

**UNIVERSIDADE FEDERAL DE SANTA MARIA  
CENTRO DE CIÊNCIAS NATURAIS E EXATAS  
DEPARTAMENTO DE ESTATÍSTICA  
CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA E MODELAGEM  
QUANTITATIVA**

**EMPREGO DA ANÁLISE DISCRIMINANTE DE  
FISHER PARA CLASSIFICAR TIPOS DE FLORESTAS  
NO BIOMA PAMPA, RS, BRASIL**

**MONOGRAFIA DE ESPECIALIZAÇÃO**

**Ricardo de Vargas Kilca**

**Santa Maria, RS, Brasil  
2009**

**EMPREGO DA ANÁLISE DISCRIMINANTE DE FISHER  
PARA CLASSIFICAR TIPOS DE FLORESTAS NO BIOMA  
PAMPA, RS, BRASIL**

**por**

**Ricardo de Vargas Kilca**

Monografia apresentada ao Curso de Especialização em Estatística e Modelagem Quantitativa, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Especialista em Estatística e Modelagem Quantitativa**.

**Orientador: Prof. Adriano Mendonça Souza**

**Santa Maria, RS, Brasil**

**2009**

## **ERRATA**

O ANEXO A não foi disponibilizado para publicação por ser parte do banco de dados do Inventário Florestal Contínuo do Estado do Rio Grande do Sul.

---

© 2009

Todos os direitos autorais reservados ao autor. A reprodução de partes ou do todo deste trabalho só poderá ser feita com autorização por escrito do autor.

Endereço: Rua Pedro Pereira, n. 108, Bairro Nsra de Lourdes, Santa Maria, RS, 97050-590  
Fone (0xx)55 3307-0167; End. Eletr: [rvkilka@gmail.com](mailto:rvkilka@gmail.com)

---

**Universidade Federal de Santa Maria  
Centro de Ciências Naturais e Exatas  
Departamento de Estatística  
Curso de Especialização em Estatística e Modelagem Quantitativa**

A Comissão Examinadora, abaixo assinada,  
aprova a Monografia de Especialização

**EMPREGO DA ANÁLISE DISCRIMINANTE DE FISHER PARA  
CLASSIFICAR TIPOS DE FLORESTAS NO BIOMA PAMPA, RS,  
BRASIL**

elaborada por  
**Ricardo de Vargas Kilca**

como requisito parcial para obtenção do grau de  
**Especialista em Estatística e Modelagem Quantitativa**

**Comissão Examinadora:**

**Adriano Mendonça Souza, Dr.**  
(Presidente/Orientador)

**Roselaine Ruviaro Zanini, Dra. (UFSM)**

**Anaelena Bragança de Moraes, Dra. (UFSM)**

Santa Maria, 10 de agosto de 2009.

## AGRADECIMENTOS

A realização deste trabalho só foi possível graças à colaboração de muitos, aos quais devo minha sincera gratidão:

- à Universidade Federal de Santa Maria, que me concedeu a oportunidade de realizar o curso de Especialização em Estatística e Modelagem Quantitativa;
- ao meu Orientador prof. Dr. Adriano Mendonça Souza que me aceitou como orientando, pela paciência nos momentos difíceis, pela confiança na realização desse estudo e por toda a logística proporcionada durante o curso;
- ao meu co-orientador prof. Dr. Solon Jonas Longhi (CCR/UFSM) pelos anos de convívio e pesquisas o que culminou mais uma vez na oportunidade de trabalhar sob sua orientação em assuntos relacionados às florestas do Rio Grande do Sul;
- a Banca Examinadora composta pelas professoras Roselaine Ruviaro Zanini e Anaelena Bragança de Moraes, que contribuíram decididamente para o aprimoramento da versão final dessa monografia;
- a SEMA por disponibilizar dados do Inventário Florestal Contínuo do Rio Grande do Sul para a realização desse estudo;
- a todos os professores da Pós - Graduação pelas ótimas aulas que tive e pelo convívio agradável ao longo do curso: Dra. Anaelena B. Moraes, Dra Roselaine R. Zanini, Msc. Denis Altieri, Dra. Luciane D. Flores, Dr. Ivanor Müller, Dr. Luis Felipe D. Lopes. Especial agradecimento ao prof. Dr. Fernando Pulgati que ministrou as aulas de Análise Multivariada e que contribuiu definitivamente para a reorganização do meu projeto inicial.
- a todos os funcionários do CCNE e Depto de Estatística, em especial à Sandra, secretária do PPG- Estatística e Modelagem Quantitativa.
- aos colegas do curso de Especialização, em especial ao amigo Laion;
- à minha família (mãe, irmãos, namorada, sogra, etc...).

## **RESUMO**

Monografia de Especialização

Programa de Pós-Graduação em Estatística e Modelagem Quantitativa

Universidade Federal de Santa Maria, RS, Brasil

### **EMPREGO DA ANÁLISE DISCRIMINANTE PARA CLASSIFICAR TIPOS DE FLORESTAS NO BIOMA PAMPA, RS, BRASIL**

AUTOR: RICARDO DE VARGAS KILCA

ORIENTADOR: ADRIANO MENDONÇA SOUZA

CO-ORIENTADOR: SOLON JONAS LONGHI

Data e Local da Defesa: Santa Maria, 10 de agosto de 2009.

A análise discriminante de Fisher (ADF) é um método estatístico multivariado de redução de dimensão que tem por objetivo criar uma transformação linear ótima das variáveis independentes que maximize a diferença entre grupos e minimize as diferenças entre os grupos. Foi usada a ADF para desenvolver um esquema de classificação utilizando atributos estruturais florestais para diferenciar tipos de fitofisionomias florestais no bioma Pampa do Rio Grande do Sul, Brasil. O estudo foi conduzido em um banco de dados que representou 5 ha do Inventário Florestal do Estado do Rio Grande do Sul. Antes da análise todas as premissas da análise discriminante foram cumpridas. O método *Stepwise* incluiu todas as oito variáveis no modelo discriminante. A primeira função discriminante (FD1) do modelo representou 74,7% de habilidade de discriminação dos grupos pelas variáveis e a segunda função discriminante (FD2) gerou 14%. A correlação canônica associada com essas duas funções foi elevada ( $r=90$ ). A FD1 esteve mais relacionada com as variáveis estruturais altura total, altura comercial e riqueza de espécies. A FD2 apresentou mais relação com as variáveis, área basal e DAP máximo. O modelo classificou todas as 50 unidades amostrais em seus devidos grupos. Essa metodologia proporcionou informações úteis e inéditas de que as fitofisionomias no Pampa podem ser classificadas quanto as suas diferenças estruturais e assim, sustentaram a hipótese de que o método pode ser empregado em outros biomas do Brasil em que sejam disponíveis dados de inventários florestais.

Palavras-chave: Análise discriminante de Fisher; florestas subtropicais; bioma Pampa.

## **ABSTRACT**

Specialist Monographs

Statistics and Quantitative Modelling Postgraduate Course

Santa Maria Federal University, RS, Brazil

### **USING FISHER'S DISCRIMINANT ANALYSIS FOR CLASSIFY FORESTS ON PAMPA BIOME, RS, BRAZIL**

AUTHOR: RICARDO DE VARGAS KILCA

ADVISOR: DR. ADRIANO MENDONÇA DE SOUZA

CO ADVISOR: DR. SOLON JONAS LONGHI

MONOGRAPH DEFENSE: Santa Maria, 10th August 2009.

Fisher's discriminant analysis (FDA) is a dimension reduction method which finds an optimal linear transformation that maximizes the independent variables between-class scatter and minimize the within-class scatter. Was used FDA to improve an existing classification scheme by identifying those structural vegetation attributes that can jointly differentiate among a set of forest in the Pampa biome, southern of Brazil. The study were conducted in data base on 5 ha tree-plots of Forest's Inventory of Rio Grande do Sul State. Before analysis all premisses for analysis was garanted. In forward stepwise analysis all variables contributes for created the Fisher discriminant functions (FDF). The FDF1 accounts 74,7% of the discriminating ability of the discriminanting variables and FDF2 accounts 14%. The canonical correlation associated with the two function was  $> 90$ . In FDF1 implying a relation with attributes: species richness, commercial height and total height of trees. The FDF2 the relationship are with attributes basal área and maximum DBH in plot. All of 50 cases werw successfully classificated by FDA. These methodology and results should provide useful information for studies conducted in the classification forests in Rio Grande do Sul State and may be applicable to classification forests the other biomes of Brazil with properties measurement issues comparable to those Inventory.

Key-words: Fisher's discriminant analysis; subtropical forests; Pampa biome.

## LISTA DE ILUSTRAÇÕES

FIGURA 1 – Abrangência do bioma Pampa nos três países da América do Sul e localização dos cinco tipos de florestas no Pampa do estado do Rio Grande do Sul .....	20
FIGURA 2 – Fotografias da fisionomia dos cinco tipos de florestas subtropicais analisadas no bioma Pampa do Rio Grande do Sul: floresta de galeria (A), savanas do sudoeste (B), floresta ciliar (C) e floresta litorânea arenosa (D) e floresta estacional semidecidual (E) .....	21
FIGURA 3 – Modelo do croqui do arranjo das 100 unidades amostrais (10 x 10 m) instalados nos cinco tipos de florestas do Pampa no Rio Grande do Sul .....	22
FIGURA 4 – Output original com a primeira análise demonstrando o número de casos inseridos no modelo .....	42
FIGURA 5 – Output original com os valores da média e desvio-padrão de cada variável independente nas cinco variáveis dependentes do modelo. São fornecidas informações sobre o número de casos utilizados para o cálculo das médias (10 casos) .....	43



FIGURA 6 – Output original da matriz de correlação entre as variáveis nos grupos .....	44
FIGURA 7 – Output original do teste de lambda de Wilks e F de igualdade de médias de grupo .....	44
FIGURA 8 – Output original do teste M de Box para a igualdade das matrizes de covariância .....	45
FIGURA 9 – Output original revelando a seleção das variáveis no modelo através da estatística lambda de Wilks .....	46
FIGURA 10 – Output original das variáveis descritoras que foram inseridas pelo critério de inclusão avançada (Forward) através da estatística de lambda de Wilks .....	47
FIGURA 11 – Output original das variáveis descritoras que foram inseridas pelo critério de eliminação retrógrada (Backward) através da estatística de lambda de Wilks .....	48
FIGURA 12 – Output original dos resultados do teste dos autovalores e suas respectivas porcentagem de variância, variância cumulativa e correlação canônica para cada função discriminante .....	49
FIGURA 13 – Output original do teste de dimensionalidade para os cinco grupos conforme a estatística lambda de Wilks e Qui-quadrado .....	49
FIGURA 14 – Output original do cálculo dos autovetores ou escores discriminantes das variáveis em cada uma das quatro funções discriminante gerada pelo modelo .....	50
FIGURA 15 – Output original dos escores discriminantes originados pela correlação entre as variáveis discriminatórias e os escores das funções discriminantes canônicas padronizadas .....	51
FIGURA 16 – Output original dos valores dos centróides de grupos .....	51

FIGURA 17 – Output original do número de casos processados para a classificação .....	52
FIGURA 18 – Output original revelando que a probabilidade inicial para a separação de grupos foi a mesma para todos os grupos .....	52
FIGURA 19 – Output original demonstrando os coeficientes de classificação de Fisher das variáveis preditoras em cada um dos grupos .....	53
FIGURA 20 – Output original com os escores discriminantes de cada caso em cada função discriminante e seus valores de distância Mahalanobis em relação ao seu centróide de grupo. O valor da estatística da distância de Mahalanobis também foi calculado para o caso em relação ao grupo original e grupo predito .....	54
FIGURA 21 – Output original do mapa perceptual bidimensional representando os valores dos escores dos centróides de grupos e dos escores discriminantes de cada caso em seus respectivos grupos .....	55
FIGURA 22 – Output original da tabela de classificação .....	56
FIGURA 23 – Output modificado com os valores das classificações dos casos no grupo teste com as probabilidades de classificação pela distância de Mahalanobis. Para os dados de validação cruzada à distância Mahalanobis quadrada é baseada nas observações .....	57

## LISTA DE TABELAS

TABELA 1 – Localização e características ambientais dos cinco tipos de florestas subtropicais no bioma Pampa do Rio Grande do Sul .....	20
TABELA 2 – Banco de dados com o tamanho amostral inicial e modificado para a realização da análise discriminante linear para o conjunto das amostras dos cinco tipos de florestas subtropicais no Pampa do Rio Grande do Sul .....	23
TABELA 3 – Critério para a análise de variância (ANOVA) para um fator .....	30
TABELA 4 – Tabela de classificação para análise discriminante .....	40

## LISTA DE ABREVIATURAS E SIGLAS

AB = média da área basal em m<sup>2</sup>;

AD = análise discriminante;

ADM = análise discriminante múltipla;

ALTCOM = altura média comercial em metros;

ALTMAX = altura máxima atingida na área amostral em metros;

ALTTOT = altura média total em metros;

ANOVA = análise de variância de um fator;

Cfa = clima temperado húmido com verão quente conforme Köppen-Geiger;

CILIAR = floresta ciliar;

DAP = média dos diâmetros dos caules amostrados em centímetros;

DAPMAX = diâmetro máximo do caule na área amostral em centímetros;

ECO/92 = Conferência das Nações Unidas para o Meio Ambiente e o Desenvolvimento no ano de 1992.

FES = floresta estacional semidecidual;

FD = floresta estacional decidual;

GALERIA = floresta de galeria;

IFC-RS = Inventário Florestal Contínuo do Rio Grande do Sul;

MANOVA = Análise de Variância Multivariada;

MMA = Ministério do Meio Ambiente;

N = número médio de indivíduos amostrados;

RESTINGA = floresta de restinga arenosa;

RIQ = riqueza ou número médio de espécies amostradas;

SAVANA = formação de savana arbórea;

SEMA = Secretaria do Meio Ambiente do Estado do Rio Grande do Sul;

SPSS = Statistical Package for the Social Sciences;

UFSM = Universidade Federal de Santa Maria

## LISTA DE ANEXOS

ANEXO A – Banco de dados com valores originais que foi utilizado para a geração da análise discriminante linear de Fisher .....	72
ANEXO B – Banco de dados com valores originais dos atributos estruturais das unidades amostrais que foram utilizadas para o teste de validação cruzada do modelo .....	74
ANEXO C – Resultados dos testes de normalidade para as variáveis independentes por grupo (valores apenas da probabilidade dos testes) .....	76
ANEXO D – Resultados da análise descritiva para as variáveis independentes por grupo .....	78
ANEXO E – Resultados dos testes ANOVA um fator, seguido dos testes de Tukey, para as oito variáveis discriminantes nos cinco grupos .....	81

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b> .....	1
<b>1.2</b>	<b>Justificativa</b> .....	3
<b>1.3</b>	<b>Objetivo Geral</b> .....	4
<b>1.4</b>	<b>Objetivos Específicos</b> .....	4
<b>2</b>	<b>REVISÃO DE LITERATURA</b> .....	5
<b>2.1</b>	<b>A análise discriminante</b> .....	5
2.1.1	Critérios para a escolha da análise discriminante .....	7
2.1.2	Objetivos da Análise Discriminante .....	7
2.1.3	A seleção dos grupos e das variáveis discriminantes .....	8
2.1.4	O tamanho da amostra .....	9
2.1.5	As premissas em comum com outros testes multivariados.....	10
2.1.6	Usando a probabilidade <i>a priori</i> para classificação .....	12
2.1.7	Validação do modelo .....	13
2.1.8	O método linear de Fisher .....	13
<b>2.2</b>	<b>A fisionomia e a estrutura florestal</b> .....	16
2.2.1	O bioma Pampa e os tipos de fisionomias florestais .....	16

2.2.2	A gestão sustentável de florestas .....	17
<b>3</b>	<b>MATERIAL E MÉTODOS</b> .....	<b>19</b>
<b>3.1</b>	<b>Áreas de estudo</b> .....	<b>19</b>
<b>3.2</b>	<b>Delineamento amostral da vegetação</b> .....	<b>21</b>
<b>3.3</b>	<b>Delineamento estatístico</b> .....	<b>24</b>
3.3.1	Passos desenvolvidos no SPSS .....	25
3.3.2	Os cálculos e o significado dos testes utilizados no SPSS .....	26
3.3.2.1	Etapa de validação .....	26
3.3.2.2	Etapa do cálculo das funções discriminantes.....	31
3.3.2.3	Etapa de classificação .....	37
<b>4</b>	<b>RESULTADOS</b> .....	<b>42</b>
<b>4.1</b>	<b>Etapa de validação</b> .....	<b>42</b>
<b>4.2</b>	<b>Etapa do cálculo das funções discriminantes</b> .....	<b>45</b>
<b>4.3</b>	<b>Etapa de classificação</b> .....	<b>52</b>
<b>5</b>	<b>CONSIDERAÇÕES FINAIS</b> .....	<b>58</b>
<b>5.1</b>	<b>O modelo discriminante de Fisher</b> .....	<b>58</b>
<b>5.2</b>	<b>A comparação entre as variáveis discriminatórias nos modelos univariado e multivariado</b> .....	<b>60</b>
<b>5.3</b>	<b>Implicações do modelo gerado para o entendimento ecológico e para a gestão sustentável das florestas no Pampa</b> .....	<b>62</b>
<b>6</b>	<b>CONCLUSÕES</b> .....	<b>65</b>
<b>7</b>	<b>REFERÊNCIAS BIBLIOGRÁFICAS</b> .....	<b>66</b>
	<b>ANEXOS</b> .....	<b>71</b>



# 1 INTRODUÇÃO

A estatística multivariada consiste em um conjunto grande de métodos estatísticos utilizado em situações nas quais as variáveis são medidas, simultaneamente, em cada unidade amostral dentro de um contexto do esforço amostral ou experimento (KENKEL *et al.*, 2002, MINGOTI, 2005). A organização básica dos dados é sumarizada em uma matriz com um conjunto das  $p$  variáveis medidas em cada  $n$  unidade amostral. Assim, a análise interpretativa dos dados multivariados considera sempre todas as variáveis, simultaneamente, mais que o conjunto das  $p$  variáveis independentemente (KENKEL *et al.*, 2002). Essa matriz pode ser não particionada ou particionada em vários grupos ou tratamentos podendo gerar uma estrutura muito complexa dos dados. A aplicação e a interpretação correta desses dados requerem o entendimento das premissas, suposições e limitações teóricas específicas de cada técnica multivariada a ser empregada que, ao mesmo tempo, está atrelada ao problema do estudo.

Em geral, os métodos multivariados são utilizados com o propósito de simplificar ou facilitar a interpretação do fenômeno que está sendo estudado através da construção de índices ou variáveis alternativas que sintetizem a informação de dados complexos: construir grupos de elementos amostrais que apresentem similaridade entre si, possibilitando a segmentação do conjunto de dados originais, investigar as relações de dependência entre as variáveis respostas associadas ao fenômeno e outros fatores (variáveis explicativas), muitas vezes com objetivo de predição; comparar populações ou validar suposições através de testes de hipóteses (MINGOTI, 2005). Assim, dois principais tipos de modelos multivariados podem ser reconhecidos: a) o modelo descritivo que envolve a descrição e sumarização da estrutura dos dados revelando a inter-relação entre as variáveis, geralmente usados para gerar hipótese sobre os mecanismos causais que produzem a estrutura subjacente dos dados; b) os modelos preditivos desempenham o mesmo objetivo de sumarização da estrutura dos dados embora utilizando muito mais testes de hipóteses, bem como, desenhos amostrais e levantamentos mais formalizados, na ajuda do diagnóstico para a tomada de decisão (REIS, 1995; LEGENDRE & LEGENDRE, 1998).

As análises multivariadas para estudos de vegetação iniciaram na década de 1930 (FISHER 1936 *apud* REIS 1995) e a partir daí inúmeras técnicas foram empregadas em estudos para os mais diferentes fins (ver GAUCH, 1982; PILLAR & ORLÓCI, 1993; LEGENDRE & LEGENDRE, 1998; VALENTIM, 2000, KENKEL *et al.*, 2002).

Existem diversas técnicas multivariadas usadas especificamente para classificação ou distinção de grupos, entre elas as mais comuns são, a análise de agrupamentos (*cluster analysis*), a análise de variância multivariada (MANOVA) e a análise discriminante (*discriminant analysis*).

A análise discriminante tem como objetivo procurar distinguir estatisticamente dois ou mais grupos de casos e/ou contribuir para determinar como podem ser alocadas novas observações dentro dos grupos (KLECKA, 1975; RENCHER, 2002; HÄRDLE & SIMAR, 2003). A técnica surgiu com aplicações nas áreas da biologia e ciências médicas, mas atualmente, encontra-se difundida em quase todas as outras áreas da ciência (HUBERTY & OLEJNIK, 2006). Atualmente, existem vários métodos de cálculo para gerar uma análise discriminante (REIS, 1995; KHATTREE & NAIK, 2000; MONGOTI, 2005) e um dos mais conhecidos é através das funções discriminantes de Fisher.

No Brasil, ainda são poucos os estudos que utilizam a análise discriminante em estudos de vegetação (SOUZA *et al.*, 2003; SILVA *et al.*, 2004; LONGHI, 1997). No estado do Rio Grande do Sul, Longhi (1997) foi o primeiro a utilizar diversas técnicas multivariadas para identificar prováveis diferenças entre fisionomias florestais de uma mesma formação fitoecológica a partir de atributos estruturais da vegetação. Entre as técnicas utilizadas pelo autor encontra-se a análise discriminante linear de Fisher. Mais tarde alguns outros estudos foram desenvolvidos com a técnica (GERHARDT *et al.*, 2001) porém, nenhum estudo buscou empregá-la para avaliar a diferença entre os tipos diferentes de florestas ocorrentes em uma determinada região.

Este estudo foi desenvolvido com a proposta de utilizar a análise discriminante linear de Fisher para comprovar a hipótese de que cinco tipos de florestas no Pampa podem diferir estatisticamente entre si quanto aos seus aspectos estruturais e verificar quais os atributos estruturais que são mais importantes para a discriminação dessas florestas. Foram avaliadas amostras de cinco tipos de florestas ocorrentes no Pampa: uma mata galeria, uma mata ciliar, uma floresta de restinga arenosa, uma floresta estacional semidecidual e uma vegetação savânica (espinal) da fronteira sudoeste. Cada floresta apresentou oito variáveis quantitativas relacionadas com a estrutura das florestas para serem comparadas: o número de espécies e indivíduos, a altura média, comercial e máxima absoluta, o diâmetro do caule à altura do peito (DAP) através dos valores médios e máximo e, finalmente, a área basal. Essas variáveis quantitativas foram utilizadas como variáveis descritoras (independentes) para distinguir os quatro tipos de florestas (variáveis dependentes) no modelo de análise discriminante. Todos os dados foram provenientes do Inventário Florestal Contínuo do Estado

do Rio Grande do Sul (SEMA/UFSM, 2002). O estudo foi desenvolvido no software SPSS e os resultados foram apresentados no formato das saídas (*output*) do programa.

## 1.2. Justificativa

Há uma escassez de estudos detalhados e comparativos que relatem as diferenças e semelhanças entre as formações vegetais que compõe os biomas do Brasil. Dentre todos os biomas brasileiros, o Pampa é um dos mais carentes de estudos em seus ecossistemas (MMA 2000). Nesse bioma, ainda não se sabe de forma exata quantos tipos de formações vegetais distintas ocorrem, sua biodiversidade total e a taxa de perda de ecossistemas naturais devido às intervenções antrópicas. Essas carências de informações mais aprimoradas e detalhadas são atualmente os principais empecilhos à elaboração de planos eficazes de conservação e de utilização racional da biodiversidade no bioma. Mesmo assim, alguns esforços estão sendo feitos para a conservação do bioma, em ampla escala, por meio da Política Nacional de Conservação da Biodiversidade (MMA 2002).

Dentre os inúmeros ecossistemas do bioma Pampa, as florestas destacam-se por serem as comunidades vegetais com menor área de abrangência e cobertura, por comportarem uma diversidade singular em relação aos demais ecossistemas, por se apresentarem entre as fitofisionomias mais ameaçadas de preservação e, por fim, revelar os menores níveis de conhecimentos científicos. Essas premissas comprovam a necessidade urgente de estudos ecológicos nesses ecossistemas.

Atualmente, os tipos de florestas no bioma são reconhecidos a partir de estudos descritivos de acordo com algumas poucas características fisionômicas e de preferência da vegetação por determinado ambiente físico. Assim, admitiu-se subjetivamente, e por muito tempo, a classificação distinta para essas florestas sem atribuir valores do quanto essas florestas diferem entre si (quanto aos diferentes atributos da vegetação), ou melhor, em que nível de confiança pode-se atribuir a diferença entre essas florestas. Mesmo sem existir qualquer estudo comparativo entre as diferentes florestas do Pampa, alguns estudos quantitativos sobre a estrutura das mesmas levantam a hipótese de que esses atributos quantitativos podem ser utilizados como importantes variáveis para a classificação das diferentes formações vegetais. Portanto, desenvolver uma técnica para classificar essas mesmas florestas através de critérios quantitativos estruturais da vegetação utilizando a

estatística multivariada é pertinente, pois, permitiria um melhor entendimento das características que determinam as diferenças entre as florestas no bioma.

Esse estudo é o primeiro a empregar a análise discriminante de Fischer para avaliar se alguns atributos da estrutura da vegetação podem revelar diferenças estatisticamente significativas entre alguns tipos de florestas ocorrentes no Pampa.

### **1.3. Objetivo geral**

Determinar por meio da análise discriminante linear de Fisher se os principais tipos de florestas no Pampa do Rio Grande do Sul formam grupos estatisticamente diferentes entre si quanto à estrutura do componente arbóreo e identificar quais os atributos estruturais da vegetação são mais eficientes para discriminar esses grupos.

### **1.4. Objetivos específicos**

O presente estudo tem como objetivos específicos:

- A) determinar se existem diferenças estatisticamente significantes entre os perfis de escore médio para um conjunto de oito variáveis estruturais das comunidades vegetais de cinco tipos de fitofisionomias florestais distintas no estado do Rio Grande do Sul.
- B) determinar quais as variáveis independentes que podem explicar o máximo de diferença nos perfis de escore médio dos grupos.
- C) avaliar o percentual de classificação correta dos casos e grupos gerados pelo modelo.
- D) interpretar os dados gerados a partir dos *outputs* fornecidos pela análise discriminante utilizando o Programa SPSS.

## 2 REVISÃO BIBLIOGRÁFICA

### 2.1 A análise discriminante

Algumas idéias associadas com a análise discriminante (AD) iniciaram desde 1920 com o estatístico Karl Pearson e G. M. Morant. Os conceitos matemáticos continuaram sendo aprimorado na década de 1930 por P. C. Mahalanobis. Mas foi R. A. Fisher (1936 *apud* HUBERTY & OLEJNIK, 2006) que propôs a idéia de traduzir distâncias multivariadas intergrupos em uma combinação linear das variáveis para classificar grupos. A extensão do método para a classificação de múltiplos grupos foi desenvolvida por C. R. Rao em 1948. A partir disso, muitas outras extensões e refinamentos das idéias de Fisher foram desenvolvidas (McLACHLAN, 2004; HUBERTY & OLEJNIK, 2006). O potencial de aplicação da AD é grande sendo utilizada em diversas áreas científicas (HUBERTY & OLEJNIK, 2006).

Quando a análise discriminante é realizada para distinguir dois grupos é denominada de análise discriminante de dois grupos, por outro lado, quando é utilizada para classificar três ou mais grupos, chama-se análise discriminante múltipla (ADM) (HAIR *et al.*, 2007). A análise discriminante é apropriada quando a variável dependente, no caso os grupos, é categórica (nominal ou não-métrica) e as variáveis independentes, as variáveis discriminantes, são métricas (HAIR *et al.*, 2007). Para revelar as diferenças de grupos, a AD permite ainda discernir o papel das variáveis discriminantes de forma individual e os efeitos coletivos na distinção dos grupos (HAIR *et al.*, 2007).

Conforme Legendre & Legendre (1998), a análise discriminante é um modelo linear multivariado da mesma forma que a análise de variância multivariada (MANOVA), a regressão linear múltipla e a análise de correlação canônica. De forma sintética, o procedimento para desenvolver uma análise segue dois passos: (1) primeiro, um teste de diferença nas variáveis explanatórias ( $X$ ), entre os grupos predefinidos. Essa parte da análise é idêntica aos testes desenvolvidos pelo MANOVA, (2) depois, verificar através de testes de hipóteses a diferença entre os grupos devido as variáveis explanatórias, essa análise prossegue até encontrar uma combinação linear (chamada funções discriminantes) das  $X$  variáveis que melhor discriminam os grupos.

Segundo Rencher (2002), o objetivo matemático da análise discriminante para a separação de grupos é desenvolver uma função linear das variáveis (funções discriminantes) para descrever as diferenças entre os grupos. Cada combinação linear constitui uma função discriminante. Assim, a análise discriminante descritiva consiste em determinar a contribuição relativa das  $p$  variáveis para separação dos grupos e encontrar no plano os pontos que podem ser projetados em um plano bi-dimensional para melhor ilustrar as configurações dos grupos. Segundo o mesmo autor, a análise preditiva do modelo discriminante está relacionada com a classificação dos casos nos grupos usando funções lineares ou quadráticas das variáveis preditoras (função de classificação) para serem empregadas na análise dos casos nos grupos. Os vetores das observações para um indivíduo ou objeto são medidos para a função de classificação encontrar o grupo para o qual o indivíduo mais se assemelha. Assim, as funções discriminantes são combinações lineares das melhores variáveis que separam os grupos (dois ou mais). Em suma, a proposta primária da análise discriminante linear é separar amostras de distintos grupos através da maximização da separação entre suas classes e minimizando a variabilidade dentro das classes.

Segundo Reis (1995), para o entendimento da técnica, supõe-se que  $n_1$  elementos amostrais procedentes, com probabilidade 1, da população de A e  $n_2$  elementos amostrais procedentes, com probabilidade 1, da população B, e que em cada um dos  $n_1+n_2 = n$  elementos amostrais tenham sido medidas  $p$  variáveis aleatórias (características). Uma análise estatística do comportamento das  $p$  características medidas permite identificar o perfil geral de cada grupo. Deste modo, se houver um novo elemento amostral, que não pertence a nenhuma das duas amostras anteriores, e cuja origem é incerta, seria possível compará-lo de algum modo com o perfil geral dos grupos A e B e classificá-lo como pertencente ao grupo cujo perfil geral fosse mais semelhante ao dele. Na análise discriminante, a comparação do elemento amostral em relação aos grupos candidatos é, em geral, feita através da construção de uma regra matemática de classificação, ou discriminação fundamentada na teoria das probabilidades. Para cada novo elemento amostral, a regra de classificação permitirá ao pesquisador decidir qual é a população mais provável de ter gerado seus valores numéricos nas  $p$  características avaliadas.

### 2.1.1. Critérios para a escolha da análise discriminante

A aplicação de uma AD se baseia em uma série de critérios a serem adotados, como: o objetivo da escolha do método, a seleção das variáveis dependentes e independentes, o tamanho necessário da amostra para estimar as funções discriminantes e a divisão da amostra para fins de validação do modelo. Recomenda-se os seguintes passos para desenvolver uma análise discriminante descritiva (BROWN & WICKER, 2000):

1. determinar se a análise discriminante pode proporcionar um resultado estatístico para descrever a sua questão de pesquisa (i.e., a análise discriminante apropriada para responder sua questão?);
2. determinar se os dados são apropriados para a realização da análise discriminante;
3. definir os grupos que deverão ser usados para a análise;
4. selecionar as variáveis que serão usadas na análise;
5. testar os dados para assegurar que as suposições da análise discriminante são encontradas. Se alguma suposição não concordar, determinar se a análise discriminante é robusta para a suposição levantada;
6. desenvolver a análise;
7. interpretar os resultados;

### 2.1.2. Objetivos da análise discriminante

Conforme Hair *et al.*, (2007) a AD pode contribuir para abordar qualquer um dos seguintes objetivos de pesquisa:

- A) determinar se existem diferenças estatisticamente significantes entre os perfis de escore médio em um conjunto de variáveis para os grupos definidos *a priori*.

- B) determinar quais das variáveis independentes explicam o máximo de diferença nos perfis de escore médio dos grupos.
- C) estabelecer procedimentos para classificar objetos (indivíduos, produtos, etc.) em grupos, com base em seus escores em um conjunto de variáveis independentes.
- D) estabelecer o número e a composição das dimensões de discriminação entre grupos formados a partir do conjunto de variáveis independentes.

Assim, o pesquisador pode usar a AD para compreender diferenças de grupos ou como classificar objetos corretamente em grupos ou classes.

### 2.1.3. A seleção dos grupos e das variáveis discriminantes

Conforme Brown & Wicker (2000), a AD requer um conjunto de dados que contenham dois ou mais grupos mutualmente exclusivos (variável dependente) e escores de duas ou mais variáveis independentes para cada caso no grupo. Os grupos podem ser construídos com base de suas características demográficas (p.ex. sexo, etnicidade, estado social), atributos pessoais (p.ex. altura, desordem psicológica, tipo de sangue), ou passado/presente de comportamento (p.ex. se assiste televisão, se tem alucinações ou fuma cigarros). Os grupos também poderiam ser construídos através de registros de outros estudos anteriores ou também na dimensão que o investigador considera seguro para representar as diferenças que existem entre os grupos. Existem outras formas de utilizar as variáveis dependentes numa AD e por isso aconselha-se consultar outras fontes bibliográficas (BROWN & WICKER, 2000; HAIR *et al.*, 2005).

A análise discriminante avalia o grau no qual as variáveis independentes diferenciam os grupos, por isso são denominadas de "variáveis discriminatórias". A efetividade da AD depende do quanto que essas variáveis independentes definem os grupos formados. Portanto, a decisão de selecionar certas variáveis com potencial discriminatório é crítico para o sucesso da análise discriminante. Diversas diretrizes podem ser seguidas na seleção das variáveis discriminatórias. Primeiro, a consulta teórica sobre o fenômeno de estudo poderia ser a forma mais apropriada para determinar quais as variáveis que apresentam maior potencial discriminante para distinguir os grupos de interesse. Segundo, os investigadores poderiam



considerar as variáveis que tem se mostrado relevante em outros trabalhos de pesquisa (BROWN & TINSLEY, 1983). Testes de ANOVA de um fator podem proporcionar algumas evidências empíricas para selecionar as variáveis discriminantes, ou então, buscar a opinião de um pesquisador experiente para a seleção das variáveis com potenciais discriminatórias. Os investigadores deveriam ter cuidado para selecionar as variáveis discriminantes que não são altamente intercorrelacionadas devido a dificuldade para determinar de forma precisa as cargas das variáveis individuais na função discriminantes. Em resumo, se as variáveis são altamente correlacionáveis entre si, pode ocorrer efeito de redundância para a discriminação de grupo. Recomenda-se que o pesquisador consulte as matrizes de correlação para selecionar as variáveis que não são correlacionadas significativamente para usar como variáveis discriminatórias. Geralmente é possível fazer a descrição das relações entre os grupos ao longo de um contínuo de variáveis fundamentais. Isso permite desenvolver escalas contínuas na qual as diferenças entre grupos nas variáveis são denotadas.

Embora duas ou mais variáveis sejam requeridas para a análise, o número de  $n$  grupos não deveria exceder o número de  $p$  variáveis pelo fato da dificuldade de interpretar os resultados da análise com o aumento da seleção das variáveis discriminatórias (BROWN & WICKER, 2000). Por esta razão, sugere-se que os pesquisadores restrinjam a escolha das variáveis discriminantes para aquelas que têm maior relevância empírica ou teórica.

#### 2.1.4. Tamanho da amostra

O investigador necessita determinar o tamanho da amostra para realizar a análise discriminante. Sugere-se que os investigadores evitem unidades amostrais muito grandes porque com o aumento do tamanho amostral, a variabilidade nas distribuições amostrais diminui, presumindo que todos outros fatores como sendo constantes. Isto poderia resultar em uma redução de sobreposição entre distribuições amostrais dentro das hipóteses nulas e alternativas, conduzindo a uma maior probabilidade de encontrar uma significância estatística (rejeitando a hipótese nula). Entretanto, amostras muito pequenas não são recomendadas pelas mesmas tendências das amostras de serem indevidamente influenciadas pelos resultados estatísticos e, desse modo, reduzir a estabilidade e generalização dos resultados (BROWN &

WICKER, 2000). Pequenas amostras devem também ser evitadas devido à necessidade dos resultados da análise discriminante confirmarem sua aplicabilidade com outras amostras através de uma validação cruzada (*cross-validation*).

Alguns autores recomendam que o tamanho total da amostra deveria ser de pelo menos 10 vezes o número de variáveis discriminantes (BROWN & TINSLEY, 1983). Outros, argumentam que a razão entre os casos e variáveis que poderia estar 5:1 (HAIR *et al.*, 2007) até 20:1 (STEVENS, 1996) no mínimo. De preferência, os casos na amostra devem estar distribuídos equitativamente através dos grupos, mas muitas vezes isto não pode ser possível. Quando os grupos de interesses diferem em número de casos deve-se então tomar procedimentos diferenciados (ver WICKER & BROWN, 2000).

A AD também requer que os dados estejam completos dentro das variáveis discriminantes e que todos os membros dos grupos sejam conhecidos. Devem ser excluídas da análise aqueles casos onde não exista escore para as variáveis discriminantes. Ademais, casos com dados faltantes podem ser sistematicamente diferentes daqueles casos que contenham dados completos (NORUSIS, 1990). Por isso, recomenda-se que o pesquisador também examine as saídas das análises para determinar se o número de casos incluídos concorda com os da amostra e do tamanho dos grupos. Se os tamanhos das amostras não satisfazem essas premissas, haverá limitações na interpretação dos resultados. Além do mais, pesquisadores aplicados podem comparar os casos com alguns dados faltantes com dados completos para ver se eles diferem sistematicamente. Se os resultados apontarem diferenças, os casos irão alterar os resultados da análise.

#### 2.1.5. As premissas em comum com outros testes multivariados

Pelo fato dos dados na AD envolverem múltiplas variáveis, as premissas requeridas são as mesmas para outras análises multivariadas. Violações nas premissas implicam no aumento do erro do tipo I: (a) observações independentes (amostras que não sejam independentes); (b) normalidade multivariada (as variáveis descritoras não serem normalmente distribuídas), e (c) homogeneidade de covariâncias das matrizes (as matrizes de

covariância populacional que não sejam iguais as das variáveis discriminantes). Para Hair *et al.* (2007), a não adoção desses critérios podem causar problemas na estimação da função discriminante e deve-se tomar medidas específicas para sanar o problema. Portanto, os pesquisadores deveriam examinar e apresentar esses resultados em suas análises.

Entretanto, evidências sugerem que AD é robusta com respeito à violação da normalidade multivariada e a homogeneidade da matriz de covariância (STEVENS, 1996). Stevens (1996) sugere uma alternativa para avaliar a premissa da normalidade multivariada a partir da análise individual das distribuições dos escores em cada variável discriminatória. Realizar essas análises são medidas paliativas e alternativas que não garantem a distribuição multivariada normal dos dados, mas reduzem os riscos de os dados não ter distribuição normal.

Os pesquisadores podem testar a violação da homogeneidade das matrizes de covariância usando o teste M de Box no SAS ou SPSS. Esse teste deverá comprovar que dentro dos grupos a variabilidade deverá ser idêntica, isto é, as matrizes de variância e covariância devem ser iguais para todos os grupos (REIS, 1995). Quando violado, Klecka (1975) relata que a pior consequência é que os casos são mais prováveis de serem classificados dentro de grupos com maior dispersão. Alternativas de alguns *softwares* permitem ao investigador adotar o procedimento de analisar separadamente as matrizes de covariância para cada grupo ao invés da matriz de covariância (*pooled within-group covariance matrix*) (BROWN & WICKER, 2000). Stevens (1996) considera que a AD é especialmente robusta para a violação da premissa de homogeneidade de covariância das matrizes se a razão do maior grupo  $n$  dividido pelo menor grupo  $n$  é menor que 1.5. Manly (2008) comenta que a falha na suposição de igualdade das matrizes de covariâncias e a distribuição normal multivariada não significam que uma análise discriminante possa ter gerado discriminações erradas para os grupos, mas podem ocorrer erros no estabelecimento da significância estatística das diferenças entre os grupos. Em resumo, a literatura indica que a violação da homogeneidade das matrizes de covariância não traz grande importância na condução da validade da análise discriminante, e mesmo quando violada há um número de alternativas que o pesquisador pode adotar para corrigir o modelo.

A premissa de independência das observações, ou do efeito de multicolinearidade, é crítica e a sua violação aumenta a probabilidade de ocorrer o erro do tipo I. A multicolinearidade denota que duas ou mais variáveis independentes estão altamente

correlacionadas, de modo que uma variável pode ser altamente prevista pela outra (s) variável (s). Para assegurar que os escores de cada variável (observações) sejam independentes (que não estão altamente correlacionadas), o pesquisador necessita examinar a matriz de correlação das variáveis discriminantes. No geral, as correlações que tiverem valores absolutos menores que 0,3 podem ser consideradas na análise (TABACHNICK & FIDELL, 1989). Por outro lado, Stevens (1996) argumenta que correlações maiores que 0,3 podem não ser problemáticas para a análise se elas não forem estatisticamente significativas, mesmo mantendo um alfa mais conservativo (5% e 10%). De qualquer forma, seria interessante que fosse realizado um exame da matriz de correlação para determinar se as observações estão correlacionadas entre os grupos ou apenas dentro dos grupos.

#### 2.1.6 Usando a probabilidade *a priori* para a classificação

Segundo Hair *et al.*, (2005), para calcular corretamente o escore de corte quando há diferentes números de grupos, o pesquisador também deve determinar se os tamanhos observados de grupos refletem as verdadeiras proporções da população ou se os tamanhos dos grupos da população devem ser considerados iguais, ou seja, cada grupo tem a mesma probabilidade de ocorrer, mesmo que os tamanhos desses grupos (objetos) sejam diferentes. No entanto, se a amostra é aleatoriamente obtida da população de forma que os grupos de fato estimam as proporções da população em cada grupo, então, as melhores alternativas das probabilidades *a priori* refletem as proporções da amostra. O pesquisador sempre deve especificar as probabilidades em todas as análises, como sendo iguais ou baseadas em tamanhos de amostras, para garantir que as suposições corretas sejam assumidas no processo de classificação. Stevens (1996) acredita que o pesquisador deve usar a probabilidade igual quando se tem a certeza teórica ou base empírica de que as diferenças no tamanho do grupo refletem a diferenças no tamanho das respectivas populações, como, por exemplo, no caso de grupos étnicos e raciais.

### 2.1.7 Validação do modelo

Muitas vezes a amostra é dividida em duas partes, uma para gerar a função discriminante e a outra para fins de validação do modelo (HAIR *et al.*, 2007), i.e, para avaliar se a performance das funções geradas foram classificadas corretamente utilizando partes derivadas da amostra principal ou de amostras extras desde que sejam confirmadas do mesmo grupo. Essa análise se chama teste de validação cruzada (*cross-validation sample*) (TINSLEY & BROWN, 1983). O teste de validação cruzada não exige uma regra geral precisa para dividir a amostra nos grupos: de análise (amostra de análise) e de teste (avaliação cruzada), mas alguns autores recomendam uma partição de 60% - 40% ou 75% - 25% entre os dois grupos (HAIR *et al.* 2007).

### 2.1.8 O método linear de Fisher

R. A. Fisher (1890–1962) na década de 1930 desenvolveu a primeira técnica de análise discriminante originalmente em estudos botânicos com objetivo de fazer distinção entre grupos de plantas com base no tamanho e tipo de folhas, para que, posteriormente, fosse possível classificar as novas espécies encontradas (FISHER, 1936 *apud* REIS, 1997). O autor utilizou a técnica que constrói funções discriminantes a partir de combinações lineares das variáveis originais para classificar os grupos (HUBERTY & OLEJNIK, 2006). Assim, Fisher desenvolveu a análise discriminante como uma técnica exploratória para: (a) descrever as características que são específicas para distinguir grupos e (b), classificar casos (i.e., indivíduos, sujeitos, participantes) dentro de grupos definidos a priori baseados em suas similaridades dos casos pertencentes aos grupos (BROWN & WICKER, 2000).

A função discriminante de Fisher obtém-se maximizando a diferença de médias dos grupos em relação à variância dentro dos grupos e produz um novo eixo resultante da combinação linear das variáveis independentes que permite encontrar uma separação máxima entre grupos previamente formados. Assim, procura-se encontrar um ou mais eixos sobre os

quais se projetarão os grupos de tal modo que seja maximizada a soma de quadrados entre os grupos relativamente à soma de quadrados dentro dos grupos (REIS, 1995).

Primeiramente, admite-se que se tenha  $p$  variáveis aleatórias,  $g$  populações normais  $p$  variadas para discriminação e que as matrizes de covariâncias populacionais sejam todas iguais. O interessante é salientar que, embora a metodologia de funções discriminantes canônicas seja semelhante a de componentes principais, uma vez que ambas as funções se fundamentam na construção de combinações lineares, o processo de construção das combinações lineares é diferente e, ao contrário das componentes principais, as funções discriminantes canônicas são correlacionadas (REIS, 1995).

MINGOTI (2001) relata um exemplo ilustrativo com dois grupos para entender melhor as seqüências de passos de cálculo de uma função discriminante de Fisher. Vamos supor que temos  $p$ -variáveis aleatórias,  $g$  populações normais  $p$ -variadas para discriminação e que as matrizes de covariâncias populacionais sejam todas iguais. Neste caso, é possível construir  $s$  combinações lineares,  $s \leq \min(g-1, p)$ , chamadas de funções discriminantes canônicas, e que são definidas por:

$$\hat{Y}_j = \hat{e}_j' X_{pxl}, j=1,2,\dots,s. \quad (1)$$

Onde:  $\hat{e}_j$  é o  $j$ -ésimo autovetor correspondente ao  $j$ -ésimo maior autovalor da matriz  $\mathbf{W}^{-1}\mathbf{B}$  e tal que  $\hat{e}_j' \mathbf{W} \hat{e}_j = 1$ , sendo as matrizes  $\mathbf{W}$  e  $\mathbf{B}$  respectivamente definidas por :

$$B_{pxp} = \sum_{i=1}^g n_i \cdot (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})' \quad (2)$$

$$W_{pxp} = \sum_{i=1}^g \sum_{k=1}^{n_i} \cdot (X_{ik} - \bar{X}_i)(X_{ik} - \bar{X}_i)' \quad (3)$$

Onde:  $\mathbf{B}$  é a variação entre grupos,  $\mathbf{W}$  é a variação intra grupos,  $X_{ik}$  o vetor de observações do elemento  $k$  que pertence à população  $i$ ,  $\bar{X}_i$  o vetor de médias amostral da população  $i$ ,  $\bar{X}$  o vetor de médias amostral, considerando todas as  $n$  observações conjuntamente,  $g$  o número de

grupos e  $n_i$  o número de elementos pertencentes à amostra da população  $i$ ,  $i= 1,2,3,\dots, g$ ,

$$\sum_{i=1}^g n_i = n.$$

As matrizes  $\mathbf{W}$  e  $\mathbf{B}$  são chamadas de matrizes de soma de quadrados e produtos cruzados dentro (within) dos grupos e entre (between) grupos e são análogas aquelas utilizadas em análise de variância multivariada. Pode ser mostrado que as combinações lineares (1) são as de maior poder de discriminação dentro do conjunto de variáveis utilizadas para a discriminação. Em termos de análise de variância, são as combinações lineares com maior separação dos grupos ou, equivalentemente, as que forneceriam o menor valor da probabilidade de significância (p-valor), considerando-se os dados observados de populações normais p-variadas e o teste de Wilks (ou F) para comparação dos vetores de médias populacionais. Assim, ocorre também uma ordenação em termos da importância das combinações lineares construídas. A primeira função discriminante canônica é a relacionada com o maior autovalor da matriz  $\mathbf{W}^{-1} \mathbf{B}$  e tem o maior poder de discriminação. A segunda função discriminante é relacionada com o segundo maior autovalor de  $\mathbf{W}^{-1} \mathbf{B}$  e tem um poder de discriminação menor do que o primeiro e maior que o terceiro autovalor, e assim sucessivamente.

Após a construção das funções discriminantes canônicas, para cada elemento amostral com vetor de observações  $x_j$ , ter-se-á um vetor com os seus escores nestas funções, isto é,  $\hat{Y}_j = [\hat{e}'_1 x_j \hat{e}'_2 x_j \hat{e}'_3 x_j \dots \hat{e}'_s x_j]$ . Além disso, ter-se-á os escores das funções discriminantes canônicas aplicadas aos vetores de média amostral observados para cada população, isto é,  $\bar{Y}_i = [\hat{e}'_1 \bar{x}_j \hat{e}'_2 \bar{x}_j \hat{e}'_3 \bar{x}_j \dots \hat{e}'_s \bar{x}_j]$ ,  $i= 1,2,3,\dots,g$ . Calcula-se então a distância Euclidiana entre os vetores  $\hat{Y}_j$  e  $\bar{Y}_i$  para todo  $i= 1,2,3,\dots,g$  sendo o correspondente elemento amostral  $j$  classificado na população cuja distância é menor.

Os autovalores são determinadas a partir da matriz  $\mathbf{W}^{-1} \mathbf{B}$ . Se os autovalores são  $\lambda_1 > \lambda_2 > \lambda_3 > \dots > \lambda_n$  então  $i$  é a razão da soma dos quadrados entre grupos e a soma dos quadrados dentro dos grupos para a  $i$ -ésima combinação linear enquanto que os elementos do correspondente autovetor  $a'_i = (a_{i1}, a_{i2}, a_{i3}, \dots, a_{in})$  são os coeficientes das variáveis  $X$  para este índice. Nesse método espera-se que as primeiras funções lineares sejam capazes de encontrar as principais diferenças entre os grupos. Existem diversos cálculos e procedimentos envolvidos durante a formação de um modelo discriminante. No entanto, esse estudo irá apresentar as equações e seus significados dos procedimentos realizados exclusivamente pelo

Software SPSS sem entrar em detalhes nas diferentes formas de análise. Para maiores informações sobre os procedimentos fundamentais para o desenvolvimento das técnicas multivariadas como operações com matrizes, as operações para sumarização dos dados, cálculos de distâncias, operações lineares, as probabilidades envolvidas podem ser encontradas descritas sucintamente em Johnson & Wichern (1999), Harris (2001), Rencher (2002) e Huberty & Olejnik (2006).

## **2.2 A fisionomia e a estrutura florestal**

A vegetação pode ser definida como uma associação de plantas crescendo juntas em um local em particular e que pode ser caracterizada pelas espécies componentes e as suas características estruturais e funcionais que caracterizam a sua aparência (ou fisionomia) de determinado tipo de comunidade vegetal (GOLDSMITH & HARRISON, 1976). A caracterização fisionômica inclui a descrição das formas de vida das plantas, a composição florística, a periodicidade foliar, etc., e são baseadas principalmente em critérios qualitativos ou empíricos (GOLDSMITH & HARRISON, 1976). A grande maioria das classificações da vegetação existentes foi realizada utilizando procedimentos e análises fisionômicas (não quantitativa restrita), o que não permite estabelecer uma certeza estatística dessas classificações. As características estruturais, por outro lado, são baseadas em critérios quantitativos e referem-se ao arranjo horizontal e vertical dos componentes da vegetação, como: a densidade de plantas, a altura das plantas, a riqueza de espécies, diversidade e a equabilidade da comunidade, a estratificação da floresta, a área basal, entre outros. A análise estrutural emprega critérios de conteúdo mensuráveis quantitativos utilizando diversos parâmetros estatísticos ou matemáticos (MULLER-DOMBOIS & ELLENBERG, 1974; PIELOU, 1975; MAGURRAN, 1988).

### **2.2.1 O Bioma Pampa e os tipos de fisionomias florestais**

O Pampa é um bioma predominantemente campestre e abrange a maior parte do estado do Rio Grande do Sul (63%), a totalidade do território do Uruguai e a parte central da Argentina (IBGE, 2004). Dentre os demais biomas brasileiros, esse destaca-se por ser o segundo em menor área ocupada na federação e o único localizado totalmente em clima sub-



tropical e temperado (IBGE, 2004; PARUELO *et al.*, 1995). Em nível de conservação apresenta o menor percentual de área protegida (COUTO, 2005) e um dos mais carentes em informações científicas o que dificulta a implantação de medidas eficazes para conservação e uso sustentável desses recursos (MMA, 2002). Segundo os estudos descritivos fisionômicos/ambientais existentes (RAMBO, 1956; REITZ *et al.*, 1983; KLEIN, 1984; WAECHTER, 1985; IBGE, 2004) podem ser classificados oito principais tipos de florestas no Pampa sul-riograndense: a) as florestas de galeria, associadas aos riachos ou rios de pequeno porte; b) as florestas ciliares, associadas aos rios de grande porte; c) as florestas de restingas arenosas, associadas ao solo arenoso e drenado da região costeira; d) as florestas de restingas paludosas associadas ao solo alagável da região costeira; e) as formações arbóreas savânicas da fronteira oeste (espinal); f) as formações dos palmares, g) as florestas estacionais decíduais, representadas por capões de mata espalhadas em toda a abrangência do bioma e, h) as florestas estacionais semidecíduais da região das escarpas da Serra dos Tapes. Existem estudos que caracterizam a estrutura de alguns desses diferentes tipos de florestas nessa região (MARCHIORI *et al.* 1984, 1985; LONGHI, 1987, LONGHI, 1991; WAECHTER & JARENKOW, 1998; SEMA/UFSM, 2002; JURINITZ & JARENKOW, 2003; DORNELLES & WAECHTER, 2004; HACK *et al.*, 2005; ROSA *et al.*, 2008; DE MARCHI & JARENKOW, 2008), porém, não existem trabalhos comparativos que busquem comprovar as diferenças entre os atributos estruturais desses tipos florestais.

### 2.2.3 Gestão sustentável das florestas

Nas últimas décadas aumentou a necessidade do estabelecimento de mecanismos de valorização econômica da diversidade biológica. Na ECO/92, a Convenção sobre Diversidade Biológica (CDB) deixou clara a necessidade urgente de definição de tais mecanismos. O artigo 11 da CDB dispôs sobre a importância de instrumentos econômicos virem a ser utilizados na gestão da conservação, tendo determinado aos países signatários a adoção, conforme o caso, de medidas econômica e socialmente racionais, que sirvam de incentivo à conservação e utilização sustentável de componentes da diversidade biológica.

Assim, a CDB busca demonstrar, como estratégia de proteção à biodiversidade, que a conservação e o uso sustentável da biodiversidade têm valor econômico e que a utilização de

critérios econômicos é relevante na sua implementação, ou seja, que é imprescindível o reconhecimento do valor econômico da biodiversidade por aqueles que participam de sua gestão (IBAMA 2001). A inclusão dos serviços coletivos pelo ecossistema florestal na classe dos bens públicos - principalmente aqueles relativos à qualidade e quantidade de água e à filtragem do ar atmosférico - poderá contribuir de forma surpreendente na valoração dos ecossistemas. Todavia, na ausência de um mercado estabelecido para esses serviços, ainda não é possível sua valoração e conseqüente “monetização”, através da fixação de um preço a ser pago por eles. No caso específico dos ecossistemas florestais, a sua valoração requer a determinação de um conjunto de procedimentos, tendo em vista a multiplicidade de bens e funções existentes e possíveis em um hectare de floresta.

Para que um processo de valorização dos ecossistemas seja colocado em prática é fundamental estudos que busquem melhor defini-los principalmente quanto as suas características similares e distintas. Um dos métodos utilizados é a realização de inventários bióticos nesses ecossistemas para contribuir à gestão dos recursos naturais proporcionados pelos ecossistemas (FAO, 2000). Nesse sentido, cresce o interesse mundial em investigar e divulgar os benefícios das florestas, principalmente nos países em desenvolvimento que apresentam grandes superfícies cobertas com essas formações (ITTO 2002, FAO, 2003). Sem dúvida, elaborar planos de manejo e conservação desses recursos naturais representam um dos maiores desafios para políticos e cientistas nesse século, visto a sua reconhecida importância para mitigar os efeitos das mudanças globais como perda da diversidade, alteração climática, fonte de alimentação, saúde e renda para a população humana (SMITH *et al.*, 1997; ITTO, 2002).

### 3 MATERIAL E MÉTODOS

#### 3.1 Áreas de estudo

Os tipos de vegetação florestal escolhidos para este estudo se localizam em diferentes regiões do Pampa do Rio Grande do Sul (Figura 1). Os tipos de florestas escolhidos para esse trabalho foram: uma mata galeria, uma mata ciliar, uma floresta de restinga arenosa, uma floresta estacional semidecidual e uma vegetação savânica arbórea (espinal) (Figura 2). A maioria das fitofisionomias estão localizadas em unidades de conservação e apresentam bom estado de conservação e preservação de suas características fisionômicas originais (Tabela 2). Essas florestas foram escolhidas do banco de dados do Inventário Florestal Contínuo do Estado do Rio Grande do Sul – IFC-RS (SEMA/UFMS, 2002).

O clima no Pampa Sul-Riograndense do tipo Cfa, na escala de Köppen, caracterizado como subtropical com chuvas bem distribuídas durante o ano e verões quentes, temperaturas médias anuais de 18°C e uma pluviosidade de 1.500mm. Na sub-região da Campanha, onde ocorrem a vegetação de savana e a floresta de galeria, o relevo é predominantemente aplainado e entalhado em derrame basáltico. Na região da Depressão Central, onde ocorre a floresta ciliar, ocorrem terrenos da era paleozóica, a topografia suave ondulada com pequena altitude em relação ao nível do mar (menos de cem metros). Na região do Escudo Sul-riograndense, com rochas ígneas do período Pré-Cambriano, a altitude não ultrapassa os 600m.n.m e o relevo é ondulado. A planície litorânea é constituída pela justaposição de cordões litorâneos (restingas), com topografia plana, baixa altitude (até 30m) e solos predominantemente arenosos. Maiores informações sobre as florestas estão expressas na Tabela 1.

Tabela 1 - Localização e características ambientais dos cinco tipos de florestas subtropicais no bioma Pampa do Rio Grande do Sul.

Cód.Fr.	Fitofisionomia	Municípios	Coordenadas	Área(ha)	Alt.(m)
1108	Floresta de galeria <sup>1</sup>	Alegrete	29°55'S e 55°46'O	150	100
1333	Floresta ciliar <sup>2</sup>	Santa Maria	29°58'S e 53°47'O	60	110
1603	Savana lenhosa <sup>3</sup>	Barra Quaraí	30°11'S e 57°27'O	96	50
1944	Floresta estacional <sup>4</sup>	Encruzilhada do	30°59'S e 52°34'O	804	61
2808	Floresta de restinga	Rio Grande	32°33'S e 52°45'	10	24

Cód. Fr.= Código do fragmento; Alt. (altitude em metros acima do nível do mar) 1. Reserva Biológica do Ibirapuitã; 2. Área de Preservação Permanente; 3. Parque Estadual do Quaraí; 4. Propriedade particular. 5. Estação Ecológica do Taim. Área = área de floresta remanescente; Alt= altitude (metros acima do nível do mar). Todos os dados foram procedentes do IFC-RS (SEMA/UFSM, 2002).



Figura 1. Abrangência do bioma Pampa nos três países da América do Sul e localização dos cinco tipos de florestas no Pampa do estado do Rio Grande do Sul.

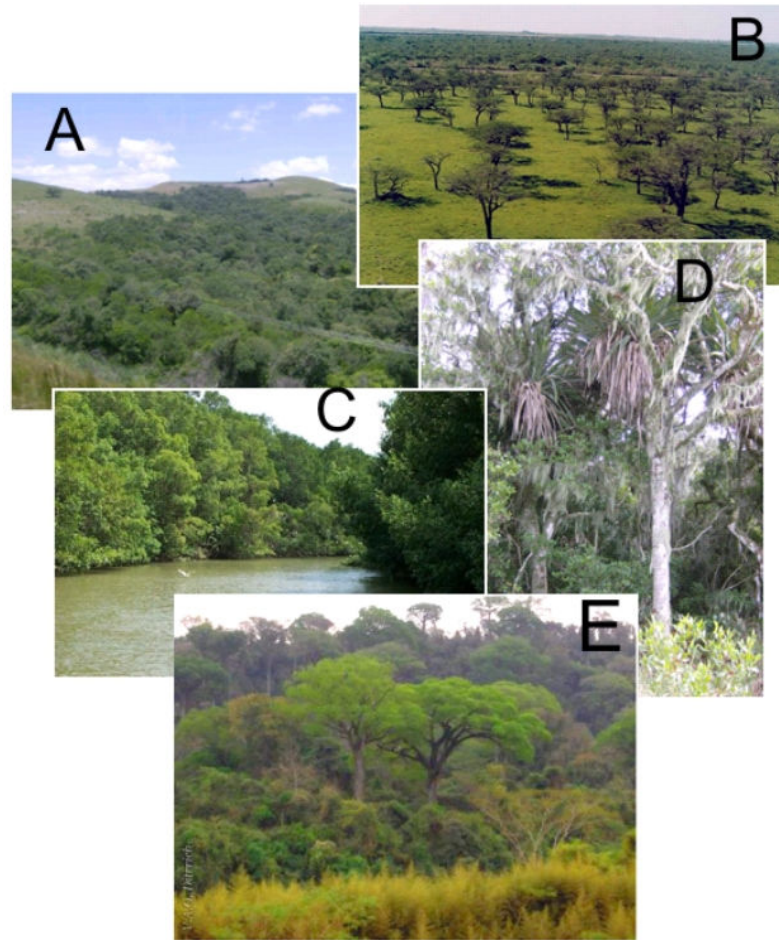


Figura 2. Fotografias da fisionomia dos cinco tipos de florestas subtropicais analisadas no bioma Pampa do Rio Grande do Sul: floresta de galeria (A), savanas do sudoeste (B), floresta ciliar (C) e floresta litorânea arenosa (D) e floresta estacional semidecidual (E).

### 3.2 Delineamento amostral da vegetação

Todas as fitofisionomias receberam o mesmo critério de amostragem da vegetação conforme o programa do Inventário Florestal Contínuo do Estado do Rio Grande do Sul (SEMA/UFSM 2002). A forma de amostragem da vegetação utilizou uma parcela permanente da forma quadrada, com 100 m de largura por 100 m de comprimento (10.000 m<sup>2</sup> de superfície ou 1ha), a qual foi dividida em 10 faixas de 10 m de largura por 100 m de comprimento (1.000 m<sup>2</sup> de superfície), e estas subdivididas em 10 subunidades de 10 m x 10 m (100 m<sup>2</sup>) totalizando 100 subunidades (Figura 3).

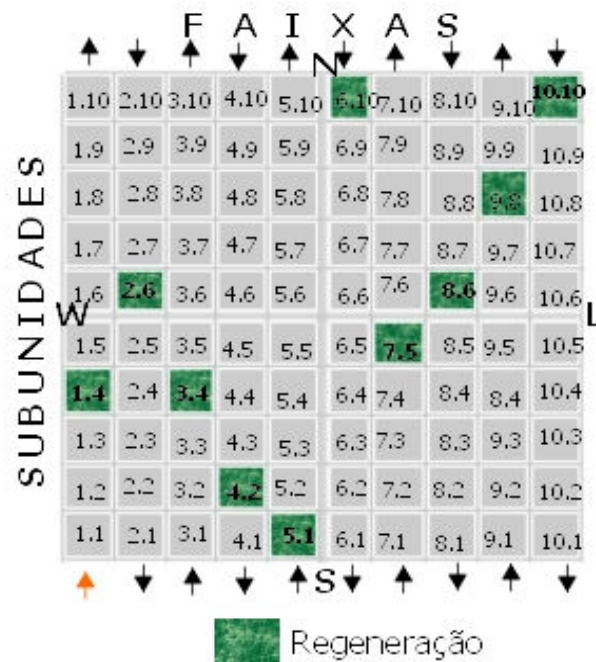


Figura 3. Modelo do croqui do arranjo das 100 unidades amostrais (10 x 10 m) instalados nos cinco tipos de florestas do Pampa no Rio Grande do Sul (Fonte da figura SEMA/UFSM, 2002).

Em cada unidade amostral todos os indivíduos arbóreos vivos, exceto lianas, com diâmetro do caule igual ou superior à 5 cm numa altura 1,3 m do solo (DAP) tiverem medidos seus diâmetros dos caules e suas alturas totais e inseridos na amostragem. Maiores detalhes sobre o processo de amostragem estão disponibilizadas no site do IFC-RS (<http://coralx.ufsm.br/ifcrs/frame.htm>).

Para a realização da análise discriminante, primeiramente o banco de dados originais sofreu alterações para que comportasse alguns pré-requisitos necessários para a análise discriminante. As 100 unidades amostrais (100m<sup>2</sup>) em cada tipo de floresta foram separadas em agrupamentos de 10 unidades amostrais (1000m<sup>2</sup>), uma “faixa” como representada na Figura 3, antes de serem calculados os parâmetros dos atributos estruturais (veja a seguir quais os atributos). Assim, cada floresta (grupo), permaneceu com 10 unidades amostrais (casos) o que, levou a um banco de dados para a amostra discriminante com 50 casos (10 em cada grupo) e oito variáveis independentes (variáveis preditoras) (Tabela 2, ANEXO A). Dessa forma, cumpriu-se uma exigência de tamanho amostral do modelo discriminante com uma proporção de 10 casos por grupo/variável. Admiti-se no presente estudo de que um

hectare de floresta pode apresentar uma considerável variação estrutural. Essa variação seria similar a que ocorre em outras florestas que pertencem a mesma fitofisionomia, e estão sobre similares condições de preservação e sítio ambiental.

Tabela 2 - Banco de dados com o tamanho amostral inicial e modificado para a realização da análise discriminante linear para o conjunto das amostras dos cinco tipos de florestas subtropicais no Pampa do Rio Grande do Sul.

Cód. flo.	Banco de dados original				Banco de dados modificado			
	Área amostral total (m <sup>2</sup> )	Nº de UA	Tamanho UA (m <sup>2</sup> )	Nº casos amostrados	Área amostral total (m <sup>2</sup> )	Nº UA amostrados	Tamanho UA (m <sup>2</sup> )	N casos
1	10.000	100	100	1351	10.000	10	1.000	10
2	10.000	100	100	904	10.000	10	1.000	10
3	10.000	100	100	141	10.000	10	1.000	10
4	10.000	100	100	870	10.000	10	1.000	10
5	10.000	100	100	730	10.000	10	1.000	10

Código da floresta (Cód. flo.): UA (unidade amostral), Floresta de galeria (1), Floresta ciliar (2), Savana lenhosa (3), Floresta estacional semidecidual (4) e Floresta de restinga arenosa (5). Banco de dados originais = estrutura do banco de dados originais no IFC-RS. BD transformados- transformação no banco de dados originais para realização do modelo discriminante.

Na etapa de validação do modelo, foram escolhidos mais 10 casos que corresponderam a duas unidades amostrais (1.000m<sup>2</sup>) de cada um dos cinco tipos florestais originais: floresta de galeria, floresta ciliar, savana lenhosa, floresta estacional semidecidual e floresta de restinga arenosa. Essas unidades amostrais foram escolhidas dentro do banco de dados originais do IFC-RS. O critério seguiu a revisão das características ambientais de cada parcela observando a condição edáfica e o estágio sucessional similar com aquelas encontradas nos grupos originais (Tabela 2). Os atributos estruturais foram calculados de maneira semelhante com a análise dos grupos originais. Os valores dos atributos estruturais em cada uma das duas unidades amostrais de cada grupo foram apresentados no Anexo (ANEXO B).

Do banco do IFC-RS foram utilizados, para cada tipo de floresta analisada, oito variáveis descritoras da estrutura das comunidades: o número de espécies, o número de indivíduos, a altura total média, a altura máxima, a altura comercial média, o DAP médio, o DAP máximo e a área basal. Cada um dos atributos foi assim especificado:

- 1) O número de espécies: o número total de espécies arbóreas contabilizados por área ( $1000\text{m}^2$ );
- 2) O número de indivíduos corresponde ao total de indivíduos arbóreas contabilizados por área amostral ( $1000\text{m}^2$ );
- 3) A altura total média (em metros) significa a média dos tamanhos totais dos indivíduos por área amostral ( $1000\text{m}^2$ );
- 4) A altura máxima (em metros) demonstra o tamanho máximo que um indivíduo arbóreo atingiu na área amostral ( $1000\text{m}^2$ );
- 5) A altura comercial média (em metros) é a altura do tronco, que corresponde ao ponto de corte na base de cada árvore até a primeira bifurcação dos seus galhos na área amostral ( $1000\text{m}^2$ );
- 6) O DAP (diâmetro do caule medido a uma altura de 1,3 m do solo) médio significa o valor médio de DAP (medido em cm) para todos os indivíduos da área amostral ( $1000\text{m}^2$ );
- 7) O DAP máximo significa o valor máximo de DAP medido para um indivíduo arbóreo na área amostral ( $1000\text{m}^2$ );
- 8) A área basal (soma das medidas de todas as áreas transversais dos troncos) total de todos os indivíduos amostrados por área amostral ( $1000\text{m}^2$ ).

Os cálculos para a obtenção da área basal e DAP estão descritas no site do IFC-RS (<http://coralx.ufsm.br/ifcrs/frame.htm>).

### **3.3 Delineamento Estatístico**

Primeiramente foram considerados como grupos do modelo discriminante (as variáveis dependentes) os tipos de vegetação, definidos por meio de estudos prévios de vegetação que reconhecem cada tipo de vegetação como um grupo distinto fisionomicamente. As variáveis independentes ou descritoras do modelo foram as oito variáveis descritas anteriormente. Os valores médios (DAP, altura média total, altura média comercial) ou os



absolutos (totais de riqueza de espécies, número de indivíduos, área basal e altura máxima) dos atributos não foram padronizados antes do início da análise porque o resultado de uma função discriminante não é afetado de nenhuma forma importante pelo escalonamento de variáveis individuais (MANLY, 2008). Além disso, o teste estatístico para normalidade multivariada não é disponível no SPSS, bem como, na maioria dos softwares estatísticos.

Para a realização da análise discriminante múltipla linear de Fisher foi utilizado o Software SPSS. Os passos, ou melhor, os comandos utilizados para a realização a análise no SPSS foram:

### 3.3.1 Passos desenvolvidos no SPSS programa DISCRIMINANT:

- A. ANALYZE/CLASSIFY – DISCRIMINANT
- B. A primeira opção no *prompt* requer a inserção das variáveis dependentes (*Grouping Variable* – os quatro tipos de florestas) e independentes do modelo (*Independents* - oito atributos da estrutura das florestas) e escolher a opção *Use Stepwise Method* (método passo-a-passo – que seleciona as variáveis independentes com melhor grau de discriminação).
- C. Definir os itens dos sub-menus STATISTICS, METHOD e CLASSIFY. No primeiro item (STATISTICS) são determinadas algumas estatísticas descritivas (*Means*) e o teste de premissa de igualdade da covariância entre grupos denominado teste M de Box e ANOVA (*Box's M, Univariate ANOVAs*). Definem-se os tipos de coeficientes da função que se deseja obter (*Fisher's*) e as matrizes de correlação e co-variância (*Withing-groups correlation, Withing groups covariance*). No item METHOD, escolher a opção de inserção de variáveis na função discriminante que é o método de lambda de Wilks (*Wilk's lambda*). O critério para a seleção das variáveis no modo *Stepwise* se dá através da escolha da probabilidade em que a variável independente é inserida ou removida do modelo. Esse nível de significância é definido na mesma caixa na opção *Use probability of F* e foi estabelecido o padrão de intervalo de 5 a 10% de significância. A opção que mostra cada passo da formação da função discriminante também foi selecionada (*Summary of steps*).

D. No último menu (CLASSIFY) se obtém os passos para a obtenção da classificação dos grupos a partir da função obtida. Assim, a probabilidade *a priori* informa quais devem ser os pesos de cada grupo nesse processo. Como todos os grupos têm o mesmo número de casos (n=10) optou-se pela opção (*All groups equal*). O método e a classificação dos casos foram calculados usando a matriz de covariância (*Withing-groups*). A tabela dos coeficientes para cada caso e a distância Mahalanobis dos casos em relação ao centróide de grupo pode ser observada na opção *Casewise Results*. A tabela de acertos e erros ocorridos na classificação foi demarcada na opção *Summary Table* e o tipo de gráfico a ser utilizado para demonstrar a classificação foi o de grupos combinados (*Combined Groups*) onde se pode observar a distâncias dos centróides dos quatro grupos (e de seus casos) em um mesmo gráfico.

Outras análises estatísticas complementares como testes ANOVA de um fator, testes de normalidade de dados, homogeneidade de variâncias foram desenvolvidas no software Statistica 7.0 (STATSOFT, 2006) (ver ANEXOS C, D e E).

Na etapa de validação dos novos casos, os 10 casos que corresponderam os atributos estruturais em duas unidades amostrais (com 1.000m<sup>2</sup> cada) de cada grupo foram inseridos no banco de dados como um grupo à parte (ANEXO B). Uma nova análise discriminante foi gerada no SPSS com 60 casos onde apenas os valores da etapa de validação cruzada dos casos no grupo teste foram apresentados com os testes de distância de Mahalanobis. Assim, esperou que os casos desse novo grupo fossem re-agrupados pelo modelo em seus verdadeiros grupos determinados *à priori*. O modelo no SPSS foi gerado com os mesmos procedimentos descrito anteriormente (item 3.2.1).

### 3.3.2 Os cálculos e o significado dos testes utilizados no SPSS

Conforme os passos descritos anteriormente no software SPSS, seguiram uma série de cálculos estatísticos que são apresentados em três etapas da análise discriminante: validação, cálculo das funções discriminantes e classificação. Cada etapa apresenta uma série de testes para validar o modelo que é apresentado na forma de *outputs*. Todos os significados desses testes estatísticos nos *outputs* são apresentados a seguir:

### 3.3.2.1 Etapa de validação:

Análise descritiva dos dados: Disponibiliza os valores das médias, com seus respectivos desvios padrão, de cada variável nos grupos. A análise também lista o número de observações que foram excluídas dentro de cada grupo e também avalia se o peso do número de observações em cada grupo é igual ao número de observações em cada grupo.

A média é calculada pela equação:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (4)$$

O desvio-padrão é calculado pela equação:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (5)$$

Matriz de correlação: O SPSS oferece a opção de análise da matriz de correlação entre as variáveis preditoras (SPSS, 2006). A verificação da matriz de correlações possibilita identificar prováveis casos de multicolinearidade que podem afetar os cálculos das cargas das variáveis individuais na função discriminante. As variáveis altamente correlacionadas não são boas preditoras e podem contribuir de forma redundante para o modelo. A correlação de Pearson é calculada como uma razão da covariância,  $Cov_j(Y_1Y_2)$ , do produto dos desvios padrão através da seguinte equação:

$$r_{jk} = \frac{\sum_{i=1}^n (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)}{\sqrt{\sum_{i=1}^n (y_{ij} - \bar{y}_j)^2 \sum_{i=1}^n (y_{ik} - \bar{y}_k)^2}} \quad (6)$$

Teste de igualdade de médias dos grupos: Quando há diversas variáveis e muitas amostras existem quatro estatísticas que são comumente usadas no SPSS para testar a hipótese de que

todas as amostras provêm de populações com o mesmo vetor médio. O critério do lambda de Wilks utiliza a razão  $F$  multivariada para testar as diferenças entre os grupos. Nesse caso a estatística selecionada foi de lambda de Wilks que é dada pela seguinte equação:

$$\Lambda = \frac{|W|}{|T|} \quad (7)$$

Onde:

$|W|$  é o determinante da matriz das somas de quadrados e de produtos cruzados dentro da amostra, e

$|T|$  é o determinante da matriz das somas totais de quadrados e de produtos cruzados.

Essencialmente, essa equação compara a variação dentro das amostras com a variação dentro e entre as amostras. Como uma informação complementar sobre as matrizes  $W$  e  $T$ , denota-se  $X_{ijk}$  como sendo o valor da variável  $X_k$  para o  $i$ -ésimo indivíduo e a  $j$ -ésima amostra; denota-se  $\bar{x}_{jk}$  como a média da variável  $K_k$  na mesma amostra; denota-se  $\bar{x}_k$  como a média global de  $X_k$  para todos os dados tomados juntos. Além disso, assume-se que há  $m$  amostras, com a  $j$ -ésima de tamanho  $n_j$ . Então o elemento na linha  $r$  e a coluna  $c$  de  $T$  é:

$$t_{rc} = \sum_{j=1}^m \sum_{i=1}^{n_j} (x_{ijr} - \bar{x}_r)(x_{ijc} - \bar{x}_c) \quad (8)$$

E o elemento na linha  $r$  e na coluna  $c$  de  $W$  é:

$$c_{rc} = \sum_{j=1}^m \sum_{i=1}^{n_j} (x_{ijr} - \bar{x}_{jr})(x_{ijc} - \bar{x}_{jc}) \quad (9)$$

Se  $\Lambda$  tem valor pequeno, este indica que a variação dentro das amostras é baixa em comparação com a variação total. Isso fornece a evidência de que as amostras não vêm de populações com o mesmo vetor de médias. Assim, quanto mais próxima de zero for o valor de lambda de Wilks, mais a média dos grupos diferem (mais a variável independente difere nos grupos) e, quanto mais próximo do valor 1, mais a média dos grupos se assemelham (o que torna os grupos mais semelhantes).

Nessa mesma análise o SPSS realiza uma análise de variância de um fator (*one-way ANOVA*) para testar se existe diferença na média de cada variável em pelo menos um dos grupos. O detalhe desse teste também foi demonstrado no ANEXO E. Esse teste ANOVA leva em conta a hipótese de que as variâncias populacionais sejam iguais e a que as variâncias totais das observações em torno da média geral  $\bar{x}$  podem ser representadas por  $\sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x})^2$ , tem-se que  $SQ_{Total} = SQ_{Tratamentos} + SQ_{Erro}$  ou *variação total = variação entre amostras + variação dentro das amostras*:

$$\sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x})^2 = n \cdot \sum_{i=1}^k (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 \quad (10)$$

A variância total ( $S_t^2$ ) é estimada a partir da hipótese de que as variâncias populacionais são iguais a  $\sigma^2$ :

$$S_t^2 = \frac{\sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x})^2}{N-1} \quad (11)$$

Onde deve-se supor que todas as médias sejam normalmente distribuídas e de mesma média e mesma variância.

A variância residual (ou entre amostras),  $S_r$ , dividida pelos seus graus de liberdade ( $N-K$ ) é uma estimativa não viesada de  $\sigma^2$  independentes se  $H_0$  for verdadeira:

$$S_r^2 = \frac{\sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x})^2}{N-K} \quad (12)$$

A variação entre amostras ( $S_e$ ) dividida pelos seus graus de liberdade ( $K-1$ ) é uma estimativa não viesada de  $\sigma^2$  apenas se  $H_0$  for verdadeira:

$$S_e^2 = n \cdot \frac{\sum_{i=1}^k (\bar{x}_i - \bar{x})^2}{K-1} \quad (13)$$

Desta forma, pode-se comparar as duas estimativas da variância através de um teste  $F$ . A tabela de classificação do ANOVA encontra-se a seguir:

Tabela 3 - Critério para a análise de variância (ANOVA) para um fator.

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	Estatística $F$
Entre amostras	$S_e$	$K-1$	$S_e^2 = \frac{S_e}{K-1}$	$F = \frac{S_e^2}{S_r^2}$
Dentro da amostra	$S_r$	$N-K$	$S_r^2 = \frac{S_r}{N-K}$	
Total	$S_t$	$N-1$		

Teste de igualdade de matrizes de covariâncias: O SPSS disponibiliza os cálculos dos coeficientes das matrizes para as variáveis independentes através da matriz de correlação entre os grupos, da matriz de covariância entre grupos, da matriz de covariância dos grupos separados e da matriz de covariância total (SPSS, 2006). O SPSS usa o Teste  $M$  de Box que é sensível a normalidade multivariada (SPSS, 2006). G.E.P Box propôs em 1949 um procedimento estatístico para testar a hipótese nula de que duas ou mais matrizes de covariância são iguais:  $H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_J$ . A equação do teste é:

$$M = \left\{ \prod_{i=1}^m |C_i|^{(n_i-1)/2} \right\} / |C|^{(n-m)/2} \quad (14)$$

Em que  $n_i$  é o tamanho da  $i$ -ésima amostra,  $C_i$  é a covariância amostral para a  $i$ -ésima amostra e  $C$  é a matriz de covariância combinada:

$$C = \sum_{i=1}^m (n_i-1) \cdot C_i \quad (15)$$

E  $\Sigma_n$  é o número total de observações.

Grandes valores de  $M$  fornecem evidência de que as amostras não provêm de populações com a mesma matriz de covariâncias. Um teste  $F$  aproximado é utilizado para saber se um valor  $M$  observado é significante. Sua fórmula é:

$$F = -2b \cdot \log(M) \quad (16)$$

Onde:

$$b = (1 - c_1 - v_1 - v_2) / v_1 \quad (17)$$

Encontrando a probabilidade de uma valor desse tamanho ou maior para uma distribuição  $F$  com  $v_1$  e  $v_2$  graus de liberdade, em que  $v_1 = p(p+1)(m+1)/2$  e  $v_2 = (v_1+2)/(c_2-c_1^2)$

sendo:

$$C_1 = (2p^2 + 3p - 1) \left\{ \sum_{i=1}^m 1/(n_i - 1) - 1/(n - m) \right\} / \{6(p+1)(m-1)\} \quad (18)$$

$$C_2 = (p-1)(p+2) \left\{ \sum_{i=1}^m 1/(n_i - 1)^2 - 1/(n - m)^2 \right\} / \{6(m-1)\} \quad (19)$$

Uma razão  $F$  significativa evidencia que as amostras vêm de populações com desvios médios diferentes, que significa populações com matrizes de covariâncias diferentes. A hipótese nula indica o contrário.

### 3.3.2.2 Etapa do cálculo das funções discriminantes

Seleção das variáveis: Quando numerosas variáveis são inseridas no modelo, a análise passo-a-passo (*Stepwise*) pode ser usada para identificar o melhor subconjunto de variáveis para classificar os objetos no modelo discriminante. A escolha da opção classificação de Fisher no SPSS fornece os coeficientes da função de classificação que podem ser usados diretamente na classificação dos grupos. O conjunto dos coeficientes é obtido para cada grupo e o caso é obtido do grupo para o qual ele teve o maior escore (SPSS, 2006). O SPSS utiliza diversas estatísticas, nesse caso foi selecionado a estatísticas lambda de Wilks. O SPSS oferece a opção de escolha do nível de significância do teste  $F$  para a variável entrar no modelo. A variável é inserida no modelo se o nível de significância foi menor que o valor  $F$  de “entrada” e é removida se o nível de significância é maior que o valor de remoção que, portanto, indica a contribuição daquela variável para a predição dos grupos. Através dessa estatística na análise discriminante *Stepwise*, as variáveis são inseridas com base no menor valor de lambda

de Wilks (SPSS, 2006). Esse método *Stepwise* é um procedimento similar com o que ocorre na regressão múltipla. Os dois modos de seleção passo-a-passo foram definidos a seguir.

No método *Forward*, ou modo de inclusão avançada, inicialmente não há variáveis prognosticadas e essas vão sendo introduzidas uma de cada vez de acordo com os critérios definidos pela estatística  $F$  no modelo de regressão múltipla. Entra primeiramente no modelo aquela variável com maior poder de discriminação entre grupos (com maior significância para  $F$ ). Esta variável é pareada com as demais, uma de cada vez, e o critério de seleção é novamente comparado. A estatística lambda de Wilks seleciona as variáveis com maior valor no teste, que representa melhor poder de discriminação. A variável que em conjunto com a primeira (co-variável) fornecer o melhor poder de discriminação é escolhida e, assim, o mesmo método irá se repetir nos próximos passos até que todas as variáveis independentes sejam ordenadas pelas suas capacidades de discriminação até que a adição de uma nova variável não significaria numa melhor discriminação entre grupos. As variáveis redundantes (com elevado correlação entre si) tendem a ser eliminadas e não entram no modelo.

O método *Backward*, ou modo de eliminação retrógrada, inicia com a inclusão de todas as variáveis no modelo de regressão múltipla. Basicamente ela funciona com o reverso da análise anterior (*Forward*). A variável com menor lambda de Wilks é a melhor candidata a ser inserida no modelo no próximo passo. A estatística  $F$  calculada para cada variável refere-se ao valor  $F$  observado for menor que o valor  $F$  de referência escolhido para a saída de variáveis. Em cada passo do algoritmo sairá a variável de menor significância. A estatística  $F$  que remove as variáveis do modelo (método *Backward*) no SPSS é:

A equação para o cálculo do lambda de Wilks é:

$$\Lambda = \prod_{i=1}^p 1 / (1 + \lambda_i) \quad (23)$$

e a equação para o teste  $F$  é:

$$F_{(i)} = \frac{1 - \Lambda / \Lambda_{(i)}}{\Lambda / \Lambda_{(i)}} \cdot \frac{gl_e - p}{J - 1} \quad (20)$$

Nos dois critérios a variável para a seleção é considerada apenas se a razão  $F$  multivariada parcial for maior que o valor específico. Se o  $F$  apresentar valor pequeno a variável não é



selecionada no modelo pois estaria relacionada com suposta igualdade entre os grupos centróides. A variável que maximiza a razão  $F$  também minimiza o lambda de Wilks. Assim pode-se considerar que existe diferença em todos os centróides de grupos e homogeneidade dentro do grupo. A ordem de seleção das variáveis nos critérios *Forward* e *Backward* não implica obrigatoriamente que estas tenham uma ordem de importância relativa como as melhores discriminatórias.

Apresentação das funções discriminantes: esse teste indica todas as funções lineares discriminantes geradas pelo modelo, geralmente o número de funções é igual ao número de grupos menos um ( $g-1$ ). Cada função atua como projeção dos dados dentro das dimensões que melhor separam ou discriminam entre os grupos e ainda essas funções são independentes ou ortogonais. Cada função discriminante é representada por um autovalor que também mede a importância relativa da função ou a importância da dimensão da classificação dos casos na variável dependente. Os autovalores refletem a porcentagem da variância explicada pela variável dependente e por isso, indicam o poder discriminatório da função discriminante. As magnitudes dos autovalores também estão relacionadas com as correlações canônicas e com a proporção da habilidade discriminante de cada função gerada. A correlação canônica ( $R$ ) é uma medida da associação entre as variáveis preditoras e os grupos. Quando o valor  $R$  é baixo indica pouca relação entre os grupos e a função e, quando é alto, indica o contrário. Por isso, também é usada para demonstrar quanto que cada função é útil para determinar a diferença entre os grupos.

A idéia matemática é determinar um conjunto de pesos para as variáveis independentes do modelo baseadas na correlação dessas variáveis que podem ser resumidas em uma variável latente  $Z$ . Onde  $Z$  é determinada pela combinação linear das variáveis  $X_{1...p}$  conforme a seguinte equação:

$$Z = b_1X_1 + b_2X_2 + \dots + b_pX_p + c, \quad (21)$$

Onde:  $b$  são os valores dos coeficientes discriminantes que maximizam as distâncias entre as médias das variáveis dependentes,  $X$  são os valores das variáveis discriminantes e  $c$  é uma constante calculada como:

$$c = -\sum b_p \bar{X}_p \quad (22)$$

onde:  $\bar{X}$  é a média da variável  $X_i$  através de todas os  $N$  casos.

O conjunto dos pesos que determinam o  $Z$  (os escores  $Z$ ) é a base para a máxima separação dos grupos. Os escores  $b_p$  são encontrados através dos cálculos dos autovalores. Os autovalores indicam a razão entre a variabilidade entre grupos e entre grupos para uma determinada função. Um maior autovalor representa a maior diferença de grupos devido as cargas das variáveis independentes na função. Seja,  $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_p \geq 0$  originados da matriz  $\mathbf{W}^{-1} \mathbf{B}$ , (que é a generalização da equação SSH/SSE) que são posteriormente obtidos da matriz  $\mathbf{B}=\mathbf{T}-\mathbf{W}$  é chamada a matriz entre amostras de somas de quadrados e produtos cruzados, porque a entrada típica é a diferença entre uma soma total de quadrados ou produtos cruzados menos o termo correspondente dentro das amostras.

O cálculo da importância relativa de cada função discriminante é dado pela equação:

$$\% \text{ var} = \frac{\lambda_i}{\sum_{j=1}^s \lambda_j} \quad (23)$$

Onde:

A correlação canônica ao quadrado indica a proporção média da variação que pode ser explicada pelo conjunto das variáveis. Assim, essa correlação pode ser calculada para cada função discriminante a partir dos autovalores calculados:

$$r_i^2 = \frac{\lambda_i}{1 + \lambda_i} \text{ sendo } i=1,2,3,\dots,s. \quad (24)$$

Teste das funções discriminantes: Essa análise oferece o teste da hipótese nula de que o valor da função discriminante é o mesmo para todos os grupos, ou seja, que a razão entre a variabilidade entre grupos para a variabilidade dentro dos grupos detectada na função 1 é maior que na função 2. Nesse caso, valores elevados do lambda de Wilks correspondem a menor poder de discriminação. O teste qui-quadrado geralmente está associado para testar a significância do teste lambda de Wilks. Se o valor de  $p$  é menor que alfa (0,05) a hipótese nula é rejeitada.

Para separar mais de três dimensões o teste estatístico calculado é o seguinte:

$$\Lambda = \prod_{i=4}^r 1 / (1 + \lambda_i) \quad (25)$$

Onde:  $\Lambda$  é uma medida inversa e  $\lambda_1, \lambda_2, \dots, \lambda_i$  são os autovalores da  $\mathbf{W}^{-1} \mathbf{B}$ .

$\lambda = 1,0$ : Não ocorreu nenhuma discriminação, ou seja, os grupos centróides são idênticos.

$\lambda = 0$ : Ocorreu máxima discriminação, ou seja, os grupos centróides estão separados.

O teste para avaliar se em pelo menos uma dimensão existe diferença para separar os grupos se dá pela transformação da equação anterior no teste  $F$ :

$$F = \frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \cdot \frac{m(s) - p(gl_h) / 2 + 1}{p(gl_h)} \quad (26)$$

Onde:

$$m = gl_e - \frac{p - gl_h + 1}{2} \quad (27)$$

e

$$s = \sqrt{\frac{p^2 (gl_h^2 - 4)}{p^2 (gl_h^2 - 5)}} \quad (28)$$

Sendo que a distribuição  $F$  aproximada com graus de liberdade  $v_1 = (p-3)(q-3)$  e  $v_2 = m(s) - (p-3)(q-3)/2 + 1$ .

A significância estatística do  $\Lambda$  pode ser testada pela conversão da aproximação da distribuição qui-quadrada:

$$\chi^2 = - \left[ n - \left( \frac{p+q}{2} \right) - 1 \right] \log_{ek} \quad (29)$$

Com  $(p-k) \times (g-k-1)$  os graus de liberdade

Coeficiente da função discriminante padronizada: os coeficientes correspondem aos valores  $b_p$  da equação (21). Eles são usados para computar o escore discriminante para o caso no qual a variável discriminatória original estão na forma não-padronizada. O escore discriminante é calculado pela multiplicação de cada variável discriminante pelo seu coeficiente correspondente mais o produto. Geralmente, o peso padronizado para a variável  $p$ ,  $b_p^*$  é calculado usando:

$$b_p^* = b_p \sqrt{d_{pp}} \quad (30)$$

Onde, o radicando é a média do elemento da diagonal da matriz de covariância. Duas matrizes podem ser consideradas como fontes do termo  $d_{ii}$ , a matriz total dos grupos e do erro da matriz de covariância. Quando há  $j$  funções discriminantes, cada caso tem seu escore em cada função. Pela média dos escores dos casos dentro de um grupo particular chega-se a média de um grupo de uma respectiva função (grupo centróide). Os coeficientes têm grande importância analítica e devem ser interpretados sem a importância do sinal matemático, pois indica apenas se a contribuição da variável é positiva ou negativa, de forma idêntica aos coeficientes  $\beta$  da regressão linear múltipla. Os coeficientes podem ser usados para calcular o escore discriminante de um dado caso. Portanto, os coeficientes indicam a contribuição relativa da associação da variável com a função discriminante.

Estrutura da matriz: essa é a estrutura canônica do modelo, também conhecida como cargas canônicas ou coeficientes discriminantes das funções discriminantes. Ela representa as correlações entre as variáveis observadas (variáveis discriminantes) do modelo com as funções discriminantes. Quando o sinal é ignorado, cada coeficiente representa uma prévia associação entre a variável e uma determinada função discriminante, sendo aqueles com maiores valores os que contribuem mais para o poder discriminatório da função. Os valores variam de 0 a 1 e quanto maior o valor maior é a capacidade da função discriminante conter a informação de determinada variável. Deve-se salientar que os coeficientes padronizados da função discriminante denotam a contribuição única de cada variável na respectiva função (quadro anterior) e os coeficientes da estrutura da matriz denotam simplesmente a correlação entre as variáveis e as funções.

Função dos grupos centróides: Essa análise apresenta a média dos escores discriminantes dos membros dos grupos em cada uma das funções discriminantes. Esses centróides são obtidos pela multiplicação da média dos grupos pelos pesos das funções discriminantes adicionadas à constante da equação (21). Cada grupo tem um centróide que sumariza a localização do grupo no espaço bi-dimensional definido pela função discriminante. Os escores dos centróides também são outra evidência que o pesquisador pode utilizar para constatar a diferença dos grupos.

### 3.3.2.3 Etapa de classificação

Resumo do teste de classificação: O *output* demonstra o número total de casos que foram classificados corretamente pelo modelo. Quando algum caso não é processado, ele aparece descrito nesse mesmo *output*.

Probabilidade de grupos: Mostra a distribuição das observações dentro dos grupos e se a probabilidade *a priori* foi a mesma para ajustar a classificação dos casos.

Classificação dos coeficientes da função: Essa análise descritiva apenas revela os coeficientes da função de Fisher de cada variável em seu grupo. O peso do coeficiente permite determinar em que grupo que uma determinada variável foi mais significativa.

Classificação dos casos: Uma tabela de comparação dos casos é fornecida com os valores das distâncias de Mahalanobis analisada pelo teste  $F$  e revela ao pesquisador se o caso pertence a determinado grupo de origem ou a outro grupo. Uma das formas de avaliar o ajuste da classificação é examinar os resultados preditivos em uma base caso-a-caso, com objetivo de entender quais as observações que foram mal classificadas nos membros do grupo. Os casos que são corretamente ou incorretamente classificados necessitam de uma medida de

similaridade de uma observação com o restante do grupo. Para calcular a probabilidade de um dado caso pertencer a um grupo, empregasse a distância Mahalanobis. Essa é uma medida de distância entre dois pontos no espaço multidimensional definidos por duas ou mais variáveis preditoras correlacionadas. Para cada grupo da amostra pode-se determinar a localização num ponto que representa a média de todas as variáveis inseridas no modelo em um espaço multivariado (centróide de grupo). Para cada caso também pode-se computar a distância Mahalanobis para cada centróide de grupo. A distância de Mahalanobis leva em conta a correlação entre todas as variáveis e as condições de distribuição normal multivariada e a igualdade de matrizes de covariâncias. Os dados em ambos os grupos a serem comparados deverão ter o mesmo número de variáveis (ou seja, o mesmo número de colunas) mas não necessariamente o mesmo número de elementos (o número de linhas pode ser diferente). Uma menor distância Mahalanobis coloca uma observação entre os dois centróides provavelmente teria uma menor probabilidade de classificação correta, mesmo que essa observação esteja mais próxima ao centróide de seu grupo. Em suma, seria uma estatística útil para determinar a similaridade entre uma amostra desconhecida e uma conhecida.

Assim, o índice de Mahalanobis é apropriado quando um ponto representa um vetor das  $p$  observações em uma unidade de análise e outro ponto representa o centróide para um grupo. Suponha que sejam  $j$  grupos de interesses a distância entre  $X_u$ , o vetor de observação da unidade  $u$ , e  $\mu_j$ , o centróide do grupo  $j$  pode ser assim representado:

$$D_{uj}^2 = \left[ (x_u - \mu_j)' \Sigma_j^{-1} (x_u - \mu_j) \right]^{1/2} \quad (31)$$

Onde:

$D_{uj}^2$  = Distância Mahalanobis

$x$  = vetor dos dados para os grupos ( $x_1, x_2, \dots, x_u$ )

$\mu_j$  = vetor dos valores médios das variáveis independentes

' = indica que o vetor pode ser transposto

$\Sigma^{-1}$  = Matriz de covariância inversa da população  $j$

Esse índice de distância irá classificar uma observação dentro de seu grupo ao qual está mais próximo. Assim, o caso  $u$  é classificado dentro do grupo  $j$  se  $D_{uj}^2$  for menor que  $D_{uj'}^2$  para todo  $j' \neq j$  e  $j, j' = 1, 2, 3, \dots, j$ .

Quando se considera que a população amostrada tem distribuição normal multivariada, os valores de  $D^2$  seguirão uma qui-quadrado com  $p$  graus de liberdade se  $x$  vier dessa distribuição.

A decisão de classificação ou alocação é baseada no princípio da máxima probabilidade onde ao caso é alocado no grupo onde seu vetor tem maior probabilidade de ocorrer. Dessa forma, assume-se que todas as  $j$  populações tenham normalidade multivariada. Seja  $f$  que denote essa função densidade comum, então a regra da máxima probabilidade irá alocar a unidade  $u$  no grupo  $j$  se a probabilidade do vetor da observação ( $X_u$ ) é maior para o grupo  $j$  do que para qualquer outro grupo. A equação é:

$$f(x_u | j) > f(x_u | j') \quad (32)$$

Para todo  $j' \neq j$ .

Mapa perceptual bidimensional: É um sistema de coordenadas retangular usado para plotar os centróides dos grupos e os valores dos escores dos casos nas duas principais funções discriminantes. Cada grupo na amostragem pode-se determinar a sua localização no espaço por um ponto que representa a sua média dos escores dos casos gerados pelo modelo discriminante. Esses pontos médios de cada grupo são chamados grupos centróides. Para cada caso da amostra também se pode computar a distância desses em relação ao seu centróide de grupo através da distância Mahalanobis. Os casos foram classificados dentro de um grupo conforme o menor valor de distância Mahalanobis. A projeção desses pontos centróides representa a idéia de separação dos grupos no espaço bidimensional das duas funções discriminantes. Assim, o mapa representa os valores obtidos nas etapas de cálculo da função dos grupos centróides (pág. 35) e classificação dos casos (item anterior).

Tabela dos resultados de classificação: também chamada de matriz de classificação, essa tabela é uma tabela de contingência que compara a designação original dos objetos (casos)

nos grupos (na linha) com a designação feita pelas funções discriminantes geradas pelo modelo (nas colunas). Nessa tabela são determinados o número e porcentagem de casos que foram corretamente classificados. Uma grande proporção de classificação correta implica em dizer que as variáveis selecionadas são boas discriminatórias e que os grupos diferem entre si devido ao conjunto dessas variáveis. A tabela de classificação pode ser construída conforme o exemplo abaixo, demonstrado para dois grupos.

Tabela 4. Tabela de classificação para análise discriminante.

GRUPOS ATUAIS	No. de	CASOS PREVISTOS NOS GRUPOS	
		Grupo 1	Grupo 2
Grupo 1	$N_1$	$n_{11}$	$n_{12}$
Grupo 2	$N_2$	$n_{21}$	$n_{22}$

Onde:

$n_{11}$  = número de casos que pertencia ao grupo 1 e foi designado no grupo 1. (i.e. foi corretamente classificado).

$n_{12}$  = número de casos que pertencia ao grupo 1 e foi designado no grupo 2 (casos incorretamente classificados)

$n_{21}$  = número de casos que pertencia ao grupo 2 e foi designado no grupo 1 (i.e. classificado incorretamente)

$n_{22}$  = número de casos que pertencia ao grupo 2 e foi designado no grupo 2 (i.e. classificado corretamente)

Número total de casos ( $NC$ ) corretamente classificados:

$$NC = n_{11} + n_{22} \quad (33)$$

Porcentagem de casos corretamente classificados

$$\% \text{ casos} = \frac{n_{11} + n_{22}}{n} \times 100 \quad (34)$$

Onde  $n$  é o total do número de casos

$$n = n_{11} + n_{12} + n_{21} + n_{22}$$



Etapa de validação com o grupo teste: a etapa de validação cruzada dos resultados na análise discriminante é muito importante, especialmente se a intenção do pesquisador é classificar outras amostras dentro dos grupos de interesse. O método onde se insere uma nova observação para a classificação do modelo é também chamada de jackknife. Nesse teste foi possível identificar se um provável caso pertence ao seu verdadeiro grupo quando se conhece apenas os valores de suas variáveis discriminatórias desse caso. A tabela com essa análise foi da classificação dos casos que utilizou a distância Mahalanobis para a re-classificação dos casos conforme a equação (26).

## 4 RESULTADOS

O modelo discriminante realizado pelo SPSS gerou um arquivo com 21 *outputs*, aqui representados por passos em três etapas diferentes da análise (validação do modelo, cálculo das funções discriminantes e classificação). Os comentários de todos os *outputs* estão representados a seguir.

### 4.1 Etapa de validação

1. Sumário dos casos processados pela análise: Na Figura 4 observa-se que todos os 50 casos (todas as 50 unidades amostrais de 1000m<sup>2</sup>), foram validados para dar prosseguimento à análise. Não houve dados faltantes. Os casos e seus valores podem ser consultados no ANEXO A.

Unweighted Cases		N	Percent
Valid		50	100,0
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Total	0	,0
Total		50	100,0

Figura 4. *Output* original com a primeira análise demonstrando o número de casos inseridos no modelo.

2. Análise descritiva dos dados: Na Figura 5 observa-se os resultados dos valores médios e o desvio-padrão das oito variáveis predictoras em cada grupo. Na mesma verifica-se também que todos os grupos apresentam os mesmos números de casos em cada variável independente, por isso tiveram o mesmo peso.

FLORESTA		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
Savana	N	14,10000	5,952590	10	10,000
	RIQ	1,80000	,918937	10	10,000
	ALTCOM	1,91500	,242728	10	10,000
	ALTTOT	3,57500	,441116	10	10,000
	ALTMAX	7,72600	1,761995	10	10,000
	AB	,49000	,201487	10	10,000
	DAP	17,74350	4,451805	10	10,000
	DAPMAX	36,00000	7,952917	10	10,000
Restinga	N	73,00000	12,657892	10	10,000
	RIQ	14,30000	2,213594	10	10,000
	ALTCOM	3,71000	,470697	10	10,000
	ALTTOT	9,38500	,397247	10	10,000
	ALTMAX	17,29400	1,955910	10	10,000
	AB	4,11080	,961961	10	10,000
	DAP	13,49650	1,010019	10	10,000
	DAPMAX	101,77900	16,745297	10	10,000
Galeria	N	135,20000	15,732486	10	10,000
	RIQ	17,90000	1,969207	10	10,000
	ALTCOM	5,57000	,265832	10	10,000
	ALTTOT	9,27500	,428985	10	10,000
	ALTMAX	17,46400	2,109862	10	10,000
	AB	2,26000	,322779	10	10,000
	DAP	12,08800	,364449	10	10,000
	DAPMAX	42,04200	9,811433	10	10,000
Ciliar	N	90,40000	21,407164	10	10,000
	RIQ	16,30000	3,653005	10	10,000
	ALTCOM	5,12500	,504012	10	10,000
	ALTTOT	8,58000	1,280451	10	10,000
	ALTMAX	14,93900	1,891058	10	10,000
	AB	2,99230	,850177	10	10,000
	DAP	14,44300	1,181205	10	10,000
	DAPMAX	62,02500	20,085882	10	10,000
FES	N	87,10000	13,908031	10	10,000
	RIQ	21,40000	2,221111	10	10,000
	ALTCOM	4,65500	,465743	10	10,000
	ALTTOT	9,51500	,864436	10	10,000
	ALTMAX	25,31800	3,466506	10	10,000
	AB	3,18800	,273610	10	10,000
	DAP	15,29400	1,176947	10	10,000
	DAPMAX	57,97100	8,795306	10	10,000
Total	N	79,96000	41,852004	50	50,000
	RIQ	14,34000	7,124577	50	50,000
	ALTCOM	4,19500	1,365809	50	50,000
	ALTTOT	8,06600	2,405997	50	50,000
	ALTMAX	16,54820	6,111249	50	50,000
	AB	2,60822	1,357561	50	50,000
	DAP	14,61300	2,831732	50	50,000
	DAPMAX	59,96340	26,652839	50	50,000

Figura 5. *Output* original com os valores da média e desvio-padrão de cada variável independente nas cinco variáveis dependentes do modelo. São fornecidas informações sobre o número de casos utilizados para o cálculo das médias (10 casos). As variáveis descritoras foram: N (número de indivíduos amostrados), RIQ (riqueza ou número total de espécies amostradas), ALTCOM (altura média comercial dos indivíduos amostrados em metros), ALTTOT (altura média total dos indivíduos amostrados em metros), ALTMAX (as maiores alturas atingidas por indivíduo na área amostral em metros), AB (área basal em m<sup>2</sup>), DAP (média dos diâmetros dos caules dos indivíduos amostrados em centímetros), DAPMAX (diâmetro máximo do caule atingido pelos indivíduos na área amostral em centímetros). Ver detalhes no item Material e Métodos.

3. Matriz de correlação: Na Figura 6, os escores de correlação entre as variáveis independentes revelaram valores baixos entre si. A única exceção ocorreu entre altura comercial e altura total com uma correlação média entre as variáveis (0,60).

	N	RIQ	ALTCOM	ALTTOT	ALTMAX	AB	DAP	DAPMAX
Correlation N	1,000	,407	,405	,418	-,184	,349	,005	-,188
RIQ	,407	1,000	,108	,333	-,143	,309	-,013	,156
ALTCOM	,405	,108	1,000	,600	-,141	,231	,348	,072
ALTTOT	,418	,333	,600	1,000	,064	,440	,351	,413
ALTMAX	-,184	-,143	-,141	,064	1,000	,141	,103	,303
AB	,349	,309	,231	,440	,141	1,000	,241	,367
DAP	,005	-,013	,348	,351	,103	,241	1,000	,072
DAPMAX	-,188	,156	,072	,413	,303	,367	,072	1,000

Figura 6. *Output* original da matriz de correlação entre as variáveis nos grupos. O significado dos símbolos que caracterizam as variáveis independentes foi demonstrado na figura 5.

4. Teste de igualdade de médias dos grupos: Os resultados do lambda de Wilks foram baixos o que indicou que as variáveis foram boas discriminadoras dos grupos. Nesse sentido, a variável altura comercial foi melhor discriminadora dos grupos, seguida de altura total e riqueza de espécies. A variável estrutural menos importante foi DAP. O teste *F* associado também revelou valores elevados e altamente significante ( $p < 0,01$ ) revelando que em pelo menos um grupo as variáveis diferem. Os valores da estatística *F* revelados na análise do SPSS foram os mesmos apresentados no ANEXO E (desenvolvido pelo pacote estatística STATISTICA 7). De forma geral, os valores indicam que as variáveis estruturais inseridas no modelo podem ser boas descritoras para a diferença entre os tipos de florestas.

	Wilks' Lambda	F	df1	df2	Sig.
N	,115	86,739	4	45	,000
RIQ	,101	100,185	4	45	,000
ALTCOM	,081	127,736	4	45	,000
ALTTOT	,093	110,043	4	45	,000
ALTMAX	,133	73,553	4	45	,000
AB	,186	49,185	4	45	,000
DAP	,544	9,428	4	45	,000
DAPMAX	,238	36,007	4	45	,000

Figura 7. *Output* original do teste de lambda de Wilks e *F* de igualdade de médias de grupo. O significado dos símbolos que caracterizam as variáveis independentes foi demonstrado na Figura 5.

5. Teste de igualdade de matrizes de covariância: O teste específico para demonstrar se as matrizes de co-variâncias são iguais foi obtida através do teste M de Box. Esse teste é uma variação do teste  $F$  e verifica a hipótese nula de igualdade de matrizes de covariância através do nível de significância obtido. O Teste  $M$  de Box, associado à estatística  $F$ , proporcionou evidências significativas para suportar a conclusão de que as matrizes de covariâncias das populações diferem:  $M= 414,087$ ,  $F(2, 3687) = 1,691$ ,  $P < 0.001$  (Figura 8).

Box's M		414,087
F	Approx.	1,691
	df1	144
	df2	3687,658
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

Figura 8. *Output* original do teste M de Box para a igualdade das matrizes de covariância.

#### 4.2 Etapa do cálculo das funções discriminantes

6. Estatística para a seleção das variáveis no modo de inclusão avançada (*Forward*): Na Figura 9 observa-se que o modelo gerou oito passos e revelou as melhores variáveis estruturais preditoras para discriminar as florestas estudadas através dos níveis de significância que alcançaram ( $p < 0,05$ ): todas aquelas relacionadas com altura das comunidades (altura comercial, altura total e altura máxima), o número de indivíduos amostrados, a riqueza de espécies, o diâmetro máximo dos caules, a área basal e o diâmetro do caule dos indivíduos arbóreos.

**Variables Entered/Removed<sup>a,c,d</sup>**

Step	Entered	Wilks' Lambda											
		Statistic	df1	df2	df3	Exact F				Approximate F			
						Statistic	df1	df2	Sig.	Statistic	df1	df2	Sig.
1	ALTCOM	,081	1	4	45,000	127,736	4	45,000	,000				
2	ALTTOT	,015	2	4	45,000	79,275	8	88,000	,000				
3	ALTMAX	,004	3	4	45,000					64,588	12	114,059	,000
4	N	,002	4	4	45,000					51,508	16	128,950	,000
5	RIQ	,001	5	4	45,000					44,569	20	136,931	,000
6	DAPMAX	,001	6	4	45,000					41,394	24	140,753	,000
7	AB	,000	7	4	45,000					36,729	28	142,039	,000
8	DAP	,000	8	4	45,000					33,477	32	141,732	,000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 16.

b. Maximum significance of F to enter is .05.

c. Minimum significance of F to remove is .10.

d. F level, tolerance, or VIN insufficient for further computation.

Figura 9. *Output* original revelando a seleção das variáveis no modelo através da estatística lambda de Wilks. O significado dos símbolos que caracterizam as variáveis independentes foi demonstrado na Figura 5.

7. Seleção das variáveis pelo método de inclusão avançada (*Forward*): Tratando-se dessa análise, na Figura 10 a seguir se observa que a variável altura comercial teve o melhor poder de discriminação dentre todas as demais para distinguir os tipos de florestas. Após seguiu as alturas total e máxima, número de indivíduos e riqueza de espécies e com as menores poder de discriminação foram selecionadas o DAP máximo, a área basal e o DAP médio dos indivíduos. A variável mais importante para o modelo é aquela que junto com outra (p-1 variável) produz a menor separação e conseqüentemente o maior valor de lambda de Wilks. O baixo valor de significância do *F* para remover a variável e o alto valor de lambda de Wilks indicou um grande poder de classificar os grupos.

**Variables in the Analysis**

Step		Tolerance	Sig. of F to Remove	Wilks' Lambda
1	ALTCOM	1,000	,000	
2	ALTCOM	,640	,000	,093
	ALTTOT	,640	,000	,081
3	ALTCOM	,607	,000	,028
	ALTTOT	,617	,000	,015
	ALTMAX	,945	,000	,015
4	ALTCOM	,590	,000	,007
	ALTTOT	,571	,000	,008
	ALTMAX	,910	,000	,008
	N	,758	,000	,004
5	ALTCOM	,560	,000	,004
	ALTTOT	,520	,000	,004
	ALTMAX	,888	,000	,003
	N	,680	,000	,003
	RIQ	,752	,000	,002
6	ALTCOM	,555	,000	,002
	ALTTOT	,405	,010	,001
	ALTMAX	,839	,000	,002
	N	,565	,000	,001
	RIQ	,723	,000	,001
	DAPMAX	,601	,000	,001
7	ALTCOM	,555	,000	,001
	ALTTOT	,401	,025	,001
	ALTMAX	,826	,000	,001
	N	,513	,000	,001
	RIQ	,717	,000	,001
	DAPMAX	,546	,004	,001
	AB	,661	,011	,001
8	ALTCOM	,529	,000	,001
	ALTTOT	,376	,016	,000
	ALTMAX	,818	,000	,001
	N	,472	,000	,001
	RIQ	,716	,000	,001
	DAPMAX	,518	,020	,000
	AB	,627	,010	,001
	DAP	,738	,016	,000

Figura 10. *Output* original das variáveis descritoras que foram inseridas pelo critério de inclusão avançada (*Forward*) através da estatística de lambda de Wilks. O significado dos símbolos que caracterizam as variáveis independentes foi demonstrado na figura 5.

8. Seleção das variáveis pelo método de eliminação retrógrada (*Backward*): Na Figura 12 observa-se que nenhuma variável foi desconsiderada do modelo, revelando que todas são boas preditoras para distinção das florestas. Novamente, aqui fica demonstrado que a variável com maior poder de discriminação foi a altura comercial (ALTCOM) e a com menor poder de discriminação foi o diâmetro dos caules dos indivíduos (DAP).

**Variables Not in the Analysis**

Step		Tolerance	Min. Tolerance	Sig. of F to Enter	Wilks' Lambda
0	N	1,000	1,000	,000	,115
	RIQ	1,000	1,000	,000	,101
	ALTCOM	1,000	1,000	,000	,081
	ALTTOT	1,000	1,000	,000	,093
	ALTMAX	1,000	1,000	,000	,133
	AB	1,000	1,000	,000	,186
	DAP	1,000	1,000	,000	,544
	DAPMAX	1,000	1,000	,000	,238
1	N	,836	,836	,000	,043
	RIQ	,988	,988	,000	,025
	ALTTOT	,640	,640	,000	,015
	ALTMAX	,980	,980	,000	,015
	AB	,947	,947	,000	,019
	DAP	,879	,879	,000	,049
	DAPMAX	,995	,995	,000	,019
2	N	,788	,603	,000	,008
	RIQ	,876	,567	,000	,007
	ALTMAX	,945	,607	,000	,004
	AB	,804	,544	,000	,009
	DAP	,847	,617	,000	,009
	DAPMAX	,781	,502	,000	,007
3	N	,758	,571	,000	,002
	RIQ	,838	,532	,000	,003
	AB	,793	,534	,000	,003
	DAP	,832	,581	,000	,003
	DAPMAX	,727	,501	,000	,002
4	RIQ	,752	,520	,000	,001
	AB	,745	,518	,000	,001
	DAP	,801	,545	,003	,002
	DAPMAX	,625	,417	,000	,001
5	AB	,728	,486	,000	,001
	DAP	,799	,496	,004	,001
	DAPMAX	,601	,405	,000	,001
6	AB	,661	,401	,011	,000
	DAP	,778	,376	,018	,001
7	DAP	,738	,376	,016	,000

Figura 11. *Output* original das variáveis descritoras que foram inseridas pelo critério de eliminação retrógrada (*Backward*) através da estatística de lambda de Wilks. O significado dos símbolos que caracterizam as variáveis independentes foi demonstrado na Figura 5.

9. Sumário da função discriminante: Pode-se constatar na Figura 12 que o primeiro autovalor ( $\lambda_{FD1} = 33,99$ ) possui 74,7% da habilidade de discriminação, o segundo autovalor ( $\lambda_{FD2} = 6,349$ ) explica 14% da discriminação e o terceiro autovalor ( $\lambda_{FD3} = 4,036$ ) explicou 8,9% da discriminação. O quarto autovalor ( $\lambda_{FD4} = 1,119$ ) explicou muito pouco a discriminação dos grupos (2,5%). No total, os três primeiros autovalores apresentaram 97,5% de habilidade de discriminação. O valor elevado da correlação canônica nas três primeiras funções discriminantes (Figura 12) indicou a maior variação compartilhada entre as variáveis e suas respectivas funções discriminantes.



**Eigenvalues**

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	33,999 <sup>a</sup>	74,7	74,7	,986
2	6,349 <sup>a</sup>	14,0	88,7	,929
3	4,036 <sup>a</sup>	8,9	97,5	,895
4	1,119 <sup>a</sup>	2,5	100,0	,727

a. First 4 canonical discriminant functions were used in the analysis.

Figura 12. *Output* original dos resultados do teste dos autovalores e suas respectivas porcentagem de variância, variância cumulativa e correlação canônica para cada função discriminante.

10. Teste de lambda de Wilks para as funções: O teste lambda de Wilk mediu o quanto que cada função separaram os casos dentro dos grupos. Os valores baixos de lambda de Wilks demonstrados na Figura 13 sugeriram que grande habilidade discriminatória principalmente para as duas primeiras funções discriminantes. O teste qui-quadrado associado indicou com elevada probabilidade (>99%) que as médias das funções listadas não são iguais para os grupos. De uma forma geral, a análise forneceu evidências que pode-se rejeitar a hipótese nula de que as centróides dos grupos são idênticos.

**Wilks' Lambda**

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 4	,000	336,490	32	,000
2 through 4	,013	185,388	21	,000
3 through 4	,094	100,621	12	,000
4	,472	31,912	5	,000

Figura 13. *Output* original do teste de dimensionalidade para os cinco grupos conforme a estatística lambda de Wilks e Qui-quadrado.

11. Coeficientes da função discriminante canônica padronizada: Os coeficientes gerados proporcionaram determinar o cálculo para a função discriminante (escore  $Z$ ) e a relação comparativa entre a variável discriminatória com a respectiva função discriminante. Dessa forma, as equações lineares de classificação do modelo, considerando as três funções discriminantes (Figura 14), foram assim descritas:

$Z1 = -0,095 * \text{número de indivíduos} + 0,548 * \text{riqueza de espécies} + 0,673 * \text{altura comercial} + 0,215 * \text{altura total} + 0,682 * \text{altura máxima atingida} - 0,005 \text{ área basal} - 0,460 - 0,358 \text{DAP máximo atingido}.$

$Z2 = -0,511 * \text{número de indivíduos} - 0,002 * \text{riqueza de espécies} - 0,609 * \text{altura comercial} - 0,757 * \text{altura total} + 0,026 * \text{altura máxima atingida} + 0,584 \text{ área basal} - 0,204 + 0,069 \text{ DAP máximo atingido}.$

$Z3 = 0,593 * \text{número de indivíduos} - 0,534 * \text{riqueza de espécies} + 0,230 * \text{altura comercial} - 0,016 * \text{altura total} - 0,541 * \text{altura máxima atingida} + 0,105 \text{ área basal} - 0,391 + 0,680 \text{ DAP máximo atingido}.$

De acordo com os resultados (Figura 14), a primeira função discriminante a quem melhor representa a separação dos grupos, foi caracterizada por três variáveis predictoras: altura comercial altura máxima, riqueza de espécies; a segunda função discriminante foi representada pelas variáveis: altura total e área basal; e a terceira função discriminante, a menos importante para classificação dos grupos foi representada pelas variáveis: número de indivíduos, DAP e DAP máximo.

Standardized Canonical Discriminant Function Coefficients				
	Function			
	1	2	3	4
N	-,095	-,511	,593	-,892
RIQ	,548	-,002	-,534	,371
ALTCOM	,673	-,609	,230	,881
ALTTOT	,215	,757	-,016	-,572
ALTMAX	,682	,026	-,541	-,318
AB	-,005	,584	,105	,541
DAP	-,460	-,204	-,391	,026
DAPMAX	-,358	,069	,680	-,016

Figura 14. *Output* original do cálculo dos autovetores ou escores discriminantes das variáveis em cada uma das quatro funções discriminante gerada pelo modelo. O significado dos símbolos que caracterizam as variáveis independentes foi demonstrado na Figura 5.

12. Estrutura da matriz: Pode-se constatar, na Figura 15, que, novamente, são revelados os pesos de cada variável na distinção dos grupos, entretanto aqui as variáveis estão ordenadas pela sua magnitude de contribuição em cada função discriminante. Esses resultados permitem uma análise subjetiva do pesquisador sobre as variáveis mais importantes em cada função. Assim, as mesmas três variáveis da análise anterior contribuíram para a função

discriminante 1 (altura comercial, riqueza de espécies e altura total), duas variáveis tiveram maior peso na função discriminante 2 (DAP máximo e área basal) e três variáveis tiveram maiores pesos na função discriminante 3 (número de indivíduos, altura máxima e DAP). O importante é observar que essa análise não permite que o pesquisador determine em qual grupo (tipo de floresta) a variável discriminatória teve maior ou menor poder discriminante, o que será possível nas análises seguintes.

**Structure Matrix**

	Function			
	1	2	3	4
ALTCOM	,540*	-,296	,416	,394
RIQ	,506*	,166	-,052	,123
ALTTOT	,494*	,394	,347	-,072
DAPMAX	,054	,641*	,341	,140
AB	,235	,578*	,272	,252
N	,429	-,257	,469*	-,373
ALTMAX	,383	,337	-,426*	-,294
DAP	-,115	-,003	-,289*	,219

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions  
Variables ordered by absolute size of correlation within function.

\*. Largest absolute correlation between each variable and any discriminant function

Figura 15. *Output* original dos escores discriminantes originados pela correlação entre as variáveis discriminatórias e os escores das funções discriminantes canônicas padronizadas. O significado dos símbolos que caracterizam as variáveis independentes foi demonstrado na Figura 5.

13. Função dos grupos centróides: Na Figura 16 verifica-se que os valores médios da distância de Mahalanobis dos 10 casos em CAD grupo. O modelo gerado demonstrou que cada tipo de floresta têm valores bem diferentes para os centróides nas funções discriminantes o que remete a um esperado distanciamento dos centróides do mapa perceptual.

**Functions at Group Centroids**

FLORESTA	Function			
	1	2	3	4
Savana	-10,138	-1,279	-1,076	-,192
Restinga	-1,059	4,036	1,808	-,462
Galeria	4,373	-2,955	1,445	-1,132
Ciliar	1,583	-,874	1,066	1,870
FES	5,241	1,071	-3,243	-,084

Unstandardized canonical discriminant functions evaluated at group means

Figura 16. *Output* original dos valores dos centróides de grupos.

### 4.3 Etapa da classificação

14. Resumo do teste de classificação: No presente exemplo, todos os 50 casos foram processados na análise discriminante na etapa de classificação.

**Classification Processing Summary**

Processed		50
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		50

Figura 17. *Output* original do número de casos processados para a classificação.

15. Probabilidade nos grupos: Na Figura 18 observou-se que a probabilidade foi padrão para todos os grupos e estes apresentam iguais números de casos com mesmo peso.

**Prior Probabilities for Groups**

FLORESTA	Prior	Cases Used in Analysis	
		Unweighted	Weighted
Savana	,200	10	10,000
Restinga	,200	10	10,000
Galeria	,200	10	10,000
Ciliar	,200	10	10,000
FES	,200	10	10,000
Total	1,000	50	50,000

Figura 18. *Output* original revelando que a probabilidade inicial para a separação de grupos foi a mesma para todos os grupos.

16. Classificação dos coeficientes da função: Na Figura 19 foram demonstrados os coeficientes da função das variáveis estruturais em relação aos seus grupos. A análise permitiu constatar que as variáveis, número de indivíduos, altura comercial e a área basal média foram mais representativas na floresta de galeria. No entanto, a riqueza de espécies, a altura máxima atingida e o DAP máximo atingido foram mais representativas para a floresta estacional semidecidual. Por fim, a altura total teve mais relação contra a floresta de restinga e o DAP foi mais representativo para a savana.

Classification Function Coefficients					
	FLORESTA				
	Savana	Restinga	Galeria	Ciliar	FES
N	,150	,040	,273	,023	-,123
RIQ	,469	1,876	3,120	3,028	4,542
ALTCOM	5,885	14,024	31,874	30,423	26,895
ALTTOT	-,558	7,402	2,514	1,552	6,062
ALTMAX	1,352	3,443	5,134	4,015	6,382
AB	-6,684	-1,421	-8,800	-4,196	-4,838
DAP	3,633	,699	,263	,761	,557
DAPMAX	,212	,144	-,052	,009	-,291
(Constant)	-47,359	-116,154	-183,791	-141,446	-205,207

Fisher's linear discriminant functions

Figura 19. *Output* original demonstrando os coeficientes de classificação de Fisher das variáveis preditoras em cada um dos grupos. O significado dos símbolos que caracterizam as variáveis independentes foi demonstrado na figura 5.

17. Classificação dos casos: Na Figura 20 observou-se os valores das distâncias Mahalanobis calculadas para todos os casos em relação ao seu centróide de grupo e a relação com os escores discriminantes em cada função. Uma análise geral permitiu constatar que todos os 50 casos oriundos de seus grupos preditos (grupos originais ou tipos de florestas) foram classificados pelo modelo discriminante e pelas distâncias de Mahalanobis, com elevadas probabilidades, de pertencerem aos seus verdadeiros grupos de origem. Entretanto, pode-se constatar que alguns casos que apresentaram elevados valores de distância de Mahalanobis ( $D > 9$ ) tiveram menor probabilidade de pertencerem ao seu grupo de origem. Isso ocorreu principalmente para os casos oriundos da floresta ciliar (três casos) e da floresta de restinga (um caso). Os valores das distâncias de Mahalanobis e dos escores discriminantes dos casos permitiram observar que a floresta ciliar foi o tipo de floresta que mais se assemelhou na estrutura com a floresta de galeria.

De qualquer forma, a probabilidades dos casos dos grupos de origem pertencerem a outros grupos, mais próximos, foram significativamente baixas ( $p < 1\%$  de confiança) (coluna do segundo grupo principal ou *Second highest group*) onde pode-se constatar pelos valores elevados das distâncias de Mahalanobis quando comprado com as distâncias do seu grupo de origem. Os valores dos escores discriminantes nas funções também demonstram que os casos do mesmo grupo tendem a apresentar escores com valores similares dentro de uma mesma função discriminante.

Casewise Statistics													
Case Number	Actual Group	Highest Group					Second Highest Group			Discriminant Scores			
		Predicted Group	P(D>d   G=g)	df	(G=g   D=d)	Squared Mahalanobis Distance to Centroid	Group	(G=g   D=d)	Squared Mahalanobis Distance to Centroid	function	function	function	function
Origin: 1	1	1	,943	4	1,000	,768	2	,000	105,620	-9,594	-1,161	-,586	-,659
2	1	1	,921	4	1,000	,925	2	,000	133,479	-10,621	-1,300	-1,876	-,417
3	1	1	1,000	4	1,000	,062	2	,000	119,865	-10,120	-1,273	-1,283	-,328
4	1	1	,821	4	1,000	1,529	2	,000	97,954	-8,952	-1,336	-,796	-,396
5	1	1	,213	4	1,000	5,820	2	,000	148,235	-10,750	-1,395	-2,840	1,332
6	1	1	,647	4	1,000	2,488	2	,000	141,549	-11,410	-1,414	-,250	,220
7	1	1	,969	4	1,000	,543	2	,000	134,990	-10,704	-1,560	-1,448	-,259
8	1	1	,899	4	1,000	1,068	2	,000	116,635	-9,683	-1,325	-1,863	-,681
9	1	1	,506	4	1,000	3,316	2	,000	103,821	-9,860	-,973	,689	-,368
10	1	1	,962	4	1,000	,610	2	,000	105,636	-9,685	-1,049	-,506	-,358
11	2	2	,785	4	1,000	1,731	4	,000	25,272	-,894	2,840	2,216	-,134
12	2	2	,379	4	1,000	4,203	4	,000	24,650	,085	2,549	1,973	-1,271
13	2	2	,051	4	1,000	9,423	4	,000	27,913	-1,082	1,532	3,417	-1,212
14	2	2	,230	4	1,000	5,607	4	,000	64,638	-2,423	4,768	1,409	-2,209
15	2	2	,203	4	1,000	5,948	4	,000	67,398	-2,106	6,087	2,452	,021
16	2	2	,940	4	1,000	,787	4	,000	43,885	-1,241	4,733	1,347	-,226
17	2	2	,038	4	1,000	10,132	4	,000	54,906	-1,425	4,057	-,658	-2,441
18	2	2	,371	4	1,000	4,265	4	,000	45,803	,294	5,495	1,633	,064
19	2	2	,359	4	1,000	4,365	4	,000	26,141	-,470	3,786	1,154	1,416
20	2	2	,246	4	1,000	5,430	4	,000	42,085	-1,329	4,517	3,138	1,370
21	3	3	,578	4	1,000	2,881	5	,000	24,345	4,990	-2,572	-,089	-1,116
22	3	3	,766	4	,992	1,833	4	,008	11,514	3,565	-2,655	,924	-,226
23	3	3	,686	4	1,000	2,271	4	,000	22,340	3,216	-3,128	2,119	-1,802
24	3	3	,548	4	1,000	3,060	4	,000	34,186	3,986	-3,545	1,386	-2,732
25	3	3	,426	4	1,000	3,858	4	,000	40,350	4,597	-3,994	2,070	-2,660
26	3	3	,996	4	1,000	,188	4	,000	23,378	4,772	-2,848	1,577	-1,138
27	3	3	,586	4	,982	2,833	4	,018	10,848	3,355	-3,014	,934	,106
28	3	3	,778	4	1,000	1,770	4	,000	28,779	5,574	-2,984	1,960	-,887
29	3	3	,917	4	1,000	,952	4	,000	16,485	4,241	-2,475	2,054	-,555
30	3	3	,702	4	1,000	2,186	4	,000	21,882	5,431	-2,334	1,510	-,309
31	4	4	,677	4	,992	2,320	3	,008	11,886	2,399	-,904	1,761	,788
32	4	4	,628	4	,994	2,595	3	,006	12,814	2,902	-,793	,516	1,130
33	4	4	,880	4	1,000	1,189	3	,000	17,869	2,369	-,363	1,512	1,538
34	4	4	,789	4	1,000	1,709	3	,000	29,594	1,664	,011	,224	2,328
35	4	4	,864	4	1,000	1,282	3	,000	21,288	2,280	-,075	1,457	1,803
36	4	4	,134	4	,996	7,038	2	,004	18,335	,261	,637	2,773	1,564
37	4	4	,324	4	1,000	4,660	2	,000	22,799	2,248	,984	1,479	1,099
38	4	4	,016	4	1,000	12,207	3	,000	29,100	2,869	-3,316	,780	3,993
39	4	4	,019	4	1,000	11,753	3	,000	41,427	-,249	-3,386	1,112	3,314
40	4	4	,023	4	1,000	11,303	3	,000	40,903	-,915	-1,535	-,958	1,139
41	5	5	,416	4	1,000	3,928	3	,000	38,931	5,656	,648	-3,413	-1,968
42	5	5	,594	4	1,000	2,787	4	,000	38,971	3,963	1,728	-3,890	,465
43	5	5	,874	4	1,000	1,227	3	,000	43,177	5,877	,873	-3,673	-,858
44	5	5	,956	4	1,000	,662	3	,000	45,688	5,435	,955	-3,925	-,466
45	5	5	,220	4	,993	5,733	3	,007	15,527	5,023	-,463	-1,486	-,580
46	5	5	,429	4	1,000	3,833	4	,000	27,350	5,084	1,704	-1,483	,474
47	5	5	,622	4	1,000	2,625	4	,000	48,452	5,922	,650	-4,093	1,039
48	5	5	,138	4	1,000	6,968	3	,000	68,316	7,327	1,387	-4,744	,430
49	5	5	,602	4	1,000	2,740	4	,000	32,041	4,477	2,070	-2,663	,823
50	5	5	,629	4	1,000	2,589	4	,000	29,695	3,650	1,163	-3,055	-,201

Figura 20. *Output* original com os escores discriminantes de cada caso em cada função discriminante e seus valores de distância Mahalanobis em relação ao seu centróide de grupo. O valor da estatística da distância de Mahalanobis também foi calculado para o caso em relação ao grupo original e grupo predito.

18. Mapa perceptual bidimensional: Na Figura 21, foi representada um sistema de configuração da classificação dos grupos no espaço bi-dimensional para as duas primeiras funções discriminantes. Nota-se que o ponto médio do centróide dos grupos encontra-se fortemente relacionado com seus escores dos casos, valores esses apresentados na figura 21. Pode-se constatar que os centróides dos grupos estiveram bem distantes entre si, o que representou uma boa classificação dos grupos. As linhas que ligam os centróides aos escores das funções discriminantes de cada grupo representam o intervalo de confiança de 95% dos casos pertencerem aquele determinado grupo. Em uma análise geral pode-se constatar que a formação savana arbórea foi a fitofisionomia mais distinta de todas as demais florestas, indicado valores das variáveis bem diferentes das demais para os outros grupos e ocupou a posição negativa nos dois eixos do plano bidimensional. O grupo com a posição próxima ao eixo zero dos dois eixos demonstrou que as variáveis apresentam valores médios similares a todos os grupos (neste caso foi a floresta ciliar). A similar localização dos escores próximos ao centróide de grupo representa pouca variabilidade nos valores das variáveis discriminatórias (no caso da Savana arbórea) e o contrário indica ampla variabilidade dos valores nas variáveis discriminatórias (como na floresta ciliar e floresta de restinga). O gráfico também revelou que as florestas que mais diferem na função 1 para o conjunto das variáveis altura comercial, riqueza e altura total ocorrem entre a savana e as demais florestas. Na função 2, representado pelas variáveis DAP máximo e área basal, separam melhor as florestas de restinga arenosa e mata de galeria.

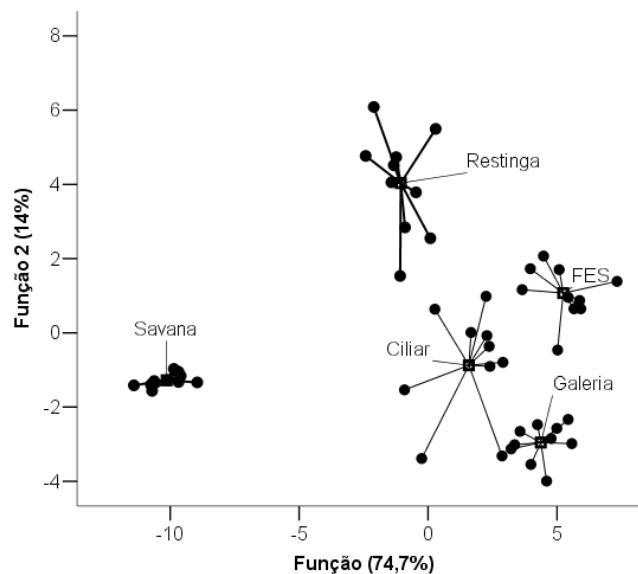


Figura 21. *Output* original do mapa perceptual bidimensional representando os valores dos escores dos centróides de grupos e dos escores discriminantes de cada caso em seus respectivos grupos.

19. Tabela dos resultados da classificação: Na Figura 22 observa-se a frequência e a classificação dos 50 casos inseridos no modelo nos seus grupos. Os valores na diagonal mostram os casos corretamente classificados e na horizontal os casos não classificados pelo modelo discriminante. A parte superior da Figura 22 revelou que todos os 10 casos que representaram um tipo de floresta foram corretamente classificados nos seus respectivos grupos de origem. Na parte inferior da Figura 22, uma análise similar mostrou a proporção de acertos da classificação em cada grupo, revelando que 100% dos casos inseridos no modelo foram devidamente classificados. Assim, as variáveis independentes validadas pelo modelo (número de indivíduos, riqueza, altura total, comercial e máxima, área basal e DAP médio) apresentaram uma boa eficiência para a separação dos tipos de florestas no Pampa.

**Classification Results<sup>a</sup>**

FLORESTA		Predicted Group Membership					Total	
		Savana	Restinga	Galeria	Ciliar	FES		
Original	Count	Savana	10	0	0	0	0	10
		Restinga	0	10	0	0	0	10
		Galeria	0	0	10	0	0	10
		Ciliar	0	0	0	10	0	10
		FES	0	0	0	0	10	10
%		Savana	100,0	,0	,0	,0	,0	100,0
		Restinga	,0	100,0	,0	,0	,0	100,0
		Galeria	,0	,0	100,0	,0	,0	100,0
		Ciliar	,0	,0	,0	100,0	,0	100,0
		FES	,0	,0	,0	,0	100,0	100,0

a. 100,0% of original grouped cases correctly classified.

Figura 22. *Output* original da tabela de classificação.

20. Etapa de validação com o grupo teste: A análise revelou que a maioria dos casos do grupo teste (validação cruzada) foi classificada corretamente pelo modelo dentro do seu grupo esperado (Figura 23). A estatística forneceu a probabilidade de ajuste dos casos aos seus respectivos grupos. Cinco casos tiveram alta probabilidade de pertencerem ( $p > 90\%$ ) aos seus grupos e três casos tiveram uma probabilidade moderada de classificação em seu grupo verdadeiro ( $50\% < p < 68\%$ ). Sob esse último aspecto, a análise estatística não permite que o pesquisador tenha certeza de que os casos com probabilidade média sejam verdadeiros daquele grupo, necessitando que o pesquisador use meios subjetivos para a interpretação. Os casos selecionados provenientes das florestas de restinga arenosa não foram re-classificados no seu grupo de origem (floresta de restinga grupo 2) e permanecendo no grupo teste, significando que esses dois casos novos das florestas de restinga apresentaram os atributos estruturais diferentes do original.



Validação cruzada (a)		Grupo principl					Segundo grupo principal			
Número dos casos	Grupo atual	Grupo predito	P(D>d   G=g)		P(D>d   G=g)	Distância Mahalanobis do centróide	Grupos	P(G= g   D=d)	Distância Mahalanobis do centróide	
			p	gl						
Savana	51	6	1(**)	,793	8	1,000	4,665	4	,000	26,534
Savana	52	6	1(**)	,298	8	1,000	9,551	6	,000	25,563
Ciliar	53	6	4(**)	,177	8	,618	11,452	3	,240	13,347
Ciliar	54	6	4(**)	,754	8	,528	5,030	6	,282	6,284
Galeria	55	6	3(**)	,447	8	,999	7,860	4	,001	22,725
Galeria	56	6	3(**)	,051	8	,681	15,431	5	,314	16,980
FES	57	6	5(**)	,003	8	,998	23,642	6	,001	36,943
FES	58	6	5(**)	,016	8	,991	18,882	6	,009	28,226
Restinga	59	6	6	,001	8	,952	27,318	5	,025	34,596
Restinga	60	6	6	,155	8	,764	11,917	5	,235	14,273

Figura 23. *Output* modificado com os valores das classificações dos casos no grupo teste com as probabilidades de classificação pela distância de Mahalanobis. Para os dados de validação cruzada à distância Mahalanobis quadrada é baseada nas observações. Na validação cruzada, cada caso é classificado pela função derivada dos demais casos. Os grupos preditos foram: 1- Savana; 2- Restinga; 3- Galeria; 4- Ciliar; 5- FES; 6- Grupo Teste. \*\* Erros de classificação.

## 5 CONSIDERAÇÕES FINAIS

### 5.1 O modelo discriminante de Fisher

Primeiramente, o estudo demonstrou que o modelo exige que seja considerada uma série de critérios prévios para a validação ou não do modelo discriminante linear de Fisher. Nesse sentido, as fitofisionomias escolhidas constituíram amostras de grupos mutuamente exclusivos (BROWN & WICKER, 2000) e que puderam ser aceitos pela análise discriminante. As variáveis independentes (atributos estruturais) também revelaram ter bom potencial de discriminação de grupos, devido a baixa correlação entre as variáveis preditoras e os resultados significativos dos testes lambda de Wilks e ANOVA de um fator para avaliar igualdade de médias das variáveis dentro dos grupos (cinco tipos de florestas). O número de oito variáveis discriminatórias contidas em cinco grupos de variáveis dependentes tendo, em cada um dos grupos 10 casos ( $n_{\text{grupo/caso}}=10$ ), somando um total de 50 casos ( $n_{\text{total}}= 50$ ) (ANEXO A), pode ser considerado suficiente porque atingiu o tamanho mínimo de casos por variável/grupo (BROWN & WICKER, 2000; HAIR *et al.*, 2007) e pela proporção de variáveis por grupo (MINGOTI, 2005). A estatística  $F$  derivada do teste lambda de Wilks também forneceu evidências significativas de que todas as variáveis puderam ser inseridas no modelo.

Geralmente as variáveis biológicas tendem a ter uma distribuição não-normal e os dados abióticos a distribuição normal (KENKEL *et al.*, 2006). Assim, o artifício de utilizar 10 unidades amostrais com 1000 m<sup>2</sup> cada em vez de 100 unidades amostrais com 100 m<sup>2</sup> reduziu muito a chance de que todas as variáveis independentes fossem não-paramétricas e que não existissem valores extremos nos casos (ANEXO B) cumprindo também as exigências de normalidade dos dados (BROWN & WICKER, 2000; HAIR *et al.*, 2007; MANLY, 2008). Outra premissa cumprida de validação da análise foi quanto a igualdade das matrizes de covariâncias, confirmada por meio do valor significativo para o teste  $M$  de Box.

O modelo discriminante gerado foi satisfatório para distinguir os cinco tipos de florestas do Pampa e todas as oito variáveis estruturais da vegetação foram consideradas com

poder de discriminação para os tipos de florestas avaliados. O método STEPWISE selecionou as variáveis com melhor poder de discriminação (em ordem de importância): o número de indivíduos por área, a riqueza de espécies por área, a altura comercial, a altura máxima atingida por um indivíduo, a área basal da comunidade, o DAP dos caules e o DAP máximo de um caule atingido na comunidade. O modelo gerou quatro funções discriminantes para separação dos grupos, desses pode-se destacar os dois primeiros com os maiores poderes de classificação dos grupos (FD1 = 74,7% e FD2= 14%). A FD3 e a FD4, apresentaram baixo poder de classificação dos grupos 8,9% e 2,5%, respectivamente.

As variáveis que apresentaram maior peso para a FD1 foram: riqueza de espécies, altura comercial e altura total. Esses atributos podem ser considerados como os mais importantes da vegetação que distinguiram as florestas do Pampa. A interpretação dessa informação revela que as florestas apresentam potenciais distintos para a conservação da biodiversidade no bioma e condições distintas de crescimento das espécies na estrutura vertical das comunidades.

A FD2 foi a segunda função mais importante para a separação dos tipos de florestas, sendo representada de forma mais evidente pelos atributos área basal e o DAP máximo dos indivíduos. Esses atributos representam uma parte da estrutura horizontal da comunidade e está relacionada também à produtividade do ecossistema florestal (SAGAR & SING, 2006). Os atributos restantes, número de indivíduo, altura máxima atingida por um indivíduo e o diâmetro do caule (DAP) foram importantes variáveis descritoras no modelo, porém, foram representativas nas FD3 e FD4, que apresentaram baixo poder de discriminação.

O modelo discriminante gerado demonstrou que as funções geradas classificaram 100% dos casos em seus grupos preditos. Cada unidade amostral de 1000m<sup>2</sup> de floresta foi corretamente classificada dentro de seu tipo florestal, revelando que a análise conjunta das variáveis permite uma forte similaridade dentro das fitofisionomias. Isso não significa que as variáveis, individualmente, diferem significativamente dentro das florestas (como será visto a seguir no teste ANOVA de um fator), pois o modelo levou em conta sempre todas as variáveis conjuntamente.

A etapa final de validação de novos casos no modelo, teste de validação cruzada, revelou uma elevada probabilidade de classificação correta para a maioria dos casos nos seus verdadeiros grupos. O problema da classificação errada dos novos casos provenientes das florestas de restinga arenosa pode estar relacionado com o fato de que não encontrou-se

amostragens nessa fitofisionomia em condições similares ao grupo original no banco de dados do IFC-RS. Esse banco de dados apresentou apenas três florestas de restinga na região do Pampa, todas em diferentes condições de preservação. Florestas de mesma fitofisionomia em diferentes idades de desenvolvimento e em diferentes graus de perturbação apresentam grandes diferenças na estrutura da comunidade (GUARIGUATA & OSTERTAG, 2001, KILCA *et al.*, no prelo). Outro fator que sustenta a hipótese de que novos casos provenientes de outras amostragens em florestas de restinga arenosa preservadas poderiam ser suficientes para uma correta classificação (teste de validação cruzada) se deve ao fato de que essas florestas geralmente não variam tanto sua estrutura quanto as florestas ciliares e matas de galeria, submetidas à condições ambientais mais variáveis.

De forma geral, esse estudo considerou ainda de forma empírica, que florestas de uma mesma fitofisionomia e em aparente similar condição de distúrbio possam ter similares estruturas entre si, apesar da clara existência de variações internas na estrutura dentro dessas florestas (ver ANEXO D). Por outro lado, diferentes tipos de florestas, que estejam mantidas preservadas e em estágio sucessional tardio, apresentam estruturas peculiares a tal ponto de poderem ser classificadas como grupos significativamente diferentes em um modelo discriminante linear de Fisher.

## **5.2 A comparação entre as variáveis discriminatórias nos modelos univariado (ANOVA) e multivariado (Análise discriminante de Fisher)**

O modelo multivariado revelou que as florestas apresentam diferenças significativas em suas estruturas a tal ponto que podem ser classificadas como fitofisionomias distintas. A abordagem univariada, no qual cada variável é independentemente examinada das demais, podem falhar na detecção de respostas conjuntas dos fatores que definem a estrutura das comunidades de plantas (KENKEL *et al.*, 2002). Entretanto, a análise multivariada permite saber como o conjunto das variáveis de interesse (no caso a estrutura geral das florestas) podem ser consideradas como preditoras para um modelo revelando se estas têm ou não um papel relevante na diferenciação dos grupos.

No caso de uma análise univariada, geralmente empregasse a análise de variância de um fator (ANOVA) gerado para as variáveis independentemente em cada um dos grupos. No

presente estudo, observou-se que o teste ANOVA refutou a hipótese de que as variáveis não diferem suas médias nos cinco grupos de florestas testados. Algumas dessas variáveis apresentaram médias que diferiram significativamente dentro de um maior número de grupos (em ordem de relevância: riqueza de espécies, área basal, altura comercial, número de indivíduos, altura máxima, DAP máximo) e outras variáveis apresentaram pouca diferença entre os grupos (altura total e diâmetro do caule) (ANEXO E).

Os atributos relacionados com a altura representam a distinta estrutura vertical nas fitofisionomias. O teste ANOVA seguido do teste de Tukey (ANEXO E) para o atributo altura comercial, revelou diferenças significativas para todas as florestas, sendo a savana arbórea com a menor estatura dos caules ( $\bar{X}_{\text{Alt.Com.}} = 1,91\text{m}$ ), seguido da floresta de restinga ( $\bar{X}_{\text{Alt.Com.}} = 3,71\text{m}$ ), da floresta estacional semidecidual ( $\bar{X}_{\text{Alt.Com.}} = 4,65\text{m}$ ), da floresta ciliar ( $\bar{X}_{\text{Alt.Com.}} = 5,12\text{m}$ ) e da floresta de galeria ( $\bar{X}_{\text{Alt.Com.}} = 5,57\text{m}$ ) (ANEXO D). As médias não diferiram apenas para a floresta ciliar com a floresta de galeria e com a FES. A altura máxima atingida, que define o limite máximo da copa em relação ao solo, que a savana teve menor média ( $\bar{X}_{\text{Alt.Tot.}} = 7,72\text{m}$ ), seguido da floresta ciliar ( $\bar{X}_{\text{Alt.Tot.}} = 17,93\text{m}$ ) e os maiores tamanhos ficaram para a floresta de galeria ( $\bar{X}_{\text{Alt.Tot.}} = 17,46\text{m}$ ), a floresta de restinga ( $\bar{X}_{\text{Alt.Tot.}} = 17,29\text{m}$ ) e a floresta semidecidual ( $\bar{X}_{\text{Alt.Tot.}} = 25,3\text{m}$ ) (ANEXO D). A altura máxima dos indivíduos amostrados foi o atributo que menos diferiu entre as florestas, as diferenças significativas foram encontradas apenas entre a savana e os demais tipos de florestas (ANEXO E). De fato, o porte das formações arbóreas de savana, com 3,57m metros de altura máxima, não podem ser comparadas com florestas que atingiram médias próximas aos 9m de altura (ANEXO D).

A riqueza de espécies foi um atributo que apresentou médias muito diferentes na maioria dos tipos de florestas do Pampa. A menor média de espécies ocorreu na formação savânica ( $\bar{X}_{\text{esp.}} = 1,8$ ), seguido da floresta de restinga ( $\bar{X}_{\text{esp.}} = 14,3$ ), floresta ciliar ( $\bar{X}_{\text{esp.}} = 16,3$ ), floresta galeria ( $\bar{X}_{\text{esp.}} = 17,4$ ) e a maior riqueza ocorreu na floresta estacional semidecidual ( $\bar{X}_{\text{esp.}} = 21,4$ ) (ANEXO D). O número de indivíduos foi outra variável preditora selecionada pelo modelo discriminante que demonstrou, através das análises de Tukey (ANEXO E), significativas diferenças entre a maioria das florestas no Pampa. As estatísticas descritivas demonstraram a menor média para a formação savânica ( $\bar{X}_{\text{ind.}} = 14,1$ ), seguido da floresta de restinga ( $\bar{X}_{\text{ind.}} = 73$ ), floresta estacional semidecidual ( $\bar{X}_{\text{ind.}} = 87,1$ ), floresta ciliar ( $\bar{X}_{\text{ind.}} = 90,4$ ) e a floresta de galeria com a maior média de indivíduos ( $\bar{X}_{\text{ind.}} = 135,2$ ) (ANEXO D). Os valores

médios de área basal tiveram a menor média para a formação savânica ( $\bar{X}_{AB} = 0,49\text{m}^2$ ), seguido da floresta de galeria ( $\bar{X}_{AB} = 2,26\text{m}^2$ ), floresta ciliar ( $\bar{X}_{AB} = 2,99\text{m}^2$ ), floresta semidecidual ( $\bar{X}_{AB} = 3,1\text{m}^2$ ) e a maior média para a floresta de restinga arenosa ( $\bar{X}_{AB} = 4,11\text{m}^2$ ) (ANEXO D). Os valores médios para DAP diferiram pouco entre as tipologias florestais. Apenas a formação de savana apresentou os maiores valores médios que foram estatisticamente diferentes dos demais tipos de florestas (ANEXO E).

Os resultados das análises de variância proporcionaram um entendimento individual de cada variável na separação dos grupos e esse resultado foi diferente daquele obtido pela análise discriminante onde considera-se a combinação linear de todas as variáveis numa única análise de separação de grupo. Houve uma semelhança de resultado nas duas abordagens quanto aos pesos dos coeficientes das funções discriminantes (*output* 19, p. 53) com os maiores valores das médias das variáveis preditoras na ANOVA (ANEXO E). De maneira geral, pôde-se constatar que as duas abordagens oferecem a informação oportuna de que as variáveis estruturais tendem a diferir entre os tipos de florestas e a escolha do método estatístico apropriado que vai proporcionar um melhor entendimento da magnitude com que essas florestas podem diferir estruturalmente.

### **5.3 Implicações do modelo gerado para o entendimento ecológico e para a gestão sustentável das florestas no Pampa**

As florestas do Pampa, de uma maneira geral, apresentam menor complexidade estrutural do que a maioria das florestas nos demais biomas brasileiros (RIZZINI, 1997) e tropicais (RICHARDS, 1952) e talvez por isso, sempre foram negligenciadas em programas de conservação da biodiversidade e valoração de seus recursos naturais. Alguns estudos descritivos revelaram a existência de formações vegetais tão simples estruturalmente que poucas informações qualitativas ou quantitativas são necessárias para evidenciar claras diferenças, é o caso principalmente da savana arbórea em relação às demais florestas (RAMBO, 1954; IBGE, 2004). Entretanto, não se sabe em que nível a maioria dos tipos de florestas do Pampa diferem entre si.

A análise discriminante empregada nesse estudo revelou que os cinco tipos de florestas avaliadas diferem significativamente entre si de acordo com oito atributos estruturais da vegetação. Esse grande número de variáveis estruturais pode ser considerado elevado dado a relativa baixa complexidade estrutural dessas formações. A análise discriminante também apontou para a ocorrência de tipos de florestas que desempenham diferentes funções no Pampa, como: aquelas que possuem maior diversidade e porte (florestas estacionais semidecíduais), aquelas que se destacam pelo grande adensamento de indivíduos (floresta de galeria), as formações de menor complexidade estrutural (savana arbórea), etc.

Essas informações apresentam um grande impacto para conservação e gestão desses recursos naturais, pois revela aos tomadores de decisão a necessidade de um planejamento específico em planos de preservação da biodiversidade, restauração de ecossistemas degradados e utilização sustentável desses recursos para cada um desses tipos de florestas. Dessa forma, fica evidente o risco de fracasso em projetos ambientais nas florestas do bioma, caso não venham a considerar a singularidade de cada um desses ecossistemas. Entretanto, seria interessante aprimorar o modelo gerado empregando a seleção de variáveis edáficas, como teores de nutrientes, variação da umidade do solo e textura dos solos nas diferentes florestas. O resultado de um modelo discriminante com dados bióticos e abióticos poderia reforçar ainda mais o entendimento da variação da estrutura e composição das espécies nas fitofisionomias florestais.

O Inventário Contínuo do Estado do Rio Grande do Sul constitui uma das melhores ferramentas para gerenciamento dos recursos florestais por constituir o maior banco de dados quantitativo das características bióticas das florestas no estado. Essas informações favorecem ao desenvolvimento de técnicas analíticas que venham a contribuir à gestão e compreensão dos ecossistemas florestais do Estado. Os dados do IFC-RS apresentam amostras de diversos tipos florestais bem como as medidas de diversos atributos estruturais da vegetação. Os atributos estruturais da vegetação apresentam informações importantes para caracterizar uma floresta. O emprego desses atributos para o desenvolvimento da análise discriminante nos diferentes tipos de florestas do Rio Grande do Sul pode ser implantado em grande (biomas) ou médio escala (bacias hidrográficas) e, como visto nesse estudo, apresenta uma grande probabilidade de ser empregado com sucesso na classificação das florestas.

Inventários florestais estão sendo implantados no Brasil de forma corrente, todos com metodologia similar e com elevado rigor amostral (NETO & BRENA, 1993; CAMPOS & LEITE, 2002; SOARES *et al.*, 2006; FELFILI *et al.*, 2005; CARVALHO & SCOLFORO, 2008; IFSC, 2008). O emprego dos dados quantitativos para realização de análise

discriminante para classificação dessas florestas apresenta uma ampla opção para suprir a enorme carência de estudos que venham a elucidar as diferenças entre as tipologias de florestas e para incrementar informações sobre as fitofisionomias no sistema oficial da vegetação brasileira (IBGE, 2004).

Um dos grandes desafios da pesquisa ecológica e florestal atual se revela na necessidade de desenvolver métodos de manejo sustentado para as florestas naturais inequívocas, programas eficientes de recuperação dos ecossistemas degradados e desenvolver formas de quantificar os serviços ambientais prestados por cada um dos tipos de florestas ocorrentes. É inegável, para tanto, que se deve primeiramente acumular conhecimentos técnico-científicos sobre diferentes aspectos da vegetação e aprimorar as técnicas de análise de dados, principalmente delineando experimentos de campo onde possam ser empregadas técnicas multivariadas que permitam análise dos vários fatores responsáveis pela manutenção de um ecossistema florestal.



## 6 CONCLUSÕES

- Esse foi o primeiro estudo que utilizou a técnica de análise discriminante linear de Fisher para classificar tipos de florestas utilizando dados da estrutura dessas florestas no Rio Grande do Sul.
- A escolha dos grupos (tipos de florestas), das variáveis (atributos estruturais) e o tamanho amostral adotado para a realização da análise foram condizentes com as premissas descritas pela bibliografia especializada para desenvolver a análise discriminante. Assim, testes para comprovar a normalidade, a inexistência de multicolinearidade e a igualdade de matrizes de covariâncias foram empregadas para as variáveis preditoras e/ou grupos discriminantes.
- O modelo discriminante gerou quatro funções discriminantes para classificar os cinco tipos de florestas, sendo que, as duas primeiras funções apresentaram um poder de 88,7% de habilidade para discriminação dos grupos: FD1 = 74,4% ( $\lambda_{FD1} = 33,99$ ) e FD2 = 14% ( $\lambda_{FD2} = 6,34$ ). As outras funções apresentaram pouco poder de predição de grupos: FD3 = 8,9% ( $\lambda_{FD3} = 4,03$ ) e FD4 = 2,5% ( $\lambda_{FD4} = 1,11$ ). O método Stepwise selecionou, através da estatística lambda de Wilks, os três atributos alturas (comercial, total e máxima) como sendo as variáveis independentes mais importantes para o modelo.
- Todas as oito variáveis (atributos estruturais) inseridas na análise foram selecionadas pelo modelo. Os atributos estruturais que estiveram mais relacionados com a FD1 foram: riqueza de espécies, altura comercial e altura total. Na FD2 os atributos mais importantes foram a área basal e o diâmetro máximo atingido pelo caule.
- O modelo de Fisher classificou 100% dos casos ( $n_{\text{casos}}=50$ ) nos respectivos grupos ( $n_{\text{grupos}}=5$ ), revelando uma alta eficiência das variáveis discriminadoras.
- O estudo demonstrou que a análise discriminante linear de Fisher tem aplicações importantes na área da ecologia e fitogeografia e poderia ser empregada para elucidar melhor as questões de diferenciação de fitofisionomias vegetais nos diferentes biomas do País.
- O teste de validação cruzada com duas amostras sorteadas de cada tipo de floresta revelou que o modelo re-classificou corretamente a maioria dos casos em seus grupos de origem. Essa classificação foi possível porque o banco de dados do IFC-RS apresentou amostras de inventários em condições similares aos grupos de origem.

## 7 REFERÊNCIAS BIBLIOGRÁFICAS

- ARAÚJO, M.M. *et al.* Análise de agrupamento da vegetação de um fragmento de floresta estacional decidual aluvial, Cachoeira do Sul, RS, Brasil. **Ciência Florestal**, v.14, n.1, p.133-147, 2004.
- BEHLING, H.; PILLAR, V.D. Late Quaternary vegetation, biodiversity and fire dynamics on the southern Brazilian highland and their implication for conservation and management of modern Araucaria forest and grassland ecosystems. **Philosophical Transactions of the Royal Society of London (Biological Sciences)**, v. 362, p. 243-251, 2007.
- BROWN, M. T., & WICKER, L. R. Discriminant analysis. In TINSLEY, H.E.A. & BROWN, S.D. (Eds.). **Handbook of applied multivariate statistics and mathematical modeling**. San Diego: Academic Press, 2000. p. 209-234.
- BROWN, M.T. & TINSLEY, H.E.A. Discriminant Analysis. **Journal of Leisure Research**, v.15, n.4, p. 290-310, 1983.
- BARROS, M.L.B.; COUTO, R.G. **Atlas de conservação da natureza brasileira- Unidades Federais**. São Paulo: Metalivros, 2004. 289 p.
- CAMIZ, S.; PILLAR, V.D. 2007. Comparison of single and complete linkage clustering with the hierarchical factor classification of variables. **Community Ecology**, v.8, n.1, p. 25-30.
- CAMPOS, J.C.C.; LEITE, H.G. **Mensuração florestal**. Viçosa, MG: Editora UFV, 2002. 407 p.
- CARVALHO, L.M.T.; SCOLFORO, J.R.S. (Orgs.). Inventário Florestal de Minas Gerais - Monitoramento da Flora Nativa 2005-2007. 1. Lavras: Editora UFLA, 2008, v. 1. 357 p.
- DE MARCHI, T.C.; JARENKOW, J.A. Estrutura do componente arbóreo de mata ribeirinha no rio Camaquã, município de Cristal, Rio Grande do Sul, Brasil. **Iheringia (Série Botânica)**, v.63, p. 241-248, 2008.
- DORNELES, L.P.P.; WAECHTER, J.L. Fitossociologia do componente arbóreo na floresta turfosa do Parque Nacional da Lagoa do Peixe, Rio Grande do Sul, Brasil. **Acta Botanica Brasilica**, v.18, n.4, p. 815-824, 2004.
- GARSON, D. Discriminant Function Analysis, 2008. Disponível em: <http://faculty.chass.ncsu.edu/garson/PA765/discrim.htm>. Acesso em: 05 de fev. 2009.
- GAUCH, H.G. **Multivariate analysis in community ecology**. Cambridge: Cambridge University Press, 1982. 298 p.
- GERHARDT, E.J.; FINGER, C.A.G.; LONGHI, J.L.; SCHUMACHER, M.W. 2001. Contribuição da análise multivariada na classificação de sítios em povoamentos de Araucaria angustifolia (Bert.) O. Ktze., baseada nos fatores físicos e morfológicos do solo e no conteúdo de nutrientes da serrapilheira. **Ciência Florestal**, v.11, n.2, p. 41-57.

GOLDSMITH, F.B.; HARRISON, C.M. Description and Analysis of Vegetation. In: CHAPMAN, S.B. (Ed.) **Methods in Plant Biology**. Oxford: Blackwell Scientific Publ., 1976, p. 85-155.

HACK, C.; LONGHI, S.J.; BOLIGON, A.A.; MUNARI, A.B.; PAULESKI, D.T. Análise fitossociológica de um fragmento de floresta estacional decidual no município de Jaguari, RS. **Ciência Rural**, v.35, n.5, p.1083-1091, 2005.

HAIR Jr., ANDERSON, R.E., TATHAM, R.L.; BLACK, W.C. **Análise Multivariada de Dados**. 5. ed. Porto Alegre: Bookman, 2005. 600 p.

HÄRDLE, W. & SIMAR, L. **Applied multivariate statistical analysis**. Heidelberg: Springer Verlag, 2003, 486 p.

HUBERTA, M.; VAN DRIESSEN, K. Fast and robust discriminant analysis. **Computational statistics & data analysis**, v. 45, n.22, p. 301-320, 2004.

IBGE (Instituto Brasileiro de Geografia e Estatística). Mapa da vegetação do Brasil, 2004. Disponível em: [http://www.ibge.gov.br/home/presidencia/noticias/noticia\\_visualiza.php?id\\_noticia=169&id\\_pagina=1](http://www.ibge.gov.br/home/presidencia/noticias/noticia_visualiza.php?id_noticia=169&id_pagina=1). Acesso em: 10 de out. de 2004.

IFSC (INVENTÁRIO FLORÍSTICO FLORESTAL DO ESTADO DE SANTA CATARINA). Inventário e análise da estrutura horizontal e vertical dos remanescentes florestais do Estado de Santa Catarina, 2008. Disponível em: <http://www.acef.org.br/boletim02ifsc.pdf>. Acesso em: 03 março de 2009.

ITTO (International Tropical Timber Organization). For services rendered: The current status and future potential of markets for the ecosystem services provided by tropical forests. ITTO Technical Series 21. Yokohama: ITTO, 2002. 73 p.

FAO (Food and Agriculture Organization of the United Nations). State of World's Forests 2003. Disponível em: <http://www.fao.org>. Acesso em: 3 de março de 2005.

FELFILI, J.M.; CARVALHO, F.A. & HAIDAR, R.F. Manual para o Monitoramento de parcelas permanentes nos biomas Cerrado e Pantanal. Brasília:UnB, 2005. 30p.

GUARIGUATA, M.R.; OSTERTAG, R. Neotropical secondary forest succession: changes in structural and functional characteristics. **Forest Ecology and Management**, v.148, p.185-206, 2001.

JOHNSON, R. A. & WICHERN, D. W. 1999. **Applied multivariate statistical analysis**. 4th ed. Upper Saddle River, New Jersey: Prentice-Hall. 815 p.

JURINITZ, C. F.; JARENKOW, J. A. Estrutura do componente arbóreo de uma floresta estacional na Serra do Sudeste, Rio Grande do Sul, Brasil. **Revista Brasileira de Botânica**, v. 26, n.4, p. 475-487, 2003.

KHATTREE, R.; NAIK, D.N. **Multivariate data reduction and discrimination with SAS software**. Cary, SAS Institute Inc., 2000. 558 p.

KENKEL, N.C.; DERKSEN, D.A; THOMAS, A.G.; WATSON, P.R. Multivariate analysis in weed science research. **Weed Science**, v.50, p. 281-292, 2002.

- KILCA, R.V.; LONGHI, S.J.; BRENA, D.A. A composição florística e a estrutura das florestas subtropicais estacionais secundárias nas escarpas do Planalto Meridional do Rio Grande do Sul. In: SCHUMACKER, M.V.; LONGHI, S.J.; BRUN, E.J. (Orgs) **Estudos ecológicos na floresta estacional decidual do Rio Grande do Sul**. No prelo.
- KLECKA, W.R. Discriminant Analysis. In: NIE, N.H.; *et al.* (Eds). **Statistical Package for the Social Sciences**. 2nd. New York: McGraw-Hill, Ch. 23, p. 434-467, 1975.
- KLEIN, R.M. Aspectos dinâmicos da vegetação do sul do Brasil. **Sellowia**, v.36, p. 5-54, 1984.
- LEITE, P. F. Contribuição ao conhecimento fitoecológico do sul do Brasil. **Ciência & Ambiente**, v.24, p. 51-73, 2002.
- LEGENDRE, L.; P. LEGENDRE. **Numerical ecology**. Amsterdam: Elsevier, 1998. 853 p.
- LIANG, Z.; ZHANG, D.; SHI, P. Robust kernel discriminant analysis and its application to feature extraction and recognition. **Neurocomputing**, v.69, p. 928-933, 2006.
- LONGHI, S.J. Aspectos fitossociológicos dos capões na região de Carovi e Tupantuba, em Santiago, RS. **Ciência Florestal**, v.1, n.1, p. 22-39, 1991.
- LONGHI, S.J.; GALVÃO, F.; CHAVES NETO, A.; KUNIYOSHI, Y.S. Aplicação da análise multivariada no agrupamento de comunidades florestais. **Cadernos da Biodiversidade**, v.2, n.2, p. 47-60, 1999.
- LONGHI, S.J. Agrupamento e análise fitossociológica de comunidades florestais na sub-bacia hidrográfica do Rio Passo Fundo - RS. 1997, 193 f. Tese (Doutorado em Ciências Florestais) – Setor de Ciências Agrárias, Universidade Federal do Paraná, Curitiba, 1997.
- McLACHLAN, G.F. **Discriminant Analysis and Statistical Pattern Recognition**. 2nd. Hoboken: John Wiley & Sons, Inc. 526 p.
- MANLY, B.F.J. 2008. **Métodos estatísticos multivariados: uma introdução**. 3ed. Porto Alegre: Bookman. 229 p.
- MARCHIORI, J.N.C.; LONGHI, S.J.; GALVÃO, L. Composição florística e estrutura do Parque de Inhanduvá no Rio Grande do Sul. **Ciência Rural**, v.15, n.4, p. 319-334, 1984.
- MARCHIORI, J.N.C.; LONGHI, S.J.; GALVÃO, L. Estrutura fitossociológica de uma associação natural de Parque Inhanduvá com Quebracho e Cina-cina, no Rio Grande do Sul. **Ciência e Natura**, v.7, n.1, p. 147-162, 1985.
- MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005. 297 p.
- MMA (MINISTÉRIO DO MEIO AMBIENTE). **Avaliação e ações prioritárias para a conservação da biodiversidade da Mata Atlântica e Campos Sulinos**. Brasília: MMA/SBF, 2000a. 40 p.
- MMA (MINISTÉRIO DO MEIO AMBIENTE). **Biodiversidade Brasileira: Avaliação e identificação de áreas e ações prioritárias para a conservação, utilização sustentável e**

repartição dos benefícios da biodiversidade nos biomas brasileiros. MMA/SBF: Brasília, 2002. 404 p.

MUELLER-DOMBOIS, D.; ELLENBERG, H. **Aims and methods of vegetation ecology**. New York: Wiley, 1974. 547 p.

NETTO, S.P.; BRENA, D.A. **Inventário florestal**. Curitiba: UFPR, 1993. 245 p.

OLIVEIRA, J.M.; PILLAR, V.D. Vegetation dynamics on mosaics of Campos and Araucaria forest between 1974 and 1999 in Southern Brazil. **Community Ecology**, v.5, n.2, p. 197-202, 2004.

OUTPUT DISCRIMINANT ANALYSIS. In: discriminant function analysis with three or more groups with SPSS, 2006. Disponível em: [https://www.u-cursos.cl/ingenieria/2008/1/IN540/2/material\\_docente/previsualizar?id\\_material=164503](https://www.u-cursos.cl/ingenieria/2008/1/IN540/2/material_docente/previsualizar?id_material=164503). Acesso em: 6 maio de 2009.

OUTPUT DISCRIMINANT ANALYSIS. In: UCLA - Annotated SPSS Output: Discriminant Analysis, 2007. Disponível em: [https://www.ats.ucla.edu/stat/spss/output/SPSS\\_discrim.htm](https://www.ats.ucla.edu/stat/spss/output/SPSS_discrim.htm). Acesso em: 12 nov. 2008.

OUTPUT DISCRIMINANT ANALYSIS. New Castle University - How to Perform and Interpret Discriminant Analysis (DA), 2006. Disponível em: <http://www.ncl.ac.uk/iss/statistics/docs/discriminant.php>. Acesso em: 23 de março de 2009.

OUTPUT DISCRIMINANT ANALYSIS. Utrecht University - SPSS - Discriminant Analysis. Disponível em: <http://www.cs.uu.nl/docs/vakken/arm/SPSS/spss6.pdf>. Acesso em: 06 de maio de 2009.

PARUELO, J. M.; LAUENROTH, W. K.; EPSTEIN, H. E.; BURKE, I. C.; AGUIAR, M. R.; SALA, O. E. Regional climatic similarities in the temperate zones of north and south America. **Journal of Biogeography**, v.22, n.4/5, p. 915-925, 1995.

PILLAR, V.D.; SOSINSKI JR., E.E. An improved method for searching plant functional types by numerical analysis. **Journal of Vegetation Science**, v.14, p. 323-332, 2003.

PILLAR, V.D.; ORLÓCI, L. **Character-Based Community Analysis: The Theory and an Application Program**, 1993. SPB Academic Publishing, The Hague. Disponível em: <http://ecoqua.ecologia.ufrgs.br>. Acesso em: 12 de março de 2007.

PIELOU, E.C. **Ecological diversity**. New York: John Wiley & Sons, 1975. 165 p.

RAMBO, B. **A fisionomia do Rio Grande do Sul**. Porto Alegre: Selbach, 1956. 417 p.

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

REITZ, R.; KLEIN, R.M.; REIS, A. Projeto Madeira do Rio Grande do Sul. **Sellowia**, v.34/35, p. 5-483, 1983.

RENCHEER, A.C. **Methods of Multivariate Analysis**. New York: Wiley, 2002. 708 p.

RICHARDS, P.W. **Tropical rain forest -an ecological study**. Cambridge: Cambridge Univ. Press, 1952. 450 p.

- RIZZINI, C.T. **Tratado de fitogeografia do Brasil**, 2 ed. Rio de Janeiro: Ambito Cultural Edições Ltda., 1997. 747 p.
- ROSA, S.F.; LONGHI, S.J.; LUDWIG, M.P. Aspectos florísticos e fitossociológicos da Reserva Capão de Tupanciretã, Tupanciretã, RS, Brasil. **Ciência Florestal**, v.18, p. 15-25, 2008.
- SAGAR, R.; SING, J.S. Tree density, basal area and species diversity in a disturbed dry tropical forest of northern India: implication for conservation. **Environment and Conservation**, v.33, n.3, p. 256-262, 2006.
- SEMA/UFMS (SECRETARIA ESTADUAL DO MEIO AMBIENTE/UNIVERSIDADE FEDERAL DE SANTA MARIA). Inventário Florestal Contínuo do Rio Grande do Sul, 2002. Disponível em: <http://coralx.ufsm.br/ifcrs/>. Acesso em: 20 de fev. de 2006.
- SHIN, P. K. S.; FONG, K. Y. S. Multiple Discriminant Analysis of Marine Sediment Data. **Marine Pollution Bulletin**, v.39, p. 285-294, 1999.
- SHIN, H. An extension of Fisher's discriminant analysis for stochastic processes. **Journal of Multivariate Analysis**, v.99, p. 1191-1216, 2008.
- SOUZA, D.R.; SOUZA, A.L.; GAMA, J.R.V.; LEITE, H.G. Emprego de análise multivariada para estratificação vertical de florestas ineqüiâneas. **Revista Árvore**, v.27, n.1, p. 59-63, 2003.
- SMITH, J; SABOGAL, C.; DE JONG, W.; KAIMOWITZ, D. Bosques secundarios como recurso para el desarrollo rural y la conservación ambiental en los Trópicos de América Latina. Pucallpa, Perú, CIFOR., 1997. 31 p. Disponível em: [http://www.cifor.cgiar.org/publications/pdf\\_files/OccPapers/OP-13.pdf](http://www.cifor.cgiar.org/publications/pdf_files/OccPapers/OP-13.pdf). Acesso em: 5 de out. de 2006.
- SOARES, C.P.B., NETO, F.P.; SOUZA, A.L. **Dendrometria e Inventário Florestal**. Viçosa: Ed. UFV, 2006. 276 p.
- SPSS (Statistical Package for the Social Sciences). **SPSS Base 15.0 User's Guide**. Chicago: SPSS Inc., 2006. 779 p.
- STEVENS, J.P. **Applied multivariate statistics for the social sciences**. 3rd. Mahwah: Laurence Erlbarum, 1996.
- TABACHNICK, B.G.; FIDELL, L.S. **Using Multivariate Statistics**. 4. ed. Needham Heights, MA: Allyn & Bacon, 2001.
- VALENTIM, J.L. Ecologia Numérica: **Uma Introdução á Análise Multivariada de Dados Ecológicos**. Editora Interciência. Rio de Janeiro, 2000. 117 p.
- WAECHTER, J.L.; JARENKOW, J.A. Composição e estrutura do componente arbóreo nas matas turfosas do Taim, Rio Grande do Sul. **Biotemas**, v.11, n.1, p. 45-69, 1998.
- WAECHTER, J.L. Aspectos ecológicos da vegetação de restinga no Rio Grande do Sul, Brasil. **Comunicações do Museu de Ciências da PUCRS** (sér. Botânica), v.33, p. 49-68, 1985.

**ANEXOS**

**ANEXO A - Banco de dados com valores originais que foi utilizado para a geração da  
análise discriminante linear de Fisher**





**ANEXO B - Banco de dados com valores originais dos atributos estruturais das unidades amostrais que foram utilizadas para o teste de validação cruzada do modelo**

Características estruturais das unidades amostrais (1.000m<sup>2</sup>) das tipologias florestais inseridas no modelo para o teste de classificação de novos casos nos grupos formados.

COD.	TIPO	N	RIQ	ALT COM	ALT TOT	ALT MAX	AB	DAP	DAP MAX
IFC-RS	FLORESTAL								
1605	SAVANA 1	14	3	2,95	4,7500	9,5	0,404	16,8500	47,8400
1605	SAVANA 2	13	4	3,192	5,907	9,5	0,574	21,71	41,57
1819	CILIAR 1	101	18	6,936	9,977	17,895	2,56	16,712	42,34
1903	CILIAR 2	101	17	5,73	9,032	15,5	2,365	16,05	46,15
2704	GALERIA 1	136	13	4,57	8,379	14,8	2,615	12,83	27,77
2406	GALERIA 2	124	19	3,61	8,539	17,6	2,535	14,697	49,02
1943	FES 1	83	22	7,393	10,52	29,3	2,74	18,261	47,43
1946	FES 2	83	23	6,84	12,014	20,89	3,703	20,733	53,479
2803	RESTINGA 1	49	10	2,685	8,757	14,3	2,376	21,639	65,029
2101	RESTINGA 2	91	16	3,84	9,503	17,5	3,037	18,143	40,74

Foram escolhidas unidades amostrais com características ambientais similares às florestas inseridas preliminarmente inseridas no modelo todas essas correspondem a florestas preservadas ou em estágio avançado de sucessão. Os cálculos médios ou totais de cada atributo foram realizados em cada unidade amostral de 1.000m<sup>2</sup> de floresta. As variáveis descritoras foram: N (número total de indivíduos amostrados), RIQ (riqueza ou número total de espécies amostradas), ALTCOM (altura média comercial em metros), ALTTOT (altura média total em metros), ALTMAX (altura máxima atingida na área amostral em metros), AB (somatório da área basal de todos os indivíduos em m<sup>2</sup>), DAP (média dos diâmetros dos caules amostrados em centímetros), DAPMAX (diâmetro máximo do caule na área amostral em centímetros).

**ANEXO C - Resultados dos testes de normalidade para as variáveis independentes por grupo (valores apenas da probabilidade dos testes)**

	Savana arbórea			Fl. restinga		
<b>Variável Independente</b>	<b>K-S</b>	<b>S-W</b>	<b>Lillief</b>	<b>K-S</b>	<b>S-W</b>	<b>Lillief</b>
N de indivíduos	>0,2	0,74	>0,2	>0,2	0,13	>0,2
Riqueza de esp.	>0,2	0,04	<0,01	>0,2	0,02	<0,2
Altura total	>0,2	0,002	<0,05	<0,05	0,17	>0,2
Altura comercial	>0,2	0,43	>0,2	>0,2	<0,01	<0,01
Altura máxima	>0,2	0,04	<0,01	>0,2	0,007	<0,05
DAP	>0,2	0,007	<0,05	>0,2	0,23	>0,2
DAP máximo	>0,2	0,43	>0,2	>0,2	0,37	>0,2
Área basal	>0,2	0,73	>0,2	>0,2	0,36	<0,2
	Fl. galeria			Fl. ciliar		
<b>Variável Independente</b>	<b>K-S</b>	<b>S-W</b>	<b>Lillief</b>	<b>K-S</b>	<b>S-W</b>	<b>Lillief</b>
N de indivíduos	>0,2	0,63	<0,2	>0,2	0,66	>0,2
Riqueza de esp.	>0,2	0,08	<0,2	>0,2	0,5	>0,2
Altura total	>0,2	0,005	<0,02	>0,2	0,19	>0,2
Altura comercial	>0,2	0,45	>0,2	>0,2	0,002	<0,05
Altura máxima	>0,2	0,96	>0,2	>0,2	0,24	>0,2
DAP	>0,2	0,005	>0,2	>0,2	0,14	<0,15
DAP máximo	>0,2	0,009	>0,2	>0,2	0,25	>0,2
Área basal	>0,2	0,54	>0,2	>0,2	0,55	<0,1
	Fl. Estacional semidecidual					
<b>Variável Independente</b>	<b>K-S</b>	<b>S-W</b>	<b>Lillief</b>			
N de indivíduos	>0,2	0,92	>0,2			
Riqueza de esp.	>0,2	0,03	<0,01			
Altura total	>0,2	0,71	>0,2			
Altura comercial	>0,2	0,85	>0,2			
Altura máxima	>0,2	0,021	<0,05			
DAP	>0,2	0,93	>0,2			
DAP máximo	>0,2	0,83	>0,2			
Área basal	>0,2	0,98	>0,2			

Testes: Kolmogorov-Smirnov (K-S), Shapiro-Willk (S-W) e Lilliefors (Lillief).

**ANEXO D - Resultados da análise descritiva para as variáveis independentes por grupo**

Estadística Descriptiva ( <b>SAVANA ARBÓREA</b> )							
	Valid N	Mean	Median	Minimum	Maximum	Variance	Std.Dev.
N	10	14,100	15,500	2,000	22,000	35,433	5,953
RIQ	10	1,800	1,500	1,000	3,000	0,844	0,919
ALTCOM	10	1,910	1,850	1,650	2,550	0,059	0,243
ALT TOT	10	3,570	3,520	3,000	4,550	0,195	0,441
ALT MAX	10	7,720	8,195	5,190	9,500	3,105	1,762
AB	10	0,490	0,490	0,215	0,847	0,040	0,201
DAP	10	17,740	17,135	11,840	27,585	19,819	4,452
DAP MAX	10	36,000	36,795	21,450	47,840	63,249	7,953
Estadística Descriptiva ( <b>RESTINGA ARENOSA</b> )							
	Valid N	Mean	Median	Minimum	Maximum	Variance	Std.Dev.
N	10	73,000	76,000	57,000	92,000	160,222	12,658
RIQ	10	14,300	15,000	11,000	18,000	4,900	2,214
ALTCOM	10	3,710	3,825	2,900	4,350	0,222	0,471
ALT TOT	10	9,385	9,300	8,850	10,250	0,158	0,397
ALT MAX	10	17,294	16,895	13,890	20,000	3,826	1,956
AB	10	4,110	4,097	2,461	5,467	0,925	0,961
DAP	10	13,497	13,320	12,100	15,630	1,020	1,010
DAP MAX	10	101,77	98,51	87,72	132,42	280,40	16,74
Estadística Descriptiva ( <b>FLORESTA DE GALERIA</b> )							
	Valid N	Mean	Median	Minimum	Maximum	Variance	Std.Dev.
N	10	135,200	133,500	116,000	165,000	247,511	15,732
RIQ	10	17,900	18,500	15,000	21,000	3,878	1,969
ALTCOM	10	5,570	5,500	5,250	6,000	0,071	0,266
ALT TOT	10	9,275	9,200	8,800	10,000	0,184	0,429
ALT MAX	10	17,464	17,645	13,890	20,790	4,452	2,110
AB	10	2,260	2,230	1,788	2,770	0,104	0,322
DAP	10	12,088	11,993	11,680	12,670	0,133	0,364
DAP MAX	10	42,042	44,270	28,890	59,650	96,264	9,811
Estadística Descriptiva ( <b>FLORESTA CILIAR</b> )							
	Valid N	Mean	Median	Minimum	Maximum	Variance	Std.Dev.
N	10	90,400	84,500	56,000	129,000	458,267	21,407
RIQ	10	16,300	15,500	12,000	22,000	13,344	3,653
ALTCOM	10	5,125	5,150	4,000	6,000	0,254	0,504
ALT TOT	10	8,580	9,050	6,200	9,800	1,640	1,280
ALT MAX	10	14,939	14,400	12,500	19,100	3,576	1,891
AB	10	2,992	3,134	1,432	4,119	0,722	0,855
DAP	10	14,443	14,135	12,860	16,680	1,395	1,181
DAP MAX	10	62,025	56,980	36,930	94,220	403,443	20,086

Estatística Descritiva (FLORESTA ESTACIONAL SEMIDECIDUAL)							
	Valid N	Mean	Median	Minimum	Maximum	Variance	Std.Dev.
N	10	87,10	86,50	61,00	111,00	193,43	13,90
RIQ	10	21,40	22,00	17,00	24,00	4,93	2,22
ALTCOM	10	4,65	4,60	4,00	5,40	0,21	0,46
ALT TOT	10	9,51	9,27	8,20	11,10	0,74	0,86
ALT MAX	10	25,31	25,79	21,29	29,39	12,01	3,46
AB	10	3,18	3,18	2,76	3,64	0,074	0,27
DAP	10	15,29	15,13	13,34	17,35	1,38	1,17
DAP MAX	10	57,97	58,96	45,20	75,26	77,35	8,795

As variáveis descritoras foram: N (número médio de indivíduos amostrados), RIQ (riqueza ou número de espécies amostradas), ALTCOM (altura comercial em metros), ALTTOT (altura total em metros), ALTMAX (altura máxima atingida na área amostral em metros), AB (área basal em m<sup>2</sup>), DAP (diâmetros dos caules amostrados em centímetros), DAPMAX (diâmetro máximo do caule na área amostral em centímetros).



**ANEXO E - Resultados dos testes ANOVA um fator, seguido dos testes de Tukey, para as oito variáveis discriminantes nos cinco grupos**

## VAR. IND.: NÚMERO DE INDIVÍDUOS

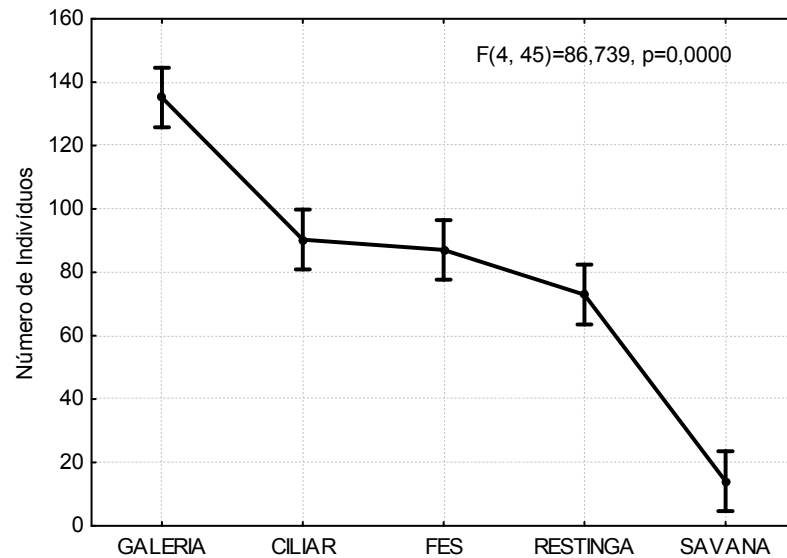


Gráfico demonstrando os valores do atributo número de indivíduos amostrados nas 10 unidades amostrais de 1000 m<sup>2</sup> em cada uma dos cinco tipos de florestas no Pampa do Rio Grande do Sul. Os valores estão representados pela média e o intervalo 95% de confiança da média (barras verticais). Onde os tipos de florestas foram: floresta de galeria (GALERIA), floresta ciliar (CILIAR), floresta estacional semidecidual (FES), floresta restinga arenosa (RESTINGA) e Savana arbórea (SAVANA). O resultado do teste ANOVA de um fator encontra-se na parte interior do gráfico.

Tabela do teste de Tukey demonstrando o valor da probabilidade para a diferença significativa das médias do número de indivíduos para os cinco tipos de florestas.

Tukey HSD test = 218,97, df = 45,000						
	Var1	{1}	{2}	{3}	{4}	{5}
1	GALERIA		<b>0,000134</b>	<b>0,000134</b>	<b>0,000134</b>	<b>0,000134</b>
2	CILIAR			0,987140	0,081976	<b>0,000134</b>
3	FES				0,225393	<b>0,000134</b>
4	RESTINGA					<b>0,000134</b>
5	SAVANA					

Os valores do teste em negrito são significativos <0,05.

### VARIÁVEL PREDITORA: NÚMERO DE ESPÉCIES

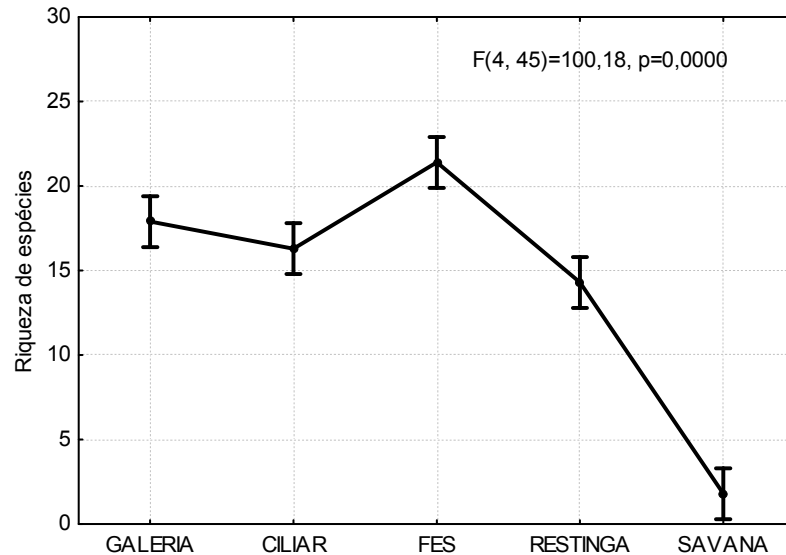


Gráfico demonstrando os valores do atributo número de indivíduos amostrados nas 10 unidades amostrais de 1000 m<sup>2</sup> em cada uma dos cinco tipos de florestas no Pampa do Rio Grande do Sul. Os valores estão representados pela média e o intervalo 95% de confiança da média (barras verticais). Onde os tipos de florestas foram: floresta de galeria (GALERIA), floresta ciliar (CILIAR), floresta estacional semidecidual (FES), floresta restinga arenosa (RESTINGA) e Savana arbórea (SAVANA). O resultado do teste ANOVA de um fator encontra-se na parte interior do gráfico.

Tabela do teste de Tukey demonstrando o valor da probabilidade para a diferença significativa das médias do número de indivíduos para os cinco tipos de florestas.

Tukey HSD test = 5,5800, df = 45,000						
	Var1	{1}	{2}	{3}	{4}	{5}
1	GALERIA		0,558735	<b>0,015057</b>	<b>0,011659</b>	<b>0,000134</b>
2	CILIAR			<b>0,000286</b>	0,335770	<b>0,000134</b>
3	FES				<b>0,000134</b>	<b>0,000134</b>
4	RESTINGA					<b>0,000134</b>
5	SAVANA					

Os valores do teste em negrito são significativos <0,05.

### VARIÁVEL PREDITORA: ALTURA COMERCIAL (m)

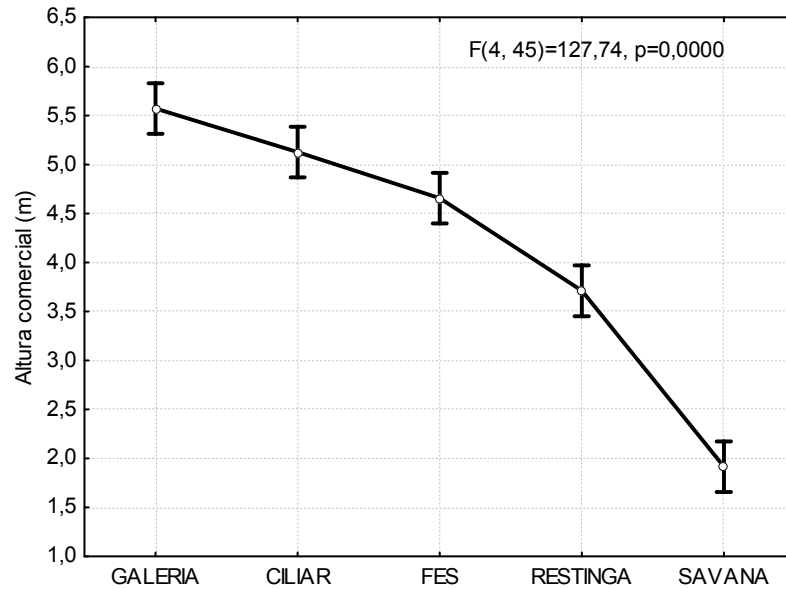


Gráfico demonstrando os valores do atributo número de indivíduos amostrados nas 10 unidades amostrais de 1000 m<sup>2</sup> em cada uma dos cinco tipos de florestas no Pampa do Rio Grande do Sul. Os valores estão representados pela média e o intervalo 95% de confiança da média (barras verticais). Onde os tipos de florestas foram: floresta de galeria (GALERIA), floresta ciliar (CILIAR), floresta estacional semidecidual (FES), floresta restinga arenosa (RESTINGA) e Savana arbórea (SAVANA). O resultado do teste ANOVA de um fator encontra-se na parte interior do gráfico.

Tabela do teste de Tukey demonstrando o valor da probabilidade para a diferença significativa das médias do número de indivíduos para os cinco tipos de florestas.

Tukey HSD test; = ,16442, df = 45,000						
	Var1	{1}	{2}	{3}	{4}	{5}
1	GALERIA		0,119908	<b>0,000213</b>	<b>0,000134</b>	<b>0,000134</b>
2	CILIAR			0,089106	<b>0,000134</b>	<b>0,000134</b>
3	FES				<b>0,000185</b>	<b>0,000134</b>
4	RESTINGA					<b>0,000134</b>
5	SAVANA					

Os valores do teste em negrito são significativos <0,05.

### VARIÁVEL PREDITORA: ALTURA TOTAL (m)

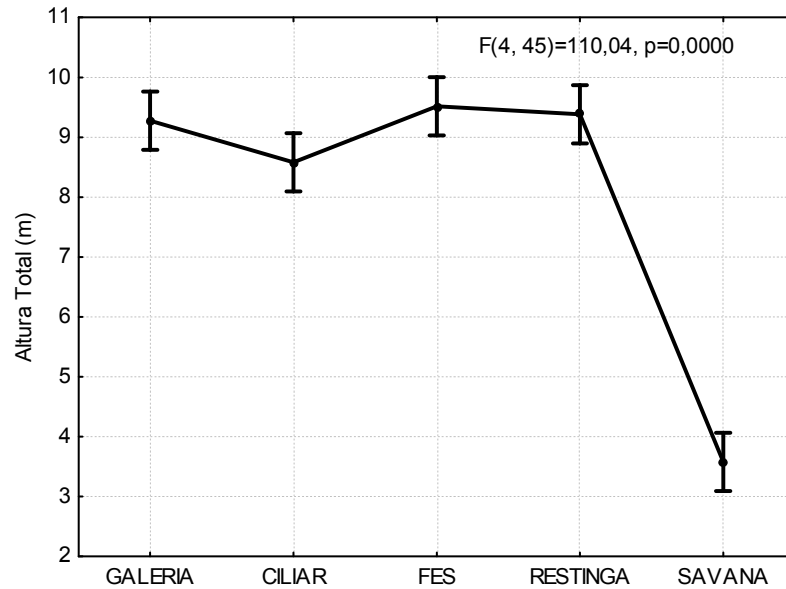


Gráfico demonstrando os valores do atributo número de indivíduos amostrados nas 10 unidades amostrais de 1000 m<sup>2</sup> em cada uma dos cinco tipos de florestas no Pampa do Rio Grande do Sul. Os valores estão representados pela média e o intervalo 95% de confiança da média (barras verticais). Onde os tipos de florestas foram: floresta de galeria (GALERIA), floresta ciliar (CILIAR), floresta estacional semidecidual (FES), floresta restinga arenosa (RESTINGA) e Savana arbórea (SAVANA). O resultado do teste ANOVA de um fator encontra-se na parte interior do gráfico.

Tabela do teste de Tukey demonstrando o valor da probabilidade para a diferença significativa das médias do número de indivíduos para os cinco tipos de florestas.

Tukey HSD test = ,58464, df = 45,000						
	Var1	{1}	{2}	{3}	{4}	{5}
1	GALERIA		0,267632	0,955064	0,997641	<b>0,000134</b>
2	CILIAR			0,064527	0,147185	<b>0,000134</b>
3	FES				0,995481	<b>0,000134</b>
4	RESTINGA					<b>0,000134</b>
5	SAVANA					

Os valores do teste em negrito são significativos <0,05.

### VARIÁVEL PREDITORA: ALTURA MÁXIMA ATINGIDA (m)

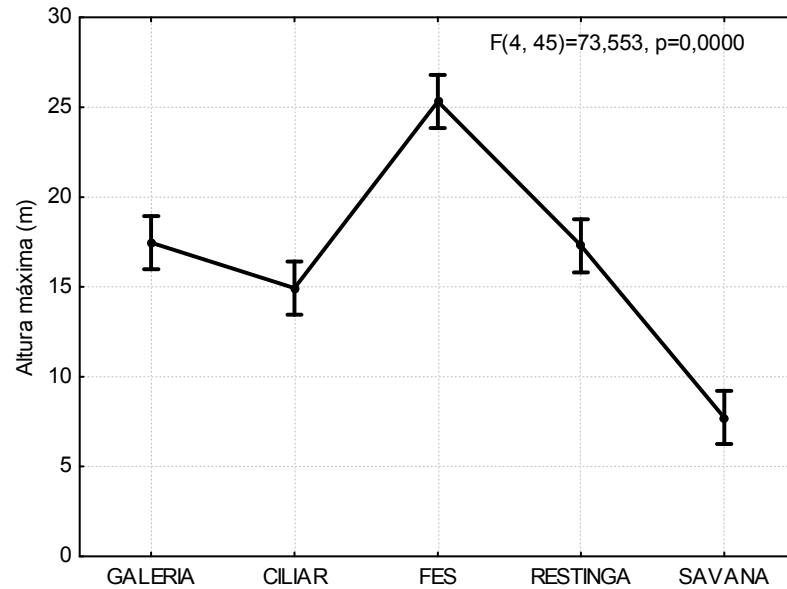


Gráfico demonstrando os valores do atributo número de indivíduos amostrados nas 10 unidades amostrais de 1000 m<sup>2</sup> em cada uma dos cinco tipos de florestas no Pampa do Rio Grande do Sul. Os valores estão representados pela média e o intervalo 95% de confiança da média (barras verticais). Onde os tipos de florestas foram: floresta de galeria (GALERIA), floresta ciliar (CILIAR), floresta estacional semidecidual (FES), floresta restinga arenosa (RESTINGA) e Savana arbórea (SAVANA). O resultado do teste ANOVA de um fator encontra-se na parte interior do gráfico.

Tabela do teste de Tukey demonstrando o valor da probabilidade para a diferença significativa das médias do número de indivíduos para os cinco tipos de florestas.

Tukey HSD test = 5,3949, df = 45,000						
	Var1	{1}	{2}	{3}	{4}	{5}
1	GALERIA		0,125842	<b>0,000134</b>	0,999849	<b>0,000134</b>
2	CILIAR			<b>0,000134</b>	0,174740	<b>0,000134</b>
3	FES				<b>0,000134</b>	<b>0,000134</b>
4	RESTINGA					<b>0,000134</b>
5	SAVANA					

Os valores do teste em negrito são significativos <0,05.

### VARIÁVEL PREDITORA: AREA BASAL (m<sup>2</sup>)

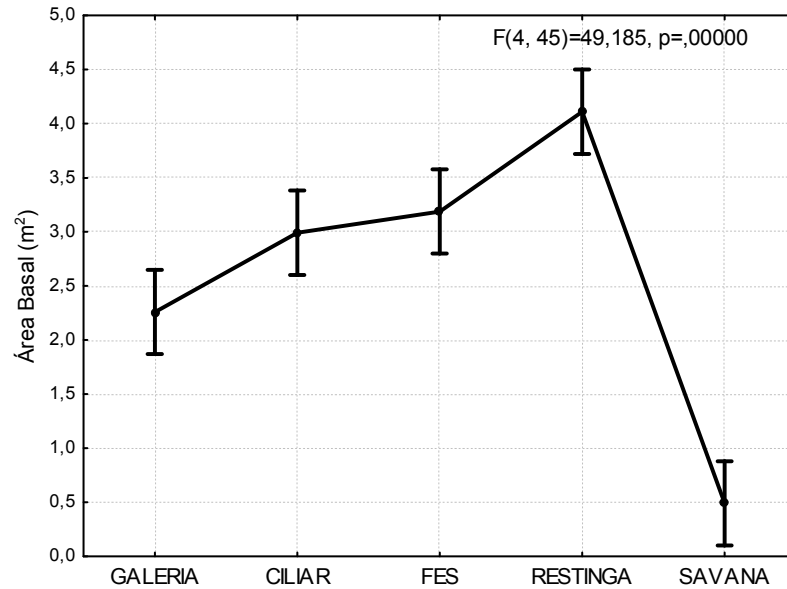


Gráfico demonstrando os valores do atributo número de indivíduos amostrados nas 10 unidades amostrais de 1000 m<sup>2</sup> em cada uma dos cinco tipos de florestas no Pampa do Rio Grande do Sul. Os valores estão representados pela média e o intervalo 95% de confiança da média (barras verticais). Onde os tipos de florestas foram: floresta de galeria (GALERIA), floresta ciliar (CILIAR), floresta estacional semidecidual (FES), floresta restinga arenosa (RESTINGA) e Savana arbórea (SAVANA). O resultado do teste ANOVA de um fator encontra-se na parte interior do gráfico.

Tabela do teste de Tukey demonstrando o valor da probabilidade para a diferença significativa das médias do número de indivíduos para os cinco tipos de florestas.

Tukey HSD test = ,37356, df = 45,000						
	Var1	{1}	{2}	{3}	{4}	{5}
1	GALERIA		0,073253	<b>0,012068</b>	<b>0,000134</b>	<b>0,000134</b>
2	CILIAR			0,951808	<b>0,001684</b>	<b>0,000134</b>
3	FES				<b>0,012706</b>	<b>0,000134</b>
4	RESTINGA					<b>0,000134</b>
5	SAVANA					

Os valores do teste em negrito são significativos <0,05.

### VARIÁVEL PREDITORA: DIÂMETRO DO CAULE (cm)

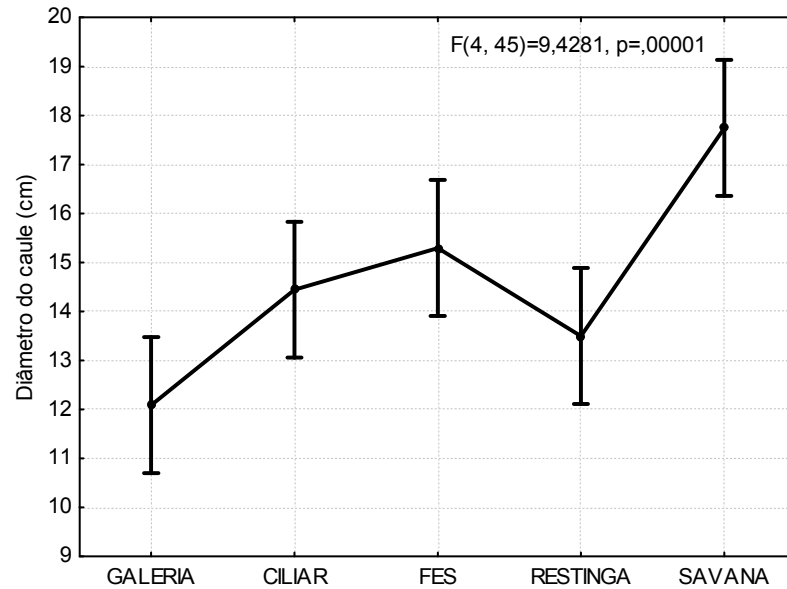


Gráfico demonstrando os valores do atributo número de indivíduos amostrados nas 10 unidades amostrais de 1000 m<sup>2</sup> em cada uma dos cinco tipos de florestas no Pampa do Rio Grande do Sul. Os valores estão representados pela média e o intervalo 95% de confiança da média (barras verticais). Onde os tipos de florestas foram: floresta de galeria (GALERIA), floresta ciliar (CILIAR), floresta estacional semidecidual (FES), floresta restinga arenosa (RESTINGA) e Savana arbórea (SAVANA). O resultado do teste ANOVA de um fator encontra-se na parte interior do gráfico.

Tabela do teste de Tukey demonstrando o valor da probabilidade para a diferença significativa das médias do número de indivíduos para os cinco tipos de florestas.

Tukey HSD test = 4,7504, df = 45,000						
	Var1	{1}	{2}	{3}	{4}	{5}
1	GALERIA		0,129750	<b>0,016053</b>	0,602600	<b>0,000137</b>
2	CILIAR			0,905324	0,866768	<b>0,012365</b>
3	FES				0,361975	0,105796
4	RESTINGA					<b>0,000819</b>
5	SAVANA					

Os valores do teste em negrito são significativos <0,05.



### VARIÁVEL PREDITORA: DIÂMETRO MÁXIMO ATINGIDO PELO CAULE (cm)

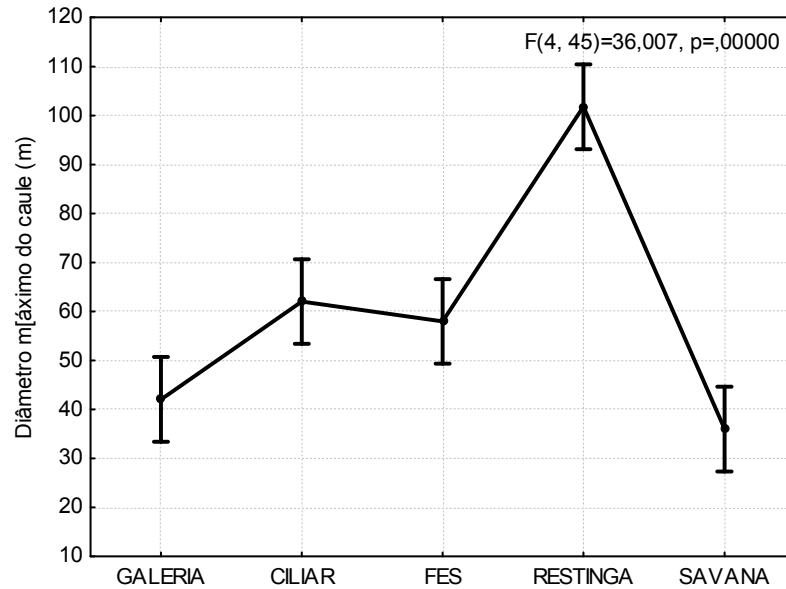


Gráfico demonstrando os valores do atributo número de indivíduos amostrados nas 10 unidades amostrais de 1000 m<sup>2</sup> em cada uma dos cinco tipos de florestas no Pampa do Rio Grande do Sul. Os valores estão representados pela média e o intervalo 95% de confiança da média (barras verticais). Onde os tipos de florestas foram: floresta de galeria (GALERIA), floresta ciliar (CILIAR), floresta estacional semidecidual (FES), floresta restinga arenosa (RESTINGA) e Savana arbórea (SAVANA). O resultado do teste ANOVA de um fator encontra-se na parte interior do gráfico.

Tabela do teste de Tukey demonstrando o valor da probabilidade para a diferença significativa das médias do número de indivíduos para os cinco tipos de florestas.

Tukey HSD test = 184,14, df = 45,000						
	Var1	{1}	{2}	{3}	{4}	{5}
1	GALERIA		<b>0,015897</b>	0,082814	<b>0,000134</b>	0,856003
2	CILIAR			0,962286	<b>0,000134</b>	<b>0,000981</b>
3	FES				<b>0,000134</b>	<b>0,006477</b>
4	RESTINGA					<b>0,000134</b>
5	SAVANA					

Os valores do teste em negrito são significativos <0,05.