UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

CARACTERIZAÇÃO TEÓRICA E APLICAÇÃO DA ANÁLISE DISCRIMINANTE

Autor: Maria Conceição de Matos Braga
Orientadora: Professora Jandyra M. G. Fachel

Monografia apresentada para a obtenção do grau de Bacharel em Estatística

MONOGRAFIA/EST
B813c

Porto Alegre, Agosto de 2003

Agradecimentos

Este trabalho surgiu como conclusão do curso de Bacharelado em Estatística, e também pela necessidade de usar a técnica de Análise Discriminante em dados do Projeto MAPEM.

Assim, essa monografia e principalmente o curso de Estatística, puderam ser concluídos graças ao interesse que gradativamente veio aumentando no decorrer do trabalho, graças à ajuda de todas as pessoas que acabaram sendo envolvidas, de uma forma ou de outra.

Agradeço à orientadora deste trabalho, professora Jandyra M. G. Fachel, ao professor João Riboldi pela sua participação, e a todos os professores do Departamento de Estatística pela dedicação e pelo desprendimento com que sempre transmitiram seus conhecimentos. Agradeço aos colegas do curso pela amizade e disponibilidade nas ocasiões mais conturbadas.

À minha mãe, Elisa e ao meu pai, Wilson agradeço pelo apoio incondicional. A João, meu marido, agradeço pelo companheirismo e compreensão. Agradeço também ao meu filho Tiago que sempre me incentivou .

Agradeço, enfim, aos amigos, aos colegas de trabalho pelas concessões especiais e a toda a minha família pelo incentivo e por terem acreditado em mim até mesmo nos momentos em que eu mesma não acreditava.

SUMÁRIO

1. Introdução 5

1.1 O que é Análise Discriminante e para que serve 5

1.2 Objetivos do Trabalho 6

1.3 Estrutura do Trabalho 7

2. Revisão Bibliográfica 9

3. Análise Discriminante 11

3.1 Objetivos e considerações da Análise Discriminante 11

3.2 Suposições para a Análise Discriminante 13

3.3 Estimação do Modelo Discriminante 14

3.3.1 Métodos de Seleção: Direto e Stepwise 14

3.3.2 Determinação do número de funções discriminantes 17

3.3.3 Interpretação dos coeficientes da função discriminante 19

3.3.4 Gráficos e rotação das funções discriminantes 20

3.3.5 Classificação dos casos 22

3.4 Métodos para Interpretação dos Resultados 25

3.5 Validação dos Resultados 28

3.6 Modelos Discriminantes 29

3.6.1 Análise Discriminante com dois grupos 30

3.6.2 Análise Discriminante com mais de dois grupos 32
3.6.3 Outros critérios discriminantes 33
3.7 Comparação da abordagem feita na Análise Discriminante com a Regressão Logística 38

4. Utilização de Softwares 41
4.1 Como proceder através do software SPSS 41
4.1.1 Como obter uma Análise Discriminante através do SPSS 42
4.1.2 Opções adicionais disponíveis através da Sintaxe 48

5. Aplicações práticas e interpretações utilizando dados do Projeto MAPEM 49
5.1 Análise Discriminante (dois grupos) 50
5.2 Análise Discriminante (três grupos) 62

6. Comentários e Sugestões 75

7. Referências Bibliográficas 77

8. Anexos 80
Anexo A: Tabelas relativas a Análise Discriminante com dois grupos 80
Anexo B: Tabelas relativas a Análise Discriminante com três grupos 82

1. Introdução

1.1 O que é Análise Discriminante e para que serve

A *Análise Discriminante* é uma técnica estatística de análise multivariada indicada quando se quer ver se um conjunto de variáveis independentes discriminam os grupos, ou seja, tem comportamento diferenciado entre dois ou mais grupos. Ela tem como objetivo identificar diferenças entre grupos, a partir de um modelo que relaciona uma variável dependente (que classifica o grupo), não métrica ou categórica, às variáveis independentes, métricas. Dessa maneira, obtém-se a combinação linear de variáveis independentes que mais discrimina entre dois ou mais grupos pré-determinados. Quando dois grupos são envolvidos a técnica é conhecida como Análise Discriminante com dois grupos. Quando a técnica envolve três ou mais grupos ela é conhecida como Análise Discriminante Múltipla. A principal diferença entre a Análise Discriminante com dois grupos e a múltipla é que no caso de dois grupos, só é possível derivar uma única função discriminante e no caso da múltipla é possível obter mais de uma função, na verdade (g – 1) funções, onde g é o número de grupos (Hair et al, 1998; p.244).

A Análise Discriminante estabelece quais as variáveis que são mais importantes para distinguir os grupos. Fornece também, a precisão da classificação feita com base nas medidas existentes. A precisão da classificação é calculada comparando-se a percentagem de casos corretamente classificados pelo modelo em relação à classificação original.

A Análise Discriminante foi introduzida por Fisher em 1936 (segundo Norusis, 1985) como uma técnica estatística útil em problemas de taxonomia.

Alguns exemplos, de como poderíamos utilizar a Análise Discriminante, seriam: pessoas que votaram no candidato do Partido dos Trabalhadores na recente eleição contra pessoas que votaram no candidato do PSDB baseada nas variáveis: renda, nível educacional e idade; um grupo experimental de animais de laboratório que recebe determinada droga A contra um grupo que recebe determinada droga B e contra um grupo que não recebe nenhuma droga, utilizando análises bioquímicas do sangue. A análise discriminante pode também ser usada para distinguir os inovadores de não inovadores de acordo com o seu perfil sócio-demográfico e perfis psicológicos. Outras aplicações incluem distinguir os compradores de marcas nacionais de compradores de marcas importadas de acordo com as variáveis idade e renda; e riscos de crédito distinguindo bons e maus pagadores, utilizando-se sempre um vetor de variáveis para classificar os grupos.

Um investigador educacional pode querer investigar quais variáveis discriminam entre os egressos do Ensino Médio que decidem (1) ir à faculdade, (2) ir a uma escola de comércio ou profissional, ou (3) não buscar nenhum treinamento ou instrução adicional. Para essa finalidade o investigador poderia coletar dados utilizando numerosas variáveis em relação aos estudantes. Após o término do ensino médio, a maioria dos estudantes enquadrar-se-ão naturalmente em uma das três categorias. A Análise Discriminante poderia então ser usada, para determinar quais variáveis melhor predizem na escolha educacional subseqüente dos estudantes.

Um investigador médico pode obter dados sobre as diferentes variáveis que se relacionam aos seus pacientes a fim de verificar quais delas predizem melhor se é provável um paciente recuperar-se completamente (grupo 1), parcialmente (grupo 2), ou de nenhum modo (grupo 3).

1.2 Objetivos do Trabalho

A Análise Discriminante é utilizada em diversas áreas tais como: Psicologia, Medicina, Educação, Sociologia, Marketing, Geologia ou qualquer outra. O primeiro trabalho de aceitação da técnica em Geociências foi desenvolvido por Potter, Shimp e Witter (1963)

discriminando argilitos de origem estuarina, marinha e de água doce a partir das concentrações de metais nas argilas.

"O objetivo geral deste trabalho é apresentar de forma consistente a estruturação básica (teórica) da Análise Discriminante e enfocar, de forma específica, uma aplicação da técnica."

A aplicação prática é na área de Geociências com dados do Projeto **MAPEM** (MONITORAMENTO AMBIENTAL EM ATIVIDADES DE PERFURAÇÃO EXPLORATÓRIA MARÍTIMA). Este Projeto é coordenado pelo Instituto de Geociências da UFRGS, tendo como um dos co-executores o Instituto de Matemática da UFRGS, com financiamento FNDCT/CTPETRO/FINEP e IBP. Em um primeiro momento usaremos uma variável dependente com dois grupos e depois uma variável dependente com três grupos, sendo que as variáveis independentes utilizadas para discriminar os grupos serão as mesmas para os dois casos. Esta aplicação proporcionará maior elucidação de alguns conceitos.

1.3 Estrutura do Trabalho

Na introdução, feita no capítulo 1, é apresentada a conceituação de Análise Discriminante, e os objetivos do trabalho.

Através da revisão bibliográfica (capítulo 2) procura-se fornecer uma relação de referências bibliográficas com os diferentes enfoques dados ao assunto.

O capítulo 3 apresenta os objetivos e as suposições que devem ser conhecidas quando propomos a aplicação da Análise Discriminante, a estimação do modelo discriminante, a interpretação, a validação dos resultados, os procedimentos matemáticos dos modelos discriminantes e a comparação da abordagem feita na Análise Discriminante com a Regressão Logística.

Apresentaremos, no capítulo 4, através do *software* SPSS® algumas instruções básicas para a utilização do mesmo na Análise Discriminante.

O capítulo 5 trata principalmente da utilização prática da Análise Discriminante. Apresentamos duas aplicações, utilizando dados do Projeto MAPEM, uma com dois grupos e

outra com 3 grupos. São apresentadas as análises feitas no SPSS e considerações quanto à interpretação dos resultados.

No capítulo 6 são feitas algumas considerações e discussões sobre o tema, bem como sobre o trabalho propriamente dito.

2. Revisão Bibliográfica

Os procedimentos matemáticos utilizados na Análise Discriminante são abordados no livro *Multivariate Data Analysis*, de Murtagh e Heck (1987). Nele os autores apresentam as descrições matemáticas dos seguintes métodos: Análise Discriminante Múltipla, Análise Discriminante Linear, Discriminação Bayesiana, Discriminação de Máxima Verossimilhança e Discriminação Não-paramétrica. Trata-se de uma linguagem complexa e bastante teórica, apesar de apresentar algumas observações práticas e fazer referências de onde se podem encontrar exemplos da utilização da técnica na área da astronomia.

Obras mais recentes, como a de Johnson e Wichern (1998), *Applied Multivariate Statiscal Analysis*, são bastante úteis na descrição da aplicação prática através dos *softwares* estatísticos mais conhecidos. A última parte do capítulo sobre Análise Discriminante apresenta vários exercícios sobre o tema. No caso deste livro, a parte teórica que trata de Análise Discriminante é bastante restrita, fornecendo uma visão geral sobre separação e classificação. Por outro lado, apresenta dados (inclusive em disquete) de um exemplo para ser utilizado no *software* SAS. Todos os comandos são relacionados e as saídas computacionais comentadas.

O manual do SPSS (*Statical Package for the Social Science,1975*), cujo capítulo sobre Análise Discriminante foi escrito por W. R. Klecka introduz o conceito e o procedimento na seleção das variáveis discriminantes. Apresenta um exemplo para dois grupos (situação simples) e um exemplo com quatro grupos. Com estes dois exemplos, discute os componentes estatísticos principais desta técnica em detalhes (determinação do número de funções discriminantes, interpretação dos coeficientes das funções discriminantes, gráficos dos escores discriminantes, rotação da função discriminante, classificação dos casos e

a seleção dos métodos). A ênfase desta referência está no uso destas estatísticas em pesquisa aplicada; sugerindo que o usuário recorra aos textos estatísticos listados ao término do capítulo para a derivação matemática de Análise Discriminante.

Por fim, a obra de Hair, Anderson, Tatham e Black (1998) – *Multivariate Data Analysis* - apresenta três exemplos ilustrativos (análise discriminante com dois grupos, análise discriminante com três grupos e regressão logística onde a variável dependente apresenta duas categorias) e os seus desenvolvimentos através do *software* SPSS. Além disso, faz uma abordagem detalhada da Análise Discriminante dividida em seis estágios: Objetivos da Análise Discriminante; Delineamento da pesquisa para a Análise Discriminante (seleção da variável dependente e das variáveis independentes, tamanho da amostra e divisão da amostra); Suposições que devem ser conhecidas quando propomos a aplicação da Análise Discriminante; Estimação das Funções Discriminantes e Avaliação do Ajuste Global; Interpretação dos Resultados e Validação destes Resultados. É um livro de linguagem objetiva, didática e de fácil interpretação, que apresenta tópicos diferenciados daqueles citados anteriormente, tais como a abordagem comparativa entre Análise Discriminante com dois grupos e a Regressão Logística. Enfim, apresenta um contexto prático, não se detêm em procedimentos matemáticos.

Encontramos em diversos *Sites da Internet* o mesmo tipo de abordagem feito em Hair et al., (1998). A relação dos *Sites* consultados para a elaboração do trabalho encontra-se na Bibliografia.

3. Análise Discriminante

3.1 - Objetivos e considerações da Análise Discriminante

Os objetivos da Análise Discriminante, descritos por Hair et al (1998; p.256), são:

1. Determinar se é estatisticamente significante a diferença que existe entre a média dos valores de um conjunto de variáveis para dois ou mais grupos pré-definidos.
2. Determinar quais as variáveis independentes que discriminam, da melhor forma, os grupos.
3. Estabelecer procedimentos para classificar os objetos (indivíduos, firmas, produtos, etc) dentro dos grupos com base nos valores do conjunto das variáveis independentes.
4. Estabelecer o número e a composição da dimensão da discriminação entre os grupos formados pelo conjunto de variáveis independentes.

Podemos notar com base nestes objetivos que a Análise Discriminante é útil quando o investigador está interessado em entender as diferenças entre os grupos ou em classificar corretamente os objetos dentro dos grupos. A Análise Discriminante para Johnson e Wichern (1998; p. 629) apresenta uma natureza bastante exploratória. É empregada freqüentemente como um procedimento de separação para investigar as diferenças observadas quando as relações causais não são bem entendidas. Procedimentos de classificação são menos exploratórios na medida que eles conduzem a regras bem definidas que podem ser usadas para nomear novos objetos.

O sucesso na aplicação da Análise Discriminante requer algumas considerações sobre alguns assuntos. Estes assuntos incluem a seleção da variável dependente e das variáveis independentes, o tamanho da amostra necessário para estimar as funções discriminantes e, opcionalmente, a divisão da amostra com a finalidade de validação.

Para a aplicação da Análise Discriminante, primeiro deve-se estipular quais são as variáveis independentes e qual é a variável dependente. Vale recordar, como mencionado anteriormente na introdução, que a variável dependente é categórica, definidora do grupo e as variáveis independentes são métricas. O número de grupos da variável dependente pode ser dois ou mais, mas estes grupos *"devem obrigatoriamente ser mutuamente exclusivos e exaustivos"* de acordo com Hair et al (1998; p.257), portanto cada observação só pode ser alocada a um único grupo, e sempre existirá um grupo para acomodar uma observação.

Em alguns casos tem-se que criar uma variável dependente categórica. Por exemplo, se temos uma variável que mede o número de refrigerantes consumidos por dia, e os indivíduos respondem em uma escala de 0 a 8 (ou mais) por dia, pode-se criar uma variável categórica com três grupos em que o objetivo seja discriminar entre usuários que consomem pouco, médio e muito. Quando três ou mais categorias são criadas surge à possibilidade de examinar somente os grupos extremos em uma Análise Discriminante com dois grupos. Este procedimento é conhecido como *"aproximação dos extremos polares"*.

Depois que a decisão foi tomada em relação à escolha da variável dependente, decide-se quais variáveis independentes devem ser incluídas na análise. As variáveis independentes, usualmente, são selecionadas de dois modos. O primeiro modo envolve identificar as variáveis através de pesquisa prévia ou modelo teórico que são bases subjacentes da questão da análise. O segundo modo é intuitivo, utiliza-se o conhecimento do pesquisador e a intuição selecionando variáveis para as quais não se tem pesquisa prévia ou teoria existente, mas que logicamente poderiam ser relacionadas a predizer os grupos.

A Análise Discriminante é bastante sensível para a relação do tamanho da amostra e o número de variáveis independentes. Muitos estudos sugerem uma relação de 20 observações para cada variável independente. Embora esta relação possa ser difícil de manter-se, na prática, o investigador tem que notar que os resultados ficam instáveis com a diminuição do tamanho da amostra em relação ao número de variáveis independentes. O tamanho mínimo recomendado são cinco observações por variável independente. Note que esta relação é aplicada para todas as variáveis consideradas na análise, até mesmo se todas as variáveis consideradas não entrarem na função discriminante.

Além do tamanho global da amostra, considera-se também o tamanho da amostra em cada grupo. O tamanho do menor grupo tem que exceder o número de variáveis independentes. Como diretriz cada grupo deveria ter pelo menos 20 observações, mas se todos os grupos excederem as 20 observações, considera-se os tamanhos relativos dos grupos. Se os grupos variarem amplamente em tamanho, isto pode causar impacto na estimação da função discriminante e na classificação das observações. Na fase de classificação, os grupos maiores têm uma chance mais alta de classificação.

3.2 - Suposições para a Análise Discriminante

É desejável conhecer certas condições para propor a aplicação da Análise Discriminante. As suposições fundamentais para derivar a função discriminante são:

- As variáveis independentes apresentam uma *distribuição normal multivariada e dispersão desconhecida (mas igual)* entre os grupos. "Dados que não satisfazem definitivamente a suposição de normalidade multivariada podem causar problemas na estimação da função discriminante. Então, é sugerido que a Regressão Logística seja usada como uma alternativa possível" Hair et al (1998; p. 259).
- *Homogeneidade de variâncias/covariâncias*: dentro de cada grupo formado pela variável dependente, a matriz de covariância de um grupo deve ser semelhante à matriz de covariância correspondente em outros grupos. Quer dizer que os grupos tem matrizes de covariâncias semelhantes.
- A *multicolinearidade* entre as variáveis independentes é outra característica dos dados que podem afetar os resultados. Multicolinearidade denota que duas ou mais variáveis independentes são altamente correlacionadas, de forma que uma variável pode ser altamente explicada pela outra variável e assim acrescenta pouco ao poder explicativo do modelo inteiro. Esta consideração fica especialmente crítica quando são empregados procedimentos stepwise.
- Supõe-se implicitamente que todas *as relações são lineares*. Não são estudadas relações não lineares na função discriminante a menos que sejam feitas transformações em variáveis específicas para representar efeitos não lineares.

- Finalmente, *outliers* podem ter um impacto significativo na precisão de classificação de qualquer resultado da Análise Discriminante. Examinam-se todos os resultados para a presença de outliers e eliminam-se verdadeiros outliers se for preciso.
- O problema da *matriz mal-condicionada* é outra suposição da análise discriminante, ou seja, as variáveis que são usadas para discriminar os grupos não podem ser redundantes. Se qualquer variável é redundante com as outras variáveis temos uma matriz de covariância das variáveis do modelo que não pode ser invertida. Por exemplo, uma variável é redundante se ela for a soma de outras três variáveis que estão no modelo. Sendo assim vamos ter alguns valores muito altos e outros muito baixos.
- Deve-se cuidar constantemente o *valor de tolerância* para cada variável, evitando assim o problema de matriz mal-condicionada. O valor de tolerância é calculado como: $1 - R^2$ da variável respectiva com todas as outras variáveis incluídas no modelo. Em geral, quando uma variável é redundante o seu valor de tolerância chega a zero.

3.3 – Estimação do Modelo Discriminante

Para derivar a função discriminante, especifica-se o método de seleção das variáveis independentes e o método determina o número de funções discriminantes. Com as funções estimadas podemos interpretar os coeficientes destas funções.

A utilização dos gráficos também vai ser abordada principalmente para verificarmos a localização relativa dos centróides dos grupos.

Vários critérios estão disponíveis para avaliar se o processo de classificação alcança significância estatística. Após, identifica-se a precisão da classificação e o seu impacto relativo na estimação do modelo global.

3.3.1 – Métodos de Seleção: Direto ou *Stepwise*

Os critérios pelos quais são selecionadas as variáveis independentes para inclusão na análise discriminante são controlados pelo usuário e são indicados pela especificação do

método. Seis métodos estão disponíveis: todas as variáveis independentes podem entrar juntas através do método direto, ou as variáveis podem entrar isoladamente ou em grupos especificados através dos cinco métodos *stepwise*.

O ***Método Direto*** é a técnica na qual todas as variáveis independentes entram juntas na análise. As funções discriminantes são criadas diretamente do conjunto das variáveis independentes, sem dar importância ao poder discriminante de cada uma destas variáveis. Ele é apropriado quando por razões teóricas o investigador deseja ter todas as variáveis independentes entrando na análise e não está interessado em ver resultado intermediário baseado em subconjuntos das variáveis independentes (Klecka; 1975, p.446). Este método requer menos tempo computacional e ocupa um menor espaço de armazenamento do que os métodos *stepwise*.

Nos ***Métodos Stepwise*** as variáveis independentes são selecionadas para entrar na análise com base no poder discriminante delas.

Em muitos exemplos o conjunto completo das variáveis independentes contém informações em excesso sobre as diferenças dos grupos, ou talvez algumas das variáveis não sejam muito úteis na discriminação entre os grupos. Sendo assim, selecionando um "próximo melhor" discriminador consecutivamente, a cada passo, será achado um conjunto reduzido de variáveis que é quase tão bom quanto, e às vezes melhor que, o conjunto completo.

Cinco critérios de seleção *stepwise* estão disponíveis no SPSS; eles serão discutidos a seguir após uma descrição das características gerais desta seleção. O processo começa escolhendo uma única variável que tem o valor mais alto no critério de seleção. Esta variável inicial é emparelhada então com cada uma das outras variáveis disponíveis, uma de cada vez, e o critério de seleção é calculado. A variável nova que, junto com a variável inicial, produz o melhor valor no critério é selecionada como uma segunda variável para entrar na equação. Estas duas são combinadas com cada uma das variáveis restantes e então, uma de cada vez, forma um trio que é avaliado no critério. O trio, com o melhor valor no critério, determina a terceira variável a ser selecionada. Este procedimento de localizar a próxima variável que produz um valor melhor no critério continua, até que todas as variáveis sejam selecionadas ou que nenhuma variável adicional forneça um nível mínimo de melhoria. Os resultados intermediários são impressos passo a passo.

Como as variáveis independentes são escolhidas para a inclusão, algumas variáveis previamente selecionadas podem perder o poder discriminante. Isto acontece porque

a informação que elas contêm sobre as diferenças dos grupos é agora avaliada incluindo a combinação das outras variáveis. Tais variáveis são redundantes e deveriam ser eliminadas. Assim, no começo de cada passo, cada uma das variáveis previamente selecionadas é testada para determinar se ainda apresenta uma contribuição suficiente para a discriminação. Uma variável que foi retirada em um passo pode reentrar em um passo posterior se satisfizer o critério de seleção naquele momento.

O usuário indica o critério de seleção *stepwise* a ser usado. Considerando que cada critério enfatiza um aspecto diferente de "separação", deve-se tomar cuidado para selecionar um critério apropriado.

Quando o método é o λ ***de Wilks***, o critério é o valor do F global para o teste de diferenças entre os grupos centróides. A variável que maximiza o valor F minimiza o λ de Wilks, uma medida de discriminação dos grupos. Este teste leva em conta as diferenças entre todos os centróides e a homogeneidade dentro dos grupos.

O método da ***distância de Mahalanobis*** busca maximizar o valor da distância de Mahalanobis entre os dois grupos mais similares. A variável que maximiza o menor valor de F entre pares de grupo é selecionada pelo método ***menor Razão F***.

Outro critério que tende a discriminar os grupos é avaliado pelo método ***Variância não explicada*** que é definido pelo valor

$$\boldsymbol{R} = \sum_{ij} \frac{1}{1 + \left(\boldsymbol{D}_{ij} / 4 \right)}, \qquad \textit{(equação 3.1)}$$

onde

D_{ij} é a distância de Mahalonobis entre os grupos i e j.

O objetivo aqui é minimizar R, que é a variância residual.

Como critério final temos o método ***V de Rao***, uma medida de distância generalizada. A variável selecionada quando acrescida às variáveis prévias é a que contribui para um maior aumento em V. Quando houver um "grande número" de casos, V tem uma distribuição assintótica qui-quadrado com um grau de liberdade de forma que podemos testar a significância estatística disto.

Segundo Hair et al (1998; p. 262) a distância de Mahalanobis e o V de Rao são os critérios mais apropriados, quando usamos o método *stepwise*, para selecionar as variáveis independentes.

Para todos os critérios anteriores, uma variável só é considerada para a seleção se sua razão do F parcial é maior que um valor especificado. A razão do F parcial mede a discriminação introduzida pela variável depois de levar em conta a discriminação alcançada pelas outras variáveis selecionadas. Este teste de F parcial é executado antes da variável ser avaliada no critério de entrada do *stepwise*. Se o F parcial é muito pequeno a variável não é considerada para inclusão, independente do seu valor no critério de entrada. Só são testadas, para a remoção, as variáveis em que a razão do F parcial é menor que determinado valor.

O uso do método *stepwise* pode resultar em uma seleção ótima do conjunto de variáveis. Cabe lembrar que a seqüência na qual são selecionadas as variáveis, necessariamente, não está relacionada com a importância delas como discriminantes.

3.3.2 - Determinação do número de funções discriminantes

O número máximo de funções discriminantes a serem derivadas é o menor valor entre: g – 1 (número de grupos da variável dependente menos um) ou k (número de variáveis independentes). Assim se houver 10 variáveis independentes e a variável dependente for "religião" com os seguintes grupos: "protestante", "católica", "espírita" e "outras" o número de funções discriminantes será (4 – 1) = 3. No caso de termos 06 grupos e 02 variáveis independentes o número de funções discriminantes será 02. A primeira função maximiza as diferenças entre os valores das categorias da variável dependente. A segunda função é ortogonal a primeira e maximiza as diferenças entre os valores das categorias da variável dependente, controlando para o primeiro fator, e assim por diante. A primeira função será a dimensão mais importante, mas as demais funções podem também representar dimensões significativas.

A importância do número de grupos originou-se nos princípios básicos da geometria. Em geral, dois pontos, no espaço, definem uma linha, três definem um plano, quatro um espaço tridimensional, etc; o número máximo de dimensões necessárias para descrever completamente um conjunto de pontos é um menos o número de pontos. Na análise discriminante, cada grupo (medido por seu centróide) é tratado como um ponto e cada função

discriminante é uma dimensão (ortogonal) que descreve a localização daquele grupo relativa aos outros (Klecka; 1975, p.442).

Uma exceção para a regra geométrica de que p pontos definem (p - 1) dimensões é a seguinte: quando três pontos estão na mesma linha ou quatro pontos estão no mesmo plano, o último ponto é situado de tal forma que não soma uma dimensão nova, ele entra no espaço já definido pelos outros pontos.

O mesmo princípio aplica-se na Análise Discriminante. Duas funções podem ser bastante adequadas para discriminar quatro grupos, mesmo sendo três o número máximo de funções discriminantes. Para isto acontecer, as duas primeiras funções podem englobar integralmente as informações das variáveis discriminantes ou, também podem deixar uma pequena parcela para a possível terceira função que sem significância estatística é ignorada. Outra razão para ignorar uma função discriminante é que a sua contribuição teórica não tem importância na prática.

O software SPSS dentro do subprograma de Análise Discriminante dispõe de duas medidas para julgar a importância das funções discriminantes. Uma destas é a porcentagem relativa do autovalor (eigenvalue) associada com a função. A soma dos autovalores é uma medida da variância total que existe nas variáveis discriminantes. Como as funções discriminantes são derivadas na ordem de importância das mesmas, podemos parar o processo sempre que a porcentagem relativa é considerada muito pequena. Não há nenhuma regra fixa para decidir o que é uma porcentagem muito pequena (Klecka;1975, p.442). A correlação canônica associada à função discriminante é considerada como uma ajuda adicional ao julgarmos a importância de uma função discriminante. Ela mede a associação entre a função discriminante e os grupos da variável dependente. Se nós invertermos a lógica, nós podemos interpretar o quadrado da correlação canônica como a proporção da variância na função discriminante que é explicada pelos grupos.

Um segundo critério para eliminar as funções discriminantes consideradas sem importância é o λ de Wilks. O λ é uma medida inversa do poder discriminante das variáveis originais. Quanto maior é o λ um número menor de informações permanece na função discriminante. O λ pode ser transformado em uma estatística qui-quadrado para um teste de significância estatística. O uso deste critério é limitado a situações nas quais os casos são uma amostra aleatória.

De acordo com Hair et al (1998; p.262) todos os softwares computacionais fornecem ao investigador as informações necessárias à averiguação do número preciso de funções para obter-se significância estatística, sem incluir as funções discriminantes que não aumentam significativamente o poder discriminante. Se uma ou mais funções discriminantes são julgadas não estatisticamente significantes, o modelo discriminante pode ser reestimado e o número de funções a serem derivadas fica limitado ao número de funções significantes. Desta maneira, a avaliação da precisão e a interpretação das funções discriminantes será fundamentada somente nas funções significantes.

3.3.3 – Interpretação dos coeficientes da função discriminante

O SPSS fornece os coeficientes padronizados das funções discriminantes. Estes coeficientes correspondem aos valores d_{ik}'s da seguinte equação:

$$D_i = d_{i1}Z_1 + d_{i2}Z_2 + \ldots\ldots + d_{ik}Z_k, \qquad \textit{(equação 3.2)}$$

sendo que,

D_i é o valor da função discriminante i ,

Z_k's são os valores padronizados das k variáveis discriminantes (independentes) usadas na análise.

Eles são usados para calcular o valor da função discriminante para um caso em que as variáveis originais discriminantes estão na forma padronizada (Z scores). O valor da função discriminante é calculado multiplicando cada variável por seu coeficiente correspondente e somando estes produtos. Haverá um valor separado para cada caso em cada função. Os coeficientes foram derivados de tal modo que os valores discriminantes produzidos estão padronizados. Isto significa que, para todos os casos da análise, o valor de uma função terá média igual a zero e desvio padrão igual a um.

Calculando a média dos valores para os casos dentro de um particular grupo, nós chegamos à média do grupo na respectiva função. Para um único grupo, as médias em todas as funções são chamadas de **centróides** dos grupos que é a localização mais típica de um caso daquele grupo no espaço da função discriminante. Uma comparação da média dos grupos em cada função nos mostra como estão os grupos distintamente separados ao longo daquela

dimensão. Devemos lembrar que as funções são organizadas em ordem de importância decrescente, de forma que uma determinada diferença entre as médias dos grupos na terceira ou quarta função não é tão significante quanto a mesma diferença na primeira função.

Quando o sinal é ignorado, cada coeficiente representa a contribuição relativa de sua variável associada àquela função. O sinal somente denota se a variável está fazendo uma contribuição positiva ou negativa.

Como as variáveis discriminantes não são usualmente apresentadas na forma padronizada, os coeficientes padronizados das funções discriminantes não são úteis para os cálculos propostos. Sendo assim, utilizamos os coeficientes não padronizados. Estes são multiplicados pelos valores originais das variáveis associadas para chegar a um valor da função discriminante. Depois de somar uma constante utilizada para ajuste, é obtido um valor que é equivalente ao calculado com os coeficientes e dados padronizados. Os coeficientes não padronizados das funções discriminantes não informam a importância relativa das variáveis já que as variáveis originais não foram ajustadas.

Outro modo para determinar quais variáveis definem, de forma mais importante, uma particular função discriminante é olhar para a matriz de estrutura. Os coeficientes de estrutura são as correlações entre as variáveis do modelo e as funções discriminantes.

Os coeficientes de estrutura devem ser usados ao interpretar-se o significado das funções discriminantes. As razões dadas são que (1) supostamente os coeficientes de estrutura são mais estáveis, e (2) eles permitem a interpretação das funções discriminantes de maneira análoga à Análise Fatorial. Porém, pesquisa utilizando Simulação de Monte Carlo (Klecka, 1975) mostra que os coeficientes das funções discriminantes e os coeficientes de estrutura são aproximadamente iguais em relação à instabilidade, a menos que o n seja bastante grande (por exemplo, se há 20 vezes mais casos que variáveis). É importante lembrar que os coeficientes das funções discriminantes denotam a única contribuição parcial para cada variável na função discriminante, enquanto que os coeficientes de estrutura denotam as correlações simples entre as variáveis e as funções.

3.3.4 – Gráficos e rotação das funções discriminantes

Utilizamos as duas primeiras funções discriminantes para representarmos graficamente os grupos. Podemos optar por delinear todos os grupos em um único gráfico ou

um gráfico separado para cada grupo, o que é especialmente útil quando há um grande número de grupos evitando assim que os grupos sobreponham-se.

Os gráficos são particularmente úteis ao estudarmos a separação e a localização dos centróides dos grupos. As interpretações podem ser feitas através dos valores dos centróides dos grupos, mas através dos gráficos é mais fácil a visualização das relações.

Outra vantagem dos gráficos está na observação do grau no qual os grupos sobrepõem-se de fato. Em situações nas quais os grupos são mais distintos eles podem ajudar a identificar casos anticonvencionais e em grupos coesos podem ilustrar os diferentes graus de coesão.

Quando a variável dependente possui apenas dois grupos e só uma função discriminante é derivada, a representação gráfica é um histograma.

Quando são derivadas duas ou mais funções discriminantes, podemos solicitar a representação gráfica através do "mapa territorial". Cada ponto no mapa é classificado de acordo com os valores das funções discriminantes. O símbolo para o centróide do grupo é representado por um asterisco (*).

As funções discriminantes são derivadas de tal forma que a primeira função separa os grupos tanto quanto possível. A segunda função os separa tanto quanto possível em uma direção ortogonal à primeira separação, a terceira função provê separação máxima em outra direção ortogonal, e assim sucessivamente. Os resultados finais são que os grupos são tão distintos quanto é possível determinar pelas variáveis discriminantes originais. As funções discriminantes podem ser consideradas com os eixos definidos em um espaço geométrico no qual, cada caso e o centróide do grupo são pontos e a orientação de espaço destes eixos é essencialmente arbitrária. Pode ser útil girar estes eixos mantendo constante a posição relativa dos casos e dos centróides dos grupos, como na análise fatorial. Um critério para uma solução de rotação é a VARIMAX que estabelece eixos nos quais os coeficientes para as variáveis discriminantes ou estão perto de 1 ou perto de 0.

Embora a rotação de eixos da função discriminante tenha sido sugerida na literatura estatística (Cooley e Lohnes, 1971, p.250), nenhum tratamento foi dado às reais conseqüências de tal rotação. Então, os usuários deveriam tratar esta característica como experimental e só deveriam empregar isto depois de adquirir uma completa compreensão da técnica.

3.3.5 – Classificação dos casos

Outro objetivo principal para a aplicação da análise discriminante é classificar, de maneira eficiente, novos casos dentro dos grupos estudados. Uma vez que o modelo foi finalizado e as funções discriminantes foram derivadas, como podemos predizer a qual grupo pertence um caso particular?

Sharma (1996) explica que a classificação consiste em dividir o espaço total discriminante em regiões mutuamente exclusivas e exaustivas e, para tanto, podem ser utilizados os seguintes métodos: (1) ponto de corte; (2) teoria da decisão estatística; (3) função de classificação e (4) distância de Mahalanobis.

O método do ***ponto de corte*** objetiva encontrar o escore discriminante ou valor de corte, que divide o espaço discriminante nas regiões supracitadas. Normalmente, o valor de corte selecionado é aquele que minimiza o número de classificações incorretas e é dada pela seguinte equação:

$$\text{Ponto de corte} = \frac{\sum_{i=1}^{n} n_i \overline{Z}_i}{\sum_{i=1}^{n} n_i}, \qquad \textit{(equação 3.3)}$$

onde,

Z_i é o escore discriminante médio para o grupo i,

n_i é o número de observações no grupo i.

O método da ***teoria da decisão estatística*** é baseado na teoria Bayesiana e consiste em minimizar os erros de classificações incorretas levando em consideração as probabilidades a priori e os custos de classificação incorreta, o que o torna um dos métodos de classificação mais gerais. Para o caso de dois grupos, este método designa uma observação para o grupo 1 se:

$$Z \geq \frac{\overline{Z}_1 + \overline{Z}_2}{2} + \ln\left[\frac{p_2 C(1/2)}{p_1 C(2/1)}\right], \qquad \textit{(equação 3.4)}$$

e uma observação é classificada no grupo 2 se:

$$Z < \frac{\overline{Z}_1 + \overline{Z}_2}{2} + \ln\left[\frac{p_2 C(1/2)}{p_1 C(2/1)}\right], \qquad \textit{(equação 3.5)}$$

onde,

Z é o escore discriminante para uma dada observação,

$\overline{Z}_1$ é o escore discriminante médio para o grupo 1 e $\overline{Z}_2$ é o escore discriminante médio para o grupo 2,

p_1 é a probabilidade a priori do grupo 1 e p_2 é a probabilidade a priori do grupo 2,

$C(1/2)$ é o custo de classificação incorreta dentro do grupo 1 de uma observação que pertence ao grupo 2 e $C(2/1)$ é o custo de classificação incorreta dentro do grupo 2 de uma observação que pertence ao grupo 1.

As ***funções de classificação*** podem ser usadas para determinar a qual grupo pertence um provável caso. Há tantas funções de classificação quanto há grupos. Cada função nos permite calcular os valores de classificação para cada caso e cada grupo, aplicando a fórmula:

$$C_i = c_{i1}V_1 + c_{i2}V_2 + \ldots + c_{ik}V_k + c_{i0}, \qquad \textit{(equação 3.6)}$$

onde,

C_i é o valor da classificação para o grupo i,

c_{ik}'s são os coeficientes de classificação,

c_{io} é a constante,

V_k são os valores originais das variáveis discriminantes.

Sempre há uma equação para cada grupo; assim se houver quatro grupos, cada caso terá quatro valores.

Calculados os valores de classificação para um caso, é fácil decidir como classificar o caso: nós o classificamos como pertencendo ao grupo para o qual o valor de

classificação é mais alto. Assim, se fôssemos estudar as escolhas de estudantes do ensino médio após o término do curso (por exemplo, freqüentar faculdade, freqüentar um curso profissional ou escola de comércio, ou arrumar um emprego) baseado em várias variáveis avaliadas no ano anterior ao término do ensino médio, poderíamos verificar como a classificação funciona para predizer o que cada estudante é provável de fazer depois da conclusão do curso. Porém, nós também gostaríamos de saber com qual probabilidade o estudante fará a escolha predita. Essas probabilidades são chamadas probabilidades *a posteriori*, e também podem ser calculadas. Porém, entender como essas probabilidades são derivadas, nos leva a considerar a distância de Mahalanobis.

Para cada caso podemos calcular a ***distância de Mahalanobis*** de cada ponto (caso) ao centróide do grupo. Classificaríamos o caso como pertencendo ao grupo para o qual o centróide do grupo está mais próximo, isto é, onde a distância de Mahalanobis é menor.

A probabilidade de que um caso pertence a um grupo particular é basicamente proporcional à distância de Mahalanobis daquele centróide do grupo (não é precisamente proporcional porque nós assumimos uma distribuição normal multivariada ao redor de cada centróide). Como calculamos a posição de cada caso a partir de nosso conhecimento anterior dos valores para aquele caso nas variáveis no modelo, estas probabilidades são chamadas probabilidades *a posteriori*. Em resumo, a probabilidade *a posteriori* é a probabilidade, baseada em nosso conhecimento dos valores de outras variáveis, que o respectivo caso pertença a um grupo particular. Alguns softwares estatísticos calculam essas probabilidades automaticamente para todos os casos.

Há um fator adicional que precisa ser considerado ao classificar casos. Às vezes, nós sabemos de antemão que há mais observações em um grupo que em qualquer outro; assim, uma probabilidade *a priori* de que um caso pertença àquele grupo é mais alta. Por exemplo, se nós sabemos de antemão que 60% dos diplomados de nossa escola secundária normalmente vão para a faculdade (20% vão para uma escola profissional, e outros 20% arrumam um emprego), então nós deveríamos ajustar nossa predição adequadamente: *a priori* é mais provável que um estudante freqüentará a faculdade do que escolher qualquer uma das outras duas opções. Você pode especificar diferentes probabilidades *a priori* que serão usadas para ajustar a classificação de casos adequadamente. Neste caso, fixaríamos as probabilidades *a priori* para ser proporcional aos tamanhos dos grupos em nossa amostra. Caso contrário,

podemos especificar as probabilidades *a priori* como sendo iguais em cada grupo. A especificação de diferentes probabilidades *a priori* pode afetar a precisão da predição.

A melhor maneira de avaliar como as funções de classificação atuais predizem corretamente a classificação dos casos nos grupos é a matriz de classificação. A matriz de classificação mostra o número de casos que foram classificados corretamente.

3.4 – Métodos para Interpretação dos Resultados

Se a função discriminante é estatisticamente significante e a precisão da classificação é aceita o foco deve ficar na interpretação do que foi descoberto. Este processo envolve examinar as funções discriminantes para determinar a relativa importância de cada variável independente na discriminação entre os grupos. De acordo com Hair et al (1998; p.272) são propostos três métodos para determinar esta relativa importância das variáveis independentes: (1) Padronização dos pesos discriminantes, (2) Cargas discriminantes e (3) Valor parcial de F.

***Padronização dos pesos discriminantes*:** O método tradicional para a interpretação das funções discriminantes examina o sinal e a magnitude dos pesos discriminantes padronizados designados para cada variável calculado nas funções discriminantes. Quando o sinal é ignorado, cada peso representa a contribuição relativa da variável associada a cada função. Variáveis independentes com grande pesos relativos contribuem mais para o poder discriminante da função do que variáveis com pequenos pesos relativos. O sinal denota apenas se a contribuição da variável é negativa ou positiva. A interpretação dos pesos discriminantes é análoga a interpretação dos "betas" na Análise de Regressão e dependem da escala de medida das variáveis.

Cargas discriminantes ou ***Coeficientes da matriz de estrutura*:** As cargas discriminantes iniciaram a serem usadas, para a interpretação, devido às deficiências na utilização dos pesos. Estas cargas referem-se às estruturas de correlações e é uma medida de correlação linear simples entre cada variável independente e a função discriminante. As cargas apresentam uma

maior validade que os pesos na interpretação do poder discriminante das variáveis independentes devido à correlação.

***Valor parcial de F*:** Como vimos anteriormente podemos utilizar dois métodos para derivar as funções discriminantes que são o direto e o *stepwise*. Se o método *stepwise* é selecionado a interpretação do poder discriminante relativo das variáveis independentes é avaliado pelo valor parcial de F. Um valor de F grande indica grande poder discriminante e através do nível de significância de F temos a associação para cada variável.

Quando temos duas ou mais funções discriminantes surgem problemas adicionais de interpretação. Primeiro, como podemos simplificar os pesos e as cargas discriminantes para facilitar o perfil de cada função? Segundo, como podemos representar o impacto de cada variável sobre as funções? Estes problemas são ambos baseados nos efeitos da discriminação total através das funções estimando o papel de cada variável no perfil de cada função separadamente. Estas duas questões vão servir para abordarmos a rotação das funções e a representação gráfica das cargas discriminantes (Hair et al; 1998, p. 273).

***Rotação das funções discriminantes*:** Basicamente a rotação preserva a estrutura original da solução discriminante tornando mais fácil a interpretação das funções. Como constou no item 3.3.4 a rotação VARIMAX é empregada como base.

***Representação gráfica das cargas discriminantes*:** Para descrever as diferenças entre os grupos através das variáveis, as cargas discriminantes e os grupos centróides podem ser representados em um espaço discriminante reduzido. Pode-se usar ou não a rotação das cargas em um gráfico, sendo que a preferência é usar a rotação. Porém, até mesmo a aproximação mais precisa envolve o que é chamado alongamento de vetores e antes de explicar o processo de alongar, temos que definir primeiro um vetor neste contexto. Um vetor é uma linha desenhada da origem (centro) de um gráfico para as coordenadas das cargas de uma variável particular. O comprimento de cada vetor é indicativo da importância relativa de cada variável em discriminar os grupos. Para obter a extensão de um vetor, multiplica-se a carga discriminante (preferencialmente depois de rotação) por seu respectivo valor de F. O processo gráfico envolve todas as variáveis significantes incluídas no modelo, mas também pode-se representar as outras variáveis cujas relações de F são significantes e que não eram

significantes na função discriminante. Este procedimento (representação gráfica) mostra a importância das variáveis colineares que não são incluídas, como no método *stepwise*. Usando a representação gráfica notamos que os vetores apontam para os grupos que tem a maior média na respectiva variável independente e ficam longe dos grupos que tem o menor valor médio. Os centróides dos grupos também estão representados neste procedimento e o processo é o mesmo, multiplicamos os centróides pelo valor aproximado de F associado com cada função discriminante. O valor aproximado de F para uma função discriminante *i* é obtido pela fórmula:

$$F_i = Autovalor_i\left(\frac{tamanho \cdot da \cdot amostra - g}{g-1}\right), \qquad \textit{(equação 3.7)}$$

onde g é o número de grupos da variável dependente.

Com o objetivo de demonstrar como fica a representação gráfica das cargas discriminantes vamos usar um exemplo do Hair (1998; p.312). Apresenta-se a tabela 3.1 com os valores das cargas discriminantes, o valor de F e também o valor desta multiplicação (extensão do vetor) que é o que utilizamos no gráfico. Na tabela constam todas as variáveis escolhidas para a análise e os centróides dos grupos. A figura 3.1 é a representação gráfica da extensão dos vetores cujo F foi significante.

Tabela 3.1 – Coordenadas dos vetores e dos centróides dos grupos em reduzido espaço discriminante

	Rotated Discriminant Function Loadings		*Univariate F Ratio*	*Reduced Space Coordinates*	
Independent Variables	*Function 1*	*Function 2*		*Function 1*	*Function 2*
X_1 Delivery speed[a]	.568	.319	23.346	13.261	7.447
X_2 Price level[a]	−.502	.672	11.674	−5.860	7.845
X_3 Price flexibility[a]	.891	−.186	33.362	29.726	−6.205
X_4 Manufacturer image	−.041	190	.961	−0.039[b]	0.183[b]
X_5 Overall service	.101	957	23.692	2.393	22.673
X_6 Salesforce image	071	224	.126	0.009[b]	0.028[b]
X_7 Product quality	−.222	.163	9.293	−2.063	1.515

	Group Centroids		*Approximate F Value*		*Reduced Space Coordinates*	
Group	*Function 1*	*Function 2*	*Function 1*	*Function 2*	*Function 1*	*Function 2*
Group 1: New task	−1.081	−1.167	55.148	15.162	−59.614	−17.694
Group 2: Modified rebuy	−.952	979	55.148	15.162	−52.500	13.327
Group 3: Straight rebuy	1.541	.472	55.148	15.162	84.992	7.156

[a]Variables entered in the stepwise solution.
[b]Vectors not plotted because of nonsignificant F ratio.

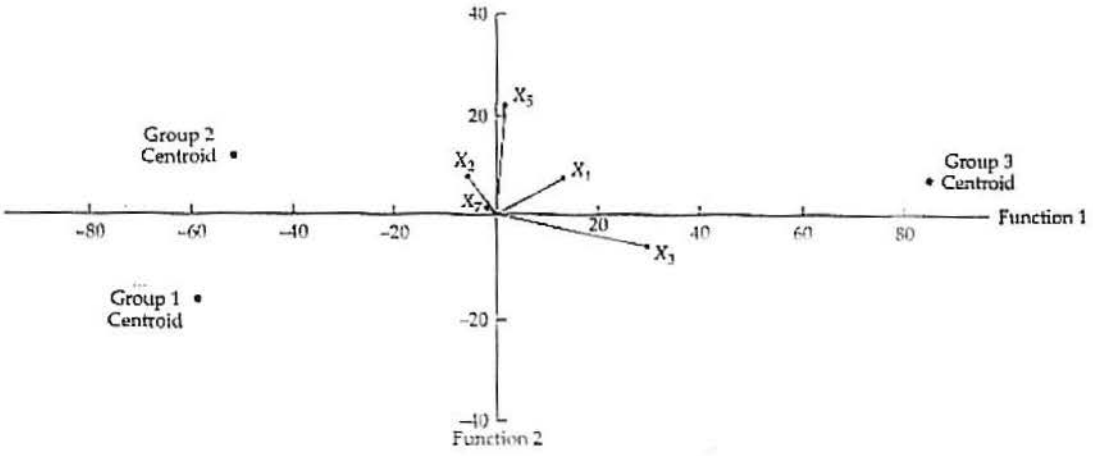


Figura 3.1 – Gráfico com a representação vetorial das variáveis em reduzido espaço discriminante

Sugerimos que o investigador empregue vários métodos disponíveis para chegar a uma interpretação com máxima precisão.

3.5 - Validação dos Resultados

A fase final de uma análise discriminante envolve a validação dos resultados discriminantes para prover garantias que estes resultados tenham validação externa como também interna. Segundo Hair et al (1998; p.275-276) podemos usar duas técnicas de validação: a validação cruzada e a análise dos perfis dos grupos.

Os procedimentos de ***validação cruzada*** ou divisão da amostra são freqüentemente utilizados na validação da função discriminante dividindo os grupos aleatoriamente em amostra de análise e amostra de validação. A validação cruzada compara a precisão de um modelo discriminante em uma amostra de validação com a precisão obtida na amostra de análise para a qual este modelo foi desenvolvido. Idealmente um tamanho de amostra grande ou uma proporção dos casos (talvez metade ou dois terços) pode ser designado como pertencendo à amostra da análise e os casos restantes podem ser designados como pertencendo à amostra de validação. Se o modelo discriminar bem tanto na amostra de análise quanto na amostra de validação, é dito que a validação cruzada é boa.

Para uma maior confiança na validação da função é sugerido por alguns investigadores que este procedimento seja seguido várias vezes. Em vez de dividir a amostra total, uma vez, em análise e validação, o investigador dividiria a amostra total em análise e validação por várias vezes, cada vez testando a validação da função com o desenvolvimento de uma matriz de classificação e calculando o percentual de casos corretamente classificados pela função discriminante. Então, para a obtenção de uma única medida calcula-se a média destes percentuais.

Foram propostos, por pesquisadores, métodos mais sofisticados baseados na estimação com subconjuntos múltiplos da amostra para validar a função discriminante. Os dois amplamente usados são o U-método e o método de Jackknife. Ambos os métodos são usados no princípio de "reamostragem, sendo que o mais usual é o "Jackknife". A diferença primária entre os dois é que o U-método focaliza-se na precisão da classificação, enquanto que o Jackknife direciona-se para a estabilidade dos coeficientes discriminantes. Ambas as aproximações são bastante sensíveis a tamanhos pequenos de amostra. Diretrizes sugerem que qualquer um destes dois métodos só seja usado quando o tamanho do menor grupo é pelo menos três vezes maior que o número de variáveis preditas, e a maioria dos investigadores propõem uma relação de cinco a um. Apesar destas limitações, ambos os métodos fornecem uma estimativa válida e consistente da taxa de precisão de classificação.

Outra técnica de validação é a ***análise dos perfis dos grupos***. Quando se tiver determinado quais as variáveis independentes que mais contribuem na discriminação entre grupos, perfila-se as características dos grupos a partir de suas médias. Desse modo, pode-se entender o perfil de cada grupo de acordo com as variáveis preditivas.

Nas aplicações práticas que estão no capítulo 5, a amostra é pequena para ser dividida em amostra de análise e amostra de validação. Sendo assim a validação foi efetivada com a mesma amostra utilizada na análise e os resultados não foram satisfatórios.

3.6 – Modelos Discriminantes

Apresentaremos o modelo com as descrições matemáticas da função linear discriminante de Fisher que é utilizada quando temos uma Análise Discriminante com dois

grupos e as descrições do método de Fisher para uma Análise Discriminante com mais de dois grupos. Descrições matemáticas de outros critérios discriminantes serão também mencionadas.

3.6.1 – Análise Discriminante com dois grupos

Para Johnson & Wichern (1998) a idéia de Fisher era transformar as observações multivariadas x em observações univariadas y (vetores), dos grupos 1 e 2.

Grupo 1	Grupo 2
$\mathbf{y}_{11}$	$\mathbf{y}_{21}$
$\mathbf{y}_{12}$	$\mathbf{y}_{22}$
.	.
.	.
.	.
$\mathbf{y}_{1n_1}$	$\mathbf{y}_{2n_2}$

A separação para estes dois vetores é avaliada em termos de diferença entre $\bar{y}_1$ e $\bar{y}_2$ em unidades de desvios ao quadrado. Isto é,

$$\text{Separação} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y} \qquad \textit{(equação 3.8)}$$

onde

$$S_Y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2} \qquad \textit{(equação 3.9)}$$

com $y_{1j} = \hat{\boldsymbol{a}}' \boldsymbol{x}_{1j}$ e $y_{2j} = \hat{\boldsymbol{a}}' \boldsymbol{x}_{2j}$

é a estimação da variância ponderada. O objetivo é selecionar a combinação linear de **x** para encontrar a máxima separação das médias $\bar{y}_1$ e $\bar{y}_2$.

A combinação linear $\hat{y} = \hat{\boldsymbol{a}}' \boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \boldsymbol{S}_C^{-1} \boldsymbol{x}$ *(equação 3.10)*

maximiza a razão $= \dfrac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$ *(equação 3.11)*

$$= \frac{(\hat{a}'\bar{x}_1 - \hat{a}'\bar{x}_2)^2}{\hat{a}'S_c\hat{a}}$$

$$= \frac{(\hat{a}'d)^2}{\hat{a}'S_c\hat{a}} \qquad \textit{(equação 3.12)}$$

sob todos os possíveis coeficientes dos vetores $\hat{a}$ onde $d = (\bar{x}_1 - \bar{x}_2)$. O máximo da razão da equação 3.11 é $D^2 = (\bar{x}_1 - \bar{x}_2)' S_c^{-1}(\bar{x}_1 - \bar{x}_2)$.

Temos então:

$$\max \frac{(\hat{a}'d)^2}{\hat{a}'S_c\hat{a}} = d'S_c^{-1}d = (\bar{x}_1 - \bar{x}_2)' S_c^{-1}(\bar{x}_1 - \bar{x}_2) = D^2$$

onde D^2 é a distância ao quadrado entre as duas médias e S_c é a matriz de covariância ponderada das populações 1 e 2. É importante lembrar que para S_c^{-1} existir é necessário que $n_1 + n_2 - 2 > p$, pois caso contrário a matriz S_c seria singular e não poderia ser invertida.

Uma regra de alocação baseada na função discriminante de Fisher é:

Aloque x_0 no grupo 1 se

$$\hat{y}_0 = (\bar{x}_1 - \bar{x}_2)' S_c^{-1} x_0 \geq \hat{m} = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_c^{-1}(\bar{x}_1 + \bar{x}_2)$$

ou $\hat{y}_0 - m \geq 0$.

Aloque x_0 no grupo 2 se:

$\hat{y}_0 < \hat{m}$

ou $\hat{y}_0 - m < 0$

3.6.2 – Análise Discriminante com mais de dois grupos

Quando há mais de dois grupos são necessárias mais de uma função discriminante para alcançar a discriminação máxima entre os grupos. Portanto, no caso de Análise Discriminante para g grupos, a equação 3.12 é transformada em:

$$\hat{e} = \frac{\hat{a}' B \hat{a}}{\hat{a}' W \hat{a}} \qquad \textit{(equação 3.13)}$$

que também pode ser escrita como:

$$\hat{a}'\left(B\hat{a} - \hat{e}W\hat{a}\right) = 0 \qquad \textit{(equação 3.14)}$$

onde a matriz $\boldsymbol{B}$ é definida como:

$$\boldsymbol{B} = \sum_{i=1}^{g} n_i \left(\bar{x}_i - \bar{x}\right)\left(\bar{x}_i - \bar{x}\right)' \qquad \textit{(equação 3.15)}$$

e a matriz W:

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(\bar{x}_{ij} - \bar{x}\right)\left(\bar{x}_{ij} - \bar{x}\right)' \qquad \textit{(equação 3.16)}$$

A matriz $\boldsymbol{B}$ possui dimensão *kxk* e sua diagonal principal é composta pela soma dos quadrados "entre grupos" para cada uma das k variáveis independentes. Os elementos fora da diagonal são análogos às somas de produtos para cada par de variáveis. A matriz W também possui dimensão *kxk* e sua diagonal principal é composta pela soma de quadrados "dentro dos grupos" para cada variável. Os elementos fora da diagonal são os mesmos da matriz $\boldsymbol{B}$. É importante destacar que a soma de quadrados "dentro dos grupos" é uma boa medida da homogeneidade dos mesmos e a soma de quadrados "entre grupos" refere-se a diferença entre as médias dos grupos. Sendo assim, o objetivo da equação 3.13 é obter o λ que maximize a razão: (soma de quadrados "entre grupos"/soma de quadrados "dentro dos grupos"), atingindo, desta forma, a máxima discriminação entre os grupos.

Na equação 3.14 devem ser examinados os valores de λ e $\hat{a}$ que solucionam a equação para determinar o valor de $\hat{a}$ que resulta no máximo λ. As soluções possíveis são encontradas em:

$$B\hat{a} - \hat{e}W\hat{a} = 0 \qquad \textit{(equação 3.17)}$$

que também pode ser escrita como a equação 3.18:

$$\left(W^{-1}B - \hat{e}I\right)\hat{a} = 0 \qquad \textit{(equação 3.18)}$$

As soluções da equação 3.18 são os autovalores λ_1, λ_2, ..., λ_s, e os autovetores $\hat{a}_1$, $\hat{a}_2$, ..., $\hat{a}_s$ associados à matriz $W^{-1}B$. Os elementos de cada vetor característico são os coeficientes procurados de cada uma das funções discriminantes lineares. As raízes características fornecem uma medida absoluta do poder discriminante de cada função linear. O número "s" de raízes características não nulas equivale ao rank da matriz B e elas são usualmente listadas em ordem decrescente , isto é, $\lambda_1 > \lambda_2 > ... > \lambda_s$. Os "s" vetores característicos são listados na mesma ordem, com $\hat{a}_1$ correspondendo a λ_1, $\hat{a}_2$ a λ_2 e assim por diante. Desta forma, o valor máximo da equação 3.13 é a raiz característica λ_1 e o vetor de coeficientes que resulta nesse valor máximo é o autovetor $\hat{a}_1$. Portanto, a função discriminante que maximiza a diferença entre as médias das populações é $y_1 = \hat{a}_1' x$.

3.6.3– Outros critérios discriminantes

Em Murtagh e Heck (1987; p.113-121) temos alguns outros critérios discriminantes que são: Discriminação de Máxima Verossimilhança, Discriminação Não-Paramétrica e Discriminação Bayesiana.

Vamos apresentar os procedimentos matemáticos destes critérios de acordo com os autores.

***Discriminação de Máxima Verossimilhança*:** Em um contexto prático deve-se calcular a média dos vetores (g_y) e a matriz de covariância (V_Y)dos dados que podem ter sido escolhidos para constituir uma amostra de uma população subjacente.

Nós usamos uma função de densidade normal multivariada para $P(x/y)$. Se todos os n objetos de x_i, forem independentes, então a distribuição é

$$£ = \prod_{i=1}^{n} P(x_i / y) \qquad \textit{(equação 3.19)}$$

Considerando £ como uma função com os parâmetros g e V desconhecidos, este é o termo da função de verossimilhança. O princípio de máxima verossimilhança diz que devemos escolher os parâmetros desconhecidos tal que £ seja maximizado. A aproximação clássica para otimizar £ é diferenciar com respeito à g e a V, e fixar os resultados iguais a zero. Fazendo isto para a expressão da normal multivariada usada previamente permite-se derivar estimativas para a média e covariância como segue.

$$\hat{g} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textit{(equação 3.20)}$$

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n} (x_i - g)(x_i - g)' \qquad \textit{(equação 3.21)}$$

Estas são usadas para prover as estimativas de máxima verossimilhança para a classificação Bayesiana.

Em uma colocação mais geral, nós poderíamos considerar uma mistura normal multivariada da seguinte forma:

$$P(x/y) = \sum_{k} w_k f_k(x/g_k, V_k) \qquad \textit{(equação 3.22)}$$

onde k é a distância fixa dos membros da mistura, w é um fator de ponderação, e a função f depende da média e estrutura de covariância dos membros da mistura. Para tais funções de densidade, uma distância iterativa é usada no lugar de uma aproximação analítica (Hand,1981).

***Discriminação Não-Paramétrica*:** Os métodos não-paramétricos dispensam as suposições relativas à função densidade de probabilidade. Dado um vetor de valores de parâmetro, x_0, a probabilidade de que qualquer ponto desconhecido caia em um local perto de x_0 pode ser definida em termos do volume relativo desta vizinhança. Se n' pontos caem nesta região, n é fixado como o total de pontos, e v é o volume da região, então a probabilidade que qualquer ponto desconhecido caia em um local perto de x_0 é n'/nv. Uma aproximação para classificação que surge fora disto é como segue.

Na aproximação do k-NN (k vizinho mais próximo), especificamos que o volume será definido pelos k NNs do ponto não classificado. Considere n_c destes k NNs para serem os

membros do grupo c, e n_y para serem os membros do grupo y. As probabilidades condicionais dos membros dos grupos c e y são então:

$$P(x_0 / c) = \frac{n_c}{nv} \qquad \textit{(equação 3.23)}$$

$$P(x_0 / y) = \frac{n_y}{nv} \qquad \textit{(equação 3.24)}$$

Conseqüentemente a regra de decisão é: nomeie para o grupo c se

$$\frac{n_c}{nv} > \frac{n_y}{nv} \qquad \textit{(equação 3.25)}$$

isto é $n_c > n_y$

Determinar os NNs requer a definição de distância: a distância Euclidiana é normalmente usada.

Uma propriedade teórica interessante da regra do NN relaciona isto como uma taxa Bayesiana de classificação errada. O último está definido como

$$1 - \max_y P(y / x_0) \qquad \textit{(equação 3.26)}$$

ou, usando anotação previamente introduzida,

$$1 - P(c / x_0) \qquad \textit{(equação 3.27)}$$

Esta é a probabilidade de classificação errada de x_0, determinando assim que ele deveria ser classificado no grupo c.

Na aproximação 1-NN, a taxa de classificação errada é o produto de: a probabilidade condicional do grupo y dado o vetor x, e um menos a probabilidade condicional do grupo y dado o vetor do NN de x:

$$\sum_{todos\,y} P(y / x)(1 - P(y / NN(x)) \qquad \textit{(equação 3.28)}$$

Esta é a probabilidade que nós atribuímos para classificar y dado que o NN não está nesta classe. Para isto pode-se mostrar que a taxa de classificação errada na aproximação 1-NN (equação 3.28) é duas vezes menor que a taxa de classificação errada Bayesiana.

Discriminação Bayesiana: Vamos partir de um exemplo para introduzirmos a Discriminação Bayesiana. Como regra geral, é melhor se tentarmos levantar o máximo de informações possíveis sobre o problema em questão. Olhando sucessivamente as características dos problemas relacionados teremos como definir as dificuldades na implementação das soluções. Superando estas dificuldades outras aproximações para o problema serão conduzidas, como será visto.

Considere um vetor de parâmetros $\boldsymbol{x}$, relativo a atributos de galáxias. Logo considera que uma amostra de galáxias que vão ser estudadas consiste em 75% espirais e 25% elípticas. Isso é

$$P(espiral) = 0{,}75$$

$$P(elíptica) = 0{,}25$$

onde P(.) denota probabilidade. Na ausência de qualquer outra informação, nós nomearíamos qualquer galáxia desconhecida então à classe de espirais. No final das contas, nós estaríamos corretos em 75% de casos, mas nós derivamos uma regra muito crua obviamente.

Considere agora que nós temos probabilidades condicionais: para um particular valor de parâmetro, x_0, nós temos

$$P(espiral / x_0) = 0{,}3$$

$$P(elíptica / x_0) = 0{,}7$$

Neste caso, nós somos levados a escolher a classe elíptica para nossa galáxia desconhecida para a qual nós medimos o parâmetro x_0. As probabilidades condicionais usadas acima são chamadas de probabilidades a priori.

Isto conduz a Regra de Bayes para designar um objeto desconhecido para um grupo c:

$P(c / x_0) > P(y / x_0)$ para todos $y \neq c$, *(equação 3.29)*

Uma dificuldade surge com a Regra de Bayes definida acima: embora nós pudéssemos tentar determinar $P(c/x)$ para todos os possíveis valores de x (ou, talvez, para uma parte destes valores), isto é incômodo. Na realidade, é normalmente mais simples derivar valores para $P(x_0/c)$, isto é a probabilidade de ter um determinado jogo de medidas, x_0, determina que nós estamos lidando com uma determinada classe, c. Tais probabilidades são chamadas de *a posteriori*. O teorema de Bayes relaciona *a priori* e *a posteriori*. Nós temos:

$$P(c/x_0) = \frac{P(x_0/c)P(c)}{\sum_{todos\cdot y} P(x_0/y)P(y)}, \qquad \textit{(equação 3.30)}$$

Substituindo a equação 3.30 em ambos os lados da equação 3.29 e cancelando o denominador comum, dá uma regra da seguinte forma: escolha o grupo c em relação a todos os grupos y, se

$$P(x_0/c)P(c) > P(x_0/y)P(y) \text{ para todos } y \neq c \qquad \textit{(equação 3.31)}$$

Esta forma da Regra de Bayes é melhor que a anterior. Mas novamente uma dificuldade surge: muitas amostras são necessárias para estimar os termos da equação 3.31. Conseqüentemente é conveniente fazer suposições de distribuição dos dados, sendo que sempre figura a distribuição normal.

A função de densidade normal multivariada é escolhida para representar melhor a distribuição de x do que um único ponto como antes. A densidade está definida como

$$(2\pi)^{-\frac{n}{2}}|V|^{-\frac{1}{2}} \exp(-\frac{1}{2}(x-g)'V^{-1}(x-g)) \qquad \textit{(equação 3.32)}$$

onde **V** é a matriz de covariância, de dimensões m x m, se m é a dimensionalidade do espaço. Se **V** for igual à matriz de identidade, indica que o **x** tem distribuição perfeitamente simétrica sem direção privilegiada de alongamento. | V | é o determinante da matriz **V**.

Assumindo que cada grupo c, tem uma distribuição normal, nós temos

$$P(x/c) = (2\pi)^{-\frac{n}{2}}|V_c|^{-\frac{1}{2}} \exp(-\frac{1}{2}(x-g_c)'V_c^{-1}(x-g_c) \qquad \textit{(equação 3.33)}$$

onde gc é o centro do grupo c, e Vc é a sua matriz de covariância.

Substituindo isto na equação 3.31 e cancelando condições comuns em ambos os lados, temos a seguinte regra: nomeie x para o grupo c se

$$\ln|V_C| + (x - g_c)'V_c^{-1}(x - g_c) - \ln P(c) < \ln|V_Y| + (x - g_y)'V_y^{-1}(x - g_y) - \ln P(y) \quad \text{para todos } y \neq c \qquad \textit{(equação 3.34)}$$

Esta expressão é simplificada definindo um "score discriminante" como:

$$\delta_c(x) = \ln|V_c| + (x - g_c)'V_c^{-1}(x - g_c) \qquad \textit{(equação 3.35)}$$

A regra se torna então: nomeie x para o grupo c se

$$\delta_c(x) - \ln P(c) < \delta_y(x) - \ln P(y) \text{ para todos } y \neq c \qquad \textit{(equação 3.36)}$$

A curva que divide entre qualquer dois grupos está definida por

$$\delta_c(x) - \ln P(c) = \delta_y(x) - \ln P(y) \text{ para todos } y \neq c \qquad \textit{(equação 3.37)}$$

A forma de uma curva definida para esta equação é quadrática. Conseqüentemente esta forma geral de Discriminação Bayesiana também será chamada de Discriminação Quadrática.

3.7 - Comparação da abordagem feita na Análise Discriminante com a Regressão Logística

Muitos autores comparam a abordagem feita na Análise Discriminante com a Regressão Logística. A Análise Discriminante é apropriada quando a variável dependente é não métrica, entretanto quando a variável dependente possui dois grupos podemos preferir por várias razões a Regressão Logística. Primeiro, para utilizar a Regressão logística não são necessárias às suposições de que a amostra tenha uma distribuição normal multivariada e que as matrizes de variância/covariância sejam iguais dento dos grupos. Ambas técnicas têm testes estatísticos diretos, habilidade para incorporar efeitos não lineares, e uma gama extensiva de

diagnósticos. Por estas e outras razões técnicas a Regressão Logística é equivalente à Análise Discriminante para dois grupos, e pode ser mais satisfatória em muitas situações.

Uma das vantagens da Regressão Logística é que nós só precisamos saber se um evento (comprar ou não, fracasso ou sucesso) aconteceu e podemos, então, usar um valor dicotômico como variável dependente. O procedimento prediz a estimativa da probabilidade que o evento vai ou não acontecer. Se a probabilidade predita for maior que 0,5, então a predição é sim, caso contrário é não. A técnica fornece a razão de chances (*odds ratio*), que pode ser expressa como:

$$\frac{\Pr ob_{(evento\text{-}ocorrer)}}{\Pr ob_{(evento\text{-}não\text{-}ocorrer)}} = e^{\beta_0+\beta_1X_1+.....+\beta_NX_N} \qquad \textit{(equação 3.38)}$$

Os coeficientes estimados $(\beta_0, \beta_1, \beta_2, ..., \beta_n)$ medem, de fato, as mudanças na relação das probabilidades, o *odds ratio*. Além disso, eles são expressos em logaritmos e precisam ser transformados de forma que o efeito relativo destes coeficientes nas probabilidades seja avaliado mais facilmente. Softwares estatísticos executam este procedimento automaticamente e fornecem os dois coeficientes, o atual e o transformado. O uso deste procedimento não muda a maneira de como interpretamos o sinal dos coeficientes. Um coeficiente positivo aumenta a probabilidade, enquanto que um valor negativo diminui a probabilidade predita.

Se β_i é positivo, sua transformação (antilog) será maior que 1, e o *odds ratio* aumentará. Este aumento acontece quando a probabilidade predita, de que um evento ocorra, aumentar e a probabilidade predita de que não ocorra reduzir. Assim o modelo tem uma probabilidade predita mais alta de ocorrência. Igualmente, se β_i for negativo, o antilog é menor que um e o *odds ratio* será diminuído. Um coeficiente de zero compara a um valor de 1,0 que resulta em nenhuma mudança no *odds ratio*. Variáveis (ou fatores) associadas a valores de exp(B) maiores do que 1, que sejam estatisticamente significantes, são interpretadas como fatores de risco para o desfecho. Variáveis associadas a valores de exp(B) menores do que 1 são variáveis "protetoras" ou maiores valores destas variáveis diminuem o risco do desfecho. Valores igual a 1 não seriam significantes.

Para representar a relação entre a variável dependente e as variáveis independentes, os coeficientes têm que representar relações não lineares de fato entre a variável dependente e as variáveis independentes. Embora o processo de transformação dos logaritmos prevê uma linearização da relação, devemos lembrar que os coeficientes na verdade representam declives diferentes na relação pelos valores da variável independente.

A Regressão Logística é semelhante à Regressão Múltipla em muitos de seus resultados, mas é diferente no método de calcular os coeficientes. Em vez de minimizar o quadrado dos desvios, a Regressão Logística maximiza a probabilidade de que o evento ocorra. Quando é usada esta técnica de estimação alternativa deve-se avaliar o ajuste do modelo de modo diferente.

A medida global que nos mostra como está ajustado o modelo é determinado pelo valor da razão de verossimilhança (-2LL). Um modelo bem-ajustando terá um valor pequeno de -2LL, sendo que o valor mínimo para -2LL é zero. O valor da razão de verossimilhança pode ser comparado entre as equações, sendo que a diferença representa a mudança no ajuste ao incluirmos ou excluirmos variáveis independentes. Softwares estatísticos têm testes automáticos para verificar a significância destas diferenças, e isto é feito pelo teste qui-quadrado. Um modelo só com a constante, que é semelhante a calcular o total da soma de quadrados usando a média, é base para a comparação.

Pode-se avaliar o ajuste do modelo global até certo ponto semelhante à avaliação na Regressão Múltipla, e, também, aplicar métodos que empregam caráter não métrico à variável dependente. Podemos usar o método das matrizes de classificação desenvolvido na Análise Discriminante para avaliar a precisão de predição em termos dos membros dos grupos.

Com os mesmos dados utilizados na aplicação da Análise Discriminante com dois grupos realizamos uma Regressão Logística tendo como objetivo a comparação. Entretanto, não tivemos resultados satisfatórios.

4. Utilização de Softwares

Neste capítulo inicialmente será apresentado um resumo do que está disponível no *software* SPSS para o procedimento da Análise Discriminante.

4.1 Como Proceder Através do Software SPSS

O Subprograma DISCRIMINANTE apresentado no SPSS foi planejado e programado por James Tuccy *(Vogelback Computing Center, North-western University)* e Willian Klecka. O programa executa a Análise Discriminante utilizando diretamente todas as variáveis escolhidas para a análise (método direto) ou utiliza o método *stepwise* que seleciona passo a passo o conjunto das variáveis escolhidas. Os critérios disponíveis para controlar a seleção da variável pelo método *stepwise* são: o λ de Wilks, a distância de Mahalanobis , o maior F mínimo entre os grupos, a maior correlação múltipla, e o maior V de Rao.

Quando há valores perdidos (*missing values*) das variáveis serão excluídas da análise todas as informações relativas ao caso, isto é, a linha inteira correspondente ao caso será excluída da análise. Há formas na literatura para imputação de casos perdidos, isto é, substituir o valor faltante por uma estimativa do seu valor. Formas comuns são: substituir o "*missing*" pela média da variável, utilizar o método da Regressão para estimar o valor "*missing*" a partir de outra variável fortemente relacionada com a variável onde há o valor perdido e ainda por outros métodos.

Os valores das funções discriminantes podem produzir análises adicionais e gráficos podem ser elaborados utilizando os valores das duas primeiras funções discriminantes. Quando mais de dois grupos estão sendo analisados, os números máximos de funções discriminantes a serem derivadas podem ser controlados por vários critérios. Matrizes de covariância utilizadas durante a análise podem ser armazenadas para serem usadas como contribuição em futuras Análises Discriminantes.

A opção *Statistics* fornece o cálculo de estatísticas adicionais. Estas incluem a média e o desvio padrão para cada grupo e para todos os casos, a matriz de covariância e a matriz de correlação dentro dos grupos, a matriz de covariância para cada grupo, um teste para a igualdade das matrizes de covariância, e vários testes F.

Há também opções que incluem a confecção de gráficos.

Na próxima seção serão apresentados os procedimentos para efetuar a Análise Discriminante, com breve nota explicativa.

4.1.1 - Como Obter uma Análise Discriminante através do SPSS

Para acessar a opção da Análise Discriminante, deve-se escolher o menu conforme a figura 4.1. Em seguida abre-se a caixa de diálogo que pode ser vista na figura 4.2.

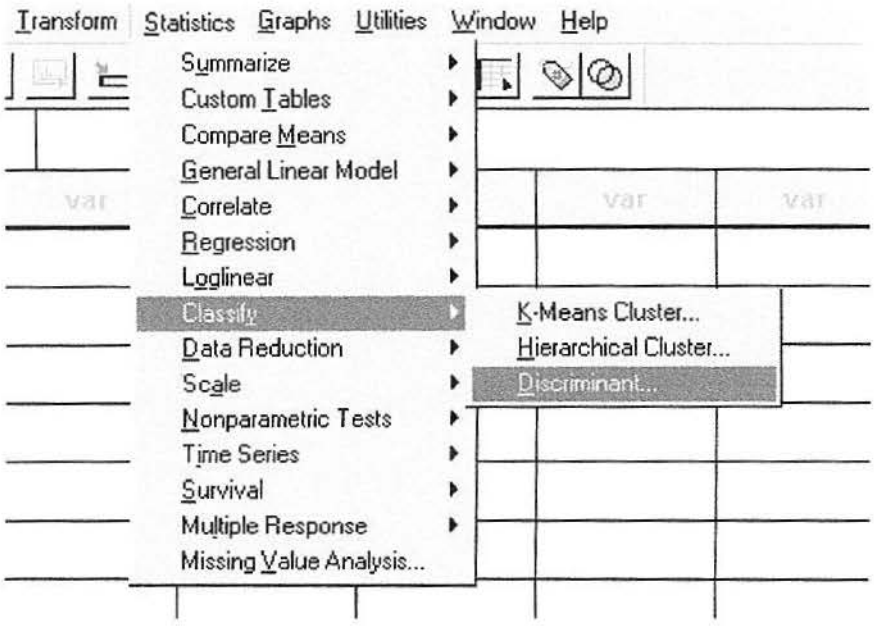


Figura 4.1: Menu para acessar a opção da Análise Discriminante

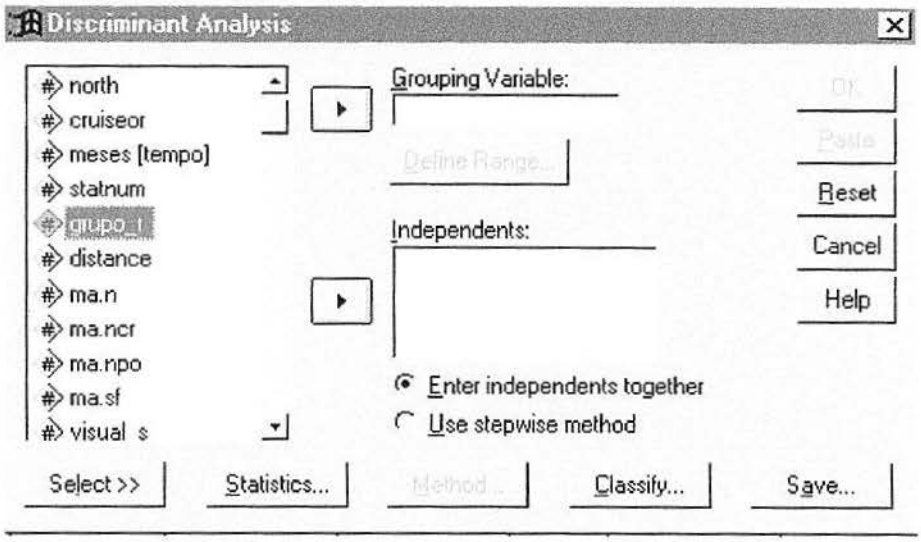


Figura 4.2: Caixa de Diálogo Principal da Análise Discriminante

Na caixa de diálogo da figura 4.2, seleciona-se a variável dependente escolhida para análise e que divide os dados em dois ou mais grupos. Depois de selecionar *Grouping Variable* use *Define Range* para definir as categorias mínima e máxima da variável de agrupamento. A seguir, selecionam-se as variáveis independentes a serem usadas na análise.

A opção *Select,* que também aparece na caixa de diálogo principal da Análise Discriminante (figura 4.2), permite limitar a análise a um subconjunto de casos.

Em *Statistics* pode-se fazer solicitações opcionais das estatísticas descritivas, coeficientes das funções e matrizes conforme figura 4.3.

Nas estatísticas descritivas a opção *means* mostra a média total, a média dos grupos e os desvios padrões para as variáveis independentes. Em *Univariate ANOVAs* é executada uma Análise de Variância para a igualdade das médias dos grupos para cada variável independente. Através do *Box's M* temos um teste de igualdade das matrizes de covariância das variáveis independentes nos grupos da variável dependente e se o teste apresentar uma significância estatística maior que o nível de significância estabelecido ($\alpha = 0{,}01$ ou $\alpha = 0{,}001$) então as matrizes de covariância são iguais, caso contrário os grupos terão diferentes matrizes de covariância e a suposição de homogeneidade não está satisfeita. No entanto esta não é uma exigência tão restritiva.

Em relação aos coeficientes das funções pode-se optar por *Fisher's* ou *Unstandardized,* sendo que em *Fisher's* temos os coeficientes das funções de classificação de Fisher usados diretamente para a classificação de novos casos e estes são nomeados ao grupo

para o qual temos o maior valor discriminante. Em *Unstandardized* exibem-se os coeficientes das funções discriminantes não padronizados.

As matrizes podem ser agrupadas como: *within-groups correlation* que é a matriz de correlação dentro dos grupos, obtida calculando a média das matrizes de covariância separadas para todos os grupos antes de calcular as correlações. Podemos também optar por *separate-groups covariance* que exibem as matrizes de covariância separadas para cada grupo. Ao optarmos por *total covariance* temos uma matriz de covariância de todos os casos como se eles fossem uma única amostra.

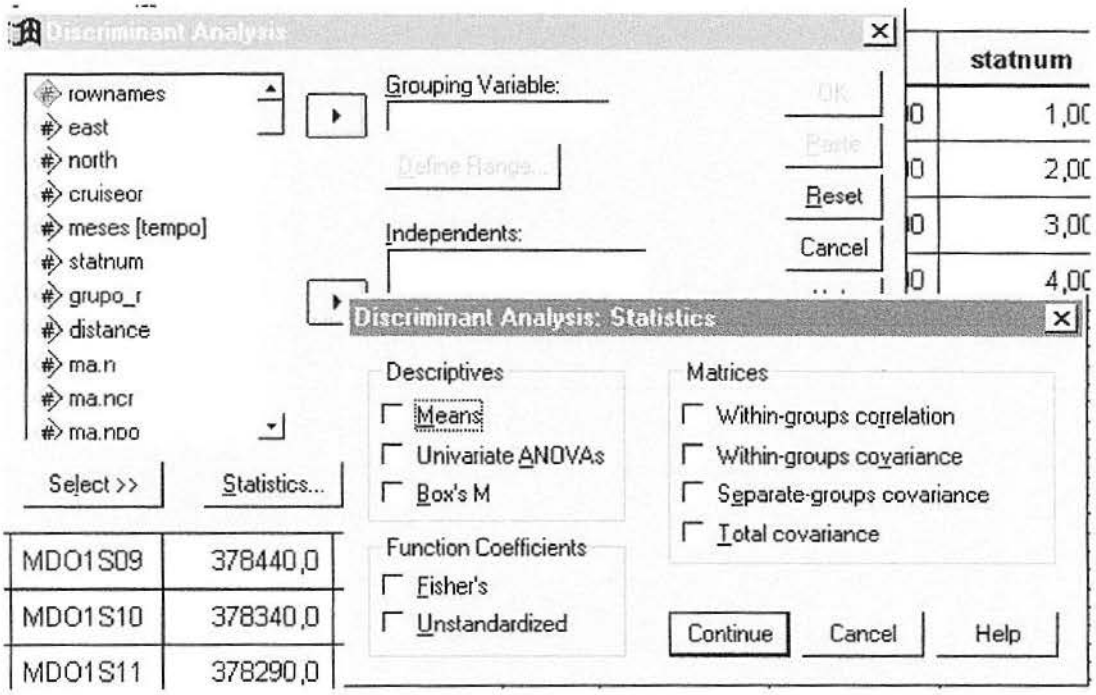


Figura 4.3: Caixa de Diálogo da Análise Discriminante para as Estatísticas

Então, clicando-se em *Classify* abre-se nova caixa, mostrada na figura 4.4, a qual permite controlar as probabilidades *a priori*, a apresentação dos resultados de classificação, o uso da matriz de covariância e a solicitação de gráficos.

Em relação às probabilidades *a priori* ao optarmos por *all groups equal* assumimos que as probabilidades *a priori* dos grupos são iguais, em *compute from group sizes* as probabilidades *a priori* são proporcionais aos tamanhos dos grupos.

Casewise results exibe os códigos do grupo original, do grupo predito, as probabilidades *a posteriori* e os valores das funções discriminantes para cada caso, sendo que podemos limitar esta apresentação apenas para os primeiros *n* casos (digamos 100 primeiros casos quando a amostra é grande). O número de casos classificados corretamente e incorretamente em cada grupo é apresentado quando solicitamos *summary table* (tabela de classificação). Com o objetivo de classificar um caso na análise usando as funções derivadas de todos os casos diferentes daquele optamos por *leave-one-out classification* que também é conhecido por U-método, conforme descrito no capítulo 3.5. Podemos substituir os valores perdidos pelas médias das variáveis preditas, durante a classificação, através de *replace missing values with mean* e desta forma utilizá-los.

A matriz de covariância, na classificação de casos, pode ser agrupada dentro dos grupos através de *within-groups* ou em grupos separados em *separate-groups*.

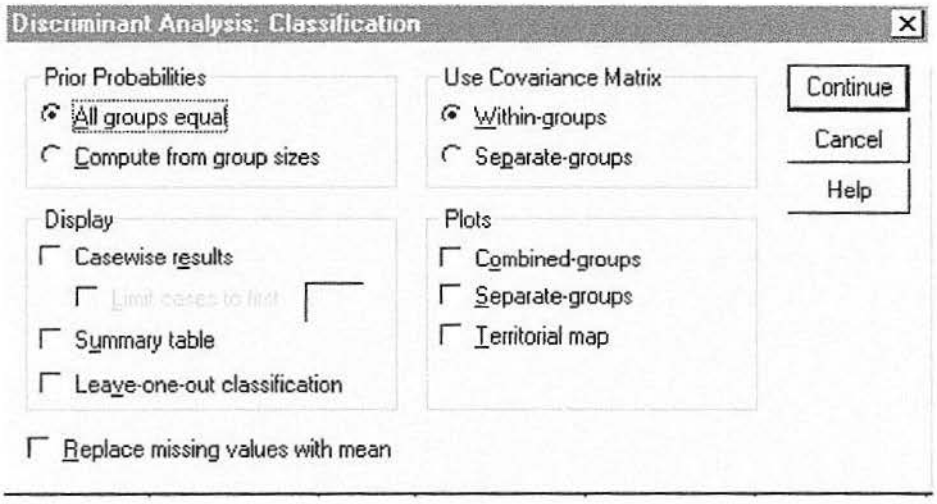


Figura 4.4: Caixa de Diálogo da Análise Discriminante para a Classificação

Ainda nesta caixa, mostrada na figura 4.4, tem-se a opção de solicitar gráficos tais como: *combined-groups, separate-groups* e *territorial map.* A opção *combined-groups* cria um gráfico de todos os grupos utilizando os valores das duas primeiras funções discriminantes, sendo que se houver só uma função é exibido um histograma. Em *separate-groups* exibe-se um gráfico com grupos separados utilizando, também, os valores das duas primeiras funções discriminantes, e é exibido um histograma se houver só uma função. O *territorial map* é um gráfico de limites usado para classificar os casos dentro dos grupos

baseado nos valores das duas funções discriminantes. Os números correspondem aos grupos nos quais são classificados os casos. O centróide, para cada grupo, é indicado por um asterisco. O mapa não é exibido se houver só uma função discriminante.

Em relação ao método podemos optar por entrar com todas as variáveis independentes que satisfaçam os critérios de tolerância simultaneamente ou utilizar o método *stepwise* para controlar a entrada e remoção da variável, conforme podemos ver na caixa de diálogo principal (figura 4.2). Quando optamos pelo método *stepwise* fica disponível a opção *method* na caixa de diálogo principal.

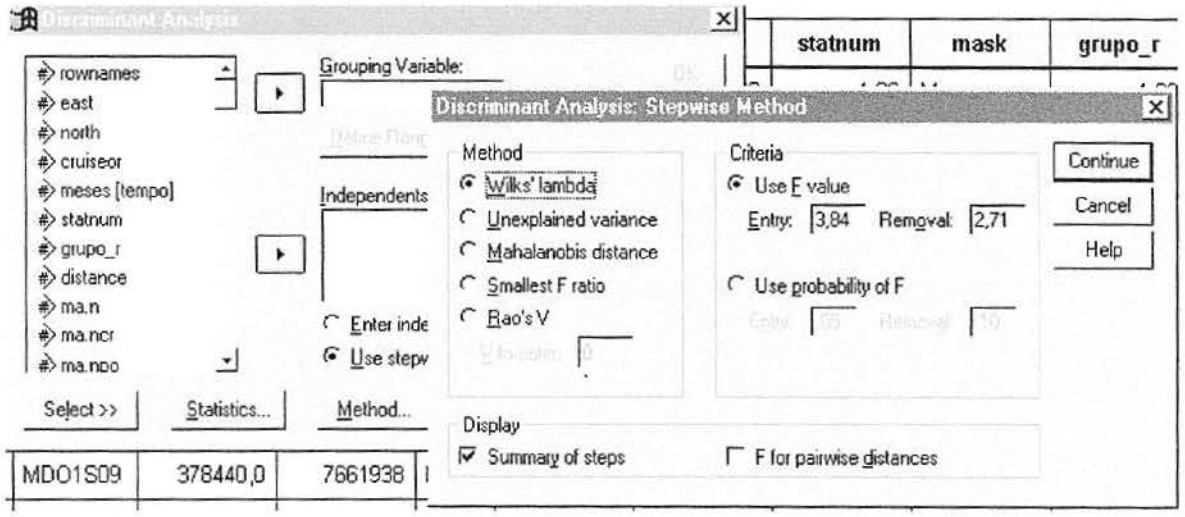


Figura 4.5: Caixa de Diálogo da Análise Discriminante para o Método

Clica-se nesta opção e temos disponíveis cinco critérios para controlar a variável (figura 4.5): o *Wilks` lambda* que escolhe as variáveis para entrada na equação com base no valor mínimo, o *Unexplained variance* que entra com a variável que minimiza a variância, o *Mahalanobis distance* que é uma medida de quanto os valores de um caso nas variáveis independentes diferem da média de todos os casos, o *Smallest F ratio* que é baseado na maximização do valor de F, e o *Rao`s V* que entra com a variável que maximiza este.

Em *summary of steps* exibe-se as tabelas para o λ de Wilks, as variáveis que entram e que são removidas, as variáveis que estão na análise e as que não estão. A tolerância e o valor da estatística usado para a seleção da variável são informados para todas as variáveis. O valor de F, o nível de significância e a tolerância mínima também são

apresentados nesta opção. Com o objetivo de mostrar a matriz com as relações de F para cada par de grupos, clicamos em *F for pairwise distances.*

Pode-se usar o valor de F ou a probabilidade de F como critério para a entrada e remoção de variáveis na análise. Quando optamos por *use F value* uma variável entra no modelo se seu valor F é maior que o valor de entrada, e é afastada do modelo se o seu valor F for menor que o valor de remoção. O valor de entrada deve ser maior que o valor de remoção e ambos os valores devem ser positivos. Entrando com mais variáveis no modelo, diminui o valor de entrada e removendo mais variáveis do modelo aumenta o valor de remoção.

Se *use probability of F* for selecionado, uma variável é incluída no modelo se o nível de significância de seu valor F é menor que o valor de entrada, e é removida se o nível de significância de seu valor F é maior que o valor de remoção. O valor de entrada deve ser menor que o valor de remoção e ambos os valores devem ser positivos. Incluindo mais variáveis, no modelo, aumenta o valor de entrada e removendo mais variáveis do modelo, diminui o valor de remoção.

Após, clica-se em *continue* e na caixa de diálogo principal (figura 4.2) seleciona-se *save* para a opção de salvar novas variáveis. A nova caixa (figura 4.6) apresenta as opções a serem salvas. Em *predicted group membership* salva-se no banco de dados o grupo com a probabilidade à posteriori maior, baseado nos valores discriminantes. Na opção *discriminant scores* temos valores calculados pela multiplicação dos coeficientes discriminantes não padronizados e os valores das variáveis independentes. Somam-se estes produtos e adiciona-se uma constante. Um valor é salvo para cada função discriminante derivada e o valor médio para todos os casos combinados é 0 e a variância dentro dos grupos 1. Com *probabilities of group membership* cria-se tantas variáveis quantos grupos existirem. A primeira variável contém a probabilidade a posteriori associada ao primeiro grupo, a segunda nova variável contém a probabilidade associada ao segundo grupo, e assim sucessivamente.

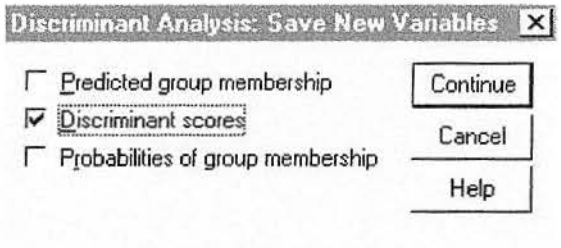


Figura 4.6: Caixa de Diálogo da Análise Discriminante para Salvar

Clique em *continue* e na caixa de diálogo principal (figura 4.2) clique em *Ok*.

O software SPSS mostrará os resultados da Análise Discriminante.

4.1.2 - Opções Adicionais Disponíveis Através da Sintaxe

Apesar da maioria das opções em Análise Discriminante estarem disponíveis diretamente nas caixas de diálogo, algumas são possíveis somente através dos comandos em sintaxe. Para isto, basta editar os comandos e subcomandos necessários na janela de sintaxe e "rodar" a execução do programa.

O comando, para o caso da Análise Discriminante, é o DISCRIMINANT, e os subcomandos que apresentam opções extras são os seguintes:

ANALYSIS – executa análises discriminantes múltiplas e controla a ordem na qual entram as variáveis.

PRIORS – especifica as probabilidades a priori (anteriores) para a classificação.

ROTATE – exibe a rotação padrão e as matrizes de estrutura.

FUNCTIONS – Limita o número de funções discriminantes extraídas.

SELECT – restringe a classificação aos casos selecionados (ou não selecionados) na análise.

MATRIX – lê e analisa uma matriz de correlação sendo que pode também escrever uma matriz de correlação para análise posterior.

Estas informações, com alguns exemplos, podem ser encontradas detalhadamente no manual do SPSS (*SPSS – Statiscal Package for the Social Science*). Por hora, é pertinente apenas que se mencione que existem essas possibilidades, sendo que, quando houver interesse, é aconselhável consultar o referido manual.

5. Aplicações práticas e interpretações utilizando dados do Projeto MAPEM

Para ilustrar a aplicação da Análise Discriminante com dois grupos e a Análise Discriminante Múltipla vamos usar dados do Projeto **MAPEM** (MONITORAMENTO AMBIENTAL EM ATIVIDADES DE PERFURAÇÃO EXPLORATÓRIA MARÍTIMA).

As variáveis independentes foram selecionadas através de pesquisa prévia com todas as áreas envolvidas no Projeto e o método escolhido foi o Método Direto em que todas as variáveis previamente selecionadas entram na análise.

O Projeto MAPEM consta do seguinte: temos três cruzeiros, cada um deles com 54 amostras. O cruzeiro 1 corresponde às 54 amostras que foram retiradas antes do início das perfurações, o cruzeiro 2 às 54 amostras retiradas após ter transcorrido um mês das perfurações e o cruzeiro 3 às 54 amostras retiradas após um período de um ano das perfurações.

Nossa aplicação será demonstrada usando as 54 observações correspondentes ao cruzeiro 2.

As 32 (trinta e duas) variáveis independentes utilizadas na análise vão ser as mesmas tanto para o caso com dois grupos como para o caso com três grupos (múltipla). São elas: MA.N, MA.NCR, MA.NPO, MA.SF, ME.N, ME.NNE, ME.SF, ME.SG, TOC, Cr, Ni, Mn, Pb, Cd, Zn, Fe, As, Ba, Co, V, Al, Cu, VA_CAR, Vágeis, Sed_car, Sed_det , Sed_filt, A1, B1, A2, B2 e TPH.

A tabela 5.1 apresenta as descrições destas variáveis e as suas unidades.

Tabela 5.1 - Variáveis independentes

Variáveis	Descrições	Unidades
MA.N	Densidade	n°ind./0,1m²
MA.NCR	Densidade de Crustáceo	n°ind./0,1m²
MA.NPO	Densidade de Polychaeta	n°ind./0,1m²
MA.SF	Número de Taxa	n°familias/0,1m²
ME.N	Densidade da meiofauna	inds./3,14 cm^2
ME.NNE	Densidade de nemátodas	inds./3,14 cm^2
ME.SF	Núm. famílias nemátodas	número/3,14 cm^2
ME.SG	Núm. gêneros nemátadas	número/3,14 cm^2
TOC	Carbono Orgânico Total	%
Cr	Cromo	mg/kg
Ni	Níquel	mg/kg
Mn	Manganês	%
Pb	Chumbo	mg/kg
Cd	Cádmio	mg/kg
Zn	Zinco	mg/kg
Fe	Ferro	%
As	Arsênio	mg/kg
Ba	Bário	mg/kg
Co	Cobalto	mg/kg
V	Vanádio	mg/kg
Al	Alumínio	%
Cu	Cobre	mg/kg
VA_CAR	Vágeis Carnívoros	n°ind/0,1m²
VÁGEIS	Vágeis Detritívoros	n°ind/0,1m²
Sed_car	Sedentários Carnívoros	n°ind/0,1m²
Sed_det	Sedentários Detritívoros	n°ind/0,1m²
Sed_filt	Sedentários Filtradores	n°ind/0,1m²
A1	Detritívoros seletivos	%
B1	Detritívoros não-seletivos	%
A2	Fitobentófagos	%
B2	Predadores/onívoros	%
TPH	Hidrocarbonetos total petróleo	

5.1 - Análise Discriminante (dois grupos)

No caso para dois grupos vamos utilizar a variável dependente Grupo_r (variável categórica com dois grupos). O grupo 1 corresponde a amostras retiradas a uma distância até 500 m da plataforma e o grupo 2 a amostras retiradas a uma distância de 2.500 m da plataforma de perfuração. O objetivo é identificar as variáveis que diferenciam

significativamente estes dois grupos. A tabela A.1 mostra as estatísticas descritivas (média e desvio padrão) das variáveis independentes para os dois grupos.

Tabela 5.2 - Teste da igualdade das médias nos Grupos

	Wilks' Lambda	F	df1	df2	Sig.
MA.N	1,000	,020	1	49	,888
MA.NCR	,966	1,720	1	49	,196
MA.NPO	,997	,161	1	49	,690
MA.SF	,994	,304	1	49	,584
ME.N	,983	,851	1	49	,361
ME.NNE	,987	,637	1	49	,428
ME.SF	,981	,969	1	49	,330
ME.SG	,990	,491	1	49	,487
TOC	,956	2,276	1	49	,138
Cr	,921	4,196	1	49	,046
Ni	,986	,684	1	49	,412
Mn	,894	5,809	1	49	,020
Pb	,931	3,654	1	49	,062
Cd	,915	4,556	1	49	,038
Zn	,997	,160	1	49	,691
Fe	,943	2,970	1	49	,091
As	,888	6,188	1	49	,016
Ba	,960	2,050	1	49	,159
Co	,999	,055	1	49	,816
V	1,000	,004	1	49	,953
Al	,983	,869	1	49	,356
Cu	,980	1,012	1	49	,319
VÁGEIS	,984	,790	1	49	,378
VA_CAR	,989	,553	1	49	,461
Sed_car	,898	5,583	1	49	,022
Sed_det	,991	,466	1	49	,498
Sed_filt	1,000	,004	1	49	,952
A1	,991	,451	1	49	,505
B1	,999	,026	1	49	,872
A2	,991	,431	1	49	,515
B2	,992	,408	1	49	,526
TPH	,936	3,341	1	49	,074

A tabela 5.2 mostra-nos que o λ de Wilks é significativo pelo teste F para as seguintes variáveis independentes: As, Mn, Sed_car, Cd e Cr ou seja a diferença entre a média dos dois grupos é significante para estas variáveis. O menor valor para o λ de Wilks determina a variável mais importante para a análise discriminante, sendo que para nossa aplicação é o arsênio.

A estatística M de Box testa a hipótese de nulidade da igualdade entre as matrizes de covariância dos dois grupos. Ele não foi executado. A Análise Discriminante é robusta e pode ser utilizada mesmo sem ter sido feito o teste.

O teste de tolerância retirou da análise as variáveis independentes Sed_filt e B2 conforme nos mostra a tabela 5.3.

Tabela 5.3 - Variáveis reprovadas Teste Tolerância[a]

	Variância dentro dos Grupos	Tolerância	Tolerância Mínima
Sed_filt	,178	,2048	,0007
B2	59,826	5,902E-14	1,699E-14

Todas as variáveis que passam no critério de tolerância entram simultâneamente.

a. Nível de tolerância mínima é ,001

A tabela 5.4 mostra os autovalores. Quanto maior o autovalor, melhor é explicada pela função discriminante a variação da variável dependente. Como nossa variável dependente (Grupo_r) tem somente dois grupos, há somente uma função discriminante. Entretanto, se houvesse mais grupos, nós teríamos funções discriminantes múltiplas e esta tabela seria apresentada em ordem decrescente de importância. A segunda coluna lista a porcentagem da variação explicada por cada função. A terceira coluna é a porcentagem cumulativa da variação explicada. A última coluna é a correlação canônica. O valor da correlação canônica é 0,893 e para interpretarmos este valor usamos o quadrado desta

correlação que é de 0,797. Então 79,7% é a proporção de variação da função discriminante que é explicada pelos grupos. Esta tabela é usada, às vezes, para decidir quantas funções são importantes.

Tabela 5.4 - Autovalores

Função	Autovalor	% da Variância	Acumulada %	Correlação Canônica
1	3,920 [a]	100,0	100,0	,893

a. Função discriminante canônica 1 foi usada na análise.

Este segundo teste do λ de Wilks serve para uma finalidade diferente do que o seu uso na tabela 5.2. Na tabela 5.5 ele testa o significado do autovalor para cada função discriminante. Nesta aplicação prática há somente um, e é significativo.

Tabela 5.5 - Wilks' Lambda

Teste das Funções	Wilks' Lambda	Qui-quadrado	df	Sig.
1	,203	54,174	30	,004

Os coeficientes padronizados da função discriminante que estão na tabela 5.6 indicam a importância relativa das variáveis independentes em predizer a variável dependente, eles são mais úteis para propósitos de interpretação.

A tabela 5.7 da matriz da estrutura mostra as correlações de cada variável independente com cada função discriminante. Neste caso, há somente uma função discriminante. Entretanto, quando a variável dependente tem mais categorias haverá mais funções discriminantes e mais colunas adicionais na tabela, uma para cada função. As variáveis independentes estão em ordem decrescente, em valor absoluto, da correlação com a

função discriminante. No nosso caso a variável As (Arsênio) tem a maior correlação com a função discriminante 1, sendo que a menor correlação é a da variável V (Vanádio).

Os coeficientes não padronizados da função discriminante encontram-se na tabela 5.8 e são mais fáceis de usar para calcular os valores Z discriminantes.

Tabela 5.6 -Coeficientes Padronizados da Função Discriminante Canônica

	Function
	1
MA.N	-1,738
MA.NCR	-2,320
MA.NPO	-2,742
MA.SF	-2,150
ME.N	,379
ME.NNE	,226
ME.SF	,474
ME.SG	-,912
TOC	-,150
Cr	,502
Ni	-,129
Mn	-,575
Pb	,527
Cd	-1,416
Zn	-1,034
Fe	,150
As	1,091
Ba	,983
Co	,091
V	,017
Al	-,005
Cu	,808
VÁGEIS	3,630
VA_CAR	1,835
Sed_car	,611
Sed_det	3,136
A1	,615
B1	,194
A2	,349
TPH	-,436

Tabela 5.7 Matriz Estrutura

	Function
	1
As	,179
Mn	,174
Sed_car	,170
Cd	-,154
Cr	,148
Pb	,138
TPH	-,132
Fe	,124
TOC	,109
Ba	-,103
MA.NCR	-,095
Cu	,073
ME.SF	,071
Al	-,067
ME.N	,067
VÁGEIS	-,064
Ni	,060
ME.NNE	,058
VA_CAR	,054
ME.SG	,051
Sed_det	,049
A1	-,048
A2	,047
B2[a]	,046
MA.SF	-,040
MA.NPO	-,029
Zn	-,029
Co	-,017
B1	-,012
Sed_filt[a]	,011
MA.N	,010
V	,004

Correlações dentro dos grupos entre as variáveis discriminantes e as funções discriminantes canônicas padronizadas
Variáveis ordenadas em ordem decrescente de correlação com a função.

a. Esta variável não foi utilizada na análise.

Tabela 5.8 Coeficientes Função Discriminante Canônica

	Função
	1
MA.N	-,150
MA.NCR	-,472
MA.NPO	-,387
MA.SF	-,493
ME.N	,016
ME.NNE	,012
ME.SF	,124
ME.SG	-,150
TOC	-,698
Cr	,066
Ni	-,030
Mn	-,002
Pb	,141
Cd	-19,907
Zn	-,085
Fe	,347
As	,173
Ba	,001
Co	,092
V	,002
Al	-,006
Cu	,248
VÁGEIS	,601
VA_CAR	,763
Sed_car	,697
Sed_det	,582
A1	,048
B1	,013
A2	,038
TPH	-,132
(Constante)	-2,855

Coeficientes não padronizados

A tabela 5.9 é usada para estabelecer o ponto do corte utilizado para classificar os casos. Se os dois grupos forem de tamanho igual, o melhor ponto do corte é a média entre os

valores centróides dos grupos das funções. Se os grupos forem desiguais, o ponto de corte ótimo é a média proporcional dos dois valores. Os casos que se encontram na função acima do ponto do corte são classificados no grupo 2 (distância de 2500 m), enquanto que aqueles que se encontram abaixo do ponto de corte são classificados como grupo 1 (distâncias < 500 m). Naturalmente, computacionalmente faz-se a classificação automaticamente, assim estes valores são para finalidades informativas.

Tabela 5.9 - Funções dos Centróides dos Grupos

	Função
GRUPO_R	1
<500m	-,709
2500m	5,315

Funções discriminantes canônicas não padronizadas avaliadas nas médias dos grupos

A tabela 5.10 apresenta os coeficientes das funções de classificação. Como nossa variável dependente apresenta 2 grupos temos então 2 funções de classificação, uma para cada grupo. Para o grupo 1 (distâncias < 500 m) temos a seguinte função de classificação:

C_1 = - 481,499 + 7,833**MA.N** – 6,708**MA.NCR** – 9,953**MA.NPO** – 6,964**MA.SF** – 0,381**ME.N** + 2,374**ME.NNE** – 2,139**ME.SF** – 6,742**ME.SG** + 53,260**TOC** – 3,189**Cr** + 2,208**Ni** – $7,12 \times 10^{-3}$**Mn** + 5,651**Pb** – 540,212**Cd** – 1,917**Zn** – 32,613**Fe** + 0,899**As** + $6,435 \times 10^{-2}$**Ba** + 32,889**Co** + 7,767**V** – 3,387**Al** + 3,435**Cu** + 4,074**VÁGEIS** + 0,155**VA_CAR** – 24,780**Sed_car** – 3,795**Sed_det** + 4,268**A1** + 4,219**B1** + 7,288**A2** + 0,812**TPH**

Tabela 5.10 Coeficientes Função de Classificação

	GRUPO_R	
	<500m	2500m
MA.N	7,833	6,929
MA.NCR	-6,708	-9,552
MA.NPO	-9,953	-12,286
MA.SF	-6,964	-9,936
ME.N	-,381	-,283
ME.NNE	2,374	2,445
ME.SF	-2,139	-1,390
ME.SG	-6,742	-7,649
TOC	53,260	49,055
Cr	-3,189	-2,789
Ni	2,208	2,028
Mn	-7,1E-03	-2,1E-02
Pb	5,651	6,500
Cd	-540,212	-660,121
Zn	-1,917	-2,430
Fe	-32,613	-30,523
As	,899	1,940
Ba	6,43E-02	7,202E-02
Co	32,889	33,441
V	7,767	7,780
Al	-3,387	-3,424
Cu	3,435	4,927
VÁGEIS	4,074	7,694
VA_CAR	,155	4,748
Sed_car	-24,780	-20,580
Sed_det	-3,795	-,292
A1	4,268	4,555
B1	4,219	4,299
A2	7,288	7,515
TPH	,812	1,669E-02
(Constante)	-481,499	-512,572

Função discriminante linear de Fisher

Para o grupo 2 (= 2500 m) utilizamos a segunda coluna da tabela 5.10 para a função de classificação, sendo ela calculada da mesma forma que a do grupo 1.

Com o objetivo de classificar um novo caso vamos usar estas duas funções. Os valores para as variáveis deste novo caso encontram-se na tabela 5.11.

Tabela 5.11 - Valores das variáveis

Variáveis	Valores
MA.N	41,000
MA.NCR	11,000
MA.NPO	25,000
MA.SF	20,000
ME.N	35,000
ME.NNE	33,000
ME.SF	11,000
ME.SG	13,000
TOC	1,200
Cr	68,800
Ni	33,700
Mn	955,000
Pb	23,800
Cd	0,330
Zn	81,100
Fe	3,600
As	24,200
Ba	999,000
Co	8,640
V	77,900
Al	3,340
Cu	23,000
VA_CAR	8,000
VÁGEIS	22,000
Sed_car	0,000
Sed_det	9,000
A1	23,330
B1	53,330
A2	16,670
TPH	2,150

Para as funções de classificação dos grupos 1 e 2 procedemos os cálculos e obtemos os valores que encontram-se na tabela 5.12.

Este caso obteve a maior classificação no **grupo 1**, logo podemos concluir que ele está classificado no grupo com distâncias < 500 m.

Tabela 5.12 - Funções de Classificação

Grupo	Valor da Função
1	474,526
2	443,046

A tabela A.2, que se encontra no Anexo A, lista o grupo original, o grupo predito baseado nas maiores probabilidades *a posteriori*, a probabilidade *a priori*, a probabilidade *a posteriori* (a chance que o caso pertença ao grupo predito, baseado nas variáveis independentes), a distância de Mahalanobis ao quadrado do caso ao centróide do grupo (valores grandes indicam outliers) e o valor da função discriminante para o caso. O caso é classificado baseado no valor discriminante em relação ao ponto de corte (não mostrado). Casos mal classificados são marcados com asteriscos. As colunas que correspondem ao segundo mais alto grupo mostram as probabilidades *a posteriori* e a distância de Mahalanobis para o caso que foi classificado baseado na segunda mais alta probabilidade *a posteriori*. Como temos somente dois grupos, o "segundo mais alto" é equivalente ao outro grupo.

Os gráficos separados por grupo utilizam os valores das duas primeiras funções discriminantes, sendo que, como temos só uma função exibimos um histograma para cada grupo.

A figura 5.1 representa o histograma do Grupo_r distâncias < 500 m e a figura 5.2 representa o histograma do Grupo_r distância igual a 2.500 m. Para termos uma boa função discriminante o gráfico deve apresentar muitos casos próximos à média e caudas com poucos casos.

Observamos, nas figuras 5.1 e 5.2 que temos um maior número de casos próximos a média e poucos casos nas caudas.

As figuras mencionadas encontram-se na próxima página.

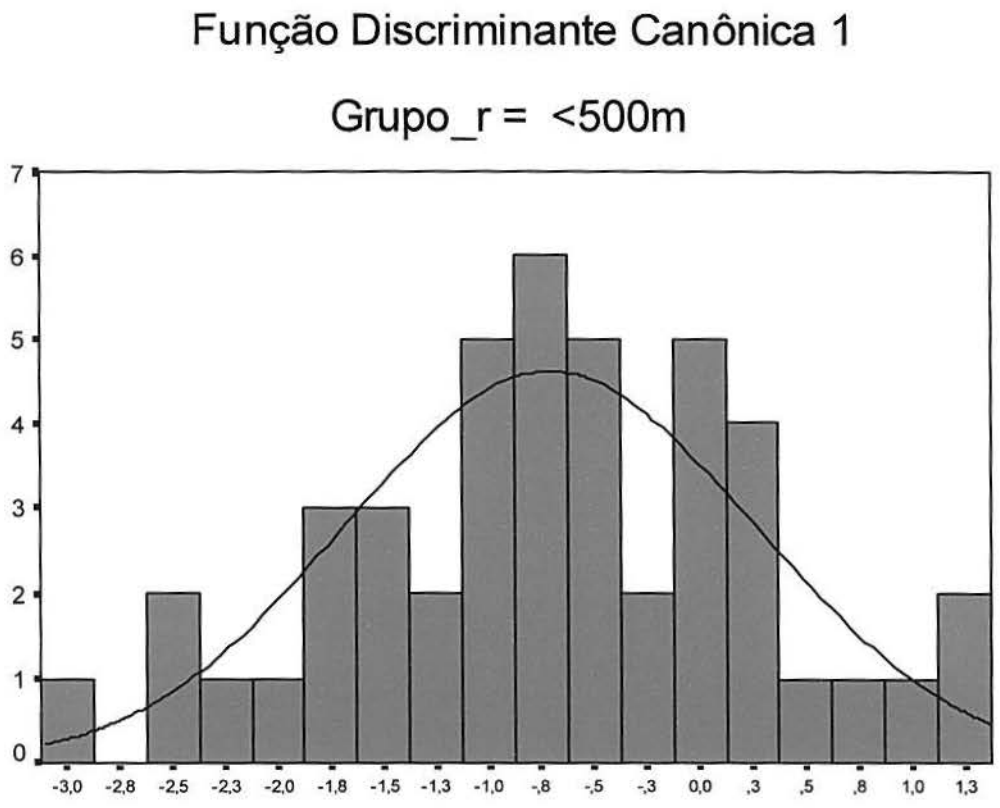


Figura 5.1 – Histograma da Função Discriminante 1 para o grupo com distâncias < 500 m

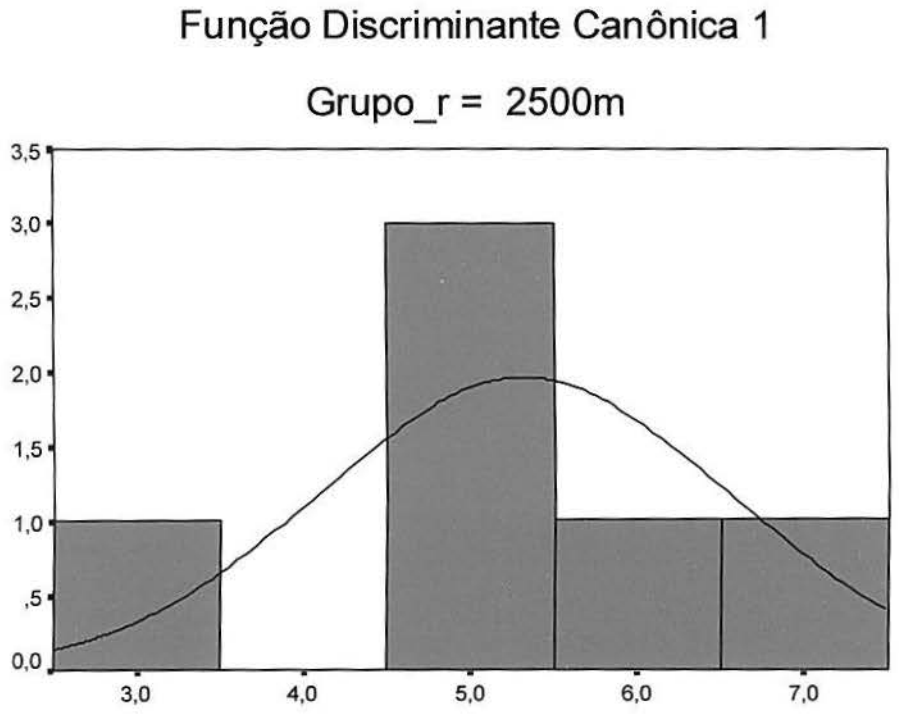


Figura 5.2 – Histograma da Função Discriminante 1 para o grupo com distância = 2.500 m

A tabela 5.13 é usada para avaliar se a função discriminante classifica bem, e se classifica igualmente bem para cada grupo da variável dependente. Aqui classificou corretamente todos os casos, um nível excelente de discriminação.

Tabela 5.13 - Resultados da Classificação [a]

	GRUPO_R	Predição dos Membros dos Grupos		Total
		<500m	2500m	
Count	<500m	45	0	45
	2500m	0	6	6
%	<500m	100,0	,0	100,0
	2500m	,0	100,0	100,0

a. 100,0% dos casos do grupo original foram classificados corretamente.

Com o objetivo de testar as variáveis que mais discriminam efetuamos uma Análise Discriminante com as variáveis independentes que tem o valor de probabilidade $p < 0,10$ na tabela 5.2 e que são: As, Mn, Sed_car, Cd, Cr, Pb, TPH e Fe. Tivemos 98% dos casos corretamente classificados o que nos dá evidências da importância destas variáveis em discriminar os grupos.

5.2 - Análise Discriminante (três grupos)

No caso para três grupos vamos utilizar a variável dependente mask_dis (variável categórica com três grupos). O grupo 1 corresponde a amostras utilizadas no controle interno, o grupo 2 são os pontos utilizados como referência e o grupo 3 é a máscara. O objetivo é identificar as variáveis que diferenciam significativamente estes três grupos. A tabela B.1, que encontra-se no anexo B, mostra as estatísticas descritivas (média e desvio padrão) das variáveis independentes para o grupo 1 (controle interno) e o grupo 2 (referência).

A tabela B.2, que encontra-se no anexo B, apresenta as estatísticas descritivas para o grupo 3 (máscara).

A tabela 5.14 mostra-nos que o λ de Wilks é significativo pelo teste F para as seguintes variáveis independentes: TPH, Fe, Mn Ba e Pb, ou seja a diferença entre a média dos três grupos é significante para estas variáveis. O menor valor para o λ de Wilks determina a variável mais importante para a análise discriminante que é o TPH.

Tabela 5.14 - Teste da igualdade das médias nos Grupos

	Wilks Lambda	F	df1	df2	Sig.
MA.N	,999	,030	2	48	,970
MA.NCR	,949	1,301	2	48	,282
MA.NPO	,989	,259	2	48	,773
MA.SF	,991	,222	2	48	,802
ME.N	,974	,630	2	48	,537
ME.NNE	,973	,665	2	48	,519
ME.SF	,979	,509	2	48	,604
ME.SG	,988	,280	2	48	,757
TOC	,950	1,272	2	48	,290
Cr	,915	2,229	2	48	,119
Ni	,951	1,231	2	48	,301
Mn	,828	4,978	2	48	,011
Pb	,846	4,380	2	48	,018
Cd	,900	2,660	2	48	,080
Zn	,995	,124	2	48	,884
Fe	,810	5,634	2	48	,006
As	,887	3,061	2	48	,056
Ba	,839	4,607	2	48	,015
Co	,969	,762	2	48	,472
V	,923	1,999	2	48	,147
Al	,962	,942	2	48	,397
Cu	,980	,500	2	48	,610
VA_CAR	,980	,499	2	48	,610
VÁGEIS	,982	,429	2	48	,653
Sed_car	,896	2,772	2	48	,073
Sed_det	,987	,319	2	48	,728
Sed_filt	,975	,607	2	48	,549
A1	,982	,432	2	48	,652
B1	,998	,055	2	48	,947
A2	,987	,311	2	48	,734
B2	,968	,787	2	48	,461
TPH	,770	7,180	2	48	,002

Embora as outras variáveis, que encontram-se na tabela 5.14, não tenham mostrado significância estatística pelo teste F, não significa que excluí-las da análise vai melhorar o poder discriminante das funções, pois na maioria das vezes o vetor das variáveis contém uma configuração não visível de forma univariada.

A estatística M de Box testa a hipótese de nulidade da igualdade entre as matrizes de covariância dos dois grupos. Ele não foi executado. A Análise Discriminante é robusta e pode ser utilizada mesmo sem ter sido feito o teste.

O teste de tolerância retirou da análise as variáveis independentes Sed_filt e B2 conforme nos mostra a tabela 5.15.

Tabela 5.15 - Variáveis reprovadas Teste tolerância [a]

	Variância dentro dos Grupos	Tolerância	Tolerância Mínima
Sed_filt	,178	,204	,001
B2	59,626	3,662E-14	1,031E-14

Todas as variáveis que passam no critério de tolerância entram simultâneamente

a. Nível de tolerância mínima é ,001.

A tabela 5.16 mostra os autovalores. Quanto maior o autovalor, melhor é explicada pela função discriminante correspondente a variação da variável dependente. Como nossa variável dependente (mask_dis) tem três grupos, há duas funções discriminantes e elas são apresentadas em ordem decrescente de importância. A segunda coluna lista a porcentagem da variação explicada por cada função. A terceira coluna é a porcentagem cumulativa da variação explicada. A última coluna é a correlação canônica. O valor da correlação canônica para a primeira função é 0,920 e para interpretarmos este valor usamos o quadrado desta correlação que é de 0,846. Então 84,6% é a proporção de variação da função discriminante 1 que é explicada pelos grupos. O valor da correlação canônica para a segunda função é 0,884,

ou seja, 78,1% é a proporção de variação da função discriminante 2 que é explicada pelos grupos.

Podemos verificar que as duas funções discriminantes são importantes.

Tabela 5.16 - Autovalores

Função	Autovalor	% da Variância	% Acumulada	Correlação Canônica
1	5,481 [a]	60,5	60,5	,920
2	3,573 [a]	39,5	100,0	,884

a. Primeiras duas funções discriminantes canônicas foram utilizadas na análise.

Na tabela 5.17 o λ de Wilks testa o significado do autovalor para cada função discriminante. As duas funções discriminantes são significativas.

Tabela 5.17 - Lambda de Wilks

Teste das Funções	Wilks' Lambda	Qui-quadrado	df	Sig.
1 por 2	,034	113,530	60	,000
2	,219	50,924	29	,007

Os coeficientes padronizados das 2 funções discriminantes estão na tabela 5.18 e indicam a importância relativa das variáveis independentes em predizer a variável dependente, eles são úteis para propósitos de interpretação.

Tabela 5.18 - Coeficientes Padronizados das Funções Discriminantes Canônicas

	Função	
	1	2
MA.N	-7,981	1,978
MA.NCR	1,341	-3,225
MA.NPO	1,830	-3,949
MA.SF	-2,024	-1,404
ME.N	-3,616	2,196
ME.NNE	3,281	-1,360
ME.SF	-,259	,656
ME.SG	1,490	-1,749
TOC	,078	-,206
Cr	,914	,109
Ni	1,359	-,809
Mn	-,312	-,464
Pb	-,434	,774
Cd	-,858	-1,147
Zn	-,288	-1,012
Fe	-1,648	,965
As	1,277	,591
Ba	-,386	1,216
Co	,102	,050
V	-1,575	,791
Al	,348	-,177
Cu	,178	,815
VÁGEIS	2,644	2,752
VA_CAR	1,508	1,299
Sed_car	1,412	-,011
Sed_det	4,496	1,288
A1	,241	,565
B1	-,482	,452
A2	-,936	,848
TPH	-1,038	,069

A tabela 5.19 da matriz da estrutura mostra as correlações de cada variável independente com cada função discriminante. Neste caso, temos duas funções discriminantes porque nossa variável dependente tem três grupos e temos 32 variáveis independentes.

Tabela 5.19 - Matriz Estrutura

	Function	
	1	2
TPH	-,233 *	-,012
Ba	-,187 *	-,003
Pb	,178 *	,050
Cd	-,112 *	-,109
V	-,106 *	,079
Ni	,097 *	,006
Al	-,082 *	-,024
B2[a]	,077 *	,002
ME.NNE	,069 *	,022
ME.N	,062 *	,037
A1	-,055 *	-,020
MA.NPO	-,044 *	-,004
MA.SF	-,037 *	-,022
Sed_filt[a]	,019 *	,015
Mn	-,030	,238 *
Fe	-,095	,228 *
As	,062	,173 *
Cr	,031	,157 *
Sed_car	,085	,146 *
MA.NCR	,011	-,122 *
TOC	,016	,120 *
ME.SF	,016	,075 *
VA_CAR	-,014	,074 *
VÁGEIS	-,011	-,069 *
Cu	,035	,063 *
Co	,058	-,061 *
Sed_det	-,002	,061 *
A2	-,004	,060 *
ME.SG	,006	,057 *
Zn	,004	-,038 *
B1	,011	-,021 *
MA.N	-,007	,017 *

Correlação dentro dos grupos entre as variáveis discriminantes e as funções discriminantes canônicas padronizadas
Variáveis ordenadas em ordem decrescente de correlação com as funções.

*. Grande correlação absoluta entre variável e função discriminante

a. Esta variável não foi usada na análise.

As variáveis independentes estão em ordem decrescente, em valor absoluto, da correlação com a função discriminante. No nosso caso as variáveis independentes que tem maior correlação com a função discriminante 1 são: TPH, Ba, Pb, Cd, V, Ni, Al, ME.NNE, ME.N, A1, MA.NPO, MA.SF. As variáveis independentes Mn, Fe, As, Cr, Sed_car, MAN.CR, TOC, ME.SF, VA_CAR, VÁGEIS, Cu, Co, Sed_det, A2, ME.SG, Zn, B1 e MA.N tem uma maior correlação com a função discriminante 2.

Tabela 5.20 - Coeficientes das Funções Discriminantes Canônicas

	Função	
	1	2
MA.N	-,683	,169
MA.NCR	,273	-,656
MA.NPO	,257	-,554
MA.SF	-,460	-,319
ME.N	-,154	,094
ME.NNE	,170	-,071
ME.SF	-,067	,170
ME.SG	,244	-,286
TOC	,360	-,948
Cr	,120	,014
Ni	,317	-,189
Mn	-,001	-,002
Pb	-,121	,215
Cd	-12,040	-16,089
Zn	-,023	-,083
Fe	-4,059	2,376
As	,200	,093
Ba	-,001	,002
Co	,104	,050
V	-,220	,110
Al	,411	-,209
Cu	,054	,247
VA_CAR	,623	,537
VÁGEIS	,434	,451
Sed_car	1,596	-,013
Sed_det	,827	,237
A1	,019	,044
B1	-,033	,031
A2	-,101	,091
TPH	-,343	,023
(Constant)	17,933	-11,959

Coeficientes não padronizados

Os coeficientes não padronizados das 2 funções discriminantes encontram-se na tabela 5.20, que encontra-se na página anterior, e são mais fáceis de usar para calcular os valores Z discriminantes.

A tabela 5.21 é usada para estabelecer os pontos do corte que são utilizados para classificar os casos. Os pontos do corte ajustam escalas dos valores discriminantes para classificar os casos como controle interno, referência e máscara. Naturalmente, o computador faz a classificação automaticamente, assim estes valores são para finalidades informativas.

Tabela 5.21 - Funções dos Centróides dos Grupos

	Função	
Mask_dis	1	2
C. Interno	2,432	-1,871
Referência	3,046	4,378
Máscara	-1,972	,126

Funções discriminantes canônicas não padronizadas avaliadas nas médias dos grupos

A tabela 5.22 apresenta os coeficientes das funções de classificação. Como nossa variável dependente apresenta 3 grupos temos então 3 funções de classificação, uma para cada grupo. Para o grupo 1 (Controle Interno) temos a seguinte função de classificação:

C_1 = - 601,226 + 17,868**MA.N** – 14,217**MA.NCR** – 16,567**MA.NPO** – 2,590**MA.SF** + 2,268**ME.N** – 0,388**ME.NNE** – 0,159**ME.SF** – 11,613**ME.SG** + 41,578**TOC** – 4,648**Cr** – 3,239**Ni** – $1{,}36\times10^{-3}$**Mn** + 8,460**Pb** – 465,555**Cd** – 2,065**Zn** + 36,965**Fe** – 1,242**As** + $8{,}028\times10^{-2}$**Ba** + 31,135**Co** + 11,227**V** – 10,099**Al** + 4,144**Cu** + 0,920**VÁGEIS** – 4,944**VA_CAR** – 45,758**Sed_car** – 13,367**Sed_det** + 4,197**A1** + 4,766**B1** + 9,042**A2** + 5,539**TPH**

Para o grupo 2 (Referência) e para o grupo 3 (Máscara) utilizamos, respectivamente, a segunda e terceira colunas da tabela 5.22 para a função de classificação, sendo ela calculada da mesma forma que a do grupo 1 (Controle interno).

Tabela 5.22 - Coeficientes da Função de Classificação

	Mask_dis		
	CI	Referência	Máscara
MA.N	17,868	18,507	21,215
MA.NCR	-14,217	-18,147	-16,727
MA.NPO	-16,567	-19,871	-18,804
MA.SF	-2,590	-4,868	-1,200
ME.N	2,268	2,759	3,135
ME.NNE	-,388	-,724	-1,279
ME.SF	-,159	,864	,477
ME.SG	-11,613	-13,250	-13,257
TOC	41,578	35,873	38,100
Cr	-4,648	-4,485	-5,149
Ni	-3,239	-4,223	-5,012
Mn	-1,36E-03	-1,408E-02	4,801E-04
Pb	8,460	9,729	9,420
Cd	-465,555	-573,500	-444,667
Zn	-2,065	-2,595	-2,127
Fe	36,965	49,319	59,586
As	-1,242	-,540	-1,939
Ba	,028E-02	9,037E-02	8,594E-02
Co	31,135	31,514	30,780
V	11,227	11,781	12,415
Al	-10,099	-11,150	-12,324
Cu	4,144	5,722	4,399
VÁGEIS	,920	4,006	-8,82E-02
VA_CAR	-4,944	-1,207	-6,616
Sed_car	-45,758	-44,858	-52,811
Sed_det	-13,367	-11,379	-16,535
A1	4,197	4,480	4,202
B1	4,766	4,940	4,974
A2	9,042	9,550	9,667
TPH	5,539	5,469	7,096
(Constant)	-601,226	-674,462	-701,332

Função Discriminante Linear de Fisher

Com o objetivo de classificar um novo caso vamos usar estas três funções. Os valores para as variáveis deste novo caso encontram-se na tabela 5.23.

Tabela 5.23 - Valores das variáveis

Variáveis	Valores
MA.N	46,000
MA.NCR	11,000
MA.NPO	24,000
MA.SF	19,000
ME.N	24,000
ME.NNE	19,000
ME.SF	7,000
ME.SG	8,000
TOC	1,300
Cr	70,000
Ni	28,200
Mn	997,000
Pb	19,100
Cd	0,110
Zn	60,500
Fe	3,830
As	21,500
Ba	342,000
Co	8,010
V	82,800
Al	1,180
Cu	17,100
VA_CAR	2,000
VÁGEIS	19,000
Sed_car	0,000
Sed_det	23,000
A1	13,330
B1	66,670
A2	6,670
TPH	1,250

Para as funções de classificação dos grupos 1, 2 e 3 procedemos os cálculos e obtemos os valores que encontram-se na tabela 5.24.

Tabela 5.24 - Funções de Classificação

Grupo	Valor da Função
1	713,319
2	730,646
3	732,125

Este caso obteve a maior classificação no **grupo 3**, logo podemos concluir que ele está classificado no grupo máscara.

A tabela B.3, que se encontra no anexo B, lista o grupo atual, o grupo predito baseado nas maiores probabilidades *a posteriori*, a probabilidade *a priori*, a probabilidade *a posteriori*, a distância ao quadrado de Mahalanobis do caso ao centróide do grupo (valores grandes indicam outliers), e o valor discriminante para o caso. O caso é classificado baseado no valor discriminante em relação ao prazo (não mostrado). Casos mal classificados são marcados com asteriscos. As colunas que correspondem ao segundo mais alto grupo mostram as probabilidades *a posteriori* e a distância de Mahalanobis para o caso que foi classificado baseado na segunda mais alta probabilidade *a posteriori*. O caso 11 foi mal classificado, ele está no grupo 3 (máscara) mas deveria ser classificado no grupo 1 (controle interno).

Na figura 5.3 temos um gráfico com os três grupos (controle interno, referência e máscara). O gráfico utiliza os valores das duas primeiras funções discriminantes, para cada estação amostral, e os valores dos centróides dos grupos.

Através do gráfico podemos observar que o ponto de cor azul (grupo 3) que está próximo do grupo 1 (cor vermelha) é o caso mal classificado.

Os grupos estão visualmente bem separados dando evidências de uma boa discriminação.

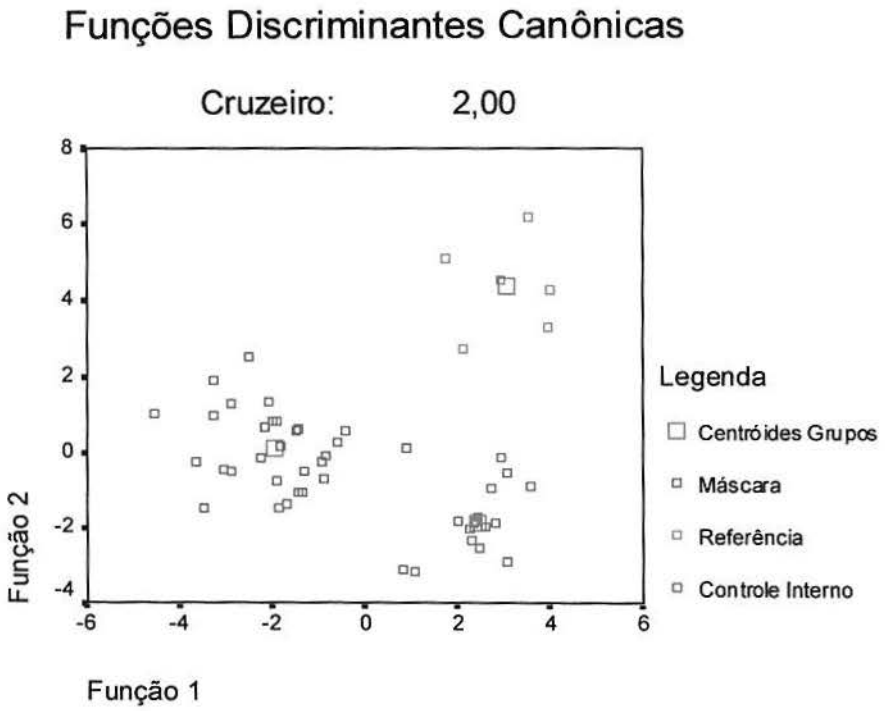


Figura 5.3 – Gráfico das Funções Discriminantes para os 3 grupos

O mapa territorial apresentado na figura B.1, que encontra-se no anexo B, é um gráfico de limites usado para classificar os casos dentro dos grupos baseado nos valores da função. Os números correspondem aos grupos nos quais são classificados os casos. A média, para cada grupo, é indicada por um asterisco dentro de seus limites. O mapa não é exibido se houver só uma função discriminante. Ele pode ser utilizado para identificar casos mal classificados e, também, *outliers*. Podemos verificar que o nosso caso mal classificado (caso 11) está no limite entre os grupos 3 e 1 pelo mapa territorial.

A tabela 5.25 apresenta os resultados da classificação. Aqui classificou corretamente 98,0% dos casos, um nível excelente de discriminação. Os grupos 1 e 2 classificaram corretamente 100,0% dos casos e o grupo 3 classificou corretamente 96,6% dos casos.

Tabela 5.25 - Resultados da Classificação [a]

	Mask_dis	Predição dos Membros dos			Total
		CI	Referência	Máscara	
Count	CI	16	0	0	16
	Referência	0	6	0	6
	Máscara	1	0	28	29
%	CI	100,0	,0	,0	100,0
	Referência	,0	100,0	,0	100,0
	Máscara	3,4	,0	96,6	100,0

a. 98,0% dos casos do grupo original foram corretamente classificados.

Com o objetivo de testar as variáveis que mais discriminam efetuamos uma Análise Discriminante com as variáveis independentes que tem o valor de probabilidade $p < 0,10$ na tabela 5.14 e que são: TPH, Fe, Mn Ba, Pb, As, Sed_car e Cd. Tivemos 76,5% dos casos corretamente classificados o que nos dá evidências da importância em utilizar todas as variáveis, previamente selecionadas, para discriminar os grupos.

6. Comentários e Sugestões

A Análise Discriminante é uma técnica estatística de análise multivariada indicada quando interessa verificar se um conjunto de variáveis discrimina os grupos, ou seja, tem comportamento diferenciado entre dois ou mais grupos. É uma ferramenta amplamente utilizada em várias áreas da ciência.

Os objetivos propostos foram alcançados, na medida em que a técnica foi caracterizada teoricamente e ilustrada sob o ponto de vista da aplicação prática na qual utilizamos dados do Projeto MAPEM. A Análise Discriminante foi apresentada formalmente e através de aplicações práticas nos capítulos 3 e 5, respectivamente.

Na revisão bibliográfica, no texto e na bibliografia foram indicadas obras, artigos e sites da internet, disponíveis ao leitor interessado em tópicos específicos.

O SPSS versão 8.0, foi o software utilizado nas aplicações práticas e uma apresentação dos principais comandos da Análise Discriminante foi apresentada no capítulo 4.

A Análise Discriminante aplicada aos dados do projeto MAPEM, utilizando o Cruzeiro 2, baseada em variáveis químicas (metais e hidrocarbonetos) e biológicas (meio fauna, macro fauna e grupo trófico) identificou poderosas diferenças entre os dois grupos que fazem parte da variável dependente Grupo_r. Escolhemos variáveis de acordo com a orientação de pesquisadores das respectivas áreas e optamos pelo método direto onde 30 das 32 variáveis escolhidas anteriormente entraram na análise. As cinco variáveis identificadas como as que mais discriminam os grupos (listadas em ordem de importância) com um valor de probabilidade $p < 0,05$ são: As, Mn, Sed_car, Cd e Cr. Também observamos que são importantes, com um valor de probabilidade $p < 0,10$ as variáveis Pb, TPH e Fe. A análise dos casos corretamente classificados revelou que nenhum caso foi mal classificado o que nos

dá grandes evidências de acerto ao classificarmos novos casos. Ao utilizarmos somente estas oito variáveis, em uma Análise Discriminante, com a mesma variável dependente obtivemos 98% dos casos corretamente classificados.

A Análise Discriminante executada, no Cruzeiro 2, com o objetivo de discriminar os 3 grupos (grupo 1: controle interno, grupo 2: referência, grupo 3: máscara) da variável mask_dis utilizou as mesmas variáveis independentes da Análise Discriminante com dois grupos. As cinco variáveis identificadas como as que mais discriminam os grupos (listadas em ordem de importância) com um valor de probabilidade $p < 0,05$ são: TPH, Fe, Mn, Ba e Pb. Também observamos que são importantes, com um valor de probabilidade $p < 0,10$ as variáveis As, Sed_car e Cd. No nosso caso as variáveis independentes que tem maior correlação com a função discriminante 1 são: TPH, Ba, Pb, Cd, V, Ni, Al, ME.NNE, ME.N, A1, MA.NPO, MA.SF e as outras variáveis independentes tem uma maior correlação com a função discriminante 2. A análise dos casos corretamente classificados revelou que 98% dos casos estavam bem classificados o que nos dá grandes evidências de acerto ao classificarmos novos casos. Ao utilizarmos somente as oito variáveis independentes com valor de probabilidade $p < 0,10$, em uma Análise Discriminante, com a mesma variável dependente obtivemos 76,5% dos casos corretamente classificados.

A primeira dimensão é tipificada por percepções mais altas do controle interno e da referência. A segunda dimensão é caracterizada melhor pela referência e pela máscara.

Como se pode perceber, o pesquisador deve ter um bom conhecimento da Análise Discriminante e dos métodos utilizados na determinação das variáveis independentes, incluídas na análise, para então poder decidir corretamente o que é mais apropriado aos objetivos da pesquisa.

Em suma, apesar do aspecto generalizado deste trabalho em apresentar conceitos teóricos e aplicações, espera-se que ele possa ser útil e amplamente utilizado em outras pesquisas. Acredita-se também que ele possa despertar o interesse de pesquisadores por este rico e fascinante assunto que é a Análise Discriminante.

7. Referências Bibliográficas

- **CHATFIELD**, C. & **COLLINS**, A. J. (1980) - *Introduction to Multivariate Analysis*. London New York: Chapman and Hall

- **COOLEY**, W. & **LOHNES**, P. R. (1971) – *Multivariate Data Analysis*. New York London Sydney Toronto: John Wiley & Sons, inc

- **HAIR**, J. F. Jr., **ANDERSON**, R. E., **TATHAM**, R. L. & **BLACK**, W. C. (1998) - *Multivariate Data Analysis*. New Jersey: Prentice Hall

- **JOHNSON**, R. A. & **WICHERN**, D. W. (1998) – *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall

- **KLECKA**, W. R. (1975) – *Discriminant Analysis* in Nie, NH et al (Eds): SPSS – Statistical Package for the Social Science. 2nd Ed. New York: Mc Graw Hill

- **MURTAGH**, F. & **HECK**, A. (1987) – *Multivariate Data Analysis*. Dordrecht, Holland: D. Reidel Publishing Company

- **NORUSIS**, M. J. (1985) – *SPSS-X Advanced Statistics Guide*. New York: Mc Graw Hill

- **PIZZOL**, S. J. S. (2002) – *Comportamento dos cafeicultores perante o risco: uma análise de três sistemas de produção da região de Marília, SP*. Dissertação apresentada para

obtenção do grau de Mestre em Economia Aplicada, Escola Superior de Agricultura "Luiz de Queiroz": São Paulo

- **PORTO**, G. S. (2000) – *A decisão empresarial de desenvolvimento tecnológico por meio da cooperação empresa-universidade*. Tese apresentada para obtenção de grau de Doutor em Administração, Faculdade de Economia, Administração e Contabilidade: São Paulo

- **POSSOLI**, S. (1992) - *Análise Multivariada*, Cadernos de Matemática e Estatística série B, n.10, UFRGS: Instituto de Matemática, Porto Alegre

- **POTTER**, P.E.; **SHIMP**, N.F. & **WITTERS**, J. (1963) - *Trace elements in marine and fresh-water argillaceous sediments*. Geochimica et Cosmochimica Acta, 27

- **RENCHER**, A.C. (1995) – *Methods of Multivariate Analysis*. New York: John Wiley & Sons

- **SHARMA**, S. (1996) - *Applied Multivariate Techniques*. New York: John Wiley & Sons

- **SICSÚ**, A. L. (1975) – *Análise Discriminante*. Dissertação apresentada para obtenção do grau de Mestre em Estatistica Aplicada, USP: Instituto de Matemática e Estatística, São Paulo

Disponível na *Internet*:

- http://www2.chass.ncsu.edu/garson/pa765/discrim.htm

- http://www.ex.ac.uk/~SEGLea/multvar2/disclogi.html.

- http://www.statsoftinc.com/textbook/stdiscan.html

- http://www.psychstat.smsu.edu/multibook/mlt03.htm.

- http://www.spssscience.co.kr/WhitePapers/classifying.newcases.htm.

- http://www.fep.up.pt/disciplinas/ce707/plano.htm.

8. Anexos

Anexo A: Tabelas relativas a Análise Discriminante com dois grupos

Tabela A.1 - Estatísticas Descritivas dos Grupos 1 e 2

Grupo_r		Média	Desvio Padrão			Média	Desvio Padrão
< 500 m	MA.N	25,289	10,759	**2.500 m**	MA.N	26,000	17,111
	MA.NCR	9,133	4,635		MA.NCR	6,333	6,890
	MA.NPO	12,733	7,316		MA.NPO	11,500	4,506
	MA.SF	13,711	4,198		MA.SF	12,667	5,574
	ME.N	37,333	23,424		ME.N	46,667	22,033
	ME.NNE	31,333	19,419		ME.NNE	38,000	17,286
	ME.SF	9,200	3,739		ME.SF	10,833	4,446
	ME.SG	12,156	5,924		ME.SG	14,000	7,127
	TOC	1,192	0,222		TOC	1,333	0,151
	Cr	63,989	7,786		Cr	70,717	5,118
	Ni	31,031	4,382		Ni	32,583	3,723
	Mn	720,422	238,045		Mn	982,333	337,678
	Pb	21,544	3,912		Pb	24,650	1,515
	Cd	0,283	0,074		Cd	0,217	0,037
	Zn	71,913	12,672		Zn	69,800	5,769
	Fe	3,337	0,402		Fe	3,662	0,649
	As	24,373	5,980		As	31,200	8,719
	Ba	750,933	814,986		Ba	270,333	29,964
	Co	7,871	1,019		Co	7,770	0,663
	V	78,976	7,701		V	79,167	3,536
	Al	3,078	0,862		Al	2,735	0,693
	Cu	19,807	3,378		Cu	21,233	1,969
	VA_CAR	2,889	2,177		VA_CAR	3,667	3,882
	VÁGEIS	13,333	5,977		VÁGEIS	11,000	6,573
	Sed_car	0,267	0,618		Sed_car	1,167	2,041
	Sed_det	8,400	5,114		Sed_det	10,000	7,403
	Sed_filt	0,156	0,424		Sed_filt	0,167	0,408
	A1	31,546	12,876		A1	27,777	13,157
	B1	46,394	13,842		B1	45,382	18,720
	A2	12,865	9,521		A2	15,497	6,083
	B2	9,196	7,842		B2	11,344	6,716
	TPH	3,550	3,481		TPH	0,928	0,368

Tabela A.2 - Estatísticas Casewise

		Grupo mais alto					Segundo Grupo mais alto			Valores Discriminantes
Estação amostral	Grupo Atual	Grupo Predito	P(D>d | G=g) p	df	P(G=g | D=d)	Distância de Mahalanobis ao quadrado do centróide	Grupo	P(G=g | D=d)	Distância de Mahalnobis ao quadrado do centróide	Função 1
1	1	1	,279	1	1,000	1,173	2	,000	50,503	-1,792
2	1	1	,323	1	1,000	,979	2	,000	25,344	,281
3	1	1	,751	1	1,000	,101	2	,000	40,205	-1,026
4	1	1	,028	1	1,000	4,819	2	,000	67,549	-2,904
5	1	1	,265	1	1,000	1,244	2	,000	50,966	-1,824
6	1	1	,727	1	1,000	,122	2	,000	32,203	-,360
7	1	1	,145	1	1,000	2,126	2	,000	55,975	-2,167
8	1	1	,805	1	1,000	,061	2	,000	33,374	-,462
9	1	1	,566	1	1,000	,329	2	,000	43,525	-1,282
10	1	1	,526	1	1,000	,402	2	,000	29,049	-,075
11	1	1	,231	1	1,000	1,437	2	,000	23,280	,490
12	1	1	,563	1	1,000	,335	2	,000	29,649	-,130
13	1	1	,734	1	1,000	,116	2	,000	40,494	-1,049
14	1	1	,060	1	,999	3,550	2	,001	17,136	1,175
15	1	1	,551	1	1,000	,355	2	,000	29,458	-,113
16	1	1	,967	1	1,000	,002	2	,000	35,792	-,668
17	1	1	,302	1	1,000	1,066	2	,000	49,787	-1,741
18	1	1	,291	1	1,000	1,116	2	,000	24,674	,348
19	1	1	,080	1	1,000	3,066	2	,000	60,442	-2,460
20	1	1	,365	1	1,000	,819	2	,000	48,006	-1,614
21	1	1	,687	1	1,000	,163	2	,000	41,303	-1,112
22	1	1	,545	1	1,000	,366	2	,000	29,357	-,103
23	1	1	,542	1	1,000	,371	2	,000	29,312	-,099
24	1	1	,196	1	1,000	1,675	2	,000	53,553	-2,003
25	1	1	,730	1	1,000	,119	2	,000	40,554	-1,053
26	1	1	,320	1	1,000	,988	2	,000	25,296	,285
27	1	1	,091	1	1,000	2,852	2	,000	59,481	-2,397
28	1	1	,975	1	1,000	,001	2	,000	35,902	-,677
29	1	1	,843	1	1,000	,039	2	,000	33,943	-,511
32	1	1	,617	1	1,000	,250	2	,000	42,552	-1,208
33	1	1	,385	1	1,000	,754	2	,000	47,495	-1,577
34	1	1	,952	1	1,000	,004	2	,000	37,005	-,768
35	1	1	,931	1	1,000	,007	2	,000	35,250	-,622
36	1	1	,893	1	1,000	,018	2	,000	34,683	-,574
37	1	1	,492	1	1,000	,472	2	,000	28,478	-,022
38	1	1	,399	1	1,000	,710	2	,000	47,147	-1,551
39	1	1	,113	1	1,000	2,517	2	,000	19,688	,878
40	1	1	,304	1	1,000	1,057	2	,000	24,956	,319
41	1	1	,708	1	1,000	,140	2	,000	40,937	-1,083
42	1	1	,824	1	1,000	,049	2	,000	33,660	-,487
43	1	1	,138	1	1,000	2,196	2	,000	20,628	,773
44	1	1	,058	1	,999	3,584	2	,001	17,060	1,185
46	1	1	,952	1	1,000	,004	2	,000	35,560	-,648
47	1	1	,939	1	1,000	,006	2	,000	35,364	-,632
48	1	1	,904	1	1,000	,015	2	,000	37,752	-,829
49	2	2	,061	1	,999	3,521	1	,001	17,198	3,438
50	2	2	,909	1	1,000	,013	1	,000	37,668	5,429
51	2	2	,722	1	1,000	,126	1	,000	40,688	5,670
52	2	2	,924	1	1,000	,009	1	,000	37,436	5,410
53	2	2	,063	1	1,000	3,460	1	,000	62,151	7,175
54	2	2	,584	1	1,000	,300	1	,000	29,988	4,768

Anexo B: Tabelas relativas a Análise Discriminante com três grupos

Tabela B.1 - Estatísticas Descritivas dos Grupos 1 e 2

Mask_dis		Média	Desvio Padrão			Média	Desvio Padrão
CI	MA.N	24,813	9,020	**Referência**	MA.N	26,000	17,111
	MA.NCR	10,063	5,434		MA.NCR	6,333	6,890
	MA.NPO	11,875	7,384		MA.NPO	11,500	4,506
	MA.SF	13,375	3,074		MA.SF	12,667	5,574
	ME.N	40,375	28,763		ME.N	46,667	22,033
	ME.NNE	34,563	24,465		ME.NNE	38,000	17,286
	ME.SF	9,000	4,131		ME.SF	10,833	4,446
	ME.SG	11,813	6,285		ME.SG	14,000	7,127
	TOC	1,168	0,246		TOC	1,333	0,151
	Cr	63,125	8,238		Cr	70,717	5,118
	Ni	32,175	5,302		Ni	32,583	3,723
	Mn	625,125	142,146		Mn	982,333	337,678
	Pb	23,131	5,062		Pb	24,650	1,515
	Cd	0,270	0,059		Cd	0,217	0,037
	Zn	72,656	11,123		Zn	69,800	5,769
	Fe	3,108	0,309		Fe	3,662	0,649
	As	24,075	5,060		As	31,200	8,719
	Ba	365,875	66,036		Ba	270,333	29,964
	Co	8,110	0,786		Co	7,770	0,663
	V	76,100	6,620		V	79,167	3,536
	Al	2,907	0,909		Al	2,735	0,693
	Cu	19,863	3,086		Cu	21,233	1,969
	VA_CAR	2,563	1,672		VA_CAR	3,667	3,882
	VÁGEIS	13,688	5,630		VÁGEIS	11,000	6,573
	Sed_car	0,313	0,602		Sed_car	1,167	2,041
	Sed_det	7,938	4,281		Sed_det	10,000	7,403
	Sed_filt	0,063	0,250		Sed_filt	0,167	0,408
	A1	29,861	12,029		A1	27,777	13,157
	B1	47,237	12,847		B1	45,382	18,720
	A2	12,034	11,488		A2	15,497	6,083
	B2	10,868	8,824		B2	11,344	6,716
	TPH	1,594	0,672		TPH	0,928	0,368

Tabela B.2 - Estatísticas Descritivas do Grupo 3

Mask_dis		Média	Desvio Padrão
	MA.N	25,552	11,752
	MA.NCR	8,621	4,144
	MA.NPO	13,207	7,365
	MA.SF	13,897	4,746
	ME.N	35,655	20,268
	ME.NNE	29,552	16,208
	ME.SF	9,310	3,577
	ME.SG	12,345	5,820
	TOC	1,205	0,210
	Cr	64,466	7,633
	Ni	30,400	3,736
	Mn	773,000	264,894
	Pb	20,669	2,844
	Cd	0,290	0,081
	Zn	71,503	13,623
	Fe	3,463	0,395
	As	24,538	6,512
	Ba	963,379	953,908
	Co	7,739	1,119
	V	80,562	7,899
	Al	3,172	0,837
	Cu	19,776	3,581
	VÁGEIS	3,069	2,419
	VA_CAR	13,138	6,249
	Sed_car	0,241	0,636
	Sed_det	8,655	5,576
	Sed_filt	0,207	0,491
	A1	32,475	13,434
	B1	45,928	14,562
	A2	13,324	8,434
	B2	8,273	7,242
	TPH	4,629	3,926

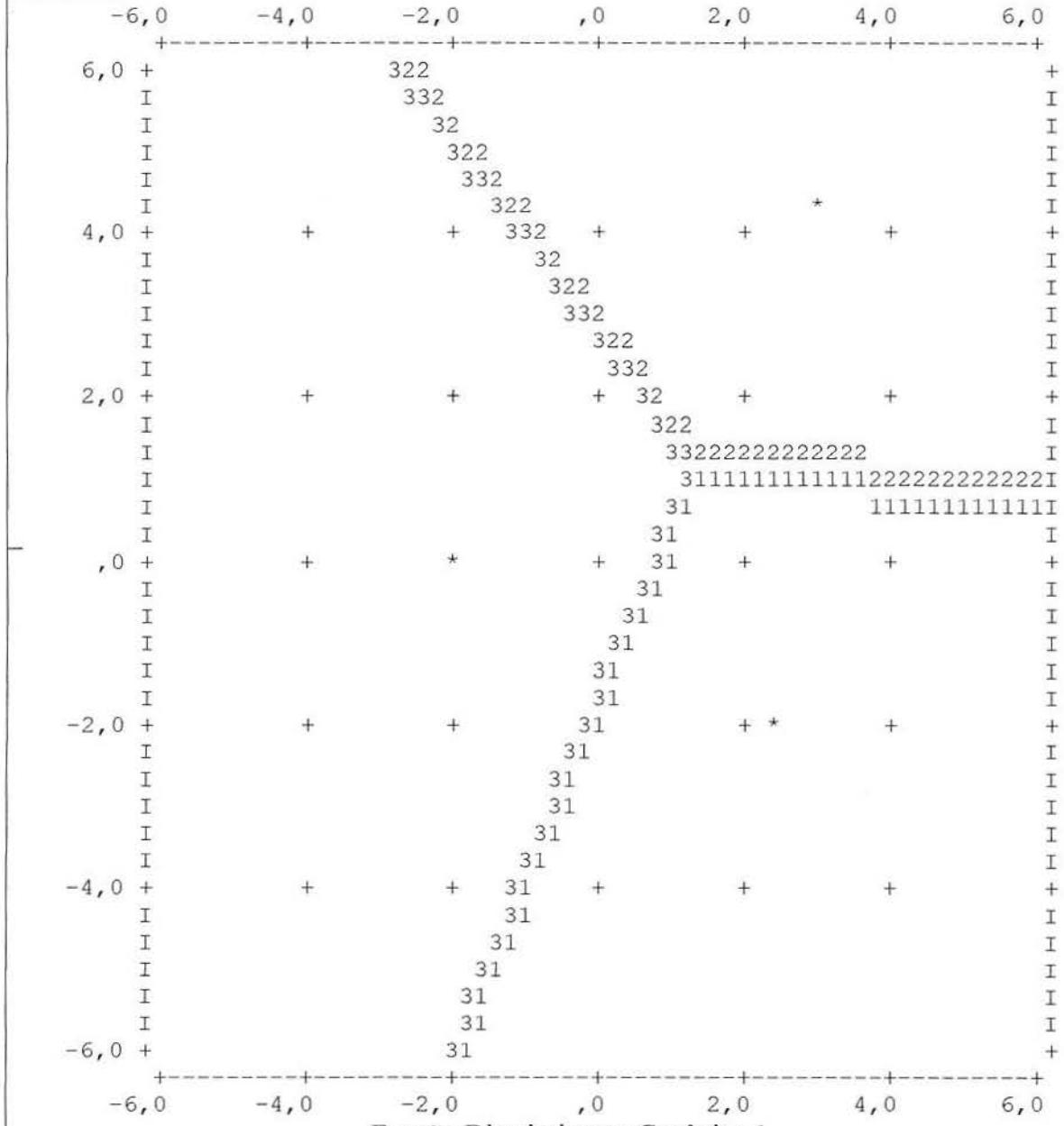
Tabela B.3 - Estatísticas Casewise

		Grupo mais alto					Segundo Grupo mais alto			Valores Discriminantes	
Estação amostral	Grupo Atual	Grupo Predito	P(D>d | G=g) p	df	P(G=g | D=d)	Distância de Mahalanobis ao quadrado do centróide	Grupo	P(G=g | D=d)	Distância de Mahalanobis ao quadrado do centróide	Função 1	Função 2
1	3	3	,453	2	1,000	1,586	1	,000	32,386	-3,083	-,467
2	3	3	,481	2	1,000	1,462	1	,000	30,604	-2,079	1,331
3	3	3	,667	2	,999	,810	1	,001	15,948	-1,315	-,488
4	3	3	,085	2	1,000	4,931	1	,000	35,162	-3,486	-1,498
5	3	3	,234	2	1,000	2,904	1	,000	39,542	-3,638	-,230
6	3	3	,838	2	1,000	,352	1	,000	27,869	-2,191	,678
7	3	3	,278	2	1,000	2,560	1	,000	18,764	-1,881	-1,471
8	3	3	,522	2	,998	1,300	1	,002	13,956	-,853	-,092
9	1	1	,468	2	1,000	1,517	3	,000	34,731	3,070	-2,924
10	3	3	,766	2	1,000	,533	1	,000	21,189	-1,437	,623
11	3	1	**,045	2	,734	6,223	3	,266	8,257	,901	,099
12	3	3	,327	2	1,000	2,236	1	,000	38,560	-2,915	1,287
13	3	3	,398	2	,995	1,842	1	,005	12,482	-,906	-,714
14	3	3	,048	2	1,000	6,058	2	,000	34,246	-2,506	2,529
15	3	3	,790	2	1,000	,470	1	,000	21,377	-1,476	,600
16	3	3	,988	2	1,000	,025	1	,000	22,161	-1,818	,154
17	3	3	,534	2	1,000	1,253	1	,000	30,295	-2,899	-,501
18	3	3	,273	2	,997	2,597	1	,003	14,274	-,430	,595
19	1	1	,125	2	,999	4,155	3	,001	18,362	,823	-3,122
20	3	3	,433	2	,999	1,676	1	,001	15,830	-1,464	-1,065
21	3	3	,934	2	1,000	,136	1	,000	24,809	-2,235	-,131
22	3	3	,761	2	1,000	,547	1	,000	27,097	-1,996	,866
23	3	3	,779	2	1,000	,499	1	,000	26,194	-1,915	,831
24	3	3	,316	2	,999	2,306	1	,001	17,467	-1,717	-1,371
25	3	3	,024	2	1,000	7,474	1	,000	57,221	-4,540	1,065
26	3	3	,089	2	1,000	4,848	2	,000	45,803	-3,257	1,914
27	1	1	,168	2	1,000	3,567	3	,000	20,075	1,062	-3,171
28	1	1	,996	2	1,000	,007	3	,000	22,832	2,354	-1,904
29	1	1	,993	2	1,000	,013	3	,000	22,901	2,427	-1,757
32	1	1	,790	2	1,000	,471	3	,000	27,053	2,485	-2,555
33	3	3	,395	2	,999	1,856	1	,001	14,881	-1,344	-1,083
34	1	1	,907	2	1,000	,195	3	,000	19,536	1,993	-1,827
35	1	1	,999	2	1,000	,002	3	,000	23,000	2,392	-1,862
36	3	3	,300	2	1,000	2,409	1	,000	40,635	-3,273	,972
37	3	3	,382	2	,997	1,926	1	,003	13,709	-,591	,267
38	3	3	,662	2	1,000	,825	1	,000	19,995	-1,905	-,780
39	1	1	,335	2	1,000	2,187	2	,000	24,101	3,058	-,531
40	1	1	,646	2	1,000	,875	3	,000	23,163	2,712	-,979
41	1	1	,892	2	1,000	,228	3	,000	24,462	2,318	-2,335
42	1	1	,935	2	1,000	,135	3	,000	26,888	2,796	-1,911
43	1	1	,321	2	1,000	2,271	2	,000	28,252	3,592	-,909
44	1	1	,195	2	1,000	3,269	2	,000	20,408	2,948	-,138
46	3	3	,543	2	,998	1,220	1	,002	13,949	-,936	-,257
47	1	1	,978	2	1,000	,045	3	,000	25,465	2,613	-1,981
48	1	1	,974	2	1,000	,052	3	,000	22,624	2,268	-2,030
49	2	2	,175	2	1,000	3,480	1	,000	21,554	2,116	2,761
50	2	2	,977	2	1,000	,046	1	,000	41,612	2,932	4,560
51	2	2	,627	2	1,000	,935	1	,000	40,543	4,010	4,298
52	2	2	,330	2	1,000	2,220	3	,000	38,793	1,753	5,118
53	2	2	,170	2	1,000	3,543	1	,000	66,311	3,526	6,198
54	2	2	,389	2	1,000	1,889	1	,000	29,372	3,940	3,334

**. Casos mal classificados

Mapa Territorial

Função Discriminante
Canônica 2



Simbolos usados no mapa territorial

Simbolo	Grupo	Label
1	1	Controle Interno
2	2	Referência
3	3	Máscara
*		Indica o centróide do grupo

Figura B.1 – Mapa territorial das Funções Discriminantes para os 3 grupos