METODOLOGIA PARA DEFINIR GRUPOS HOMOGÊNEOS DE PROPRIEDADES RURAIS[1]

HUMBERTO ANGELO[2], LUIZ HERNAN RODRIGUES CASTRO[3]
e ROBERTO TUYOSHI HOSOKAWA[4]

RESUMO – Buscou-se, com este trabalho, conhecer os parâmetros: mata nativa, capoeira, reflorestamento, agricultura de subsistência, pastagem e população rural de 110 imóveis de tamanho variando de 26 a 52 ha, localizados no município de Porto Vitória, Estado do Paraná, Brasil, com o objetivo de propor uma metodologia para identificar grupos homogêneos de imóveis e as relações dos referidos parâmetros, em cada grupo. Utilizou-se como base metodológica a Análise dos Componentes Principais, seguida da aplicação do método de "Cluster Analysis" e Análise de Variância. Os resultados obtidos com a aplicação da metodologia, permitiram constatar, estatisticamente, seis grupos homogêneos de propriedades rurais significativamente heterogêneos entre si. Em geral, verificou-se relações significativas entre os parâmetros em cada grupamento de propriedades.

Termos para indexação: componentes principais, conglomerados, propriedade rural, cobertura florestal.

A METHODOLOGY OF DEFINE HOMOGENOUS GROUPS OF FARMS

ABASTRACT – Data from 110 farms located in the country of Porto Vitória, State of Paraná - Brazil, has been collected according to the following variables: natural forest, secondary forest, forest plantations, agriculture (subsistence crops), pastures and rural population. These variables were used to identify homogeneous groups of farms. Principal Component Analysis has been initially used, followed by cluster analysis and analysis of variance. Through these methodologies it was possible to establish six homogeneous groups of farms with significant differences among then.

Index terms: principal component, clusters, farms, forest cover.

INTRODUÇÃO

As propriedades rurais constituem a célula do desenvolvimento econômico e social, dada sua relevância na produção de gêneros alimentícios, na fixação do homem no campo, na geração de empregos e rendas no meio-rural. No aspecto ecológico, contribuem na preservação e conservação do meio a partir do momento que harmonizam suas atividades agropecuárias com as florestas e as matêm dentro do seu limite.

Diante do exposto, torna-se importante o estudo das relações entre os fatores: cobertura florestal, agricultura de subsistência, pecuária e população rural, sem desconhecer que outros influenciem o comportamento das florestas nas propriedades rurais.

---

[1] Recebido em 1º de julho de 1986.
Aceito para publicação em 03 de março de 1988.

[2] Agrº, M.S. em Engenharia Florestal, Professor Visitante da Universidade de Brasília - Departamento de Engenharia Florestal - Campus Universitário - CEP 70000 - Brasília, DF.
[3] Estatístico, Ph.D., Pesquisador do Centro de Pesquisa Agropecuária do Cerrado (CPAC/EMBRAPA) - BR 20 (Rodovia BSB/Fortaleza) Km 18 - CEP 70023 - Planaltina, DF.
[4] Engº Florestal, Ph.D. em Economia Florestal, Professor Titular da UFPR - Escola de Floresta - Rua Bom Jesus, 650 - CEP 80000 - Curitiba, PR.

A tentativa de agrupar as propriedades rurais em conglomerados homogêneos, em função dos vários parâmetros que afetam seu desenvolvimento e as caracterizam, são de grande utilização no planejamento e na definição de uma política rural mais eficiente (Angelo 1985).

Quando se trata do estudo de vários parâmetros observados ou medidos sobre um mesmo indivíduo tem-se recorrido aos métodos de análise multivariados (Judes *et al.* 1984; Angelo 1985 e Moreira 1985), com o intuito de uma melhor explicação da estrutura da massa de dados.

Este trabalho tem como objetivos: a) identificar conglomerados homogêneos de propriedades rurais e as variáveis que os afetam; b) verificar as relações estruturais , em cada grupo, entre os parâmetros: mata nativa, reflorestamento, capoeira, agricultura de subsistência, pastagem e população rural.

MATERIAL E MÉTODOS

Obtenção e preparo de dados

Neste estudo foram utilizados os 110 estabelecimentos rurais de tamanho variando de 26 a 52 ha, localizados no município de Porto Vitória, região sul do Estado do Paraná, cujas coordenadas geográficas são 51°00' e 51°30' de longitude sul e 26°30' e 27°00' de latitude oeste de Greenwich.

Os parâmetros estudados em cada propriedade foram áreas em hectares de mata nativa (Y_1), capoeira (Y_2), reflorestamento (Y_3), agricultura de subsistência (X_1), pastagens (X_2) e a população rural (X_3).

Os dados foram coletados nas fichas cadastrais das propriedades rurais oriundas do Cadastro Técnico de Imóveis Rurais, realizado pelo convênio Instituto de Terras Cartografia e Florestal do Paraná, Fundação Universidade Federal do Paraná e República Federal da Alemanha.

As áreas de cobertura florestal primitiva bem como as áreas de mata secundária que já atingiram um alto estádio de desenvolvimento foram, para efeito de estudo, denominadas áreas de mata nativa. A capoeira refere-se ao somatório das áreas em regeneração da cobertura florestal arbórea na propriedade, destacando-se os bracatingais.

Quanto às áreas reflorestadas consideraram-se aquelas plantadas com *Pinus* spp, *Araucaria angustifolia* e *Eucalyptus* spp.

A agricultura de subsistência refere-se às áreas cultivadas com culturas de milho, arroz, feijão e mandioca, mais as áreas com pomares.

Como pastagens, foram consideradas apenas as plantadas.

Quanto à população rural, refere-se ao número total de residentes na propriedade rural.

Modelos de análise

Face aos objetivos do presente estudo, e dada a matriz de observações, justifica-se um estudo da estrutura dos dados para detectar grupos homogêneos que permitam mais eficiente explicação da massa de dados.

A princípio, realizou-se a análise de componentes principais, que antecede a formação de conglomerados da segunda parte. Nas duas últimas partes, são realizadas as análises estrutural e de variância, aplicadas a cada grupo.

Análise em componentes principais

Considere-se as variáveis Y_1, Y_2, Y_3, X_1, X_2, X_3 normalmente distribuídas com o vetor de médias u e matriz covariância Σ.

A análise em componentes principais procura definir combinações lineares das variáveis Y_1, Y_2, Y_3, X_1, X_2, X_3 (denominadas componentes principais), tal que cada combinação explique ao máximo a variância generalizada das variáveis e seja linearmente independente entre si (menor número de variáveis não correlacionada), para facilitar o estudo das relações entre elas e determinar os fatores responsáveis pelas variações entre os conglomerados.

O primeiro problema, nesta análise, é determinar a primeira componente principal, aquela que explica a maior variabilidade global das variáveis. A solução para este problema algébrico é equivalente, usando notação matricial, a determinar os autovalores ou raízes características (λ_i) e os autovetores associados (vetores característicos) da matriz covariância do respectivo componente principal, enquanto os elementos do autovetor fornecem os coeficientes para obter os componentes principais (Carvalho, 1979).

Os valores próprios (λ_i) têm as seguintes características, tal que $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5 > \lambda_6$ e estes valores são encontrados ao dar solução à equação $| R - \lambda_I | = 0$, onde I(6x6) é a matriz identidade de tamanho 6 e R(6 x 6) é a matriz correlação de tamanho 6. E cada raiz característica (λ_i) tem um vetor próprio associado.

Os valores transformados (Zi) para a análise de conglomerados, são encontrados pela seguinte função (Judes *et al.* 1984):

$$Z_i = \frac{\sqrt{\lambda_i}\,(\text{vetor próprio})\,(X_i - X_i)\,\sigma\, xi}{\sqrt{\lambda_i (NV)}}$$

Onde: Z_i = valores transformados
X_i = valor da variável i;
X_i = média da variável i;
σxi= desvio padrão da variável i;
NV = número de variáveis;
λi = valores próprios (raízes características)

Ao se obter os valores Z_i têm-se a seguinte característica: r(Zi, Zj) = 0 onde r = correlação, para $i \neq j$.

Análise de conglomerado ("cluster analysis")

Precedida muitas vezes pela análise dos componentes principais, a análise de conglomerado tem por finalidade proporcionar várias partições na massa de dados (que são as propriedades rurais) visando identificar grupos hierárquicos, ascendentes, excludentes das observações.

Existem vários métodos para a formação de grupos, cuja escolha e aplicação, em cada caso, depende da natureza das variáveis e dos objetivos do estudo. O algoritmo utilizado neste estudo, para identificar grupos hierárquicos foi o método de Ward (Anderberg 1973 e Everitt 1977), também conhecido como "variância mínima. Este método consiste em agregar em cada etapa dois grupos que conservam o máximo de dispersão entre eles, com a minimização da dispersão dentro dos mesmos e tem como função de grupamento a distância Euclidiana e o critério de grupamento é dado pelo valor do incremento que se obtém na matriz de dispersão da soma dos quadrados do erro (Gama 1980).

Neste ítem, introduziu-se alguns conceitos, os das matrizes π, D e E e das medidas de correlação múltipla quadrática (RSQ) e a correlação múltipla quadrática semi-parcial (SPRSQ). Estas medidas auxiliam na escolha do número de grupos (g) ótimo.

A matriz π se refere a variância total fixa dos g grupamentos e pode ser escrita como a soma da variância dentro de todos os g grupos (D) mais a variação entre os g grupamentos (E). O RSQ é a proporção da variância explicada pela variação entre os g grupos. Na notação matricial, a formação dos g conglomerados é dada por:

$\pi = D + E$ ou $D/\pi + E/\pi = 1$

da qual se obtém:

$RSQ = 1 - D/\pi = E/\pi$

Entre os vários métodos de especificação do nível de agregação, tem-se o (SPRSQ) que significa o decréscimo da variância entre os grupos, causados pela aglutinação de dois grupos (Judes et al. 1984).

Esta medida é expressa como proporção da variância total pela seguinte relação:

$SPRSQ = l_{ij}/\pi$

onde lij é o incremento na matriz D pela união dos grupos Gi e Gj.

O número de grupos (g) ótimo será feito com base no dendrograma (diagrama de árvores) juntamente com as informações das estimativas das correlações RSQ e SPRSQ. Este diagrama de árvores assim como o "método de Ward" são obtidos através do PROC TREE e PROC CLUSTER (METHOD = WARD), respectivamente, os quais são programas do sistema SAS - Statistical Analysis System (SAS Institute 1982).

Análise estrutural dos grupos

Primeiramente, foi calculada a média geral (média ponderada dos grupos) dos parâmetros, em seguida tomou-se as médias das variáveis em cada grupo e verificou-se entre elas, aquelas que são maiores que a média geral do parâmetro correspondente. Na seqüência, para cada variável i, procurou-se determinar a proporção

(em porcentagem) de observações no grupo que estão acima do valor da média geral i, de tal forma que, permitiu-se a construção de uma tabela que visualiza as variáveis que afetam os grupos (média do parâmetro i no grupo maior que sua média geral e a proporção das observações no grupo acima da média geral).

Em seguida, foram comparados os valores médios das variáveis entre grupos, utilizando-se teste de Duncan, ao nível de significância de 5% (Snedecor 1967).

RESULTADOS E DISCUSSÃO

Componentes principais

Os dados usados para análise de grupamentos são resultantes dos componentes principais. A matriz dos vetores próprios e dos valores próprios dos seis componentes são apresentados na Tabela I.

TABELA 1. Matriz dos vetores próprios dos quais se obtêm os componentes principais para a formação de grupos de propriedades do município de Porto Vitória, PR, 1985.

Variável	Vetores próprios modificados[1]						Valor próprio		
								Percentagem	
	1	2	3	4	5	6	i	Simples	Acum.
Y_1	0,88	−0,05	0,25	−0,17	0,11	0,36	1,6	27,2	
Y_2	−0,33	−0,51	−0,72	0,06	0,23	0,24	1,2	20,3	47,5
Y_3	0,09	0,62	−0,12	0,68	0,37	0,06	1,1	19,0	66,5
X_1	−0,30	0,72	−0,32	−0,37	−0,33	0,20	1,0	16,0	82,5
X_2	−0,61	−0,19	0,55	0,38	−0,28	0,26	0,8	12,6	95,1
X_3	−0,53	0,14	0,38	−0,44	0,60	0,04	0,3	5,0	100,0

Fonte: Dados e análise da pesquisa.
[1] Os valores estão multiplicados pelo correspondente $\sqrt{\lambda i}$, i = 1, ...6.

Como conseqüência da propriedade de ortogonalidade, cada componente pode ser interpretada separadamente, como segue:

1. comparação da mata nativa e reflorestamento com as demais variáveis explicando 27,2% das variações;
2. comparação do reflorestamento, agricultura de subsistência e população rural com as demais, explicando 20,3% das variações;
3. comparação da área de pastagem, população rural e mata nativa com as demais variáveis, o que explica 19% da variabilidade dos dados;
4. comparação do reflorestamento e pastagem com as demais variáveis explicando 16,0%;

5. comparação da agricultura de subsistência e pastagem com as demais variáveis, explicando 12,6% da variância total e
6. média geral de todas as variáveis, explicando apenas 5% da variação dos dados.

Observa-se que os quatro primeiros componentes principais explicam 82,5%, o que, segundo Morrison (1967), é um valor bastante significativo.

Análise de conglomerado

Sobre os valores resultantes dos componentes principais, ou seja, para todos os Z_{ih} (h = i, 2...6 e i = 1, 2...110) da Matriz Z (110x6) para as propriedades, o método de Ward foi aplicado para obter os grupamentos dos imóveis rurais.

A proximidade entre os pontos permite a formação de grupos homogêneos de indivíduos (propriedades rurais) e para tal o referido método foi empregado.

As medidas usadas como critérios na formação de grupos estão relacionadas com a correlação múltipla quadrática (RSQ) e a correlação múltipla quadrática semiparcial (SPRSQ) que são mostradas na Tabela 2.

TABELA 2. Mudanças na composição da variância na formação de conglomerados de propriedades rurais do município de Porto Vitória, PR, 1985.

Número de Conglomerados	RSQ	SPRSQ	Freqüência de novos Conglomerados
1	0	0,126	110
2	0,126	0,115	106
3	0,241	0,100	66
4	0,341	0,094	40
5	0,435	0,078	28
6	0,513	0,055	38

Fonte: Dados e análise da pesquisa.

À medida que aumenta o número de grupos, verificam-se diminuições proporcionais da variância entre os grupos, ou seja, no valor SPRSQ. As mudanças no valor da variância ocorrem em sentido oposto as experimentadas dentro do grupo. A escolha do número de conglomerados é realizada ao nível em que essa perda de variância seja mínima e estável, conforme especificações metodológicas (Judes *et al.*, 1984).

No dendrograma das observações da (Figura 1) observam-se possíveis grupos de propriedades rurais que poderiam ser consideradas em função dos níveis de SPRSQ. Para SPRSQ = 0,055, tem-se a formação de seis conglomerados, os quais foram estudados.

FIGURA 1. Dendrograma (diagrama de árvores) das propriedades rurais.

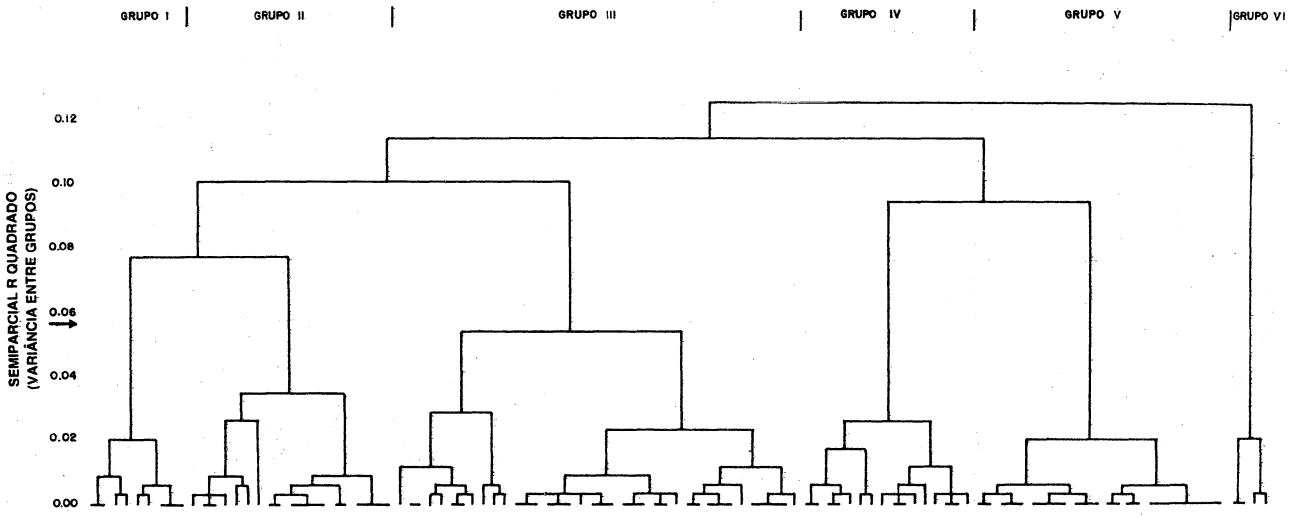

FIG. 1. Dendrograma mostrando vários grupos de propriedades rurais do município de Porto Vitória (PR) em função da partição da variância entre grupo, 1985.

Análise estrutural dos grupos

Esta análise foi feita considerando a média de cada parâmetro em cada grupo. A Tabela 3 mostra as médias das variáveis para cada grupo, e também, a proporção de propriedades rurais nos grupamentos acima da média geral de cada parâmetro. Os quadros hachurados mostram que a média da variável no grupo é maior ou igual ao seu valor médio geral e que esta influenciou a formação do grupo. Segundo Moreira (1985), este tipo de análise é um complemento bastante útil, pois ajuda a detectar os grupos que possuem as médias acima da média geral de cada parâmetro e a percentagem de observações acima da média geral para os grupos com média acima dela.

TABELA 3. Quadro de análise dos seis grupos de propriedades rurais.

Grupo	Y_1	Y_2	Y_3	X_1	X_2	X_3
I	10,64	0,51	0,13	3,19	29,87 100%	1,44
II	6,04	5,01	0,11	20,07 84,47%	9,78 52,63%	0,74
III	8,96	5,02	0,23	10,46 50%	10,84 57,90%	6,55 94,74%
IV	7,33	23,00 100%	0,06	6,70	6,23	1,43
V	26,14 100%	2,77	0,00	3,35	1,82	0,45
VI	6,07	3,69	5,14 100%	8,35	8,65	2,00
Média Geral	12,00	6,73	0,31	9,36	9,50	2,89

Fonte: Dados e análise da pesquisa.

Com 9 propriedades, o grupo I é caracterizado pelo fato de que a média da área de pastagem está acima da média geral.

No conglomerado II, formado por 19 imóveis rurais, são freqüentes as áreas de agricultura de subsistência e pastagens com médias acima da geral. Estas áreas estão relacionadas positivamente no grupo, caracterizando sua formação.

O Grupo III caracteriza-se pela relação positiva dos parâmetros agricultura de subsistência, pastagem e a população rural, os quais possuem médias acima da geral. Este resultado, confirma, estatisticamente, os preconizados por Guerrero (1981) e Konzen & Richter (1982), quando concluíram que, a agricultura de subsistência e a produção animal estão associados com a fixação do homem no meio-rural, nos pequenos e médios estabelecimentos. Este conglomerado, formado por 38 propriedades rurais, constitui o mais numeroso entre os seis grupos estudados.

Nos conglomerados IV e V, com um total de 40 imóveis, predominam somente as áreas de capoeira e de mata nativa, respectivamente. Com 24 propriedades, o grupo V destaca-se pelas extensas áreas cobertas por matas nativas e ausência absoluta de reflorestamento. Isto mostra que, estando Porto Vitória em um dos últimos pólos madeireiros do Estado do Paraná, o reflorestamento com as espécies florestais de rápido crescimento ainda não está competindo com as matas nativas no município. Formado por apenas 4 propriedades rurais, o grupo VI constitui um conglomerado atípico, dedicando-se basicamente ao reflorestamento.

A tabela 4 apresenta, de maneira sucinta, informações das relações entre os parâmetros estudados nos seis grupos. Observa-se que não há relações positivas entre os três tipos florestais estudados, mas sim um antagonismo entre mata nativa e reflorestamento no conglomerado V. Nos grupos IV, V e VI observa-se uma tendência dos mesmos se caracterizem por um determinado tipo de cobertura florestal.

As médias dos parâmetros em diferentes grupos e as comparações dos valores médios pelo teste de Ducan são apresentados na Tabela 4.

TABELA 4. Médias das variáveis e teste de Ducan aplicado entre os grupos. Porto Vitória, Estado do Paraná, 1985.

Grupos	Y_1	Y_2	Y_3	X_1	X_2	X_3
I	10,64 b	0,51 b	0,13 b	3,19 c	29,87 a	1,44 b
II	6,04 b	5,01 b	0,11 b	20,07 a	9,78 b	0,74 b
III	8,96 b	5,02 b	0,23 b	10,46 b	10,84 b	6,55 a
IV	7,33 b	23,00 a	0,06 b	6,70 bc	6,23 bc	1,43 b
V	26,14 a	2,77 b	0,00 b	3,35 c	1,82 c	0,45 b
VI	6,07 b	3,69 b	5,14 a	8,35 bc	8,65 b	2,00 b

Fonte: Dados e análise da pesquisa.

Y_1 = área de mata nativa em ha; Y_2 = área de capoeira em ha; Y_3 = área reflorestada em ha; X_1 = área de agricultura em ha; X_2 = área de pastagem em ha; X_3 = população rural.

– Médias seguidas pela mesma letra, em cada coluna, não diferem estatisticamente, entre si, ao nível de 5% de probabilidade, pelo teste de Ducan.

As propriedades dos grupos I, II, IV, V e VI apresentaram maiores valores médios de pastagem, agricultura de subsistência, capoeira, mata nativa e reflorestamento, respectivamente, em comparação aos demais grupos, tendo sido estatisticamente significativa essa diferença.O grupo III apresentou população rural diferente, estatisticamente, dos demais grupos. Essa diferença estatística pode ser explicada pela relação positiva verificada na análise de grupamento (Tabela 3) da população rural com agricultura de subsistência.

CONCLUSÕES E SUGESTÕES

a) Com base na metodologia aplicada a 110 propriedades rurais, indentificaram-se seis grupos homogêneos de propriedades, quanto aos parâmetros estudados;
b) Pela análise estrutural dos grupos, verificou-se que não há, estatisticamente, relações positivas entre os três tipos de cobertura florestal estudados, mas sim, uma relação antagônica entre a área de mata nativa e a reflorestada no grupo V;
c) De acordo com análise de variância, os grupos definidos de propriedades rurais tendem a se especializar, ou seja, há a predominância de pelo menos um dos parâmetros em cada grupo;
d) A metodologia adotada permitiu uma melhor identificação da estrutura das propriedades rurais, recomendando-se porém, outros estudos para melhor conhecer as possíveis causas destas relações em cada grupo já definido.

REFERÊNCIAS

ANDERBERG, M. R. **Cluster analysis for applications.** New York, Academic Press, 1973. 359p.

ANDERSON, T. W. **An introduction to multivariate statistical analysis.** New York, J. Wiley, 1974. 374p.

ANGELO, H. **Cobertura florestal na propriedade rural:** um método de análise. Curitiba, UFPR, 1985. 84p. Tese Mestrado.

CARVALHO, J. P. de. **Algebra linear:** introdução. 2. ed. Brasília, UnB, 1979. 176p.

EVERITT, B. **Cluster analysis.** London, Heineman Educational, 1977. 121p.

GAMA, M. de P. **Bases da análise de grupamentos:** Cluster analysis. Brasília, UnB, 1980. 229p. Tese Mestrado.

GUERRERO, S. J. Transição energética do Brasil: a opção da cana-de-açúcar e o futuro do programa de biomassa, In: SEMINÁRIO DO DEPARTAMENTO DE ECONOMIA RURAL, Viçosa, UFV, 1981. **Anais.** Viçosa, UFV, 1981. 16p.

JUDES, L. A. et al. **Fundamentos teóricos e aplicações da análise de dados:** subsídios para o Programa de avaliação sócio-econômica da pesquisa agropecuária do Projeto II - EMBRAPA/BIRD. Brasília, EMBRAPA, 1984.

KONZEN, O. G. & RICHTER, H. V. Estrutura da produção e da renda agrícola em diferentes grupos de estabelecimentos rurais no Brasil: subsídios para política agrícola. **R. Econ. rural, 20** (2):237-67, 1982.

MOREIRA, A. M. **Metodologia para definir padrões pluviométricos caso:** cerrados brasileiros. Brasília, UnB, 1985. 120p. Tese Mestrado.

MORRISON, D. F. **Multivariate statistical methods.** New York, McGraw-Hill, 1967. 338p.

STATISTICAL ANALYSIS SYSTEM INSTITUTE, Cary, EUA. **SAS user's quide:** statistics. Cary, 1982. 584p.

SNEDECOR, G. W. **Statistical methods.** Ames, Iowa State University Press, 1967. 593p.