Unidade V – Análise Discriminante

5.1 Introdução

A análise discriminante é uma técnica estatística multivariada que busca a separação (discriminação) de indivíduos (observações) e, ou, a alocação (classificação) de novos indivíduos em grupos previamente definidos, com base em variáveis mensuradas nos indivíduos que compõem cada um dos grupos.

Essa técnica é frequentemente utilizada para definição de regras para designar novos indivíduos aos grupos e para investigação de diferenças observadas, quando os relacionamentos causais não são bem entendidos.

A análise discriminante permite avaliar se os grupos diferem entre si, em termos do conjunto das variáveis mensuradas em seus indivíduos, e se o conhecimento prévio dessas variáveis permite designar um novo indivíduo a um dos grupos, com um risco mínimo de erro.

A análise discriminante pode ser empregada com as seguintes finalidades:

a) Construir regras para a alocação de indivíduos aos grupos, com base em funções lineares das variáveis.
b) Identificar as variáveis que contribuem para a discriminação dos grupos.
c) Determinar o número de funções discriminantes necessário para descrever o modelo de agrupamento.
d) Estimar as probabilidades de classificações corretas.
e) Alocar observações para grupos.

5.2 Modelo Teórico

No caso geral, haverá *m* amostras aleatórias de diferentes grupos, com tamanhos $n_1, n_2, \cdots, n_m$, e os valores estarão disponíveis para as p variáveis $X_1, X_2, \cdots, X_p$ para cada membro da amostra. Os dados para uma análise discriminante podem ser organizados conforme a Tabela 5.1. A aplicação da análise discriminante requer a existência de uma variável de classificação pré-estabelecida. Nesse caso, considera-se um conjunto de *n* unidades de observação classificadas em *m* subconjuntos ou grupos, para os quais foram computados valores em *p* variáveis aleatórias.

De acordo com Manly e Alberto (2017), os dados não precisam ser padronizados para ter média zero e variância unitária antes do início da análise, como é usual com ACP e AF. Isso porque o resultado da uma análise discriminante não é muito afetada pelo dimensionamento de variáveis individuais.

Tabela 5.1. Organização dos dados para análise discriminante com *n* casos, *p* variáveis e *m* grupos.

Casos	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\cdots$	$\boldsymbol{X_p}$	**Grupo**
1	x_{111}	x_{112}	$\cdots$	x_{11p}	1
2	x_{211}	x_{212}	$\cdots$	x_{21p}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_1	$x_{n_1 11}$	$x_{n_1 12}$	$\cdots$	$x_{n_1 1p}$	1
1	x_{121}	x_{122}	$\cdots$	x_{12p}	2
2	x_{221}	x_{222}	$\cdots$	x_{22p}	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_2	$x_{n_2 21}$	$x_{n_2 22}$	$\cdots$	$x_{n_2 2p}$	2
1	x_{1m1}	x_{1m2}	$\cdots$	x_{1mp}	m
2	x_{2m1}	x_{2m2}	$\cdots$	x_{2mp}	m
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_m	$x_{n_m m1}$	$x_{n_m m2}$	$\cdots$	$x_{n_m mp}$	m

Com o objetivo de estabelecer critérios (regras) de acesso ao modelo de agrupamento, Fisher (1936) sugeriu a transformação das observações multivariadas para observações univariadas, de maneira que essas últimas fossem o mais separadas possível. Para tanto, esse autor propôs o uso de combinações lineares das variáveis originais para criar variáveis univariadas, de maneira que essas maximizassem a razão das somas de quadrados entre os grupos e a soma de quadrados dentro dos grupos. O modelo apresenta o pressuposto de que as variáveis possuam distribuição normal multivariada e que os grupos apresentem matrizes de variância iguais.

O teorema central do limite assegura a robustez da técnica para quase todos os tipos de distribuição, cuja variância seja independente da média. No caso de não normalidade, pode-se recorrer a transformação de dados, convencionalmente usada e recomenda em estatística univariada.
O procedimento matemático inicia pela computação das médias e dos desvios-padrão para cada grupo e da média e desvio-padrão para o conjunto de todos os indivíduos, considerando as *p* variáveis. Esses parâmetros geram as matrizes de dispersão intergrupos (**B**) e intragrupos (**W**), conforme as equações matriciais:

$$\mathbf{B} = \sum_{i=1}^{g} (\bar{X}_i - \bar{X})\,(\bar{X}_i - \bar{X})',$$

onde

$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \qquad \text{e} \qquad \bar{X} = \frac{\sum_{i=1}^{g} n_i X_i}{\sum_{i=1}^{g} n_i} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{g} n_i}$$

$$\mathbf{W} = \sum_{i=1}^{g}\sum_{j=1}^{n_i} (\bar{x}_{ij} - \bar{X}_i)\,(\bar{x}_{ij} - \bar{X}_i)' = \sum_{i=1}^{g} (n_i - 1)\mathbf{S}_i = \left(\sum_{i=1}^{g} n_i - g\right)\mathbf{S}$$

Onde

S é a matriz de covariâncias amostral combinada: $\mathbf{S} = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2 + \cdots + (n_g-1)\mathbf{S}_g}{n_1+n_2+\cdots+n_g-g}$

em que

- g = número de grupos;
- n_i = número de indivíduos no i-ésimo grupo;
- $\mathbf{S}_i$ = matriz de variância-covariância amostral do i-ésimo grupo; e
- $\mathbf{S}$ = matriz de variância-covariância amostral combinada, ponderada pelo número de indivíduos de cada grupo.

Da análise de variância multivariada, tem-se que a matriz da soma de quadrados e produtos total (**T**) é igual a soma da matriz de soma de quadrados e produtos intergrupos (**B**) com a matriz de soma de quadrados intragrupos (**W**).
Considerando o espaço dimensional inicial, a matriz **B** expressa os desvios dos centroides dos grupos em relação ao grande centroide; a matriz **W** reflete os desvios dos indivíduos em relação aos centroides dos respectivos grupos; e **T** congrega os desvios dos indivíduos em relação ao grande centroide.
De posse das matrizes **B** e **W**, pode ser solucionada a equação $(\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I})\mathbf{V} = 0$, sujeita a restrição $\mathbf{V}'\mathbf{V} = 1$, onde λ são as raízes características ou autovalores da matriz $\mathbf{W}^{-1}\mathbf{B}$; **V** são os vetores característicos ou autovetores de $\mathbf{W}^{-1}\mathbf{B}$; e **I** é uma matriz identidade. Se **V** são os autovetores que maximizam a razão $\mathbf{V}'\mathbf{B}\mathbf{V}/\mathbf{V}'\mathbf{W}\mathbf{V}$, então as combinações lineares $y = \mathbf{V}'\mathbf{X} = (\bar{X}_i - \bar{X})\mathbf{S}^{-1}\mathbf{X}$ são as funções discriminantes de Fisher, ou variáveis canônicas, enquanto os autovalores associados valem $\lambda = \mathbf{V}'\mathbf{B}\mathbf{V}/\mathbf{V}'\mathbf{W}\mathbf{V}$. O número de autovalores reais será igual a g-1 ou p, o menor deles.
As funções discriminantes são derivadas em ordem de importância decrescente. A primeira representa a melhor combinação linear possível das variáveis iniciais, ou seja, ela extrai o máximo possível da variância intergrupos existente no espaço inicial; a segunda extrai o máximo possível da variância remanescente, com restrição de ser ortogonal à primeira; e assim, sucessivamente, são extraídos vetores mutuamente ortogonais, até esgotar-se a variância contida na matriz $\mathbf{W}^{-1}\mathbf{B}$.
As combinações lineares $\mathbf{V}_1'\mathbf{X}$, $\mathbf{V}_2'\mathbf{X}$, ..., $\mathbf{V}_k'\mathbf{X}$ são a primeira, a segunda e a *k*-ésima função discriminante, associadas a *k* autovetores. Essas funções discriminantes geram os escores discriminantes e os centroides dos grupos, que representados graficamente num espaço bidimensional, resultam em mapas territoriais dos grupos (SOUZA, 1989).

5.3 Critérios para a seleção de variáveis discriminantes

Em algumas situações, inicialmente, o número de variáveis pode ser muito grande. Nesse caso, é obviamente desejável selecionar um número relativamente menor de variáveis, que contenha tanta informação quanto a coleção original (JOHNSON e WICHERN, 1988).
Geralmente, podem ser usadas três modalidades para seleção de variáveis: **forward entry; stepwise selection; backward elimination**. O método **stepwise** é o mais usado, pois combina as feições do **forward selection** e do **backward elimination**. No método **stepwise**, a primeira variável incluída na análise possui o maior valor aceitável para o critério de seleção. Após a inclusão da primeira variável, o valor do critério é redefinido para todas as variáveis não incluídas no modelo, e a variável com o maior valor aceitável de critério é reavaliada para

determinar se ela satisfaz o critério de remoção. Assim, em cada passo é examinada a possibilidade de inclusão de novas variáveis no modelo, bem como da remoção daquelas já incluídas. A seleção de variáveis termina quando nenhuma das variáveis satisfaz os critérios de inclusão ou remoção (SPSS, 1990).

Entretanto, os resultados de qualquer método de seleção de variáveis devem ser interpretados com cautela, pois não há garantia de que o subconjunto de variáveis selecionado é o melhor, independente do critério de seleção utilizado. O problema de seleção de variáveis é ampliado quando existem grandes correlações entre as variáveis ou entre combinações lineares das variáveis.

Dentre os critérios para seleção de variáveis, pode-se destacar: **Lambda de Wilk**; **V de Rao**; **Mahalanobis**; **teste de F** e variância entre grupos não explicada.

5.3.1. Lambda de Wilk

A estatística Lambda de Wilk **(Λ)** expressa a relação entre a variância intragrupos e a variância total, e pode ser calculada de duas maneiras (FERREIRA e LIMA, 1978):

a) Em função dos autovalores da matriz $\mathbf{W}^{-1}\mathbf{B}$

$$\Lambda = \sum_{j=1}^{n} \frac{1}{1+\lambda_j}$$

b) Como uma razão entre discriminantes

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|}$$

A significância da estatística Lambda de Wilk para que uma variável seja incluída ou removida do modelo discriminante pode ser baseada no **teste de F**. O valor de **F**, em função de Lambda de Wilk, para o modelo com **p** variáveis já incluídas, é:

$$F_{(g-1,n-g-p)} = \left(\frac{n-g-p}{g-1}\right)\left(\frac{1-\Lambda_{p+1}/\Lambda_p}{\Lambda_{p+1}/\Lambda_p}\right)$$

onde

n = número total de observações;

g = número de grupos; e

Λ_p e Λ_{p+1} = lambdas antes e após a inclusão de nova variável ao modelo.

Quanto maior o poder discriminatório da variável, menor será o seu índice, sendo os valores oscilantes entre $\mathbf{0} < \Lambda \leq \mathbf{1}$. Um valor de **lambda** igual a **1** ocorre quando todas as médias dos grupos são iguais. Valores próximos de **zero** indicam que a variabilidade intragrupos é pequena comparada com a variabilidade total, ou seja, quando a maioria da variabilidade total é atribuída a diferenças entre as médias dos grupos. Assim, num processo de seleção de variáveis, a cada passo a variável que apresenta o menor valor de Lambda de Wilk seria a escolhida (SPSS, 1990).

5.3.2 V de Rao

Também conhecida como **Lawley-Hotelling**, é definida como:

$$V = (n-g)\sum_{i=1}^{p}\sum_{j=1}^{p} W_{ij*} \sum_{k=1}^{g} n_k(\bar{X}_{ik}-\bar{X}_i)(\bar{X}_{jk}-\bar{X}_j)$$

onde

p = número de variáveis no modelo;

g = número de grupos;

n_k = tamanho da amostra no k-ésimo grupo;

$\bar{X}_{ik}$ = média da i-ésima variável para o k-ésimio grupo;

$\bar{X}_i$ = média da i-ésima variável para todos os grupos combinados; e

$\bar{X}_j$ = média da j-ésima variável para todos os grupos combinados; e

W_{ij*} = elemento da matriz inversa de variância-covariância intragrupos.

Quanto maior a diferença entre as médias dos grupos, maior o valor de **V de Rao**. Portanto, uma maneira de avaliar a contribuição de uma dada variável é verificar o quanto ela incrementa **V de Rao**, quando incluída ao modelo. Um teste de significância para a alteração em **V de Rao** pode ser baseado na distribuição de qui-quadrado (χ^2), pois a distribuição de **V** segue a distribuição de χ^2, com $p(g$-1) graus de liberdade.

5.3.3 Distância de Mahalanobis (D^2)

A distância de Mahalanobis é uma medida generalizada na distância entre dois grupos. Assim, a distância entre dois grupos **a** e **b** é definida como:

$$D_{ab}^2 = (n-g)\sum_{i=1}^{p}\sum_{j=1}^{p} W_{ij*}(\bar{X}_{ia}-\bar{X}_{ib})(\bar{X}_{ja}-\bar{X}_{jb})$$

onde

p = número de variáveis no modelo;
$\bar{X}_{ia}$ = média para a i-ésima variável no grupo a;
W_{ij*} = elemento da matriz inversa de variância-covariância intragrupos.

Quando a distância de Mahalanobis é usada como critério para seleção de variáveis, ela é calculada primeiro, sendo a variável que apresentar o maior $\mathbf{D^2}$ para os dois grupos mais próximos (menor $\mathbf{D^2}$ inicialmente) a selecionada para inclusão no modelo.

5.3.4 F Entre Grupos

Na seleção de variáveis, a cada passo a variável escolhida para inclusão é aquela com maior valor de F. Nesse caso, o resultado pode diferir do critério anterior, pois aqui $\mathbf{D}_{ab}^2$ é ponderada pelos tamanhos das amostras dos grupos.

$$F = \frac{(n-1-p)n_1 n_2}{p(n-2)(n_1+n_2)} D_{ab}^2$$

em que

n = número total de observações;
p = número de variáveis no modelo; e
n_k = tamanho da amostra no k-ésimo grupo.

5.3.5 Soma da Variância não Explicada

A distância de Mahalanobis **(D^2)** e o quadrado do coeficiente de correlação **(R^2)** são proporcionais, quando se trata da análise discriminante entre apenas dois grupos, ou seja, $\mathbf{R^2 = cD^2}$. Para cada par de grupos ***a*** e ***b***, a variância não explicada pela regressão é $1-\mathbf{D}_{ab}^2$.
A soma da variância não explicada para todos os pares de grupos pode ser usada como critério para seleção de variáveis, isto é, a variável eleita para inclusão é aquela que minimiza a soma da variância não explicada pelo modelo.

5.4 Critérios para a seleção de funções discriminantes

As funções lineares discriminantes, conforme já mencionado, são obtidas a partir da extração dos **autovetores** da matriz $\mathbf{W^{-1}B}$, e constituem combinações das variáveis iniciais que maximizam a razão entre as dispersões intergrupos e intragrupos **(B/W)**. O número máximo de funções discriminantes que podem ser extraídas é igual a **(**g-1**)** ou p, o que for menor. Contudo, nem todas as funções discriminantes possíveis têm poder discriminatório significativo. Alguns critérios podem ser usados para avaliar a importância relativa de cada função discriminante na diferenciação dos grupos. Dentre eles, temos a porcentagem relativa dos **autovalores (λ)**; o coeficiente de correlação canônica (R); e o teste de qui-quadrado **(χ^2)** (FERREIRA e LIMA, 1978).

5.4.1 Porcentagem Relativa dos Autovalores

Associado a cada função temos um autovalor **(λ)**, que é diretamente proporcional ao montante da variância total intergrupos por ela explicada, constituindo, portanto, uma medida de poder discriminatório. É um critério empírico e baseia-se no autovalor da função discriminante **j**, expresso em porcentagem.

$$\text{Porcentagem Relativa dos Autovalores} = \frac{\lambda_j}{\sum_{i=j}^{m}\lambda_j} \times 100$$

Pode-se considerar também o percentual acumulado de variação explicado pelas funções, tomando-se a razão entre a soma dos autovalores de cada uma delas e a soma de todos os autovalores (SPSS, 1990). Uma vez que as primeiras funções normalmente concentram a maior proporção da variação total, em geral tomada como acima de 80%, obtém-se um modelo com espaço dimensional mais simplificado e reduzido, cujos eixos coordenados são os escores relativos às primeiras funções discriminantes, ou variáveis canônicas (CRUZ e REGAZZI, 1994).

5.4.2 Coeficiente de Correlação Canônica

O coeficiente de correlação canônica expressa o grau de associação existente entre uma dada função e a discriminação dos grupos (FERREIRA e LIMA, 1978). Quando elevado ao quadrado, expressa a proporção da variabilidade total explicada pelas diferenças entre os grupos. Portanto, pode constituir-se num critério para comparar o mérito das funções, ou seja, a porcentagem da variabilidade intergrupos atribuída a cada uma delas (SPSS, 1990).

Portanto, o coeficiente de correlação canônica é um indicador da habilidade de uma função para discriminar os grupos (SOUZA, 1989), e pode ser calculado pela expressão (COOLEY e LOHNES, 1971).

$$R = \left[\frac{\lambda_j}{1+\lambda_j}\right]^{1/2}$$

5.4.3 Teste de Qui-Quadrado

O teste de qui-quadrado permite medir o poder discriminatório das funções, indicando se a informação discriminante que ainda resta, após a extração das(s) primeira(s) função(ões), tem significância estatística (FERREIRA e LIMA, 1978).

A estatística qui-quadrado pode ser calculada a partir da **Lambda de Wilk** (COOLEY e LOHNES, 1971), ou seja:

$$\chi^2 = \left(N - \frac{p+g}{2} - 1\right)\mathbb{l}_n\Lambda', \text{com } (p-k)(g-k-1) \text{ graus de liberdade}$$

onde

- p = número de variáveis discriminantes;
- g = número de grupos;
- k = número de funções geradas;
- n = número máximo de funções discriminantes;
- N = número total de elementos; e
- Λ = estatística Lambda de Wilk.

5.5 As funções de classificação

Uma vez geradas as funções discriminantes, as funções de classificação podem ser obtidas de vários métodos, conforme as características de distribuição das populações (JOHNSON e WICHERN, 1988). Entretanto, as equações de classificação geralmente são calculadas à partir da matriz de variâncias-covariâncias amostral combinada **(S)** e dos centroides dos agrupamentos (SOUZA, 1989).

Tomando as funções discriminantes de Fisher como base para alocação, uma regra de classificação razoável é aquela que atribui um indivíduo **X** ao grupo *k*, se o quadrado da distância entre **X** e a média do grupo *k* for menor de que o quadrado da distância entre **X** e a média do grupo k_i. Assim, se somente **r ≤ S** dos discriminantes são usados para alocação, a regra é (JOHNSON e WICHERN, 1988):

Alocar **X** para ***k*** se

$$\sum_{j=1}^{r}\left(\hat{Y}_j - \bar{Y}_{kj}\right)^2 = \sum_{j=1}^{r}\left[\mathbf{V}_j'(\boldsymbol{X} - \bar{X}_k)\right]^2 \leq \sum_{i=1}^{r}[\mathbf{V}_i'(\boldsymbol{X} - \bar{X}_i)]^2 \qquad \text{para todo } i \neq k$$

sendo

$\mathbf{V}_i'$ = autovetor que maximiza a razão **V'BV/V'WV**.

Exemplo

O uso da técnica estatística multivariada de análise discriminante será demonstrado na área florestal.

O objetivo será identificar as características ambientais que permitem separar sítios florestais para a essência *Araucaria angustifolia* e gerar um modelo matemático para classificação de novas unidades amostrais.

Foram usadas informações de 21 parcelas amostrais de povoamentos plantados de *A. angustifolia*, coletadas por HOOGH e DIETRICH (1979), nos estados do Rio Grande do Sul, Santa Catarina, Paraná, São Paulo e Minas Gerais (Quadro 1).

Quadro 1 - Localização e tipo de solo das parcelas estudadas (extraído de HOOGH e DIETRICH, 1979)

Nº	Perfil	Localidade	Tipo de Solo
1	SC 28	Três Barras - SC	Latossolo Vermelho-escuro
2	SC 30	Três Barras - SC	Latossolo Vermelho-amarelo
3	PR 32	Teixeira Soares - PR	Latossolo Vermelho-amarelo
4	PR 37	Teixeira Soares - PR	Latossolo Vermelho-escuro

5	PR 44	Ponta Grossa - PR	Latossolo Vermelho-escuro
6	PR 68	Telêmaco Borba - PR	Latossolo Vermelho-escuro
7	PR 73	Telêmaco Borba - PR	Latossolo Vermelho-escuro
8	PR 83	Jussara - PR	Latossolo Vermelho-escuro
9	RS 85	S. Frco. Paula - RS	Cambissolo Húmico
10	RS 87	S. Frco. Paula - RS	Laterítico Bruno
11	RS 107	Passo Fundo - RS	Latossolo Vermelho-escuro
12	RS 113	Passo Fundo - RS	Laterítico Bruno
13	PR 123	Cascavel - PR	Latossolo Roxo
14	SC 142	Caçador - SC	Latossolo Bruno
15	SC 146	Chapecó - SC	Latossolo Roxo
16	SP 152	Capão Bonito - SP	Latossolo Vermelho-escuro
17	SP 153	Capão Bonito - SP	Latossolo Vermelho-escuro
18	MG 187	Passa Quatro - MG	Latossolo Vermelho-amarelo
19	SP 216	Caieiras - SP	Latossolo Vermelho-amarelo
20	PR 240	Renascença - PR	Latossolo Roxo
21	SC 246	Rio Negrinho - SC	Cambissolo Húmico

O índice de sítio foi usado como indicador do potencial produtivo de cada local. Três classes de qualidade de sítio foram definidas. A amplitude de classe foi obtida considerando a amplitude total dos índices de sítio (12,6 m), dividido por três, ou seja, 4,2 m (Quadro 2).

Optou-se por usar somente características ambientais de caráter mais estável, relativas a pedologia, clima e geografia (Quadro 2). As variáveis relacionadas com a fertilidade do solo não foram usadas, devido a acentuada diferença de idade dos povoamentos, por ocasião da amostragem.

Quadro 2 - Dados originais obtidos de HOOGH e DIETRICH (1979)

Nº	Latitude	Longitude	Altitude	Precipitação	Temperatura	SiO_2 A	SiO_2 B	Al_2O_3 A
1	26,20	50,32	790	1.341	16,3	13,9	14,5	20,1
2	26,20	50,32	783	1.341	16,3	24,8	28,6	17,5
3	25,45	50,58	799	1.442	17,2	12,1	13,9	19,8
4	25,45	50,58	849	1.442	17,2	10,2	10,8	21,6
5	25,22	50,03	853	1.402	17,6	5,4	8,4	10,7
6	24,30	50,62	864	1.422	18,2	4,7	8,2	6,6
7	24,30	50,62	841	1.422	18,2	9,2	9,2	18,6
8	23,62	52,47	398	1,413	21,3	5,7	6,4	6,1
9	29,35	50,32	921	2,250	14,5	7,5	11,8	11,4
10	29,35	50,32	885	2.250	14,5	11,8	11,8	16,3
11	28,27	52,20	738	1.658	17,6	9,1	9,8	17,0
12	28,27	52,20	682	1.658	17,6	10,8	10,8	16,3
13	25,00	53,35	770	1.500	18,2	8,6	9,2	24,2
14	26,75	51,22	1086	1.951	16,8	9,5	5,5	19,8
15	27,10	52,70	614	1.900	17,3	23,8	25,2	18,6
16	23,93	48,50	673	1.405	18,9	6,3	9,2	6,6
17	23,93	48,50	650	1.405	18,9	15,2	17,7	12,0
18	22,38	44,97	1035	1.437	17,5	14,7	16,9	13,7
19	23,37	46,77	761	1.460	18,7	13,2	13,9	14,7
20	26,10	53,02	650	1.809	17,7	15,6	15,6	19,1
21	26,42	49,48	913	1.271	16,4	11,50	12,9	11,8

Continua...

Quadro 2, Cont.

Nº	Al_2O_3 B	Fe_2O_3 A	Fe_2O_3 B	Ki A	Ki B	Kr A	Ki B
1	25,2	9,80	11,60	1,2	1,0	0,9	0,8
2	24,2	11,20	19,40	2,4	2,0	1,7	1,3
3	24,9	6,80	6,80	1,0	0,9	0,9	0,8
4	24,9	9,00	9,80	0,8	0,7	0,6	0,6
5	15,5	3,80	5,40	0,9	0,9	0,7	0,8
6	19,1	2,80	4,80	1,2	0,7	1,0	0,6

7	21,9	9,40	10,40	0,8	0,7	0,6	0,5
8	7,6	3,80	4,40	1,6	1,4	1,1	1,0
9	18,3	9,80	9,80	1,1	1,1	0,7	0,8
10	17,0	18,00	18,00	1,2	1,2	0,7	0,7
11	24,7	10,40	11,40	0,9	0,7	0,7	0,5
12	22,1	13,00	13,00	1,1	0,8	0,7	0,6
13	20,1	28,20	27,80	0,6	0,8	0,3	0,4
14	22,9	21,00	20,20	0,8	0,4	0,5	0,3
15	17,0	17,50	10,10	2,2	2,5	1,4	1,5
16	10,7	2,60	3,30	1,6	1,5	1,3	1,2
17	12,2	6,40	7,90	2,2	2,5	1,6	1,8
18	20,9	4,00	5,20	1,8	1,4	1,5	1,2
19	19,6	6,60	9,80	1,5	1,2	1,2	0,9
20	23,8	21,60	22,70	1,4	1,1	0,8	0,7
21	12,7	4,90	4,10	1,7	1,7	1,3	1,4

Continua...

Quadro 2, Cont.

Nº	Prof. A	Areia A	Areia B	Silte A	Silte B	Argila A	Argila B	Altura Dominante	Classe Sítio
1	48	3	3	15	15	82	82	15,7	2
2	50	16	8	16	22	68	70	10,2	1
3	52	25	14	21	20	56	66	13,6	2
4	76	28	28	10	10	62	62	19,5	3
5	40	52	52	20	20	29	29	8,8	1
6	46	60	55	20	18	20	27	16,6	2
7	60	37	32	13	11	50	57	18,9	3
8	52	77	74	6	1	17	25	19,8	3
9	74	37	16	19	26	44	58	16,0	2
10	17	25	25	21	21	54	54	17,9	3
11	63	20	10	16	8	65	82	16,6	2
12	37	18	12	17	10	65	78	14,5	2
13	36	16	26	10	12	74	62	18,7	3
14	37	6	8	16	16	78	76	15,8	2
15	57	10	6	14	14	76	80	19,1	3
16	34	70	60	8	14	22	26	10,7	1
17	20	32	30	16	12	52	58	9,3	1
18	41	52	36	4	8	44	56	13,5	2
19	52	48	38	10	6	42	56	21,9	3
20	19	20	8	4	10	76	82	20,8	3
21	101	38	42	18	22	44	36	11,2	1

Procedeu-se a análise discriminante com as 21 parcelas, pré-classificadas nas três classes de sítio. Utilizou-se o método stepwise, tendo como critério de seleção de variáveis a maximização da Distância Generalizada de Mahalanobis (D^2) entre as duas mais próximas.

As análises estatísticas são descritas por COOLEY e LOHNES (1971) e SPSS (1999). BRAGA (1997) descreve com mais detalhes essa metodologia estatística.

Um modelo discriminante com duas funções lineares foi derivado (Quadros 3 e 4), obtendo-se 100% de classificações corretas das parcelas nas três classes de qualidade de sítio (Quadro 5).

Quadro 3 - Resumo do procedimento de Stepwise

Passo	Entrada	Wilks' Lambda	Sig. (%)	D^2	Sig. (%)	Entre Classes		Acertos (%)
1	Al_2O_3 A	0,69798	3,93	0,51	16,57	2	3	57,14
2	Al_2O_3 B	0,53599	2,77	1,61	7,41	2	3	66,67

3	Fe_2O_3	A	0,36760	0,95	2,56	5,91	2	3	61,90
4	Fe_2O_3	B	0,28996	0,93	3,90	4,13	2	3	80,95
5	Areia	A	0,17187	0,19	4,96	4,38	2	3	95,24
6	Silte	B	0,11892	0,12	7,22	2,83	2	3	100,00

Quadro 4 - Funções discriminantes

Variáveis	Funções	
	1	2
Al_2O_3 horiz. A	-0,1524856	0,4850249
Al_2O_3 horiz. B	0,4255511	-0,3323927
Fe_2O_3 horiz. A	0,5560556	-0,5543184E-01
Fe_2O_3 horiz. B	-0,4164315	0,1336054
Areia horiz. A	0,5214411E-01	0,7114489E-01
Silte horiz. B	-0,9842278E-01	-0,2504084E-01
Constante	-7,371230	-3,936616

Quadro 5 - Resumo dos resultados de classificação das parcelas nas classes de sítio usando o modelo discriminante

Classe de Sítio Atual	Número de Casos	Classificação Prevista 1	2	3
1	5	5	0	0
		100%	0%	0%
2	8	0	8	0
		8%	100%	8%
3	8	0	0	8
		0%	0%	100%

Total geral de classificações corretas: 100%.

A primeira função explicou 58,12% da variância total envolvida no modelo, com uma correlação canônica de 83,08% (Quadro 6).
A segunda função explicou 41,88% da variância, com uma correlação de 78,49% (Quadro 6).
O modelo com as duas funções apresentou significância estatística de 0,10% (Quadro 6). Com a remoção da primeira função, o modelo apresentou significância de 1,11% (contribuição da segunda função).

Quadro 6 - Estatística das funções do modelo discriminante

Função	Variância	Correl. Canônica	Remoç. Função	Wilks' Lambda	Qui-quad.	GL	Sig. (%)
			0	0,1189	33,004	12	0,10
1	58,12	0,8308	1	0,3839	14,841	5	1,11
2	41,88	0,7849					

Para a classificação de novas parcelas nas três classes de sítio, pode-se usar as funções de classificação de Fischer (Quadro 7), considerando o maior escore obtido.

Quadro 7 - Funções de classificação de Fisher

Variáveis	Classes de Sítio		
	1	2	3
Al_2O_3 horiz. A	3,312118	2,164546	3,467604

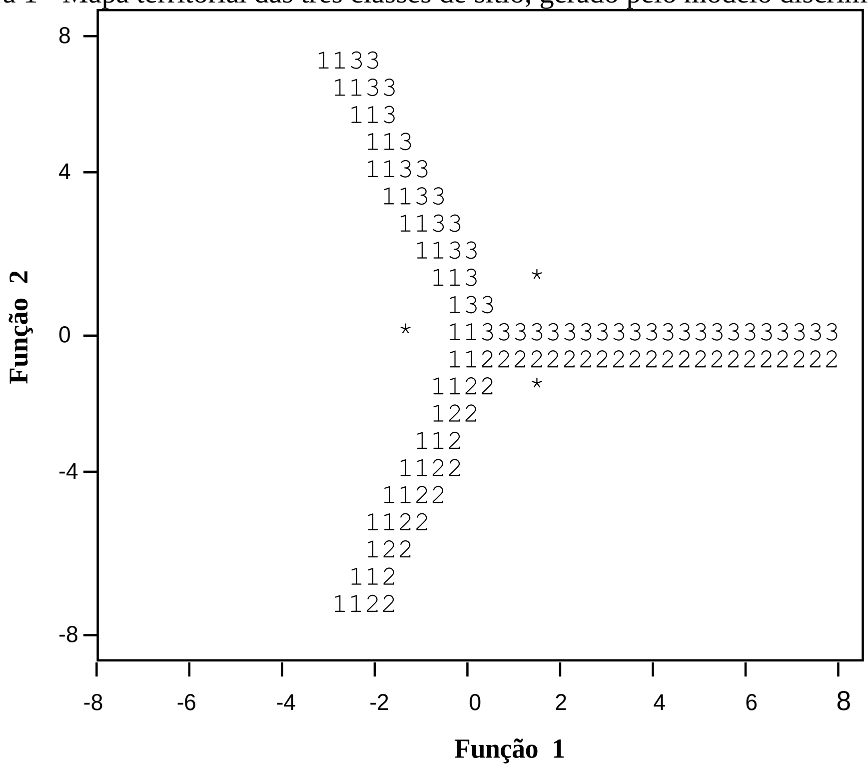
Al_2O_3 horiz. B	1,709598	3,537183	2,645179
Fe_2O_3 horiz. A	2,896468	4,777591	4,630342
Fe_2O_3 horiz. B	-1,708904	-3,239253	-2,881472
Areia horiz. A	1,403871	1,477175	1,668543
Silte horiz. B	-0,7971284	0,5116682	0,4440645
Constante	71,27365	-88,02296	-98,62636

O mapa territorial (Figura 1) delimita as regiões pertinentes a cada classe de sítio, no espaço bi-dimensional gerado pelo modelo discriminante (Quadro 4). Os centroides dos grupos representativos das três classes de sítio, encontram-se no Quadro 8.

Quadro 8 - Coordenadas dos Centroides

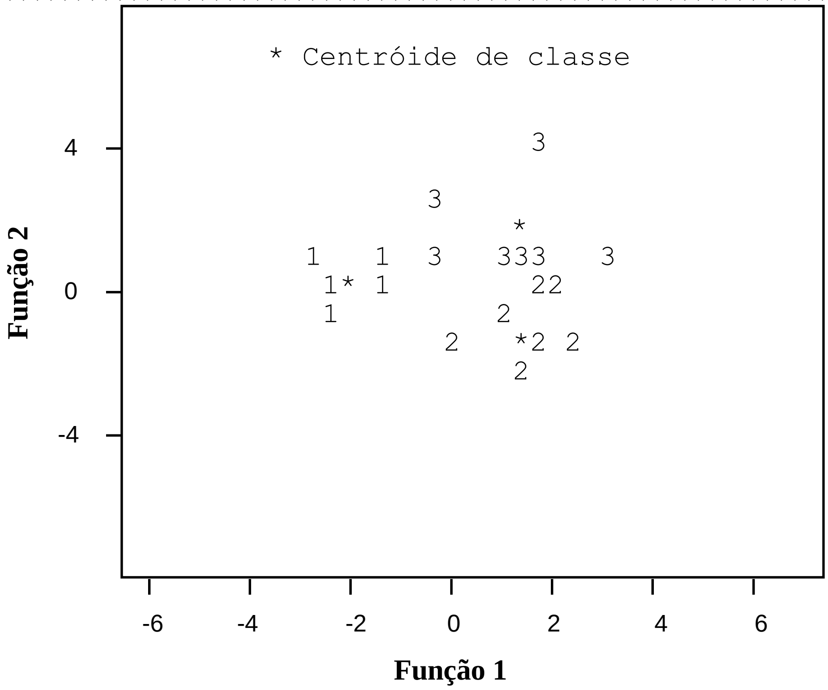
Classes Sítio	Função 1	Função 2
1	-2,47198	0,00205
2	0,77094	-1,34442
3	0,77404	1,34314
*	Centroides da Classe	

Figura 1 - Mapa territorial das três classes de sítio, gerado pelo modelo discriminante.



A Figura 2 mostra a dispersão das parcelas em torno dos centroides das classes.
Cada função discriminante expressa a influência distinta de um complexo de variáveis ambientais, facilitando a interpretação dos fatores envolvidos na definição da capacidade produtiva dos sítios. As correlações combinadas intragrupos permitem identificar em qual função a variável exerce seu papel principal.

Figura 2 - Dispersão das parcelas em torno dos centroides.



No presente caso, na função 1 (Quadro 4) atuam basicamente os teores de Al_2O_3 nos horizontes A e B e Fe_2O_3 no horizonte A, e a fração areia do horizonte A (Quadro 9). A função 2 (Quadro 4) envolve principalmente os teores de Fe_2O_3 e níveis de silte no horizonte B. Cabe ressaltar, entretanto, que os níveis de silte no horizonte B do solo também exercem papel importante na função 1 (Quadro 9).

Quadro 9 - Correlações entre as variáveis e as funções discriminantes

Variáveis	Funções	
	1	2
Al_2O_3 horiz. B	**0,37718**	-0,26856
Al_2O_3 horiz. A	**0,30227**	0,13815
Fe_2O_3 horiz. A	**0,29355**	0,26020
Areia horiz. A	**-0,16887**	0,08906
Silte horiz. B	-0,27321	**-0,28903**
Fe_2O_3 horiz. A	0,19657	**0,21199**

Apesar do caráter holístico do modelo discriminante, acarretando pouco valor à contribuição individual isolada das variáveis, pode-se notar que os sítios de maior potencial produtivo para *Araucaria angustifolia* apresentaram a seguinte tendência para os horizontes superficial e sub-superficial do solo (Quadro 10): - relativamente maiores teores totais de óxidos de ferro e de alumínio e menores quantidades das funções areia e silte, ou seja, relativamente mais argilosos.

Quadro 10 - Médias das variáveis selecionadas nas três classes de sítio

Variáveis	Classes de Sítio		
	1	2	3
Al_2O_3 horiz. A	11,72	15,58	17,40
Al_2O_3 horiz. B	15,06	22,26	18,99
Fe_2O_3 horiz. A	5,78	9,70	14,26
Fe_2O_3 horiz. B	8,02	10,35	14,12
Areia horiz. A	41,60	27,62	32,62
Silte horiz. B	18,00	15,12	10,62

Em linhas gerais, essas características denotam solos mais velhos e mais intemperizados.
Concluindo, a análise discriminante possibilitou a seleção de seis características de solo que permitem separar três classes de qualidade de sítios para *A. angustifolia* com elevada precisão, rapidez e facilidade.
As seis características de solo refletem e expressam direta e indiretamente fatores ambientais determinantes da capacidade produtiva dos sítios para *A. angustifolia*.