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

APLICATIVO COMPUTACIONAL PARA OBTENÇÃO DE PROBABILIDADES "a priori" DE CLASSIFICAÇÃO ERRÔNEA EM EXPERIMENTOS AGRONÔMICOS.

CARLOS ROBERTO PEREIRA PADOVANI

Tese apresentada à Faculdade de Ciências Agronômicas da Unesp - Câmpus de Botucatu, para obtenção do título de Doutor em Agronomia (Energia na Agricultura).

BOTUCATU - SP
Junho – 2007

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

APLICATIVO COMPUTACIONAL PARA OBTENÇÃO DE PROBABILIDADES "a priori" DE CLASSIFICAÇÃO ERRÔNEA EM EXPERIMENTOS AGRONÔMICOS.

CARLOS ROBERTO PEREIRA PADOVANI

Orientador: Prof. Dr. Flávio Ferrari Aragon

Tese apresentada à Faculdade de Ciências Agronômicas da Unesp - Câmpus de Botucatu, para obtenção do título de Doutor em Agronomia (Energia na Agricultura).

BOTUCATU - SP
Junho - 2007

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

Padovani, Carlos Roberto Pereira, 1975-
P124a Aplicativo computacional para obtenção de probabilidades “a priori” de classificação errônea em experimentos agronômicos / Carlos Roberto Pereira Padovani. – Botucatu : [s.n.], 2007.
vii, 70 f. : il. color., tabs.

Tese (Doutorado) -Universidade Estadual Paulista, Faculdade de Ciências Agronômicas, Botucatu, 2007
Orientador: Flávio Ferrari Aragon
Inclui bibliografia

1. Estruturas de dados (Computação). 2. Algoritmos de computador. 3. Função discriminante linear. 4. Análise multivariada. 5. Distância de Mahalanobis. I. Aragon, Flávio Ferrari. II. Universidade Estadual Paulista “Júlio de Mêsquita Filho” (Campus de Botucatu). Faculdade de Ciências Agronômicas. III. Título.

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

CERTIFICADO DE APROVAÇÃO

TÍTULO: **APLICATIVO COMPUTACIONAL PARA OBTENÇÃO DE PROBABILIDADES "a priori" DE CLASSIFICAÇÃO ERRÔNEA EM EXPERIMENTOS AGRONÔMICOS.**

ALUNO: CARLOS ROBERTO PEREIRA PADOVANI

ORIENTADOR: PROF. DR. FLÁVIO FERRARI ARAGON

Aprovado pela Comissão Examinadora

PROF. DR. FLÁVIO FERRARI ARAGON

PROF. DR. ADRIANO WAGNER BALLARIN

PROF. DR. LUÍS FERNANDO NICOLOSI BRAVIN

PROF. DR. RUI VIEIRA DE MORAES

PROFa DRa SANDRA FIORELI DE ALMEIDA P. SIMEÃO

Data da Realização: 27 de julho de 2007.

"A MENTE QUE SE ABRE A UMA IDÉIA JAMAIS VOLTARÁ AO SEU TAMANHO ORIGINAL."

ALBERT EINSTEIN

DEDICO...

AO MEU PAI, CARLOS, PRINCIPAL PERSONAGEM DESTA CONQUISTA...

A MINHA MÃE, SILVIA, QUE SEMPRE ESTEVE AO MEU LADO...

A MINHA FUTURA ESPOSA, JULIANA, QUE FOI ACIMA DE TUDO, MINHA FIEL COMPANHEIRA E CUMPLICE DE MAIS UMA JORNADA...

Agradeço...

ao Prof. Dr. Flávio Ferrari Aragon pela valiosa orientação, incentivo e paciência em mais jornada científica;

ao Analista de Sistemas Luis Fernando Gaido, pela competência e interesse no desenvolvimento do aplicativo computacional ;

ao Diretor da FATEC, Prof. Dr. Roberto Antonio Colenci, por ter propiciado a possibilidade de cursar o programa de doutorado;

ao Diretor de Serviço José Roberto Sperandim por possibilitar o uso da tecnologia da FATEC para o desenvolvimento do programa de doutorado;

aos familiares e amigos, pelo incentivo e torcida;

a Profa. Cristiane, pelo auxilio na elaboração do abstract;
aos funcionários da Seção de Pós-graduação pela simpatia e serviços prestados;

aos professores do Depto de Bioestatística pelo apoio;

a "TODOS" que de forma direta ou indireta contribuíram para a conclusão deste trabalho.

SUMÁRIO

Página

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

Nas Ciências Agronômicas, encontram-se várias situações em que são observadas diversas variáveis respostas nas parcelas ou unidades experimentais. Nestas situações, um caso de interesse prático à experimentação agronômica é o que considera a construção de regiões de similaridade entre as parcelas para a discriminação entre os grupos experimentais e ou para a classificação de novas unidades experimentais em uma dessas regiões. Os métodos de classificação ou discriminação exigem, para sua utilização prática, uma quantidade considerável de retenção de informação da estrutura de variabilidade dos dados e, principalmente, alta fidedignidade e competência nas alocações de novos indivíduos nos grupos, mostradas nas distribuições corretas destes indivíduos. Existem vários procedimentos para medir o grau de decisão correta (acurácia) das informações fornecidas pelos métodos classificatórios. Praticamente, a totalidade deles utilizam a probabilidade de classificação errônea como o indicador de qualidade, sendo alguns destes freqüentistas (probabilidade estimada pela freqüência relativa de ocorrências – métodos não paramétricos) e outros baseados nas funções densidade de probabilidade das populações (métodos paramétricos). A principal diferença entre esses procedimentos é a conceituação dada ao cálculo da probabilidade de classificação errônea. Pretende-se, no presente estudo, apresentar alguns procedimentos para estimar estas probabilidades, desenvolver um software para a obtenção das estimativas considerando a distância generalizada de Mahalanobis como o procedimento relativo à da função densidade de probabilidade para populações com distribuição multinormal . Este *software* será de acesso livre e de fácil manuseio para pesquisadores de áreas aplicadas, completado com o manual do usuário e com um exemplo de aplicação envolvendo divergência genética de girassol.

Palavras-chave: probabilidade de classificação errônea, função discriminante linear, análise multivariada, algoritmo computacional, distância de Mahalanobis.

COMPUTATIONAL APPLICATION FOR OBTAINING PROBABILITIES "a priori" OF ERRONEOUS CLASSIFICATION IN AGRONOMIC EXPERIMENTS. Botucatu, 2007, 69 p.Tese (Doutorado em Agonomia/Energia na Agricultura) – Faculdade de Ciências Agrnômicas, Universidade Estadual Paulista.

Author: Carlos Roberto Pereira Padovani

Adviser: Flávio Ferrari Aragon

In the Agronomical Sciences, mainly in studies involving biomass production and rational use of energy, there are several situations in which several variable answers in the parts or experimental units are observed. In these situations, a case of practical interest to the agronomical experimentation is that one which considers the construction of similarity regions among parts and or the classification of new experimental units. The classification methods demand, for their utilization, a considerable quantity for utilization of their information retention of data and, mostly, high fidelity and competence in the new individual allocations. There are several procedures to measure accuracy degree of the information supplied by the discrimination method. Practically all of them use the miss-classification probability (erroneous classification) like the quality indicator. The main difference among these evaluation methods is the characterization of the miss-classification probability.

Therefore, the aim is to present some estimate procedures of the miss-classification probabilities involving repetition frequency and distribution methods and to develop a software to obtain their estimate, which is accessible and easy handling for researchers of applied areas, complementing the study with user's manual and examples in the rational energy application and biomass energy.

Key-words: miss-classification probability, discriminating function, multivariate analysis.

1. INTRODUÇÃO

Nas Ciências Agronômicas, encontram-se várias situações em que são observadas diversas variáveis respostas nas parcelas ou unidades experimentais. Nestas situações, a utilização de técnicas multivariadas constitui-se, atualmente, em uma das melhores alternativas de análise exploratória e procedimento inferencial dos dados, pois permite combinar as múltiplas informações contidas na unidade experimental envolvendo as variações intra e intervariáveis numa estrutura matricial de dispersão, apresentada pelas variâncias e covariâncias dos dados, que será utilizados em todos os procedimentos teóricos do desenvolvimentos analítico.

Um caso de interesse prático bastante comum à experimentação agronômica é o que considera a construção de regiões de similaridade entre as parcelas e/ou a classificação de novas unidades experimentais. Em algumas situações experimentais, o pesquisador necessita de um critério objetivo para decidir entre dois ou mais grupos de indivíduos, previamente definidos sob características comuns para todos os elementos do grupo, em qual dos grupos seria alocado um novo indivíduo considerando as características comuns avaliadas e, principalmente, a minimização de uma alocação casual. É evidente que todo processo de tomada de decisão, traz consigo um possível erro de decisão. Do ponto de vista estatístico, o objetivo do estabelecimento de um critério qualquer de decisão quanto à alocação de um novo indivíduo deve ser construído de tal forma que o número de classificações errôneas seja

minimizado, ou seja, o erro de dizer que o elemento amostral pertence a uma população quando na verdade, ele pertence a outra deve ser estritamente casual.

Quanto aos critérios de alocação estabelecidos pelos métodos de classificação, estes exigem, para sua utilização, uma quantidade considerável de retenção de informação dos dados e, principalmente, alta fidedignidade e competência nas alocações de novos indivíduos.

Quanto aos procedimentos para medir o grau de acurácia das informações fornecidas pelo método de discriminação, principalmente referente a alocações de novos indivíduos, a totalidade deles utiliza a probabilidade de classificação errônea como o indicador de qualidade. A principal diferença entre esses procedimentos é a conceituação da probabilidade que pode ser realizada de maneira freqüentista (número de alocações erradas) ou por distribuição probabilística (geralmente baseada na normalidade dos dados – distribuição gaussiana).

Neste sentido, pretende-se apresentar alguns procedimentos de estimação das probabilidade de classificação errônea envolvendo métodos freqüentistas e probabilísticos para a discriminação linear de Fisher e desenvolver um *software* para obtenção de suas estimativas considerando a situação em que os dados são provenientes da distribuição normal de probabilidade (método da distância de Mahalanobis), que seja de acesso livre e de fácil manuseio para pesquisadores de áreas aplicadas, complementando o estudo com o manual do usuário e um exemplo de discriminação genética entre variedades de girassol.

2. REVISÃO DA LITERATURA

Sir Ronald Aylmer Fisher (1890 – 1962) nasceu em Londres no dia 17 de fevereiro de 1890 e bacharelou-se em Astronomia pela Universidade de Cambridge em 1912. Em 1919, após trabalhar dois anos como estatístico e mais quatro como professor de matemática e física quântica em escolas públicas, recebeu o convite para exercer suas atividades profissionais na Estação Experimental de Agricultura de Rothamstead, Inglaterra, onde permaneceu ate 1933.

Durante esse período, o envolvimento cotidiano com problemas da área agrícola despertou em Fisher a necessidade de desenvolver vários métodos de análise de dados e, principalmente, os modelos probabilísticos para os delineamentos experimentais. Destacam-se entre as técnicas Fisherianas a análise de variância, testes de hipóteses, método da máxima verossimilhança, análise de covariância e os métodos multivariados para a interpretação de resultados da coleta de dados.

O estudo de discriminação entre duas populações, apresentado por Fisher em 1935, e publicado pela primeira vez em 1936, foi originalmente desenvolvido na Botânica, com o propósito de distinguir grupos de plantas com base no tamanho e tipo de folhas, para que, posteriormente, fosse possível classificar as novas espécies encontradas. Pioneiro no estudo das

técnicas multivariadas de discriminação e classificação, Fisher é considerado “O Arquiteto da Análise Multivariada” (RAO, 1964).

O método consistia em construir função linear das variáveis aleatórias $X_i,\ldots,X_p$ que descreviam com propriedades os indivíduos ou objetos em grupos mutuamente exclusivos, cujos coeficientes deviam ser calculados para que a função maximizasse a “distância entre as populações” definida pelo quociente da diferença entre as médias dos grupos relativamente aos desvios padrão no interior de cada grupo.

Não obstante a função discriminante tenha sido sugerida, a principio, para o trabalho específico com duas populações, seu uso foi estendido para a discriminação entre mais de duas. Sendo o valor numérico da função discriminante um escalar, este não esgota toda a informação a respeito da configuração (disposição geométrica) de mais de duas populações, a menos que estas sejam colineares (os centros das populações estão sobre uma reta). Frente à certeza de que o conhecimento da configuração das populações trata de fato importante no problema de discriminação, Fisher sugeriu, em 1938, um teste para verificar a colinearidade de *k* populações.

O desenvolvimento teórico conceitual dado à discriminação sob o ponto de vista da função discriminante de Fisher baseou-se na maximização da distância entre as populações sem levar em consideração o aspecto de minimizar as probabilidades de erro no processo discriminatório. No segmento do aperfeiçoamento da técnica, Welch (1939) abordou o problema de discriminação entre duas populações considerando as probabilidades de erro de classificação a partir das idéias introduzidas por Neyman e Pearson (1933) sobre noções de erro na aplicação de técnicas estatísticas.

Arseven e Kshirsagar (1975) mostram em seu artigo que para a discriminação entre *k* populações normais multivariadas com a mesma matriz de variância-covariância e diferentes vetores de médias para classificação de uma nova observação em uma delas, existem dois procedimentos na literatura, ambos de Fisher, que sob certas condições são equivalentes. Um é a generalização da construção de funções discriminantes para duas populações, enquanto o outro envolve a maximização da razão das somas de quadrados e de produtos da variação entre e dentro de grupos. Esta envolve a eliminação de variáveis canônicas que não interferem na capacidade de discriminação.

Crocci (1979) apresenta uma extensa revisão sobre alguns procedimentos para a obtenção da probabilidade de má classificação baseando-se na função discriminante linear de Fisher para duas populações. Complementa o estudo com a construção da função linear e a estimação das taxas de erros dos diversos procedimentos considerando 12 caracteres quantitativos mensurados em duas espécies de abelhas *(Partamona testacea e Partamona pseudomusarum).* A comparação numérica das probabilidades de má classificação mostrou que o procedimento de Wald, usando a distância generalizada de Mahalanobis, foi o que apresentou a menor taxa de classificação errônea.

Woodward e Elliott (1983) apresentam dados sobre um levantamento de 26 pacotes estatísticos específicos para microcomputadores. O levantamento destinava-se a apresentar a estatísticos e demais membros da comunidade científica as disponibilidades de *softwares* disponíveis à utilização em pesquisas, bem como envolvimento da comunidade na depuração de sérios erros que muitos deles continham, pelo fato de terem sido desenvolvidos por empresas que não contavam com a assessoria de estatísticos.

Basnet (1993) utiliza a técnica dos componentes principais para estudar relações entre fatores ligados ao meio ambiente e padrões de distribuição de árvores em florestas úmidas subtropicais. O estudo desenvolvido em Porto Rico, na Floresta Experimental de Luquilo, considera dados envolvendo variáveis geológicas, edáficas e ambientais. Os eixos discriminantes indicam que a associação entre os fatores edáficos e a geologia local complementados com alterações exógenas estabelecem o padrão de desenvolvimento florestal.

Piassi *et al.* (1995) consideram oito grupos genéticos de aves de postura observados em relação a características de importância econômica para idade de produção de ovos (peso corporal, consumo alimentar, idade ao primeiro ovo, taxa de postura, peso médio do ovo, massa do ovo, massa do ovo/unidade do tamanho metabólico e viabilidade) avaliadas em dois momentos sucessivos de oito semanas. Foram utilizados vários procedimentos multivariados de análise estatística: MANOVA (análise de variância multivariada), análise de agrupamento tendo como coeficiente de similaridade a distância generalizada de Mahalanobis o método de otimização de Tocher considerando a matriz de distância entre pares de genótipos e, finalmente, a análise canônica. Em relação à análise canônica, os dois primeiros eixos discriminantes respondem por mais de 92% da variação total observada em ambos os períodos estudados e a representação bidimensional nos eixos canônicos indicaram dois grupos distintos, com alto grau

de divergência genética. O grau de divergência genética entre os seis grupos genéticos remanescentes foi baixo.

Lucio *et al.* (1999) investigam o regime climático da precipitação, temperatura e umidade relativa do ar à superfície na região metropolitana de Belo Horizonte - MG. Para o estudo, utilizaram-se dados coletados durante 30 anos consecutivos. No plano analítico, para os valores observados, foram considerados os seguintes métodos estatísticos: ajuste de modelo quadrático de regressão pelo procedimento dos mínimos quadrados, técnica da análise multivariada (MANOVA), análise discriminante linear de Fisher. Os resultados mostraram que, na caracterização do clima de Belo Horizonte, nenhuma das variáveis consideradas pode ser julgada como de baixa contribuição, ou seja, todas contribuem significativamente para a discussão das variações climáticas da região.

Morgano *et al.* (1999) utilizam duas metodologias diferentes de preparação de amostra para determinação da concentração de minerais em sucos de frutas por espectrometria de emissão óptica em plasma indutivamente acoplado, para aplicação mais adequada nas análises das frutas abacaxi, acerola, caju, goiaba, manga, maracujá e uva; complementada com métodos quimiométricos, análise de componentes principais e análise hierárquica por agrupamento, na busca de uma melhor interpretação dos resultados. Os teores minerais Ca, P, Na, K, Mq, Zn, Fe, Mn e Cu não diferiram significativamente para as duas metodologias empregadas e a discriminação canônica estabeleceu que composição de minerais nas diferentes variedades de sucos de frutas diferiram entre si, mostrando que o primeiro eixo canônico identifica, principalmente, os sucos de manga e o caju e segundo, os sucos de abacaxi e uva. Além disso, as regiões de discriminação mostram os sucos de goiaba e acerola como os mais semelhantes. O suco de abacaxi apresentou os maiores teores de cálcio e magnésio; o suco de uva os maiores níveis de cobre e o suco de maracujá, os maiores níveis de fósforo, potássio e zinco.

Peroni *et al.* (1999) utilizam metodologia multivariada no estudo da diversidade inter e intra-específica presente num sistema itinerante de uma propriedade rural de agricultores tradicionais autóctones da região de Cananéia, litoral do estado de São Paulo e também, as etnovariedades de mandiocas presentes no sistema. A aplicação da técnica de discriminação canônica, juntamente com a análise de agrupamento em 21 caracteres morfológicos das folhas, caule e raiz da mandioca obtidos *in situ* mostrou-se eficiente para diferenciar, com baixo risco de erro de classificação, as etnovariedades encontradas e apontar que

a diversidade tanto inter como intra-específica é elevada, revelando que nas roças são cultivadas espécies conjugadas, isto é, existem espécies de propagação vegetativa em que as partes utilizadas para o consumo são as raízes e tubérculos, assim como espécies de propagação por sementes das quais se utilizam os grãos e frutos para consumo.

Ribeiro *et al.* (1999), tendo por finalidade explorar a divergência genética entre populações de coqueiro-gigante-do-brasil (*Cocos nucifera* L.), utilizaram técnicas de análise multivariada envolvendo a distância generalizada de Mahalanobis e a discriminação linear canônica de Fisher, com vistas ao auxilio aos melhoristas na definição da manutenção desses germoplasmas em bancos ativos, na conservação e avaliação, bem como permitir identificar aquelas populações com maior potencial a ser explorado na obtenção de híbridos. Foram coletados três frutos com idade aproximadamente de 12 meses em 96 plantas de cada uma das cinco populações estudadas (localizações geográficas): Pacatuba-SE; Praia do Forte – BA; Merepe-PE; Santa Rita – PE e São José do Mipibu – RN. Os frutos foram mantidos em galpões com ventilação livre durante 21 dias para complementação da maturação e secagem da fibra, e em seguida procedeu-se a análise dos componentes do fruto utilizando-se características de pesos, porcentagens, diâmetros e índice polar equatorial. A partir da dispersão gráfica dos escores dos eixos canônicos discriminantes de Fisher e as distâncias generalizadas de Mahalanobis entre os centróides das populações, classificou-se como sendo a menor divergência genética a observada entre as populações Pacatuba e Merepe. A população de São José do Mipibu e de Praia Grande do Forte apresentam distâncias intermediárias. Dos 19 caracteres avaliados, apenas quatro são selecionados (peso de noz, peso de albúmen, diâmetro equatorial e porcentagem de albúmen no fruto sem água) como tendo a maior contribuição para a discriminação entre populações no contexto da informação total contida no modelo de classificação.

Fonseca *et al*. (2000) avaliam o desempenho das três principais raças suínas utilizadas nos programas de melhoramento com relação às características reprodutivas e às respectivas divergências genéticas. As informações coletadas foram submetidas a vários procedimentos quantitativos de análise de dados: variáveis canônicas, análise de variância multivariada (MANOVA), teste de Roy e distância generalizada de Mahalanobis. A discussão foi complementada com a construção da função linear discriminante de Fisher para as três raças estudadas . Os resultados obtidos demonstraram que as raças Landrace e Large White apresentam semelhanças genéticas quando comparadas à raça Duroc.

Cardim *et al.* (2001) objetivando estudar a variabilidade genética ultra-específica em *Oncidium varicosun lindl*. (Orchidaceae-Oncidiine) em Minas Gerais, consideram 22 caracteres florais, medidos a partir de "vouchers" (fichas florais) de cinco populações amostradas em localidades de Alfenas, Estiva, Pedra do Coração, Pedra Branca e Pouso Alegre. Todos os caracteres florais para o estudo morfométrico e taxonômico foram medidos em milímetros e, posteriormente, submetidos à procedimentos analíticos de testes estatísticos univariados e multivariados. No contexto multivariado, foram utilizadas as técnicas da análise discriminante canônica para verificar como as populações relacionam-se entre si considerando a estrutura completa das correlações residuais existentes entre as características observadas e a do principio de conglomeração baseada na distância genética de Mahalanobis para constituição de fenogramas entre as populações. Com alto grau de precisão, foi possível estabelecer a separação entre as populações e também verificar que o padrão de variabilidade genética observado não parece estar relacionado com a diferença de latitude. Entretanto, parece estar associado a diferenças de longitudes, embora este único fator possa não ser suficiente para explicá-lo.

Bellaver *et al.* (2002) apresentam uma alternativa multivariada para avaliar o efeito da lisina (LM) e da energia metabolizáveis (EM) sobre o desempenho e deposição de tecidos na carcaça em frangos de corte, mantendo-se a relação de aminoácidos indicada na formulação de dietas experimentais, a qual propicia dietas equilibradas com base no conceito de proteína total. Os tratamentos (dietas) resultaram de um esquema fatorial 3 x 4 (3 níveis de energia metabolizável e 4 níveis de lisina metabolizável) mais um tratamento com as exigências supridas por aminoácidos totais, distribuídos em oito blocos de peso inicial. Foram considerados 2808 pintos machos de linhagem comercial de um dia de idade, distribuídos em 104 boxes, com 27 unidades por parcela. A primeira função discriminante canônica de Fisher, constituída a partir do ganho de peso, consumo de ração e proteína bruta na carcaça juntamente com desempenho e a deposição de tecidos, permitiu estimar com alta probabilidade correta, a exigência mínima de lisina em 1,18% da dieta, quando esta tiver 3000 kcal de EM/kg de ração e, em 1,22% de LM, ao se aumentar para 3100 ou 3200 kcal/kg de EM/kg de ração.

Caixeta *et al.* (2002) avaliaram a qualidade da madeira após a secagem natural e a identificação dos fenótipos mais promissores para auxiliar o estabelecimento de um programa de melhoramento florestal utilizando-se de procedimentos multivariados. Para a verificação da possibilidade de classificação errônea dos fenótipos nas três classes de qualidade,

foram considerados 44 fenótipos superiores obtidos em povoamentos de eucaliptos adaptados para as condições ambientais da região noroeste do estado de Minas Gerais a partir de valores médios das porcentagens de índice de rachadura, encurvamento, encanoamento; quino e nó. A avaliação conjunta das características observadas nos fenótipos, segundo o método de análise discriminante "stepwise", tendo com critério de seleção das variáveis a maximização da distância generalizada de Mahalanobis entre duas classes mais próximas combinada com o modelo discriminante de Fisher, objetivando maximizar a diferenciação entre grupos, indicou uma melhor definição na alocação dos fenótipos dentro das três classes de qualidade. Os fenótipos da classe I foram considerados os mais indicados para um programa de melhoramento, uma vez que apresentaram o mais baixo percentual de defeitos (6,57%) em relação às outras duas classes (16,57%, para classe II e 28,34% para classe III).

Fonseca *et al.* (2002) utilizam procedimentos multivariados para comparar o desempenho de características de carcaça de dois híbridos de frangos de corte desenvolvidos pela UFV – Viçosa – MG com dois híbridos comerciais. Para a comparação dos quatro grupos foram consideradas características relativas ao peso de abate, peso da carcaça, peso do peito, peso da contracoxa, peso da coxa e rendimento da carcaça. A técnica da MANOVA complementada pelo teste de Roy e Bose mostrou diferença significativa ($P<0,05$) entre os grupos, sendo que para os pesos da carcaça e do peito os produtos comerciais foram superiores. Com a utilização da função discriminante linear de Fisher, identificam-se dois grandes grupos, um formado pelos híbridos da UFV e outro pelos híbridos comerciais, com expressiva separação nas regiões de classificação (mínima probabilidade de má-classificação) que enfatizam as variáveis peso da carcaça e peso do peito como as mais contributivas para a separação.

Martinello *et al.* (2002) apresentam um modelo para estimar a divergência genética e classificar os caracteres morfoagronômicos de maior importância na evidenciação de 39 acessos do gênero *Abelmoschus,* incluindo genótipos de espécies univariadas de diferentes origens. Considerando a técnica dos componentes principais, foi possível sumarizar os 13 descritores quantitativos em três eixos canônicos, obtidos a partir da matriz de correlações genotípicas que explicaram, conjuntamente, 76,7% da variação total dos dados. Com uma taxa elevada de classificação correta e, conseqüentemente, uma baixa probabilidade de erro de má-classificação, verificou-se que as técnicas estatísticas empregadas para as características

morfométricas estudadas foram capazes de estabelecer a diversidade genética e permitir a discriminação genotípica.

Pires *et al.* (2002) consideram três raças suínas Landrace, Large White e Duroc – quanto à avaliação de desempenho por meio da técnica da análise de variância multivariada e da função discriminante linear de Fisher, usando os testes do maior autovalor de Roy e da união-interseção de Roy para as comparações múltiplas, complementam a pesquisa como estudo da divergência genética por meio da análise canônica. Foram incluídas no estudo seis características de desempenho: peso do leitão ao nascimento, peso do leitão aos 21 dias, peso do leitão aos 70 dias, ganho de peso médio diário, idade para atingir 100kg e espessura de toucinho. Concluíram que, na discriminação das raças quanto ao desempenho, a raça Large White apresentou uma pequena superioridade em relação à Landrace, e ambas bem superiores à Duroc. Os resultados justificaram a utilização das raças Landrace e Large White para a obtenção de fêmeas FI, para um posterior acasalamento com machos Duroc, visando a obtenção de animais híbridos com efeito heterótico expressivo e para haver complementaridade entre as características.

Sant`Anna e Malinovski (2002) procuram por meio de técnicas multivariadas, classificar os principais fatores humanos relacionados com a atividade de corte com motosserra, considerando a produtividade individual (m^3/dia), idade (anos), experiência na atividade (meses), capacidade aeróbica (ml/O_2/kg/min), peso corporal (kg), estatura (m), circunferência do braço (cm), circunferência da perna (cm), diâmetro do úmero (cm), diâmetro no fêmur (cm), dobra cutânea do tríceps (mm), dobra cutânea suprailíaca (mm), dobra cutânea subescapular (mm), dobra cutânea da perna (mm), dobra cutânea abdominal (mm), gordura corporal (%), índice de massa corporal (kg/m^2) e três componentes do somatotipo do ser humano: endomorfia, mesomorfia e ectomorfia de 29 operadores de uma empresa que atuavam em áreas montanhosas. Dos 20 fatores avaliados, a técnica dos componentes principais associada a análise de agrupamento permitiu indicar 13 deles como sendo os mais importantes na classificação dos operadores, assim como produziu um modelo consistente de classificação que possibilita indicar, com alto grau de confiança, os grupos típicos de operadores de motossera da empresa.

Silva *et al.* (2002) discutem a discriminação geográfica de águas minerais do estado de São Paulo utilizando a técnica da análise dos componentes principais obtidos a partir da matriz de dados padronizados de pH e teores de Ba, Ca, K, Mg, Na. Com o auxílio dos dois

primeiros eixos canônicos, que acumulam 78,6% da variância total dos dados observados, estabeleceram-se subsídios para a construção de modelos de previsão ou classificação que permitem prever a origem de amostras de água mineral do estado, evidenciando a potencialidade de rastreabilidade.

Alves *et al.* (2003) utilizam a técnica dos componentes principais para selecionar caracteres botânico-agronômicos avaliados em banco ativo de germoplasma de cupuaçuzeiro de material coletado em pomares caseiros, pequenas propriedades e áreas silvestres de dez localidades do estado do Amazonas, seis no Pará e duas no Amapá. Os descritores quantitativos foram apresentados por grupos que envolviam caracteres de folha (14 variáveis), flor (18 variáveis), fruto (16 variáveis) e agronômicos (5 variáveis), fornecendo um vetor multidimensional de resposta de ordem 53. A técnica de seleção dos descritores que deveriam permanecer para construção da função de classificação de cada espécie envolveu duas etapas. Na primeira, o descarte aconteceu dentro de cada grupo e, em seguida, foi realizada análise conjunta para a seleção final. Após a seleção das variáveis dentro de grupo, foi comprovada a eficiência do descarte, resultando para a análise conjunta sete variáveis da folha (50%), dez da flor (56%), nove do fruto (56%), e três agronômicas (60%). Na segunda análise, das 29 variáveis pré-selecionadas, foi possível descartar mais 10 variáveis, que apresentavam redundância intergrupo e, portanto, muito pouco contributivas para a classificação.

Andrade *et al*. (2003) consideram 39 variedades de canas-de-açúcar com o objetivo de comparar características na alimentação de ruminantes quanto a aspectos de degradabilidade, composição química e nutricional. O experimento foi desenvolvido em terreno preparado com a aplicação de calcário dolomítico, na base de 4500kg/ha e aplicando-se no sulco 60 e 100kg de K_2O e P_2O_5/ha, respectivamente. Para a análise de agrupamento e discriminação canônica foram considerados nove caracteres, a saber, produção de matéria seca, porcentagem de carboidratos totais não estruturais, degradabilidade potencial da matéria seca, fibra insolúvel em detergente neutro, fibra insolúvel em detergente ácido, celulose, hemicelulose e lignina, das 39 variedades de canas colhidas aos 12 meses. Levando-se em conta a qualidade, composição e degradabilidade, os procedimentos multivariados discriminaram 9 grupos de variedades; destes, 5 destacando-se como os mais importantes para a alimentação animal. Ademais, foi evidenciado que, em canas com elevada porcentagem de fibra insolúvel em detergente neutro e baixa porcentagem de carboidratos totais disponíveis, ocorreu maior degradabilidade da fibra.

Assis *et al.* (2003) avaliam seis diferentes espécies de braquiária, em que foram consideradas características vegetativas, reprodutivas e de pilosidade. Para cada um dos grupos de caracteres morfológicos, foram estabelecidas as funções discriminantes para as seis espécies (*B. brizantha, B. decumbens, B. humidicola, B. jubata, B. ruziziensis e B. dictyoneura*). Considerou-se, ainda, a análise de consistência, em que os dados foram submetidos a uma nova classificação, para o estabelecimento das probabilidades de má-classificação. Verificou-se que os grupos de caracteres morfológicos estudados, os vegetativos e os reprodutivos, foram os mais importantes para discriminação de espécies. Porém, cada caractere teve sua responsabilidade apurada na discriminação, ou seja, para as espécies *B. decumbens e B. humidicola*, a única responsável foi a pilosidade; nas espécies *B. jubata e B. dictyoneura*, o vegetativo e, para as espécies *B. brizantha e B. ruziziensis*, o reprodutivo.

Gatti *et al.*(2003) utilizam as técnicas de componentes principais e análise de grupamentos para identificar as variáveis que explicam a variação entre indivíduos dentro da população de aspargos e determinar um critério de seleção de plantas superiores para compor novas populações. Considerando um ensaio instalado no campo experimental da Faculdade de Ciências Agrárias da Universidade Nacional de Rosário, Santa Fé, Argentina, composto por 1200 plantas, foram observadas as seguintes variáveis respostas: número de turiões por plantas, peso médio do turião, diâmetro do turião; produção total por planta; produção comercial por planta; número de dias para início da colheita; número de hastes por planta; altura da haste principal e peso seco da massa verde. Os resultados dos procedimentos estatísticos permitiram identificar cinco grupos distintos para os dois sexos e selecionar as melhores características para atuar como progenitores de um novo conjunto gênico para obter uma população melhorada de aspargo branco tanto quanto os caracteres vegetativos quanto produtivos.

Martel *et al.* (2003) consideram três técnicas estatísticas multivariadas (análise de componentes principais, análise discriminante e análise de agrupamentos) com o objetivo de caracterizar morfometricamente raças e populações de pupunha. Foram examinadas pupunheiras (*Bactris gasipaes* Kunth) ao longo dos rios Amazonas e Solimões, que apresentam grande variabilidade genética; porém, ainda não totalmente caracterizadas. Os seguintes descritores morfológicos foram avaliados: número de espigas por cacho, comprimento da ráquis, distância morfológica dos frutos, peso dos frutos, adensamento dos frutos, espessura das cascas, facilidades para descascar os frutos, peso das cascas, texturas da polpa, sabor dos frutos,

espessura da polpa, peso das sementes e teores de água, óleo e fibras. A análise de agrupamento possibilitou a formação de três grupos de afinidades morfométricas indicados pelas raças Solimões, Putumayo e Pará. Os dois principais eixos discriminantes permitiram a representação bidimensional das 16 populações de pupunha das três raças estudadas, possibilitando, assim, a visualização gráfica da caracterização morfométrica. As três técnicas multivariadas em conjunto definem uma diferenciação das raças, mostrando que, para a seleção morfométrica das pupunhas, destacam-se os seguintes descritores: número de espigas, comprimento da raquis, peso do fruto, espessura e peso das cascas, facilidade para descascar os frutos, sabor dos frutos, espessura da polpa, distância morfológica dos frutos e peso da semente.

Miranda *et al.* (2003) objetivam avaliar o potencial de melhoramento e a divergência genética de cultivares tropicais de milho-pipoca por meio de técnicas multivariadas. Neste sentido, foram utilizados oito ensaios conduzidos durante dois anos, com quatro épocas de semeadura e escolhido para estudo o que apresentava a menor interação dos cultivares com os ambientes, escolha resultante dos procedimentos analíticos multivariados empregados na pesquisa. Na avaliação da divergência genética entre as cultivares utilizou-se o procedimento de discriminação relativo à análise dos eixos canônicos com os grupos formados com base na distância generalizada de Mahalanobis, a partir de estatísticas calculadas das seguintes características principais observadas no vetor de resposta da parcela experimental: altura da planta; altura da espiga; empalhamento; número de espigas doentes; capacidade de expansão e produtividade de grãos; prolicifidade e variáveis derivadas das principais. As cultivares foram classificadas em três grupos probabilisticamente dissimilares e o potencial de melhoramento das cultivares foi determinado pelas médias das características com maior potencial de discriminação juntamente com as posições das cultivares em diferentes grupos obtidos da divergência genética.

Rodrigues & Ando (2003) avaliam 65 variedades de arroz-de-sequeiro (*Oryza sativa* L.) quanto à sensitividade à radiação gama, para discriminar os grupos Índica e Japônica. Inicialmente, todos os caracteres foram submetidos a procedimentos de análise univariada para detectar a existência de variabilidade entre os grupos. Após a análise univariada, realizaram–se duas análises multivariadas; a primeira para a visualização da discriminação dos grupos em espaço bidimensional, colocada em prática pela construção de gráficos com as duas primeiras variáveis canônicas de cada dosagem, e a segunda pela função discriminante para classificar as variedades em grupos, com probabilidade de classificação errada menor que 5%. Os

resultados mostraram que a radiossensitividade foi eficiente para discriminar as variedades de arroz dos grupos Índica e Japônica, sendo que a Japônica é mais sensível à radiação gama do que as do grupo Índica.

Sanches *et al*. (2003) utilizam a técnica da discriminação da análise de componentes principais para a avaliação da qualidade de banana nanicão quanto à submissão a duas condições de temperatura de armazenamento (ambiente sem controle-testemunha e, 13 ± 1^{o}C, com controle de umidade ajustada para 90± 2,5%) e três tipos de embalagens (madeira "torito", com capacidade para 18 kg de frutas; madeira chamada de ½ caixa, com capacidade de 13kg e papelão de capacidade para 18 kg de banana). Para discriminação das combinações dos dois fatores foram avaliados a coloração da casca, porcentagem de imperfeições, massa fresca, acidez total titulável, pH, sólidos solúveis totais e porcentagem de sacarose. O procedimento multivariado, com probabilidade máxima de 5% de erro de afirmação, permitiu concluir que em qualquer temperatura, a melhor embalagem para o condicionamento das frutas é a madeira de ½ caixa, mostrando que a redução do número de frutas por embalagem se faz necessária para obter maior área de ventilação e diminuição de danos mecânicos, preservando a qualidade por mais tempo.

Souza *et al.* (2003) empregam a análise multivariada, em particular as técnicas de agrupamento discriminante, para estratificação vertical de florestas ineqüianas considerando dados coletados em 10 parcelas permanentes de 20m x 50m cada, de um experimento instalado na mata da Silvicultura, no município de Viçosa (20º45'S e 42º55'S), estado de Minas Gerais. Consideram-se, na análise da estrutura vertical, os dados das alturas totais das árvores amostrais com diâmetro de tronco (dap) igual ou maior que 5,0cm abordados pela distância euclidiana na técnica de discriminação e, pelo método de ligação complementar, na técnica de agrupamento. Obteve-se, como resultado, que as técnicas multivariadas são viáveis para estratificação vertical de floresta ineqüiana quando se utiliza a distribuição de alturas com classes com amplitudes comuns de 1m e, também, que a diversidade de espécies e das estruturas fitossociológicas e paramétrics por meio da estratificação vertical são úteis nas análises estruturadas de florestas ineqüiana.

Fonseca e Fonseca (2004) descrevem uma alternativa multivariada para a caracterização das fases de desenvolvimento do mosaico sucessional de um trecho estacional semidecidual, por meio de variáveis estruturais. Foram instaladas 200 parcelas de 100m^{2}(10 x 10)

no centro do fragmento, formando uma malha de 100 x 200m, na qual se procedeu à análise estrutural (levantamento fitossociológico acrescido do estudo das variáveis porcentagem de cobertura, altura do dossel e cobertura por lianas), em 100 parcelas sorteadas. A técnica dos componentes principais juntamente com a análise de classificação hierárquica ascendente permitiu caracterizar as fases do desenvolvimento do mosaico sucessional; porém, algumas ressalvas devem ser consideradas frente ao caráter gradativo do desenvolvimento do mosaico, fato que dificulta a classificação em poucas fases. Assim, embora o método seja consistente, as taxas de erro devem ser controladas pois as situações intermediárias características de transição sempre estarão presentes. A incrementação de outras variáveis na discriminação é um fator que pode ser considerado para a majoração da taxa de classificação correta.

Leal *et al*.(2004) mostram a importância da determinação da divergência genética na produção da cultura de feijão-de-vagem no contexto da agricultura do Rio de Janeiro – RJ. O uso da análise multivariada, em que diversos caracteres podem ser dimensionados simultaneamente, apresenta-se bastante vantajoso, como identificador de fontes de variabilidade genética. Com o auxílio das variáveis canônicas, foi possível estimar os dois primeiros escores, que foram utilizados para a disposição dos genótipos em gráficos cartesianos de dispersão, fato que possibilitou um exame visual da divergência genética da cultura. Os resultados apontam excelentes perspectivas para trabalhos futuros na busca de explorar a variabilidade encontrada entre os acessos de feijão-de-vagem estudados .

Nanni *et al.* (2004) objetivam desenvolver e avaliar um método para discriminação das classes de solos a partir de suas respostas espectrais, utilizando-se de um sensor em laboratório. A matriz de dados utilizada na análise estatística foi composta por 22 bandas e 13 diferenças de altura do fator de refletância para as duas camadas de solo amostradas em cada ponto de amostragem. Para a diferenciação e caracterização dos solos foi utilizada a técnica da análise discriminante de Fisher com o objetivo de desenvolver e validar o método para determinação da classe do solo avaliado em função de suas respostas espectrais. As conclusões obtidas demonstram que a análise estatística aplicada à resposta espectral das amostras obtidas no laboratório permite a discriminação dos solos e o teste estatístico é eficiente, com taxa de acerto médio acima de 91%, e erro global de 8,8%.

Pereira *et al*. (2004) estudam a variabilidade existente entre 36 acessos de taro do Banco de Germoplasma de Hortaliças da Universidade Federal de Viçosa-MG. Os

dados foram submetidos à análise de similaridade por variáveis canônicas e de agrupamento pelo método de Tocher, adotando a distância generalizada de Mahalanobis como estimativa de semelhança genética. Os resultados mostram que os cruzamentos visando aumento em produtividade, entre os acessos de um mesmo grupo, devem ser evitados devido a elevada similaridade expressada entre os mesmos.

Santos *et al.* (2004) investigam distinguir grupos sócioecológicos por meio de procedimentos multivariados que envolvem análise de agrupamento, técnica dos componentes principais e função discriminante linear com custos idênticos de má classificação e probabilidades conhecidas de várias populações. Foram utilizados dados do tratamento sem intervenção de 10 anos de monitoramento do Ensaio de Produção Sustentável em Floresta Secundária de Transição, implantado em 1986 pela Companhia Vale do Rio Doce em Rio Vermelho e Serra Azul de Minas, MG, classificados em três grupos ecológicos de espécies florestais: pioneiras, secundárias iniciais e secundárias tardias. O vetor de resposta para cada unidade experimental apresentava o número de árvores por hectare, número de ingressos, número de árvores mortas, área basal, volume, diâmetro médio, incrementos em diâmetro, em área basal e em volume, índice de valor de importância, e regeneração natural. A utilização das análise de componentes principais, de agrupamento e discriminante linear permitiu identificar que as espécies arbóreas devem ser classificadas em maior número (>3) de grupos ecológicos visando melhorar a probabilidade de má-classificação e reduzir a subjetividade da maioria dos pesquisadores na pré-classificação.

Soares *et al.*(2004) avaliam 58 populações de *Roylenchulus reniformis* de amostras de solo e raízes de diferentes culturas e inoculadas em plantas de algodoeiro. Para o estudo morfométrico das populações, os dados foram submetidos a técnicas de análises multivariadas de agrupamento e componentes principais, considerando para a verificação da similaridade entre as populações, a distância euclidiana média. Os resultados mostram a espécie *R. reniformis* como predominante nos agroecossistemas brasileiros.

Souza e Queiroz (2004) objetivam determinar, quantitativamente, utilizando técnicas de análise uni e multidimensionais, a capacidade que alguns caracteres morfológicos apresentam para discriminar o nível de ploidia em plantas de melancia. No estudo, foram consideradas três linhagens diplóides, três linhagens tetraplóides e três híbridos triplóides avaliados quanto ao número de cloroplastos por estômato foliar, largura e comprimento foliar,

relação largura / comprimento, diâmetro do caule, dímetro da corola em flores masculinas e femininas, diâmetro do ovário, peso, diâmetro transversal e longitudinal do fruto, relação diâmetro transversal / diâmetro longitudinal e espessura média da casca. O dendrograma baseado na distância generalizada de Mahalanobis e as regiões de discriminação dispostas no plano cartesiano relativo aos dois primeiros eixos canônicos permitiram diferenciar eficientemente, com baixa taxa de erro de classificação, as plantas diplóides, triplóides e tetraplóides e indicar o número de cloroplastos por estômato foliar como aquele que apresentou maior importância relativa para a formação dos grupos de divergência, enquanto, que os caracteres espessura média da casca e peso médio dos frutos foram os menos contributivos.

Barbosa *et al.* (2005a, 2005b), consideram 367 suínos (fêmeas e machos castrados) abatidos ao atingirem 64,79 (+/-) 5,06 kg provenientes de uma geração F2 (foram formadas duas famílias pelo cruzamento de dois varrões da raça nativa Piau com 18 fêmeas originadas de linhagem desenvolvidas pelo acasalamento de animais das raças Landrace, Large White e Pietnan. A geração F1 foi acasalada *inter si* para produção da geração F2), após jejum de 18 horas, com livre acesso a água fresca e submetidos à insensibilização elétrica. A sangria foi realizada imediatamente após a insensibilização, pela punção do coração, por meio de inserção sob a axila esquerda do animal. No primeiro estudo, objetivou-se utilizar procedimentos multivariados para classificar a contribuição das variáveis na avaliação de características de carcaça com vista ao descarte das menos contributivas ou redundantes em decorrência de possíveis associações. A análise de componentes principais com os eixos canônicos discriminantes de grupos (conglomerados) de informação de variabilidade, utilizando-se a matriz de correlação de todas variáveis observadas, foi a técnica multivariada empregada. Com base nos eixos discriminantes, conclui-se que metade das variáveis analisadas foram relativamente invariantes ou redundantes e, portanto podem ser descartadas em experimentos futuros. No segundo estudo, complementar ao primeiro, buscou-se nas técnicas multivariadas de classificação e associação auxílios para a interpretação das relações entre as variáveis originais e as canônicas. O coeficiente de correlação canônica classificou as variáveis peso aos 77 dias de idade, número de tetas, idade ao abate, peso das costelas e peso da meia-carcaça direita como as mais importantes para explicar a associação entre característica de desempenho e carcaça. Além disso, pode-se considerar que, para as características do desempenho, houve predomínio absoluto do

discriminador peso aos 77 dias de idade, enquanto que, para as características da carcaça, o melhor discriminador foi a idade ao abate.

Khoury Jr *et al.* (2005) discutem, por meio de técnicas de análise multivariada, a capacidade de discriminar defeitos em tábuas de eucalipto, utilizando-se as características de percentis de imagens digitais coloridas. Foram usados 492 blocos contendo os 12 defeitos estudados e madeira limpa, retirados das imagens de 40 tábuas de eucaliptos amostradas aleatoriamente. As características de percentis do histograma das bandas do vermelho, verde e azul, coletados em dois tamanhos de blocos de imagens, foram utilizadas para a construção das funções discriminantes linear e quadrática. As menores taxas de classificação errônea foram 19 e 24% para funções discriminantes lineares com os escores das variáveis canônicas para tamanho de bloco de 64 x 64 e 32 x 32 pixels, respectivamente. As características de percentis foram satisfatórias para discriminar defeitos e madeira limpa em imagens digitais, podendo ser empregadas em sistemas de visão artificial. Com a inserção de outros tipos de classificadores não-paramétricos e posição do defeito, respectivamente, acredita-se que a probabilidade de má-classificação (apesar da magnitude de erro ser considerada baixa) deve ser minorada.

Padovani e Aragon (2005) objetivam desenvolver *software* estatístico para a discriminação e a classificação multivariada em experimentos agronômicos. A metodologia estatística utilizada no pacote computacional consiste da proposta por Fisher, a qual envolve a maximização da relação da estrutura de variação intra e intervariáveis das populações. Associando-se a obtenção dos dois primeiros eixos canônicos, referentes aos dois eixos discriminantes, estabeleceram-se procedimentos para o cálculo da probabilidade de classificação errônea e da quantidade de informação retida na redução do espaço paramétrico. Como exemplo de utilização do sistema computacional de fácil manuseio, denominado “FISHER”, foram considerados dados referentes ao estudo da avaliação do grau de adaptabilidade de diferentes variedades de girassol (*Helianthus annuus*) provenientes de varias origens da América do Sul. O procedimento gráfico mostrou-se bastante eficiente na discriminação das variedades de girassol, com alta taxa de retenção de informação no sistema redimensionado (bidimensional) e baixa porcentagem de classificação errônea de novos elementos.

Souza *et al.* (2005) discutem a divergência genética entre 31 genótipos de melancia por meio de procedimentos multivariados envolvendo a análise de variáveis

canônicas (discriminação canônica) e de técnicas de Tocher e Ward na análise de agrupamento, utilizando-se, como medida de similaridade, a distância generalizada de Mahalanobis. As características observadas para o estudo da divergência foram: o número de dias para o aparecimento da primeira flor masculina, o número de dias para o aparecimento da primeira flor feminina, número de gemas da base da planta até primeira flor masculina, número de gemas da base da planta até a primeira flor feminina, comprimento da rama principal, número médio de frutos por planta, peso médio de fruto, teor de sólidos solúveis, diâmetro transversal do fruto, diâmetro longitudinal do fruto e espessura média da casca. Os resultados dos procedimentos analíticos permitiram formar três grupos pelos métodos de agrupamento e quatro grupos pela dispersão gráfica de Fisher baseada nas duas primeiras componentes canônicas, sendo este último procedimento aquele que melhor representou a divergência genética entre as linhagens de melancia.

Dal'Col Lúcio *et al.*(2006) procurando reduzir o número de variáveis utilizadas para a explicação da variabilidade total das características morfológicas de sementes de espécies florestais exóticas e buscando um modelo consistente e com alta acurácia de discriminação de lotes de sementes, utilizam procedimentos multivariados envolvendo discriminação canônica e análise de agrupamento a partir de 218 análises de sementes de espécies florestais no Centro de Pesquisas Florestais e Conservação do Solo da FPAGRO, localizado em Santa Maria-RS. Considerando-se o banco de dados das informações quantitativas e categorizadas mensuradas nas sementes, foram estabelecidos os grupos similaridade das características morfológicas, destacadas as principais variáveis para o processo de seleção e apresentados os perfis de discriminação dos grupos de cada espécie florestal exótica estuda.

Souza e Souza (2006) apresentam um método objetivando a estratificação em classes homogêneas de estoque volumétrico da floresta ombrófila densa de terra firme não explorada, a partir de dados coletados na unidade de manejo florestal da Fazenda Tracajás, situada no município de Paragominas, PA. O método proposto envolve a utilização de técnicas de análise multivariada (agrupamento e discriminação linear) nos volumes estimados de fuste comercial das árvores selecionadas para corte por talhão, em ordem crescente. A matriz de dados desses volumes teve como elemento genérico X_{ij}, representando o i-ésimo volume classificado no j-ésimo talhão. Os resultados revelaram que o método mostrou-se eficiente na estratificação de florestas tropicais naturais e permite maior precisão (erro de classificação

errônea baixo) das estimativas do inventário florestal e, conseqüentemente, redução de tempo e recursos humanos e financeiros na execução dos levantamentos feitos por amostragem.

Vianna *et al.* (2006) apresentam um novo método de extração de vetores de características de imagens textuais, obtidos a partir do cálculo de distâncias tomadas do contorno das imagens até os pontos de observação dispostos ao redor da mesma feito por meio de modelos neurais e de figuras de mérito estatísticos tais como a análise discriminante de Fisher e distâncias entre os centros de massas. A discriminação Fisheriana foi utilizada como uma medida de avaliação do poder que cada técnica de representação possui para separar as classes de dígitos existentes; entendendo a função discriminante como sendo a diferença máxima entre as médias de cada classe, normalizada pelo espalhamento total do conjunto de amostras. Os resultados da técnica multivariada possibilitaram concluir que o método de representação que apresenta a melhor separação entre as classes de dígitos é o método do quadrado seguidos em ordem de qualidade de diferenciação, pelos métodos do octógono, raio, diâmetro e matriz de bits. Finalizam seu trabalho realizando uma comparação entre os resultados das avaliações obtidos com critério discriminante de Fisher, o critério das distâncias entre os centros de massas e os apresentados pelos modelos neurais, concluindo-se que os melhores desempenhos ocorreram neste último procedimento.

3. DESENVOLVIMENTO METODOLÓGICO

3.1 Considerações Gerais

A caracterização de um procedimento analítico estatístico com dados multivariados pode ser entendida como a possibilidade de análise simultânea das variáveis observadas nas parcelas ou unidades experimentais, com pleno aproveitamento de toda a estrutura de variação envolvida no fenômeno biológico. Embora a idéia esteja apresentada de forma muito simplista, não é difícil imaginar a complexidade da estrutura de variabilidade envolvida nos dados, pois há que se considerar a variação dentro de cada característica observada e a variação entre as características. A dispersão dos dados experimentais deve ser apresentada por medidas de variabilidade envolvendo variâncias e covariâncias. Esta multiplicidade de características pode ser também estendida à forma de abordagem de dados observacionais (MURTEIRA,1993).

Os métodos e procedimentos multivariados mostram-se como alternativas extremamente eficientes e poderosas quando a situação exige uma combinação de múltiplas informações procedentes de uma parcela experimental (ou seja, de um vetor observacional), com a finalidade de associar ou predizer fenômenos biológicos baseando-se em

um complexo de variáveis importantes para o desenvolvimento do plano experimental (DILLON; GOLDSTEIN, 1984).

Embora exista o reconhecimento da eficiência dos procedimentos multivariados como métodos que propiciam o enriquecimento das conclusões oriundas dos dados experimentais e que possibilitam maior acurácia e fidedignidade dos modelos estatísticos associados ao fenômeno biológico, para seu uso, necessita-se dispor de resultados computacionais e alguns conhecimentos básicos de manuseio dos programas existentes. Nas últimas três décadas, a utilização das técnicas multivariadas nas diversas áreas do conhecimento, em especial na área de ciências agrárias, graças ao avanço tecnológico no campo computacional, tornou-se realidade frente ao desenvolvimento de pacotes estatísticos genéricos, destacando-se entre os mais conhecidos o *SPSS, SAS, BMDP,* que foram criados na década de 80, para *mainframes*, posteriormente adaptados para microcomputadores. Esses pacotes foram gerados para funcionarem no sistema operacional Windows, o BMDP também em UNIX, sendo todos desenvolvidos por fabricantes americanos (MORETTIN; BUSSAB, 2003).

Conforme descreve Rencher (1995), a Análise Multivariada pode ser conceituada como conjunto de métodos e técnicas que permitem a análise simultânea de dados observados para um ou mais conjuntos de indivíduos (populações ou amostras) caracterizados por duas ou mais variáveis correlacionadas entre si. Esta conceituação imprescinde que os tipos de métodos e técnicas permitam analisar relações simultâneas entre as variáveis e entre os indivíduos, ou seja, dentro de variáveis.

Em estudos biológicos, é comum considerar um número elevado de variáveis aleatórias (respostas) correlacionadas entre si. Essa estrutura do vetor de observação não deixa de refletir a estrutura biológica que é um sistema altamente integrado, no qual os caracteres e/ou componentes estão internamente relacionados por meio de suas dependências ou interdependências (JOHNSON; WICHERN, 1998).

Inicialmente, as teorias e conceitos desenvolvidos na Análise Multivariada, por vários autores, tinham por base a distribuição multinormal de probabilidades e consistiam em generalizações dos procedimentos univariados. Mais recentemente, incorporou-se à literatura especializada a análise exploratória de dados multidimensionais mostrando que a terminologia multivariada não se limita apenas a examinar relações entre duas ou mais variáveis que seguem uma distribuição normal multivariada.

Tomando a cronologia das principais obras de referência em Análise Multivariada, segundo Reis (1997), a conceituação dos procedimentos multidimensionais teve seu início no começo do século XX, a partir dos relatos de Pearson (1901), Fisher (1928), Hotelling (1931), Wilks (1932) e Bartlett (1937). Na seqüência, numa fase intermediária, destacam-se as obras de Kendall (1957, 1975), Anderson (1958, 1984), Morrison (1967, 1976) e Mardia, Kent e Bibby (1979). Mais recentemente, têm-se algumas menções voltadas às áreas aplicadas e outras já incorporando programas computacionais estatísticos para aplicação de métodos multivariados, destacando-se Chatfield e Colins (1980), Dillon e Goldstein (1984), Monly (1986), Hair, Anderson, Tathan e Black (1987), Everit e Dunn (1991). Deve ser acrescentada nesta cronologia a obra de Johnson e Wichern, publicada em 1982, a qual aprofunda a argumentação matemática na discussão dos procedimentos. Atualmente, existe uma incontável bibliografia de procedimentos multivariados concernentes a todas as áreas do conhecimento científico.

Segundo Kendal (1950), classificam-se as técnicas da Análise Multidimensional ou Multivariada em:

I. Análise de Interdependência: estuda as relações de um conjunto de variáveis entre si.

1) Análise de Agrupamento.
2) Análise de Componentes Principais.
3) Análise de Fatores.

II. Análise de Dependência: estuda a dependência de uma ou mais variáveis em relação às outras.

1) Análise Discriminante.
2) Análise de Variância.
3) Análise de Medidas Repetidas.
4) Análise de Regressão.
5) Análise de Correlação Canônica

3.2 Funções Discriminantes

Os termos “discriminar” e “classificar” foram introduzidos na Estatística por *Sir Ronald Aylmer Fisher* no primeiro tratamento moderno dos problemas de separação de conjuntos na década de 30 (Johnson *et al.* 1998). Dadas duas populações $\prod_1$ e $\prod_2$ de observações multivariadas $(\underset{\sim}{X})$, a idéia de Fisher foi transformar estas observações multivariadas em observações univariadas (Y *)*, de tal modo que as populações transformadas (grupos) estivessem separadas tanto quanto possível para facilitar a indicação (pertinência) de novos indivíduos a uma destas populações.

A Análise Discriminante consiste numa técnica multivariada empregada com objetivo de diferenciar populações ou na classificação de um novo indivíduo em uma de várias populações, considerando a estrutura multidimensional dos dados observados. Seu emprego foi originalmente em Botânica, cuja aplicação teve como objetivo fazer a diferenciação de grupos de plantas com base no tamanho e tipo de folhas, para que, posteriormente, fosse possível classificar as novas espécies encontradas (Fisher, 1936). Em geral, o propósito conceitual da Análise Discriminante consiste em encontrar a separação máxima entre as populações considerando a maximização da diferença entre as médias das populações relativamente aos desvios-padrão no interior de cada população, sem perder a estrutura de covariância das variáveis observadas.

Resumindo, a Análise Discriminante constitui-se em um método estatístico para classificar ou alocar indivíduos em uma das populações estudadas, considerando o vetor de respostas de dados com estrutura multidimensional e identificar as divergências entre (explicar diferenças) populações. Para isso, são determinadas combinações lineares dessas variáveis que discriminam entre grupos definidos “a priori”, de tal modo que seja minimizada a probabilidade de classificação errônea “a posteriori”.

Na literatura especializada as análises discriminantes e de agrupamento têm-se complementadas em várias aplicações práticas como ferramentas do procedimento de mineração de dados (“Data Mining”), nas diversas áreas do conhecimento, considerando a primeira como método supervisionado, enquanto a última, não-supervisionado. Essa diferenciação conceitual torna-se marcante quando evidenciamos os direcionamentos de propósito das duas técnicas. Na análise discriminante, deseja-se formar grupos homogêneos (regiões de classificação) a partir do conhecimento “a priori” a quais populações pertencem as unidades amostrais, ou seja, coleta-se uma amostra aleatória de cada população multivariada (p-

dimensional) e estabelece-se o critério de classificação nos grupos. Na análise de agrupamentos a constituição dos grupos homogêneos na amostra em estudo é efetivada sem o conhecimento "a priori" da alocação das unidades experimentais multivariadas nos grupos (BARROSO; ARTES, 2003).

No contexto da discriminação entre populações, um outro procedimento multivariado tem tido um destaque de uso bastante relevante. Este procedimento trata da Análise de Componentes Principais. Deve ser ressaltado que a análise dos componentes principais trata de uma técnica estatística que por meio de transformações lineares, projeta o conjunto original p-dimensional em um conjunto com número menor ($k < p$) de variáveis não-correlacionadas, que explica substancialmente as informações da variação dos dados do conjunto inicial. Para a construção das componentes, não se requer qualquer suposição sobre a distribuição probabilística do vetor de dados e utiliza-se apenas a estrutura de variação intra-inter variáveis em estudo, dada ou pela matriz de covariância ou pela matriz de correlação. Geometricamente, as componentes principais são as coordenadas das unidades amostrais no sistema de eixos cartesianos obtido pela rotação do sistema de eixos original, na direção de variabilidade máxima dos dados (TABACHNICK; FIDELL, 2001). A consistência geométrica da técnica de componentes principais possibilita a construção de gráficos bi e tridimensionais, mais acentuada no bidimensional, como um procedimento interessante e muito utilizado na discriminação de populações e classificação de novos indivíduos, por meio dos eixos canônicos.

Torna-se interessante enfocar a interpretação espacial da análise discriminante para um melhor entendimento prático da técnica de classificação. Para isto, suponha um sistema de pontuação em que a cada unidade multidimensional experimental faz-se corresponder um escore resultante de uma média ponderada dos valores assumidos pelas variáveis componentes do vetor resposta da respectiva unidade experimental. Uma vez obtida essa pontuação total da parcela, o escore pode ser transformado em probabilidades "a posteriori" da unidade experimental pertencer a cada um dos grupos de estudo, cuja decisão do qual mais provável pode ser feita considerando as probabilidades.

No campo da experimentação agronômica a técnica da análise discriminante tem-se mostrado bastante prática em situações em que as *p* variáveis observadas, no vetor de resposta, são quantitativas. Não obstante, pode-se utilizar a análise discriminante em situações em que o vetor de resposta apresenta variáveis mistas (quantitativas e qualitativas –

nominais e ordinárias), porém, nestas conjunturas, é mais comum utilizar a regressão logística, as árvores de classificação ou as redes neurais artificiais. Uma discussão interessante e abrangente sobre comparações entre esses procedimentos analíticos, enfocados a partir de dados de concessão de créditos a consumidores por bancos, supermercados, lojas comerciais e assemelhados pode ser vista em (OHTOSHI, 2003).

O método de discriminação para mais de duas populações (discriminação múltipla) foi desenvolvido à semelhança da construção da função discriminante linear de Fisher para duas populações (discriminação simples).

No caso da discriminação simples, a maximização da diferença de médias das populações equivale a estabelecer uma taxa mínima de classificação errônea, cujo novo eixo resultante da combinação linear das variáveis observadas permite encontrar uma separação máxima entre as populações previamente definidas.

Na discriminação múltipla, o procedimento será similar , o qual consiste em encontrar os eixos sobre os quais projetarão as populações de tal modo que seja maximizada a soma de quadrados entre as populações relativamente à soma de quadrados dentro das populações. Para o desenvolvimento teórico da construção das funções lineares para discriminação múltipla torna-se fundamental a utilização de Álgebra de Matrizes e a estruturação da soma de quadrados em termos de Formas Quadráticas.

Graybill (1983) e Noble e Daniel (1988) mostram que a Álgebra Linear pode ser extensivamente utilizada no desenvolvimento de teorias e aplicações estatísticas. Um conjunto de dados multivariados (p-dimensional), considerando g grupos experimentais, pode ser apresentado num arranjo matricial na seguinte forma:

$$Y = \begin{bmatrix} y_{111} & y_{112} & \cdots & y_{11P} \\ \vdots & \vdots & & \vdots \\ y_{1n_1 1} & y_{1n_1 2} & \cdots & y_{1n_1 P} \\ y_{211} & y_{212} & \cdots & y_{21P} \\ \vdots & \vdots & & \vdots \\ y_{2n_2 1} & y_{2n_2 2} & \cdots & y_{2n_2 P} \\ \vdots & \vdots & & \vdots \\ y_{g11} & y_{g12} & \cdots & y_{g1p} \\ \vdots & \vdots & & \vdots \\ y_{gn_g 1} & y_{gn_g 2} & \cdots & y_{gn_g p} \end{bmatrix}$$

ou genericamente,

$$Y = \left[\quad y_{hij} \right]$$

onde **Y** é de ordem **n x p**, sendo $\mathbf{n = n_1 + \cdots + n_g}$; $\mathbf{h = 1, \cdots, g}$ (índice de população); $\mathbf{i = 1, \cdots, n_h}$ (repetições dentro da população) e $\mathbf{j = 1, \cdots p}$ (número de variáveis observadas).

O elemento genérico y_{hij} refere-se ao valor numérico observado na j-ésima resposta da i-ésima unidade experimental (parcela) da h-ésima população.

Uma alternativa interessante para a matriz **Y** é sua representação em submatrizes $\mathbf{Y_h}$ associadas ao grupo populacional em consideração, descritas por

$$Y_h = \begin{bmatrix} y_{h11} & y_{h12} & \cdots & y_{h1p} \\ \vdots & \vdots & & \vdots \\ y_{hn_h1} & y_{hn_h1} & \cdots & y_{hn_hp} \end{bmatrix} = \left[\underset{\sim}{Y}'_{hi} \right]$$

onde $\mathbf{Y_h}$ é uma matriz de ordem $\mathbf{n_h x\ p}$, para $\mathbf{h = 1, \cdots, g}$ e $\underset{\sim}{Y}`_{hi}$ o vetor linha correspondente a i-ésima observação da h-ésima população.

A partir desta consideração pode-se representar $\boldsymbol{Y}$ por:

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 (\boldsymbol{n}_1 \times \boldsymbol{p}) \\ \vdots \\ \boldsymbol{Y}_h (\boldsymbol{n}_h \times \boldsymbol{p}) \\ \vdots \\ \boldsymbol{Y}_g (\boldsymbol{n}_g \times \boldsymbol{p}) \end{bmatrix}$$

onde, nas linhas de $\mathbf{Y_h}$ estão alocados os vetores respostas das unidades da h-ésima população, ou seja, cada vetor-linha da submatriz representa a resposta multidimensional de uma unidade experimental e, cada vetor coluna, as respostas na população *h*, observadas em uma dada variável aleatória (j =1,...p).

Para $n_1,\ldots,n_h$ indivíduos alocados em g amostras (grupos) das populações $\pi_1,\ldots,\pi_g$, respectivamente, e caracterizados por p variáveis, algumas matrizes de interesse da estrutura de variabilidade dos dados multivariados podem ser calculadas.

A matriz da soma total de quadrados e produtos cruzados apresenta a estrutura de variação total dos dados dada por:

$$T = \sum_{h=}^{g}\sum_{i=1}^{n_h}\left(\underset{\sim}{Y_{hi}} - \underset{\sim}{\overline{Y}}\right)\left(\underset{\sim}{Y_{hi}} - \underset{\sim}{\overline{Y}}\right)',$$

onde $\underset{\sim\ hi}{Y}$ é o i-ésimo vetor linha da submatriz Y_h e $\underset{\sim}{\overline{Y}}$, o vetor geral de médias.

Para cada grupo, a matriz de soma de quadrados e produtos cruzados responsável pela variação dentro do grupo *h* pode ser estabelecida por:

$$W_h = \sum_{i=1}^{n_k}\left(\underset{\sim\ hi}{Y} - \underset{\sim}{\overline{Y_h}}\right)\left(\underset{\sim}{Y_{hi}} - \underset{\sim}{\overline{Y}_h}\right)',$$

com Y_{hi} especificado anteriormente e $\underset{\sim\ h}{\overline{Y}}$, o vetor de médias do grupo h.

Para instituir um indicador da variação conjunta dentro dos *g* grupos basta calcular:

$$W = \sum_{h=1}^{g} W_h$$

a matriz da soma de quadrados e produto dentro de grupos.

Para a variação entre grupos, a matriz soma de quadrados e produtos cruzados pode ser especificada por:

$$B = \sum_{h=1}^{g} n_h\left(\underset{\sim\ h}{\overline{Y}} - \underset{\sim}{\overline{Y}}\right)\left(\underset{\sim\ h}{\overline{Y}} - \underset{\sim}{\overline{Y}}\right)'$$

Alternativamente, tem-se, para a matriz de soma de quadrados e produtos entre grupos (B), seguinte expressão:

$$B = T - W.$$

Uma observação interessante à respeito das matrizes ***T,W e B,*** consiste no emprego destas na técnica da análise de variância multivariada para inferir sobre o teste de hipóteses da igualdade dos vetores das *g* populações. Estas matrizes permitem estabelecer quatro estatísticas para o teste de hipóteses: princípio da união-insterseção de Roy (maior raiz

característica de $\boldsymbol{B}(\boldsymbol{B}+\boldsymbol{W})^{-1}$), razão de verossimilhança de Wilks$\left(\frac{\mathbf{det}(\boldsymbol{W})}{\mathbf{det}(\boldsymbol{B}+\boldsymbol{W})}\right)$, traço de Pilai $\left(\boldsymbol{tr}\,\boldsymbol{B}(\boldsymbol{B}+\boldsymbol{W})^{-1}\right)$ e traço de Lawley-Hotelling $\left(\boldsymbol{tr}\,\boldsymbol{B}\boldsymbol{W}^{-1}\right)$.

Outra utilização importante das matrizes de somas de quadrados e produtos está na construção de regiões gráficas de discriminação. Fisher (1938) propõe uma técnica de separação de duas populações que consiste em transformar a observação multivariada (vector das respostas experimentais) em uma observação univariada tal que, os valores transformados relativos as observações provenientes das populações $\pi_1 \boldsymbol{e}\, \pi_2$, estivessem separados tanto quanto possível. Na seqüência , Fisher estabelece a extensão do método para várias populações visando buscar uma técnica exploratória interessante de separação com propósitos de inspeção visual e descrição gráfica das regiões das populações. Os propósitos ficam bem estabelecidos quando se consegue uma representação razoável da estrutura de variabilidade no novo sistema referencial.

Para a extensão são consideradas duas matrizes de variação, uma indicando a variação dentro das populações representada pela *W*, definida anteriormente; e outra indicando a variação entre populações, representada por B, cujas expressões nas descritas anteriormente.

Os eixos discriminantes de Fisher são determinados conforme detalhes na seqüência:

1. sejam $\lambda_1, \lambda_2, \ldots, \lambda_s \rangle 0$ os $s \le \mathbf{min}(g-1, p)$ autovalores não nulos de $W^{-1}B$ e $\underset{\sim}{e}_1, \underset{\sim}{e}_2, \ldots, \underset{\sim}{e}_s$ os correspondentes autovalores condicionados a restrição $\underset{\sim}{e}_m' \, S \, \underset{\sim}{e}_m = 1, \quad m = 1, \ldots, s$;
2. o vetor de coeficientes $\underset{\sim}{\ell}$ que maximiza a razão $\frac{\underset{\sim}{\ell}' B \underset{\sim}{\ell}}{\underset{\sim}{\ell}' W \underset{\sim}{\ell}}$ é dado por $\underset{\sim}{\ell}_1 = \underset{\sim}{e}_1$;
3. a combinação linear $\underset{\sim}{\ell}_1' \underset{\sim}{Y}$ é chamada de primeira discriminante amostral (primeiro eixo discriminante). A escolha de $\underset{\sim}{\ell}_2 = \underset{\sim}{e}_2$

produz a segunda discriminante amostral $\underset{\sim}{\ell}_2' \underset{\sim}{Y}$. Em continuidade, $\underset{\sim}{\ell}_k' \underset{\sim}{Y}$ é a k-ésima discriminante amostral $k \leq s$.

Duas observações interessantes devem ser feitas para os eixos discriminantes:

1. para usar a matriz S (ou alternativamente, $S = \frac{1}{n-g} W$) como estimativa de $\sum$ tornar-se necessário que as matrizes de covariâncias das *g* populações sejam iguais;
2. $\underset{\sim}{\ell}_k \; S \; \underset{\sim}{\ell}_t' = 1$, se $k = t \leq s$ e $\underset{\sim}{\ell}_k' \; S \; \underset{\sim}{\ell}_t = 0$, nos demais casos.

As discriminantes amostrais ou eixos discriminantes permitem estabelecer uma regra de classificação de um novo indivíduo em uma única das g populações. Esta regra de classificação baseada nos $r \leq s$ primeiros eixos discriminantes, denominada de Procedimento de Classificação de Fisher Fundamentado nas Discriminantes Amostrais, estabelece que: o indivíduo $\underset{\sim}{Y}$ deve ser alocado em π_h se $\sum_{j=1}^{r} \left[\underset{\sim}{\ell}_j' \left(\underset{\sim}{Y} - \underset{\sim}{\bar{Y}}_h \right) \right]^2 \leq \sum_{j=1}^{r} \left[\underset{\sim}{\ell}_j' \left(\underset{\sim}{Y} - \bar{Y}_{h'} \right) \right]^2$ para todo $h' \neq h, \; h = 1, \ldots, g$.

3.3 Critério Geral de Classificação

Seja $f_h(y)$ a função densidade de probabilidade associada à população π_h, com $h = 1, \ldots, g$, e as seguintes considerações:

1. $p_h \rightarrow$ a probabilidade "a priori" da observação pertencer à população π_h;
2. $c(k|h) \rightarrow$ o custo de classificação de um indivíduo de π_h em π_k (para *k=h, c(h, h)=0*);

3. $P(k|h)$ a probabilidade de se classificar um indivíduo em π_k quando na verdade ele é de π_h, ou seja,

$$P(k|h) = \int_{R_k} f(\underset{\sim}{y})\, d\underset{\sim}{y},$$

onde R_k constitui o conjunto dos $\underset{\sim}{y}$ classificados em π_k.

Segundo Barroso e Artes (2003), tem-se que o custo esperado de erro (CEE) ao classificar $\underset{\sim}{y}$ de π_1 em $\pi_2, \pi_3, \ldots ou\ \pi_g$ é obtido como:

$$\text{CEE(1)} = \sum_{k=2}^{g} P(k|1)c(k|1)$$

Este custo esperado CEE(1) ocorre com probabilidade p_1, e assim, semelhantemente até o custo CEE(g), com probabilidade p_g.

Nestas condições, o custo esperado de erro classificatório (custo total) é dado por:

$$CEET = \sum_{h=1}^{g} p_h\ CEE(h) = \sum_{h=1}^{g} p_h \left[\sum_{k=1, k\neq h}^{g} P(k|h)c(k|h) \right].$$

A regra de classificação consiste em determinar as regiões (conjunto) $R_1, R_2, \ldots, R_g$ que minimizem o custo total esperado de erro classificatório (CEET). Ou seja, o procedimento indica em alocar $\underset{\sim}{y}$ na população $k = 1,2,\ldots,g$ para qual

$$\sum_{h=1, h\neq k}^{g} p_h f(\underset{\sim}{y}) c(k \mid h)$$

é menor (eventualmente, se ocorrer um empate, $\underset{\sim}{Y}$ pode ser classificado em qualquer uma das populações para as quais o empate ocorre).

Considerando os custos de má-classificação todos iguais (sem perda da generalidade pode-se considerar todos os custos de má-classificação igual à unidade), aloca-se $\underset{\sim}{Y}$ na população π_k em que $\sum_{h=1, h\neq k}^{g} p_h f_h(\underset{\sim}{y})$ é menor.

Se considerar que essa quantidade será menor quando o termo excluído $p_h f_h(\underset{\sim}{y})$ for maior, tem-se a seguinte regra de classificação para o critério geral pelo custo esperado de erro mínimo, com custos iguais por falhas na classificação:

alocar $\underset{\sim}{Y}$ em π_k se $p_k f_k(\underset{\sim}{y}) \rangle p_h f_h(\underset{\sim}{y})$ para todo $h \neq k$.

3.4 Classificação para Populações Normais.

Os testes multivariados de normalidade são em geral baseados em distribuições marginais das variáveis. Uma das possibilidades interessante, alternativa aos testes com distribuições marginais, consiste em utilizar testes indiretos. Um dos procedimentos mais utilizados refere-se aos testes de assimetria e curtose (MARDIA, 1970), embora os resultados estabeleçam condições necessárias à normalidade, mas não suficientes. Ou seja, não existem garantias de que a não rejeição da hipótese de distribuição simétrica e mesocúrtica possa ser devida a normalidade multivariada, mas sim a normalidade univariada para cada variável.

Neste sentido, considerando os momentos centrados na média amostral, de 2ª (m2), 3ª (m3) e 4ª (m4) ordens, tem-se para uma amostra de tamanho *n* as seguintes estimativas para a assimetria (g_1) e curtose (g_2), com os respectivos erros-padrão:

$$Assimetria \begin{cases} g_1 = m_3 / (m_2)^{3/2} \\ ep(g_1) = (6/n)^{1/2} \end{cases}$$

$$Curtose \begin{cases} g_2 = m_4 / (m_2)^2 - 3 \\ ep(g_2) = (24/n)^{1/2} \end{cases}$$

Os testes estatísticos assintóticos para verificar se a distribuição é simétrica e mesocúrtica são dados por, respectivamente:

$$z_1 = g_1 / ep(g_1) \approx N(0,1)$$

e

$$z_2 = g_2 / ep(g_2) \approx N(0,1),$$

com a regra de decisão habitual (MARDIA, 1970).

Seja $\underset{\sim}{Y} \sim N_p\left(\underset{\sim}{\mu_h}, \sum_h\right)$, para $h = 1, \ldots, g$, ou ainda,

$$f_h(\underset{\sim}{y}) = \frac{1}{(2\pi)^{p/2}\left|\sum_h\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\underset{\sim}{y}-\underset{\sim h}{\mu}\right)' \sum_h^{-1}\left(\underset{\sim}{y}-\underset{\sim h}{\mu}\right)\right].$$

Mingoti (2005) descreve que, considerando, *c(h|h) =0* e *c(k|h)=1*, $k = 1,\ldots,g$, $h = 1,\ldots,g$, $k \neq h$ (custos de má-classificação iguais e unitários), o critério de classificação anterior expresso no logaritmo da inequação, indica a seguinte regra:

alocar $\underset{\sim}{Y}$ em π_k se

$$Ln\left(p_k\, f_h(\underset{\sim}{y})\right) = Ln(p_k) - \frac{p}{2}Ln(2\pi) - \frac{1}{2}Ln\left(\sum\nolimits_k\right) - \frac{1}{2}\left(\underset{\sim}{y}-\underset{\sim k}{\mu}\right)' \sum\nolimits_k^{-1}\left(\underset{\sim}{y}-\underset{\sim}{\mu_k}\right) = \max_h Ln\left(p_h\, f_h(\underset{\sim}{y})\right).$$

A constante $\frac{p}{2}\ell n(2\pi)$ é a mesma para todas as populações, portanto, pode ser ignorada.

Considerando os estimadores de máxima verossimilhança de μ_h e $\sum_h$, tem se a função discriminante, discriminante quadrática representada pelo escore quadrado ($Q_h(\underset{\sim}{y})$)de classificação:

$$Q_h(\underset{\sim}{y}) = -\frac{1}{2}\ell n|S_h| - \frac{1}{2}\left(\underset{\sim}{y}-\underset{\sim h}{\bar{y}}\right)' S_h^{-1}\left(\underset{\sim}{y}-\underset{\sim h}{\bar{y}}\right) + \ell n p_h.$$

Para várias populações normais, a regra de classificação consiste em indicar $\underset{\sim}{Y}$ em π_k se $Q_k(\underset{\sim}{y}) = \mathbf{max}\left(Q_1(\underset{\sim}{y}),\ldots,Q_g(\underset{\sim}{y})\right)$.

No caso em que as matrizes de covariância são homogêneas, isto é, as matrizes de todas as populações são iguais $\left(\sum_1 = \sum_2 = \ldots = \sum_g = \sum\right)$, os termos que dependem exclusivamente de $\sum_h$ passam a ser constantes para as g populações e, portanto, podem ser ignorados. Nesse sentido, o escore de classificação passa a ser linear, dado por

$$d_h\left(\underset{\sim}{y}\right) = \underset{\sim}{\bar{y}}'_h\, S^{-1}\, \underset{\sim}{y} - \frac{1}{2}\underset{\sim h}{\bar{y}}'\, S^{-1}\, \underset{\sim h}{\bar{y}} + \ell n p_h, \text{ com}$$

$$S = \frac{\left[\sum_{h=1}^{g}(n_h - 1)S_h\right]}{(n-g)}.$$

A regra de classificação pela função discriminante linear para o critério geral do custo esperado de erro mínimo, com custos iguais por falhas na classificação é dada por :

$$\text{alocar } \underset{\sim}{Y} \text{ em } \pi_k \text{ se } d_k(\underset{\sim}{y}) = \mathbf{max}\left(d_1(\underset{\sim}{y}),\ldots,d_g(\underset{\sim}{y}) \right)$$

3.5 Probabilidade de Classificações Errôneas

Em situações práticas que avaliam somente duas populações, existem dois erros de classificação que precisam ser considerados. O primeiro refere-se à classificação do indivíduo em π_1 quando este na realidade pertence a π_2; o outro, na situação contrária, ou seja, classificar em π_2 quando o indivíduo pertence a π_1. Esta abordagem pode ser generalizada para um número *g* de populações, onde as possibilidades de erros de classificação tornam-se bem mais numerosas.

Do ponto de vista da acurácia do procedimento de classificação, quanto menor forem estas probabilidades, mais eficiente será a qualidade da função discriminante estabelecida para o estudo de alocação de novos indivíduos.

Existem alguns métodos utilizados para se estimarem as probabilidades de classificação errôneas, destacando-se procedimentos com conjecturas paramétricas (envolvendo a distribuição normal de probabilidades) ou não-paramétricas (freqüências de ocorrência).

Em relação aos métodos não-paramétricos, três procedimentos são os mais usuais: método da Ressubstituição, método de Colocação de Elementos à Parte e método de Lachenbruch (MINGOTI, 2005).Quanto ao método paramétricos(distribuição multinormal dos dados) o mais comum trata-se de desenvolver o cálculo das probabilidades da distância generalizada de Mahalanobis entre os centróides das populações.

3.5.1 Método da Ressubstituição

Os escores discriminantes de cada indivíduo observado das *g* populações são calculados, sendo a regra de discriminação utilizada para classificar todos os $n_1 + n_2 + \ldots + n_g$ elementos observados. Uma porcentagem alta de acerto na classificação dos

elementos amostrais em relação à população a que de fato pertencem indica a boa qualidade da função discriminante. Em resumo, os mesmos indivíduos participam da construção da função discriminante e da estimação dos erros de classificação.

3.5.2 Método de Colocação de Elementos à Parte

O conjunto total de elementos $n_1 + n_2 + \ldots + n_g$, é repartido em duas partes com números diferentes de indivíduos, uma fração maior utilizada na construção da função discriminante, e outra, a menor, para a estimação das probabilidades de erros de classificação. Uma vantagem muito interessante em relação ao anterior, deve-se ao fato que as estimativas das probabilidades de erros são imparciais (não-viesadas). A desvantagem consiste na redução do tamanho amostral original para a construção da regra de discriminação, o que pode diminuir a acurácia do procedimento se as amostras não forem grandes.

Johnson & Wichern (2002) recomendam deixar à parte de 25 a 50% dos elementos para a estimação das probabilidades de erro, consideração limitante para o emprego do método em pequenas amostras. Porém, deve-se destacar que este método é melhor que o anterior quando se tem grandes amostras.

3.5.3 Método de Lachenbruch

Conhecido na literatura como método de validação cruzada("cross-validation") ou de Lachenbruch consiste em, partindo das $n_1 + n_2 + \ldots + n_g$ observações das g populações, considerar os seguintes passos (LACHENBRUCH & MICKEY, 1968):

I. Retira-se um indivíduo do total de observados e utilizam-se as $n_1 + n_2 + \ldots + n_g - 1$ unidades experimentais restantes para construir a função discriminante;
II. classifica-se o indivíduo retirado sob a regra construída; anotando se a classificação foi correta ou não;

III. retorna-se o indivíduo que foi retirado ao conjunto original e retira-se uma outra unidade experimental diferente da primeira e, em seguida, os passos anteriores(I e II) são repetidos.

Os três passos devem ser repetidos para todos as $n_1 + n_2 + \ldots + n_g$ unidades experimentais e as probabilidades de erros são determinadas (freqüências de ocorrência de classificações incorretas).

Timm (2002) indica que as estimativas deste método são aproximadamente não viciadas e melhores que o método da ressubstituição para populações normais e não normais.

Uma consideração importante para descrever a qualidade da função discriminante consiste em determinar a estimativa da probabilidade global de acerto. O valor desta probabilidade associado às estimativas de probabilidades de erros de má-classificação estabelecem um indicador freqüentista da acurácia da regra de classificação.

Ou seja, considerando *Erro(k|h)*, o erro da unidade ser proveniente da população *h*, mas classificada na população k pela regra de discriminação utilizada, para $h e k = 1, \ldots, g \; e \; h \neq k$, tem-se:

$$P(k \mid h) = \frac{n_{hk}}{n_h},$$

onde n_{hk} é o número de elementos da população *h* classificados incorretamente em *k*.

Uma forma prática para determinar as probabilidades de classificação pode ser estabelecer pela "matriz de confusão" (BARROSO; ARTES, 2003) descrita na tabela de dupla entrada (Tabela 1)

Tabela 1: Distribuição dos indivíduos segundo as populações de Origem e classificação

População Verdadeira	População Classificada (Regra)				Total
	π_1	π_2	...	π_g	
π_1	n_{11}	n_{12}	...	n_{1g}	n_1
π_2	n_{21}	n_{22}	...	n_{2g}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
π_g	n_{g1}	n_{g2}	...	n_{gg}	n_g
Total	$\hat{n}_1$	$\hat{n}_2$	...	$\hat{n}_g$	n

Onde

$n_{hk} \longrightarrow$ número de elementos de π_h classificados em π_k;

$n_h \longrightarrow$ número de elementos de π_h;

$\hat{n}_h \longrightarrow$ número de elementos classificados em π_h;

$n \longrightarrow$ número total de elementos.

A estimativa da probabilidade global de acerto pode ser expressa por:

$$P(acerto) = \frac{\sum_{h=1}^{g} n_{hh}}{n},$$

enquanto que a estimativa da probabilidade global de erro torna-se expressa na unidade complementar descrita por:

$$P(erro\ global) = \frac{n - \sum_{h=1}^{g} n_{hh}}{n}.$$

Deve ser lembrado que essa taxa de erro é geral e engloba todos os possíveis tipos de erro de classificação sob a regra em consideração.

3.5.4. Método da Distância de Mahalanobis

Quando o vetor de observações para as populações apresenta distribuição de probabilidades normal p-variada homocedásticas, torna-se mais interessante obter as estimativas pelo método paramétrico.

Para o caso de duas populações multinormais homocedásticas com custos iguais por falhas na classificação, tem-se:

$\underset{\sim}{Y}$ será classificado em π_1 se $\frac{f_1(\underset{\sim}{y})}{f_2(\underset{\sim}{y})} \geq 1$ e

$\underset{\sim}{Y}$ será classificado em π_2 se $\frac{f_1(\underset{\sim}{y})}{f_2(\underset{\sim}{y})} < 1$.

Porém ,

$$\frac{f_1(\underset{\sim}{y})}{f_2(\underset{\sim}{y})} = \exp\left[(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1} \underset{\sim}{y} - \frac{1}{2}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1}\left(\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}\right)\right].$$

Aplicando-se a função crescente $\log_e$ a ambos os lados, tem-se

$$\log_e \frac{f_1(\underset{\sim}{y})}{f_2(\underset{\sim}{y})} = (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1} \underset{\sim}{y} - \frac{1}{2}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1}\left(\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}\right)$$

e, portanto, $\underset{\sim}{Y}$ será classificado em π_1 se

$$(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1} \underset{\sim}{y} \geq \frac{1}{2}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1}\left(\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}\right)$$

e $\underset{\sim}{Y}$ será classificado em π_2 se

$$(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1} \underset{\sim}{y} < \frac{1}{2}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1}\left(\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}\right).$$

Seja $FDL(\underset{\sim}{y}) = (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1} \underset{\sim}{y}$ então

$$E\left[FDL(\underset{\sim}{y}) / \pi_h\right] = (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1} \underset{\sim}{\mu_h} \text{ e}$$

$$Var\left[FDL(\underset{\sim}{y}) / \pi_h\right] = (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1} Var(\underset{\sim}{y}) \sum^{-1} (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2}) = (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1} (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2}) = \delta_{12}^2 \quad \text{onde} \quad \delta_{12}^2$$

é a distância generalizada de Mahalanobis e $FDL(\underset{\sim}{y})$ normal p-variada sob π_1 e π_2 (GODOI, 1985).

As probabilidades de classificação errônea podem ser obtidas por:

$$P(\pi_1 / \pi_2) = P(FDL(\underset{\sim}{y}) \geq \frac{1}{2}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1} (\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}) / N((\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}) \sum^{-1} \underset{\sim}{\mu_2}; \delta_{12}^2)) = P\left(Z \geq \frac{1}{2}\delta_{12}\right)$$

$$= 1 - \Phi\left(\frac{1}{2}\delta_{12}\right);$$

$$P(\pi_2 / \pi_1) = P\left(FDL(\underset{\sim}{y}) < \frac{1}{2}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})' \sum^{-1} (\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}) / N(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2}) \sum^{-1} \underset{\sim}{\mu_1}; \delta_{12}^2\right)$$

$$= P\left(Z < -\frac{1}{2}\delta_{12}\right) = \Phi\left(-\frac{1}{2}\delta_{12}\right)$$

À semelhança do cálculo freqüentista, tem –se a seguinte tabela de probabilidades de classificação errônea (Tabela 2):

Tabela 2: Probabilidade de classificação errônea segundo populações

População Verdadeira	População Classificada	
	π_1	π_2
π_1	...	$1-\Phi\left(\frac{1}{2}\delta_{12}\right)$
π_2	$\Phi\left(-\frac{1}{2}\delta_{12}\right)$	...

Uma observação interessante e prática consiste na simplificação dos valores da tabela pela propriedade da simetria da distribuição normal, onde tem-se,

$$\Phi\left(-\frac{1}{2}\delta_{12}\right)=1-\Phi\left(\frac{1}{2}\delta_{12}\right).$$

Generalizando para g populações e utilizando-se os estimadores de máxima verossimilhança de μ, $\sum$ e δ_{hk}, probabilidades de classificação errônea ficam conforme Tabela 3:

Tabela 3: Probabilidades de classificação errônea para populações Normais e homocedasticas

População Verdadeira	População Classificada			
	π_1	π_2		π_g
π_1	...	$\Phi\left(-\frac{1}{2}\hat{\delta}_{12}\right)$	...	$\Phi\left(-\frac{1}{2}\hat{\delta}_{1g}\right)$
π_2	$\Phi\left(-\frac{1}{2}\hat{\delta}_{12}\right)$	...	...	$\Phi\left(-\frac{1}{2}\hat{\delta}_{2g}\right)$
$\vdots$	$\vdots$			$\vdots$
π_g	$\Phi\left(-\frac{1}{2}\hat{\delta}_{1g}\right)$	$\Phi\left(-\frac{1}{2}\hat{\delta}_{2g}\right)$	...	...

onde $\hat{\delta}_{h\,k}=\left(\underset{\sim h}{\bar{y}}-\underset{\sim k}{\bar{y}}\right)'S^{-1}\left(\underset{\sim h}{\bar{y}}-\underset{\sim k}{\bar{y}}\right)$; $h = 1,...,g$, $k = 1,..., g$ e $h \neq k$.

Portanto, a probabilidade estimada (distância de Mahalanobis) de classificação errônea para a população $\pi_h\left(h=1,\ldots,g\right)$ é dada por:

$$P\left(Classificação\ Errônea\ \pi_h\right) \leq \sum_{\substack{k=1 \\ k \neq h}}^{g} \Phi\left(-\frac{1}{2}\hat{\delta}_{hk}\right).$$

A probabilidade estimada geral de classificação errônea é dada por

$$P\left(Erro\ Classificação\right) = \sum_{h=1}^{g}\left(Classificação\ Errônea\ \pi_h\right) \text{ ou}$$

$$P\left(Erro\ Classificação\right) = \sum_{h=1}^{g}\sum_{\substack{k=1 \\ k \neq h}}^{g} \Phi\left(-\frac{1}{2}\hat{\delta}_{hk}\right).$$

4. PROGRAMA COMPUTACIONAL

O estudo foi complementado com a elaboração de um programa computacional em linguagem de alto nível (PHP – *Hipertext Preprocessor*), para obtenção das estimativas das probabilidades de má-classificação, que seja de fácil acesso e simples manuseio para pesquisadores das áreas biológicas. Finalizando o desenvolvimento do software, foi construído o manual do usuário e apresentado um exemplo de aplicação envolvendo a discriminação genética de variedades de girassol a partir de indicadores morfoagronômicos.

4.1- Linguagem de Programação PHP

O desenvolvimento do programa computacional para a metodologia estatística apresentada foi baseado na linguagem PHP - *Hipertext Preprocessor*. Trata-se de uma linguagem destinada a servidor *WEB*. As vantagens do uso do PHP são dadas por sua gratuidade, independência de plataforma, rapidez e segurança (ALVAREZ, 2006). A independência de plataforma possibilita a compatibilidade com qualquer sistema operacional.

O PHP permite programar pequenos *scripts* que podem ser inseridos dentro do código HTML – *Hypertext Markup*, código este largamente utilizado para a construção de páginas para Internet. O HTML pode ser interpretado por qualquer navegador (Browser).

Uma linguagem com essas características implica em seu processamento ser realizado pelo servidor de *WEB* e que o resultado será enviado, via Internet, para que o navegador interprete e mostre para o usuário o resultado. Este porém, só recebe o necessário para a visualização do resultado, ou seja, todo o código PHP fica guardado no servidor *WEB*.

A linguagem PHP foi criada em 1994 por Rasmus Lerdorf (2006), mas como o PHP está desenvolvido utilizando a política de código aberto, ao longo a história sofreu alterações de vários desenvolvedores. Atualmente, o PHP se encontra na versão 4, modernizada para suprir as necessidades das aplicações *WEB*.

4.2 Manual do Usuário

4.2.1 Entrada de Dados

A entrada de dados no Software PROBABILIDADE_ERRONEA é realizada por meio da leitura de dados construídos no aplicativo *Microsoft Excel*. A planilha deve trazer somente os valores numéricos relativos às observações por indivíduos, sendo que nas colunas estarão representadas as variáveis (características) e nas linhas as mensurações realizadas em cada indivíduo. Após a introdução dos valores, o arquivo deve ser gravado utilizando–se a opção “salvar como” do menu “Arquivo”. Quando a janela para salvar for apresentada coloque um nome sugestivo a seu processamento e na opção “Salvar como tipo” deve-se escolher, obrigatoriamente, na caixa de seleção, a opção “CSV (separado por virgulas)”, de acordo com a Figura 1. Após este procedimento, o arquivo está apto para ser lido pelo software.

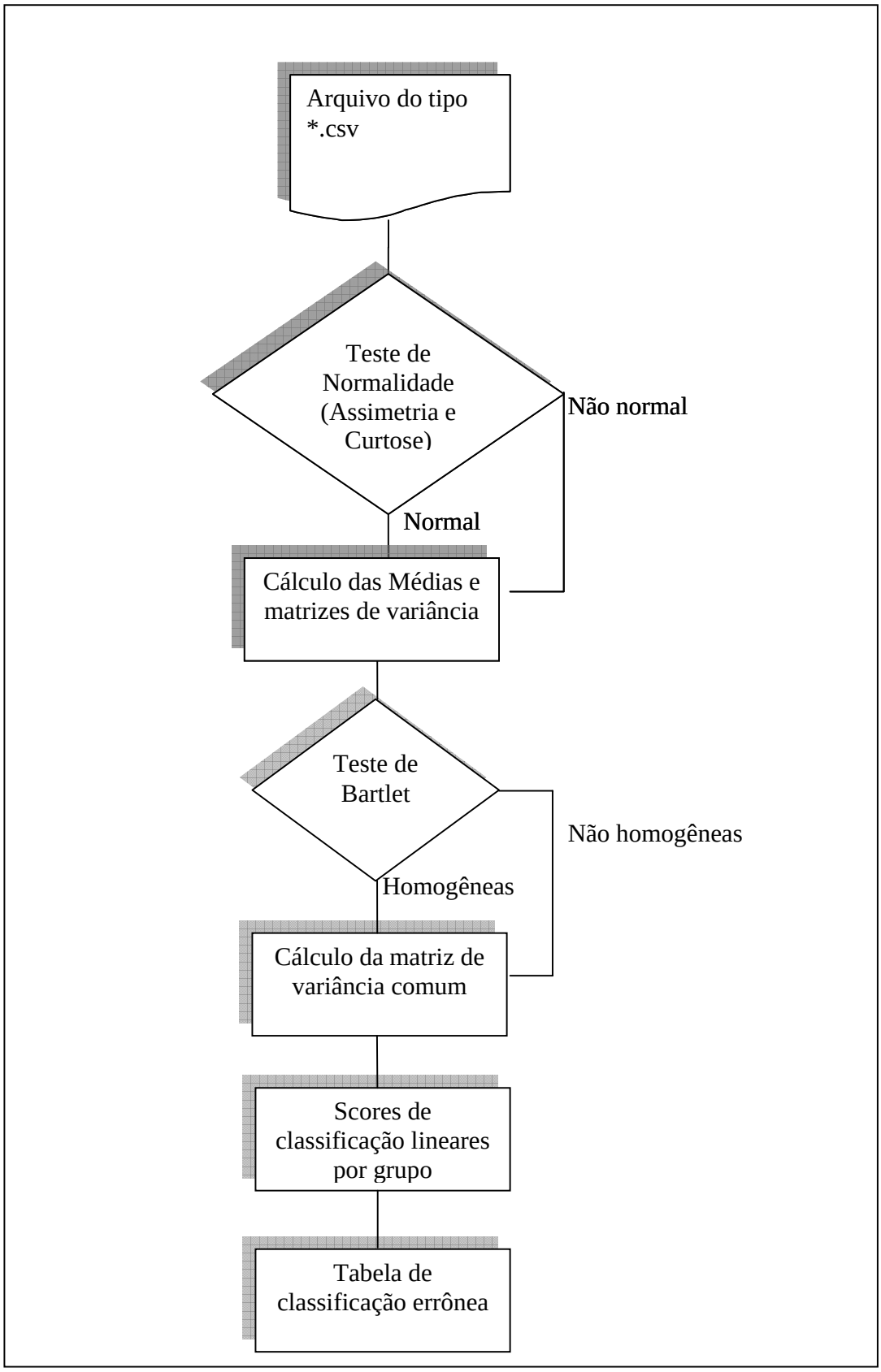


Figura 1 – Diagrama de Fluxo de Dados

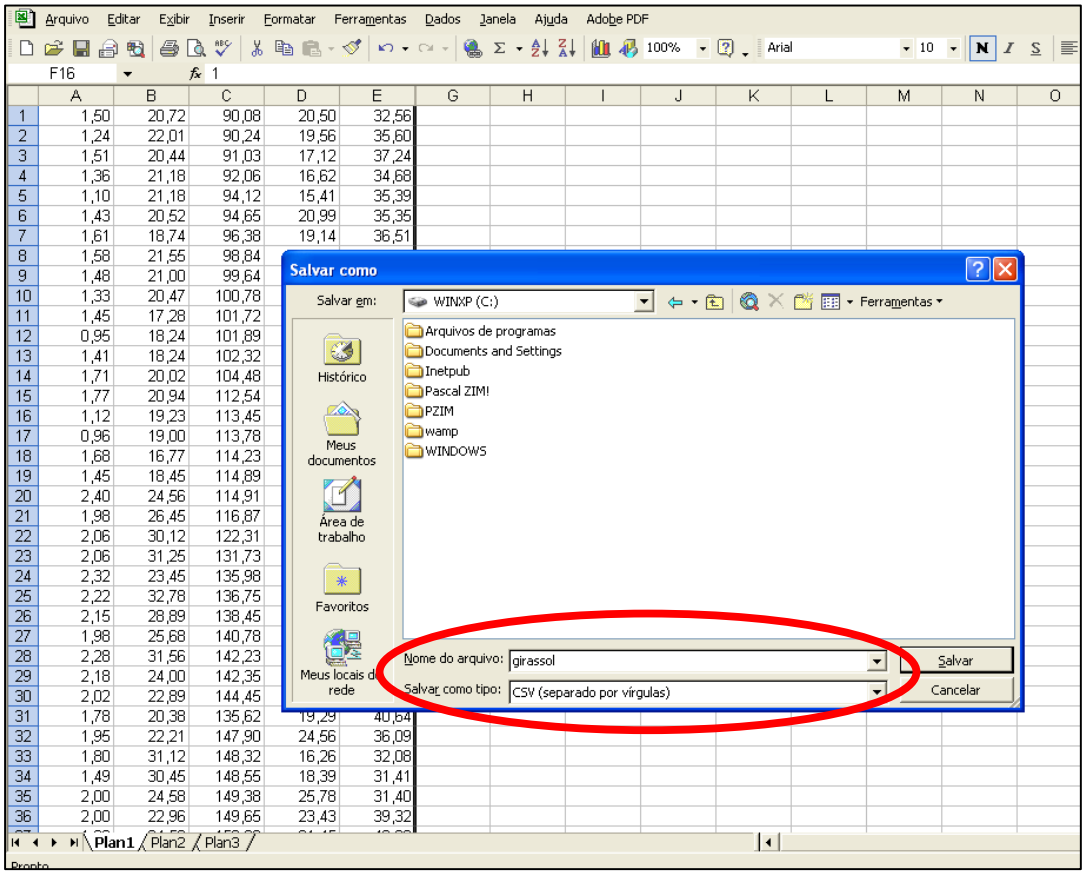


Figura 2 – Janela "Salvar Como" do Ms. Excel

4.2.2 Acesso ao Software

Como descrito anteriormente, uma característica marcante da linguagem é sua independência de plataforma e seu uso pela Internet. Dessa maneira o software será executado, diretamente da Internet; para isso, basta acessar o site: www.padovani.pro.br. Estando na página, "navegue" até o link Doutorado. Encontrando o link clique com o mouse sobre ele e a tela do software surgirá (Figuras 2 e 3).

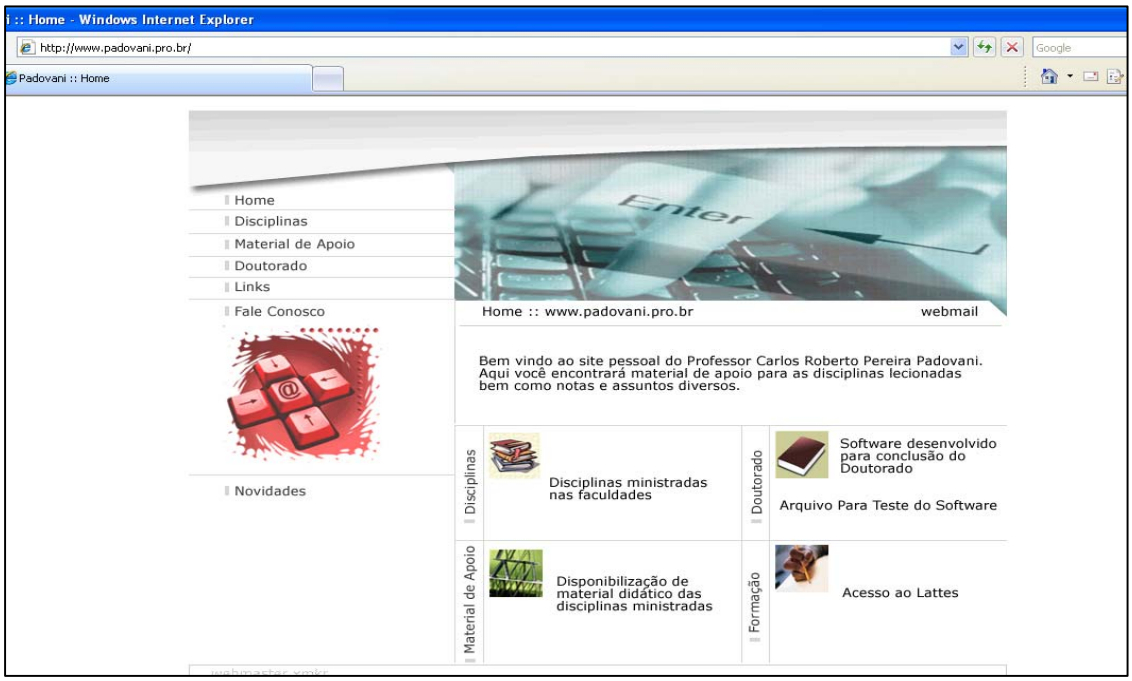


Figura 3- Página indicativa para o Software.

Figura 4 – Tela inicial do Software

4.2.3 Entrada de parâmetros

A Figura 3 apresenta a tela inicial do software para digitação dos dimensionamentos de variáveis (características), grupos, unidades experimentais e pesos dos grupos, bem como a seleção do arquivo de dados. Em seguida, esses dados serão submetidos ao servidor para o processamento. Observe que ao apontar o mouse sobre o ponto de interrogação, uma breve mensagem de orientação é mostrada.

As informações iniciais são introduzidas começando pelo nível de significância - α - para os testes de homogeneidade e normalidade, que é selecionado entre cinco possíveis opções em uma caixa de seleção. Na seqüência, informa-se o número de características (variáveis), o número de grupos, quantidade de elementos por grupo, o peso de cada grupo para a probabilidade “a priori” de pertinência e, por fim, o arquivo de dados.

A navegação entre os campos é feita utilizando-se a tecla TAB. As teclas ENTER e as setas de direcionamento, não produzem resultados. Os campos também são personalizados, ou seja, só são permitidas entradas de números.

Teoricamente, não há restrições quanto à quantidade de variáveis e/ou grupos a serem introduzidos para realização dos procedimentos, pois nas linguagens de programação de última geração, como é o caso do PHP, a alocação de memória é realizada de maneira flutuante. Talvez o grande limitante neste aspecto, seja o fator tempo de processamento, uma vez que o processamento é feito remotamente. Este fator pode ser alterado, pois depende exclusivamente dos serviços de hospedagem.

Para que o processamento possa ser efetuado, basta um “clique” com o mouse no botão “PROCESSAR” (Figura 5).

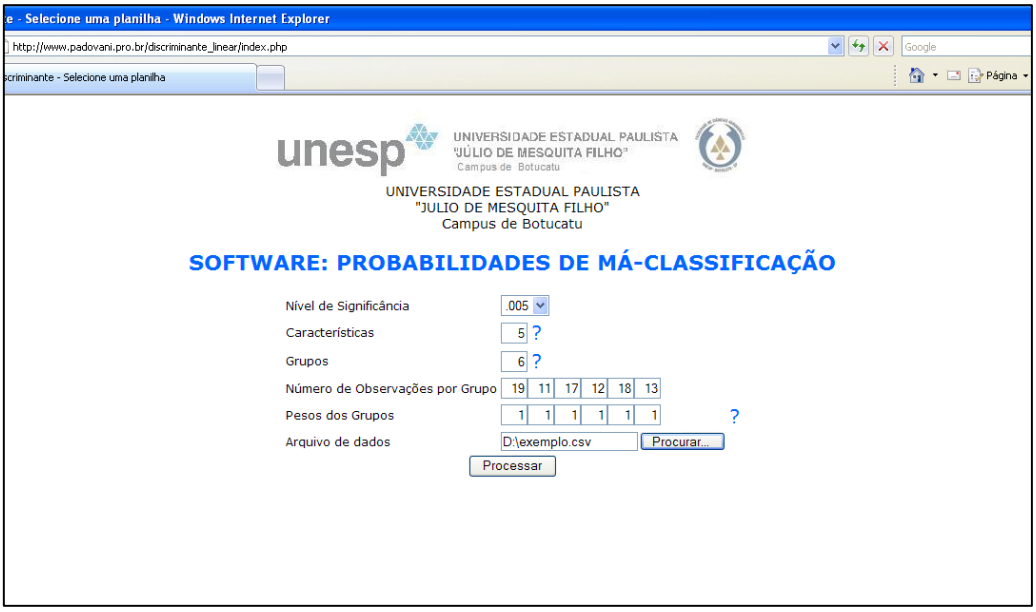


Figura 5 - Tela inicial do software com os campos preenchidos.

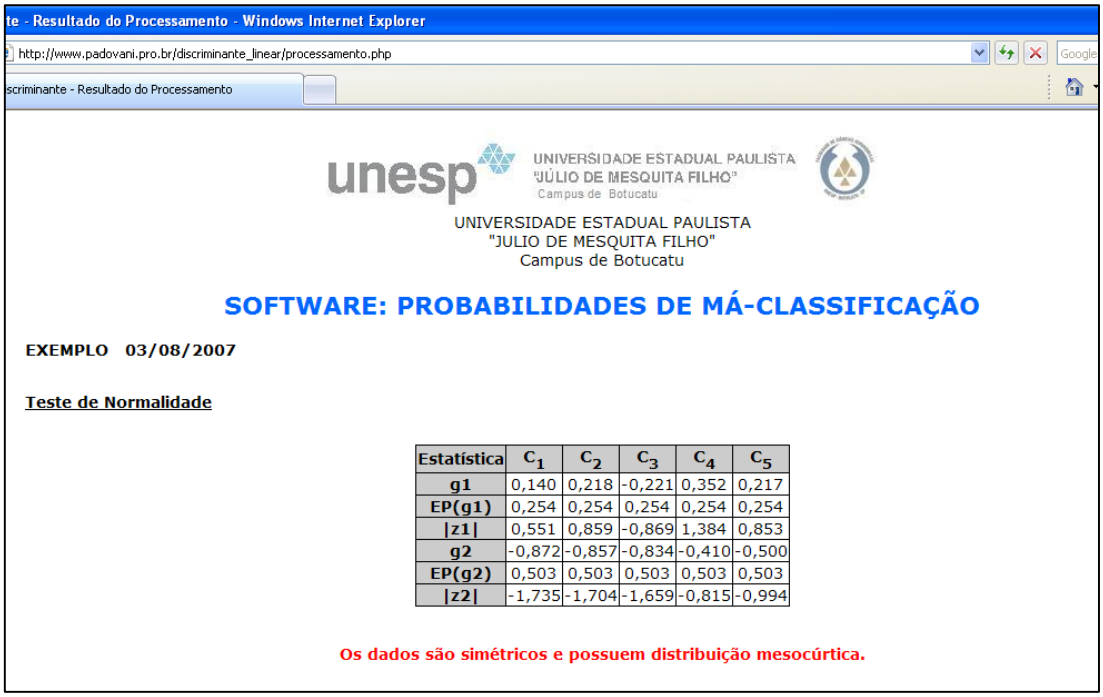
4.2.4 Saída dos resultados

A saída dos resultados em uma nova página contempla os seguintes itens:

- Mensagem do teste de normalidade(Assimetria e Curtose).
- Vetor de médias por grupo.
- Matriz de covariância S_i.
- Matriz de variação conjunta S.
- Mensagem do teste de homogeneidade.
- Matriz de covariância conjunta inversa.
- Matriz de classificação dos indivíduos pela função discriminante linear e respectivas freqüências percentuais de classificação correta.
- Escores de classificação lineares para os grupos.
- Sumário de classificações incorretas.
- Tabela de probabilidade de classificação errônea.

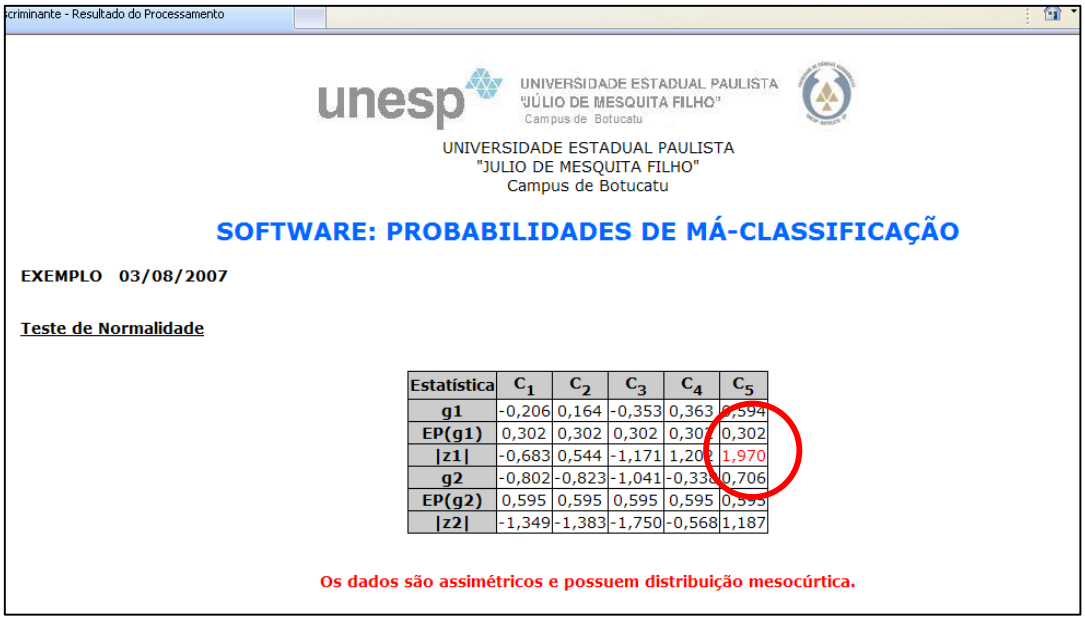
As Figuras a seguir mostram, de maneira fragmentada, os pontos de maior relevância do Software, destaque para a Figura 9 que mostra a tabela de probabilidade de classificação errônea.

Para a impressão dos resultados clique no comando imprimir do Menu Arquivo seu navegador.



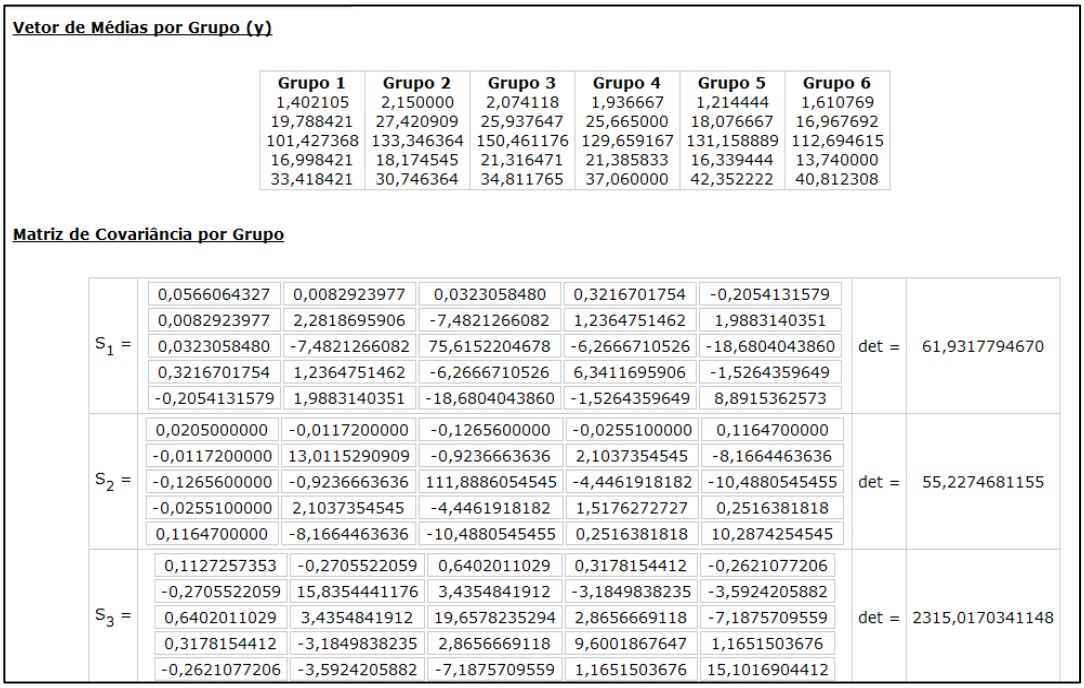
Estatística	C_1	C_2	C_3	C_4	C_5
g1	0,140	0,218	-0,221	0,352	0,217
EP(g1)	0,254	0,254	0,254	0,254	0,254
|z1|	0,551	0,859	-0,869	1,384	0,853
g2	-0,872	-0,857	-0,834	-0,410	-0,500
EP(g2)	0,503	0,503	0,503	0,503	0,503
|z2|	-1,735	-1,704	-1,659	-0,815	-0,994

Figura 6- Tela de Resultado do Software – destaque para o teste de normalidade.



Estatística	C_1	C_2	C_3	C_4	C_5
g1	-0,206	0,164	-0,353	0,363	0,594
EP(g1)	0,302	0,302	0,302	0,302	0,302
|z1|	-0,683	0,544	-1,171	1,202	1,970
g2	-0,802	-0,823	-1,041	-0,334	0,706
EP(g2)	0,595	0,595	0,595	0,595	0,595
|z2|	-1,349	-1,383	-1,750	-0,568	1,187

Figura 7- Tela de Resultado do Software – destaque para o teste de normalidade, demonstrando que os dados são assimétricos.



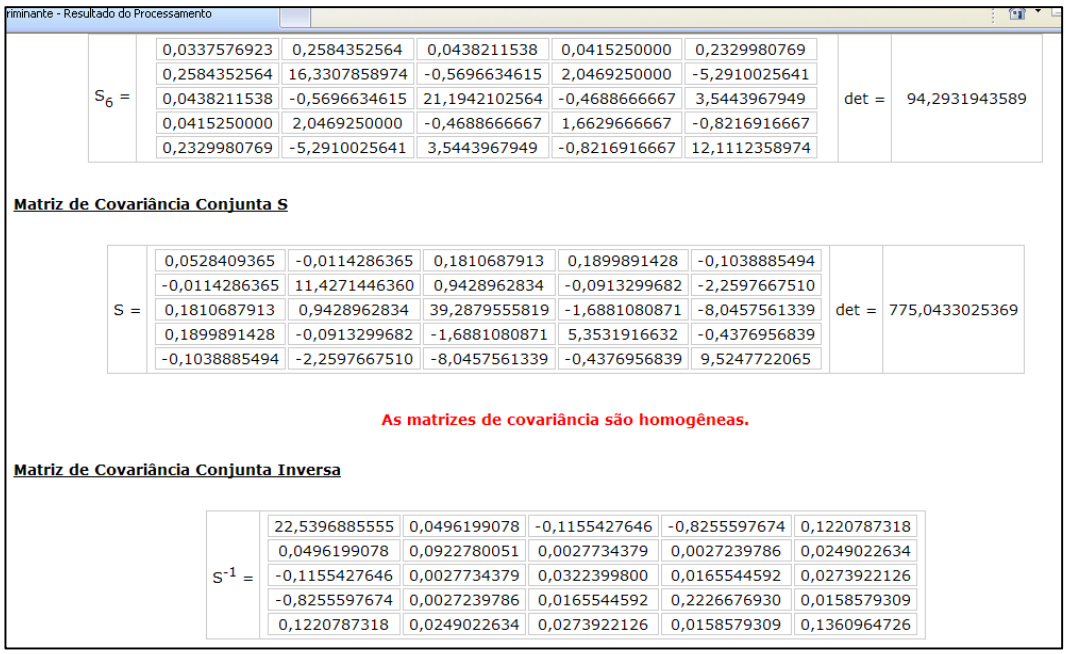
Vetor de Médias por Grupo (y)

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
1,402105	2,150000	2,074118	1,936667	1,214444	1,610769
19,788421	27,420909	25,937647	25,665000	18,076667	16,967692
101,427368	133,346364	150,461176	129,659167	131,158889	112,694615
16,998421	18,174545	21,316471	21,385833	16,339444	13,740000
33,418421	30,746364	34,811765	37,060000	42,352222	40,812308

Matriz de Covariância por Grupo

S_1 =	0,0566064327	0,0082923977	0,0323058480	0,3216701754	-0,2054131579	det =	61,9317794670
	0,0082923977	2,2818695906	-7,4821266082	1,2364751462	1,9883140351		
	0,0323058480	-7,4821266082	75,6152204678	-6,2666710526	-18,6804043860		
	0,3216701754	1,2364751462	-6,2666710526	6,3411695906	-1,5264359649		
	-0,2054131579	1,9883140351	-18,6804043860	-1,5264359649	8,8915362573		
S_2 =	0,0205000000	-0,0117200000	-0,1265600000	-0,0255100000	0,1164700000	det =	55,2274681155
	-0,0117200000	13,0115290909	-0,9236663636	2,1037354545	-8,1664463636		
	-0,1265600000	-0,9236663636	111,8886054545	-4,4461918182	-10,4880545455		
	-0,0255100000	2,1037354545	-4,4461918182	1,5176272727	0,2516381818		
	0,1164700000	-8,1664463636	-10,4880545455	0,2516381818	10,2874254545		
S_3 =	0,1127257353	-0,2705522059	0,6402011029	0,3178154412	-0,2621077206	det =	2315,0170341148
	-0,2705522059	15,8354441176	3,4354841912	-3,1849838235	-3,5924205882		
	0,6402011029	3,4354841912	19,6578235294	2,8656669118	-7,1875709559		
	0,3178154412	-3,1849838235	2,8656669118	9,6001867647	1,1651503676		
	-0,2621077206	-3,5924205882	-7,1875709559	1,1651503676	15,1016904412		

Figura 8- Tela de Resultado do Software – destaque para as matrizes de médias e covariâncias.



riminante - Resultado do Processamento

S_6 =	0,0337576923	0,2584352564	0,0438211538	0,0415250000	0,2329980769	det =	94,2931943589
	0,2584352564	16,3307858974	-0,5696634615	2,0469250000	-5,2910025641		
	0,0438211538	-0,5696634615	21,1942102564	-0,4688666667	3,5443967949		
	0,0415250000	2,0469250000	-0,4688666667	1,6629666667	-0,8216916667		
	0,2329980769	-5,2910025641	3,5443967949	-0,8216916667	12,1112358974		

Matriz de Covariância Conjunta S

S =	0,0528409365	-0,0114286365	0,1810687913	0,1899891428	-0,1038885494	det =	775,0433025369
	-0,0114286365	11,4271446360	0,9428962834	-0,0913299682	-2,2597667510		
	0,1810687913	0,9428962834	39,2879555819	-1,6881080871	-8,0457561339		
	0,1899891428	-0,0913299682	-1,6881080871	5,3531916632	-0,4376956839		
	-0,1038885494	-2,2597667510	-8,0457561339	-0,4376956839	9,5247722065		

As matrizes de covariância são homogêneas.

Matriz de Covariância Conjunta Inversa

S^{-1} =	22,5396885555	0,0496199078	-0,1155427646	-0,8255597674	0,1220787318
	0,0496199078	0,0922780051	0,0027734379	0,0027239786	0,0249022634
	-0,1155427646	0,0027734379	0,0322399800	0,0165544592	0,0273922126
	-0,8255597674	0,0027239786	0,0165544592	0,2226676930	0,0158579309
	0,1220787318	0,0249022634	0,0273922126	0,0158579309	0,1360964726

Figura 9- Tela de Resultado do Software – destaque para a mensagem em vermelho.

Matriz de classificação dos indivíduos pela função discriminante linear e respectivas freqüências percentuais

População Classificada Pela Função Linear	População Original					
	G1	G2	G3	G4	G5	G6
G1	19					
G2		10				
G3		1	16			
G4			1	12	2	
G5					16	
G6						13
Total Indivíduos	19	11	17	12	18	13
Indivíduos Corretos	19	10	16	12	16	13
% Correta	100,0	90,9	94,1	100,0	88,9	100,0

Scores de Classificação Lineares para os Grupos

	Coeficientes Lineares	Termo Independente
d_1		
	10.9121829608 y_1 3.05540851953 y_1 4.3596993309 y_1 4.89040298633 y_1 8.25995275287 y_1	-440.350898331
d_2		
	23.1630494358 y_2 3.82201856836 y_2 5.26979804369 y_2	-615.707624115

Figura 10- Tela de Resultado do Software - destaque pra a tabela de freqüências percentuais e os escores lineares.

Sumário das Classificações Incorretas (Método de Reclassificação)

Grupo	Indivíduo	Grupo Classificado
G2	11	G3
G3	1	G4
G5	13	G4
G5	15	G4

Tabela de Probabilidade de Classificação Errônea (Distribuição Gaussiana)

	Classificação					
Original	G_1	G_2	G_3	G_4	G_5	G_6
G_1	X	0,000598	0,000000	0,000291	0,000121	0,017003
G_2	0,000598	X	0,011911	0,072145	0,001589	0,004269
G_3	0,000000	0,011911	X	0,043633	0,004145	0,000046
G_4	0,000291	0,072145	0,043633	X	0,017003	0,004269
G_5	0,000121	0,001589	0,004145	0,017003	X	0,007976
G_6	0,017003	0,004269	0,000046	0,004269	0,007976	X

Figura 11- Tela de Resultado do Software - destaque para as tabelas de classificação incorreta e classificação errônea.

4.3. Exemplo da área agronômica

Consideraram-se para ilustrar o programa computacional sobre a discriminação de Fisher, dados adaptados de Messetti (2000), relativos a experimentos desenvolvidos por pesquisadores da EMBRAPA, na região de Londrina – PR, no ano agrícola de 1996/97, cujo objetivo destinava-se a avaliar o grau de adaptabilidade de variedades de girassol (*Helianthus annuus*) de diferentes origens. No contexto do presente trabalho, foram escolhidas seis variedades (G1: Corona; G2: Pehuen; G3 Comangir; G4: Talenay; G5: Maribondo; G6:3Granains), todas tendo como país de origem a Argentina e Chile, avaliadas sob os seguintes caracteres quantitativos: **altura da planta** (mensurada em metros, do nível do solo até a inserção do capítulo), **diâmetro do caule** (em mm, medição feita a 5 cm do nível do solo); **altura do capítulo** (em cm, medida realizada do nível do solo até a inserção do capítulo); **diâmetro do capítulo** (em cm, realizado no ponto de maturação fisiológica) e **teor percentual de óleo** (análise feita em laboratório).

Os vetores respostas observados nas unidades experimentais encontram-se no APÊNDICE 1.

5. RESULTADOS E DISCUSSÃO

As medidas descritivas envolvendo tendência central (média) e variabilidade (variância e covariância) dos caracteres agronômicos nas diferentes variedades de girassol (grupo) estão apresentadas na Tabela 4 e nas matrizes de dispersão.

Tabela 4: Vetor de médias segundo variedade

Variedade	Característica Quantitativa (Variável)				
	Alt. Planta	Diâm. caule	Alt. capítulo	Diâm. capítulo	Teor de Óleo
Corona	1,402105	19,788421	101,427368	16,998421	33,418421
Pehuen	2,150000	27,420909	133,346364	18,174545	30,746364
Comangir	2,074118	25,937647	150,461176	21,316471	34,811765
Talenay	1,936667	25,665000	129,659167	21,385833	37,060000
Maribondo	1,21444	18,076667	131,158889	16,339444	42,352222
3Grnains	1,610769	16,967692	112,694615	13,740000	40,812308

Matriz de Dispersão

- **Grupo 1 – Corona**

$$S_1 = \begin{bmatrix} 0,0566 & 0,0082 & 0,0323 & 0,3216 & -0,2054 \\ & 2,2818 & -7,4821 & 1,2364 & 1,9883 \\ & & 75,6152 & -6,2666 & 18,6804 \\ & & & 6,3411 & -1,5264 \\ & & & & 8,8915 \end{bmatrix}$$

- **Grupo 2 – Pehuen**

$$S_2 = \begin{bmatrix} 0{,}2050 & -0{,}0117 & -0{,}1265 & -0{,}0255 & 0{,}1164 \\ & 13{,}0115 & -9{,}2366 & 2{,}1037 & -8{,}1664 \\ & & 111{,}8886 & -4{,}4461 & -10{,}4880 \\ & & & 1{,}5176 & 0{,}2516 \\ & & & & 10{,}2874 \end{bmatrix}$$

- **Grupo 3 – Comangir**

$$S_3 = \begin{bmatrix} 0{,}1127 & -0{,}2705 & 0{,}6402 & 0{,}3178 & -0{,}2621 \\ & 15{,}834 & 3{,}4354 & -3{,}1849 & -3{,}5924 \\ & & 19{,}6578 & 2{,}8656 & -7{,}1875 \\ & & & 9{,}6001 & 1{,}1651 \\ & & & & 15{,}1016 \end{bmatrix}$$

- **Grupo 4 – Talenay**

$$S_4 = \begin{bmatrix} 0{,}0506 & -0{,}5221 & -0{,}0353 & 0{,}2037 & 0{,}0240 \\ & 15{,}5663 & 1{,}8507 & -6{,}3792 & 0{,}4295 \\ & & 6{,}0584 & -0{,}8973 & -4{,}3907 \\ & & & 6{,}2947 & -1{,}8091 \\ & & & & 9{,}0651 \end{bmatrix}$$

- **Grupo 5 – Maribondo**

$$S_5 = \begin{bmatrix} 0{,}0502 & 0{,}5341 & 0{,}3552 & 0{,}1822 & -0{,}1331 \\ & 21{,}4172 & 8{,}6936 & 4{,}1273 & -5{,}3642 \\ & & 25{,}8270 & -1{,}2069 & -4{,}2009 \\ & & & 5{,}7356 & -0{,}6205 \\ & & & & 11{,}5185 \end{bmatrix}$$

- **Grupo 6 – 3Grnains**

$$S_6 = \begin{bmatrix} 0{,}0337 & 0{,}2584 & 0{,}0438 & 0{,}0415 & 0{,}2329 \\ & 16{,}3307 & -0{,}5696 & 2{,}0469 & -5{,}2910 \\ & & -0{,}5696 & 21{,}1942 & -0{,}4688 \\ & & & 1{,}6629 & -0{,}8216 \\ & & & & 12{,}1112 \end{bmatrix}$$

- **Matriz de Variação Conjunta S**

$$S = \begin{bmatrix} 0{,}0528 & -0{,}0114 & 0{,}1810 & 0{,}1899 & -0{,}1038 \\ & 11{,}4271 & 0{,}9428 & -0{,}0913 & -2{,}2597 \\ & & 39{,}2879 & -1{,}6881 & -8{,}0457 \\ & & & 5{,}3531 & -0{,}4376 \\ & & & & 9{,}5247 \end{bmatrix}$$

- **Matriz de Variação Conjunta Inversa** S^{-1}

$$S^{-1} = \begin{bmatrix} 22{,}5396 & 0{,}0496 & -0{,}1155 & -0{,}8255 & 0{,}1220 \\ & 0{,}0922 & 0{,}0027 & 0{,}0027 & 0{,}0249 \\ & & 0{,}0322 & 0{,}0165 & 0{,}0273 \\ & & & 0{,}2226 & 0{,}0158 \\ & & & & 0{,}1360 \end{bmatrix}$$

A seguir, são apresentadas as funções discriminantes lineares de cada variedade de girassol (Tabela 5), cujos coeficientes são as cargas que cada variável acumula na composição do descritor genético. O termo independente pode ser considerado como a correção comum que se faz no total de cada indivíduo de um mesmo grupo.

Tabela 5: Coeficientes das funções discriminantes lineares segundo grupo

Variável	Grupos					
	Corona	**Pehuen**	**Comangir**	**Talenay**	**Maribondo**	**3Grnains**
T. Independente	-440,9121	-615,707	-761,3810	-662,361	-660,2176	-537,2343
Alt. Planta	10,0554	23,1630	17,3040	16,8131	4,7981	17,7662
Diâm. caule	3,0554	3,8220	3,8386	3,8051	3,1912	3,0119
Alt. apítulo	4,3596	5,2697	5,9896	5,3968	5,5689	4,8396
Diâm. capítulo	4,8904	5,0416	6,1476	5,9671	5,5277	4,2886
Teor de Óleo	8,2599	9,0706	10,0963	9,8100	10,2142	9,4784

A partir das funções discriminantes pode-se construir a Tabela 6 que expressa, por meio da distribuição normal de probabilidades, as respectivas probabilidades de má-classificação das variedades.

Tabela 6: Probabilidade de Classificação Errônea

	Classificação					
Original	**Corona**	**Pehuen**	**Comangir**	**Talenay**	**Maribondo**	**3Grnains**
Corona	**XXX**	0,00060	0,00000	0,00030	0,00012	0,01743
Pehuen	0,00060	**XXX**	0,01222	0,07215	0,00164	0,00427
Comangir	0,00000	0,01222	**XXX**	0,04363	0,00415	0,00048
Talenay	0,00030	0,07215	0,04363	**XXX**	0,01743	0,00440
Maribondo	0,00012	0,00164	0,00415	0,01743	**XXX**	0,008198
3Grnains	0,01743	0,00427	0,00048	0,00440	0,008198	**XXX**

Observa-se, pelos valores encontrados, que a maior probabilidade de erro de classificação ocorre entre as variedades Pehuen e Talenay, enquanto que, entre as variedades Corona e Comangir é muito pouco provável cometer erro na classificação de um novo indivíduo.

As probabilidades acumuladas de erros de má-classificação em cada variedade resultam, percentualmente, em: Corona (1,845%); Pehuen (9,088%); Comangir (6,048%); Talenay (13,701%); Maribondo (3,154%) e 3Grnains (3,478%).

6. CONCLUSÃO

Quanto ao algoritmo computacional, o sistema foi desenvolvido para ser executado utilizando a Internet, pois o mesmo fica hospedado em um servidor WEB. Teoricamente, não há restrições quanto à quantidade de variáveis e/ou grupos a serem introduzidos para realização dos procedimentos classificatórios, pois nas linguagens de programação de última geração, como é o caso do PHP, a alocação de memória é realizada de maneira flutuante. O *software* resultante comunica-se de forma interativa com o usuário, é auto-explicativo, de fácil utilização por qualquer pesquisador da área aplicada e de acesso livre.

O *software* oferece aos usuários os vetores de médias das populações com as respectivas matrizes de covariância, os testes de assimetria, curtose e homocedasticia, a matriz de classificação dos indivíduos, as funções discriminantes lineares e todas as probabilidades de má-classificação entre pares de populações, consideradas sob a distribuição normal de probabilidades.

A matriz das probabilidades de má-classificação permite avaliar a qualidade das funções discriminantes construídas, quer se considere em termos de erros de classificação como em capacidade de discriminação.

Em relação à discriminação genética das variedades de girassol, verificou-se que a espécie Corona foi a que apresentou menor probabilidade acumulada de má-

classificação, constituindo-se na espécie mais diferenciada nos aspectos morfoagronômicos considerados. Em oposição, a espécie Talenay foi a que apresentou maior probabilidade acumulada e com maior possibilidade de má-classificação em relação à variedade Pehuen.

REFERÊNCIAS BIBLIOGRÁFICAS

ABREU, F.B; LEAL,N.R; ROGRIGUES,R; JÚNIOR, A.T.A; SILVA, D.J.H. Divergência genética entre os acessos de feijão-de-vagem de hábito de crescimento indeterminado. Horticultura Brasileira, 22:(3) 547-552. Setembro, 2004.

ALVAREZ, M.A. **Introdução a Programação PHP.** Disponível em: Htp://www.criarweb.com/artigos/70.php – acessado em 29/05/07.

ALVES, M.A; GARCIA. A.A.F; CRUZ, E.D; FIGUEIRA.A. Seleção de descritores botânicos para caracterização de germoplasma de cupuaçuzeiro. Pesquisa Agropecuária Brasileira(38):7 807-818, 2003.

ANDRADE, J. B; JUNIOR, E.F; POSSENTI, R,A; OTSUK, I,P; ZIMBACK, L; LANDELL, M.G.A. Seleção de 39 variedades de cana-de-açúcar para alimentação animal. Brazilian Journal of Veterianry Research and Animal Science 40: 287-296, 2003.

ARSEVEN, E.; KSHIRSAGAR,A. M. A note on the equivalency of two discrimination procedures. The American Statistication, v.29, n.1,p.38-39, 1975.

ASSIS, G.M.L.; EUCLYDES, R.F.; CRUZ, C.D.; VALLE, C.B. Discriminação de espécie de *Brachiaria* baseada em diferentes grupos de caracteres morfológicos. Revista Brasileira de Zootecnia,32(3):576-584, 2003.

BARBOSA, L; LOPES, P.S; REGAZZI, A.J; CUIMARÃES, S.E.F; TORRES,R.A. Estudo da associação entre características de desempenho e de carcaça de suínos por meio de correlação canônica. Revista Brasileira de Zootecnia,34(6):2218-2224, 2005a.

BARBOSA, L; LOPES, P.S; REGAZZI, A.J; CUIMARÃES, S.E.F; TORRES,R.A. Avaliação de características de carcaça de suínos utilizando-se da análise dos componentes principais.Revista Brasileira de Zootecnia,34(6):2209-2217, 2005b.

BARROSO, L,P; ARTES, R. (2003) Análise Multivariada. Minicurso do 10º SAEGRO e 48ª RBRAS, UFLA, Lavras, MG, 151p.

BASNET, K. Controls of environmental factors on pattern of montane rain forests in Puerto Rico. Tropical Ecology, 34(1):51-63, 1993

BELLAVER,C; GUIDONI, A.L; BRUM, P.A.R; ROSA,P.S. Estimativas exigências de lisina e de energia metabolizável em frangos de corte de 1 a 21 dias de idade, utilizando-se uma variável multivariada canônica.Revista Brasileira e Zootecnia(31):1 71-78, 2002.

BERNARD, M. M. The secular varaitions of skull characteres in four series of egiptian Skulls. Annals of Eugenics, v.6,1935.

CAIXETA, R.P; TRUGILHO, P.F; LIMA, J.T; ROSADO, S.C.S. Classificação de *eucalyptus* relacionados com a qualidade da madeira após a secagem natural.Scientia Forestalis(61) 49-58, 2002.

CARDIM, D.C; CARLINI-GARCIA, L.A; MONDIN, M; MATINS, M; ANN VEASEY, E; ANDO, A. Variabilidade intra-específica em cinco populações de *Oncidium varicosum Lindl.* (Orchidaceae – Oncidiinae) em Minas Gerais. Revista Brasileira de Botânica(24):4 553-560, 2001.

CROCCI, A.J. Problemas de classificação em uma da varias populações. Estudo da probabilidade de má classificação, baseado na função discriminante linear. Dissertação (Mestrado em Medicina -Área de Concentração: Bioestatística). Faculdade de Medicina de Ribeirão Preto USP, Ribeirão Preto, SP, 64p.,1979.

DAL'COL LÚCIO, A; FORTES, F.O; STORK, L; FILHO, A.C. Abordagem multivariada em analise de sementes de espécies florestais exóticas. Revista Cerne (12):1 27-37, 2006.

DILLON, W.R.; GOLDSTEIN, M. Multivariate Analysis: Methods and Applications, 2d ed., John Wiley, New York, 462p,1984.

FISHER, R.A. The use multiple measurements in toxonomic problems, Annals of Eugenics,7: 179 - 188,1936.

FONSECA, R.C.B; FONSECA, I.C.B. Utilização de métodos multivariados na caracterização do mosaico sucessional em florestas semidecídual. Revista Árvore,28:(3) 351-359. Julho, 2004.

FONSECA, R; PIRES, A.V; LOPES, P.S; TORRES, R.A; EUCLYDES, R.F. Estudo da divergência genética entre raças suínas utilizando técnicas de análise multivariada. Arquivo Brasileiro de Medicina Veterinária e Zootecnia , 52:(4) 403-409.Agosto 2000.

FONSECA. R; TORRES FILHO , R.A; PEIXOTO, J.O; PIRES, A.V; CARNEIRO, P.L.S; SOUZA, G.H; BUENO, R.S; LOPES, P.S; EUCLYDES, R.F. Avaliação de frangos de corte utilizando técnicas de análise multivariada: I – características de carcaça. Arquivo Brasileiro de Medicina Veterinária e Zootecnia(54):5 300-3007, 2002.

GATTI, I; LÓPES ANIDO, F; PICARDI, L; COINTRY,E. Seleccion de progenitores em espárrago. Horticultura Brasileira (21):2 162-165, 2003.

GODOI, D.R.M. Análise estatística multimensional. 1º SEGRO e 30º RBRAS, Piracicaba – SP, 187p., 1985.

GRAYBILL, F.A. Introduction to matrices witch apiplications in statistics. Wadsworth Flublising Company, Belmont, 372p, 1983.

JOHNSON, R.A.; WICHERN, D.W. Applied multivariate statistical analysis. 4rd ed Prentice-Hall International, New Jersey, 642p., 1998.

KENDALL, M.G. Factor analysis. Journal the Royal Statistical Society.Serie B, 12: 60-94,1950.

KHOURY JR. J,K; PINTO, F.A.C; SANTOS, N.T; LUCIA, R.M.D; MAEDA, E,E. Análise discriminante paramétrica para reconhecimento de defeitos em tábuas de eucalipto utilizando imagens digitais. Revista Árvore(29):2 99-309, 2005.

LACHEMBRUCH, P.A. An almost unbiased method of obtaining confídence intervals for the probability of misclassification in discriminant analysis. Biometrics,23:639-645,1967.

LACHENBRUCH, P.A.; MICKEY, M.R. Estimation of rates in discriminant analysis. Technometrics, 10(1):1,11.

LUCIO, P.S; TOSCAO, E.M.M.; ABREU, M.L. Caracterização de séries climatológicas pontuais via análise de correspondência . Estudo de caso: Belo Horizonte – MG (Brasil). Revista Brasileira de Geofísica,17:(2-3)192-207 Julho de 1999.

MARDIA,K.V. Measures of multivariate skewness and kurtosis with applications. Biometrika, 57(3):519-530,1970.

MARTEL, J.H.I.; FERRAUDO, A.S.; MÔRO, J.R.; PERECIN,D. Estatística multivariada na discriminação re raças amazônicas de pupunheiras (Bactris gasipaes Kunth) em Manaus (Brasil). Revista Brasileira de Fruticultura, 25:(1)115-118, 2003.

MARTINELLO, G.E; LEAL, N.R; AMARAL JÚNIOR, A.T; PREIRA.M.G; DAHER, R.F. Divergência genética em acessos de quiabeiro com base em marcadores morfológicos. Horticultura Brasileira, 20:(1)52-58. Março, 2002.

MESSETTI, A.V.L. Estudo da semelhança de genótipos de girassol *(Helianthus annuus L.)* com uso da distância generalizada de Mahalanobis na análise de agrupamento. Dissertação (Mestrado

Agronomia - Área de Concentração: Energia na Agricultura). Universidade Estadual Paulista, Botucatu, SP, 86p., 2000.

MINGOTI, S.A. Análise de dados através de métodos de estatística multivariada. Uma abordagem aplicada. Ed. UFMG,Belo Horizonte, MG, 295p.,2005.

MIRANDA, G.V; COIMBRA, R.R; GODOY, C.L; SOUZA,L.V; GUIMARÃES, L.J.M; MELO, A.V. Potencial de melhoramento e divergência genética de cultivares de milho-pipoca. Pesquisa Agropecuária Brasileira(38):6 681-688, 2003.

MORETTIN,P.A.; BUSSAB,W.O.(2003) Estatistica Básica, 5ªed. Editora Saraiva, São Paulo,526p.

MORGANO, M.A; QUEIROZ, S.C.N; FERREIRA, M.M.C. Determinação dos teores de minerais em sucos de frutas por espectrometria de emissão óptica em plasma indutivamente acoplado (ICP-OES). Ciência e Tecnologia de Alimentos(19):3 130-142, 1999.

MURTEIRA, B.J. Análise exploratória de dados. Lisboa. McGraw-Hill, 1993, 350p.

NANNI, M.R; DEMATTÊ, J.A.M; FIORIO, P.R. Análise discriminante dos solos por meio da resposta espectral no nível terrestre. Pesquisa Agropecuária Brasileira,39:(10) 995-1006. Outubro, 2004.

NOBLE,B.; DANIEL,J.W.(1988) Aplied Linear Álgebra,3rd ed., Prentice Hall, New Jersey,386p.

OHTOSHI, C. Uma comparação de regressão logística, árvore de classificação e redes neurais: Analisando dados de crédito. Dissertação de Mestrado. IME – USP, 89p.

OKAMOTO, M. An asymptotic expansion for the distribution of the linear discriminant function. Annals of Mathematial Statistics, 34:1286-1301, 1963.

PADOVANI, C.R.P; ARAGON, F.F. Processo gráfico para o método de discriminação de Fisher, baseado nos eixos canônicos. Energia na Agricultura(20):1 1-10, 2005.

PEREIRA, F.H.F; PUIATTI, M; MIRANDA, G.V; SILVA, D.J.H; FINGER, F.L. Divergência genética entre acessos de taro. Horticultura Brasileira, 22:(1) 55-60. Março, 2004.

PERONI, N; MARTINS, P.S; ANDO, A. Diversidade inter-intra-específica e uso de análise multivariada para morfologia da mandioca (*Manihot esculenta Crantz*): um estudo de caso.Scientia Agrícola(56):3 200-214, 1999.

PIASSI, M.; SILVA, M.A.; REGAZZI, A.Z.; TORRES, R.A.; SOARES,P.R.; JÚNIOR, R.A.A.T. Estudo da divergência entre oito grupos de aves de postura, por meio de técnicas de análisse multivariada. Revista da Sociedade Brasileira de Zootecnia. 24(5);715-727,1995a.

PIRES, A. V.; et al. Genetic divergence study among Duroc, Landrace and Large White swine breeds using techniques of multivariate analisys. Arch. Latinoam. Prod. Anim., v.10, n.2, p.81-85, 2002.
RAO,C.R. Sir Ronald Aimer Fisher – The architect of multivariate analysis.Biometrics, v.20,p.286-300,1964.

REIS, E. Estatística Multivariada Aplicada. Edições Sílabo Ltda. Lisboa, 343p,1997.

RENCHER,A.c. Methods of Multivariate Analysis, Whiley, 1995

RIBEIRO, F.E; SOARES, A.R; RAMALHO, M.A.P. Divergência genética entre populações de cocqueiro-gigante-do-brasil. Pesquisa Agropecuária Brasileira(34):9 1615-1622, 1999.

RODRIGUES, L.R.F; ANDO.A; Uso da sensitividadde à radiação gama na discriminação de variedade de arroz-de-sequeiro dos grupos Índica e Japônica. Bragantia, 62:(2) 179-188. 2003.

SANCHES, J; LEAL, P.A.M; SARAVALI, J.H; ANTONIALI, S.Análise de componentes principais para avaliação da qualidade de banana "nanicão" refrigerada e em diferentes embalagens. Revista Brasileira de Fruticultura (25):2 220-223, 2003.

SANT'ANNA, C.M; MALINOVSKI, J.R. Uso da análise multivariada de fatores humanos em operadores de motosserra. Revista Cerne (8):2 98-104, 2002.

SANTOS, J.H.S; FERREIRA, R.L.C; SILVA J.A.A; SOUZA, A.L; SANTOS, E.S; MEUNIER, I.M.J. Distinção de grupos ecológicos de espécies florestais por meio de técnicas multivariadas. Revista Árvore(28):3 387-396, 2004.

SILVA, F.V; KAMOGAWA, M.Y; FERREIRA, M.M.C; NÓBREGA, J.A; NOGUEIRA, A.R.A. Discriminação geográfica de águas minerais do Estado de São Paulo através da análise exploratória. Eclética Química, 27:(esp)91-102. 2002.

SOARES, P.L.M; SANTOS, J.M; FERRAUDO, A.S. Estudo morfométrico comparativo de 58 populações brasileiras de *Rotylenchulus reniformis* (Nemata: *Rotylenchulinea*). Fitopatologia Brasileira,29:(4) 419-424. Agosto, 2004.

SOUZA, A.L; SOUZA, D.R. Análise multivariada para estratificação volumétrica de uma floresta ombrófila densa de terra firme, Amazônia Oriental. Revista Árvore(30):1 49-54, 2006.

SOUZA, D.R.; SOUZA, A.L.; GAMA, J.R.V.; LEITE, H.G. Emprego de Análise multivariada para estratificação vertical de florestas ineqüiâneas. Revista Arvore,27(1):59-63,2003.

SOUZA, F.F;QUEIROZ, M.A. Avaliação de caracteres morfológicos úteis na identificação de plantas poliplóides de melancia. Horticultura Brasileira(22):3 516-520, 2004.

SOUZA, F.F;QUEIROZ, M.A; DIAS,R.S.C. Divergência genética em linhagens de melancia. Horticultura Brasileira(23):2 179-183, 2005.

TABACHNICK, B,G; FIDELL, L.S; (2001) Using Multivariate Statistics, 4th ed. Allyn & Bacon, Boston, 966p.

TIMM, N.H. Applied multivariate analysis. Springer Verlag, New York, 365p.,2002.

VIANNA, G.K; RODRIGUES, R.J; THOMÉ, A. C.G. Extração de características para o reconhecimento de dígitos cursivos - uma nova abordagem. Núcleo de Computação Eletrônica da UFRJ – LabIC, Rio de Janeiro – RJ Disponível em: <http://www.labic.nce.ufrj.br/downloads/sbrn_2000.pdf >.Acessado em 25 jul. 2006.

WELCH, B. L. Note on discrimination function. Biometrika, v.31, n.1/2, p.218-220,1939.

WOODWARD,W.A.; ELLIOT,A.C. A survey of statistical package on microcomputers. Computational Statistics & Data analisys, v.1, p.191-200, 1983.

ZHANG, M. Program MZEF. http://argon.cshl.org/genefinder. Acessado em 13/10/06.

APÊNDICE

QUADRO1. Vetores de resposta das características quantitativas do girassol.

Variedade	Parcela	Alt. Planta	Diâm. Caule	Alt. Capítulo	Diâm. Capítulo	Teor de Óleo
CORONA	1	1,50	20,72	90,08	20,50	32,56
CORONA	2	1,24	22,01	90,24	19,56	35,60
CORONA	3	1,51	20,44	91,03	17,12	37,24
CORONA	4	1,36	21,18	92,06	16,62	34,68
CORONA	5	1,10	21,18	94,12	15,41	35,39
CORONA	6	1,43	20,52	94,65	20,99	35,35
CORONA	7	1,61	18,74	96,38	19,14	36,51
CORONA	8	1,58	21,55	98,84	20,45	32,37
CORONA	9	1,48	21,00	99,64	12,71	36,83
CORONA	10	1,33	20,47	100,78	14,20	36,78
CORONA	11	1,45	17,28	101,72	15,62	30,12
CORONA	12	0,95	18,24	101,89	13,54	35,99
CORONA	13	1,41	18,24	102,32	16,00	32,81
CORONA	14	1,71	20,02	104,48	17,45	33,47
CORONA	15	1,77	20,94	112,54	20,00	28,16
CORONA	**16**	1,12	19,23	113,45	14,45	32,69
CORONA	**17**	0,96	19,00	113,78	15,18	31,14
CORONA	**18**	1,68	16,77	114,23	17,25	28,78
CORONA	**19**	1,45	18,45	114,89	16,78	28,48
CORONA	**V. Mínimo**	**0,95**	**16,77**	**90,08**	**12,71**	**28,16**
CORONA	**V. Máximo**	**1,77**	**22,01**	**114,89**	**20,99**	**37,24**
CORONA	**Média**	**1,40**	**19,79**	**101,43**	**17,00**	**33,42**
CORONA	**D. Padrão**	**0,24**	**1,51**	**8,70**	**2,52**	**2,98**
PEHUEN	1	2,40	24,56	114,91	18,37	36,46
PEHUEN	2	1,98	26,45	116,87	19,26	30,86
PEHUEN	3	2,06	30,12	122,31	19,47	32,31
PEHUEN	4	2,06	31,25	131,73	18,25	26,12
PEHUEN	5	2,32	23,45	135,98	16,23	31,86
PEHUEN	6	2,22	32,78	136,75	19,32	26,05
PEHUEN	7	2,15	28,89	138,45	16,01	26,84
PEHUEN	8	1,98	25,68	140,78	17,58	31,78
PEHUEN	9	2,28	31,56	142,23	19,54	30,78
PEHUEN	10	2,18	24,00	142,35	17,65	32,41
PEHUEN	11	2,02	22,89	144,45	18,24	32,74
PEHUEN	**V. Mínimo**	**1,98**	**22,89**	**114,91**	**16,01**	**26,05**
PEHUEN	**V. Máximo**	**2,40**	**32,78**	**144,45**	**19,54**	**36,46**
PEHUEN	**Média**	**2,15**	**27,42**	**133,35**	**18,17**	**30,75**
PEHUEN	**D. Padrão**	**0,14**	**3,61**	**10,58**	**1,23**	**3,21**

QUADRO2. Vetores de resposta das características quantitativas do girassol.

Variedade	Parcela	Alt. Planta	Diâm. Caule	Alt. Capítulo	Diâm. Capítulo	Teor de Óleo
COMANGIR	1	1,78	20,38	135,62	19,29	40,64
COMANGIR	2	1,95	22,21	147,90	24,56	36,09
COMANGIR	3	1,80	31,12	148,32	16,26	32,08
COMANGIR	4	1,49	30,45	148,55	18,39	31,41
COMANGIR	5	2,00	24,58	149,38	25,78	31,40
COMANGIR	6	2,00	22,96	149,65	23,43	39,32
COMANGIR	7	1,69	24,56	150,23	21,45	42,26
COMANGIR	8	2,47	25,51	150,47	19,65	32,84
COMANGIR	9	2,45	31,11	150,92	22,30	41,39
COMANGIR	10	2,14	29,78	151,38	17,98	33,45
COMANGIR	11	1,68	32,16	152,15	21,10	34,64
COMANGIR	12	2,54	29,54	152,62	24,98	30,29
COMANGIR	13	2,36	23,45	153,21	18,54	30,78
COMANGIR	14	2,00	23,45	153,92	17,78	35,17
COMANGIR	15	2,38	22,45	154,19	20,45	31,58
COMANGIR	**16**	2,58	20,45	154,31	25,79	35,36
COMANGIR	**17**	1,95	26,78	155,02	24,65	33,10
COMANGIR	**V. Mínimo**	**1,49**	**20,38**	**135,62**	**16,26**	**30,29**
COMANGIR	**V. Máximo**	**2,58**	**32,16**	**155,02**	**25,79**	**42,26**
COMANGIR	**Média**	**2,07**	**25,94**	**150,46**	**21,32**	**34,81**
COMANGIR	**D. Padrão**	**0,34**	**3,98**	**4,43**	**3,10**	**3,89**
TALENAY	1	1,64	27,45	126,72	19,71	44,56
TALENAY	2	1,99	25,12	125,94	24,11	38,39
TALENAY	3	2,25	19,45	126,94	22,26	39,76
TALENAY	4	1,78	20,44	128,12	23,78	33,26
TALENAY	5	2,03	30,45	127,48	18,46	38,12
TALENAY	6	1,89	28,45	131,18	20,21	37,24
TALENAY	7	1,99	24,56	131,15	19,14	36,01
TALENAY	8	2,26	24,23	130,43	23,54	35,98
TALENAY	9	1,63	29,87	131,19	23,30	33,70
TALENAY	10	2,22	20,23	132,48	25,00	36,32
TALENAY	11	1,75	29,17	133,12	18,25	35,00
TALENAY	12	1,81	28,56	131,16	18,87	36,38
TALENAY	**V. Mínimo**	**1,63**	**19,45**	**125,94**	**18,25**	**33,26**
TALENAY	**V. Máximo**	**2,26**	**30,45**	**133,12**	**25,00**	**44,56**
TALENAY	**Média**	**1,94**	**25,67**	**129,66**	**21,39**	**37,06**
TALENAY	**D. Padrão**	**0,23**	**3,95**	**2,46**	**2,51**	**3,01**

QUADRO3. Vetores de resposta das características quantitativas do girassol.

Variedade	Parcela	Alt. Planta	Diâm. Caule	Alt. Capítulo	Diâm. Capítulo	Teor de Óleo
MARIBONDO	1	1,13	13,45	123,93	16,52	44,36
MARIBONDO	2	1,09	14,45	124,82	18,01	40,95
MARIBONDO	3	1,09	15,21	125,94	17,01	40,95
MARIBONDO	4	1,06	19,47	127,02	18,80	48,28
MARIBONDO	5	1,17	18,79	126,08	16,52	41,60
MARIBONDO	6	1,09	17,12	125,92	13,95	45,61
MARIBONDO	7	1,08	12,89	127,42	14,71	44,56
MARIBONDO	8	1,31	16,78	128,02	14,18	44,22
MARIBONDO	9	1,27	22,56	130,61	17,99	40,16
MARIBONDO	10	1,29	29,45	138,03	14,00	38,54
MARIBONDO	11	0,94	18,49	131,48	17,25	39,57
MARIBONDO	12	1,35	17,15	134,42	15,04	39,92
MARIBONDO	13	1,97	23,89	133,45	20,27	42,87
MARIBONDO	14	1,15	12,49	135,42	13,26	36,58
MARIBONDO	15	1,41	25,14	135,94	18,89	37,21
MARIBONDO	**16**	1,27	16,47	138,54	17,53	44,34
MARIBONDO	**17**	1,09	18,13	136,78	18,86	44,75
MARIBONDO	**18**	1,10	13,45	137,04	11,32	47,87
MARIBONDO	**V. Mínimo**	**0,94**	**12,49**	**123,93**	**11,32**	**36,58**
MARIBONDO	**V. Máximo**	**1,97**	**29,45**	**138,54**	**20,27**	**48,28**
MARIBONDO	**Média**	**1,21**	**18,08**	**131,16**	**16,34**	**42,35**
MARIBONDO	**D. Padrão**	**0,22**	**4,63**	**5,08**	**2,39**	**3,39**
3GRNAINS	1	1,70	25,15	113,15	15,41	38,39
3GRNAINS	2	1,79	14,45	112,28	11,16	44,42
3GRNAINS	3	1,19	13,79	114,23	12,98	41,13
3GRNAINS	4	1,54	17,49	111,09	13,78	38,35
3GRNAINS	5	1,64	25,17	110,15	13,78	38,35
3GRNAINS	6	1,72	16,24	108,99	16,54	39,45
3GRNAINS	7	1,31	12,89	111,03	13,26	37,38
3GRNAINS	8	1,75	18,97	108,18	13,26	37,38
3GRNAINS	9	1,68	15,21	109,78	14,34	46,65
3GRNAINS	10	1,68	14,12	110,25	13,81	45,15
3GRNAINS	11	1,53	13,45	111,11	12,65	39,24
3GRNAINS	12	1,59	17,18	121,90	14,00	38,54
3GRNAINS	13	1,82	16,47	122,89	13,65	46,13
3GRNAINS	**V. Mínimo**	**1,19**	**12,89**	**108,18**	**11,16**	**37,38**
3GRNAINS	**V. Máximo**	**1,82**	**25,17**	**122,89**	**16,54**	**46,65**
3GRNAINS	**Média**	**1,61**	**16,97**	**112,69**	**13,74**	**40,81**
3GRNAINS	**D. Padrão**	**0,18**	**4,04**	**4,60**	**1,29**	**3,48**