

IPEF, n.36, p.51-56, ago.1987

METODOLOGIA PARA DEFINIR GRUPOS HOMOGÊNEOS DE PROPRIEDADES RURAIS COM COBERTURA FLORESTAL

HUMBERTO ANGELO

Universidade de Brasília
70.000 - Brasília - DF

LUIZ HERNAN RODRIGUES CASTRO

EMBRAPA/CPAC
77.240 - Planaltina - DF

ROBERTO TUYOSHI HOSOKAWA

Universidade Federal do Paraná
80.000 - Curitiba - PR

ABSTRACT - Data from 110 farms located in the county of Porto Vitória, state of Parana - 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, analysis of variance. This methodology was possible to establish six different groups of farms as confirmed by using analysis of variance techniques.

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 52ha, 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. E de um modo geral verificou-se relações significativas entre os parâmetros em cada grupamento de propriedades.

INTRODUÇÃO

Um dos principais problemas encontrados na definição de uma política florestal voltada aos anseios e às necessidades dos agricultores reside na falta de conhecimento das relações da cobertura florestal com as demais atividades de produção, subsistência, sociais, culturais, econômicas, ecológicas, políticas e outras que participem das decisões dos produtores. Tal fato tem concorrido para uma série de distúrbios no meio rural, como por exemplo a diminuição acelerada das florestas e o êxodo do homem do meio rural. Acrescenta-se a esta justificativa a relevância das propriedades rurais na produção de

gêneros alimentícios, na fixação do homem no campo e a carência de pesquisa conduzida nesta ótica.

Diante do exposto, torna-se importante o estudo das relações entre os fatores cobertura florestal, agricultura de subsistência, pecuária e a população rural, sem desconhecer que outros influenciam o comportamento das florestas nas propriedades rurais. A fim de alcançar os objetivos desta proposta, esboça-se uma metodologia apropriada para este estudo, visando identificar grupos homogêneos de imóveis rurais e as variáveis que os afetam, bem como verificar para cada grupamento de propriedades as relações 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 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 estágio de desenvolvimento foram, para efeito de estudo, denominadas áreas de mata nativa. A capoeira refere-se do somatório das áreas em regeneração da cobertura florestal arbórea na propriedade, destacando os bracingais.

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.

A área de pastagem foi considerada 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 estado da estrutura dos dados para detectar grupos homogêneos que permitam mais eficiente explicação da massa de dados.

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

Análise em componentes principais

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

A análise dos 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 existente 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, desenvolvida dos dados originais. Do autovalor sai a variância do respectivo componente principal, enquanto os elementos do autovetor fornecem os coeficientes para obter os componentes principais (CARVALHO, 1979).

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

Os valores transformados (Z_i) para a análise de conglomerados são encontrados pela seguinte função:

$$Z_i = \frac{\sqrt{\lambda_i} (\text{vetor próprio})(x_i - \bar{x}_i)/\sigma_{x_i}}{\sqrt{\lambda_i} (NV)}$$

Onde:

Z_i = valores transformados;

x_i = valor da variável i ;

\bar{x}_i = média da variável i ;

σ_{x_i} = desvio padrão da variável i ;

NV = número de variáveis

λ_i = valores próprios (raízes características)

Ao se obter os valores Z_i tem-se a seguinte característica:

$r(Z_i, Z_j) = 0$ onde r = correlação, para $i \neq j$.

Análise de conglomerados ("Cluster analysis")

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

Entre os autores que descrevem a "Cluster Analysis" estão ANDERBERG (1973), GAMA (1980) e JUDEZ et alii (1984).

O algoritmo adotado para identificar os grupos hierárquicos foi o método de Ward, 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; MOREIRA, 1985).

Detalhes de aplicações do algoritmo de Ward são mencionados nos trabalhos de ANGELO (1985), GAMA (1980) e MOREIRA (1985).

Neste item, introduziram-se alguns conceitos, os das matrizes π , D e E e das medidas 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 variação dos grupamentos e pode ser escrita como a soma da variação dentro de todos os g grupos (D) e a variação entre os g grupamentos (E), que resulta na fórmula:

$$\pi = D + E$$

Esta decomposição permite, para cada uma das possíveis participações da massa de dados, determinar a soma das matrizes D e E, obtendo-se para cada uma das partições possíveis um valor para π .

A correlação múltipla quadrática (RSQ) é soma dos quadrados entre todos os grupos dividido pela soma total dos quadrados corrigidos e pode ser calculado pela fórmula: $RSQ = E/\pi = 1 - (D/\pi)$ e em cada passo do processo (desde $g = n$ até $g = 1$).

A correlação múltipla quadrática semi-parcial (SPRSQ), é a soma dos quadrados entre os grupos que se uniram para um novo grupamento dividido pela soma total dos quadrados corrigidos e pode ser calculada por:

$$SPRSQ = I_{ij}/\pi$$

onde I_{ij} é o incremento na matriz D pela união dos grupos G_i e G_j . Também, SPRSQ significa o decréscimo na RSQ causado pela aglutinação dos grupos.

O número de grupos (g) ótimo será feito de acordo com um dendrograma (diagrama de árvores) que juntamente com as informações das correlações RSQ e SPRSQ, que aparecem no diagrama, servirão de apoio para escolher o melhor g. 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 - do SAS INSTITUTE.

Análise estrutural dos grupos

Primeiramente, foi calculada a média geral (média ponderada dos grupos) dos parâmetros, em seguida tomam-se as médias das variáveis em cada grupo e verifica-se - entre elas aquelas que são maiores - a que média geral do parâmetro correspondente. Na seqüência, para cada variável i, procura-se determinar a proporção (em porcentagem) de observações no grupo que estão acima do valor da média geral i, de tal forma que permite 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).

Os efeitos dos grupos sobre as variáveis foram avaliados por meio do teste "F" da análise de variância, a nível de 5% de probabilidade. As comparações dos valores médios das variáveis entre grupos foi baseado no teste de Duncan, para médias, também ao nível de 5% de probabilidade (SNEDECOR, 1967).

RESULTADOS E DISCUSSÃO

Componentes principais

Os dados usados para análise de grupamentos são resultantes dos componentes principais. A matriz dos vetores próprios ou vetores característicos dos seis componentes são apresentados na Tabela 1. Os valores próprios ou variâncias estão contidos na Tabela 2.

Tabela 1 - Matriz dos vetores próprios ou vetores característicos.

Variável	Vetor Característico (Componente)*					
	1	2	3	4	5	6
Y1	0,88	-0,05	0,25	-0,17	0,11	0,36
Y2	-0,33	-0,51	-0,72	0,06	0,23	0,24
Y3	0,09	0,62	-0,12	0,68	0,37	0,06
X1	-0,30	0,72	-0,32	-0,37	-0,33	0,20
X2	-0,61	-0,19	0,55	0,38	-0,28	0,26
X3	-0,53	0,14	0,38	-0,44	0,60	0,04

* Os valores estão multiplicados pelo correspondente $\sqrt{\lambda_i}$, $i=1, \dots, 6$.

Tabela 2 - Valores próprios e percentagem da variância explicada.

Componente principal	Valor próprio (λ_i)	Percentagem da Variância Total (%)	Variância Acumulada (%)
1	1,6	27,2	27,2
2	1,2	20,3	47,5
3	1,1	19,0	66,5
4	1,0	16,0	82,5
5	0,8	12,6	95,1
6	0,3	5,0	100,0
Total	6,0	100,0	100,0

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 e 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 qual explica 19% da variabilidade dos dados;

4 - Comparação do reflorestamento e pastagem com as demais variáveis explicando 16,0%;

5 - Comparação da agricultura da 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 Conglomerados - sobre os valores resultantes dos componentes principais, ou seja, para todos os Z_{ih} ($h = 1, 2 \dots 6$ e $i = 1, 2 \dots 110$) da Matriz Z (110 x 6) para as propriedades, o método de Ward foi aplicado para obter os agrupamentos dos imóveis rurais.

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

Os parâmetros usados como critérios na formação de grupos estão relacionados com a correlação múltipla quadrática (RSQ) e a correlação múltipla quadrática semi-parcial (SPRSQ) que são mostrados na Tabela 3.

Tabela 3 - SPRSQ e RSQ para os seis primeiros grupos.

Nº de Grupos	SPRSQ	RSQ
6	0,055	0,513
5	0,078	0,435
4	0,094	0,341
3	0,100	0,241
2	0,115	0,126
1	0,127	0,000

Os grupos de propriedades rurais podem ser detectados no dendrograma de árvores da Figura 1, no qual o "corte" escolhido para este trabalho corresponde a seis agrupamentos e tem como valor para o parâmetro SPRSQ = 0,055. Observa-se, à medida que diminui o número de grupos, o parâmetro SPRSQ que diminui o número de grupos, o parâmetro SPRSQ aumenta, enquanto RSQ decresce. A matriz de variação dentro de todos os seis grupos é $D = 1 - 0,513 = 0,487\pi$ (é uma parcela de 0,487 da variação total π); e a matriz de variação entre todos os seis grupos é $E = 0,513\pi$ (0,513 da variação total).

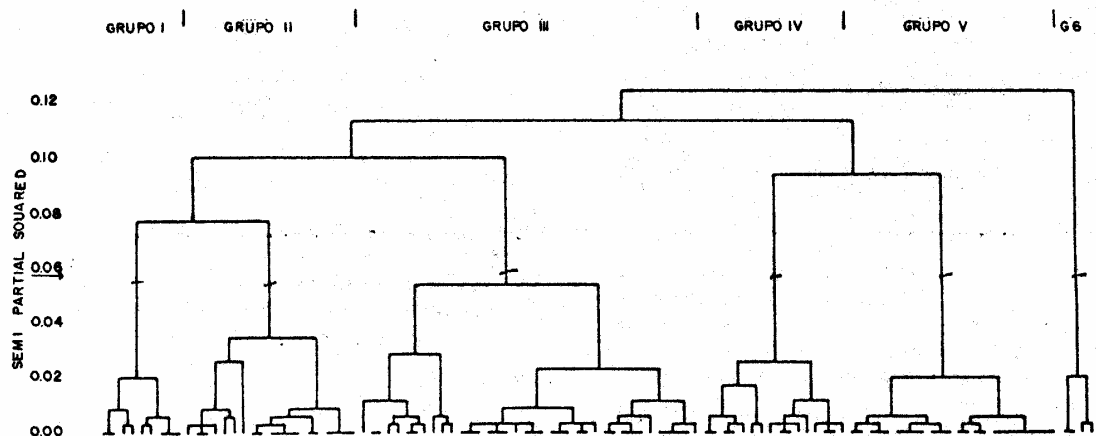


Figura 1 - Dendrograma (diagrama de árvores) das propriedades rurais.

Análise Estrutural dos Grupos

Esta análise foi feita considerando a média de cada parâmetro em cada grupo. A Tabela 4 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 4 - Quadro de análise dos seis grupos de propriedades rurais.

Grupo	Y ₁	Y ₂	Y ₃	X ₁	X ₂	X ₃
1	10,64	0,51	0,13	3,19	29,87 100%	1,44
2	6,04	5,01	0,11	20,07 84,47%	9,78 52,63%	0,74
3	8,96	5,02	0,23	10,46 50%	10,84 57,90%	6,55 94,74%
4	7,33	23,00 100%	0,06	6,70	6,23	1,43
5	26,14 100%	2,77	0,00	3,35	1,82	0,45
6	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

O grupo 1 é caracterizado pela média da área de postagem estar acima da média geral.

No grupo 2 as médias das variáveis agricultura de subsistência e área de pastagem estão acima da média geral, as quais estão relacionadas positivamente neste grupo, bem como caracterizando sua formação.

O grupo 3 caracteriza-se pelos parâmetros agricultura de subsistência, área de pastagem e a população rural, os quais possuem médias acima da média geral. Este resultado confirma, estatisticamente, os preconizados por GUERREIRO (1981) e KONZEN & RICHTER (1982), quando concluíram que a agricultura de subsistência e a produção animal estão relacionadas com a fixação do homem no meio rural, nos pequenos e médios estabelecimentos.

Nos grupos 4, 5 e 6 predominam somente as áreas de capoeira, de mata nativa e de reflorestamento, respectivamente. Estes grupos são caracterizados pelas médias dos referidos parâmetros que estão acima da média geral para a variável correspondente.

Análise de Variância

A Tabela 5 mostra os valores de F da análise de variância. Os dados demonstram que os grupos tiveram efeito significativo nas variáveis consideradas separadamente.

Tabela 5 - Valor e teste de F para grupos de propriedades rurais.

Variáveis	Fonte de variação: grupos (Valor F)	Teste
Y ₁	21,28	**
Y ₂	29,33	**
Y ₃	60,05	**
X ₁	18,88	**
X ₂	23,10	**
X ₃	31,15	**

** Significativo a 1% de probabilidade.

Tabela 6 - Médias das variáveis

Grupos	Y ₁	Y ₂	Y ₃	X ₁	X ₂	X ₃
1	10,64 b	0,51 b	0,13 b	3,20 c	29,87 a	1,44 b
2	6,04 b	5,00 b	0,11 b	20,07 a	9,78 b	0,74 b
3	8,97 b	5,03 b	0,23 b	10,47 b	10,84 b	6,55 a
4	7,33 b	23,00 a	0,06 b	6,70 bc	6,24 bc	1,44 b
5	26,14 a	2,78 b	0,00 b	3,36 c	1,83 c	0,46 b
6	6,04 b	3,70 b	5,14 a	8,35 bc	8,66 b	0,74 b

Y₁ = área mata nativa em ha;

X₁ = área agrícola em ha;

Y₂ = área de capoeira em ha;

X₂ = área de pastagem em ha;

Y₃ = área reflorestada em ha;

X₃ = população rural

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

As médias dos parâmetros em diferentes grupos e as comparações dos valores médios pelo teste de Duncan são apresentados na Tabela 6.

As propriedades dos grupos 1, 2, 4, 5 e 6 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 3 apresentou população rural diferente, estatisticamente, dos demais grupos. Essa diferença estatística pode ser explicada pela relação positiva verificada na análise de agrupamento (Tabela 4) da população rural com agricultura de subsistência.

CONCLUSOES E RECOMENDAÇÕES

Com base nos resultados obtidos e nas análises efetuadas, conclui-se que:

1. Com a aplicação da metodologia proposta foi possível identificar seis grupos homogêneos de propriedades rurais, porém heterogêneos entre si;

2. 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;

3. Pela análise estrutural dos grupos, verificou-se que não há estatisticamente relações positivas entre os três tipos de cobertura florestal estudado, mas sim uma relação antagônica entre a área de mata nativa e a reflorestada no grupo 5.

4. A metodologia adotada permitiu uma melhor identificação da estrutura das propriedades rurais, recomendando-se porém outros estudos para melhor conhecer as possíveis causas destas estruturas em cada grupo já definido;

5. Com a análise de componente principal e a Análise de Conglomerados foi possível atingir os propósitos deste trabalho e recomenda-se a aplicação destas técnicas às futuras pesquisas no campo das ciências agrárias;

6. Recomenda-se um estudo mais amplo sobre os diversos fatores que afetam e se relacionam com a cobertura florestal nas propriedades rurais, de tal forma que esta metodologia possa ser aplicada em outras regiões.

REFERÊNCIAS BIBLIOGRÁFICAS

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

ANOERSON , 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, 1985. 84p. (Tese-Mestrado-UFPR).

- BARBOSA, J.F. **Critério de classificação de financeiras com o emprego de função discriminante linear**. Brasília, 1978. 162p. (Tese-Mestrado-UNB).
- CARVALHO, J.P. **Álgebra linear**: introdução. 2. ed. Brasília, Editora Universidade de Brasília, 1979. 176p.
- GAMA, M. P. **Bases da análise de grupamento** ("Cluster analysis"). Brasília, 1980. 229p. (Tese-Mestrado -UNB).
- GUERREIRO, S.J. Transição energética do Brasil: a opção da cana-de-açúcar e o futuro do programa de biomassa. In: SEMINARIO DEPARTAMENTO DE ECONOMIA RURAL, Viçosa, UFV, 1981. **Anais**. Viçosa, UFV, 1981. 16p.
- JUDEZ, L.A. et. alii. **Fundamentos teóricos e aplicações da análise de dados**: subsídios para o programa de avaliação sócio-econômico da Pesquisa Agropecuária do Projeto II - EMBRAPA/BIRD. Brasília, 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. **Revista de Economia Rural**, 20(2): 237-67, 1982.
- MOREIRA, A.M. **Metodologia para definir padrões pluviométricos-caso**: cerrados brasileiros. Brasília, 1985. 120p. (Tese-Mestrado-UNB).
- MORRISON, D.F. **Multivariate statistical methods**. New York, McGraw-Hill, 1967. 338p.
- SNEDECOR, G.W. **Statistical methods**. Ames, Iowa State University Press, 1967. 593p.