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

APLICATIVO COMPUTACIONAL DA FUNÇÃO DISCRIMINANTE QUADRÁTICA PARA UTILIZAÇÃO EM CIÊNCIAS EXPERIMENTAIS

SANDRA FIORELLI DE ALMEIDA PENTEADO SIMEÃO

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
Março - 2007

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

APLICATIVO COMPUTACIONAL DA FUNÇÃO DISCRIMINANTE QUADRÁTICA PARA UTILIZAÇÃO EM CIÊNCIAS EXPERIMENTAIS

SANDRA FIORELLI DE ALMEIDA PENTEADO SIMEÃO

Orientador: Prof. Dr. Carlos Roberto Padovani

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

BOTUCATU - SP
Março - 2007

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

Simeão, Sandra Fiorelli de Almeida Penteado, 1965-
S589a Aplicativo computacional da função discriminante quadrática para utilização em ciências experimentais / Sandra Fiorelli de Almeida Penteado Simeão. – Botucatu : [s.n.], 2006.
xi, 143 f. : il. color., gráfs., tabs.

Tese (Doutorado)-Universidade Estadual Paulista, Faculdade de Ciências Agronômicas, Botucatu, 2006
Orientador: Carlos Roberto Padovani
Inclui bibliografia.

1. Análise discriminante. 2. Programação quadrática. 3. Análise multivariada. 4. Algoritmo computacional. 5. Função discriminante quadrática. I. Padovani, Carlos Roberto. II. Universidade Estadual Paulista "Júlio de Mesquita Filho" (Campus de Botucatu). Faculdade de Ciências Agronômicas. III. Título.

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

CERTIFICADO DE APROVAÇÃO

TÍTULO: **"APLICATIVO COMPUTACIONAL DA FUNÇÃO DISCRIMINANTE QUADRÁTICA PARA UTILIZAÇÃO EM CIÊNCIAS EXPERIMENTAIS"**

ALUNA: SANDRA FIORELLI DE ALMEIDA PENTEADO SIMEÀO

ORIENTADOR: PROF. DR. CARLOS ROBERTO PADOVANI

Aprovado pela Comissão Examinadora

PROF. DR. CARLOS ROBERTO PADOVANI

PROF. DR. ADRIANO WAGNER BALLARIN

PROF. DR. FLAVIO EKkÀRI ARAGON

PROF. DR. JOSÉ CARLOS MARTINEZ

PROFA. DRA.~MARIE QSHÍIWA

Data da Realização: 19 de dezembro de 2006

Deus - princípio e fim - obrigada por me permitir alcançar mais este objetivo em minha vida!
Maria - Rainha da Paz - Apoio Incondicional - obrigada pelo suporte!

"Nenhuma grande vitória é conseguida sem pequenas vitórias sobre nós mesmos"

Antonio, meu Pai, de quem herdei a determinação e a organização.
Neide, minha Mãe, de quem herdei a fé e a perseverança.
O fruto do meu trabalho é também seu fruto, pois me deram a vida e a coragem para vivê-la.

Marido e Amigo Vitor, em quem encontrei o respeito, a humildade e a verdade.
Este trabalho também é um produto seu, pois sem seu apoio crítico, cumplicidade, paciência e compreensão, ele não se concretizaria.

Filho João Vitor, bênção divina, sua inocência e carinho me mostram, a cada dia, o amor puro e sincero.
Que você possa se orgulhar de sua mãe no futuro.

A vocês, motivos de minha existência, dedico este trabalho:
a grande vitória sobre mim mesma ...

AGRADECIMENTOS

Ao Professor Dr. Carlos Roberto Padovani, pela competência e extremo profissionalismo com que conduziu a sua orientação, e ainda, pela paciência, disponibilidade e amizade durante o desenvolvimento deste trabalho.

Aos familiares, pelo incentivo e torcida.

Às companheiras de jornada: Maria José Lourenção Brighenti, Regina Célia Baptista Belluzzo, Ivete Maria Baraldi e Rosária Helena Ruiz Nakashima, por me suportarem "naqueles" momentos críticos e me apoiarem incondicionalmente.

À amiga Nilza Regina da Silva pelo companheirismo, cumplicidade, troca de idéias e convivência que, incansavelmente, colaborou em todos os instantes dessa caminhada.

Às Profas. Léa Sílvia Braga de Castro Sá e Marileide Dias Esqueda, pelas correções, sugestões e adequações.

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

Aos colegas da PRAc - Pró-Reitoria Acadêmica da Universidade do Sagrado Coração, pelo apoio e idéias.

Aos alunos do curso de Matemática da USC, por permitirem compartilhar os novos conhecimentos adquiridos, assim como os novos anseios.

Ao Professor Dr. Adalberto José Crocci, pela doação dos primeiros textos sobre o assunto pesquisado.

A Sra. Silvia Padovani, que gentil e carinhosamente, me acolheu em sua residência.

A todos os professores e funcionários da FCA - Botucatu, especialmente a Marilena, Marlene e Kátia (Secretaria da Pós-Graduação), por tantos nós desfeitos.

Às companheiras de luta comum, Ana Vergínia e Marie, pelas “trocas”.

Aos funcionários da Biblioteca da FCA, pela atenção e profissionalismo.

À Secretária do Departamento de Bioestatística - Elizabete - pela simpatia e colaboração.

A “TODOS” os que me acompanharam e comigo conviveram nesta jornada, agradeço-lhes com profundo respeito: a torcida, as orações e, principalmente, a cumplicidade.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

Aspectos teóricos relacionados à Análise Discriminante Multivariada - Linear e Quadrática - foram discutidos, por meio de um extenso levantamento histórico da função discriminante, com seus primórdios no trabalho de *Fisher* e sua posterior evolução, enfocando o intenso desenvolvimento das técnicas classificatórias discriminantes com o advento dos computadores. Foi dada ênfase aos *softwares* estatísticos desenvolvidos para *PC*, que realizam a análise discriminante, e que representam uma grande contribuição para pesquisadores e usuários desta técnica. Considerando a dificuldade existente quanto a aplicativos computacionais acessíveis a pesquisadores da área de ciências agrárias, elaborou-se um programa que realiza a análise discriminante quadrática com as respectivas freqüências de classificação correta, bem como o manual explicativo do usuário. Verificou-se que a função discriminante quadrática trata de um procedimento bastante útil nas ciências agrárias, como, por exemplo, em estudos nas áreas de solos, cultivos diversos (soja, milho, cana de açúcar, pupunha, braquiária, frutas), criação de animais e classificação e seleção de madeiras; porém, subutilizada frente à dificuldade de programas computacionais de fácil manuseio e acesso a pesquisadores das áreas aplicadas. Os procedimentos estudados e discutidos foram ilustrados com exemplos de aplicação, utilizando dados experimentais agronômicos de espécies de *Girassóis* e *Eucalyptus*, submetidos ao aplicativo desenvolvido.

Palavras-chave: análise discriminante, função discriminante quadrática, taxa de classificação incorreta, algoritmo computacional.

SOFTWARE OF THE QUADRATIC DISCRIMINANT FUNCTION FOR USE IN EXPERIMENTAL SCIENCES. Botucatu, 2006, 154 p. Tese (Doutorado em Agronomia/Energia na Agricultura) - Faculdade de Ciências Agronômicas, Universidade Estadual Paulista.
Author: SANDRA FIORELLI DE ALMEIDA PENTEADO SIMEÃO
Adviser: CARLOS ROBERTO PADOVANI

SUMMARY

A large historical study of the discriminant function has allowed a discussion on theoretical aspects related to the Multivaried Discriminant Analysis – Linear and Quadratic, showing its past in the work of *Fisher* and its later evolution, emphasizing the wide development of classificatory discriminant techniques with the happening of the computers, and specific statistic softwares which practice the discriminant analysis, representing a big contribution to researches and users of this technique. Considering the difficulty in relation to accessible softwares to researches of the agrarian area, a software which performs a linear and quadratic discriminant analysis was built with its frequencies of correct classification, as well as an explicative manual to users. The quadratic discriminant was studied as being a very useful process in agrarian sciences. Some examples of this usefulness is in studies of the ground, diversified cultivation (soybean, corn, sugarcane, pejibaye, *brachiaria decumbens* fruits), animal creation and wood selection, and classification; however, misused in relation to the difficulties of easy handing and access to researchers of applied areas. The studied and discussed procedures were illustrated with applications, using agronomic experimental data of *Sunflower* and *Eucalyptus*, submitted to developed software.

Keywords: discriminant analysis, quadratic discriminant function, wrong classification rate, computing algorithm.

1 INTRODUÇÃO

Na pesquisa agronômica muitos experimentos são conduzidos de maneira a se obter várias mensurações (medidas de várias características biológicas) na mesma unidade experimental ou parcela. Além disso, existem algumas situações em que o vetor observacional constitui-se de uma mesma variável resposta considerada ao longo de um período experimental, composto por vários momentos ou tempos de mensuração.

Todas estas situações são exemplos típicos de problemas que imprescindem de técnicas multivariadas na análise dos dados de observação; porém, não raro são avaliadas e discutidas sem qualquer consideração da estrutura de dependência dos dados. Deve ser destacado que o termo "multivariado" não é usado de maneira consistente na literatura. Vários textos utilizam a análise multivariada como uma simples junção multifatorial de procedimentos univariados.

A Análise Multivariada constitui-se no ramo da Estatística que objetiva o resumo, a representação e a interpretação de dados observados a partir de

populações onde cada unidade experimental envolve a mensuração de diversas variáveis. O interesse em medir um número expressivo de características em cada unidade experimental deve-se ao fato que, algumas vezes, isoladamente, as variáveis podem não conseguir caracterizar, de maneira adequada, o conteúdo biológico da parcela, ou ainda, em situações e que informações importantes sobre a estrutura de variabilidade dos dados não devem ser negligenciadas ou irrelevadas.

Sem exagero afirmativo, pode-se dizer que muitos dos processos de experimentação agronômica são multivariados, e que a falta de um procedimento multidimensional na análise dos dados observacionais, com certeza, deve sobrestar em muito a acurácia e fidedignidade da discussão biológica dos resultados.

Na área agronômica, freqüentemente, faz-se necessária a distinção estatística entre dois ou mais grupos de indivíduos, previamente definidos a partir de características conhecidas para todos os membros dos grupos, ou seja, deseja-se discriminar grupos de indivíduos definidos *a priori* com base em um critério pré-definido, a partir da informação recolhida sobre os indivíduos desses grupos. Muito comum também é a necessidade de locar um "novo indivíduo" em uma de várias populações conhecidas. A técnica multivariada que atende estas necessidades é conhecida como Análise Discriminante. Esta, além de determinar subsídios para classificação de indivíduos em grupos, consiste numa maneira interessante de análise exploratória das populações consideradas.

Basicamente, a Análise Discriminante constitui-se em um conjunto de processos estatísticos com a finalidade de alocar um novo indivíduo X a uma de g populações distintas, previamente conhecidas, admitindo-se que X realmente pertença a uma dessas g populações. O tratamento estatístico dado a esse problema de locação reside no fato que os dados utilizados são os valores de um conjunto de p variáveis aleatórias ($X_1, \ldots, X_p$) (JOHNSON e WICHERN, 1998).

Levando-se em consideração que o problema de classificação pode ser encarado como o estabelecimento de uma regra de decisão estatística, torna-se fundamental

que a construção do procedimento ou critério de classificação seja tal que minimize a probabilidade de classificação errônea de um indivíduo em uma população Π_i , quando ele realmente pertencer à população $\Pi_{i'}$, ($i \neq i'$) $i, i' = 1, 2, \ldots , g$.

Os procedimentos de classificação baseados em populações que possuem distribuições normais multivariadas são predominantes na Estatística devido à sua simplicidade e alta eficiência para uma ampla variedade de modelos populacionais. Se a homogeneidade entre as matrizes de covariâncias das populações é verificada, a função a ser ajustada é uma função de primeiro grau, denominada "Função Discriminante Linear de Fisher". Caso a homogeneidade não se comprove, então a função adequada é a "Função Discriminante Quadrática" (JOHNSON e WICHERN, 1998).

Em muitas situações agronômicas, mesmo não existindo a homogeneidade entre as matrizes de covariâncias, a função linear é aplicada para se efetuar a discriminação, uma vez que a utilização da função quadrática é considerada complexa pelos pesquisadores, pois os procedimentos matemáticos envolvidos são mais elaborados. Existem alguns *softwares* que atendem a esta solicitação como o *SAS, STATISTICA* e o *SPSS*. Entretanto, a área agronômica se ressente da falta de um aplicativo mais específico que contemple suas necessidades.

Dessa maneira, os objetivos norteadores deste trabalho são relacionados a seguir.

- discutir aspectos teóricos relativos à Análise Discriminante Multivariada (Linear e Quadrática), com ênfase à sua utilização na experimentação agronômica;
- desenvolver e disponibilizar um programa computacional com características simples de manuseio e respectivo manual explicativo do usuário;

- ilustrar os procedimentos discutidos com exemplos de aplicação envolvendo dados experimentais agronômicos.

O texto foi subdividido, para melhor compreensão, em seis capítulos, apresentados a seguir.

Capítulo 1 - *Introdução* - define a motivação para o trabalho, o objetivo do mesmo e sua organização, destacando as aplicações da análise multivariada, especificamente da análise discriminante, em diversos processos das ciências agronômicas.

Capítulo 2 - *Revisão da Literatura* - apresenta um levantamento bibliográfico que fornece um posicionamento histórico com uma ampla visão do desenvolvimento teórico, bem como referências do uso da função discriminante quadrática nas aplicações em ciências agrárias. Contempla também uma relação de *softwares* estatísticos que realizam a análise discriminante.

Capítulo 3 - *Desenvolvimento Metodológico* - embasamento teórico - revisão da metodologia estatística para determinação da função discriminante - enfocando os itens de maior interesse para a compreensão do estudo.

Capítulo 4 - *Programa Computacional DISCRIMINANTE* - descrição dos componentes do programa desenvolvido, bem como características de seu funcionamento - Manual do Usuário.

Capítulo 5 - *Resultados e Discussão* - abordagem prática da técnica da análise discriminante aplicada a dois conjuntos de dados - *Girassóis* e *Eucalyptus* - apresentando os resultados classificatórios comparativos das funções linear e quadrática para ambos.

Capítulo 6 - *Conclusões* - apresenta as conclusões obtidas a partir da análise dos referenciais teóricos e da abordagem prática, com sugestões de trabalhos futuros baseados nestas técnicas classificatórias.

2 REVISÃO DA LITERATURA

2.1 Evolução histórica e teórica da função discriminante

O procedimento, que ficou conhecido como *Análise Discriminante*, tem como idéia básica substituir o conjunto original das diversas mensurações (variáveis) por um único valor *Y*, tal como na igualdade 2.1, definido como uma combinação linear destas (REIS, 1997).

$$Y_i = a_{i0} + a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p \qquad (2.1)$$

Esta técnica, a partir da generalização do processo, permitiu que novas situações experimentais fossem consideradas, ou seja, discriminar indivíduos entre mais de dois grupos.

O problema da discriminação é bastante freqüente em diversas áreas científicas, como pode ser constatado pelos exemplos descritos a seguir.

i) Botânica: certa planta deve ser classificada em uma das espécies pré-determinadas ($\Pi_1,...,\Pi_g$), baseando-se apenas nas suas características morfológicas (X).
ii) Medicina: com base nos sintomas do paciente e em alguns exames preliminares (X), um médico precisa decidir se o paciente é portador de certa doença (Π_1) ou não (Π_2).
iii) Finanças: um analista de crédito deve decidir, com base nas informações (X) de uma ficha cadastral, se concede ou não crédito ao cliente. O crédito será concedido se o cliente for considerado bom pagador (Π_1).
iv) Arqueologia: um crânio (X) é descoberto em escavações e deseja-se saber se pertenceu a um ser do sexo masculino (Π_1) ou feminino (Π_2).
v) Política: satisfação de eleitores que votaram no partido vencedor das últimas eleições (Π_1) contra aqueles que votaram nos outros partidos.
vi) Mercado consumidor: grupo de consumidores de marcas diferentes para um mesmo produto.

Esta técnica de análise multivariada é empregada para descobrir as características que distinguem os membros de um grupo das de outro, de modo que, conhecidas as características de um novo indivíduo, se possa prever a que grupo pertence.

Ao tomar várias características combinadas matematicamente, a análise discriminante tem em vista encontrar uma ou mais dimensões que maximizem a distinção entre os grupos. Pressupõe-se que as variáveis independentes (X_i) têm distribuição conjunta multivariada enquanto a dependente (Y_i) é fixa e de tipo nominal (assume os valores 1, 2, ..., g, de acordo com o número de grupos).

A solução para o problema de discriminação entre duas populações foi originalmente desenvolvida na Botânica e tratou da *função discriminante* proposta por Fisher em 1935 e publicada pela primeira vez em 1936 (FISHER, 1936). Sua aplicação teve como objetivo fazer a distinção de 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).

Fisher sugeriu o uso de uma função linear das variáveis aleatórias $X_1,...,X_p$, que caracterizavam os indivíduos ou objetos de modo exaustivo em grupos mutuamente exclusivos, com base num conjunto de variáveis independentes, cujos coeficientes deviam ser calculados de forma 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 (*within*) de cada grupo.

Embora a função discriminante tenha sido sugerida para o trabalho específico com duas populações, seu uso foi estendido (BERNARD, 1935) para a discriminação entre mais de duas. Sendo um escalar, a função discriminante não esgota 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).

Ciente de que o conhecimento da configuração das populações era importante no problema de discriminação, Fisher sugeriu, em 1938, um teste para verificar a colinearidade de k populações.

O enfoque dado à discriminação sob o ponto de vista da função discriminante de Fisher era baseado em maximizar a distância entre as populações, não levando em consideração o problema de minimização das probabilidades de erro no processo discriminatório.

Welch (1939) abordou o problema de discriminação entre duas populações, considerando as probabilidades de erro de classificação. Para tanto, baseou-se nas idéias introduzidas por Neyman e Pearson (1933) sobre noções de erro na aplicação de técnicas estatísticas.

A solução básica proposta por Welch consistia em particionar o espaço amostral em duas regiões R_1 e R_2 de tal forma que, se um elemento pertencesse a R_1, seria classificado como sendo da população 1 ou, caso contrário, da população 2. A escolha da partição devia ser feita de modo a minimizar a probabilidade de erro de classificação. Dois casos distintos foram considerados: as probabilidades "a priori" conhecidas e as probabilidades "a priori" desconhecidas. Em ambos os casos, ficou demonstrado que as partições eram definidas com auxílio da razão de verossimilhança. Uma das desvantagens do trabalho de Welch era a restrição para o caso de duas populações.

Nos anos seguintes, os trabalhos de Jackson (1943) e Beall (1945) sugeriram aproximações para a determinação da função discriminante, uma vez que o cálculo dos coeficientes b_i era bastante elaborado.

Von Mises (1945) apresentou um trabalho sobre a discriminação entre k ($k \geq 2$) populações, abordando o problema de uma forma um pouco distinta de Welch, pois trabalhou com as probabilidades de classificação correta associadas a cada população.

Para os casos em que as populações se distribuíam normalmente, mas possuíam diferentes matrizes de covariâncias, Smith (1947) provou que as funções discriminantes quadráticas eram mais adequadas, significando que a superfície discriminadora era uma quádrica no espaço p-dimensional. Por meio de dois exemplos numéricos, para os quais foram calculados os dois tipos de funções discriminantes – linear (considerando-se a variância comum) e quadrática (sem considerar variância comum) – comprovou-se que a quadrática era mais adequada apresentando a melhor discriminação com probabilidade de erro menor que a função linear.

Além da sugestão da função quadrática, Smith fornece, por meio de um teorema, a importância da função discriminante linear, o qual estabelece sua suficiência com respeito a duas populações em consideração, isto é, nenhuma perda de informação resulta quando as medidas múltiplas são reduzidas a uma simples função linear.

Em 1948 um novo conceito para os problemas de discriminação foi introduzido por Rao. A modificação sugerida foi a partição do espaço amostral da discriminação em $k+1$ regiões: $R_0,R_1,...,R_k$, sendo que R_0 era denominada região de dúvida. Até então as k populações eram divididas em k regiões $R_1,...,R_k$. A utilização do novo conceito dava-se da seguinte forma: o aparecimento de R_0 era motivado pela construção de $R_1,...,R_k$ sob limites pré-estabelecidos das probabilidades de classificação errada. Se o elemento a ser classificado pertencesse a R_0, nenhuma decisão seria tomada quanto à sua classificação. Rao estendeu para k populações a solução proposta por Welch quando as probabilidades "a priori" eram conhecidas. No caso destas últimas não serem conhecidas, a sugestão era de que fossem consideradas iguais.

Tais processos discriminatórios, construídos sob a consideração das probabilidades de erro, apresentavam o inconveniente de assumirem como conhecidas as distribuições conjuntas das p variáveis $X_1,...,X_p$, que caracterizavam os elementos das populações. Nem sempre estas distribuições eram conhecidas e a saída óbvia era a substituição, nas regras discriminantes dos parâmetros populacionais, por suas estimativas. Muitos estudiosos ocuparam-se do estudo das distribuições amostrais das estatísticas utilizadas nas regras discriminatórias com a finalidade de estudar as probabilidades de erro e, nem sempre, tais distribuições eram de fácil utilização.

Hoel e Peterson, em artigo publicado em 1949, abrem uma nova linha de estudos na Análise Discriminante. Neste trabalho, a estimação dos parâmetros populacionais, das probabilidades "a priori" e a discriminação são estudadas simultaneamente. A idéia básica é a obtenção de um conjunto de estimadores que maximize a probabilidade de classificação correta.

Um estudo sobre a extensão da função discriminante linear para *k* populações foi apresentado por Bryan (1951) com a introdução das "funções discriminantes múltiplas". Estas funções que, originalmente, foram construídas para classificar um elemento em uma de *k* populações, tiveram vital importância no estudo das direções em que variavam as populações (configuração das populações).

Comprovando o estudo de Smith, de anos atrás, quanto à suficiência da função discriminante para duas populações, Rao (1962) generalizou os resultados para um conjunto mais amplo de alternativas populacionais.

Apesar da comprovação matemática de que a função quadrática era a mais adequada para a discriminação entre duas populações com diferentes matrizes de covariâncias, Anderson e Badahur (1962) propuseram a construção de um modelo linear que minimizava a probabilidade de má classificação por meio de um procedimento linear *minimax*, como também discutiram alguns modelos lineares *bayesianos*. De acordo com os autores, a dificuldade de utilização da função quadrática refletia-se na análise das regiões de discriminação, as quais não eram, necessariamente, formas quadráticas definidas positivas, dependendo muito de suposições de normalidade e especialmente da forma da distribuição normal relativamente à distância do seu centro.

Ainda em 1962, Lubischew reforça positivamente o uso de funções discriminantes lineares na taxonomia de espécies de *Chaetocnema* (Pulga-do-arroz), afirmando que seu uso será tão mais efetivo quanto melhor for a seleção das características medidas. Para tanto, as características deverão possuir, individualmente, um alto coeficiente de discriminação – *K*, relativo à variabilidade inter e intra-específica, bem como, tomada aos pares, uma alta correlação intra-específica contra a inter-específica. Enfatiza a importância da avaliação gráfica da discriminação por meio do gráfico de dispersão e elipses correlacionadas, sugerindo, inclusive, um novo critério chamado *rank de discriminação*.

Devido a um problema proposto por Burnaby (1966), um novo rumo nas pesquisas sobre Análise Discriminante foi adotado. Até então os problemas tratados

resumiam-se às classificações de elementos em uma de *k* populações. O problema proposto por Burnaby tratava-se da classificação de um elemento em um de dois conjuntos, cada um dos quais constituído por uma mescla de diferentes populações em proporções desconhecidas. Rao também abordou o assunto propondo, ainda em 1966, uma solução para o mesmo.

Conforme a década de 70 se aproximava, algumas investidas em testes computacionais para os problemas de discriminação foram acontecendo, como fica comprovado no trabalho de Horton et al. (1968), que sugere a utilização de dois programas: o primeiro para análise multivariada da covariância e o segundo, que realizava a análise canônica. Ambos trabalhavam com dados agrupados de solos, visando à classificação quanto a três classes microtopográficas. Tais procedimentos foram utilizados, pois os autores necessitavam de um método de análise que discriminasse entre os grupos, o que não era obtido por meio da análise univariada. Com a aplicação dos programas, os mesmos conseguiram unir as vantagens da análise multivariada com a economia do número de variáveis necessárias.

Gilbert (1969) empreendeu uma investigação sobre os efeitos da desigualdade das matrizes de covariâncias na função discriminante linear de Fisher (LDF - *linear discriminant function*) quando usada para discriminação e estimação de riscos. Suas conclusões foram baseadas na comparação entre os resultados relativos às probabilidades de má classificação obtidos para a função linear e a forma quadrática equivalente, quando os parâmetros para as duas populações envolvidas no estudo eram conhecidos e todas as correlações consideradas iguais. Como o autor suspeitava, a forma quadrática, devido ao uso das discrepâncias nas variâncias, mostrou uma probabilidade de má classificação diminuída em relação à forma linear, indicando que a LDF é adequada para classificação, mas não satisfatória para a estimação de riscos.

Em 1972, dez anos após o falecimento de Sir Ronald Aylmer Fisher, Rao realizou uma extensa revisão das contribuições do matemático e biólogo que mudou o rumo das pesquisas em estatística multivariada, apresentando as recentes tendências do trabalho nesta área. Para tanto, discutiu alguns dos métodos multivariados mais novos que

pareciam ser de utilização imediata para pesquisadores de vários campos, bem como revisou brevemente alguns dos desenvolvimentos teóricos mais importantes.

Em relação à discriminação, Rao enfatizou o grande progresso nas pesquisas quanto aos testes de ajuste nas funções discriminantes, como também sua aplicação satisfatória na análise estatística de dados relacionados a crescimento e sua predição. Enfatizou também a grande revolução na pesquisa científica com o uso do computador, não só na estatística de maneira geral, mas especialmente na análise multivariada, com o desenvolvimento de novas técnicas assistidas computacionalmente.

A década de 70 indicou um forte esforço dos pesquisadores na utilização de aplicativos computacionais auxiliares na resolução de seus problemas estatísticos. Os equipamentos disponíveis nesta época eram de grande porte - *mainframes* - e sua utilização era bastante restrita aos centros de pesquisa, como pode ser constatado no trabalho de Schucany et al. (1972), que descreveram uma coletânea de aplicativos estatísticos utilizados nestes equipamentos.

Para a discriminação entre k populações normais multivariadas com a mesma matriz de covariâncias e diferentes vetores de médias e para classificação de uma nova observação para alguma delas, dois procedimentos, ambos de Fisher, são conhecidos na literatura. Um é a generalização da construção de funções discriminantes para duas populações, enquanto o outro envolve a maximização da razão entre grupos e dentro dos grupos das somas de quadrados e produtos. Arseven e Kshirsagar (1975) provaram em seu artigo que ambos os procedimentos, generalizados para k populações, são equivalentes sob certas condições. Estas envolvem a eliminação de variáveis canônicas que não interferem na capacidade de discriminação.

Com o objetivo de testar a melhor função discriminante, Kronmal e Wahl (1977) efetuaram um estudo comparativo para duas populações multivariadas com distribuição normal e diferentes matrizes de covariâncias, envolvendo as funções discriminantes linear de Fisher e quadrática. Os autores apontaram que o uso da função linear,

mesmo quando a restrição de igualdade de matrizes de covariâncias não se satisfazia, era mais comum entre os pesquisadores devido a sua simplicidade de forma e conceito, pois o uso da quadrática era questionável quando de sua construção baseada em parâmetros estimados, situação em que seu desempenho era superado pela linear. No entanto, os resultados retratados no artigo foram bastante esclarecedores quanto ao uso de uma ou outra, baseado em importantes considerações: o tamanho da amostra (quanto maior, aumentam as vantagens do uso da quadrática); dimensão e diferenças de covariâncias pequenas permitiam o uso de ambas, porém quando tais diferenças eram grandes, o desempenho da quadrática era extremamente melhor que o da linear. Concluindo, os autores afirmaram que pequenas amostras com substanciais diferenças de covariâncias e presença de grandes dimensões numéricas eram situações desconfortáveis; pois, apesar do melhor desempenho da função linear sobre a quadrática, a alta probabilidade de má classificação contribuía para inviabilizar seu uso, conforme já havia constatado Gilbert (1969).

Também com o objetivo de comparar o desempenho de métodos discriminantes, três deles foram testados por Dunn e Marks (1974): discriminação quadrática, linear e a melhor linear. Os três métodos foram aplicados na classificação de indivíduos em duas populações multivariadas normalmente distribuídas, porém com matrizes de covariâncias diferentes. A comparação foi executada usando amostras e assintoticamente. Os parâmetros que foram variáveis no estudo incluem a distância entre as populações, matrizes de covariâncias, número de dimensões, tamanho da amostra e as probabilidades *a priori* originais das populações. As razões assintóticas indicaram que, para grandes amostras de distribuições normais, a função quadrática é muito melhor que a função linear para grandes valores de λ; para pequenos valores é fracamente melhor. Para pequenas amostras, entretanto, a função quadrática tem desempenho muito pior que a linear, para pequenos valores de λ e esta tendência é incrementada com o número de parâmetros. Para pequenos valores de λ, a função melhor linear tem desempenho um pouco melhor que a linear; para valores maiores de λ, é notoriamente melhor que a linear, mas neste intervalo, a quadrática ainda é usualmente melhor.

Condições de êxito ou falha no desempenho da função linear de Fisher sob condições não ótimas, foram apontadas no trabalho de Krzanowski (1977). Este destacou que o principal atrativo na utilização da função linear era a simplicidade da técnica, bem como a facilidade de seu manuseio computacional, dando ênfase à disponibilidade de seus resultados para pesquisadores poderem comparar e avaliar o sucesso ou fracasso da FDL sob as condições não ótimas: matrizes de covariâncias diferentes, ausência de normalidade (dados contínuos não normais, dados discretos, mistura de variáveis discretas e contínuas).

A pesquisa, envolvendo análise discriminante para problemas de alta dimensão com o objetivo de mostrar as mudanças nas probabilidades de classificação correta quando esta dimensão é aumentada, foi a motivação de Van Ness (1979) em seu artigo. Para tanto, considerou duas populações normais com médias e matrizes de covariâncias conhecidas, sendo que, para estas, $\Sigma_1 = I$ e $\Sigma_2 = \frac{1}{2}I$, ou seja, diferentes. Tais dados foram submetidos a seis algoritmos diferentes: discriminante linear, discriminante quadrático – executado, a princípio, como nivelador para os outros, e depois com os mesmos dados – dois algoritmos de Bayes com núcleos gaussianos e o algoritmo da média entrelaçada (*average linkage*). Os resultados, apresentados de maneira gráfica, mostraram particularidades específicas para cada caso, conforme os parâmetros de Δ, p e γ eram alterados; porém, a conclusão geral foi de que os algoritmos não paramétricos são completamente estáveis para altas dimensões, mantendo fortemente seu desempenho.

Do ponto de vista computacional, a década de oitenta teve a presença do computador consolidada, mais especificamente do microcomputador, sendo publicado o trabalho de Woodward e Elliott (1983), trazendo os resultados de um levantamento de 26 aplicativos estatísticos específicos para microcomputadores. O objetivo deste trabalho foi apresentar à comunidade de estatísticos e cientistas as disponibilidades de *softwares* para microcomputadores visando à utilização em suas pesquisas, bem como seu envolvimento 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.

Ainda no mesmo ano - 1983 - Ivor Francis, do Departamento de Estatística da Universidade Cornell - USA, também publicou um *survey* sobre *softwares* de estatística, porém dando um enfoque bastante geral e abordando aqueles utilizados nos anos de 79 e 80. Segundo o autor, os aplicativos computacionais transformaram-se em importantes e permanentes ferramentas para a prática da estatística, tal o número de pesquisadores que neles confiavam para suas análises. Relata que no passado - 20 anos atrás - a invasão dos programas computacionais foi vista pelos estatísticos como uma praga que iria prejudicar seus ambientes de trabalho; porém, com o passar dos anos, não somente muitos estatísticos aprenderam a conviver com o "invasor", como também experimentaram um proveitoso sinergismo com os mesmos. Assim, o que o motivou no desenvolvimento deste artigo foi o interesse em levantar quais eram os programas, seus produtores, seus detalhes e características - algoritmos implementados, tempo de processamento, acurácia dos resultados, existência de manuais de usuário e outros. Para tanto, o autor coletou uma grande quantidade de dados dos desenvolvedores dos aplicativos e comparou com as experiências dos usuários dos mesmos, organizando as respostas de maneira a tornar a informação compreensível para outros usuários em potencial. Como resultado desse levantamento, 15 programas foram eleitos pelos usuários, destacando-se os itens padrão de documentação e o nome do aplicativo atribuído pelo desenvolvedor, ou seja, significante ao propósito do programa.

Também em 1983, Kane, Bayne e Beauchamp sumarizaram os resultados de um estudo comparativo dos desempenhos de três métodos classificatórios, as funções discriminantes: linear, linear logística e quadrática. Cada regra de classificação foi comparada ao procedimento de máxima probabilidade para três tipos de distribuição bivariada. As probabilidades de má classificação teóricas das amostras das funções discriminantes foram calculadas diretamente e usadas para a comparação dos diferentes procedimentos em termos de tendência e variação. Como conclusão, os autores mostraram a importância que a especificação do tipo de função discriminante pode ter no procedimento de análise classificatória, fazendo recomendações para os estatísticos das áreas aplicadas.

Com o intuito de apresentar outro método de discriminação que minimizasse os problemas das técnicas linear e quadrática, Friedman (1989) propôs o método

conhecido como Análise Discriminante Regularizada - RDA. Seu estudo enfocou conjuntos de dados nos quais o número de variáveis é comparável ao número de parâmetros a serem estimados; a estimativa da matriz de covariâncias é altamente incerta e os autovalores, se pequenos, são levados a extremos muito baixos, se grandes, levados a extremos muito altos. Para conjuntos de dados mal condicionados, se o número de variáveis é maior que o número de objetos do grupo todo e menor que o número total de objetos, QDA (análise discriminante quadrática) não pode ser aplicada, pois a matriz de covariâncias é singular. Se o número de variáveis é maior que o número total de objetos, ambos, QDA e LDA (análise discriminante linear) não podem ser utilizados, uma vez que Σ_g e Σ_{pooled} são singulares.

Dessa maneira, Friedman mostrou que estes problemas podem ser contornados pela aplicação da regularização apresentada em 2.2, em que I é a matriz identidade; o parâmetro $\lambda \in [0,1]$ controla os graus que podem ser utilizados para a matriz *pooled* de covariâncias e o parâmetro de regularização adicional $\gamma \in [0,1]$, controla os graus de "encolhimento" dos autovalores; e $\Sigma_g(\lambda) = (1-\lambda)\Sigma_g + \lambda\Sigma_{pooled}$.

$$\Sigma_g(\lambda,\gamma) = (1-\gamma)\Sigma_g(\lambda) + \frac{\gamma}{p} tr\left[\Sigma_g(\lambda)\right] I \qquad (2.2)$$

Quando $\lambda = 1$ e $\gamma = 0$, RDA é idêntica a LDA. Por outro lado, se $\lambda = 0$ e $\gamma = 0$, o resultado é QDA. Caso contrário, RDA é indicada. Como conclusão final, Friedman observou que QDA só é viável se a razão de objetos para variáveis é maior. LDA então é um método a ser escolhido. Quando o tamanho da amostra é muito pequeno para LDA ou quando as matrizes de covariâncias são muito diferentes e o *pooled* não pode ser usado, RDA pode ser empregado como método alternativo.

Ainda tratando com dados de alta dimensionalidade, Schott (1993) enfatiza que um dos objetivos comuns de muitas técnicas multivariadas é obter uma redução de dimensionalidade enquanto, ao mesmo tempo, retém a grande maioria da informação relevante contida no conjunto de dados original. Esta redução não somente provê uma

descrição parcimoniosa dos dados, mas em muitos casos, incrementa a confiança nas subseqüentes análises dos dados. O autor empenhou-se em determinar a dimensão mínima necessária para a aplicação da análise discriminante quadrática em populações normais com matrizes de covariâncias heterogêneas. Para tanto, foram usadas simulações para investigar a adequação da aproximação *Qui-Quadrado* e para comparar as probabilidades de má classificação da discriminação quadrática sob o conjunto de dados com dimensão reduzida e sob o conjunto completo original. Concluiu que reduzir a dimensionalidade antes de aplicar a discriminação quadrática diminui o tamanho necessário da amostra e permite que a função quadrática seja preferida, por suas respostas mais acuradas, ao invés da linear.

Grouven et al. (1995), ao enfatizarem a importância da vasta aplicação da análise discriminante linear e quadrática nas pesquisas biológicas e médicas, oferecem à comunidade científica um aplicativo computacional, escrito em Linguagem *Borland Pascal 7.0*, que executa a discriminação, linear ou quadrática, considerando imparcialidades e predições que ajudam a reduzir os erros de aproximações em grande parte das situações das áreas pesquisadas. Os autores destacam que a discriminação linear (LDA) e a quadrática (QDA) estão disponíveis em alguns aplicativos de análise estatística, como o *BMDP, SAS* e *SPSS*, entretanto, versões modificadas da LDA e QDA, chamadas linear imparcial (ULDA) e quadrática imparcial (UQDA) e ainda, linear preditiva (PLDA) e quadrática preditiva (PQDA), foram propostas uma vez que melhoram os resultados, sob certas condições, em termos da acurácia da classificação. Assim, o programa apresentado pelos autores fornece este tipo de resultado, ilustrado por meio da análise das respostas obtidas de dados reais de pacientes que foram alocados para um de vários estágios de anestesia com base em medidas eletroencefalométricas.

Utilizando as conclusões das pesquisas de Friedman (1989), Wu et al. (1996) compararam os três classificadores: análise discriminante linear (LDA), quadrática (QDA) e regularizada (RDA) para dados químicos de alta dimensionalidade obtidos por Espectroscopia no Infravermelho Próximo (NIR). Os autores concluíram que, em diversos casos, RDA reduz-se a LDA ou a QDA dependendo de alguns parâmetros, podendo inclusive

apresentar-se com melhor desempenho. Nos casos em que pequenos ganhos na qualidade da classificação foram importantes, a aplicação do RDA foi considerada mais útil.

Retomando o trabalho apresentado no ano anterior, Grouven et al. (1996) novamente enfatizam a grande utilização da análise discriminante linear e quadrática nas pesquisas biológicas e médicas. Para tanto, um novo aplicativo computacional é apresentado pelos autores, também escrito em Linguagem *Borland Pascal*, que executa a discriminação, linear ou quadrática, considerando a possibilidade do processamento dos dados com diferentes custos de má-classificação para os diversos grupos em estudo. Apesar de outros conhecidos aplicativos estatísticos, *BMDP, SAS* e *SPSS*, muito utilizados, realizarem a análise discriminante, os mesmos não apresentam a possibilidade de alteração dos custos, o que, na opinião dos autores, fornece resultados menos acurados e mais distantes da realidade.

Os estudos de Hase et al. (2000) levaram à proposta de uma nova função discriminante quadrática baseada na demonstração de que os autovalores da matriz de covariâncias obtida a partir de amostras são estimadores oblíquos, possibilitando seu uso retificado na construção de uma nova matriz de covariâncias, a qual será usada para a determinação da função. Trata-se de uma função bastante específica, apresentando resultados favoráveis, demonstrados pelo Método de Monte Carlo, para casos de pequenas amostras.

Cooke (2004) também dedicou um artigo ao estudo do desempenho da função discriminante quadrática sob certos limites inferiores. Enfatiza que a função quadrática é freqüentemente usada para separar duas classes de pontos em um hiperespaço multidimensional. Quando as duas classes são normalmente distribuídas, isto resulta em uma separação ótima. Em alguns casos, entretanto, a suposição da normalidade é muito pobre e o erro de classificação é incrementado. Assim, o autor deduz um limite superior teórico para o erro de classificação adequado à superfície de decisão quadrática. Este limite é rigoroso quando as classes de médias e covariâncias e a superfície discriminante quadrática satisfazem certas condições específicas de simetria e, para este caso, a função quadrática que minimiza o pior erro possível de classificação é a função discriminante linear. Os resultados numéricos indicaram que para uma distribuição Gaussiana bem separada, o uso da máxima probabilidade

discriminante somente dá uma pequena melhora na razão de erro quando comparada com a função linear. Por outro lado, a função linear dá uma significante melhora no limite superior do erro para o caso das classes serem não-Gaussianas, o que a torna mais robusta. Enquanto seria conveniente encontrar um limite de erro estrito para o problema não simétrico, um problema de significância mais prática é o efeito do erro de estimação. Na prática, as médias e covariâncias de cada classe não serão conhecidas exatamente, mas devem ser estimadas de um conjunto de amostras. Um método para proceder com o erro de estimação é encontrar um intervalo de confiança para a média e covariância estimada e então escolher o caso que dá o pior limite de erro para o conjunto de parâmetros possível. Para este caso, o pior cenário ocorre para um valor de variância maior, os quais são mais fechados para a superfície de decisão e os valores numéricos para o limite de erro.

Hua et al. (2005) realizaram uma pesquisa com o objetivo de encontrar um método essencialmente analítico de produzir uma curva de erro como uma função do número de variáveis, de forma que a curva possa minimizar a determinação do número ótimo de variáveis. Foi usada uma aproximação normal para a distribuição da função discriminante quadrática estimada. A exatidão da média e da variância estimadas para a função colabora na previsão de como a matriz de covariâncias afeta o número de variáveis ótimas. Os autores usaram um procedimento de estimativa da média e variância e compararam com as variáveis ótimas usando a aproximação normal para estimar a função discriminante com otimização obtida pela simulação da distribuição real desta função. Este tipo de otimização proporciona um imenso ganho computacional em comparação com a obtida via simulação. Concluíram que a imparcialidade destes estimadores garante a boa estimativa para grandes amostras, porém não se enquadrando para pequenas.

2.2 Utilização da análise discriminante nas Ciências Agrárias

As técnicas de classificação são largamente utilizadas na experimentação agronômica e zootécnica tal é a diversidade de problemas que necessitam do estabelecimento de regras de separação entre espécies, identificação de grupos de caracteres

que mais contribuem para esta separação e atribuição de novos indivíduos para populações existentes. Dentre as diversas técnicas de classificação, a análise discriminante é apontada como uma das mais adequadas e completas, sendo comprovada pelas recentes referências publicadas.

O trabalho de Adams et al. (2000), que trata da identificação de deficiências de Mn, Zn, Cu e Fe em soja, beneficiou-se do uso da análise discriminante quadrática para determinar se medidas de fluorescência e reflectância poderiam ser utilizadas para discriminar a efetiva relação marginal entre a deficiência dos metais micronutrientes e o crescimento de mudas de soja em culturas. As conclusões dos autores foram de que os métodos discriminantes são apropriados para determinação de campos de "*stress*" das plantas, adicionalmente às medidas espectrais necessárias, pela redução da análise a um ou dois parâmetros.

Gröger e Gröhsler (2001) empregaram a discriminação quadrática como modelo para diferenciação e avaliação na criação de arenques (*Clupea harengus L.*) na zona de transição entre os mares do Norte e Báltico (Europa), onde se encontravam, temporariamente misturadas, duas populações dos peixes: *spring spawner* e *autumn spawner*. A medida de separação - ICES, comumente utilizada pelos pesquisadores da área, foi substituída pelos autores por outras, baseadas em características da coluna vertebral dos animais. Estas foram sujeitas à avaliação e comparação com a usual. Para tanto, duas técnicas foram utilizados: a linear, tomando variâncias invertidas e a análise discriminante quadrática. Concluíram que a discriminação quadrática forneceu resultados bastante superiores ao outro método implementado.

Ainda no campo experimental, Banowetz et al. (2002) trataram das mudanças ocorridas na composição biológica do solo, avaliando medidas dos efeitos das práticas de utilização, bem como da saúde do mesmo. Embora algumas dessas mudanças possam ser caracterizadas usando procedimentos de isolamento microbiológico, foi estimado que menos que 10% dos microorganismos do solo são propensos à cultura usando as técnicas existentes na literatura. Assim, aproximações alternativas baseadas na análise estrutural dos

componentes do solo foram desenvolvidas para caracterizar as mudanças ocorridas nas comunidades biológicas. Pelo uso supervisionado de classificadores - *FAME* (*Fatty acid methyl ester*) e *LH-PCR* (*length heterogeneity-polymerase chain reaction*) - procedimentos de análise discriminante foram empregados. A comparação entre os resultados fornecidos pela discriminação linear, quadrática e estimação não-paramétrica da densidade, demonstrou que hipóteses minimizadas sobre a distribuição dos dados melhoraram a capacidade da análise *FAME* responder às diferenças para grupos específicos de microorganismos.

Sabe-se que os programas de melhoramento animal necessitam de um constante acompanhamento das características economicamente importantes de raças de animais e/ou linhagem, para que assim possam ser planejados os melhores cruzamentos. Com essa proposta, Pires et al. (2002) submeteram três raças suínas – *Landrace, Large White* e *Duroc* – à avaliação de desempenho por meio 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. O estudo da divergência genética foi feito por meio da análise por variáveis canônicas. 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 100 kg e espessura de toucinho. A raça Large White apresentou uma pequena superioridade em relação à Landrace, e ambas foram bem superiores à Duroc. Os resultados justificaram a utilização das raças Landrace e Large White para a obtenção de fêmeas F1, para um posterior acasalamento com machos Duroc, visando à obtenção de animais híbridos com efeito heterótico expressivo e para haver complementaridade entre as características.

Também utilizando a análise discriminante linear de Fisher, com os objetivos de estabelecer funções de discriminação entre seis espécies de braquiária, verificar a consistência das funções estabelecidas e identificar os grupos de caracteres que mais contribuem na discriminação das espécies, Assis et al. (2002) pesquisaram a correta classificação dos acessos envolvidos nos programas de hibridação para gêneros de *Brachiaria*, constituídos por cerca de 100 espécies. Tal estudo é de grande interesse para o melhoramento genético das espécies forrageiras. Foram analisados 301 acessos, pertencentes a seis diferentes

espécies de braquiária, nos quais foram avaliadas características vegetativas, reprodutivas e de pilosidade. Foram realizadas análises discriminantes para cada um dos três grupos de caracteres morfológicos, sendo estabelecidas funções para as seis espécies. Após a obtenção dos resultados, verificou-se que os caracteres vegetativos e reprodutivos mostraram ser os mais eficientes, enquanto os de pilosidade foram os menos eficientes na classificação e discriminação das espécies.

Martel et al. (2003) apresentam uma aplicação bastante prática, enfocando o uso da análise discriminante linear aplicada a 15 descritores morfológicos na tentativa de caracterizar, morfometricamente, três raças de pupunheiras encontradas ao longo dos rios Amazonas e Solimões, que apresentam grande variabilidade genética ainda não totalmente caracterizada. As três análises em conjunto permitiram uma discriminação das raças, mostrando também que os descritores mais importantes nessa seleção foram: número de espigas, comprimento da ráquis, peso do fruto, espessura das cascas, facilidade para descascar os frutos, peso das cascas, sabor dos frutos, espessura da polpa, distância morfológica dos frutos e peso da semente. Neste trabalho, além da análise discriminante, utilizada numa tentativa de classificar os locais que contêm as raças, outros métodos de classificação foram empregados: análise de agrupamento, que tem como objetivo dividir um grupo original de observações em vários grupos, seguindo algum critério de similaridade ou dissimilaridade, sendo os resultados apresentados em um dendrograma (diagrama em forma de árvore que mostra a subdivisão dos grupos formados, buscando máxima homogeneidade entre os indivíduos no grupo e máxima heterogeneidade entre os grupos); análise de componentes principais, complementando a análise de agrupamento, com objetivo de tentar explicar a estrutura de variância e covariância das variáveis originais, construindo, mediante processo matemático, um conjunto menor de combinações lineares das variáveis originais que preserve a maior parte da informação fornecida por essas variáveis. Como resultados desta análise foram obtidas duas funções discriminantes que conseguiram reter 100% da variância inicial. As análises foram processadas no software *STATISTICA* versão 6.0. As técnicas estatísticas multivariadas mostraram, em conjunto, ser um método eficiente de discriminação das raças *Pará, Putumayo* e *Solimões*.

As pesquisas realizadas por Moshou et al. (2003) utilizaram, como classificadores, a discriminação quadrática e redes neurais (supervisionadas e não supervisionadas), para avaliar as medidas de fluorescência à clorofila para maçãs das espécies *Jonagold* e *Cox* armazenadas sob diferentes condições, que levam ao aparecimento de lanosidades - *mealiness*. As conclusões encontradas referem-se à diferença no nível de fluorescência geral registrado para as duas variedades, bem como à constatação de que as duas técnicas de classificação forneceram resultados semelhantes, sendo suplantadas somente por avançados algoritmos de *data mining*, em testes bastante incipientes, visando automatizar a seleção das frutas.

A disponibilidade de dados em alta dimensionalidade, para aplicações em sensoriamento remoto, abriu novas possibilidades, até então não disponíveis com a utilização de dados tradicionais em baixa dimensionalidade como dados LandSat-TM e SPOT, por exemplo. Dados com resolução espectral muito alta permitem a separação de classes espectralmente muito semelhantes. Com dados em alta dimensionalidade, é possível separar classes que possuem idêntica resposta espectral, isto é, vetores médios iguais, desde que as matrizes de covariâncias difiram suficientemente entre si. Para tanto, Erbert e Haertel (2003) enfatizam a importância e os ótimos resultados da análise discriminante quadrática (ADQ), ao permitir a implementação de um dos classificadores mais utilizados no processo de classificação de imagens digitais em sensoriamento remoto, que é o da Máxima Verossimilhança Gaussiana (MVG).

Visando desenvolver e avaliar um método para discriminação das classes de solos, em uma área no sudoeste do Estado de São Paulo, a partir de suas respostas espectrais, Nanni et al. (2004) fizeram uso de equações discriminantes que foram desenvolvidas para 18 classes de estudo. Os resultados demonstraram que as classes de solos podem ser individualizadas e distinguidas pela análise discriminante, pois esta registrou índices de acerto acima de 80% de determinação da classe de solo avaliada. Sendo assim, concluiu-se, com o uso da análise discriminante, que o método sugerido auxilia na discriminação de classes de solos pela sua reflectância, devido às interações físicas com a energia eletromagnética.

Ainda em 2004, Zandonadi et al. mostraram a aplicação da análise discriminante quadrática para a classificação final de imagens de plantas de milho atacadas pela lagarta elasmo (*Elasmopalpus lignosellus*), complementarmente ao uso de um algoritmo baseado em técnicas de processamento de imagens digitais. Tais técnicas têm sido utilizadas no intuito de melhorar a eficiência dos métodos de controle dessa praga, visando evitar os reflexos de sua atuação na produtividade da cultura de milho. As imagens das culturas são obtidas por meio de sistemas de visão artificial (SVA), que têm sido considerados um dos mais promissores recursos utilizados na agricultura de precisão para aplicação de insumos com base em informação coletada em tempo real.

Os sistemas de visão artificial também têm sido propostos para automação das etapas de classificação e seleção de madeira serrada na indústria madeireira, conforme comprova o trabalho de Khoury Junior et al. (2005). Os autores realizaram análises discriminantes linear e quadrática para classificação de defeitos e madeira isenta de defeitos em imagens digitais de tábuas de eucalipto, utilizando-se as características de percentis de imagens coloridas. As características de percentis do histograma das bandas do vermelho, verde e azul, retiradas de dois tamanhos de blocos de imagens, foram utilizadas para desenvolvimento e teste das funções discriminantes. As características foram analisadas com seus valores originais, escores dos componentes principais e escores das variáveis canônicas, sendo que as funções discriminantes com os escores das variáveis canônicas tenderam a apresentar menores erros de classificação do que com as variáveis originais e com os escores dos componentes principais. As funções discriminantes quadráticas tenderam a exibir erros de classificação global menores do que as funções lineares, porém a função linear apresentou melhor distribuição de erros entre as classes de defeitos. Para a realização das análises utilizou-se o *Software SAS*.

Ainda em 2005, Galvão et al., publicaram um trabalho, também baseado em dados coletados por meio de visão artificial – sensor hiperespectral Hyperion (242 bandas) a bordo do satélite Earth Observing-1 (EO-1) – com o objetivo de contribuir com as pesquisas relativas à utilização de dados hiperespectrais para a discriminação de variedades de cana-de-açúcar, que, segundo os autores, são praticamente inexistentes na literatura. No Brasil,

assim com em outros países produtores de cana-de-açúcar (*Saccharum sp.*), variedades têm sido continuamente desenvolvidas e testadas com os objetivos de aumentar a produtividade, obter uma maior resistência às pragas e doenças e uma melhor adaptação às variações de clima, tipos de solos, técnicas de corte ou manejo. As pesquisas relativas à utilização do sensoriamento remoto em áreas de plantio de cana-de-açúcar têm abordado questões importantes como classificação e mapeamento, manejo e estimativa de produtividade. O uso deste instrumento (ou de similares) possibilita a aquisição de dados com resolução espectral suficiente relacionados com conteúdos de clorofila, água nas folhas e de lignina/celulose, que podem ser parâmetros importantes na diferenciação das variedades de cana. Visando diferenciar as variedades com maior similaridade espectral, a análise discriminante quadrática foi aplicada, com a finalidade de também reduzir a dimensionalidade dos dados e possibilitar a subseqüente classificação da imagem, buscando maximizar a separação dos grupos (variedades). A verificação da eficiência da função discriminante para diferenciar as variedades foi feita com um outro conjunto de *pixels* (vinte por variedade), obtendo-se uma exatidão total de classificação de 87,5%, e ainda, por meio da comparação da verdade de campo com a classificação resultante da análise discriminante, concluiu-se acerca do bom desempenho da função discriminante e dos dados hiperespectrais na discriminação das variedades.

A publicação recente de Carreiras et al. (2006) buscou fazer, por meio do uso de um conjunto de dados multitemporal do sensor *SPOT-4 VEGETATION (VGT)*, uma avaliação da extensão da área de agricultura e pasto na Amazônica Legal Brasileira no ano de 2000. Adicionalmente, os autores realizaram um estudo discriminatório da vegetação local: floresta tropical, cerrado savana e nascentes de rios artificiais e naturais. Para tanto, submeteram os dados coletados a quatro algoritmos de classificação: a análise discriminante quadrática, árvores simples de classificação, árvores probabilísticas de classificação e *k* vizinhos mais próximos. A comparação com os mapas existentes indicou que a cobertura de agricultura e pasto ocorria, a princípio, em áreas previamente ocupadas por floresta tropical primária, seguida de cerrado savana, floresta de transição e outros tipos de vegetação. Os estudos concluíram que a aquisição dos dados, via sensor, é bastante adequada para o estudo discriminatório de cobertura de regiões tropicais e também que, apesar da análise

discriminante quadrática ser extremamente conhecida na literatura relativa aos métodos de classificação utilizando dados remotos, os melhores resultados de classificação foram obtidos pelo algoritmo de árvores probabilísticas de classificação.

2.3 *Softwares* estatísticos que realizam análise discriminante

É quase impossível discutir a aplicação de técnicas multivariadas sem uma discussão do impacto do computador. A ampla aplicação de computadores (primeiro de grande porte e depois computadores pessoais - *PC*) para processar bancos de dados grandes e complexos tem incentivado, significativamente, o uso de métodos estatísticos multivariados. A teoria estatística para técnicas multivariadas foi desenvolvida bem antes do surgimento de computadores, mas essas técnicas permaneceram quase desconhecidas fora da área de estatística teórica até o momento em que o poder computacional tornou-se disponível para executar seus cálculos cada vez mais complexos.

Os avanços tecnológicos contínuos em computação, particularmente em computadores pessoais, têm oferecido, a qualquer pesquisador interessado, rápido acesso a todos os recursos necessários para abordar problemas multivariados de praticamente qualquer tamanho. Mesmo para pesquisadores com forte qualificação quantitativa, a disponibilidade de aplicativos computacionais para análise multivariada tem facilitado a complexa manipulação de matrizes de dados que há muito tempo dificultavam o desenvolvimento de técnicas multivariadas.

Programas estatísticos não são mais primeiramente desenvolvidos para sistemas de grande porte e então adaptados para computadores pessoais; em vez disso, eles agora são inicialmente desenvolvidos para microcomputadores. Talvez a categoria de programas estatísticos de mais rápido crescimento seja a dos aplicativos estatísticos projetados especificamente para tirar proveito da flexibilidade do computador pessoal.

Ressalta-se que os computadores e os *softwares* estatísticos são mais do que um meio de evitar cálculos repetitivos. A sua capacidade para representar graficamente os dados em análise e funcionar em modo interativo com o investigador, faz deles instrumentos essenciais em qualquer etapa do processo de investigação, desde que o pesquisador esteja munido de conhecimentos mínimos que lhe permitam aplicar os métodos adequados à qualidade dos dados observados e a interpretação correta dos resultados.

Dentre os aplicativos mais conhecidos, destacam-se o *SPSS*, *SAS, SYSTAT* ou *BMDP*, que foram criados na década de 80, para *mainframes*, e foram posteriormente adaptados para *PC*. São relacionados, na seqüência, os aplicativos mais comuns, que realizam o procedimento da discriminação:

- *SPSS* (*Statistical Package for Social Sciences*): um dos mais antigos *softwares* na área de estatística existente e utilizado como ferramenta de referência em muitas áreas das Ciências Biológicas e da Saúde;
- *Statistica*: talvez o mais moderno e mais completo aplicativo na área. Muitos pesquisadores consideram o *Statistica* como a melhor ferramenta estatística existente. É muito completo e possui vasto material de apoio, além de *helpfiles* extremamente bem estruturados e escritos.
- *P-STAT*: providencia ferramentas poderosas para administradores, analistas, pesquisadores e estudantes desenvolverem seus trabalhos. Suas aplicações mais comuns são nas análises demográfica e biométrica, controle de qualidade, administração, pesquisa de mercado e muitos outros campos de estudo. É utilizado em instalações acadêmicas, comerciais e governamentais.
- *SAS* (*Statistical Analysis System*): é um sistema moderno que fornece, além dos cálculos estatísticos, ferramentas de análise e gerenciamento dos dados. Conta com alguns específicos produtos adicionais: *SAS/GRAPH, SAS/ETS, SAS/FSP* e *SAS/IMS-DL/I*.

- *SYSTAT* ou *BMDP* (*BIOMEDICAL COMPUTER PROGRAMS*): fornece uma biblioteca com mais de 40 flexíveis, modernas e compatíveis rotinas com um conjunto de instruções básico. Muito poderoso em seus cálculos precisos, possui uma interface de fácil manipulação.
- *MINITAB, SIGMASTAT*: realizam a discriminação linear e quadrática.
- *SAEG, EVIEWS, S-PLUS* e *CAPTURE E 2CAPTURE*: realizam somente a discriminação linear.
- *GENES*: software destinado à análise e processamento de dados binários, geralmente obtidos de estudos de Genética Molecular por meio de diferentes modelos biométricos. Conta com procedimentos uni e multivariados, realizando processamento com ênfase na estimação de parâmetros genéticos, porém não realiza a discriminação quadrática (http://www.ufv.br/dbg/genes/genes.htm).
- *MZEF (Michael Zhang's Exon Finder)*: programa baseado em algoritmo preditivo que utiliza análise discriminante quadrática para reconhecimento de padrões estatísticos multivariados, para identificar uma das mais freqüentes classes de *exões* internos do DNA (ZHANG, 2006).

3 DESENVOLVIMENTO METODOLÓGICO

3.1 Fundamentação Teórica

3.1.1 Metodologia estatística para determinação da função discriminante

A análise discriminante constitui-se em uma técnica multivariada interessante para pesquisas com o objetivo de verificar a pertinência a grupos, seja de indivíduos, animais, empresas, produtos, conjuntos de elementos, ou qualquer outra situação que possa ser avaliada em uma série de variáveis independentes.

Para a aplicação da análise discriminante é necessário que os grupos para os quais cada elemento amostral possa ser classificado sejam predefinidos, ou seja, conhecidos *a priori*, considerando-se todas as suas características observadas. Este conhecimento permite a elaboração de uma função matemática chamada de regra de classificação ou discriminação, utilizada para alocar novos elementos amostrais nos grupos já existentes. Portanto, o número de grupos já é conhecido previamente, mas a regra de

classificação é elaborada utilizando-se procedimentos que, em geral, vão além do uso de distâncias matemáticas (REIS, 1997).

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 e WICHERN, 1998). Dadas duas populações Π_1 e Π_2 de observações multivariadas, a idéia de Fisher foi transformar estas observações multivariadas em observações univariadas, 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.

Esquematicamente, é possível apresentar a definição da regra de discriminação conforme a Figura 3.1.

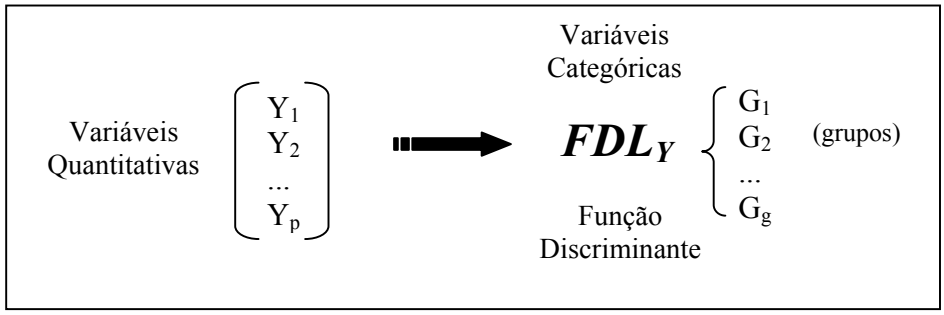

Figura 3.1 - Representação esquemática da determinação da Função Discriminante

A apresentação de dados observacionais, obtidos a partir de experimentos agronômicos, pode ser descrita, genericamente, por meio da estrutura básica de estudos com planejamento longitudinal (ANDERSON, 2003; MORRISON, 2005) que corresponde a uma matriz bidimensional de respostas.

Considerando-se *g* grupos e *p* variáveis, as respostas observadas podem ser organizadas de acordo com a Tabela 3.1.

Tabela 3.1 - Estrutura genérica de valores observados em um experimento

Grupo Experimental	Unidade Experimental (Parcela)	Variável Resposta				Vetor Resposta
		V_1 (Y_1)	V_2 (Y_2)	…	V_p (Y_p)	
1	1	y_{111}	y_{112}	…	y_{11p}	$\underset{\sim}{y'}_{11}$
1	2	y_{121}	y_{122}	…	y_{12p}	$\underset{\sim}{y'}_{12}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	n_1	y_{1n_11}	y_{1n_12}	…	y_{1n_1p}	$\underset{\sim}{y'}_{1n_1}$
2	1	y_{211}	y_{212}	…	y_{21p}	$\underset{\sim}{y'}_{21}$
2	2	y_{221}	y_{222}	…	y_{22p}	$\underset{\sim}{y'}_{22}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	n_2	y_{2n_21}	y_{2n_22}	…	y_{2n_2p}	$\underset{\sim}{y'}_{2n_2}$
g	1	y_{g11}	y_{g12}	…	y_{g1p}	$\underset{\sim}{y'}_{g1}$
g	2	y_{g21}	y_{g22}	…	y_{g2p}	$\underset{\sim}{y'}_{g2}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
g	n_g	y_{gn_g1}	y_{gn_g2}	…	y_{gn_gp}	$\underset{\sim}{y'}_{gng}$

A observação y_{ijk} é expressa pelos índices: $i = 1,...,g$ (grupo), $j=1,...,n_i$ (parcela) e $k = 1,...,p$ (variável).

Considerando-se n_1, n_2, ..., n_g amostras aleatórias de g populações Π_1, Π_2, ..., Π_g, respectivamente, tem-se as estatísticas descritivas relativas às medidas de posição e variabilidade dos dados relacionadas na Tabela 3.2, sendo:

$\underset{\sim ij}{y}$ = vetor de observações das p variáveis para a parcela $j = 1, 2, ..., n_i$) do grupo $i = 1, 2, ..., g$.

$\underset{\sim i}{\bar{y}}$ = vetor de médias amostrais para o grupo $i = 1, 2, ..., g$.

S_i = matriz de covariâncias amostral para o grupo $i = 1, 2, ..., g$.

Tabela 3.2 - Estatísticas descritivas para amostras de grupos independentes

Grupo	Resposta	Estatística Descritiva
Grupo 1	$\underset{\sim}{y}_{11}, \underset{\sim}{y}_{12}, \ldots, \underset{\sim}{y}_{1n1}$	$\overline{\underset{\sim}{y}}_{1} = \frac{1}{n_1}\sum_{j=1}^{n1} \underset{\sim}{y}_{1j}$ $S_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n1} (\underset{\sim}{y}_{1j} - \overline{\underset{\sim}{y}}_{1})(\underset{\sim}{y}_{1j} - \overline{\underset{\sim}{y}}_{1})'$
Grupo 2	$\underset{\sim}{y}_{21}, \underset{\sim}{y}_{22}, \ldots, \underset{\sim}{y}_{2n2}$	$\overline{\underset{\sim}{y}}_{2} = \frac{1}{n_2}\sum_{j=1}^{n2} \underset{\sim}{y}_{2j}$ $S_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n2} (\underset{\sim}{y}_{2j} - \overline{\underset{\sim}{y}}_{2})(\underset{\sim}{y}_{2j} - \overline{\underset{\sim}{y}}_{2})'$
...	...	...
Grupo g	$\underset{\sim}{y}_{g1}, \underset{\sim}{y}_{g2}, \ldots, \underset{\sim}{y}_{gng}$	$\overline{\underset{\sim}{y}}_{g} = \frac{1}{n_g}\sum_{j=1}^{ng} \underset{\sim}{y}_{gj}$ $S_g = \frac{1}{n_g - 1}\sum_{j=1}^{ng} (\underset{\sim}{y}_{gj} - \overline{\underset{\sim}{y}}_{g})(\underset{\sim}{y}_{gj} - \overline{\underset{\sim}{y}}_{g})'$

Quando as populações são distintas a acurácia da discriminação é bastante alta, porém, quanto mais as populações se aproximam, mas difícil passa a ser o procedimento de alocação correta dos indivíduos às populações. Uma alternativa interessante consiste em verificar se as populações envolvidas no estudo são distintas, hipótese equivalente a inferir sobre a igualdade de vetores de médias das g populações, ou seja, avaliar $H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \ldots = \underset{\sim}{\mu}_g = \underset{\sim}{\mu}$.

Para verificar se esta hipótese pode ser aceita ou não, alguns pressupostos relativos à estrutura dos dados observados devem estar presentes nos vetores de respostas das amostras. Ou seja, as observações devem ser independentes, o conjunto das p variáveis dependentes deve seguir uma distribuição normal multivariada (isto é, qualquer combinação linear das variáveis dependentes deve seguir uma distribuição normal) e as matrizes de covariâncias devem ser iguais (homogêneas) para todos os grupos de tratamento.

3.1.1.1 Independência

A mais básica, porém mais séria violação de uma suposição ocorre quando há falta de independência entre as observações. Existem diversas situações experimentais, bem como não-experimentais, nas quais essa suposição pode ser facilmente violada. Por exemplo, um efeito temporalmente ordenado (correlação serial) pode ocorrer se forem tomadas medidas ao longo do tempo, mesmo a partir de diferentes respondentes. Outro problema comum seria reunir informação em grupos, de modo que uma experiência em comum faça com que um subconjunto de indivíduos tivesse respostas que de algum modo fossem correlacionadas. Finalmente, efeitos estranhos e não medidos podem afetar os resultados, criando dependência entre os respondentes.

Apesar de não existirem testes com uma certeza absoluta de detectar todas as formas de dependência, é necessário proceder-se a uma exploração de todos os efeitos possíveis e corrigi-los quando encontrados. Assim, quando se suspeita que exista dependência, deve-se usar um nível de significância mais baixo.

3.1.1.2 Normalidade

A normalidade multivariada considera que o efeito conjunto de duas variáveis é distribuído normalmente. Esta suposição é inerente à maioria das técnicas multivariadas, porém não existe teste direto para a normalidade multivariada, recorrendo-se, então, ao teste da normalidade univariada de cada variável. Pela teoria, o fato de se demonstrar que todas as distribuições univariadas são normais não implica necessariamente que o vetor aleatório $\underset{\sim}{Y}$ tem distribuição normal multivariada (ANDERSON, 2003). No entanto, na prática, quando a distribuição univariada é normal, a chance de se estar com um vetor normal multivariado é muito grande.

Segundo Johnson e Wichern (1998, 2002), os testes de normalidade univariada, para o caso multivariado, têm como principal objetivo verificar a normalidade de

distribuições marginais. Dentre eles tem-se o exame do histograma e das caudas da distribuição como também a verificação por meio de gráficos. Embora úteis, as verificações gráficas são limitadas, uma vez que têm utilidade apenas nos casos em que o ajuste de uma determinada distribuição teórica a um conjunto de dados é graficamente óbvio, ou ainda quando existem dados muito discrepantes em relação à distribuição proposta.

Existem alguns testes estatísticos para a verificação da normalidade, muito utilizados por pesquisadores, como o de *Kolmogorov-Smirnov, Lilliefors e Shapiro – Wilk*, além dos mais comuns: assimetria e curtose.

3.1.1.2.1 Teste de Kolmogorov-Smirnov (KS)

Segundo Conover (1980), Siegel (1981) e Campos (1983) este teste verifica o grau de concordância entre a distribuição de um conjunto de valores amostrais (observados) e determinada distribuição teórica específica, determinando-se os valores amostrais que podem vir de uma população com aquela distribuição teórica, ou seja, este teste tenta especificar a distribuição de freqüência acumulada que ocorreria sob a teórica e a compara com a distribuição de freqüência acumulada observada, sendo a teórica representada no que se esperaria sob H_0, determinando o ponto em que as distribuições, teórica e observada, apresentam maiores divergências, indicando se a distribuição amostral (observada) nessa diferença máxima pode ser atribuída ao acaso.

Admitindo que $F(x_0) = \int_{-\infty}^{X_0} f(x)dx = P(X \leq x_0)$, sendo *f(x)* a função de densidade de probabilidade da distribuição normal, as hipóteses serão: H_0: $F = F_0$ e H_α: $F \neq F_0$ para ao menos um valor de *X*. Rejeita-se H_0 se a estatística do Teste, *T*, for maior ou igual a *d* (os valores de "*d*" são tabelados). A estatística do teste é: $T = sup \mid F(X) - S(X) \mid$. Pode ainda a hipótese alternativa ser H_α: $F < F_0$ ou H_α: $F > F_0$.

O teste de Kolmogorov-Smirnov usa toda a informação presente no grupo de dados (DANIEL, 1995).

3.1.1.2.2 Teste de Lilliefors

Segundo Conover (1980) o teste de Lilliefors amplia o uso da média e da variância estimada por meio dos dados amostrais, sendo: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ *e* $S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$, obtendo a variável reduzida $Z_i = \frac{X_i - \bar{x}}{S}, i = 1,2,3,...,n$.

A estrutura do teste é análoga ao de Kolmogorov-Smirnov, que é definida como a distância máxima entre a função de distribuição de X_i e a função distribuição normal, mas calculado a partir de Z_i, ao invés da variável original. As hipóteses são:

H_0: A amostra aleatória tem distribuição normal com média e variância desconhecida

H_α: A função de distribuição de X não é normal.

A aceitação de H_0 não indica que a distribuição seja normal, mas apenas é uma razoável apresentação da distribuição desconhecida.

A estatística do teste é semelhante à de Kolmogorov-Smirnov, pois $T = sup\ |F(X) - S(X)\ |$, sendo $F(X)$ a função da distribuição normal e $S(X)$ é obtida da amostra normalizada.

3.1.1.2.3 Teste de Shapiro-Wilk

É um teste quantitativo para normalidade, e mede a relação de linearidade entre os dados e os escores normais. Shapiro-Wilk é o coeficiente de correlação

entre os dados de X_i e os valores dos dados de Z_i (*Transformation*). A estatística *W* testa se uma amostra aleatória vem de uma distribuição normal específica. Valores pequenos de *W* evidenciam a normalidade.

A estatística do teste, *W*, é obtida por: $W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\left(\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2\right)}$.

Neste teste, se *W* é significante, a hipótese de que a respectiva distribuição é normal deve ser rejeitada. Apesar do Teste de Shapiro-Wilk ser usado apenas para n ≤ 50 alguns autores atestam que pode ser usado para n > 50 (CONOVER, 1980).

3.1.1.3 Homogeneidade

A suposição da igualdade de matrizes de covariâncias é extremamente importante para o teste de igualdade de vetores de médias em populações normais multivariadas. Alguns estudiosos sobre a robustez das propriedades do procedimento da *MANOVA* afirmam que a desigualdade das matrizes é uma séria violação com comprometimentos severos aos resultados do teste.

Segundo Morrison (2005), o teste estatístico para a verificação da homogeneidade é realizado por meio do cálculo do coeficiente *M*, de acordo com a expressão 3.1.

$$M = (n - g)\ln|S| - \sum_{i=1}^{g} (n_i - 1)\ln|S_i| \qquad (3.1)$$

sendo:

$n_i =$ número de unidades experimentais da i-ésima população

$S = \dfrac{\sum_{i=1}^{g} (n_i - 1)S_i}{n - g}$ matriz de covariâncias amostral conjunta (*pool*)

Box (1949) mostrou que se o fator de escala C^{-1} (3.2) é introduzido, a quantidade MC^{-1} é, aproximadamente, distribuída como uma variável com distribuição *Qui-Quadrado* com $\frac{1}{2}(g-1)p(p+1)$ graus de liberdade.

$$C^{-1} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(g-1)}\left(\sum_{i=1}^{g} \frac{1}{(n_i - 1)} - \frac{1}{(n-g)}\right) \tag{3.2}$$

3.1.2 Teste de hipótese da igualdade dos vetores de médias

Para o caso univariado a hipótese nula, relativa ao teste estatístico de que as médias das *g* populações são idênticas, pode ser expressa por: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$.

A hipótese nula para o caso multivariado, à semelhança do univariado, refere-se à igualdade entre os centróides dos grupos populacionais, representados pelos vetores de médias populacionais, de acordo com a expressão (3.3).

$$H_0 : \underset{\sim 1}{\mu} = \underset{\sim 2}{\mu} = \cdots = \underset{\sim g}{\mu} = \underset{\sim}{\mu} \quad com \quad \underset{\sim i}{\mu} = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \cdots \\ \mu_{ig} \end{bmatrix} \; i = 1,2,..., g \tag{3.3}$$

A hipótese alternativa refere-se ao fato que, pelo menos um dos vetores das médias das populações, seja diferente, e pode ser expressa por $H_1 : \underset{\sim i}{\mu} \neq \underset{\sim i'}{\mu} \; com \; i \neq i'$.

3.1.2.1 Teste de vetores de médias de duas populações independentes

Segundo Johnson e Wichern (1998, 2002) o teste T^2 *de Hotelling* foi desenvolvido para testar se dois vetores de médias são iguais, seguindo a mesma analogia de procedimento do teste *t* de *Student* univariado. O teste T^2 permite comparar a resposta do vetor média da população 1 com a do vetor média da população 2, com tamanhos amostrais n_1 e n_2, diferentes entre si. A partir dos valores amostrais calculam-se estatísticas $\overline{X}_i$ e S_i, que são os estimadores imparciais dos parâmetros populacionais $\underset{\sim i}{\mu}$ e Σ_i, respectivamente, vetor de médias e matriz de covariâncias da população *i* (i = 1, 2).

Os pressupostos básicos para a realização do teste de hipóteses são: os dados de ambas as amostras, de tamanhos n_1 e n_2, devem ser provenientes de populações com distribuição normal multivariada e as respectivas matrizes de covariâncias são homogêneas, ou seja, iguais (populações homocedásticas).

Isto é, para o teste da hipótese $H_0 : \underset{\sim 1}{\mu} - \underset{\sim 2}{\mu} = \underset{\sim 0}{\delta}$, considera-se que:

$$\underset{\sim 1}{\overline{Y}} \sim N\left(\underset{\sim}{\mu_1}, \frac{1}{n_1}\Sigma_1\right), \; \underset{\sim 2}{\overline{Y}} \sim N\left(\underset{\sim}{\mu_2}, \frac{1}{n_2}\Sigma_2\right) \text{ e } \Sigma_1 = \Sigma_2 = \Sigma.$$

Então: $\mathrm{E}\left(\underset{\sim 1}{\overline{Y}} - \underset{\sim 2}{\overline{Y}}\right) = \mathrm{E}\left(\underset{\sim 1}{\overline{Y}}\right) - \mathrm{E}\left(\underset{\sim 2}{\overline{Y}}\right) = \underset{\sim 1}{\mu} - \underset{\sim 2}{\mu} = \underset{\sim 0}{\delta}$ e

$$\mathrm{Var}\left(\underset{\sim 1}{\overline{Y}} - \underset{\sim 2}{\overline{Y}}\right) = \mathrm{Var}\left(\underset{\sim 1}{\overline{Y}}\right) + \mathrm{Var}\left(\underset{\sim 2}{\overline{Y}}\right) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma .$$

Como $\Sigma_1 = \Sigma_2$, tem-se S_p como a matriz de covariâncias amostral conjunta (3.4), a qual estima a matriz comum de covariâncias populacional Σ.

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{3.4}$$

Considerando, sem perda de qualidade, $\underset{\sim}{\delta}_0 = \underset{\sim}{0}$, sob a veracidade de H_0, a estatística do teste de hipóteses da igualdade dos vetores de médias, é dada pela expressão 3.5.

$$T^2 = \left[\underset{\sim}{\overline{Y}}_1 - \underset{\sim}{\overline{Y}}_2\right]\left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)S_p\right]^{-1}\left[\underset{\sim}{\overline{Y}}_1 - \underset{\sim}{\overline{Y}}_2\right] \sim T^2_{(p, n_1+n_2-p-1)} \tag{3.5}$$

Ou, equivalentemente, $\frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2_{(p, n_1+n_2-p-1)} = F_{(p, n1+n2-p-1)}$, como distribuição F de *Snedecor* com p e (n_1+n_2-p-1) graus de liberdade.

A regra de decisão para o teste de hipóteses é a habitual, ou seja, se $F_{calc} \geq F_{(\alpha, p, n_1+n_2-p-1)}$, rejeita-se H_0. Caso contrário, a hipótese H_0 não deve ser rejeitada.

3.1.2.2 Teste de vetores de médias de *g > 2* populações independentes

À semelhança do modelo univariado, Johnson e Wichern (1998, 2002) consideram que os g grupos possuem matriz comum de covariâncias Σ. A técnica da Análise da Variância, usada para comparar vetores de médias de g populações é construída a partir do modelo de observação multivariado (vetor p-dimensional), $\underset{\sim}{Y}_{ij} = \underset{\sim}{\mu} + \underset{\sim}{\tau}_i + \underset{\sim}{\varepsilon}_{ij}$, $j = 1, 2, ..., n_i$ e $i = 1, 2, ..., g$, sendo $\underset{\sim}{\varepsilon}_{ij}$ vetor aleatório de erros independentes, distribuído como

multinormal $N_p = \left(\underset{\sim}{0}, \Sigma\right)$. O parâmetro $\underset{\sim}{\mu}$ é o vetor de médias comum a todas as observações $\underset{\sim}{\tau_i}$, efeito do tratamento *i* (grupo) com $\sum_{i=1}^{g} n_i \underset{\sim}{\tau}_i = \underset{\sim}{0}$.

De acordo com o modelo multivariado, cada componente do vetor de observação $\underset{\sim}{Y}_{ij}$ satisfaz o modelo univariado $Y_{ijk} = \mu_k + \tau_{ik} + \varepsilon_{ijk}$.

Os componentes do erro casual $\underset{\sim}{\varepsilon}_{ij}$ do vetor $\underset{\sim}{Y}_{ij}$ são não correlacionados e a matriz de covariâncias Σ é a mesma para todos os grupos.

Sendo assim, consideram-se as matrizes *W* e *B* (3.6), construídas a partir dos dados observados, chamadas de matrizes de soma de quadrados e produtos cruzados dentro (*Within*) dos grupos e entre (*Between*) as médias dos grupos.

$$W_{(pxp)} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(\underset{\sim}{Y}_{ij} - \overline{\underset{\sim}{Y}}_{i} \right)\left(\underset{\sim}{Y}_{ij} - \overline{\underset{\sim}{Y}}_{i} \right)' \qquad (3.6)$$

$$B_{(pxp)} = \sum_{i=1}^{g} n_i \left(\overline{\underset{\sim}{Y}}_{i} - \overline{\underset{\sim}{Y}}\right)\left(\overline{\underset{\sim}{Y}}_{i} - \overline{\underset{\sim}{Y}}\right)'$$

Nas expressões, $\underset{\sim}{Y}_{ij}$ é o vetor de observações do elemento amostral *j* que pertence à população *i*; $\overline{\underset{\sim}{Y}}_{i}$ é o vetor amostral de médias da população *i*; $\overline{\underset{\sim}{Y}}$ o vetor amostral geral de médias, considerando-se todas as *n* observações conjuntamente e n_i o número de parcelas amostradas na população *i*, *i*= 1, 2, 3, ..., *g*, $\sum_{i=1}^{g} n_i = n$. A Tabela 3.3 expressa a formulação genérica de uma Análise de Variância Multivariada (*MANOVA*).

Tabela 3.3 - Tabela da Análise de Variância Multivariada - MANOVA

Fonte de Variação	Matriz Soma de Quadrados e Produtos	Graus de Liberdade
Tratamento	$\boldsymbol{B}$	$g - 1$
Resíduo (Erro)	$\boldsymbol{W}$	$\sum_{i=1}^{g} n_i - g = n - g$
Total	$\boldsymbol{T = B + W}$	$\sum_{i=1}^{g} n_i - 1 = n - 1$

3.1.2.3 Avaliação da significância estatística dos resultados (testes multivariados)

Para a avaliação do teste de hipóteses da igualdade dos vetores de médias, existem quatro critérios estatísticos (MORRISON, 2005) que podem ser utilizados no procedimento inferencial, sendo os mais usuais os determinados pelo princípio da *União-Intersecção de Roy* e da *Razão de Verossimilhança*.

No presente estudo serão considerados os quatro especificados como: Lâmbda de *Wilks* (Razão de Verossimilhança), Traço de *Pillai*, Traço de *Lawley-Hotelling* e Maior Raiz de *Roy* (Princípio da *União-Intersecção de Roy*).

Para o estabelecimento dos testes estatísticos para a hipótese H_0 serão consideradas as seguintes definições: $\theta_v = \frac{\lambda_v}{1+\lambda_v}, v = 1, \ldots, s$; $s = \min(g-1, p)$; $m_1 = \frac{|g-p-1|-1}{2}$ e $m_2 = \frac{n-g-p-1}{2}$, sendo λ_v a v-ésima raiz característica de $W^{-1}B$ ou raiz de $|B - \lambda W| = 0$.

3.1.2.3.1 Lâmbda de *Wilks* (Razão de Verossimilhança)

Para o teste da igualdade dos vetores de médias, são utilizadas as variâncias generalizadas associadas às matrizes de variação entre e dentro de grupos. A estatística do teste é dada pela razão das variâncias generalizadas $\Lambda = \frac{\det\{W\}}{\det\{B+W\}} = \prod_{i=1}^{s} \frac{1}{1+\lambda_z} = \prod_{i=1}^{s} (1-\theta_z)$.

Quando o tamanho da amostra é grande, utiliza-se a aproximação de Bartlett (1938), conforme a expressão 3.7.

$$-\left(n-1-\frac{p+g}{2}\right)\ln\Lambda = -\left(n-1-\frac{p+g}{2}\right)\ln\left(\frac{\det\{W\}}{\det\{B+W\}}\right) \sim \chi^2_{p(g-1)} \tag{3.7}$$

A regra de decisão do teste de hipótese da igualdade dos vetores de médias é a habitual, isto é, rejeita-se H_0, no nível de significância α, se a desigualdade 3.8 for satisfeita.

$$-\left(n-1-\frac{p+g}{2}\right)\ln(\Lambda) > \chi^2_{p(g-1)} \tag{3.8}$$

O valor $\chi^2_{p(g-1)}$ corresponde ao quantil de ordem $100(1-\alpha)\%$ da distribuição *qui-quadrado* com p(g-1) graus de liberdade.

A distribuição probabilística de Λ, para algumas situações especiais é descrita na Tabela 3.4.

Tabela 3.4 - Distribuição de Lâmbda de *Wilks*

Número de Variáveis	Número de Grupos	Distribuição Exata para dados normais multivariados
p = 1	g ≥ 2	$\left(\frac{\sum_{i=1}^{g} n_i - g}{g - 1}\right)\left(\frac{1 - \Lambda}{\Lambda}\right) \sim F_{(g-1,\,n-g)}$
p = 2	g ≥ 2	$\left(\frac{\sum_{i=1}^{g} n_i - g - 1}{g - 1}\right)\left(\frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}}\right) \sim F_{[2(g-1),\,2(n-g-1)]}$
p ≥ 2	g = 2	$\left(\frac{\sum_{i=1}^{g} n_i - p - 1}{p}\right)\left(\frac{1 - \Lambda}{\Lambda}\right) \sim F_{(p,\,n-p-1)}$
p ≥ 1	g = 3	$\left(\frac{\sum_{i=1}^{g} n_i - p - 2}{p}\right)\left(\frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}}\right) \sim F_{[2p,\,2(n-g-2)]}$

3.1.2.3.2 Traço de *Pillai*

Outro critério para o teste de H_0, trata-se do Traço de *Pillai*, sendo sua estatística expressa pela igualdade 3.9, sendo θ_z a z-ésima raiz característica de $B(B+W)^{-1}$.

$$PI = trB\left(B+W\right)^{-1} = \sum_{v=1}^{s} \frac{\lambda_v}{1+\lambda_v} = \sum_{z=1}^{s} \theta_z \tag{3.9}$$

A aproximação *F* para a estatística de *Pillai* é indicada pela expressão 3.10.

$$\frac{2m_2+s+1}{2m_1+s+1}\cdot\frac{PI}{s-PI}\sim F_{[s(2m_1+s+1),s(2m_2+s+1)]} \tag{3.10}$$

3.1.2.3.3 Traço de *Lawley-Hotelling*

O Traço de *Lawley-Hotelling* é a estatística determinada pela igualdade 3.11, cuja aproximação *F* é dada pela expressão 3.12, sendo λ_v a raiz característica de $W^{-1}B$.

$$LH = tr\left(W^{-1}B\right) = \sum_{v=1}^{s}\lambda_v \tag{3.11}$$

$$\frac{2\left(sm_2+1\right)}{s^2\left(2m_1+s+1\right)}LH\sim F_{[s(2m_1+s+1),2(sm_2+1)]} \tag{3.12}$$

3.1.2.3.4 Maior Raiz de Roy (Princípio da União-Intersecção de Roy)

No princípio da união-intersecção de *Roy*, a hipótese H_0 é testada por meio da seguinte estatística: $\theta_1 = max\theta_v$, que corresponde à maior raiz característica $B(B+W)^{-1}$. A distribuição probabilística de θ_1 é aproximada a *F* de *Snedecor* com os parâmetros *s*, m_1 e m_2, ou $\lambda_1 = \max\lambda_z$, que corresponde à maior raiz característica de $W^{-1}B$.

Se $s = 1$, então $F = \frac{1-\theta_1}{\theta_1}\cdot\frac{m_2+1}{m_1+1}$, com $(2m_1 + 2)$ e $(2m_2 + 2)$ graus de liberdade ou $F = \frac{1}{\lambda_1}\cdot\frac{m_2+1}{m_1+1}$.

3.2 Função Discriminante Linear de Fisher

Fisher (1936) introduziu a idéia da construção de funções discriminantes a partir de combinações lineares das variáveis originais, algo similar ao que é feito na técnica de análise de componentes principais e análise fatorial.

Segundo Johnson e Wichern (1998, 2002), a função discriminante linear consiste, basicamente, em separar duas classes de objetos ou fixar um novo objeto em uma das duas classes.

Conforme já abordado, o objetivo de Fisher, ao criar essa regra de reconhecimento de padrões e classificação, foi transformar observações multivariadas em univariadas, tal que as populações Π_1 e Π_2 fossem separadas em relação às médias das duas populações tanto quanto possível.

Sendo $\underset{\sim}{\mu}_1$ a média das observações da população Π_1 e $\underset{\sim}{\mu}_2$ a média da população Π_2 e, considerando a matriz de covariâncias, Σ, comum para ambas as populações, Fisher selecionou uma combinação linear que maximizasse a razão entre a "soma de quadrados das distâncias entre as médias das populações e Y ", bem como a "variância de Y " (equação 2.1) de acordo com 3.13.

$$\underset{\sim}{C}'\underset{\sim}{X} = Var(\underset{\sim}{C}'\underset{\sim}{X}) = \frac{\left[\underset{\sim}{C}'\left(\underset{\sim 1}{\mu} - \underset{\sim 2}{\mu}\right)\right]^2}{\underset{\sim}{C}'\Sigma\,\underset{\sim}{C}} \tag{3.13}$$

A razão expressa em 3.13 é maximizada por $C = \Sigma^{-1}(\mu_1 - \mu_2)$, tendo-se, então, a igualdade 3.14, que é conhecida como função discriminante linear populacional.

$$FDL\,(\underset{\sim}{Y}) = \underset{\sim}{C}'\,\underset{\sim}{X} = \left(\underset{\sim 1}{\mu} - \underset{\sim 2}{\mu} \right)' \Sigma^{-1}\, \underset{\sim}{X} \tag{3.14}$$

Tendo em vista que os parâmetros $\underset{\sim 1}{\mu}$, $\underset{\sim 2}{\mu}$ e Σ não são conhecidos, utilizam-se seus estimadores, ou seja, $\underset{\sim 1}{\overline{y}}$, $\underset{\sim 2}{\overline{y}}$, sendo a matriz de covariâncias conjunta (estimada) dada por 3.15, respectivamente.

$$S_{pool} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{3.15}$$

Dessa forma, a função discriminante linear amostral fica determinada por 3.16.

$$FDL\left(\underset{\sim}{Y}\right) = \left(\underset{\sim 1}{\overline{y}} - \underset{\sim 2}{\overline{y}} \right)' S_{pool}^{-1}\, \underset{\sim}{Y} \tag{3.16}$$

A estimativa do ponto médio amostral é dada pela expressão 3.17.

$$\hat{m} = \frac{1}{2}\left[\left(\underset{\sim 1}{\overline{y}} - \underset{\sim 2}{\overline{y}} \right)' S_{pool}^{-1} \left(\underset{\sim 1}{\overline{y}} + \underset{\sim 2}{\overline{y}} \right) \right] \tag{3.17}$$

Obtém-se, desta forma, a seguinte regra de classificação:

Alocar $\underset{\sim}{y}$ a Π_1 se $\left(\underset{\sim 1}{\overline{y}} - \underset{\sim 2}{\overline{y}} \right)' S_{pool}^{-1}\, \underset{\sim}{Y} \geq \hat{m}$, caso contrário, alocá-lo a Π_2.

Ainda segundo Johnson e Wichern (1998, 2002), o procedimento pode ser generalizado para o caso de g populações $\Pi_1,...,\Pi_g$, sendo $g \geq 2$, em que está associada a cada população Π_i uma distribuição normal multivariada e, ainda, supondo-se a igualdade das matrizes de covariâncias, as funções ou escores discriminantes lineares são definidos de acordo com a expressão 3.18.

$$d_i(\underset{\sim}{y}) = \underset{\sim}{\mu}_i^{'} \Sigma^{-1} \underset{\sim}{y} - \frac{1}{2} \underset{\sim}{\mu}_i^{'} \Sigma^{-1} \underset{\sim}{\mu}_i + \ln p_i \qquad (3.18)$$

sendo:

$d_i(\underset{\sim}{y})$ = escore de classificação do i-ésimo grupo;

Σ^{-1} = inversa da matriz comum de covariâncias;

$\underset{\sim}{\mu}_i$ = vetor de médias do i-ésimo grupo;

$\underset{\sim}{y}$ = vetor de observações do indivíduo que se deseja classificar;

p_i = probabilidade "a priori" de que um indivíduo pertença à população i.

Um estimador $\hat{d}_i(\underset{\sim}{y})$, do escore discriminante linear $d_i(\underset{\sim}{y})$, pode ser estabelecido a partir dos estimadores de mínimos quadrados de $\underset{\sim}{\mu}_i$ e $\sum$ baseado no vetor de médias amostrais e na estimativa conjunta – pool – da matriz $\sum$, conforme indicado em 3.19.

$$S_{pool} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + ... + (n_g - 1)S_g}{n_1 + n_2 + ... + n_g - g} \qquad (3.19)$$

Nestas condições o estimador pode ser descrito como a igualdade 3.20.

$$\hat{d}_i(\underset{\sim}{y}) = \underset{\sim i}{\overline{y}}' S_{pool}^{-1} \underset{\sim}{y} - \frac{1}{2} \underset{\sim i}{\overline{y}}' S_{pool}^{-1} \underset{\sim i}{\overline{y}} + \ln p_i \qquad (3.20)$$

Conseqüentemente, um novo indivíduo será classificado como pertencente à população para a qual se tem o maior escore de classificação, ou seja, o indivíduo desconhecido ($\underset{\sim}{y}$) pertencerá à população Π_i se, e somente se, a igualdade 3.21 for verificada.

$$\hat{d}_i(\underset{\sim}{y}) = máx\left[\hat{d}_1\left(\underset{\sim}{y}\right), \hat{d}_2\left(\underset{\sim}{y}\right), ..., \hat{d}_g\left(\underset{\sim}{y}\right)\right] \qquad (3.21)$$

A técnica para discriminar indivíduos pertencentes a diferentes grupos é considerada "a melhor" quando permite a minimização de erros de classificação incorreta. No entanto, este fato só é verdadeiro quando, segundo Reis (1997), se verificam os pressupostos seguintes:

1. os grupos deverão ser retirados de populações que seguem uma distribuição normal multivariada para as p variáveis discriminantes;
2. dentro dos grupos a variabilidade deverá ser idêntica, isto é, as matrizes de covariâncias deverão ser iguais para todos os grupos - homocedasticia;
3. o número de observações em cada grupo é pelo menos dois ($n_i \geq 2$);
4. o número de variáveis discriminantes (p) poderá ser qualquer, desde que verifique a seguinte condição: ser no máximo o número total de observações (n) menos dois ($0 < p < n-2$);
5. nenhuma das variáveis discriminantes poderá ser combinação linear das restantes.

3.3 Função Discriminante Quadrática

Considerando o caso de duas populações, Π_1 e Π_2, com distribuição normal, mas com diferentes matrizes de covariâncias, $\Sigma_1 \neq \Sigma_2$, à semelhança do problema de Behrens-Fisher (MEHTA e SRINIVASAN, 1970), a regra de classificação é apresentada como: atribua o indivíduo à população Π_1 se a expressão 3.22 for satisfeita; atribua para Π_2, caso contrário (3.23).

$$R_1 : -\frac{1}{2}\underset{\sim}{y}'(\Sigma_1^{-1} - \Sigma_2^{-1})\underset{\sim}{y} + (\underset{\sim}{\mu}'_1\Sigma_1^{-1} - \underset{\sim}{\mu}'_2\Sigma_2^{-1})\underset{\sim}{y} - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (3.22)$$

$$R_2 : -\frac{1}{2}\underset{\sim}{y}'(\Sigma_1^{-1} - \Sigma_2^{-1})\underset{\sim}{y} + (\underset{\sim}{\mu}'_1\Sigma_1^{-1} - \underset{\sim}{\mu}'_2\Sigma_2^{-1})\underset{\sim}{y} - k < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (3.23)$$

sendo:

$$k = \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}(\underset{\sim}{\mu}'_1\Sigma_1^{-1}\underset{\sim 1}{\mu} - \underset{\sim}{\mu}'_2\Sigma_2^{-1}\underset{\sim 2}{\mu}),$$

$c(1|2)$ = custo de má classificação de um indivíduo de Π_2 ser classificado na população Π_1

$c(2|1)$ = custo de má classificação de um indivíduo de Π_1 ser classificado na população Π_2

As regiões de classificação são definidas por funções quadráticas de $\underset{\sim}{y}$. O termo quadrático $-\frac{1}{2}\underset{\sim}{y}'(\Sigma_1^{-1} - \Sigma_2^{-1})\underset{\sim}{y}$ desaparece quando $\Sigma_1 = \Sigma_2$ e as regiões definidas pela redução deste são dadas por expressões lineares, como verificado na equação 3.18.

Na prática, a regra de classificação é implementada pela substituição dos parâmetros populacionais $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$, Σ_1 e Σ_2 por seus estimadores $\underset{\sim}{\bar{y}}_1$, $\underset{\sim}{\bar{y}}_2$, S_1 e S_2, respectivamente. Assim, $\underset{\sim}{y}$ é atribuído para Π_1, se a equação 3.24 for satisfeita.

$$-\frac{1}{2}\underset{\sim}{y}'(S_1^{-1}-S_2^{-1})\underset{\sim}{y}+(\underset{\sim}{\bar{y}}'_1 S_1^{-1}-\underset{\sim}{\bar{y}}'_2 S_2^{-1})\underset{\sim}{y}-k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (3.24)$$

Caso contrário $\underset{\sim}{y}$ é atribuído para Π_2.

Generalizando a regra de classificação para g populações normais $\Pi_1,...,\Pi_g$, $g \geq 2$, cuja função densidade de probabilidade pode ser expressa como $f_i(\underset{\sim}{y}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_i|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\underset{\sim}{y}-\underset{\sim i}{\mu})'\Sigma_i^{-1}(\underset{\sim}{y}-\underset{\sim i}{\mu})\right]$, $i = 1,2,...,g$, o logaritmo natural desta função conduz a uma forma quadrática na definição das regiões de classificação $R_1,...,R_g$. Dessa maneira, $\ln f_i(\underset{\sim}{y}) = -\frac{p}{2}\ln(2\pi)-\frac{1}{2}\ln|\Sigma_i|-\frac{1}{2}(\underset{\sim}{y}-\underset{\sim i}{\mu})'\Sigma_i^{-1}(\underset{\sim}{y}-\underset{\sim i}{\mu})$.

Assim sendo, um escore de discriminação quadrática para a *i-ésima* população, considerando os custos de má classificação iguais, é definido conforme a expressão 3.25.

$$d_i^Q(\underset{\sim}{y}) = \ln p_i - \frac{p}{2}\ln(2\pi)-\frac{1}{2}\ln|\Sigma_i|-\frac{1}{2}(\underset{\sim}{y}-\underset{\sim i}{\mu})'\Sigma_i^{-1}(\underset{\sim}{y}-\underset{\sim i}{\mu})+, \; i = 1,2,...,g \qquad (3.25)$$

O escore quadrático $d_i^Q(\underset{\sim}{y})$ é composto pela variância generalizada $|\Sigma_i|$, pela probabilidade "a priori" p_i, e pelo quadrado da distância de $\underset{\sim}{y}$ à média populacional

$\underset{\sim}{\mu_i}$. É importante observar, entretanto, que as diferenças da função de distância, orientação e extensão da constante elipsoidal, devem ser usadas para cada população.

Portanto, utilizando os escores quadráticos, a regra de classificação é definida desta forma: atribua $\underset{\sim}{y}$ para Π_k, se a expressão 3.26 for satisfeita.

$$d_k^Q(\underset{\sim}{y}) = máx\left[d_1^Q\left(\underset{\sim}{y}\right), d_2^Q\left(\underset{\sim}{y}\right),..., d_g^Q\left(\underset{\sim}{y}\right)\right], i = 1,2,..., g \qquad (3.26)$$

Na prática, o escore de discriminação quadrático estimado, $\hat{d}_k^Q(\underset{\sim}{y})$, é dado por 3.27.

$$\hat{d}_i^Q(\underset{\sim}{y}) = -\frac{1}{2}\ln|S_i| - \frac{1}{2}(\underset{\sim}{y} - \underset{\sim i}{\overline{y}})' S_i^{-1}(\underset{\sim}{y} - \underset{\sim i}{\overline{y}}) + \ln p_i \qquad (3.27)$$

Dessa maneira, a regra de classificação para várias populações normais com matrizes de covariâncias diferentes é: confira $\underset{\sim}{y}$ para Π_k, de acordo com a veracidade da expressão 3.28.

$$\hat{d}_k^Q(\underset{\sim}{y}) = máx\left[\hat{d}_1^Q\left(\underset{\sim}{y}\right), \hat{d}_2^Q\left(\underset{\sim}{y}\right),..., \hat{d}_g^Q\left(\underset{\sim}{y}\right)\right], \ i = 1,2,..., g \qquad (3.28)$$

3.4 Probabilidades de classificações incorretas

Após a construção da função discriminante, é necessário avaliar a sua qualidade. Para cada elemento amostrado das populações avaliadas, calcula-se o escore numérico da função discriminante construída e a análise destes escores permitirá que se faça

uma avaliação da qualidade da função em termos de erros de classificação e capacidade de discriminação. Se a função é adequada, espera-se que os escores de uma população sejam bem diferenciados dos escores de outra população.

Para o caso de duas populações normais multivariadas e independentes, por exemplo, pode ser realizada a comparação das médias dos escores das duas populações por meio do teste de Hotelling (JOBSON, 1996) para dois grupos independentes. Uma outra análise importante está relacionada com os erros de classificação:

- ERRO 1: o elemento amostral pertence à Π_1, mas a regra de classificação o classifica como sendo proveniente de Π_2.
- ERRO 2: o elemento amostral pertence à Π_2, mas a regra de classificação o classifica como sendo proveniente de Π_1.

Denotando as probabilidades de ocorrência destes erros respectivamente por P(Erro1) = p(2/1) e P(Erro2) = p(1/2), quanto menor forem estas probabilidades, melhor será a função de discriminação.

Uma melhor visualização do erro pode ser dada com o cálculo da taxa aparente de erro e apresentação da tabela genérica das freqüências dos erros de classificação, conforme a Tabela 3.5, que é a comparação entre acertos e erros na classificação dos novos elementos provenientes das *g* populações, sendo:

n_1 Número total de indivíduos em Π_1

n_2 Número total de indivíduos em Π_2

n_g Número total de indivíduos em Π_g

n_{11} Número de indivíduos de Π_1 classificados corretamente como de Π_1

n_{21} Número de indivíduos de Π_1 classificados incorretamente como de Π_2

n_{g1} Número de indivíduos de Π_1 classificados incorretamente como de Π_g

n_{12} Número de indivíduos de Π_2 classificados incorretamente como de Π_1

n_{22} Número de indivíduos de Π_2 classificados corretamente como de Π_2

n_{g2} Número de indivíduos de Π_2 classificados incorretamente como de Π_g

n_{1g} Número de indivíduos de Π_g classificados incorretamente como de Π_1

n_{2g} Número de indivíduos de Π_g classificados incorretamente como de Π_2

n_{gg} Número de indivíduos de Π_g classificados corretamente como de Π_g

n_{cor1} Número de indivíduos classificados corretamente em Π_1

n_{cor2} Número de indivíduos classificados corretamente em Π_2

n_{corg} Número de indivíduos classificados corretamente em Π_g

$\%n_1$ Porcentagem correspondente ao número de indivíduos classificados corretamente em Π_1

$\%n_2$ Porcentagem correspondente ao número de indivíduos classificados corretamente em Π_2

$\%n_g$ Porcentagem correspondente ao número de indivíduos classificados corretamente em Π_g

Tabela 3.5 – Modelo da tabela genérica das freqüências dos erros de classificação

População classificada pela regra	**População de origem**			
	Π_1	Π_2	**...**	Π_g
Π_1	n_{11}	n_{12}	...	n_{1g}
Π_2	n_{21}	n_{22}	...	n_{2g}
...	...	...	...	
Π_g	n_{g1}	n_{g2}	...	n_{gg}
***N* Total**	n_1	n_2	...	n_g
***N* Corretos**	n_{cor1}	n_{cor2}	...	n_{corg}
% Correta	$\%n_1$	$\%n_2$	...	$\%n_g$

Segundo Mingoti (2005) existem vários procedimentos para se estimar as probabilidades de classificações incorretas: o método da ressubstituição, o método da colocação de elementos à parte para classificação e o método de Lachenbruch.

Considerando-se duas populações, o método da ressubstituição prevê que os escores de cada elemento amostral observado das populações 1 e 2 sejam calculados, sendo a regra de discriminação utilizada para classificar os $n = n_1 + n_2$ elementos da amostra conjunta. Quando a função discriminante é de boa qualidade, espera-se que ela apresente uma grande porcentagem de acerto na classificação dos elementos amostrais em relação à população a que de fato pertencem. Portanto, neste método, os mesmos elementos amostrais participam da regra de classificação e da estimação dos erros de classificação. Este procedimento de estimação do erro aparente de classificação é consistente, mas viciado (JOHNSON e WICHERN, 2002) e tende a subestimar os verdadeiros valores das probabilidades para elementos que não pertencem à amostra conjunta utilizada para a construção da regra de discriminação, isto é, novos elementos amostrais. No entanto, pode servir como uma etapa inicial de avaliação, pois se os valores das probabilidades de erros forem muito elevados, é sinal de que a regra de classificação deve ser reformulada. O vício deste procedimento tende a zero quando as quantidades amostrais n_1 e n_2 são grandes.

No método da colocação de elementos à parte para classificação, também conhecido por *Holdout method*, a amostra conjunta $n = n_1 + n_2$ de elementos é repartida em duas partes, uma que vai servir para a construção da regra de discriminação (amostra de treinamento) e outra que vai ser utilizada para a estimação das probabilidades de classificação incorretas (amostra de validação). Inicialmente, selecionam-se, aleatoriamente, alguns indivíduos das amostras das populações 1 e 2, deixando-os à parte da amostra original de $n_1 + n_2$ elementos. Para cada um destes elementos, sabe-se a qual população ele pertence e, portanto, eles servirão para testar a função discriminante construída a partir dos elementos amostrais restantes. A regra de discriminação estimada é utilizada para classificar os elementos que foram colocados à parte inicialmente, e as proporções de classificações incorretas são calculadas da mesma forma como a descrita no método da ressubstituição. As estimativas das probabilidades de classificações incorretas obtidas por este procedimento não são viciadas. No entanto, a desvantagem do método é a redução do tamanho da amostra original para a estimação da regra de discriminação, o que poderá diminuir acentuadamente a confiabilidade da regra de classificação construída, se as amostras não forem grandes. Para que este método seja eficiente, é recomendável que se deixe, "à parte", de 25 a 50% dos

elementos da amostra original (JOHNSON e WICHERN, 2002), logo não pode ser empregado em amostras pequenas. Este método é melhor que o da ressubstituição para amostras grandes.

O método de Lachenbruch, também conhecido como de validação cruzada ou *pseudo-jackknife*, consiste nos seguintes passos:

- Passo 1: retira-se um vetor de observações da amostra conjunta e utilizam-se os (n_1 + n_2 - 1) elementos amostrais restantes para construir a função de discriminação;
- Passo 2: utiliza-se a regra de discriminação construída no passo 1 para classificar o elemento que ficou à parte da construção da regra de discriminação, verificando se a regra conseguiu acertar na sua real procedência ou não;
- Passo 3: retorna-se o elemento amostral que foi retirado no passo 1 à amostra original e retira-se outro elemento amostral diferente do primeiro. Os passos 1 e 2 são repetidos.

Os passos 1, 2 e 3 devem ser repetidos para todos os (n_1 + n_2) elementos da amostra conjunta. As estimativas deste método são aproximadamente não viciadas e melhores que o método da ressubstituição para populações normais e não normais.

4 PROGRAMA COMPUTACIONAL *DISCRIMINANTE*

4.1 Linguagem de programação *PHP*

A metodologia estatística relativa à análise discriminante para classificação de grupos foi implementada em um programa computacional desenvolvido na linguagem de programação *PHP* - *Hipertext Preprocessor*. Trata-se de uma linguagem denominada "do lado do servidor", fácil de utilizar e com algumas vantagens como sua gratuidade, independência de plataforma, rapidez e segurança (ALVAREZ, 2006). É independente de plataforma, visto que existe um módulo de *PHP* para quase qualquer servidor *Web*. Isto faz com que qualquer sistema possa ser compatível com a linguagem já que permite levar o site desenvolvido em *PHP* de um sistema a outro sem praticamente trabalho algum. Permite programar *scripts* que se incrustam dentro do código *HTML* - *HyperText Markup Language*, que significa *Linguagem de Formatação de Hipertexto*, ou seja, linguagem utilizada para produzir páginas na *Web,* que é o código interpretado pelos navegadores, como o *Internet Explorer, Netscape, Mozilla Firefox* e outros.

Uma linguagem “do lado do servidor” é aquela que se executa no servidor *Web* antes da página ser enviada ao usuário, por meio da Internet. As páginas que se executam no servidor podem realizar acessos a bases de dados, conexões em rede e outras tarefas para criar a página final que será vista pelo usuário. Este somente recebe uma página com o código *HTML* resultante da execução do script *PHP*. Como a página resultante contém unicamente código *HTML*, é compatível com todos os navegadores.

Foi criada originalmente em 1994 por Rasmus Lerdorf (2006), mas como o *PHP* está desenvolvido em política de código aberto, ao longo de sua história teve muitas contribuições de outros desenvolvedores. Atualmente *PHP* se encontra em sua versão 4, atualizada para cobrir as necessidades das aplicações *Web* atuais. Está preparada para realizar muitos tipos de aplicações *Web* graças à extensa livraria de funções com a qual está dotada, que suporta desde cálculos matemáticos complexos até tratamento de conexões de rede.

4.2 Manual do Usuário

4.2.1 Entrada de dados

A entrada de dados no *Software DISCRIMINANTE* é realizada por meio da leitura de arquivos de dados construídos no aplicativo *Microsoft Excel*, seguindo o modelo dos exemplos especificados nos Anexos 1 e 2. As planilhas devem trazer somente os números relativos às observações por indivíduos, sendo que nas colunas estarão representadas as variáveis e nas linhas as mensurações para cada indivíduo. Após a introdução dos valores o arquivo deverá ser gravado utilizando-se a opção “Salvar Como”. Quando a janela “Salvar Como” é apresentada, deve-se colocar um nome sugestivo para o arquivo e na opção “Salvar como Tipo” deve-se escolher, obrigatoriamente, na caixa de seleção, a opção “CSV (separado por vírgulas)”, de acordo com o visualizado na Figura 5.1. Após este procedimento, o *Software DISCRIMINANTE* estará apto para fazer a leitura do arquivo.

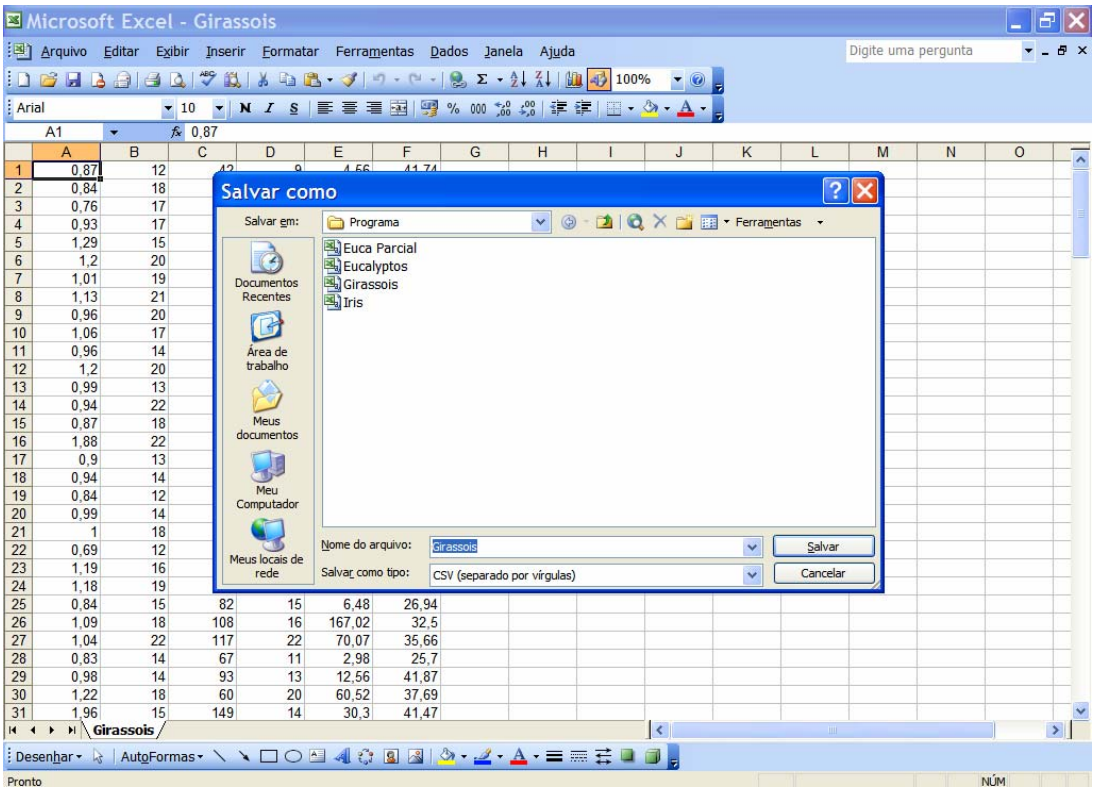

Figura 5.1 - Janela "Salvar Como" do *Microsoft Excel*

4.2.2 Acesso ao *Software*

Como verificado, uma característica marcante da linguagem de programação *PHP* é a independência de plataforma, podendo ser executada em qualquer ambiente. Dessa maneira, o *Software DISCRIMINANTE* é executado diretamente na Internet, que neste caso trata-se da página da Faculdade de Ciências Agronômicas - Campus de Botucatu (FCA), acessada mediante a digitação, na barra de endereços do Navegador, da *URL:* *http://www.fca.unesp.br/*. Estando nesta página, deve-se procurar o *link INTRANET*, conforme observado na Figura 5.2.

Figura 5.2 - Página inicial da FCA com o destaque para o *link INTRANET*

Quando o *link INTRANET* é acionado, uma outra tela é exibida onde se deve escolher a opção *SOFTWARES* - Figura 5.3. No momento que a opção é escolhida, a tela apresentada na Figura 5.4 relaciona os *softwares* disponíveis, entre os quais o programa *DISCRIMINANTE*.

Para que a execução do *DISCRIMINANTE* seja iniciada, basta apenas dar um clique com o *mouse* sobre a palavra e imediatamente as ações são transferidas para a página inicial do programa - Figura 5.5, que traz uma breve explicação a respeito do seu conteúdo, passando-se para a tela seguinte ao clicar o botão *OK*.

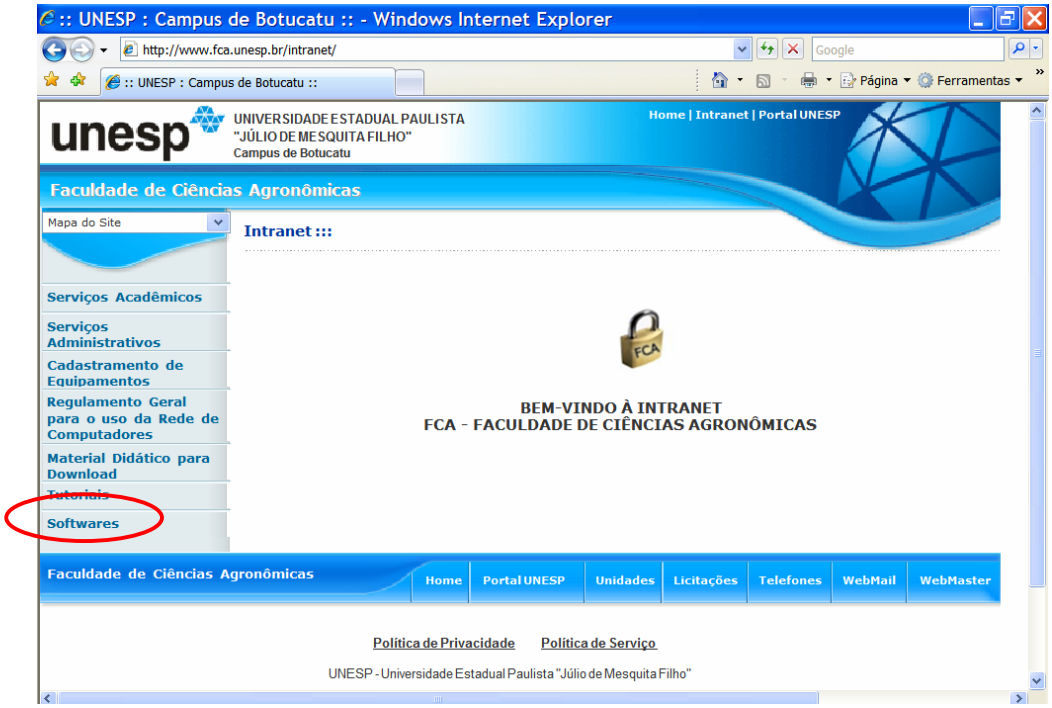

Figura 5.3 - Página *INTRANET* com destaque para a opção *SOFTWARES*

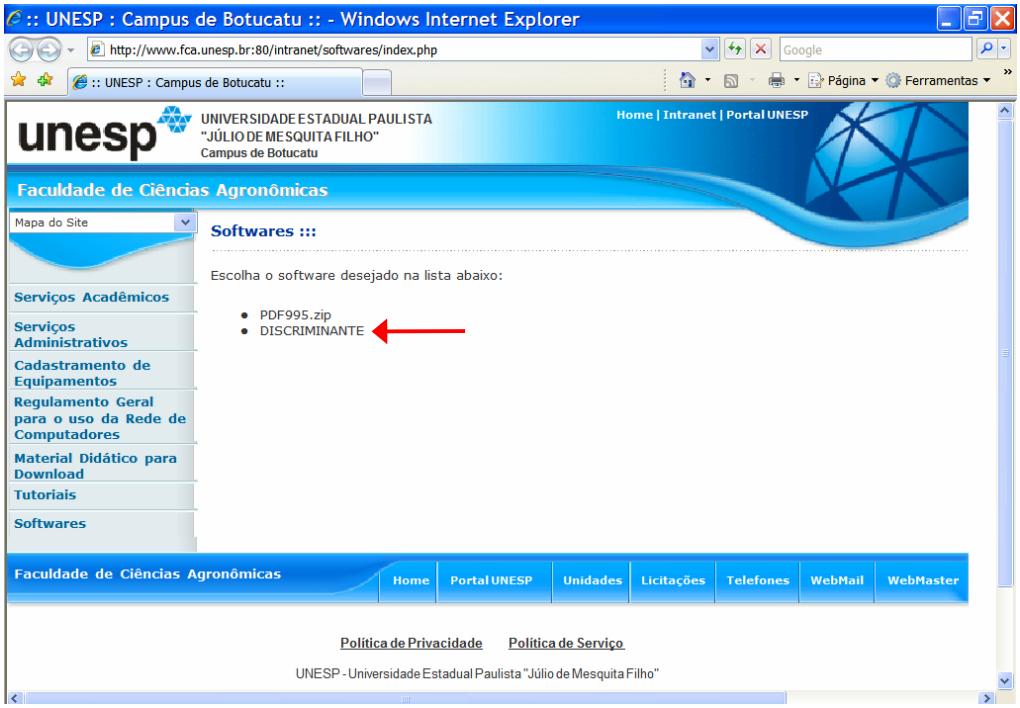

Figura 5.4 - Página *SOFTWARES* com destaque para o *link DISCRIMINANTE*

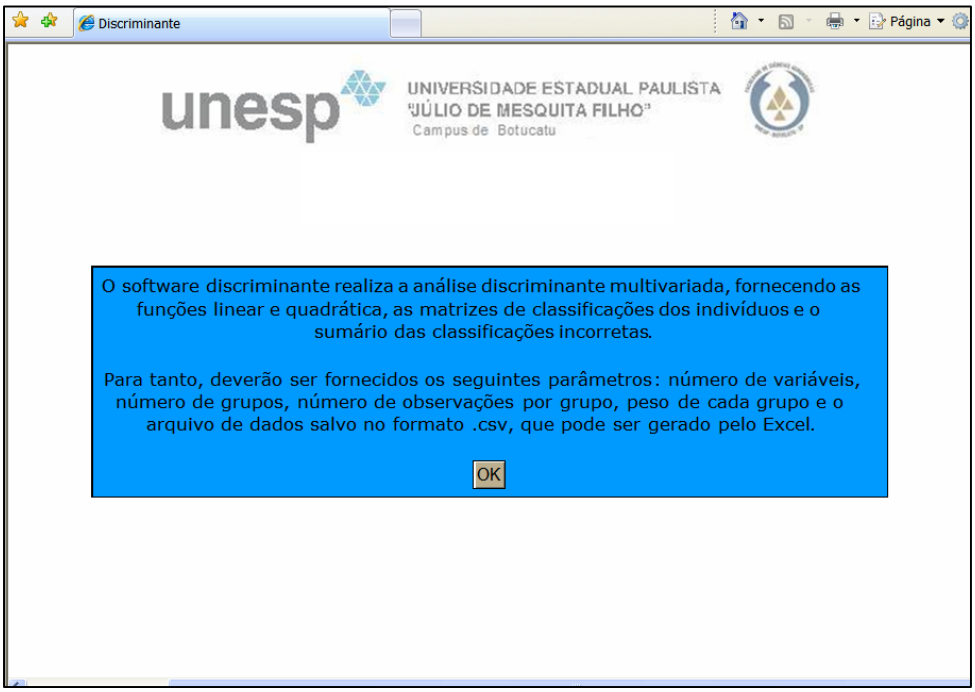

Figura 5.5 - Página inicial do *Software DISCRIMINANTE*

4.2.3 Entrada de parâmetros

A Figura 5.6 apresenta a tela que é visualizada pelo usuário para a digitação dos parâmetros, bem como a seleção do arquivo de dados. Em seguida, estes itens serão enviados ao servidor que, remotamente, realizará o processamento. Observa-se que, ao colocar o cursor sobre os campos a serem digitados, o *software* fornece uma pequena ajuda sobre as características do campo em questão.

As informações iniciais são introduzidas nesta página começando pelo nível de significância - α - para os testes de homogeneidade e normalidade, que é selecionado entre 5 opções em uma caixa de seleção. Também devem ser informados, nesta tela inicial, a quantidade de características observadas no conjunto de dados - variáveis, bem como o número de grupos envolvidos no procedimento classificatório. Tais escalares irão definir as

dimensões das matrizes envolvidas nos cálculos, assim como o total de escores lineares e quadráticos a serem calculados.

A navegação entre os campos é feita utilizando-se a tecla *TAB*. As teclas *ENTER* e setas não produzem resultados. Os campos também são consistidos, tais que não são permitidas digitações de caracteres diferentes de números.

Teoricamente, não haveria 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 da *PHP*, a alocação de memória é realizada de maneira flutuante. Talvez o grande limitante neste aspecto, seja a própria característica do *Software* ser executado remotamente, pois existem restrições de tempo de processamento nos serviços de hospedagem de páginas, que podem ser contornados pela alteração do atributo de tempo de processamento, se for julgado necessário.

Figura 5.6 - Página inicial do *Software DISCRIMINANTE*

Após a escolha dos parâmetros introdutórios, os números de observações por grupos devem ser informados, assim como os pesos de cada grupo, que correspondem às probabilidades de má classificação individuais. Estes valores são escalares inteiros e positivos, distribuídos proporcionalmente entre os grupos, que se não forem informados, serão assumidos iguais - Figura 5.7.

Em seguida, o nome do arquivo de dados deverá ser informado ou selecionado dentre as pastas do dispositivo de arquivamento utilizado pelo usuário.

O *software* proporciona a facilidade de submeter um novo indivíduo à classificação pelas funções - linear e quadrática - que serão definidas, bastando que a caixa de seleção "Novo Indivíduo" seja ativada por meio de um clique com o *mouse*. Quando esta opção é escolhida, abrem-se campos para a digitação das características deste novo indivíduo.

Figura 5.7 - Página de entrada dos parâmetros, nome do arquivo de dados e características do novo indivíduo

4.2.4 Processamento

Para que o processamento seja efetuado é necessário apenas um "clique" com o *mouse* no botão PROCESSAR -  ; em poucos instantes uma nova tela (página) é exibida trazendo os resultados dos procedimentos classificatórios quadrático e linear.

Os cálculos efetuados na fase de processamento são os seguintes:

- Determinação de vetores e matrizes:
  - Cálculo dos vetores $\underset{\sim i}{\bar{y}}$, vetores de médias amostrais para os grupos i = 1, 2, ..., g;
  - Definição das matrizes de covariâncias para os grupos i=1, 2,.., g;
- Construção da matriz de variação conjunta S e verificação da homogeneidade entre as matrizes de covariâncias, por meio da realização do teste de Bartlet;
  - O teste envolve uma pesquisa de valores críticos na Tabela do χ^2, que é realizada automaticamente pelo *Software*;
  - Uma mensagem é mostrada na tela de saída informando qual a melhor opção de classificação, baseada nos resultados deste Teste. Mesmo assim, as duas funções, linear e quadrática, são determinadas;
- Verificação da normalidade dos dados envolvidos, por meio da realização do teste de Kolmogorov-Smirnov;

  - O teste envolve uma pesquisa de valores nas tabelas de área sob a curva normal padronizada e valores críticos para o teste de Kolmogorov-Smirnov para parâmetros estimados, realizadas automaticamente pelo *Software*;
  - O teste é realizado separadamente por variável dentro de cada grupo, por esse motivo as respostas sobre a normalidade das variáveis são apresentadas individualmente em um quadro sinótico. Mesmo não havendo normalidade, o processamento dá prosseguimento, cabendo ao pesquisador ações para contornar tal situação;
- Determinação dos escores de classificação lineares e quadráticos;
- Substituições dos dados originais nas funções classificatórias para montagem das matrizes de classificações dos indivíduos e cálculo das respectivas taxas percentuais de classificação, bem como construção do sumário das classificações incorretas;
- Substituição das características do novo indivíduo nas duas funções para apresentação do grupo classificado em ambos os casos.

4.2.5 Saída dos resultados

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

- Vetor de médias de cada grupo.
- Matrizes de covariâncias - S_i.

- Matriz de variação conjunta - *S*.
- Mensagem sobre a adequacidade da função.
- Quadro sinótico da normalidade.
- Matriz de classificações dos indivíduos pela função discriminante quadrática e respectivas taxas percentuais de classificação.
- Escores de classificação quadráticos para os grupos.
- Sumário das Classificações Incorretas - Função Quadrática.
- Classificação do Novo Indivíduo pela Função Quadrática (somente se esta opção é escolhida pelo usuário).
- Matriz de classificações dos indivíduos pela função discriminante linear e respectivas taxas percentuais de classificação.
- Escores de classificação lineares para os grupos.
- Sumário das Classificações Incorretas - Função Linear.
- Classificação do Novo Indivíduo pela Função Linear (somente se esta opção é escolhida pelo usuário).

A Figura 5.8 mostra a saída parcial, enquanto a Figura 5.9 dá ênfase à mensagem relacionada à homogeneidade das matrizes e indicação do procedimento classificatório adequado. Na Figura 5.10 está destacado o item relacionado à classificação do novo indivíduo. Os resultados, na íntegra, são apresentados nos Apêndices 1 a 4.

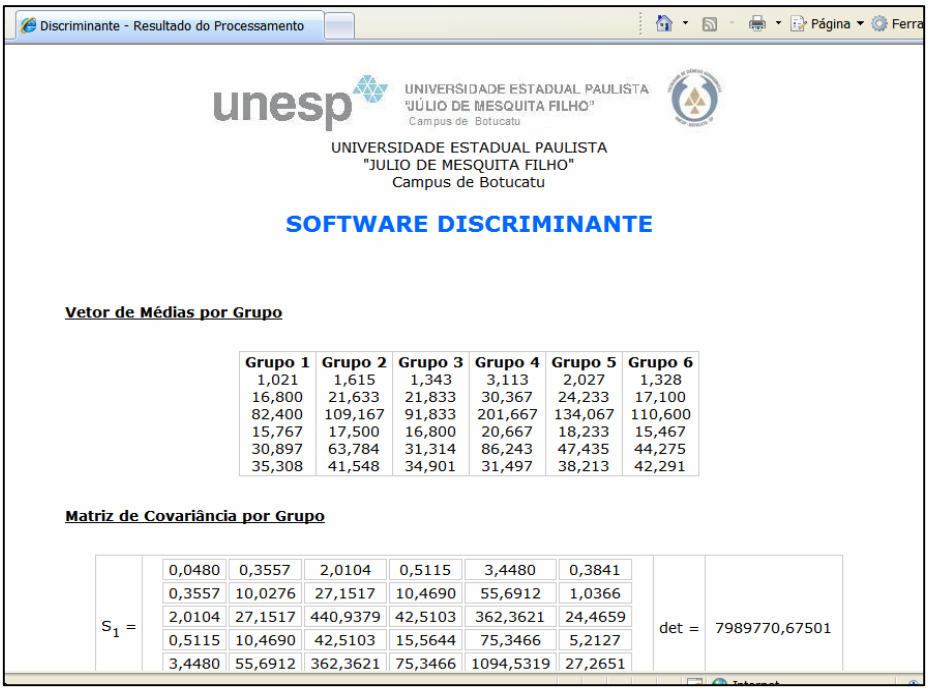

Figura 5.8 - Página dos resultados do *Software DISCRIMINANTE*

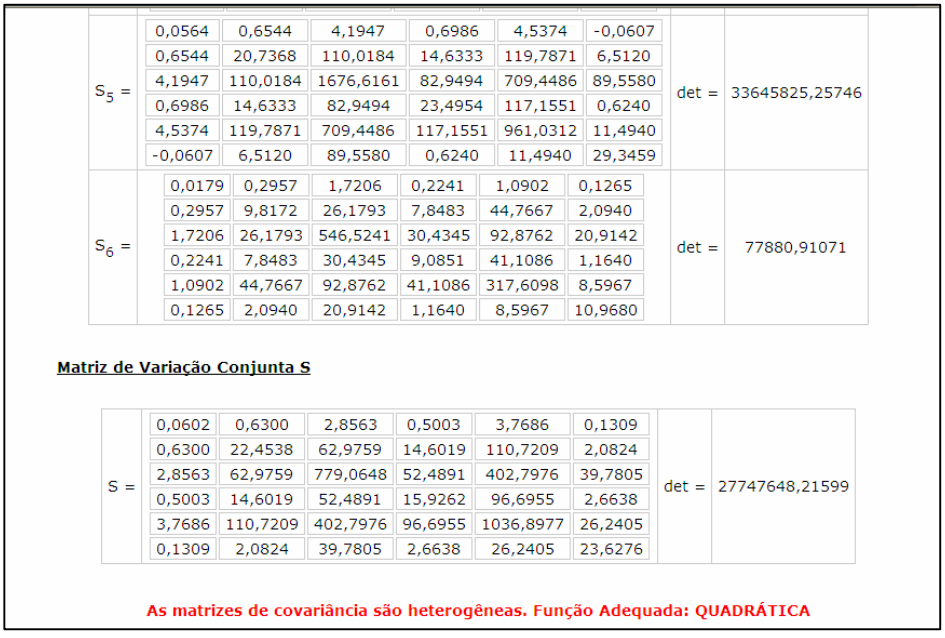

Figura 5.9 - Página dos resultados do *Software DISCRIMINANTE* - destaque para a mensagem em vermelho

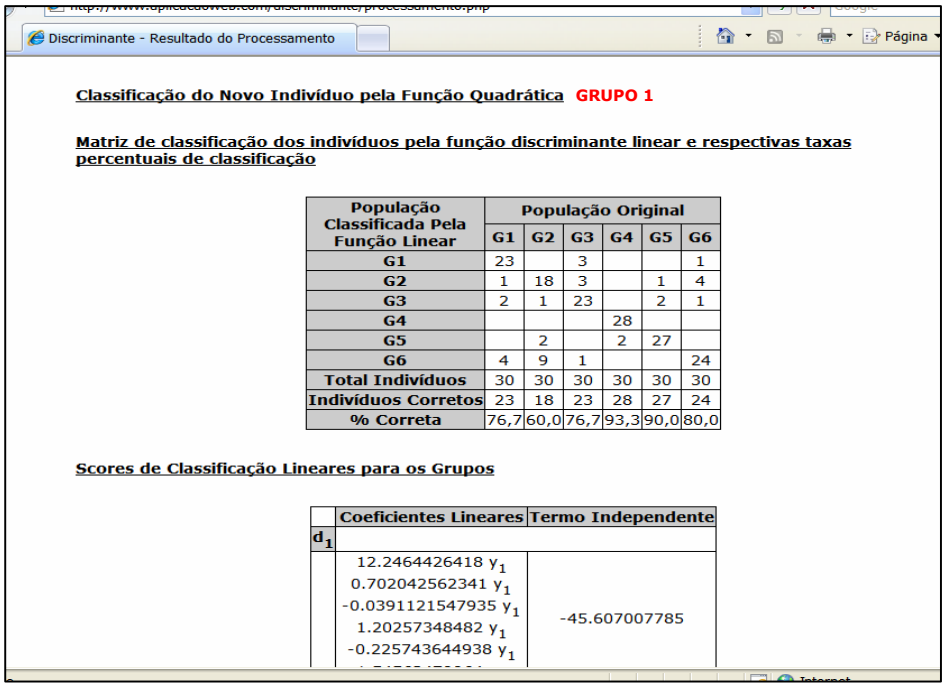

Figura 5.10 - Resultados do *Software DISCRIMINANTE* - destaque para a classificação do novo indivíduo

Para a impressão dos resultados basta apenas selecionar a opção Imprimir no Menu Arquivo do navegador.

Esquematicamente, o *software DISCRIMINANTE* pode ser representado por meio do diagrama de blocos da Figura 5.11. Trata-se de uma visão simplificada da realidade, mas que colabora com a utilização do programa, visto tratar-se de uma forma bastante acessível de, rapidamente, identificar todas as precedências e interdependências entre os processos, valorizando o objetivo de disponibilizar um programa com características simples de manuseio para os pesquisadores das áreas aplicadas.

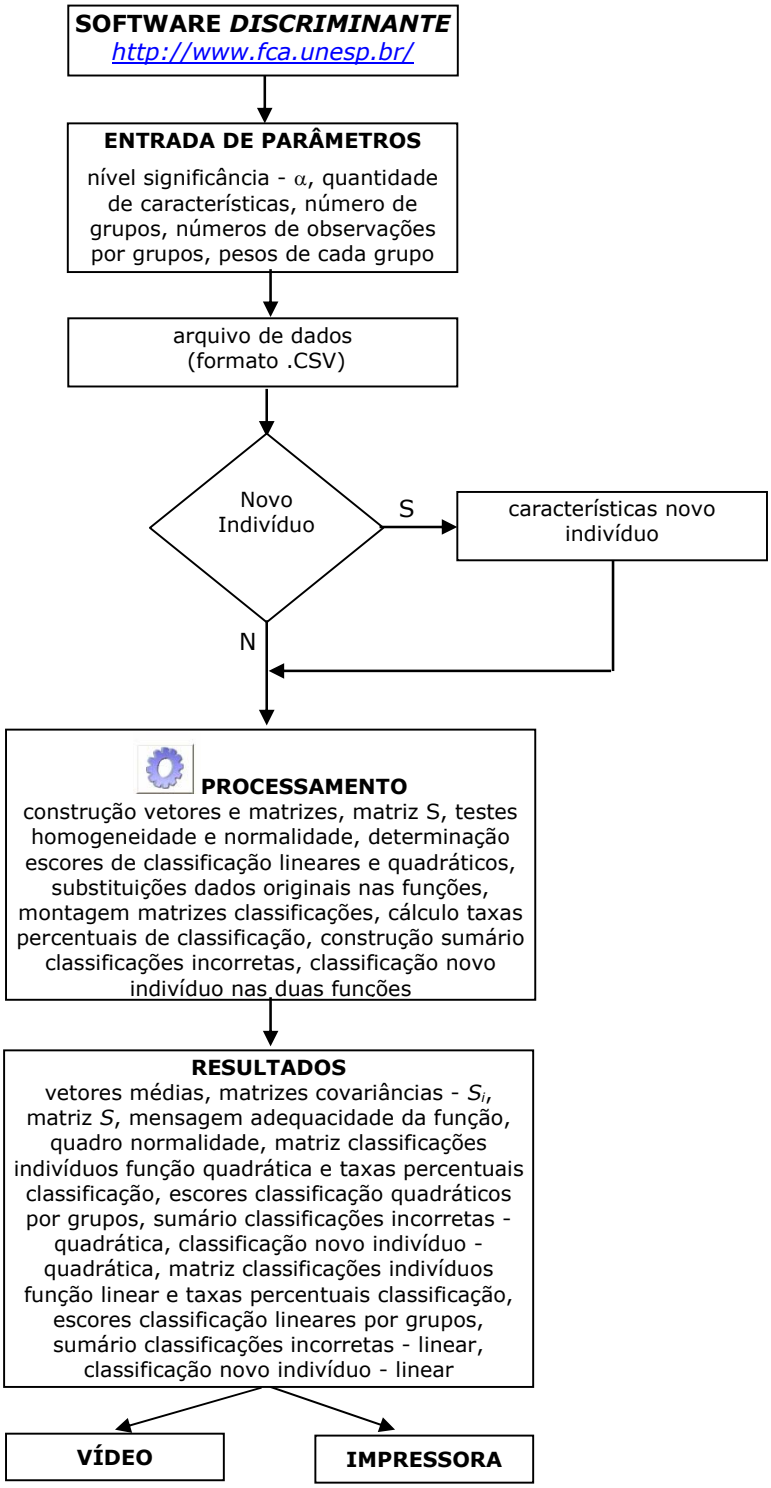

Figura 5.11 - Diagrama de blocos do *Software DISCRIMINANTE*

4.3 Exemplos da área agronômica

A utilização da função discriminante como técnica de classificação pode ser feita depois de dado o passo inicial, isto é, depois de se ter encontrado um conjunto de variáveis que permita a discriminação significativa dos grupos, para os quais se conhece a distribuição por grupos. É então possível estimar um conjunto de funções que permitirão a classificação de novos casos, cujo agrupamento seja inicialmente desconhecido.

Dois conjuntos de dados foram submetidos ao programa *DISCRIMINANTE*:

- *GIRASSOIS* (MESSETTI, 2000): dados normais coletados por pesquisadores da EMBRAPA na região de Londrina - PR, no ano agrícola de 1996/97, com o objetivo de avaliar o grau de adaptabilidade de 12 populações de girassol de diferentes origens. Dentre as 12 espécies, 6 foram selecionadas para constituição da base de dados: *3GRNAINS, TALENAY, CORONA, COMANGIR, PEHUEN* e *MARIBONDO*. Todos os grupos são constituídos de 30 indivíduos avaliados sob 6 características: X3 - altura da planta (metros), X5 - diâmetro do caule (mm), X9 - altura do capítulo (cm), X10 - diâmetro do capítulo (cm), X11 - peso de 1000 aqüênios (gramas) e X12 - teor de óleo (%). (Anexo 1).

- *EUCALYPTUS* (XAVIER, 1998): dados originais obtidos por meio de sensoriamento remoto para estimar o Índice de Área Foliar (IAF), o Diâmetro à Altura do Peito (DAP), a Altura e a Idade de diferentes materiais genéticos de *Eucalyptus*, dos 12 aos 78 meses de idade. A área teste utilizada foi a regional Aracruz da Aracruz Celulose S.A., município de Aracruz, Espírito Santo. A imagem TM utilizada para o processamento digital foi a do satélite Landsat-

5. Foram realizadas as seguintes etapas no processamento das imagens, visando também a normalidade dos dados: correção dos níveis digitais, correção atmosférica, conversão para reflectância, cálculo dos Índices de Vegetação, NDVI, Índice de Vegetação da Diferença Normalizada ("Normalized Difference Vegetation Index") e SAVI, Índice de Vegetação de Ajuste do Solo ("Soil-Adjusted Vegetation Index") e, cálculo dos valores de proporção da vegetação (Pveg), proporção de solo (Psol) e proporção de sombra (Psom). Os grupos foram divididos de acordo com o comportamento do IAF, que foi diferenciado nos vários materiais genéticos estudados - CL1 (17 indivíduos), CL2 (25 indivíduos), CL3 (12 indivíduos), CL4 (14 indivíduos) e CL5 (26 indivíduos), levando-se em consideração o manejo do tipo reforma - o *eucalyptus* é plantado por meio de mudas, podendo estas serem mudas de sementes ou da parte vegetativa. A seqüência das características observadas e introduzidas na planilha de dados é: IDADE, IAF, DAP, ALTURA, SAVI, NDVI E MIS (Anexo 2).

5 RESULTADOS E DISCUSSÃO

Os dois conjuntos de dados - Girassóis e *Eucalyptus* - submetidos à análise classificatória por meio do *Software DISCRIMINANTE*, foram avaliados em dois momentos distintos, atendendo aos procedimentos para se estimarem as probabilidades de classificações incorretas, abordados no Capítulo 3, ou seja, o método da ressubstituição e o método da colocação de elementos à parte para classificação.

5.1 Girassóis

As listagens computacionais dos resultados obtidos nas análises classificatórias dos dados dos seis grupos de Girassóis, considerando, inicialmente, o método da ressubstituição e, posteriormente o método da colocação de elementos à parte para classificação, são apresentadas, integralmente, nos Apêndices 1 e 2, respectivamente.

5.1.1 Girassóis - Método da Ressubstituição

De maneira sintética, as Tabelas 5.1 e 5.2 apresentam os resultados das classificações dos indivíduos pela utilização da função discriminante quadrática e respectivas taxas percentuais de classificação por grupos, bem como a taxa percentual total. Na seqüência, as Tabelas 5.3 e 5.4 trazem os resultados classificatórios sumarizados quando da aplicação da função discriminante linear.

Tabela 5.1 - Tabela de classificações dos indivíduos pela função discriminante quadrática e respectivas taxas percentuais de classificação

População Classificada pela função Quadrática	População Original					
	G1	**G2**	**G3**	**G4**	**G5**	**G6**
G1	25	0	2	0	0	1
G2	0	23	2	0	1	3
G3	3	2	24	0	2	0
G4	0	0	0	29	0	0
G5	1	0	0	1	27	0
G6	1	5	2	0	0	26
Total Indivíduos	30	30	30	30	30	30
Indivíduos Corretos	25	23	24	29	27	26
% Corretos	83,3	76,7	80,0	96,7	90,0	86,7

Tabela 5.2 - Tabela contendo os escores de classificação quadráticos

	Coeficientes Quadráticos	Coeficientes Mistos	Coeficientes Lineares	Termo Independente
d^{Q}_{1}	-17,91639	0,304757028	8,5862141	-52,2753412
	-0,183308	0,044144923	3,2401584	
	-0,001769	0,597865704	0,147682	
	-0,141919	0,035460851	-1,5871091	
	-0,000795	0,246617415	-0,1306709	
	-0,017775	-0,002589275	1,0979553	
		0,240322593		
		0,002750169		
		-0,02981669		
		0,006285468		
		0,000716264		
		0,000594918		

		0,00277473		
		0,022968743		
		-0,000152343		
$\mathbf{d^Q_2}$				
	-21,548	1,026014806	75,94627	-142,10377
	-0,177473	0,221094533	-0,1127479	
	-0,00249	-2,600391895	-0,3593301	
	-0,447382	0,192654746	7,1332472	
	-0,004153	-0,468374224	-0,869539	
	-0,029734	0,00170722	2,7918553	
		0,332809138		
		0,018516642		
		-0,025445493		
		0,022156221		
		-0,001876723		
		0,005800307		
		0,043287178		
		0,008309113		
		0,003249167		
$\mathbf{d^Q_3}$				
	-31,38187	0,72409993	35,081041	-75,3235645
	-0,081981	0,074554064	1,9207763	
	-0,001102	-0,596899045	-0,0043496	
	-0,126584	0,306014123	2,134288	
	-0,005174	0,773609061	-0,8520241	
	-0,025796	0,002707863	0,9926958	
		0,075073642		
		0,01319667		
		-0,035429883		
		-0,000537393		
		0,000365859		
		0,001290601		
		0,020379902		
		0,019845575		
		0,002892684		
$\mathbf{d^Q_4}$				
	-5,247905	0,251712619	25,479348	-115,520432
	-0,042274	0,014477654	0,5566887	
	-0,000949	0,144354238	0,0227655	
	-0,105043	0,008254571	0,9740115	
	-0,000716	-0,224289403	-0,2422871	
	-0,05368	0,001986875	3,4620341	
		0,05095813		

		0,000410761		
		-0,008320673		
		0,0063446		
		-0,000224045		
		0,004537511		
		0,009943849		
		-0,024323355		
		0,005310226		
d^{Q}_{5}				
	-17,28428	0,687896944	52,459102	-99,292541
	-0,10975	0,038603403	1,3351624	
	-0,000567	0,528775397	-0,08155	
	-0,058484	-0,011316619	0,6145766	
	-0,002968	-0,348756075	-0,3974735	
	-0,025432	0,000899337	1,5058624	
		-0,006561217		
		0,023790578		
		0,038207169		
		-0,000784319		
		0,000601478		
		0,003123977		
		0,013132049		
		0,002286873		
		-0,005092302		
d^{Q}_{6}				
	-75,94146	5,637286188	83,232998	-125,723084
	-0,320861	0,295126158	-2,0668953	
	-0,001545	-1,177907579	-0,2140533	
	-0,226755	-0,218146371	2,8778651	
	-0,005225	0,409205607	-0,2622651	
	-0,051862	-0,009221724	3,5985268	
		0,270811415		
		0,038303698		
		0,016303655		
		0,013593031		
		-0,000659141		
		0,003320304		
		0,021482312		
		-0,032742476		
		0,002371856		

Tabela 5.3 - Tabela de classificações dos indivíduos pela função discriminante linear e respectivas taxas percentuais de classificação

População Classificada pela função Linear	População Original					
	G1	G2	G3	G4	G5	G6
G1	23	0	3	0	0	1
G2	1	18	3	0	1	4
G3	2	1	23	0	2	1
G4	0	0	0	28	0	0
G5	0	2	0	2	27	0
G6	4	9	1	0	0	24
Total Indivíduos	30	30	30	30	30	30
Indivíduos Corretos	23	18	23	28	27	24
% Corretos	76,7	60,0	76,7	93,3	90,0	80,0

Tabela 5.4 - Tabela contendo os escores de classificação lineares

	Coeficientes Lineares	Termo Independente
d_1	8,586	-44,33
	3,240	
	0,148	
	-1,587	
	-0,131	
	1,098	
d_2		
	75,946	-134,97
	-0,113	
	-0,359	
	7,133	
	-0,870	
	2,792	
d_3		
	35,081	-68,03
	1,921	
	-0,004	
	2,134	
	-0,852	
	0,993	
d_4		
	25,479	-106,34
	0,557	
	0,023	

	0,974	
	-0,242	
	3,462	
d_5		
	52,459	-17,43
	1,335	
	-0,082	
	0,615	
	-0,397	
	1,506	
d_6		
	83,233	-120,09
	-2,067	
	-0,214	
	2,878	
	-0,262	
	3,599	

Portanto, pode-se concluir que, das 180 observações de espécies de Girassóis distribuídas em 6 grupos de 30 parcelas, 143 (79,4%) foram classificadas acertadamente pela utilização da análise discriminante linear, enquanto que 154 (85,6%) indivíduos foram atribuídos corretamente aos seus grupos por meio da discriminação quadrática.

5.1.2 Girassóis - Método da colocação de elementos à parte para classificação

Similarmente ao item anterior, o conjunto de dados dos Girassóis foi submetido ao *software DISCRIMINANTE;* porém, para a aplicação do método da colocação dos elementos à parte, 67% dos dados (parcelas 1 a 20 de cada grupo) foram selecionados para a definição das funções de classificação linear e quadrática - amostra de treinamento, enquanto os dados restantes - 33% (parcelas 21 a 30 de cada grupo) - aguardaram para posterior substituição nas funções - amostra de validação. Tais resultados estão relacionados nas Tabelas 5.5 e 5.6, relativos à função quadrática e linear, respectivamente.

Tabela 5.5 - Tabela de classificações dos indivíduos pela função discriminante quadrática e respectivas taxas percentuais de classificação

População Classificada pela função Quadrática	População Original					
	G1	G2	G3	G4	G5	G6
G1	9	0	0	0	0	0
G2	0	4	1	0	0	2
G3	0	1	7	1	0	0
G4	0	0	0	8	0	0
G5	0	2	0	1	10	0
G6	1	3	2	0	0	7
Total Indivíduos	10	10	10	10	10	10
Indivíduos Corretos	9	4	7	8	10	7
% Corretos	90,0	40,0	70,0	80,0	100,0	70,0

Tabela 5.6 - Tabela de classificações dos indivíduos pela função discriminante linear e respectivas taxas percentuais de classificação

População Classificada pela função Linear	População Original					
	G1	G2	G3	G4	G5	G6
G1	8	0	1	0	0	1
G2	1	3	2	0	0	3
G3	0	0	6	0	2	0
G4	0	0	0	9	0	0
G5	0	1	0	1	8	0
G6	1	6	1	0	0	6
Total Indivíduos	10	10	10	10	10	10
Indivíduos Corretos	8	3	6	9	8	6
% Corretos	80,0	30,0	60,0	90,0	80,0	60,0

Analisando-se ambas as tabelas, pode-se concluir que das 60 observações de espécies de Girassóis que ficaram à parte na determinação das funções, isto é, 10 parcelas por grupo, 40 (66,7%) foram classificadas acertadamente pela utilização da função discriminante linear, enquanto que 45 (75%) dos indivíduos foram atribuídos corretamente aos seus grupos por meio da discriminação quadrática. Dessa maneira, verificam-se que os percentuais de classificações corretas foram inferiores àqueles determinados pelas funções que

utilizaram a totalidade dos dados na sua determinação, de acordo com o previsto por Mingoti (2005).

5.2 *Eucalyptus*

Os resultados, na íntegra, obtidos nas análises classificatórias dos dados dos cinco grupos de *Eucalyptus* considerando, inicialmente, o método da ressubstituição e, posteriormente o método da colocação de elementos à parte para classificação, são apresentados, por meio de suas listagens computacionais, nos Apêndices 3 e 4, respectivamente.

5.2.1 *Eucalyptus* - Método da Ressubstituição

As Tabelas 5.7 e 5.8 representam os valores finais obtidos das classificações dos indivíduos pela utilização da função discriminante quadrática e respectivas taxas percentuais de classificação por grupos, bem como a taxa percentual total.

Em seguida, as Tabelas 5.9 e 5.10 apresentam a totalização dos dados classificados quando da utilização da função discriminante linear.

Tabela 5.7 - Tabela de classificações dos indivíduos pela função discriminante quadrática e respectivas taxas percentuais de classificação

População Classificada pela função Quadrática	População Original				
	G1	G2	G3	G4	G5
G1	9	0	0	0	0
G2	1	15	0	0	0
G3	4	4	11	0	4
G4	3	5	0	14	0
G5	0	1	1	0	22
Total Indivíduos	17	25	12	14	26
Indivíduos Corretos	9	15	11	14	22
% Corretos	52,9	60,0	91,7	100,0	84,6

Tabela 5.8 - Tabela contendo os escores de classificação quadráticos

	Coeficientes Quadráticos	Coeficientes Mistos	Coeficientes Lineares	Termo Independente
d^Q_1	-0,0327503	-0,58135952	5,951900371	-1375,278727949
	-13,709284	0,1167813	124,6751557	
	-0,7043466	0,03987109	-1,324619607	
	-0,160854	-0,08861869	-2,119759253	
	-0,539931	-0,00792893	-11,82713279	
	-0,2123223	-0,01165799	26,37999137	
	-0,2605757	1,16358791	11,27587025	
		0,40758104		
		3,02610756		
		-1,28369252		
		-1,43525279		
		0,42130255		
		0,26433982		
		-0,14790185		
		0,19870999		
		-0,08662679		
		0,05124364		
		-0,0181608		
		0,41890376		
		0,22202511		
		0,00366529		
d^Q_2				
	-0,0200255	0,01826206	-6,456037137	-31813,596721206
	-6,9962183	0,03834295	-389,3782583	
	-0,8833227	0,08969539	105,7377907	
	-0,2928047	-0,07555962	-50,1460117	
	-0,6107939	0,00950243	-77,92693413	
	-0,6055528	0,23951009	96,36110087	
	-27,398165	0,36780976	1798,755631	
		-0,37642647		
		1,49612805		
		-0,47025922		
		12,7548449		
		0,59220313		
		0,10394426		
		-0,06681637		
		-2,92757021		
		-0,07338067		
		0,17693839		
		1,18794317		

		0,9767806		
		1,14999993		
		-0,90168107		
d^{Q}_{3}				
	-0,0120212	-0,04587553	-151,6225963	-2569263,38548816
	-16,730929	0,24088626	6911,32588	
	-5,7979387	-0,04590307	6887,275865	
	-0,7025155	-0,1223027	-2280,551548	
	-2,9626197	0,06056546	-4125,662548	
	-4,3221138	4,57658078	5189,086868	
	-2393,2684	-10,0139741	166372,2398	
		3,88635929		
		11,5283891		
		-11,8780153		
		-187,991635		
		3,857034		
		6,75637808		
		-8,71163471		
		-191,982658		
		-2,34720088		
		3,03525687		
		63,2714037		
		6,36039419		
		113,722453		
		-4786,53687		
d^{Q}_{4}				
	-0,0095563	0,31548729	47,63592213	-964466,464996161
	-11,590377	0,02347761	1314,955737	
	-0,4583233	0,02985311	54,63401037	
	-0,0874737	-0,10422426	-119,6644095	
	-0,9532252	0,01984592	354,0725935	
	-0,4373443	-1,41798415	-12,79714864	
	0,87404647	0,14903189	52266,60121	
		-0,78639929		
		3,33713078		
		-1,04560035		
		-39,2033284		
		0,1704067		
		-0,66504129		
		0,31886915		
		-1,54027958		
		0,21699952		
		-0,05716989		

		3,61550699		
		0,8230725		
		1,74809294		
		-1583,46001		
d^{Q}_{5}				
	-0,0113261	-0,09646549	34,53653028	-1066269,09329288
	-3,8156871	0,07721304	7,703538896	
	-0,9891656	0,0411445	422,0465026	
	-0,2047758	0,0378043	-256,005589	
	-0,869963	0,00834305	-11,76390074	
	-0,0256385	-1,1352933	-29,77978154	
	-940,68419	0,73036265	61582,13059	
		0,05820375		
		0,96227863		
		0,06583764		
		-1,04041082		
		0,56422799		
		0,04263527		
		-0,00329307		
		-12,6079945		
		-0,06394613		
		-0,02619708		
		7,89992676		
		0,01359494		
		2,25856306		
		1,026534		

Tabela 5.9 - Tabela de classificações dos indivíduos pela função discriminante linear e respectivas taxas percentuais de classificação

População Classificada pela função Linear	População Original				
	G1	**G2**	**G3**	**G4**	**G5**
G1	13	2	0	1	0
G2	2	13	2	2	3
G3	1	6	7	0	2
G4	1	2	0	10	3
G5	0	2	3	1	18
Total Indivíduos	17	25	12	14	26
Indivíduos Corretos	13	13	7	10	18
% Corretos	76,5	52,0	58,3	71,4	69,2

Tabela 5.10 - Tabela contendo os escores de classificação lineares

	Coeficientes Lineares	Termo Independente
d_1	14,247	-63,45
	0,661	
	-0,126	
	2,761	
	-0,340	
	2,050	
d_2		
	21,391	-80,39
	0,940	
	-0,125	
	2,112	
	-0,288	
	2,316	
d_3		
	18,507	-72,79
	1,029	
	-0,138	
	2,622	
	-0,372	
	2,088	
d_4		
	49,085	-115,97
	0,726	
	-0,003	
	1,484	
	-0,357	
	1,588	
d_5		
	30,416	-87,67
	0,981	
	-0,091	
	2,015	
	-0,367	
	2,115	
d_6		
	19,404	-74,17
	0,500	
	-0,104	
	2,457	
	-0,307	
	2,277	

A análise das Tabelas 5.7 e 5.9 permite concluir que 75,5%, ou seja, 71 observações de espécies de *eucalyptus*, distribuídas em 5 grupos totalizando 94 parcelas, apresentaram classificação adequada quando da aplicação do procedimento quadrático. Por outro lado, 61 indivíduos (64,9%) tiveram sucesso no processo classificatório pela utilização da função linear.

5.2.2 *Eucalyptus* - Método da colocação de elementos à parte para classificação

Em um segundo momento, o conjunto de dados dos *Eucalyptus*, submeteu-se ao *software DISCRIMINANTE*, para a aplicação do método da colocação dos elementos à parte. Das 94 parcelas iniciais, 75% (13, 19, 9, 10 e 19 dados dos grupos 1 a 5, respectivamente) foram selecionados como amostra de treinamento. Os 25% restantes constituíram a amostra de validação. Os resultados deste procedimentos estão relacionados nas Tabelas 5.11 e 5.12, relativos à função quadrática e linear, respectivamente.

Tabela 5.11 - Tabela de classificações dos indivíduos pela função discriminante quadrática e respectivas taxas percentuais de classificação

População Classificada pela função Quadrática	População Original				
	G1	G2	G3	G4	G5
G1	9	0	0	0	0
G2	1	14	0	1	3
G3	1	2	8	0	2
G4	1	1	1	9	1
G5	0	0	0	0	13
Total Indivíduos	13	19	9	10	19
Indivíduos Corretos	9	14	8	9	13
% Corretos	69,2	73,9	88,9	90,0	68,4

Tabela 5.12 - Tabela de classificações dos indivíduos pela função discriminante linear e respectivas taxas percentuais de classificação

População Classificada pela função Linear	População Original				
	G1	G2	G3	G4	G5
G1	10	0	0	0	0
G2	0	13	0	1	3
G3	2	4	7	1	2
G4	1	1	0	8	3
G5	0	1	2	0	11
Total Indivíduos	13	19	9	10	19
Indivíduos Corretos	10	13	7	8	11
% Corretos	76,9	68,4	77,8	80,0	57,9

A conclusão obtida, por meio da análise das Tabelas 5.11 e 5.12, é que para a função linear, houve um aparente ganho percentual - de 64,9% para 70% - indicando que não houve perdas no processo classificatório, enquanto que a função quadrática, praticamente apresentou o mesmo percentual.

Dessa maneira, a conclusão geral é que, embora os valores percentuais dos acertos dos procedimentos linear e quadrático estejam relativamente próximos, deve ser considerado que, na prática experimental, todo ganho de precisão de respostas torna-se fator primordial para a fidedignidade das conclusões e, extremamente fundamental para futuras decisões com alto grau de confiança.

6 CONSIDERAÇÕES FINAIS

Como mencionado na Introdução deste trabalho, o objetivo do mesmo foi a discussão dos aspectos teóricos relacionados à Análise Discriminante Multivariada (Linear e Quadrática), com ênfase à sua utilização na experimentação agronômica; o desenvolvimento e disponibilização de um programa computacional - *DISCRIMINANTE* - com respectivo manual explicativo do usuário e ilustração dos procedimentos discutidos com exemplos de aplicação envolvendo dados experimentais agronômicos.

Dessa maneira, no presente estudo, realizou-se um extenso levantamento histórico da função discriminante, com seus primórdios no trabalho de *Fisher* (1936) e sua posterior evolução, enfocando o intenso desenvolvimento das técnicas classificatórias discriminantes com o advento dos computadores. Foi dada ênfase aos *softwares* estatísticos desenvolvidos para *PC*, que realizam a análise discriminante, e que representam uma grande contribuição para pesquisadores e usuários desta técnica.

Com relação às aplicações na área de Ciências Agrárias, a revisão bibliográfica cobriu uma recente relação de trabalhos que utilizam as funções discriminantes linear e, em especial, a quadrática, mostrando que se trata de uma técnica bastante empregada; porém, em algumas situações, apresentando certos descuidos metodológicos no estudo de suas ramificações: solos, cultivos diversos (soja, milho, cana de açúcar, pupunha, braquiária, frutas), criação de animais e classificação e seleção de madeiras. Também mostrou que a discriminação quadrática está, atualmente, despertando nas pesquisas que envolvem a coleta de dados por meio de sistemas de visão artificial.

Como complemento ao estudo teórico, dois conjuntos de dados agronômicos - espécies de Girassóis e de *Eucalyptus* - foram submetidos à análise discriminante, feita por meio do programa *DISCRIMINANTE*. Para ambos os casos foi possível constatar a não homogeneidade entre as matrizes de covariâncias dos grupos, o que permitiu, para efeitos didáticos, a comparação das discriminações linear e quadrática (mais adequada para esta situação).

A pesquisa desenvolvida aponta para possibilidades de trabalhos futuros, descritos a seguir:

- Implementação computacional da discriminação quadrática gráfica;
- Estender o algoritmo computacional *DISCRIMINANTE*, para que realize o procedimento de Lachenbruch ou validação cruzada para se estimarem as probabilidades de classificações incorretas;
- União, em um único aplicativo, de técnicas de seleção das variáveis que mais contribuem para uma classificação adequada, previamente à aplicação da técnica classificatória.

REFERÊNCIAS

ADAMS, M. L. et al.. Toward the discrimination of manganese, zinc, copper and iron deficiency in bragg Soybean using spectral detection methods. **Agronomy Journal**, Madison, v. 92, p. 268-274, 2000.

ALVAREZ, M. A. **Introdução à programação em PHP**. Disponível em: http://www.criarweb.com/artigos/70.php . Acesso em: 12 set. 2006.

ANDERSON, T. W. **An introduction to multivariate statistical analysi**s. 3. ed. New York: John Wiley & Sons, 2003. 752 p.

ANDERSON, T. W.; BADAHUR, R. R. Classification into two multivariate normal distributions with different covariance matrices. **The Annals of Mathematical Statistics**, Beachwood, v. 33, n. 2, p. 420-431, 1962.

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

ASSIS, G. M. L. et al.. *Brachiaria* species discrimination based on different groups of morphological traits. **Revista Brasileira de Zootecnia**, Viçosa, v. 32, n. 3, p. 576-584, 2003.

BANOWETZ, G. M. et al.. High resolution characterization of soil biological communities by nucleic acid and fatty acid analyses. **Soil Biology & Biochemistry**, Madison, v. 34, p. 1853-1860, 2002.

BARTLETT, M. S. Further aspectsof the theory of multiple regression. **Proceedings of the Cambridge Philosophical Society**, Cambridge, v. 34, p. 33-40, 1938.

BERNARD, M. M. The secular variations of skull characteres in four series of egiptian skulls. **Annals of Eugenics,** London, v. 6, 1935.

BEALL, G. Approximate methods in calculating discriminant functions. **Psychometric**, Champaign, v. 10, n. 3, p. 205-217, 1945.

BOX, G. E. P. A general distribution theory for a class of likelihood criteria. **Biometrika**, London, v. 36, p. 317-346, 1949.

BRYAN, J. G. The generalized discriminant function: mathematical foundation and computational routine. **Harvard Educational Review**, Cambridge, v. 21, 1951.

BURNABY, T. P. Growth invariant discriminant functions and generalized distances. **Biometrics**, Arlington, v. 22, n. 1, p. 96-110, 1966.

CAMPOS, H. **Estatística experimental não-paramétrica**. Piracicaba: ESALQ/USP, 1983. 349 p.

CARREIRAS, J. M. B. et al.. Assessing the extent of agriculture/pasture and secondary succession forest in the Brazilian Legal Amazon using SPOT VEGETATION data. **Remote Sensing of Environment**, New York, v. 101, p. 283-298, 2006.

CONOVER, J. W. **Practical nonparametric statistics**. New York: John Wiley & Sons, 1980.

COOKE, T. A lower bound on the performance of the quadratic discriminant function. **Journal of Multivariate Analysis**, New York, v. 89, p. 371-383, 2004.

COOLEY, W. W.; LOHNES, P. R. **Multivariate data analysis.** New York: John Wiley & Sons, 1971, 363 p.

DANIEL, W. W. **Bioestatistics: a foundation for analysis in the Health Sciences**. New York: John Wiley & Sons, 1995.

DUNN, O. J.; MARKS, S. Discriminant functions when covariance natrices are unequal. **Journal of the American Statistical Association**, Alexandria, v. 69, n. 346, p. 555-559, 1974.

ERBERT, M.; HAERTEL, V. Estudo sobre técnicas de regularização da matriz covariância no processo de classificação de dados em alta dimensionalidade. In: Simpósio Brasileiro de Sensoriamento Remoto, 11., 2003, Belo Horizonte. **Anais ...** Belo Horizonte: INPE, 2003. p. 1061-1068.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, London, v. 7, n. 2, p. 179-188, 1936.

FISHER, R. A. The statistical utilization of multiple measurements. **Annals of Eugenics**, London, v. 8, n. 4. 1938.

FRANCIS, I. A survey of statistical software. **Computational Statistics & Data Analysis**, Voorburg, v. 1, p. 17-27, 1983.

FRIEDMAN, J. H. Regularized discriminant analysis. **Journal of the American Statistical Association**, Alexandria, v. 84, p. 165-175, 1989.

GALVÃO, L. S.; FORMAGGIO, A. R.; TISOT, D. A. Discriminação de variedades de cana-de-açúcar com dados hiperespectrais do sensor Hyperion/Eo-1. **Revista Brasileira de Cartografia**, Rio de Janeiro, n. 57/01, 2005.

GILBERT, E. S. The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. **Biometrics**, Arlington, v. 25, p. 505-515, 1969.

GRÖGER, J.; GRÖHSLER, T. Comparative analysis of alternative statistical models for differentiation of herring stocks based on meristic characters. **Journal of Applied Ichthyology**, Oxford, v. 17, p. 207-219, 2001.

GREEN, P. E.; CARROL, J. D. **Mathematical tools for applied multivariate analysis**. New York: Academic, 1976. 375 p.

GROUVEN, U. et al.. A PC program for unbiased and predictive linear and quadratic discriminant analysis. **Computers in Biology and Medicine**, New York, v. 25, n. 4, p. 425-430, 1995.

GROUVEN, U.; BERGEL, F.; SCHULTZ, A. implementation of linear and quadratic discriminant analysis incorporating costs of misclassification. **Computer Methods and Programs in Biomedicine**, New York, v. 49, p. 55-60, 1996.

HASE, H. et al.. A quadratic discriminant function based on bias rectification of eigenvalues. **Scripta Technica Systems and Computers in Japan**, Tokyo, v. 31, n. 9, p. 28-38, 2000.

HOEL, P. G.; PETERSON, R. P. A solution to the problem of optimum classification. **Annals of Mathematical Statistics**, Beachwood, v. 20, n. 3, p. 433-438, 1949.

HORTON, I. F.; MOORE, A. W.; RUSSELL, J. S. Multivariate-covariance and canonical analysis: a method for selecting the most effective discriminators in a multivariate situation. **Biometrics**, Arlington, v. 24, n. 4, p. 845-858, 1968.

HUA, J.; XIONG, Z.; DOUGHERTY, E. R. Determination of the optimal number of features for quadratic discriminant analysis via the normal approximation to the discriminant

distribution. **Journal of the Pattern Recognition Society**, New York, v. 38, p. 403-421, 2005.

JACKSON, R. W. B. Aproximate multiple regression weights. **Journal of Experimental Education**, Washington, v. 11, p. 221-225, 1943.

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

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis.** 5. ed. New Jersey: Prentice-Hall, 2002. 767 p.

KANE, V. E., BAYNE, C. K., BEAUCHAMP, J. J. Assessment of Fisher and logistic linear and quadratic discrimination models. **Computational Statistics & Data Analysis**, Voorburg, v. 1, p. 257-273, 1983.

KHOURY JR, J. K. et al.. Análise discriminante paramétrica para reconhecimento de defeitos em tábuas de eucalipto utilizando imagens digitais. **Revista Árvore**, Viçosa, v. 29, n. 2, p. 299-309, 2005.

KRONMAL, R. A.; WAHL, P. W. Discriminant functions when covariances are unequal and sample sizes are moderate. **Biometrics**, Arlington, v. 33, n. 3, p. 479-484, 1977.

KRZANOWSKI, W.J. The performance of Fisher's linear discriminant function under non-optimal conditions. **Technometrics**, v. 19, n. 2, p. 191-200, 1977.

LACHENBRUCH, P. A. **Discriminant analysis**. New York: Hafner Press, 1975. 128 p.

LERDORF, R. Do You PHP? Disponível em: http://www.oracle.com/technology/pub/articles/php_experts/rasmus_php.html. Acesso em: 06 out. 2006.

LINDEMAN, R. H.; MERENDA, P. F.; GOLD, R. Z. **Introduction to bivariate and multivariate analysis.** USA: Scott, Foresman and Company, 1980, 443 p.

LUBSCHEW, A. A. On the use of discriminant functions in taxonomy. **Biometrics,** Arlington, v. 18, n. 4, p. 455-477, 1962.

MARTEL, J. H. I. et al.. Estatística multivariada na discriminação de raças amazônicas de pupunheiras (*Bactris gasipaes Kunth*) em Manaus (Brasil). **Revista Brasileira de Fruticultura**, Jaboticabal, v. 25, n. 1, p. 115-118, 2003.

MEHTA, J. S.; SRINIVASAN, R. On the Behrens-Fisher problem. **Biometrika**, London, v. 57, p. 649-655, 1970.

MESSETTI, A. V. L. Estudo da semelhança de genótipos de Girassol (*Helianthus annuus L.*) com o uso da distância generalizada de Mahalanobis na análise de agrupamento. 86 p. Dissertação (Mestrado em Energia na Agricultura) - Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu, 2000.

MINGOTI, S. A. Análise **de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005. 297 p.

MORRISON, D. F. **Multivariate statistical methods**. 4. ed. Belmont: Thomson Brooks/Cole, 2005. 480 p.

MOSHOU, D. et al.. Apple mealiness detection using fluorescence and self-organising maps. **Computers and Electronics in Agriculture**, Budapest, v. 40, p. 103-114, 2003.

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**, Brasília, v. 39, n. 10, p. 995-1006, 2004.

NEYMAN, J.; PEARSON, E. S. The testing of statistical hypotheses in relation to probabilities a priori. **Proceedings of the Cambridge Philosophical Society**, Cambridge, v. 24, p. 492-510, 1933.

PIRES, A. V. et al.. Genetic divergence study among Duroc, Landrace and Large White swine breeds using techniques of multivariate analysis. **Archivos Latinoamericanos de Producción Animal**, Maracaibo, v. 10, n. 2, p. 81-85, 2002.

RAO, C. R. The utilization of multiple measurements in problems of biological classification. **Journal of the Royal Statistical Society**, London, Series B, v. 10, n. 2, p. 159-203, 1948.

RAO, C. R. Use of discriminant and allied functions in multivariate analysis. **Sankhyã - The Indian Journal of Statistics**, Kolkata, v. A 24, p. 149-154, 1962.

RAO, C. R. Sir Ronald Aylmer Fisher: the architect of multivariate analysis. **Biometrics**, Arlington, v. 20, p. 286-300, 1964.

RAO, C. R. Discriminant function between composite hypotheses and related problems. **Biometrika**, London, v. 53, n. 3/4, p. 339-345, 1966.

RAO, C. R. Recent trends of research work in multivariate analysis. **Biometrics**, Arlington, v. 28, p. 3-22, 1972.

REIS, E. **Estatística multivariada aplicada**. Lisboa: Edições Silabo, 1997. 343 p.

SCHOTT, J. R. Dimensionality reduction in quadratic discriminant analysis. **Computational Statistics & Data Analysis**, Voorburg, v. 16, p. 161-174, 1993.

SCHUCANY, W. R.; MINTON, P. D.; SHANNON JR, S. A survey of statistical packages, **Computing Surveys**, New York, v. 4, p. 65-79, 1972.

SMITH, C. A. B. Some examples of discrimination. **Annals of Eugenics**, London, v. 13, p. 228-237, 1947.

SHAPIRO, S. S.; WILK, M. B.; CHEN, H. J. A comparative study of various tests for normality. **Journal of America Statistics Association**, Alexandria, p. 1343-1372, 1968.

SIEGEL, S. **Estatística não paramétrica.** São Paulo, McGraw Hill do Brasil, 1981. 312 p.

STEEL, R. G. D.; TORRIE, J. H. **Principles and procedures of statistics**. New York: McGraw Hill Book, 1960. 481 p.

VAN NESS, J. On the effects of dimension in discriminant analysis for unequal covariance populations. **Technometrics**, Alexandria, v. 21, n. 1, p. 119-127, 1979.

VON MISES, R. On the classification of observation data into distinct groups. **Annals of Mathematical Statistics**, Beachwood, v. 16, n. 1, p. 68-73, 1945.

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

WOODWARD, W. A.; ELLIOTT, A. C. A survey of statistical packages on microcomputers. **Computational Statistics & Data Analysis**, Voorburg, v. 1, p. 191-200, 1983.

WU, W. et al.. Comparision of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data. **Analytica Chimica Acta**, New York, v. 329, p. 257-265, 1996.

XAVIER, A. C. Estimativa de propriedades biofísicas de plantações de eucaliptos a partir de dados Landsat-Tm. 116 p. Dissertação (Mestrado em Sensoriamento Remoto) - Ministério da Ciência e Tecnologia, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 1998.

ZANDONADI, R. S. et al.. Avaliação visual e sistema de visão artificial na classificação de imagens de plantas atacadas pela lagarta Elasmo (*Elasmopalpus lignosellus*) aliados à tecnologia de aplicação de agrotóxicos. In: SIMPÓSIO INTERNACIONAL DE TECNOLOGIA DE APLICAÇÃO DE AGROTÓXICOS, 3., 2004, Botucatu. **Anais ...** Botucatu: FEPAF, 2004. p. 68-71.

ZHANG, M. **Programa MZEF**. Disponível em: http://rulai.cshl.org/software/index1.htm. Acesso em: 9 ago. 2006.

ANEXO 1 - Dados GIRASSÓIS

Dados *Girassóis*

X3 (altura da planta - m)	X5 (diâmetro do caule - mm)	X9 (altura do capítulo - cm)	X10 (diâmetro do capítulo - cm)	X11 (peso de 1000 aqüênios - gramas)	X12 (teor de óleo - %)
0,870	12	42,000	9,0	4,560	41,740
0,840	18	75,000	19,0	35,700	36,610
0,760	17	56,000	16,0	12,000	34,660
0,930	17	86,000	16,0	22,760	31,520
1,290	15	105,000	15,0	21,230	45,660
1,200	20	105,000	24,0	66,080	37,830
1,010	19	100,000	17,0	31,360	36,970
1,130	21	101,000	18,0	40,690	41,710
0,960	20	50,000	15,0	18,000	31,250
1,060	17	51,000	12,0	6,750	38,920
0,960	14	64,000	10,0	4,170	18,140
1,200	20	114,000	19,0	38,490	33,830
0,990	13	90,000	14,0	17,310	35,830
0,940	22	87,000	19,0	40,000	33,890
0,870	18	74,000	16,0	5,280	29,460
1,880	22	105,000	23,0	76,230	40,600
0,900	13	80,000	10,0	6,640	29,230
0,940	14	86,000	15,0	17,080	34,970
0,840	12	82,000	12,0	15,000	37,890
0,990	14	92,000	15,0	25,310	42,760
1,000	18	72,000	19,0	41,450	40,000
0,690	12	67,000	9,0	4,720	39,420
1,190	16	53,000	17,0	19,680	33,100
1,180	19	108,000	17,0	36,800	32,900
0,840	15	82,000	15,0	6,480	26,940
1,090	18	108,000	16,0	167,020	32,500
1,040	22	117,000	22,0	70,070	35,660
0,830	14	67,000	11,0	2,980	25,700
0,980	14	93,000	13,0	12,560	41,870
1,220	18	60,000	20,0	60,520	37,690
1,960	15	149,0	14	30,30	41,470
1,400	18	114,0	16	27,62	36,870
1,680	26	100,0	21	97,00	32,670
1,390	17	78,0	14	35,56	42,090
1,350	18	108,0	17	54,20	45,980
1,890	30	127,0	22	128,45	45,650
1,700	25	100,0	17	41,83	39,300
1,440	33	125,0	26	118,64	34,730
1,900	28	150,0	23	98,28	44,190
2,140	35	139,0	24	137,50	33,900
1,140	12	82,0	12	15,62	45,430
1,750	29	114,0	21	114,06	46,270
1,950	29	117,0	22	125,75	35,870
1,230	14	90,0	12	23,15	47,190

1,690	30	107,0	24	106,46	45,240
1,650	17	94,0	11	18,73	38,560
1,880	30	142,0	22	105,37	44,100
1,400	23	133,0	20	57,00	48,270
1,360	21	130,0	17	68,98	43,060
1,450	13	102,0	12	17,00	46,210
1,830	30	131,0	21	91,66	42,620
1,590	14	94,0	13	25,78	43,320
1,640	26	102,0	22	105,06	45,890
1,740	28	115,0	23	114,11	46,050
1,560	11	76,0	10	12,99	34,740
1,490	13	62,0	12	18,00	34,650
1,500	16	86,0	15	40,13	41,240
1,480	15	102,0	14	27,35	38,710
1,870	20	114,0	17	44,59	42,870
1,400	13	92,0	11	12,36	39,300
1,54	27	112,0	19	49,230	40,390
1,59	22	109,0	18	52,790	39,700
1,10	24	99,0	19	33,050	32,590
1,49	23	97,0	16	26,290	34,810
1,37	26	124,0	17	36,990	39,860
1,58	26	137,0	21	75,630	33,120
1,50	17	78,0	15	22,420	35,290
1,48	22	62,0	14	43,230	37,400
0,96	19	77,0	20	14,440	26,740
1,30	23	122,0	20	29,000	37,450
1,16	19	74,0	17	31,550	36,250
1,38	24	62,0	17	29,500	28,810
1,30	19	92,0	14	12,540	36,030
1,11	12	52,0	14	9,310	40,030
1,44	26	94,0	19	43,070	37,130
1,56	21	83,0	17	13,820	33,430
1,28	21	68,0	19	24,070	35,520
1,52	25	108,0	20	55,070	36,600
1,14	18	65,0	14	20,300	37,670
1,04	20	68,0	15	15,390	27,330
1,38	21	89,0	16	27,840	40,790
1,43	25	140,0	16	30,270	43,400
1,28	20	86,0	15	18,000	32,090
1,47	28	86,0	17	40,000	34,420
1,44	23	78,0	20	65,000	42,360
1,40	25	97,0	21	32,480	32,400
1,73	28	126,0	21	64,560	34,350
1,01	20	45,0	12	3,180	25,530
1,41	13	138,0	12	16,400	40,010
0,90	18	87,0	9	4,000	15,540
3,290	32	221,0	24	129,210	36,380
3,780	31	170,0	25	93,990	24,410
2,720	32	199,0	16	32,480	32,290
2,520	18	164,0	14	30,570	31,440

3,400	31	201,0	18	62,100	25,070
2,270	31	198,0	24	100,810	34,260
2,760	25	107,0	14	27,850	23,890
3,580	34	227,0	22	124,720	33,020
3,600	28	219,0	22	122,600	39,270
2,950	30	187,0	15	87,220	32,240
3,460	38	246,0	27	40,150	26,230
3,420	30	171,0	21	98,990	36,400
3,200	32	212,0	24	73,020	37,120
3,450	38	247,0	27	137,710	29,440
2,860	30	215,0	21	93,600	34,420
3,400	32	227,0	20	115,990	31,250
3,270	32	185,0	23	132,700	32,390
3,600	38	191,0	18	10,680	29,360
3,000	30	237,0	20	79,300	30,720
2,750	21	218,0	16	36,560	30,520
3,180	27	196,0	21	95,140	27,790
2,730	24	203,0	18	45,210	30,400
2,980	32	236,0	26	140,530	33,850
3,000	28	170,0	22	87,330	29,080
3,050	26	189,0	18	71,260	33,210
2,480	33	183,0	20	101,990	33,890
3,400	35	225,0	26	148,110	31,600
3,120	38	200,0	24	159,990	32,000
3,200	32	221,0	19	75,000	33,030
2,970	23	185,0	15	32,480	29,940
1,660	21	56	13	21,430	37,340
2,270	25	176	19	68,780	36,320
1,980	22	60	14	13,110	29,140
2,040	28	152	18	53,000	45,850
2,180	27	164	17	18,230	45,130
1,920	19	74	11	13,430	36,620
2,090	25	164	19	72,940	38,990
2,170	20	175	15	30,720	37,010
2,270	28	130	23	95,240	35,250
2,010	20	86	14	26,990	40,580
1,730	26	110	21	62,340	36,000
1,640	21	144	16	18,990	38,860
1,990	23	163	16	28,000	44,950
2,200	23	58	16	33,080	29,270
2,040	22	149	16	42,990	41,480
1,770	21	130	16	37,730	36,530
2,020	25	171	19	60,990	39,990
2,460	32	202	25	103,440	34,390
2,150	27	173	17	48,000	47,220
1,690	20	104	14	23,410	35,340
2,420	34	158	26	90,040	34,310
1,960	18	110	11	9,460	24,680
1,690	20	120	15	36,720	36,040
2,130	31	173	20	66,710	45,170

1,590	23	134	19	28,270	41,670
1,910	25	138	14	49,070	46,560
2,090	20	155	23	27,390	33,460
2,190	18	67	26	23,410	39,470
2,390	33	171	32	136,460	35,660
2,170	30	155	22	82,670	43,120
1,500	20	88	18	54,310	31,200
1,440	18	107	16	35,000	44,950
1,270	14	65	13	31,850	45,620
1,410	17	110	15	46,000	41,620
1,200	17	114	17	56,720	41,760
1,400	18	134	17	67,350	40,470
1,150	14	104	15	42,000	40,220
1,250	14	113	16	42,560	42,240
1,380	14	105	14	31,530	42,280
1,360	16	131	15	44,200	46,350
1,470	22	131	22	48,100	43,120
1,450	19	152	15	25,630	45,670
1,130	18	104	18	69,070	41,220
1,470	22	94	18	68,680	46,080
1,300	14	128	13	17,800	43,560
1,260	17	72	13	40,770	39,310
1,370	17	131	17	37,780	41,920
1,300	19	118	18	53,000	41,020
1,320	15	122	15	30,490	43,740
1,340	14	113	12	20,300	43,050
1,420	17	124	15	45,470	43,580
1,490	21	142	15	65,440	43,950
1,290	15	98	12	39,320	45,180
1,460	19	150	18	59,080	43,330
1,030	14	60	12	30,600	43,860
1,520	23	124	19	76,230	46,120
1,250	19	120	19	47,420	38,530
1,100	10	90	8	10,310	35,670
1,420	22	94	19	75,380	44,680
1,090	14	80	10	15,860	38,430

OBS: as linhas destacadas correspondem às parcelas que foram separadas para a aplicação do método da colocação de elementos à parte para classificação (amostras de validação).

ANEXO 2 - Dados *EUCALYPTUS*

Dados *Eucalyptus*

IDADE	IAF	DAP	ALTURA	SAVI	NDVI	MIS
22	3,04	10,84	18,1	37,53	88,91	32,77
26	2,84	10,7	15,3	35,74	87,39	32,73
18	2,94	7,97	9,8	38,48	89,27	32,80
40	2,81	14,34	20,5	35,80	88,17	32,80
41	2,48	14,38	22,4	36,61	88,01	39,43
60	2,01	15,53	23,8	29,64	81,14	32,77
60	1,72	14,84	26,6	29,02	81,36	32,80
64	1,82	15,56	31,2	27,98	81,29	32,83
64	1,87	16,05	27	29,23	83,92	32,83
21	3,28	10,61	18	37,73	88,7	32,77
36	2,61	14,51	23,8	36,97	87,66	32,80
45	3,14	14,42	25	34,83	86,4	32,77
26	3,19	12,25	18,8	40,96	89,36	32,77
44	2,85	14,94	24,6	35,67	86,97	32,77
56	2,24	12,19	22,9	32,97	84,61	32,73
36	3,3	14,49	21,9	35,60	80,39	32,80
56	2,29	15,08	25	32,38	84,5	32,83
40	3,17	12,35	18,3	39,63	89,6	32,73
68	3,06	14,89	25,1	32,37	83,37	32,87
15	3,47	7,74	9,6	39,83	86,81	32,80
26	3,66	10,73	15,9	41,98	89,53	32,77
68	2,84	15,31	26,6	33,37	84,75	32,80
61	2,48	14,91	24	33,03	84,69	32,83
67	2,65	14,1	24,5	33,95	84,57	32,80
44	2,42	14,49	20,3	33,21	82,11	32,77
45	2,85	13,56	20,1	36,06	86,66	32,80
40	3,21	13,37	17,06	38,91	88,5	32,77
70	2,28	14,55	26,3	32,35	82,78	32,80
67	2,26	14,55	26,8	31,92	84,17	32,77
76	2,75	15,29	26,8	34,08	85,07	32,83
69	2,81	14,94	28,3	33,56	84,87	32,80
24	3,67	10,31	13,8	37,16	84,58	32,77
24	4,31	9,72	15,1	43,65	90	33,67
56	2,64	14,03	22,5	36,13	86,27	32,83
55	2,98	14,2	25,5	36,91	86,73	32,80
21	3,42	11,31	16,2	40,79	89,24	32,83
22	3,18	10,72	13,9	43,02	87,76	32,77
30	2,54	9,67	14,6	37,04	87,67	32,73
58	2,88	15,16	24,8	37,20	86,86	32,77
53	2,16	12,56	23	34,16	84,55	32,80
21	3,57	9,79	13,4	38,84	86,2	32,87
32	3	13,41	22,1	38,89	88,77	32,77
15	2,67	7,36	8,5	38,48	88,97	32,80
65	2,54	15,09	28,1	35,11	85,13	32,83
65	2,46	15,82	26,3	36,26	86,25	32,77
65	2,58	16,62	25,7	35,98	85,22	32,77
85	2,46	14,58	22,6	34,86	84,99	32,80
24	4,02	12,12	15,34	40,72	86,8	32,77
25	3,03	10,06	16,4	39,42	89,48	32,80

56	3,24	14,23	25,7	38,83	88,02	32,80
32	2,85	11,06	18,1	38,85	88,8	32,80
30	3,07	11,74	18,2	39,09	88,67	32,73
26	3,35	11,74	17,2	41,75	89,54	32,80
56	2,81	13,7	19,9	39,17	86,47	32,80
15	2,95	8,26	9,8	37,90	83,44	32,73
63	2,17	18,23	22,6	30,84	83,1	32,83
19	2,62	10,09	11,84	37,53	88,07	32,77
20	2,66	11,38	12,8	38,69	88,6	32,77
35	2,14	13,52	19	36,62	86,6	32,80
40	2,17	14,65	21,9	35,29	85,57	32,83
70	2,31	17,4	24	31,58	85,65	32,77
71	2,28	16,34	20,5	32,34	84,97	32,80
22	2,81	9,5	14,3	37,68	86,81	32,83
35	3,25	14,22	15,54	37,30	86,98	32,77
65	3,01	13,12	18	36,26	86,38	32,77
57	2,48	14,98	27,9	34,30	85,67	32,80
65	2,52	16,17	25,7	35,13	85,5	32,77
37	2,55	13,56	19,9	36,75	87,5	32,77
15	1,94	7,95	8,41	37,43	87,17	32,80
76	3,47	16,03	26	39,02	87,46	32,80
41	2,7	13,28	17,2	39,43	89,23	32,80
56	3,54	14,33	21,7	39,64	88,24	32,83
68	2,66	16,26	24	37,99	88,25	32,73
68	3,32	16,24	29,2	38,81	87,6	32,80
68	2,79	16,54	27,6	39,49	87,53	32,83
21	3,48	9,9	13,2	39,76	87,89	32,77
33	3,25	13,45	17,65	37,24	87,51	32,80
35	3,57	13,78	19,34	37,69	87,23	32,80
69	3,04	14,96	22,3	37,97	85,99	32,77
36	3,03	12,93	18,2	39,17	66,96	32,77
36	2,27	12,87	19,6	38,80	88,44	32,80
44	3,28	15,27	19,8	39,20	88,33	32,77
45	3,21	14,03	18,2	39,66	89,35	32,79
62	3,13	14,73	22	40,07	87,73	32,80
62	3	14,44	21	39,81	87,21	32,80
22	3,45	10,45	11,3	38,25	88,29	32,80
22	3,13	9,61	13,4	37,78	88,22	32,80
44	2,77	14,27	22,3	37,65	76,73	32,80
78	3,15	16,44	25,5	39,21	87,67	32,73
58	3,24	15,31	24,8	39,30	87,88	32,77
59	3,25	15,7	20,4	39,69	87,84	32,77
61	3,01	15,98	24,9	39,37	88,16	32,77
60	3,03	14,45	23,5	39,30	88,7	32,77
66	2,52	15,11	22	38,12	86,05	32,80

OBS: as linhas destacadas correspondem às parcelas que foram separadas para a aplicação do método da colocação de elementos à parte para classificação (amostras de validação).

APÊNDICE 1 - Resultados do processamento do conjunto de dados de GIRASSÓIS - Método da Ressubstituição

UNIVERSIDADE ESTADUAL PAULISTA
"JULIO DE MESQUITA FILHO"
Campus de Botucatu

SOFTWARE DISCRIMINANTE

GIRASSOIS 20/02/2007

Vetores de Médias por Grupo

$$\begin{bmatrix} 1{,}021 \\ 16{,}800 \\ 82{,}400 \\ 15{,}767 \\ 30{,}897 \\ 35{,}308 \end{bmatrix} \quad \begin{bmatrix} 1{,}615 \\ 21{,}633 \\ 109{,}167 \\ 17{,}500 \\ 63{,}784 \\ 41{,}548 \end{bmatrix} \quad \begin{bmatrix} 1{,}343 \\ 21{,}833 \\ 91{,}833 \\ 16{,}800 \\ 31{,}314 \\ 34{,}901 \end{bmatrix} \quad \begin{bmatrix} 3{,}113 \\ 30{,}367 \\ 201{,}667 \\ 20{,}667 \\ 86{,}243 \\ 31{,}497 \end{bmatrix} \quad \begin{bmatrix} 2{,}027 \\ 24{,}233 \\ 134{,}067 \\ 18{,}233 \\ 47{,}4347 \\ 38{,}213 \end{bmatrix} \quad \begin{bmatrix} 1{,}328 \\ 17{,}100 \\ 110{,}600 \\ 15{,}467 \\ 44{,}275 \\ 42{,}291 \end{bmatrix}$$

Matrizes de Covariâncias por Grupo

$$S_1 = \begin{bmatrix} 0{,}0480 & 0{,}3557 & 2{,}0104 & 0{,}5115 & 3{,}4480 & 0{,}384 \\ 0{,}3557 & 10{,}0276 & 27{,}1517 & 10{,}4690 & 55{,}691 & 1{,}037 \\ 2{,}0104 & 27{,}1517 & 440{,}9379 & 42{,}5103 & 362{,}362 & 24{,}466 \\ 0{,}5115 & 10{,}4690 & 42{,}5103 & 15{,}5644 & 75{,}347 & 5{,}213 \\ 3{,}4480 & 55{,}691 & 362{,}362 & 75{,}3466 & 1094{,}5319 & 27{,}265 \\ 0{,}384 & 1{,}037 & 24{,}466 & 5{,}213 & 27{,}265 & 33{,}5850 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 0{,}0588 & 1{,}1278 & 3{,}1826 & 0{,}6347 & 6{,}389 & -\,0{,}197 \\ 1{,}1278 & 54{,}0333 & 106{,}0632 & 33{,}9138 & 299{,}153 & -\,0{,}574 \\ 3{,}1826 & 106{,}0632 & 498{,}4885 & 70{,}3966 & 572{,}013 & 19{,}261 \\ 0{,}6347 & 33{,}9138 & 70{,}3966 & 23{,}0862 & 195{,}212 & 1{,}248 \\ 6{,}389 & 299{,}153 & 572{,}013 & 195{,}212 & 1825{,}2425 & 4{,}474 \\ -\,0{,}197 & -\,0{,}574 & 19{,}261 & 1{,}248 & 4{,}474 & 20{,}9135 \end{bmatrix}$$

$$S_3 = \begin{bmatrix} 0{,}0430 & 0{,}4422 & 2{,}7726 & 0{,}2944 & 2{,}704 & 0{,}676 \\ 0{,}4422 & 15{,}9368 & 42{,}6954 & 7{,}9310 & 51{,}284 & 2{,}681 \\ 2{,}7726 & 42{,}6954 & 661{,}7299 & 27{,}1379 & 227{,}838 & 52{,}022 \\ 0{,}2944 & 7{,}9310 & 27{,}1379 & 9{,}1310 & 39{,}2643 & 5{,}361 \\ 2{,}7038 & 51{,}2838 & 227{,}8383 & 39{,}2643 & 340{,}0334 & 45{,}193 \\ 0{,}6759 & 2{,}6813 & 52{,}0216 & 5{,}3610 & 45{,}193 & 33{,}5747 \end{bmatrix}$$

$$S_4 = \begin{bmatrix} 0,1368 & 0,9044 & 3,2572 & 0,6386 & 4,4436 & -0,143 \\ 0,9044 & 24,1713 & 65,7471 & 12,8161 & 93,643 & 0,744 \\ 3,2572 & 65,7471 & 850,0920 & 61,5057 & 452,247 & 32,463 \\ 0,6386 & 12,8161 & 61,5057 & 15,1954 & 112,087 & 2,373 \\ 4,4436 & 93,643 & 452,247 & 112,0869 & 1682,9376 & 60,420 \\ -0,143 & 0,744 & 32,463 & 2,373 & 60,420 & 13,3786 \end{bmatrix}$$

$$S_5 = \begin{bmatrix} 0,564 & 0,6544 & 4,1947 & 0,6986 & 4,5374 & -0,061 \\ 0,6544 & 20,7368 & 110,0184 & 14,6333 & 119,787 & 6,512 \\ 4,1947 & 110,0184 & 1676,6161 & 82,9494 & 709,449 & 89,558 \\ 0,6986 & 14,6333 & 82,9494 & 23,4954 & 117,155 & 0,624 \\ 4,5374 & 119,787 & 709,449 & 117,1551 & 961,0312 & 11,494 \\ -0,061 & 6,512 & 89,558 & 0,624 & 11,494 & 29,3459 \end{bmatrix}$$

$$S_6 = \begin{bmatrix} 0,0179 & 0,2957 & 1,7206 & 0,2241 & 1,0902 & 0,127 \\ 0,2957 & 9,8172 & 26,1793 & 7,8483 & 44,767 & 2,094 \\ 1,7206 & 26,1793 & 546,5241 & 30,4345 & 92,876 & 20,914 \\ 0,2241 & 7,8483 & 30,4345 & 9,0851 & 41,109 & 1,164 \\ 1,0902 & 44,767 & 92,876 & 41,1086 & 317,6098 & 8,597 \\ 0,127 & 2,094 & 20,914 & 1,164 & 8,597 & 10,9680 \end{bmatrix}$$

Matriz de Variação Conjunta *S*

$$S = \begin{bmatrix} 0,0602 & 0,6300 & 2,8563 & 0,5003 & 3,7686 & 0,1309 \\ 0,6300 & 22,4538 & 62,9759 & 14,6019 & 110,7209 & 2,0824 \\ 2,8563 & 62,9759 & 779,0648 & 52,4891 & 402,7976 & 39,7805 \\ 0,5003 & 14,6019 & 52,4891 & 15,9262 & 96,6955 & 2,6638 \\ 3,7686 & 110,7209 & 402,7976 & 96,6955 & 1036,8977 & 26,2405 \\ 0,1309 & 2,0824 & 39,7805 & 2,6638 & 26,2405 & 23,6276 \end{bmatrix}$$

As matrizes de covariância são heterogêneas. Função Adequada: QUADRÁTICA

Quadro sinótico Teste de Normalidade

	V1	V2	V3	V4	V5	V6
G1	S	S	S	S	S	S
G2	S	S	S	S	S	S
G3	S	S	S	S	S	S
G4	S	S	S	S	S	S
G5	S	S	S	S	S	S
G6	S	S	S	S	S	S

Matriz de classificações dos indivíduos pela função discriminante quadrática e respectivas taxas percentuais de classificação

População Classificada pela função Quadrática	População Original					
	G1	G2	G3	G4	G5	G6
G1	25	0	2	0	0	1
G2	0	23	2	0	1	3
G3	3	2	24	0	2	0
G4	0	0	0	29	0	0
G5	1	0	0	1	27	0
G6	1	5	2	0	0	26
Total Indivíduos	30	30	30	30	30	30
Indivíduos Corretos	25	23	24	29	27	26
% Correta	83,3	76,7	80,0	96,7	90	86,7

Escores de classificação quadráticos para os grupos

	Coeficientes Quadráticos	Coeficientes Mistos	Coeficientes Lineares	Termo Independente
d^Q_1				
	$-17{,}91639\ y_1^2$	$0{,}304757028\ y_1y_2$	$8{,}5862141\ y_1$	-52,2753412
	$-0{,}183308\ y_2^2$	$0{,}044144923\ y_1y_3$	$3{,}2401584\ y_2$	
	$-0{,}001769\ y_3^2$	$0{,}597865704\ y_1y_4$	$0{,}147682\ y_3$	
	$-0{,}141919\ y_4^2$	$0{,}035460851\ y_1y_5$	$-1{,}5871091\ y_4$	
	$-0{,}000795\ y_5^2$	$0{,}246617415\ y_1y_6$	$-0{,}1306709\ y_5$	
	$-0{,}017775\ y_6^2$	$-0{,}002589275\ y_2y_3$	$1{,}0979553\ y_6$	
		$0{,}240322593\ y_2y_4$		
		$0{,}002750169\ y_2y_5$		
		$-0{,}02981669\ y_2y_6$		

	0,006285468 y_3y_4		
	0,000716264 y_3y_5		
	0,000594918 y_3y_6		
	0,00277473 y_4y_5		
	0,022968743 y_4y_6		
	-0,000152343 y_5y_6		
d^Q_2			
-21,548 y_1^2	1,026014806	75,94627	-142,10377
-0,177473 y_2^2	0,221094533	-0,1127479	
-0,00249 y_3^2	-2,600391895	-0,3593301	
-0,447382 y_4^2	0,192654746	7,1332472	
-0,004153 y_5^2	-0,468374224	-0,869539	
-0,029734 y_6^2	0,00170722	2,7918553	
	0,332809138		
	0,018516642		
	-0,025445493		
	0,022156221		
	-0,001876723		
	0,005800307		
	0,043287178		
	0,008309113		
	0,003249167		
d^Q_3			
-31,38187	0,72409993	35,081041	-75,3235645
-0,081981	0,074554064	1,9207763	
-0,001102	-0,596899045	-0,0043496	
-0,126584	0,306014123	2,134288	
-0,005174	0,773609061	-0,8520241	
-0,025796 y_6^2	0,002707863	0,9926958	
	0,075073642		
	0,01319667		
	-0,035429883		
	-0,000537393		
	0,000365859		
	0,001290601		
	0,020379902		
	0,019845575		
	0,002892684		
d^Q_4			
-5,247905	0,251712619	25,479348	-115,520432

-0,042274	0,014477654	0,5566887	
-0,000949	0,144354238	0,0227655	
-0,105043	0,008254571	0,9740115	
-0,000716	-0,224289403	-0,2422871	
-0,05368 y_6^2	0,001986875	3,4620341	
	0,05095813		
	0,000410761		
	-0,008320673		
	0,0063446		
	-0,000224045		
	0,004537511		
	0,009943849		
	-0,024323355		
	0,005310226		
d^Q_5			
-17,28428	0,687896944	52,459102	-99,292541
-0,10975	0,038603403	1,3351624	
-0,000567	0,528775397	-0,08155	
-0,058484	-0,011316619	0,6145766	
-0,002968	-0,348756075	-0,3974735	
-0,025432 y_6^2	0,000899337	1,5058624	
	-0,006561217		
	0,023790578		
	0,038207169		
	-0,000784319		
	0,000601478		
	0,003123977		
	0,013132049		
	0,002286873		
	-0,005092302		
d^Q_6			
-75,94146	5,637286188	83,232998	-125,723084
-0,320861	0,295126158	-2,0668953	
-0,001545	-1,177907579	-0,2140533	
-0,226755	-0,218146371	2,8778651	
-0,005225	0,409205607	-0,2622651	
-0,051862 y_6^2	-0,009221724	3,5985268	
	0,270811415		
	0,038303698		
	0,016303655		

0,013593031
-0,000659141
0,003320304
0,021482312
-0,032742476
0,002371856

Sumário das Classificações Incorretas

Grupo	Indivíduo	Grupo Classificado
G1	5	G6
	9	G3
	10	G3
	16	G5
	23	G3
G2	2	G3
	4	G6
	5	G6
	7	G3
	11	G6
	14	G6
	19	G6
G3	2	G2
	6	G2
	9	G1
	11	G1
	25	G6
	29	G6
G4	6	G5
G5	11	G3
	23	G2
	25	G3
G6	1	G2
	13	G1
	14	G2
	26	G2

Classificação do Novo Indivíduo pela Função Quadrática: GRUPO 1

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

População Classificada pela função Linear	População Original					
	G1	G2	G3	G4	G5	G6
G1	23	0	3	0	0	1
G2	1	18	3	0	1	4
G3	2	1	23	0	2	1
G4	0	0	0	28	0	0
G5	0	2	0	2	27	0
G6	4	9	1	0	0	24
Total Indivíduos	30	30	30	30	30	30
Indivíduos Corretos	23	18	23	29	27	26
% Correta	76,7	60,0	76,7	93,3	90,0	80,0

Escores de classificação lineares para os grupos

	Coeficientes Lineares	Termo Independente
d_1		
	8,586	-44,33
	3,240	
	0,148	
	-1,587	
	-0,131	
	1,098	
d_2		
	75,946	-134,97
	-0,113	
	-0,359	
	7,133	
	-0,870	
	2,792	
d_3		
	35,081	-68,03
	1,921	
	-0,004	
	2,134	
	-0,852	
	0,993	
d_4		
	25,479	-106,34
	0,557	
	0,023	

	0,974	
	-0,242	
	3,462	
d_5		
	52,459	-17,43
	1,335	
	-0,082	
	0,615	
	-0,397	
	1,506	
d_6		
	83,233	-120,09
	-2,067	
	-0,214	
	2,878	
	-0,262	
	3,599	

Sumário das Classificações Incorretas

Grupo	Indivíduo	Grupo Classificado
G1	5	G6
	9	G3
	10	G3
	16	G2
	20	G6
	26	G6
	29	G6
G2	1	G5
	2	G6
	5	G6
	8	G3
	10	G5
	11	G6
	14	G6
	18	G6
	19	G6
	20	G6
	28	G6
	30	G6
G3	2	G2
	8	G2
	9	G1
	11	G1
	14	G1

	25	G2
	28	G6
G4	6	G5
	26	G5
G5	1	G2
	11	G3
	25	G3
G6	1	G3
	4	G2
	6	G2
	26	G2
	27	G1
	29	G2

Classificação do Novo Indivíduo pela Função Linear: GRUPO 1

APÊNDICE 2 - Resultados do processamento do conjunto de dados de GIRASSÓIS - Método da colocação de elementos à parte para classificação

UNIVERSIDADE ESTADUAL PAULISTA
"JULIO DE MESQUITA FILHO"
Campus de Botucatu

SOFTWARE DISCRIMINANTE

GIRASSOIS PARCIAL 20/02/2007

Vetores de Médias por Grupo

$$\begin{bmatrix} 1{,}028 \\ 16{,}900 \\ 82{,}250 \\ 15{,}700 \\ 25{,}232 \\ 35{,}674 \end{bmatrix} \quad \begin{bmatrix} 1{,}618 \\ 23{,}150 \\ 115{,}050 \\ 18{,}350 \\ 71{,}075 \\ 41{,}853 \end{bmatrix} \quad \begin{bmatrix} 1{,}342 \\ 21{,}700 \\ 89{,}150 \\ 17{,}250 \\ 31{,}885 \\ 35{,}308 \end{bmatrix} \quad \begin{bmatrix} 3{,}164 \\ 30{,}650 \\ 202{,}100 \\ 20{,}550 \\ 81{,}513 \\ 31{,}506 \end{bmatrix} \quad \begin{bmatrix} 2{,}014 \\ 23{,}750 \\ 132{,}050 \\ 16{,}950 \\ 43{,}6420 \\ 38{,}313 \end{bmatrix} \quad \begin{bmatrix} 1{,}339 \\ 16{,}950 \\ 111{,}800 \\ 15{,}850 \\ 43{,}157 \\ 42{,}270 \end{bmatrix}$$

Matrizes de Covariâncias por Grupo

$$S_1 = \begin{bmatrix} 0{,}0585 & 0{,}3751 & 2{,}8084 & 0{,}5331 & 3{,}5140 & 0{,}523 \\ 0{,}3751 & 11{,}1474 & 23{,}6053 & 10{,}8105 & 46{,}844 & 1{,}161 \\ 2{,}8084 & 23{,}6053 & 432{,}5132 & 52{,}3947 & 278{,}702 & 37{,}455 \\ 0{,}5331 & 10{,}8105 & 52{,}3947 & 16{,}0105 & 72{,}049 & 6{,}453 \\ 3{,}5140 & 56{,}844 & 278{,}702 & 72{,}049 & 394{,}1579 & 43{,}439 \\ 0{,}523 & 1{,}161 & 37{,}455 & 6{,}453 & 6{,}453 & 36{,}9798 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 0{,}0779 & 1{,}3204 & 3{,}6896 & 0{,}7141 & 7{,}806 & -0{,}478 \\ 1{,}3204 & 52{,}2395 & 79{,}7289 & 32{,}5237 & 297{,}089 & -12{,}5631 \\ 3{,}6896 & 79{,}7289 & 460{,}0500 & 57{,}7184 & 479{,}181 & -4{,}871 \\ 0{,}7141 & 32{,}5237 & 57{,}7184 & 22{,}3447 & 191{,}726 & -6{,}4598 \\ 7{,}806 & 297{,}089 & 479{,}181 & 191{,}726 & 1886{,}0568 & -60{,}291 \\ -0{,}478 & -12{,}563 & -4{,}871 & -6{,}459 & -60{,}291 & 23{,}8076 \end{bmatrix}$$

$$S_3 = \begin{bmatrix} 0{,}0392 & 0{,}3833 & 2{,}2965 & 0{,}0842 & 2{,}107 & 0{,}323 \\ 0{,}3833 & 13{,}9053 & 62{,}8368 & 5{,}5000 & 46{,}061 & 0{,}579 \\ 2{,}2965 & 62{,}8368 & 574{,}5553 & 35{,}5921 & 274{,}439 & 25{,}535 \\ 0{,}0842 & 5{,}5000 & 35{,}5921 & 5{,}4605 & 22{,}748 & -1{,}272 \\ 2{,}1068 & 46{,}0609 & 274{,}4388 & 22{,}7478 & 294{,}0624 & 18{,}274 \\ 0{,}3226 & 0{,}5787 & 25{,}5351 & -1{,}2720 & 18{,}724 & 16{,}0186 \end{bmatrix}$$

$$S_4 = \begin{bmatrix} 0,1694 & 1,1836 & 3,8175 & 0,8235 & 5,7244 & -0,161 \\ 1,1836 & 25,0816 & 79,3000 & 13,2553 & 57,416 & -1,295 \\ 3,8175 & 79,3000 & 1090,8316 & 74,8368 & 501,931 & 40,720 \\ 0,8235 & 13,2553 & 74,8638 & 16,8921 & 100,394 & 2,921 \\ 5,7244 & 57,416 & 501,931 & 100,3938 & 1625,8445 & 78,629 \\ -0,161 & -1,295 & 40,720 & 2,921 & 78,629 & 18,3134 \end{bmatrix}$$

$$S_5 = \begin{bmatrix} 0,507 & 0,4895 & 4,9998 & 0,4134 & 3,4398 & 0,013 \\ 0,4895 & 12,0921 & 88,8026 & 10,3026 & 71,864 & 3,524 \\ 4,9998 & 88,8026 & 2048,4711 & 91,0553 & 650,595 & 116,330 \\ 0,4134 & 10,3026 & 91,0553 & 11,4184 & 82,701 & -0,210 \\ 3,4398 & 71,864 & 650,595 & 82,7012 & 701,1903 & -6,431 \\ 0,013 & 3,524 & 116,330 & -0,210 & -6,431 & 24,3655 \end{bmatrix}$$

$$S_6 = \begin{bmatrix} 0,0114 & 0,1547 & 0,5407 & 0,0687 & -0,0613 & 0,003 \\ 0,1547 & 6,7868 & 6,6737 & 4,8868 & 23,967 & -1,203 \\ 0,5407 & 6,6737 & 446,2737 & 12,2316 & -42,561 & 20,088 \\ 0,0687 & 4,8868 & 12,2316 & 5,6079 & 24,537 & -1,637 \\ -0,0613 & 23,967 & -42,561 & 24,5369 & 228,7361 & -15,098 \\ 0,003 & -1,203 & 20,088 & -1,637 & -15,098 & 10,9739 \end{bmatrix}$$

Matriz de Variação Conjunta *S*

$$S = \begin{bmatrix} 0,0678 & 0,6511 & 3,0254 & 0,4395 & 3,7550 & 0,0371 \\ 0,6511 & 20,2088 & 56,8246 & 12,8798 & 90,5401 & -1,6328 \\ 3,0254 & 56,8246 & 842,1158 & 53,9715 & 357,0478 & 39,2097 \\ 0,4395 & 12,8798 & 53,9715 & 12,9557 & 82,3591 & -0,0189 \\ 3,7550 & 90,5401 & 357,0478 & 82,3591 & 855,0080 & 9,8286 \\ 0,0371 & -1,6328 & 39,2097 & -0,0189 & 9,8286 & 21,7431 \end{bmatrix}$$

As matrizes de covariância são heterogêneas. Função Adequada: QUADRÁTICA

Quadro sinótico Teste de Normalidade

	V1	V2	V3	V4	V5	V6
G1	S	S	S	S	S	S
G2	S	S	S	S	S	S
G3	S	S	S	S	S	S
G4	S	S	S	S	S	S
G5	S	S	S	S	S	S
G6	S	S	S	S	S	S

Matriz de classificações dos indivíduos pela função discriminante quadrática e respectivas taxas percentuais de classificação

População Classificada pela função Quadrática	População Original					
	G1	G2	G3	G4	G5	G6
G1	18	0	3	0	0	1
G2	0	14	1	0	0	1
G3	1	1	16	0	1	0
G4	0	0	0	19	0	0
G5	0	0	0	1	19	0
G6	1	5	0	0	0	18
Total Indivíduos	20	20	20	2	20	20
Indivíduos Corretos	18	14	16	19	19	18
% Correta	90,0	70,0	80,0	95,0	95,0	90,0

Escores de classificação quadráticos para os grupos

	Coeficientes Quadráticos	Coeficientes Mistos	Coeficientes Lineares	Termo Independente
d^Q_1				
	-17,91639	0,304757028	8,5862141	-52,2753412
	-0,183308	0,044144923	3,2401584	
	-0,001769	0,597865704	0,147682	
	-0,141919	0,035460851	-1,5871091	
	-0,000795	0,246617415	-0,1306709	
	-0,017775	-0,002589275	1,0979553	
		0,240322593		

	0,002750169		
	-0,02981669		
	0,006285468		
	0,000716264		
	0,000594918		
	0,00277473		
	0,022968743		
	-0,000152343		
d^{Q}_{2}			
-21,548	1,026014806	75,94627	-142,10377
-0,177473	0,221094533	-0,1127479	
-0,00249	-2,600391895	-0,3593301	
-0,447382	0,192654746	7,1332472	
-0,004153	-0,468374224	-0,869539	
-0,029734	0,00170722	2,7918553	
	0,332809138		
	0,018516642		
	-0,025445493		
	0,022156221		
	-0,001876723		
	0,005800307		
	0,043287178		
	0,008309113		
	0,003249167		
d^{Q}_{3}			
-31,38187	0,72409993	35,081041	-75,3235645
-0,081981	0,074554064	1,9207763	
-0,001102	-0,596899045	-0,0043496	
-0,126584	0,306014123	2,134288	
-0,005174	0,773609061	-0,8520241	
-0,025796	0,002707863	0,9926958	
	0,075073642		
	0,01319667		
	-0,035429883		
	-0,000537393		
	0,000365859		
	0,001290601		
	0,020379902		
	0,019845575		
	0,002892684		

d^Q_4			
-5,247905	0,251712619	25,479348	-115,520432
-0,042274	0,014477654	0,5566887	
-0,000949	0,144354238	0,0227655	
-0,105043	0,008254571	0,9740115	
-0,000716	-0,224289403	-0,2422871	
-0,05368	0,001986875	3,4620341	
	0,05095813		
	0,000410761		
	-0,008320673		
	0,0063446		
	-0,000224045		
	0,004537511		
	0,009943849		
	-0,024323355		
	0,005310226		
d^Q_5			
-17,28428	0,687896944	52,459102	-99,292541
-0,10975	0,038603403	1,3351624	
-0,000567	0,528775397	-0,08155	
-0,058484	-0,011316619	0,6145766	
-0,002968	-0,348756075	-0,3974735	
-0,025432	0,000899337	1,5058624	
	-0,006561217		
	0,023790578		
	0,038207169		
	-0,000784319		
	0,000601478		
	0,003123977		
	0,013132049		
	0,002286873		
	-0,005092302		
d^Q_6			
-75,94146	5,637286188	83,232998	-125,723084
-0,320861	0,295126158	-2,0668953	
-0,001545	-1,177907579	-0,2140533	
-0,226755	-0,218146371	2,8778651	
-0,005225	0,409205607	-0,2622651	
-0,051862	-0,009221724	3,5985268	
	0,270811415		

0,038303698
0,016303655
0,013593031
-0,000659141
0,003320304
0,021482312
-0,032742476
0,002371856

Sumário das Classificações Incorretas

Grupo	Indivíduo	Grupo Classificado
G1	5	G6
	15	G3
G2	2	G6
	4	G6
	5	G6
	7	G3
	14	G6
	19	G6
G3	6	G2
	11	G1
	19	G1
	20	G1
G4	6	G5
G5	11	G3
G6	1	G1
	14	G2

Classificação do Novo Indivíduo pela Função Quadrática: GRUPO 1

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

População Classificada pela função Linear	População Original					
	G1	G2	G3	G4	G5	G6
G1	15	0	3	0	0	0
G2	1	10	2	0	0	2
G3	2	1	15	0	2	1
G4	0	0	0	19	0	0
G5	0	2	0	1	18	0
G6	2	7	0	0	0	17
Total Indivíduos	20	20	20	20	20	20
Indivíduos Corretos	15	10	15	19	15	17
% Correta	75,0	50,0	75,0	95,0	90,0	85,0

Escores de classificação lineares para os grupos

	Coeficientes Lineares	Termo Independente
d_1		
	14,247	-63,45
	0,661	
	-0,126	
	2,761	
	-0,340	
	2,050	
d_2		
	21,391	-80,39
	0,940	
	-0,125	
	2,112	
	-0,288	
	2,316	
d_3		
	18,507	-72,79
	1,029	
	-0,138	
	2,622	
	-0,372	
	2,088	

d_4		
	49,085	-115,97
	0,726	
	-0,003	
	1,484	
	-0,357	
	1,588	
d_5		
	30,416	-87,67
	0,981	
	-0,091	
	2,015	
	-0,367	
	2,115	
d_6		
	19,404	-74,17
	0,500	
	-0,104	
	2,457	
	-0,307	
	2,277	

Sumário das Classificações Incorretas

Grupo	Indivíduo	Grupo Classificado
G1	5	G6
	9	G3
	10	G3
	16	G2
	20	G6
G2	1	G5
	2	G6
	4	G6
	5	G6
	8	G3
	11	G6
	14	G6
	16	G5
	18	G6
	20	G6
G3	2	G2
	8	G2

	9	G1
	11	G1
	14	G1
G4	6	G5
G5	1	G3
	11	G3
G6	1	G3
	14	G2
	16	G2

Classificação do Novo Indivíduo pela Função Linear: GRUPO 1

APÊNDICE 3 - Resultados do processamento do conjunto de dados de *Eucalyptus* - Método da Ressubstituição

UNIVERSIDADE ESTADUAL PAULISTA
"JULIO DE MESQUITA FILHO"
Campos de Botucatu

SOFTWARE DISCRIMINANTE

EUCALYPTUS 20/02/2007

Vetores de Médias por Grupo

$$\begin{bmatrix}42{,}059\\2{,}614\\13{,}453\\22{,}041\\34{,}537\\85{,}768\\33{,}176\end{bmatrix}\quad\begin{bmatrix}46{,}080\\2{,}970\\12{,}866\\20{,}582\\36{,}721\\86{,}244\\32{,}829\end{bmatrix}\quad\begin{bmatrix}45{,}333\\2{,}923\\12{,}843\\20{,}170\\38{,}212\\87{,}362\\32{,}789\end{bmatrix}\quad\begin{bmatrix}46{,}587\\2{,}566\\13{,}673\\18{,}841\\35{,}586\\86{,}060\\32{,}786\end{bmatrix}\quad\begin{bmatrix}50{,}192\\3{,}047\\14{,}012\\20{,}519\\38{,}8405\\86{,}602\\32{,}787\end{bmatrix}$$

Matrizes de Covariâncias por Grupo

$$S_1 = \begin{bmatrix}251{,}6838 & -7{,}2521 & 29{,}2748 & 68{,}8724 & -54{,}1621 & -38{,}094 & -0{,}238\\ -7{,}2521 & 0{,}2854 & -0{,}6654 & -1{,}7140 & 1{,}756 & 1{,}029 & -0{,}062\\ 29{,}2748 & -0{,}6654 & 5{,}1655 & 10{,}2868 & -5{,}489 & -4{,}504 & 0{,}419\\ 68{,}8724 & -1{,}7140 & 10{,}2868 & 25{,}0263 & -13{,}9418 & -10{,}419 & 0{,}217\\ -54{,}1621 & 1{,}756 & -5{,}489 & -13{,}9418 & 14{,}0255 & 9{,}775 & 0{,}813\\ -38{,}094 & 1{,}029 & -4{,}504 & -10{,}419 & 9{,}775 & 9{,}9171 & 0{,}898\\ -0{,}238 & -0{,}062 & 0{,}419 & 0{,}217 & 0{,}813 & 0{,}898 & 2{,}6007\end{bmatrix}$$

$$S_2 = \begin{bmatrix}380{,}7433 & -6{,}9621 & 38{,}9220 & 99{,}6615 & -57{,}709 & -27{,}974 & -0{,}626\\ -6{,}9621 & 0{,}2672 & -0{,}7371 & -1{,}8400 & 1{,}447 & 0{,}744 & 0{,}049\\ 38{,}9220 & -0{,}7371 & 4{,}9174 & 11{,}1359 & -5{,}740 & -2{,}608 & -0{,}100\\ 99{,}6615 & -1{,}8400 & 11{,}1359 & 28{,}9489 & -14{,}579 & -6{,}529 & -0{,}159\\ -57{,}709 & 1{,}447 & -5{,}740 & -14{,}579 & 12{,}2838 & 6{,}916 & 0{,}219\\ -27{,}974 & 0{,}744 & -2{,}608 & -6{,}529 & 6{,}916 & 5{,}0026 & 0{,}1114\\ -0{,}626 & 0{,}049 & -0{,}100 & -0{,}159 & 0{,}219 & 0{,}1114 & 0{,}0317\end{bmatrix}$$

$$S_3 = \begin{bmatrix} 503{,}333 & -6{,}0921 & 51{,}8361 & 107{,}8800 & -39{,}6736 & -33{,}2424 & 0{,}0980 \\ -6{,}0921 & 0{,}2086 & -0{,}3820 & -1{,}0245 & 0{,}8098 & 0{,}3393 & -0{,}0029 \\ 51{,}8361 & -0{,}3820 & 7{,}0724 & 14{,}1568 & -3{,}4595 & -3{,}7365 & -0{,}0048 \\ 107{,}8800 & -1{,}0245 & 14{,}1568 & 32{,}5744 & -7{,}9393 & -6{,}7549 & 0{,}0153 \\ -39{,}6736 & 0{,}8098 & -3{,}4595 & -7{,}9393 & 4{,}7745 & 2{,}978 & -0{,}010 \\ -33{,}2424 & 0{,}3393 & -3{,}7365 & -6{,}7549 & 2{,}978 & 3{,}0519 & -0{,}003 \\ 0{,}0980 & -0{,}0029 & -0{,}0048 & 0{,}0153 & -0{,}010 & -0{,}003 & 0{,}0007 \end{bmatrix}$$

$$S_4 = \begin{bmatrix} 428{,}4396 & -2{,}7230 & 53{,}8489 & 91{,}1710 & -43{,}2434 & -13{,}802 & 0{,}121 \\ -2{,}7230 & 0{,}1181 & -0{,}5728 & -1{,}0639 & 0{,}530 & 0{,}145 & -0{,}006 \\ 53{,}8489 & -0{,}5728 & 8{,}9322 & 13{,}6983 & -6{,}415 & -1{,}708 & 0{,}031 \\ 91{,}1710 & -1{,}0639 & 13{,}6983 & 29{,}4296 & -9{,}7869 & -2{,}673 & 0{,}063 \\ -43{,}2434 & 0{,}530 & -6{,}415 & -9{,}7869 & 6{,}1561 & 2{,}419 & -0{,}030 \\ -13{,}802 & 0{,}145 & -1{,}708 & -2{,}673 & 2{,}419 & 2{,}4659 & -0{,}0095 \\ 0{,}121 & -0{,}006 & 0{,}031 & 0{,}063 & -0{,}030 & -0{,}0095 & 0{,}0010 \end{bmatrix}$$

$$S_5 = \begin{bmatrix} 338{,}2415 & 0{,}7917 & 37{,}8140 & 82{,}7242 & 5{,}6314 & 10{,}892 & -0{,}098 \\ 0{,}7917 & 0{,}1548 & 0{,}1838 & 0{,}3244 & 0{,}096 & 0{,}171 & -0{,}00022 \\ 37{,}8140 & 0{,}1838 & 5{,}1452 & 10{,}5411 & 0{,}651 & 0{,}605 & -0{,}012 \\ 82{,}7242 & 0{,}3244 & 10{,}5411 & 24{,}7947 & 1{,}3093 & 0{,}584 & -0{,}015 \\ 5{,}6314 & 0{,}096 & 0{,}651 & 1{,}3093 & 0{,}7206 & 0{,}497 & -0{,}001 \\ 10{,}892 & 0{,}171 & 0{,}605 & 0{,}584 & 0{,}497 & 21{,}3681 & 0{,}0040 \\ -0{,}098 & -0{,}00022 & -0{,}012 & -0{,}015 & -0{,}001 & 0{,}0040 & 0{,}0006 \end{bmatrix}$$

Matriz de Variação Conjunta *S*

$$S = \begin{bmatrix} 367{,}7214 & -4{,}1095 & 40{,}6529 & 89{,}1443 & -34{,}9372 & -17{,}4570 & -0{,}2094 \\ -4{,}1095 & 0{,}2099 & -0{,}3977 & -0{,}9952 & 0{,}9104 & 0{,}4967 & 0{,}0008 \\ 40{,}6529 & -0{,}3977 & 5{,}8788 & 11{,}5638 & -3{,}7165 & -2{,}0544 & 0{,}0487 \\ 89{,}1443 & -0{,}9952 & 11{,}5638 & 27{,}5951 & -8{,}4809 & -4{,}6950 & 0{,}0032 \\ -34{,}9372 & 0{,}9104 & -3{,}7165 & -8{,}4809 & 7{,}5256 & 4{,}4833 & 0{,}1993 \\ -17{,}4570 & 0{,}4967 & -2{,}0544 & -4{,}6950 & 4{,}4833 & 9{,}8715 & 0{,}1907 \\ -0{,}2094 & 0{,}0008 & 0{,}0487 & 0{,}0032 & 0{,}1993 & 0{,}1907 & 0{,}4765 \end{bmatrix}$$

As matrizes de covariância são heterogêneas. Função Adequada: QUADRÁTICA

Quadro sinótico Teste de Normalidade

	V1	V2	V3	V4	V5	V6	V7
G1	S	S	S	S	S	S	S
G2	S	S	S	S	S	S	S
G3	S	S	S	S	S	S	S
G4	S	S	S	S	S	S	S
G5	S	S	S	S	S	S	S

Matriz de classificações dos indivíduos pela função discriminante quadrática e respectivas taxas percentuais de classificação

População Classificada pela função Quadrática	População Original				
	G1	G2	G3	G4	G5
G1	9	0	0	0	0
G2	1	15	0	0	0
G3	4	4	11	0	4
G4	3	5	0	14	0
G5	0	1	1	0	22
Total Indivíduos	17	25	12	14	26
Indivíduos Corretos	9	15	11	14	22
% Correta	52,9	60,0	91,7	100,0	84,6

Escores de classificação quadráticos para os grupos

	Coeficientes Quadráticos	Coeficientes Mistos	Coeficientes Lineares	Termo Independente
d^Q_1				
	-0,0327503	-0,58135952	5,951900371	-1375,278727949
	-13,709284	0,1167813	124,6751557	
	-0,7043466	0,03987109	-1,324619607	
	-0,160854	-0,08861869	-2,119759253	
	-0,539931	-0,00792893	-11,82713279	
	-0,2123223	-0,01165799	26,37999137	
	-0,2605757	1,16358791	11,27587025	
		0,40758104		
		3,02610756		

		-1,28369252		
		-1,43525279		
		0,42130255		
		0,26433982		
		-0,14790185		
		0,19870999		
		-0,08662679		
		0,05124364		
		-0,0181608		
		0,41890376		
		0,22202511		
		0,00366529		
d^{Q}_{2}				
	-0,0200255	0,01826206	-6,456037137	-31813,596721206
	-6,9962183	0,03834295	-389,3782583	
	-0,8833227	0,08969539	105,7377907	
	-0,2928047	-0,07555962	-50,1460117	
	-0,6107939	0,00950243	-77,92693413	
	-0,6055528	0,23951009	96,36110087	
	-27,398165	0,36780976	1798,755631	
		-0,37642647		
		1,49612805		
		-0,47025922		
		12,7548449		
		0,59220313		
		0,10394426		
		-0,06681637		
		-2,92757021		
		-0,07338067		
		0,17693839		
		1,18794317		
		0,9767806		
		1,14999993		
		-0,90168107		
d^{Q}_{3}				
	-0,0120212	-0,04587553	-151,6225963	-2569263,385488163
	-16,730929	0,24088626	6911,32588	
	-5,7979387	-0,04590307	6887,275865	
	-0,7025155	-0,1223027	-2280,551548	
	-2,9626197	0,06056546	-4125,662548	

-4,3221138	4,57658078	5189,086868	
-2393,2684	-10,0139741	166372,2398	
	3,88635929		
	11,5283891		
	-11,8780153		
	-187,991635		
	3,857034		
	6,75637808		
	-8,71163471		
	-191,982658		
	-2,34720088		
	3,03525687		
	63,2714037		
	6,36039419		
	113,722453		
	-4786,53687		
d^Q_4			
-0,0095563	0,31548729	47,63592213	-964466,464996161
-11,590377	0,02347761	1314,955737	
-0,4583233	0,02985311	54,63401037	
-0,0874737	-0,10422426	-119,6644095	
-0,9532252	0,01984592	354,0725935	
-0,4373443	-1,41798415	-12,79714864	
0,87404647	0,14903189	52266,60121	
	-0,78639929		
	3,33713078		
	-1,04560035		
	-39,2033284		
	0,1704067		
	-0,66504129		
	0,31886915		
	-1,54027958		
	0,21699952		
	-0,05716989		
	3,61550699		
	0,8230725		
	1,74809294		
	-1583,46001		
d^Q_5			
-0,0113261	-0,09646549	34,53653028	-1066269,093292882

-3,8156871	0,07721304	7,703538896
-0,9891656	0,0411445	422,0465026
-0,2047758	0,0378043	-256,005589
-0,869963	0,00834305	-11,76390074
-0,0256385	-1,1352933	-29,77978154
-940,68419	0,73036265	61582,13059
	0,05820375	
	0,96227863	
	0,06583764	
	-1,04041082	
	0,56422799	
	0,04263527	
	-0,00329307	
	-12,6079945	
	-0,06394613	
	-0,02619708	
	7,89992676	
	0,01359494	
	2,25856306	
	1,026534	

Sumário das Classificações Incorretas

Grupo	Indivíduo	Grupo Classificado
G1	1	G3
	3	G3
	4	G4
	10	G3
	11	G4
	13	G3
	15	G2
	17	G4
G2	4	G3
	6	G4
	8	G4
	9	G4
	10	G5
	18	G3
	21	G4
	22	G3
	23	G4
	25	G3
G3	12	G5

G5	6	G3
	8	G3
	11	G3
	26	G3

Classificação do Novo Indivíduo pela Função Quadrática: GRUPO 1

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

População Classificada pela função Linear	População Original				
	G1	G2	G3	G4	G5
G1	13	2	0	1	0
G2	2	13	2	2	3
G3	1	6	7	0	2
G4	1	2	0	10	3
G5	0	2	3	1	18
Total Indivíduos	17	25	12	14	26
Indivíduos Corretos	13	13	7	10	18
% Correta	76,5	52,0	58,3	71,4	69,2

Escores de classificação lineares para os grupos

	Coeficientes Lineares	Termo Independente
d_1		
	0,057	-1490,95030373
	2,112	
	1,607	
	1,566	
	0,484	
	8,267	
	65,963	
d_2		
	0,252	-1472,520557044
	2,882	
	1,123	
	1,293	
	1,162	
	8,099	
	65,154	

d_3		
	0,283	-1480,29571325
	0,403	
	1,334	
	1,200	
	1,810	
	8,104	
	64,793	
d_4		
	0,184	-1465,87385467
	-0,473	
	2,911	
	0,591	
	1,155	
	8,176	
	64,832	
d_5		
	0,319	-1479,26003995
	0,371	
	2,257	
	0,786	
	2,163	
	7,931	
	64,634	

Sumário das Classificações Incorretas

Grupo	Indivíduo	Grupo Classificado
G1	3	G2
	4	G4
	13	G3
	15	G2
G2	1	G3
	4	G3
	8	G4
	9	G4
	10	G5
	12	G1
	17	G3
	19	G3
	20	G5
	21	G3

	22	G3
	25	G1
G3	2	G2
	4	G5
	5	G5
	6	G2
	12	G5
G4	1	G2
	9	G2
	11	G5
	12	G1
G5	1	G4
	6	G3
	8	G2
	9	G4
	10	G2
	13	G3
	18	G4
	19	G2

Classificação do Novo Indivíduo pela Função Linear: GRUPO 1

APÊNDICE 4 - Resultados do processamento do conjunto de dados de *EUCALYPTUS* - Método da colocação de elementos à parte para classificação

---

UNIVERSIDADE ESTADUAL PAULISTA
"JULIO DE MESQUITA FILHO"
Campus de Botucatu

SOFTWARE DISCRIMINANTE

EUCALYPTOS PARCIAL 20/02/2007

Vetores de Médias por Grupo

$$\begin{bmatrix} 40{,}231 \\ 2{,}596 \\ 13{,}231 \\ 21{,}562 \\ 34{,}655 \\ 86{,}275 \\ 33{,}298 \end{bmatrix} \quad \begin{bmatrix} 49{,}263 \\ 2{,}996 \\ 13{,}176 \\ 21{,}198 \\ 36{,}257 \\ 86{,}016 \\ 32{,}844 \end{bmatrix} \quad \begin{bmatrix} 48{,}000 \\ 2{,}872 \\ 12{,}993 \\ 20{,}749 \\ 37{,}612 \\ 87{,}073 \\ 32{,}793 \end{bmatrix} \quad \begin{bmatrix} 39{,}000 \\ 2{,}536 \\ 13{,}359 \\ 17{,}228 \\ 35{,}577 \\ 85{,}979 \\ 32{,}790 \end{bmatrix} \quad \begin{bmatrix} 46{,}263 \\ 3{,}066 \\ 13{,}529 \\ 19{,}479 \\ 38{,}801 \\ 86{,}770 \\ 32{,}793 \end{bmatrix}$$

Matrizes de Covariâncias por Grupo

$$S_1 = \begin{bmatrix} 296{,}1923 & -8{,}5857 & 38{,}4531 & 85{,}9596 & -67{,}3013 & -49{,}8780 & 0{,}6947 \\ -8{,}5857 & 0{,}3159 & -0{,}9755 & -2{,}2302 & 2{,}1537 & 1{,}5957 & -0{,}0736 \\ 38{,}4531 & -0{,}9755 & 6{,}2065 & 12{,}9846 & -7{,}3949 & -5{,}5174 & 0{,}6688 \\ 85{,}9596 & -2{,}2302 & 12{,}9846 & 31{,}7809 & -18{,}1123 & -13{,}5521 & 0{,}5477 \\ -67{,}3013 & 2{,}1537 & -7{,}3949 & -18{,}1123 & 17{,}8912 & 12{,}9531 & 1{,}0248 \\ -49{,}8780 & 1{,}5957 & -5{,}5174 & -13{,}5521 & 12{,}9531 & 10{,}1676 & 0{,}9257 \\ 0{,}6947 & -0{,}0736 & 0{,}6688 & 0{,}5477 & 1{,}0248 & 0{,}9257 & 3{,}3957 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 393{,}8713 & -8{,}2440 & 39{,}8655 & 103{,}9873 & -61{,}3559 & -31{,}0722 & -1{,}0084 \\ -8{,}2440 & 0{,}2859 & -0{,}9434 & -2{,}1686 & 1{,}6891 & 0{,}9382 & 0{,}0615 \\ 39{,}8655 & -0{,}9434 & 4{,}7919 & 11{,}0177 & -6{,}5026 & -3{,}1596 & -0{,}1484 \\ 103{,}9873 & -2{,}1686 & 11{,}0177 & 29{,}4228 & -15{,}7965 & -7{,}6676 & -0{,}2402 \\ -61{,}3559 & 1{,}6891 & -6{,}5026 & -15{,}7965 & 13{,}0548 & 7{,}9978 & 0{,}3243 \\ -31{,}0722 & 0{,}9382 & -3{,}1596 & -7{,}6676 & 7{,}9978 & 5{,}8394 & 0{,}1723 \\ -1{,}0084 & 0{,}0615 & -0{,}1484 & -0{,}2402 & 0{,}3243 & 0{,}1723 & 0{,}0410 \end{bmatrix}$$

$$S_3 = \begin{bmatrix} 593,7500 & -6,7600 & 64,9013 & 136,0050 & -43,7462 & -35,8825 & 0,0262 \\ -6,7600 & 0,2568 & -0,4255 & -1,1847 & 0,8872 & 0,2972 & -0,0029 \\ 64,9013 & -0,4255 & 9,3031 & 18,7155 & -4,1434 & -4,5127 & -0,0131 \\ 136,0050 & -1,1847 & 18,7155 & 42,8160 & -8,9533 & -7,9994 & 0,0085 \\ -43,7462 & 0,8872 & -4,1434 & -8,9533 & 4,3779 & 2,8965 & -0,0098 \\ -35,8825 & 0,2972 & -4,5127 & -7,9994 & 2,8965 & 3,1964 & 0,0038 \\ 0,0262 & -0,0029 & -0,0131 & 0,0085 & -0,0098 & 0,0038 & 0,0004 \end{bmatrix}$$

$$S_4 = \begin{bmatrix} 468,8889 & -4,9167 & 69,5256 & 96,4267 & -60,2489 & -18,1433 & 0,2533 \\ -4,9167 & 0,1467 & -0,7854 & -1,4256 & 0,7196 & 0,1869 & -0,0071 \\ 69,5256 & -0,7854 & 11,8746 & 15,9668 & -8,9000 & -2,2456 & 0,0485 \\ 96,4267 & -1,4256 & 15,9668 & 25,0697 & -12,5808 & -3,3268 & 0,0973 \\ -60,2489 & 0,7196 & -8,9000 & -12,5808 & 8,4840 & 3,2001 & -0,0371 \\ -18,1433 & 0,1869 & -2,2456 & -3,3268 & 3,2001 & 3,2610 & -0,0064 \\ 0,2533 & -0,0071 & 0,0485 & 0,0973 & -0,0371 & -0,0064 & 0,0011 \end{bmatrix}$$

$$S_5 = \begin{bmatrix} 374,9825 & 1,2366 & 43,0803 & 96,1447 & 6,1165 & 8,0244 & -0,0063 \\ 1,2366 & 0,1898 & 0,2588 & 0,4813 & 0,0804 & 0,0375 & 0,0009 \\ 48,0803 & 0,2588 & 6,0239 & 12,4733 & 0,7121 & 0,5267 & -0,0028 \\ 96,1447 & 0,4813 & 12,4733 & 29,0399 & 1,5388 & 0,6997 & 0,0077 \\ 6,1165 & 0,0804 & 0,7121 & 1,5388 & 0,8027 & -0,1866 & 0,0037 \\ 8,0244 & 0,0375 & 0,5267 & 0,6997 & -0,1866 & 23,5984 & 0,0264 \\ -0,0063 & 0,0009 & -0,0028 & 0,0077 & 0,0037 & 0,0264 & 0,0005 \end{bmatrix}$$

Matriz de Variação Conjunta *S*

$$S = \begin{bmatrix} 405,5950 & -5,0383 & 47,6831 & 101,3811 & -41,4482 & -22,5192 & -0,1144 \\ -5,0383 & 0,2420 & -0,5308 & -1,2222 & 1,0964 & 0,6272 & 0,0024 \\ 47,6831 & -0,5308 & 6,9301 & 13,4166 & -4,7110 & -2,6140 & 0,0867 \\ 101,3811 & -1,2222 & 13,4166 & 30,7977 & -10,1360 & -5,8767 & 0,0512 \\ -41,4482 & 1,0964 & -4,7110 & -10,1360 & 8,8540 & 5,3541 & 0,2737 \\ -22,5192 & 0,6272 & -2,6140 & -5,8767 & 5,3541 & 10,8740 & 0,2255 \\ -0,1144 & 0,0024 & 0,0867 & 0,0512 & 0,2737 & 0,2255 & 0,6386 \end{bmatrix}$$

As matrizes de covariância são heterogêneas. Função Adequada: QUADRÁTICA

Quadro sinótico Teste de Normalidade

	V1	V2	V3	V4	V5	V6	V7
G1	S	S	S	S	S	S	S
G2	S	S	S	S	S	S	S
G3	S	S	S	S	S	S	S
G4	S	S	S	S	S	S	S
G5	S	S	S	S	S	S	S

Matriz de classificações dos indivíduos pela função discriminante quadrática e respectivas taxas percentuais de classificação

População Classificada pela função Quadrática	População Original				
	G1	G2	G3	G4	G5
G1	9	0	0	0	0
G2	2	17	0	0	1
G3	0	0	9	0	0
G4	1	1	0	10	0
G5	1	1	0	0	18
Total Indivíduos	13	19	9	10	19
Indivíduos Corretos	9	17	9	10	18
% Correta	69,2	89,5	100,0	100,0	94,7

Escores de classificação quadráticos para os grupos

	Coeficientes Quadráticos	Coeficientes Mistos	Coeficientes Lineares	Termo Independente
d^Q_1				
	-0,084419256	-0,376722089	20,370076309	-4129,70395
	-12,889774791	0,584193623	-47,546645413	
	-1,631175775	-0,024904794	-71,819532206	
	-0,140513632	-0,281306215	10,737158501	
	-0,694635784	-0,130194942	-8,225827558	
	-0,761025417	0,035723313	102,459601610	
	-0,222045581	0,206633533	-8,426098587	
		0,526378804		
		1,444727316		

	1,298045729		
	-1,396802237		
	0,578600971		
	1,053982591		
	0,493223813		
	-0,018344571		
	-0,130563719		
	-0,100377595		
	0,014629120		
	0,545350050		
	0,172920935		
	0,224132197		
d^{Q}_{2}			
-0,028177794	0,079351572	-13,128101558	-32594,165757121
-10,758935604	0,006087332	-280,257939132	
-1,028167599	0,135455337	1,807954322	
-0,331494743	-0,184315759	-32,468318167	
-1,739652748	0,112221018	-253,205242516	
-1,377003371	0,296705007	245,243282927	
-24,081263149	-1,262635658	1687,422094865	
	-0,005608866		
	3,066207537		
	-1,335393111		
	10,999540566		
	0,486429487		
	-0,592770386		
	0,582311007		
	-0,308626271		
	0,011771799		
	0,074253932		
	0,811192409		
	2,860184899		
	4,286547982		
	-3,739832301		
d^{Q}_{3}			
-0,037103267	6,415149504	-121,148258137	-38969725,991295
-431,466352555	1,107364327	-81998,651652699	
-28,548298623	-0,310861169	54600,719253252	
-3,164756280	-2,997183983	-18753,476949062	
-83,950304068	2,070767213	29119,197127504	

-39,526878877	0,938247196	7315,596930684	
-34648,958051805	-53,988039599	2321454,450430152	
	13,975743666		
	378,093430705		
	-234,882684805		
	2769,216708139		
	18,837168109		
	30,754415490		
	-42,025094969		
	-1574,870677644		
	-8,380296882		
	12,917707437		
	542,953140941		
	106,991546019		
	-1015,084877182		
	-109,871603076		
d^Q_4			
-0,026086288	0,308448430	93,737457545	-935373,8564281
-10,596432833	0,053378771	208,340477087	
-0,572197458	0,053634322	182,924873230	
-0,311926748	-0,300747899	-507,883794712	
-1,446300080	0,073294797	387,420176595	
-0,383713054	-2,736326828	-124,680381803	
-870,648505606	1,567065923	57023,659543644	
	-1,934069268		
	3,137378360		
	-1,052552335		
	-5,347979856		
	0,509314427		
	-0,313100820		
	0,234537075		
	-5,839990761		
	0,261176478		
	-0,102051024		
	15,679263246		
	1,014356709		
	-11,231094057		
	4,666388302		
d^Q_5			

-0,010801943	-0,063134835	39,857530931	-1358901,0377056
-2,999864766	0,045570315	-322,975882747	
-0,955912160	0,051051098	764,503494703	
-0,241008250	0,040209500	-449,923396379	
-0,786860008	0,006679438	-452,525822474	
-0,024349731	-1,293435088	-110,145930727	
-1287,973183434	0,688610844	83631,041389357	
	-0,014234059		
	0,455525206		
	0,008857150		
	9,661063527		
	0,657060575		
	0,136101528		
	0,033714387		
	-23,293748288		
	-0,118861134		
	-0,034183431		
	13,895898027		
	-0,043894126		
	15,692935659		
	3,535802372		

Sumário das Classificações Incorretas

Grupo	Indivíduo	Grupo Classificado
G1	3	G5
	4	G4
	6	G2
	13	G2
G2	8	G4
	10	G5
G5	8	G2

Classificação do Novo Indivíduo pela Função Quadrática: GRUPO 1

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

População Classificada pela função Linear	População Original				
	G1	G2	G3	G4	G5
G1	10	0	0	0	0
G2	0	13	0	1	3
G3	2	4	7	1	2
G4	1	1	0	8	3
G5	0	1	2	0	11
Total Indivíduos	13	19	9	10	19
Indivíduos Corretos	10	13	7	8	11
% Correta	76,9	68,4	77,8	80,0	57,9

Escores de classificação lineares para os grupos

	Coeficientes Lineares	Termo Independente
d_1		
	0,106	-1192,16867461
	0,422	
	0,748	
	1,540	
	0,136	
	8,066	
	48,986	
d_2		
	0,346	-1170,71153026
	2,191	
	0,240	
	1,167	
	0,669	
	7,867	
	48,254	
d_3		
	0,374	-1176,82005693
	-0,730	
	0,365	
	1,107	
	1,316	
	7,877	
	47,897	
d_4		
	0,232	-1161,5667119

	-2,187	
	2,444	
	0,336	
	0,812	
	7,897	
	47,863	
d_5		
	0,384	-1174,69778473
	-0,784	
	1,427	
	0,646	
	1,638	
	7,724	
	47,709	

Sumário das Classificações Incorretas

Grupo	Indivíduos	Grupo Classificado
G1	3	G3
	4	G4
	13	G3
G2	1	G3
	4	G3
	8	G4
	10	G5
	17	G3
	19	G3
G3	4	G5
	6	G5
G4	1	G2
	9	G3
G5	1	G4
	6	G3
	8	G2
	9	G4
	10	G2
	13	G3
	18	G4
	19	G2

Classificação do Novo Indivíduo pela Função Linear: GRUPO 1

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo