISTA DE QUADROS

Página

LISTAS DE TABELAS

LISTAS DE FIGURAS

Página

UTILIZAÇÃO DE TÉCNICAS MULTIVARIADAS NA AVALIAÇÃO DA DIVERGÊNCIA GENÉTICA DE POPULAÇÕES DE GIRASSOL (*Helianthus annuus L.*). Botucatu, 2007. 87f. Tese (Doutorado em Agronomia – Área de Concentração em Energia na Agricultura) – Faculdade de Ciências Agronômicas, Universitade Estadual Paulista "Júlio de Mesquita Filho".

Autora: ANA VERGÍNIA LIBOS MESSETTI

Orientador: CARLOS ROBERTO PADOVANI

RESUMO

Este trabalho foi desenvolvido com os objetivos de avaliar a divergência genética de 12 populações de girassol do Banco de Germoplasma da EMBRAPA /Soja de Londrina por meio de técnicas multivariadas; divulgar tópicos recentes e interessantes das técnicas multivariadas que não são explorados nos trabalhos científicos de melhoramento de plantas e orientar a escolha de populações para cruzamentos nos programas de melhoramento genético da cultura de girassol. O modelo experimental constitui-se de delineamento bloco casualizado envolvendo 12 variedades de girassol avaliadas sob cinco caracteres morfoagronômicos. Por meio da análise univariada foi verificada diferença significativa ($p<0,05$) dos tratamentos para todos caracteres. A aplicação dos componentes principais permitiu a redução bidimensional, com a explicação de 82,5% da variação total. O número de componentes foi avaliado pelo critério de Kaiser e critério "Scree-test". A visualização da divergência genética proporcionada pelos escores das duas primeiras variáveis canônicas, evidenciaram grupos geneticamente diferentes. Ambas técnicas apontaram concordância nos resultados. Com base nas estimativas da distância Mahalanobis e distância euclideana foi realizada a análise de agrupamento adotando-se três algoritmos hierárquicos. Para determinar o número de grupos adotou-se o dendrograma, a análise do nível de fusão e a análise do comportamento de similaridade. Para validação utilizou-se o critério de Wilks dentro de cada grupo e gráficos multivariados auxiliaram na interpretação dos resultados. Pode-se concluir pela existência da divergência genética, detectando-se quatro grupos geneticamente diferentes e caracterizado pelos escores médios.

THE USE OF MULTIVARIATE TECHNIQUES IN THE EVALUATION OF GENETIC DIVERGENCE IN SUNFLOWER *(Helianthus annuus L.)* POPULATIONS. Botucatu, 2007. 87 pages. Thesis (PhD in Agronomy – Major Area: Agriculture Energy) - Faculdade de Ciências Agronômicas, Universidade Estadual Paulista "Júlio Mesquita Filho".

Author: ANA VERGINIA LIBOS MESSETTI

Advisor: CARLOS ROBERTO PADOVANI

SUMMARY

The objective of this work was to evaluate genetic divergence in 12 sunflower populations from EMBRAPA/ Londrina Soybean Germplasm Bank, using multivariate techniques, to discuss recent and interesting topics related to the multivariate techniques don´t found in plant improvement scientific papers, and to offer guidelines on how to choose populations for sunflower genetic improvement crossing programs. The experiment included a totally block casualized design, with twelve sunflower varieties, evaluated according to 5 morphoagronomics traits. The univariate analysis showed a significant difference ($p<0,05$) among treatments for all the traits. Application of main components allowed for a bi-dimensional reduction, with 82,5% of the total variation. The number of components were evaluated by the Kaiser and "Scree-test" criteria. Genetic divergence visualization provided by the two first canonical variables showed genetically different groups. Both techniques showed the same results. Based on Mahalanobis and Euclidean distance estimates, a clustering analysis was carried out using three hierarchicals algorithms. A dendrogram, a fusion level analysis and a similarity behavior analysis were conducted to determine the number of groups. Validation used the Wilks criteria inside each group, while multivariate graphs helped with data interpretation. Results from this study showed genetic divergence in four groups characterized by average/mean scores.

Keywords: genetic divergence, techniques multivariate, sunflowers

1 INTRODUÇÃO

O girassol é uma planta de uso diversificado, muito utilizado na alimentação humana e animal, na produção de combustível, na adubação verde em rotação de culturas, em floriculturas, entre outros. Devido a essa versatilidade, tem despertado interesse econômico em vários países, inclusive no Brasil. Para se ter idéia da necessidade de produção, o girassol responde por cerca de 13% de todo óleo vegetal produzido no mundo e vem aumentando o índice de crescimento de produção (UNITED STATES, 2003). A demanda mundial por óleo de girassol vem crescendo, em média, 1,8% ao ano. A demanda interna cresce em média 13% ao ano. Para suprir essa carência, o país importa o óleo principalmente da Argentina.

Por outro lado, o girassol está entre as espécies oleaginosas em estudo para viabilizar a produção de biodiesel no estado do Paraná. Órgãos governamentais estão empenhados em substituir o óleo diesel por óleo de origem vegetal, contribuindo para a redução do nível de poluição do ambiente, tornando-se assim uma importante alternativa de geração de renda para agricultura (GUERRA; PICKSIUS, 2005).

A cultura de girassol no Brasil necessita de um estudo mais aprofundado, no sentido de angariar informações que orientem os programas de melhoramento e superem os níveis de produtividade atuais. Um método utilizado pelos “melhoristas” é a avaliação da divergência genética, que pode ser estimada por meio de diversos caracteres de interesse da planta.

Os estudos brasileiros envolvendo girassol, até o presente momento descritos na literatura, na sua maioria, utilizaram-se de técnicas univariadas, não explorando a riqueza dos resultados das técnicas multivariadas, que direcionam melhoristas a concentrar esforços nas combinações mais promissoras, ou seja, nos materiais com maior divergência genética, maior heterose (aumento do vigor ou da fertilidade dos híbridos em relação às linhagens cauzadas para produzi-las), e consequentemente, maior produtividade.

As técnicas multivariadas podem ser utilizadas em estudos de divergência genética, tais como análise de componentes principais, análise de agrupamento, análise de variáveis canônicas, análise discriminantes, análise de variância multivariada, pois proporcionam enriquecimento das informações extraídas dos dados experimentais.

Na década de 80, a EMBRAPA/Soja de Londrina – PR traçou diretrizes de pesquisa para a cultura de girassol, em âmbito nacional, realizando e conduzindo ensaios em diversas regiões do país, para definir os cultivares apropriados por regiões, visando uma boa adaptação da planta quanto ao solo, clima, altitude, épocas de plantio, resistência às pragas e doenças e alta produtividade.

Diante desse contexto, este trabalho tem por objetivos: 1) avaliar a divergência genética de 12 populações de girassol do Banco de Germoplasma da EMBRAPA/Soja Londrina-PR, por meio de técnicas multivariadas denominadas análise de componentes principais, análise de agrupamento, análise de variáveis canônicas e análise de variância multivariada, envolvendo 5 caracteres morfoagronômicos das plantas; 2) divulgar tópicos recentes e interessantes dessas metodologias, que são pouco explorados nos trabalhos científicos de melhoramento de plantas e 3) comparação dos métodos multivariados para orientar a escolha de populações para cruzamentos nos programas de melhoramento genético da cultura de girassol (EMBRAPA).

2 REVISÃO DE LITERATURA

2.1 Histórico e Aspectos Fisiológicos do Girassol

O girassol cultivado, *Helianthus annuus* L., pertence à ordem Synandrales, família Compositae, subfamília Tubuliflorae, tribo Helianthease e gênero *Helianthus*. A denominação *Helianthus* vem do grego *hélios* significando sol e *anthos*, flor (ASTAFEIEF, 1997).

O desenvolvimento do girassol é uma seqüência de alterações morfológicas e fisiológicas na planta, denominada de fases fenológicas, separadas em dois estádios:

1. **Fase vegetativa** que caracteriza-se com início da emergência das plântulas e finaliza com o aparecimento da inflorescência (botão floral).
2. **Fase reprodutiva** que caracteriza-se com início do aparecimento da inflorescência e finaliza com a maturação da planta.

Para melhor esclarecimento das diferentes fases de desenvolvimento do girassol, as Figuras 1 e 2 apresentam um esquema baseado na descrição do desenvolvimento da planta definida por Schneiter e Miller (1981).

Semeadura		Botão floral	Início do florescimento	Final da floração	Maturação fisiológica	Colheita
↓		↓	↓	↓	↓	↓

Emergência	N° de folhas maiores que 4cm	Desenvolvimento da inflorescência	Floração	Enchimento de aquênios	Maturação fisiológica
VE	VI, V2, VN...	RI, R2, R3, R4	R5, R6	R7, R8	R9
Fase vegetativa			Fase reprodutiva		

Figura 1 – Descrição esquemática da fase vegetativa e reprodutiva do girassol.

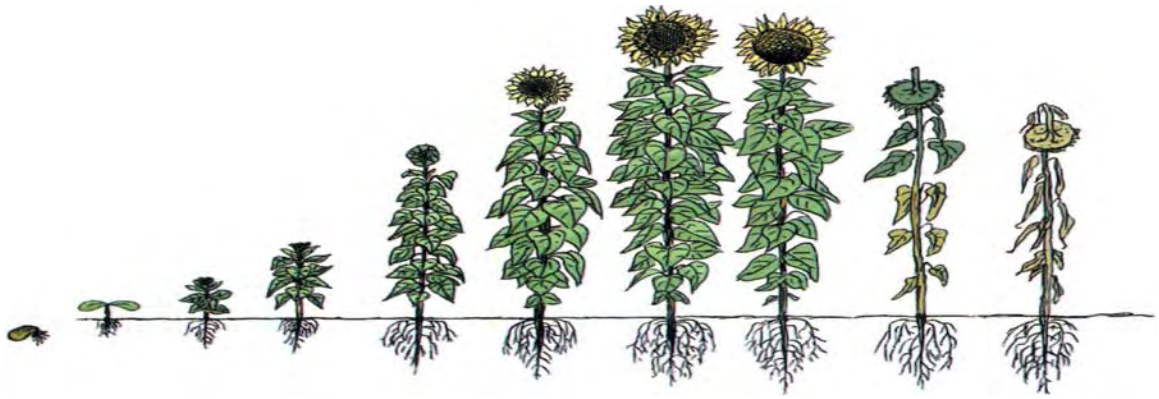


Figura 2 – Representação esquemática das fases de desenvolvimento do girassol.

Fase vegetativa:
VE: Emergência das plântulas, primeiro par de folhas menores que 4 centímetros. (Figura 3a)
V(N): Aparecimento de folhas verdadeiras e definidas pelo número de folhas, com o mínimo de 4 (quatro) centímetros (Figura 3b).

Fase reprodutiva:
Fase R1: A inflorescência circundada pela bráctea imatura torna-se visível.
Fase R2: O internódio abaixo da base da inflorescência alonga-se de 0,5 a 2 cm acima da folha
Fase R3: O internódio abaixo do botão reprodutivo alonga-se a mais de 2 cm acima da folha mais próxima da inflorescência (Figura 3c).
Fase R4: A inflorescência começa a abrir. Pequenas flores liguladas são visíveis e amarelas.
Fase R5: As flores liguladas estão completamente expandidas e todo disco das flores está

visível. É o início da antese (Figura 3d).

Fase R6: A antese está completa e as flores liguladas perderam a turgidez e estão murchando.

Fase R7: O dorso do capítulo torna-se amarelo claro.

Fase R8: O dorso do capítulo torna-se amarelo para castanho, porém as brácteas permanecem verdes.

Fase R9: As brácteas adquirem a coloração entre amarela e castanha. O dorso torna-se castanho. Ocorre a maturação fisiológica.

Fase VE (a) Fase VN (b) Fase R3 (c) Fase R5 (d)

Figura 3 – Detalhes das fases vegetativa e reprodutiva do girassol (LEITE, 2005).

O girassol teve como centro de origem a América do Norte. Evidências arqueológicas indicam que os índios norte-americanos já o cultivavam no Arizona e Novo México há cerca de 3000 a.C. Os primeiros contatos entre a América do Norte e a Europa ocorreram através da Espanha, onde exploradores espanhóis fizeram as primeiras introduções de girassol na Europa. A primeira documentação da presença da planta na Europa foi em meados de 1568, pelo herbarista Dodonaeus. O girassol era cultivado, até então, como planta ornamental, e a descoberta da utilização do girassol como planta oleaginosa ocorreu na Inglaterra em 1716, segundo Fick (1978), quando Artur Bunyan patenteou um processo de extração de óleo de girassol. O início da produção de óleo de girassol em escala comercial ocorreu em 1830, na Rússia, conforme citado em Skoric (1992). A Rússia tem sido o segundo maior produtor de óleo de girassol no mundo, e os méritos pela produção são creditados aos programas de melhoramento, que aumentaram a produtividade e a porcentagem de óleo nas sementes.

No Brasil, presume-se que cultivos esporádicos foram iniciados na região Sul, na época da colonização, quando os imigrantes europeus trouxeram hábitos

alimentares desta planta. A primeira referência ocorreu no ano 1924. Segundo Melo (1992), na década de 60 o Brasil possuía uma área de 360 ha com rendimento médio de 833 kg/ha. Até 1969 houve um aumento gradativo na área de plantio e no rendimento obtido, que passaram para 15300 ha e 1180 kg/ha, respectivamente. Na década de 70, foi constatado um período de declínio devido à falta de tecnologia de produção, à falta de organização da comercialização e às doenças fúngicas – altenária (*Alternaria helianthi*, Hausf.), esclerotínia (*Sclerotinia sclerotiorum*, Lib.), e ferrugem (*Puccinia helianthi* Schw.).

Nos últimos anos, a área de plantio do girassol aumentou, e a produção no Brasil passou de 27,5 mil ton na safra de 1998 para 115 mil ton na de 1999. O estado de Goiás é o principal responsável pelo aumento da área plantada (BUZZETTI, 1999).

Atualmente, a média da produção mundial é 26 milhões de ton/ano, sendo a Argentina o maior produtor e o maior exportador de óleo de girassol, seguido da Rússia, Estados Unidos, França, Romênia e Espanha.

O girassol cultivado está situado entre as quatro mais importantes culturas anuais de óleo comestível do mundo, precedido apenas pela soja, palma e canola. O óleo de girassol nos últimos anos teve bom crescimento no mercado consumidor devido a riqueza em ácidos graxos polinsaturados que atuam na prevenção de doenças cardiovasculares e no controle do nível de colesterol no sangue de acordo com Souza (2001).

No Brasil, o Instituto Agronômico de Campinas (IAC) e a EMBRAPA são as instituições mais citadas em programas de melhoramento genético. O girassol possui grande variabilidade genética, e o conhecimento dessa variabilidade permite o direcionamento das estratégias de melhoramento genético, visando a obtenção de materiais que apresentam diferentes características de interesse agronômico, tais como produtividade, resistência a doenças, controle de pragas, precocidade, altura reduzida das plantas e adaptação às diferentes regiões brasileiras.

2.2 Técnicas Multivariadas e a Divergência Genética: Aspectos gerias

Os estudos das técnicas de análise multivariada não são recentes. Como citado em Anderson (1958), Adrian, em 1808, estudou a função densidade da distribuição normal bivariada, com seguimento pelos pesquisadores Gauss (1823), Bravais (1846) e Laplace (1911).

As técnicas multivariadas estavam completamente construídas, na teoria, por volta da década de 30. Na obra de Reis (1997), a história da estatística multivariada divide-se em três momentos do século passado. No início do século, alguns pesquisadores destacados contribuíram grandemente para o fundamento teórico multivariado, entre estes, Pearson (1901), Fisher (1928), Wilks (1932), Hotteling (1933) e Bartlett (1937). Após a década de 50, são citadas outras obras consideradas como clássicas, destacando-se Rao (1952), Kendall (1957, 1975), Anderson (1958, 1984), Morrison (1967, 1976) e Mardia, Kent, Bibby (1979).

No momento atual, Chatfield e Collins (1980), Dillon e Goldstein (1984), Hair Jr. et al. (1987, 2005), Johnson e Wichern (1988) e Everitt e Dunn (1996), Mingotti (2005) são autores que visam as aplicações dos métodos multivariados e as interpretações dos resultados, mostrando a necessidade atual dos pesquisadores buscarem resultados mais precisos.

No século XXI nota-se que em todas as áreas houve avanços tecnológicos e conseqüentemente grande demanda pelo conhecimento, envolvendo técnicas mais sofisticadas e rigorosas para executar as análises multivariadas de dados.

Especificamente, nos estudos de divergência genética, Falconer (1981) comentou que há quatro maneiras de se avaliar a divergência: estudos genealógicos, diversidade ecogeográfica, análise dialélica e técnicas multivariadas. Por dispensarem a obtenção prévia das combinações híbridas (análise dialélica), os melhoristas exploram os métodos preditivos da divergência a partir de técnicas multivariadas. Atualmente, tem sido tema de muitos trabalhos envolvendo várias culturas de interesse agrícola.

A literatura de Cruz e Regazzi (1997) sobre modelos biométricos é muito citada devido a orientação teórica, aplicação e interpretação dos parâmetros

multivariados. O autor disponibilizou um material para orientação da aplicação dos métodos de análises de dados resultantes de ensaios de genética e melhoramento de plantas.

Segundo Duarte (1998), o estudo de divergência genética por técnicas multivariadas tem merecido grande ênfase por serem empregadas tanto em caracteres morfológicos e agronômicos, como em marcadores moleculares. Os marcadores moleculares surgiram com o advento das técnicas de biologia molecular e tem sido bastante úteis para estudos de genética e melhoramento. Entre os vários tipos de marcadores moleculares destacam-se o RFLP ("Restriction Fragment Length Polymorphism"), o RAPD ("Random Amplified Polymorphic DNA"), o SSRs ("Simple Sequence Repeat") e o AFLP ("Amplified Fragment Length Polymorphism").

O objetivo central dos melhoristas tem sido classificar os genótipos em grupos, facilitando a escolha de genitores para hibridações, tendo como base medidas estatísticas como a distância D^2 de Mahalanobis e a distância euclideana. Neste sentido, Moura (2003) apresentou os métodos mais explorados nos estudos de melhoramento: análise de variáveis canônicas, análise de componentes principais, análise de agrupamento e métodos aglomerativos (medidas de dissimilaridades). A escolha do método mais adequado tem sido determinada pela precisão desejada do pesquisador, pela facilidade da análise e pela forma como os dados foram obtidos.

Nos programas de melhoramento genético, diversos trabalhos utilizaram técnicas multivariadas nas diferentes culturas e regiões do Brasil. Especificamente, para a cultura de girassol as técnicas multivariadas são pouco exploradas. Mantêm-se, tradicionalmente, as técnicas univariadas como revistos nos 50 trabalhos publicados no simpósio nacional sobre a cultura de girassol realizado em Londrina – EMBRAPA (REUNIÃO..., 2005).

Camarano (1997) citou seu trabalho como o primeiro nacional a abordar a divergência genotípica entre populações de girassol através de técnicas multivariadas. Os experimentos foram instalados em Goiânia e Goianésia, estado de Goiás, e foram observados 11 caracteres de 10 populações distintas de girassol. A análise de agrupamento e a dispersão gráfica por variáveis canônicas formaram grupos homogêneos de acordo com as regiões de origem.

As técnicas multivariadas podem ser utilizadas no estudo da divergência genética, pois permitem combinar as múltiplas informações contidas na unidade experimental, de modo que seja possível executar uma seleção com base num complexo de variáveis, proporcionando ainda, enriquecimento das informações extraídas dos dados experimentais.

2.3 Análise de Agrupamento

2.3.1 Breve Histórico

Tryon (1932, “apud” LEITE, 2000) desenvolveu um procedimento chamado análise de agrupamento, que a partir dos trabalhos de Pearson (1901) e Spearman (1904) do início do século, foi aperfeiçoado por inúmeros autores os quais desenvolveram estudos visando à construção de um algoritmo denominado V-Análise, ou seja, análise de agrupamento de variáveis.

Sokal e Sneath (1963) contribuíram de forma grandiosa para essa metodologia no livro “Principles of Numerical Taxonomy”, voltado para área biológica. A partir dessa época houve uma revolução nos trabalhos científicos centrados em dois motivos. O primeiro deve-se ao avanço tecnológico: as técnicas, que na época eram consideradas inviáveis pela grande quantidade de cálculos, passaram a ser viáveis. O segundo diz respeito à importância da classificação biológica no meio científico (taxonomia numérica), em que ocorre a junção de várias informações sobre o mesmo indivíduo. No Brasil, nesta década, Cunha (1969) propôs uma configuração taxonômica do grupo de abelhas sociais sem ferrão os “Meliponinae”. Definiu 76 caracteres da morfologia externa que foram determinados sobre 55 espécies para investigar as relações de semelhança fenética existentes. Pisani (1973) investigou as repercussões dos acasalamentos recorrentes recíprocos sobre algumas variáveis associadas à produtividade comercial da ave.

Na década de 80, podem-se citar alguns autores que contribuíram de forma significante para a evolução da técnica, como Gama (1980), Van Laar (1987) e Johnson e Wichern (1988).

Entre autores nacionais, destacam-se Bussab, Miazaki e Andrade (1990), que detalharam as particularidades das técnicas e os algoritmos de agrupamento com aplicações numéricas simples aos principiantes. Descrevem também os principais aplicativos computacionais para a utilização dessa análise.

2.3.2 Metodologia

Hair Jr. et al. (2005) definiram a análise de agrupamento como uma técnica multivariada que tem por finalidade agrupar indivíduos em dois ou mais grupos com base na similaridade dos indivíduos em relação a um conjunto de caracteres que eles possuem. A técnica classifica indivíduos semelhantes de modo que os grupos reflitam elevada homogeneidade interna (dentro do grupo) e elevada heterogeneidade externa (entre grupos).

Para o desenvolvimento da metodologia, Reis (1997) apresentou cinco etapas:

- seleção de indivíduos a serem agrupados;
- definição das variáveis a partir das quais será obtida a informação ao agrupamento dos indivíduos;
- definição de uma medida de semelhança ou distância;
- escolha de um critério de agregação dos indivíduos denominado de algoritmo de partição;
- interpretação e validação dos resultados.

O ponto de partida consiste em selecionar indivíduos para classificá-los em um pequeno número de grupos mutuamente excludentes.

De acordo com Aldenderfer e Blashfield (1984), um dos fatores que

mais influencia o resultado da análise de agrupamento é a escolha das variáveis. A seleção das variáveis a serem incluídas na análise de agrupamento deve ter argumento baseado em uma teoria, suposição ou o conhecimento da importância de analisa-lás.

Um aspecto importante refere-se a padronização das variáveis. Everitt e Dunn (1996) citaram que há muita controvérsia em relação à estandardização da variável com média nula e variância unitária, pois aquelas variáveis, que deveriam ser as melhores discriminantes para diferença entre grupos, são modificadas e conseqüentemente reduzem a capacidade de distinguir as espécies de forma natural. Somente o conhecimento profundo do assunto traz a decisão correta.

A sugestão dada por Bussab, Andrade e Miazaky (1990) seria aplicar a análise de agrupamento no conjunto de dados originais e, posteriormente, nos dados estandardizados.

2.3.2.1 Medidas de dissimilaridades

O conceito de similaridade ou dissimilaridade é fundamental na análise de agrupamento. Existem diversas medidas de dissimilaridade para medir a relação entre dois indivíduos. Essas medidas definem critérios para avaliar se dois indivíduos estão próximos ou distantes, e distinguir se pode fazer parte de um mesmo grupo ou não.

Aldendefer e Blashfield (1985) classificaram as medidas de (dis)semelhanças em quatro categorias: coeficientes de similaridades ou dissimilaridades; coeficientes de correlação; coeficientes de associação e medidas de semelhança probabilística. Em geral, nos trabalhos publicados, basicamente três medidas predominam na análise de agrupamento: coeficientes de dissimilaridade (ou similaridade), coeficientes de correlação e coeficientes de associação.

Coeficiente de dissimilaridade

O estabelecimento de uma medida de dissimilaridade entre dois indivíduos constitui-se o ponto de partida para várias técnicas multivariadas. O primeiro passo na análise de agrupamento é transformar a matriz de dados em uma matriz de dissimilaridade. Para isso existem várias medidas utilizadas como coeficientes de dissimilaridades, quando os caracteres são morfológicos. Cormack (1971) apresentou diversas medidas, entre essas, a mais utilizada é denominada distância euclideana.

A distância euclideana, distância euclideana média e a distância de Mahalanobis são frequentemente utilizadas nos trabalhos científicos de melhoramento genético, e os resultados medem a distância genética dos cultivares. Segundo Cruz e Regazzi (1997), a distância D^2 de Mahalanobis tem muita utilidade pelo fato de ter grande analogia com outras técnicas multivariadas.

Messetti (2000) apresentou a distância generalizada de Mahalanobis, cujo cálculo, ao contrário de outras distâncias, envolve a estrutura de variabilidade, logo necessita de repetições para estimar as médias originais e a matriz de covariâncias residuais entre características mensuradas. Esta é recomendada por medir objetivamente a posição multidimensional de cada indivíduo em relação ao centro médio das observações e tem propriedades estatísticas que viabilizam testes de significância.

A vantagem da distância euclideana, segundo Cruz e Carneiro (2003), é que não necessita da existência de informações em nível de repetições, estimando-se apenas a média padronizada. A desvantagem é o fato de ser alterada com as mudanças de escala de medidas, com o número de caracteres estudados, além de desprezar parâmetros que envolvam o grau de correlações entre as variáveis. Para solucionar o problema do número de caracteres envolvidos recomendou a distância euclideana média.

Carvalho et al. (2003) ressaltou que os três coeficientes de dissimilaridade têm sido muito utilizados nas estimativas da divergência genética entre cultivares. A distância euclideana pode ser estimada tomando-se por base dados sem repetições, como geralmente ocorre em Banco Ativo de Germoplasma.

Coeficiente de correlação

Reis (1997) descreveu duas vantagens ao aplicar o coeficiente de correlação. O coeficiente é caracterizado por ser de fácil interpretação geométrica, além da insensibilidade às diferenças de escalas de variáveis, tornando um resultado adimensional - o cálculo da média de todas as variáveis para cada indivíduo realiza de forma natural de padronização das variáveis.

Rosa Neto (2006) relatou que esse coeficiente expressa a similaridade dos dois indivíduos relativos à relação linear. Quanto maior o coeficiente, maior a proximidade entre indivíduos, e mais linearmente estarão relacionados os indivíduos. Se subtrair o valor (1,0) um do módulo do coeficiente de correlação, essa transformação define uma nova medida de dissimilaridade entre indivíduos.

Coeficiente de associação

Os coeficientes de associação definem o grau de dissimilaridade entre os indivíduos, segundo variáveis dicotômicas, como as geradas por marcadores moleculares dominantes como o RAPD- “Random Amplified Polymorphic DNA” - (Polimorfismo de DNA amplificado ao acaso), e o AFLP- “Amplified Fragment Length Polymorphism”- (Polimorfismo de comprimento de fragmento amplificado). As quatro possíveis observações de comparação entre dois genótipos são classificados na presença (1) e ausência (0) da banda no gel de eletroforese.

Sokal e Sneath (1963) citam vários coeficientes de associação, sendo os mais utilizados, os de Jaccard e Sorensen-Dice ou Nei e Li. Alguns trabalhos realizam as comparações dos coeficientes de associação, como Meyer (2002), que comparou oito coeficientes para avaliar a divergência genética: Jaccard, Sorensen-Dice, Anderbeg, Ochiai, Simple Matching, Rogers e Tanimoto, Ochiai II e Russel e Rao sendo este último não recomendado para estudar a divergência genética em que os caracteres envolvidos são dados de marcadores moleculares.

Nessa mesma linha de pesquisa, Emygdio et al. (2003) trabalharam com cultivares de feijão e avaliaram a eficiência de nove coeficientes de similaridade de

Jaccard, Sorensen-Dice, Russel e Rao, Ochiai, Coincidência simples, Roger e Tanimoto, Hamann, Kulczynski 2, Yule e Phi, comparando-os, quanto aos dendrogramas, às projeções no espaço bidimensional e aos números de grupos formados. Os coeficientes de Yule, Russel e Rao, foram os mais discordantes em relação aos demais. Outras medidas podem ser vistas em Everitt, Landau e Leese (2001).

2.3.2.2 Critério de agregação ou algoritmos de agrupamento

Os dois métodos de agrupamento mais utilizados no melhoramento de plantas são a técnica hierárquica e técnica não hierárquica.

A **técnica hierárquica** subdividide-se em agrupamentos divisivos e aglomerativos. Nos hierárquicos aglomerativos, o processo se inicia com a matriz de similaridade, a qual é utilizada para identificar o par de indivíduos mais semelhantes entre si. Os dois indivíduos se agrupam e são considerados um único indivíduo. Em seguida, identifica-se o novo par mais semelhante e formará outro grupo, e assim novos grupos serão formados de acordo com suas similaridades até que todos estejam reunidos num único grupo. Os algoritmos mais empregados na hierárquica aglomerativa e apresentados em trabalhos de melhoramento genético são: método do vizinho mais próximo, método do vizinho mais distante, método das médias dos grupos, método dos centróides. Alguns trabalhos aplicaram esta técnica: Totti (1997), Messetti (2000), Melo (2000), Ferreira (2001), Moura (2003) e Rosa Neto (2006)

Os hierárquicos divisivos, de maneira inversa, parte de um único grupo e finaliza com todos indivíduos separadamente.

A técnica não hierárquica ou métodos de partição são métodos usados para agrupar genótipos dentro de uma classificação simples de K grupos, em que K é especificado “a priori” ou é determinado como parte do método de agrupamento.

Rojas, Barriga e Figueroa (2000) aplicaram a técnica não hierárquica, e ressaltaram que seu uso está relacionado ao grande número de genótipos envolvidos no agrupamento. Dois critérios podem ser adotados neste método. O primeiro baseia-se na minimização da soma de quadrados dentro dos grupos, equivalente à maximização da

dispersão entre os grupos. O segundo critério consiste em maximizar as distâncias de Mahalanobis entre os grupos.

Souza (2004) estudou 233 variedades de soja quanto a concentração de isoflavonóides. A análise de agrupamento, através do método das K médias, foi mais indicado devido ao grande número de dados observados. A validação dos nove grupos foi estabelecida pela análise de variância por variável.

Para finalizar, uma boa sugestão foi apresentada por Hair Jr. et al. (2005) que sugeriram a combinação de ambas técnicas (hierárquica e não hierárquica). A técnica hierárquica estabelece o número de grupos para serem aplicados “a priori” na técnica não hierárquica. Outra vantagem é visualizar os centróides dos grupos na técnica hierárquica, para que estes valores sejam aplicados como semente inicial exigida na técnica não hierárquica.

A seguir serão abordados dois tópicos (itens 2.3.2.3 e 2.3.2.4) pouco explorados nos trabalhos de melhoramento genético. Em termos de técnicas multivariadas, muito se evoluiu no decorrer dos anos, e há trabalhos tratando de assuntos recentes e inovadores dentro de técnicas usuais, com o propósito de simplificar e auxiliar a interpretação dos fenômenos biológicos.

A divulgação desses tópicos virá a contribuir e enriquecer os resultados finais dos trabalhos científicos de diversas culturas agrícolas utilizadas em melhoramento genético.

2.3.2.3 Definição do número de grupos

Alguns pesquisadores das técnicas de agrupamento recomendam aplicar mais de um método sobre o mesmo conjunto de dados e comparar os grupos formados para apresentar um melhor resultado. Aldenderfer e Blashfield (1984) indicaram reaplicar a metodologia numa amostra menor do mesmo conjunto de dados. Se a solução final não é estável, não se deve generalizar os resultados.

Everitt, Landau e Leese (2001) apresentaram uma alternativa para obter o número adequado de grupos. A técnica também apresentada por Calinski e Harabasz (1974), sugere fornecer valores para g (número de grupos), o qual corresponde ao máximo valor de C(g), onde C(g) é dado por:

$$C(g) = \frac{\frac{traço\ ou\ tr\ (B)}{g-1}}{\frac{traço\ ou\ tr\ (W)}{n-g}}, \quad \text{com}$$

B - matriz de dispersão entre grupos (equação 1 – item 3.2.5.2)

W - matriz de dispersão dentro dos grupos. (equação 2- item 3.2.5.2)

Na definição quanto ao número ideal de grupos, muitos trabalhos utilizam o tracejo da linha de Fenon, como em Ferreira (2001). A linha intercepta os ramos formados pelo dendrograma, paralela ao eixo horizontal, onde o número de ramos interceptados é o número de grupos originados, e a locação da linha de Fenon é feita em função da necessidade do pesquisador.

Frei (2006) comentou que não existe um procedimento padrão para resolver esta questão, mas para uma solução satisfatória utilizam-se vários procedimentos: dendrogramas, aplicação de vários métodos, divisão do conjunto de dados em duas amostras ou comparação de várias resoluções usando K médias (k=2, k=3...).

Rosa Neto (2006) refez detalhadamente o exemplo do livro do Bussab, Andrade e Miazaky (1990). Como exemplo ilustrativo, a análise de agrupamento foi aplicada aos dados moleculares de 40 estirpes de ribózio isolados de nódulos de feijão e decidiu o número ideal de grupos por meio da análise do comportamento do nível de similaridade.

2.3.2.4 Validação e interpretação dos agrupamentos

Para certificar-se de que os agrupamentos realmente diferem entre si, é necessário validar os agrupamentos. Validar significa certificar-se de que realmente os grupos

diferem. A proposta mais antiga deve-se a Sokal e Rohlf (1962), denominada **coeficiente de correlação cofenética**, sendo a idéia básica realizar uma comparação de distâncias efetivamente observadas entre os indivíduos, e as distâncias previstas a partir do processo de agrupamento. É a medida de validação mais utilizada nos métodos de agrupamentos hierárquicos.

Barroso e Artes (2003) propuseram quatro alternativas:

- a correlação cofenética;
- a aplicação da análise de variância multivariada para verificar se existe diferença estatisticamente significante entre os vetores médios dos grupos;
- dois gráficos multivariados: **gráfico silhueta,** para verificar se o indivíduo está mais próximo dos indivíduos do seu próprio grupo ou dos indivíduos do grupo vizinho;
- **gráfico de perfil,** no eixo das abscissas indicam-se as variáveis e no eixo das ordenadas, as escalas de medidas. A média é representada por um ponto no eixo cartesiano e, unindo-se os pontos, obtêm-se os perfis de cada grupo.

Hair Jr. et al. (2005) discutiram a importância da validação dos agrupamentos. Primeiro, por ser uma técnica exploratória, caracterizada como descritiva, sem base teórica e não inferencial. Não se pode generalizar conclusões de uma amostra para população. Logo, como método exploratório, a idéia é gerar hipóteses, mais que testá-las, sendo a validação um passo muito importante dessa técnica para não comprometer análises posteriores.

2.4 Análise de Componentes Principais

2.4.1 Breve histórico

A **análise de componentes principais** citado em Morrison (1976) foi desenvolvida primeiramente por Karl Pearson (1901). Essa teoria foi reformulada por Hotteling (1933), na avaliação das habilidades dos alunos de resolverem problemas de

aritmética e a velocidade com que os textos eram lidos. Na psicologia moderna, as variáveis que apresentavam uma maior influência foram chamadas de fatores mentais, mais tarde denominadas de componentes. A análise desses componentes que maximizavam a variância dos dados originais foi denominada por Hotteling de Análise de Componentes Principais.

Thurstone (1931) e Hotelling (1933) estiveram trabalhando na mesma linha de pesquisa. Rao (1966) contribuiu de maneira notável, pois sugeriu um grande volume de idéias concernente a aplicações, interpretações e extensões dessa metodologia. Gower (1966) discutiu algumas relações entre componentes principais e outras técnicas estatísticas. Finalmente, Jeffers (1967) deu um impulso de maneira prática, discutindo a complexidade da aplicação de componentes principais.

2.4.2 Metodologia

A análise dos componentes principais é um método estatístico multivariado que transforma um conjunto de variáveis, inicialmente correlacionadas entre si, num outro conjunto de variáveis não correlacionadas, que resultam de combinações lineares das variáveis originais. Essas combinações lineares são chamadas de componentes principais.

O objetivo desta metodologia não é explicar as correlações entre as variáveis, mas apenas encontrar funções matemáticas entre as variáveis iniciais que expliquem o máximo possível da variação existente nas variedades e que permitam uma redução no espaço paramétrico para simplificar a interpretação de resultados, que são de grande interesse em estudos de melhoramento.

Os componentes podem ser derivados da matriz de covariâncias ou da matriz de correlação. Reis (1997) orienta que, se os componentes forem estimados através da matriz de correlação, e caso as variáveis não estejam correlacionadas, deve-se testar a validade da aplicação dessa análise por meio de um dos três testes: Teste de esfericidade da Bartlett; Estatística de Kaiser-Meyer-Olkin (KMO) ou Matriz antiimagem.

Uma questão importante dentro da concepção geral dessa metodologia é a diferença nas escalas de medidas dos diversos caracteres agronômicos envolvidos no estudo. Van Laar (1991) demonstrou a importância de padronizar as variáveis antes de gerar os componentes. Indicou a padronização quando as medidas das variáveis estão em escalas diferentes ou quando a análise de agrupamento é aplicada posteriormente à análise de componentes principais. Segundo Tabachnick e Fidell (2001), essa metodologia não requer suposição sobre a forma da distribuição multivariada, mas, se existe a normalidade, a análise é engrandecida.

O primeiro passo da metodologia consiste em detectar o primeiro componente, aquele que explica a maior variabilidade global das variáveis. A solução é algébrica, equivale a extrair os autovalores "eigenvalues", λ_i, de uma matriz, os quais expressam a variância de cada um dos componentes. Os autovetores "eigenvectores" orientam os componentes no espaço dos caracteres, e as coordenadas dos autovetores são compreendidas como coeficientes das variáveis originais para a formação do componente principal (JOHNSON; WICHERN, 1992).

Os componentes são calculados em ordem decrescente de importância. O primeiro componente principal, explica a maior parte da variabilidade entre os dados, e essa variância corresponde ao maior autovalor da matriz de correlação ou matriz de variâncias e covariâncias. O segundo componente explica a maior parte da variabilidade restante (menor que a explicada pelo primeiro componente) e, assim, sucessivamente. A importância de cada componente é dada pela percentagem de variância total que este absorve. Segundo Cruz e Carneiro (2003) os primeiros componentes principais em estudos de divergência genética têm sido utilizados quando eles envolvem 80% da variação total.

Barroso e Artes (2003) apresentaram três objetivos da metodologia: redução da dimensionalidade dos dados, obtenção de combinações interpretáveis das variáveis e descrição da estrutura de correlação das variáveis. O principal objetivo e o mais utilizado desta técnica é a redução da dimensionalidade das "p" variáveis envolvidas no estudo em umas poucas "k" variáveis, sem perda substancial de informação. A interpretação e a visualização dos resultados podem ser facilitadas quando ocorre a passagem de um espaço multidimensional, proporcionado pelas "p" variáveis, para um espaço bi ou tridimensional,

mantendo um elevado grau de explicação. Outra vantagem da redução refere-se ao fato de os componentes principais não estarem correlacionados, podendo ser interpretados independentemente.

Diversos autores apresentam a análise de componentes principais, destacando-se, Anderson (1958), Morrison (1976), Johnson e Wichern (1988), Van Laar (1991), Reis (1997) e Mingotti (2005).

2.4.2.1 Critérios para definir o número de componentes

Em síntese, a análise de componentes principais busca reduzir o espaço paramétrico, mas uma dificuldade encontrada nesta técnica consiste em determinar o número de componentes principais que deve ser utilizado na redução desse espaço paramétrico. Esse é outro tópico muito abordado na estatística e pouco utilizado na área biológica.

Silva (2005) apresentou quatro critérios para escolha do número de componentes: o critério de Kaiser, o diagrama de autovalores “scree test”, os fatores interpretáveis e o critério de simulação de Lèbart. Elaborou um programa computacional para gerar componentes, de fácil manuseio, e acessível aos pesquisadores da área agronômica.

O critério de Kaiser (1958), também denominado critério da raiz latente, sugere manter na análise os componentes principais correspondentes ao número de autovalores maiores ou iguais à média das variâncias das variáveis no estudo, quando a análise incorpora a matriz de covariâncias. Ou, seguindo a mesma idéia, selecionar somente os componentes principais, correspondentes aos autovalores maiores que um (1,0), quando a análise incorpora a matriz de correlação.

Horn e Engstrom (1979) discutiram problemas envolvendo a aplicação do critério de Kaiser, referentes à magnitude dos resultados. Por exemplo, encontrar um autovalor igual a 1,01 e ser retido na análise. Ou um autovalor igual a 0,99 e ser descartado da análise. Como decidir o número de componentes frente a esses valores?

O critério proposto por Cattell (1966) é o gráfico “Scree-plot”, em que observa-se o número de componentes que se deve excluir da análise. O gráfico, conforme Everitt e Der (2006), descreve no eixo das abscissas os números das ordens dos componentes (ordenados por magnitude decrescente) e, no eixo das ordenadas, os correspondentes autovalores. Comumente, a diferença entre os primeiros autovalores é grande e diminui entre os últimos. A sugestão é optar pelo número de componentes observados no eixo das abscissas, quando a variação do segmento gráfico passa a ser pequena.

Em relação ao critério de Fatores Interpretáveis, Van Laar (1991) indicou o resumo do complexo multivariado logo nos primeiros componentes, desde que estes absorvam 70% ou mais da variância total. Silva (2005) apresentou o critério de Fatores Interpretáveis e fez uma boa discussão referente aos trabalhos que utilizaram 70, 80 até 90% da variância total.

Resumindo Mingotti (2005) diz que o objetivo da análise de componentes principais consiste em sintetizar as informações das “p” variáveis originais, em um número menor de funções lineares dessas. Reforça que a utilidade prática do método diminui com o aumento do número de componentes utilizados, pois quanto mais componentes, maior será a dificuldade para a discussão biológica dos resultados.

Qualquer que seja o critério adotado para definir o número de componentes principais no estudo, é aconselhável adotar o bom-senso, e verificar se existe algum componente relevante sendo descartado no processo.

A técnica de componentes principais vem sendo utilizada em diversas áreas, em especial para avaliação da divergência genética entre genótipos ou populações de diferentes cultivos com base em caracteres quantitativos, como serão abordados a seguir.

Strapasson (1997) selecionou os descritores botânico-agronômicos mais representativos para caracterizar acessos das espécies *Paspalum guenoarum* e *Paspalum plicatulum* (Capim), do grupo Plicatula, por meio de componentes principais. O método foi utilizado para selecionar descritores e descrever a variabilidade presente na coleção de acessos do germoplasma estudado.

Agong, Schittenhelm e Friedt (2000) avaliaram a diversidade genética de 26 espécies de tomates, baseado na variação morfológica, agronômica e no tratamento bioquímico. O experimento em blocos casualizados foi conduzido no Centro Federal de

Pesquisa na Agricultura, na Alemanha. Em seu trabalho, agrupou as espécies, utilizando a técnica de componentes principais, e verificou que foram claramente separadas quanto as características do fruto.

Alves (2005) caracterizou e comparou a estrutura genética de sete populações de cupuaçuzeiro, uma planta nativa da Amazônia. Para selecionar algumas das 53 variáveis observadas, utilizou a técnica da análise de componentes principais, obtendo 64% de redução das variáveis. A partir daí, obteve seis grupos geneticamente diferentes, dos 31 acessos avaliados, considerando como medida de divergência a distância euclideana média e o método de ligação da média.

A aplicação de componentes principais para avaliação da divergência genética é evidenciada em trabalhos com germe de trigo por Gou e Song (1991); alfafa por Annicchiarico (1992); cacau por Dias (1994) e linhagens de milho por Meyer (2002).

2.5 Análise de Variáveis Canônicas

A análise multivariada, com base em variáveis canônicas, foi relatada primeiramente por Rao (1952). É um processo alternativo aos componentes principais nas situações em que dispõem-se de dados experimentais com informações de repetições, de modo que estimam-se médias e matriz de dispersão residual entre dados. De forma geral, a técnica serve para avaliar o grau de diversidade entre os genótipos, quando plotados em gráficos de dispersão, desde que a concentração da variabilidade total entre as primeiras variáveis canônicas esteja acima de 80% (CRUZ; CARNEIRO, 2003).

Camarano (1997) avaliou a similaridade de 10 populações de girassol por meio da divergência genética em eixos canônicos. As estimativas das variâncias atingiram mais de 80% nos quatro experimentos realizado em Goiás, o que justificou a utilização de um único eixo para ilustrar a disposição das cultivares num espaço unidimensional.

Melo (2000) utilizou as variáveis canônicas e afirmou que possuem vantagem em relação aos componentes principais, por considerar a estrutura de covariância residual, e ser invariante com respeito à transformação não singular dos caracteres originais.

Cruz e Regazzi (1997) identificaram Y_n como a variável canônica de menor importância relativa, dada por: $Y_n = a_1x_1 + a_2 x_2 + ... a_nx_n$, em que $x_1, x_2, ..., x_n$ são variáveis originais padronizadas. Identifica-se a variável de menor importância como aquela associada ao maior dos coeficientes a_1, a_2 ,....., a_n. A segunda variável de menor importância é identificada, utilizando o mesmo critério, pelos coeficientes da variável canônica Y_{n-1} e, assim, sucessivamente.

Ferreira (2001) ressaltou que análises de variáveis canônicas podem ser utilizadas com o objetivo de identificar e descartar variáveis de menor importância na divergência entre tratamentos. Identificam-se os caracteres de menor importância entre os cultivares em estudo, como sendo aqueles cujos coeficientes de ponderação são de maior magnitude, em valor absoluto, nas últimas variáveis canônicas.

Neves (2003) investigou a divergência genética de cultivares de arroz (moderno e tradicional), utilizando diferentes medidas de dissimilaridades da análise de agrupamento. As análises de variáveis canônicas permitiram a visualização dos diferentes cultivares pela redução das dimensões do conjunto de dados, preservando a maior parte das informações biológicas. Ressaltou ainda que a dispersão por variáveis canônicas pode ser utilizada para caracterizar coleções de germoplasma e, em decisões de melhoramento, para explorar o vigor híbrido ou minimizar a depressão por endogamia.

Miranda et al. (2003) avaliaram nove cultivares tropicais de milho de pipoca por meio de técnicas multivariadas. As análises empregadas foram agrupamento com base na distância de Mahalanobis e dispersão gráfica por variáveis canônicas. As duas primeiras variáveis canônicas foram suficientes para representar 96,5% da variância total. Descartou três caracteres dos oito obtidos, e quanto aos resultados das duas técnicas, considerou concordância parcial, obtendo quatro grupos geneticamente diferentes.

Na literatura existem diversos trabalhos que fazem uso da análise de variáveis canônicas para o estudo da divergência genética, como descritos na seqüência. Cruz (1990) aplicou essa metodologia para selecionar genótipos de milho; Reis et al. (1999) estudaram a divergência genética com trigo; Messetti (2000) avaliou o grau de divergência genética entre populações de girassol; Lal, Sharma e Singh (2001) trabalharam com camomila; Adugna e Labuschagne (2003) com óleo de linhaça; Benin (2003) com aveia.

2.6 Análise de Variância Multivariada

2.6.1 Breve histórico

O primeiro passo foi dado por Wishart (1928), que trabalhou com a distribuição normal multivariada. Hotelling (1933) verificou que a distribuição T^2 é uma extensão da distribuição t de Student para normal multivariada.

Historicamente, Wilks (1932), por meio do método da razão de verossimilhança "likelihood ratio method", obteve uma generalização da análise de variância aplicada a várias variáveis. A estatística Λ (lâmbda) de Wilks fornece testes de significância para análise multivariada, limitado na época pela dificuldade de cálculo para valores exatos. As distribuições assintóticas foram as alternativas para a análise de variância multivariada, sendo a aproximação às distribuições χ^2 e F as mais utilizadas.

Seguindo o curso histórico, Bartlett (1934) aplicou o teste de significância para duas variáveis, e Hotelling (1935) verificou a utilidade dessa metodologia em testes de independência para vários grupos de variáveis.

A análise de variância multivariada segue a mesma restrição da análise univariada quanto à homogeneidade de variâncias. Box (1950) definiu o teste M, que determina se deve ou não rejeitar a hipótese de igualdade de matrizes de covariâncias populacionais, utilizando o método do quociente de verossimilhanças como generalização do teste Bartlett (1937). Rao (1952) faz um breve histórico da análise multivariada, mostrando o ponto inicial dos trabalhos, visavam generalizar a análise de variância univariada em multivariada para qualquer tipo de delineamento experimental.

2.6.2 Metodologia

A análise de variância multivariada é uma extensão da análise de variância univariada, diferindo em alguns aspectos como o grande número de variáveis envolvidas no experimento e o propósito de avaliar as diferenças entre médias de grupos (MARDIA; KENT; BIBBY, 2003).

No modelo univariado, testa-se a diferença entre as médias de um caráter em diversos grupos, considerando a pressuposição de que as variâncias entre os grupos são homogêneas. No modelo multivariado, testa-se a hipótese de que as populações têm o mesmo vetor de médias, contra a alternativa que pelo menos um vetor difere significativamente dos demais. Johnson e Wichern (1992) colocaram que é equivalente a testar se os centróides dos grupos são distintos, considerando a pressuposição de que as matrizes de covariâncias desses grupos são homogêneas. Esse teste pode ser realizado utilizando-se a Análise de Variância Multivariada, abreviadamente denominada de MANOVA.

2.6.2.1 Verificação dos pressupostos

Para os procedimentos de testes multivariados semelhantes à análise de variância univariada, há alguns pressupostos a serem atendidos: as matrizes de covariâncias devem ser iguais para todos os grupos de tratamentos, e o conjunto de “p” variáveis dependentes deve seguir a distribuição normal multivariada.

Em geral, o teste de Box verifica as igualdades das matrizes de covariâncias e os níveis de significância para a estatística do teste. Segundo Hair Jr. et al. (2005), é um teste sensível a desvios da normalidade e requer a verificação da normalidade univariada de todas as variáveis do processo, anteriormente à aplicação do teste de Box.

Quanto às pressuposições do modelo, Mardia, Kent e Bibby (1979) estudaram o efeito da não normalidade, mostrando que os resultados dos testes não são

afetados por heterogeneidade das matrizes de variâncias e covariâncias, quando o número n de repetições for grande e igual para os tratamentos.

Quanto à normalidade, os métodos estatísticos pressupõem que cada vetor de variáveis seja proveniente de uma população normal multivariada. A proposta de Reis (1997) é testar a normalidade para cada variável, embora isso não implique que todas as variáveis em conjunto mantenham a normalidade. O autor descreveu como construir o gráfico Q-Q, o qual pode ser utilizado para avaliar a normalidade de determinada distribuição. A hipótese de normalidade é plausível quando o resultado no gráfico se aproxima de uma linha reta. A normalidade univariada constitui-se em condição necessária para a normalidade multivariada, mas não suficiente.

Moreira (2003) testou a normalidade por meio de teste de Shapiro-Wilk e teste de Kolmogorov-Smirnov. Os caracteres foram submetidos à transformação dos dados através log(x). Quanto ao pressuposto da homogeneidade utilizou-se o teste de Bartlett.

Ferreira e Cantelmo (2005) comparou o desempenho do teste multivariado de normalidade de Shapiro Wilk com o desempenho do teste de assimetria e curtose, utilizando simulação Monte Carlo. Avaliou as taxas de erro tipo I e o poder dos testes. O teste de Shapiro Wilk teve fraco desempenho, com altas taxas do erro tipo I, e o poder de teste semelhante ao da assimetria e curtose. O teste de assimetria e curtose apresentou melhor desempenho principalmente quando n>50.

2.6.3. MANOVA

Para avaliação da significância da hipótese nula referente a comparação dos vetores de médias de grupos, podem ser utilizados os testes de Wilks, Pillai, Hotteling-Lawley e Roy. O teste mais utilizado é o de Wilks, mas vale ressaltar que os quatro testes são competidores.

Geralmente os softwares estatísticos geram os quatro testes para análise de variância multivariada, o mais popular é o teste de Wilks. Harris (1975) justificou que os determinantes são mais fáceis de computar do que os autovalores, sendo indicado o critério de Wilks quando os autovalores são aproximadamente iguais.

Reis (1997) fez a seguinte explanação quanto a dois pontos importantes: robustez e potência do teste. De forma geral, combinando a robustez (não violação dos pressupostos) e a potência de teste (probabilidades "versus" erro tipo I e erro tipo II) é possível encontrar a seguinte ordenação, para situações em que estejam presentes mais do que um autovalor não nulo: Pillai $\geq$ Wilks $\geq$ Hotelling $\geq$ Roy.

Hair Jr et al. (2005) afirmaram que o teste de Roy é o teste estatístico mais poderoso se todos os pressupostos são inicialmente atendidos e as medidas dependentes são representativas de uma única dimensão de efeitos. Consideraram os outros testes semelhantes por envolverem todas as raízes características no cálculo, com aproximação da estatística F e χ^2.

Demétrio (1985) mostrou de forma simples as semelhanças e as diferenças entre análises de variância univariada e multivariada, utilizando 22 variedades de cana-de- açúcar. Oshiiwa (2001) considerou o mesmo delineamento e desenvolveu um programa computacional para microcomputadores, de fácil acesso e manuseio por pesquisadores da área agronômica, usando a análise de dados experimentais, a fim de ilustrar os dados agronômicos.

Melo (2000) estimou a divergência genética de dez cultivares de milho por meio de 25 caracteres morfoagronômicos e marcadores moleculares. Utilizou o critério de Wilks, num delineamento fatorial para testar a hipótese de igualdade dos efeitos dos cultivares

de milho, e a interação de cultivares e locais. Concluiu pela diferença significativa dos efeitos, locais e interação, embora a interação se apresentasse significativa somente para alguns caracteres.

Daoyu e Lawes (2000) trabalharam com melhoramento genético da fruta Kiwui. Inicialmente, partindo dos resultados obtidos da análise de variância multivariada, verificou-se a diferença significativa entre os vetores de médias das seis populações da fruta Kiwui envolvidas no estudo. Posteriormente, aplicou a análise discriminante para identificar os caracteres que apresentaram melhor desempenho para diferenciar as populações, visando aumentar a produtividade e melhorar a extração de vitamina C da fruta.

Ferreira (2001) quantificou a divergência fenética entre 20 clones de palma forrageira do Banco de Germoplasma da Empresa Pernambucana de Pesquisa Agropecuária, envolvendo oito caracteres. Por meio da técnica da variância multivariada, e adotando o critério de Wilks, detectou diferença significativa entre os vetores de médias de todos os clones de palma forrageira envolvidos no trabalho. Posteriormente, aplicou outras técnicas multivariadas de interesse.

Ledo (2002) aplicou a análise de variância multivariada para cruzamentos dialélicos, com objetivo de selecionar genótipos superiores de milho. Ressaltou a vantagem desta metodologia comparada com a tradicional, pois há a possibilidade de estimar as matrizes de covariâncias dos efeitos genéticos do modelo, às correlações fenotípicas e genotípicas e, consequentemente, obter informações para orientação de programas de melhoramento.

Nos trabalhos de melhoramento genético, a análise de variância multivariada geralmente antecede as outras técnicas multivariadas, como Sousa (2003) utilizou a metodologia para a verificação preliminar da existência de variabilidade genética, por meio da comparação dos vetores de médias de várias populações de guanazeiro.

As técnicas multivariadas tem contribuído, de forma significativa, para o desenvolvimento científico em diversas áreas. No estudo de divergência genética, observa-se que a maioria dos trabalhos se restringe ao uso da estatística multivariada exploratória. Vale ressaltar a importância da estatística inferencial multivariada para realmente validar os resultados, sendo esta uma das propostas desta tese.

Na Tabela 1 está apresentado um resumo da revisão de literatura.

Quadro 1 – Resumo da revisão de literatura.

Análise	Autor	Cultura	Observações
Agrupamentos	Camarano 1997	Girassol	Distância Mahalanobis Técnicas hierárquicas
	Duarte 1998	Feijão	Coeficiente de Associação Técnicas Hierárquicas
	Messetti 2000	Girassol	Distância Mahalanobis Técnicas hierárquicas
	Meyer 2000	Milho	Coeficiente de Associação Técnicas hierárquicas
	Ferreira 2001	Palma Forrageira	Distância euclideana média Distância Mahalanobis
	Souza 2004	Soja	Técnicas não hierárquicas K médias
Componentes Principais	Dias 1994	Cacau	Agrupamento 25 acessos por Componentes Principais
	Strapasson 1997	Capim	Descarte dos descritores Descreve variabilidade
	Agong 2000	Tomates	Agrupamento por dispersão gráfica C.P.
	Alves 2003	Cupuaçuzeiro	64% descartes variáveis
	Silva 2005	Milho	Quatro critérios para selecionar número de componentes
Variáveis Canônicas	Reis 1999	Trigo	Agrupamento 94 acessos por eixos canônicos
	Melo 2000	Milho	Dispersão gráfica 10 cultivares eixo canônicos
	Moura 2003	Guanazeiro	Agrupamento 93 cultivares três eixos canônicos
	Miranda 2003	Milho pipoca	Agrupamento nove cultivares por dois eixos canônicos
MANOVA	Melo 2000	Milho	Critério Wilks
	Daoyu 2000	Fruta Kiuwi	Critério de Wilks Análise discrimante
	Ferreira 2001	Palma forrageira	Critério Wilks
	Ledo 2002	Milho	Critérios Wilks Cruzamentos dialélicos
	Sousa 2003	Guanazeiro	Critério Wilks
	Moura 2003	Jaborandi	Critério de Wilks
	Moreira 2004	Tomate	Pressupostos:Teste Shapiro-Wilk, Kolmogorov e Teste de Bartlett

3 MATERIAL E MÉTODOS

3.1 Material

Objetivando conservar, multiplicar, caracterizar, avaliar os acessos do Banco de Germoplasma de Girassol, e divulgar as informações disponíveis por meio de catálogos, pesquisadores da EMBRAPA realizaram experimentos no ano agrícola de 2000, na região de Londrina – PR, localizada à latitude de 23^0 23'S, longitude 51^0 11'W e altitude de 566 m.

Os experimentos das pesquisas foram dispostos seguindo o seguinte planejamento de semeadura:

a) para linhagens (gerações) foram constituídas três fileiras de 6,0m de comprimento, com plantas espaçadas entre si de 0,30 m, em espaçamento de 0,70m entre as fileiras;

b) para populações e variedades foram constituídas três fileiras, realizando autofecundações e SIB (fecundações cruzadas), respectivamente.

No processo de adubação da área experimental foram aplicados 500 kg/ha de formulação 40-80-80 de NPK, sendo a colheita realizada quando as plantas atingiram a maturação, para obter melhor qualidade de germinação e vigor das sementes.

Para o estudo da semelhança dos genótipos, foram consideradas 12

populações de girassol e dois híbridos em cada experimento. Os cinco caracteres quantitativos foram observados em campo, durante às fases de desenvolvimento e a avaliação dos aquênios feita em laboratório.

Nas Tabelas 2 e 3, descritas a seguir, estão apresentadas as variáveis respostas (com as respectivas unidades de medidas) e as populações de girassol, com os devidos tamanhos amostrais.

Tabela 1 – Variáveis estudadas e suas respectivas unidades de medida.

Indicação	Variável	Unidade de medida
X_1	Altura da planta	Centímetros
X_2	Tamanho do capítulo	Centímetros
X_3	Curvatura do caule	Nota*
X_4	Peso dos aqüênios	Gramas / planta
X_5	Floração inicial	Número de Dias

* Classificada entre nota 1 (pouca curvatura) a 7 (muita curvatura).
A nota ideal para bom aproveitamento da produção está entre 3 e 4.

Tabela 2 – Especificação das populações (linhagens) de girassol.

População	Nome	Número de amostras
P1 – Variedade 1	Sundak	178
P2 – Variedade 2	Sunfola	191
P3 – Variedade 3	Local Blue	201
P4 – Variedade 4	Klein A	193
P5 – Variedade 5	Belensky	191
P6 – Variedade 6	Majak	201
P7 – Variedade 7	Armovensky	189
P8 – Variedade 8	Gigante	160
P9 – Variedade 9	PI 343804	198
P10 – Variedade 10	PI 170389	189
P11 – Variedade 11	PI 170407	180
P12 – Variedade 12	Guayacan	190

Fonte: Banco Ativo de Germoplasma. EMBRAPA/ Soja - Londrina.

3.2 Métodos

Os dados foram submetidos à análise de variância univariada para verificar a existência de diferenças entre as populações de girassol. Em seguida, foi realizada a análise de componentes principais e variáveis canônicas para inspeção gráfica da divergência genética. A análise de agrupamento foi realizada por intermédio da distância de Mahalanobis e distância euclideana.

A análise de variância univariada, considerando um experimento em blocos casualizados, com 12 populações para verificar o efeito das médias das populações e o efeito das gerações para cada variável isoladamente. O modelo matemático representativo é:

$$Y_{ij} = \mu + t_i + b_j + e_{ij} \quad \text{onde;}$$

Y_{ij} Valor observado na parcela da população i na repetição j;

μ Média geral do caráter;

T_i efeito fixo ao tratamento que foi aplicado na parcela;

b_j efeito devido ao bloco j em que se encontra a parcela;

e_{ij} efeito aleatório devido aos fatores não controlados.

3.2.1 Medidas descritivas para amostras multivariadas

O estudo apresentará as medidas descritivas para amostras multivariada dos dados. Da análise exploratória constarão as medidas descritivas de posição e variabilidade da estrutura multivariada representadas pelos vetores de médias, matrizes de covariâncias, matrizes de correlações necessárias para o desenvolvimento do método.

*Vetor de médias amostrais***.** O vetor de médias μ será estimado pelo vetor $\underset{\sim}{\overline{X}}$ de médias amostrais, em que $\overline{X_i}$ representa a média da i-ésima variável com i=1,2,...,p, será dado por:

$$\underset{\sim}{\overline{X}}_{px1} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \ldots \\ \ldots \\ \overline{X}_p \end{bmatrix}, \text{ sendo } \quad \overline{X}_i = \sum_{j=1}^{n} X_n \quad i = 1,2,....,p.$$

Matriz de covariâncias amostrais – A matriz de covariâncias $\sum_{pp}$ será estimada pela matriz simétrica de covariâncias amostrais S_{pp} representada por:

$$S_{pxp} = \begin{bmatrix} s_{11} & s_{12} & ..... & s_{1p} \\ s_{21} & s_{22} & ..... & s_{2p} \\ .... & .... & ..... & .... \\ s_{p1} & .... & ..... & s_{pp} \end{bmatrix} = [\, s_{ii'} \,]$$

O elemento genérico $s_{ii'}$ da matriz S é dado por:

$$s_{ii'} = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_{ij} - \overline{X}i \right)\left(X_{i'j} - \overline{X}i' \right) \qquad \text{para} \quad i\,,\, i' = 1,2,.....,p \quad \text{e} \quad i < i'.$$

Matriz de correlações amostrais – A matriz de correlação $\Re$ será estimada pela matriz simétrica de correlação amostral R_{pxp} representada por:

$$R = \begin{bmatrix} 1 & r_{12} & ..... & r_{1p} \\ r_{21} & 1 & ..... & r_{2p} \\ ... & ... & ..... & .... \\ r_{p1} & r_{p2} & ..... & 1 \end{bmatrix} \text{ sendo } \; r_{ii'} = \frac{s_{ii'}}{\sqrt{s_{ii}\; s_{i'i'}}} \quad \text{para } i\,,\, i' = 1,2,.....,p \text{ e } i < i'.$$

As próximas análises estatísticas incluem a aplicação de quatro metodologias: técnicas de componentes principais com base na matriz de correlação, análise de agrupamento, análise de variáveis canônicas e análise de variância multivariada.

3.2.2 Análise de Componentes Principais

A análise de componentes principais é um método muito utilizado na estimativa da divergência genética, pois resume o conjunto de variáveis originais em poucos componentes lineares, que permitem informar-se a respeito do comportamento dos genótipos em um espaço bi ou tridimensional.

Para não comprometer a análise de componentes principais, quando se utiliza a matriz de correlação R, Reis (1997) indica um dos três testes para verificar se as variáveis estão correlacionadas: Teste de esfericidade de Bartlett; Estatística de Kaiser-Meyer-Olkin (KMO) ou Matriz antiimagem.

O teste de esfericidade de Bartlett verifica a hipótese da matriz de correlação populacional (P) ser igual à matriz identidade (I), ou seja, admitir ausência de associação linear (variáveis não correlacionadas) entre as características morfométricas estudadas.

i) Hipótese de nulidade do teste: H_0: P = I

ii) Sob a hipótese de ausência de associação linear (H_0 verdadeira), a estatística do teste é dada por: $-\left[n-1-\frac{1}{6}(2p+5)\right]\ln|R|$

A estatística do teste tem distribuição assintótica χ^2 com [½ p (p-1)] graus de liberdade, sob a veracidade de H_0 , com a regra de decisão habitual.

iii) Quando se rejeita a hipótese nula, a hipótese alternativa mostra que existem variáveis que apresentam correlações significativas, indicando a continuidade do procedimento.

3.2.2.1 Componentes Principais

A técnica de componentes principais transforma as p variáveis originais contidas no vetor $\underset{\sim}{X}'$ = [X_1, X_2, ...X_p], mediante uma matriz ortogonal ou

matriz dos elementos dos autovetores, em p novas variáveis (futuros componentes principais) descritas no vetor $\underset{\sim}{Y}$ =[Y_1, Y_2,.....,Y_p] mutuamente não correlacionadas entre si.

Se $\underset{\sim}{X}$ é um vetor de variáveis aleatórias com média μ e matriz de covariâncias Σ simétrica, positiva semidefinida, então a transformação dos componentes principais é dada por:

$$\underset{\sim}{X} \rightarrow \underset{\sim}{Y} = A'(\underset{\sim}{X} - \mu)$$

onde A é uma matriz ortogonal. A matriz de covariância do vetor componente principal ($\underset{\sim}{Y}$) é representada por Var ($\underset{\sim}{Y}$) = $A' \Sigma A = \Lambda$, sendo Λ uma matriz diagonal contendo as raízes características de Σ, ou seja, $\Lambda = diagonal\{\lambda_1, \lambda_2, ..., \lambda_p\}$.

Para o conjunto de componentes principais, utiliza-se a ordem de magnitude decrescente para as raízes características frente à sua importância na informação da variação dos dados, isto é, o primeiro componente explica o máximo possível da variância dos dados originais, o segundo o máximo possível da variância não explicada, e assim por diante. A última raiz corresponde ao componente que menor contribuição dá para a explicação da variância total dos dados originais, isto é:

$$\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq ..... \geq \text{Var}(Y_p).$$

3.2.2.2 Obtenção dos componentes principais

Para estimar os coeficientes de ponderação das variáveis em cada componente e a respectiva variância, deve-se inicialmente, encontrar as raízes características da matriz S (matriz de covariância amostral) ou da matriz R (matriz de correlação amostral).

Os coeficientes de ponderação das variáveis são estimados pelos elementos dos vetores característicos (autovetores) correspondentes, os quais não serão os mesmos nas duas opções. O primeiro componente principal é dado por: $Y_1 = a_{11}X_1 + a_{21}X_2 + ... + a_{p1}X_p = \underset{\sim}{a}_1' \underset{\sim}{X}$, que é combinação linear das variáveis originais com a maior variância, sendo esta representada por $Var(Y_1) = \underset{\sim}{a}_1 ' S \underset{\sim}{a}_1$. Para obter o primeiro componente principal deve-se, determinar o vetor $\underset{\sim}{a}_1$, de forma que a variância Y_1 seja maximizada, sujeito à restrição no conjunto de soluções de $\underset{\sim}{a}_1$, por meio de: $\underset{\sim}{a}'_1 \underset{\sim}{a}_1 = 1$ (soma de quadrados dos coeficientes igual a 1).

Expressando a variância Y_1 e incorporando a restrição pelo multiplicador λ de Lagrange, tem-se pela derivação em relação ao vetor a_1, o seguinte sistema linear homogêneo: $(S - \lambda I)\underset{\sim}{a}_1 = \underset{\sim}{0}$.

A solução do sistema deve ser tal que $\underset{\sim}{a}_1 \neq \underset{\sim}{0}$ (solução não trivial). Assim, é necessário que o determinante de $(S - \lambda I)$ seja nulo, para que o sistema se torne indeterminado, e que a solução seja escolhida dentre aquelas que satisfaçam a condição $\underset{\sim}{a}'_1 \underset{\sim}{a}_1 = 1$.

O primeiro componente principal está associado ao maior autovalor de S, ou seja λ_1, aquele que maximiza a variância de Y_1. O vetor solução correspondente a λ_1 será denominado de vetor próprio, autovetor ou vetor característico de S associado a λ_1, cujos elementos serão os coeficientes da função linear das variáveis originais, denominada primeiro componente principal (Y_1). As demais componentes principais são estimados de maneira análoga à descrita anteriormente.

3.2.2.3 Decomposição da variância total

Seja A a matriz de vetores próprios A = [$a_1, a_2, \ldots . a_p$] de $\sum$ e $\underset{\sim}{Y}$ = [Y_1, Y_2, ...,Y_p]' o vetor dos p componentes principais. Considerando $\underset{\sim}{Y} = A' \underset{\sim}{X}$, a matriz de covariâncias será dada por: Var (Y) = $A' \sum A = \Lambda$.

Sendo A uma matriz ortogonal e pela propriedade cíclica do traço, tem-se:

$$\mathrm{tr}(\Lambda) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \mathrm{Var}(Y_i) \quad \text{e}$$

$$tr\,(\,\Lambda\,) = tr\,(\,A'\,\Sigma\,A\,) = tr\,(\,\Sigma\,A'\,A\,) = tr\,(\,\Sigma\,) = \sum_{i=1}^{p} Var\,(X_i\,),$$

ou seja, a soma das variâncias das p variáveis originais X_i é igual à soma das variâncias dos p componentes principais.

A importância relativa de um componente é avaliada pela porcentagem da variância total explicada pelo j-ésimo componente principal, expressa por: $\dfrac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$.

A variabilidade acumulada explicada pelos m $\leq$ p primeiros componentes principais é dada por: $\dfrac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$. 100% .

Em estudos de divergência genética, Cruz e Regazzi (1997), comentam que é desejável que a variância acumulada nos dois primeiros componentes principais exceda 80%. Assim sendo, o gráfico de dispersão cujos eixos são os componentes principais representará cada população de girassol, de forma aceitável.

3.2.2.4 Indicação para o número de componentes principais

Uma decisão que deve ser tomada durante o procedimento consiste em estabelecer o número de componentes que deve ser considerado na análise. A seguir serão apresentados dois critérios para decidir quanto ao número de componentes.

O **critério de Scree-plot** (CATTELL, 1966) trata-se de um procedimento gráfico, no qual no eixo das abscissas representa-se a ordem numérica dos componentes e, no eixo das ordenadas os correspondentes autovalores. A sugestão é utilizar como número de componentes o número correspondente à ordem, em que a variação do segmento de reta no gráfico passa a ser pequena. Iniciando com o primeiro autovalor, os ângulos de inclinação decrescem rapidamente no início e depois lentamente se aproximam de uma reta horizontal.

Para o **critério de Kaiser** (1958) ou análises dos autovalores, a idéia é que qualquer autovalor individual deve explicar a variância de pelo menos uma variável se o mesmo for mantido para interpretação. Cada variável contribui com o valor um (1) do autovalor total. Os autovalores maiores que um (1) são significantes e menores que um (1,0) não são significantes e portanto descartados.

Neste caso, o aconselhável é excluir os componentes relativos aos autovalores menores que 1, quando se considera a matriz de correlação; e autovalores menores que a média quando se considera a matriz de covariâncias (MARDIA; KENT; BIBBY, 2003).

Do ponto de vista biológico, estabelecido o número de componentes principais retido na análise, certo que a variabilidade disponível foi maximizada, sugere-se que a técnica proporcionou uma simplificação dos resultados. Logo, pode-se avaliar o comportamento das populações num espaço bi ou tri-dimensional.

Na área biológica, a possibilidade de descartar variáveis que pouco contribuem para a discriminação do material genético, reduz custo e tempo despendidos na experimentação agronômica. Portanto a variável que apresenta maior coeficiente de ponderação (elemento do autovetor) no componente de menor autovalor é considerada de menor importância para explicar a divergência genética, sendo passível de descarte. Mardia, Kent e Bibby (1979) orientam não descartar duas variáveis com base no mesmo componente,

mais correto é identificar a importância relativa dos caracteres no outro componente de variância maior e, assim, sucessivamente.

3.2.3 Análise de agrupamento

Os métodos de análise de agrupamento são procedimentos de estatística multivariada, que têm por objetivo agrupar um conjunto de indivíduos isolados, em grupos homogêneos. O primeiro passo desta metodologia é a conversão da matriz de dados numa matriz de dissimilaridade, calculada das relações entre todos os possíveis pares de populações.

3.2.3.1 Coeficiente de dissimilaridade ou critério de semelhança

Um conceito importante para as técnicas de agrupamento é a escolha de um critério para avaliar se dois indivíduos estão próximos, ou que quantifique o quanto os indivíduos são semelhantes e podem fazer parte de um mesmo grupo. Esta medida é denominada coeficiente de similaridade, e quanto maior o valor do coeficiente, mais semelhantes serão os indivíduos; ou coeficiente de dissimilaridade, quanto maior o valor do coeficiente, menos semelhantes serão os indivíduos.

Os métodos de agrupamento exigem que os coeficientes respeitem as propriedades métricas como apresentadas em Aldenderfer e Blashfield (1984).

Sejam l,m dois pontos que representam medidas sobre dois indivíduos. Uma função real d (l,m) é definida como função de distância, e denominada métrica, se satisfaz as seguintes propriedades: I) simetria, $d(l,m) = d(m,l)$;

II) não negatividade, $d(l,m) \geq 0$;

III) $d(l,m) = 0$, se somente se $l = m$;

IV) $d(l,m) \leq d(l,s) + d(s,m)$, desigualdade triangular.

3.2.3.1.1 Coeficientes de dissimilaridades para atributos quantitativos

Existem diversas métricas que podem ser utilizadas como distâncias entre indivíduos observacionais. As mais destacadas estão descritas a seguir.

Distância euclideana - A métrica mais conhecida para indicar a proximidade entre dois indivíduos l e k é a distância euclideana, dada por:

$d_{l,k} = [\sum_{i=1}^{p} (X_{li} - X_{ki})^2]^{1/2}$ em linguagem matricial: $d_{l,k} = [(\underset{\sim}{X}_l - \underset{\sim}{X}_k)'(\underset{\sim}{X}_l - \underset{\sim}{X}_k)]^{1/2}$

onde $\underset{\sim}{X}_l$ e $\underset{\sim}{X}_k$ são dois vetores amostrais, cotejados nas variáveis pertencentes ao universo de observações.

Distância euclideana média - O valor da distância euclideana aumenta quando novas variáveis são incorporadas às originais. Uma maneira de contornar esse problema é dividir esse valor pela raiz quadrada do número de caracteres, isto é:

$$\Delta_{l,k} = \frac{1}{\sqrt{p}} d_{l,k}$$

Essa distância é apenas um reescalonamento da anterior, possuindo as mesmas propriedades e, portanto, produzindo os mesmos resultados se submetidos às técnicas de análise de agrupamentos.

Distância Generalizada de Mahalanobis - Um coeficiente de dissimilaridade, definido por Mahalanobis (1936), é a distância generalizada entre dois grupos l e k, dada matricialmente por:

$$D^2_{l,k} = \left(\underset{\sim}{\bar{x}}_l - \underset{\sim}{\bar{x}}_{k'}\right)' S^{-1} \left(\underset{\sim}{\bar{x}}_l - \underset{\sim}{\bar{x}}_k\right)$$

sendo S_{pxp} a matriz comum de covariâncias das unidades amostrais referentes aos grupos l,k, e $\underset{\sim}{\bar{x}}_l$ e $\underset{\sim}{\bar{x}}_k$ seus respectivos vetores de médias. Pode-se também usar a matriz de correlação amostral R no lugar da matriz de dispersão S, principalmente, quando as variáveis não estão mensuradas na mesma unidade métrica.

Coeficiente de correlação linear simples - Sokal e Sneath (1963) utilizam como coeficiente de similaridade entre dois indivíduos (l e k), para caracterizar as relações entre os caracteres, o coeficiente de correlação momento produto de Pearson definido por:

$$r_{l,k} = \frac{\sum_{j=1}^{p}\left(x_{lj} - \overline{x_l}\right)\left(x_{kj} - \overline{x_k}\right)}{\sqrt{\left[\sum_{j=1}^{p}\left(x_{lj} - \overline{x_l}\right)^2\right]\left[\sum_{j=1}^{p}\left(x_{kj} - \overline{x_k}\right)^2\right]}}.$$

Ao contrário da distância, quanto maior for o valor absoluto de $r_{l,k}$ (mais próximo da unidade) mais similares serão os indivíduos. A transformação desse coeficiente define uma nova medida de dissimilaridade entre dois indivíduos dado por: $d_{l,k} = 1 - \left|r_{l,k}\right|$

3.2.3.2 Algoritmo de agrupamento

Os algoritmos utilizados na formação dos grupos podem ser classificados em métodos hierárquicos e não hierárquicos.

Os métodos hierárquicos aglomerativos são formados a partir de uma matriz de similaridade, onde identifica-se o par de indivíduos que mais se parecem. Nesse instante o par é agrupado formando um único indivíduo. Esse processo requer uma nova matriz de similaridade. Em seguida, identifica-se o par mais semelhante que formará o novo grupo e, assim, sucessivamente, até que todos os indivíduos fiquem reunidos num só grupo.

Os métodos apresentados a seguir são de interesse para a formação de conglomerados, e eles diferem entre si pela função usada para o cálculo repetido dos coeficientes de similaridade entre os novos agrupamentos formados e os candidatos potenciais à futura admissão nos agrupamentos seguintes. Para formalizar esta etapa considera-se os agrupamentos l, k contendo r_l , r_k indivíduos, em que esses números são maiores ou, no mínimo, iguais a 1. Se os agrupamentos l, k se unem, isto é indicado como (l, k) com $r_{l,k} = r_l + r_k$ indivíduos.

"Single Linkage" - "Method SLM" - Método do vizinho mais próximo

Uma população P_j candidata-se a um agrupamento quando apresenta uma distância a este agrupamento igual à sua menor distância com relação aos membros do agrupamentos. A distância entre dois agrupamentos L e K será dada por:

$$d_{L,K} = \min_{\substack{l \in L \\ k \in K}} d_{l,k}$$

"Complete Linkage" - "Method CLM" - Método do vizinho mais distante

Uma população P_j candidata-se a um agrupamento quando apresenta uma distância a este agrupamento igual à sua maior distância com relação aos membros do agrupamentos. A distância entre dois agrupamentos L e K será dada por:

$$d_{L,K} = \max_{\substack{l \in L \\ k \in K}} d_{l,k}$$

"Unweighted Pair-group Average" – "Método UPGMA" - Método não ponderado de agrupamento aos pares por médias aritméticas.

Define-se a distância entre dois agrupamentos como a média entre os valores individuais de um dos grupos com os de outro grupo.

$$d_{L,K} = \frac{1}{r_l r_k} \sum_{\substack{l \in L \\ k \in K}} d_{l,k}$$

3.2.3.3 Definição do número de grupos

O número de grupos pode ser definido “a priori” quando se tem algum conhecimento a respeito dos dados, ou pode ser definido“a posteriori” com base nos resultados da análise. A seguir serão adotados três critérios para definir o número de grupos: O dendrograma, a análise de comportamento de fusão e a análise de comportamento do nível de similaridade.

Dendrograma - Os resultados dos algoritmos apresentados na técnica hierárquica aglomerativa se combinam até que seja estabelecido um diagrama de árvore denominado dendrograma, no qual no eixo das abscissas se posicionam os indivíduos e no eixo das ordenadas, o nível de ligação após aplicação da metodologia.

Análise do comportamento do nível de fusão (MINGOTTI, 2005)- Quando o procedimento passa do estágio k para (k+1), e a similaridade entre os grupos que estão sendo agrupados vai decrescendo e, conseqüentemente, a distância entre os grupos vai crescendo, pode-se apresentar um gráfico do número de grupos “versus” o nível de distância (fusão) do agrupamento em cada estágio do processo. Se visualmente ocorrem “*pontos de salto”*, o momento indica o número de grupos final. Quando o segmento de reta no gráfico se estabiliza, ou não apresenta variações significativas, tem-se a partição ótima. Caso existam vários pontos de salto, sugere-se aplicar outro procedimento.

Análise do comportamento do nível de similaridade (MINGOTTI, 2005) **-** Este critério visa observar o comportamento do nível de similaridade em cada k estágio do agrupamento. Se C_i e C_l são grupos unidos num determinado estágio, o nível de similaridade é dado por:

$$S_{il} = \left[1 - \frac{d_{il}}{\max\left\{d_{jk,}\ j,k = 1,2,....n\right\}}\right]100\ \%$$

no qual max ($d_{jk,}$ j, k =1,2,....n) é a maior distância entre os n indivíduos da matriz de distância D_{nxn} do primeiro estágio do processo de agrupamento.

Para estabelecer o número de grupos ideal, detectam-se pontos onde há um descréscimo acentuado na similaridade dos conglomerados unidos, pontos que indicam a interrupção do algoritmo utilizado.

3.2.3.4 Validação e interpretação dos agrupamentos

No contexto de agrupamento, a validação do método significa certificar-se de que os grupos realmente diferem entre si. Neste sentido, alguns gráficos multivariados podem auxiliar nesta etapa, como visto (BARROSO; ARTES, 2003).

Gráfico de silhueta - O gráfico de silhueta é recomendado para verificar a qualidade dos agrupamentos. Esse procedimento é puramente descritivo e a idéia é verificar se um indivíduo está mais próximo dos indivíduos do seu próprio grupo ou de grupos vizinhos. O valor da silhueta no ponto i é dado por:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Seja G(i) o grupo que contém o indivíduo i, e n_{Gi} o número de indivíduos pertencentes ao grupo. Define-se;

$$a(i) = \frac{\sum_{i' \in G(i); i' \neq i} d_{i;i'}}{n_{G(i)} - 1}$$

em que $d_{ii'}$ é a distância euclideana entre os indivíduos i e i'

Na seqüência, determina-se a distância média entre o elemento i e os elementos de cada grupo diferente do grupo i. Define-se como grupo H(i) aquele com a menor distância média encontrada entre os elementos dos grupos e o indivíduo i. Seja $n_{H(i)}$ o número de indivíduos pertencentes a esse grupo H(i), agora denominado vizinho do indivíduo i, tem-se; então:

$$b(i) = \frac{\sum_{i' \in H(i); i' \neq i} d_{i;i'}}{n_{H(i)}}.$$

O valor de s (i) varia entre 1 e -1. Valores próximo de 1 indicam ótima alocação do indivíduo i, pois b (i) >> a (i) e valores negativos indicam péssima alocação. Valores negativos sugerem péssima alocação dos indivíduos, indicando que o indivíduo i está mais próximo de outro indivíduo não pertencente ao seu grupo.

Gráfico de perfil - O gráfico de perfil como procedimento para verificar a qualidade dos agrupamentos formados, enriquecem a interpretação. No eixo das abscissas indicam-se as variáveis e no eixo das ordenadas, as escalas de medidas. Cada média é representada por um ponto nos eixos cartesianos. Unindo-se os pontos, observam-se os perfis dos grupos, onde ficam em evidência as diferenças dos grupos por variável.

.

3.2.4 Análise de variáveis canônicas (eixos canônicos)

A análise de variáveis canônicas teve como objetivo avaliar a similaridade das populações por intermédio de uma dispersão gráfica em eixos cartesianos. Essa metodologia consiste num processo alternativo para avaliação do grau de similaridade genotípica entre populações, levando em consideração a matriz T de covariâncias genotípicas, e a matriz E de covariâncias residuais entre os caracteres avaliados.

Quando se utiliza este procedimento, é comum a transformação das variáveis originais em variáveis padronizadas e não correlacionadas, de modo que a matriz de dispersão se iguala à identidade (CRUZ; REGAZZI, 1997). A técnica consiste em transformar o conjunto de p variáveis originais em um conjunto que é função linear dos $X_{i's..}$

Foram verificadas as seguintes propriedades:

a) Se Y_j é uma variável canônica então: $Y_j = a_1X_1 + a_2X_2 + ...a_pX_p = \underset{\sim}{a}'X$ e

b) Se Y_j' é outra variável canônica então: $Y'_{j'} = a_1X_1 + a_2X_2 + ...a_pX_p = \underset{\sim}{b}'X$ então

$\sum_{j}\sum_{j'} a_j a_{j'} \sigma_{jj'} = \sum_{j}\sum_{j'} b_j b_{j'} \sigma_{jj'} = 1$

$\sum\sum a_j b_{j'} \sigma_{jj'} = 0$, ou seja, as variáveis canônicas são não correlacionadas.

c) A primeira variável canônica Y_1 apresenta a maior variância, a segunda Y_2 a segunda maior e, assim, sucessivamente.

As variâncias e os coeficientes de ponderação das variáveis canônicas foram estimados pela resolução do sistema de equação matricial: $[T - \lambda E]\, \underset{\sim}{a} = \underset{\sim}{0}$.

A importância relativa de cada variável canônica é dada pela razão entre a variância por ela explicada e o total da variância disponível. Quando há, nas primeiras variáveis, a concentração de grande proporção da variância total, geralmente acima de 80%, torna-se viável o estudo da divergência genotípica por meio das distâncias geométricas entre populações em gráfico de dispersão, cujas coordenadas são os escores relativos às primeiras variáveis canônicas (CRUZ ; CARNEIRO, 2003).

O estudo de descarte de variáveis baseia-se no princípio de que a importância relativa das variáveis canônicas decresce da primeira para a última, pois as últimas variáveis são responsáveis pela explicação de uma fração mínima da variância total disponível. Assim, o caráter que apresentou maior coeficiente de ponderação, em módulo, nas últimas variáveis canônicas, foi considerado de menor importância para explicar a variabilidade em estudo, portanto, passível de descarte.

3.2.5 Análise de variância multivariada

A utilização da análise de variância multivariada (MANAVA) fez-se com o objetivo de comparar os vetores de médias dos grupos detectados nas análises anteriores. A hipótese nula da igualdade dos vetores de médias foi testada não apenas para uma variável, mas simultaneamente para o conjunto de p variáveis.

3.2.5.1 Os pressupostos sobre a estrutura de dados

Os pressupostos subjacentes à análise de variância multivariada (MANAVA) são generalizações dos pressupostos da análise de variância simples (ANAVA):

- as populações têm matriz comum de covariâncias $\sum$ (homogeneidade);
- as populações têm distribuição multinormal (normalidade).

Homogeneidade - Pode-se utilizar o teste M de Box (1950) para verificar a hipótese de igualdade de matrizes de covariâncias populacionais. Reis (1997) sugeriu verificar o pressuposto da normalidade antes de aplicar o teste M. Caso seja violado, fazer a transformação dos dados.

O teste M foi definido utilizando-se o método do quociente de verossimilhança e pressupondo-se que os vetores de médias dos grupos são desconhecidos.

I) Hipóteses estatísticas:

$H_0: \sum_1 = \sum_2 = ... = \sum_g$ (Homogeneidade das matrizes de covariâncias)

H_1: Existe $\sum_g \neq \sum_{g'}$ com $g \neq g'$ (Heterogeneidade das matrizes de covariâncias)

II) Fixado o nível de significância α, a estatística MC , em que

$M = (n-g) \ln |S| - \sum_{i=1}^{g} (n_i - 1) \ln |S_i|$, com $n = n_1 + n_2 + ... + n_i$;

S é a matriz comum de covariância, S_i a matriz de covariância do grupo i.

$$C = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(g-1)} \left[\sum_{l=1}^{k} \frac{1}{n_1 - 1} - \frac{1}{n - g} \right] ,$$

sob a veracidade de H_0, μC^{-1} têm distribuição quiquadrado com $\frac{1}{2}$(g-1) p (p+1) graus de liberdade.

Normalidade - A qualidade da inferência depende do modo como a população se aproxima de uma distribuição normal multivariada. Hair Jr. et al. (2005) sugeriram testar a normalidade univariada através do teste de assimetria e curtose, embora não se garanta a multinormalidade.

Teste de assimetria e curtose – Seja uma amostra da populações (x_1, x_2, ...,x_n), retirada de alguma distribuição $F_0(x)$ desconhecida. Estabelece-se o confronto com a distribuição normal F(X), para verificar se é razoável estudar a amostra por meio da distribuição normal, admitindo como a verdadeira função de distribuição da amostra selecionada.

i) Hipótese nula: H_0: $F_0(X) = F(X)$, a distribuição populacional segue as características da distribuição normal.

Hipótese alternativa: H_1: $F_0(X) \neq F(X)$, a distribuição populacional não segue as características da distribuição normal.

ii) O teste estatístico para assimetria e curtose é realizado utilizando

$$Z = \frac{G_1}{\sqrt{\frac{6}{N}}} \quad \text{e} \quad Z = \frac{G_2 - 3}{\sqrt{\frac{24}{N}}}, \qquad \text{Se X ~ normal então:}$$

$G_1 \sim N\left(0, \left(\sqrt{\frac{6}{N}}\right)^2\right)$ $G_1 = \frac{m_3}{\sqrt{m_2^3}}$ coeficiente de assimetria;

$G_2 \sim N\left(3, \left(\sqrt{\frac{24}{N}}\right)^2\right)$ $G2 = \frac{m_4}{m_2^2}$ coeficiente de curtose;

m_3 : terceiro momento centrado na média populacional;

m_2 : segundo momento centrado na média populacional;

m_4 : quarto momento centrado na média populacional.

iii)Conclusão: Se o valor calculado exceder o valor crítico indica a rejeição da suposição de normalidade da distribuição no nível de significância de α, definido “a priori”.

Teste de Lilliefors para normalidade (BARBIN, 2003) - A idéia foi introduzida por Kolmogorov (1933), para adaptação de uma específica e conhecida distribuição F(X) a dados provenientes de uma distribuição desconhecida $F_0(X)$. Lilliefors (1967) introduziu uma modificação no teste de Kolmogorov-Smirnov, considerando a média $\bar{x}$ e a variância para a variável reduzida Z_i estimadas a partir dos valores observados da variável X.

i) Hipóteses $H_{0:}$ A distribuição populacional segue as características da distribuição normal.

H_1: A distribuição populacional não segue as características da distribuição normal

ii) Estatística Teste

$D = \max \left|F(Z_i) - S(Z_i)\right|$ ou $D = \max \left|F(Z_i) - S(Z_{i-1})\right|$, em que

$F(Z_i)$ são as probabilidades acumuladas da variável normal reduzida $Z_i = \frac{X_i - \bar{x}}{s}$,

$S(Zi) = \frac{k}{n}$, onde k é o número de observações menores ou iguais que $X_{i,}$

iii) Conclusão se $D_{calc} > D_{tab}$, rejeita-se H_o, sendo α o nível de significância definido "a priori".

3.2.5.2 Teste de Wilks - Teste de igualdade de g vetores de médias

A hipótese nula refere-se à igualdade entre os centróides dos grupos populacionais (um centróide é um vetor de dimensão p das médias de um grupo).

i) H_0: $\underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \ldots\ldots \underset{\sim}{\mu}_g = \underset{\sim}{\mu}$ as populações têm vetores de médias iguais.

H_1: existe pelo menos duas populações com vetores de médias diferentes.

Na análise de variância multivariada adotou-se o teste de Wilks para a avaliação da diferença entre os vetores de médias dos tratamentos. O teste resulta do quociente entre os determinantes das matrizes de somas de quadrados e produtos cruzados dentro de grupos e total, ou seja: $\Lambda = \frac{|W|}{|B+W|}$, onde

B é a Matriz das Somas de Quadrados e Produtos entre grupos;

$$B = \sum_{i=1}^{g} n_i \left(\underset{\sim i}{\bar{x}} - \underset{\sim}{\bar{x}}\right)\left(\underset{\sim i}{\bar{x}} - \underset{\sim}{\bar{x}}\right)' \quad \textbf{(1)}$$

W é a Matriz de Somas de Quadrados e Produtos dentro de grupos;

$$W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\underset{\sim ij}{x} - \underset{\sim}{\bar{x}_i}\right)\left(\underset{\sim ij}{x} - \underset{\sim}{\bar{x}_i}\right)' \quad \textbf{(2)}$$

T é a Matriz de Somas de Quadrados e Produtos Total.

$$T = W + B = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\underset{\sim ij}{x} - \underset{\sim}{\bar{x}}\right)\left(\underset{\sim ij}{x} - \underset{\sim}{\bar{x}}\right)'$$

A Tabela 3 resume o quadro de resultados do teste Λ de Wilks (MARDIA; KENT; BIBBY, 2003).

Tabela 3 - Análise de variância multivariada para comparar vetores de médias dos grupos (MANAVA).

Fontes de variação	**Graus de Liberdade**	**Matriz de somas dos quadrados e produtos**	**Estatística do teste**
Entre os grupos	g-1	B	$\Lambda = \dfrac{\lvert W \rvert}{\lvert W + B \rvert}$
Dentro dos grupos	n-g	W	
Total	n-1	T	

$$\Lambda = \frac{\lvert W \rvert}{\lvert W + B \rvert} \sim \Lambda\,(p, n\text{-}g, g\text{-}1) \text{ distribuição de } \Lambda \text{ de Wilks.}$$

p = número de variáveis;

n = número de indivíduos;

g = número de grupos.

Observa-se que, quanto maior for a semelhança dos resultados dos dois determinantes, menores serão as diferenças entre os grupos (B) e mais o valor de Λ se aproximará de 1. Ao contrário, se as diferenças entre os grupos forem elevadas quando comparadas com à variabilidade dentro dos grupos, o valor de Λ tenderá a aproximar-se de 0. Assim, a estatística de Wilks é uma medida inversa do grau de diferenciação entre os grupos: quanto menor o seu valor, maior esse grau de diferenciação.

A distribuição Λ de Wilks pode ser aproximada à distribuição χ^2 ou F.

Aproximação de Bartlett (Johnson e Wichern, 1992)

$$-\left[\, n-1-\frac{(p+g)}{2}\right]\ln\Lambda \;\sim \chi^2_{\,p\,(g-1)} \qquad \text{e} \qquad \frac{(n-g)-p+1}{p}\,\frac{1-\Lambda}{\Lambda} \sim F_{(p;\, n-g-p+1)}$$

Aproximação Millon- Rao

$$r = n-1-\frac{1}{2}(p+g) \qquad \text{e} \qquad t^2=\frac{p^2(g-1)^2-4}{p^2+(g-1)^2-5}$$

$$\frac{r\,t-\frac{1}{2}p(g-1)+1}{p(g-1)}\left[\left(\frac{1}{\Lambda}\right)^{\frac{1}{t}}-1\right] \sim \quad F_{(p\,(g-1)\,;\, r\,t-\frac{1}{2}p(g+1)+1)}$$

3.2.6 PROGRAMAS COMPUTACIONAIS

Os softwares utilizados nos procedimentos deste trabalho foram os seguintes:

BioEstat 4.0 – Análises univariadas, análise de agrupamento, análise de componentes principais e as distâncias genéticas.

Statistica 6.0 – Análise de agrupamento, análise de componentes principais e análise de variância multivariada (Critério de Wilks, Pillai's, Hotelling, Roy's). Na verificação do pressuposto para normalidade foi aplicado o teste de Lilliefors, e o teste de Box na verificação do pressuposto para homogeneidade das matrizes de covariâncias.

Programa GENES – O software de livre acesso, desenvolvido pelo professor e pesquisador Cosme Damião Cruz e financiado pelo CNPQ. Desenvolvido primordialmente para pesquisa genética, possui vários procedimentos para estatística univariada e multivariada. O diferencial deste programa esta na execução da análise de variáveis canônicas, sendo necessário determinar anteriormente o vetor de médias e a matriz de covariâncias residuais (Cruz e Carneiro, 2003).

4 RESULTADOS E DISCUSSÃO

4.1 Análise de Variância Univariada e Matriz de Correlações

Os resultados da técnica de análise de variância univariada para as variáveis altura, tamanho do capítulo, curvatura, peso dos aqüênios e dias para floração inicial foram significativos ($p< 0,05$) entre populações e não significativos ($p>0,05$) entre as gerações, exceto para a variável floração inicial (Tabela 4). Camarano (1997) estudou a divergência genética de 10 populações de girassol, encontrando diferenças significativas, semelhantes a deste trabalho, para altura, floração inicial e tamanho do capítulo, dentre outras não semelhantes. Isso reforça o indicativo de que existe divergência genética para esses caracteres na cultura de girassol.

Tabela 4 - Análise de variância das variáveis avaliadas nas populações de girassol.

FV	gl	X1 p valor Altura da planta	X2 p valor Tamanho capítulo	X4 p valor Peso dos Aquênios	X5 p valor Floração inicial
Blocos	2	0,179	0,5943	0,5263	0,02465
Populações	11	0,00001	1,21E-08	1,12E-13	9,5E-22

No intuito de verificar a intensidade das associações lineares entre as variáveis, estimou-se a matriz de correlação (Apêndice 1). O teste t, detectou resultados significativos ($p < 0,05$) entre as variáveis: altura da planta (X_1) e dias para floração inicial (X_5); peso dos aquênios (X_4) e dias para floração inicial (X_5); altura da planta (X_1) e peso dos aquênios (X_4); tamanho do capítulo (X_2) e peso dos aquênios (X_4) no valor de 0,89; 0,76; 0,67 e 0,60 respectivamente. As demais correlações não foram significativas ($p > 0,05$). O tamanho do capítulo apresentou uma correlação positiva de fraca intensidade entre as variáveis, altura da planta (X_1) e dias para floração inicial (X_5) no valor de 0,165 e 0,26 respectivamente. A variável curvatura (X_3) foi responsável pelas correlações negativas de baixa intensidade entre as seguintes variáveis: tamanho do capítulo (-0,189); peso dos aquênios (-0,28) e dias para floração inicial (- 0,03), respectivamente (Apêndice 1). As correlações encontradas foram de forte, média e baixa intensidade. Caso só ocorressem correlações de baixa intensidade, os procedimentos multivariados ficariam comprometidos, pois a estrutura de dependência entre as variáveis estaria bastante enfraquecida.

4.2 Análise de Componentes Principais

Na avaliação da divergência genética das 12 populações de girassol, com base na técnica de componentes principais, foi estimada a matriz de correlação entre as cinco variáveis originais. Este procedimento equivale a estimar a matriz de covariâncias com variáveis padronizadas.

Devido a matriz de correlação estar incluída na análise, antes de gerar os componentes principais, optou-se em aplicar o teste de esfericidade de Bartlett (REIS, 1997). O resultado obtido para o teste apresentou um valor de 39,656, bastante superior à distribuição do qui-quadrado com 10 graus de liberdade (18,3), para um nível de significância de 0,05. Portanto, rejeita-se a hipótese nula, logo há correlações significativas entre as variáveis. Neste sentido, torna-se imperativo o prosseguimento da metodologia. Caso o resultado fosse contrário, a análise de componentes principais estaria comprometida.

As variâncias de cada componente estão apresentadas pelo autovalor da matriz de correlação R, na Tabela 5. Os resultados evidenciam que os dois primeiros componentes acumularam uma percentagem satisfatória, com 82,5% da variabilidade total dos dados. O primeiro componente explica 55% e o segundo 27,5% da variação total. Os demais componentes absorveram apenas 17,5 %. Segundo Cruz e Carneiro (2003), quando a variabilidade dos dois primeiros componentes principais envolvem mais de 80%, torna-se satisfatório o estudo da divergência genética por meio da dispersão gráfica cujos eixos discriminantes serão os dois primeiros componentes principais.

Tabela 5 – Estimativas das variâncias (autovalores) associadas a matriz de correlação, e respectivas porcentagens de explicação da variação total.

Componente Principal	Autovalor	Autovalor Acumulado	% da variância	% da variância acumulada
CP1	2,750	2,75	55,0	55,0
CP2	1,376	4,126	27,5	82,5
CP3	0,720	4,85	14,4	96,9
CP4	0,126	4,97	2,50	99,4
CP5	0,027	5,00	0,60	100,0

Para a definição do número de componentes principais, utilizados na discriminação gráfica, foram utilizados dois critérios. O primeiro critério, denominado critério de Kaiser, que consiste em selecionar autovalores maiores que o valor um (1,0), com isso, notou-se a redução do espaço paramétrico da quinta para a segunda dimensão, pois os dois

primeiros componentes apresentaram autovalores iguais a $\lambda_1 = 2{,}750$ e $\lambda_2 = 1{,}376$ (Tabela 5). O segundo critério, denominado de "Scree-Plot" (Figura 4), foi construído com base na matriz de correlação, e apresentou forte decaimento entre o segundo e o terceiro autovalor, sugerindo o uso de dois ou três componentes.

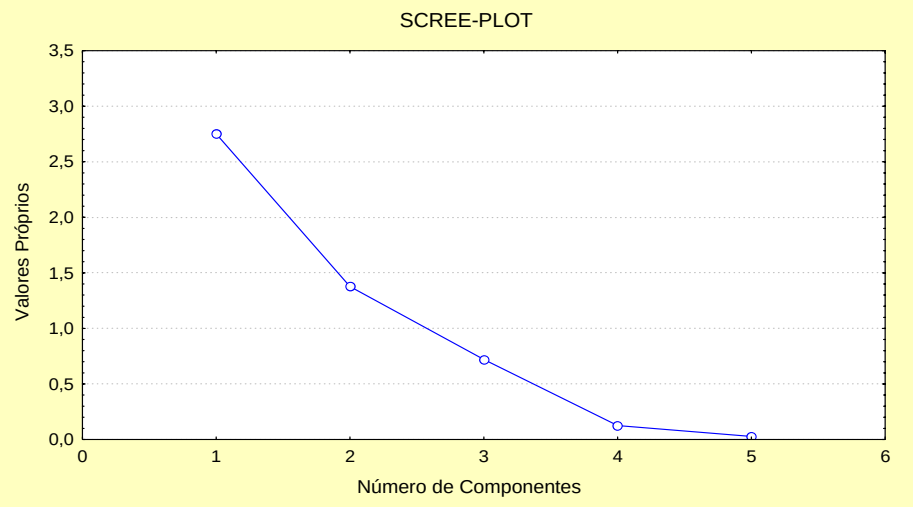


Figura 4 – Scree- Plot da Matriz de Correlação.

Norman e Streiner (1994) denominam o gráfico de "Scree Test de Cattell" e o definiram como uma ferramenta poderosa que depende simplesmente da crítica visual para definição final quanto ao número de componentes.

Confrontando-se os resultados da visualização gráfica (Figura 4), dos valores da variância acumulada (Tabela 5) e a análise dos autovalores pelo critério de Kaiser, optou-se por utilizar apenas os dois primeiros componentes principais. Como já relatado, os componentes principais podem ser utilizados com alta capacidade de buscar diversidades genéticas nas populações. Diante dos resultados encontrados, optou-se por prosseguir a análise para obtenção dos componentes e a análise gráfica, considerando os dois primeiros componentes principais (Tabela 6).

Tabela 6- Coeficientes de ponderação das variáveis morfoagronômicas do girassol.

Variável	**CP1 Y_1**	**CP2 Y_2**	**CP3 Y_3**	**CP4 Y_4**	**CP5 Y_5**
X_1 (altura da planta)	0,325	-0,450	-0,762	-0,316	0,111
X_2 (tamanho do capítulo)	- 0,030	0,744	-0,563	0,152	-0,325
X_3 (curvatura da planta)	0,557	-0,229	0,022	0,753	-0,262
X_4 (peso dos grãos)	0,558	0,143	0,308	-0,556	-0,512
X_5 (Dias p/ floração)	0,520	0,413	0,088	-0,004	0,742

O primeiro componente principal caracterizou-se numa composição envolvendo as respostas da curvatura da planta (X_3), peso dos grãos (X_4) e floração inicial (X_5). Em relação ao segundo componente, o tamanho do capítulo (X_2) acumulou-se aos dias para floração e contrasta com a altura da planta (X_1).

Os escores relativos a cada população, em cada componente, para a representação gráfica bidimensional, são apresentados na Tabela 6. De forma ilustrativa, como exemplo, tomou-se a população 1 nos componentes 1 e 2:

$$Y_{11} = 0{,}325\ (\ \overline{X}_1\ ) - 0{,}03\ (\overline{X}_2) + 0{,}557\ (\overline{X}_3) + 0{,}558\ (\overline{X}_4) + 0{,}52\ (\overline{X}_5) = -0{,}33$$

$$Y_{12} = -0{,}45(\overline{X}_1) + 0{,}744\ (\overline{X}_2) - 0{,}229\ (\overline{X}_3) + 0{,}143\ (\overline{X}_4) + 0{,}413(\overline{X}_5) = -1{,}30$$

Substituindo as médias das variáveis da população P1, logo essas coordenadas localizarão o centróide da população1 na dispersão gráfica (Figura 5). De maneira análoga, seguem os cálculos para as demais populações.

Não se constituiu como objetivo descartar variáveis no presente exemplo, mas, por interesse didático, foi mostrado que a importância das variáveis pode ser analisada pelo coeficiente de ponderação (Tabela 6). Para isto pode-se utilizar da seguinte técnica: localiza-se o maior coeficiente absoluto de ponderação 0,742 no último componente (CP5) que possui o menor autovalor. Logo, a primeira variável selecionada para descarte seria a floração inicial (X_5). Para a segunda variável procede-se de maneira análoga: localiza-se o

maior coeficiente absoluto de ponderação no penúltimo componente (CP4), cujo valor é 0,753. A segunda variável selecionada para descarte seria a curvatura (X_3).

Segundo Cruz e Regazzi (1997), variáveis dispensáveis são aquelas invariantes entre as populações, como no caso da variável curvatura (X_3) com baixa variabilidade (CV= 3,68% - apêndice), ou variáveis redundantes por estarem correlacionadas com outros caracteres, como no caso da variável floração inicial (X_5), que apresentou a mais alta correlação (r = 0,89) com a variável altura da planta (X_1).

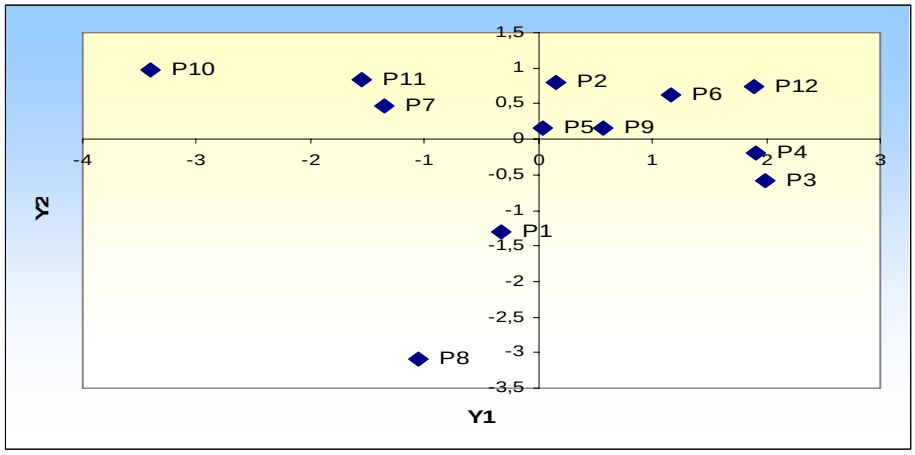
A possibilidade de descarte de variáveis que pouco contribuem para discriminação de germoplasma é fundamental, por ocasionar a redução do tempo destinado ao estudo, economias de recursos materiais e mão-de-obra.

As coordenadas da dispersão gráfica obtidas por meio dos escores dos dois primeiros componentes estão apresentadas na Tabela 7. A divergência genética pode ser representada nesse espaço bidimensional para facilitar a interpretação geométrica.

Tabela 7 – Escores relativos das populações de girassol, obtidos em relação aos dois primeiros componentes principais.

CP	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
Y1	-0,33	0,15	1,99	1,90	0,04	1,16	-1,35	-1,05	0,56	-3,4	-1,55	1,89
Y2	-1,30	0,80	-0,58	-0,19	0,15	0,63	0,47	-3,09	0,15	0,97	0,84	0,75

Pela inspeção visual do gráfico (Figura 5), utilizando os escores dos dois primeiros componentes principais, podem ser identificados de 4 a 5 grupos. Os resultados das próximas análises foram confrontados com este gráfico, dando suporte às futuras decisões, quanto ao número final de grupos.



F**igura 5** - Dispersão das populações de girassol em relação aos escores dos dois componentes principais.

4.3 Análise de Agrupamento

Com base nas estimativas das distâncias euclideana e distância generalizada de Mahalanobis entre duas populações, a partir dos dados padronizados (Apêndice), foi realizada a análise de agrupamento pelos três algoritmos da técnica hierárquica. A utilização dessas distâncias e esses algoritmos ocorreu por estes serem muito utilizados em trabalhos científicos na área agronômica envolvendo diversas culturas (Quadro 1). Observando-se a matriz de distância euclideana, as maiores distâncias se referem à população P10 em relação a P3 (5,99), P4 (5,84), P12 (5,59) e P6 (5,01), respectivamente. A menor distância foi entre P7 e P11 (0,67). Quanto à distância de Mahalanobis, a menor distância continua entre as populações P7 em relação a P1 (0,001) e P11 (0,001). A maior distância novamente aconteceu entre P10 e P12 (0,98).

4.3.1 Determinação do número de grupos

Para a determinação final do número de grupos foram abordados três critérios atuais na área da estatística.

O primeiro critério refere-se ao dendrograma. Na construção dos gráficos foram considerados três métodos da técnica hierárquica, o “Single Linkage”, “Complete Linkage” e “Average Linkage”, contemplados com as seguintes medidas de dissimilaridade: distânia euclideana e distância de Mahalanobis nos dados padronizados.

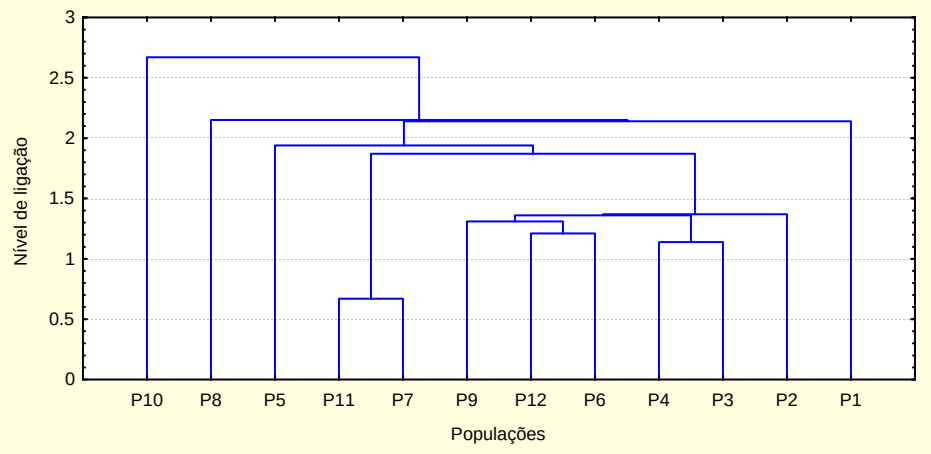


Figura 6 – Dendrograma resultante da análise de agrupamento das populações de girassol obtido do algoritmo “Single Linkage”, baseado na distância euclideana.

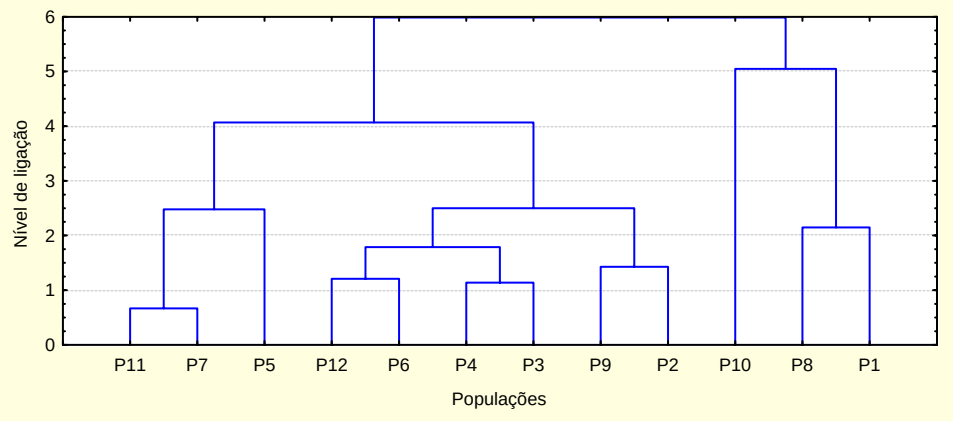


Figura 7 – Dendrograma resultante da análise de agrupamento das populações de girassol obtido do algoritmo “Complete Linkage”, baseado na distância euclideana.

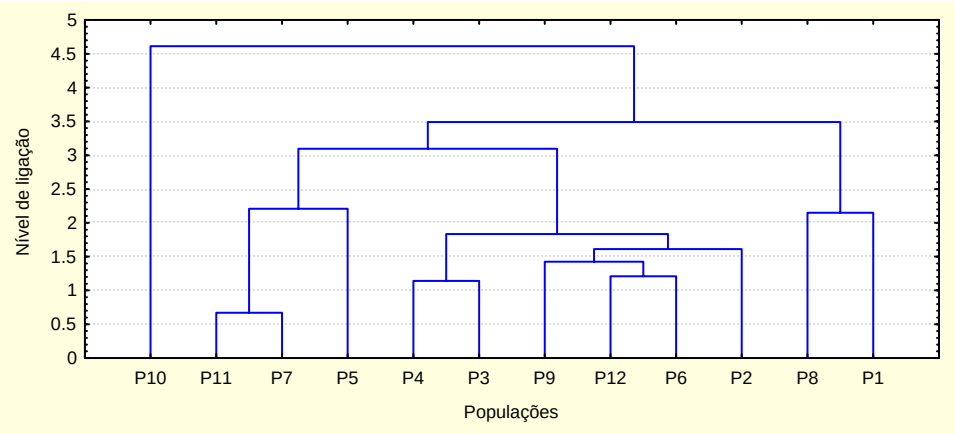


Figura 8 – Dendrograma resultante da análise de agrupamento das populações de girassol obtido do algoritmo "Average Linkage" baseado na distância euclideana.

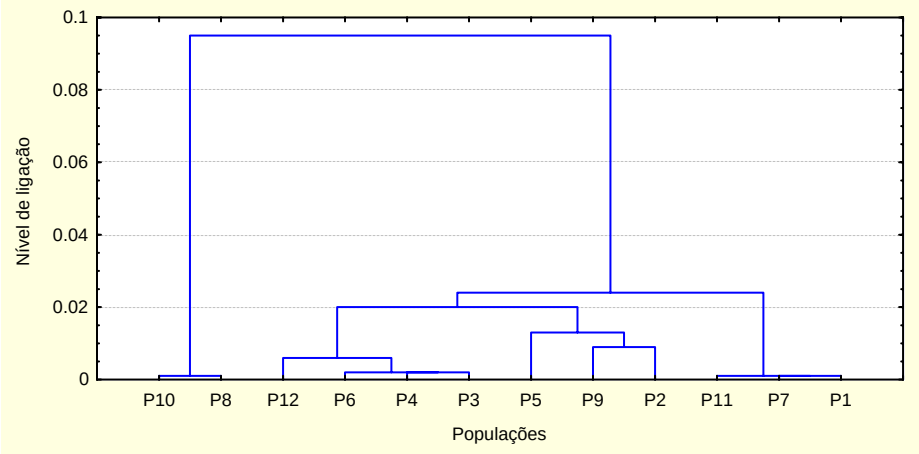


Figura 9 – Dendrograma resultante da análise de agrupamento das populações de girassol obtido do algoritmo "Single Linkage" baseado na distância Mahalanobis.

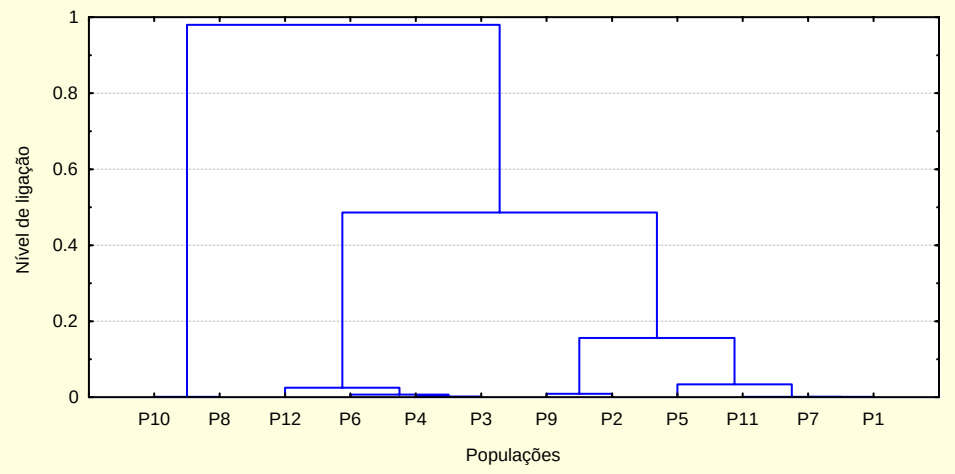


Figura 10 – Dendrograma da análise de agrupamento das populações de girassol obtido do algoritmo "Complete Linkage" baseado na distância de Mahalanobis.

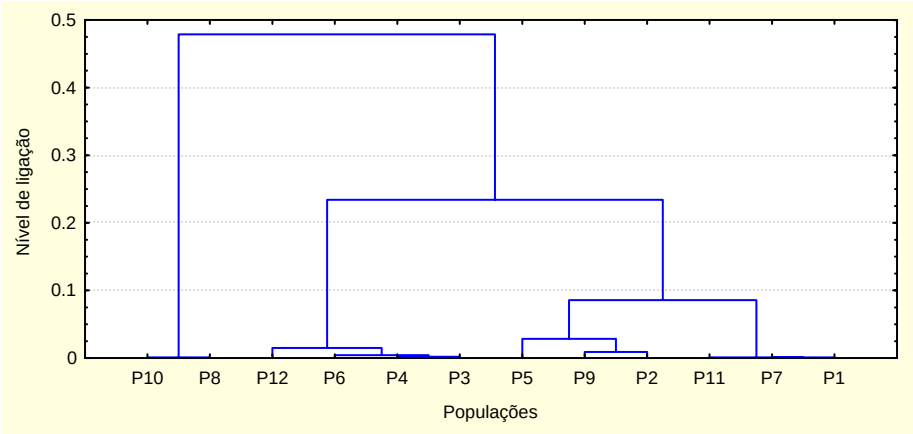


Figura 11 – Dendrograma resultante da análise de agrupamento das populações de girassol obtido do algoritmo "Average Linkage" baseado na distância Mahalanobis.

Para aplicação dos critérios seguintes, tornou-se necessário decidir quanto a uma métrica (distância) e a um algoritmo que obtiveram grupos geneticamente diferentes. Confrontando os resultados da dispersão gráfica pelos componentes principais (Figura 5), as estimativas das matrizes de dissimilaridades (Apêndices) e os dendrogramas (Figuras 6 a 11), notou-se existir uma tendência de populações mais similares se aproximarem, e de as menos similares se repelirem, em função da magnitude da divergência morfoagronômica. As diferentes medidas de dissimilaridades (euclideana e Mahalanobis) conduziram a diferentes padrões de agrupamento. Os agrupamentos mais estáveis em relação aos algoritmos utilizados foram mais consistentes na distância de Mahalanobis, o algoritmo "Average Linkage", constituindo o seguinte agrupamento:

Grupo 1 formado pelas populações P3, P4, P6, P12.

Grupo 2 formado pelas populações P2, P5, P9.

Grupo 3 formado pelas populações P1, P7, P11.

Grupo 4 formado pelas população P8, P10.

Neste sentido, o segundo critério utilizado na determinação do número de grupos foi a análise do comportamento de fusão (Figura 12). O gráfico da análise mostra o número de grupos em função o nível de fusão do agrupamento em cada estágio do procedimento e permite indicar o número final de grupos. Para isto, basta visualizar onde ocorre a mudança no segmento de reta.

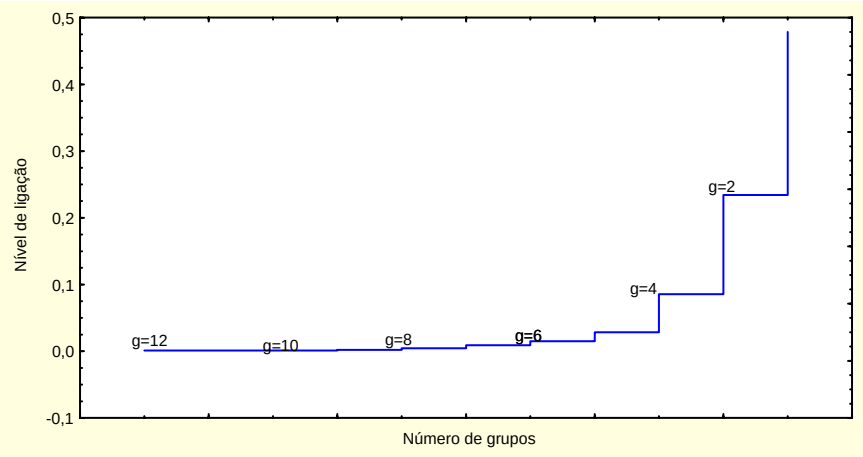


Figura 12- Gráfico da análise do comportamento do nível de fusão.

Na inspeção visual do gráfico da análise do comportamento do nível fusão, o segmento do traçado manteve-se estável no intervalo de 12 a 8 grupos, com pouca alteração no intervalo de 8 a 6 grupos e alterações significativas após o quarto grupo. O início da alteração no ponto de grupo igual a quatro indicou finalizar com quatro agrupamentos.

O terceiro critério para a determinação do número de grupos foi a análise do comportamento do nível de similaridade. A Tabela 8 apresenta o resultado do comportamento do nível de similaridade ($S_{l,k}$) no qual observou-se um decréscimo mais acentuado entre o 8^0 e 9^0 passo do procedimento, concluindo-se que o último agrupamento deve ser interrompido com três ou quatro grupos.

Tabela 8 - Nível de similaridade em relação à fusão das populações de girassol baseando-se na distância de Mahalanobis e algoritmo “Average Linkage”.

Passo	N^0 grupos	Fusão	Distância	$S_{l,k}$ Nível de ligação
1	11	P11 P7	0,001	99,9 %
2	10	P1 P11 P7	0,001	99,9 %
3	9	P8 P10	0,001	99,9 %
4	8	P3 P4	0,002	99,7 %
5	7	P3 P4 P6	0,0045	99,5%
6	6	P2 P9	0,009	99,0 %
7	5	P3 P4 P6 P12	0,015	98,4 %
8	4	P2 P9 P5	0,0285	97,1 %
9	3	P1 P7 P11 P2 P9 P5	0,085	91,0 %
10	2	P1P7 P11 P2 P9 P5 P3 P4 P6P12	0,24	75,5 %
11	1	Todos	0,47	52,0 %

Para compreender a operacionalização do procedimento considere-se a primeira fusão. Ou seja, no primeiro passo ocorreu a fusão entre as populações P7 e P11, pois estas apresentaram a menor distância. Substituindo-se os valores na fórmula (3.2.3.3.3), obtém-se o seguinte nível de similaridade $S_{7.11}$:

$$S_{(i=7,\ l=11)} = \left(1 - \frac{d_{i,l}}{\max\{d_{jk} = 1,2,.....12\}}\right). 100\ \%;\ \text{onde}\ \max\ d_{12,10} = 0{,}98$$

$$S_{P7,\ P11} = [\ 1 - \frac{0{,}001}{0{,}98}\ ]\ (100)\ \% = 0{,}999\ (\ 100\ )\% = 99{,}9\ \%$$

O procedimento de cálculo para as demais fusões acontece de maneira análoga e o resultado final pode ser representado graficamente.

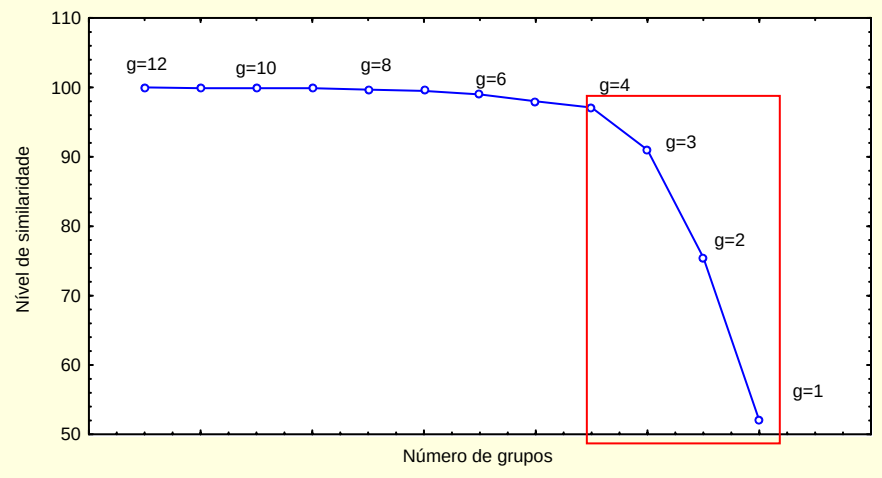


Figura 13 - Gráfico do nível de similaridade “versus” número de grupos**.**

A visualização da Figura 13 apresenta um forte decaimento nos últimos quatro pontos, indicando ser possível finalizar a análise com quatro grupos diferentes. Rosa Neto (2006) utilizou os resultados da análise do comportamento do nível de similaridade e os dendrogramas para decidir quanto ao número final de grupos nos dados de 40 estirpes de ribózio retiradas dos nódulos de feijão.

Um dos critérios utilizados para validar a análise de agrupamento consiste nos gráficos multivariados.

O gráfico silhueta averigua a qualidade dos agrupamentos. O procedimento confirma se uma população está mais próxima dos elementos do seu próprio grupo ou de elementos de grupos vizinhos (BARROSO; ARTES, 2003).

A Tabela 9 apresenta o resumo dos cálculos do valor silhueta, utilizando-se as distâncias euclideana e de Mahalanobis nos dados padronizados. O apêndice 2 apresenta, de forma didática, os cálculos detalhados para a construção do gráfico.

Tabela 9- Resumo dos cálculos e valores da silhueta para distâncias euclideana e Mahalanobis.

Pop	**a (i)**	**b (i)**	**S (i)**	**Pop**	**a(i)**	**b(i)**	**s(i)**
	Distância euclideana				Distância Mahalanobis		
P8	5,05	3,4	-0,356	P10	0,0010	0,100	0,990
P10	5,05	3,33	-0,307	P8	0,0005	0,120	0,995
P9	2,25	1,70	-0,135	P11	0,0010	0,090	0,989
P5	2,65	2,41	-0,05	P7	0,0010	0,080	0,988
P1	2,52	2,42	-0,03	P1	0,0010	0,073	0,986
P2	1,83	1,96	0,03	P3	0,0050	0,120	0,958
P6	1,49	1,69	0,04	P4	0,0060	0,085	0,929
P3	1,51	2,05	0,106	P12	0,0150	0,170	0,912
P12	1,39	2,36	0,181	P2	0,0095	0,080	0,881
P4	1,33	2,29	0,182	P6	0,0120	0,070	0,829
P7	1,54	2,23	0,196	P9	0,0240	0,047	0,489
P11	1,66	2,40	0,202	P5	0,0250	0,027	0,074

O valor da silhueta existe ($-1 \leq s(i) \leq +1$), os valores positivos indicam boa alocação no grupo, ao contrário dos valores negativos. O valor da silhueta para distância de Mahalanobis indicou boa alocação dos grupos, ao contrário da distância euclideana. Na construção dos gráficos silhueta, ordena-se as populações em ordem decrescente segundo o valor da silhueta encontrado. Cada população foi representada por uma barra horizontal, cujo comprimento é o valor da silhueta, como mostram as figuras 14 e 15 a seguir.

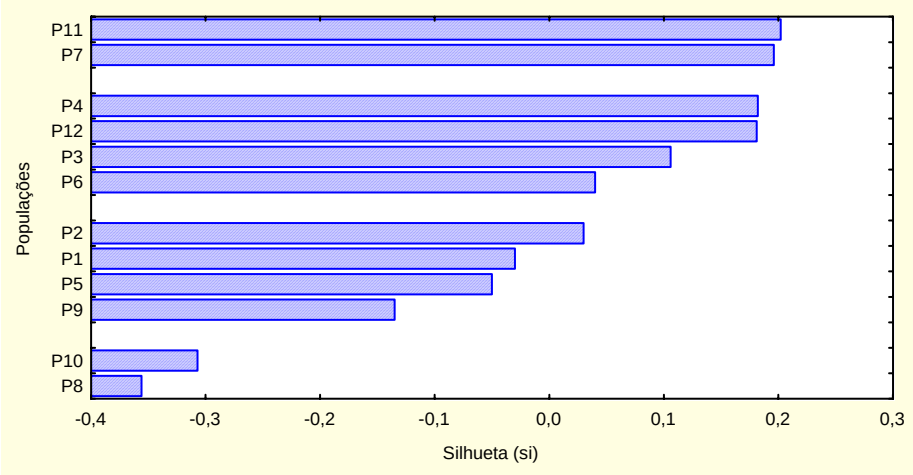


Figura 14 – Gráfico silhueta das populações empregando a distância euclideana.

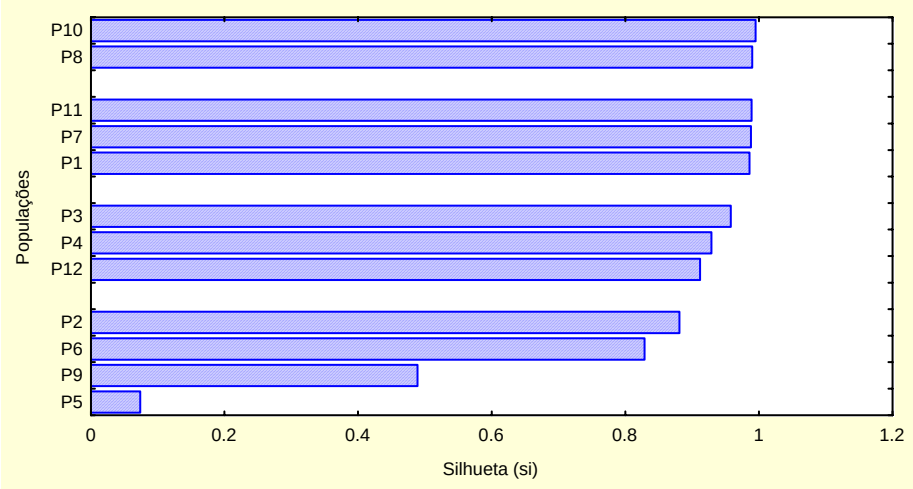


Figura 15 - Gráfico silhueta das populações empregando a distância Mahalanobis.

No gráfico silhueta da Figura 14, gerado pela distância euclideana, somente a população P1 posicionou-se num grupo diferente do estabelecido. As demais mantiveram-se estáveis nos grupos. No gráfico silhueta da Figura 15, gerado pela distância de Mahalanobis, somente a população P6 posicionou-se no grupo 2, por centésimos de diferença. Portanto pela distância de Mahalanobis foi possível verificar a qualidade dos agrupamentos e reconhecer um bom padrão de similaridade.

Para interpretação, utilizou-se o gráfico de perfil (Figura 16). Este apresentou a descrição dos caracteres em cada grupo, auxiliando nitidamente a interpretação dos grupos quanto aos valores médios dos caracteres. A visualização das variáveis tamanho do capítulo (X_2), curvatura (X_3), não contribuiram para discriminar os grupos.

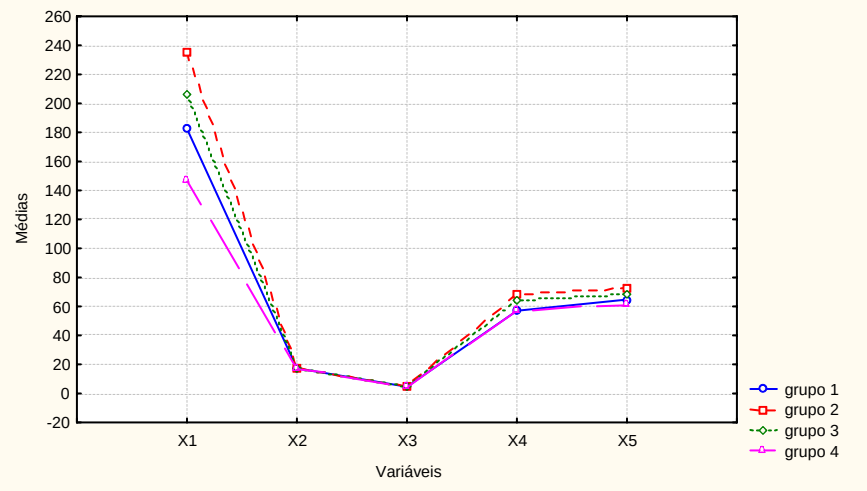


Figura 16- Perfis médios de agrupamento para solução de quatro grupos.

O **grupo 2** (P2, P5, P9) caracterizou-se pelas mais altas médias; o **grupo 3** (P1, P7, P11) caracterizou-se pelo segundo grupo com os valores médios mais altos; o **grupo 1** (P3, P4, P6, P12) caracterizou-se pelo terceiro grupo com os valores médios mais altos. O **grupo 4** (P8, P10) caracterizou-se pelas mais baixas médias envolvendo em todos momentos três variáveis discriminatórias: X_1-altura da planta; X_4- peso dos aqüênios por planta e X_5- número de dias para floração inicial, com exceção das variáveis X_3-curvatura e X_2- tamanho do capítulo.

4.3.2 Análise de variáveis canônicas

A análise da dispersão por variáveis canônicas constituiu-se num procedimento alternativo à análise de componentes principais As estimativas das variâncias extraídas para as variáveis canônicas e os escores relativos estão apresentados na Tabela 10.

No estudo, observou-se que as duas primeiras variáveis canônicas absorveram 86,5 % da variação acumulada, indicando a utilização dos dois primeiros eixos canônicos para a dispersão gráfica. De acordo com Cruz (1990), este resultado indica a existência da divergência genética entre essas populações. Assim, pode-se afirmar que pelo menos uma população, das 12 em estudo, apresenta divergência genética entre as demais.

Camarano (1997) utilizou essa metodologia para seis variáveis e agrupou 10 populações de girassol por meio de um eixo canônico, concluindo que o número pequeno de variáveis pode reduzir o número dos eixos canônicos.

Tabela 10 – Estimativas de variâncias (autovalores) associadas às variáveis canônicas, importâncias relativas e escores obtidos dos caracteres avaliados nas populações de girassol.

	Autovalores λ_i	Variância acumulada (%)	X_1	X_2	X_3	X_4	X_5
VC1	**26,8**	**67,0**	1,30	0,40	-0,50	0,67	0,95
VC2	**7,8**	**19,5**	1,75	-0,30	-1,94	0,79	-0,42

A dispersão gráfica para as duas primeiras variáveis canônicas (Figura 17) mostrou a posição geométrica das populações. A esse respeito, constata-se uma transposição da divergência genética do espaço p-dimensional (p=5) para bidimensional. Conforme pode-se observar na Figura 17, quatro grupos foram identificados: o grupo1 (P3,P4,P6,P12), o grupo 2 (P2, P5 e P9), o grupo 3 (P1, P7, P11) e o grupo 4 (P8, P10). O grupo 4 caracterizou-se pelas variedades para o mercado confeiteiro (girassol in natura).

Este resultado apresentou boa concordância com os da análise de agrupamento para a distância de Mahalanobis nos algoritmos “Single Linkage” (Figura 9) e “Average Linkage” (Figura 11). Quanto aos dos agrupamentos obtidos pela distância euclideana no algoritmo “Complete Linkage” (Figura 7) e “Average Linkage” (Figura 8), a população P1 posicionou-se no grupo 4 (P8, P10), e a população P5 posicionou-se no grupo 3 (P1, P7, P11), as quais não estão muito distantes no gráfico (Figura 17).

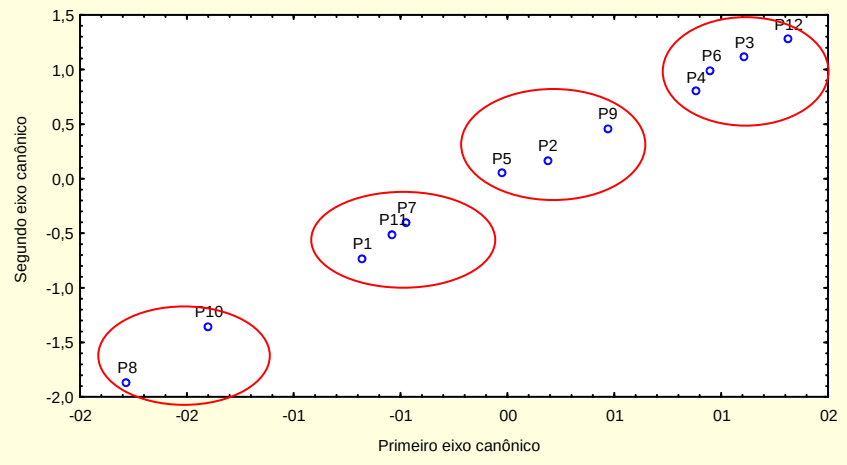


Figura 17 – Dispersão das populações em relação aos primeiros eixos canônicos.

A divergência genética foi expressa pela inspeção visual das técnicas, utilizando os escores dos dois primeiros componentes principais (Figura 5) e os dois primeiros eixos canônicos (Figura 17), apresentando satisfatória concordância dos agrupamentos. Neste contexto, uma análise comparativa com as medidas de dissimilaridades entre cada par de variedades (Apêndices) identificou-se que o grupo 4 (P8, P10) foi mais divergente em relação ao grupo 1 (P3, P4, P6, P12). O grupo 2 (P1, P7, P11) foi menos divergente em relação ao grupo1 (P3, P4, P6, P12), tanto pela distância euclideana, como pela distância de Mahalanobis.

De forma semelhante, Miranda et al. (2003), utilizando as duas primeiras variáveis canônicas, representou 96,5% da variabilidade genética de nove cultivares de milho de pipoca, mostrou que o resultado teve grande concordância com a análise de agrupamento.

Uma conclusão lógica seria a recomendação dos cruzamentos entre os materiais mais divergentes, uma vez que é esperado o maior efeito heterótico entre populações geneticamente mais contrastantes (FALCONER, 1981).

4.3.3 Análise de Variância Multivariada

A técnica da análise de variância multivariada foi aplicada no intuito de verificar a homogeneidade interna dos agrupamentos, ou seja, a variação genética intragrupos (dentro dos grupos). Os resultados obtidos da análise de componentes principais, análise de agrupamento e análise por variáveis canônicas foram a formação de quatro grupos geneticamente diferentes como mostra a Tabela 11.

Tabela 11 – Agrupamentos formados das populações de girassol estabelecidos pela distância Mahalanobis e do algoritmo "Average Linkage".

Grupos	Populações			
Grupo 1	**P3**- Local Blue;	**P4**- Klein A;	**P6**- Majak;	**P12**- Guayacan
Grupo 2	**P2**- Sunfola;	**P5**- Belensky **;**	**P9**- PI34804	
Grupo 3	**P1**- Sundak;	**P7**- Armovensk;	**P11**- PI70389	
Grupo 4	**P8**- Gigante;	**P10**- PI170389		

Os pressupostos para a análise de variância multivariada foram analisados para cada grupo. Quanto à normalidade, o teste de Lilliefors foi aplicado nas 60 combinações (cinco variáveis nas doze populações), onde 50 destas foram significativas ($p< 0,01$) e 10 apresentaram normalidade, como mostra a Tabela 12.

Tabela 12 – Combinações que apresentaram normalidade pelo teste Lilliefors.

Variáveis	**X1**				**X2**		**X4**			**X5**
Populações	P2	P3	P6	P9	P3	P2	P3	P6	P1	P3
DM	0,06	0,07	0,05	0,05	0,06	0,04	0,05	0,04	0,06	0,06
Vcritíco	0,08	0,08	0,08	0,08	0,07	0,08	0,08	0,08	0,08	0,08
p-value	p>0,01	p>0,01	p>0,01	p>0,01	p>0,01	p>0,01	p>0,01	p>0,01	p>0,01	p>0,01

O teste estatístico da assimetria e curtose foi aplicado nas 60 combinações dos experimentos e resultaram na rejeição da hipótese de nulidade em 46 combinações. Uma alternativa foi transformar os dados utilizando a raiz quadrada, de modo a torná-los normalizados.

Verificada a normalidade dos dados, aplicou-se o teste de Box para verificação da igualdade das matrizes de covariâncias dos grupos.

Tabela 13 – Teste de Box- Igualdade das matrizes de covariâncias.

Grupos	Teste M de Box	p-value
Grupo 1	172,36	0,000015
Grupo 2	39,74	0,00059
Grupo 3	166,70	0,00001
Grupo 4	152,26	0,00001

Os resultados apresentados na Tabela 13 mostram que não houve igualdade das matrizes de covariâncias para os quatro grupos, mas segundo Mardia et al. (1979) esse pressuposto não é afetado quando o número de repetições for grande. Neste ponto, torna-se difícil uma rica discussão pois poucos trabalhos científicos, na área de melhoramento genético, verificaram os pressupostos de homogeneidade e normalidade.

Por meio da análise de variância multivariada aplicada para cada grupo foram evidenciadas diferenças significativas entre os vetores de médias das populações de girassol, conforme mostra a Tabela 14.

Tabela 14 - Resultado do teste de Wilks aplicado nos quatro grupos.

Grupo	Teste Λ	Valor	F	P value
Grupo 1	Wilks	0,125	0,387	0,877
Grupo 2	Wilks	0,003	14843	0,000067
Grupo 3	Wilks	0,25	0,500	0,750
Grupo 4	Wilks	0,0017	150,3	0,062

Nos grupos 1, 3 e 4 não houve a rejeição da hipótese nula, ou seja, os vetores de médias não são estatisticamente diferentes. No grupo 2, houve a rejeição da hipótese nula e, portanto, os vetores de médias são estatisticamente diferentes.

Num enfoque diferente, a análise de variância multivariada foi aplicada de forma inédita na área agronômica, posteriormente a técnicas exploratórias multivariadas, no sentido confirmatório da divergência genética dos grupos selecionados, apoiando-se na estatística inferencial. Daoyu e Lawes (2000) e Ferreira (2001) aplicaram essa metodologia com o objetivo de detectar a diferença entre os vetores de médias de todas variedades e, assim, justificar-se da necessidade da busca da divergência genética ou da redução dimensional e descartes de variáveis. Ambos objetivos são válidos como sugestão para um programa de melhoramento genético.

A solução final, nesta discussão, diz respeito a escolha das populações mais promissoras para garantir o sucesso no programa de melhoramento. O resultado da análise de variância detectou diferenças significativas entre as populações para algumas variáveis. Portanto o material em estudo possui diversidade para selecionar populações e iniciar cruzamentos envolvendo populações que se posicionaram em grupos diferentes. Miranda et al. (2003) orientou realizar os cruzamentos entre populações que apresentaram ampla divergência nas características de interesse.

A sugestão de orientação do melhoramento genético para a EMBRAPA /Soja de Londrina, visando auxiliar pesquisadores na identificação de materiais mais promissores para programa de melhoramento genético, a orientação seria realizar o cruzamento entre os grupos mais divergentes, apresentando o seguinte resultado:

Grupo 1 (P3, P4, P6, P12) em relação ao grupo 4 (P8, P10) ou;

Grupo 2 (P2, P5, P9) em relação ao grupo 4 (P8, P10) pelas distâncias euclideana e Mahalanobis.

A orientação para não realizar o cruzamento são os grupos menos divergentes, apresentando o seguinte resultado:

Grupo 1 (P3, P4, P6, P12) em relação ao grupo 2 (P2, P5, P9) pela distância euclideana;

Grupo 2 (P2, P5, P9) em relação ao grupo 3 (P1, P7, P11) pela distância de Mahalanobis.

5 CONCLUSÃO

Os resultados dos procedimentos exploratórios e analíticos permitiram as seguintes conclusões:

- As técnicas multivariadas indicaram divergência genética das populações de girassol.
- A inspeção visual de gráficos de dispersão, por intermédio dos escores em eixos cartesianos, representados pelos dois primeiros componentes principais e pelas duas primeiras variáveis canônicas, apresentaram resultados semelhantes nos agrupamentos formados, mostrando satisfatória concordância dos grupos formados.
- Os gráficos multivariados utilizados na análise de agrupamento auxiliaram na interpretação e definição final quanto ao número de grupos.
- Os resultados dos três métodos (análise de componentes principais, análise de agrupamento e análise de variáveis canônicas) foram semelhantes, mantendo-se um padrão de agrupamento estável.
- A análise de variância multivariada identificou que as populações de girassol dentro de cada grupo possuem vetores de médias iguais, exceto no grupo 2.

REFERÊNCIAS

ADUGNA, W.; LABUSCHAGNE, M. T. Cluster and canonical variate analysis in multilocation trials of linseed. **Journal of Agricultural Science**, Cambridge, v. 140, n. 3, p. 297-304, May 2003.

AGONG, S. G., SCHITTENHELM, S.; FRIEDT, W. Genotypic variation of Kenyan tomato (*Lycopersicon esculentum L.*). **Plant Genetic Resources Newsletter,** Rome., n. 123, p. 61-67, 2000.

ALDENDERFER, M. S.; BLASHFIELD, R. K. **Cluster Analysis.** Beverly Hills, CA: Sage Publications**,** 1984.

ALDENDERFER, M. S.; BLASHFIELD, R. K. **Cluster Analysis.** Beverly Hills, CA: Sage University. 1985. (Papers, 44).

ALPERT, M. I.; PETERSON, R. A. On the interpretation of canonical analysis. **Journal of Marketing Research**, Chicago**,** v. 9, n. 2, p. 187-192, May 1972.

ALVES, R. M. **Caracterização genética de populações de cupuaçuzeiro, *Theobroma grandiflorum Schum*., por marcadores microsatélites e descrição botâncio-agronômicos**. 2005. Tese (Doutorado em Agronomia - Genética e Melhoramento de Plantas) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo – ESALQ/USP, Piracicaba.

ALVES, R. M. et al. Seleção de descritores botânico-agronômicos para a caracterização de germoplasma de cupuaçuzeiro. **Pesquisa Agropecuária Brasileira**, Brasília, DF, v.38, n.7, p.807-818, jul. 2003.

ANDENBERG, M. R. **Cluster analysys for aplications**. New York: Academic Press, 1973.

ANDERSON, T. W. **An introduction to multivariate statistical analysis**. New York: John Wiley, 1958.

_____. **An introduction to multivariate statistical analysis**. 2. ed. New York: John Wiley, 1984.

ANDERSON, R. E.; BLACK, W. C.; HAIR, J. F. Jr., TATHAM, R. L. **Análise multivariada de dados.** 5. ed. Porto Alegre: Bookman, 2005.

ANNICCHIARICO, P. Cultivar adaptation and recommendation from alfafa trials in northern Italy. **Journal of Genetic and Breeding**, Rome, v. 46, n. 3, p. 269-277, 1992.

ASTAFEIEF, N. C. **Identificação, em hidroponia, de genótipos de girassol (*Helianthus annuus annuus* –L) tolerantes ao alumínio**. 1997. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Universidade Estadual de Londrina, Londrina.

BARBIN, D. **Planejamento e Análise Estatística de Experimentos Agronômicos**. Arapongas: Editora Midas, 2003.

BARROSO, L .P.; ARTES, R. **Análise multivariada**: Minicurso do 10 Simpósio de Estatística Aplicada à Experimentação Agronômica – RBRAS, 48 Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria- SEAGRO. Lavras: UFLA, 2003.

BARTLETT, M. S. Properties of sufficiency and statistical tests. **Proceedings of the Royal Society of London., Serie A**, London, v. 160, n. 901, p. 268-282, May 1937.

BENIN, G. et al. Comparações entre medidas de dissimilaridade e estatísticas multivariadas como critérios no direcionamento de hibridações em aveia. **Ciência Rural**, Santa Maria, v. 33, n. 4, p. 657-662, jul./ago. 2003.

BOX, G. E. P. A General distribution theory for class of likelihood criteria. **Biometrika**. v. 36, n. 3/4, p. 317-346, Dec. 1946.

BOX, G. E. P. Problems in the analysis of growth and wear curves. **Biometrics**, v. 6, n. 4, p. 362-389, Dec. 1950.

BUSSAB, W. O.; ANDRADE, D. F.; MYAZAKY, E. S**.** Introdução à análise de agrupamentos. **Associação Brasileira de Estatística**, São Paulo: IME/USP, 1990.

BUZZETTI, A. R. Cresce a produção de girassol. **Óleos e Grãos**, São Bernardo do Campo, v. 8, n. 46, p. 34-38, 1999.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics**, Hamilton, Ontário, v. 3, n. 1, p. 1-27, 1974.

CAMARANO, L. F. Estudo da divergência genotípica entre populações de girassol (Helianthus annuus L.). 1997. Dissertação (Mestrado em Agronomia, Área de Concentração: Genética e Melhoramento de Plantas) - Universidade Federal de Goías, Goiânia.

CAMPOS, H. **Estatística experimental não paramétrica**. 4. ed. Piracicaba: ESALQ, 1983.

CARVALHO, L. P. et al. Análise da diversidade genética entre acessos de banco ativo de germoplasma de algodão. **Pesquisa agropecuária Brasileira**, Brasília, v. 38, n. 10, p. 1149-1155, 2003.

CATTELL, R. B. The screen test for the number of factor. **Multivariate Behavioral Research**, Mahwah, NJ, v. 1, p.140-161, 1966.

CHATFIELD, C.; COLLINS, A. J. **Introduction to multivariate analysis.** London: Chapman & Hall, 1980.

CORMACK,R. A review of classification. **Journal of the Royal Statistical Society**. Serie A, London, n. 134, p. 321-367, 1971.

CRUZ, C. D. **Aplicações de algumas técnicas multivariadas no melhoramento de plantas**. 1990. Tese (Doutorado em Melhoramento Genético) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo - ESALQ/USP, Piracicaba, SP.

CRUZ, C. D.; REGAZZI, A. J. **Modelos biométricos aplicados ao melhoramento genético**. 2. ed. Viçosa: Editora UFV, 1997.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: Editora UFV, 2003. v. 2.

CUNHA, R.A. **Contribuição ao estudo da taxonomia dos Heliponinae** (*Hymenoptera – Apidae)*. 1969. Tese (Doutorado) – Faculdade de Filosofia, Ciências e Letras de Rio Claro, Rio Claro.

DEMÉTRIO, C.G.B**. Análise multidimensional para dados de cana-de-açucar**. 1985. Tese (Doutorado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba- SP.

DIAS, L. A S. **Divergência genética e fenética multivariada na predição de híbridos e preservação de germoplasma de cacau (*Theobroma cacao L.*)**. Piracicaba, 1994. Tese (Doutorado em Agronomia, Área de concentração: Genética e Melhoramento de Plantas) - Escola Superior de Agricultura "Luiz de Queiroz",Universidade de São Paulo, São Paulo.

DILLON, W. R.; GOLDSTEIN, M. **Multivariate analysis**: methods and applications. New York: John Wiley & Sons, 1984.

DUARTE, J.M. Estudo da divergência genética em raças de feijão por meio de marcadores RAPD. 1998. 78f. Tese (Doutorado em Agronomia, Área de Concentração em Genética e Melhoramento de Plantas) - Universidade Federal de Lavras.

DUARTE, J. M.; SANTOS, J. B. dos; MELO, L. C. Comparison of similarity coefficients based on RAPD markers in the common bean. **Genetics and Molecular Biology**, Ribeirão Preto, v. 22, n. 3, p. 427-432, Sept. 1999.

EMYGDIO, B. M.; ANTUNES, I. F.; CHOER, E.;NEDEL, J.L.. Eficiência de coeficientes de similaridade em genótipos de feijão mediante marcadores RAPD. **Pesquisa Agropecuária Brasileira**, Brasília, v. 38, n. 2, p. 243-250, fev. 2003.

EVERITT, B. S. **Cluster analysis**. 2nd. ed. New York: Halsted Press, 1980.

EVERITT, B. S.; DUNN, G. (Ed.). **Applied multivariate data analysis**. 4th. ed. London: Arnold, 1996.

EVERITT, B. S.; LANDAU, S.; LEESE, M. **Cluster Analysis**. 4th. ed. London: Arnold, 2001.

EVERITT, B. S.; DER, G. **Statistical Analysis of Medical Data Using SAS**. New York: Chapman & Hall/CRC, 2006.

FALCONER, D.S**. Introduction to quantitative genetics.** 2nd. ed. London: Longman, 1981.

FERREIRA, C. A. **Utilização de técnicas multivariadas na avaliação da divergência fenética entre clones de palma forrageira (*Opuntia e Nopalea*)**. 2001. Dissertação (Mestrado em Biometria) - Universidade Federal Rural de Pernanbuco, Recife.

FERREIRA, D. F., CANTELMO, N. F. Desempenho de testes de normalidade multivariada avaliado por simulação Monte Carlo. In: **51ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria**, 51., 2006, Botucatu. Departamento de Bioestatística –Instituto de Biociências. Botucatu-SP. 2006.

FICK, G. N. Breedings and genetics. In: CARTER, J. F. (Ed.). **Sunflower science and technology**. Madison: American Society of Agronomy, 1978. p. 279-338.

FISHER, R.A. The general sampling distribuition of the multiple correlation coefficient. **Proceedings of the Royal Society of London**, Serie A, London, v.121, n. 788, p. 654-673, Dec. 1928.

FLURY, B.; RIEDWYL, H**. Multivariate statistics**: a practical approach. New York: Chapman and Hall, 1988.

FREI, F. **Introdução à análise de agrupamento**: teoria e prática. São Paulo: Editora Unesp, 2006.

GAMA, M. de P. **Bases da análise de agrupamento (Cluster Analysis)**. 1980. Dissertação (Mestrado em Estatística e Métodos Quantitativos) – Universidade de Brasília, Brasília.

GUERRA, E. P.; PICKSIUS, A. Avaliação de genótipos de girassol de ensaio conduzido na PUC-PR. In: Reunião Nacional de Pesquisa de Girassol, 16.; Simpósio Nacional Sobre Cultura do Girassol, 4., Londrina. **Anais...**. Londrina: EMBRAPA/ Soja, 2005. p.74-75.

GODOI, C. R.de M. **Análise estatística multidimensional**. Piracicaba: ESALQ/USP, 1985.

GOU, R. L., SONG, C. J. A study on genetic divergence of quantitative characters and their cluster in winter wheat parents. **Acta Agriculturae Boreali Sinica,** China, v. 6, n. 3, p. 1-6, 1991.

GOWER, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. **Biometrika**, London, v. 53, n. ¾, p. 325-338, Dec. 1966.

HARRIS, R.J. **A primer of multivariate statistics**. New York: Academic Press, 1975.

HAIR JR, J. F. et al. **Multivariate Data Analysis**. 5th. ed. Upper Saddle River: Prentice Hall, 1998.

HAIR JR, J.F. et al. **Análise multivariada de dados.** 5. ed. Porto Alegre: Bookman, 2005.

HORN, J. L.; ENGSTROM, R. Cattell's scree test in relation to Bartlett's chi square test and other observations on the number of factors problem. **Multivariate Behavioral Research**, Fort Worth, Tex., US, v. 14, p. 283-300, 1979.

HOTELLING, H. Analysis of a complex of statistical variables in to principal components. **Journal of Education Psychology**, Columbia, v. 24, p. 417-441, 1933.

_____. The most predictable criterion. **Journal of Education Psychology**, Columbia, v. 26, p.139-142, 1935.

_____. Simplified calculation of principal components. **Psychometrika**, n. 1, p. 27-35, 1936.

JEFFER, J. N. R. Two cases studies in the application of principal component analysis. **Applied Statistics**., v. 16, n. 3, p. 225-236, 1967.

JOHNSON, R. A; WICHERN, D.W. **Applied multivariate statistical analysis**. 2nd. ed. Englewood Cliffs, NJ: Prentice Hall, 1988.

_____. **Applied multivariate statistical analysis**. 3th. ed. Englewood Cliffs, NJ: Prentice Hall, 1992.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, Massachussetts, v. 23, n. 1, p.187-200, Jul. 1958.

KENDALL, M. G. **A course in multivariate analysis**. New York: Hafner Publishing, 1957.

_____. **Multivariate Analysis.** New York: Hafner Press, 1975.

LAL, R. K.; SHARMA, J. R.; SINGH, N. Genetic variability and diversity pattern in chamomille (*Chamomilla recutita*). **Journal of Medicinal and Aromatic Plant Sciences**. v. 23, n. 2, p. 53-58, 2001.

LEDO, C. A S. **Análise de variância multivariada para cruzamentos dialélicos**. Lavras, 2002. 126f. Tese (Doutorado em Agronomia, Área de Concentração em Genética e Melhoramento de Plantas) - Universidade Federal de Lavras.

LEITE, R. M. V. L. C.; BRIGHENTI, A. M; CASTRO, C. **Girassol no Brasil**. EMBRAPA/Soja Londrina. 2005. 641p.

LEITE, L. M. R. M. **Uso de técnicas multivariadas no estudo morfométrico de *Didelphis albiventris* Lund, 1841 (Marsupialia, Didelphidae) no estado de Pernambuco**. 2000.

Dissertação (Mestrado em Biometria, Área de concentração: Métodos Estatísticos Aplicados as Ciências Biológicas) – Universidade Federal Rural de Pernambuco, Recife.

LILLIEFORS, H. W. On the Kolmogorov- Smirnov Test for normality with mean and variance unknown. **Journal of the American Statistical Association**, Washington, n. 62, p. 399-402, 1967.

MacQUEEN, J. Some methods for classification and analysis of multivariate observations. In: LeCAM, L. M.; NEYMAN, J. (Ed.). **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. Berkeley: University of California Press, 1967. v. 1, p. 281-97.

MAHALANOBIS, P. C. On the generalizad distance to statistics. **Proceedings of the National Institute of Science of India**, v. 12, p. 49-55, 1936.

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

_____. **Multivariate Analysis.** 6th London: Academic Press**,** 2003.

MELO, F. H. **Desempenho recente da agricultura brasileira.** São Paulo: EDUSP, 1992.

MELO,W. M. C**. Divergência genética e capacidade de combinação entre híbridos de milho**. Lavras. 2000. Tese (Doutorado em Agronomia – Genética e Melhoramento de Plantas) – Universidade Federal de Lavras, Lavras.

MESSETTI, A. V. L. **Estudo da semelhança de genótipos de girassol (*Helianthus annuus L.*) com o uso da distância generalizada de Mahalanobis na análise de agrupamento.** 2000. Dissertação (Mestrado em Agronomia – Energia na Agricultura). Faculdade de Ciências Agronômicas - Universidade Estadual Paulista, Botucatu-SP.

MEYER, A. S. **Comparação de coeficientes de similaridade usados em análise de agrupamento com dados de marcadores moleculares dominantes**. 2002. Dissertação (Mestrado em Agronomia – Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

MINGOTTI, S. A. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: UFMG, 2005.

MIRANDA, G. V. et al. Potencial de melhoramento e divergência genética de cultivares de milho de pipoca. **Pesquisa Agropecuária Brasileira**, Brasília, v. 38. n. 6, p. 681-688, 2003.

MOHAMMADI, S. A; PRASANNA, B. M. Analysis of genetic diversity in crop plants-salient statistical tools and considerations. **Crop Science**, Madison, v. 43, p. 1235-1248, 2003.

MOREIRA, G. R. et al. Divergência genética e subcoleção representativa de populações da traça-do-tomateiro. **Pesquisa Agropecuária Brasileira**, Brasília, v. 39, n. 5. p. 437-443, 2004.

MORRISON, D. F. **Multivariate statistical methods.** New York: McGraw-Hill, 1967.

_____. **Multivariate statistical methods**. 2nd. ed. New York: McGraw-Hill Book, 1976.

MOURA, E. F. **Divergência genética entre acessos de Jaborandi** *(Pilocarpus microphyllus*). 2003. Dissertação (Mestrado em Agronomia. – Genética e Melhoramento de Plantas). Universidade Federal de Lavras, Lavras – MG.

MOREIRA, G. R. et al. Divergência genética e subcoleção representativa de populações da traça-do-tomateiro. **Pesquisa Agropecuária Brasileira**, Brasília, v.39, n.5, p. 437-443, maio 2004.

NORMAN, G. R.; STREINER, D. L. Biostatistics: the bare essentials. St Louis: Mosby Year Book, 1994.

OSHIIWA, M. **Desenvolvimento de programa computacional para delineamento em blocos casualizados com respostas multidimensionais e sua aplicação em ensaios agronômicos**. 2001. Dissertação (Mestrado em Agronomia – Energia na Agricultura). Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu-SP.

PEARSON, K. On lines and planes of closet fit to systems of points in space. **Philosophical Magazine**, v. 2, p. 559-572. 1901.

PEARSON, K. On lines and planes of closet fit to systems of points in space. **London, Edinburg and Dublin Philosophical Magazine and Journal of Science**, v. 2, p. 559-572, Sixth Series, Jul./Dec. 1901.

PISANI, J.F. **Análise estatística multidimensional aplicada a problemas de acasalamentos recíprocos em avicultura.** 1973**.** Tese (Doutorado em Ciências) – Departamento de Genética, Evolução e Bioestatística da Faculdade de Filosofia, Ciências e Letras de Rio Claro, Rio Claro.

RAO, R. C. **Advanced statistical methods in biometric research**. New York: J. Willey, 1952.

_____. **Advanced statistical methods in biometric research**. New York: Hafner Press, 1974.

REIS, E. **Estatística multivariada aplicada**. Lisboa: Silabo, 1997.

REIS, W.P.; VELLO, N.A.; FERREIRA, D.F. et al. Associação entre as estimativas dos coeficientes de parentesco e de técnicas multivariadas como medida de divergência genética em cultivares de trigo. **Ciências e Agrotécnica**, Lavras, v.23, n.2, p.258-263, abr. 199.

ROJAS, W.; BARRIGA, P.; FIGUEROA, H. Multivariate analysis of the genetic diversity of Bolivian quinua germplasm. **Plant Genetic Resources Newsletter,** Bolívia, n. 122, p. 16-23, 2000.

ROSA NETO, E.A.R. **Algoritmos aglomerativos para análise de agrupamentos**: exemplos e aplicações. 2006. Trabalho de Conclusão de Curso (Graduação em Matemática, Habilitação Bacharelado). – Universidade Estadual de Londrina, Londrina, PR. 2006.

REUNIÃO NACIONAL DE PESQUISA DE GIRASSOL, 16.; SIMPÓSIO NACIONAL SOBRE A CULTURA DO GIRASSOL, 4., 2005, Londrina. **Anais...** Londrina: EMBRAPA/Soja, 2005.

SCHNEITER, A. A.; MILLER, J. F. Description of Sunflower Growth Stages. **Crop Science**, Madison, v. 21, p. 901-903, 1981.

SCREMIN, M. A. A. **Método para a seleção do número de componentes principais com base na lógica difusa**. 2003. Tese (Doutorado em Engenharia de Produção) – Universidade Federal de Santa Catarina, Florianópolis.

SEBER, G. A. F. **Multivariate observations**. New York: John Wiley & Sons, 1984.

SILVA, N. R. **Aplicativo computacional para utilização de componentes principais em experimentação Agronômica**. 2005. Dissertação (Mestrado em Agronomia) – Universidade Estadual Paulista, Botucatu.

SILVA, E. L. **Utilização de técnicas multivariadas no estudo morfométrico da Albacora laje (*Thunnus albacares Bonnaterre*, 1788) do Atlântico**. 2000. Dissertação (Mestrado em Biometria) – Universidade Federal Rural de Pernambuco, Recife.

SKORIC, D. Achievements and future directions of sunflower breeding. **Field Crops Research**, Amsterdam, n. 30, p. 231-270, 1992

SNEATH, P. H. A.; SOKAL, R. R. **Numerical taxonomy:** the principles and pratice of numerical classification. San Francisco: W.H. Freeman and Company, 1973.

SOUSA, N. R. **Variabilidade genética e estimativas de parâmetros genéticos em germoplasma de guaranazeiro**. 2003. Tese (Doutorado em Agronomia – Genética e Melhoramento de Plantas) - Universidade Federal de Lavras, Lavras.

SOUZA, R. F. **Análise da variabilidade genética em acessos e cultivares de girassol (***Helianthus annuus annuus* **–L.) através de marcadores isoenzimáticos e de microssatélites.** 2001. Tese (Doutorado em Ciências: Genética) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo - FMRP-USP, Ribeirão Preto.

SOUZA, S. R. **Classificação de variedades da soja quanto à concentração de isoflavonóides utilizando análise estatística multivariada.** 2004. Monografia (Especialização em Estatística) – Universidade Estadual de Londrina, Londrina.

SOKAL, R. R.; ROHLF, F. J. The comparison of dendograms by objective methods. **Taxonomy**, Washington, v. 11, p. 33-40, Feb. 1962.

SOKAL, R. R.; SNEATH, P. H. A. **Principles of numerical taxonomy**. San Francisco: W. H. Freeman, 1963.

SPEARMAN, C. General intelligence objectivelly determined and measured. **American Journal of Psychology**, Chicago, v. 15, p. 201-293, 1904.

STEWART, D.; WILLIAM, L. A general canonical correlation index. **Psychological Bulletin** Washington, v. 70, p.160-63, 1968.

STRAPASSON, E. **Seleção de descritores na caracterização de germoplasma de *Paspalum* através de componentes principais**. 1997. Dissertação (Mestrado em Agronomia – Estatística e Experimentação Agronômica) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, ESALQ/USP, Piracicaba.

TABACHNICH, B.G.; FIDELL, L. S. **Using multivariate statistics**. 4th ed. Boston: Allyn & Bacon, 2001.

TOTTI, R. **Utilização de métodos de agrupamento hierárquico em acessos de paspalum (Graminea poaceae).** 1998. Dissertação (Mestrado em Agronomia – Estatística e Experimentação Agronômica) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, ESALQ/USP, Piracicaba.

VAN LAAR, A. Multivariate analysis a way to better understanding of complexity. **South African Forestry Journal**, Pretoria, v. 141, p. 34-41, 1987.

_____. Forest biometry. Sappi Forests: Stellenbosch, 1991.

WARD, J. H. Hierarchical grouping to optimize an objetive function. **Journal of the American Statistical Association**, Washington, v. 58, p. 236-244, 1963.

WILKS, S. S. Certain generalization in the analysis of variance**. Biometrika**, London, v. 24, n. 3/4, p. 471-494, 1932.

DAOYU, Z.; LAWES, G. S. Manova and discriminant analysis of phenotypic data as a guide for parent selection in Kiwifruit (Actinida deliciosa) breeding. **Euphytica**, Wageningen, v. 114, n. 2, p. 151-157, Jul. 2000.

APÊNDICE

APÊNDICE 1 – Medidas descritivas multivariadas

Vetor de Médias

$$\overline{X} = \begin{bmatrix} 200{,}15 & 17{,}20 & 4{,}5 & 62{,}80 & 68{,}80 \end{bmatrix}$$

Matriz de Variâncias e Covariâncias

$$S = \begin{bmatrix} 1120{,}75 & 3{,}50 & 3{,}30 & 149{,}0 & 165{,}4 \\ 3{,}50 & 0{,}40 & -0{,}035 & 2{,}50 & 0{,}910 \\ 3{,}30 & -0{,}035 & 0{,}09 & -0{,}55 & -0{,}056 \\ 149{,}0 & 2{,}50 & -0{,}55 & 45{,}17 & 28{,}50 \\ 165{,}4 & 0{,}91 & -0{,}056 & 28{,}50 & 30{,}75 \end{bmatrix}$$

$s_1 = 33{,}48$ $\quad CV_1 = 16{,}73\ \%$

$s_2 = 0{,}63$ $\quad CV_2 = 3{,}68\ \%$

$s_3 = 0{,}30$ $\quad CV_3 = 6{,}67\ \%$

$s_4 = 6{,}72$ $\quad CV_4 = 10{,}70\ \%$

$s_5 = 5{,}55$ $\quad CV_5 = 8{,}06\ \%$

Matriz de Correlação

$$R = \begin{bmatrix} 1{,}00 & 0{,}165 & 0{,}33 & 0{,}66 & 0{,}89 \\ 0{,}165 & 1{,}00 & -0{,}189 & 0{,}60 & 0{,}26 \\ 0{,}33 & -0{,}189 & 1{,}00 & -0{,}28 & -0{,}03 \\ 0{,}66 & 0{,}60 & -0{,}28 & 1{,}00 & 0{,}76 \\ 0{,}89 & 0{,}26 & -0{,}03 & 0{,}76 & 1{,}00 \end{bmatrix}$$

APÊNDICE 2 – Análise de agrupamento.

Medidas de dissimilaridades entre populações de girassol, obtidas pela distância de Mahalanobis acima da diagonal principal e distância euclideana abaixo da diagonal principal.

0	0,07	0,35	0,30	0,02	0,26	0,001	0,10	0,13	0,11	0,001	0,45
2,29	0	0,10	0,08	0,01	0,06	0,08	0,40	0,009	0,37	0,09	0,16
2,58	2,49	0	0,002	0,20	0,007	0,40	0,80	0,05	0,80	0,40	0,006
2,83	1,97	1,14	0	0,15	0,002	0,30	0,80	0,03	0,80	0,30	0,014
2,82	2,23	3,28	2,89	0	0,13	0,03	0,25	0,04	0,24	0,03	0,26
2,77	1,37	1,79	1,49	2,40	0	0,28	0,73	0,02	0,70	0,29	0,025
2,42	1,92	3,87	3,57	1,94	2,89	0	0,10	0,14	0,10	0,001	0,47
2,15	4,28	4,18	4,70	4,18	4,54	3,91	0	0,50	0,001	0,10	0,97
2,14	1,43	1,93	2,03	3,08	1,31	2,85	3,88	0	0,50	0,15	0,09
4,20	3,98	5,99	5,84	4,98	5,01	3,14	5,05	4,29	0	0,09	0,98
2,65	1,87	4,07	3,70	2,48	3,03	0,67	4,24	2,85	2,67	0	0,48
3,30	2,05	1,61	1,36	3,52	1,21	3,86	5,11	1,54	5,59	3,87	0

Resumo detalhado dos cálculos para a construção do gráfico silhueta (Tabela 9)

Considere os quatro grupos abaixo e os valores da distância euclideana:

Grupo 1- P8, P10; **Grupo 2** - P4, P3, P6. P12; **Grupo 3**- P2, P5, P9 **Grupo 4**- P1,P7, P11

Para ilustração do valor de s(i) tomou-se a população P1 em relação as demais:

Substituindo os valores na equação (3.2.3.4.3), tem-se:

População 1 com populações do próprio grupo (4): a(i) = $\frac{d_{1,7}+d_{1,11}}{2} = \frac{2,4+2,65}{2} = 2,52$

População 1 com grupo 1: b(i) = $\frac{d_{1,8}+d_{1,10}}{2} = \frac{2,15+4,2}{2} = 3,35$

População1com grupo 2: b(i) = $\frac{d_{1,3}+d_{1,4}+d_{1,6}+d_{1,12}}{2} = \frac{2,58+2,83+2,77+3,30}{4} = 2,87$

População1 com grupo3: b(i)= $\frac{d_{1,2}+d_{1,5}+d_{1,9}}{2} = \frac{2,3+2,14+2,82}{4} = 2,42$

Substituindo na fórmula s(i) = -0,03

De maneira análoga têm-se os cálculos para as demais populações.

APÊNDICE 3 – Análise de Variância Multivariada.

Matrizes do grupo1

$$B_1 = \begin{bmatrix} 392{,}1 & & & & \\ -11{,}1 & 1{,}26 & & & \\ -3{,}2 & -0{,}32 & 0{,}26 & & \\ 49{,}6 & -2{,}35 & -0{,}48 & 12{,}12 & \\ 69{,}8 & 2{,}31 & -1{,}3 & -7{,}15 & 59{,}9 \end{bmatrix} \quad W_1 = \begin{bmatrix} 3394 & 380{,}5 & 28{,}05 & 2651{,}7 & 278{,}8 \\ 380{,}5 & 45{,}8 & 2{,}6 & 313{,}8 & 45{,}6 \\ 28{,}05 & 2{,}6 & 0{,}8 & 18{,}18 & -6{,}22 \\ 2651{,}7 & 313{,}8 & 18{,}18 & 2176 & 306{,}37 \\ 278{,}8 & 45{,}6 & -6{,}22 & 306{,}4 & 171{,}6 \end{bmatrix}$$

Matrizes do grupo 2

$$B_2 = \begin{bmatrix} 579{,}3 & & & & \\ -38{,}3 & 2{,}62 & & & \\ -15{,}16 & 1{,}03 & 0{,}40 & & \\ 43{,}16 & -1{,}08 & -0{,}50 & 36{,}54 & \\ 201{,}8 & -14{,}4 & 36{,}54 & -5{,}99 & 83{,}59 \end{bmatrix} \quad W_2 = \begin{bmatrix} 368{,}9 & & & & \\ 44{,}5 & 12{,}12 & & & \\ -6{,}43 & -2{,}9 & 0{,}918 & & \\ 554{,}5 & 150{,}0 & -37{,}18 & 1914{,}8 & \\ 226{,}4 & 44{,}88 & -10{,}59 & 609{,}9 & 242{,}5 \end{bmatrix}$$

Matrizes do grupo 3

$$B_3 = \begin{bmatrix} 116{,}6 & & & & \\ -6{,}17 & 0{,}48 & & & \\ 8{,}82 & -0{,}49 & 0{,}67 & & \\ -73{,}44 & 3{,}59 & -5{,}50 & 46{,}77 & \\ -67{,}02 & 2{,}95 & -4{,}96 & 43{,}28 & 40{,}69 \end{bmatrix} \quad W_3 = \begin{bmatrix} 640{,}8 & & & & \\ 49{,}7 & 22{,}67 & & & \\ -45{,}37 & -10{,}45 & 6{,}13 & & \\ 913{,}7 & 157{,}9 & -110{,}5 & 2811 & \\ 528{,}4 & 107{,}1 & -65{,}8 & 1172 & 725 \end{bmatrix}$$

Matrizes do grupo 4

$$B_4 = \begin{bmatrix} 639{,}6 & & & & \\ -64{,}7 & 6{,}55 & & & \\ 22{,}3 & -2{,}25 & 0{,}77 & & \\ -503{,}9 & 51{,}0 & -17{,}57 & 397{,}07 & \\ -61{,}95 & 6{,}27 & -2{,}16 & 48{,}8 & 6{,}0 \end{bmatrix} \quad W_4 = \begin{bmatrix} 5400 & & & & \\ 585 & 64{,}8 & & & \\ 33{,}5 & 4{,}6 & 1{,}59 & & \\ 2924 & 317{,}8 & -0{,}26 & 2096 & \\ 690 & 74{,}44 & -2{,}22 & 538{,}2 & 141{,}8 \end{bmatrix}$$