Presidente da República
Fernando Henrique Cardoso

Secretário de Estado de Planejamento e Avaliação
Edward Amadeo

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Sérgio Besserman Vianna

Diretor de Planejamento e Coordenação
Nuno Duarte da Costa Bittencourt

ÓRGÃOS TÉCNICOS SETORIAIS

Diretoria de Pesquisas
Maria Martha Malard Mayer

Diretoria de Geociências
Trento Natali Filho

Diretoria de Informática
Paulo Roberto Ribeiro da Cunha

Centro de Documentação e Disseminação de Informações
David Wu Tai

Escola Nacional de Ciências Estatísticas
Kaizô Iwakami Beltrão

Presidência da República
Secretaria de Estado de Planejamento e Avaliação
Instituto Brasileiro de Geografia e Estatística - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 58 número 210 julho/dezembro 1997

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 58, n. 210, p. 1-108, jul./dez. 1997

Instituto Brasileiro de Geografia e Estatística - IBGE
Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro - RJ - Brasil

Revista Brasileira de Estatística, ISSN 0034-7175

Órgão oficial do IBGE e da Associação Brasileira de Estatística - ABE.

Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais, através de divulgação de artigos inéditos.

Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para os órgãos produtores de estatísticas.

Os originais para publicação deverão ser submetidos em três vias (que não serão devolvidas) para:

Pedro Luis do Nascimento Silva
Editor Responssável - RBEs - IBGE,
Av. República do Chile, 500 - Centro
20031-170 - Rio de Janeiro, RJ.

Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Impressão
Centro de Documentação e Disseminação de Informações-CDDI/ IBGE, em meio digital, em 1999.

Capa
Renato J. Aguiar - Divisão de Criação - DIVIC/CDDI

Ilustração da Capa
Marcos Balster - Divisão de Criação - DIVIC/CDDI

Revista brasileira de estatística/IBGE, - v. 1, n.1 (jan./mar. 1940)- , - Rio de Janeiro: IBGE, 1940-

v.

Trimestral (1940-1986), semestral (1987-).
Continuação de: Revista de economia e estatística.
Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v.50 (1980-1989).
Co-edição com a Associação Brasileira de Estatística a partir do v. 58.
ISSN 0034-7175 = Revista brasileira de estatística.

1. Estatística - Periódicos. I. IBGE. II. Associação Brasileira de Estatística.

IBGE. CDDI. Div. de Biblioteca e Acervos Especiais CDU 31 (05)
RJ-IBGE/88-05 (rev. 98) PERIÓDICO

Impresso no Brasil/Printed in Brazil

SUMÁRIO

NOTA DO EDITOR

Apresentamos neste número da RBEs a nova política editorial definida pelo conselho editorial. Esta reformulação tem por objetivo tornar a revista um veículo para divulgação de trabalhos aplicados de Estatística, com ênfase nos resultados e não somente na metodologia.

A editoração da revista tem sido feita usando o processador de textos "*Word for Windows*", e a tarefa é bastante agilizada quando os autores submetem suas contribuições nesse formato. Encorajamos, portanto, que artigos destinados à submissão para a RBEs sejam preparados nesse padrão.

As submissões de artigos aumentaram bastante. Temos no momento 25 artigos em estágios diversos do processo de avaliação e editoração, de onde certamente sairá material suficiente para os próximos dois números da revista. Temos também a iminência de duas reuniões promovidas pela ABE, a saber a 6ª Escola de Modelos de Regressão e a Escola de Séries Temporais e Econometria, das quais esperamos obter mais algumas submissões. Portanto, para continuar o ritmo de recuperação da revista, solicito a todos que continuem empenhados em submeter ou encorajar a submissão de artigos, bem como a manutenção da colaboração como avaliadores, onde a melhoria que ainda falta é nos prazos de remessa dos pareceres.

Saudações,

Pedro Luis do Nascimento Silva

Editor Responsável

Análise Espacial de Indicadores de Qualidade de Vida para o Município do Rio de Janeiro

Ana Maria Lima de Farias (UFF)*

Maria Tereza Serrano Barbosa (Uni-Rio)**

RESUMO

Nesse artigo são definidos indicadores que permitem classificar as regiões administrativas do Município do Rio de Janeiro em grupos homogêneos quanto a características demográficas, físicas, econômicas e de segurança. A técnica de análise fatorial é utilizada inicialmente para reduzir a dimensão do conjunto de variáveis; em seguida o método de agrupamento das K-médias é aplicado sobre notas definidas a partir dos escores fatoriais para identificar os grupos. Os resultados mostram a diversidade dos grupos para as quatro dimensões do estudo, salientando a importância da determinação de prioridades públicas por região administrativa.

1. INTRODUÇÃO

Este trabalho tem por objetivo a construção de indicadores que permitam qualificar as diversas Regiões Administrativas (RAs) do Município do Rio de Janeiro. A partir de dados disponíveis em publicações do IBGE e do IplanRIO, técnicas estatísticas multivariadas são utilizadas para sintetizar as informações disponíveis para cada RA.

O estudo encontra-se dividido em quatro partes, de acordo com as variáveis envolvidas. Na primeira parte, denominada Dimensão Demográfica, são analisadas variáveis demográficas, que caracterizam as RAs com relação ao movimento populacional entre os anos de 1980 e 1991. Na segunda parte, denominada Dimensão Física, são analisadas variáveis que caracterizam a ocupação do solo em cada RA, contrapondo indicadores de ambiente interno e externo às residências. Na terceira parte, denominada Dimensão Renda, são analisadas variáveis econômicas que caracterizam as RAs em termos de concentração de renda e renda média. Como a questão da segurança tem se tornado um fator importante na

* Endereço para correspondência: Rua: Pires de Almeida, 49/201 - Laranjeiras - RJ - 22240-150 - E-mail: tereza@malaria.procc.fiocruz.br.

** Endereço para correspondência: Rua: Álvaro Ramos 451/202 - Botafogo - RJ - 22280-110 - E-mail: amlima@netgate.com.br.

aferição da qualidade de vida dos indivíduos, na quarta parte do estudo, denominada Dimensão Segurança, são analisadas variáveis relativas à criminalidade em geral.

Para cada uma dessas quatro dimensões é apresentada uma análise descritiva dos indicadores, e em seguida a técnica de análise fatorial é aplicada aos dados. O objetivo principal da análise fatorial é identificar um pequeno número de variáveis não observáveis, chamadas fatores, que possam representar as relações entre várias variáveis inter-relacionadas. Esses fatores devem resumir o comportamento de uma ou mais variáveis e devem ser interpretados a partir delas. Finalmente, uma análise de agrupamento aplicada aos escores fatoriais permitiu a identificação de quatro grupos de RAs com comportamentos semelhantes para cada uma das dimensões do estudo. A análise dos dados foi feita utilizando-se o programa SPSS, versão 5 para Windows.

Na Seção 2, é apresentada a divisão do Município do Rio de Janeiro em regiões administrativas, vigente no ano de 1991. Na Seção 3, os indicadores utilizados são definidos e analisados. Na Seção 4, são apresentados os resultados da análise espacial dos indicadores, enquanto algumas conclusões são dadas na Seção 5. No apêndice são apresentadas as tabelas dos dados utilizados.

2. AS REGIÕES ADMINISTRATIVAS

Em 1991, o Município do Rio de Janeiro encontrava-se dividido em 26 regiões administrativas, agrupadas em cinco Áreas de Planejamento (AP), sendo cada RA um aglomerado de bairros. Essa divisão é apresentada na Tabela 1 e Figura 1 (Anexo 1).

As favelas da Rocinha, Jacarezinho, Morro do Alemão e Maré, a partir de 1993, passaram a constituir quatro novas RAs. Nesse estudo, porém, será adotada a divisão de 1991 por questões de consistência dos dados. Assim, a favela da Rocinha está alocada na RA da Lagoa, a do Jacarezinho na RA de Inhaúma, a da Maré na RA de Ramos e a do Morro do Alemão, parte na RA de Ramos e parte na RA de Inhaúma.

Nesse estudo não foram consideradas as RAs de Paquetá, Centro e Guaratiba por não existirem dados para algumas das variáveis, conforme pode ser visto nas Tabelas A1 a A4 (Anexo 2). As duas primeiras RAs foram excluídas também em função de sua peculiar situação diante do estudo que foi sendo feito. A ilha de Paquetá é uma região residencial, com uma pequena população, considerada um dos pontos turísticos do município. O centro do Rio de Janeiro, por sua vez, é uma região basicamente comercial, que vem apresentando um declínio no tamanho da população residente.

3. DEFINIÇÃO E ANÁLISE DESCRITIVA DOS INDICADORES

3.1 Dimensão Demográfica

Os indicadores utilizados foram (Tabela A1 do Anexo 2):

TVPOP91 taxa de crescimento populacional entre 1980 e 1991;

DEN91 densidade populacional em 1991;

FAV%80 porcentagem da população residente em favelas sobre a população total em 1980;

FAV%91 porcentagem da população residente em favelas sobre a população total em 1991;

TVFAV91 taxa de crescimento populacional em favelas entre 1980 e 1991; e

DENFAV91 densidade populacional em favelas em 1991.

As RAs da Barra da Tijuca e de Santa Cruz apresentam as maiores taxas de crescimento populacional (140,61% e 72,30%, respectivamente), bem superiores ao crescimento global, que se deu a uma taxa de 7,62%. No entanto, essas duas RAs ainda apresentam baixa densidade populacional (5,58 e 15,54 hab./ha, respectivamente), quando comparada com a densidade global, que era de 48,93 hab./ha em 1991. Copacabana, embora ainda apresente uma alta densidade, teve um decrescimento populacional de 20,75%, o mesmo acontecendo com Botafogo, que teve um decrescimento de 14,93%.

Com relação à população residente em favelas, a maior taxa de crescimento populacional se deu na RA de Campo Grande, que apresenta uma densidade populacional de 385,88 hab./ha, próxima à média global, que é de 390,70 hab./ha. Barra da Tijuca, embora tenha apresentado a segunda maior taxa de crescimento populacional em favelas (265,42%), ainda apresenta densidade populacional bem abaixo da média: 171,21 hab./ha. As RAs de Ramos, São Cristóvão e Portuária são as que apresentam maior percentual de população favelada, sendo que Ramos tem as favelas com a maior densidade: 812,50 hab./ha.

Observando a matriz de correlação apresentada a seguir, verifica-se que a maior correlação negativa (-0,5698) se dá entre DEN91 (densidade populacional) e TVPOP91 (taxa de crescimento populacional global), indicando que as RAs com alta densidade, em média, tiveram um decrescimento populacional na década de 80. O diagrama de dispersão da Figura 2 (Anexo 1) descreve o comportamento conjunto dessas duas variáveis. É interessante observar que, restringindo a análise às favelas, o comportamento entre essas duas variáveis se mantém, apresentando ainda uma correlação negativa, porém menor (-0,3753), entre DENFAV91 (densidade populacional nas favelas) e TVFAV91 (taxa de crescimento populacional nas favelas), conforme exibido na Figura 3 (Anexo 1).

Matriz de Correlação para a Dimensão Demográfica

	TVPOP91	DEN91	FAV%80	FAV%91	TVFAV91	DENFAV91
TVPOP91	1,0000					
DEN91	-0,5698	1,0000				
FAV%80	-0,2099	0,0735	1,0000			
FAV%91	-0,2045	-0,0081	0,9636	1,0000		
TVFAV91	0,6259	-0,4907	-0,5004	-0,3863	1,0000	
DENFAV91	-0,4923	0,2712	0,6968	0,7884	-0,3753	1,0000

Na Figura 4 (Anexo 1) o diagrama de dispersão para as taxas de crescimento populacional global e nas favelas ilustra a alta correlação (0,6259) entre essas variáveis. Tal correlação positiva pode indicar que, nas RAs com alta taxa de crescimento, o crescimento tenha se dado principalmente nas favelas. Podemos ver também que praticamente todas as RAs tiveram um crescimento populacional em favelas maior que o crescimento populacional global (os pontos se encontram acima da linha de igualdade). Essa situação é refletida também no fato de que a participação da população favelada na população total em 1991 é maior que em 1980 em quase todas as RAs.

3.2 Dimensão Física

Os indicadores utilizados foram (Tabela A2 do Anexo 2):

VERDEHAB. área verde, em m^2, por habitante;

VERDE% percentual de área verde em relação à área territorial total;

CONST.% percentual de área construída com relação à área territorial líquida; e

CONSTDEN habitantes por m^2 de área residencial construída.

Aqui é importante salientar o conceito de área verde utilizado, que engloba as áreas de praças, parques, largos, jardins e outras áreas de lazer, não incluídas as praias do município. Sendo assim, a RA da Tijuca é a que apresenta a maior concentração de área verde, em função do Parque Nacional da Floresta da Tijuca. No entanto, com relação à porcentagem de área construída (CONST.%), que dá uma idéia do nível de verticalização da RA, ela é a segunda com maior valor, precedida apenas por Copacabana. Com relação à área residencial por habitante (CONSTHAB.), temos Barra da Tijuca e Lagoa que são as RAs mais favorecidas, seguidas de Copacabana e Botafogo. Santa Tereza apresenta indicadores desfavoráveis com relação ao total de área verde, mas, sob o ponto de vista de uma análise qualitativa, não se pode esquecer de sua proximidade com a Floresta da Tijuca.

A alta correlação negativa entre as variáveis CONST.% e CONSTDEN está ilustrada na Figura 5(Anexo 1). As RAs que apresentam baixos percentuais de área construída com relação à área territorial líquida tendem a ter, em média, poucos m^2 construídos por habitante. As RAs de Santa Cruz e Copacabana têm comportamentos opostos: a primeira tem um

percentual de área construída pequeno, indicando um baixo nível de verticalização, mas o espaço construído por habitante é pequeno. Já em Copacabana, embora o nível de verticalização seja alto, os espaços por habitante estão acima da média. Barra da Tijuca apresenta índices de verticalização semelhantes às RAs da área rural do município, mas com indicadores de área construída por habitante bastante mais favoráveis.

Matriz de Correlação para a Dimensão Física

	VERDEHAB.	VERDE%	CONST.%	CONSTDEN
VERDEHAB.	1,00000			
VERDE%	0,88800	1,00000		
CONST.%	0,37395	0,47695	1,00000	
CONSTDEN	-0,34436	-0,45293	-0,62494	1,00000

3.3 Dimensão Renda

Os indicadores utilizados foram (Tabela A3 do Anexo 2):

RMSM renda média em salários mínimos;

MENOS1% porcentagem de pessoas de baixa renda;

MAIS20% porcentagem de pessoas de alta renda; e

INDGINI índice de Gini.

Barra da Tijuca e Lagoa são as RAs com maior renda média, seguidas de Copacabana, Botafogo e Tijuca. Os conceitos de baixa e alta renda compreendem as famílias cujos chefes ganhavam menos de um salário mínimo e mais de vinte salários mínimos, respectivamente. A maioria das RAs apresenta um percentual de pessoas de baixa renda superior a 15%. Barra da Tijuca, Lagoa, Copacabana e Botafogo são as regiões que apresentam maiores percentuais de pessoas de alta renda; para a maioria das outras RAs, esse percentual é inferior a 5%. O índice de Gini é uma medida de concentração de renda com valores próximos de 1 indicando uma má distribuição da renda. Analisando os dados, pode-se ver que as RAs do município têm um comportamento semelhante, com índices superiores a 0,5, em geral. Vale lembrar a característica do município, onde, em praticamente todas as RAs, convivem pessoas de alta e baixa rendas, caracterizando uma distribuição bastante deficiente da renda.

A análise da matriz de correlação indica que a maior correlação positiva se dá entre as variáveis MAIS20% e RMSM, o que reflete o fato de a média ser muito influenciada por valores extremos. Já a variável INDGINI é a que apresenta as menores correlações.

Matriz de Correlação para a Dimensão Renda

	RMSM	MENOS1%	MAIS20%	INDGINI
RMSM	1,0000			
MENOS1%	-0,8138	1,0000		
MAIS20%	0,9949	-0,7659	1,0000	
INDGINI	0,5184	-0,3428	0,4978	1,0000

3.4 Dimensão Segurança

Todos os indicadores utilizados refletem o comportamento da variável para cada 100 000 habitantes (Tabela A4 do Anexo 2). São eles:

HOMITRAN homicídios no trânsito;

HOMIC outros homicídios;

ROUMORTE roubo com morte;

FUREMVEI furto em veículos;

FVEIMOTO furto de veículos e motos;

FURRESID furto em residências;

FURCOMER furto no comércio;

FURTRAN furto de transeuntes;

ROUEMVEI roubo em veículos;

RVEIMOTO roubo de veículos e motos;

ROURESID roubo em residências;

ROUCOMER roubo no comércio; e

ROUTRAN roubo de transeuntes.

As RAs de São Cristóvão, Madureira, Anchieta e Santa Cruz destacam-se pelos altos índices de homicídios em geral, enquanto Botafogo, Lagoa e Tijuca são as regiões com maiores índices de furto em/de veículos e motos. Vale a pena destacar a Barra da Tijuca pelo seu alto índice de homicídios no trânsito e também como a região onde ocorrem mais roubos e furtos em residências. Já a RA de Copacabana lidera nos furtos de transeuntes. Embora a RA do Centro não faça parte do estudo pelos motivos anteriormente explicados, cabe ressaltar que os altos valores para os indicadores de segurança decorrem do fato de se estar trabalhando com a população residente, que é pequena quando comparada com a população que freqüenta o centro.

A análise da matriz de correlação indica que praticamente todas as correlações são positivas, com as correlações negativas apresentando valores pequenos e a maioria deles relacionados à variável HOMIC. Por outro lado, as maiores correlações positivas se dão entre as variáveis FUREMVEI e FURRESID (furto em veículos e residências) e FURTRAN e ROUTRAN (furto e roubo de transeuntes).

Matriz de Correlação para a Dimensão Segurança

	HOMITRAN	HOMIC	ROUMORTE	FUREMVEI	FVEIMOTO	FURRESID	FURCOMER
HOMITRAN	1,0000						
HOMIC	0,6369	1,0000					
ROUMORTE	0,3938	0,5182	1,0000				
FUREMVEI	0,1539	-0,2981	0,0673	1,0000			
FVEIMOTO	0,1132	-0,2854	0,1190	0,7918	1,0000		
FURRESID	0,2082	-0,1130	0,2252	0,8924	0,6564	1,0000	
FURCOMER	0,5450	0,2184	0,3837	0,6356	0,5898	0,5900	1,0000
FURTRAN	0,1187	-0,0603	0,1311	0,6934	0,5032	0,5823	0,7459
ROUEMVEI	0,0623	-0,1318	0,3020	0,6012	0,6810	0,4201	0,6646
RVEIMOTO	0,3792	0,2777	0,5485	-0,0371	0,2988	0,0064	0,3140
ROURESID	0,1184	-0,1806	0,2700	0,5924	0,5491	0,7000	0,3906
ROUCOMER	0,6416	0,2302	0,4391	0,5873	0,7237	0,5405	0,8972
ROUTRAN	0,3087	0,3356	0,2842	0,4455	0,3790	0,3632	0,7974

	FURTRAN	ROUEMVEI	RVEIMOTO	ROURESID	ROUCOMER	ROUTRAN
FURTRAN	1,0000					
ROUEMVEI	0,5792	1,0000				
RVEIMOTO	-0,1736	0,2230	1,0000			
ROURESID	0,1839	0,4596	0,2962	1,0000		
ROUCOMER	0,5335	0,6472	0,5096	0,4465	1,0000	
ROUTRAN	0,8318	0,5136	0,0384	0,0343	0,5968	1,0000

4. ANÁLISE ESPACIAL DOS INDICADORES

A técnica de análise fatorial foi aplicada aos quatro grupos de indicadores separadamente, com o objetivo de sintetizar o comportamento de tais indicadores, mas respeitando a realidade diversificada do município.

O objetivo principal da análise fatorial é descrever, sempre que possível, as relações de covariâncias entre várias variáveis em termos de poucas, mas não observáveis variáveis aleatórias, chamadas fatores.

Assim, o modelo de análise fatorial postula que um vetor aleatório X de dimensão p, com média μ e matriz de covariância Σ, é linearmente dependente em poucas variáveis aleatórias $F_1, \cdots, F_m$, chamadas fatores, e p fontes adicionais de variação $\varepsilon_1, \cdots, \varepsilon_p$ chamadas erros ou fatores específicos. Em notação matricial, o modelo é

$$X - \mu = LF + \varepsilon$$

O elemento (i, j) da matriz L é chamado carga (*loading*) da i-ésima variável no j-ésimo fator. As hipóteses do modelo ortogonal de análise fatorial são:

$$E(F) = 0 \qquad Cov(F) = I_m$$
$$E(\varepsilon) = 0 \qquad Cov(\varepsilon) = \Psi = diag(\psi_1, \cdots, \psi_p)$$

e resultam na seguinte estrutura de covariâncias

$$\Sigma = LL' + \Psi$$
$$Cov(X, F) = L$$

de modo que

$$\sigma_{ii} = \mathrm{Var}(X_i) = l_{i1}^2 + \cdots + l_{im}^2 + \psi_i$$

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = l_{i1} l_{j1} + \cdots + l_{im} l_{jm}$$

e

$$\mathrm{Cov}(X_i, F_j) = l_{ij}$$

Segue, então, que a variância de X_i pode ser decomposta em 2 parcelas:

$$\sigma_{ii} = h_i^2 + \psi_i$$

onde $h_i^2 = l_{i1}^2 + \cdots + l_{im}^2$ representa a contribuição dos m fatores comuns e é chamada comunalidade (*communality*) da variável X_i.

É interessante notar que, para qualquer matriz ortogonal T, as matrizes $L^* = LT$ e $F^* = T'F$ têm as mesmas propriedades estatísticas de L e T, respectivamente. Essa ambigüidade do modelo de análise fatorial fornece a base teórica para a prática da rotação de fatores, uma vez que matrizes ortogonais estão associadas à rotação (e reflexão) do sistema de coordenadas.

Um dos métodos de estimação da matriz L baseia-se na decomposição espectral da matriz de covariância amostral S. Tal método de extração dos fatores é chamado de método das componentes principais porque as cargas dos fatores resultantes diferem das primeiras componentes principais apenas pelos fatores de escala.

Obtida a matriz estimada de cargas $\hat{L}$, é possível que os fatores resultantes não sejam facilmente interpretáveis, tornando necessária a rotação dos fatores. O método de rotação denominado VARIMAX procura identificar uma matriz rotacionada $\hat{L}^*$ de modo a minimizar o número de variáveis com valores altos das cargas em cada fator, o que pode facilitar a interpretabilidade dos fatores.

Para maiores detalhes sobre o modelo de análise fatorial, sugere-se a leitura de Johnson&Wichern(1982).

A técnica de análise fatorial foi aplicada aos quatro grupos de indicadores definidos nas seções anteriores, usando-se o método das componentes principais para extração dos fatores. Por questões de interpretabilidade e visualização dos resultados, optou-se por trabalhar com dois fatores. Além disso, a rotação dos fatores pelo método VARIMAX facilitou a interpretação dos mesmos.

Para as dimensões demográfica, renda e física, os percentuais da variância total explicados pelos dois fatores foram todos superiores a 80%, caindo para 66% na dimensão segurança (Tabela 2 do Anexo 1).

Os percentuais das variâncias de todas as variáveis (*communalities*) explicados pelos dois fatores são bastante altos para todas as dimensões, com a dimensão segurança apresentando os piores resultados novamente (Tabela 3 do Anexo 1).

Analisando as matrizes dos fatores apresentadas na Tabela 3, pode-se ver que, na dimensão demográfica, os valores altos do primeiro fator dizem respeito às variáveis relacionadas com a composição das favelas: FAV%91, FAV%80 e DENFAV91; assim, esse fator será denominado **Favelas**. Como os sinais desses fatores são positivos, valores altos para os escores fatoriais indicarão forte presença de favelas na região administrativa. Já para o segundo fator, os valores altos, com sinal positivo, ocorrem nas variáveis referentes ao crescimento da população (global e nas favelas) e com sinal negativo, na densidade populacional. Assim, esse fator será denominado **PopGeral**.

Na dimensão física, o primeiro fator apresenta valores altos nas variáveis relativas à área verde e, assim, denominando-o por **Verde**, valores altos dos escores relativos a ele indicarão forte presença de área verde na RA. Com relação ao segundo fator, escores baixos indicarão a predominância da variável CONSTDEN, que indica um alto número de habitantes por m^2 construído, e escores altos corresponderão a valores altos de CONST.%, indicando um alto nível de verticalização; sendo assim, esse fator será denominado **Constr**.

Na dimensão renda, os valores altos do primeiro fator ocorrem nas variáveis RMSM, MENOS1% e MAIS20%, que são variáveis que mostram a composição da população com relação à renda; assim, esse fator será denominado **Renda**. Já o segundo fator apresenta valor alto apenas para a variável INDGINI e, portanto, receberá a denominação **Concent**.

Finalmente, para a dimensão segurança, o primeiro fator tem valores altos nas variáveis que envolvem furtos e roubos em geral, sem ocorrência de morte. Já o fator 2 tem valores altos nas variáveis envolvendo crimes com morte. Assim, o fator 1 será denominado **CrimGeral** e o fator 2, **CrimMorte**. Como os escores são positivos, valores altos dos fatores tendem a refletir uma situação desfavorável para a RA em questão.

Por questões de padronização, optou-se por trabalhar com notas variando de 0 a 100, definidas a partir dos escores fatoriais das RAs. Essas notas refletem a situação das RAs com relação ao fator em questão, notas baixas indicando uma situação desfavorável e notas altas, uma situação favorável. Em alguns casos, tornou-se necessária a inversão do sinal dos escores para manter tal interpretação.

Na dimensão demográfica, o fator **Favelas** indica a forte presença de favelas na RA e, portanto, valores altos desse fator correspondem as piores notas; para o fator **PopGeral**, valores baixos correspondem a notas baixas, indicando a predominância da variável densidade populacional na composição do fator. Na dimensão física, as notas maiores para o fator **Verde** correspondem a valores altos dos escores fatoriais, indicando a forte presença de área verde na RA. Por outro lado, para o fator **Constr.**, as notas altas vão refletir a existência de maiores espaços construídos por habitante, correspondendo, portanto, a valores baixos

dos escores fatoriais. Atribuindo notas às regiões administrativas de acordo com os escores fatoriais na dimensão renda, RAs com valores altos do fator **Renda** tenderão a ter notas altas, enquanto RAs com valores altos do fator **Concent.** terão notas baixas (quanto maior o índice de Gini, pior a distribuição da renda). Na dimensão segurança, como os escores são positivos, valores altos de ambos os fatores tendem a refletir uma situação desfavorável para a RA em questão e, assim, correspondem às notas mais baixas. Na Tabela 4 (Anexo 1) são apresentadas as notas das RAs para cada uma das dimensões do estudo, enquanto nas Figuras 6 a 9 (Anexo 1) temos as respectivas distribuições das RAs segundo essas notas fatoriais.

Num segundo momento, o agrupamento das RAs com base nas notas fatoriais foi feito usando-se o método das K-médias. Esse método é um dos métodos de partição mais usados para definir o agrupamento de n objetos em K grupos. Partindo de um agrupamento inicial, cada objeto é realocado no grupo cuja média (centróide) se encontra mais próxima, em um contexto usual de distância euclidiana. Uma das dificuldades associadas à aplicação desta técnica multivariada diz respeito ao número de agrupamentos que melhor explica o comportamento da matriz de dados analisada. Nesse estudo, o número de grupos foi fixado em quatro, levando-se em conta uma divisão quase "natural" do Município do Rio de Janeiro em quatro regiões: zona sul, zona oeste e zona norte dividida em dois setores, conforme a maior proximidade com a zona sul ou com a Baixada Fluminense.

Para maiores detalhes sobre análise de agrupamento sugere-se a leitura de Bussab et al.(1990) e Johnson & Wichern(1982).

Nas Figuras 6 a 9 (Anexo 1) os quatro grupos aparecem delineados por uma linha e na Tabela 5 (Anexo 1) são apresentadas as coordenadas dos centros dos grupos para cada uma das dimensões.

Na dimensão demográfica, o Grupo 1, formado pelas RAs de Ramos, Portuária e São Cristóvão, é aquele com as piores notas do fator FAVELAS, indicando a presença de favelas muito densas e com altas taxas de crescimento populacional. Já o Grupo 4 reúne as RAs com alta densidade demográfica, mas com baixos percentuais de moradores em favelas. O Grupo 3, embora apresente uma situação pouco favorável com relação à presença de favelas, é um grupo onde as RAs ainda têm baixa densidade populacional.

Na dimensão física, Copacabana e Tijuca têm comportamentos extremos, a primeira apresentando baixos índices de área verde, mas com espaços residenciais amplos, enquanto a segunda apresenta alta concentração de área verde, conforme já visto. É interessante notar aí que, no Grupo 2, Botafogo e Lagoa apresentam um comportamento um pouco diferente das demais RAs; essas duas regiões ficariam em um grupo separado, caso tivéssemos trabalhado com cinco grupos e não quatro (isso também acarretaria uma realocação das RAs de Jacarepaguá, Ilha e Portuária).

Os Grupos 3 (Lagoa e Barra) e 4 (Copacabana, Botafogo, Tijuca e Vila Isabel) apresentam um comportamento semelhante com relação ao fator Renda, mas o primeiro

apresenta uma pior distribuição da renda, medida pelo índice de Gini. Comportamento análogo têm os Grupos 1 e 2, mas com renda média inferior.

Madureira e São Cristóvão, compondo o Grupo 3 na dimensão segurança, são as RAs com os piores indicadores envolvendo crimes com morte. Já Copacabana, Botafogo e Lagoa (Grupo 4) apresentam o pior comportamento com relação aos crimes, sem morte (Figuras de 10 a 13 do Anexo 1) .

5. CONCLUSÃO

Pretendemos que eventuais intervenções oriundas do poder público ou reclamos da cidadania voltados para a melhoria da qualidade de vida no Município do Rio de Janeiro possam se utilizar, em seus diagnósticos ou propostas, das análises expostas até aqui. Sem pretender substituir sínteses e conclusões produzidas por especialistas melhor qualificados, ousamos, quase como um exercício que pode e deve ser ampliado, salientar alguns aspectos em cada uma das dimensões do estudo.

Segurança

Dada a impossibilidade da determinação exata dos motivos da violência e, portanto, do remédio que a elimine, o tratamento epidemiológico da violência parece particularmente útil na definição das políticas públicas que visem a minimizá-la. Neste sentido, as análises permitem identificar áreas de risco e uma tipologia espacial de ilícitos, cada um deles reclamando soluções específicas. Assim, ressalvando a desigualdade do ponto de vista da consistência dos dados utilizados, observamos, por exemplo, que:

- São Cristóvão e Madureira destacam-se nos registros de homicídios e roubos com morte, caracterizando-se talvez como as regiões de maior índice de violência;

- Barra da Tijuca, Tijuca e Rio Comprido também se destacam no eixo do crime com morte, sendo que a Barra da Tijuca entrou neste grupo devido ao seu alto índice de homicídios no trânsito; e

- Copacabana, Lagoa e Botafogo constituíram um grupo que poderia ser o de menor índice de criminalidade com morte no município, mas, em compensação, o de pior performance com relação aos crimes sem morte.

Física

Em se tratando da necessidade de incremento na qualidade de vida, a cidadania carioca parece, cada vez mais, interessada em considerar as questões ambientais, incluindo-se aí tanto as demandas de preservação da natureza quanto as intervenções planejadas para o embelezamento, desobstrução, conservação e conforto das vias de circulação pública, assim como para a ampliação e melhoria das áreas voltadas propriamente para o lazer. A análise da dimensão física buscou, portanto, articular a variável presença de área verde

com fatores, tais como verticalização e espaço vital disponível por habitante. Neste sentido observamos:

- o destaque da RA da Tijuca, no que diz respeito à sua área verde devido à presença da Floresta da Tijuca;
- Copacabana com a maior área construída por habitante e ao mesmo tempo a mais verticalizada e com pouquíssima área verde; e
- o contraste nas RAs do Grupo 4 (Figuras 7 e 11 do Anexo 1), que ao mesmo tempo é o menos verticalizado, mas tem pouquíssima área verde e pequena área construída por habitante.

Demografia e Renda

Os interessados na definição de prioridades de investimento nas chamadas obras de infra-estrutura, na definição dos modos, volume e prioridades de prestação de serviços nas áreas da saúde e da educação, enfim, os interessados no enfrentamento da pobreza urbana visando a minimizar o dualismo e a fragmentação da metrópole em várias *cidades* podem aproveitar as seguintes observações:

- São Cristóvão e Ramos destacam-se por apresentarem uma alta proporção da população residente em favelas, por terem as favelas mais densas e por estarem entre as piores rendas médias do município;
- Botafogo e Copacabana apresentaram decrescimento populacional entre 1980 e 1991, caracterizando um provável envelhecimento da população e alta densidade populacional, mas renda média elevada e baixa proporção de população residente em favelas;
- Barra da Tijuca e Lagoa apresentam as maiores rendas médias, mas diferem completamente em relação ao comportamento demográfico. A Barra teve a maior taxa de crescimento populacional global e a segunda maior de moradores em favelas, embora ainda apresente densidade populacional (global e em favelas) inferior à média do município. Enquanto isso, a RA da Lagoa teve um decrescimento da população em geral, mas crescimento da população moradora em favelas, resultando, por conseguinte, em favelas muito densas.

A utilização de técnicas multivariadas permitiu entender e ressaltar aspectos que tornam RAs tão distintas, como a da Barra da Tijuca e Santa Cruz ou Lagoa e Anchieta, "semelhantes". No entanto, pretender que tais técnicas possam descrever semelhanças entre as RAs nas quatro dimensões em conjunto, apesar de ser um caminho estatisticamente natural, não nos parece apropriado, uma vez que minimizaria a conclusão mais importante do estudo, que deixa claro que as intervenções do poder público devem levar em conta a realidade diversificada do Município do Rio de Janeiro.

REFERÊNCIAS

Bussab,W.O., Miazaki,E.S. e Andrade,D.F. *Introdução à Análise de Agrupamentos.* Texto de minicurso para o 9º SINAPE, IME-USP, 1990

IBGE *Censo Demográfico 1991 - Rio de Janeiro*

IplanRIO *Anuário Estatístico da Cidade do Rio de Janeiro*, 1992-1993

Johnson,R.A. and Wichern,D.W. *Applied Multivariate Statistical Analysis.* Prentice Hall, New Jersey, 1982

Norusis,M.J. *SPSS for Windows, Professional Statistics, Release 5.* SPSS Inc., 1992

AGRADECIMENTOS

Esse trabalho foi realizado com suporte financeiro do IplanRIO. As autoras agradecem a Eugênia Vitória Câmera Loureiro, Alcides José de Carvalho Carneiro pela colaboração na definição e análise dos indicadores e a Oswaldo Gonçalves Cruz por sua inestimável ajuda no momento da utilização do programa MAPINFO para a elaboração dos mapas.

ABSTRACT

In this article we define variables which allow us to classify the administrative sectors of the city of Rio de Janeiro in homogeneous groups concerning demographic, physical, economic and security characteristics. Factorial analysis is first applied to reduce the dimension of the set of the variables; then, K-means method of clustering analysis is applied on grades defined over factorial scores to identify the clusters. Results show the diversity of the groups for the four dimensions of the study, emphasizing the importance of determining public priorities by administrative sector.

Anexo 1
Tabelas e Figuras

Tabela 1
Bairros por Região Administrativa
Município do Rio de Janeiro - 1991

RA	Bairros	RA	Bairros	RA	Bairros
1 Portuária	Saúde	13 Méier	S.Francisco Xavier	19 Santa Cruz	Paciência
	Gamboa		Rocha		Santa Cruz
	Santo Cristo		Riachuelo		Sepetiba
	Caju		Sampaio	20 Ilha do Governador	Ribeira
2 Centro	Centro		Engenho Novo		Zumbi
3 Rio Comprido	Catumbi		Lins de Vasconcelos		Cacuia
	Rio Comprido		Méier		Pitangueiras
	Cidade Nova		Todos Santos		Praia da Bandeira
	Estácio		Cachambi		Cocota
4 Botafogo	Flamengo		Engenho de Dentro		Bancários
	Glória		Água Santa		Freguesia
	Laranjeiras		Encantado		Jardim Guanabara
	Catete		Piedade		Jardim Carioca
	Cosme Velho		Abolição		Tauá
	Botafogo		Pilares		Monero
	Humaitá	14 Irajá	Vila Cosmos		Portuguesa
	Urca		Vicente de Carvalho		Galeão
5 Copacabana	Leme		Vila da Penha		Cidade Universitária
	Copacabana		Vista Alegre	21 Paquetá	Paquetá
6 Lagoa	Ipanema		Irajá	22 Anchieta	Guadalupe
	Leblon		Colégio		Anchieta
	Lagoa	15 Madureira	Campinho		Parque Anchieta
	Jardim Botânico		Quintino		Ricardo de Albuquerque
	Gávea		Cavalcanti	23 Santa Tereza	Santa Tereza
	Vidigal		Engenheiro Leal	24 Barra da Tijuca	Joá
	São Conrado		Cascadura		Itanhangá
7 São Cristóvão	São Cristóvão		Madureira		Barra da Tijuca
	Mangueira		Vaz Lobo		Camorim
	Benfica		Turiacu		Vargem Pequena
8 Tijuca	Praça da Bandeira		Rocha Miranda		Vargem Grande
	Tijuca		Honório Gurgel		Recreio dos Bandeirantes
	Alto da Boa Vista		Osvaldo Cruz		Grumari
9 Vila Izabel	Maracanã		Bento Ribeiro	25 Pavuna	Coelho Neto
	Vila Isabel		Marechal Hermes		Acari
	Andaraí	16 Jacarepaguá	Jacarepaguá		Barros Filho
	Grajaú		Anil		Costa Barros
10 Ramos	Manguinhos		Gardênia Azul		Pavuna
	Bonsucesso		Cidade de Deus	26 Guaratiba	Guaratiba
	Ramos		Curicica		Barra de Guaratiba
	Olaria		Freguesia		Pedra de Guaratiba
11 Penha	Penha		Pechincha		
	Penha Circular		Taquara		
	Brás de Pina		Tanque		
	Cordovil		Praça Seca		
	Parada de Lucas		Vila Valqueire		
	Vigário Geral	17 Bangu	Deodoro		
	Jardim América		Vila Militar		
12 Inhaúma	Higienópolis		Campo dos Afonsos		
	Jacaré		Jardim Sulacap		
	Maria da Graça		Magalhães Bastos		
	Del Castilho		Realengo		
	Inhaúma		Padre Miguel		
	Engenho da Rainha		Bangu		
	Tomás Coelho		Sen. Camará		
		18 Campo Grande	Santíssimo		
			Campo Grande		
			Sen.Vasconcelos		
			Inhoaiba		
			Cosmos		

Fonte: Anuário Estatístico da Cidade do Rio de Janeiro - 92/93 - IplanRIO

Tabela 1.2.1 - pp. 15-17

Tabela 2
Variância explicada por 2 fatores

Dimensão	Fator	Variância	% da Variância	% Acumulado da Variância
Demográfica	1	3,29001	54,83	54,83
	2	1,60432	26,74	81,57
Física	1	2,59308	64,83	64,83
	2	0,93040	23,26	88,09
Renda	1	3,02818	75,70	75,70
	2	0,71401	17,85	93,55
Segurança	1	6,03857	46,45	46,45
	2	2,59634	19,97	66,42

Tabela 3
Percentual da variância explicada, matriz dos fatores e matriz dos escores fatoriais

Dimensão	Variável	% da Variância Explicada (Communality)	Matriz rotacionada dos fatores		Matriz dos escores fatoriais	
			Fator 1	Fator 2	Fator 1	Fator 2
Demográfica	FAV%91	0,98388	0,99144	-0,03061	0,40084	0,13287
	FAV%80	0,92742	0,95744	-0,10357	0,37599	0,08995
	DENFAV91	0,76468	0,80594	-0,33933	0,27867	-0,05498
	DEN91	0,75104	-0,06265	-0,86436	-0,15529	-0,45743
	TVPOP91	0,77987	-0,18557	0,86338	0,05364	0,41969
	TVFAV91	0,68743	-0,38517	0,73422	-0,04736	0,32279
Física	VERDEHAB	0,95353	0,96178	0,16887	0,60260	-0,19805
	VERDE%	0,94485	0,92111	0,31048	0,52627	-0,07456
	CONSTDEN	0,81945	-0,19308	-0,88440	0,16973	-0,61843
	CONST%	0,80565	0,23736	0,86563	-0,13291	0,58876
Renda	RMSM	0,97120	0,92713	0,33410	0,35752	0,00577
	MENOS1%	0,84502	-0,91421	-0,09609	-0,45137	0,27259
	MAIS20%	0,93655	0,91142	0,32537	0,35276	0,00201
	INDGINI	0,98942	0,22888	0,96800	-0,28677	1,05744
Segurança	FUREMVEI	0,88858	0,93744	-0,09896	0,19578	-0,10982
	FVEIMOTO	0,73808	0,85886	0,02101	0,17041	-0,06149
	FURRESID	0,69745	0,83498	0,01602	0,16603	-0,06132
	FURTRAN	0,62017	0,78419	0,07229	0,15134	-0,03753
	FURCOMER	0,87774	0,77441	0,52729	0,11291	0,12267
	ROUEMVEI	0,60943	0,76659	0,14756	0,14178	-0,00975
	ROUCOMER	0,88030	0,71981	0,60181	0,09599	0,15316
	ROURESID	0,40006	0,63165	0,03284	0,12394	-0,03913
	ROUTRAN	0,52200	0,58344	0,42615	0,08275	0,10254
	HOMIC	0,77543	-0,27137	0,83773	-0,12153	0,31528
	HOMITRAN	0,65661	0,13469	0,79904	-0,03706	0,26917
	ROUMORTE	0,55397	0,15009	0,72900	-0,02836	0,24340
	RVEIMOTO	0,41507	0,07460	0,63992	-0,03635	0,21824

Tabela 4
Notas fatoriais por dimensão e região administrativa

Região Administrativa	DIMENSÃO							
	Demográfica		Física		Renda		Segurança	
	Favelas	PopGeral	Verde	Constr	Renda	Concent	CrimGeral	CrimMorte
1 Portuária	14	46	14	34	15	85	75	57
3 Rio Comprido	39	35	9	50	0	6	20	46
4 Botafogo	84	21	43	63	87	78	30	95
5 Copacabana	100	0	0	100	93	85	30	100
6 Lagoa	51	37	48	60	74	0	0	87
7 São Cristóvão	10	41	11	43	1	47	53	0
8 Tijuca	72	42	100	46	75	80	20	57
9 Vila Izabel	77	33	7	62	66	87	67	84
10 Ramos	0	39	14	29	6	54	86	75
11 Penha	68	33	17	26	15	65	86	52
12 Inhaúma	36	38	15	26	15	70	72	82
13 Méier	78	34	10	44	39	76	72	73
14 Irajá	58	35	16	27	32	79	88	84
15 Madureira	79	35	16	27	21	77	70	23
16 Jacarepaguá	74	63	13	31	19	30	84	95
17 Bangu	78	47	25	7	16	79	96	85
18 Campo Grande	78	72	19	12	15	67	90	73
19 Santa Cruz	85	70	21	0	8	85	97	68
20 Ilha do Governador	51	52	15	33	34	42	74	91
22 Anchieta	88	51	19	16	22	100	99	39
23 Santa Tereza	45	40	7	55	26	26	64	93
24 Barra da Tijuca	70	100	12	48	100	34	40	36
25 Pavuna	47	45	19	10	18	96	100	64

Tabela 5
Agrupamento das RA's pelo método das K-Médias

Grupo	Dimensão Demográfica			Dimensão Física			Dimensão Renda			Dimensão Segurança		
	Nº de	Centro do grupo		Nº de	Centro do grupo		Nº de	Centro do grupo		Nº de	Centro do grupo	
	RA's	Favelas	PopGeral	RA's	Verde	Constr	RA's	Renda	Concent	RA's	Crimgeral	CrimMorte
1	3	8,00	42,00	1	0,00	100,00	11	19,64	79,91	15	83,33	74,33
2	14	59,92	38,92	8	18,38	53,13	6	14,33	34,17	3	26,67	46,33
3	4	79,00	71,20	1	100,00	46,00	2	87,00	17,00	2	61,50	11,50
4	2	92,00	10,50	13	17,15	21,38	4	80,25	82,50	3	20,00	94,00

Figura 1
Áreas de Planejamento e Regiões Administrativas
Município do Rio de Janeiro, 1991

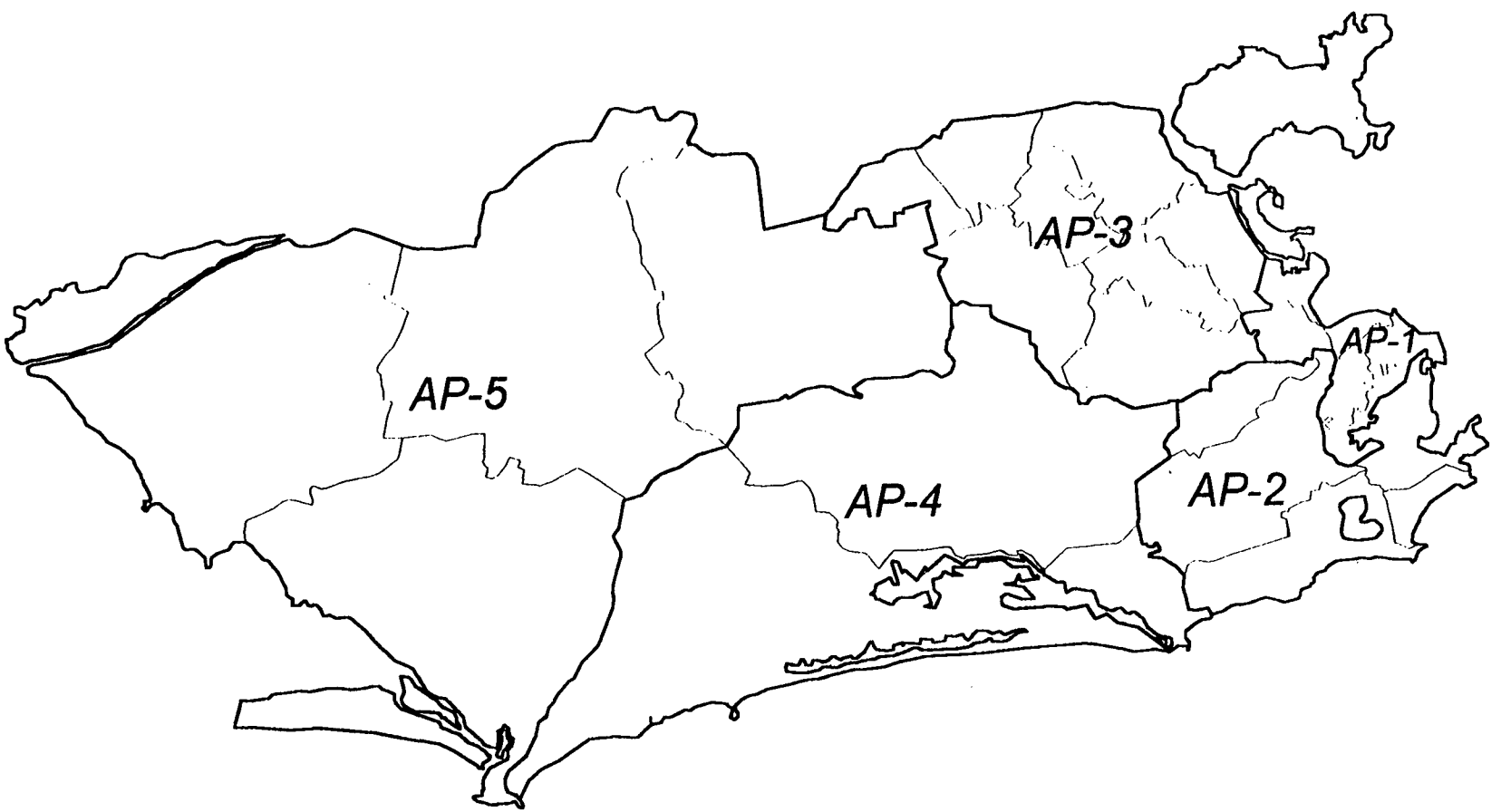


Figura 2

Diagrama de Dispersão: Crescimento x Densidade Populacional

Município do Rio de Janeiro

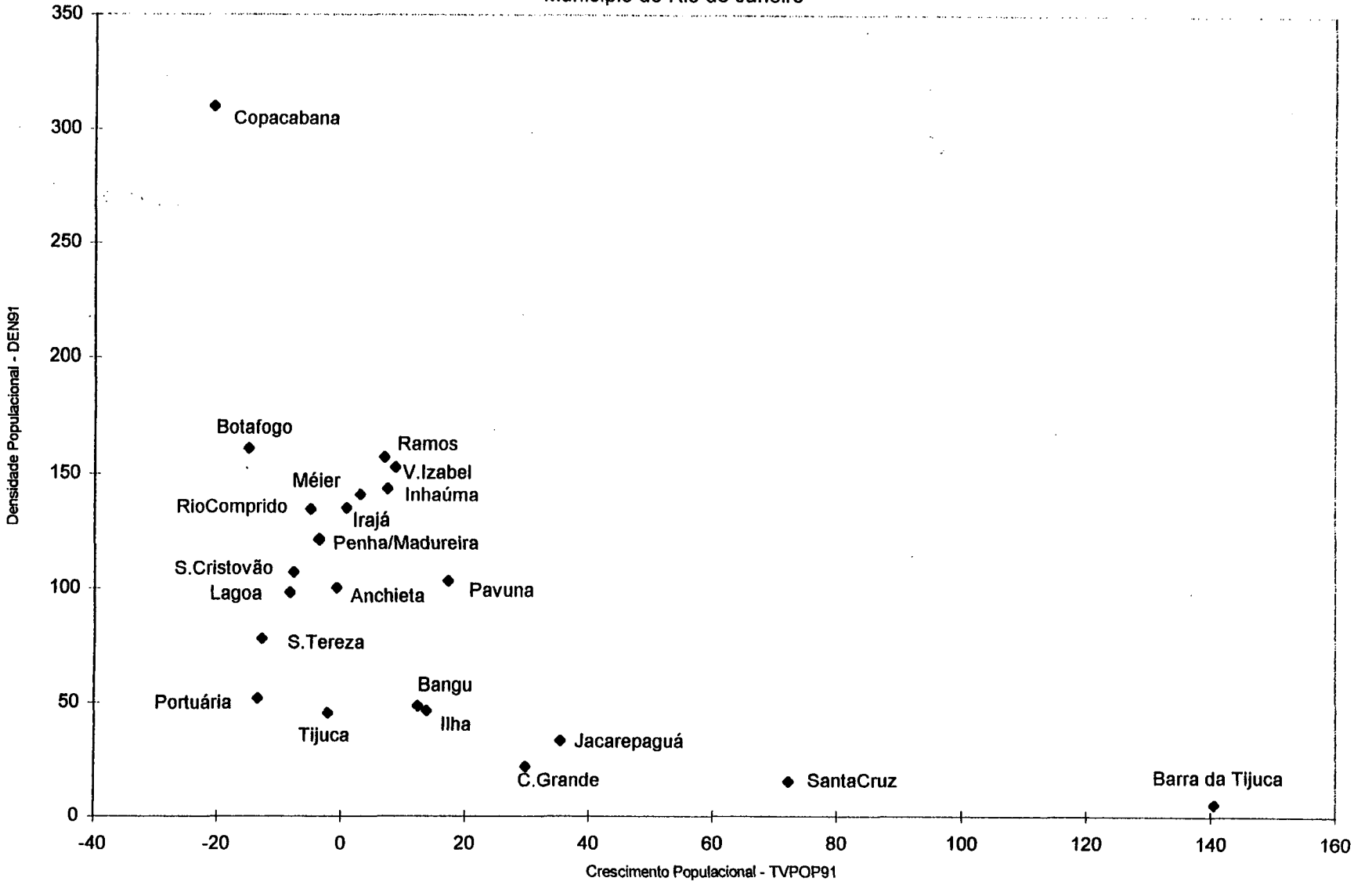


Figura 3

Diagrama de Dispersão: Crescimento Populacional x Densidade Populacional em Favelas
Município do Rio de Janeiro

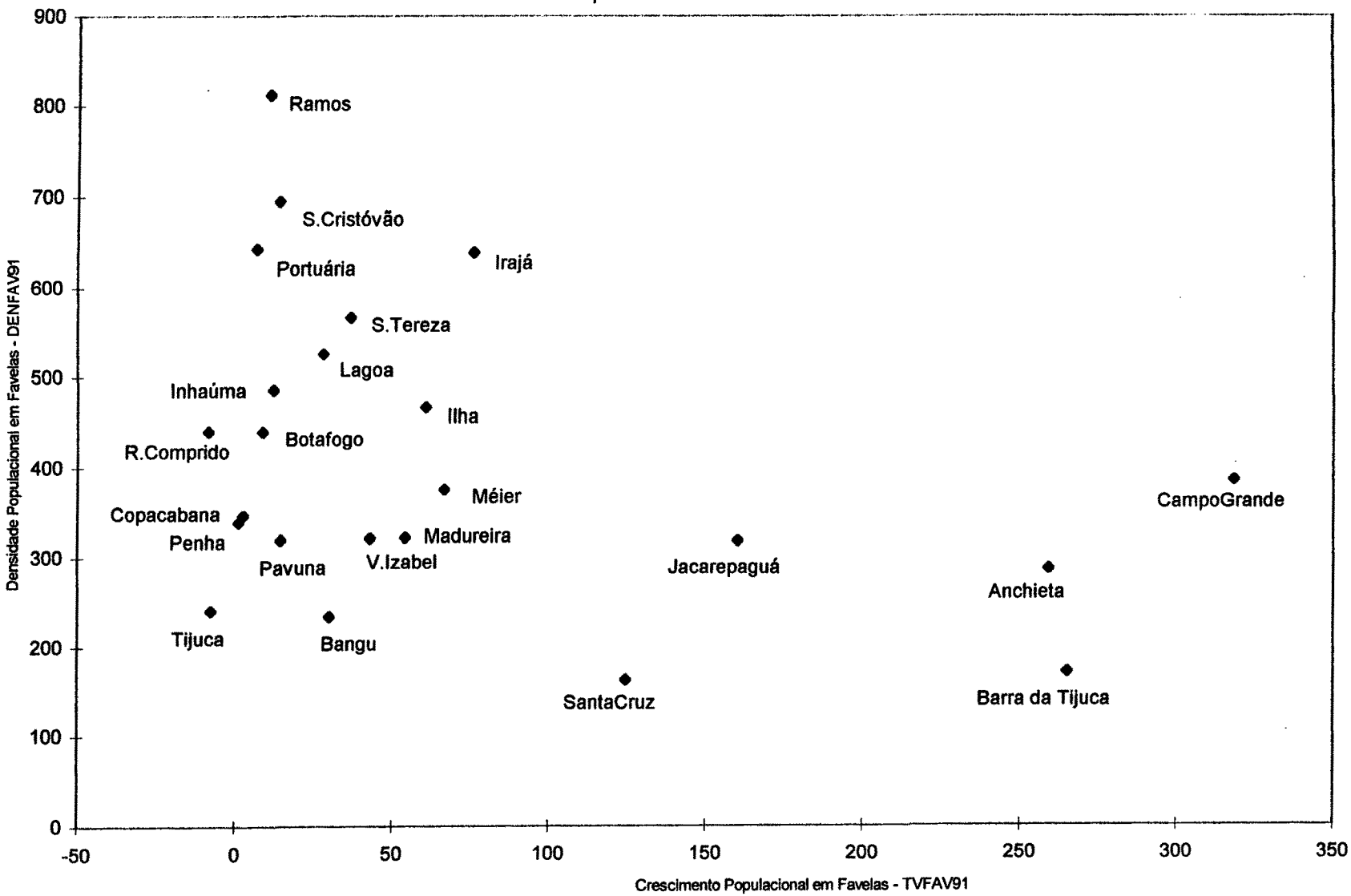


Figura 4

Diagrama de Dispersão: Crescimento Populacional Global X Crescimento Populacional em Favelas
Município do Rio de Janeiro

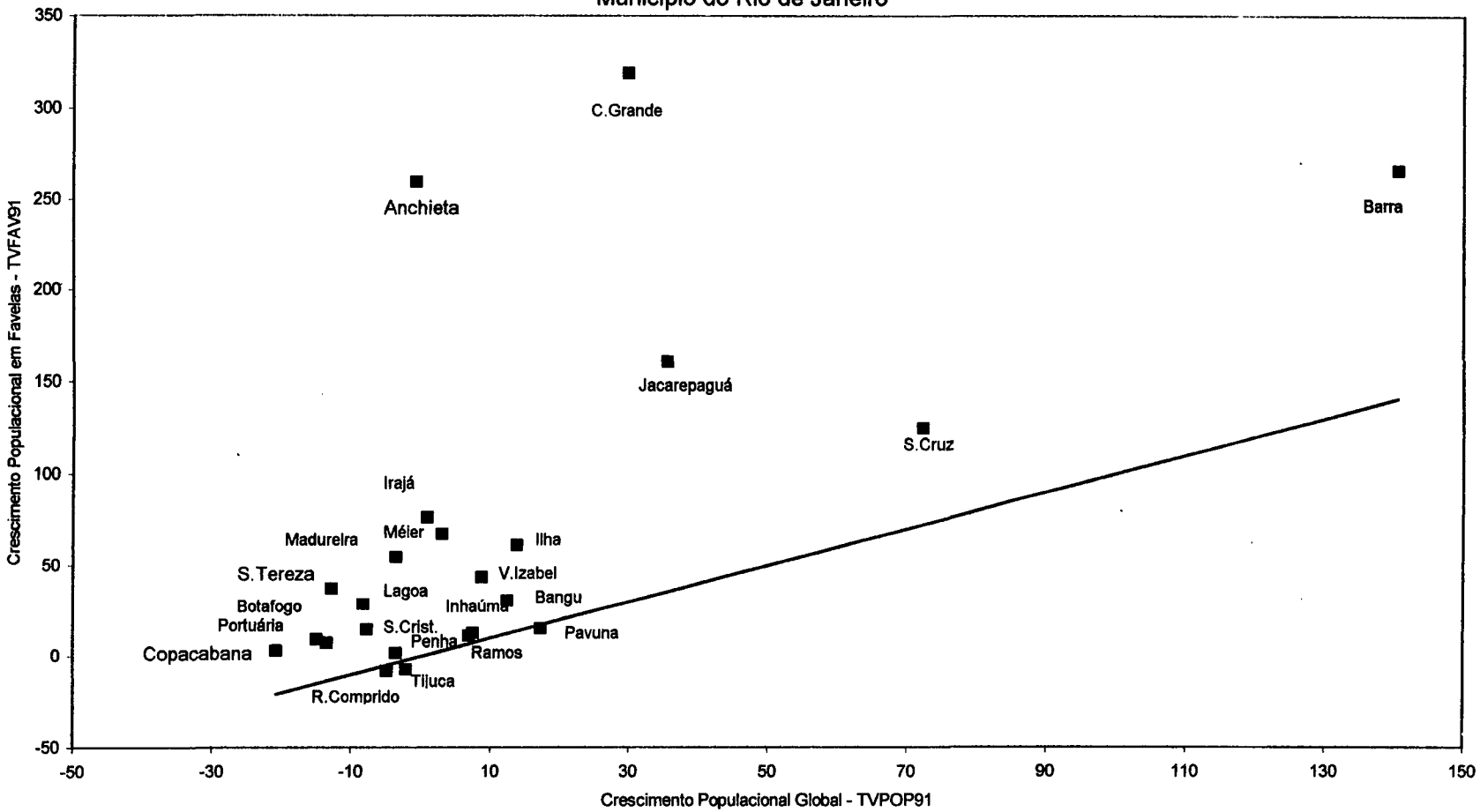


Figura 5
Diagrama de Dispersão: Percentual de Área Construída X Hab. Por área construída
Município do Rio de Janeiro

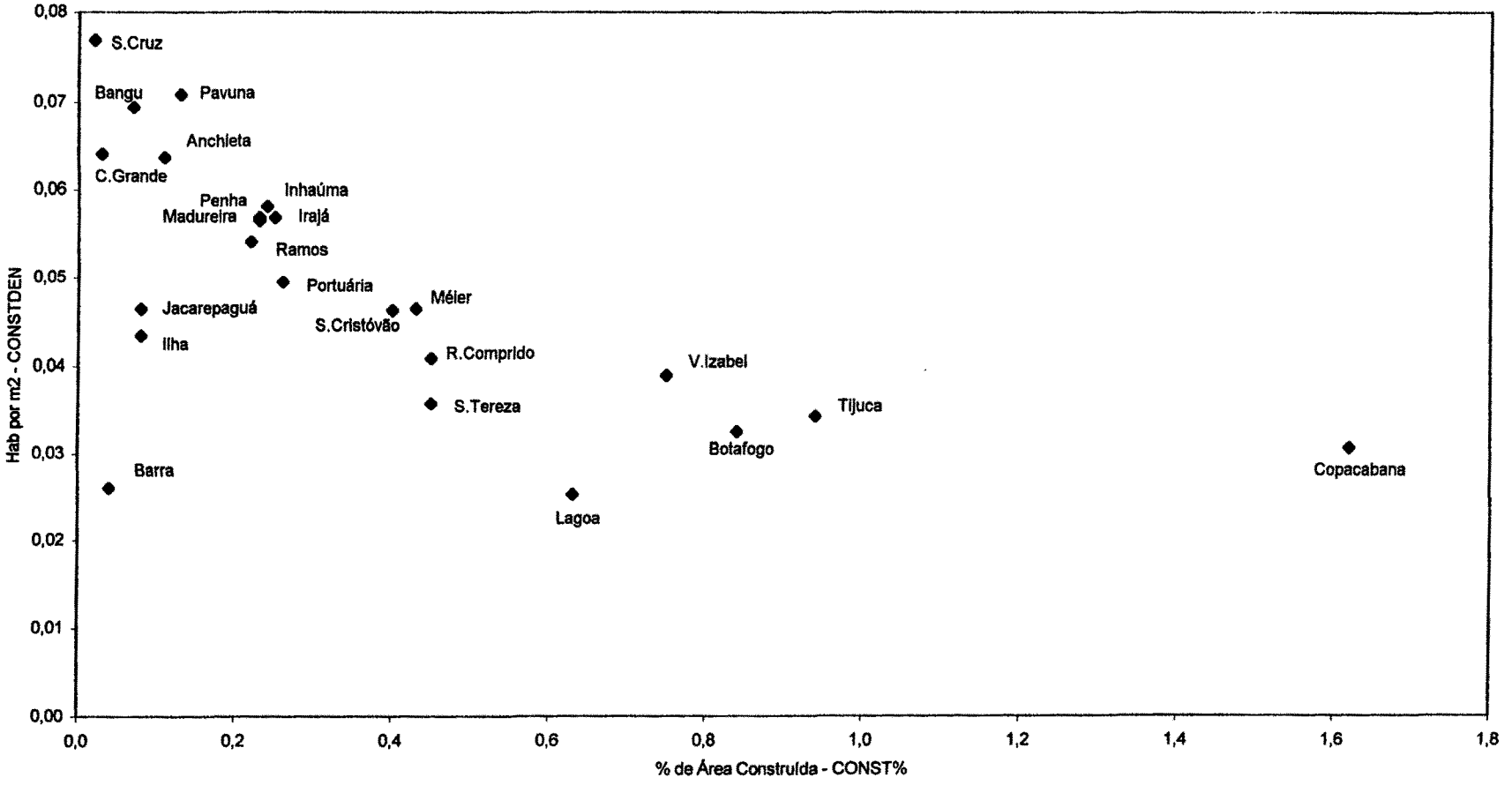


Figura 6

Dimensão Demográfica: Distribuição das RA's segundo as notas fatoriais

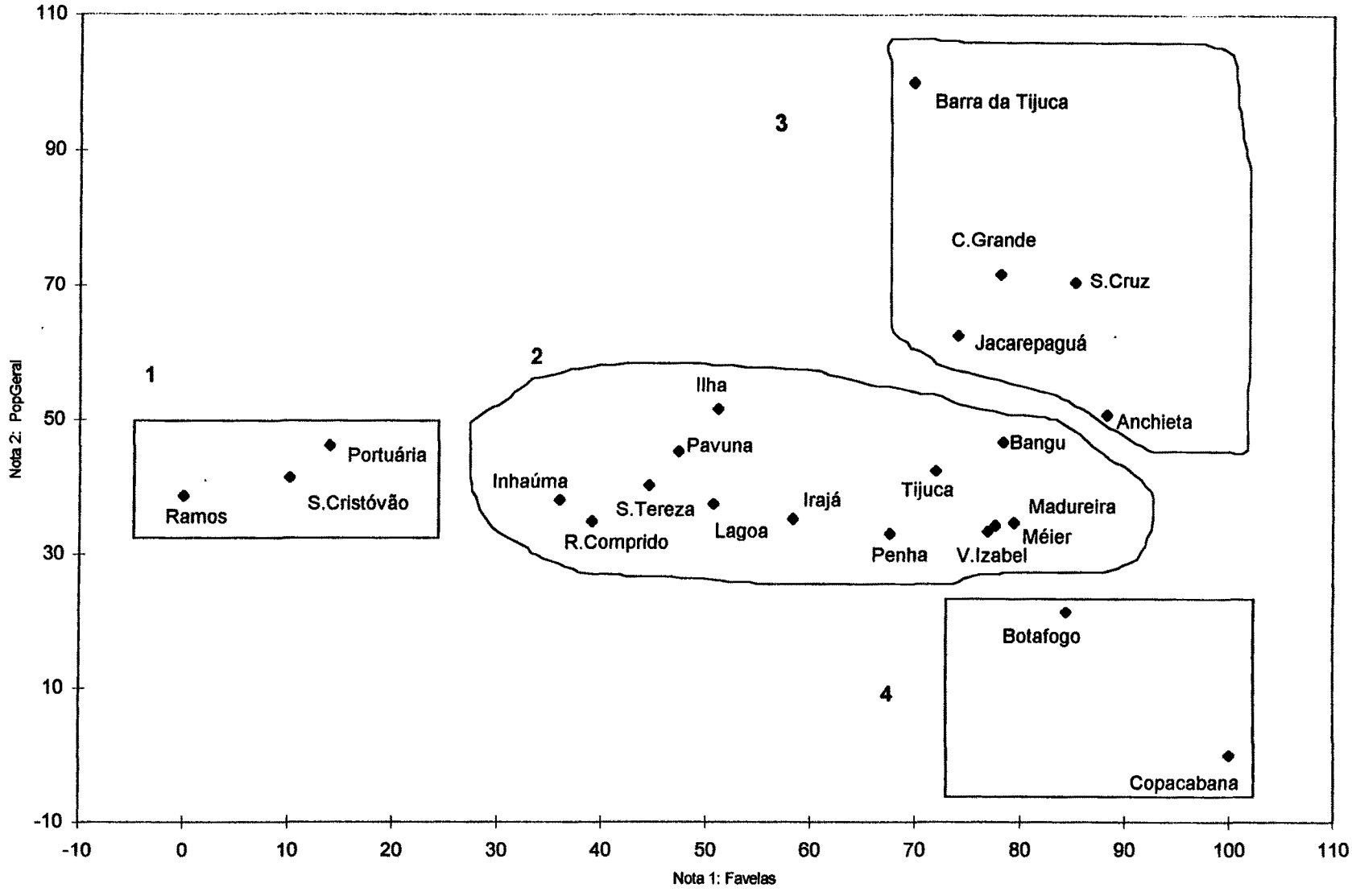


Figura 7

Dimensão Física: Distribuição das RA's segundo as notas fatoriais

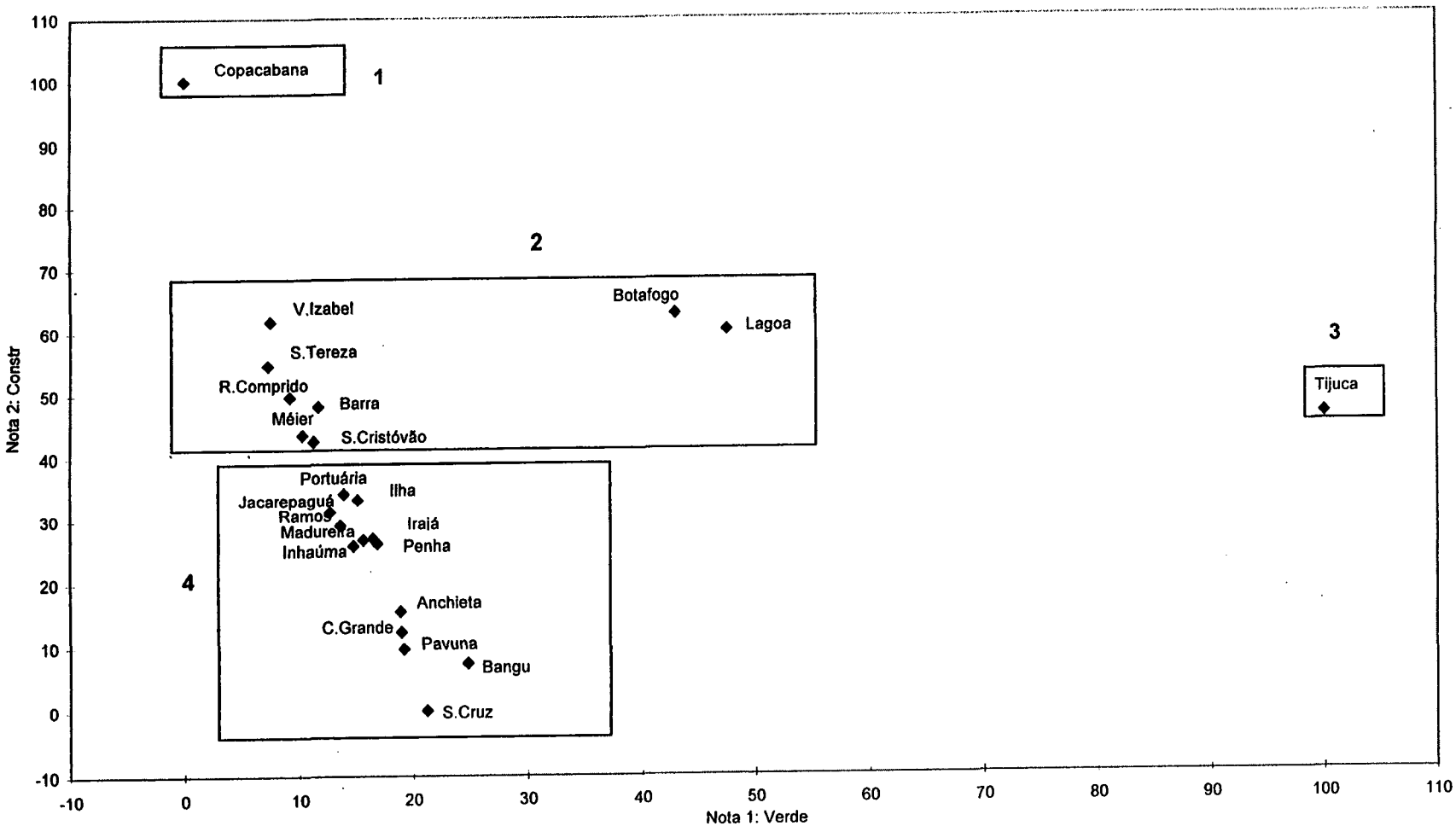


Figura 8

Dimensão Renda: Distribuição das RA's segundo as notas fatoriais

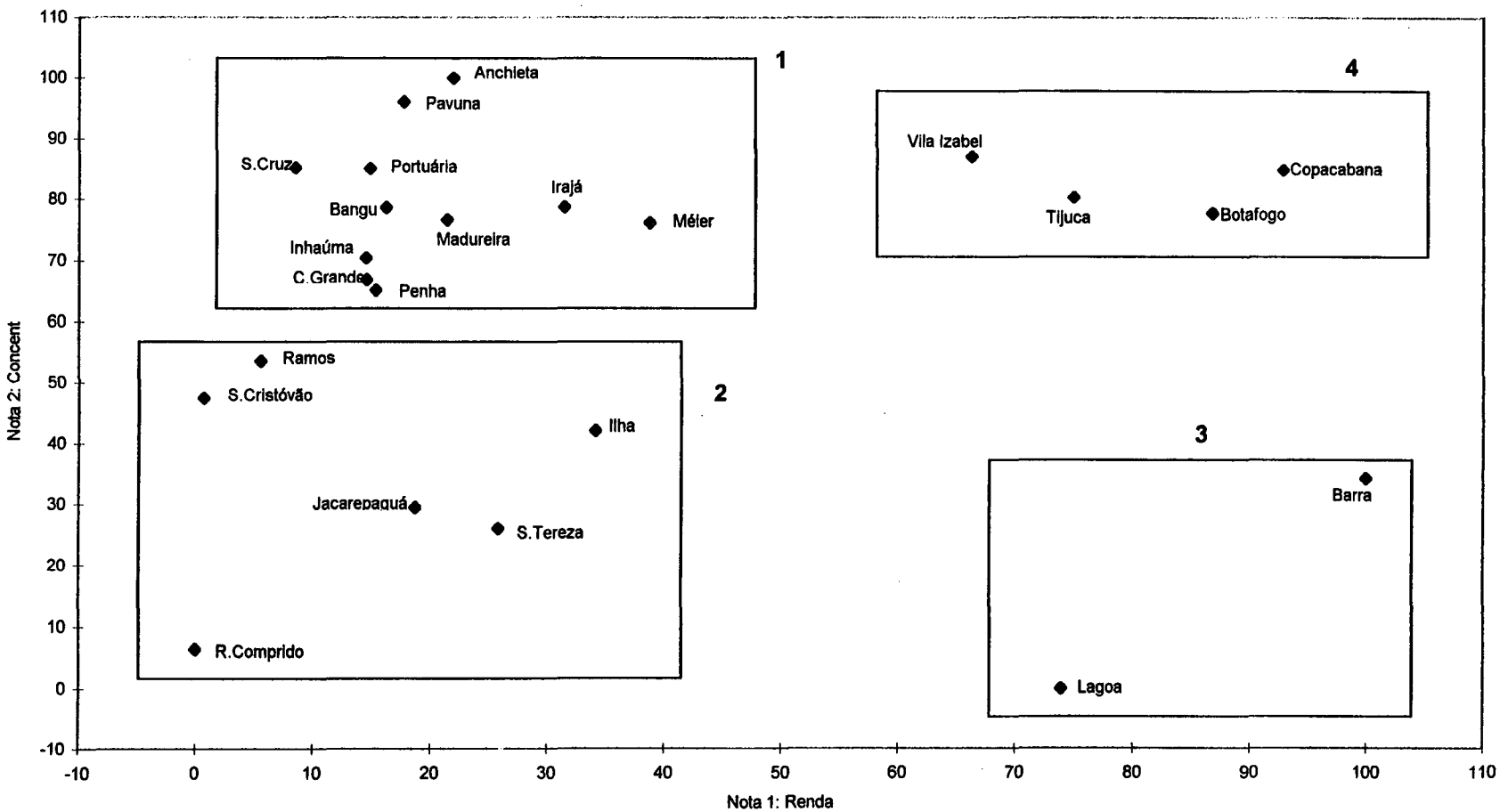


Figura 9

Dimensão Segurança: Distribuição das RA's segundo as notas fatoriais

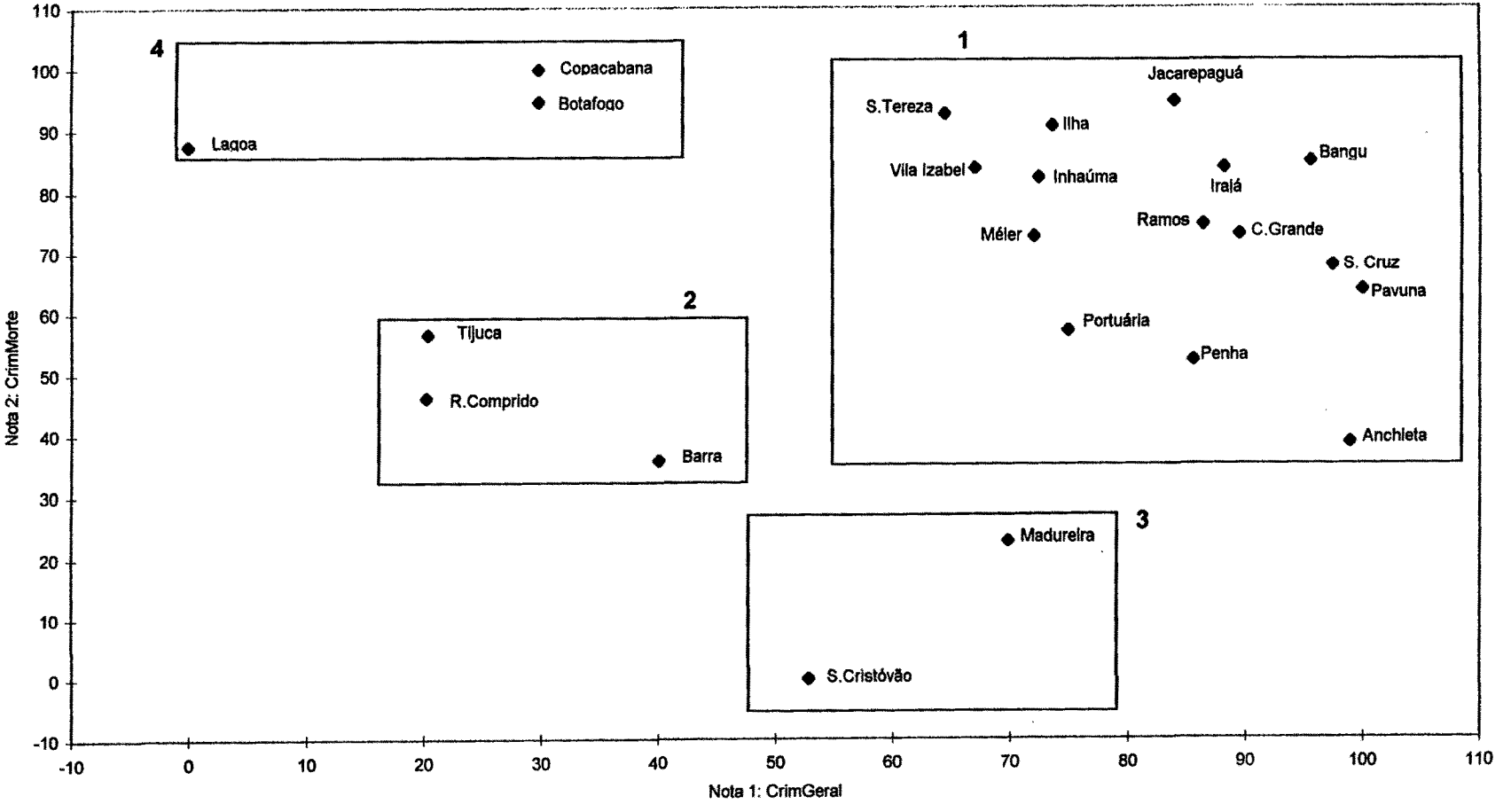


Figura 10

Dimensão Demográfica: Agrupamento das RA's segundo as notas fatoriais

Figura 11

Dimensão Física: Agrupamento das RA's segundo as notas fatoriais

Figura 12

Dimensão Renda: Agrupamento das RA's segundo as notas fatoriais

Figura 13

Dimensão Segurança: Agrupamento das RA's segundo as notas fatoriais

Anexo 2
Tabelas dos Dados

Tabela A1

População residente, crescimento populacional e densidade populacional total e em favelas por Região Administrativa do Município do Rio de Janeiro

RA	População Total [1]					População em Favelas [2]						
	1980	1991	Taxa de Crescimento (%)	Área (ha) 1991	Densidade Populacional 1991 (hab/ha)	1980	1991	Participação na população total (%) 1980	Participação na população total (%) 1991	Taxa de Crescimento (%)	Área (ha) 1991	Densidade Populacional 1991 (hab/ha)
	POP80	POP91	TVPOP91	AREA91	DEN91	FAV80	FAV91	FAV%80	FAV%91	TVFAV91	AREFAV91	DENFAV91
1 Portuária	50.906	44.070	-13,43	845	52,13	18.697	20.037	36,73	45,47	7,17	31	643,66
2 Centro	61.088	48.713	-20,26	640	76,11							
3 Rio Comprido	86.542	82.307	-4,89	611	134,66	29.518	27.066	34,11	32,88	-8,31	61	440,74
4 Botafogo	295.261	251.177	-14,93	1.558	161,20	13.816	15.059	4,68	6,00	9,00	34	440,45
5 Copacabana	213.809	169.446	-20,75	547	310,06	7.291	7.501	3,41	4,43	2,88	22	346,95
6 Lagoa	239.263	219.522	-8,25	2.232	98,37	44.489	57.091	18,59	26,01	28,33	108	526,33
7 São Cristóvão	86.542	79.852	-7,73	746	107,05	34.205	39.085	39,52	48,95	14,27	56	695,96
8 Tijuca	198.537	194.402	-2,08	4.259	45,64	28.593	26.484	14,40	13,62	-7,38	110	241,60
9 Vila Izabel	183.265	199.290	8,74	1.303	152,90	20.449	29.284	11,16	14,69	43,21	91	321,38
10 Ramos	264.716	283.126	6,95	1.800	157,32	127.789	141.992	48,27	50,15	11,11	175	812,50
11 Penha	325.805	314.299	-3,53	2.579	121,87	54.800	55.573	16,82	17,68	1,41	164	339,50
12 Inhaúma	193.448	207.958	7,50	1.447	143,72	65.072	73.136	33,64	35,17	12,39	150	486,70
13 Méier	402.166	414.522	3,07	2.937	141,13	33.075	55.202	8,22	13,32	66,90	147	376,01
14 Irajá	208.719	210.643	0,92	1.558	135,19	23.955	42.221	11,48	20,04	76,25	66	639,32
15 Madureira	386.982	373.187	-3,56	3.078	121,23	30.958	47.805	8,00	12,81	54,42	148	322,48
16 Jacarepaguá	315.624	427.491	35,44	12.786	33,44	22.376	58.255	7,09	13,63	160,35	183	318,21
17 Bangu	529.432	595.352	12,45	12.237	48,65	49.256	64.158	9,30	10,78	30,25	273	234,85
18 Campo Grande	292.715	380.057	29,84	17.167	22,14	9.458	39.584	3,23	10,42	318,52	103	385,88
19 Santa Cruz	147.630	254.363	72,30	16.373	15,54	7.036	15.814	4,77	6,22	124,76	97	162,78
20 Ilha do Governador	173.083	197.022	13,83	4.223	46,65	30.578	49.197	17,67	24,97	60,89	105	466,76
21 Paquetá	2.545	3.254	27,86	147	22,14							
22 Anchieta	142.540	141.539	-0,70	1.411	100,33	3.975	14.285	2,79	10,09	259,37	50	287,25
23 Santa Tereza	50.907	44.378	-12,83	570	77,86	9.699	13.300	19,05	29,97	37,13	23	567,41
24 Barra da Tijuca	40.726	97.990	140,61	17.567	5,58	4.609	16.842	11,32	17,19	265,42	98	171,21
25 Pavuna	152.721	179.078	17,26	1.735	103,23	46.105	52.929	30,19	29,56	14,80	166	319,74
26 Guaratiba	45.818	60.715	32,51	15.173	4,00	1.267	893	2,77	1,47	-29,52	1	992,22
Total	5.044.972	5.413.038	7,30	110.356	49,05	715.799	961.900	14,19	17,77	134,38	2.462	390,70
Média	201.799	216.522	10,55	4.414	97,37	31.122	41.822	17,15	21,48	68,83	107	415,12

Fonte: Anuário Estatístico da Cidade do Rio de Janeiro - 92/93 - IplanRIO

(1) Tabela 2.1.1 - pp. 125-127

(2) Tabela 3.2.7 - pp. 297-306

Tabela A2

Área verde e área construída, total e per capita,
por Região Administrativa do Município do Rio de Janeiro

RA	Área Verde			Área Construída		
	m^2 (1)	%	m^2/hab	Densidade Construída (4)	m^2 construído por habitante (5)	Habitante por m^2 construído
	VERDE (2)	**VERDE%** (3)	**VERDEHAB**	**CONST%**	**CONSTHAB**	**CONSTDEN**
1 Portuária	32.440	0,0384	0,7361	0,26	20,2	0,04950
2 Centro	190.664	0,2979	3,9140	1,74	20,1	0,04975
3 Rio Comprido	16.880	0,0264	0,3465	0,45	24,5	0,04082
4 Botafogo	1.495.483	2,4468	18,1696	0,84	30,8	0,03247
5 Copacabana	41.863	0,0269	0,1667	1,62	32,7	0,03058
6 Lagoa	2.020.474	3,6971	11,9240	0,63	39,4	0,02538
7 São Cristóvão	25.546	0,0114	0,1164	0,40	21,6	0,04630
8 Tijuca	5.748.081	7,7062	71,9842	0,94	29,2	0,03425
9 Vila Izabel	65.426	0,0154	0,3366	0,75	25,7	0,03891
10 Ramos	38.227	0,0293	0,1918	0,22	18,5	0,05405
11 Penha	203.165	0,1129	0,7176	0,23	17,6	0,05682
12 Inhaúma	41.637	0,0161	0,1325	0,24	17,2	0,05814
13 Méier	55.325	0,0382	0,2660	0,43	21,5	0,04651
14 Irajá	118.422	0,0403	0,2857	0,25	17,6	0,05682
15 Madureira	163.965	0,1052	0,7784	0,23	17,7	0,05650
16 Jacarepaguá	111.885	0,0363	0,2998	0,08	21,5	0,04651
17 Bangu	1.303.836	0,1020	3,0500	0,07	14,4	0,06944
18 Campo Grande	380.794	0,0311	0,6396	0,03	15,6	0,06410
19 Santa Cruz	184.572	0,0108	0,4856	0,02	13,0	0,07692
20 Ilha do Governador	233.130	0,0142	0,9165	0,08	23,0	0,04348
21 Paquetá	9.350	0,0022	0,0475	0,10	26,5	0,03774
22 Anchieta	91.867	0,0218	0,4663	0,11	15,7	0,06369
23 Santa Tereza	5.135	0,0349	1,5781	0,45	28,0	0,03571
24 Barra da Tijuca	196.530	0,1393	1,3885	0,04	38,3	0,02611
25 Pavuna	67.741	0,1188	1,5265	0,13	14,1	0,07092
26 Guaratiba	101.112	0,0058	1,0319	0,01	17,4	0,05747

Fonte: Anuário Estatístico da Cidade do Rio de Janeiro 92/93 - IplanRIO

(1) Tabela 3.1.2 - pp. 278-280

(2) Inclui praças, parques, largos, jardins e outros

(3) Percentual com relação à area total

(4) Tabela 3.1.3 - pp. 281-283; Área total construída dividida por Área territorial líquida

(5) Tabela 3.2.23- p. 293

Tabela A3

Renda média, pessoas de baixa renda e de alta renda, índice de Gini
por Região Administrativa do Município do Rio de Janeiro

RA	Renda Média em Salários Mínimos (3)	Pessoas de baixa renda (1)		Pessoas de renda alta (2)		Índice de Gini
		Total	% da população global	Total	% da população global	
	RMSM	MENOS1	MENOS1%	MAIS20	MAIS20%	INDGINI
1 Portuária	2,66	9.563	21,6996	205	0,4652	0,4973
2 Centro	4,57	5.234	10,7446	797	1,6361	0,5084
3 Rio Comprido	4,43	16.805	34,4980	2.084	4,2781	0,5883
4 Botafogo	12,17	11.629	14,1288	48.274	58,6511	0,5266
5 Copacabana	12,25	3.840	1,5288	32.082	12,7727	0,5214
6 Lagoa	16,27	22.188	13,0944	60.561	35,7406	0,6078
7 São Cristóvão	3,10	19.330	8,8055	646	0,2943	0,5370
8 Tijuca	10,05	10.174	12,7411	26.735	33,4807	0,5230
9 Vila Izabel	8,82	14.328	7,3703	19.891	10,2319	0,5129
10 Ramos	2,94	62.671	31,4471	2.064	1,0357	0,5325
11 Penha	3,23	59.542	21,0302	2.812	0,9932	0,5234
12 Inhaúma	3,12	41.779	13,2928	1.452	0,4620	0,5160
13 Méier	5,53	49.788	23,9414	13.255	6,3739	0,5195
14 Irajá	4,15	28.497	6,8747	2.988	0,7208	0,5148
15 Madureira	3,47	65.971	31,3189	2.908	1,3805	0,5119
16 Jacarepaguá	5,20	66.213	17,7426	16.370	4,3865	0,5681
17 Bangu	2,97	120.375	28,1585	3.377	0,7900	0,5066
18 Campo Grande	3,20	74.652	12,5391	3.201	0,5377	0,5206
19 Santa Cruz	2,44	63.708	16,7627	787	0,2071	0,4930
20 Ilha do Governador	6,22	22.248	8,7466	10.887	4,2801	0,5590
21 Paquetá	4,42	605	0,3071	73	0,0371	0,5598
22 Anchieta	3,04	29.259	14,8506	631	0,3203	0,4817
23 Santa Tereza	5,12	4.542	139,5821	1.493	45,8820	0,5787
24 Barra da Tijuca	18,11	6.370	4,5005	31.296	22,1112	0,5733
25 Pavuna	2,63	38.911	87,6808	440	0,9915	0,4848
26 Guaratiba	2,35	20.658	21,0817	397	0,4051	0,5615

Fonte: IBGE e IplanRIO

(1) Moradores em domicílios cujo chefe de família ganha menos de 1 Salário Mínimo

(2) Moradores em domicílios cujo chefe de família ganha mais de 20 Salários Mínimos

(3) Salário mínimo utilizado: Cr$36.161,60

Tabela A4

Indicadores de Segurança, por 100.000 habitantes, por Região Administrativa do Município do Rio de Janeiro [1]

RA	Delegacias da Polícia Civil	Homicídios		Roubo com morte	Furtos					Roubos				
		No trânsito (2)	Outros (3)		em veículos	de veículos e motos	em residências	no comércio	transeunte	Em veículos	de veículos e motos	em residência	no comércio	transeunte
		HOMITRAN	HOMIC	ROUMORTE	FUREMVEI	FVEIMOTO	FURRESID	FURCOMER	FURTRAN	ROUEMVEI	RVEIMOTO	ROURESID	ROUCOMER	ROUTRAN
1 Portuária	2	15,88	131,61	2,27	88,50	93,03	61,27	106,65	478,78	0,00	97,57	4,54	40,84	485,59
2 Centro	1,3,4	47,21	145,75	8,21	877,09	1161,90	82,11	1346,66	2465,46	26,68	410,56	16,42	529,63	2001,51
3 Rio Comprido	5,6	25,51	105,70	6,07	364,49	462,90	199,25	212,62	471,41	10,93	249,07	47,38	122,71	364,49
4 Botafogo	9,10	12,74	30,66	0,40	421,22	1171,68	170,80	97,14	225,74	4,38	242,86	29,86	96,74	179,55
5 Copacabana	12	4,72	15,34	2,36	367,08	512,26	205,38	108,59	616,13	2,95	77,90	34,23	69,64	201,83
6 Lagoa	13,14,15	18,68	29,61	3,19	674,19	974,39	225,95	132,11	486,06	9,57	247,81	55,12	111,15	269,22
7 São Cristóvão	17	70,13	174,07	5,01	154,03	576,07	115,21	191,60	340,63	5,01	569,80	25,05	170,32	384,46
8 Tijuca	18,19	19,03	74,59	4,12	327,67	1148,14	151,75	134,77	342,59	10,29	703,70	58,13	134,77	242,80
9 Vila Izabel	20	9,03	42,15	3,01	91,32	322,14	81,29	46,16	84,30	6,02	335,69	50,68	44,16	99,85
10 Ramos	21	15,54	67,46	1,06	45,56	200,97	46,27	54,04	50,15	2,12	350,02	13,42	59,69	83,36
11 Penha	22,38	16,23	70,00	4,77	63,00	298,76	63,00	72,86	24,82	1,59	880,37	21,64	56,95	92,27
12 Inhaúma	23	3,85	58,67	1,44	64,44	334,68	52,90	61,07	73,09	3,85	442,88	53,38	47,61	75,02
13 Méier	24,25,26	3,86	59,59	4,34	70,68	396,84	86,61	66,34	62,96	1,21	403,11	47,04	77,20	83,95
14 Irajá	27	3,80	64,56	0,95	38,45	197,49	54,12	41,30	25,64	0,95	374,57	24,69	34,66	47,00
15 Madureira	28,29,30,40	23,31	144,16	8,57	84,14	541,82	129,69	92,98	112,81	3,48	873,02	43,41	101,02	189,18
16 Jacarepaguá	32	8,89	38,36	0,23	54,50	203,51	68,54	36,26	33,22	0,23	173,80	27,37	36,02	43,28
17 Bangu	33,34	11,09	75,25	0,67	24,19	112,37	40,98	22,34	22,51	0,34	162,93	14,28	24,86	60,97
18 Campo Grande	35	22,10	80,25	2,37	31,84	152,87	69,73	35,26	32,10	0,00	217,60	30,00	45,78	63,41
19 Santa Cruz	36	26,73	124,63	1,97	13,76	46,00	54,65	25,16	66,05	0,79	106,15	25,95	27,52	45,21
20 I.Governador	37	7,61	50,25	1,52	108,62	392,34	102,02	48,22	128,92	1,02	135,52	23,35	48,73	76,13
21 Paquetá														
22 Anchieta	31	31,79	128,59	9,18	24,73	101,74	66,41	30,38	29,67	1,41	348,31	19,78	45,22	67,12
23 Santa Tereza	7	0,00	103,65	2,25	272,66	317,72	234,35	22,53	85,63	0,00	90,13	51,83	20,28	90,13
24 Barra da Tijuca	16	69,39	104,09	4,08	418,41	579,65	251,05	153,08	151,04	0,00	622,51	74,50	127,56	119,40
25 Pavuna	39	29,60	108,89	0,56	22,90	88,23	29,04	36,86	34,06	0,56	317,74	6,14	50,82	56,96
26 Guaratiba														

Fonte: Anuário Estatístico da Cidade do Rio de Janeiro - 92/93 - IplanRIO

(1) Tabela 2.7.7 - pp. 257-260 - dados totalizados por RA

(2) Inclui homicídios culposos por colisão e atropelamento

(3) Inclui homicídios dolosos consumados e tentados

Interpretando a Função Desvio em Modelos Lineares Generalizados

Renato Martins Assunção*

RESUMO

Apresentamos uma interpretação simples e intuitiva para a função desvio em alguns dos modelos lineares generalizados mais usados. Esta interpretação baseia-se no fato dessa função poder ser vista como uma soma ponderada de erros relativos com pesos que compensam por erros absolutos entre observação e ajuste.

1. INTRODUÇÃO

Introduzidos por Nelder e Wedderburn(1972) e posteriormente expandidos por McCullagh e Nelder(1989), os modelos lineares generalizados (MLG) tornaram-se um dos principais instrumentos estatísticos para a análise de dados do tipo regressão. Deste modo, esses modelos passaram a ser ensinados rotineiramente em cursos de pós-graduação em estatística e em outras áreas tais como economia. Essa difusão foi estimulada pela disponibilidade de vários pacotes estatísticos que oferecem rotinas para o ajuste de modelos MLGs e de vários livros-textos, inclusive em português, tais como: McCullagh and Nelder(1989), Cordeiro(1986) e Cordeiro e Paula(1989).

A exposição desses modelos usualmente apresenta em grande generalidade a classe MLG e conceitos associados e, posteriormente, estuda alguns casos particulares como a regressão logística, por exemplo.

Embora seja apresentado como uma generalização da soma dos resíduos ao quadrado de regressão linear, o conceito de *função desvio* (deviance, em inglês) é um dos menos intuitivos para o usuário do MLG. Em contraste com a simplicidade funcional e fácil visualização gráfica dos resíduos $r_i = y_i - \hat{y}_i$ em regressão linear, o desvio em modelos mais complexos mostra-se menos palpável e, em geral, é justificado e assimilado por meio de

* Endereço para correspondência: Deptº de Estatística - UFMG - C.Postal 702 - 30161-970-Belo Horizonte - MG. - E-mail: assuncao@est.ufmg.br.

argumentos teóricos. Hastie(1987), por exemplo, mostrou a íntima relação da função desvio com a distância de Kullback-Leibler. A importância da função desvio para testar o modelo e diagnosticar possíveis violações impõe um esforço extra para apresentá-la intuitivamente. Deste modo, o usuário pode sentir-se à vontade para manipulá-la e interpretá-la. Nesta nota, mostramos como a função desvio nos modelos de regressão logística e de Poisson pode ser interpretada de modo extremamente simples e intuitivo, facilitando sua compreensão e uso.

2. MODELOS LINEARES GENERALIZADOS

Em MLG com parâmetro de dispersão constante, observa-se n realizações $y_1,\ldots,y_n$, cada uma com um vetor associado de p covariáveis $x=(x_1,\ldots,x_p)$. A resposta y segue uma função de probabilidade ou densidade da forma

$$f(y_i \mid \theta_i) = \exp(y_i\theta_i - c(\theta_i))g(y_i).$$

A definição do MLG segue da distribuição acima e de uma relação explícita entre a média $\mu_i = E(y_i)$ e as covariáveis através da função de ligação $g(\mu_i) = \Sigma_r x_{ir}\beta_r, i = 1,\ldots,n.$ Um exemplo comum é a distribuição binomial $b(n_i, p_i)$ com ligação logística para a qual $\theta_i = g(p_i) = \log(p_i/(1-p_i))$ e $c(\theta_i) = n_i\log(1+\exp(\theta_i))$.

A função log-verossimilhança do MLG é dada por

$$L(y;\theta) = L(y_1,\ldots,y_n;\theta_1,\ldots,\theta_n) = \sum_{i=1}^{n}\log f(y_i \mid \theta_i) = \sum_{i=1}^{n}(y_i\theta_i - c(\theta_i)) + \sum_{i=1}^{n} g(y_i),$$ onde o último termo não depende de $\theta_1,\ldots,\theta_n$.

A estatística de qualidade global do ajuste do MLG é dada pela função desvio D definida como

$$D = 2\{L(y;y) - L(y;\hat{\mu})\}.$$

onde $L(y;\hat{\mu})$ é o máximo da log-verossimilhança do modelo ajustado e $L(y;y)$ é o máximo da log-verossimilhança quando tomamos $\mu_i = y_i$. Para testar a adequação de um MLG, compara-se D com quantis de uma distribuição qui-quadrado com $n-p$ graus de liberdade.

3. FUNÇÃO DESVIO

Alguns dos modelos lineares generalizados mais populares são aqueles onde a resposta possui distribuição de Bernoulli (regressão logística), distribuição binomial ou distribuição de Poisson. Nós vamos discutir em detalhes a interpretação da função desvio para o caso binomial e, a seguir, de modo breve, os outros dois casos que são similares ao primeiro.

Caso Binomial

A função desvio para dados do tipo binomial é dada por

$$D = 2\sum_{i=1}^{n}\left\{ y_i \log(\frac{y_i}{\hat{y}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{y}_i}) \right\}.$$

Uma forma intuitiva de apresentar essa expressão é a seguinte. O termo $\log(y_i/\hat{y}_i)$ leva em conta a diferença *relativa* do ajuste em relação a y_i. Assim, D tende a ser grande se a diferença relativa entre o observado e o ajustado é grande. No entanto, uma diferença relativa de, digamos, 10% quando temos valores pequenos de y_i tem conseqüências muito diversas de quando temos valores grandes de y_i. No segundo caso, quando y_i é grande, a diferença absoluta entre y_i e $\hat{y}_i$ tenderá a ser bem maior do que quando y_i é pequeno. Deste modo, os termos $\log(y_i/\hat{y}_i)$ deveriam ser ponderados por um fator crescente em y_i. Isto é exatamente o que aparece na primeira parte da fórmula da função desvio.

Com relação à segunda parte da função desvio, vamos começar com o termo $\log[(n_i - y_i)/(n_i - \hat{y}_i)]$. Como antes, esse termo expressa diferenças relativas, mas agora levando em conta a razão entre os números observados e estimados de fracassos ao invés do número de sucessos. Isso é razoável por dois motivos. O primeiro é a arbitrariedade da rotulação dos eventos como sucesso e fracasso, o que nos leva a repetir o raciocínio anterior de ponderação das diferenças relativas, agora em relação ao número de fracassos.

O outro motivo é que a primeira parte da função desvio, $\sum y_i \log(y_i/\hat{y}_i)$, não é suficiente para captar desvios do modelo. Suponha que $p_i \approx 0$ e que observamos $y_i \approx 0$. Um ajuste muito ruim poderia produzir $\hat{y}_i \approx n_i$. Nesse caso, o termo $y_i \log(y_i/\hat{y}_i)$ tenderá a ser pequeno (e negativo). Por outro lado, $(n_i - y_i)\log(n_i - y_i/n_i - \hat{y}_i)$ tenderá a ser positivo e grande, apontando o ajuste ruim.

Deste modo, a fórmula D para o caso binomial pode ser interpretada como uma soma ponderada das diferenças relativas entre o número de sucessos ou fracassos e seus valores estimados.

Caso Logístico

O caso logístico é um caso particular da binomial, com $n_i = 1$ e $y_i = 0$ ou 1. No entanto, a interpretação de D como soma ponderada é menos convincente e, neste caso, deve-se enfatizar apenas a característica de D medir a discrepância entre y_i e $\hat{p}_i = \hat{y}_i$ em termos da diferença relativa e não como uma diferença absoluta.

Caso Poisson

Quando os dados têm distribuição de Poisson com função de ligação $g(\mu_i) = \log(\mu_i)$, a função desvio é então dada por

$$D = 2\sum_{i=1}^{n} y_i \log(\frac{y_i}{\hat{y}_i}) - 2\sum_{i=1}^{n}(y_i - \hat{y}_i).$$

A situação mais comum na prática é aquela em que as covariáveis incluem um termo constante. Neste caso, a função desvio é dada por

$$D = 2\sum_{i=1}^{n} y_i \log(\frac{y_i}{\hat{y}_i}).$$

Novamente, o termo $\log(y_i/\hat{y}_i)$ caracteriza a discrepância entre observação e ajuste em termos relativos. No entanto, embora tanto $y_i = 2$ e $\hat{y}_i = 1$ quanto $y_i = 200$ e $\hat{y}_i = 100$ produzam $\log(y_i/\hat{y}_i) = 0.69$, não podemos considerar a falta de ajuste nestes dois casos do mesmo modo. É claro que a diferença entre $y_i = 200$ e $\hat{y}_i = 100$ deveria ser mais ponderada do que aquela entre $y_i = 2$ e $\hat{y}_i = 1$. É exatamente este papel de peso que exerce o fator y_i na fórmula de D.

4. EXEMPLO

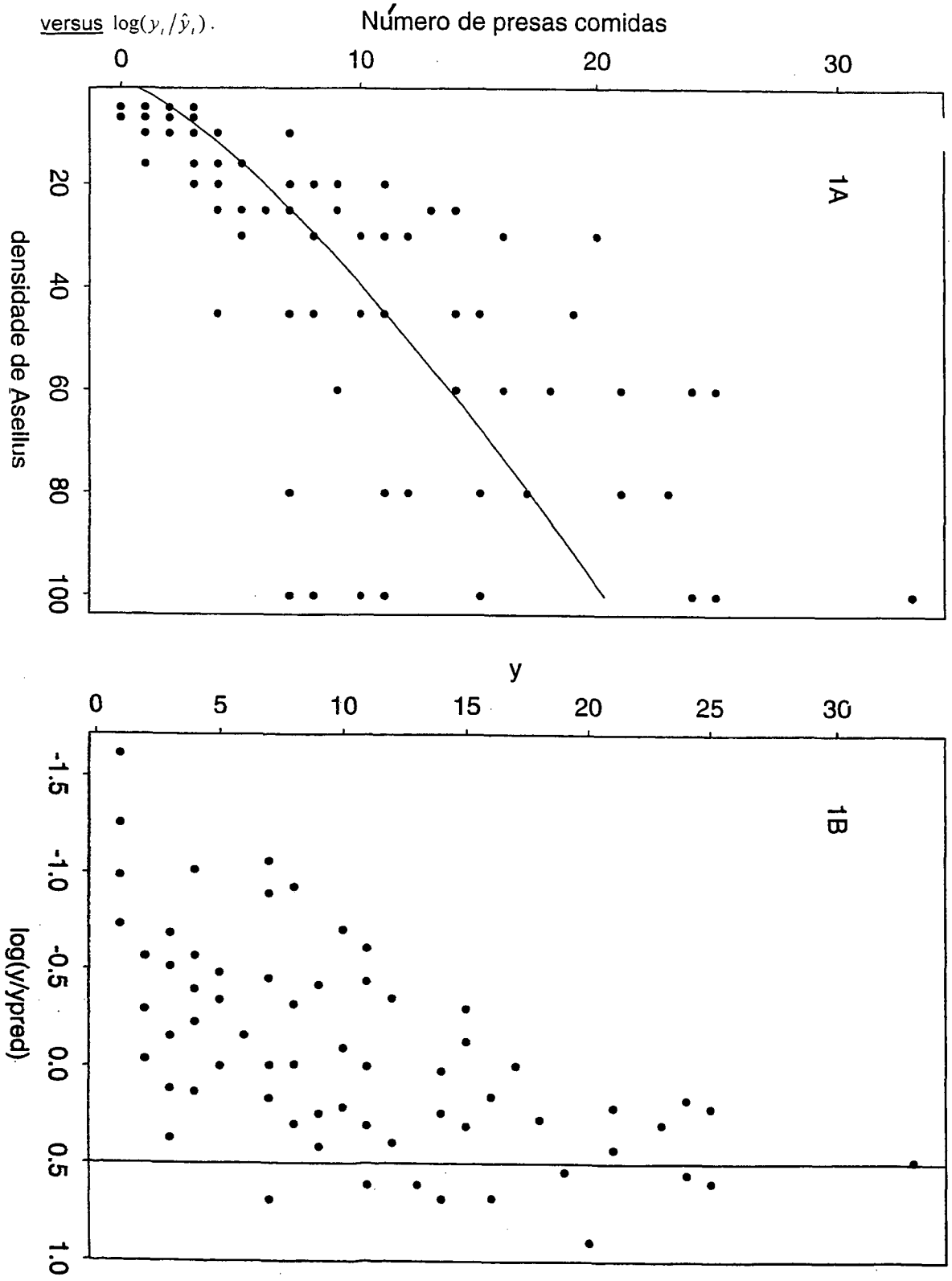
Em ecologia, é comum postular que o número de indivíduos y_i mortos por um predador individual é função da densidade N_i dos predados disponíveis inicialmente e a relação é chamada resposta funcional (Trexler et al., 1988). Numerosos modelos mecânicos e fenomenológicos são usados para descrever respostas funcionais e um deles, proposto por Ivlev(1961), supõe que existe uma taxa de saciedade decrescente de modo que y_i aproxima-se de um valor limite máximo à medida que N_i cresce.

Para analisar este problema, Hassell et al(1977) realizaram um experimento utilizando *Notonecta glauca*, um inseto aquático predador do crustáceo *Asellus aquaticus*. *Notonecta* individuais foram expostos a densidades de *Asellus* variando de cinco a cem e deixados alimentando-se por 72 horas, com reposição das presas a cada 24 horas. Foram feitas oito replicações em cada nível de densidade. O gráfico da Figura 1A mostra relação entre o número de *Asellus* mortos e sua densidade.

Para determinar a resposta funcional, pode-se modelar as contagens y_i como um MLG onde elas são independentes e possuem distribuição de Poisson com média μ_i onde $\mu_i = \alpha N_i^{\beta}$. Espera-se que $\beta < 1$. Sem a intenção de uma análise detalhada, o ajuste do MLG com a função de ligação canônica, produz a estimativa $\hat{\beta} = 0.76$ com desvio padrão estimado igual a 0.046. A função desvio é igual a 173.45 com 86.graus de liberdade. A linha sólida no gráfico da Figura 1A mostra o valor de $\hat{\mu} = \exp(-0.49 + 0.76\log(x))$ onde x é o valor da densidade de *Asellus*.

Legenda da Figura 1

1A: número observado de presas comidas versus densidade de presas no exemplo *Asellus-Notonecta* e o ajuste por MLG da curva de resposta funcional. 1B: gráfico de y_i versus $\log(y_i/\hat{y}_i)$.



O gráfico da Figura 1B mostra os valores de y_i versus $\log(y_i/\hat{y}_i)$. A linha vertical neste gráfico mostra os casos em que os valores observados são $\exp(0.5) = 1.65$ vezes aqueles preditos pelo modelo. No entanto, para efeito da função desvio, esta diferença relativa constante é ponderada diferencialmente dependendo do valor de y_i, o qual encontra-se na escala vertical.

5. CONCLUSÃO

Esta nota apresentou uma interpretação simples e intuitiva para a estatística desvio em alguns dos mais usados modelos lineares generalizados. Essa interpretação baseia-se na observação de que D pode ser visto como uma soma ponderada (dentre várias possíveis) de erros relativos com pesos que compensam por erros absolutos entre observação e ajuste.

REFERÊNCIAS

Cordeiro, G. M. (1986). *Modelos Lineares Generalizados.* VII SINAPE - Simpósio Nacional de Probabilidade e Estatística, Campinas, SP.

Cordeiro, G. M. e Paula, G. A. (1989). *Modelos de Regressão para Análise de Dados Univariados.* 17° Colóquio Brasileiro de Matemática, IMPA, RJ.

Hastie, T. (1987) A closer look at the deviance. *The American Statistician*, 41, 16-20.

Ivlev, V. S. (1961) *The experimental Ecology of the Feeding of Fishes.* Yale University Press, New Haven, CT.

Juliano, S. A. (1993) Nonlinear Curve Fitting: Predation and Functional Response Curve. In *Design and Analysis of Ecological Experiments* (eds Scheiner, S. M. e Gurevitch, J.), Chapman and Hall, New York, pp 159-182.

McCullagh, P. e Nelder, J. (1989). *Generalized Linear Models*, segunda edição. London: Chapman and Hall.

Nelder, J. e Wedderburn, R. (1972) Generalized linear models. *Journal of the Royal Statistical Society*, Ser. A, 135, 370-384.

Trexler, J. C., McCulloch, C. E., and Travis, J. (1988) How can the functional response best be determined ? *Oecologia*, 76, 206-214.

AGRADECIMENTOS

Desejo agradecer as sugestões feitas pelos pareceristas e pelo editor os quais melhoraram o artigo. Este trabalho foi desenvolvido com apoio financeiro da Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG - e do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq.

ABSTRACT

This paper presents a simple and intuitive interpretation to deviance in some of the most used generalized linear models. The deviance can be interpreted as a weighted sum of the relative difference between observation and fitted value with weights allowing for absolute differences.

Análise Estatística de Dados de Pesquisas por Amostragem: Problemas no Uso de Pacotes-Padrões

Djalma Galvão Carneiro Pessoa*

Pedro Luis do Nascimento Silva*

Renata Pacheco Nogueira Duarte*

RESUMO

Em análises descritivas das estimativas de quantidades populacionais de interesse, como médias, totais e razões, as agências produtoras de dados estatísticos consideram os pesos e o desenho amostral utilizados na obtenção destes dados. Mas a modelagem e a análise dos dados das pesquisas por amostragem são feitas, em geral, por usuários que trabalham fora destas agências e não levam em conta, na análise, os pesos e, principalmente, o desenho amostral. Os pacotes estatísticos usuais assumem que os dados são IID (independentes e identicamente distribuídos), ou seja, provenientes de uma amostragem aleatória simples com reposição - AASC. As pesquisas por amostragem, em geral, possuem um desenho amostral mais complexo, o que pode influenciar, principalmente, a estimação da variância das estimativas pontuais. O objetivo deste trabalho é estudar esta influência na análise feita por Leote(1996), em dados do Suplemento Trabalho da Pesquisa Nacional por Amostra de Domicílios - PNAD de 1990, do Estado do Rio de Janeiro, em que se utilizou um pacote estatístico usual para o ajuste de um modelo logístico aos dados. Foi feito o ajuste de um modelo semelhante, usando-se um pacote estatístico que leva em conta o desenho amostral e os pesos da PNAD 90 e os resultados foram comparados com aqueles obtidos, sem levar tais aspectos em conta. Além disso, discutiu-se o impacto em estudos analíticos de não se considerar aspectos importantes do desenho amostral, como estratificação e conglomeração, e as dificuldades encontradas por usuários de dados de pesquisas por amostragem (como a PNAD), quando querem fazer modelagem dos dados, já que, em geral, desconhecem características importantes do desenho amostral.

1. INTRODUÇÃO

Este artigo trata de problema de grande importância para os usuários de dados obtidos através de pesquisas amostrais por agências produtoras de dados estatísticos. No dia-a-dia destas agências são feitas análises descritivas com estimativas de totais, médias e razões,

* Endereço para correspondência: IBGE - Av. República do Chile, 500 - 10º andar - Centro - 20031-170 - E-mail: pedrosilva@ibge.gov.br.

onde são devidamente considerados os pesos e o desenho da amostra que geraram os dados.

É comum, porém, a utilização destes dados também na construção e ajuste de modelos em análises secundárias. Nestas análises secundárias, usualmente feitas por analistas que trabalham fora das agências produtoras dos dados, tais aspectos não são incorporados na análise. Em geral, são utilizados pacotes estatísticos que assumem hipóteses básicas válidas somente quando os dados são obtidos através de amostras aleatórias simples com reposição - AASC. Tais pacotes estatísticos não consideram os seguintes aspectos relevantes no caso de amostras complexas:

1) probabilidades distintas de seleção das unidades;

2) conglomeração das unidades;

3) estratificação;

4) não-resposta e outros ajustes.

Estimativas pontuais de parâmetros da população são influenciadas pelos pesos distintos das observações. Além disso, as estimativas de variância são influenciadas pela conglomeração, estratificação e pesos. Ao ignorar estes aspectos, os pacotes tradicionais de análise podem produzir resultados incorretos, tanto para as estimativas pontuais como para as respectivas variâncias.

O objetivo deste artigo é analisar as conseqüências desta simplificação na análise dos dados. Isto é feito considerando-se uma análise de dados obtidos na **Pesquisa Nacional por Amostra de Domicílios - PNAD**, Suplemento de Trabalho de 1990, contida em monografia do curso de Estatística Aplicada de 1993 da ENCE, e que posteriormente foi publicada na coleção Relatórios Técnicos da ENCE por **Leote(1996).** Nas análises estatísticas de Leote(1996) foi utilizado o pacote estatístico **GLIM** (Cordeiro, 1986), desenvolvido para ajustar modelos lineares generalizados. As análises foram replicadas neste trabalho com pequenas alterações. Para isto utilizou-se a função **glm** do **S-Plus** (Venables e Ripley, 1994), que produz resultados equivalentes ao **GLIM**. A escolha do **S-Plus** deveu-se ao fato de este conter funções que agilizam a seleção de modelos. Após a seleção de um modelo adequado no S-Plus, foram feitas análises dos mesmos dados através do **SUDAAN (Shah et alii, 1993),** que é um pacote de análise estatística especialmente destinado à análise de dados obtidos por amostras complexas e que permite considerar na análise tanto os pesos como o desenho amostral utilizado.

A Seção 2 contém uma descrição dos dados e discute o objetivo da análise. A Seção 3 resume a seleção do modelo, seguindo os mesmos passos usados em Leote(1996). A Seção 4 descreve o desenho amostral da PNAD, que foi considerado na análise feita na Seção 6 através do pacote SUDAAN. A Seção 5 apresenta os fundamentos teóricos para incorporação dos pesos e do desenho amostral na análise, usados inclusive nas análises do SUDAAN. Na Seção 7 é discutido o impacto em estudos analíticos, resultante da não consideração de aspectos importantes do desenho amostral, tais como estratificação e

conglomeração. Finalmente, são discutidas as dificuldades encontradas pelos usuários dos dados de pesquisas como a PNAD para modelagem, já que estes por vezes desconhecem aspectos importantes do desenho amostral.

2. DESCRIÇÃO DO PROBLEMA

O estudo de Leote(1996) foi feito com o objetivo de obter um perfil sócio-econômico das pessoas ocupadas no setor informal da economia na área urbana do Rio de Janeiro. A definição adotada de setor informal é amplamente discutida em Leote(1996) e os dados utilizados são relativos a pessoas que:

- moravam em domicílios urbanos do Estado do Rio de Janeiro;
- trabalhavam em atividades mercantis (não foram incluídos trabalhadores domésticos);
- na semana da pesquisa estavam trabalhando ou não estavam trabalhando por estarem de férias, licença, etc., mas tinham trabalho;
- desenvolviam atividades não-agrícolas.

As pessoas que trabalhavam em locais com até cinco pessoas ocupadas foram classificadas no setor informal, independente da posição de ocupação delas, enquanto as que trabalhavam em locais com mais de cinco pessoas ocupadas foram classificadas no setor formal. O trabalho refere-se ao trabalho principal da pessoa entrevistada. Para a variável renda considerou-se a soma dos rendimentos de todos os trabalhos.

Foi considerada uma amostra de 6 507 pessoas (após a exclusão de nove registros considerados atípicos), classificadas de acordo com as variáveis descritas no Quadro 1, todas tratadas como fatores na análise.

Quadro 1
Descrição das variáveis explicativas

Fatores	Níveis	Descrição dos níveis
Sexo (sx)	sx(1)	Homens
	sx(2)	Mulheres
Anos de estudo (ae)	ae(1)	Até 4
	ae(2)	De 5 a 8
	ae(3)	9 ou mais
Horas trabalhadas (ht) (em todos os trabalhos, por semana)	ht(1)	Menos de 40
	ht(2)	De 40 a 48
	ht(3)	Mais de 48
Idade em anos completos (id)	id(1)	Até 17
	id(2)	De 18 a 25
	id(3)	De 26 a 49
	id(4)	50 ou mais
Rendimento médio mensal (re) (de todos os trabalhos, em salários mínimos)	re(1)	Menos de 1
	re(2)	De 1 a 5
	re(3)	Mais de 5

Os fatores acima foram tomados como explicativos e a variável resposta foi o indicador de pertinência ao setor informal da economia. Foi ajustado um modelo logístico (Agresti, 1990) para explicar a probabilidade de uma pessoa pertencer ao setor informal da economia.

Leote(1996) utilizou três níveis para a variável idade. Neste trabalho introduzimos mais um nível para esta variável, com a finalidade de isolar o efeito na faixa de 18 a 25 anos. Contudo, os objetivos básicos da modelagem foram preservados.

3. SELEÇÃO DO MODELO

Para a seleção do modelo foi usada a função glm do S-Plus, aplicada aos dados da tabela de contingência obtida por cruzamento da variável resposta com todas as variáveis explicativas descritas no Quadro 1. O modelo final selecionado foi escolhido passo a passo, incluindo em cada passo as interações que produziam maior decréscimo do desvio residual, considerando a perda de graus de liberdade. O modelo selecionado foi

$$\log\left(\frac{p_{ijklm}}{1-p_{ijklm}}\right) = \mu + \beta_i^{sx} + \beta_j^{ae} + \beta_k^{ht} + \beta_l^{id} + \beta_m^{re} + \beta_{il}^{sx.id} + \beta_{ik}^{sx.ht} + \beta_{jk}^{ae.ht} + \beta_{kl}^{ht.id} + \beta_{km}^{ht.re} \quad (1)$$

onde p_{ijklm} é a probabilidade de pertencer ao setor informal correspondente à combinação de níveis das variáveis explicativas, sendo i=1, 2 o nível de **sx**; j=1, 2, 3 o nível de **ae**; k=1, 2, 3 o nível de **ht**; l=1, 2, 3, 4 o nível de **id**; e m=1, 2, 3 o nível de **re**.

Os efeitos foram adicionados seqüencialmente na ordem da Tabela 1. Depois de introduzidos os efeitos principais, as interações de dois fatores foram introduzidas na ordem definida pela função **step** do S-Plus.

Tabela 1
Análise de desvios

Efeito	Diminuição nos g.l.	Diminuição no Desvio	g.l.	Desvio Residual
Nulo			179	1305,309
sx	1	24,917	178	1280,392
ae	2	328,040	176	952,352
ht	2	491,319	174	461,032
id	3	37,059	171	423,974
re	2	105,982	169	317,992
ht:re	4	48,632	165	269,360
ht:id	6	36,560	159	232,800
sx:id	3	20,109	156	212,690
sx:ht	2	13,106	154	199,585
ae:ht	4	20,284	150	179,301

O p-valor do teste de nulidade das interações não incluídas no modelo é 0,0515, aceitando-se a hipótese de nulidade destes efeitos ao nível $\alpha = 5\%$. O modelo obtido difere do selecionado em Leote(1996) só pela inclusão de mais um efeito, referente à interação ae:ht.

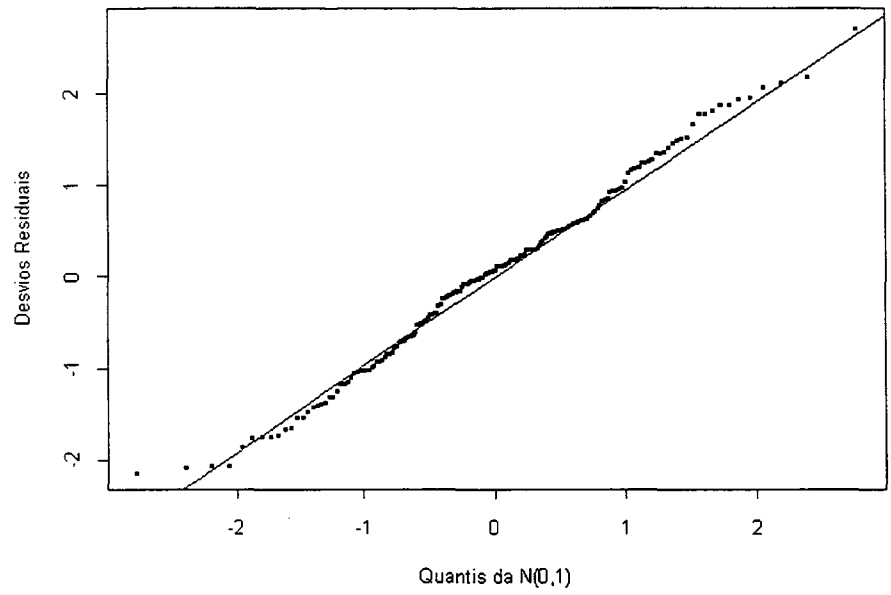
As estimativas dos coeficientes do modelo (1) e respectivos desvios-padrões obtidos função glm do S-Plus são dadas na Tabela 2. Na estimação dos coeficientes foi utilizada ametrização que toma como referência o último nível de cada variável explicativa.

Tabela 2
Estimativas dos coeficientes e respectivos desvios-padrões

Efeito	Coeficiente	Desvio-Padrão	Valor t
Intercepto	-0,5137	0,2694	-1,9071
sx	0,1562	0,2284	0,6839
ae1	0,7398	0,1649	4,4863
ae2	0,4971	0,1590	3,1270
ht1	-0,3859	0,3124	-1,2354
ht2	-0,6976	0,2676	-2,6067
id1	-0,2432	0,4917	-0,4945
id2	-0,7237	0,3136	-2,3077
id3	0,2266	0,2336	0,9698
re1	0,2926	0,2453	1,1930
re2	0,0618	0,1453	0,4250
ht1.re1	1,5306	0,3323	4,6057
ht2.re1	0,3361	0,2836	1,1852
ht1.re2	0,4981	0,2212	2,2516
ht2.re2	-0,1116	0,1784	-0,6257
ht1.id1	-1,4077	0,5148	-2,7344
ht2.id1	-0,3969	0,4648	-0,8539
ht1.id2	-0,1289	0,3512	-0,3671
ht2.id2	-0,1056	0,2858	-0,3695
ht1.id3	-0,2162	0,2525	-0,8563
ht2.id3	-0,5334	0,2006	-2,6585
sx.id1	0,8700	0,3350	2,5969
sx.id2	0,2936	0,2262	1,2980
sx.id3	-0,2629	0,1859	-1,4141
sx.ht1	-0,7375	0,2115	-3,4876
sx.ht2	-0,0933	0,1821	-0,5124
ae1.ht1	0,7919	0,2388	3,3165
ae2.ht1	0,7354	0,2256	3,2597
ae1.ht2	0,0292	0,1963	0,1487
ae2.ht2	0,0867	0,1886	0,4596

O Gráfico 1 apresenta um comportamento aceitável dos resíduos padronizados, sem a presença de pontos atípicos, indicando um bom ajuste do modelo.

Gráfico 1
Gráfico normal dos resíduos



A Tabela 3 mostra que, em geral, as proporções mais baixas de informal ajustadas pelo modelo correspondem a pessoas com ae no nível 3 (nove ou mais anos de estudo), ht no nível 2 (de 40 a 48 horas semanais trabalhadas) e re nos níveis 2 (1 a 5 salários mínimos) e 3(mais de 5 salários mínimos).

Tabela 3
Ajustes com menores proporções no setor informal

sx	ae	ht	id	re	Probabilidade de informal ajustada
2	3	2	2	2	0,110
2	3	2	2	3	0,115
1	3	1	2	3	0,115
2	3	1	1	2	0,120
2	3	2	1	2	0,120
1	3	2	3	2	0,146
2	3	1	2	3	0,148
1	3	2	2	2	0,150
1	3	1	3	3	0,150
1	3	2	3	3	0,152

A Tabela 4 mostra que, em geral, as proporções mais altas de informal ajustadas pelo modelo correspondem a re na faixa 1 (menos de um salário mínimo), ht na faixa 1 (menos de 40 horas semanais trabalhadas) e ae nos níveis 1 (até quatro anos de estudo) e 2 (de cinco a oito anos de estudo).

Tabela 4
Ajustes com maiores proporções no setor informal

sx	ae	ht	id	re	Probabilidade de informal ajustada
2	2	1	2	1	0,786
1	1	1	2	1	0,788
1	2	1	3	1	0,790
2	1	1	2	1	0,832
1	1	1	3	1	0,835
1	1	1	4	1	0,867
2	2	1	4	1	0,896
2	2	1	3	1	0,897
2	1	1	4	1	0,921
2	1	1	3	1	0,922

Vamos designar por **vantagem** (*odds* em inglês; Agresti, 1990, p.14) do informal a razão $V = p/(1-p)$, onde p é probabilidade de trabalhar no setor informal da economia. Para fins de comparação, vamos calcular alguns intervalos de confiança através do S-Plus, que serão recalculados posteriormente pelo SUDAAN. No modelo *logit* é de interesse estudar a razão entre vantagens, quando se passa de um nível a outro de um fator explicativo.

Vamos considerar, por exemplo, o que ocorre com a vantagem do informal quando se passa do nível 1 para o nível 2 e do nível 2 para o 3 de anos de estudo. Devido à interação entre ae e ht existente no modelo, a análise deve ser feita para níveis fixos de ht. Os resultados são apresentados na Tabela 5.

Tabela 5
Intervalos de confiança de 95% para razões de vantagens, variando-se os níveis de ae para níveis fixos de ht

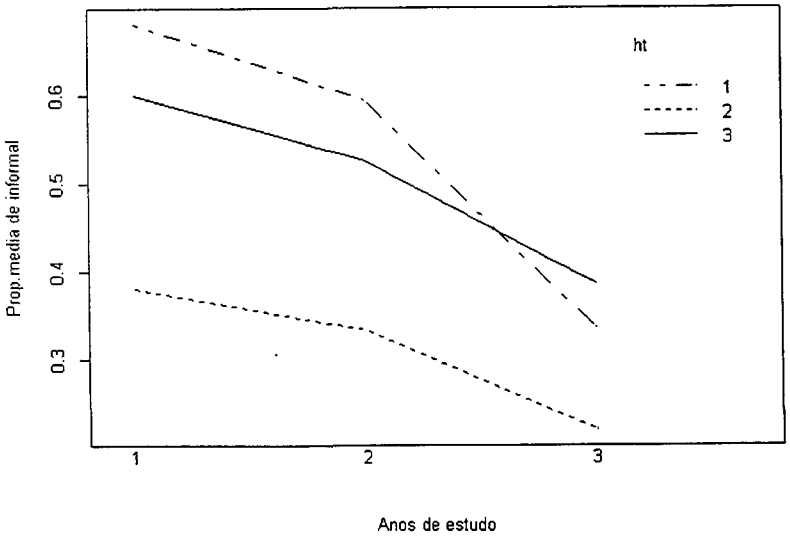
ht	Mudança de nível de ae	Razão de vantagens estimada	Intervalo de 95% para razão de vantagens
1	de 1 para 2	0,741	(0,530; 1,036)
1	de 2 para 3	0,291	(0,213; 0,399)
2	de 1 para 2	0,831	(0,697; 0,991)
2	de 2 para 3	0,558	(0,457; 0,680)
3	de 1 para 2	0,785	(0,586;1,050)
3	de 2 para 3	0,608	(0,445; 0,831)

De acordo com a Tabela 5, linha 1, a razão de vantagens do informal quando se passa do nível 1 para o nível 2 de ae, fixando ht=1, é de 0,741, cujo intervalo de confiança de 95% é (0,530;1,036). Como esse intervalo contém o valor 1, devemos concluir que não há diferença significativa nas vantagens de informal para ae=2 e ae=1, fixado ht=1.

Em todas as faixas de horas trabalhadas, a razão de vantagens é significativamente diferente de 1, ao nível α=5%, quando se passa do nível 2 para o 3 de anos de estudo. Uma análise semelhante revela que não há diferença significativa na vantagem de informal quando se passa do nível 1 para o 2 de anos de estudo.

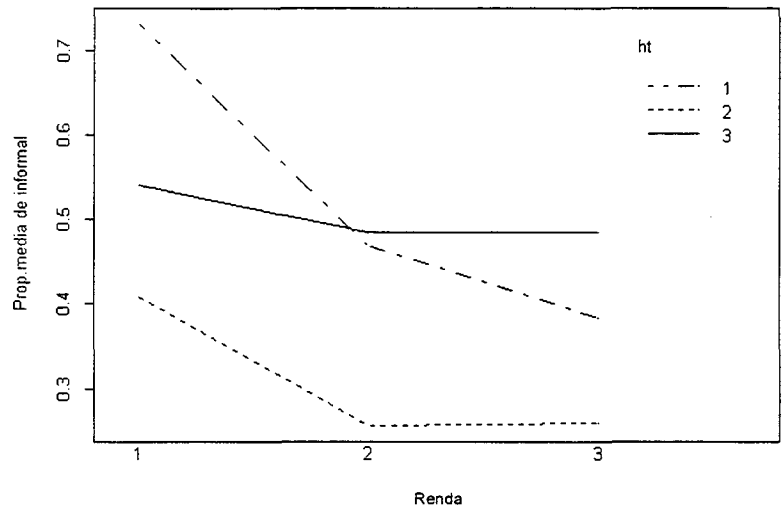
O Gráfico 2 mostra que a proporção de pessoas no mercado informal diminui quando aumenta o número de anos de estudo, qualquer que seja o número de horas semanais trabalhadas. Por outro lado, na faixa de 40-48 de horas trabalhadas, há predomínio de trabalhadores no setor formal.

Gráfico 2
Interação entre anos de estudo e horas trabalhadas



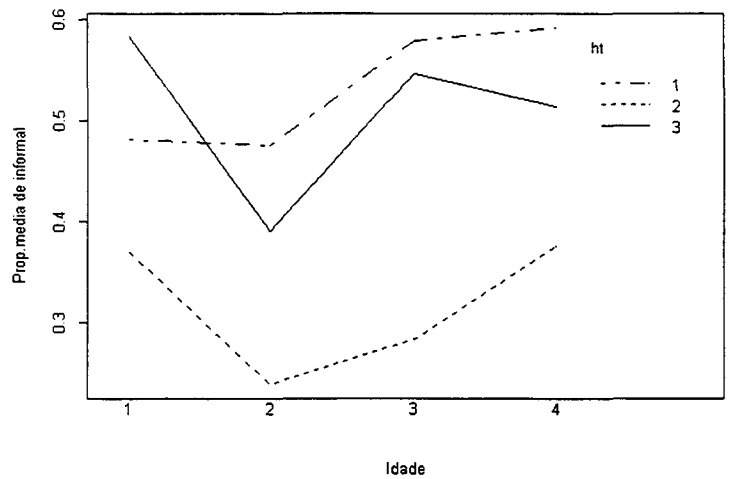
O Gráfico 3 mostra a interação entre renda e horas trabalhadas. Diminui a proporção no mercado informal quando aumenta a renda. Notam-se ainda proporções baixas no mercado informal na faixa de 40-48 horas trabalhadas e nos níveis mais altos de renda.

Gráfico 3
Interação entre renda e horas trabalhadas



O Gráfico 4 mostra a interação entre idade e horas trabalhadas. Para os três níveis de horas trabalhadas nota-se um mínimo na proporção de informal na faixa 2 de idade, entre 18 e 25 anos. Além disso, a maior proporção de pessoas no mercado informal se verifica para a faixa de idade até 17 anos e mais de 48 horas trabalhadas.

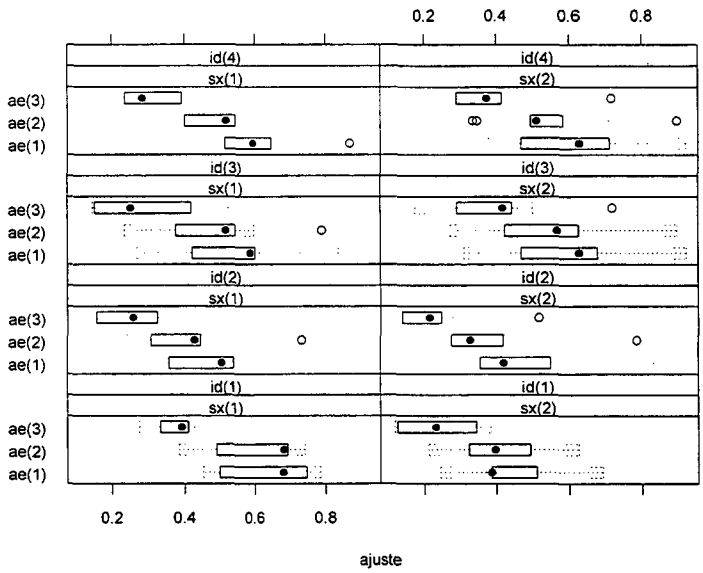
Gráfico 4
Interação entre idade e horas trabalhadas



O Gráfico 5, obtido pela utilização da função *boxplot* do S-Plus, resume os valores ajustados pelo modelo 1 através de painéis múltiplos condicionais. Cada painel mostra os *boxplots* das proporções ajustadas para as três faixas de escolaridade (até 4 anos, de 5 a 8 anos e 9 anos ou mais), para cada sexo (homens, mulheres) e para cada faixa etária (até 17 anos, de 18 a 25 anos, de 26 a 49 anos e 50 anos ou mais). No eixo vertical estão representadas as faixas de escolaridade e no eixo horizontal as proporções ajustadas.

Por exemplo, o painel superior da esquerda mostra os *boxplots* das proporções ajustadas nas três faixas de escolaridade para pessoas do sexo masculino e com idade acima de 49 anos. As variáveis do painel são proporções ajustadas e escolaridade e as variáveis condicionantes são sexo e idade. Nota-se que, para qualquer combinação de sexo e idade, a proporção no mercado informal diminui quando aumenta o número de anos de estudo. É interessante ainda observar que, tanto para homens como para mulheres da faixa até 17 anos de idade (id(1)), a proporção no mercado informal para as faixas de menos de 5 anos e entre 5 e 8 anos de estudo (ae(1) e ae(2), respectivamente) parece ser a mesma, diferentemente do que ocorre nas outras combinações de sexo e idade.

Gráfico 5
Proporção de pessoas no setor informal em função de anos de estudo, controlando por sexo e faixa etária



4. DESENHO AMOSTRAL DA PNAD

O desenho amostral da PNAD 1990 consistiu de amostragem estratificada de conglomerados com múltiplos estágios de seleção. Foi fixada a fração amostral para a região metropolitana e para a área não metropolitana de cada Unidade da Federação. Alguns municípios, em função do tamanho populacional ou por pertencerem à região metropolitana, foram considerados como estratos certos e denominados de auto-representativos. Os demais municípios foram estratificados segundo o critério de pertinência a uma mesma microrregião geográfica, porém formando estratos de mesmo tamanho em relação à população total (segundo o Censo de 1980). Estes municípios foram denominados não-auto-representativos. Em cada um desses estratos o desenho amostral consistiu em três estágios de seleção. As unidades primárias de amostragem foram os municípios, e a seleção destes dentro de cada

estrato foi feita com probabilidades proporcionais à população total de cada município no Censo de 1980.

Para a seleção no segundo estágio, tanto nos municípios auto-representativos como nos não-auto-representativos selecionados, os setores urbanos foram colocados em ordem crescente de numeração, seguidos dos setores rurais também ordenados da mesma forma. Foi atribuída a cada setor uma medida de tamanho igual ao número de domicílios particulares do setor no Censo de 1980. Foi então feita uma seleção sistemática de setores com probabilidades proporcionais a esta medida de tamanho. Em cada um dos municípios não-auto-representativos foram selecionados cinco setores e, portanto, o intervalo para seleção dos setores neste estágio foi obtido dividindo-se o total acumulado do número de domicílios no município por 5. Nos municípios auto-representativos, o intervalo para seleção de setores foi calculado pelo produto do inverso da fração amostral desejada pelo número de domicílios a selecionar por setor.

A cada ano, antes do início da pesquisa, é feita uma nova listagem dos domicílios pertencentes aos setores selecionados. Assim, no último estágio, os domicílios foram selecionados diretamente a partir desta listagem. A seleção dos domicílios foi sistemática, com partida aleatória. O intervalo de seleção em cada setor foi fixado de forma a se obter uma amostra autoponderada dentro de cada Unidade da Federação, isto é, uma amostra para a qual todos os domicílios (e conseqüentemente pessoas) tivessem a mesma probabilidade de inclusão.

Como se pode observar dessa descrição, o desenho amostral da PNAD apresenta todos os aspectos de um desenho amostral complexo, incluindo estratificação (geográfica), seleção de unidades primárias (municípios ou setores, nos municípios auto-representativos) ou secundárias (setores nos municípios não-auto-representativos) com probabilidades desiguais, conglomeração (de domicílios em setores, e de pessoas nos domicílios) e seleção sistemática sem reposição de unidades. (Mais detalhes sobre o desenho amostral da PNAD podem ser encontrados em IBGE(1981). Nesse caso, fica difícil admitir *a priori* com confiança as hipóteses usuais de modelagem das observações amostrais como IID (independente e identicamente distribuídos). Por esse motivo foram considerados métodos alternativos de modelagem e ajuste, que são discutidos no Capítulo 5 a seguir.

5. AJUSTE DE MODELOS CONSIDERANDO O DESENHO AMOSTRAL

5.1- Método de Máxima Verossimilhança

Esta seção descreve resumidamente o método de Máxima Verossimilhança -MV- comumente empregado para ajuste de modelos paramétricos, quando se ignora o desenho amostral e pesos. Essa descrição servirá de base à apresentação na próxima seção do

método de Máxima Pseudo-verossimilhança -MPV-, que foi utilizado aqui para incorporar pesos e outros aspectos do plano amostral no processo de inferência.

Seja $y_i=(y_{i1}, ..., y_{iR})'$ o vetor Rx1 das variáveis de pesquisa observadas para o elemento *i*, gerado por um vetor aleatório Y_i, para i = 1, ..., n, onde n é o tamanho da amostra selecionada. Suponha que os vetores aleatórios Y_i, para i = 1, ..., n, são independentes e identicamente distribuídos com distribuição comum $f(y;\theta)$, onde $\theta=(\theta_1,...,\theta_K)'$ é o vetor Kx1 de parâmetros desconhecidos de interesse. Sob essas hipóteses, a verossimilhança amostral é dada por $l(\theta)=\prod_{i=1}^{n} f(y_i;\theta)$ e a correspondente log-verossimilhança por $L(\theta)=\sum_{i=1}^{n}\log[f(y_i;\theta)]$.

Sob condições de regularidade, igualando-se as derivadas parciais de $L(\theta)$ com relação a cada componente de θ a 0, temos um sistema de equações $\sum_{i=1}^{n} u_i(\theta)=0$, onde $u_i(\theta)=\partial\log[f(y_i;\theta)]/\partial\theta$ é o vetor Kx1 dos escores do elemento i. A solução $\hat{\theta}$ deste sistema é o estimador de Máxima Verossimilhança de θ. A variância assintótica do estimador $\hat{\theta}$ é dada por $V(\hat{\theta})=[J(\theta)]^{-1}$ e um estimador consistente dessa variância é dado por $\hat{V}(\hat{\theta})=[J(\hat{\theta})]^{-1}$, onde $J(\theta)=\sum_{i=1}^{n}\partial u_i(\theta)/\partial\theta$ e $J(\hat{\theta})=J(\theta)|_{\hat{\theta}=\theta}$.

Para uma discussão mais detalhada do método de Máxima Verossimilhança para ajuste de modelos paramétricos regulares, veja, por exemplo, Garthwaite, Jollife e Jones(1995).

5.2- Método de Máxima Pseudo-verossimilhança

O material apresentado nesta seção se baseia em Nascimento Silva(1996, Capítulo 6).

Suponha agora que os vetores observados y_i das variáveis de pesquisa do elemento i são gerados por vetores aleatórios Y_i, para $i\in U$, onde $U=\{1,\cdots,N\}$ é o conjunto de rótulos dos elementos da população. Suponha também que $Y_1,..., Y_N$ são IID com densidade $f(y;\theta)$. Se todos os elementos da população finita U fossem conhecidos, as funções de verossimilhança e de log-verossimilhança populacionais seriam dadas por:

$$\ell_U(\theta)=\prod_{i\in U} f(y_i;\theta)$$

$$L_U(\theta)=\sum_{i\in U}\log[f(y_i;\theta)].$$

Sob condições de regularidade, igualando-se as derivadas parciais de $L_U(\theta)$ com relação a cada componente de θ a 0, temos um sistema de equações $\sum_{i\in U} u_i(\theta)=0$, onde $u_i(\theta)=\partial\log[f(y_i;\theta)]/\partial\theta$ é o vetor Kx1 dos escores do elemento i, $i\in U$. A solução θ_U deste sistema é o estimador de Máxima Verossimilhança de θ no caso de um censo. Podemos

considerar θ_U como uma quantidade desconhecida da população finita, sobre a qual se deseja fazer inferências baseadas em informações da amostra. Sob certas condições de regularidade, $\theta_U - \theta = o_p(1)$.

Seja $T(\theta) = \sum_{i \in U} u_i(\theta)$ a soma dos escores, que é um vetor de totais populacionais. Para estimar este vetor de totais, pode-se usar um estimador linear ponderado da forma $\hat{T}(\theta) = \sum_{i \in s} w_i u_i(\theta)$, onde w_i são pesos propriamente definidos. O estimador de Máxima Pseudo-verossimilhança $\hat{\theta}_{MPV}$ será a solução das equações de Máxima Pseudo-verossimilhança dadas por

$$\hat{T}(\theta) = \sum_{i \in s} w_i u_i(\theta) = 0 . \tag{2}$$

Através da linearização de Taylor podemos obter a variância assintótica, sob o plano amostral, do estimador $\hat{\theta}_{MPV}$ e seu estimador correspondente, dados respectivamente por:

$$V_p(\hat{\theta}_{MPV}) = [J(\theta_U)]^{-1} V_p\left[\sum_{i \in s} w_i u_i(\theta_U)\right][J(\theta_U)]^{-1} \tag{3}$$

$$\hat{V}(\hat{\theta}_{MPV}) = [\hat{J}(\hat{\theta}_{MPV})]^{-1} \hat{V}\left[\sum_{i \in s} w_i u_i(\hat{\theta}_{MPV})\right][\hat{J}(\hat{\theta}_{MPV})]^{-1}, \text{ onde}$$

$$J(\theta_U) = \left.\frac{\partial T(\theta)}{\partial \theta}\right|_{\theta=\theta_U} = \sum_{i \in U} \left.\frac{\partial u_i(\theta)}{\partial \theta}\right|_{\theta=\theta_U},$$

$$\hat{J}(\hat{\theta}_{MPV}) = \left.\frac{\partial \hat{T}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{MPV}} = \sum_{i \in s} w_i \left.\frac{\partial u_i(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{MPV}},$$

$V_p\left[\sum_{i \in s} w_i u_i(\theta_U)\right]$ é a matriz de variância (do desenho) do estimador do total populacional dos escores e $\hat{V}\left[\sum_{i \in s} w_i u_i(\hat{\theta}_{MPV})\right]$ é um estimador consistente para esta variância.

Muitos modelos paramétricos, com muitos planos amostrais e estimadores de totais diferentes, podem ser ajustados, resolvendo-se as equações de Máxima Pseudo-verossimilhança, satisfeitas algumas condições de regularidade. Os estimadores de MPV não serão únicos, já que existem diversas maneiras de se definir os pesos w_i. Os pesos mais usados são os do estimador de Horwitz-Thompson para totais, dados pelo inverso da probabilidade de inclusão do indivíduo *i*, ou seja, $w_i = \pi_i^{-1}$, para $i \in s$.

Assim, um procedimento padrão para ajustar um modelo paramétrico regular $f(y;\theta)$ pelo método da Máxima Pseudo-verossimilhança seria dado pelos passos abaixo:

- Resolver $\sum_{i\in s} \pi_i^{-1} u_i(\theta) = 0$ e calcular o estimador pontual $\hat{\theta}_\pi$ do parâmetro θ no modelo $f(y;\theta)$.

- Calcular a matriz de variância estimada $\hat{V}(\hat{\theta}_\pi) = [\hat{J}(\hat{\theta}_\pi)]^{-1} \hat{V}_\pi \left[\sum_{i\in s} \pi_i^{-1} u_i(\hat{\theta}_\pi) \right] [\hat{J}(\hat{\theta}_\pi)]^{-1}$, onde

$$\hat{V}_\pi \left[\sum_{i\in s} \pi_i^{-1} u_i(\hat{\theta}_\pi) \right] = \sum_{i\in s} \sum_{j\in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} [u_i(\hat{\theta}_\pi)][u_j(\hat{\theta}_\pi)]' \quad \text{e} \quad [\hat{J}(\hat{\theta}_\pi)] = \left. \frac{\partial \hat{T}(\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}_\pi} = \sum_{i\in s} \pi_i^{-1} \left. \frac{\partial u_i(\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}_\pi}$$

- Usar $\hat{\theta}_\pi$ e $\hat{V}(\hat{\theta}_\pi)$ para calcular intervalos de confiança ou estatísticas de teste baseadas na distribuição normal e utilizá-los para fazer inferência sobre os componentes de θ.

Sob probabilidades iguais de seleção , os pesos π_i^{-1} serão constantes e o estimador pontual $\hat{\theta}_\pi$ será idêntico ao estimador de Máxima Verossimilhança ordinário em uma amostra de observações IID com distribuição $f(y;\theta)$, mas o mesmo não é verdade quando se trata da variância do estimador.

Vantagens de se usar o procedimento de MPV

1 - O procedimento proporciona estimativas "baseadas no desenho" para a variância dos estimadores dos parâmetros, as quais são razoavelmente simples de calcular e são consistentes sob "condições fracas" no desenho amostral e na especificação do modelo. Mesmo quando o estimador pontual coincide com o estimador usual de Máxima Verossimilhança, a estimativa da variância obtida pelo procedimento de MPV pode ser preferível aos estimadores usuais da variância baseados no modelo, que ignoram o desenho amostral.

2 - O procedimento fornece estimativas "robustas", no sentido de que em muitos casos a quantidade θ_U da população finita permanece um "alvo" válido para inferência, mesmo quando o modelo especificado por $f(y;\theta)$ não proporcione uma descrição adequada para a distribuição das variáveis de pesquisa.

Este procedimento requer conhecimento de informações detalhadas sobre os elementos da amostra, tais como pertinência a estratos e conglomerados ou unidades primárias de amostragem, e suas probabilidades de inclusão.

Desvantagens do método de MPV

1 - As propriedades para pequenas amostras não são conhecidas. Este pode não ser um problema muito grande em análises que usam os dados de pesquisas feitas pelas agências oficiais de estatística, desde que em tais análises seja utilizada a amostra inteira, ou, no caso de subdomínios estudados separadamente, que sejam usadas amostras suficientemente grandes.

2 - Não podem ser utilizados métodos usuais de diagnóstico e outros procedimentos da inferência clássica, tais como gráficos de resíduos e testes estatísticos de Razões de Verossimilhança.

5.3- Modelo de regressão logística

No modelo de regressão logística, a equação de verossimilhança populacional é dada por:

$$\sum_{i\in U}[y_i - p(x'_i\beta)]\ x_i = 0, \qquad (4)$$

onde x_i é o vetor de variáveis explicativas da i-ésima observação e

$$p(x_i'\beta) = P(Y_i = 1 \mid X_i = x_i) = \exp(x_i'\beta)/[1+\exp(x_i'\beta)].$$

O estimador de MPV do vetor de coeficientes β no modelo (4) é obtido como solução da equação de pseudo-verossimilhança:

$$\sum_{i\in s} w_i[y_i - p(x_i'\beta)]\ x_i = 0, \qquad (5)$$

onde w_i é o peso da i-ésima observação.

A matriz de covariância do estimador de MPV de β pode ser obtida conforme indicado na Seção 5.2., bastando notar que $u_i(\beta) = [y_i - p(x_i'\beta)]x_i$. Para maiores detalhes, o leitor interessado pode consultar Binder(1983), que aborda o problema da estimação da matriz de covariância dos estimadores de MPV na família de modelos lineares generalizados, que inclui o modelo de regressão logística.

5.4- Correção do χ^2

Na análise de dados categóricos a estatística X^2 de Pearson tem distribuição assintótica χ^2, quando supomos amostragem multinomial. Quando os dados não são obtidos por amostragem aleatória simples com reposição, $X^2 \sim \sum_k \lambda_k X_k$ onde os λ_k's são autovalores da matriz de efeitos do desenho, denominados efeitos de desenho generalizados, e os X_k são variáveis aleatórias IID com distribuição χ^2 com 1 grau de liberdade. Rao e Scott(1981) utilizaram uma correção de Satterthwaite para a estatística X^2 de Pearson dividindo-a por $\bar{\lambda}$, média dos autovalores. Para testar hipóteses de nulidade dos efeitos e interações do modelo de regressão logística, a PROC LOGISTIC do pacote SUDAAN utiliza as estatísticas F de Wald e a estatística F com ajuste de Satterthwaite, descritas na página 21 de Shah et alii(1993).

6. COMPARAÇÃO DE AJUSTES

Nesta seção vamos apresentar o ajuste obtido pela PROC LOGISTIC do pacote SUDAAN do modelo selecionado na Seção 3. A análise na Seção 3 foi feita utilizando-se o pacote estatístico S-Plus, e nela não foram considerados nem os pesos das unidades nem o desenho amostral. É um exemplo de uma análise que, tipicamente, seria feita por usuário de uma agência produtora dos dados estatísticos. A comparação entre as análises contidas nesta seção e as da Seção 3 permitirá avaliar o impacto dos pesos e do desenho amostral sobre os resultados.

6.1- Ajuste do SUDAAN

Apresentamos na Tabela 6 as estimativas dos efeitos principais e interações do modelo selecionado e seus respectivos desvios-padrões, calculadas pela PROC LOGISTIC do pacote SUDAAN. Para facilitar a comparação incluímos na tabela os valores correspondentes estimados pelo S-Plus.

As estimativas calculadas pelo pacote SUDAAN são feitas pelo método de Máxima Pseudo-verossimilhança, resolvendo a equação (5) da Seção 5.3. As estimativas dos desvios-padrões são obtidas pelo método de linearização descrito em 5.2, na equação (3), considerando os escores tal como apresentados na Seção 5.3. Para esses cálculos, os estimadores de variância considerados levaram em conta os pesos das observações, mas utilizaram uma aproximação que consiste em considerar que as unidades primárias de amostragem foram selecionadas com reposição, especificando a opção WR do pacote SUDAAN. [Veja Shah et alii(1993, p. 4) e Wolter(1985, eq. 7.7.2)].

Tabela 6
Estimativas dos coeficientes e dos respectivos desvios-padrões, obtidas pelo S-Plus e pelo SUDAAN

Efeitos	Ajuste no SUDAAN		Ajuste no S-Plus	
	Coeficiente	Desvio-Padrão	Coeficiente	Desvio-Padrão
Intercepto	-0,515	0,260	-0,514	0,269
sx	0,148	0,222	0,156	0,228
ae1	0,745	0,165	0,740	0,165
ae2	0,496	0,156	0,497	0,159
ht1	-0,377	0,317	-0,386	0,312
ht2	-0,697	0,275	-0,698	0,268
id1	-0,239	0,540	-0,243	0,492
id2	-0,729	0,302	-0,724	0,314
id3	0,227	0,231	0,227	0,234
re1	0,286	0,277	0,293	0,245
re2	0,065	0,144	0,062	0,145
ht1.re1	1,529	0,356	1,531	0,332
ht2.re1	0,338	0,320	0,336	0,284
ht1.re2	0,490	0,233	0,498	0,221
ht2.re2	-0,115	0,183	-0,112	0,178
ht1.id1	-1,420	0,605	-1,408	0,515
ht2.id1	-0,413	0,506	-0,397	0,465
ht1.id2	-0,124	0,354	-0,129	0,351
ht2.id2	-0,109	0,279	-0,106	0,286
ht1.id3	-0,220	0,248	-0,216	0,253
ht2.id3	-0,537	0,205	-0,533	0,201
sx.id1	0,878	0,348	0,870	0,335
sx.id2	0,300	0,231	0,294	0,226
sx.id3	-0,259	0,190	-0,263	0,186
sx.ht1	-0,736	0,206	-0,737	0,211
sx.ht2	-0,089	0,185	-0,093	0,182

Na Tabela 7 são apresentadas as probabilidades de significância dos testes de nulidade dos efeitos do modelo. Todos os efeitos incluídos no modelo são significativos, nos níveis usuais de significância. A PROC LOGISTIC do pacote SUDAAN não inclui testes para os efeitos principais, por não ser possível separar tais efeitos das interações. Os resultados da Tabela 7 não são diretamente comparáveis aos da Tabela 1 de análise de desvios obtida pela função glm do S-Plus. A coluna de p-valores da Tabela 7, obtida pela PROC LOGISTIC do pacote SUDAAN, utiliza a estatística de Wald, corrigida pelo efeito do plano amostral,

considerando-se os autovalores da matriz de deficiência para o modelo. Maiores detalhes são encontrados em Shah et alii(1993).

Tabela 7
Testes de hipóteses de nulidade dos efeitos do modelo

Contraste	Graus de liberdade	Graus de liberdade ajustados	Estatística F ajustada	P-valor da estatística F ajustada
Modelo global	30	26,132	37,510	0,000
Bondade do ajuste	29	25,692	28,179	0,000
ht:re	4	3,946	6,040	0,000
ht:id	6	5,764	4,110	0,001
sx:id	3	2,969	7,168	0,000
sx:ht	2	1,993	9,166	0,000
ae:ht	4	3,959	4,814	0,001

Os testes da Tabela 7 indicam a significância de todas as interações de dois fatores que entraram no modelo selecionado na Seção 2. O teste de qualidade global de ajuste, na primeira linha da Tabela 7, indica a necessidade de serem introduzidas novas interações. É importante ressaltar que na análise feita pela PROC LOGISTIC do pacote SUDAAN foram utilizados dados individuais (resposta 0-1), enquanto que na análise feita pela função glm do S-Plus foram usados os dados tabelados, conforme explicado na Seção 3.

Para comparação, incluímos na Tabela 8 alguns intervalos de confiança calculados tanto pela função glm do S-Plus como pela PROC LOGISTIC do pacote SUDAAN. Na construção destes intervalos foi necessário utilizar estimativas pontuais dos efeitos, bem como a matriz de covariância estimada dos estimadores dos efeitos do modelo. Deste modo, as discrepâncias verificadas entre os intervalos apresentados nessa tabela sumarizam, ao mesmo tempo, tanto as diferenças nas estimativas pontuais dos efeitos como nas variâncias e covariâncias das estimativas.

Tabela 8
Intervalos de confiança de 95% para razões de vantagens, variando-se os níveis de id para níveis fixos de ht e sx

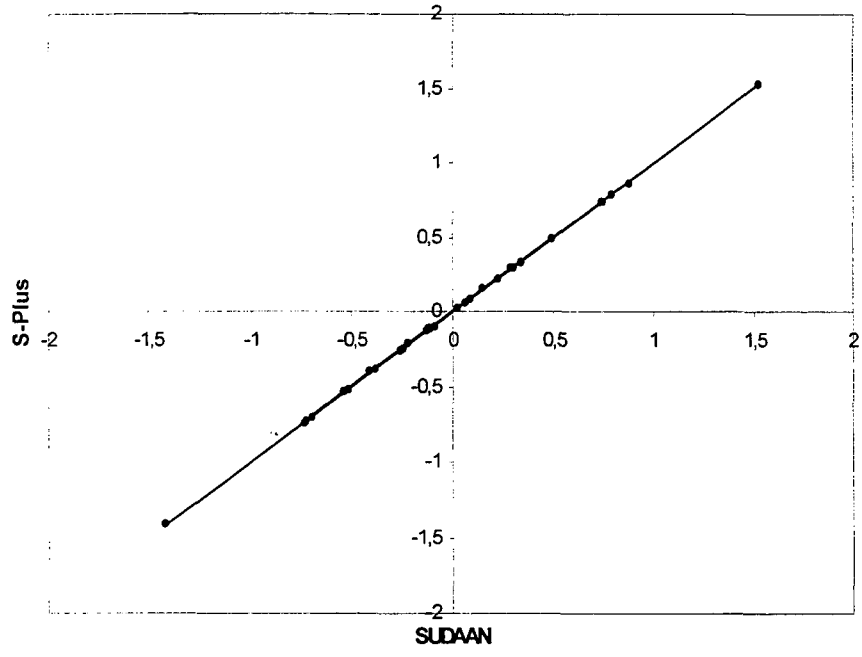
ht	sx	Mudança de nível de id	Razão de vantagens estimada		Intervalo de 95% para razão de vantagens	
			S-Plus	SUDAAN	S-Plus	SUDAAN
1	1	1 para 2	1,248	1,256	(0,678;2,298)	(0,659;2,393)
1	1	2 para 3	1,359	1,351	(0,894;2,064)	(0,907;2,014)
1	1	3 para 4	1,287	1,287	(0,873;1,898)	(0,863;1,918)
1	2	1 para 2	2,222	2,239	(1,090;4,526)	(1,012;4,953)
1	2	2 para 3	2,370	2,363	(1,583;3,550)	(1,615;3,458)
1	2	3 para 4	0,990	0,993	(0,674;1,453)	(0,667;1,479)
2	1	1 para 2	0,465	0,466	(0,318;0,680)	(0,316;0,686)
2	1	2 para 3	0,967	0,969	(0,772;1,210)	(0,758;1,240)
2	1	3 para 4	1,768	1,766	(1,413;2,212)	(1,382;2,258)
2	2	1 para 2	0,828	0,830	(0,443;1,547)	(0,428;1,609)
2	2	2 para 3	1,686	1,696	(1,234;2,305)	(1,224;2,348)
2	2	3 para 4	1,359	1,363	(0,958;1,929)	(0,947;1,962)
3	1	1 para 2	0,348	0,344	(0,147;0,823)	(0,133;0,886)
3	1	2 para 3	1,483	1,487	(0,977;2,251)	(0,984;2,247)
3	1	3 para 4	1,037	1,033	(0,741;1,450)	(0,742;1,437)
3	2	1 para 2	0,618	0,613	(0,236;1,622)	(0,224;1,672)
3	2	2 para 3	2,587	2,601	(1,591;4,205)	(1,661;4,073)
3	2	3 para 4	0,797	0,797	(0,504;1,260)	(0,506;1,254)

Além dos ajustes comparados acima, foram feitos os seguintes ajustes com a utilização do S-Plus: 1) dados individuais (resposta 0-1) considerando os pesos; 2) dados da tabela estimada considerando os pesos; e 3) dados individuais com pesos normalizados. Em todas estas análises, como esperado, as estimativas pontuais dos efeitos coincidiram com as obtidas pela PROC LOGISTIC do pacote SUDAAN.

6.2- Impacto de pesos diferentes nas estimativas pontuais

O Gráfico 6 apresenta uma comparação entre as estimativas pontuais dos parâmetros do modelo (1) da Seção 3, calculadas usando-se a função glm do S-Plus e pela PROC LOGISTIC do pacote SUDAAN, apresentadas na Tabela 6. Pode-se notar que, neste caso, há estreita concordância entre as estimativas pontuais obtidas pelos dois pacotes, já que todos os pontos do gráfico estão praticamente sobre a bissetriz, apesar de o S-Plus não considerar nem os pesos nem o desenho amostral.

Gráfico 6
Comparação entre as estimativas pontuais dos parâmetros obtidas pelo S-Plus e pelo SUDAAN



A concordância das estimativas pontuais dos coeficientes pode ser explicada, em grande parte, pela pequena variabilidade dos pesos das unidades amostrais. Como se pode observar na Tabela 9, que apresenta a distribuição das freqüências dos pesos, há essencialmente dois pesos distintos, um para moradores da região metropolitana e outro para moradores do interior (os quatro valores distintos são resultado de arredondamento diferenciado efetuado para garantir calibração com o total da população projetada correspondente). A variação dos pesos seria bem maior caso fossem consideradas as amostras da PNAD nos outros estados da federação, o que não foi feito para o presente exercício, que considerou somente o Estado do Rio de Janeiro.

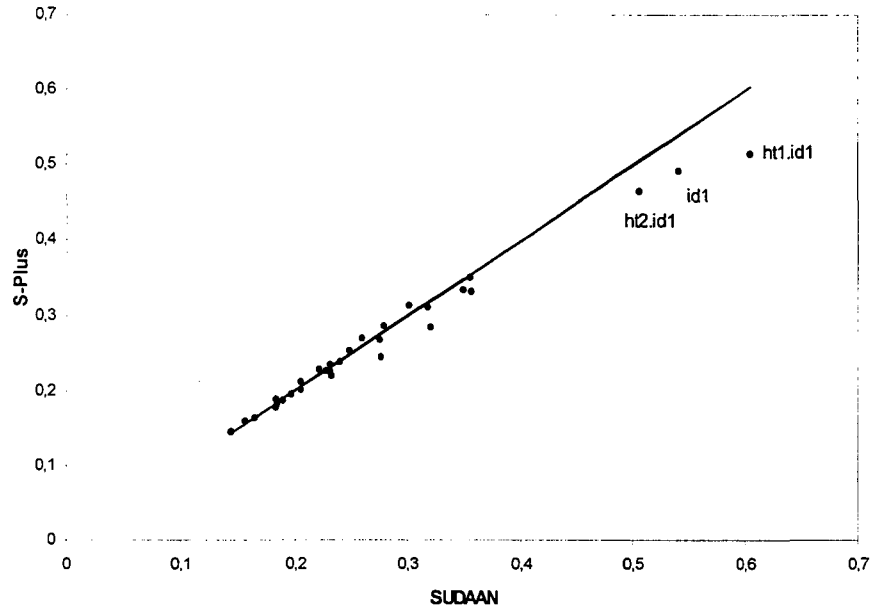
Tabela 9
Distribuição de freqüências dos pesos das pessoas na amostra

Valor do peso	Freqüência
674	127
675	784
711	3288
712	2308

6.3- Impacto do desenho nas estimativas de precisão

O Gráfico 7 apresenta comparação entre os desvios-padrões estimados das estimativas pontuais dos parâmetros calculados usando-se a função glm do S-Plus (não considerando pesos e desenho) e a PROC LOGISTIC do SUDAAN (considerando pesos e desenho amostral). Pode-se observar que, neste caso, há um maior distanciamento dos pontos em relação à bissetriz. As maiores diferenças entre os dois métodos ocorrem na estimação dos desvios-padrões das estimativas dos parâmetros do primeiro nível de idade (até 17 anos) e da interação deste com horas trabalhadas (tanto no nível de menos de 40 horas semanais como no nível de 40 a 48 horas semanais trabalhadas).

Gráfico 7
Comparação entre os desvios-padrões estimados das estimativas pontuais dos parâmetros obtidos pelo S-Plus e pelo SUDAAN



Esta diferenciação maior no caso dos desvios-padrões já era esperada. Quando não levamos em conta os pesos nem o desenho amostral na estimação dos parâmetros, podemos até chegar em uma estimativa pontual dos coeficientes bem próxima de quando levamos ambos em conta, mas as estimativas dos desvios-padrões são mais sensíveis a esta diferença entre as análises. A tendência revelada pelo Gráfico 7 é de subestimação dos desvios-padrões pelo S-Plus ao ignorar o desenho e a variação dos pesos.

7- CONSIDERAÇÕES FINAIS

Na seleção do modelo no exercício acima utilizou-se a função glm do S-Plus. Feita a seleção, o mesmo modelo foi ajustado através da PROC LOGISTIC do SUDAAN. O propósito foi imitar uma situação onde o modelo já tivesse sido selecionado e ajustado por usuário da

agência produtora dos dados, sem considerar os pesos e o desenho amostral, tal como é usual. Outra possibilidade seria repetir o processo de seleção usando-se a PROC LOGISTIC do SUDAAN. Isto poderia ser feito passo a passo, incluindo efeitos e interações que melhorassem mais a qualidade de ajuste, tal como foi feito automaticamente pela função *step* do S-Plus. Este procedimento possibilitaria comparar a seleção de modelos quando são considerados os pesos e o desenho amostral na análise. Considerando-se a maior dificuldade de seleção de modelos através do SUDAAN, preferiu-se usá-lo aqui apenas para ajustar um modelo já selecionado. Diferentemente dos pacotes mais usados de análise estatística, tais como SAS, S-Plus, BMDP, etc., o SUDDAN não possui ferramentas usuais de diagnóstico de ajuste de modelos, como gráficos de resíduos padronizados, etc., tornando mais difícil seu uso para a etapa de seleção de modelos.

Por outro lado, ao construir e ajustar modelos a partir de dados produzidos pelo IBGE, o usuário precisa incorporar as informações sobre os pesos e o desenho amostral utilizado. Em geral os pesos são considerados na produção dos resultados publicados da pesquisa, sendo possível, mesmo utilizando os pacotes tradicionais, produzir estimativas pontuais "corretas". Porém, para construir intervalos de confiança e testar hipóteses sobre os parâmetros do modelo, seria preciso o conhecimento das estimativas de variâncias e covariâncias das estimativas obtidas a partir do desenho amostral utilizado. Mesmo contando com esta informação não é simples incorporá-la à análise nos pacotes usuais. Exceto em casos simples, incorporar pesos e desenho na análise não é uma tarefa simples sem o uso de pacotes especializados. Isto porque estes pacotes utilizam metodologia geral, tal como método de Máxima Pseudo-verossimilhança e linearização para estimação de matrizes de covariância.

Para concluir, apesar de não pretendermos examinar aqui os resultados substantivos da inserção no mercado informal de trabalho, impressionou-nos detectar, com base no Gráfico 4, que a maior proporção de pessoas no mercado informal é de pessoas jovens (idade até 17 anos) que trabalham mais de 48 horas semanais. Se a alta taxa de inserção no mercado informal para essa faixa etária era esperada, a combinação desta com elevado número de horas trabalhadas por semana surpreende, revelando um aspecto negativo da inserção de jovens no mercado de trabalho.

REFERÊNCIAS

Agresti, A. (1990). *Categorical Data Analysis.* Nova Iorque: John Wiley & Sons.

Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 51, 279-292.

Cordeiro, G. (1986). *Modelos Lineares Generalizados.* 7° SINAPE, Campinas, S.P.

Garthwaite, P.H.; Jollife, I.T. e Jones, B. (1995). *Statistical Inference.* Londres: Prentice Hall.

Leote, R.M.D. (1996). *Um perfil sócio-econômico das pessoas ocupadas no setor informal na área urbana do Rio de Janeiro.* Relatórios Técnicos n° 02/96, Escola Nacional de Ciências Estatísticas.

IBGE (1981). *Metodologia da Pesquisa Nacional por Amostra de Domicílios na Década de 70.* Rio de Janeiro: IBGE, Série Relatórios Metodológicos, volume 1.

Nascimento Silva, P.L. do (1996). *Utilizing Auxiliary Information for Estimation and Analysis in Sample Surveys.* Tese de Doutorado, Universidade de Southampton.

Pfeffermann, D. (1993). The Role of Sampling Weights when Modeling Survey Data. *International Statistical Review*, 61, 317-337.

Rao, J.N.K. e Scott, A.J. (1981). The Analysis of Categorical Data from Complex Surveys: Chi- Square Tests of Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association,* 76, p.221-230.

Shah, B.V. et al (1993) *Statistical Methods an Mathematical Algorithms used in SUDAAN.* Research Triangle Institute, Relatório Técnico.

Skinner, C.J.; Holt, D. e Smith, T.M.F. eds (1989). *Analysis of Complex Surveys.* Chichester: John Wiley & Sons.

Venables, W.N. e Ripley, B.D. (1994). *Modern Applied Statistics with S-Plus.* Nova Iorque: Springer-Verlag.

Wolter, K. M. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.

ABSTRACT

When performing descriptive analysis of estimates of finite population quantities of interest, such as means, totals, ratios and proportions, statistical agencies consider the weights and sample design adopted to obtain the data. However, users of such data outside the statistical agencies frequently carry out modelling and analysis without considering the weights, and principally the sample design. Statistical software usually assumes that the observations are IID (independent and identically distributed), that is, that they were generated by simple random sampling with replacement. Sample surveys, in general, use more complex designs, and this can affect the estimates, particularly of variances of point estimates. In this report we study the effect of ignoring sample weights and design in the analysis carried out by Leote(1996), who used data from the supplement on labour of the 1990 PNAD (a national household sample survey) for the State of Rio de Janeiro. She fitted logistic models to data on participation of the informal economy using standard statistical software. Here a similar model is fitted using software that incorporates both the complex survey design and its weights, and results of both modelling exercises are compared. The impact of ignoring important aspects of sample design, such as stratification, clustering and unequal weights is discussed. The difficulties faced by external users of sample survey data (such as those from PNAD) for modelling and analysis are also highlighted, because they often ignore such important features of the design.

Análise Bayesiana para Modelos Não Lineares de Crescimento

Josmar Mazucheli*

Jorge Alberto Achcar**

RESUMO

Considerando alguns modelos não lineares de crescimento, cujo comportamento é sigmóide (Logístico, Gompertz, Tipo-Weibull, Morgan-Mercer-Flodin e Richards), apresentamos uma análise Bayesiana usando a priori não informativa para os parâmetros e o método de Laplace para aproximação de integrais. Também apresentamos algumas técnicas Bayesianas de discriminação de modelos e ilustramos com um conjunto de dados.

Palavras-Chaves: Modelos de Crescimento, Análise Bayesiana e Método de Laplace.

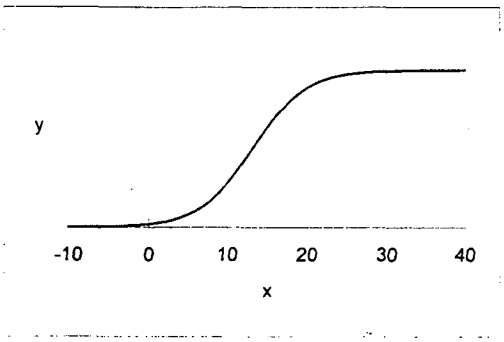
1. INTRODUÇÃO

A análise de dados gerados a partir de fenômenos de crescimento (dados de crescimento) é uma tarefa muito comum em diversas áreas de investigação científica. Por exemplo, biólogos podem ter interesse em descrever o crescimento de organismos e entender seus mecanismos fundamentais de crescimento; químicos podem ter interesse em avaliar a produção de uma reação química ao longo do tempo; em agronomia existem interesses óbvios em conhecer como as plantas crescem e a velocidade que crescem; o comportamento do crescimento de tumores submetidos a um determinado tratamento é de grande interesse em medicina; cientistas sociais podem ter interesse em avaliar o crescimento populacional ao longo de um período de tempo. Nestes exemplos, podemos observar que a variável resposta basicamente depende do tempo, porém são comuns situações onde isso não ocorre. Um exemplo é o caso em que a variável resposta é o ganho de peso, o qual, possivelmente, depende da quantidade de nutrientes presentes em uma determinada dieta.

Em geral, em aplicações, temos interesse em modelar uma variável resposta que aumenta com o tempo e quando os dados são colocados num gráfico observamos uma forma em "S" (Figura 1) com certa tendência a estabilização.

* Endereço para correspondência: Univ. Estadual de Maringá - Maringá, PR. Av. Colombo, 5790. CEP. 87020-900 - Deptº. de Estatística, Bl. F67.
** Universidade de São Paulo - ICMSC-SP, CP. 668 - 13560-970 - São Carlos - SP.

Figura 1 - Gráfico da variável resposta y versus uma variável independente x



Para modelar esse tipo de comportamento, a literatura apresenta vários modelos de crescimento sigmóide (por exemplo Seber e Wild, 1989 ou Ratkowsky, 1990), dos quais destacamos os modelos: Logístico, Gompertz, Tipo-Weibull, Morgan-Mercer-Flodin e o modelo Richards.

$$
\begin{aligned}
y_i &= \alpha[1+\exp(\beta-\gamma x_i)]^{-1} && +\varepsilon_i && \text{(Logístico)} \\
y_i &= \alpha\exp[-\exp(\beta-\gamma x_i)] && +\varepsilon_i && \text{(Gompertz)} \\
y_i &= \alpha-\beta\exp\left(-\gamma x_i^{\delta}\right) && +\varepsilon_i && \text{(Tipo Weibull)} \\
y_i &= \left[\beta\gamma+\alpha\, x_i^{\delta}\right]\left[\gamma+x^{\delta}\right]^{-1} && +\varepsilon_i && \text{(M.M.F.)} \\
y_i &= \alpha[1+\exp(\beta-\gamma x_i)]^{-\frac{1}{\delta}} && +\varepsilon_i && \text{(Richards)}
\end{aligned}
\qquad (1)
$$

onde ε_i, (i=1,2,...,n), são considerados variáveis aleatórias independentes e identicamente distribuídas, normais com média zero e variância constante σ^2, isto é $\varepsilon_i \sim N(0,\sigma^2)$.

Para as inferências de interesse nessa classe de modelos não lineares, em geral, é considerado um processo iterativo, por exemplo o método de Gauss-Newton, para encontrar os estimadores de mínimos quadrados e a construção de testes de hipóteses ou intervalos de confiança são baseados em resultados aproximados. Para avaliar a precisão desses resultados, usualmente o estatístico utiliza medidas para avaliar a extensão do comportamento não linear, dentre elas podemos citar: medidas de curvatura de Bates e Watts (Bates e Watts, 1980), medida de vício de Box (Box, 1971), medidas de assimetria (Lowry e Morton, 1983) ou alguns estudos de simulação (Ratkowsky, 1983).

Nas aplicações de modelos não lineares, também é comum o uso do método de máxima verossimilhança para a estimação dos parâmetros do modelo e resultados assintóticos usuais; para os estimadores de máxima verossimilhança são usados na construção de testes de hipóteses e intervalos de confiança.

Quando os resultados inferenciais obtidos não são muitos precisos, usualmente o estatístico considera diferentes parametrizações que podem melhorar a precisão dos resultados obtidos (Bates e Watts, 1988). Apesar disso, muitas vezes o pesquisador não quer

formas transformadas dos parâmetros originais, pois, em geral, esses parâmetros possuem interpretações físicas relacionadas ao experimento.

Assim, uma análise Bayesiana dos modelos não lineares, dados em (1), apresentada a seguir, pode ser de grande interesse prático.

Na Seção 3, é apresentado o método de Laplace para aproximações de integrais Bayesianas, na Seção 4, são apresentadas algumas estratégias Bayesianas de discriminação, pois, em geral, muitos modelos não lineares podem ser propostos para a análise de um mesmo fenômeno. Na Seção 5, um exemplo é apresentado considerando um conjunto de dados cujo comportamento é sigmóide, e na Seção 6, são apresentadas algumas conclusões.

2. UMA ANÁLISE BAYESIANA PARA OS MODELOS DE CRESCIMENTO

Sejam (x_i, y_i), (i=1,2,...,n), pares de observações cuja relação funcional é um modelo de regressão não linear, usualmente escrito na forma:

$$y_i = f(x_i;\theta) + \varepsilon_i, \tag{2}$$

onde y_i é a variável resposta, x_i é uma variável independente, $\theta^t = (\theta_1,...,\theta_p)$ é um vetor de parâmetros e ε_i é o erro aleatório. Além disso, em geral, supomos que ε_i e (i=1,2,...,n) são variáveis aleatórias independentes com distribuição normal, $\varepsilon_i \sim N(0,\sigma^2)$.

A matriz de informação de Fisher para θ e σ^2 (Seber e Wild, 1989) é dada por:

$$I(\theta,\sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{F}^t(\theta)\mathbf{F}(\theta) & 0 \\ 0^t & \frac{n}{2\sigma^4} \end{bmatrix}, \tag{3}$$

onde $\mathbf{F}(\theta)$ é uma matriz de ordem $(n_x n)$ dada por: $\mathbf{F}(\theta) = [\partial f(x_i;\theta)/\partial\theta_j]$ (i=1,2,...,n); (j=1,2,...,p).

Uma densidade a priori não informativa para os parâmetros θ e σ^2 (Seber e Wild, 1989) usando a regra de Jeffreys e assumindo independência a priori entre θ e σ^2 é dada por:

$$\pi(\theta,\sigma^2) \propto \{\det[\mathbf{F}^t(\theta)\mathbf{F}(\theta)]\}^{1/2}\sigma^{-1}. \tag{4}$$

Considerando a densidade a priori não informativa (4), a densidade a posteriori conjunta para θ e σ^2 é dada por:

$$\pi(\theta,\sigma^2/D) \propto \left|F^t(\theta)F(\theta)\right|^{\frac{1}{2}}\sigma^{-(n+1)}\exp\left[-\frac{1}{2\sigma^2}S(\theta)\right], \tag{5}$$

onde $S(\theta) = \sum_{i=1}^{n}[y_i - f(x_i,\theta)]^2$, e D representa um conjunto de dados observados.

Vale ressaltar que na classe de modelos considerada, modelos não lineares de crescimento sigmóide, além a priori não informativa de Jeffreys dada em (4), densidades a priori informativas podem ser sugeridas uma vez que os parâmetros possuem interpretações. Por exemplo, nos modelos dados em (1), o parâmetro α refere-se a assíntota $y=\alpha$ e pode ser considerado como o máximo valor que a variável resposta y atinge. Interpretações dos demais parâmetros podem ser encontradas, por exemplo, em Mazucheli, (1995) ou Ratkowsky, (1983).

Observar que considerando os modelos de crescimento Logístico e Gompertz dados em (1) $\theta^t = (\alpha, \beta, \gamma)$ e considerando os modelos Tipo-Weibull, M.M.F. e Richards, $\theta^t = (\alpha, \beta, \gamma, \delta)$ em geral precisamos utilizar um método de integração numérica ou de aproximação para obter as quantidades a posteriori de interesse (densidades a posteriori marginais, momentos a posteriori e densidades preditivas de interesse).

Um método muito utilizado para resolver integrais em inferência Bayesiana, especialmente quando o número de parâmetros não é muito grande, é dado pelo método de Laplace (Tierney e Kadane, 1986 ou Tierney, Kass e Kadane, 1989). Outras aplicações do método de Laplace em modelos não lineares de crescimento cujo comportamento é exponencial são tratadas em Ehlers e Gamerman (1996).

3. O MÉTODO DE LAPLACE

Supor que temos interesse em calcular momentos a posteriori para funções $g(\theta)$ positiva da forma:

$$E\{g(\theta)\} = \frac{\int g(\theta)\pi(\theta)L(\theta/D)d\theta}{\int \pi(\theta)L(\theta/D)d\theta} \qquad (6)$$

onde $\theta \in \Re^p$, $\pi(\theta)$ é uma densidade a priori e $L(\theta/\boldsymbol{D})$ é a função de verossimilhança para θ dado um vetor de dados D.

O método de aproximação para momentos a posteriori introduzido por Tierney e Kadane (1986), é baseado nas aproximações de Laplace para o numerador e denominador de (6). Essas aproximações assumem que a principal contribuição para as integrais está vindo de um máximo no interior do espaço paramétrico Θ. O método de Laplace para aproximação de integrais é usado para resolver integrais na forma:

$$I = \int f(\theta)\exp\{-nh(\theta)\}d\theta, \qquad (7)$$

onde $-nh(\theta)$ é uma função com máximo em $\hat{\theta}$ e que satisfaz algumas condições usuais de regularidade.

Para aproximar integrais da forma (7) o método de Laplace assume uma expansão de h e f em série de Taylor em torno do máximo $\hat{\theta}$ de -h (Tierney e Kadane, 1986 ou Kass, Tierney e Kadane, 1990).

Com θ unidimensional, a aproximação de Laplace para (7) é dada por:

$$\hat{I} \cong \left(\frac{2\pi}{n}\right)^{\frac{1}{2}} \sigma f(\hat{\theta}) \exp\{-nh(\hat{\theta})\}, \tag{8}$$

onde $\sigma = \{h''(\hat{\theta})\}^{-\frac{1}{2}}$. No caso m-dimensional, ou seja, com $\theta \in \Re^p$, a aproximação de Laplace é dada por:

$$\hat{I} \cong (2\pi)^{\frac{p}{2}} \{\det[n\Sigma_h^{-1} h(\hat{\theta})]\}^{-\frac{1}{2}} f(\hat{\theta}) \exp(-nh\hat{\theta}), \tag{9}$$

onde $\Sigma_h^{-1} h(\hat{\theta})$ é a matriz hessiana de h calculada em $\hat{\theta}$.

Podemos considerar várias escolhas para f em (7). Quando f=1 o método de Laplace é conhecido como método de Laplace totalmente exponencial (Tierney e Kadane, 1986 ou Tierney, Kass e Kadane, 1989).

3.1 Modelos Logístico e Gompertz

De (4), uma densidade a priori não informativa para $\theta^t = (\alpha, \beta, \gamma)$ e σ^2 é dada por:

$$\pi_j(\theta, \sigma^2) = \{\alpha^2 A_j^{\frac{1}{2}}(\beta, \gamma)\sigma^1, \tag{10}$$

onde $A_j(\beta,\gamma) = b_{j1}b_{j4}b_{j6} + 2b_{j2}b_{j3}b_{j5} - b_{j3}^2 b_{j4} - b_{j1}b_{j5}^2 - b_{j2}^2 b_{j6}$; j=1: modelo logístico; j=2: modelo Gompertz e os b_{ji} são dados por:

$$b_{11} = \sum_{i=1}^{n}(1+a_{1i})^{-2} \qquad b_{12} = \sum_{i=1}^{n} a_{1i}(1+a_{1i})^{-3}$$

$$b_{13} = \sum_{i=1}^{n} x_i a_{1i}(1+a_{1i})^{-3} \qquad b_{14} = \sum_{i=1}^{n} a_{1i}^2(1+a_{1i})^{-4}$$

$$b_{15} = \sum_{i=1}^{n} x_i a_{1i}^2(1+a_{1i})^{-4} \qquad b_{16} = \sum_{i=1}^{n} x_i^2 a_{1i}^2(1+a_{1i})^{-4}$$

$$b_{21} = \sum_{i=1}^{n} \exp(-2a_{2i}) \qquad b_{22} = \sum_{i=1}^{n} a_{2i}\exp(-2a_{2i}).$$

$$b_{23} = \sum_{i=1}^{n} x_i a_{2i}\exp(-2a_{2i}) \qquad b_{24} = \sum_{i=1}^{n} a_{2i}^2\exp(-2a_{2i})$$

$$b_{25} = \sum_{i=1}^{n} x_i a_{2i}^2\exp(-2a_{2i}) \qquad b_{26} = \sum_{i=1}^{n} x_i^2 a_{2i}^2\exp(-2a_{2i})$$

e $a_{ji} = \exp(\beta - \gamma x_i)$.

Considerando-se a densidade a priori (10), a densidade a posteriori conjunta para α, β, γ e σ^2 é dada por:

$$\pi_j(\alpha,\beta,\gamma,\sigma^2/D) \propto \alpha^2 A_j^{\frac{1}{2}}(\beta,\gamma)\sigma^{-1}\sigma^{-(n+1)} \exp\left[-\frac{1}{2\sigma^2}S_j(\alpha,\beta,\gamma)\right], \quad (11)$$

onde $$S_1(\alpha,\beta,\gamma) = \sum_{i=1}^{n}\left[y_i - \alpha(1+a_{1i})^{-1}\right]^2 \quad \text{e}$$

$$S_2(\alpha,\beta,\gamma) = \sum_{i=1}^{n}\left[y_i - \alpha\exp(a_{2i})\right]^2 .$$

Integrando (11) com relação a σ, obtém-se a densidade a posteriori conjunta para α, β e γ dada por:

$$\pi_j(\alpha,\beta,\gamma/D) \propto \frac{\alpha^2 A_j^{\frac{1}{2}}(\beta,\gamma)}{\left[S_j(\alpha,\beta,\gamma)\right]^{\frac{n}{2}}}. \quad (12)$$

Para determinar as densidades a posteriori marginais para α, β e γ usamos o método de Laplace. Por exemplo, para encontrarmos uma aproximação de Laplace para a densidade marginal para α, consideramos $f_j(\beta,\gamma) = A_j^{\frac{1}{2}}(\alpha,\beta,\gamma)$, e $-nh_{j\alpha}(\beta,\gamma) = -\frac{n}{2}\log\left[S_j(\alpha,\beta,\gamma)\right]$, j=1,2 em (7).

De forma similar, encontramos as densidades a posteriori marginais para β e γ, respectivamente, como também para os parâmetros dos modelos Tipo-Weibull, M.M.F. e Richards (Mazucheli, 1995).

Também considerando-se a priori não-informativa (10), a densidade preditiva para uma observação futura y* num dado valor fixado x* da variável independente, é dada por:

$$\pi(y^*/y) = \int\int\int\int f_j(y^*/\theta,\sigma^2)\,\pi_j(\theta,\sigma^2/D)d\sigma\, d\alpha\, d\beta\, d\gamma\ , \quad (13)$$

onde:

$$f_1(y^*/\theta,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left[y^* - \frac{\alpha}{1+\exp(\beta-\gamma x^*)}\right]^2\right\},$$

$$f_2(y^*/\theta,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}[y^* - \alpha\exp(-\exp(\beta-\gamma x^*))]^2\right\},$$

e,

$$\pi_j(\theta,\sigma^2/y) \propto \alpha^2 A_j^{\frac{1}{2}}(\beta,\gamma)\sigma^{-(n+1)}\exp\left[-\frac{1}{2\sigma^2}S_j(\alpha,\beta,\gamma)\right].$$

Para aproximar a integral múltipla dada em (13), também usamos o método de Laplace e aproximações similares são obtidas para os modelos Tipo-Weibull, M.M.F. e Richards.

4. DISCRIMINAÇÃO DE MODELOS DE CRESCIMENTO

Para discriminar modelos de crescimento sob o enfoque clássico, usualmente o estatístico considera medidas de vício (Box, 1971), medidas de curvatura (Bates e Watts, 1980), estudos de simulação ou medidas de assimetria.

Se as suposições básicas de erros independentes e identicamente distribuídos normais com média zero e variância constante σ^2 estiverem satisfeitas, o melhor modelo possível, dentre todos os propostos, é aquele que apresenta um comportamento mais próximo do comportamento linear, uma vez que a validade das inferências (previsões, intervalos de confiança, testes de hipóteses, etc.) dependem fundamentalmente desse comportamento. Se o comportamento for distante do comportamento linear, as inferências podem não ser confiáveis, principalmente para pequenas amostras e, muitas vezes, parametrizações alternativas podem reverter esse fato.

Sob o enfoque Bayesiano, a discriminação pode ser conduzida através das muitas estratégias que são propostas na literatura. Dentre elas, destacamos:

(i) Fator de Bayes (Jeffreys, 1939);

(ii) Critério baseado no conceito de entropia (Box e Hill, 1967);

(iii) Pseudo Fator de Bayes (Geisser e Eddy, 1979); e

(iv) Fator de Bayes a Posteriori (Aitkin, 1991).

Num artigo recente, Gelfand e Dey (1994) definem uma densidade preditiva genérica que engloba muitas das estratégias Bayesianas existentes.

O Fator de Bayes é definido por:

$$FB = \frac{\overline{L}_1^{(FB)}}{\overline{L}_2^{(FB)}}, \tag{14}$$

onde $\overline{L}_j^{(FB)} = \int L\ (\theta_j / D, M_j)\, \pi\ (\theta_j)\, d\theta_j$, (j = 1,2), é interpretado como a média a priori da função de verossimilhança $\overline{L}_j^{(FB)} = L(\theta_j / D, M_j)$ em relação a priori $\pi(\theta_j)$ e ao modelo M_j.

Um dos problemas relacionados ao Fator de Bayes (14) para discriminar dois modelos é o uso de densidades a priori impróprias, pois $\overline{L}_j^{(FB)}$ também será imprópria, e (14) não pode ser usado como critério de comparação. Outra restrição relacionado ao uso do Fator de Bayes é verificado quando $n \to \infty$, que pode implicar em $FB \to \infty$, dando evidências ao modelo M_1, sendo este melhor ou não (paradoxo de Lindley).

No sentido de contornar os problemas associados ao Fator de Bayes, várias modificações são sugeridas na literatura como as citadas em (iii) e (iv).

Gelfand e Dey (1994) introduzem a densidade preditiva,

$$f(y_1 / y_2, M_j) = \frac{\int\int L(\theta_j, \sigma_j^2 / y_1, M_j)\, L(\theta_j, \sigma_j^2 / y_2, M_j)\, \pi(\theta_j, \sigma_j^2)\, d\sigma_j d\theta_j}{\int\int L(\theta_j, \sigma_j^2 / y_2, M_j)\, \pi(\theta_j, \sigma_j^2)\, d\sigma_j d\theta_j}, \tag{15}$$

onde, $\mathbf{y} = (y_1,...,y_n)$ é o vetor de variáveis resposta; $\mathbf{y}_1 = (y_1,...,y_{n_1})$, $\mathbf{y}_2 = (y_1,...,y_{n_2})$ são subconjuntos arbitrários de y, de tamanhos n_1 e n_2 , respectivamente; $L(\theta_j,\sigma_j^2 / \mathbf{y}_1,M_j)$ é a função de verossimilhança do j-ésimo modelo de crescimento definidos em (1), (j=1,...,5), dado n_1 observações; $L(\theta_j,\sigma_j^2 / \mathbf{y}_2,M_j)$ é a função de verossimilhança do j-ésimo modelo dado n_2 observações e $\pi(\theta_j,\sigma_j^2)$ é a densidade a priori para o j-ésimo modelo.

Alguns casos especiais de (15) são:

(i) $\mathbf{y}_1 = \mathbf{y} = (y_1,...,y_n)$ e $\mathbf{y}_2 = \varnothing$, resultando na densidade a posteriori marginal de θ_j e σ_j^2. Neste caso o denominador de (15) é ignorado (Gelfand e Dey, 1994).

(ii) $\mathbf{y}_1 = \{y_r\}$ e $\mathbf{y}_2 = \mathbf{y} - \{y_r\}$ produzindo a densidade $f(y_r / \mathbf{y}_{(r)},M_j)$ conhecida como densidade cross-validation, (Geisser, 1975 ou Stone, 1974), onde $\mathbf{y}_{(r)} = (y_1, y_2,..., y_{r-1}, y_{r+1},..., y_n)$. Geisser e Eddy (1979) consideram $\Pi_{r=1}^{n} f(y_r / \mathbf{y}_{(r)}, M_j)$.

(iii) $\mathbf{y}_1 = \mathbf{y}_2 = \mathbf{y} = (y_1,...,y_n)$ o qual resulta na densidade a posteriori preditiva proposta por Aitkin (1991). A idéia é ponderar a densidade conjunta de y com relação a posteriori e não com relação a priori.

(iv) $\mathbf{y}_1 = (y_1,...,y_{n_1})$, $\mathbf{y}_2 = \{y_{n_1+1}\}$, calculando (15) seqüencialmente, isto é, até que $\mathbf{y}_1 = (y_1,...,y_{n-1})$ e $\mathbf{y}_2 = \{y_n\}$, obtêm-se as probabilidades utilizadas no critério de discriminação proposto por Box e Hill (1967).

A partir de (i) obtém-se Fator de Bayes definido em (14). A partir de (ii), obtém-se o Pseudo Fator de Bayes (Geisser e Eddy, 1979), dado por:

$$PSFB = \frac{\Pi_{r=1}^{n} f(y_r / \mathbf{y}_{(r)},M_j)}{\Pi_{r=1}^{n} f(y_r / \mathbf{y}_{(r)},M_{j^*})}. \quad (j \neq j^*) \tag{16}$$

A partir de (iii), obtém-se o Fator de Bayes a Posteriori (Aitkin, 1991) dado por:

$$FBPO = \frac{f(\mathbf{y} / \mathbf{y},M_j)}{f(\mathbf{y} / \mathbf{y},M_{j^*})}, \quad (j \neq j^*) \tag{17}$$

e a partir de (iv) obtém-se as probabilidades utilizadas no critério baseado no conceito de entropia proposto por Box e Hill (1967).

Usando o método de Laplace e os modelos de crescimento dados em (1), podemos obter formulas simples para os critérios introduzidos acima (Mazucheli, 1995).

5. UM EXEMPLO

Os dados da Tabela (1), introduzidos por Heyes e Brown (Heyes e Brown, 1956) e também analisados por Ratkowsky (1983), referem-se à quantidade de água presente em células de raízes de feijão (variável dependente y) obtida em diferentes pontos das raízes (variável independente x).

Tabela 1 - Dados de Heyes e Brown

x	y
0.5	1.3
1.5	1.3
2.5	1.9
3.5	3.4
4.5	5.3
5.5	7.1
6.5	10.6
7.5	16.0
8.5	16.4
9.5	18.3
10.5	20.9
11.5	20.5
12.5	21.3
13.5	21.2
14.5	20.9

A partir de uma análise gráfica, verificamos uma forma similar à dada na Figura (1). Portanto, decidimos analisar os dados considerando um dos cinco modelos dados em (1): Logístico, Gompertz, Tipo-Weibull, Morgan-Mercer-Flodin e Richards.

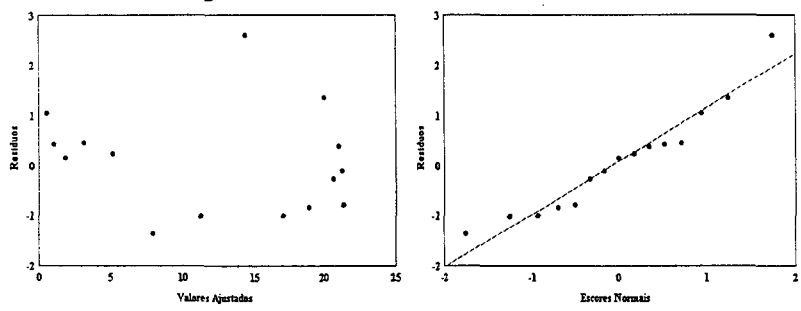
Usando o método de Gauss-Newton, com critério de convergência $\eta \leq 1e-10$ ou 50 iterações e valores iniciais obtidos pelas técnicas sugeridas por Ratkowsky (1983), obtemos as estimativas de mínimos quadrados dos parâmetros (Tabela 2). Ressaltamos que os mesmos valores iniciais usados no método de Gauss-Newton foram utilizados para obtenção a posteriori marginais.

Tabela 2 - Estimativas de Mínimos Quadrados

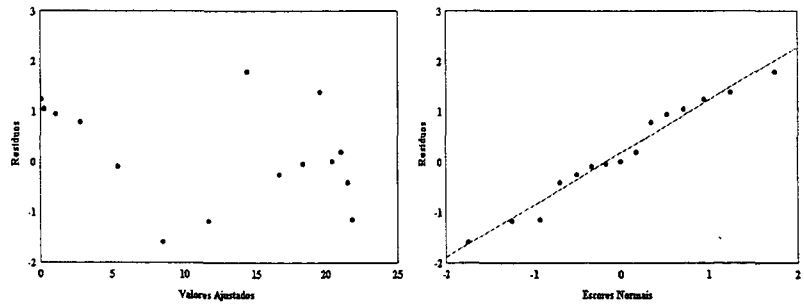
Modelo	Estimativas					Iterações
	α	β	γ	δ	σ^2	
Logístico	21.5089	3.9573	0.6222	-	0.5175	5
Gompertz	22.5066	2.1063	0.3881	-	1.0492	9
Weibull	21.1036	19.8147	0.0018	3.1796	0.4952	4
M.M.F.	22.0772	1.6531	5586.05	4.5601	0.5791	7
Richards	21.2040	5.6919	0.7772	1.6189	0.5021	7

Na Figura (2), temos os gráficos de resíduos considerando os cinco modelos de crescimento (ver 1). Observar que, podemos ter dúvidas sobre qual modelo melhor se ajusta aos dados.

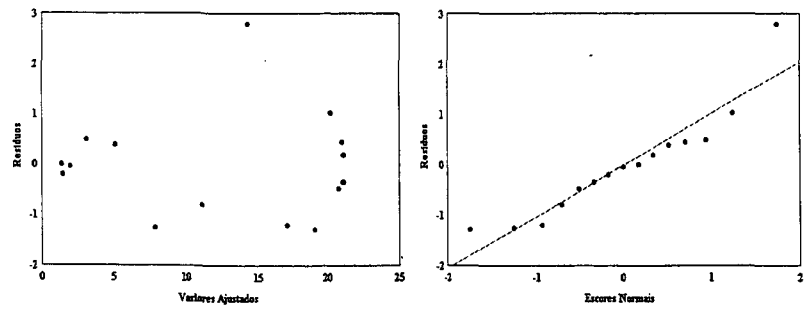
Figura 2 - Gráficos de Resíduos e Escores Normais



Modelo Logístico

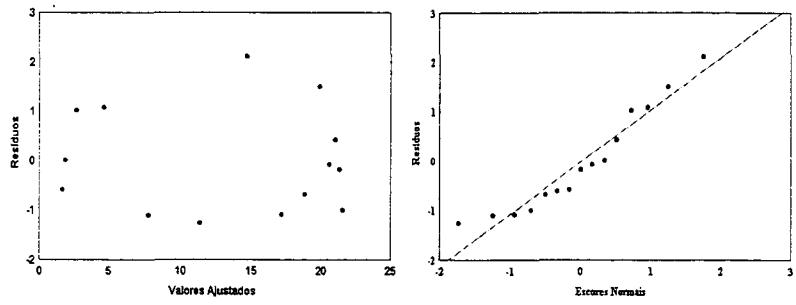


Modelo Gompertz

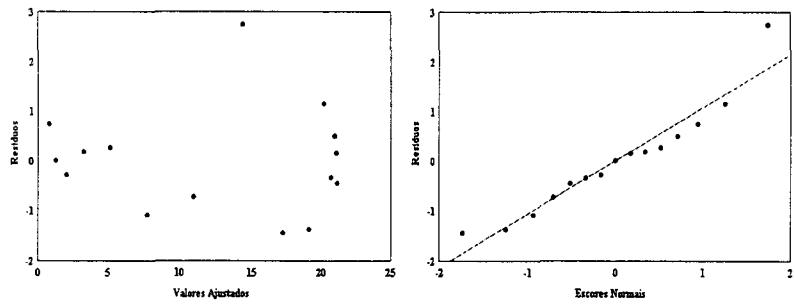


Modelo Tipo-Weibull

Figura 2 - Gráficos de Resíduos e Escores Normais (continuação)



Modelo M.M.F.



Modelo Richards

A partir das medidas de curvatura de Bates e Watts (1980) (Tabela 3), e comparando as medidas "IN" (curvatura intrínseca) e "PE" (curvatura paramétrica), com os valores críticos $1/(2\sqrt{F})$ para modelos de três parâmetros, e $1/(2\sqrt{F})$ para os modelos de quatro parâmetros, podemos observar que só o modelo Richards apresenta curvatura intrínseca significativa, indicando assim que para o conjunto de dados da Tabela (1), a curvatura do espaço de estimação não é razoavelmente pequena (por exemplo Bates e Watts, 1988). Com respeito à curvatura paramétrica, o modelo Logístico é o modelo que apresenta menor não linearidade devido ao efeito dos parâmetros, apesar desta ser, de certa forma, significativa. Vale ressaltar que uma alta não linearidade paramétrica (significativa) indica que pelo menos um parâmetro, no modelo, comporta-se não linearmente e o cálculo do vício e porcentagens, através da medida proposta por Box (1971), pode revelar qual, ou quais, são estes parâmetros. Na Tabela 4, é apresentado o vício e porcentagem de vício de cada parâmetro dos modelos em consideração.

Tabela 3 - Medidas de Curvatura (*: Não Significativa ; **: Significativa)

Modelo	Curvaturas Máximas	
	IN	PE
Logístico	0.1070*	0.3717**
Gompertz	0.2318*	0.8799**
Weibull	0.2319*	13.2534**
M.M.F.	0.2102*	24.9338**
Richards	0.2948**	4.2680**

Tabela 4 - Medida de Vício e Porcentagem de Vício = (100 Vício)/$\hat{\theta}$

Modelo	Estimativa do Vício				Porcentagem de Vício			
	α	β	γ	δ	α	β	γ	δ
Logístico	0.0154	0.0173	0.0028	-	0.0714	0.4381	0.4439	-
Gompertz	0.0608	0.0225	0.0038	-	0.2703	1.0659	0.9759	-
Weibull	0.0217	0.0311	0.0002	0.0222	0.1028	0.1571	11.4855	0.6975
M.M.F.	0.0524	0.0131	3091.382	0.0367	0.2374	0.7942	55.3410	0.8051
Richards	0.0389	0.2522	0.0329	0.1030	0.1835	4.4310	4.2322	6.3632

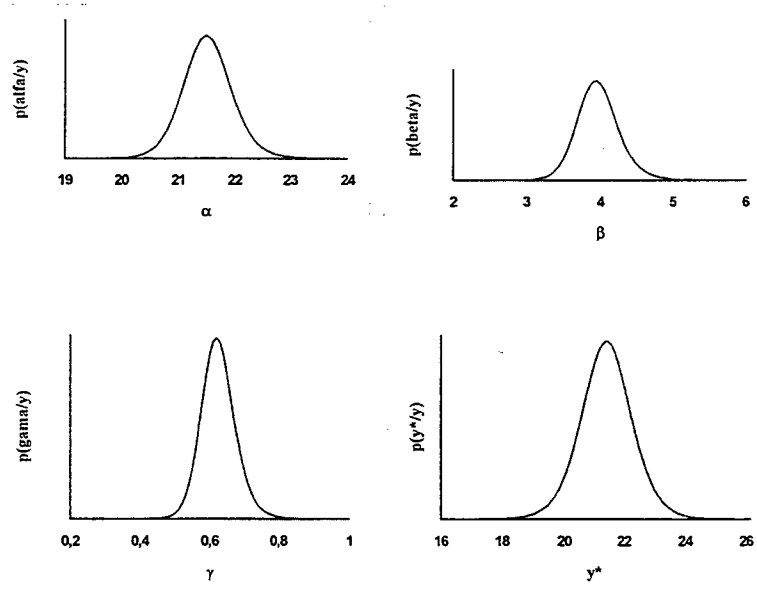
Observando a Tabela 4, nota-se que no modelo Logístico todos o parâmetros apresentam vício não significativo, ou seja, as porcentagens são inferiores a 1% em valor absoluto, sendo estas, em sua maioria, inferiores às porcentagens de vício dos demais modelos. Isso indica que o comportamento não linear do modelo Logístico pode ser pequeno em termos práticos, uma vez que seus parâmetros são mais estáveis e é possível que a medida que o tamanho da amostra aumente a curvatura devido ao efeito de parâmetros e os vícios se tornem desprezíveis.

Para o modelo Gompertz, vícios e porcentagens são todos maiores que do modelo Logístico estando assim de acordo com a não linearidade significativa devido ao efeito de parâmetros. Para o modelo Tipo-Weibull, podemos verificar que o parâmetro γ é o maior responsável pela extensão do comportamento não linear, sendo que o mesmo fato é verificado no modelo Morgan-Mercer-Flodin. E dentre os modelos especificados a partir de 4 parâmetros, o modelo Richards é, para o conjunto de dados em consideração, o modelo com maior número de parâmetros com vício significativo.

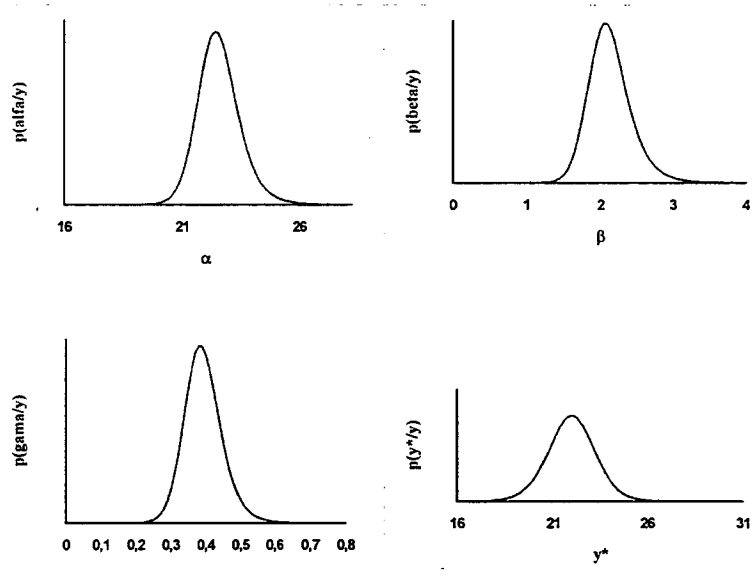
Alternativas e críticas a respeito da medida de vício proposta por Box (1971) são discutidas, dentre outros, em Ratkowsky (1990). Dentre essas alternativas pode-se citar o T-Plot de Hills e Smith (1993) a qual é uma alternativa gráfica. Estudos revelam que parâmetros com vício significativo são facilmente detectados a partir do T-Plot de Hills e Smith (por exemplo Mazucheli, 1995).

Considerando-se a priori não informativa (4) e o método de Laplace (Seção 2.1), obtemos gráficos das densidades a posteriori e preditivas (um valor $x^*=15$ da variável independente) de interesse (Figura 3).

Figura 3 - Densidades a Posteriori Marginais e Preditivas em x*=15

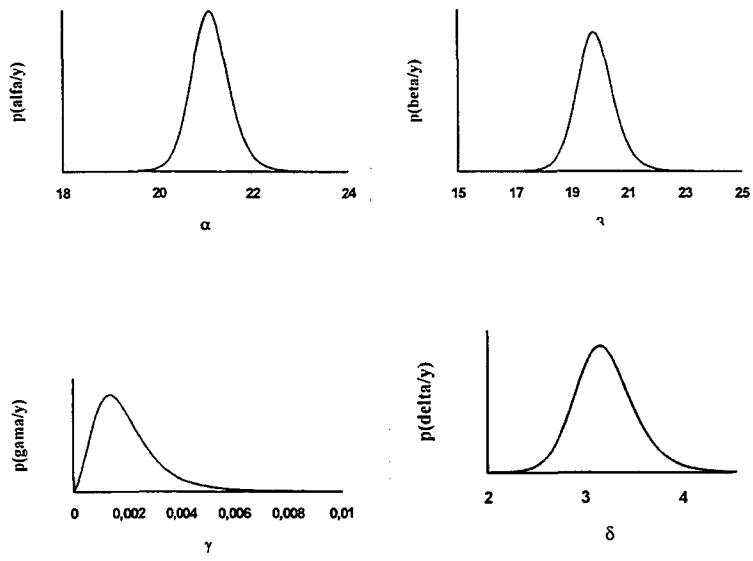


Modelo Logístico

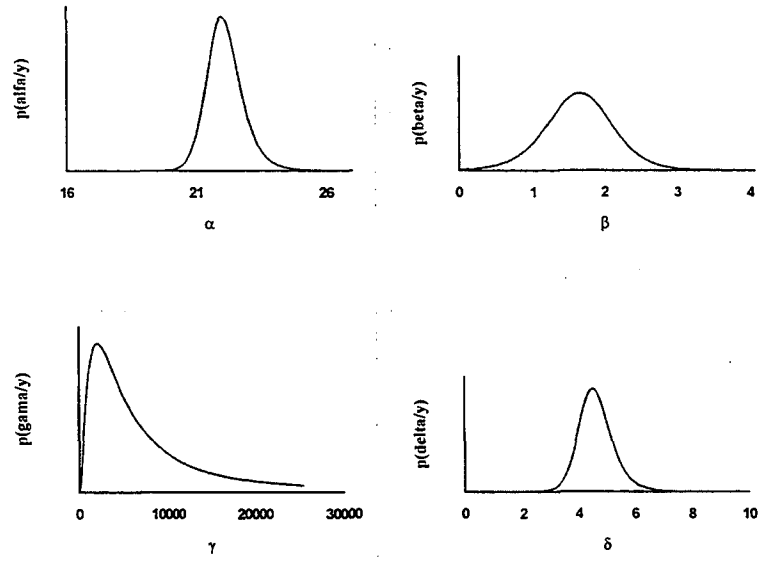


Modelo Gompertz

Figura 3 - Densidades a Posteriori Marginais e Preditivas em x*=15

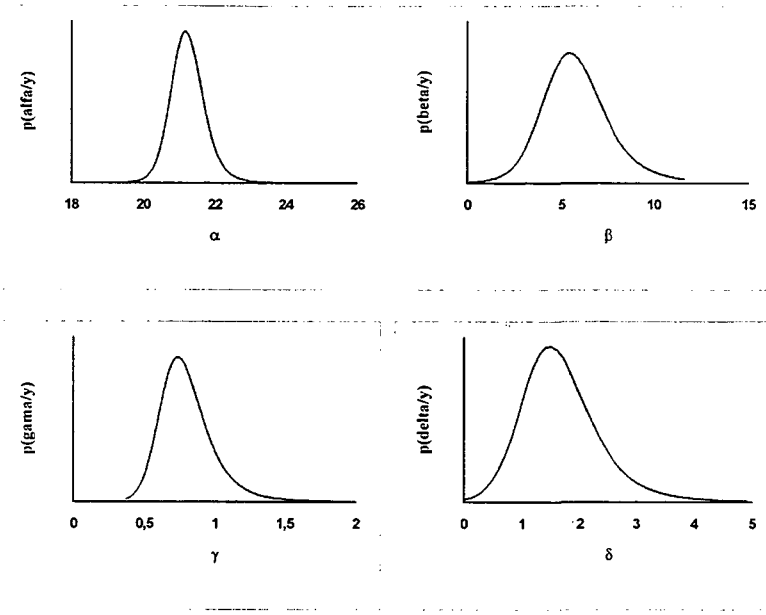


Modelo Tipo-Weibull

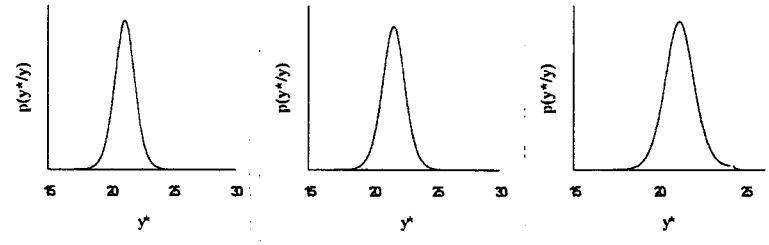


Modelo Morgan-Mercer-Flodin

Figura 3 - Densidades a Posteriori Marginais e Preditivas em x*=15 (continuação)



Modelo Richards



Modelo Tipo-Weibull **Modelo M.M.F.** **Modelo Richards**

Observar que parâmetros com vício significativo apresentam, em geral, gráfico a posteriori marginal bastante assimétrico.

Sob o enfoque Bayesiano, podemos discriminar os cinco modelos de crescimento propostos, como discutidos em (4).

Inicialmente, considerando-se a técnica proposta por Box e Hill (1967), assumindo mesma probabilidade a priori para todos os modelos, isto é, $\pi_{j,n_1} = 0.2$ e $n_1 = 12$ observações, temos na Tabela 5 as probabilidades a posteriori, em cada etapa seqüencial. Para calcular esses valores, usamos o método de Laplace de integração.

Tabela 5 - Probabilidades a Posteriori dos 5 Modelos

n	$\pi_{n,1}$	$\pi_{n,2}$	$\pi_{n,3}$	$\pi_{n,4}$	$\pi_{n,5}$
12	0.2329	0.0927	0.2408	0.1959	0.2377
13	0.2680	0.0587	0.2384	0.2050	0.2298
14	0.2861	0.0294	0.2613	0.1786	0.2447
15	0.2687	0.0132	0.3145	0.1291	0.2745
Modelo	Logístico	Gompertz	Weibull	M.M.F.	Richards

A partir da Tabela 5, observar que existe maior evidência para o modelo Tipo-Weibull.

Considerando comparações somente entre modelos de mesma dimensão a partir do procedimento proposto por Box e Hill (1967) e usando o método de Laplace, temos na Tabela 6, as probabilidades a posteriori para cada modelo. Observar que entre os modelos Logístico e Gompertz, claramente, ficamos com o modelo Logístico. Comparando-se os três modelos Tipo-Weibull, Morgan-Mercer-Flodin, e Richards, o melhor modelo seria dado pelo modelo Tipo-Weibull (Tabela 6).

Tabela 6 - Probabilidades a Posteriori dos Modelos de Mesma Dimensão

n	$\pi_{n,1}$	$\pi_{n,2}$	$\pi_{n,3}$	$\pi_{n,4}$	$\pi_{n,5}$
12	0.7153	0.2847	0.3571	0.2904	0.3524
13	0.8203	0.1797	0.3541	0.3045	0.3413
14	0.9069	0.0931	0.3817	0.2609	0.3574
15	0.9530	0.0470	0.4380	0.1798	0.3822
Modelo	*Logístico*	*Gompertz*	*Tipo-Weibull*	*M.M.F.*	*Richards*

Definindo M_1: modelo Logístico, M_2: modelo Gompertz, M_3: modelo Tipo-Weibull, M_4: modelo Morgan-Mercer-Flodin e M_5: modelo Richards, nas Tabelas 7 e 8, temos os resultados da discriminação Bayesiana, usando o Fator de Bayes, o Pseudo Fator de Bayes e o Fator de Bayes a Posteriori, introduzidos na Seção (4) e considerando o método de Laplace para aproximação de integrais. Note que o modelo M_i é preferível ao modelo M_j, para $i \neq j$, se a razão $M_i/M_j \geq 1$. Outras possíveis formas para decidir a respeito de um modelo com relação a outro são dadas, por exemplo, em Kass e Raftery (1995) e Jeffreys (1961).

Tabela 7 - Discriminação Bayesiana Baseada no Fator de Bayes

Procedimento	M_1 vs. M_2	M_1 vs. M_3	M_1 vs. M_4	M_1 vs. M_5
F. Bayes	63.8777	0.3042	0.6861	0.3587
F.Bayes a Posteriori	200.4549	0.5287	0.7107	0.5867
P. Fator Bayes	58.0387	0.0003	0.4804	0.4939

Na Tabela 8, temos os resultados de discriminação Bayesiana para modelos de mesma dimensão.

Tabela 8 - Discriminação Bayesiana - Modelos de Mesma Dimensão

	F. Bayes	F. Bayes Post.	P. Fator Bayes
M_3 vs. M_4	2.2557	3.2355	4414.8696
M_3 vs. M_5	1.1792	1.1096	13402.248

A partir dos resultados das Tabelas 7 e 8, novamente concluímos que o modelo Tipo-Weibull seria destacado como o melhor, mas quando comparamos modelos de mesma dimensão, o modelo M_1 (Logístico) é claramente superior. Problemas relacionados à discriminação de modelos de diferentes dimensões como também o uso a priori impróprias são discutidos, dentre outros em Berger e Pericchi (1996) e Gelfand e Dey (1994).

6. CONCLUSÕES

Neste artigo, apresentamos análises clássica e Bayesiana para alguns modelos de crescimento sigmóide discutidos na literatura. Aspectos de discriminação clássica e Bayesiana também foram tratados. Sob o ponto de vista Bayesiano, apesar de nos últimos anos ter havido grande aplicação de métodos baseados em simulação estocástica (simulação via cadeias de Markov), observamos que o uso do método de Laplace pode ser uma alternativa apropriada para a obtenção das quantidades a posteriori de interesse. Uma das vantagens do uso do método de Laplace é a obtenção de expressões analíticas que são de grande interesse para os pesquisadores das mais diversas áreas.

Como sugestão relacionada às metodologias de discriminação de modelos é razoável recomendar que todas as técnicas sejam utilizadas e decidir pelo modelo cujas evidências são mais freqüentes. Claro que utilizar todas as técnicas pode resultar num alto tempo computacional.

REFERÊNCIAS

Aitkin, M. (1991). Posterior Bayes Factor. J. R. Stat. Soc., B, 53, 111-142.

Bates, D. M.; Watts, D. G. (1980). Relative Curvature Measures of Nonlinearity. J. R. Stat. Soc., B, 42 ,1-25.

Bates, D. M.; Watts, D. G. (1988). Nonlinear Regression Analysis and its Applications. Wiley: New York. Beale, E. M. L. (1960). Confidence Regions in Nonlinear Estimation. J. R. Statist. Soc., B, 22 ,41-76.

Berger, J. O.; Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. J. Am. Stat. Assoc., vol. 91, 109-12.

Box, G. E. P.; Hill, W. J. (1967). Discrimination Among Mechanistic Models, Technometrics, 9, 57-71.

Box, M. J. (1971). Bias in Nonlinear Estimation. J. R. Stat. Soc., B, 33, 171,201.

EHLERS, R. S.; GAMERMAN , D. (1996). Analytic Approximation Methods in Bayesian Dynamic Non-Linear Models. Braz. Journ. of Prob. and Statistics, Vol 10, 1, 87-101.

Geisser, S.; Eddy, W. F. (1979). A Predictive Approach to Model Selection. J. Am. Stat. Assoc., 74, 153-160.

Gelfand, A. E.; Dey, D. K. (1994). Bayesian Model Choice: Asymptotic and Exact Calculations. J. R. Stat. Soc., B, 3, 501-514.

HEyes, K. K.; Brown, R. (1956). Growth and Cellular Differential, in F. L. Milthorpe (ED.). The Growth of Leaves, Butterworth, London.

HILLS, S. E.; SMITH, A. F. M. (1993). Diagnostics of Posterior Nonnormality in Bayesian Inference. Biometrica, 80, 1, 61-74.

Kass, R. E.; Raftery, A. E. (1995). Bayes Factor. J. Am. Stat. Assoc., Vol. 90, 773-795.

KASS, R. E.; TIERNEY, L.; KADANE, J. B. (1990).The Validity of Posterior Expansions Based on Laplace's Method. In Essays in Honor of George A. Barnard, Ed. J. Hodges, North- Holland, Amsterdam, 473-488.

JEFFEYS, H. (1939). Theory of Probability. .(1rd ED.) Oxford University Press.

JEFFEYS, H. (1961). Theory of Probability .(3rd ED.). Oxford University Press.

LOWRY, R. K.; MORTON, R. (1983). An Asymmetry Measure for Estimators in Non-Linear Regression Models. Proc. 44th Session Inst., Madrid, Contributed Papers, Vol. 1, 351-354.

Mazucheli, J. (1995). Análise Bayesiana e Discriminação de Modelos Não Lineares. Dissertação de Mestrado. ICMCS-USP.

Ratkowsky, D. A. (1983). Nonlinear Regression Models. Marcel Dekker: New York.

RATKOWSKY, D. A. (1990). Handbook of Nonlinear Regression Models. Marcel Dekker: New York.

Seber, G. A. F.; Wild, C. J. (1989). Nonlinear Regression. Wiley: New York.

STONE, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions (with Discussion). J. R. Stat. Soc. B, 36, 111-147.

Tierney, L.; Kadane, J. B. (1986). Accurate Approximations For Posterior Moments and Marginal Densities. J. Am. Stat. Assoc., 81, 82-86.

Tierney, L.; KASS, r. E.; Kadane, J. B. (1989). Approximate Marginal Densities For Non-Linear Functions. Biometrics, 76, 425-433

AGRADECIMENTOS

Este trabalho é resultado de pesquisa em nível de mestrado desenvolvida pelo primeiro autor no ICMSC/USP com apoio financeiro do CNPq. Os autores agradecem os árbitros e ao editor pelos valiosos comentários e sugestões.

Aspectos computacionais tratados neste artigo podem ser obtidos com o primeiro autor.

ABSTRACT

Considering some non-linear growth models, which comportment is sigmoidal (Logistic, Gompertz, Weibull Type, Morgan-Mercer-Flodin and Richards), we present a Bayesian analysis using a non-informative prior for the parameters and the Laplace's method for approximations of integrals. We also present some Bayesian techniques of discrimination of models and we illustrate with a data set.

Key Words: Growth Models, Bayesian Analysis, Laplace's Method.

Comparação do Efeito de Dois Critérios de Avaliação do Estado de Ruína de uma Seguradora, sobre o Percentual de Retenção em um Programa de Resseguro de Quota-Parte

Ary Elias Sabbag Junior*

RESUMO

Um dos aspectos mais importantes, quando do estabelecimento de um programa de resseguro de quota-parte, é a determinação do percentual de responsabilidade a ser assumido pela companhia seguradora. Normalmente nesta determinação são levados em conta apenas aspectos empíricos e julgamentos baseados em aspectos patrimoniais e financeiros, não se dando a devida atenção a modelos matemáticos baseados na teoria do risco, os quais poderiam auxiliar as companhias seguradoras na otimização do seu processo decisório. Neste trabalho é estabelecido um modelo sob o enfoque no qual a avaliação da condição de ruína da seguradora é feita em um único ponto de teste, conseqüentemente após um tempo finito. Dentro deste contexto, o modelo proposto determina o percentual de retenção da companhia cedente levando em conta a probabilidade de ruína aceitável pela mesma ao fim de um certo período. Também é apresentada a derivação do modelo de retenção, no caso em que se considera uma avaliação contínua em um horizonte de planejamento infinito. Finalmente são discutidas as implicações do uso de um modelo ou outro, sendo também apresentadas algumas simulações que comparam as retenções estabelecidas pelos dois enfoques.

1. INTRODUÇÃO

O resseguro é uma classe de negócios que tem por objetivo estabelecer um balanço necessário no negócio de seguros, em face deste ser constantemente influenciado por flutuações aleatórias, políticas monetárias e catástrofes. Desta forma, pode-se considerar que um programa de resseguro tem a função primordial de possibilitar a uma companhia seguradora o alcance de uma homogeneização e a obtenção de um limite para as responsabilidades assumidas pela mesma junto aos seus segurados. Assim, segundo este enfoque, Dirube (1991) argumenta que se pode assumir que o resseguro tem como alvo

* Endereço para correspondência: Rua: Prof. Brandão 218/202 - Alto da Rua XV - 80040 -100 - Curitiba - PR.

principal o fornecimento de condições que permitam a estabilização do comportamento da carteira de riscos assumidos, através da cobertura dos desvios ou desequilíbrios que afetam a freqüência, intensidade, distribuição temporal ou a quantia individual dos sinistros que ocorram na referida carteira. Uma das modalidades mais adotadas de contrato de resseguro, é a de Quota-Parte. Neste tipo de contrato todos os seguros em uma certa classe de negócio são ressegurados segundo um percentual uniforme e definido. Desta forma um percentual de cada um dos riscos assumidos pela companhia seguradora é cedido à resseguradora, sendo que a proporção da importância segurada de cada risco assumida pela resseguradora determina o quanto esta receberá de prêmio e o quanto assumirá nas indenizações relativas aos sinistros que venham a ocorrer. A resseguradora, por sua vez, repassa para a companhia seguradora uma comissão de resseguro, retornando para esta o prêmio cobrado associado a despesas administrativas e lucro. O principal problema que se apresenta com este tipo de resseguro é que o mesmo não vem a ser uma ferramenta capaz de homogeneizar totalmente os valores em risco de uma carteira, não resultando portanto em uma redução nas flutuações dos resultados dos seus negócios. Apesar deste fato, segundo Pfeiffer (1990), esta modalidade é muito aplicada em função da mesma ser recomendável no caso de uma companhia estar iniciando em um ramo para o qual não existem informações suficientes, no caso de necessidade de ajuda financeira quando de um forte incremento de sua produção ou ainda quando esta perdeu o controle da sua sinistralidade e não é possível sanear sua situação sem por em perigo as relações com a sua organização de vendas e com seus clientes. Outro aspecto associado a este tipo de contrato é o fato, conforme Dirube (1991), do mesmo ser bem favorável para os resseguradores, em função destes receberem participação em toda a carteira de um ramo, em igual medida para riscos bem, regularmente ou mal tarifados. Este fato permite uma segurança maior para a resseguradora devido a sua sorte estar associada integralmente à sorte da companhia cedente, sendo este outro motivo que leva a uma grande utilização desta modalidade de resseguro. Dadas as vantagens e desvantagens citadas, o resseguro de quota-parte geralmente é combinado com outros tipos de contratos, os quais permitem uma estabilização maior da carteira, como por exemplo, excesso de danos ou excedente de responsabilidade. Apesar deste fato e dada a importância do programa de resseguro em questão, na seqüência deste trabalho a análise se prenderá à determinação da retenção, da companhia cedente, em um programa de quota-parte puro.

No item seguinte é apresentada a descrição do problema, juntamente com algumas definições consideradas importantes.

2. DESCRIÇÃO DO PROBLEMA

Em um programa de quota-parte puro, um dos aspectos mais importantes é a determinação do percentual do risco a ser assumido pela companhia seguradora. Apesar desta constatação, Dirube (1991) afirma que normalmente na referida determinação, são levados em conta apenas aspectos empíricos e julgamentos baseados em dados patrimoniais e financeiros, não se levando em conta modelos matemáticos baseados na teoria do risco, em função das dificuldades no estabelecimento dos mesmos. No referido trabalho é citado Carter (1979), que argumenta que modelos estabelecidos com base matemática poderiam auxiliar, de maneira bem útil, as companhias no seu processo decisório. Objetivando-se a adoção de tais modelos, neste trabalho é efetuada uma abordagem de um programa puro de quota-parte, no qual o prêmio considerado é o líquido de custos, despesas e comissões.

A abordagem do problema é estabelecida segundo o enfoque de estabilização do resultado líquido de uma companhia, conforme tratamento dado por Straub (1988). Dentro deste contexto, a retenção será estabelecida levando-se em conta dois critérios de avaliação do estado de ruína de uma companhia. Conforme Bühlmann (1970), para definição destes critérios tem-se que um caminho aleatório associado a reservas livres de uma companhia será aceitável se suas ordenadas são positivas nos pontos para os quais se está interessado em avaliar o estado da companhia. A forma de seleção dos pontos de interesse caracteriza o critério de avaliação. Desta forma, os critérios possíveis são :

1) horizonte de planejamento finito, caso discreto;
2) horizonte de planejamento finito, caso contínuo;
3) horizonte de planejamento infinito, caso discreto; e
4) horizonte de planejamento infinito, caso contínuo.

Na sequência abordar-se-á o problema de retenção considerando-se a probabilidade de ruína segundo os enfoques 1 e 4, e adotar-se-á o conceito de que ocorrerá ruína quando o caminho aleatório, associado a reservas livres, for inaceitável. Esta terminologia (ruína) será adotada ao longo do texto, apesar da impropriedade da mesma.

Em relação ao primeiro critério, o horizonte finito a ser considerado será o período de um ano e o aspecto discreto ficará por conta de ser efetuada uma única avaliação no final do referido período. Quanto ao quarto critério, estar-se-á interessado em todos os pontos no intervalo de tempo $[0, \infty)$.

Feitos os comentários acima, no próximo item é apresentado o modelo proposto para o primeiro critério de avaliação. A seguir é apresentado um resumo da derivação do modelo

referente ao quarto critério e por último é estabelecida uma comparação entre os dois enfoques.

3. MODELO PARA DETERMINAÇÃO DO PERCENTUAL DE RETENÇÃO, COM UM ÚNICO PONTO DE AVALIAÇÃO DO ESTADO DE RUÍNA DA COMPANHIA

Para implementação da solução considerou-se o montante líquido total de indenizações ocorridas em um ano (X), como tendo uma distribuição de Poisson Composta Mista, com valor esperado (μ_X), variância (σ_X^2) e coeficiente de assimetria (γ_X), dados por:

$$\mu_X = n.\, a_1 = P$$

$$\sigma_X^2 = n.a_2 + n^2.m^2.s_q^2$$

$$\gamma_X = \frac{n.a_3 + 3.\, n^2.m.\, a_2.\, s_q^2 + n^3.\, m^3.\, g_q \cdot s_q^3}{s_X^3}$$

onde n é o número esperado de sinistros ;

$a_j = E(Z^j)$;

$m = a_1$;

Z variável aleatória que define o montante indenizado de um sinistro ;

σ_q = desvio padrão da variável de estrutura ; e

γ_q = coeficiente de assimetria da variável de estrutura .

Observando as equações acima percebe-se a incorporação de considerações relativas a oscilações a curto prazo (σ_q e γ_q). Este fato, implicará em considerar-se a suposição de independência, das referidas oscilações, em períodos consecutivos. Outro aspecto a ser salientado é que apesar do período adotado no presente trabalho ter sido de um ano, não há qualquer restrição para períodos de qualquer extensão.

Para o cálculo da probabilidade de ruína adotou-se a aproximação "Normal Power", ou seja, considerou-se X como tendo distribuição NP (μ_X , σ_X , γ_X), estando o detalhamento desta aproximação apresentado em Beard et alli (1984).

Desta forma, dado que a seguradora adote um programa de resseguro de quota parte e tenha como admíssivel uma probabilidade de ruína ε, tem-se que a retenção (α_{NP}) necessária da seguradora deverá ser tal que :

$$P[\, U+(1+\delta).\mu_X-(1+\delta_R).\mu_X.(1-\alpha_{NP})-\alpha_{NP}.X \geq 0\ ]=1-\varepsilon \qquad (1)$$

onde U são os recursos da companhia em um momento t=0 (p.e. no início do ano em estudo), disponíveis para absorver flutuações em suas operações técnicas :

δ é o carregamento de segurança da seguradora; e

δ_R é o carregamento de segurança da resseguradora .

Observando-se a probabilidade acima, pode-se perceber o estabelecimento da condição que ao final de um ano, as reservas livres da seguradora tenham probabilidade igual a 1 - ε de serem maiores ou iguais a zero, ou equivalentemente, o montante líquido destinado a cobrir as indenizações tenha uma probabilidade ε de ser insuficiente para cobrir as mesmas.

A probabilidade dada em (1) é equivalente a:

$$P\left[\, X \leq \frac{1}{\alpha_{NP}}.[U+(1+\delta).\mu_X-(1+\delta_R).(1-\alpha_{NP}).\mu_X\,]\,\right] = 1-\varepsilon$$

Dos resultados da aproximação NP obtém-se a equação para retenção dada por:

$$\alpha_{NP}=\frac{\frac{U}{\mu_X}+(\delta-\delta_R)}{\frac{\sigma_X}{\mu_X}.\left[y_\varepsilon+\frac{\gamma_X}{6}.(y_\varepsilon^2-1)\right]-\delta_R} \qquad (2)$$

onde y_ε é o valor de uma variável normal padronizada para o qual a área acima dele é igual a ε.

4. SOLUÇÃO PARA O PROBLEMA DE RETENÇÃO SOB O ENFOQUE DE HORIZONTE DE PLANEJAMENTO INFINITO E CONTÍNUO

O desenvolvimento a ser apresentado é feito conforme Beard et alli (1984), considerando-se desta forma o processo como estacionário, não se levando em conta flutuações a curto ou a longo prazo nem mudanças no carregamento de segurança da seguradora.

Feitas as considerações acima, seja Δ uma unidade de tempo. Sejam também Δ, $2.\Delta$, ..., $t.\Delta$, ... os tempos em que o estado de risco do processo é testado. Considerando-se que os sinistros agregados durante o período $((t-1).\Delta , t.\Delta]$ são denotados por X(t), tem-se que o lucro da seguradora, para cada período, será dado pela equação:

$$Y(t) = (1+ \delta) . E(X(t)) - X(t). \qquad (3)$$

No modelo a ser adotado, as variáveis Y(t), definidas acima, são consideradas mutuamente independentes e assumem valores negativos com probabilidade maior que zero. Definindo o lucro acumulado até o tempo t por :

$$W(t) = Y(1) + Y(2) + ... + Y(t)$$

tem-se que a probabilidade de ruína relativa ao período $(0, T.\Delta]$ fica sendo dada por

$$\Psi_T = \Psi_T(U) = 1 - Pr\{W(t) \geq -U, para\ t = 1, 2, ..., T\}$$

onde, como antes, U representa a reserva de riscos no início do período.

Para estimação de Ψ_T considere-se t_1 como sendo o tempo da primeira ruína. Com isto as realizações do processo estocástico podem ser separadas em dois grupos, ou seja, um grupo com todas as realizações que levam a ruína durante os primeiros T períodos (a probabilidade de obter uma realização pertencente a este grupo é Ψ_T) e o outro grupo com todas as outras realizações que levam a ruína subseqüentemente, ou seja, após $(0, T.\Delta]$.

Levando em consideração t_1, como definido acima, a função geratriz de momentos de W(t) será dada por :

$$M_T(s) = \Psi_T . E\{\exp(s.W(T)) \mid t_1 \leq T\} + (1 - \Psi_T) . E\{\exp(s.W(T)) \mid t_1 > T\} \qquad (4)$$

A solução para o cálculo de Ψ_T é obtida encontrando-se um valor para a variável auxiliar s tal que $M_T(s)=1$. Para tanto, pode ser provado que a função M é estritamente convexa, sendo crescente na origem (s=0) onde assume valor 1. Este fato associado a que $M_T(s) \rightarrow \infty$ quando $s \rightarrow -\infty$ leva a conclusão de que existe uma solução s = - R tal que $M_T(-R) = 1$.

Substituindo-se s por - R em (4), obtém-se a expressão :

$$\Psi_T.E\{\exp(-R.W(t_1)) \mid t_1 \leq T\} + (1- \quad _T).E\{\exp(-R.W(T)) \mid t_1 > T\} = 1 \qquad (5)$$

Considerando o fato de que t_1 é definido como sendo o tempo da primeira ruína, tem-se que a quantidade $W(t_1)$ corresponde ao lucro (negativo) no momento da ruína, sendo portanto sempre $\leq -U$. Substituindo $W(t_1)$ por - U e associando-se isto ao fato de que o segundo termo de (5) é sempre positivo ou igual a zero, obtém-se o resultado :

$$\Psi_T \leq e^{-R.U} \qquad (6)$$

Esta fórmula é um dos principais resultados da teoria do risco, sendo válida para qualquer número inteiro positivo T, valendo portanto também quando T tende a infinito.

Fica ainda a questão de como encontrar R em (6). Para tanto, deve-se considerar que:

$$M_T(s) = M(s)^T \qquad (7)$$

onde $M(s) \equiv M_{Y(t)}(s) = E\left\{e^{(s.Y(t))}\right\}$.

Substituindo s por - R em M(s) e considerando a definição de Y(t) em (3) obtém-se o resultado:

$$e^{-(1+\delta).P.R}.E(e^{R.X(t)}) = 1 \qquad (8)$$

onde P é o prêmio de risco, ou seja , o valor esperado de indenizações.

Na expressão acima, para o caso do processo de risco ser um Poisson Composto, pode-se expressar a função geratriz de momentos de X(t) em termos da função geratriz de momentos de Z, ou seja:

$$E\left(e^{R.X(t)}\right) = e^{n.\left[\int_0^{\infty} e^{R.Z} dS(Z)\right]} \qquad (9)$$

onde S(Z) é a função de distribuição de probabilidade da variável aleatória Z.

Substituindo (9) em (8) obtém-se a expressão:

$$\int_0^{\infty} e^{R.Z} dS(Z) = 1 + (1+\delta).a_1.R \qquad (10)$$

Expandindo em série de Taylor $e^{R.Z}$ e integrando termo a termo, tem-se que (10) é transformada para a forma

$$24.a_1.\delta = 12.a_2.R + 4.a_3.R^2 + a_4.R^3 + ... \qquad (11)$$

Desta forma, usando-se o primeiro ou os dois primeiros termos de (11) obtém-se a solução para R, ou seja:

$$R = \frac{2.\delta}{r_2.a_1} \qquad \text{(usando o primeiro termo de (11))}$$

$$R = \frac{1}{a_1}\sqrt{\frac{6.\delta}{r_3} + \frac{9.r_2^2}{4.r_3^2}} - \frac{3.r_2}{2.r_3.a_1} \text{ (usando os dois primeiros termos de (11))}$$

onde $r_i = \frac{a_i}{a_1^i}$.

A questão da continuidade fica por conta de que os resultados acima valem quando $\Delta \to 0$, conforme provado em Beard et alli (1984).

No caso da companhia seguradora adotar um programa de resseguro de quota-parte, retendo α de cada risco, ocorrerá que o lucro da mesma, para cada período, passará a ser :

$$\tilde{Y}(t) = (1+\delta).E(X(t)) - (1+\delta_R).E(X(t)).(1-\alpha) - \alpha.X(t) \qquad (12)$$

Neste contexto, o cálculo da probabilidade de ruína de uma companhia, quando a mesma retém α de cada risco, se dará por resolver-se a equação (7), substituindo-se $Y(t)$ por $\tilde{Y}(t)$, ou seja, resolvendo-se :

$$E\left\{e^{-R.\tilde{Y}(t)}\right\} = 1 \qquad (13)$$

Aplicando o logaritmo neperiano em (13) e expandindo em série de Taylor este resultado, obter-se-á o valor de R. Para o caso de considerarem-se os três primeiros termos da referida expansão, o valor obtido será dado por:

$$R = \frac{2.\mu_X[(\delta - \delta_R) + \alpha.\delta_R]}{\alpha^2.\sigma_X^2} \tag{14}$$

Definindo ε como a probabilidade de ruína admissível da companhia seguradora, e igualando-a ao lado direito da desigualdade dada em (6), tem-se que a probabilidade de ruína verdadeira da companhia será menor que a probabilidade de ruína admissível. Desta forma, substituindo (14) em (6) obtém-se o valor α, o qual fica sendo dado por:

$$\alpha = \frac{\delta_R + \sqrt{\delta_R^2 + 2.\frac{\sigma_X^2}{\mu_X}\left(\frac{-\ln\varepsilon}{U}\right).(\delta - \delta_R)}}{\frac{\sigma_X^2}{\mu_X}.\left(-\frac{\ln\varepsilon}{U}\right)} \tag{15}$$

5. COMPARAÇÃO ENTRE OS DOIS ENFOQUES

Observando-se as equações (2) e (15) percebe-se o estabelecimento de duas regras para a determinação da retenção de uma companhia seguradora em um programa de quota-parte. Estas regras compreendem situações extremas para o problema, abordando a questão da retenção sob os enfoques de um ano e infinito.

Na questão da comparação das duas equações, se faz necessário levar em conta a filosofia adotada na derivação das mesmas. Desta forma, em relação à equação (2), não se deve perder de vista o fato de que o pressuposto básico considerado para seu estabelecimento foi o de uma avaliação anual do estado de ruína da companhia. Este período de um ano pode coincidir ou não com o ano fiscal, mas a questão operacional que fica por trás deste estabelecimento é a necessidade da revisão das retenções ao final de cada período selecionado. Além do motivo conceitual, esta necessidade tem como fundamento não estarem sendo consideradas questões relativas a tendências e variações nas probabilidades básicas a longo prazo, sendo consideradas apenas variações a curto prazo. Este fato terá um impacto grande sobre a probabilidade de ruína e em conseqüência nas retenções, no caso da não reavaliação periódica dos componentes do modelo.

Quanto à abordagem infinita, esta tem merecido uma considerável atenção na literatura atuarial. Como um exemplo deste interesse, pode-se citar uma publicação da Swiss

Re (1992), onde são enfatizados aspectos operacionais da retenção sob o referido enfoque, visando-se difundir a utilização do mesmo. Um dos problemas do enfoque de tempo infinito é que a probabilidade de ruína sofre uma influência muito acentuada de δ. Por sua vez δ é grandemente afetado por correções de tarifa, diferentes tipos de tendências e por outras influências cíclicas (Beard et alli (1984)). Estes fatos levam a uma falta de estabilidade para δ, o que por sua vez influencia grandemente a retenção sob o enfoque infinito. Outro aspecto desta abordagem é a questão de não serem levadas em conta tendências das várias variáveis envolvidas no problema. Sobre isto, Beard et alli (1984) argumentam que algumas das tendências não apresentam influências fatais nas probabilidades de insolvência, e por conseguinte nas retenções sob esta abordagem. Esta argumentação tem como fundamento o fato de que para algumas tendências (por exemplo, o número de sinistros) apenas o começo do processo tem influência sobre a probabilidade de ruína, sendo esta uma justificativa para considerar-se a probabilidade de ruína num tempo infinito como um limite superior aceitável para aproximações finitas. Infelizmente, conforme citado no mesmo trabalho, também existem algumas tendências e oscilações que a longo prazo influenciam na probabilidade de ruína, sendo as referidas influências de difícil quantificação.

Em relação a similaridades dos modelos, seguindo Straub (1988), pode-se perceber que os mesmos termos de (15) estão presentes em (2), ou seja, a reserva inicial em unidades monetárias do prêmio de risco, o coeficiente de variação do montante total de indenizações, a probabilidade de ruína tolerada pela organização e a diferença entre os carregamentos de segurança da seguradora e da resseguradora. Outro aspecto a ser observado, é que estes termos influenciam no mesmo sentido as retenções em ambos enfoques. A diferenciação das abordagens fica por conta da intensidade das referidas influências, podendo-se perceber serem as retenções sob o enfoque infinito menores do que as retenções estabelecidas pelo enfoque anual. Nesta questão, fica clara a influência da forma de avaliação do estado de ruína da companhia seguradora. Estas observações podem ser melhor visualizadas no quadro abaixo, onde são apresentadas algumas simulações de retenções sob os enfoques discreto finito (α_{NP}) e contínuo infinito (α).

Tabela 1
Comparação das retenções obtidas em um programa de resseguro de quota-parte, segundo os enfoques discreto finito e contínuo infinito

			$\delta = 0,03$		$\delta = 0,05$	
$\varepsilon =$	$U/\mu_X =$	$\sigma_X/\mu_X =$	$\delta_R = 0,01$	$\delta_R = 0,03$	$\delta_R = 0,01$	$\delta_R = 0,03$
0,01	0,05	0,10	α_{NP}=0,269 α =0,219	α_{NP}=0,208 α =0,065	α_{NP}=0,346 α =0,306	α_{NP}=0,292 α =0,243
		0,15	α_{NP}=0,177 α =0,144	α_{NP}=0,133 α =0,029	α_{NP}=0,228 α =0,201	α_{NP}=0,187 α =0,154
	0,10	0,10	α_{NP}=0,462 α =0,317	α_{NP}=0,417 α =0,130	α_{NP}=0,539 α =0,439	α_{NP}=0,500 α =0,367
		0,15	α_{NP}=0,304 α =0,206	α_{NP}=0,267 α =0,058	α_{NP}=0,354 α =0,288	α_{NP}=0,320 α =0,228
0,05	0,05	0,10	α_{NP}=0,415 α =0,276	α_{NP}=0,336 α =0,100	α_{NP}=0,533 α =0,382	α_{NP}=0,471 α =0,313
		0,15	α_{NP}=0,271 α =0,180	α_{NP}=0,210 α =0,044	α_{NP}=0,349 α =0,251	α_{NP}=0,294 α =0,196
	0,10	0,10	α_{NP}=0,711 α =0,400	α_{NP}=0,672 α =0,200	α_{NP}=0,830 α =0,551	α_{NP}=0,807 α =0,479
		0,15	α_{NP}=0,465 α =0,259	α_{NP}=0,420 α =0,089	α_{NP}=0,542 α =0,360	α_{NP}=0,504 α =0,292

Obs.: em todas simulações considerou-se $\gamma_X = 0,5$

6. CONCLUSÃO

O objetivo principal deste trabalho foi o da apresentação de dois modelos, que permitem o estabelecimento da retenção de uma seguradora, em um programa de resseguro de quota-parte, levando-se em conta a teoria do risco. Paralelamente, procurou-se estabelecer as distinções conceituais entre as duas abordagens, preocupando-se em evidenciar a importância dos aspectos teóricos existentes na derivação dos modelos. Como última colocação, fica implícito que quanto à questão da escolha de um enfoque ou outro, deve-se considerar esta decisão como conseqüência do planejamento estratégico da empresa. Assim, quando se optar por um planejamento a curto prazo estar-se-á admitindo retenções maiores do que aquelas obtidas quando da adoção de um planejamento que leve em conta questões que tratam da continuação da vida da companhia a longo prazo.

REFERÊNCIAS

Beard, R. E. , T. Pentikäinen e E. Pesonen (1984). *Risk Theory : The Stochastic Basis of Insurance*. Chapman and Hall, Great Britain.

Bühlmann, Hans (1970). *Mathematical Methods in Risk Theory*, Springer-Verlag, Berlin.

Carter, R. L. (1979). *El Reaseguro*, Editorial Mapfre, Madrid.

Dirube, Ariel Fernández (1991). *Manual de Reaseguros*,2a Edição, Volume 2, General RE, Buenos Aires.

Pfeiffer, Christoph (1990). *Introduction to Reinsurance*, Cologne Reinsurance Company Ltd., Cologne.

Straub, Ervin (1988). *Non-Life Insurance Mathematics*, Springer-Verlag, Zürich.

Swiss Re (1992). *Setting retentions : Fundamental considerations*, Swiss Reinsurance Company, Zurich.

ABSTRACT

One of the most important aspects in the establishment of a quota share reinsurance program is to determine the percentage of liability to be assumed by the insurance company. Often in this determination only empirical and judgement based on financial and assets aspects are taken into account, and proper attention is not given to mathematical models based on the risk theory, which could help the insurance company to optimise its decision process. In this paper a model is established that focuses on evaluating the risk of ruin of the insurance company .in only one test point, consequently after a finite time. In this context, the proposed model determines the retention percentage of the ceding company taking into account the ruin probability acceptable at the end of a certain period. The derivation of a retention model is also presented in which we consider a continuous evaluation in a finite planning horizon. Finally, implications of the use of one and other model are discussed, presenting some simulations which compare the established retention by the two approaches.

Revista Brasileira de Estatística - RBEs

Política Editorial

A Revista Brasileira de Estatística (RBEs) objetiva promover a Estatística relevante para aplicação em questões sociais, interpretadas amplamente para incluir questões educacionais, de saúde, demográficas, econômicas, legais, de políticas públicas e de estatísticas oficiais, entre outras. A revista pretende apresentar artigos num formato que permita fácil assimilação pelos membros da comunidade científica em geral. Os artigos devem incluir aplicações práticas como assunto central. Essas aplicações deverão ter conteúdo estatístico substancial. As análises deverão ser exaustivas e bem apresentadas, mas o emprego de métodos estatísticos inovadores não é essencial para publicação.

Artigos contendo exposição de métodos são aceitáveis, desde que estes sejam relevantes para as áreas cobertas pela revista, auxiliem na compreensão do problema e contenham interpretação clara das expressões matemáticas apresentadas. A apresentação de aplicações ilustrativas envolvendo dados adequados é requerida. Tratamentos algébricos extensos devem ser evitados.

A RBEs tem periodicidade semestral e publicará também artigos escritos a convite e resenhas de livros, bem como artigos abordando os diversos aspectos de metodologias relevantes para órgãos produtores de estatísticas, incluindo:

a) planejamento de pesquisas;
b) avaliação e mensuração de erros em pesquisas;
c) uso e combinação de fontes alternativas de informação; integração de dados;
d) novos desenvolvimentos em metodologia de pesquisa;
e) crítica e imputação de dados;
f) amostragem e estimação;
g) disseminação e confiabilidade de dados;
h) análise de dados;
i) análise de séries temporais;
j) modelos e métodos demográficos;
k) modelos e métodos econométricos.

Todos os artigos submetidos serão avaliados quanto à qualidade e relevância por dois especialistas indicados pelo Comitê Editorial da RBEs. Os artigos submetidos deverão ser inéditos e não deverão ter sido simultaneamente submetidos a qualquer outro periódico nacional. O processo de avaliação é do tipo duplo cego, isto é, os artigos são avaliados sem identificação da autoria, e os comentários dos avaliadores também são repassados aos autores sem identificação.

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBEs

Os artigos submetidos para publicação deverão ser remetidos em 3 vias (que não serão devolvidas) para:

Pedro Luis do Nascimento Silva
Editor Responsável
Revista Brasileira de Estatística - RBEs
Av. República do Chile 500, 10o. andar
Rio de Janeiro – RJ – 20031-170
Tel: +55 - 21 - 514 0470
Fax: +55 - 21 - 514 4785
E-mail: pedrosilva@ibge.gov.br

Os artigos submetidos à RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

Para cada artigo publicado, serão fornecidas gratuitamente 20 separatas.

Instruções para preparo de originais:

1. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se para cada um a filiação e endereço para correspondência. Agradecimentos a colaboradores e instituições e auxílios recebidos devem figurar também nesta página.

2. A segunda página do original deve conter resumos em português e em inglês (*Abstract*), destacando os pontos relevantes do artigo. Cada resumo deve ser datilografado seguindo o mesmo padrão do restante do texto, em um único parágrafo, sem fórmulas, com no máximo 150 palavras.

3. O artigo deve ser dividido em seções numeradas progressivamente, com títulos concisos e apropriados. Todas as seções e subseções devem ser numeradas e receber título apropriado.

4. A citação de referências no texto e a listagem final das referências devem ser feitas de acordo com as normas da ABNT.

5. As tabelas e gráficos devem ser precedidas de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas seqüencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Deve ser citada a fonte dos dados utilizados em tabelas e gráficos. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração tais como as demais seções do trabalho.

6. Gráficos e diagramas para publicação devem ser incluídos nos arquivos com os originais do artigo, sempre que possível. Quando isto não ocorrer, devem ser traçados em papel branco, como nitidez e boa qualidade, para permitir que a redução seja feita mantendo qualidade. Fotocópias não serão aceitas. É fundamental que não existam erros quer no desenho quer nas legendas ou títulos.

7. Serão preferidos originais processados pelo editor de texto *Word for Windows*.

Se o assunto é Brasil;
procure o IBGE

http://www.ibge.gov.br
http://www.ibge.org

atendimento
0800 21 81 81

ISSN 0034-7175