Presidente da República
Fernando Henrique Cardoso

Ministro do Planejamento, Orçamento e Gestão
Guilherme Gomes Dias

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Sérgio Besserman Vianna

Diretor Executivo
Nuno Duarte da Costa Bittencourt

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Maria Martha Malard Mayer

Diretoria de Geociências
Guido Gelli

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

Ministério do Planejamento, Orçamento e Gestão
Instituto Brasileiro de Geografia e Estatística - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 61 número 216 julho/dezembro 2000

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 61, n. 216 p. 1-90, jul./dez. 2000

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 responsável – 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.

Editoração
Helem Ortega da Silva - Departamento de Metodologia - DPE

Impressão
GráficaDigital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2002

Capa
Renato J. Aguiar – Gerência de Criação – CDDI

Ilustração da Capa
Marcos Balster – Gerência de Criação – 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.

I. 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 mais um número da RBEs, com quatro artigos. Abrindo o número, Abramovitz e Mendes discutem uma aplicação de métodos estatísticos robustos para a estimação e ajuste de modelos para avaliação do risco de crédito em instituições financeiras, considerando dados de empresas brasileiras. Em seguida, Junqueira e Achcar apresentam uma análise Bayesiana para modelagem da variação extrabinomial e aplicam métodos de simulação estocástica para a estimação e ajuste dos modelos considerados. Januzzi propõe dois índices aproximativos do status socioeconômico das ocupações brasileiras empregando dados dos Censos Demográficos de 1980 e 1991, bem como das PNADs da década de 1990, e com base nestes avalia os níveis de precarização dessas ocupações. Por último, Andrade e Gilardoni apresentam uma abordagem Bayesiana para combinar resultados de estudos de câncer entre espécies, onde mais uma vez são empregados métodos de simulação estocástica para a estimação e ajuste dos modelos considerados.

Mais uma vez, a variedade de contextos e aplicações da Estatística é bem refletida nesta combinação de artigos, cujas áreas de aplicação vão desde a análise da precarização das ocupações e relações de trabalho, passando pela análise do risco de crédito a empresas brasileiras, até a preocupação com a estimação dos riscos de câncer para diferentes espécies.

O Conselho Editorial da RBEs está em processo de ampliação e reformulação, estando prevista a incorporação de novos membros e a substituição do editor responsável até o final do ano de 2002. Nesta oportunidade, manifesto aos membros atuais do Conselho Editorial minha sincera gratidão pela colaboração recebida nestes quase cinco anos de atuação conjunta. Agradeço mais uma vez a dedicada colaboração dos revisores que tem avaliado artigos submetidos à RBEs.

Aos autores em potencial, convidamos a submeter seus trabalhos, que serão avaliados com base na política editorial em vigor.

Saudações,

Pedro Luis do Nascimento Silva

Editor Responsável

Sobre a não-existência dos estimadores de máxima verossimilhança: uma aplicação na estimação do risco de crédito

Lilian Abramovitz*

Beatriz Vaz de Melo Mendes**

RESUMO

A inadimplência de credores pode resultar em grandes perdas para instituições financeiras, as quais utilizam-se de diversas abordagens na tentativa de reduzir seus riscos.

Uma maneira de se modelar rigorosa e objetivamente a qualidade do crédito de um cliente é através do modelo de regressão logística, cujos parâmetros são geralmente estimados por máxima verossimilhança. Entretanto, existe uma classe de situações, relacionadas com o fenômeno da superposição, onde as estimativas únicas e finitas dos parâmetros não podem ser obtidas. Por outro lado, as observações responsáveis pela superposição podem ser atípicas e exercer grande influência sobre o ajuste. Neste caso, um procedimento robusto que desse pesos menores aos *outliers* deveria ser utilizado. Neste trabalho estimamos os risco de crédito de 56 firmas brasileiras utilizando dados de 1994. Propomos procedimentos para identificar os subconjuntos de observações influentes e para regular o estimador robusto. Utilizamos várias medidas para comparar a performance dos estimadores clássico e robusto. O procedimento robusto de estimação se mostra superior, pois além de possuir maior informação entrópica, permite atribuir pesos menores às observações atípicas enquanto preserva a superposição dos dados.

Palavras-chave: Superposição; Regressão Logística; e Estimação Robusta.

* Endereço para correspondência: Deptº de Métodos Estatísticos - UFRJ- Rio de Janeiro - RJ - e-mail: lilian@dme.ufrj.br.

** Deptº de Métodos Estatísticos - UFRJ - Rua: Marques de Santos, 22, apt. 1204, CEP. 22221-080 - Rio de Janeiro - RJ, e-mail: bmendes@dme.ufrj.br.

1. Introdução

No mercado financeiro, risco pode ser definido como uma estimativa para as possíveis perdas de uma instituição devidas às incertezas relacionadas com suas atividades diárias. Risco é um conceito multidimensional que cobre quatro grandes grupos: risco de mercado, risco operacional, risco de crédito e risco legal. Risco de Crédito é o risco resultante da incerteza de que a parte credora irá cumprir com suas obrigações contratuais.

A inadimplência de credores pode resultar em grandes perdas para instituições financeiras. Devido a isto o gerenciamento do risco de crédito vem tornando-se cada vez mais importante para as empresas. A análise de crédito envolve a habilidade de se fazer uma decisão de concessão de crédito dentro de um cenário de incertezas, de constantes mutações e informações incompletas (Schrikel, 1994). O objetivo final da análise de crédito é reduzir o risco potencial de inadimplência.

O tema análise de crédito é diretamente associado aos bancos comerciais. Sua importância é fundamental, pois a lucratividade e permanência dos bancos no mercado estão intimamente relacionadas aos critérios de seleção de seus clientes, já que o resultado do não recebimento de uma operação equivale à perda do valor emprestado e dos juros devidos. Por outro lado, restrições excessivas que limitam a concessão do crédito podem acabar transferindo clientes a bancos mais agressivos no que diz respeito à exposição aos riscos.

Existem diversas abordagens para a tarefa de tentar reduzir estes riscos. Um banco, por exemplo, pode procurar mensurar seu risco de crédito conhecendo seus possíveis devedores. Com esta finalidade, muitas instituições financeiras empregam analistas de crédito para estudar a posição financeira de seus clientes e tendências econômicas gerais. Outras utilizam vários modelos e medidas de risco para classificar um cliente em potencial. Esses incluem a história de crédito deste cliente, uma avaliação do risco interno dada por agências especializadas que geram uma ordenação dos clientes (baseadas em escores discretos ou contínuos), e estimativas ou previsões geradas por algum modelo estatístico. Modelos confiáveis fornecem uma maneira efetiva em termos de custo para viabilizar e tornar mais rápido o processo de aprovação de um crédito, além de servir para monitorar o risco de crédito.

Uma maneira de se modelar rigorosa e objetivamente a qualidade do crédito de um cliente é através do modelo estatístico de regressão logística (para outras abordagens veja, por exemplo, Altman (1968), ou Almeida e Dumontier (1996)). Neste caso, a "saúde financeira" das firmas seria codificada em 1, caso a firma estivesse em inadimplência, ou em 0, caso contrário. As várias informações extraídas (Pereira; 1995) das Demonstrações Contábeis e Financeiras (Balanço Patrimonial, Demonstração de Resultado do Exercício, Demonstrações de Lucros ou Prejuízos Acumulados) das firmas poderiam ser utilizadas como variáveis explicativas na estimação das probabilidades de inadimplência.

A estimação dos parâmetros do modelo é geralmente feita pelo método da máxima verossimilhança. Entretanto, existe uma classe de situações onde o processo de estimação falha na convergência, fazendo com que estimativas únicas e finitas dos parâmetros não possam ser obtidas. A existência dessas estimativas está relacionada com o fenômeno da superposição (Albert e Anderson, 1984), o qual ocorre quando não é possível achar nenhuma combinação linear das covariáveis que separe completamente os 0's e os 1's. A interpretação

geométrica deste resultado é que o estimador de máxima verossimilhança existe somente se não há um hiperplano separando sucessos e falhas, onde o próprio hiperplano pode conter tanto sucessos como falhas. Conjunto pequeno de dados e/ou correlação alta entre as variáveis independentes favorecem a não ocorrência da superposição.

Por outro lado, os dados podem possuir observações discrepantes capazes de exercer grande influência sobre o ajuste. Notemos que no caso da regressão logística, uma observação pode ser considerada discrepante se é um *outlier* no espaço dos fatores ou se pertence à região do espaço dos fatores ocupada pelos 0's, sendo um 1 (e vice-versa). Alguns autores (por exemplo, Pregibon (1981), e Johnson (1985)) propuseram procedimentos destinados a identificar as observações influentes. Uma maneira (robusta) de contornar este problema é identificar essas observações, retirá-las do conjunto de dados, e ajustar novamente o modelo ao conjunto de dados sem as observações. Contudo, no caso da regressão logística, retirar observações influentes não é garantia de bons resultados, principalmente quando as observações influentes não forem pontos extremos no espaço dos fatores. Pode acontecer que o conjunto completo de dados apresente superposição, mas o conjunto reduzido não. Nesta situação , nem mesmo um estimador robusto irá produzir estimativas finitas.

Neste trabalho iremos usar os estimadores robustos propostos por Künsch, Stefansky e Carrol (1989) que se caraterizam por atribuir pesos menores aos *outliers*. Veremos que é possível calibrar este estimador de tal forma que dê pesos menores às observações influentes e ainda produza estimativas finitas e confiáveis. Notemos que as mesmas observações que poderiam ser rotuladas como *outliers* são aquelas responsáveis pela {não}-separação dos dados. Dar peso zero a todas as observações discrepantes pode conduzir à indeterminação do modelo. É, portanto, necessário saber quantas e quais observações poderiam ter seu peso diminuído. Para tanto propomos uma versão do algoritmo de Christmann e Rousseeuw (1999).

Os dados utilizados neste trabalho contêm variáveis referentes à situação financeira de 56 firmas brasileiras no ano de 1994, extraídas de informações contidas nas Demonstrações Contábeis e Financeiras de cada firma, as quais têm obrigação de torná-las públicas.

Utilizamos várias medidas para comparar a performance dos estimadores clássico e robusto. Dentre as medidas gráficas destacamos o Gráfico do Perfil Acumulativo da Precisão de tipo I e de tipo II. Outras medidas de performance utilizadas são a Taxa Aparente de Erro, a Razão entre Precisões e a Razão da Informação Entrópica. O procedimento robusto de estimação se mostra superior, pois além de possuir maior informação entrópica, permite atribuir pesos menores às observações atípicas enquanto preserva a superposição dos dados.

Na seção 2 revemos o modelo de regressão logística e o procedimento de estimação por máxima verossimilhança. Na seção 3 definimos alguns estimadores robustos para este modelo. Na seção 4 estudamos o problema da não existência dos estimadores de máxima verossimilhança.

Na seção 5 fazemos a análise do risco de crédito das firmas brasileiras e, finalmente, na seção 6 damos nossas conclusões.

2. Regressão logística

Considere as variáveis aleatórias $(\mathbf{x}', y)$, onde $\mathbf{x}' = (x_0, x_1, \ldots x_p)$ é o vetor de variáveis explanatórias com $x_0 \equiv 1$ e y é a resposta. O modelo linear de regressão usual assume que

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i, i = 1, \ldots, n, \tag{1}$$

onde $\mathbf{x}_i'\beta = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p$ e onde os erros ε_i são variáveis aleatórias independentes e identicamente distribuídas com esperança zero e variância constante.

Quando y assume apenas dois valores, em geral codificados como 1 ("sucesso") e 0 ("fracasso"), temos que $E[y] = \mu$ é a probabilidade de sucesso. Assim como no modelo linear usual, gostaríamos de modelar $E[y] = \mu = \mathbf{x}'\boldsymbol{\beta}$ e investigar o efeito das variáveis explanatórias na probabilidade de sucesso. Entretanto, neste caso, o modelo linear usual não parece adequado, já que os parâmetros $\boldsymbol{\beta}$ deveriam ser estimados de tal maneira que os valores ajustados $\hat{E}[y_i]$ resultassem em números restritos ao intervalo $[0,1]$. Um outro agravante é que a variância de y depende de μ, fato que também deveria ser levado em consideração pelo modelo. O modelo de regressão logística é um membro da classe dos modelos lineares generalizados especialmente idealizado para modelar dados binomiais (McCullagh e Nelder, 1989).

Um elemento fundamental na descrição de um modelo linear generalizado é a função de ligação $g(.)$ que descreve como a média $E[y]$ depende da função linear das variáveis explanatórias. No caso da regressão logística temos

$$g(\mu) = \log it(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \mathbf{x}'\boldsymbol{\beta}. \tag{2}$$

Notemos que a função de ligação $\log it$ garante que μ esteja no intervalo $[0,1]$.

O segundo elemento importante na descrição de um modelo linear generalizado é a função de variância $V(\mu)$ que tem como objetivo descrever como a variância de y dada por $\text{var}(y) = \tau V(\mu)$, onde τ é uma constante, depende da média μ. No caso da ligação $\log it$ temos que $V(\mu) = \mu(1-\mu)$.

Os estimadores de máxima verossimilhança (EMV) dos parâmetros $\boldsymbol{\beta}$ de um modelo de regressão logística são os valores em $\Re^{p+1}$ que maximizam a função

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\{y_i \log(\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}) + (1-y_i)\log(1-\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})})\}. \tag{3}$$

Uma maneira de encontrar os EMV é diferenciando (3) com respeito a $\boldsymbol{\beta}$ e igualando estas equações a zero (Hosmer e Lemeshow, 1989). A solução do sistema pode ser obtida através do método de Mínimos Quadrados Iterativamente Ponderados (Chambers e Hastie, 1992).

Os EMV podem ser alternativamente definidos como os valores que minimizam a função de dispersão

$$\sum_{i=1}^{n} D_i(\boldsymbol{\beta}), \tag{4}$$

onde as *deviances* $D_i(\boldsymbol{\beta})$ são iguais ao negativo do lado direito de (3).

3. Estimação robusta no modelo de regressão logística

3.1 Procedimentos robustos

Do ponto de vista prático, os métodos robustos tentam extrair a melhor informação possível dos dados, ajudando os analistas a identificar o padrão sugerido pela maioria das observações, identificando erros grosseiros, possíveis misturas de distribuições ou estruturas diferentes. Em particular, no problema de regressão linear, um procedimento robusto procura achar o hiperplano que melhor ajusta a maioria dos dados conseqüentemente identificando observações e/ou estruturas discrepantes, chamados de *outliers* (Mendes, 1999).

Um conjunto de dados pode possuir observações capazes de exercer grande influência sobre o ajuste. A abordagem robusta caracteriza-se por identificar se essas observações são atípicas e minimizar sua influência (por exemplo, Pregibon (1981), Johnson (1985), Hampel et al. (1986)). Contudo, conforme veremos neste trabalho, retirar observações influentes não é garantia de bons resultados no caso da regressão logística. A solução mais adequada seria usar um estimador robusto que ao mesmo tempo identificasse as observações influentes e as ponderasse para não exercerem tanta influência no ajuste.

Pregibon (1982) propôs robustificar os EMV através da aplicação de uma função ρ às *deviances*. A função ρ deveria ter taxa de crescimento inferior à da função identidade usada em (4), e os estimadores resultantes seriam obtidos através da minimização de

$$\sum_{i=1}^{n} \rho(D_i(\beta)). \tag{5}$$

Contudo esses estimadores não são consistentes, são assintoticamente viciados (Copas, 1988), e podem não ser robustos em relação aos pontos de alavanca.

Stefanski, Carroll e Ruppert (1986) obtiveram estimadores robustos de influência limitada para o modelo linear generalizado. Esses estimadores dependem de um estimador robusto para a matriz de covariâncias de $\mathbf{x}$. Carroll e Pederson (1993) mostraram que alguns desses estimadores podem ser severamente viciados.

Bianco e Yohai (1996) propuseram uma classe de M-estimadores para regressão logística que podem ser considerados como uma versão Fisher-consistente (definição em (8)) dos estimadores de Pregibon (1982). Eles são definidos como os valores que minimizam

$$\sum_{i=1}^{n}[\rho(D_i(\beta)) + G(\frac{\exp(\mathbf{x}_i'\beta)}{1+\exp(\mathbf{x}_i'\beta)}) + G(1-\frac{\exp(\mathbf{x}_i'\beta)}{1+\exp(\mathbf{x}_i'\beta)})], \tag{6}$$

onde ρ é limitada, não decrescente e diferenciável, e com $G(t) = \int_0^t \psi(-\log u)du$, onde $\psi(t) = \rho'(t)$. Eles provaram consistência e normalidade assintótica desses estimadores e quantificaram o vício assintótico sob contaminações.

Künsch it et al. (1989) propuseram os estimadores de influência limitada condicionalmente não-viciados. Esses estimadores são capazes de limitar a influência de pontos discrepantes através de pesos e serão os estimadores utilizados na análise dos dados deste trabalho.

3.2. Estimadores robustos condicionalmente não viciados de influência limitada

Os M-estimadores são definidos implicitamente por uma equação da forma

$$\sum_{i=1}^{n}\psi(y_i, \mathbf{x}_i, \hat{\beta}_n) = 0, \tag{7}$$

onde ψ é uma função assumindo valores em $\Re^{p+1}$. Por exemplo, os EMV são M-estimadores não robustos.

O estimador $\hat{\beta}_n$ é Fisher-consistente se a equação de estimação (7) for não viciada, isto é,

$$E_\beta[\psi(y|\mathbf{x},\beta)] = \int\int\psi(y|\mathbf{x},\beta)P_\beta(dy|\mathbf{x})F(d\mathbf{x}) = 0\forall\beta \tag{8}$$

onde $P_\beta(y|\mathbf{x})$ é a distribuição da resposta y dado $\mathbf{x}$ e $F(.)$ é a distribuição do vetor aleatório $\mathbf{x}$. O estimador $\hat{\beta}_n$ é condicionalmente Fisher-consistente (Künsch et al. (1989)) se satisfaz

$$E_\beta[\psi(y|\mathbf{x},\beta)|\mathbf{x}] = \int\int\psi(y|\mathbf{x},\beta)P_\beta(dy|\mathbf{x}) = 0\forall\beta\, e\, \mathbf{x}. \tag{9}$$

Na regressão linear generalizada, os estimadores de máxima verossimilhança são condicionalmente Fisher-consistentes quando a distribuição de $\mathbf{x}$ não depende de β. O conceito de consistência condicional de Fisher é expressivo, pois não depende do valor de $\mathbf{x}$ ser aleatório e se ele o for, o conceito não envolve sua distribuição. Contudo, as propriedades de um M-estimador Fisher-consistente dependem da matriz $(x_{ij}), i = 1,\ldots,n;\ j = 1,\ldots,p$ das variáveis explicativas, como ilustra a função de influência, definida a seguir.

A função de influência ou curva de influência (CI) de um M-estimador é dada por

$$CI_\psi(y_0, \mathbf{x}_0, \beta) = D(\psi,\beta)^{-1}\psi(y_0, \mathbf{x}_0, \beta$$

onde

$$D(\psi,\beta) = -\frac{\partial}{\partial \boldsymbol{\theta}} \int\int \psi(y|\mathbf{x},\boldsymbol{\theta}) P_{\beta}(dy|\mathbf{x}) F(d\mathbf{x})|\boldsymbol{\theta} = \beta .$$

A função de influência mede o efeito padronizado no estimador de uma contaminação infinitesimal em $(y_0, \mathbf{x}_0)$ e fornece a matriz de covariâncias assintóticas. Sob condições de regularidade, $n^{1/2}(\hat{\beta}_h - \beta)$ tem distribuição assintótica normal com média 0 e matriz de covariâncias $V(\psi,\beta)$, dada por:

$$\begin{aligned} V(\psi,\beta) &= E_{\beta}[CI_{\psi}(y,\mathbf{x},\beta) CI_{\psi}(y,\mathbf{x},\beta)^T] \\ &= D(\psi,\beta)^{-1} W(\psi,\beta) D(\psi,\beta)^{-T}, \end{aligned}$$

onde

$$W(\psi,\beta) = E_{\beta}[\psi(y,\mathbf{x},\beta)\psi(y,\mathbf{x},\beta)^T].$$

Os M-estimadores de influência limitada caracterizam-se por minimizarem a variância assintótica sujeitos a um limite na função de influencia. Por ser a função de influência um vetor, ela é reduzida para uma medida escalar, o coeficiente auto padronizado de sensitividade, $s(\psi)^2$, definido como

$$\begin{aligned} s(\psi)^2 &= \sup_{y,x} \sup_{\lambda \neq 0} \frac{(\lambda^T CI\psi)^2}{\lambda^T V(\psi)\lambda} \\ &= \sup_{y,x} \psi(y,\mathbf{x},\beta)^T W(\psi,\beta)^{-1} \psi(y.\mathbf{x},\beta) \end{aligned}$$

A medida de sensitividade $s(\psi)^2 \geq p$ mede a influência máxima que uma observação pode ter sobre uma combinação linear dos parâmetros, padronizada pelo desvio padrão desta combinação.

Os M-estimadores condicionalmente não viciados, de influência limitada (veja detalhes da definição em Künsch et al. (1989)) satisfazem (9), minimizam $V(\psi,\beta)$ e ao mesmo tempo possuem $s(\psi) \leq b$, onde $b > 0$ é a constante do M-estimador de Huber (Hampel et al., 1986). Estes estimadores são a solução de um sistema de equações que pode ser visto em Marazzi (1993). A constante b está relacionada com o valor dos pesos atribuídos às observações influentes. Quando $b \uparrow$ os pesos $\uparrow 1$ e o procedimento robusto aproxima-se dos EMV baseado no conjunto completo de dados.

No caso da regressão logística os pesos ajudam a identificar os *outliers*. Entretanto, pode acontecer que observações rotuladas como *outliers* sejam responsáveis pela não-separação dos dados. Neste caso, dar peso zero a todas as observações discrepantes pode conduzir à indeterminação do modelo. Usamos a constante b

como uma constante reguladora dos pesos e para diagnóstico do modelo. A estratégia é escolher um valor inicial grande para b e ir diminuindo este valor gradativamente, conforme ilustraremos na seção 5.

4. Sobre a não-existência dos EMV

O problema da existência e unicidade das estimativas de máxima verossimilhança em modelos log-lineares foi estudado por vários autores, inclusive Anderson (1972), Haberman (1974) e Wedderburn (1976). O problema da não-existência dos mesmos, abordado em Silvapulle (1981), foi investigado por Albert e Anderson (1984) e Santner e Duffy (1986), que mostraram que os estimadores de máxima verossimilhança dos coeficientes do modelo de regressão logística existem (são finitos e únicos) somente se há superposição dos dados. Lesaffre e Albert (1989) estudaram a não existência dos EMV para o caso de mais de dois grupos e mostraram que neste caso pode ocorrer o que definem como separação parcial.

Assim, no caso da regressão logística, podem ocorrer três categorias mutuamente exclusivas e exaustivas de configurações dos pontos na amostra: separação completa, separação quase completa e superposição, que irão resultar na existência ou não-existência dos estimadores de máxima verossimilhança (Albert e Anderson, 1984).

Diz-se que o conjunto de dados é completamente separado quando existe um vetor $\beta \in \Re^{p+1}$ tal que:

$$\mathbf{x}_i'\beta > 0 \text{ se } y_i = 1 \quad e \quad \mathbf{x}_i'\beta < 0 \text{ se } y_i = 0 \tag{10}$$

para $i = 1,\ldots,n$. Isto significa que existe um vetor β que aloca corretamente todas as observações aos seus grupos.

Um conjunto de dados é quase completamente separado se existe um vetor $\beta \in \Re^{p+1} \setminus \{0\}$ tal que:

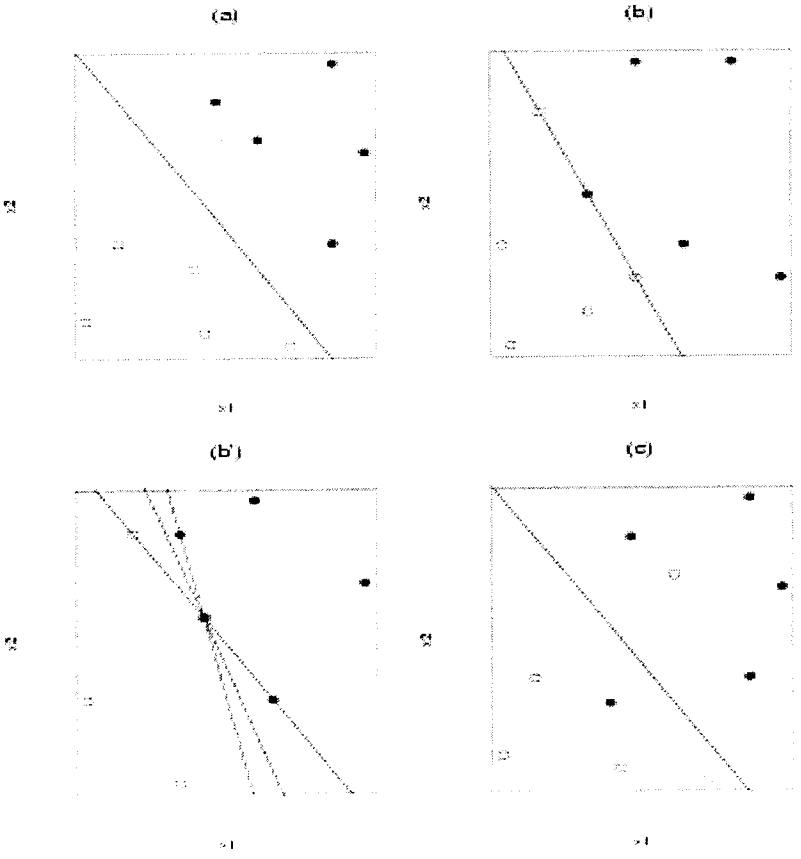
$$\mathbf{x}_i'\beta \geq 0 \text{ se } y_i = 1 \quad e \quad \mathbf{x}_i'\beta \leq 0 \text{ se } y_i = 0 \tag{11}$$

para todo $i = 1,\ldots,n$ e se existe $j \in \{1,\ldots,n\}$ tal que $\mathbf{x}_j'\beta = 0$. Neste caso, teremos uma submatriz de linhas $\mathbf{x}_i'$ linearmente dependentes pertencendo a um subespaço linear de dimensão $d \leq p-1$. De fato, o posto da submatriz é igual a $d+1$. Se $d = p-1$ o hiperplano de separação é único.

Um conjunto de dados apresenta superposição quando não existe separação completa ou separação quase completa.

A Figura 1 (de Albert e Anderson, 1984) ilustra essas três situações quando $p = 2$. As observações do grupo "0" são representadas por círculos vazios e as do grupo "1" por círculos cheios. Na figura (a) temos o caso de separação completa. Separação quase completa acontece em (b) e (b'). Em (b) temos $d = 1$, e o hiperplano de separação é único. Em (b') o hiperplano de separação não é único e na interseção das três linhas existem três observações, uma de "0" e duas de "1". Em (c) vemos a superposição.

Figura 1 - Possíveis configurações de pontos na amostra no caso de duas variáveis, x_1 e x_2. Na figura, os "0"'s são representados pelos círculos vazios e os "1"'s são representados pelos círculos cheios (a) Separação completa, (b) Separação quase completa, (b') Separação quase completa: no ponto de interseção das três linhas, existem três observações, uma de "0" e duas de "1", e (c) Superposição



Quando não existe superposição, a solução do processo iterativo não é única e se localiza no limite (infinito) do espaço paramétrico. Se o conjunto de dados é deste tipo o que se pode inferir? Do ponto de vista robusto, e de acordo com a definição de ponto de ruptura (Mendes, 1999), temos uma situação onde teria ocorrido *breakdown* (ruptura) do estimador de máxima verossimilhança. De acordo com Albert e Anderson (1984), estas estimativas apenas indicam a ocorrência da separação ou quase separação dos dados. Neste caso, dificuldades computacionais devem ser esperadas no processo iterativo na procura dos EMV. Em geral, as técnicas de otimização numérica não detectam este problema e continuam iterando até que o número limite de iterações seja atingido. É preciso fazer algum tipo de verificação numérica deste processo iterativo.

Uma abordagem baseada em técnicas de programação linear (Albert e Anderson, 1984) consiste em observar as estimativas dos desvios- padrões dos EMV para detectar a ocorrência do problema. Esta abordagem está implementada no software SAS (Statist. Sc. Inc.(1995)) e baseia-se no fato de que qualquer método de

convergência usado para maximizar a log-verossimilhança deve resultar numa solução que acarreta na separação completa, se esta solução existe. (Para uma comparação entre os procedimentos implementados por vários softwares veja Heinze, 1999). No processo de otimização, se a convergência é atingida em oito ou menos iterações, nenhum teste é feito para detectar separação completa ou quase completa. Após a oitava iteração, a probabilidade de resposta observada é calculada para cada observação. Se esta probabilidade é igual a um para todas as observações, conclui-se que há separação completa dos dados e o processo iterativo é interrompido. Se a separação completa dos dados ainda não foi determinada e identifica-se uma observação com uma probabilidade de resposta observada muito grande (≥ 0.95), há dois casos a serem considerados. Primeiramente, existe separação completa no conjunto de dados, e a observação em questão é um ponto atípico dentro do grupo a que pertence. O processo iterativo, se não impedido de continuar, irá parar quando um máximo for alcançado. No segundo caso, existe separação quase completa e a matriz de dispersão assintótica é ilimitada. Se algum dos elementos da diagonal da matriz de dispersão para os vetores de observações padronizado (todas as variáveis explanatórias são padronizadas para média zero e variância um) exceder 5 000, a separação quase completa é declarada e o processo iterativo é interrompido.

A separação ou quase separação dos dados pode também ocorrer no processo de obtenção de estimativas robustas. Muitos estimadores robustos são construídos de forma que pontos discrepantes sejam excluídos ou recebam pesos menores. Entretanto, pode acontecer que o conjunto com todos os dados apresente superposição, mas o conjunto de dados reduzido não. Nesta situação, o estimador robusto aplicado ao conjunto de dados reduzido não existe (Künsch, Stefansky e Carrol, 1989). É, portanto, necessário saber quantas e quais observações poderiam ter seu peso diminuído. O problema da identificação dessas observações foi investigado por Albert e Anderson (1984) e está relacionado com os conceitos dados a seguir.

Denotamos por n_{spp} o menor número de observações cuja retirada destrói a superposição dos dados. Num modelo de regressão logística, o n_{spp} é o menor número de observações que precisa ser retirado para tornar a estimativa de máxima verossimilhança inexistente. Do mesmo modo, denotamos por n_{comp} o menor número de observações cuja retirada resulta em separação completa. Por definição, sempre temos que $n_{spp} \leq n_{comp}$.

Voltando à Figura 1, notemos que em (a) $n_{spp} = n_{comp} = 0$. A estimativa de máxima verossimilhança de β não existiria nesse caso, devido à separação completa. Em (b) $n_{spp} = n_{comp} = 3$. Em (b') $n_{spp} = 0$ e $n_{comp} = 1$. Em (c) $n_{spp} = n_{comp} = 2$.

Fica claro que para dimenções maiores de $\mathbf{x}$ torna-se complicado determinar visualmente a separação completa dos dados e para isso usaremos um algoritmo desenvolvido por Christmann e Rousseeuw (1999) e modificado por nós. Este algoritmo obtém a combinação linear $\mathbf{x}_i\boldsymbol{\mu}$ e fornece apenas as quantidades n_{comp} ou n_{spp}. Com a nossa extensão, identificamos quais são as (n_{comp}) observações que se retiradas ao mesmo tempo resultam na separação completa dos dados (a função do S-Plus pode ser obtida das autoras por e-mail). Pode

acontecer que para um mesmo n_{comp}, tenhamos subconjuntos diferentes de observações. Isto é, o conjunto de observações a ser retirado de forma a resultar na separação completa dos dados pode não ser único.

5. Risco de crédito de firmas brasileiras

Nesta seção, motivados pelo problema da análise do risco de crédito de firmas brasileiras, comparamos os procedimentos de estimação clássico e robusto para o modelo logístico e investigamos o efeito da configuração dos dados nas soluções obtidas.

Os dados utilizados contêm medidas referentes à situação financeira de 56 firmas brasileiras no ano de 1994, extraídas de informações (públicas e que podem ser encontradas em jornais) contidas nas suas Demonstrações Contábeis e Financeiras (Balanço Patrimonial, Demonstração do Resultado do Exercício, Demonstrações de Lucros e Prejuízos Acumulados). Em geral, os analistas de crédito consideram essas variáveis como bastante informativas para se avaliar a saúde financeira de firmas.

As 56 empresas analisadas estão enumeradas a seguir. As empresas inadimplentes estão em negrito.

(1) ACESITA
(2) A/c COS VILLARES CONS.
(3) ALCATEL
(4) ARACRUZ
(5) AUTOLATINA
(6) BOMPREÇO
(7) BURI
(8) CACIQUE CONS.
(9) CAEMI.CONS.
(10) CAMARGO CORRÊA
(11) CARGILL
(12) CBPO
(13) CCE CONSOLIDADO
(14) CNO
(15) CONFAB
(16) CONSTRAN
(17) COPAS
(18) COPENE
(19) COPESUL
(20) CORAL
(21) CPC ONSOLIDADO
(22) FERRO-LIGAS CONS.
(23) CST
(24) ELDORADO
(25) ELETROPAULO
(26) ELIZABETH IND. TÊXTIL
(27) ELUMA CONS.
(28) EMBRAER}
(29) ENCOL CONS.
(30) ESCELSA
(31) ESTEVES IRMÃOS
(32) FERTIBRÁS
(33) FIBRA
(34) FOSFÉRTIL
(35) FRIGOBRÁS
(36) GERDAU CONS.
(37) GRADIENTE ELET. CONS.
(38) IKPCCONS
(39) IOCHPE-MAXION CONS.
(40).**ITAP CONS.**
(41) LACTA
(42) LOJAS AMER. CONS.
(43) LATASA
(44) MANNESMANN
(45) MAPPIN CONS.
(46) MARCOPOLO
(47) MESBLA CONS.
(48) MINASGÁS
(49) MONTREAL EMP. CONS.}
(50) PARANAPANEMA
(51) PERDIGÃO CONS.
(52) PERNAMBUCANAS RJ
(53) RHODIA
(54) SALGEMA
(55) SENDAS
(56) TROMBINI

As informações selecionadas dos balanços foram:

- pl : patrimônio líquido (em US$Mil)
- ap : ativo permanente (em US$Mil)
- ll : lucro líquido (em US$Mil)
- lo : lucro operacional (em US$Mil)
- $deprec$: depreciação (em US$Mil)
- rol : receita operacional líquida (em US$Mil)
- $eblcp$: empréstimo bancário líquido de curto prazo (em US$Mil)
- $ebllp$: empréstimo bancário líquido de longo prazo (em US$Mil)
- $cccppas$: conta com controladas/coligadas de curto prazo de passivo (em US$Mil) (Transferências que a empresa faz às empresas que a têm como controlada ou coligada)
- $cccppat$: conta com controladas/coligadas de curto prazo de ativo (em US$Mil) (Transferências que a empresa recebe de suas controladas ou coligadas)
- $afincp$: aplicações financeiras de curto prazo (em US$Mil)

Com estas informações foram calculados os oito índices geralmente usados pelos analistas de crédito, que são:

1ª variável explicativa: $x_1 = \frac{pl}{ap}$; participação do capital próprio no ativo permanente, ou seja, como o capital próprio financia o ativo permanente.

2ª variável explicativa: $x_2 = \frac{ll}{pl}$; coeficiente do retorno sobre o patrimônio líquido.

3ª variável explicativa: $x_3 \frac{(lo+deprec)}{rol}$; parte da receita operacional líquida que seria o lucro operacional real (lo + deprec).

4ª variável explicativa: $x_4 \frac{(eblcp+ebllp)}{pl}$; quanto do patrimônio líquido está comprometido com endividamento bancário.

5ª variável explicativa: $x_5 \frac{(eblcp+ebllp)+cccppas-cccpat}{pl}$; é x_4 considerando o resultado das contas com coligadas/controladas.

6ª variável explicativa: $x_6 \frac{eblcp+(cccppas-cccpat)}{pl}$; é x_5 somente para o curto prazo.

7ª variável explicativa: $x_7 \frac{(eblcp+afincp)+(cccppas-cccpat)}{pl}$; x_6 deduzidas as aplicações financeiras de curto prazo.

8ª variável explicativa: $x_8 = \frac{eblcp}{rol}$; participação do endividamento bancário líquido de curto prazo na receita operacional líquida.

Em resumo, os três primeros índices refletem como a empresa se financia e os cinco seguintes o endividamento bancário.

Inicialmente, foi ajustado um modelo completo contendo todas as variáveis explicativas. Tanto para o ajuste clássico quanto para o ajuste robusto poucas variáveis foram significativas. Contudo, vale a pena lembrar que existe a possibilidade da empresa ter "maquiado" seu balanço, o que acarretaria em análises bastante equivocadas. Este problema só poderia ser contornado com a validação ou correção das informações por meio de auditoria, e não iremos tratar dele aqui. Além disto, observamos uma correlação alta entre algumas variáveis, se calculadas pelo método clássico (covariância amostral) ou pelo método robusto (elipsóide de volume mínimo, Rousseeuw e Leroy, 1987). Por exemplo, a correlação entre as variáveis x_4 e x_5 é 0.991 pelo método clássico e 0.999 pelo método robusto.

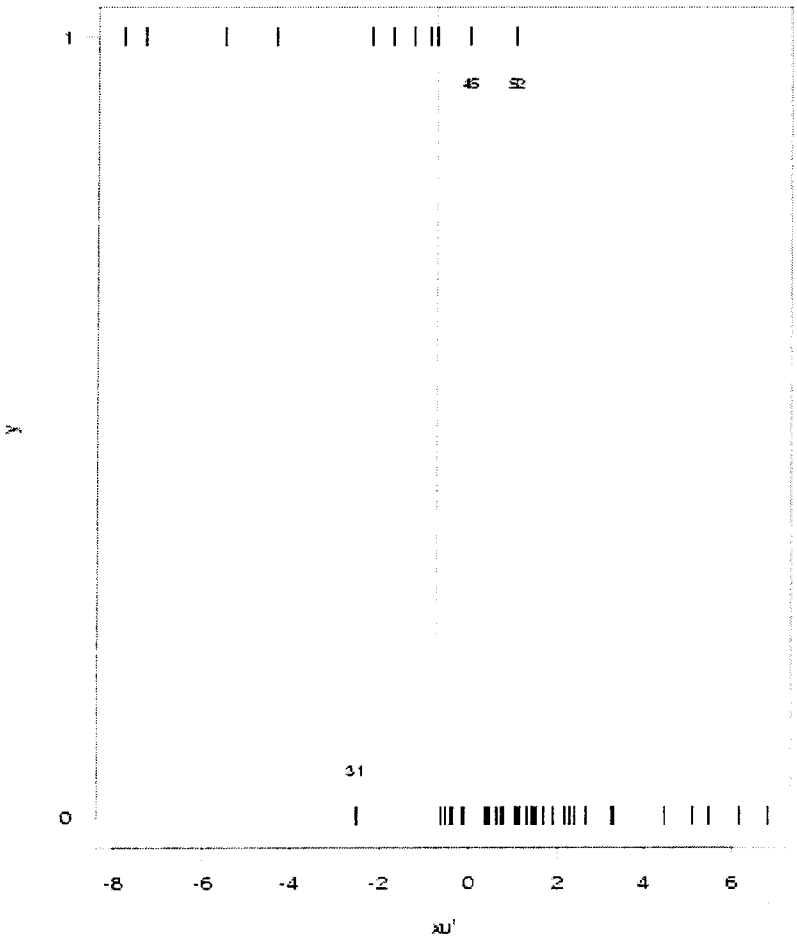
A partir de um subconjunto básico dos dados, formado por variáveis consideradas indispensáveis pelos economistas consultados e, levando-se em conta o problema da multicolinearidade, foram testados todos os outros modelos possíveis. O modelo reduzido final foi

$$\log it(\mu_i) = \beta_0 + \beta_2 \mathbf{x}_{2i} + \beta_3 \mathbf{x}_{3i} + \beta_4 \mathbf{x}_{4i} + \beta_8 \mathbf{x}_{8i}. \qquad (12)$$

O ajuste clássico resultou nas seguintes estimativas (desvios padrões): $\hat{\beta}_0 = -0.781(0.719)$, $\hat{\beta}_2 = -12.564(5.998)$, $\hat{\beta}_3 = -15.761(7.482)$, $\hat{\beta}_4 = -5.489(3.028)$ e $\hat{\beta}_8 = -22.9(14.618)$. O valor da deviance é 24.88, menor que o valor esperado de uma qui-quadrado com 50 graus de liberdade, e, portanto, não traz suspeitas sobre a adequacidade do modelo. Excetuando-se o intercepto, todas os outros coeficientes são significantes ao nível de 10%.

Em seguida, utilizando a extensão que fizemos do algoritmo de Christmann e Rousseeuw (1999), identificamos as (n_{comp}) observações influentes que se retiradas em conjunto, resultam na separação completa dos dados. O (n_{comp}) encontrado foi igual a 3 e os subconjuntos contendo os índices das observações responsáveis pela separação são {31, 45, 52}, {17, 31, 52} e {29, 45, 52}. Observamos que o índice 52, referente à firma Pernambucanas, está presente em todos os subconjuntos encontrados. De fato, quando retiramos um desses subconjuntos de observações e ajustamos o modelo, tanto pelo método clássico quanto pelo método robusto, notamos instabilidade numérica dos algoritmos, o modelo fica indeterminado, apresentando desvios padrões extremamentes inflados ou pequenos, troca de sinal dos coeficientes, deviance igual a zero, etc.

Figura 2 - Plot de $\mathbf{y}$, *versus* a combinação linear $\mathbf{x\mu}'$ com $\boldsymbol{\mu}$ fornecendo o menor $\mathbf{n}_{comp}$ para os dados

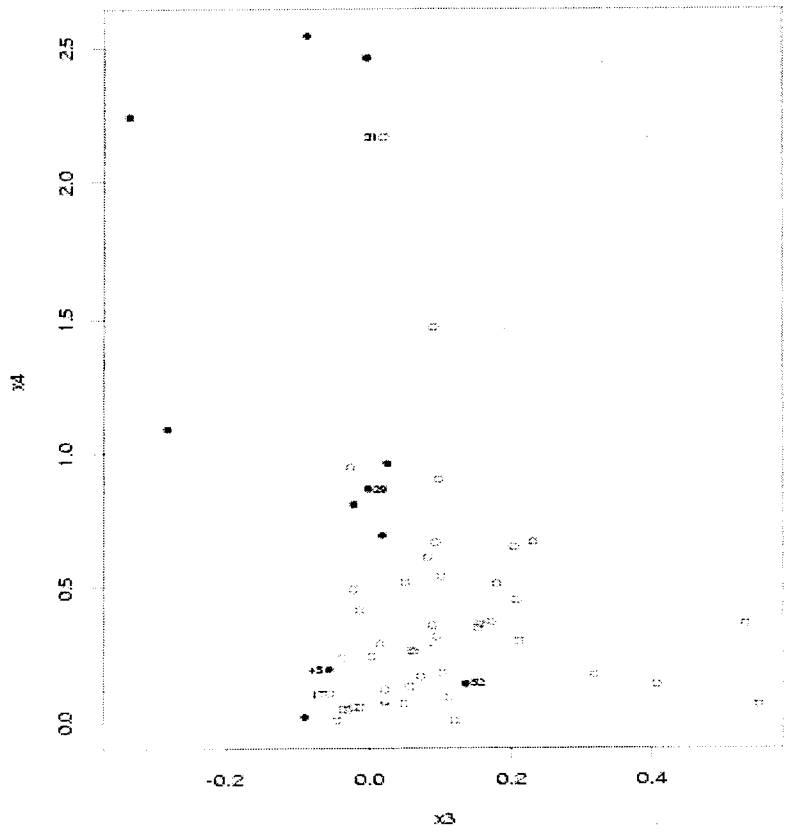


A Figura 2 foi construída a partir do *output* do programa utilizado. No eixo vertical temos a resposta y_i e no horizontal temos a combinação linear $\mathbf{x\mu}'$ onde $\boldsymbol{\mu}$ é a direção que permitiu a identificação dos subconjuntos. Observe que a linha tracejada separa completamente os "0"'s dos "1"'s, a menos das observações de índices $\{31, 45, 52\}$.

De fato, as observações 31, 45 e 52 além de influentes são outliers. Isto pode ser visto na Figura 3 que mostra a dispersão entre as variáveis x_3 e x_4. As firmas não inadimplentes (inadimplentes) estão representadas como círculos vazios (cheios). Nota-se nesta figura que a firma Esteves Irmãos (31) é um outlier na direção de

x_4. Já as observações 17 e 29 são apenas influentes. Veremos mais tarde que o método clássico as classifica erroneamente e o robusto não.

Figura 3 - Gráfico de dispersão entre as variáveis $\mathbf{X}_3$ e $\mathbf{X}_4$



Iremos agora usar o procedimento robusto para estimar os parâmetros do modelo (12). Conforme já vimos, este estimador dá pesos menores aos outliers e ainda permite um certo controle sobre esta atribuição de pesos, já que a mesma se reduz à escolha do valor apropriado da constante reguladora b.

Na escolha do melhor valor de b, devemos começar com um valor arbitrário, alto, e diminuir este valor passo-a-passo, examinando os pesos atribuídos às observações de modo que nenhum dos subconjuntos detectados seja retirado, e de tal forma que as estimativas dos desvios padrões não inflem demasiadamente. Além disto, sabemos que $b > \sqrt{4}$ já que $p = 4$. A Tabela 1 exemplifica este processo, onde a escolha final para b foi $b = 3.45$.

As estimativas (desvios-padrão) fornecidas pelo procedimento robusto (Marazzi, 1997) regulado com $b = 3.45$ foram: $\hat{\beta}_0 = -2.24(1.275)$, $\hat{\beta}_2 = -11.58(7.114)$, $\hat{\beta}_3 = -30.46(19.316)$, $\hat{\beta}_4 = -6.91(3.871)$ e $\hat{\beta}_8 = -15.98(15.298)$.

O valor da deviance é 19.74 indicando adequacidade do modelo. As variáveis são significativas, porém à níveis um pouco mais altos que na estimação clássica. Entretanto, pelas razões já citadas, preferimos manter todas as variáveis no modelo, mesmo porque a esta altura já sabemos que o conjunto de dados é problemático no sentido de que aquelas observações responsáveis pela superposição também são *outliers*, o que provoca um aumento no valor dos desvio- padrões dos estimadores.

Tabela 1- Pesos atribuídos aos *outliers*

Valor de *b*	17	29	31	45	52	outras
5	x	x	x	x	x	x
4	x	x	x	x	0,5997	x
3,5	x	x	x	0,7160	0,3934	x
3,45	x	x	0,1700	0,3578	0,1445	x
3,4	x	x	0,1179	0,2436	0,0990	x
3,3	x	x	0,0092	0,0178	0,0077	x
3,2	x	x	0,0009	0,0016	0,0008	x
3,1	x	x	0,0004	0,0006	0,0003	x
3	x	x	0,0002	0,0003	0,0002	x

A seguir, verificaremos a qualidade dos ajustes obtidos com as estimações clássica e robusta do modelo (12), os quais, por simplicidade, chamaremos de modelo clássico e de modelo robusto. Começaremos com a análise gráfica da Figura 4.

A Figura 4 mostra os resíduos do modelo clássico (lado esquerdo) e robusto (lado direito). Notamos que o estimador robusto ajusta bem a maioria das firmas, destacando mais os *outliers*, observações de índices 31, 45 e 52 (firmas Esteves Irmãos, Mappin e Pernambucanas, respectivamente).

Neste processo de validação e comparação de modelos, o ideal seria usar parte do conjunto de dados para estimação (o conjunto de treinamento) e reservar outra parte para validação. Contudo, conforme ocorre geralmente, não temos um número suficiente de firmas inadimplentes para compor os dois subconjuntos. Independentemente do modelo e método de estimação utilizado, modelos de previsão de inadimplência podem errar de duas maneiras, gerando os erros de tipo I e de tipo II.

Dizemos que se comete um erro de tipo I quando o modelo indicar um risco baixo (digamos, a probabilidade estimada de inadimplência $\hat{p} < 0.5$) quando, de fato, o risco é alto (a resposta correspondente é "1", inadimplência). Esta é a situação correspondente à clientes de cotação alta na praça que, contudo, não irão cumprir suas obrigações contratuais. Por outro lado, o modelo pode indicar um risco alto ($\hat{p} < 0.5$) quando de fato ele é baixo, o que seria um erro de tipo II. Um bom modelo deve balancear o número de erros cometidos dos

dois tipos e ao mesmo tempo diferenciar o risco relativo de crédito através do espectro completo das medidas de qualidade de crédito de todos os credores.

Consideremos uma regra de classificação para as firmas baseada nas probabilidades estimadas de inadimplência $\hat{p}$ e com ponto de corte 0.5 (para um estudo do efeito do ponto de corte numa regra de classificação para dados binomiais veja Mendes, 1988). O ajuste clássico resulta em uma Taxa Aparente de Erro (*TAE*) (número total de classificações errôneas dividido pelo tamanho da amostra) de 6/56. O ajuste robusto resulta em apenas 3/56 de classificações erradas. Este resultado pode ser visualizado na Figura 5, que mostra os valores ajustados pelos EMV (círculos) e os estimadores robustos (triângulos). Como anteriormente, os símbolos cheios representam os "1" 's e os símbolos vazios os "0" 's.

Figura 4 - Resíduos clássicos (esquerda) e robustos (direita)

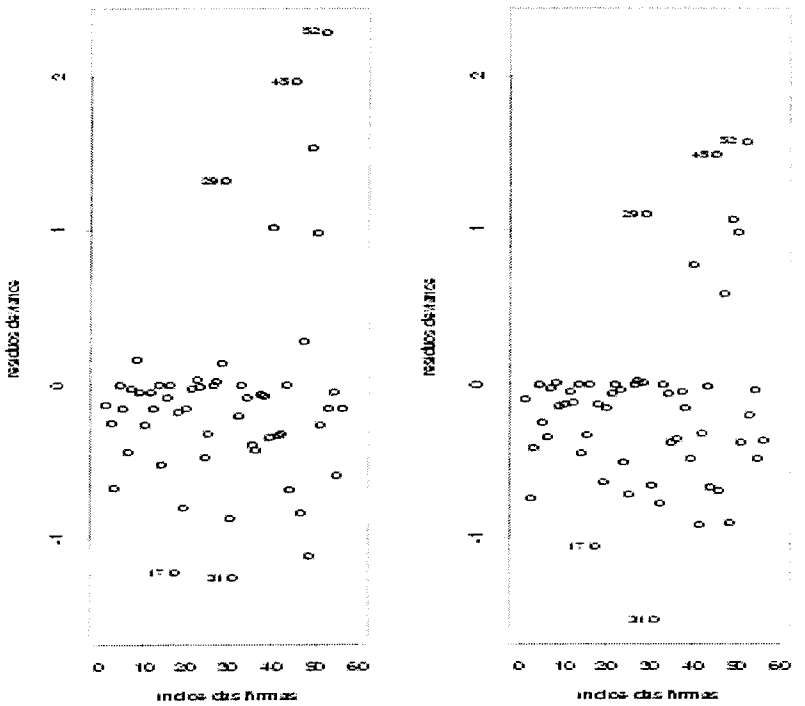
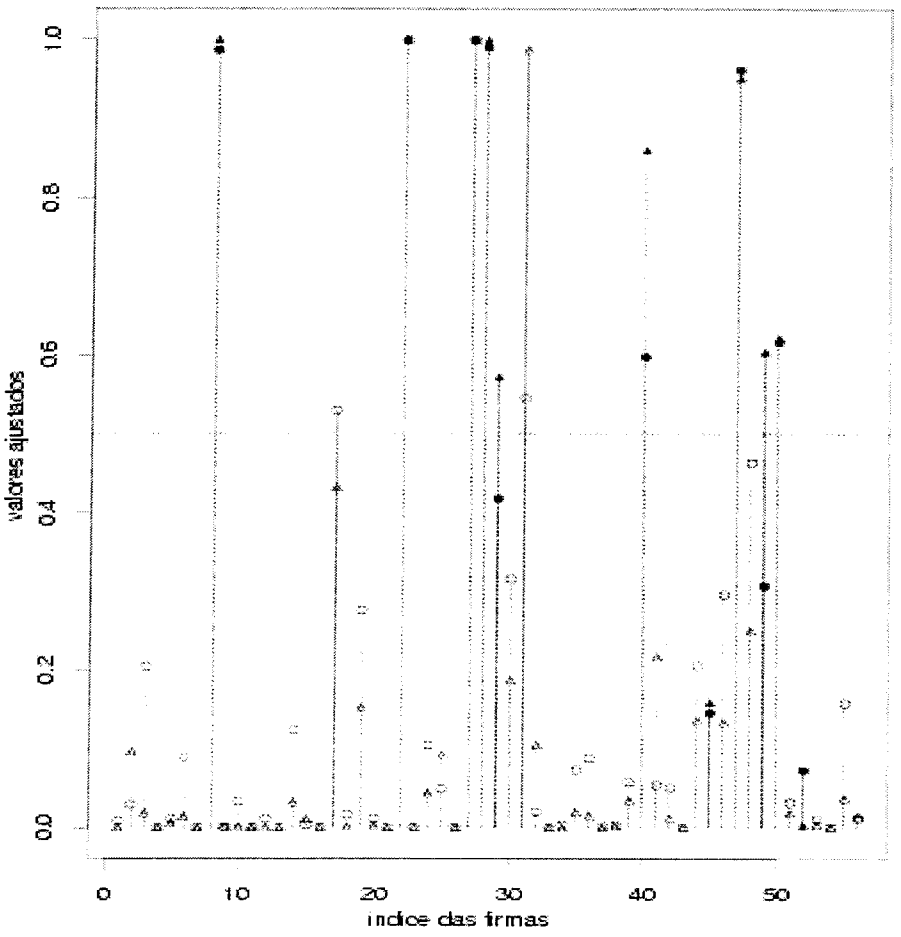


Figura 5 - Valores ajustados pelos EMV (círculos) e os estimadores robustos (triângulos). Símbolos cheios representam os "1"'s. Os vazios representam os "0"'s



A *TAE* assume implicitamente que os dois tipos de erro são igualmente importantes. Podemos também calcular a *TAE* de tipo I (*TAEI*) e de tipo II (*TAEII*). Pode-se ver na Figura 5 que duas firmas tipo "1" (Encol, 29, e Montreal, 49) foram erroneamente classificadas como adimplentes pelo ajuste clássico, enquanto o robusto as classificou corretamente como "1". Além disso, uma observação tipo "0" (Copas, 17) foi erroneamente classificada como inadimplente pelo ajuste clássico, enquanto o robusto a classificou corretamente como "0". Ambos os métodos classificaram erroneamente as firmas Esteves Irmãos, Mappin e Pernambucanas. Assim, vemos que a *TAEI* e a *TAEII* para os modelos clássico e robusto foram respectivamente (4/11, 2/11) e (2/45, 1/45), e notamos que o modelo robusto foi superior em relação aos dois tipos de erro. É interessante notar que as

outras duas observações responsáveis pela superposição (17 e 29), e portanto influentes, foram corretamente classificadas pelo procedimento robusto por não serem *outliers*.

As probabilidades de inadimplência estimadas $\hat{p}$ podem ser utilizadas para ordenar os credores quanto ao risco que oferecem, na medida que elas representam uma medida de quanto uma firma possui as características (valores das variáveis explicativas) de uma firma inadimplente. Duas medidas gráficas de performance, baseadas na ordenação dos $\hat{p}$, são as curvas *Perfil Acumulativo da Precisão* - PAP - e a *Razão da Informação Entrópica* - RIE. Essas medidas, usadas por Keenan e Sobehart (2000) para comparar modelos de risco de crédito, também indicam se existe ou não redundância na informação.

A curva PAP coloca no eixo horizontal a fração $x\%$ do total de firmas ordenadas pelo seu risco ($\hat{p}$). A curva PAP de tipo I tem no eixo vertical o percentual $y(x)$ de inadimplentes cujo $\hat{p}$ é maior ou igual que o $\hat{p}$ da fração x correspondente. A curva PAP do tipo II tem no eixo vertical o percentual $z(x)$ dos inadimplentes. Assim, a curva PAP do tipo I representa a fração acumulativa de inadimplentes para diferentes percentuais na escala do risco, e a de tipo II é o seu complemento. Notemos que no caso de alocação completamente aleatória (modelo não informativo) essas duas curvas coincidiriam com a reta $y(x) = x$.

Uma propriedade muito útil de uma curva PAP é a de que ela revela informação a respeito da capacidade preditiva do modelo ao longo de todo o intervalo [0,1]. A Figura 6 mostra a curva PAP para os dois modelos. A linha vertical (pontuada) representa o percentual de firmas inadimplentes na amostra. As duas retas sólidas representam a situação ideal quanto aos erros de tipo I e II. A linha tracejada e a pontuada são respectivamente as curvas PAP para os modelos clássico e robusto. Podemos observar que os dois modelos se alternam ao longo do $[0,1]$ na tarefa de melhor discriminar as firmas. Este fato é confirmado pelo cálculo de uma outra medida, a *Razão entre Precisões* (*RP*), definida como

$$RP = \frac{1}{1-f}(2\int_0^1 y(x)dx - 1) = \frac{1}{f}(1 - 2\int_0^1 z(x)dx)$$

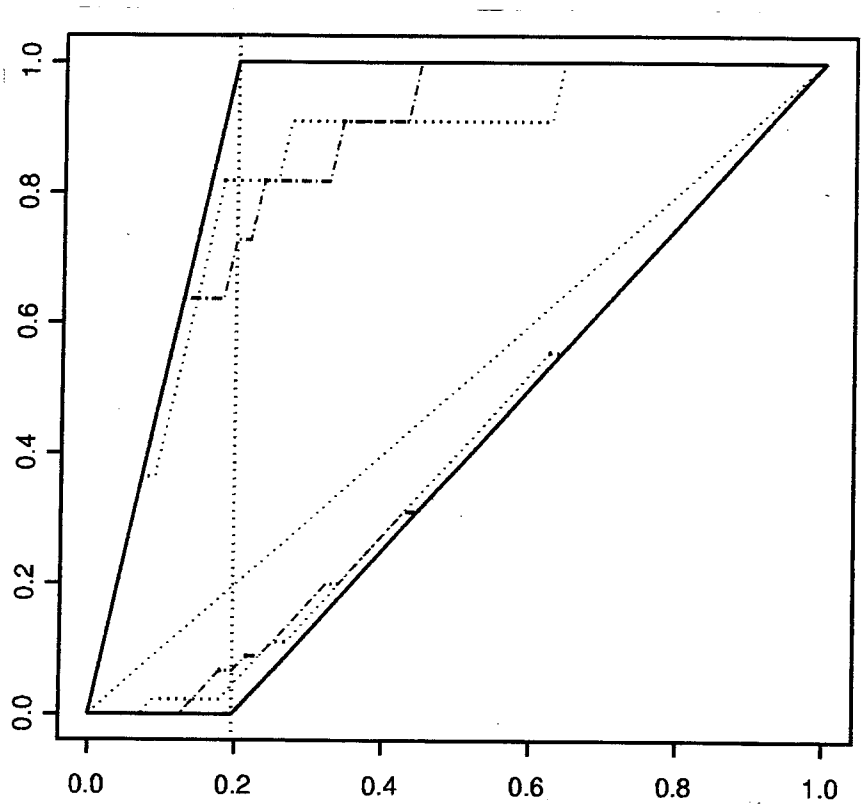
onde f é a fração de inadimplentes. Geometricamente, *RP* é igual à razão entre duas áreas: a área entre a curva PAP obtida e a reta $y = x$, e a área entre a curva PAP ideal e a reta $y = x$. Um modelo com *RP* perto de zero quase não possui vantagem sobre um modelo que assinalasse aleatoriamente as probabilidades $\hat{p}$. Nos casos dos modelos aqui comparados obtivemos os valores 0.9010 e 0.8808 para as curvas de tipo I e modelos clássico e robusto respectivamente, e os valores 0.7878 e 0.7677 no caso do tipo II, indicando uma superioridade muito pequena do modelo clássico quanto à capacidade de discriminar ao longo da amostra. Mas existiria alguma diferença entre esses modelos quanto à incerteza inerente aos mesmos? A medida utilizada a seguir examina a qualidade de modelos sob esse aspecto entrópico.

A informação entrópica é uma medida que resume o total de incerteza que o conjunto de $\hat{p}'s$ de um modelo representa. Em primeiro lugar notemos que a informação "extra", adicional, que um banco precisa, e que não é fornecida pelo modelo (ou pelos $\hat{p}'s$) pode ser definida como

$$-\log_2(\hat{p}),$$

onde $\log_2$ é a função logaritmo na base 2. Quando $\hat{p}$ é 1 (ou 0), temos certeza do evento, e não haveria nenhuma informação relevante que não se tivesse conhecimento. Por outro lado, se $\hat{p}$ é 1/2, a informação requerida é a máxima possível.

Figura 6 - As curvas PAP de tipo I e de tipo II para os modelos clássico (tracejada) e robusto (pontuada)



A informação entrópica de um evento com probabilidade p é definida como

$$H_0 = -[p\log_2(p) + (1-p)\log_2(1-p)], \tag{13}$$

a qual é máxima quando $p = 1/2$, o que significa o estado de absoluta ignorância.

A informação entrópica de um modelo que fornece a estimativa $\hat{p}_j$ para a firma j é definida como

$$IE = \frac{1}{n}\sum_{j=1}^{n} h_j , \tag{14}$$

onde

$$h_j = -[\hat{p}_j \log_2(\hat{p}_j) + (1-\hat{p}_j)\log_2(1-\hat{p}_j)] \tag{15}$$

Vemos então que um modelo que carregue mais informação sobre a ocorrência ou não de um conjunto de eventos (aqui, inadimplência de firmas) terá IE menor que o de um outro modelo contendo menos informação. Os modelos clássico e robusto produziram, respectivamente, os valores *IE* iguais a 0.3205 e 0.2319, onde vemos uma maior redução da incerteza do modelo robusto traduzida pela redução da entropia. Para compararmos modelos usamos a medida *Razão da Informação Entrópica - RIE* definida como

$$RIE = 1 - \frac{IE}{H_0}, \tag{16}$$

onde H_0 é a informação entrópica da amostra, isto é, a expressão (13) avaliada quando p é igual à proporção de inadimplentes na amostra. Um modelo com *RIE* igual a zero teria nenhum poder preditivo. Por outro lado, um modelo que fizesse previsões perfeitas teria *RIE* igual a 1. Os modelos clássico e robusto produziram, respectivamente, os valores da medida *RIE* iguais a 0.5515 e 0.6756, confirmando que sob este aspecto o modelo robusto é melhor.

Finalmente, a Tabela 2 resume os resultados das comparações dos modelos sob os diferentes aspectos. Notamos de uma maneira geral uma superioridade do modelo robusto, que conta ainda com as vantagens de identificar os pontos atípicos, atribuir pesos menores aos mesmos, e permitir o controle para que não ocorra o *breakdown* dos estimadores.

Tabela 2 - Resumo dos resultados das comparações dos modelos

Medida de Qualidade de Ajuste		M. Clássico	M.Robusto
TAE:	Taxa Aparente de Erro		√
TAEI:	Taxa Aparente de Erro Tipo I		√
TAEII:	Taxa Aparente de Erro Tipo II		√
RP:	Razão entre Precisões	√	
RIE: :	Razão da Informação Entrópica		√

6 . Conclusões

Risco de Crédito pode ser definido como o risco resultante da incerteza de que um credor irá cumprir com suas obrigações contratuais. Neste trabalho usamos o modelo de regressão logística para estimar a probabilidade de inadimplência de firmas brasileiras.

Estudamos o problema da não existência dos estimadores de máxima verossimilhança para este conjunto de dados à luz dos conceitos de separação completa, separação quase completa e superposição dos dados. Construímos um algoritmo para identificar as observações que, se retiradas ao mesmo tempo, resultam na separação completa dos dados. A função do S-PLUS pode ser obtida das autoras por e-mail.

Concluímos que o conjunto de dados é problemático no sentido de que aquelas observações responsáveis pela superposição também são *outliers*, o que provoca um aumento no valor dos desvios-padrões dos estimadores. O procedimento robusto forneceu melhores resultados, uma vez que aplicou pesos menores mas não excluiu estes pontos.

Vale ressaltar que existem outros modelos que podem ser testados. Uma sugestão seria utilizar outras funções de ligações para dados binários, como a probit. Poderíamos também fazer transformações nos dados, como calcular as componentes principais.

Utilizamos várias medidas para comparar a performance dos estimadores clássico e robusto. Dentre as medidas gráficas destacamos o Gráfico do Perfil Acumulativo da Precisão de tipo I e de tipo II. Outras medidas de performance utilizadas foram a Taxa Aparente de Erro, a Razão entre Precisões e a Razão da Informação Entrópica. O procedimento robusto de estimação se mostrou superior pois além de possuir maior informação entrópica, permitiu atribuir pesos menores às observações atípicas ao mesmo tempo que preservava a superposição dos dados.

Entendemos que a contribuição principal deste artigo foi a de aliar conceitos e procedimentos relacionados com a existência dos estimadores de máxima verossimilhança no modelo de regressão logística e métodos robustos, tendo como motivação a análise do risco de crédito de firmas brasileiras.

Referências bibliográficas

ALBERT,A., E ANDERSON, J.A. (1984); On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. ***Biometrika*** **71, 1-10.**

ALMEIDA, F. C. E DUMONTIER, P. (1996); O uso de Redes Neurais na Avaliação de Riscos de Inadimplência. ***Revista de Administração FEA/USP*** **31, 1, 52-63.**

ALTMAN, E. (1968); Financial Ratios, Discriminant Analysis and the Prediction of Corporation Bankruptcy. ***Journal of Finance*** **23, 4, 589-609.**

ANDERSON, J.A. (1972); Separate Sample Logistic Discrimination. ***Biometrika*** **59, 19-35.**

BIANCO, A.M., YOHAI, V.J. (1996); Robust Estimation in the Logistic Regression Model. *Robust Statistics, Data Analysis and Computer Intensive Methods*, Lecture Notes in Statistics, 109. Helmut Rieder ed., Springer Verlag.

CARROLL, R.J., PEDERSON S. (1993); On Robustness in the Logistic Regression Model. *Biometrika* 55, 693-706.

CHAMBERS J.M., HASTIE T.J. (1992); *Statistical Models in S.* Wadsworth and Brooks/Cole.

CHRISTMANN A., ROUSSEEUW.P.J. (1999); Measuring Overlap in Logistic Regression. *Technical Report, University of Antwerp.*

COPAS, J.B. (1998); Binary Regression Models for Contaminated Data. *Journal Royal Statistical Association Society B.* 50, 225-265.

HABERMAN, S.J. (1974); *The Analysis of Frequency Data.* University of Chicago Press.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J., STAHEL, W. A. *Robust Statistics: The Approach Based On Influence Functions.* John Willey and Sons, Inc., 1986.

HEINZE, G. (1999); Facing Separation in Logistic Regression. *www.dkfz-heidelberg.de/\-biostatistics/ISCB-GMDS-99.*

HOSMER, D.W., LEMESHOW, S. (1989); *Applied Logistic Regression.* John Wiley and Sons, New York.

JOHNSON W. (1985); Influence Measures for Logistic Regression: Another Point of View. *Biometrika* 72, 59-66.

LESAFFRE, E., ALBERT, A. (1989); Partial Separation in Logistic Regression. *Journal of Royal Statist. Society B,* 51, No. 1, 109-116.

KEENAN, S., SOBEHART, J. (2000); A Credit Risk Catwalk. *Risk,* Vol. 3, No. 7, 84-88.

KÜNSCH, H.R., STEFANSKI, L.A., CARROL, R.J. (1989); Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, With Applications to Generalized Linear Models. *Journal of the American Statistical Association,* 84, 460-466.

MARAZZI, A. (1993); *Algorithms, Routines and S Functions for Robust Statistics, The Fortran Library ROBETH with an Interface to S-Plus.* Wadsworth Brooks/Cole Statistics/ Probability Series, Pacific Grove.

MARAZZI, A. (1997); S-Plus Robeth e Robglm Libraries. *www.hospvd.ch/\ iumsp/\ download/\ robeth-en.htm.*

MCCULLAGH,P., NELDER, J.A. (1989); *Generalized Linear Models.* Chapman and Hall, London.

MENDES, B.V.M. (1988); Estudo do Ponto de Corte em uma Regra de Classificação para Dados Binários e uma Aplicação em Pneumologia. *Revista Brasileira de Estatística.* 49, 31-48.

Mendes, B.V.M. (1999); *Regressão Robusta: Conceitos, Aplicações e Aspectos Computacionais.* Mini-curso da 6a. Escola de Modelos de Regressão. Brasília.

PEREIRA, C. G. (1995); Análise de Crédito Bancário: Um sistema Especialista com Técnicas Difusas para os Limites da Agência. *Tese de Mestrado,* Programa de Pós-Graduação em Engenharia de Produção, UFSC.

PREGIBON, D. (1981); Logistic Regression Diagnostics. *The Annals of Statistics* 9, 705-724.

PREGIBON, D. (1982); Resistant Fits for Some Commonly Used Logistic Models with Medical Applications. *Biometrics* 38, 485-498.

ROUSSEEUW, P.J., LEROY, A.M. (1987); *Robust Regression and Outlier Detection.* Wiley, New York.

SANTNER, T.J., DUFFY, D.E. (1986); A Note no A. Alberto and J.A. Anderson's Conditions for Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika* 73, 755-758.

SAS INSTITUTE INC. (1995); *A Tutorial on Logistic Regression.* SAS, Inc., Carry, NC, USA.

SCHRIKEL, W.K. (1994); *Análise de Crédito: Concessão e gerência de empréstimos.* São Paulo, Atlas.

SILVAPULLE, M.J. (1981); On the Existence of Maximum Likelihood Estimates for the Binomial Response Models *J. R. Statist. Soc. B* **43, 310-325.**

STEFANSKI, L.A.; CARROL, R.J.; RUPPERT, D. (1986); Optimally Bounded Score Functions for Generalized Linear Models with Applications to Logistic Regression. *Biometrika* **73, 413-425.**

WEDDERBURN, R.V.M. (1976); On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models. *Biometrika* **63, 27-32.**

Agradecimentos

As autoras agradecem aos órgãos brasileiros de suporte à pesquisa FAPERJ e PRONEX/CNPq. Também agradecem ao economista H. A. de Lima pela ajuda na interpretação dos \índices utilizados.

ABSTRACT

In order to reduce and control their risks, financial institutions try different approaches when modeling the credit quality of their clients. The logistic regression model is often chosen to rigorously estimate the probability of default of their (potential) borrowers, and the estimation method is usually the maximum likelihood procedure. However, there is a class of situations, related to data overlapping, for which the maximum likelihood estimates fail to converge. On the other hand, the overlapping observations may also be atypical and have a large influence on the fit. In this case a robust procedure downweighting those observations should be called for. In this work we estimate the credit risk of 56 Brazilian firms using data from 1994. We propose a procedure for identifying the influent subsets and for tuning the robust estimator. Several measures of performance are used to compare the classical and the robust methods. The robust estimation procedure is able to downweight the atypical observations while preserving the data overlapping, and seems to perform better due to its larger information entropy.

Uso de métodos Bayesianos na modelagem da variabilidade extrabinomial

Juliano José Guimarães Junqueira*

Jorge Alberto Achcar*

RESUMO

Em muitas aplicações da distribuição binomial, pode se ter uma variabilidade observada maior ou menor do que a variabilidade esperada. Essa variabilidade é chamada variabilidade extrabinomial. Alguns modelos são introduzidos na literatura para ajustar a variabilidade extrabinomial: modelos betabinomial, binomial correlacionado e modelos de misturas de distribuições binomiais. Neste artigo, esses modelos são analisados sob o enfoque Bayesiano, utilizando métodos de Monte Carlo em cadeias de Markov (MCMC). Também são considerados modelos na presença de covariáveis. Dois exemplos com dados reais são introduzidos.

Palavras-chave: variabilidade extrabinomial, análise Bayesiana, métodos MCMC, covariáveis.

1. Introdução

A distribuição binomial é comumente usada para dados de contagens de y_i sucessos num total de n_i ensaios independentes, sendo que cada ensaio admite duas respostas possíveis, sucesso ou fracasso. Contudo, em muitas aplicações pode-se ter uma variabilidade observada dos dados maior do que a variabilidade esperada quando se considera a suposição ordinária de uma distribuição binomial $b(n_i, p)$ para Y_i (variável aleatória

* Endereço para correspondência: Universidade de São Paulo ICMC - C. Postal 668 -13560-970, São Carlos, SP.

que representa o número de sucessos em n_i ensaios), $i = 1,....,N$. Essa variabilidade é chamada variabilidade extrabinomial (por exemplo, Skellam, 1948 ; Altham, 1978 ou Rudolfer, 1990).

Alguns modelos tem sido propostos na literatura para ajustar a variabilidade extrabinomial. Entre eles destacam-se os modelos betabinomial, binomial correlacionado e misturas de distribuições binomiais.

Com o modelo betabinomial (Skellam, 1948), a variabilidade extrabinomial é originada pela distribuição de probabilidade na probabilidade de sucesso p da distribuição binomial. A distribuição beta com parâmetros α e β é a distribuição escolhida para p e a função densidade de probabilidade para a variável Y_i, dados n_i, α e β, é dada por,

$$f(y_i;n_i,\alpha,\beta)=\binom{n_i}{y_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)}{\Gamma(\alpha+\beta+n_i)}I_{\{0,1,\ldots,n_i\}}(y_i) \qquad (1)$$

(densidade betabinomial).

O modelo binomial correlacionado (Altham,1978 ou Kupper & Hasemam, 1978) assume que a fonte da variabilidade extrabinomial é dada pela correlação δ entre as variáveis binárias $X_{i1}, X_{i2},..., X_{in_i}$ de Y_i. Neste caso a função densidade de probabilidade de Y_i, dados n_i, δ e p, é dada por,

$$f(y_i;n_i,\delta,p)=\binom{n_i}{y_i}p^{y_i}(1-p)^{n_i-y_i}\left\{1+\frac{\delta}{2p(1-p)}\left[(y_i-n_ip)^2+y_i(2p-1)-n_ip^2\right]\right\}. \qquad (2)$$

Uma terceira possibilidade de modelagem para ajustar a variabilidade extrabinomial é considerar que a contagem de sucessos y_i é supostamente gerada por uma mistura finita de distribuições binomiais (Hsiao, 1994), ou seja, a observação y_i origina-se de uma entre J categorias, mas a fonte exata é desconhecida. Considerando o caso particular $J = 2$ (mistura de duas binomiais), temos,

$$f(y_i;n_i,\theta_1,\theta_2,\pi)=\pi\binom{n_i}{y_i}\theta_1^{y_i}(1-\theta_1)^{n_i-y_i}+(1-\pi)\binom{n_i}{y_i}\theta_2^{y_i}(1-\theta_2)^{n_i-y_i}. \qquad (3)$$

Uma introdução à variabilidade extra para vários modelos é introduzida por Hinde e Demetrio(2000).

Neste artigo, apresenta-se uma análise Bayesiana desses modelos usando métodos de Monte Carlo em Cadeia de Markov (MCMC). Em particular, são utilizados os algoritmos Gibbs Sampling (por exemplo, Gelfand & Smith, 1990) e Metropolis-Hastings (por exemplo, Smith & Roberts, 1993) para obter estimadores de Monte

Carlo das quantidades *a posteriori* de interesse dos parâmetros. O critério de Gelman & Rubin (Gelman & Rubin, 1992) é utilizado para o diagnóstico de convergência das amostras geradas por esses algoritmos.

2. Análise Bayesiana do modelo Betabinomial

Assumindo o modelo Betabinomial (1), considera-se as seguintes distribuições *a priori* para os parâmetros:

$$\begin{aligned} &\text{i)}\ \ \alpha \sim \Gamma(a_1, b_1);\ \ a_1, b_1 \text{ conhecidos} \\ &\text{ii)}\ \ \beta \sim \Gamma(a_2, b_2);\ \ a_2, b_2 \text{ conhecidos} \end{aligned} \tag{4}$$

em que $\Gamma(a,b)$ denota uma distribuição gama com média a/b e variância a/b^2.

A escolha da distribuição gama para os parâmetros é relacionada à sua flexibilidade para incorporar opinião *a priori* e é definida para valores não-negativos.

Considerando que as distribuições *a priori* para os parâmetros são independentes, a densidade *a posteriori* conjunta para α e β é dada por,

$$\pi\left(\alpha,\beta / \underset{\sim}{y}, \underset{\sim}{n}\right) \propto \alpha^{a_1-1} e^{-b_1\alpha} \beta^{a_2-1} e^{-b_2\beta} \left\{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right\}^N \prod_{i=1}^{N}\left[\frac{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)}{\Gamma(\alpha+\beta+n_i)}\right] \tag{5}$$

sendo $\underset{\sim}{y} = (y_1,\ldots,y_N)$ e $\underset{\sim}{n} = (n_1,\ldots,n_N)$.

A partir de (5), as distribuições condicionais *a posteriori* de α e β são encontradas.

Essas distribuições são dadas por,

$$\text{i)}\quad \pi\left(\alpha / \beta, \underset{\sim}{y}, \underset{\sim}{n}\right) \propto \alpha^{a_1-1} e^{-b_1\alpha} \left\{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\right\}^N \prod_{i=1}^{N}\left[\frac{\Gamma(\alpha+y_i)}{\Gamma(\alpha+\beta+n_i)}\right]$$

$$\propto \alpha^{a_1-1} e^{-b_1\alpha} \exp\left\{N\log[\Gamma(\alpha+\beta)] - N\log[\Gamma(\alpha)] + \sum_{i=1}^{N}\log[\Gamma(\alpha+y_i)] - \sum_{i=1}^{N}\log[\Gamma(\alpha+\beta+n_i)]\right\}$$

ou $\pi\left(\alpha/\beta,\underset{\sim}{y},\underset{\sim}{n}\right)\propto\Gamma(a_1,b_1)\psi_1(\alpha,\beta)$

em

que $\psi_1(\alpha,\beta)=\exp\left\{N\log[\Gamma(\alpha+\beta)]-N\log[\Gamma(\alpha)]+\sum_{i=1}^{N}\log[\Gamma(\alpha+y_i)]-\sum_{i=1}^{N}\log[\Gamma(\alpha+\beta+n_i)]\right\}$

ii) $\pi\left(\beta/\alpha,\underset{\sim}{y},\underset{\sim}{n}\right)\propto\beta^{a_2-1}e^{-b_2\beta}\left\{\frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)}\right\}^{N}\prod_{i=1}^{N}\left[\frac{\Gamma(\beta+n_i-y_i)}{\Gamma(\alpha+\beta+n_i)}\right]$

$$\propto\beta^{a_2-1}e^{-b_2\beta}\exp\left\{N\log[\Gamma(\alpha+\beta)]-N\log[\Gamma(\beta)]+\sum_{i=1}^{N}\log[\Gamma(\beta+n_i-y_i)]-\sum_{i=1}^{N}\log[\Gamma(\alpha+\beta+n_i)]\right\}$$

ou $\pi\left(\beta/\alpha,\underset{\sim}{y},\underset{\sim}{n}\right)\propto\Gamma(a_2,b_2)\psi_2(\alpha,\beta)$

em que

$$\psi_2(\alpha,\beta)=\exp\left\{N\log[\Gamma(\alpha+\beta)]-N\log[\Gamma(\beta)]+\sum_{i=1}^{N}\log[\Gamma(\beta+n_i-y_i)]-\sum_{i=1}^{N}\log[\Gamma(\alpha+\beta+n_i)]\right\}$$

(6)

Como as distribuições condicionais para α e β não apresentam formas de distribuições conhecidas, o algoritmo Metropolis-Hastings é utilizado para realizar a simulação das amostras dos parâmetros e fazer as inferências desejadas.

3. Análise Bayesiana do modelo Binomial Correlacionado

Assumindo o modelo binomial correlacionado (2), considera-se as seguintes distribuições *a priori* para os parâmetros:

(i) $\delta \sim U(a,b)$; a,b conhecidos

(7)

(ii) $p \sim \text{Beta}(c,d)$; c,d conhecidos

em que $U(a,b)$ denota uma distribuição uniforme no intervalo (a,b) e $\text{Beta}(c,d)$ denota uma distribuição beta com média $\frac{c}{c+d}$ e variância $\frac{cd}{(c+d)^2(c+d+1)}$.

Assumindo independência *a priori* para os parâmetros, encontra-se a densidade a posteriori conjunta para δ e p dada por,

$$\pi\left(\delta, p / \underset{\sim}{y}, \underset{\sim}{n}\right) \propto p^{c+\sum_{i=1}^{N} y_i - 1} (1-p)^{d+\sum_{i=1}^{N}(n_i - y_i)-1} \prod_{i=1}^{N} \left\{1 + \frac{\delta}{2p(1-p)}\left[(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2\right]\right\} \quad (8)$$

em que $\underset{\sim}{y} = (y_1, ..., y_N)$ e $\underset{\sim}{n} = (n_1, ..., n_N)$.

Portanto, as distribuições condicionais *a posteriori* para os parâmetros δ e p são dadas por,

i)

$$\pi\left(\delta / p, \underset{\sim}{y}, \underset{\sim}{n}\right) \propto \prod_{i=1}^{N} \left\{1 + \frac{\delta}{2p(1-p)}\left[(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2\right]\right\}$$

ou $\pi\left(\delta / p, \underset{\sim}{y}, \underset{\sim}{n}\right) \propto U(a,b)\psi(\delta, p)$

ii)

$$\pi\left(p / \delta, \underset{\sim}{y}, \underset{\sim}{n}\right) \propto p^{c+\sum_{i=1}^{N} y_i - 1} (1-p)^{d+\sum_{i=1}^{N}(n_i - y_i)-1} \prod_{i=1}^{N} \left\{1 + \frac{\delta}{2p(1-p)}\left[(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2\right]\right\}$$

ou $\pi\left(p/\delta, \underset{\sim}{y}, \underset{\sim}{n}\right) \propto \text{Beta}\left(c+\sum_{i=1}^{N} y_i, d+\sum_{i=1}^{N}(n_i - y_i)\right)\psi(\delta, p)$

em que $\psi(\delta, p) \propto \prod_{i=1}^{N}\left\{1+\frac{\delta}{2p(1-p)}\left[(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2\right]\right\}$ (9)

Como as distribuições condicionais para δ e p não apresentam formas de distribuições conhecidas, o algoritmo Metropolis-Hastings é utilizado para realizar a simulação das amostras dos parâmetros e fazer as inferências desejadas.

4. Análise Bayesiana do modelo de mistura de duas distribuições Binomiais

Para simplificar a análise Bayesiana do modelo de mistura de duas distribuições binomiais (3), são introduzidos no modelo, variáveis não observáveis Z_i, $(i = 1,....,N)$, também chamadas de variáveis latentes.

A variável latente Z_i, $(i = 1,....,N)$ é definida como:

$$Z_i = \begin{cases} 1 \text{ se a } i\text{-ésima observação foi gerada da primeira distribuição binomial} \\ 0 \text{ se a } i\text{-ésima observação foi gerada da segunda distribuição binomial} \end{cases}$$

ou seja, Z_i tem uma distribuição de Bernoulli com probabilidade de sucesso dada por,

$$h_i = P(Z_i = 1/y_i, n_i, \theta_1, \theta_2, \pi) = \frac{\pi\binom{n_i}{y_i}\theta_1^{y_i}(1-\theta_1)^{n_i - y_i}}{\pi\binom{n_i}{y_i}\theta_1^{y_i}(1-\theta_1)^{n_i - y_i} + (1-\pi)\binom{n_i}{y_i}\theta_2^{y_i}(1-\theta_2)^{n_i - y_i}} \quad (10)$$

A função de verossimilhança simplificada com a introdução das variáveis latentes $(Z_i, i = 1,....,N)$ é dada por,

$$L\left(\underset{\sim}{n}, \theta_1, \theta_2, \pi / \underset{\sim}{y}, \underset{\sim}{z}\right) = \pi^{\sum_{i=1}^{N} z_i}(1-\pi)^{N - \sum_{i=1}^{N} z_i}\prod_{i=1}^{N}\left[\left(\binom{n_i}{y_i}\theta_1^{y_i}(1-\theta_1)^{n_i - y_i}\right)^{z_i}\left(\binom{n_i}{y_i}\theta_2^{y_i}(1-\theta_2)^{n_i - y_i}\right)^{1-z_i}\right]. \quad (11)$$

Uma análise Bayesiana para o modelo de mistura de duas distribuições binomiais é desenvolvida assumindo as seguintes distribuições *a priori* para os parâmetros:

$$\begin{aligned} &\text{i) } \theta_1 \sim \text{Beta}(a_1, b_1)\ ; \\ &\text{ii) } \theta_2 \sim \text{Beta}(a_2, b_2)\ ; \text{ e} \\ &\text{iii) } \pi \sim \text{Beta}(c, d). \end{aligned} \tag{12}$$

Observar que os parâmetros θ_1, θ_2 e π são definidos no intervalo [0,1], que leva à escolha da *a priori* beta, além de sua grande flexibilidade para incorporar opinião *a priori*.

Assumindo que as distribuições *a priori* para os parâmetros θ_1, θ_2 e π são independentes, a densidade *a posteriori* conjunta para θ_1, θ_2 e π, é dada por,

$$\begin{aligned} &g\left(\theta_1, \theta_2, \pi / \underset{\sim}{n}, \underset{\sim}{y}, \underset{\sim}{z}\right) \propto \theta_1^{a_1-1}(1-\theta_1)^{b_1-1}\theta_2^{a_2-1}(1-\theta_2)^{b_2-1}\pi^{c-1}(1-\pi)^{d-1}\pi^{\sum_{i=1}^{N} z_i}(1-\pi)^{N-\sum_{i=1}^{N} z_i} \\ &\prod_{i=1}^{N}\left[\left(\binom{n_i}{y_i}\theta_1^{y_i}(1-\theta_1)^{n_i-y_i}\right)^{z_i}\left(\binom{n_i}{y_i}\theta_2^{y_i}(1-\theta_2)^{n_i-y_i}\right)^{1-z_i}\right] \\ &\propto \theta_1^{a_1+\sum_{i=1}^{N} y_i z_i-1}(1-\theta_1)^{b_1+\sum_{i=1}^{N} z_i(n_i-y_i)-1}\theta_2^{a_2+\sum_{i=1}^{N}(1-z_i)y_i-1}(1-\theta_2)^{b_2+\sum_{i=1}^{N}(1-z_i)(n_i-y_i)-1}\pi^{c+\sum_{i=1}^{N} z_i-1}(1-\pi)^{d+N-\sum_{i=1}^{N} z_i-1}. \end{aligned} \tag{13}$$

Portanto, as distribuições condicionais necessárias para o uso dos métodos MCMC são dadas por,

$$\begin{aligned} &\text{i) } g\left(\theta_1 / \theta_2, \pi, \underset{\sim}{n}, \underset{\sim}{z}, \underset{\sim}{y}\right) \propto \theta_1^{a_1+\sum_{i=1}^{N} y_i z_i-1}(1-\theta_1)^{b_1+\sum_{i=1}^{N} z_i(n_i-y_i)-1}; \\ &\text{ii) } g\left(\theta_2 / \theta_1, \pi, \underset{\sim}{n}, \underset{\sim}{z}, \underset{\sim}{y}\right) \propto \theta_2^{a_2+\sum_{i=1}^{N}(1-z_i)y_i-1}(1-\theta_2)^{b_2+\sum_{i=1}^{N}(1-z_i)(n_i-y_i)-1}; \text{ e} \\ &\text{iii) } g\left(\pi / \theta_1, \theta_2, \underset{\sim}{n}, \underset{\sim}{z}, \underset{\sim}{y}\right) \propto \pi^{c+\sum_{i=1}^{N} z_i-1}(1-\pi)^{d+N-\sum_{i=1}^{N} z_i-1}. \end{aligned} \tag{14}$$

Resultados similares são obtidos para misturas finitas de mais de duas distribuições binomiais.

Como as distribuições condicionais para os parâmetros θ_1, θ_2 e π apresentam formas de distribuições beta, o algoritmo Gibbs Sampling é usado para realizar a simulação das amostras dos parâmetros e fazer as inferências desejadas.

O algoritmo computacional para gerar amostras dos parâmetros θ_1, θ_2 e π baseado na introdução de variáveis latentes é dado pelas seguintes etapas:

i) dados os valores iniciais $\theta_1^{(0)}$, $\theta_2^{(0)}$ e $\pi^{(0)}$, gerar uma amostra $(z_1, z_2, ..., z_N)$ a partir da distribuição de Bernoulli p/ z_i com probabilidade de sucesso,

$$h_i = \frac{\pi^{(0)}\binom{n_i}{y_i}\left(\theta_1^{(0)}\right)^{y_i}\left(1-\theta_1^{(0)}\right)^{n_i-y_i}}{\pi^{(0)}\binom{n_i}{y_i}\left(\theta_1^{(0)}\right)^{y_i}\left(1-\theta_1^{(0)}\right)^{n_i-y_i} + \left(1-\pi^{(0)}\right)\binom{n_i}{y_i}\left(\theta_2^{(0)}\right)^{y_i}\left(1-\theta_2^{(0)}\right)^{n_i-y_i}}, \quad i = 1,...,N.$$

ii) gerar $\theta_1^{(1)}$ a partir da distribuição $\text{Beta}\left(a_1 + \sum_{i=1}^{N} y_i z_i, b_1 + \sum_{i=1}^{N} z_i (n_i - y_i)\right)$;

iii) gerar $\theta_2^{(1)}$ a partir da distribuição $\text{Beta}\left(a_2 + \sum_{i=1}^{N} y_i (1-z_i), b_2 + \sum_{i=1}^{N} (1-z_i)(n_i - y_i)\right)$; e

iv) gerar $\pi^{(1)}$ a partir da distribuição $\text{Beta}\left(c + \sum_{i=1}^{N} z_i, d + N - \sum_{i=1}^{N} z_i\right)$.

Repetir (i)-(iv) até gerar uma amostra de tamanho S.

5. Discriminação dos modelos

A verossimilhança marginal é uma ferramenta de grande importância numa análise Bayesiana quando há interesse na construção de testes de hipóteses e na discriminação dos modelos. Para um modelo representado por M, a verossimilhança marginal é definida na forma,

$$p(\underset{\sim}{y}/M) = \int L(\underset{\sim M}{\theta}, M)\,\pi(\underset{\sim M}{\theta}/M)\,\partial\underset{\sim M}{\theta} \tag{15}$$

sendo $L(\underset{\sim M}{\theta}, M)$ a função de verossimilhança para o modelo M,

$\pi(\underset{\sim M}{\theta}/M)$ a distribuição *a priori* conjunta para os parâmetros do modelo M,

e $\underset{\sim M}{\theta}$ o vetor de parâmetros do modelo M.

Quando nosso objetivo é a comparação de dois modelos M_1 e M_2, utilizamos o fator de Bayes B_{12}, que é definido como a razão das verossimilhanças marginais desses modelos, ou seja:

$$B_{12} = \frac{P\left(\underset{\sim}{y}/M_1\right)}{P\left(\underset{\sim}{y}/M_2\right)}. \tag{16}$$

Desta forma, o fator de Bayes é um valor numérico baseado nos dados, que está a favor de um determinado modelo estatístico e contra um outro modelo. Um interpretação do fator de Bayes sugerido por Kass e Raftery (1995) é dada na Tabela 1 a seguir:

Tabela 1: Interpretação do fator de Bayes

B_{12}	$2\log B_{12}$	Evidência a favor de M_1
<1	<0	Negativa (apóia M_2)
$1 \mapsto 3$	$0 \mapsto 2$	Vale ser mencionada
$3 \mapsto 20$	$2 \mapsto 6$	Positiva
$20 \mapsto 150$	$6 \mapsto 10$	Forte
>150	>10	muito forte

Pela Tabela 1, concluímos que quando a verossimilhança marginal do modelo M_1 for maior que a verossimilhança marginal do modelo M_2, ou seja quando $B_{12} > 1$, o modelo M_1 apresenta um melhor ajuste aos dados.

Para o caso de r modelos, M_1, M_2,...., M_r, escolhemos aquele modelo que apresentar a maior verossimilhança marginal.

6. Variabilidade extrabinomial na presença de covariáveis

Se variáveis explanatórias estão presentes, usualmente é considerada uma regressão logística para modelagem. A variação residual, contudo, pode ser maior do que a variação esperada, o que caracteriza a existência da variabilidade extrabinomial (por exemplo, Williams, 1982).

Na presença de um vetor $\underset{\sim}{x} = (x_1,....,x_k)$ de covariáveis, o modelo de regressão logística é dado por,

$$y_i / n_i, y_i, \underset{\sim}{x_i} \sim b(n_i, p_i) \tag{17}$$

sendo $p_i = \dfrac{e^{\underset{\sim}{\beta}'\underset{\sim}{x}_i}}{1+e^{\underset{\sim}{\beta}'\underset{\sim}{x}_i}}$, e $\underset{\sim}{\beta}'\underset{\sim}{x}_i = \beta_0 + \sum_{l=1}^{k}\beta_l x_{li}$, $i = 1,...,N$.

Assume-se os mesmos valores para as covariáveis $\underset{\sim}{x}_i = (x_{1i},...,x_{ki})$ para as n_i observações na distribuição $b(n_i, p)$, $i = 1,....,N$.

A variação residual contudo pode ser maior do que a variação esperada assumindo o modelo Binomial, quando todas as covariáveis do modelo são ajustadas.

Para modelar a variação extrabinomial na presença de covariáveis, assume-se o modelo Binomial Correlacionado (2), isto é,

$$P(Y_i = y_i / n_i, \underset{\sim}{x}_i, \underset{\sim}{\beta}, \delta) = \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i} A_i\left(n_i, y_i, \underset{\sim}{x}_i, \underset{\sim}{\beta}, \delta\right) \tag{18}$$

sendo $A_i\left(n_i, y_i, \underset{\sim}{x}_i, \underset{\sim}{\beta}, \delta\right) = 1 + \dfrac{\delta}{2p_i(1-p_i)}\left[(y_i - n_i p_i)^2 + y_i(2p_i - 1) - n_i p_i^2\right]$,

$$p_i = \frac{e^{\underset{\sim}{\beta}'\underset{\sim}{x}_i}}{1+e^{\underset{\sim}{\beta}'\underset{\sim}{x}_i}}, \text{e } \underset{\sim}{\beta}' = (\beta_0, \beta_1,...., \beta_k), \quad i = 1,....,N.$$

Observar que no modelo (18), assume-se que a correlação intraclasse entre as variáveis binárias $U_{i1},....,U_{1n_i}$ de $Y_i = \sum_{l=1}^{n_i} U_{il}$ é homogênea e independente do valor das covariáveis $\underset{\sim}{x}_i$, $i = 1,...,N$.

Assumindo independência *a priori* entre os parâmetros, as seguintes distribuições *a priori* para os parâmetros são consideradas:

i) $\beta_0 \sim N(\mu_0, \sigma_0^2)$; μ_0, σ_0^2 conhecidos;

ii) $\beta_l \sim N(\mu_l, \sigma_l^2)$; μ_l, σ_l^2 conhecidos; e (19)

iii) $\delta \sim U(a,b)$; a, b conhecidos; e

sendo que $l = 1,2,...,k$, e $N(u, \sigma^2)$ denota uma distribuição normal com média u e variância σ^2.

Assumindo as distribuições *a priori* dadas em (19), utiliza-se os métodos MCMC para realizar a simulação dos parâmetros e obter as inferências desejadas.

7. Alguns exemplos

7.1 – Um exemplo com dados genéticos

Nesta aplicação, considera-se os dados genéticos introduzidos por Skellam (1948), onde o objetivo é a associação secundária de cromossomos numa espécie de couve-flor e repolho, chamada Brassica.

As unidades observadas são 337 núcleos, sendo que em cada núcleo existem três pares de cromossomos durante a meiose. Cada par pode mostrar ou não, uma associação entre esses cromossomos, ou seja em cada núcleo podemos ter 0, 1, 2 ou 3 pares de cromossomos associados. As freqüências observadas com 0, 1, 2 e 3 pares associados nesses núcleos são dadas por 32, 103, 122 e 80, respectivamente.

Observa-se que neste exemplo, a variável Y_i representa o número de pares de cromossomos associados no i-ésimo núcleo, $i = 1,....,337$. Assume-se para Y_i os modelos binomial, betabinomial (1) , binomial correlacionado (2) e mistura de duas distribuições binomiais (3), utilizando os métodos clássicos e Bayesianos.

Na Tabela 2 são apresentados os resultados das análises clássicas dos modelos, considerando os dados genéticos.

Tabela 2 – Resultados das análises Clássicas dos modelos para os dados genéticos

Modelo	Parâmetro	EMV	Int. Conf. (95%)
Binomial	p	0,5806	(0,5502 ; 0,6110)
Beta-binomial	α	6,1198	(0,9563 ; 11,2833)
	β	4,4186	(0,6945 ; 8,1427)
Binomial Correlacionado	p	0,5809	(0,5480 ; 0,6139)
	δ	0,0883	(0,0214 ; 0,1553)
Mistura de 2 Distribuições Binomiais	θ_1	0,5138	(0,3738 ; 0,6538)
	θ_2	0,8941	(0,3644 ; 1)
	π	0,8242	(0,2911 ; 1)

Na análise Bayesiana são considerados os seguintes valores para os hiperparâmetros: modelo beta-binomial ($a_1 = 90$, $b_1 = 15$, $a_2 = 70$ e $b_2 = 16$), modelo binomial correlacionado ($a = 30$, $b = 20$, $c = 0$ e $d = 0{,}15$) e modelo de mistura de duas distribuições binomiais ($a_1 = 30$, $b_1 = 30$, $a_2 = 20$, $b_2 = 2$, $c = 18$ e $d = 4$).

Utilizando os métodos MCMC foram geradas para todos os modelos cinco cadeias de 2000 valores para cada parâmetro, a partir das distribuições condicionais *a posteriori*. Para cada cadeia descartamos as primeiras 1000 iterações e consideramos a 10ª, 20ª,....., iterações. A convergência do algoritmo foi verificada graficamente e usando o critério de Gelman e Rubin (1992).

Para a geração das amostras foi utilizado o software MATLAB.

As estimativas Bayesianas de todos os modelos considerando os dados genéticos encontram-se na Tabela 3. Também são dados na Tabela 3 os valores do critério de Gelman e Rubin (GR) para todos os parâmetros.

Tabela 3 - Sumários *a posteriori* para os dados genéticos

Modelo	Parâmetro	Média	DP	Int. Cred. (95%)	GR
Binomial	p	0,5805	0,0155	(0,5501 ; 0,6108)	....
Beta-	α	6,0839	0,5028	(5,0783 ; 7,0830)	1,0016
binomial	β	4,3901	0,3759	(3,7198 ; 5,2018)	1,0014
Binomial	p	0,0883	0,0294	(0,0309 ; 0,1429)	1,0002
Correlacionado	δ	0,5806	0,0166	(0,5487 ; 0,6110)	1,0026
Mistura de 2	θ_1	0,5159	0,0241	(0,4691 ; 0,5624)	1,0010
Distribuições	θ_2	0,9067	0,0529	(0,7801 ; 0,9845)	1,0020
Binomiais	π	0,8316	0,0574	(0,6860 ; 0,9239)	1,0033

A Tabela 4 apresenta o valor da verossimilhança marginal exata do modelo Binomial e as estimativas de Monte Carlo dos outros modelos, obtidas para os dados genéticos.

Tabela 4 – Discriminação dos modelos para os dados genéticos

Modelo	Estimativa da Ver. Marginal
Binomial	$0{,}2149 \times 10^{-192}$
Betabinomial	$1{,}4615 \times 10^{-190}$
Binomial Correlacionado	$1{,}1294 \times 10^{-190}$
Mistura de 2 distribuições Binomiais	$1{,}4958 \times 10^{-190}$

Observa-se que os modelos betabinomial, binomial correlacionado e mistura de 2 distribuições binomiais tiveram um comportamento semelhante no ajuste aos dados genéticos, ao passo que ambos superaram o ajuste do modelo binomial, indicando a presença da variabilidade extrabinomial nesses dados.

Observar que os intervalos de confiança assintóticos clássicos (Tabela 2) para alguns parâmetros têm comprimentos muito grandes quando comparados com os intervalos Bayesianos.

7.2 – Um exemplo com covariáveis

Considerar o experimento analisado por Crowder (1978), em que uma quantidade de sementes é colocada numa placa coberta com um extrato numa dada diluição. Os números de sementes que germinaram e não germinaram são anotados. Foram considerados dois tipos de sementes **(O. aegyptiaca 75 e O. aegyptiaca 73)**, dois tipos de extratos (feijão e pepino) e algumas réplicas para as combinações.

Apresentamos os dados desse experimento na Tabela 5. Para cada réplica, temos o número total de sementes (n_i), o número de sementes que germinaram (y_i) e a proporção de sementes que germinaram (y_i / n_i).

Tabela 5 – Dados de Crowder (1978)

(I) O. aegyptiaca 75						(II) O. aegyptiaca 73					
feijão			pepino			feijão			pepino		
n_i	y_i	y_i / n_i	n_i	y_i	y_i / n_i	n_i	y_i	y_i / n_i	n_i	y_i	y_i / n_i
39	10	0,26	6	5	0,83	16	8	0,50	12	3	0,25
62	23	0,37	74	53	0,72	30	10	0,33	41	22	0,54
81	23	0,28	72	55	0,76	28	8	0,29	30	15	0,50
51	26	0,51	51	32	0,63	45	23	0,51	51	32	0,63
39	17	0,44	79	46	0,58	4	0	0	7	3	0,43
			13	10	0,77						

Numa inspeção dos dados da Tabela 5, observa-se uma significativa heterogeneidade entre as proporções das réplicas e é considerada na análise dos dados a presença de duas covariáveis x_{1i} e x_{2i}: x_{1i} assume os valores -1 para o tipo **O. aegyptiaca 75** e 1 para o tipo **O. aegyptiaca 73** e x_{2i} assume os valores -1 para o extrato de feijão e 1 para o extrato de pepino.

Assumir para a variável Y_i (número de sementes que germinaram em n_i sementes), os modelos de regressão logística (17) sem assumir a presença de variabilidade extrabinomial e o modelo binomial correlacionado (18), considerando as covariáveis x_{1i}, x_{2i} e a interação $x_{1i}\, x_{2i}$:

$$p_i = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{1i} x_{2i}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{1i} x_{2i}}}, \text{ para } i = 1,....,21. \qquad (20)$$

Na Tabela 6 apresenta-se os resultados das análises clássicas dos modelos, considerados com covariáveis.

Tabela 6 – Resultados das análises clássicas dos modelos para os dados da Tabela 5

Modelo	Parâmetro	Média	Int. Conf (95%)
Regressão Logística	β_0	-0,0206	(-0,1708 ; 0,1296)
	β_1	-0,1216	(-0,2718 ; 0,0286)
	β_2	0,4646	(0,3144 ; 0,6148)
	β_3	-0,1945	(-0,3447 ; -0,0443)
Binomial Correlacionado	δ	-0,0562	(-0,2410 ; 0,1286)
	β_0	-0,1557	(-0,3422 ; 0,0307)
	β_1	0,4214	(0,2279 ; 0,6150)
	β_2	-0,2003	(-0,3843 ; -0,0163)
	β_3	0,02085	(0,0033 ; 0,0384)

Na análise Bayesiana são considerados os seguintes valores para os hiperparâmetros: modelo de regressão logística ($\mu_0 = 0{,}05$, $\sigma_0^2 = 0{,}2$, $\mu_1 = -0{,}2$, $\sigma_1^2 = 0{,}2$, $\mu_2 = 0{,}3$, $\sigma_2^2 = 0{,}2$, $\mu_3 = -0{,}3$ e $\sigma_3^2 = 0{,}2$) e binomial correlacionado ($\mu_0 = -0{,}05$, $\sigma_0^2 = 0{,}1$, $\mu_1 = -0{,}15$, $\sigma_1^2 = 0{,}1$, $\mu_2 = 0{,}42$, $\sigma_2^2 = 0{,}1$, $\mu_3 = -0{,}2$, $\sigma_3^2 = 0{,}1$, $a = 0{,}015$ e $b = 0{,}025$).

Os métodos MCMC são usados para gerar amostras das distribuições *a posteriori* marginais dos parâmetros.

Nos dois modelos são geradas cinco cadeias de 2000 valores para cada parâmetro, utilizando as distribuições condicionais *a posteriori*. Para cada cadeia descarta-se as primeiras 1000 iterações e considera-se a 10^a, 20^a,....., iterações. Usando o critério de Gelman e Rubin (1992), foi observada a convergência do algoritmo Gibbs Sampling (valores do critério Gelman e Rubin, GR são dados na Tabela 7).

As estimativas Bayesianas dos modelos considerando a presença de covariáveis são dadas na Tabela 7.

Tabela 7 – Sumários *a posteriori* para o exemplo com covariáveis

Modelo	Parâmetro	Média	DP	IC (95%)	GR
Regressão Logística	β_0	-0,0144	0,0691	(-0,1501 ; 0,1265)	1,0002
	β_1	-0,1268	0,0711	(-0,2638 ; 0,0144)	1,0005
	β_2	0,4398	0,0648	(0,3163 ; 0,5666)	1,0006
	β_3	-0,2155	0,0704	(-0,3604 ; -0,0865)	1,0010
Binomial Correlacionado	δ	-0,0503	0,0658	(-0,1878 ; 0,0702)	0,9998
	β_0	-0,16	0,0635	(-0,2857 ; -0,0447)	1,0001
	β_1	0,423	0,0626	(0,3109 ; 0,5532)	1,0007
	β_2	-0,1998	0,0615	(-0,3256 ; -0,0704)	1,0004
	β_3	0,0201	0,0028	(0,0152 ; 0,0248)	0,9999

A Tabela 8 apresenta as estimativas de Monte Carlo da verossimilhança marginal obtidas para este exemplo com covariáveis.

Tabela 8 – Discriminação dos modelos para o exemplo com covariáveis

Modelo	Estimativa da Ver. Marginal
Regressão Logística	$4{,}1349 \times 10^{-25}$
Binomial Correlacionado	$6{,}6823 \times 10^{-24}$

Observa-se pelos resultados da Tabela 8 que o modelo binomial correlacionado na presença de covariáveis, ajustou-se melhor aos dados do que o modelo de regressão logística. Também observa-se que considerando o modelo binomial correlacionado, as covariáveis x_{1i}, x_{2i} e a interação $x_{1i}\,x_{2i}$ tem efeitos significativos.

8. Conclusões

O uso dos modelos para ajustar a variabilidade extrabinomial é necessário em muitas aplicações, quando a variação observada é maior do que a variação esperada a partir da suposição binomial. Também, em muitas aplicações, pode-se ter a presença de covariáveis.

O uso dos métodos MCMC é apropriado para obter sumários *a posteriori* de interesse desses modelos. Além disso, o uso de métodos clássicos usuais baseados na aproximação normal dos estimadores de máxima verossimilhança pode não ser apropriado pois o tamanho amostral pode não ser adequado na parametrização original.

Outra vantagem do método Bayesiano é relacionado à discriminação de diferentes modelos usando a distribuição preditiva.

Um outro ponto importante: observamos que o uso de métodos MCMC não exigem conhecimento computacional sofisticado para gerar amostras das distribuições condicionais dos parâmetros.

Referências bibliográficas

ALTHAM, P.M.E (1978). Two Generalizations of the Binomial Distribution. *Applied Statistics*, n.27, p.162-167.

CROWDER, M.J. (1978). Beta-Binomial Anova for Proportions. *Applied Statistics*, n.27, p.34-37.

GELFAND, A.E.; SMITH, A.F.M. (1990). Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, n.85, p.398 – 409.

GELMAN, A.E.; RUBIN, D. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Sciences*, n.7, p.457-472.

HINDE,J.P.;DEMETRIO,C.G.B. (2000). Overdispersion:models and estimation,notas da 20ª International Biometric Conference,Berkeley,CA,E.U.A.,110 p.

HSIAO, C.K. (1994). Bayesian Tests of Extra-Binomial Variability with Emphasis on the Boundary Case, *PhD thesis*, Carnegie-Mellon University, U.S.A.

KASS, R.E.; RAFTERY, A.E. Bayes Factors and Model Uncertainty. *Journal of the American Statistical Association*, n.90, p.773-775, 1995.

KUPPER, L.L.; HASEMAN, J.K. (1978). The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments. *Biometrics*, n.34, p.69–76.

RUDOLFER, S.,M. (1990). A Markov Chain Model of Extra-Binomial Variation. *Biometrika*, n.77, p.255-264.

SKELLAM, J.G. (1948). A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable between the Sets of Trials. *Journal of the Royal Statistical Society*, B,10, p.257-261.

SMITH, A.F.M.; ROBERTS, G.O. (1993). Bayesian Methods via the Gibbs Sampler and related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society*, B, n.55, p.3-23.

WILLIAMS, D.A. (1982). Extra-Binomial Variation in Logistic Linear Models. *Applied Statistics*, p.31, n.144-148.

Abstract

In many applications of the binomial distribution, observed variability may be larger or smaller than expected. This phenomenon is called extrabinomial variability. Some models are available in the literature to cope with extrabinomial variability: betabinomial models, correlated binomial and mixtures of binomial distributions. In this paper, these models are analysed under a Bayesian perspective, using Markov Chain Monte Carlo (MCMC) methods. Models containing covariates are also considered. Two examples with real data are examined.

Status socioeconômico das ocupações brasileiras: índices aproximativos para 1980, 1991 e anos de 1990[1]

Paulo de Martino Jannuzzi*

RESUMO

Status socioeconômico de uma ocupação tem sido tratado na Pesquisa Social como um conceito relacionado ao prestígio social conferido pela população ou especialistas às ocupações ou como a posição relativa da ocupação em um *ranking* classificatório das mesmas, ordenado por algum indicador socioeconômico simples ou composto como rendimento proporcionado, nível de qualificação típico das ocupações ou mesmo ambas dimensões. Apostando na premissa que a hierarquização socioeconômica unidimensional das ocupações ainda é relevante, válida e útil na construção de classificações agregadas de ocupações - para uso em estudos sobre a estrutura ocupacional em uma perspectiva comparativa regional ou intertemporal e em estudos de mobilidade social, propõe-se aqui dois índices aproximativos para o status socioeconômico das ocupações levantadas nos Censos Demográficos de 1980, de 1991 e nas PNADs dos anos de 1990, tomadas conjuntamente. Com base nesta última fonte, computa-se diferentes indicadores do nível de precarização das ocupações brasileiras. Por fim, com base nesses resultados, apresenta-se uma escala socioocupacional para uso em estudos de Estratificação Socioeconômica e Mobilidade Social no Brasil.

Palavras-chave: Ocupação, classificação ocupacional, escala socioocupacional, mobilidade social.

1.Introdução

Status socioeconômico de uma ocupação tem sido tratado na Pesquisa Social como um conceito relacionado ao prestígio social conferido pela população às ocupações ou como a posição relativa da ocupação

[1] Este texto corresponde a uma versão revisada e ampliada daquela submetida e aprovada para apresentação no 7º Encontro Nacional de Estudos do Trabalho, em outubro de 2001, Salvador BA (Jannuzzi 2001a). Para desenvolvimento deste trabalho, contou-se com recursos para compra de material disponibilizado pela FAPESP (Proc. N. 00/09046-3) e bolsa de estudos e pesquisa da Fundação Ford (Grant 990-1161). Agradeço a Luiz Marcelo Ferreira Carvano/ENCE/IBGE pela produção e tratamento das bases de microdados utilizadas. Agradeço as correções e sugestões de aprimoramento do trabalho aos parecedistas anônimos.

* Endereço para correspondência: Professor visitante na Escola Nacional de Ciências Estatísticas(ENCE)/IBGE pelo convênio "Bolsa de estudos e pesquisa em Estatísticas Públicas" financiado pela Fundação Ford. Professor PUC-Campinas. E-mail: pjannuzzi@mpc.com.br ou pjannuzzi@ibge.gov.br .

em um *ranking* classificatório das mesmas, ordenado por algum indicador socioeconômico simples ou composto como rendimento proporcionado, nível de qualificação típico das ocupações, escolaridade da mão-de-obra nela alocada ou mesmo uma combinação destas dimensões (Valle Silva 1978, Jorrat & Acosta 1992, Bukstein 1997). Nesta perspectiva, por exemplo, Médico, Magistrado, Professor seriam consideradas ocupações de maior *status* que as de Trabalhador rural, Pedreiro ou Empregado doméstico já que na percepção subjetiva da sociedade seriam dotadas de maior prestígio social, assim como também apresentam indicadores objetivos de rendimento e escolaridade mais elevados.

Embora seja um conceito usual e de longa tradição na Sociologia americana e inglesa, o conceito e suas medidas parecem não ter tido um lugar muito destacado na agenda da Pesquisa Social no Brasil. O pioneiro e clássico trabalho "Posição social das ocupações" de Valle Silva (1978) e sua atualização posterior (Valle Silva 1985) constituem-se nas principais – para não dizer únicas- iniciativas mais conhecidas de computação de um índice de *status* socioeconômico para o conjunto das ocupações levantadas nos censos e pesquisas nacionais no País. É curioso que, tendo tido papel fundamental na construção das escalas socioocupacionais empregadas em três estudos clássicos da nossa Pesquisa Social-estudo da inserção de migrantes no mercado de trabalho metropolitano de Martine & Peliano (1978), de mobilidade social no Brasil de Pastore (1979) e de análise da mobilidade ocupacional da força de trabalho no Brasil realizado por técnicos do IBGE (1982), este tipo de trabalho não tenha sido replicado por outros autores em outras oportunidades. Tal fato talvez seja reflexo da baixa atividade de pesquisa no País no campo de estudos de Mobilidade e Estratificação Social, como apontado por Valle Silva (1999) e Vianna et al. (1998), e também conseqüência da descontinuidade dos estudos voltados à construção de sistemas classificatórios de ocupações no Brasil e da dificuldade de acesso e manipulação de informações desagregadas por categorias ocupacionais nas bases de dados mais adequadas para esses estudos como os Censos Demográficos e a Relação Anual de Informações Sociais (Médici 1989).

Naturalmente não se pode deixar de mencionar que se os trabalhos ancorados em medidas de *status* socioeconômicos para as ocupações não constituíram um programa de pesquisa progressivo nos anos de 1970 e décadas seguintes - no conceito empregado por Blaug (1999) para designar uma linha articulada e dinâmica de estudos por uma comunidade de pesquisadores - o mesmo não se pode dizer quanto aos estudos sobre sistemas mais abrangentes de classificação ocupacional a partir da década de 1980[2]. Emblemático neste sentido foi o trabalho de Jorge et al. (1984), que propunha um sistema de classificação ocupacional adaptável a diferentes necessidades de agregação das ocupações, baseados em critérios de similaridade interna das mesmas quanto ao tipo de controle da atividade produtiva, à propriedade dos meios de produção, às formas de inserção produtiva no processo de trabalho, ao nível de qualificação técnica exigido e ao setor de atividade da ocupação. Nessa perspectiva, não só era desnecessário como não fazia sentido computar medidas de *status* socioeconômico das ocupações, já que o critério de agregação de ocupações obedecia outra lógica substantiva, lógica essa que não pretendia chegar a um escalonamento unidimensional das ocupações (como a visada pelas medidas de *status*). Assim, já a partir do final da década de 1980 outros pesquisadores passaram a propor sistemas de classificação

[2] Tributários da tradição de hierarquização de grupos ocupacionais em termos de uma medida sintética de *status* baseado em rendimento e escolaridade são os trabalhos de Scalon (1999) e Jannuzzi (1999).

ocupacional empregando, em boa medida – com maior ou menor abrangência -, os critérios elencados por Jorge et al. (1984) como ilustram os trabalhos de Médici (1989), Valle Silva (1992), SEADE (1992), Oliveira (1993), Matos (1994), Barros et al. (1997), Ribeiro & Lago (2000).

Diferentemente destes últimos, o presente trabalho procura contribuir na linha anteriormente descrita de estudos classificatórios de ocupações com bases em medidas de *status* socioeconômico. Apostando na premissa que a hierarquização socioeconômica unidimensional das ocupações ainda é relevante, válida e útil na construção de classificações agregadas de ocupações - para uso em estudos sobre a estrutura ocupacional em uma perspectiva comparativa regional ou intertemporal e em estudos de mobilidade social – assim como para análises de mudanças de posição relativa das ocupações no espectro ocupacional propõe-se aqui dois índices aproximativos para o *status* socioeconômico das ocupações levantadas nos Censos Demográficos de 1980, de 1991 e nas PNADs dos anos de 1990, tomadas conjuntamente[3].

Para tanto, depois da apresentação da metodologia empregada para computação dos índices e de uma breve discussão sobre os resultados, avalia-se a validade dos mesmos como *proxies* do *status* socioeconômico das ocupações brasileiras, assim como outras propriedades. Em outra seção, mostram-se a relevância e a utilidade destes índices para estudos voltados à análise das mudanças estruturais do mercado de trabalho, mais especificamente à análise da precarização dos postos de trabalho, através da comparação dos mesmos com indicadores de vulnerabilidade das ocupações. Por fim, com base nesses resultados, apresenta-se uma escala socioocupacional para uso em estudos de Estratificação Socioeconômica e Mobilidade Social no Brasil, motivação primária para a realização do presente trabalho[4].

2. Método para cômputo dos índices socioeconômicos das ocupações brasileiras

Como discutido em trabalho anterior (Jannuzzi 1999), há várias metodologias sugeridas e empregadas para construção de um índice socioeconômico para ocupações e classificações socioocupacionais. Diferenciam-se pelas variáveis consideradas como critérios de ordenamento socioeconômico, pelas métricas com que as mesmas são expressas, pelas variáveis-controle consideradas, pela maior ou menor cobertura da fonte de dados usada para extrair as características ocupacionais, pelas técnicas estatísticas usadas e por outras decisões

[3] Para garantir maior precisão das estimativas mais recentes dos índices, consolidou-se em um só arquivo as PNADs entre 1992 a 1999. Se, por um lado, tal solução permite diminuir o problema da variabilidade amostral das estimativas, por outro, certamente, não garante que as mesmas estejam livres da tendenciosidade derivada do desenho amostral da pesquisa. Anualmente a PNAD coleta informação em cerca de 120 mil domicílios, sorteados em setores censitários aleatoriamente escolhidos no começo da década. A cada ano diferentes domicílios são selecionados neste conjunto de setores, tendo-se o cuidado de se acrescentar as novas edificações. Assim, as amostras anuais da PNAD não são, de fato, independentes.

[4] A construção da referida escala socioocupacional é uma das atividades necessárias e previstas para a realização da pesquisa "Mobilidade social e migração dos trabalhadores rurais no Brasil " (Jannuzzi 2000), cujo objetivo é o de analisar o padrão e intensidade da mobilidade socioocupacional da força de trabalho no Brasil, no contexto das mudanças estruturais da base produtiva e da conjuntura do mercado de trabalho na década de 1990, procurando identificar as especificidades do processo para distintos grupos sociodemográficos (homens, mulheres, chefes, cônjuges, migrantes, naturais, negros, não negros, grupos segundo coortes de nascimento, grupos segundo escolaridade, grupos segundo localidade de residência).

operacionais de menor alcance. Tal como em outras práticas de pesquisa nas Ciências Sociais, estas decisões metodológicas envolvem escolhas pragmáticas e preferências subjetivas do pesquisador, que podem conspirar contra a inteligibilidade, clareza interpretativa e validade do indicador produzido (em representar operacionalmente o *constructo* "*status* socioeconômico").

Uma das propostas metodológicas para construção de índices socioeconômicos de ocupações, que procura privilegiar a simplicidade técnica e facilidade interpretativa, garantindo- em tese- a validade de *constructo* da medida final, é a empregada por Jorrat & Acosta (1992) na computação de indicador semelhante para as ocupações na Argentina[5]. Por esta metodologia, o índice computado para uma ocupação corresponde a uma medida de posição relativa da mesma em um intervalo de 0 a 100. Mais precisamente o índice socioeconômico posicional (ISEP) corresponde à porcentagem de indivíduos ocupados cujos níveis médios combinados de rendimento e escolaridade são menores ou iguais aos da ocupação considerada. Assim, um ISEP de 82 % para uma determinada ocupação- como é o caso dos Protéticos em 1991 - significa que o conjunto de trabalhadores aí classificados apresenta um nível médio combinado de rendimento e escolaridade superior a 82% da população ocupada. Uma ocupação com ISEP de 25% - como os Serventes de Pedreiro em 1991- tem *status* socioeconômico mais baixo, já que os trabalhadores aí classificados teriam um nível combinado médio de rendimento e escolaridade superior a um contingente menor de ocupados (25%).

Para construção do ISEP relativo às ocupações registradas em cada base de microdados - Censos de 1980 e 1991 e PNADs dos anos 1990-, computou-se, primeiramente, para cada ocupação, as medianas de escolaridade e rendimento do trabalho principal das pessoas ocupadas de 15 a 64 anos, trabalhando 40 ou mais horas, com rendimentos válidos e escolaridade conhecida. Depois, calculou-se medidas de posição relativa das ocupações, segundo a escolaridade mediana observada, com base na distribuição de freqüências acumuladas das pessoas ocupadas em cada ocupação, ordenadas segundo nível de escolaridade. Repetiu-se o mesmo procedimento usando como critério de ordenamento o rendimento mediano. O ISEP é, então, calculado como média aritmética das duas medidas de posição relativas, daí o porquê ele representaria o *status* médio combinado de rendimento e escolaridade.

O uso da mediana como medida de tendência central para cômputo do ISEP justifica-se pela sua característica de menor sensibilidade a dados extremos como tipicamente ocorre com a coleta de informações sobre rendimento. Seu emprego também como estimativa da tendência central da escolaridade deve-se à necessidade de garantir compatibilidade metodológica com o tipo de estimativa de rendimento e, sobretudo, compatibilidade da própria variável nas três bases de dados, já que a todo indivíduo com 17 ou mais anos de estudo atribui-se o código 17. A consideração da faixa etária de 15 a 64 anos para cômputo das medidas teve o objetivo de privilegiar a parcela - majoritária - da população ocupada com inserção ocupacional mais claramente definida seja em termos de condição de atividade e ocupação, seja em termos de categoria ocupacional. Já a

[5] A metodologia empregada pelos autores é a proposta em estudo clássico de Nam & Powers na década de 1960.

consideração dos ocupados com 38 ou mais horas teve o objetivo de compatibilizar a duração da jornada sobre a qual se referia a remuneração[6] .

Para garantir maior precisão das estimativas mais recentes dos índices, consolidou-se em um só arquivo as PNADs entre 1992 a 1999. Se, por um lado, tal solução permite diminuir o problema da variabilidade amostral das estimativas, por outro, certamente, não garante que as mesmas estejam livres da tendenciosidade derivada do desenho amostral da pesquisa. Anualmente a PNAD coleta informação em cerca de 120 mil domicílios, sorteados em setores censitários aleatoriamente escolhidos no começo da década. A cada ano diferentes domicílios são selecionados neste conjunto de setores, tendo-se o cuidado de se acrescentar as novas edificações. Assim, as amostras anuais da PNAD não são, de fato, independentes. Os valores monetários do Censo de 1991 e PNADs dos anos de 1990 foi deflacionados para setembro de 1999 com base no INPC-Brasil e, para o Censo 1980, empregou-se o índice proposto por Ferreira & Barros (1999).

Uma outra estratégia para construção de um índice de *status*, que procura incorporar a "distância socioeconômica" entre as ocupações e não apenas a sua posição ordinal relativa, baseia-se na utilização de métodos multivariados de "redução" de dados, empregada por Scalon (1999) e por Ribeiro & Lago (2000) em grupos ocupacionais já agregados anteriormente por outros critérios substantivos. Nessa metodologia, duas ocupações com índices próximos devem apresentar níveis também próximos de rendimento e escolaridade medianos. Uma diferença grande entre os índices computados para duas ocupações, ao contrário, reflete um "distanciamento" significativo entre os níveis de rendimento e/ou escolaridade das mesmas. Do ponto de vista metodológico, a idéia básica nessa metodologia é computar um índice a partir do escore fatorial da primeira componente principal, obtida através da aplicação da análise de componentes principais sobre as duas dimensões socioeconômicas anteriormente explicitadas – rendimento e escolaridade medianos. Como estas dimensões são, em geral, altamente correlacionadas, a primeira componente principal tem capacidade de representar a maior parte da variabilidade do conjunto de dados (em termos de rendimento e escolaridade entre as ocupações), o que garante o emprego dos escores referentes a cada ocupação como uma medida sintética das duas variáveis. Como os escores podem variar em um intervalo amplo, com valores positivos e negativos, mediante uma transformação matemática simples, pode-se fazer correspondê-los a uma medida entre 0 e 1 (ou 0 a 100). Aplicando-se a técnica sobre as três bases de dados separadamente, tomando-se os escores fatoriais sobre a primeira componente principal (cujo poder explicativo da variabilidade foi de 84%, 86 % e 87%, para os Censos de 1980, 1991 e PNADs dos anos de 1990, respectivamente) e transformando-os para o intervalo de 0 a 1 obteve-se estimativas do Índice Socioeconômico Distancial - ISED - para cada ocupação em 1980, 1991 e 1996[7].

No anexo ao final do sexto, são trazidos o ISEP e o ISED, rendimento mediano, escolaridade mediana e totais de ocupados para cada ocupação em 1980, 1991 e anos de 1990.

---

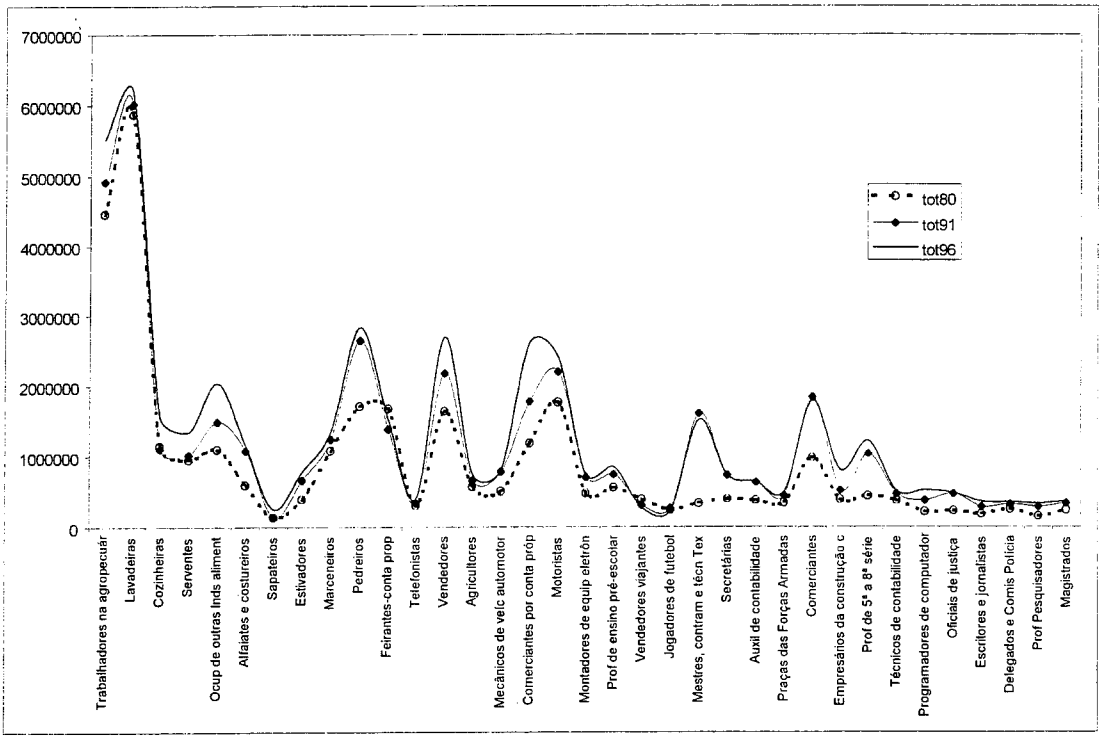
[6] O emprego destes critérios excluiu uma parcela de 25% das 42 271 526 pessoas ocupadas de 10 anos ou mais registradas no Censo de 1980 e 26 % das 55 293 317 ocupadas em 1991.

[7] Para garantir comparabilidade temporal do ISED, na transformação do escore fatorial no índice, tomou-se a função ISED = (Escore – Escore-min)/(Escore-max - Escore-min), onde Escore-min e Escore-max foram tomados como o menor e o maior valor de escore fatorial das ocupações nos três momentos (Escore-min = - 1,62 relativo a Lavadeiras e Passadeiras em 1980 e Escore-max = 13,95, correspondente a Magistrados em 1991).

3. Uma breve discussão sobre os resultados dos Índices

Como se poderia esperar, as ocupações de maior *status* – em qualquer dos três momentos ou índice - são as de nível superior, cargos da Alta Administração Pública e em postos de direção. Refletindo a estreiteza do cume da pirâmide social brasileira, estas ocupações apresentam ISEPs acima de 90 % . Na base do índice, com ISEP inferior a 20%, estão as ocupações manuais, na Agropecuária, Extrativismo e Serviços Domésticos, isto é, ocupações de baixa escolaridade e rendimento. Valores intermediários de ISEP correspondem às ocupações técnicas, de escritório, no comércio e semiqualificadas. Análise análoga pode ser realizada com o ISED, com resultados muito semelhantes, embora, neste caso, a diferença entre os índices de duas ocupações pode acrescentar um conhecimento adicional sobre a proximidade do nível médio combinado de rendimento e escolaridade.

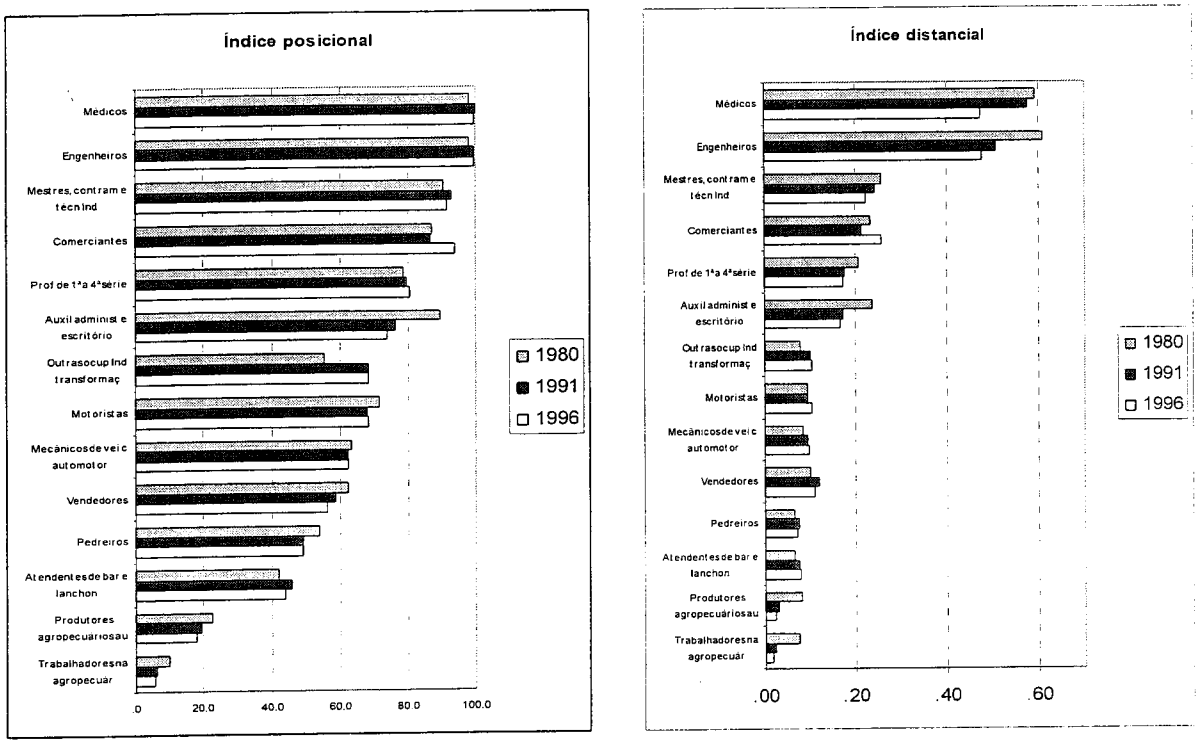
Gráfico 1 - Distribuição dos ocupados por ocupações ordenadas, segundo ISEP
Brasil 1980 - 1996



O comportamento dos índices ao longo do tempo é condizente com o que se poderia esperar. Quanto ao ISEP, pode se observar uma elevação sistemática da posição socioeconômica das ocupações de alto *status* ao longo do tempo e diminuição daquelas de baixo *status*, já que o índice é afetado não apenas por variações diferenciais de rendimento e escolaridade, mas também da distribuição de mão-de-obra na estrutura ocupacional. Assim, com a diminuição do pessoal ocupado como, por exemplo, Produtor Autônomo, há uma diminuição do *status* desta categoria e de todas as ocupações localizadas imediatamente acima; com aumento da parcela da mão-de-obra nas ocupações de médio-alto *status* – seja pela criação de novas ocupações, novos postos de

trabalho ou mobilidade ascendente – as ocupações imediatamente acima passam a ter um ISEP maior. O ISED é muito mais sensível a variações de rendimento ao longo do período, já que escolaridade é uma variável em geral crescente. A queda dos níveis de rendimento identificada no Censo Demográfico 1991 – causados, sobretudo, pela recessão induzida pelo Plano Collor e insucessos seguidos da política econômica em baixar os elevados níveis de inflação - revela-se através de níveis mais baixos do ISED de quase todas as ocupações, em relação aos valores levantados em 1980. Em meados da década de 1990, o *status* de várias categorias ocupacionais aumentou, como resultado da recuperação parcial no nível real dos rendimentos do trabalho, do aumento dos níveis de escolaridade da população ocupada e da própria seletividade crescente do mercado empregador na contratação de mão-de-obra de maior nível educacional.

Gráfico 2 : *Status* socioeconômico de ocupações selecionadas
Brasil 1980 - 1996



Outro aspecto a salientar é a forte correlação entre os índices nos três momentos em análise e entre as duas medidas computadas, ainda que neste último caso a relação não é linear. As ocupações de maior *status* têm rendimento mediano bem maior que as das demais ocupações, "distanciando-as" em termos do ISED das demais. Se tomar o coeficiente de correlação ordinal de Spearman, de forma a eliminar a não-linearidade da correspondência entre as ocupações de mais alto *status*, verifica-se que os dois índices tendem a classificar as ocupações nas mesmas posições relativas em um *ranking* ordinal.

Gráfico 3 - ISED x ISEP Brasil 1991

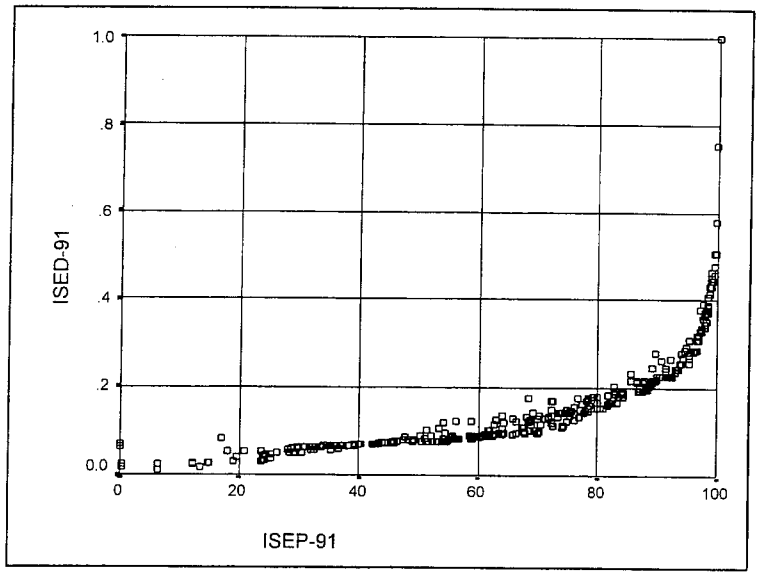


Tabela 1 - Matriz de correlação entre ISEPs e ISEDs
Brasil 1980 - 1996

Coeficiente Pearson						
	ISEP-1980	ISEP-1991	ISEP-1996	ISED-1980	ISED-1991	ISED-1996
ISEP-80	1,00	0,95	0,92	0,78	0,76	0,76
ISEP-91	0,95	1,00	0,95	0,76	0,79	0,79
ISEP-96	0,92	0,95	1,00	0,72	0,75	0,81
ISED-80	0,78	0,76	0,72	1,00	0,95	0,91
ISED-91	0,76	0,79	0,75	0,95	1,00	0,96
ISED-96	0,76	0,79	0,81	0,91	0,96	1,00
Coeficiente Spearman						
	ISEP-1980	ISEP-1991	ISEP-1996	ISED-1980	ISED-1991	ISED-1996
ISEP-80	1,00	0,96	0,93	0,97	0,95	0,92
ISEP-91	0,96	1,00	0,95	0,91	0,99	0,95
ISEP-96	0,93	0,95	1,00	0,88	0,95	0,98
ISED-80	0,97	0,91	0,88	1,00	0,92	0,89
ISED-91	0,95	0,99	0,95	0,92	1,00	0,95
ISED-96	0,92	0,95	0,98	0,89	0,95	1,00

4. Validade, robustez e consistência interna dos índices

Para uso metodologicamente consistente de um indicador socioeconômico na Pesquisa Social, esta medida deve gozar de uma série de propriedades (Kidder 1987, Rose et al. 2000, Jannuzzi 2001b). Dentre estas propriedades destacam-se a validade de *constructo*, isto é, o grau de proximidade entre o conceito idealizado e a medida operacional construída, a confiabilidade ou robustez do indicador, que pode lhe conferir a replicabilidade ao longo do tempo e, no caso de um indicador composto (índice), formado a partir de outros indicadores ou variáveis, há que se garantir também sua consistência interna, isto é, a concordância ou associação entre suas variáveis constituintes.

A aderência do ISEP e ISED com relação à replicabilidade pode ser atestada pela Tabela 1 e gráficos já apresentados: os índices são fortemente correlacionados nos três momentos analisados e suas mudanças de nível não são inconsistentes, já que podem ser explicadas com base nas modificações contextuais (queda da renda, aumento da escolaridade, mudanças na distribuição dos ocupados pela estrutura ocupacional, etc).

Pela forma como as medidas foram construídas era de se esperar que o ISEP e o ISED fossem internamente consistentes. Afinal, dentre as regularidades empíricas mais freqüentes reportadas nos Estudos do Trabalho figura, certamente, a associação entre escolaridade e rendimento. Aqui não é diferente, como se pode verificar pelos níveis elevados de correlação dos indicadores construídos com dados do Censo de 1991, na Tabela 2. A correlação dos índices, sobretudo do ISED, com o rendimento e escolaridade medianos das ocupações em 1991 é de magnitude elevada e próxima, revelando que as duas variáveis constituintes se fazem representar de forma relativamente equilibrada pelos indicadores.

Tabela 2 - Matriz de correlação entre os índices e outras variáveis socioeconômicas Brasil 1991

Coeficiente Pearson						
	ISEP-1991	ISED-1991	Escolaridade	Rendimento	Contribuintes Previdência	Cat. Valle Silva 9 cl
ISEP-1991	1,00	0,79	0,87	0,65	0,65	-0,83
ISED-1991	0,79	1,00	0,90	0,95	0,45	-0,82
Escolaridade	0,87	0,90	1,00	0,72	0,50	-0,87
Rendimento	0,65	0,95	0,72	1,00	0,39	-0,68
Contrib. Prev.	0,65	0,45	0,50	0,39	1,00	-0,47
Cat. VS 9 cl	-0,83	-0,82	-0,87	-0,68	-0,47	1,00
Coeficiente Spearman						
	ISEP-1991	ISED-1991	Escolaridade	Rendimento	Contribuintes Previdência	Cat. Valle Silva 9 cl
ISEP-1991	1,00	0,99	0,93	0,97	0,58	-0,87
ISED-1991	0,99	1,00	0,95	0,94	0,57	-0,87
Escolaridade	0,93	0,95	1,00	0,81	0,49	-0,85
Rendimento	0,97	0,94	0,81	1,00	0,61	-0,81
Contrib. Prev.	0,58	0,57	0,49	0,61	1,00	-0,48
Cat. VS 9 cl	-0,87	-0,87	-0,85	-0,81	-0,48	1,00

Legenda: Cat..VS 9 cl: Classificação ocupacional hierarquizada de Valle Silva (1992) em 9 classes.

A validade dos dois índices como *proxies* do *status* das ocupações na estrutura ocupacional ou social pode ser inferida a partir da regularidade e coerência do seu comportamento frente a outras variáveis socioeconômicas ou características ocupacionais. Algumas regularidades já foram apresentadas e discutidas anteriormente: por construção, os dois índices assinalam *status* mais elevados às ocupações de maior remuneração e qualificação (medida pela escolaridade); por construção, a correlação entre os índices e a escolaridade e/ou rendimento são elevados. Com tais associações era de se esperar, como mostra a Tabela 2, uma correlação significativa entre os índices e um indicador de nível de proteção social típico da ocupação (aqui representado pela proporção de contribuintes à Previdência).

Por fim, como mostra a última coluna da Tabela 2, os índices estão fortemente correlacionados com a escala hierárquica de ocupações em 9 categorias propostas por Valle Silva (1992)[9]. Ou seja, visto de outra forma (Tabela 3), os diferenciais de *status* socioeconômico estimados pelos índices das categorias ocupacionais de Valle Silva (1992) – de 9 ou 18 classes- revelam-se consistentes com o padrão esperado em termos da posição destas categorias ocupacionais quando se considera o setor de atividade a que se referem, a condição de proprietário, empregado e autônomo, o nível de qualificação, etc. Vide, neste sentido que Profissionais (1), Administradores e Gerentes (2) e Proprietários empregadores (3) apresentam índices de *status* mais elevados que, por exemplo, as categorias de Ocupações Manuais não-qualificadas (8) e de Trabalhadores rurais (9).

Tabela 3 - Estatísticas dos Índices de *status*, segundo categorias de Valle Silva (1992) Brasil 1991

Cat. VS	Cat .VS		ISEP- 91		ISED- 91	
9 classes	18 classes	Denominação	Média	Dv. Pad.	Média	Dv. Pad.
1 -Profissionais	1	Profissionais liberais	97,8	4,00	0,49	0,21
	3	Profissionais	97,2	2,00	0,35	0,07
2- Adm./Gerentes	2	Dirigentes	94,3	3,80	0,28	0,06
	4	Ocupações administrativas	91,9	3,10	0,24	0,03
3-Prop.empreg.	6	Proprietários empregadores	86,0	5,40	0,22	0,05
4-Não-manual	5	Não manual de rotina	78,5	9,80	0,18	0,04
	8	Técnicos e supervisores	81,3	12,90	0,19	0,07
5-Prop. cta. prop.	7	Empresários conta própria	74,2	7,00	0,13	0,04
6-Manual qualif.	9	Trabalhadores ind moderna	69,8	7,70	0,12	0,03
	11	Trabalhadores serviços	60,8	15,30	0,11	0,03
7-Empregador rur.	15	Proprietários agropecuários	68,4	12,70	0,12	0,04
8-Man. não-qual.	10	Trabalhad. ind. tradicional	45,1	14,00	0,07	0,02
	12	Trabalhadores domésticos	34,8	12,20	0,06	0,02
	13	Vendedores ambulantes	45,9	12,10	0,07	0,02
	14	Artesãos	8,8	12,80	0,04	0,02
9-Empreg. rurais	16	Técnicos e adm. agropecuar.	59,7	24,90	0,12	0,08
	17	Produt. Agropec. autonomos	19,2	.	0,03	.
	18	Trabalhadores rurais	22,0	12,30	0,04	0,02

Legenda: Cat. VS 9 cl: Classificação ocupacional hierarquizada de Valle Silva (1992) em 9 classes.
Cat. VS 18 cl: Classificação ocupacional de Valle Silva (1992) em 18 classes.

[9] A Escala de 9 classes de Valle Silva (1992) associa ao código 1 à categoria de maior rendimento e escolaridade; daí o sentido negativo dos coeficientes de correlação.

5. Uma validação adicional: índices de *status* e indicadores de precarização das ocupações

Em um contexto de aumento dos níveis de desemprego, do "conta-propismo", de perda de ocupações formais, da persistência de elevados níveis de rotatividade de pessoal e de sobrejornada semanal de trabalho, das tentativas e efetivos sucessos de desregulamentação do mercado de trabalho, enfim do que se convencionou denominar de precarização dos postos de trabalho poderia-se argumentar que índices de *status* socioeconômico, baseados unicamente nos rendimentos e escolaridade das ocupações, seriam medidas pouco sensíveis para captar ou pouco adequadas para retratar o que de importante e atual está acontecendo com a estrutura ocupacional brasileira nas últimas décadas. A crítica é certamente procedente, mas não inteiramente válida, já que não se pode esquecer que o rendimento é uma variável de ajuste que tem acompanhado – com maior ou menor aderência - as tendências acima.

Tabela 4 - Matriz de correlação entre Índices de *status* e medidas de precarização Brasil 1996

Coeficiente Pearson	ISEP	ISED	Contribuintes previdência	Ocupados com até 3 anos na ocupação	Ocupados com jornada semanal maior 48 horas	Fator relacionado ao risco de desemprego
ISEP	1,00	0,81	0,61	-0,44	-0,13	-0,28
ISED	0,81	1,00	0,47	-0,47	-0,15	-0,27
Contribuintes Previdência	0,61	0,47	1,00	-0,21	-0,34	0,06
Ocup. < 3 anos	-0,44	-0,47	-0,21	1,00	-0,06	0,35
Jorn. > 48 h	-0,13	-0,15	-0,34	-0,06	1,00	-0,18
Risco desemp.	-0,28	-0,27	0,06	0,35	-0,18	1,00
Coeficiente Spearman						
	ISEP	**ISED**	**Contribuintes previdência**	**Ocupados com até 3 anos na ocupação**	**Ocupados com jornada semanal maior 48 horas**	**Fator relacionado ao risco de desemprego**
ISEP	1,00	0,98	0,56	-0,52	-0,20	-0,49
ISED	0,98	1,00	0,54	-0,49	-0,24	-0,47
Contribuintes Previdência	0,56	0,54	1,00	-0,27	-0,41	-0,03
Ocup. < 3 anos	-0,52	-0,49	-0,27	1,00	0,03	0,65
Jorn. > 48 h	-0,20	-0,24	-0,41	0,03	1,00	-0,17
Risco desemp.	-0,49	-0,47	-0,03	0,65	-0,17	1,00

Há certamente necessidade de se propor e desenvolver índices e escalas de precarização e vulnerabilidade da mão-de-obra, como os propostos em Seade (1992) e Sabóia (1999), que possam servir instrumentos analíticos para aprofundar o entendimento das transformações correntes do mercado de trabalho. Embora não seja o objetivo deste texto propor uma medida síntese, neste sentido, traz-se, em anexo, um conjunto de indicadores de

precariedade (ou vulnerabilidade) ocupacional, computados através da série dos anos de 1990 da PNAD, e que se prestam a validar – em alguma medida, como mostrado em seguida- os índices de *status* anteriormente apresentados[9].

Como se pode observar na Tabela 4, há uma correlação significativa (especialmente com respeito ao Coeficiente de Spearman)[10] dos índices de *status* e algumas das dimensões operacionalizáveis da precarização do posto de trabalho, como o nível de formalização ou proteção social do trabalho (indicado pela proporção de contribuintes à Previdência), a rotatividade do pessoal ocupado (indicado pela proporção de ocupados com até 3 anos na ocupação) e risco ao desemprego (calculado como a razão de desempregados egressos de cada ocupação pelo total de ocupados nas mesmas). A jornada excessiva de trabalho (proporção de ocupados com jornada superior a 48 horas semanais) é, entre os indicadores de precarização, o menos associado com as medidas de *status*.

Assim, se por estes resultados não se pode negar a inadequação do ISEP e ISED como *proxies* do grau de vulnerabilidade das ocupações, não se pode deixar de reconhecer-lhes alguma capacidade de retratá-las.

6. Uma proposta de agregação de ocupações para estudos de mobilidade social

As classificações ocupacionais agregadas- orientadas segundo a similaridade das tarefas desenvolvidas nas ocupações ou setor de atividade das mesmas, grau de controle, forma de inserção no processo produtivo, nível de qualificação ou habilidades requeridas, como as classificações que agrupam ocupações em Técnicas, Manuais, Diretivas ou em Operários industriais, Trabalhadores do Comércio, Proprietários ou ainda em Ocupações qualificadas, semiqualificadas, de baixa qualificação - cumprem uma função metodológica básica nos estudos empíricos em Economia e Sociologia do Trabalho pela possibilidade que conferem ao analista na investigação das características e mudanças não apenas da estrutura ocupacional, mas também da estrutura social de uma sociedade. Como argumentam diferentes autores, a ocupação é um *"constructo operacional"* para identificação da posição dos indivíduos - e suas famílias - na estrutura social de uma sociedade, assertiva válida seja em uma perspectiva marxista – em que as relações de classe estariam determinadas pelas relações de produção, manifestada pela posse ou não dos meios de produção - seja em uma perspectiva weberiana – em que a estrutura de classes resultaria das desigualdades de oportunidades dos indivíduos frente ao mercado (Valle Silva 1992, Bukstein 1997, Jorrat 1998, Scalon 1999, Rose et al. 2000).

Um tipo particular de classificação ocupacional, de larga tradição nos estudos desta natureza por parte de autores anglo-saxões (Desrosières & Thévenot 1996), é aquela que procura agrupar as ocupações, segundo suas

[9] Para cômputo desses indicadores procedeu-se à junção das PNADs dos anos de 1990, como descrito anteriormente. Vale observar que o erro amostral das estimativas de alguns dos indicadores pode ser extremamente elevado, pela rarefação das ocupações na amostra considerada.

[10] Pela magnitude do erro amostral das estimativas dos indicadores de precarização o Coeficiente de Spearman é uma medida mais adequada (robusta) para avaliar a correlação dos mesmos com os índices de *status*.

similaridades em termos de nível de rendimento proporcionado e de escolaridade requerida (ou observada), denominadas de escalas socioeconômicas de ocupações ou ainda escalas socioocupacionais. Com base em uma medida sintética – derivada da combinação do rendimento e escolaridade por algum método – as ocupações são agrupadas e hierarquizadas em estratos ou categorias de diferentes *status* socioeconômico.

Assim, com base na ordenação das ocupações proporcionada pelos índices socioeconômicos obtidas para 1991, nos indicadores de precarização dos postos de trabalhos dos anos 1990, na tipologia de grupos ocupacionais de Valle Silva (1992) organizados, segundo diversos critérios substantivos (diferenças e similaridades das ocupações em relação ao caráter urbano/rural, manual/não-manual, aos setores de atividade - serviços, indústrias modernas e tradicionais - nível de qualificação - técnicos , profissionais de nível superior-, nível de controle e autonomia – empregador, empregado e conta-própria), na proposta metodológica de estratificação social segundo grupos ocupacionais de Goldthorpe (1992) e não perdendo de vista as restrições operacionais inerentes ao uso de uma pesquisa amostral como a PNAD (em termos de qualidade da informação captada sobre ocupação e do erro amostral das estimativas) desenvolveu-se uma escala socioeconômica de cinco categorias para as ocupações brasileiras[11].

Quadro 1: Algumas ocupações típicas dos estratos socioocupacionais

Estrato socioocupacional	Ocupações típicas
1 Alto	**Médico, Engenheiro, Professor universitário, Empresários, Gerentes e postos superiores na administração pública (Juízes, Promotores, Delegados, Oficiais das forças armadas, etc.).**
2 Médio-alto	**Técnicos de contabilidade e administração, Mestre e Contramestres na indústria, Professores de ensino fundamental e médio, Corretores de imóveis, Inspetores de polícia, Carteiros, Comerciantes (proprietários) e Agricultores.**
3 Médio	**Torneiro mecânico, Montadores de equipamentos elétricos, Vendedores, Operadores de caixa, Comerciantes conta-própria, Professores de ensino pré-escolar, Motoristas, Inspetores de alunos, Auxiliares de enfermaria, Auxiliares administrativos e de escritório, Policiais e Praças das forças armadas.**
4 Médio-baixo	**Ocupações da indústria de alimentos, ocupações da indústria do têxtil, pedreiros, Pintores, Garçons, Vigias, Porteiros, Estivadores, Vendedores ambulantes.**
5 Baixo	**Trabalhadores rurais na condição de empregados ou autônomos (produtores meeiros ou parceiros), além das ocupações urbanas de baixo *status* como a de Serventes de pedreiro, Lavadeiras, Empregados domésticos e Lixeiros.**

O primeiro grupo social da escala compreende os indivíduos na condição de grandes proprietários e as pessoas ocupadas em postos de comando, de direção ou com especialização técnica superior. Reúne, portanto, as ocupações com rendimento e escolaridade mais elevadas, isto é, de maior *status* (ISEP médio de 96%) e menor grau de precarização. Este estrato corresponderia em alguma medida ao que Goldthorpe (1992) denominou de classes de colarinho branco na classificação socioocupacional da população inglesa, nas quais se enquadrariam os indivíduos com maior poder de comando da atividade produtiva, de delegação de autoridade ou aqueles com alto grau de conhecimento especializado. Na classificação de ocupações adotada pelo IBGE são típicos

[11] A codificação da escala socioocupacional para cada ocupação é apresentada na primeira coluna da tabela no anexo (cat).

representantes deste estrato social os indivíduos que exercem ocupações como de Médico, Engenheiro, Professor Universitário, Empresários, Gerentes e postos superiores na Administração Pública (Juízes, Promotores, Delegados, Oficiais das Forças Armadas, etc.). Rendimento mediano acima de R$ 3 000,00 (em valores de setembro de 1999), escolaridade de 16 anos ou mais, tempo médio de trabalho acima de 10 anos, baixo risco ao desemprego são atributos não incomuns das ocupações aqui enquadradas.

O segundo estrato (ISEP de 85%) reúne, em boa medida, o que Goldthorpe denomina de 'pequena burguesia', isto é, os pequenos proprietários, chefes e supervisores e empregados qualificados de escritório e técnicos de média especialização na indústria e serviços. Técnicos de contabilidade e administração, Mestre e Contramestres na indústria, Professores de Ensinos fundamental e médio, Corretores de imóveis, Inspetores de polícia, Carteiros, Comerciantes (proprietários) e Agricultores são algumas das ocupações enquadradas neste grupo. Com menor poder de comando ou qualificação, estas ocupações apresentam um rendimento médio bem menor, ainda que com relação aos indicadores de escolaridade e precariedade as diferenças não sejam tão expressivas. Compare-se, por exemplo, os indicadores de duas ocupações com relativa similaridade funcional como a de Professor universitário (estrato socioocupacional alto) e Professor de ensino médio (estrato médio-alto): o rendimento médio dos primeiros (Professor universitário) é cerca de R$ 2 120, duas vezes e meia maior que os dos últimos (R$ 799); a escolaridade mediana é de 16 anos entre os professores de ensino superior contra 15 anos dos professores de ensino médio; a proporção de contribuintes à Previdência Pública é superior a 94% nos dois grupos; o risco ao desemprego é inferior a 3% para os dois grupos.

O terceiro estrato (ISEP de 69%) reúne a grande maioria de ocupações no Comércio, Serviços e postos qualificados da indústria, ao que se poderia associar - ainda que remotamente - à 'elite de colarinho azul' referida pelo autor. Ocupações típicas deste grupo são as de Torneiro mecânico, Montadores de equipamentos elétricos, Vendedores, Operadores de caixa, Comerciantes conta-própria, Professores de ensino pré-escolar, Motoristas, Inspetores de alunos, Auxiliares de enfermaria, Auxiliares administrativos e de escritório, Policiais e Praças das forças armadas. Os indicadores de precarização ocupacional são significativamente piores: a rotatividade é muito elevada (46% dos ocupados estavam empregados há até 3 anos), um quarto dos ocupados aí enquadrados têm uma jornada superior a 48 horas semanais e o risco ao desemprego é maior.

Tabela 5: Indicadores socioeconômicos dos estratos socioocupacionais Brasil 1991 e 1996

Estrato sócio-ocupacional	ISEP 1991 Média	ISEP 1991 Desvio-Padrão	ISED 1991 Média	ISED 1991 Desvio-padrão	Contribuintes previdência (%)	Ocupados com até 3 anos na ocupação (%)	Ocupados com jornada semanal maior 48 horas (%)	Fator relacionado ao risco de desemprego	Rendimento (R$ de set./2000)	Escolaridade (anos)
1 Alto	95,7	3,9	0,34	0,11	79,8	27,0	18,1	3,4	1386,47	13,2
2 Médio-alto	85,2	7,1	0,20	0,04	73,2	32,1	18,9	5,9	687,19	10,5
3 Médio	67,2	7,5	0,12	0,03	60,8	46,7	25,1	9,3	369,86	7,5
4 Médio-baixo	45,2	9,9	0,07	0,01	43,8	49,9	19,9	11,4	251,35	4,59
5 Baixo	14,8	7,5	0,03	0,01	12,2	38,5	22,0	9,7	121,13	2,48

O quarto grupo socioocupacional (ISEP de 45%), mais heterogêneo, compreende os empregados em prestação de serviços de baixa qualificação nos serviços, construção civil e indústria tradicional – algo que talvez se poderia associar, com alguma boa vontade - à classe operária' de Goldthorpe. Aqui estão reunidas as ocupações da Indústria de alimentos, da Indústria têxtil, Pedreiros, Pintores, Garçons, Vigias, Porteiros, Estivadores. Alguns indicadores de precarização chegam a ser ainda piores que os do estrato anterior: menos da metade contribui para Previdência, metade tem sobre jornada e o risco ao desemprego é mais elevado.

O último grupo socioocupacional (ISEP de 18%) reúne os trabalhadores rurais na condição de empregados ou autônomos (produtores meeiros ou parceiros), além das ocupações urbanas de baixo *status* como a de Serventes de pedreiro, Lavadeiras, Empregadas domésticas e Lixeiros. Compreende pois a parcela da mão-de-obra – empregada ou autônoma - de remuneração e escolaridade mais baixa, inseridos em postos de trabalho de elevada precariedade[12]. O caso dos Empregados Domésticos (código 807 na classificação do IBGE) é sintomático neste sentido: pelos dados das PNADs dos anos de 1990, o rendimento mediano era de R$ 136 (em valores de setembro de 1999), somente um quarto contribuía para Previdência Pública, 65% estavam no trabalho atual há menos de 3 anos, quase um terço tinha jornada semanal superior a 48 horas, 22% eram o risco de se vir a se desempregar.

Dessas observações apreende-se que o poder discriminatório dos níveis de rendimento e escolaridade na classificação das ocupações vai decrescendo dos estratos socioocupacionais mais elevados em direção aos mais baixos, crescendo, em contrapartida, a importância das medidas de precarização ocupacional como rotatividade, contribuição à Previdência, risco ao desemprego.

[12] O risco ao desemprego e a rotatividade do conjunto de ocupações aí inseridos só não são maiores certamente pelas características da mobilidade ocupacional dos trabalhadores rurais.

Tabela 6: Correspondência entre a escala de Valle Silva (1992) e estratos socioocupacionais Brasil 1991

Cat. VS 9 classes	Cat. VS 18 classes		Escala socioocupacional 1-A	2-MA	3-M	4-MB	5-B	Total
1 – Profissionais	1	Profissionais liberais	100,0					100,0
	3	Profissionais	86,7	13,3				100,0
2- Adm./Gerentes	2	Dirigentes	40,0	60,0				100,0
	4	Ocupações administrativas	29,7	70,3				100,0
3-Prop. empreg.	6	Proprietários empregadores	22,8	77,2				100,0
4-Não-Manual	5	Não manual de rotina	0,5	32,4	67,1			100,0
	8	Técnicos e supervisores	2,9	55,4	39,0	2,7		100,0
5-Prop. cta. prop.	7	Empresários conta própria		5,1	94,9			100,0
6-Manual qualif.	9	Trabalhadores ind. moderna		2,3	97,7			100,0
	11	Trabalhadores serviços		0,2	76,4	21,3	2,2	100,0
7-Empregdor rural	15	Proprietários agropecuários	1,6	97,8		0,6		100,0
8-Man. não qual.	10	Trabalhad. ind. tradicional		0,4	0,3	87,9	11,4	100,0
	12	Trabalhadores domésticos			4,8	45,8	49,4	100,0
	13	Vendedores ambulantes			12,4	81,5	6,1	100,0
	14	Artesãos				55,5	44,5	100,0
9-Empreg. rurais	16	Técnicos e adm. agropecuar.	0,6	8,9		90,6	.	100,0
	17	Produt. agropec. autônomos					100,0	100,0
	18	Trabalhadores rurais				4,1	95,9	100,0
	total		5,5	12,6	26,8	31,3	23,8	100,0

Legenda: Cat. VS 9 cl: Classificação ocupacional hierarquizada de Valle Silva (1992) em 9 classes
Cat. VS 18 cl: Classificação ocupacional de Valle Silva (1992) em 18 classes

Como era de se esperar pela forma de construção, a escala socioocupacional apresenta-se com alto grau de associação com o sistema classificatório de Valle Silva, tomado com 9 ou 18 grupos ocupacionais. Para citar alguns exemplos, como mostrado na Tabela 5, o conjunto de ocupações classificadas como ocupações "Profissionais" do autor insere-se completamente dentro do primeiro nível da escala (Alto); as "Ocupações Não - Manuais de Rotina" e "Manuais Qualificadas" enquadram-se em ampla maioria no estrato médio; as ocupações da "Indústria tradicional" estão concentrados no estrato médio-baixo.

Uma evidência adicional do grau de concordância da escala proposta com a de Valle Silva é que na realização de uma análise discriminante confirmatória, tendo como variáveis preditoras o ISEP, o ISED e a classificação hierarquizada de nove grupos de Valle Silva, a proporção de ocupações corretamente assignadas aos seus estratos atingiu 83 %. Assim, talvez se pudesse tomar a presente escala como uma proposta de agregação ainda mais condensada (em cinco categorias) para o sistema classificatório de 9 ou 18 grupos do autor.

Com base da estratificação socioocupacional proposta à estrutura social, teria apresentado uma mudança significativa nos últimos 20 anos, com forte redução do grupo situado na base (como resultado da desruralização da mão-de-obra) e aumento dos segmentos médios, como se pode verificar na Tabela 6. Aparentemente, as mudanças na estrutura socioocupacional foram mais marcantes na década de 1980 que nos últimos 10 anos, algo que os dados do Censo 2000 permitirão verificar com maior clareza.

Tabela 7 - Evolução dos estratos socioocupacionais da população ocupada
Brasil 1980 – 1996

	Estrato socioocupacional	1980	1991	1996
1	Alto	4,5	5,5	5,9
2	Médio-alto	9,9	12,6	11,9
3	Médio	22,3	26,8	27,3
4	Médio-baixo	27,1	25,8	26,3
5	Baixo	36,2	29,3	28,6
	Total	100,0	100,0	100,0

6. Considerações finais

A necessidade de dispor de uma escala socioocupacional para estudar a Mobilidade Social de trabalhadores rurais e de outros segmentos sociodemográficos no País ao longo do Século XX, passível de ser adaptada às características e limitações dos levantamentos amostrais retrospectivos em que informações relativas a esta temática foram investigados, levou a necessidade de computação de medidas de *status* socioeconômico para as ocupações brasileiras (Jannuzzi 2000). Ainda que estes últimos – os índices - sejam a questão central deste trabalho – daí porque figurem no título- a maior utilidade deles no projeto de pesquisa que suscitou seu desenvolvimento foi chegar à escala socioocupacional apresentada na seção anterior. Tal observação é pertinente para justificar a brevidade da exploração analítica das medidas tão trabalhosamente computadas.

De qualquer forma, o esforço operacional de obter informações detalhadas das mais de 300 ocupações a partir dos microdados dos Censos Demográficos 1980 e 1991 e das PNADs de 1992 a 1999, assim como o esforço metodológico de construção dos índices ISEP e ISED e dos indicadores de precarização das ocupações pareceu suficientemente grande para justificar a organização deste texto e deixar uma contribuição de valor ao menos empírico a outros pesquisadores interessados em temas correlatos. De fato, os resultados apresentados em anexo trazem um conjunto de informações que podem vir a ser úteis na análise do comportamento do rendimento, da escolaridade, do pessoal ocupado, do grau de precarização das diversas denominações ocupacionais empregadas pelo IBGE.

Por fim, mas não menos importante: a maior validação que se pode conferir aos índices de status e à escala socioocupacional de cinco grupos é a utilidade instrumental dos mesmos em estudos da Estrutura ocupacional, Estratificação social ou Mobilidade Social. Parodiando os clássicos da Metodologia da Pesquisa Social... que estes recursos analíticos "sejam julgados mais pelas conseqüências e resultados de sua utilização que pelas supostas limitações teóricas de sua concepção".

Referências bibliográficas

BARROS, R. *et al. A desigualdade da pobreza: estratégias ocupacionais e diferenciais por gênero.* (Texto para discussão n. 453). Rio de Janeiro: IPEA, 1997, 25p.

BLAUG, M. *Metodologia da economia.* São Paulo: EDUSP, 1999.

BUKSTEIN,G. Medidas de prestígio ocupacional: aproximación a la medición del *status* socioeconómico de las ocupaciones. *Estudios del Trabajo*, Buenos Aires, 14: 93-114, 1997.

DESROSIÈRES, A. & THÉVENOT, L. *Les catégories socioprofessionnelles.* Paris: La Decouverte, 1996.

FERREIRA, H.G. & BARROS, R. *The slippery slope: explaining the increase in extreme poverty in urban Brazil: 1976-1996* (Texto para discussão 404). Departamento de Economia/PUC-Rio, 1999.

GOLDTHORPE, J.H. *Social mobility and class structure in modern britain.* New York: Oxford University Press, 1992.

IBGE. *Força de trabalho no Brasil: uma análise de mobilidade ocupacional.* Rio de Janeiro, 1982.

JANNUZZI, P. M. Construção de uma escala sócio-ocupacional: notas metodológicas. *Revista Brasileira de Estatística*, Rio de Janeiro, 60(214):7-24, 1999.

_____. *Mobilidade Social e Migração dos Trabalhadores Rurais no Brasil.* (Projeto de pesquisa em Estatísticas Públicas – Convênio ENCE/FORD), 2000.

_____. *Status* socioeconômico das ocupações brasileiras: medidas aproximativas para 1980, 1991 e anos 90. *Anais do 7° Encontro Nacional de Estudos do Trabalho* (cd-rom), outubro de 2001, Salvador BA,2001a.

_____. *Indicadores sociais no Brasil: conceitos, fonte de dados e aplicações.* Campinas: Alínea, 2001b.

JORGE, A. *et al.* Categorias sócio-ocupacionais: uma perspectiva para análise da força de trabalho e da distribuição de rendimentos no Brasil. *Anais do IV Encontro Nacional de Estudos Populacionais*, São Paulo, ABEP, vol.1,p.77-110, 1984.

JORRAT, J.R. & ACOSTA,L.R. Aproximaciones a la medición del *status* socioeconómico de las ocupaciones en Argentina. *Estudios del Trabajo*, Buenos Aires, 4: 79-106, 1992.

JORRAT, J.R. Modelos predominantes de estrutura de clases y su rendimiento empírico: un estudio del Area Metropolitana de Buenos Aires. *Estudios del Trabajo*, Buenos Aires, 16: 3-48,1998.

KIDDER, L.H. *Métodos de pesquisa nas relações sociais - Selltiz, Wrightsman e Cook.* São Paulo, EPU, 1987.

MARTINE, George & PELIANO, J.C. *Migrantes no mercado de trabalho metropolitano.* Rio de Janeiro: IPEA, 1978 (Série Estudos para o Planejamento n.19).

MATOS, R.E.S. *Distribuição espacial da população de Minas Gerais e tendências de desconcentração nas áreas de influência de Belo Horizonte.* Belo Horizonte, CEDEPLAR/FACE/UFMG, 1994 (Tese de Doutorado).

MÉDICI, André C. *Urbanização e estrutura ocupacional: alternativas metodológicas para uma investigação.* Rio de Janeiro: IEI/UFRJ, 1989 (Texto para discussão IEI/UFRJ n.202).

OLIVEIRA, J.S. O *Traço da desigualdade social no Brasil.* Rio de Janeiro: IBGE, 1993.

PASTORE, J. *Desigualdade e mobilidade social no Brasil.* São Paulo: T.A. Queiroz/EDUSP,1979.

RIBEIRO, L.C.Q. & LAGO. L. C. O espaço social das grandes metrópoles brasileiras: São Paulo, Rio de Janeiro e Belo Horizonte. *Revista Brasileira de Estudos Urbanos e Regionais*, Recife, 3:111-130, novembro, 2000.

ROSE, D. *et al. Towards a european socio-economic classificacion : final report to Eurostat of expert group.* University of Essex, 2000 (mimeo).

SABOIA, J. L. M. Um Novo Indice para o Mercado de Trabalho Urbano no Brasil *Anais do VI Encontro de Estudos do Trabalho.* Belo Horizonte: ABET, 1999. v.1. (cd-rom).

SEADE. *Pesquisa de Condições de Vida: Mercado de trabalho.* São Paulo, 1992.
SCALON, M.C. *Mobilidade social no Brasil: padrões e tendências.* Rio de Janeiro: IUPERJ/Revan, 1999.

VALLE SILVA, N. *Posição social das ocupações.* Rio de Janeiro: IBGE, 1978 (Mimeo).

_____. *Atualização da escala socioeconômica de ocupações para 1980.* Rio de Janeiro: LNCC, 1985 (Relatório de pesquisa e desenvolvimento).

_____. *Uma proposta de classificação das ocupações brasileiras.* LNCC, 1992 (Mimeo).

_____. Mobilidade social. In: MICELI, S. (org) *O que ler na ciência social brasileira (1970-1995).* São Paulo,:Ed. Sumaré, 1999.

VIANNA, L.W. *et al.* Doutores e teses em Ciências Sociais. *Dados – Revista de Ciências Sociais*, Rio de Janeiro, 41(3): 453-516, 1998.

Abstract

Social economic status of an occupation has been treated in the Social Research as a concept related to the social prestige granted by the population or experts or as the position of the occupation in a ranking ordered by some simple or composite variable such as income, level of qualification of the occupations or both dimensions. As such ranking of occupations still seems appropriate, valid and useful in the construction of classification systems - to be used in studies on the occupational structure, labor market comparative analysis over different regions and social mobility - this paper presents two measures of occupational status, based on data collected in the Demographic Censuses of 1980 and 1991 and in the National Household Surveys executed in the nineties. It is also presented some characteristics of the occupations. Finally, it is presented an aggregate classification of brazilian occupations to be used on Social Stratification and Social Mobility in Brazil.

Key words: Occupation, occupational groups, social social economic scale, social mobility.

ANEXO

ANEXO : Índices de *status*, rendimento, escolaridade, total de ocupados e indicadores de precarização das ocupações brasileiras: 1980, 1991 e 1996

Cat. Ocu.	Cód. ocupacional	Denominação	Censo 1980					Censo 1991					PNADs anos 1990 (1996)									
			ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	% Contr. Prev. Publ.	Jornada semanal média (horas)	Tempo médio ocup. (anos)	% pessoal temp<3 anos	% pessoal jorn>48 h/sem	Fat. Rel. risco de desocup. no ano
2	1	Agricultores	59.8	0.08	532.00	3	161	60.9	0.09	363.75	4	181 254	63.7	0.10	515.77	4	186 274	21.1	46	19	9.2	40.4	1.2
2	2	Criadores de gado bovino	82.3	0.16	1197.00	4	70 359	73.2	0.13	682.02	4	123 902	68.4	0.14	838.12	4	117 372	30.4	46	20	7.2	43.7	0.7
2	3	Avicultores e criad. peq. animais	84.5	0.20	1596.00	4	3 995	79.8	0.16	727.49	6	7 772	76.3	0.18	1059.9	5	7 483	39.6	50	13	20.1	56.5	0.7
2	4	Criadores de outros animais	82.4	0.18	1330.00	4	1 016	61.9	0.09	363.75	4	4 601	80.5	0.19	953.88	7	10 682	28.8	47	20	6.8	45.9	0.2
2	5	Prop. em ativ. agrop. não espec.	84.5	0.20	1596.00	4	4 731	79.8	0.16	727.49	6	2 422	70.6	0.17	1133.7	4	1 510	37.4	45	22	3.5	41.3	5.2
1	6	Empresários da extr. veg. e pesca	67.5	0.12	957.00	3	5 020	77.4	0.16	909.36	4	5 261	65.6	0.11	635.92	4	7 535	25.3	51	10	32.9	55.9	3.4
1	7	Empresários da extração mineral	84.5	0.20	1596.00	4	4 460	75.3	0.14	727.49	4	7 243	69.2	0.16	1031.5	4	8 249	46.9	48	11	16.9	37.5	2.1
1	8	Empresários da Ind. transformação	88.1	0.26	1596.00	7	157	87.9	0.22	909.36	8	254 383	87.0	0.25	1093.4	10	319 451	75.1	51	9	23.8	48.4	1.9
1	9	Empresários da construção civil	82.5	0.18	1330.00	4	36 751	89.2	0.25	1136.70	8	32 266	78.7	0.17	913.05	6	104 545	58.2	48	12	17.1	41.1	1.8
2	10	Comerciantes	87.2	0.23	1383.00	7	287	86.6	0.22	909.36	8	762 982	94.0	0.26	1093.4	11	725 911	76.6	53	9	24.8	55.4	2.4
2	11	Hoteleiros e donos de pensão	84.6	0.20	1596.00	4	8 936	82.6	0.20	909.36	7	12 546	84.9	0.22	1093.4	8	11 910	66.0	55	11	15.7	56.9	0.3
1	12	Empresários nos transportes	85.5	0.25	1862.00	5	18 094	89.6	0.28	1364.05	8	20 985	87.2	0.27	1500.0	8	36 686	69.7	50	10	21.3	49.4	1.9
2	13	Vendedores viajantes	91.2	0.27	1596.00	8	123	73.9	0.11	418.31	5	6 927	75.6	0.14	671.45	6	5 245	24.7	50	9	21.8	56.4	0.3
2	14	Feirantes	.	.	.	.	.	78.2	0.14	682.02	5	4 645	82.6	0.18	750.00	8	6 725	47.1	51	11	21.4	53.9	2.9
2	15	Outros Proprietários	.	.	.	.	.	90.6	0.26	1045.77	10	287 715	96.1	0.28	1275.5	11	321 965	72.2	48	9	25.6	42.3	1.5
1	20	Altos dirigentes públicos	95.8	0.33	1596.00	11	7 890	95.4	0.27	1045.77	11	20 471	97.0	0.34	1749.4	11	19 032	60.5	35	5	45.9	17.3	2.0
1	21	Direi., assess. e chefes serv. Púb.	94.9	0.33	1330.00	13	83 580	94.2	0.26	818.43	12	148 973	94.0	0.25	893.67	12	249 321	88.7	39	10	25.7	7.1	3.3
4	30	Admin. e gerent. agropecuária	47.6	0.06	354.00	3	98 284	45.3	0.07	259.17	4	97 463	41.5	0.06	272.00	4	124 146	53.5	52	8	30.4	49.3	1.4
1	31	Admin. e gerent. extr veg e pesca	68.7	0.10	532.00	4	2 104	93.2	0.24	818.43	11	2 069	63.9	0.10	515.77	4	766	51.4	48	7	45.1	30.8	1.7
1	32	Admin. e gerent. extração mineral	96.0	0.37	2075.00	11	2 523	96.5	0.29	1136.70	11	3 628	96.5	0.29	1300.0	11	3 669	69.7	48	9	28.7	37.6	1.0
1	33	Admin. e gerent Ind transformação	96.4	0.41	2414.00	11	96 715	96.8	0.32	1364.05	11	141 564	95.5	0.28	1257.4	11	199 095	87.5	45	7	34.0	22.6	6.4
1	34	Admin. e gerent. Ind const civil	96.0	0.37	2021.00	11	12 866	97.6	0.36	1364.05	14	14 906	97.1	0.31	1312.1	13	16 933	81.3	46	7	37.6	26.4	8.4
2	35	Admin. e gerent. comércio	86.6	0.21	957.00	8	144	90.5	0.22	682.02	11	294 033	88.0	0.21	688.91	11	336 221	80.3	49	6	42.2	32.2	8.6
2	36	Admin. e gerent. hotéis	85.4	0.19	745.00	8	9 829	89.3	0.21	591.09	11	13 179	82.5	0.18	600.00	10	13 975	77.3	54	6	36.2	52.1	5.9
1	37	Admin. e gerent. transportes	95.8	0.33	1596.00	11	13 847	96.0	0.28	1100.33	11	18 462	95.2	0.27	1200.0	11	24 962	87.4	47	7	34.7	28.2	6.3
1	38	Admin. e gerent. Emp. financ./imobil	96.7	0.43	2660.00	11	56 955	97.7	0.39	1818.73	12	78 996	98.2	0.43	2119.7	14	100 875	95.5	42	12	17.4	16.8	4.9
1	39	Out. Admin. e ger. não classif. nat.	93.7	0.30	1330.00	11	69 891	96.3	0.29	1136.70	11	162 825	95.0	0.26	1093.4	11	127 440	75.0	43	7	37.9	19.7	6.0
2	40	Chefes e encar. seção serv. Adm.	88.7	0.25	1330.00	8	477	86.5	0.21	682.02	10	576 742	90.2	0.22	794.90	11	539 378	93.5	44	8	30.2	14.0	6.9
1	50	Técnicos e fiscais trib. e arrec	95.0	0.30	1383.00	11	38 900	97.4	0.33	1364.05	12	43 552	98.4	0.39	1712.0	15	48 359	92.5	40	11	17.2	5.6	2.5
2	51	Inspetores do trabalho	95.1	0.30	1383.00	11	2 706	90.0	0.22	677.48	11	1 400	91.5	0.22	800.00	11	911	85.6	42	13	4.5	3.5	
2	52	Assistentes administrativos	91.1	0.25	809.00	11	120	89.6	0.22	627.46	11	149 441	88.9	0.21	708.59	11	225 552	94.2	37	10	19.6	1.9	4.5
2	53	Pagad.e caixas (excl. serv./comércio)	90.3	0.24	771.00	11	119	91.8	0.22	682.02	11	143 597	91.0	0.22	798.92	11	152 161	92.7	39	9	21.6	7.3	8.4
3	54	Almoxarifes e armazenistas	75.2	0.14	426.00	7	114	70.2	0.14	318.28	8	131 362	68.7	0.13	378.39	8	130 247	92.5	44	5	50.0	10.4	15.3
3	55	Expedidores e confer. de materiais	72.7	0.10	426.00	5	72 774	66.3	0.12	318.28	7	89 825	73.3	0.13	386.82	8	95 565	91.2	45	5	49.7	16.5	14.1
3	56	Datilógrafos	76.6	0.19	399.00	10	52 093	72.5	0.17	286.45	11	24 242	72.5	0.17	356.39	11	15 208	81.8	35	8	31.8	2.6	7.1
2	57	Operadores de teleimpressoras	86.8	0.21	617.00	10	3 659	78.4	0.17	391.03	10	5 102	84.6	0.18	502.96	11	2 979	96.8	39	6	42.7	5.9	10.4
2	58	Operadores de máq. proc. de dados	90.8	0.25	798.00	11	27 170	87.2	0.19	454.68	11	127 752	81.5	0.18	450.00	11	155 424	84.2	38	4	53.1	4.4	11.3
2	59	Secretárias	85.7	0.22	532.00	11	262	79.1	0.18	341.01	11	466 740	76.3	0.17	382.69	11	367 970	78.2	39	5	48.8	4.7	10.9
2	60	Auxil. de contabilidade	86.3	0.22	532.00	11	78 346	81.7	0.18	363.75	11	128 524	79.6	0.17	393.62	11	216 673	82.8	41	5	46.6	4.9	10.1
3	61	Operadores de máquinas copiadoras	65.9	0.13	319.00	7	4 365	62.8	0.13	259.17	8	7 530	61.8	0.12	283.44	8	12 556	69.6	41	5	55.3	9.7	15.5
2	62	Arquivistas	79.3	0.17	426.00	9	24 407	78.1	0.17	363.75	10	19 077	79.9	0.17	396.81	11	12 901	79.9	38	7	38.9	4.1	13.6
3	63	Recepcionistas	67.5	0.14	319.00	8	119	64.0	0.14	227.34	9	208 965	66.1	0.16	285.33	11	413 597	75.4	40	4	59.7	7.8	15.3
3	64	Auxil. administ. e escritório	89.8	0.24	701.00	11	116	76.5	0.18	318.28	11	1 377 931	74.1	0.17	380.00	11	1 238 308	83.8	40	5	46.3	5.4	11.1
1	101	Engenheiros	98.0	0.61	3564.00	16	98 577	99.6	0.51	2273.41	16	142 739	99.7	0.48	2331.7	16	146 547	90.6	42	9	27.8	14.5	3.0
1	102	Arquitetos	97.4	0.49	2394.00	16	9 386	98.7	0.41	1591.39	16	18 301	98.6	0.40	1700.6	16	25 916	72.0	40	8	28.9	16.7	2.5
2	103	Agrimensores	92.3	0.27	1064.00	11	4 227	92.1	0.22	682.02	11	3 325	95.2	0.29	1059.9	14	2 872	67.8	42	13	17.1	16.2	2.8
1	104	Cartógrafos	92.3	0.27	1064.00	11	533	98.5	0.39	1500.45	15	403	99.1	0.44	2119.7	15	983	90.3	38	8	34.0	4.7	14.0
2	111	Desenhistas	92.4	0.27	1064.00	11	80 001	92.6	0.22	682.02	11	112 052	89.3	0.22	763.33	11	94 424	68.6	41	6	45.0	11.1	7.5
2	112	Técnicos de edific. e agrimensura	85.9	0.19	798.00	8	26 025	89.9	0.22	636.55	11	26 433	89.5	0.22	787.24	11	23 028	79.2	43	9	35.6	12.9	5.1
3	113	Outras ocup. auxil. engenharia	67.9	0.10	395.00	5	16 000	58.7	0.09	272.81	5	10 635	63.4	0.10	353.39	6	13 213	76.6	44	8	49.5	19.2	13.2
1	121	Químicos	97.8	0.55	2926.00	16	5 791	99.4	0.48	2047.16	16	12 111	98.7	0.42	1984.1	15	16 226	95.7	42	9	23.0	9.3	3.7
1	122	Farmacêuticos	96.1	0.39	1543.00	15	2 736	96.8	0.31	909.36	15	5 401	95.5	0.30	1027.8	15	8 285	92.6	37	5	39.1	14.5	3.0

Cat.	Cód. Ocu.	Denominação ocupacional	Censo 1980					Censo 1991					PNADs anos 1990 (1996)									
			ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	% Contr. Prev. Publ.	Jornada semanal média (horas)	Tempo médio ocup. (anos)	% pessoal temp<3 anos	% pessoal jorn>48 h/sem.	Fat. rel. risco de desocup. no ano
1	123	Físicos	97.9	0.57	2944.00	17	470	99.5	0.46	1818.73	17	595	99.4	0.46	2186.8	16	651	100.0	42	11	0.0	13.9	
1	124	Geólogos mineralogistas	97.9	0.61	3724.00	15	2 713	99.7	0.51	2273.41	16	3 469	99.9	0.49	2400.0	16	3 319	90.4	41	10	16.1	11.5	0.6
1	125	Out. especial em química e física	95.1	0.30	1383.00	11	908	98.5	0.38	1416.33	15	973	97.4	0.34	1400.0	14	1 007	91.7	39	14	6.6	0.0	2.6
2	131	Técnicos químicos	89.0	0.22	755.00	10	23 690	89.4	0.22	613.82	11	45 262	88.4	0.21	688.91	11	43 381	93.4	42	8	30.1	8.1	6.8
2	132	Práticos de farmácia	81.8	0.17	576.00	8	8 926	82.0	0.17	545.62	8	5 501	85.3	0.19	536.42	11	4 880	80.2	44	6	51.8	20.3	3.9
2	133	Técnicos de meteorologia	81.3	0.19	435.00	10	229	84.1	0.19	412.22	11	380	75.5	0.14	442.16	8	53	82.9	33	10			
1	141	Agrônomos	97.1	0.48	2394.00	15	12 105	98.4	0.37	1364.05	15	22 368	98.2	0.38	1640.1	15	20 069	88.8	42	10	18.9	13.4	2.2
1	142	Biologistas	96.7	0.40	1596.00	15	1 999	97.9	0.34	1136.70	15	4 759	97.3	0.33	1271.8	15	6 755	85.8	38	7	30.7	9.9	2.2
1	143	Farmacologistas	96.7	0.40	1596.00	15	2 233	97.9	0.35	1182.17	15	3 447	97.1	0.33	1255.4	15	3 243	91.6	34	8	23.6	4.7	
1	144	Veterinários	96.8	0.43	1915.00	15	6 199	98.2	0.38	1318.58	16	12 467	97.5	0.34	1215.7	16	13 124	78.0	41	8	30.2	21.5	0.9
1	151	Médicos	98.0	0.59	3192.00	17	65 222	99.9	0.58	2705.36	17	94 798	99.7	0.47	2186.8	17	93 857	83.2	36	10	24.4	17.1	1.4
1	152	Dentistas	96.9	0.45	2128.00	15	35 059	98.6	0.41	1636.85	15	58 474	98.8	0.43	2000.0	15	51 841	81.9	35	12	19.2	12.6	0.9
1	153	Enfermeiros diplomados	95.3	0.37	1330.00	15	8 389	96.6	0.31	909.36	15	16 635	92.6	0.26	699.77	15	31 518	92.8	38	7	29.5	5.9	3.3
1	154	Outros especialistas em medicina	93.1	0.34	1052.00	15	4 777	96.8	0.31	909.36	15	12 367	95.4	0.30	1000.0	15	13 506	84.7	35	6	40.7	5.1	4.2
1	161	Acadêmicos de hospital	89.3	0.30	585.00	15	515	85.5	0.23	386.48	15	131	54.3	0.19	113.37	15	517	17.5	36	1	91.2	3.2	3.1
3	162	Enfermeiros não diplomados	68.4	0.15	346.00	8	228	68.0	0.13	295.54	8	279 870	69.2	0.15	340.00	10	319 780	89.0	40	7	30.8	6.7	5.9
2	163	Técnicos de reabilitação	81.0	0.17	532.00	8	7 544	88.5	0.20	500.15	11	15 663	93.0	0.26	765.38	14	18 556	62.9	34	6	39.1	13.2	5.1
2	164	Ortoptistas e óticos	84.8	0.18	638.00	8	899	75.3	0.14	386.48	8	1 905	80.2	0.17	408.00	11	770	95.8	34	10	24.2	3.9	7.1
2	165	Oper. equip. médicos e odont.	81.0	0.17	532.00	8	6 061	87.0	0.19	454.68	11	6 742	86.4	0.20	593.52	11	9 740	89.1	35	8	28.4	5.5	3.2
4	166	Parteiras	45.9	0.07	266.00	4	1 020	29.5	0.06	165.96	4	1 353	21.4	0.05	137.78	4	1 358	74.8	45	12	10.0	29.1	6.7
2	167	Protéticos	82.5	0.16	638.00	7	11 104	81.7	0.16	522.88	8	17 602	85.0	0.19	515.77	11	23 129	47.4	42	11	31.5	22.7	4.1
2	168	Técnicos em análise clínica	87.1	0.22	585.00	11	10 556	88.6	0.20	522.88	11	12 955	86.2	0.19	570.66	11	19 324	89.3	38	8	28.7	4.3	3.4
1	171	Matemáticos e Atuários	97.6	0.54	3011.00	15	229	99.1	0.45	1905.12	16	99	61.6	0.10	342.39	6	81	100.0	34	6	63.5	0.0	
1	172	Estatísticos	96.7	0.42	1809.00	15	2 225	98.0	0.36	1314.03	15	1 360	98.4	0.41	1900.0	15	793	100.0	38	8	10.4	2.9	11.2
1	173	Analistas de Sistemas	97.8	0.60	3574.00	15	15 395	98.9	0.43	1818.73	15	58 705	98.9	0.43	2000.0	15	72 057	89.6	42	7	38.1	8.7	4.0
1	181	Economistas	97.5	0.50	2660.00	15	28 377	99.1	0.46	2046.07	15	20 998	98.5	0.41	1900.0	15	31 136	91.8	41	9	26.6	9.7	3.9
1	182	Contadores	93.9	0.30	1330.00	11	120	97.5	0.33	1091.24	15	137 343	97.9	0.34	1300.0	15	129 142	85.8	41	9	25.0	12.3	4.3
1	183	Técnicos de administração	97.1	0.48	2394.00	15	14 419	98.3	0.37	1364.05	15	22 811	98.1	0.37	1530.8	15	25 112	94.5	40	10	23.3	5.9	5.6
2	191	Técnicos de contabilidade	91.8	0.27	1011.00	11	37 330	92.0	0.22	682.02	11	52 515	87.5	0.21	684.79	11	52 038	83.7	41	9	27.7	8.0	3.4
2	192	Técnicos de estatística	90.1	0.24	770.00	11	15 940	92.3	0.22	682.02	11	21 559	87.6	0.21	684.79	11	29 386	76.1	38	7	41.4	5.5	11.0
2	193	Agentes Censitários	95.3	0.32	1489.00	11	14 525	89.3	0.21	609.27	11	16 621	72.0	0.16	342.39	11	4 858	70.4	39	0	97.6	15.0	54.3
2	194	Programadores de computador						93.4	0.24	818.43	11	50 387	91.8	0.23	850.31	11	71 282	77.9	41	5	48.5	10.4	5.6
1	201	Sociólogos, antropólogos	96.8	0.42	1862.00	15	1 726	98.5	0.38	1454.98	15	1 626	98.1	0.35	1417.2	15	1 876	95.2	37	10	19.4	10.0	2.4
1	202	Psicólogos	96.9	0.41	1596.00	16	5 074	98.2	0.35	1136.70	16	9 941	97.8	0.35	1271.8	16	8 569	81.7	32	7	32.5	4.8	3.7
1	203	Geógrafos e demógrafos	97.0	0.47	2287.00	15	595	98.3	0.37	1364.05	15	673	97.3	0.33	1271.8	15	510	100.0	36	14	6.6	0.0	3.9
1	204	Assistentes sociais	93.8	0.35	1170.00	15	16 665	94.5	0.28	727.49	15	29 114	94.5	0.27	800.00	15	34 635	83.9	35	8	30.8	3.6	5.5
1	205	Outros cientistas sociais	96.0	0.38	1416.00	15	204	98.0	0.37	1334.49	15	332	94.8	0.28	848.14	15	411	59.8	37	6	33.1	12.7	7.6
1	211	Prof. Pesquisadores	97.2	0.48	2234.00	16	490	98.8	0.43	1705.06	16	1 344	98.0	0.35	1381.0	15	2 981	54.4	32	7	41.1	4.3	3.1
1	212	Prof. de ensino superior	97.6	0.51	2500.00	16	32 798	99.3	0.44	1818.73	16	50 045	99.4	0.45	2119.7	16	57 740	95.3	35	11	20.8	7.4	2.4
1	213	Prof. de ensino do 2º grau	95.4	0.37	1330.00	15	59 295	94.7	0.29	786.60	15	112 318	94.4	0.27	798.92	15	121 843	89.6	32	10	21.8	5.2	3.2
2	214	Prof. de 5ª a 8ª série	92.7	0.31	867.00	14	42 809	91.4	0.25	545.62	14	61 535	91.2	0.24	600.00	14	139 609	88.7	30	9	23.0	3.8	3.7
2	215	Prof. de 1ª a 4ª série	78.9	0.21	425.00	11	82 851	79.9	0.18	345.56	11	193 872	80.8	0.18	423.00	11	284 419	79.9	28	8	28.5	2.0	5.4
2	216	Prof. de ens. 1º grau (sem espec.)	87.2	0.22	585.00	11	55 073	87.5	0.19	454.68	11	156 416	85.1	0.19	524.83	11	39 928	77.5	29	8	31.4	2.5	4.3
3	217	Prof. de ensino pré-escolar	73.3	0.20	365.00	11	4 305	72.2	0.17	272.81	11	22 424	67.9	0.16	312.71	11	38 992	65.8	27	5	45.4	1.8	10.3
2	218	Prof. e instr. de form. profissional	86.7	0.20	692.00	9	20 419	86.9	0.19	454.68	11	31 680	85.9	0.19	565.43	11	46 833	51.8	31	5	47.7	9.2	7.6
2	219	Prof. de ensino não especificado	89.3	0.23	638.00	11	37 164	91.6	0.24	580.87	13	101 583	90.7	0.22	580.24	13	39 749	55.8	27	5	52.0	5.0	7.6
1	221	Orientadores e técnicos de ensino	93.7	0.35	1099.00	15	30 004	94.2	0.28	682.02	15	42 398	93.1	0.27	741.91	15	67 278	89.1	35	9	26.9	4.7	4.4
3	222	Inspetores de alunos	69.8	0.13	372.00	7	16 659	69.1	0.14	272.81	9	29 012	64.0	0.14	285.33	10	34 790	87.9	36	7	34.5	2.5	8.7
1	231	Magistrados	98.2	0.77	5160.00	16	2 881	99.9	1.00	5910.86	16	5 393	99.9	0.89	5576.3	16	7 382	93.3	41	9	19.3	22.9	0.3
1	232	Procuradores, promotores públicos	97.9	0.59	3298.00	16	4 867	99.9	0.75	4092.14	16	6 491	99.9	0.79	4770.8	16	6 655	93.3	40	8	20.4	14.1	3.5
1	233	Advogados e defensores públicos	97.1	0.47	2128.00	16	61 386	99.0	0.44	1818.73	16	102 494	98.9	0.43	1934.1	16	131 485	65.0	39	10	25.0	16.4	2.2
1	241	Tabeliães e oficiais de registro	90.7	0.26	1064.00	10	4 189	95.4	0.27	1000.30	11	2 762	95.1	0.26	1093.4	11	6 726	77.2	39	18	13.1	6.5	1.2
2	242	Escrivães de cartório	91.3	0.25	825.00	11	17 674	92.4	0.22	682.02	11	27 858	94.6	0.26	975.36	12	42 088	91.6	39	10	17.5	4.0	1.6
1	243	Oficiais de justiça	89.5	0.23	798.00	10	8 724	96.5	0.29	1136.70	11	12 490	97.1	0.32	1369.6	13	16 728	95.0	40	9	19.0	5.3	1.1
2	244	Outras Ocup. auxil. da justiça	90.9	0.25	798.00	11	12 591	93.1	0.24	797.97	11	16 543	96.8	0.29	1275.5	12	22 392	92.5	37	8	28.0	1.5	3.1

Cat.	Cod. Ocu.	Denominação ocupacional	Censo 1980					Censo 1991					PNADs anos 1990 (1996)									
			ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	% Contr. Prev. Publ.	Jornada semanal média (horas)	Tempo médio ocup. (anos)	% pessoal temp<3 anos	% pessoal jorn>48 h/sem.	Fat. rel. risco de desocup. no ano
2	251	Sacerdotes	74.8	0.23	319.00	13	21 029	68.3	0.18	227.34	12	**24 597**	79.3	0.17	386.82	11	**38 740**	31.4	44	10	27.6	39.0	1.3
2	252	Religiosos por conta própria	.	.	.	.	.	73.5	0.14	363.75	8	**6 715**	86.8	0.20	635.92	11	**5 579**	19.4	32	15	11.4	22.7	0.4
1	261	Escritores e jornalistas	96.2	0.34	1596.00	12	19 624	97.2	0.33	1045.77	15	**30 075**	97.0	0.32	1200.0	15	**27 952**	70.9	40	6	45.3	18.5	7.6
1	271	Escultores e pintores	79.3	0.15	532.00	7	5 266	82.4	0.17	454.68	9	**10 969**	77.9	0.17	451.30	10	**12 014**	18.1	33	9	31.8	15.1	7.9
4	272	Artesãos de objetos	45.9	0.07	266.00	4	12 903	47.5	0.08	190.97	5	**49 713**	52.9	0.09	250.00	6	**29 415**	8.6	34	8	35.6	17.0	7.5
2	273	Decoradores e cenógrafos	82.8	0.16	692.00	7	10 643	81.6	0.16	500.15	8	**14 187**	80.8	0.17	570.00	9	**12 532**	42.7	37	7	41.3	14.2	11.0
3	274	Fotógrafos	76.7	0.11	532.00	5	37 296	75.1	0.14	377.39	8	**44 741**	73.9	0.13	393.62	8	**41 690**	41.3	39	8	34.5	17.8	6.6
2	275	Músicos e compositores	82.6	0.16	638.00	7	5 409	80.4	0.15	454.68	8	**14 084**	79.6	0.15	529.93	8	**13 532**	17.4	24	8	33.1	6.2	5.2
2	276	Artistas de cinema, televisão	84.8	0.18	638.00	8	4 586	80.9	0.15	454.68	8	**8 285**	79.5	0.16	500.00	9	**5 722**	19.9	25	5	46.1	9.8	7.1
4	277	Artistas de circo	62.6	0.08	426.00	4	616	36.1	0.06	190.97	4	**3 117**	57.0	0.09	353.39	5	**1 054**	23.6	28	6	21.8	9.0	13.5
2	278	Locutores e comentaristas	84.8	0.18	638.00	8	5 419	78.3	0.17	386.48	10	**5 797**	82.2	0.18	479.35	11	**8 393**	57.2	34	4	56.8	12.0	9.0
1	279	Produtores e diret. espetáculos	95.9	0.33	1596.00	11	3 906	97.3	0.33	1136.70	14	**13 876**	96.5	0.28	1289.4	11	**8 542**	53.1	41	6	45.8	28.4	7.4
2	280	Cinegraf. e operadores de câmaras	84.8	0.18	638.00	8	1 514	82.4	0.17	454.68	9	**3 430**	79.4	0.15	515.77	8	**4 799**	53.2	37	4	47.5	14.2	8.4
3	281	Operadores equip. som e cenog.	74.3	0.15	404.00	8	4 650	65.8	0.13	272.81	8	**7 776**	67.1	0.13	340.12	8	**9 732**	46.7	35	4	53.7	13.8	8.5
3	282	Outros oper. est. rádio e televisão	77.7	0.16	479.00	8	2 090	75.3	0.14	388.75	8	**4 273**	68.3	0.13	353.39	8	**2 958**	83.9	45	7	33.5	21.4	6.5
3	283	Operadores proj. cinemat.	50.9	0.07	319.00	4	1 107	63.8	0.11	309.18	6	**612**	59.9	0.12	270.00	8	**458**	71.5	42	3	72.6	12.8	
1	291	Bibliotecários	93.1	0.32	1011.00	14	9 653	94.3	0.27	712.15	14	**11 217**	93.0	0.27	736.93	15	**11 613**	88.2	35	9	28.1	2.0	4.5
1	292	Arquivologistas e museólogos	93.7	0.33	1064.00	14	237	95.4	0.31	909.36	15	**410**	95.2	0.29	900.00	15	**426**	78.9	36	13	16.7	0.0	15.9
1	293	Outras ocup. técnicas e artist.	96.4	0.38	1755.00	13	4 710	97.4	0.33	1182.17	14	**11 768**	96.8	0.32	1271.8	14	**12 717**	78.5	39	9	22.3	6.6	4.5
5	301	Produtores agropecuários autôn.	22.4	0.08	210.00	0	3 969	19.2	0.03	136.40	2	**3 454 966**	18.0	0.02	148.37	2	**3 182 561**	4.7	44	19	12.9	32.5	2.5
2	302	Técnicos de agropecuária	90.9	0.25	798.00	11	19 690	88.4	0.19	463.78	11	**31 203**	21.7	0.05	147.67	4	**57 011**	20.4	31	10	29.0	7.4	10.2
4	303	Tratoristas agrícolas	41.0	0.05	266.00	3	191	34.5	0.06	190.97	4	**201 079**	35.8	0.06	228.26	4	**283 265**	48.7	50	6	42.7	47.7	4.0
5	304	Trabalhadores na agropecuária	9.7	0.08	160.00	0	4 344	6.1	0.02	90.94	2	**4 821 211**	5.9	0.02	100.00	2	**5 350 193**	8.7	33	9	35.2	16.5	7.5
4	305	Trabalhadores na aqüicultura	.	.	.	.	.	42.8	0.07	227.34	4	**3 966**	28.3	0.05	180.00	4	**5 931**	14.3	23	8	45.8	14.1	2.5
5	321	Caçadores	39.6	0.04	299.00	2	352	6.1	0.01	90.94	1	**176**		0.11			**351**	-	25	2	77.3	4.9	4.9
5	322	Pescadores	29.5	0.01	223.00	1	115	25.3	0.04	177.33	2	**186 704**	21.4	0.02	154.73	2	**275 658**	5.9	40	10	26.2	31.5	3.9
4	331	Madeireiros	38.0	0.02	319.00	1	39 995	32.1	0.06	218.25	3	**53 339**	26.2	0.04	210.00	3	**55 094**	37.5	47	4	64.6	31.8	8.7
5	332	Lenhadores	31.9	0.09	239.00	0	33 395	24.1	0.03	159.14	2	**34 795**	14.8	0.02	141.36	2	**35 912**	12.9	44	5	54.6	27.9	6.1
5	333	Carvoeiros (fabricantes)	33.5	0.09	266.00	0	32 864	24.0	0.03	150.05	2	**19 976**	20.8	0.02	148.42	2	**24 134**	10.0	44	5	56.8	32.7	6.3
5	334	Seringueiros	33.7	0.09	266.00	0	52 265	17.0	0.08	181.87	0	**35 661**	21.5	0.04	148.38	3	**43 663**	30.2	41	3	63.8	22.9	11.8
4	335	Ervateiros	35.8	0.03	266.00	2	1 707	45.8	0.08	272.81	4	**6 751**	11.4	0.02	130.38	2	**944**	-	47	4	59.1	60.0	0.0
5	336	Apanhadores de prod. vegetais	0.2	0.07	106.00	0	68 822	0.1	0.07	81.84	0	**57 906**	.1	0.06	52.48		**102 990**	0.7	31	12	25.5	5.0	3.9
5	341	Mineiros	56.1	0.07	426.00	3	10 455	29.0	0.05	190.97	3	**5 270**	43.0	0.07	282.71	4	**3 604**	59.6	44	3	58.9	24.6	13.2
5	345	Canteiros e marroeiros	35.9	0.03	266.00	2	33 882	28.4	0.05	181.87	3	**37 717**	22.8	0.03	171.20	2	**42 363**	20.5	43	6	45.4	21.1	4.8
4	351	Operad. máquinas de ext. e benef.	47.9	0.06	372.00	3	14 664	54.1	0.08	277.36	4	**16 417**	51.0	0.07	322.35	4	**17 984**	83.2	46	5	45.8	21.4	7.8
2	361	Trabalh. extr. petróleo e gás	88.0	0.23	1205.00	8	6 145	82.8	0.18	682.02	8	**9 667**	95.2	0.27	1200.0	11	**4 938**	98.9	55	10	25.1	42.0	5.4
4	371	Garimpeiros	34.2	0.02	266.00	1	40 667	32.5	0.06	227.34	3	**163 003**	23.1	0.04	171.20	3	**107 486**	2.7	48	6	41.3	41.2	7.2
5	381	Salineiros	16.4	0.08	186.00	0	4 483	13.4	0.02	136.40	1	**3 733**	5.8	0.07	127.18		**1 320**	36.9	41	5	48.2	4.1	22.0
3	391	Soldadores poços (exc. petróleo/gás)	62.7	0.08	426.00	4	5 042	54.1	0.08	272.81	4	**4 667**	51.0	0.07	322.35	4	**7 271**	53.1	45	6	48.5	20.6	5.5
2	401	Mestres e técn. empresas ext. min.	95.3	0.32	1489.00	11	3 587	93.8	0.26	909.36	11	**5 227**	82.9	0.19	765.38	9	**4 342**	76.0	45	10	32.1	12.6	8.4
2	402	Mestres, contram. e técn. Ind. transp.	90.6	0.26	1436.00	8	66 467	93.1	0.24	818.43	11	**34 363**	91.6	0.23	826.69	11	**38 297**	91.6	44	8	35.1	10.7	7.8
2	403	Mestres, contram. e técn. Ind. têxtil	81.0	0.13	851.00	4	15 262	76.7	0.13	504.70	6	**13 396**	78.3	0.14	481.84	8	**7 800**	84.3	43	8	33.4	7.6	5.4
3	404	Mestres de construção civil	80.4	0.12	798.00	4	82 109	72.3	0.11	545.62	4	**83 783**	65.4	0.11	623.56	4	**80 250**	80.7	46	7	41.8	21.4	8.5
2	405	Mestres e técn. empresas serv. urb.	94.2	0.30	1330.00	11	42 324	93.6	0.24	818.43	11	**45 580**	92.4	0.24	913.05	11	**45 293**	93.9	42	10	27.8	8.0	4.9
2	406	Outros mestres, contram. e técn.	90.4	0.24	1064.00	9	15 707	93.8	0.26	909.36	11	**17 884**	92.5	0.25	989.50	11	**7 389**	91.9	41	9	27.5	10.0	2.4
3	411	Forneiros metalúrgicos	67.2	0.09	479.00	4	37 769	60.6	0.09	363.75	4	**39 543**	65.4	0.10	424.07	5	**40 916**	88.9	44	6	43.5	8.2	10.8
3	412	Laminadores	68.7	0.10	532.00	4	9 235	62.7	0.09	403.78	4	**9 566**	70.9	0.11	500.00	5	**10 244**	91.7	44	8	28.1	9.0	7.7
3	413	Trefiladores	68.7	0.10	532.00	4	3 190	69.9	0.10	427.40		**5 026**	66.4	0.10	479.35	5	**9 334**	89.9	46	5	47.2	11.7	10.3
3	414	Modeladores e macheiros	68.4	0.09	512.00	4	16 653	62.3	0.09	372.84	4	**18 848**	64.3	0.09	400.00	5	**17 316**	89.5	43	5	49.6	4.9	14.6
3	415	Galvanizadores e recob de metais	66.9	0.09	444.00	4	8 564	65.7	0.09	409.21	4	**7 015**	65.5	0.10	425.51	5	**10 900**	85.5	45	5	44.5	10.9	12.3
3	416	Afiadores e amoladores	76.4	0.11	638.00	4	5 313	65.7	0.09	409.21	4	**7 623**	65.6	0.10	445.14	5	**6 556**	77.0	43	8	41.2	11.6	8.2
3	417	Polidores e esmerilhadores	67.3	0.09	479.00	4	35 023	62.3	0.09	372.84	4	**31 988**	59.5	0.09	386.82	5	**35 495**	83.8	44	4	51.2	7.0	13.1
2	418	Ferramenteiros	84.2	0.18	1064.00	6	33 276	85.6	0.22	909.36	8	**37 185**	84.2	0.19	890.29	8	**23 678**	90.7	45	7	37.9	9.6	5.7
3	419	Estampadores mecânicos	58.7	0.08	383.00	4	2 992	61.2	0.09	297.82	5	**2 761**	59.1	0.08	399.46	4	**3 113**	83.7	47	4	58.6	22.6	14.3
3	420	Operadores de prensa mec. metais	67.0	0.09	449.00	4	38 729	62.6	0.09	391.03	4	**32 756**	66.2	0.10	470.16	5	**41 203**	92.6	44	4	47.6	6.3	8.9

Cat.	Cód. Ocu.	Denominação ocupacional	Censo 1980					Censo 1991					PNADs anos					1990	(1996)			
			ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	% Contr Prev Publ	Jornada semanal média (horas)	Tempo médio ocup. (anos)	% pessoal temp<3 anos	% pessoal jorn>48 h/sem.	Fat. Rel. risco de desocup. no ano
3	421	Fresadores e furadores	79.9	0.12	745.00	4	20 758	78.5	0.16	591.09	7	29 505	80.7	0.15	570.66	8	34 591	93.8	44	5	42.8	7.4	12.6
3	422	Torneiros mecânicos	76.6	0.11	638.00	4	140	75.9	0.14	454.68	7	136 277	79.7	0.15	529.93	8	114 257	83.0	44	6	43.7	11.7	9.0
3	423	Ajustadores e montad. mecânicos	74.7	0.10	559.00	4	103	77.5	0.14	463.78	7	91 160	79.1	0.15	509.32	8	110 852	83.9	44	6	40.5	9.6	11.3
3	424	Mecânicos de veic. automotores	63.2	0.08	426.00	4	340	62.3	0.09	313.73	5	602 156	62.3	0.10	342.39	6	600 548	45.9	46	6	43.0	25.7	7.4
3	425	Mecânicos sem especificação	75.4	0.10	585.00	4	271	72.3	0.12	409.21	6	312 415	72.3	0.13	451.30	7	322 375	66.6	44	6	41.4	17.1	7.9
3	426	Soldadores	67.6	0.09	479.00	4	168	60.3	0.09	350.11	4	178 809	64.5	0.09	400.00	5	169 992	76.5	45	5	51.9	12.3	11.0
3	427	Rebitadores e mont. estr. metálicas	67.1	0.09	452.00	4	14 961	66.3	0.10	327.37	5	21 501	63.6	0.10	356.45	6	24 656	77.9	45	4	61.4	16.3	13.4
3	428	Caldeireiros	79.5	0.11	665.00	4	22 037	74.1	0.11	454.68	5	17 107	70.6	0.11	490.77	5	20 322	92.3	45	5	52.9	12.6	9.5
3	429	Ferreiros e serralheiros	62.2	0.08	420.00	4	106	58.9	0.09	272.81	5	132 120	61.4	0.10	340.12	6	155 150	41.6	45	6	44.1	24.1	7.7
3	430	Funileiros	74.9	0.10	563.00	4	31 871	74.1	0.11	454.68	5	35 950	62.9	0.10	342.39	6	22 165	45.0	44	5	60.0	20.6	9.4
3	431	Lanterneiros de veículos	68.8	0.10	532.00	4	46 555	64.1	0.09	318.28	5	99 850	56.6	0.09	348.14	5	124 221	32.5	47	7	39.7	32.6	6.0
4	441	Cardadores e penteadores	45.9	0.07	266.00	4	5 612	51.6	0.08	272.81	4	14 595	47.0	0.08	255.65	5	16 458	88.8	44	5	43.8	8.3	19.0
4	442	Maçaroqueiros, bobinadores	49.7	0.07	293.00	4	18 462	43.2	0.07	240.98	4	12 959	47.4	0.08	270.78	5	9 632	92.9	43	3	53.7	2.8	14.5
4	443	Fiandeiros	50.4	0.07	298.00	4	39 941	39.5	0.07	227.34	4	23 995	46.7	0.08	254.44	5	15 444	90.7	43	5	46.3	4.6	13.7
5	444	Rendeiros	0.0	0.07	43.00	0	8 052	0.1	0.02	45.47	2	6 942	9.1	0.02	34.24	3	5 600	0.3	30	13	20.4	7.7	6.1
4	445	Urdidores e remetedores	54.7	0.08	372.00	4	6 205	45.8	0.08	272.81	4	6 156	52.2	0.08	317.96	5	6 330	87.0	43	4	55.2	2.4	26.4
5	446	Cordoeiros	29.8	0.03	221.00	2	1 588	12.0	0.02	90.94	2	728	28.3	0.05	182.61	4	1 258	20.3	33	8	31.7	7.8	10.8
4	447	Tecelões	51.1	0.07	319.00	4	106	46.0	0.08	272.81	4	81 713	51.8	0.08	309.46	5	55 664	83.8	44	5	47.7	10.1	18.1
4	448	Tapeceiros	63.7	0.08	426.00	4	11 940	67.5	0.10	350.11	5	14 500	50.3	0.08	294.77	5	8 443	34.1	36	7	45.4	13.1	9.4
5	449	Redeiros	0.1	0.07	64.00	0	8 440	0.2	0.02	81.84	2	12 057	9.3	0.03	70.68	3	11 013	3.7	33	9	28.5	11.6	9.1
4	450	Alvejadores e tint. têxteis	63.7	0.08	426.00	4	6 886	56.7	0.08	318.28	4	9 336	53.8	0.08	368.47	4	7 750	93.6	44	5	52.3	4.3	13.5
4	451	Estampadores têxteis	54.7	0.08	372.00	4	10 263	59.1	0.09	272.81	5	18 588	56.5	0.09	295.22	6	16 396	71.3	43	4	62.8	9.2	11.3
4	452	Acabadores de pano	49.8	0.07	293.00	4	22 483	55.7	0.09	260.17	5	30 843	57.1	0.09	296.57	6	28 666	86.5	43	4	54.7	5.3	16.2
4	461	Correeiros e seleiros	46.0	0.07	266.00	4	7 127	29.9	0.06	181.87	4	11 265	39.8	0.07	216.85	5	12 387	37.9	41	7	49.2	18.1	7.5
4	462	Curtidores	49.7	0.07	277.00	4	11 898	40.1	0.07	227.34	4	17 589	40.5	0.06	254.44	4	19 526	85.6	44	4	57.5	8.0	6.6
4	470	Alfaiates e costureiros	46.7	0.07	266.00	4	449	38.3	0.07	209.15	4	706 918	42.5	0.07	228.26	5	695 886	32.3	37	10	36.0	13.9	10.6
4	471	Auxil. de costuras	41.3	0.06	223.00	4	25 566	47.8	0.09	163.69	6	45 716	47.1	0.09	176.90	7	57 112	47.8	38	2	75.3	8.4	24.3
4	472	Calceiros e camiseiros	51.2	0.07	319.00	4	761	69.9	0.10	422.85	4	623	51.0	0.07	322.35	4	1 341	28.8	38	15	18.3	25.6	11.0
3	473	Modelistas e cortadores	63.8	0.08	426.00	4	13 574	64.2	0.12	295.54	7	19 710	62.6	0.11	326.22	7	23 509	76.3	42	4	48.3	8.7	12.6
4	474	Bordadeiras e cerzideiras	39.7	0.06	213.00	4	31 890	28.6	0.06	156.87	4	34 224	21.5	0.05	141.72	4	53 689	8.2	28	9	29.1	7.9	7.4
5	475	Chapeleiros de palha	0.0	0.06	18.00	0	8 152	0.0	0.06	18.19	0	5 742	.0	0.06	21.87		4 910	0.8	31	14	23.8	11.0	4.8
4	476	Chapeleiros (exc. palha)	33.9	0.05	221.00	3	1 096	29.9	0.06	181.87	4	526	28.4	0.06	130.00	5	1 224	22.0	38	4	65.5	6.4	6.9
4	477	Sapateiros	51.3	0.07	319.00	4	47 160	39.3	0.07	218.25	4	23 419	35.2	0.06	226.75	4	25 827	21.5	43	12	27.6	29.0	5.2
4	478	Trabalhadores fab. calçados	50.0	0.07	293.00	4	129	51.0	0.08	227.34	5	211 112	40.8	0.07	218.68	5	235 632	70.5	43	3	66.1	9.6	13.7
4	479	Bolseiros e cinteiros	45.1	0.07	261.00	4	7 048	49.2	0.08	204.61	5	3 650	47.1	0.08	215.41	6	8 652	57.7	42	3	63.6	13.0	14.5
4	481	Marceneiros	59.7	0.08	399.00	4	189	46.5	0.08	272.81	4	325 479	50.8	0.08	296.01	5	361 173	42.3	45	6	45.8	22.3	6.6
4	482	Carpinteiros	48.5	0.06	372.00	3	404	44.7	0.07	254.62	4	333 668	43.8	0.07	283.44	4	285 176	44.7	45	7	47.7	21.7	9.3
5	483	Tanoeiros	60.0	0.08	399.00	4	722	24.0	0.04	136.40	3	128	14.8	0.03	130.00	3	355	22.7	49	20	22.3	21.6	0.0
4	484	Operadores de máq. marcenaria	50.2	0.07	293.00	4	21 591	35.8	0.06	190.97	4	18 715	35.3	0.06	226.75	4	25 302	67.5	43	3	70.5	11.9	16.0
4	485	Serradores	41.4	0.05	266.00	3	72 135	37.0	0.07	204.61	4	76 939	33.3	0.06	217.27	4	97 146	61.7	47	4	60.3	23.6	8.0
4	486	Prep. compensados e aglomerados	49.7	0.07	277.00	4	4 846	39.1	0.07	209.15	4	3 614	32.7	0.06	205.44	4	6 618	87.8	45	3	65.2	11.3	11.4
4	487	Estofadores e capoteiros	63.8	0.08	426.00	4	22 676	58.6	0.09	272.81	5	46 308	51.6	0.08	300.00	5	55 161	37.0	44	6	45.9	23.3	7.7
4	488	Colchoeiros	47.3	0.07	266.00	4	1 515	32.2	0.06	181.87	4	1 557	39.5	0.07	204.00	5	2 360	62.0	41	2	70.3	12.4	13.2
4	489	Lustradores de madeira	51.4	0.07	319.00	4	16 643	43.2	0.07	241.83	4	12 627	48.3	0.08	280.00	5	13 459	48.3	43	5	55.0	11.1	14.9
5	490	Cesteiros e esteireiros	23.1	0.02	160.00	2	5 079	14.8	0.03	104.58	2	2 723	.3	0.02	85.03	2	5 247	1.4	29	11	27.7	9.2	8.8
3	501	Montadores de equip. elétricos	54.8	0.08	372.00	4	25 469	67.6	0.11	331.92	6	23 162	68.4	0.13	353.39	8	38 348	82.4	44	4	52.5	12.4	12.2
3	502	Montadores de equip. eletrônicos	66.0	0.11	344.00	6	15 615	68.6	0.13	341.01	7	17 139	68.5	0.13	369.00	8	21 077	82.2	43	4	63.5	8.6	17.4
3	503	Repar. de equip. elét. eletrônicos	69.1	0.10	532.00	4	143	80.0	0.15	454.68	8	314 898	78.6	0.14	490.77	8	233 770	70.6	43	7	40.1	13.8	7.4
3	504	Montadores apar de áudio e video	64.5	0.09	351.00	5	8 920	74.4	0.14	363.75	8	5 518	67.1	0.13	340.12	8	7 018	86.2	44	4	49.2	8.1	23.7
3	505	Reparar apar. de áudio e video	69.4	0.10	532.00	4	29 146	69.8	0.13	363.75		40 175	74.5	0.13	402.75	8	89 837	40.4	43	8	33.2	21.7	4.8
3	506	Eletricistas de instalações	68.1	0.09	505.00	4	111	69.4	0.10	363.75		116 041	71.0	0.12	410.02	7	177 005	52.4	44	7	40.2	20.8	6.8
3	507	Instal. e repar. de equip. de telec.	82.6	0.16	638.00	7	12 038	77.1	0.15	409.21	8	14 184	78.9	0.14	494.75	8	22 215	84.6	44	5	56.8	13.0	8.8
3	508	Instal. e repad. de linhas elétric.	76.9	0.11	638.00	4	31 417	77.8	0.15	491.06	7	55 516	80.4	0.15	565.43	8	70 615	88.1	44	8	36.5	11.3	7.2
2	509	Operadores inst. de prod. en elétric.	80.8	0.13	745.00	5	19 013	82.9	0.18	682.02	8	12 681	85.8	0.22	848.14	10	13 333	98.9	42	11	17.6	7.1	3.9
4	511	Armadores de concreto	55.7	0.07	412.00	3	59 989	51.5	0.08	272.81	4	47 249	46.7	0.07	319.57	4	47 791	66.8	46	4	63.4	22.9	18.1

Cat.	Cód. Ocu.	Denominação ocupacional	Censo 1980					Censo 1991					PNADs anos					1990	(1996)			
			ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	% Contr. Prev. Publ.	Jornada semanal média (horas)	Tempo médio ocup. (anos)	% pessoal temp<3 anos	% pessoal jorn>48 h/sem.	Fat. Rel. risco de desocup. no ano
4	512	Pedreiros	53.8	0.06	404.00	3	1 143	49.1	0.08	272.81	4	**1 836 229**	48.9	0.07	320.00	4	**1 888 563**	25.2	46	8	38.9	26.3	7.8
5	513	Serventes de pedreiro	32.8	0.03	229.00	2	744	25.3	0.05	163.69	3	**637 168**	26.3	0.05	170.06	4	**937 450**	24.9	43	2	75.6	15.1	20.3
4	514	Pintores e caiadores	59.0	0.08	383.00	4	209	52.3	0.08	272.81	4	**293 524**	46.3	0.07	311.78	4	**314 422**	23.5	44	7	40.7	21.2	11.8
4	515	Estucadores	56.1	0.07	426.00	3	8 410	56.4	0.08	318.28	4	**15 969**	53.2	0.08	350.00	4	**70 022**	26.8	45	7	39.3	23.9	7.6
4	516	Ladrilheiros e taqueiros	69.4	0.10	532.00	4	20 834	70.1	0.10	454.68	4	**20 877**	60.1	0.08	423.95	4	**28 379**	34.3	46	9	33.3	27.6	5.7
4	517	Encanadores	64.0	0.08	426.00	4	116	58.9	0.08	331.92	4	**115 327**	55.5	0.08	381.66	4	**113 109**	64.5	43	8	39.4	15.4	10.0
4	518	Vidraceiros (colocadores de vidro)	54.8	0.08	372.00	4	9 594	59.2	0.09	272.81	5	**15 431**	57.0	0.09	296.19	6	**19 647**	52.1	44	4	51.2	15.7	6.9
4	519	Calceteiros e asfaltadores	44.9	0.05	298.00	3	12 053	36.5	0.06	190.97	4	**14 606**	41.3	0.06	263.60	4	**16 043**	45.7	45	6	52.1	21.3	8.1
4	520	Calafates	64.2	0.08	426.00	4	9 694	56.0	0.08	318.28	4	**7 228**	51.3	0.07	323.26	4	**8 568**	30.6	46	12	26.7	27.7	8.9
4	521	Oper. máquinas de const. civil	56.3	0.07	426.00	3	85 321	54.4	0.08	285.54	4	**94 916**	53.5	0.08	352.35	4	**123 938**	80.7	48	6	44.0	28.9	7.3
4	531	Linguiceiros e salsicheiros	51.4	0.07	319.00	4	2 215	39.2	0.07	216.79	4	**3 814**	47.1	0.08	257.88	5	**4 305**	68.0	42	4	61.6	13.4	9.1
4	532	Charqueadores	44.5	0.05	293.00	3	1 902	37.1	0.07	204.61	4	**1 492**	39.8	0.06	240.31	4	**2 119**	81.1	45	2	66.4	19.2	6.8
4	533	Magarefes	41.6	0.05	266.00	3	29 924	36.9	0.07	204.61	4	**44 091**	41.2	0.06	259.63	4	**51 776**	69.6	43	4	57.9	17.9	9.6
4	534	Ocup. da Ind. de laticínios	47.4	0.07	266.00	4	7 200	36.1	0.06	190.97	4	**16 611**	41.2	0.06	260.76	4	**26 071**	56.9	40	5	52.3	17.1	6.4
4	535	Doceiros e confeiteiros	51.5	0.07	319.00	4	23 237	43.1	0.07	236.43	4	**37 054**	54.3	0.09	275.56	6	**59 458**	27.1	35	5	50.5	18.4	9.0
4	536	Macarroneiros e pasteleiros	50.3	0.07	293.00	4	3 428	37.4	0.07	209.15	4	**4 800**	55.9	0.09	283.44	6	**8 256**	44.6	38	4	55.9	20.4	6.9
4	537	Padeiros	47.5	0.07	266.00	4	87 122	34.1	0.06	190.97	4	**101 482**	40.0	0.06	250.00	4	**139 203**	42.4	45	4	59.2	33.3	9.7
5	538	Farinheiros e moleiros	22.8	0.01	170.00	1	8 688	12.0	0.03	104.58	2	**9 915**	.2	0.00	65.60	1	**29 441**	4.9	35	6	47.8	14.5	11.4
4	539	Ocup. da Ind. de açúcar	41.1	0.04	346.00	2	18 326	35.2	0.06	227.34	3	**15 363**	51.3	0.07	328.02	4	**16 517**	78.4	49	6	43.6	33.0	8.1
4	540	Ocup. da Ind. de bebidas	52.7	0.08	346.00	4	11 567	45.8	0.07	263.72	4	**18 710**	49.8	0.08	283.44	5	**19 661**	59.9	45	6	45.6	19.2	10.5
4	541	Ocup. da Ind. do café	52.7	0.08	346.00	4	3 653	37.4	0.07	209.15	4	**4 526**	47.4	0.08	271.41	5	**6 895**	70.1	45	7	42.1	11.7	13.7
4	542	Ocup da Ind. do pescado	33.8	0.05	221.00	3	4 228	33.5	0.06	181.87	4	**4 516**	33.4	0.06	218.68	4	**3 967**	62.4	46	4	56.3	25.6	34.8
4	543	Ocup. da Ind. de chá, mate e cacau	44.9	0.06	309.00	3	1 593	54.3	0.08	236.43	5	**859**	59.7	0.11	304.73	7	**1 739**	79.7	45	3	68.2	11.3	23.3
4	544	Ocup. da Ind. de oleaginosos	27.4	0.04	170.00	3	13 006	18.2	0.05	104.58	4	**16 441**	21.6	0.05	147.39	4	**14 801**	86.6	44	4	56.0	11.4	25.2
4	545	Ocup. de outras Ind. alimentares	45.1	0.07	261.00	4	32 888	33.7	0.06	181.87	4	**19 770**	37.3	0.06	236.25	4	**25 584**	44.4	42	4	53.0	21.9	14.1
3	551	Linotipistas	82.7	0.16	645.00	7	2 529	69.9	0.13	363.75		**1 420**	79.9	0.15	529.93	8	**1 464**	72.2	43	7	33.1	10.9	6.2
3	552	Tipógrafos	75.9	0.13	505.00	6	40 399	74.5	0.14	363.75	8	**48 150**	74.9	0.14	420.00	8	**38 558**	75.2	43	6	40.0	12.4	10.3
3	553	Clicheristas e gravadores	82.2	0.14	638.00	6	4 758	79.4	0.15	454.68	8	**8 000**	71.2	0.12	413.35	7	**11 481**	66.3	43	5	49.0	11.5	13.3
3	554	Impressores	76.8	0.11	532.00	5	25 232	71.6	0.13	386.48	7	**34 738**	75.5	0.14	437.36	8	**41 480**	84.8	44	5	47.8	10.8	9.0
3	555	Revisores na Ind. gráfica	85.1	0.20	532.00	10	1 547	73.6	0.14	363.75	8	**2 853**	79.9	0.17	395.80	11	**2 227**	80.1	38	5	44.4	11.4	14.6
3	556	Encadernadores e cartonadores	51.5	0.07	319.00	4	17 316	62.0	0.10	272.81	6	**18 559**	62.6	0.11	322.35	7	**23 343**	79.0	43	4	57.8	9.1	17.5
3	557	Outras ocup. da Ind. gráfica	67.4	0.10	372.00	5	14 473	62.7	0.12	272.81	7	**19 908**	65.9	0.12	320.00	8	**28 339**	48.3	42	3	60.9	14.2	15.6
4	561	Vidreiros e ampoleiros	64.3	0.08	426.00	4	15 579	61.2	0.09	309.18	5	**15 032**	70.8	0.12	407.36	7	**12 651**	83.1	44	5	45.4	6.5	13.7
4	562	Ceramistas e louceiros	49.6	0.07	276.00	4	21 324	39.4	0.07	218.25	4	**18 714**	39.8	0.06	243.77	4	**21 765**	60.5	43	5	43.6	12.9	9.1
4	563	Pintores cerâmicos	50.3	0.07	293.00	4	5 696	55.8	0.09	272.81	5	**5 976**	67.1	0.12	382.69	7	**5 127**	61.5	42	5	50.1	14.5	12.6
5	564	Oleiros	29.5	0.03	221.00	2	132	24.1	0.04	136.40	3	**109 534**	18.5	0.04	148.01	3	**152 539**	23.1	44	4	58.3	21.5	6.0
2	571	Inspetores de qualidade	82.9	0.16	692.00	7	77 354	81.9	0.17	543.55	8	**89 606**	81.0	0.17	581.69	9	**86 485**	95.7	44	7	35.4	6.9	12.5
2	572	Ourives e relojoeiros	69.5	0.10	532.00	4	19 474	61.9	0.10	272.81	6	**24 764**	56.7	0.10	275.56	7	**21 314**	38.4	43	9	35.1	19.0	8.4
4	573	Lapidadores	51.6	0.07	319.00	4	5 010	45.9	0.08	181.87	5	**9 194**	49.6	0.09	228.26	6	**6 129**	38.1	45	6	45.4	20.5	9.8
4	574	Borracheiros	51.6	0.07	319.00	4	40 146	42.4	0.07	227.34	4	**67 045**	40.9	0.06	257.88	4	**80 573**	37.1	51	5	48.4	42.7	6.4
4	575	Vulcanizadores e recauchutadores	74.9	0.10	564.00	4	4 319	58.7	0.08	327.37	4	**6 082**	65.6	0.10	438.21	5	**7 636**	90.6	46	7	38.3	9.7	9.7
4	576	Foguisteiros	27.4	0.04	186.00	3	1 216	23.8	0.05	109.12	4	**2 265**	9.3	0.03	91.31	3	**1 983**	3.4	39	7	41.8	15.8	17.1
5	577	Vassoureiros	33.5	0.05	213.00	3	2 025	19.7	0.04	113.67	3	**4 471**	21.4	0.05	137.78	4	**5 922**	14.4	34	7	49.7	10.3	9.1
4	578	Marmoristas	54.8	0.08	372.00	4	10 061	39.2	0.07	218.25	4	**19 294**	45.8	0.07	309.46	4	**28 853**	73.5	46	4	56.6	14.1	9.7
4	579	Preparadores de fumo	33.6	0.05	218.00	3	9 417	28.2	0.06	136.40	4	**6 891**	39.8	0.07	214.86	5	**3 604**	91.9	44	3	62.9	6.3	134.1
4	580	Charuteiros e cigarreiros	39.7	0.06	213.00	4	2 250	68.5	0.12	334.14	7	**1 511**	76.3	0.14	480.61	8	**2 718**	91.8	44	5	30.5	16.8	16.9
4	581	Pintores à pistola	64.4	0.08	426.00	4	84 147	63.8	0.09	318.28	5	**127 074**	54.9	0.09	328.02	5	**133 320**	46.5	45	5	50.6	20.8	9.5
4	582	Operadores de empilhadeira	75.8	0.10	585.00	4	11 731	69.9	0.10	431.95	4	**28 195**	66.3	0.10	476.94	5	**51 177**	93.7	46	5	44.9	12.3	10.6
4	583	Foguistas(exc. embarc. e trens)	55.5	0.07	408.00	3	26 286	54.2	0.08	281.90	4	**32 387**	45.3	0.07	296.76	4	**39 514**	83.1	47	5	45.6	24.0	9.2
4	584	Embaladores de mercadorias	44.8	0.07	255.00	4	162	48.3	0.08	190.97	5	**178 244**	49.9	0.09	228.26	6	**262 810**	73.9	44	3	68.6	12.6	21.9
4	585	Ocup. da Ind. do papel e papelão	54.9	0.08	372.00	4	31 026	54.6	0.08	286.45	4	**25 142**	63.4	0.10	353.39	6	**32 021**	87.4	44	5	51.5	8.6	13.4
4	586	Ocup. da Ind. art. borracha e plás.	53.9	0.08	362.00	4	49 388	60.5	0.09	281.90	5	**73 968**	60.3	0.10	322.35	6	**132 504**	89.0	44	3	61.2	6.9	18.5
4	587	Ocup. da Ind. art. cimento e fibro.	51.7	0.07	319.00	4	12 080	39.2	0.07	217.34	4	**7 143**	33.4	0.06	218.68	4	**11 448**	47.4	43	3	63.3	16.3	8.8
2	588	Supervisores segurança no trabalho	92.6	0.28	1170.00	11	8 711	93.7	0.25	905.18	11	**17 353**	92.5	0.24	935.95	11	**21 010**	97.6	44	7	35.8	12.8	6.7
4	589	Outras ocup. Ind. transformação	55.4	0.08	372.00	4	271	68.6	0.10	363.75	5	**440 959**	68.2	0.11	386.82	6	**329 175**	78.0	43	5	45.9	11.7	10.4

Cat.	Cod. Ocu.	Denominação ocupacional	Censo 1980					Censo 1991					PNADs anos					1990	(1996)			
			ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc	Total	% Contr. Prev. Publ.	Jornada semanal média (horas)	Tempo médio ocup. (anos)	% pessoal temp<3 anos	% pessoal jorn>48 h/sem.	Fat. Rel. risco de desocup. no ano
3	601	Comerciantes por conta própria	78.1	0.11	638.00	4	778	64.0	0.09	409.21	4	1 060 841	61.4	0.09	386.82	5	1 669 103	33.7	51	6	38.8	52.4	3.9
3	602	Vendedores	62.2	0.10	266.00	6	1 074	58.7	0.12	227.34	8	1 766 367	56.4	0.11	236.84	8	2 192 043	50.1	43	3	61.8	22.0	14.8
3	603	Operadores de caixa	65.4	0.14	266.00	8	136	54.2	0.12	203.12	8	233 508	61.3	0.14	250.00	10	300 678	73.7	46	3	63.3	24.0	18.9
3	604	Repositores de mercadorias	56.9	0.08	242.00	5	35 412	50.2	0.09	181.87	6	83 331	51.0	0.10	224.36	7	101 666	76.4	48	2	72.4	25.9	17.3
3	605	Demonstradores	70.3	0.13	391.00	7	4 900	74.7	0.15	331.92	9	5 407	65.5	0.15	306.15	10	7 687	65.8	39	2	71.2	6.6	30.6
3	611	Feirantes	64.7	0.08	426.00	4	92 032	53.9	0.08	272.81	4	96 378	47.2	0.08	257.88	5	94 458	13.9	36	7	36.3	22.7	5.9
4	612	Aguadeiros	16.4	0.08	186.00	0	1 190	40.0	0.07	227.34	4	2 244	33.4	0.06	218.68	4	2 315	1.2	33	5	51.8	13.1	12.3
4	613	Doceiros, sorveteiros e baleiros	41.6	0.05	266.00	3	17 691	37.3	0.07	209.15	4	88 302	32.4	0.06	200.00	4	111 942	6.1	35	4	56.7	22.1	9.8
5	614	Quitandeiros e fruteiros	45.0	0.06	319.00	3	14 932	29.3	0.05	190.97	3	46 968	32.6	0.06	200.00	4	52 508	6.6	38	6	43.6	27.1	6.5
4	615	Tripeiros, peixeiros e leiteiros	45.0	0.06	319.00	3	7 873	42.5	0.07	227.34	4	21 143	41.0	0.06	257.88	4	22 783	7.2	35	6	43.8	19.8	6.4
4	616	Bilheteiros	45.1	0.06	319.00	3	7 875	42.3	0.07	227.34	4	10 259	43.0	0.07	282.71	4	8 321	11.7	40	7	43.5	21.3	7.8
4	617	Outras Ocup. no comércio ambulante	65.1	0.08	426.00	4	163	53.2	0.08	272.81	4	494 060	48.9	0.08	282.71	5	486 301	9.3	34	5	49.5	19.4	9.4
4	621	Vendedores de jornais e revistas	65.3	0.08	426.00	4	11 093	68.2	0.12	318.28	7	16 392	66.2	0.12	322.35	8	19 976	31.8	41	4	63.6	27.4	15.2
2	631	Pracistas e viajantes comerciais	87.1	0.22	1064.00	8	108	87.6	0.21	682.02	10	97 209	87.0	0.20	644.71	11	191 213	60.8	43	4	54.1	19.4	10.9
1	632	Representantes comerciais	91.5	0.27	1596.00	8	36 586	95.3	0.26	909.36	11	97 373	92.7	0.26	1059.9	11	115 338	57.7	43	7	35.6	27.0	2.2
1	633	Propagandistas	94.3	0.30	1330.00	11	8 616	93.9	0.26	909.36	11	6 035	75.9	0.16	386.82	10	13 914	39.0	43	2	80.2	25.8	8.4
1	641	Corretores de seguros	90.5	0.24	1064.00	9	10 388	92.5	0.22	682.02	11	18 800	89.6	0.22	793.62	11	32 112	58.6	41	5	49.7	13.6	5.4
2	642	Corretores de imóveis	87.4	0.22	1064.00	8	59 192	87.1	0.20	682.02	9	67 059	91.3	0.22	798.92	11	76 013	36.7	45	8	33.0	28.4	4.5
1	643	Corretores de títulos e valores	95.1	0.30	1383.00	11	5 440	93.6	0.24	818.43	11	7 249	95.3	0.28	1235.5	11	5 117	68.7	40	5	47.0	10.6	3.2
2	644	Avaliadores e leiloeiros	90.9	0.26	1117.00	10	1 481	92.9	0.24	768.41	11	3 830	86.2	0.19	570.66	11	7 060	67.5	40	6	50.7	9.0	7.6
2	645	Outros agentes e corretores	85.9	0.19	798.00	8	14 050	89.0	0.21	545.62	11	39 341	86.1	0.19	568.57	11	66 520	56.7	42	5	50.9	18.3	12.2
2	646	Compradores	90.5	0.24	1064.00	9	19 419	93.5	0.24	818.43	11	35 247	89.5	0.22	766.96	11	34 456	77.5	44	7	28.8	17.7	7.8
1	711	Aviadores civis	96.2	0.41	2394.00	11	3 613	97.3	0.38	1818.73	11	5 587	97.5	0.40	2186.8	11	5 703	78.2	41	8	26.5	24.4	4.9
2	712	Comissários de bordo	95.9	0.34	1702.00	11	1 534	96.6	0.32	1362.56	11	3 051	97.1	0.32	1483.8	12	2 729	100.0	35	8	27.0	13.0	5.9
1	721	Oficias de marinha mercante	96.0	0.38	2128.00	11	2 838	96.6	0.32	1364.05	11	3 350	92.6	0.25	992.03	11	2 030	65.2	53	13	17.8	42.8	2.8
3	722	Mestres de embarcação	69.5	0.10	532.00	4	7 648	60.7	0.09	363.75	4	9 661	46.7	0.07	317.96	4	15 640	45.7	54	7	45.7	49.6	4.9
3	723	Maquinistas de embarcação	83.8	0.17	942.00	6	2 344	78.3	0.15	545.62	7	1 964	81.5	0.17	656.04	8	2 163	82.1	50	9	10.3	29.5	2.8
3	724	Foguistas de embarcação	81.1	0.13	851.00	4	1 638	74.0	0.11	454.68	5	958	73.0	0.13	656.04	5	1 134	94.3	51	8	38.1	40.1	
3	725	Marinheiros civis	75.8	0.10	585.00	4	7 541	74.8	0.12	431.95	6	9 454	76.2	0.14	476.94	8	9 090	75.4	50	6	45.3	38.0	11.1
3	726	Taifeiros	69.6	0.10	532.00	4	3 285	70.3	0.12	372.84	6	1 782	57.2	0.09	367.53	5	2 736	60.4	53	5	57.9	42.8	7.3
4	727	Barqueiros e canoeiros	39.7	0.04	319.00	2	4 503	36.5	0.06	250.08	3	6 417	39.8	0.06	241.96	4	6 795	40.8	49	7	41.0	43.3	9.8
3	731	Guindasteiros	79.5	0.11	660.00	4	14 604	62.5	0.09	384.93	4	19 599	71.2	0.11	515.77	5	23 234	95.5	46	8	31.5	17.0	8.1
4	732	Estivadores	65.4	0.08	426.00	4	27 942	43.0	0.07	227.34	4	33 269	51.1	0.07	322.35	4	20 729	62.2	46	10	23.8	30.5	8.3
2	741	Agentes de estrada de ferro	80.0	0.12	745.00	4	12 232	84.2	0.18	545.62	9	9 084	88.5	0.21	706.54	11	6 237	99.6	43	12	5.2	5.9	2.1
2	742	Condutores e chefes de trem	80.6	0.12	798.00	4	1 690	82.7	0.18	636.55	8	602	91.7	0.23	847.15	11	559	100.0	43	15	19.3	19.3	12.4
3	743	Maquinistas de trem	80.6	0.12	798.00	4	13 378	82.3	0.17	550.17	8	12 699	83.9	0.18	803.64	8	7 722	95.7	45	12	11.7	18.5	10.0
3	744	Foguistas de trem	76.4	0.10	622.00	4	462	73.9	0.14	409.21	7.5	227	31.9	0.10	368.47		77	100.0	40	9	0.0	0.0	98.1
3	745	Guarda-freios	69.6	0.10	532.00	4	701	62.5	0.09	386.48	4	156	63.3	0.09	494.75	4	384	100.0	44	10	20.3	21.3	
3	746	Manobreiros e sinaleiros	69.6	0.10	532.00	4	6 678	74.7	0.12	418.31	6	4 527	79.9	0.18	838.95	7	2 973	94.7	42	13	7.8	9.2	7.8
3	751	Motoristas	71.6	0.10	532.00	4	1 347	67.8	0.10	418.31	4	1 692 672	68.5	0.10	481.84	5	1 832 458	66.5	51	6	43.6	44.3	6.0
3	752	Trocadores	51.9	0.07	319.00	4	94 983	61.4	0.09	309.18	5	131 034	63.0	0.11	328.02	7	153 184	88.3	48	3	56.6	28.3	11.5
5	753	Carroceiros e tropeiros	28.5	0.08	213.00	0	38 988	23.7	0.03	136.40	2	44 248	14.7	0.02	136.96	2	46 327	9.8	41	7	37.9	24.5	4.1
2	761	Inspet. e despach. transportes	73.7	0.10	532.00	4	24 692	80.8	0.15	454.68	8	30 161	79.5	0.15	515.77	8	38 544	95.6	46	7	32.5	22.8	7.7
3	762	Trabalhad. de cons. ferrovias	49.2	0.06	372.00	3	12 469	61.7	0.09	363.75	4	4 698	46.9	0.07	418.65	3	4 190	90.4	42	13	16.9	11.9	13.0
2	771	Agentes postais e telegráficos	86.9	0.21	634.00	10	8 169	88.8	0.21	545.62	11	4 410	91.8	0.23	848.14	11	5 328	97.5	42	14	19.2	4.5	1.4
3	772	Postalistas	83.1	0.18	532.00	9	8 541	78.2	0.18	323.74	11	10 680	82.2	0.18	480.00	11	14 459	93.6	41	8	36.6	2.5	5.7
3	773	Telegrafistas e radioteleg	85.1	0.18	692.00	8	8 369	83.4	0.18	454.68	10	4 229	80.5	0.18	422.29	11	4 216	80.1	39	8	37.8	5.3	4.0
3	774	Telefonistas	71.7	0.15	378.00	8	37 826	56.1	0.12	227.34	8	58 842	64.7	0.15	295.22	10	77 278	81.6	37	5	51.8	5.6	14.0
3	775	Carteiros	78.0	0.16	511.00	8	19 691	76.0	0.15	363.75	9	15 968	77.4	0.17	450.00	10	29 402	95.3	42	8	35.9	4.2	6.1
4	801	Arrumadeiras	81.2	0.14	957.00	4	25 190	23.8	0.05	105.17	4	12 361	21.4	0.05	137.78	4	17 812	24.3	39	3	67.8	17.4	20.3
4	802	Babás	.	.	.	.	.	38.4	0.07	113.67	5	33 485	28.5	0.06	131.21	5	127 802	12.9	42	1	83.6	27.6	39.6
4	803	Cozinheiras	.	.	.	.	.	28.1	0.06	136.40	4	39 974	25.2	0.05	158.98	4	49 173	34.5	45	4	56.2	32.4	23.3
4	804	Faxineiras	.	.	.	.	.	36.3	0.06	190.97	4	129 102	33.0	0.06	211.97	4	161 814	11.6	28	5	51.0	7.4	15.4
5	805	Lavadeiras	18.1	0.04	133.00	3	1 713	14.9	0.03	109.12	2	45 221	15.0	0.03	135.79	3	32 872	5.0	21	6	45.6	4.5	16.1
4	806	Governantas e mordomos	.	.	.	.	.	51.3	0.08	272.81	4	5 053	56.6	0.09	296.01	6	7 945	65.2	47	8	32.5	36.7	8.3

Cat.	Cód. Ocu.	Denominação ocupacional	Censo 1980					Censo 1991					PNADs anos 90 (1996)									
			ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	ISEP	ISED	Rendim.	Esc.	Total	% Contr. Prev. Publ.	Jornada semanal média (horas)	Tempo médio ocup. (anos)	% pessoal temp<3 anos	% pessoal jorn>48 h/sem	Fat. Rel. risco de desocup. no ano
5	807	Empregados domést. não especial	.	.	.	.	.	21.0	0.05	104.58	4	2 279 717	18.3	0.05	135.79	4	2 720 087	25.0	45	3	64.6	32.2	22.4
4	808	Outras ocup. do serv. doméstico	.	.	.	.	.	31.8	0.06	181.87	4	43 047	34.9	0.06	219.20	4	114 122	33.9	46	4	57.9	36.3	13.7
3	811	Donos de pensão por conta própria	80.6	0.12	798.00	4	10 057	70.0	0.10	454.68		5 053	56.2	0.08	386.82	4	4 585	22.0	54	10	25.2	64.0	5.1
4	812	Camareiros(exc. no serv. dom.)	41.4	0.06	223.00	4	21 248	30.0	0.06	181.87	4	29 532	32.7	0.06	200.00	4	39 335	79.3	48	3	66.2	25.3	20.9
4	813	Cozinheiros(exc. no serv. dom.)	38.4	0.05	255.00	3	204	35.3	0.06	190.97	4	408 618	34.0	0.06	218.68	4	528 695	64.7	42	4	52.7	21.0	15.4
4	814	Garçons	52.2	0.07	319.00	4	98 812	54.1	0.08	227.34	5	163 123	53.5	0.09	254.44	6	176 131	45.1	46	3	65.4	36.1	18.9
4	815	Atendentes de bar e lanchonete	41.9	0.06	226.00	4	171	45.6	0.08	181.87	5	224 572	44.0	0.08	171.20	6	310 591	42.8	41	3	59.9	23.1	14.1
3	816	Govern. e Mordomos(exc.serv. dom.)	.	.	.	.	.	67.3	0.11	318.28	6	1 067	74.6	0.13	404.56	8	958	100.0	46	4	53.7	5.9	11.3
2	817	Maître de hotel	.	.	.	.	.	82.0	0.17	545.62	8	1 013	82.8	0.18	791.60	8	402	100.0	47	5	49.6	36.8	
2	818	Maitre no serv. de alimentação	.	.	.	.	.	79.0	0.17	682.02	7	1 631	76.7	0.15	750.00	6	1 641	88.1	46	5	40.4	36.1	
3	821	Cabeleireiros	65.5	0.08	426.00	4	48 741	71.2	0.13	363.75	7	133 031	73.6	0.13	386.82	8	183 588	26.1	42	7	33.6	32.3	4.9
4	822	Barbeiros	65.6	0.08	426.00	4	35 153	56.3	0.08	318.28	4	32 174	53.1	0.08	349.89	4	30 757	34.8	47	18	16.9	44.5	1.5
3	823	Maquiladores. depil. e estetic.	81.0	0.17	532.00	8	3 079	79.3	0.15	454.68	8	9 382	85.3	0.19	529.93	11	12 668	36.4	36	6	41.3	16.9	7.6
4	824	Manicuros e pedicuros	47.7	0.07	266.00	4	31 910	54.4	0.10	227.34	6	54 869	52.7	0.10	228.26	7	65 681	11.3	28	7	36.0	9.3	6.5
4	825	Lavadeiras e passadeiras	13.4	0.00	133.00	1	124	29.6	0.06	172.78	4	62 326	27.5	0.05	171.20	4	92 680	21.0	27	7	42.2	6.3	11.3
5	826	Engraxates	23.2	0.02	183.00	2	2 741	24.5	0.05	159.14	3	2 420	11.4	0.02	127.83	2	2 026	2.8	29	3	69.4	8.4	17.7
3	831	Jogadores de futebol	81.1	0.17	532.00	8	2 839	75.4	0.15	397.85	8	5 370	75.3	0.14	423.95	8	3 856	49.5	31	3	52.8	8.6	16.4
3	832	Lutadores e outros atletas prof.	65.7	0.08	426.00	4	1 047	76.0	0.13	454.68	6	1 727	59.9	0.12	264.97	8	1 456	22.4	31	3	55.7	9.3	4.6
2	833	Juizes de esportes	73.8	0.10	532.00	4	801	72.3	0.15	318.28	9	283	82.3	0.18	570.66	10	121	13.1	15	4	49.5	0.0	3.9
2	834	Técnicos de esportes	93.7	0.33	1117.00	14	8 428	92.3	0.26	682.02	14	30 290	91.5	0.24	623.56	14	33 557	61.7	30	5	44.0	6.9	4.5
4	841	Porteiros	52.8	0.08	346.00	4	87 609	45.6	0.07	259.17	4	138 566	45.5	0.07	296.76	4	181 862	89.2	44	5	47.4	17.4	8.9
4	842	Ascensoristas	52.4	0.07	319.00	4	9 092	43.2	0.07	243.91	4	7 065	48.3	0.08	278.82	5	6 810	98.7	37	7	31.0	3.4	10.0
4	843	Vigias	45.4	0.06	319.00	3	233	43.8	0.07	250.08	4	439 336	42.3	0.07	278.48	4	533 612	78.5	49	5	49.6	33.2	9.3
4	844	Serventes	35.1	0.05	223.00	3	678	30.9	0.06	181.87	4	719 201	29.2	0.06	193.41	4	889 049	80.1	39	5	47.2	8.1	12.1
4	845	Contínuos	55.8	0.10	223.00	6	132	51.2	0.10	181.87	7	181 426	46.8	0.09	174.94	7	191 420	59.6	37	2	78.7	3.9	21.8
4	851	Out. prop. agrop. conta própria	.	.	.	.	.	45.8	0.07	261.44	4	1 815	60.2	0.08	425.15	4	1 484	18.5	50	9	33.3	59.0	
2	852	Out. prop. serv. conta própria	.	.	.	.	.	84.4	0.18	682.02	8	46 579	83.2	0.19	684.79	10	146 543	42.3	45	7	39.3	36.5	3.3
1	861	Oficiais das Forças Armadas	96.7	0.44	2341.00	13	33 419	97.7	0.36	1364.05	14	41 572	97.3	0.35	1589.8	13	57 017	41.2	43	15	11.3	14.4	3.7
3	862	Praças das Forças Armadas	82.8	0.17	612.00	8	277	83.3	0.17	454.68	9	374 263	81.6	0.18	547.83	10	421 963	49.9	44	8	28.0	17.8	3.4
3	863	Oficiais e Praças Bombeiros	83.3	0.17	745.00	7	15 569	84.3	0.19	531.98	10	29 284	86.5	0.20	600.00	11	29 635	88.6	49	10	15.0	31.3	0.8
1	864	Delegados e Comis. Polícia	94.5	0.30	1353.00	11	11 234	98.0	0.37	1364.05	15	15 587	98.3	0.39	1696.3	15	14 646	93.7	46	13	13.1	21.9	3.0
2	865	Investigadores de Polícia	87.5	0.22	1064.00	8	24 765	89.8	0.22	636.55	11	45 665	88.5	0.21	706.78	11	49 605	94.5	44	11	11.5	14.6	0.9
3	866	Guardas-civis e Inspet. Tráfego	66.9	0.09	446.00	4	31 957	67.4	0.10	345.56	5	35 668	86.0	0.19	567.00	11	28 103	91.7	44	8	30.5	16.5	3.6
3	867	Carcereiros	83.8	0.18	798.00	7	6 386	84.2	0.19	513.79	10	11 519	86.5	0.20	600.00	11	18 100	95.2	44	8	21.6	10.5	1.5
2	868	Datiloscopistas	90.1	0.24	769.00	11	1 380	89.3	0.21	594.68	11	1 969	89.1	0.21	741.91	11	3 388	96.8	39	10	9.1	3.3	5.9
3	869	Guardas-vigias de org. part.	53.1	0.08	346.00	4	120	63.3	0.09	318.28	5	233 415	66.9	0.12	381.66	7	209 655	91.2	48	4	56.4	26.5	10.4
4	911	Aprendizes	54.2	0.10	221.00	6	16 889	61.5	0.10	272.81	6	23 658	55.3	0.10	238.53	7	33 667	37.1	36	2	74.8	7.4	11.6
4	912	Bilheteiros no serv. diversões	47.7	0.07	266.00	4	1 352	52.1	0.09	190.97	6	1 721	60.2	0.12	275.56	8	3 289	52.2	40	4	65.2	26.1	38.3
3	913	Bombeiros (exc. Corpo Bombeiros)	53.9	0.08	356.00	4	6 294	68.8	0.11	363.75	6	5 396	68.1	0.13	350.00	8	4 049	91.9	43	6	41.5	4.4	4.3
4	914	Capatazes	67.9	0.09	479.00	4	22 539	56.3	0.08	318.28	4	7 860	60.2	0.08	425.15	4	14 499	81.6	45	9	38.7	20.9	5.5
4	915	Dedetizadores	52.4	0.07	319.00	4	2 832	50.4	0.08	227.34	4,5	3 823	55.0	0.09	328.02	5	3 455	34.5	36	6	49.7	14.3	12.5
4	916	Guardadores de automóveis	52.4	0.07	319.00	4	1 914	39.2	0.07	209.15	4	2 998	30.2	0.06	193.41	4	7 606	8.2	39	4	54.4	26.6	6.2
3	917	Guardas sanitários	62.0	0.08	414.00	4	16 146	72.1	0.14	322.82	8	37 161	60.9	0.12	280.00	8	59 676	78.3	38	7	37.4	1.4	4.1
2	918	Inspetores e Fiscais	78.9	0.13	532.00	6	40 588	71.7	0.13	391.03	7	44 936	75.4	0.14	424.07	8	38 094	86.2	41	8	32.8	11.3	7.6
5	919	Jardineiros (exc. lavoura)	34.2	0.03	255.00	2	55 515	30.4	0.05	190.97	3	52 154	35.4	0.06	226.75	4	81 069	55.3	40	6	45.1	10.2	8.1
5	920	Lixeiros	29.9	0.03	221.00	2	52 537	23.9	0.03	145.50	2	120 421	21.0	0.02	153.08	2	125 950	75.0	41	5	46.3	7.9	10.4
4	921	Lubrificadores	59.4	0.08	383.00	4	20 656	55.0	0.08	295.54	4	17 874	52.2	0.07	342.39	4	15 455	85.3	47	6	46.2	26.6	8.5
3	922	Oper. tratamento e bomb. de água	67.1	0.09	473.00	4	11 845	70.0	0.10	454.68		20 067	73.1	0.12	482.24	6	22 184	94.5	44	10	21.9	12.8	3.4
4	923	Oper. máquinas(exc. Agrop./c civil)	66.1	0.08	426.00	4	267	59.1	0.08	336.46	4	42 478	63.5	0.10	353.39	6	41 087	87.3	45	6	41.4	14.5	10.9
4	924	Trabalhadores braçais. sem espec.	39.4	0.05	255.00	3	492	32.8	0.06	181.87	4	543 045	31.2	0.06	193.41	4	883 907	36.7	42	3	63.8	21.0	10.3
5	925	Trabalhadores de conser. rodovias	34.0	0.03	233.00	2	14 302	26.2	0.05	172.78	3	3 772	23.9	0.04	180.52	3	3 895	93.2	42	9	28.7	3.8	
4	926	Babás (exclusive no serv. dom.)	57.2	0.08	372.00	4	924	53.4	0.11	204.61	7	38 356	49.2	0.09	198.57	7	75 123	61.2	38	3	55.5	13.6	12.0
		Média das ocupações	68.6	0.16	755.7	6.5		65.9	0.16	537.6	7.4		65.2	0.15	589.3	7.9							
		Média do pessoal ocupado	48.4	0.11	445.7	3.7		50.0	0.11	342.3	5.5		49.2	0.10	376.1	6.0							

Notas: 1. Valores de rendimentos em R$ de setembro de 1999, deflacionados pelo INPC (anos de 1990) e deflator de Ferreira & Barros (1999) para 1980.

2. Indicadores apresentados para PNAD foram obtidos através do agrupamento das edições anuais da pesquisa de 1992 a 1999, com totais estimados para 1996 com emprego da Contagem Populacional de 1996 e incorporação de estimativa de pessoal por ocupação na Região Norte (exclusive Tocantins), segundo o Censo de1991. Desta forma, além de estarem sujeitas a erro amostral superior que as produzidas a partir dos Censos, estas estimativas podem ter uma tendenciosidade não facilmente comensurável.

3. Legenda: ISEP = Índice de *status* socioeconômico posicional;
   ISED = Índice de *status* socieconômico distancial ;
   Rendim. = Rendimento mediano dos ocupados na ocupação descrita;
   Esc. = Escolaridade mediana dos ocupados na ocupação descrita; e
   Total = Total de ocupações referem-se à população de 15 a 64 anos, com 38 ou mais horas no trabalho principal, com rendimento e escolaridade declarados nas pesquisas.

Abordagem Bayesiana para combinar resultados de estudos de câncer entre espécies via amostrador de Gibbs

José Ailton Alencar Andrade*

Gustavo L. Gilardoni[1]

RESUMO

Neste trabalho estudamos um modelo hierárquico proposto por DuMouchel e Harris para a extrapolação de risco de câncer entre espécies e substâncias. No trabalho original os autores estimaram as quantidades de interesse usando métodos bayesianos empíricos, os quais, ao menos em teoria, poderiam subestimar a incerteza presente nas estimativas finais. Aqui nos propomos um enfoque completamente bayesiano que usa o Amostrador de Gibbs para simular amostras da distribuição o *a posteriori* dos parâmetros.

1. Introdução

Resultados de experimentos com animais (ratos, camundongos, etc.) servem de suporte para grande parte das hipóteses que surgem nos estudos biomédicos, em particular os estudos de câncer, nos quais o interesse principal é estudar o potencial cancerígeno de substâncias em humanos. Experimentos laboratoriais são bastante

* Endereço para correspondência: Universidade de Brasília - e-mail: j.a.andrade@shef.ac.uk e gilardon@unb.br.

[1] Este trabalho faz parte da dissertação de mestrado do primeiro autor realizada sob a orientação do segundo autor no Departamento de Estatística da Universidade de Brasília. Parte do trabalho do segundo autor foi financiada por um projeto da Fundação de Amparo à Pesquisa do Distrito Federal - FAPDF.

confiáveis do ponto de vista estatístico. No entanto, essa metodologia de pesquisa se torna inviável quando se estuda câncer em humanos, pois não é possível submeter pessoas a testes em que as substâncias podem causar câncer. Baseado nesse argumento surge a idéia de *extrapolação interespécie e interagente*, que consiste essencialmente em estender (extrapolar) de forma conveniente o risco de câncer de animais para homens e o potencial cancerígeno entre substâncias. Para facilitar a leitura, daqui em diante a extrapolação interespécie e interagente seria tratada somente como extrapolação interespécie.

A extrapolação é feita em dois momentos: de altas para baixas doses e de animais para humanos. A primeira extrapolação consiste, basicamente, em aumentar a dosagem em animais de forma que a resposta à substância aconteça em tempo hábil, pois se for dosada proporcionalmente à dosagem humana, o animal pode levar anos para desenvolver a doença. Vale enfatizar que quando se trabalha com grandes doses deve-se ficar atento a um dos axiomas da toxicologia: *a dose faz o veneno*, ou seja, até água pode se tornar tóxica dependendo da dose, portanto, a dosagem deve seguir um rigoroso estudo para que se possa determinar a *tolerância máxima* (Freedman e Zeisel, 1988). A segunda extrapolação possível é devido às similaridades entre animais e humanos. Os ratos e camundongos possuem muitas similaridades com o homem (por exemplo, ambos são vertebrados e mamíferos, além disso, são o particularmente úteis como unidades experimentais, visto que têm baixo custo com alimentação, manuseio, etc. Muitos pesquisadores acham insuficientes esses argumentos, pois a extrapolação animal/homem traz complicadores adicionais no que diz respeito à sensibilidade entre espécies, sexo e indivíduos. É necessário que essas diferenças sejam quantificadas através de seus embasamentos biológicos, o que leva à procura de evidências em estudos epidemiológicos. Freedman e Zeisel (1988) discutem vários estudos dessa natureza, ressaltando algumas evidências que servem para validar as extrapolações quantitativas e qualitativas. Na extrapolação quantitativa ressaltam-se aspectos técnicos pertinentes à dosagem, tais como o modelo matemático usado para estabelecer a dose. Esse modelo se baseia nas diferenças entre humanos e animais, pesos, tempo de vida, etc. Já na extrapolação qualitativa, utilizam-se apelos intuitivos baseados em evidências de estudos anteriores. De acordo com esses estudos existem fortes indícios de similaridades entre ratos e humanos, assim como o fato de que substâncias cancerígenas em homens são quase sempre malignas em ratos. Para maiores detalhes veja Freedman e Zeisel (1988). Para que a extrapolação seja possível é necessário que estejam disponíveis estudos comuns as duas espécies. Como experimentos com humanos raramente estão acessíveis, então são usados estudos epidemiológicos como sendo a "ponte" que liga as duas espécies.

No presente problema tem-se uma coleção de estudos, nos quais foram utilizados modelos lineares de dose-resposta que são casos particulares do modelo de regressão de Poisson. Assim, o número de respostas positivas (tumores, células mutantes, etc.) n_j satisfaz $n_j \sim$ Poisson $(N_j(\zeta + \xi d_j))$ e $E(n_j) = N_j(\zeta + \xi d_j)$, onde N_j é o número de experimentos (ratos, células numa cultura, etc.) realizados para cada uma das várias doses d_j. O modelo de dose-resposta pode ser resumido pela expressão $n_j = N_j(\zeta + \xi d_j) + \varepsilon_j$, onde ζ, ξ são parâmetros desconhecidos e ε_j é o erro correspondente à dose d_j. O modelo acima sugere que o número de respostas positivas está relacionado linearmente com o número de

unidades experimentais não afetadas e a dose propriamente dita. O interesse central nesse problema é o *potencial cancerígeno* da substância representado pelo parâmetro ξ. Maiores detalhes podem ser encontrados em Harris (1981, 1983) e DuMouchel e Harris (1983, Apêndice A).

O problema aqui abordado é essencialmente o mesmo proposto por DuMouchel e Harris (1983). Porém, DuMouchel e Harris usaram métodos bayesianos empíricos para estimar os parâmetros do modelo, o que poderia subestimar a incerteza presente nas estimativas finais. Aqui faremos uso do Amostrador de Gibbs para calcular não só as estimativas dos parâmetros, mas suas distribuições *a posteriori*, intervalos de confiança, etc.

Na Tabela 1, reproduzida de DuMouchel e Harris (1983), têm-se os resultados de experimentos envolvendo nove substâncias e cinco espécies, isto é, cada célula não vazia da tabela constitui um estudo. Assim, 34 dos experimentos envolvem animais e apenas três envolvem humanos. Para cada um dos estudos disponíveis na tabela são fornecidas três quantidades: o potencial cancerígeno estimado $(\hat{\xi})$, o seu coeficiente de variação estimado (a razão entre o erro estimado de $\hat{\xi}$ e o próprio $\hat{\xi}$) e o logaritmo natural do potencial cancerígeno estimado. O objetivo do estudo é extrapolar os resultados obtidos com animais para humanos, isto é, seria possível preencher as células vazias daquela tabela através dos resultados obtidos do modelo hierárquico abaixo. Além disso, teoricamente é possível obter estimativas mais exatas, também, das células não vazias, visto que uma das suposições do problema é que os experimentos da Tabela 1 estão interligados.

Os dados da Tabela 1 foram obtidos a partir de experimentos laboratoriais independentes, exceto os estudos com humanos (linha 1) que são estudos epidemiológicos da incidência de câncer de pulmão em grupos de indivíduos expostos. Os dados para câncer de pulmão em humanos foram obtidos por Hammond et al. (1976) (emissões de piche), Lloyde (1971) (forno de coque) e Khan (1966) (fumantes de cigarros). Os experimentos com animais são resultados de experimentos laboratoriais, nominalmente: extratos de diclorometano em emissões de piche, forno de coque, quatro diferentes emissões de diesel de motor, uma de gasolina, poliaromático benzohidrocarbono e fumaça de cigarro. Estes experimentos laboratoriais foram conduzidos sob condições idênticas. Vale ressaltar que, para um ensaio ser incluído na tabela, teve que satisfazer três exigências: a) foi considerado como sendo confiável e reproduzível; b) a estimação da relação dose-resposta foi feita com uma amostra grande; e c) o experimento foi considerado válido para o fim de medição de carcinogenesis ou mutagenesis. Em todos os casos a estimação dos parâmetros dos modelos de dose-resposta foi feita por máxima verossimilhança - Harris (1981, 1983) e DuMouchel (1981).

Quando se trabalha com extrapolação interespécie surgem algumas questões relativas aos erros envolvidos no processo (erro experimental e entre experimentos), bem como o uso adequado dos dados da Tabela 1. Sendo assim, isolam-se formalmente os erros dentro de cada modelo de dose-resposta e o erro entre experimentos, usando para isso um sistema hierárquico de distribuições (Lindley e Smith, 1972), por meio do qual serão estimados os erros envolvidos na extrapolação e os potenciais cancerígenos das substâncias em humanos.

Para que seja possível utilizar de forma adequada os dados disponíveis, é preciso estabelecer um mecanismo comum ou hipóteses que gerem os dados da Tabela 1. Uma das hipóteses necessárias é que a razão

entre duas inclinações de dose-resposta (potência relativa) é preservada entre espécies. Essa hipótese surge naturalmente da Tabela 1, devido ao fato de que os potenciais foram medidos nas mesmas unidades, e então, as razões entre agentes são livres de unidades e portanto comparáveis.

O resto do artigo se organiza como segue. Na seção 2 discutimos brevemente o modelo e o enfoque bayesiano empírico adotado por DuMouchel e Harris, assim como a implementação do Amostrador de Gibbs para o modelo considerado, incluindo o critério de convergência usando cadeias paralelas sugerido por Gelman e Rubin (1992). Os resultados da implementação do amostrador e as conclusões finais são apresentados respectivamente nas seções 3 e 4.

Tabela 1 - Dados do problema

Estudo	Emissões de piche	Emissões de forno de coque	Emissões motores a diesel				Emissões de motores a gasolina	Benzo-pirene	Fumantes de cigarros
			A	B	C	D			
Câncer de	1,64	4,40							0,03
pulmão	1,41	0,34							0,15
(Humanos)	0,50	1,48							-3,46
Início de tumores	0,54	2,10	0,53	0,16		0,01	0,03	85,28	0,00
de pele	0,04	0,04	0,04	0,22		0,82	0,04	0,03	1,30
(Ratos)	-0,63	0,74	-0,64	-1,86		-4,51	-3,61	4,45	-5,88
Aumento de	2,07	0,86	0,65	0,07	0,13	0,04	0,20	540,00	0,58
transf. viral	0,18	0,10	0,15	0,33	0,18	0,59	0,12	0,04	0,08
(Células SHE)	0,73	-0,15	-0,44	-2,70	-2,06	-3,24	-1,59	6,29	-0,54
Mutagêneses - MA	0,31	0,73	1,66	0,27	2,55	0,16	0,35		0,59
(Ratos 5178Y)	0,39	0,21	0,31	0,43	0,16	0,24	0,11		0,23
	-1,17	-0,32	0,15	-1,31	0,93	-1,86	-1,06		-0,53
Mutagêneses + MA	9,56	9,96	1,87	0,76	1,01	0,05	0,99		0,45
(Ratos 5178Y)	0,16	0,07	0,26	0,14	0,20	0,43	0,10		0,13
	2,26	2,30	0,63	-0,27	0,01	-3,02	-0,01		-0,79

Dados de Estudos Epidemiológicos e Experimentos Laboratoriais envolvendo Nove Substâncias. Em cada célula não vazia tem-se: potencial cancerígeno (isto é, inclinação da reta de dose-resposta), o coeficiente de variação e logaritmo do potencial cancerígeno.

2. Modelo

Suponha que os experimentos estão interligados, ou seja, a célula (k,l) pode ser alcançada por alguma outra célula (k',l') através de uma série de movimentos entre experimentos observados (células não vazias).

Na prática, a não ser que uma forte *a priori* seja usada, experimentos não conectados não podem ser usados para estimar outro parâmetro e devem ser analisados separadamente. O modelo proposto por DuMouchel e Harris (1983) considera que $y_{kl} \sim N(\theta_{kl}, c_{kl}^2)$, onde y_{kl} é o logaritmo da inclinação estimada da reta de dose-resposta para o (kl)-ésimo estudo, θ_{kl} é o verdadeiro logaritmo da inclinação da curva de dose-resposta e c_{kl} são os coeficientes de variação obtidos da Tabela 1. Essa suposição é motivada pelo fato de que os y_{kl} foram estimados pelo método de máxima verossimilhança e com amostras relativamente grandes. (Observe que para uma variável aleatória positiva y com esperança e variância dadas, respectivamente, por μ e σ, $Var[\log(y)] \simeq Var[\mu^{-1}(y-\mu)] = \sigma^2/\mu^2$, o quadrado do coeficiente de variação de y).

A estrutura do modelo é introduzida assumindo que $\theta_{kl} = \mu + \alpha_k + \gamma_l + \delta_{kl}$ (k=1,...,K, l=1,...,L), onde μ é a média global, α_k o efeito da espécie, γ_l o efeito do agente e δ_{kl} representa um efeito de interação agente-espécie. A hipótese de que a potência relativa de dois agentes é preservada em média entre espécies é assegurada assumindo que dado σ, δ_{kl} tem distribuição *a priori* com média zero e variância σ^2.

Resumindo, condicionado a σ, $y_{kl}=\mu+\alpha_k+\gamma_l+\delta_{kl}+\varepsilon_{kl}$. k=1,...,K e l=1,...,L onde μ, α_k e γ_l são parâmetros cujas distribuições *a priori* serão estabelecidas posteriormente, e assumindo independência condicional $\delta_{kl}|\sigma \sim N(0,\sigma^2)$ e $\varepsilon_{kl} \sim N(0,c^2{}_{kl})$. Para descrever a especificação da distribuição *a priori* é conveniente re-escrever o modelo na forma matricial $E(\theta|\beta,\sigma^2)=X\beta$, onde $\beta=(\mu,\alpha_1,...,\alpha_{K-1},\gamma_1,...,\gamma_{L-1}$ e a matriz de delineamento X tem posto completo. *A priori* assume-se que $\beta|\sigma^2 \sim N(b,V)$ e que σ tem densidade $\pi(\sigma)$. Finalmente, denotando por Y o vetor de observações e por C a matriz diagonal com elementos c_{kl}^2, pode-se escrever o modelo em forma hierárquica como:

$$\begin{aligned} Y|\theta,\beta,\sigma &\sim N(\theta,C) \\ \theta|\beta,\sigma &\sim N(X\beta,\sigma^2 I) \\ \beta|\sigma &\sim N(b,V) \\ \log\sigma &\sim \pi(\sigma) \end{aligned} \qquad (1)$$

No modelo (1)) informação difusa *a priori* sobre β pode ser considerada fazendo $V = tV_0$ com $t \to \infty$. Neste caso, a densidade *a posteriori* de σ satisfaz que

$$\pi(\sigma \mid Y) \alpha \pi p(\sigma) \mid W \mid^{\frac{1}{2}} \mid X'WX \mid^{-\frac{1}{2}} \exp\{-\frac{1}{2} Y'SY\}, \qquad (2)$$

onde $W=(C-\sigma^2 I)^{-1}$ e $S = W-WX(X'WX)^{-1}X'W$. Assumindo que $\log(\sigma)$ tem distribuição uniforme no intervalo [log(0.05),log(5.0)], DuMouchel e Harris (1983) usaram métodos bayesianos empíricos para aproximar as características da distribuição *a posteriori* de θ. Num dos enfoques usados por eles, usaram os dados Y para estimar por máxima verossimilhança σ e β. Seja $\hat{\sigma}_{MLE}$ o estimador de $\hat{\sigma}$ e $\hat{\beta}_{MLE}=(X'\hat{W}X)^{-1}X'\hat{W}Y$, onde $\hat{W} = (C-\hat{\sigma}^2_{MLE}I)^{-1}$. Considerando $\sigma=\hat{\sigma}_{MLE}$ e $\hat{\beta}_{MLE}$ como se fossem os verdadeiros valores de σ e β obtém-se então uma aproximação para a distribuição *a posteriori* de θ dada pela distribuição $N(\hat{\theta}_{MLE},\hat{C}_{MLE})$, onde $\hat{\theta}_{MLE}=\hat{C}_{MLE}[C^{-1}Y+\hat{\sigma}^{-2}_{MLE}X\hat{\beta}_{MLE}]$ e $\hat{C}_{MLE}=[C^{-1}+\hat{\sigma}^{-2}_{MLE}I]^{-1}$. De forma análoga obtém-se as estimativas *a posteriori* dos outros parâmetros. O problema da estimação bayesiana empírica é que existe um erro que não está sendo considerado, visto que estamos estimando alguns parâmetros e aceitando-os como se fossem os valores reais, e isto pode potencialmente subestimar a incerteza presente nas estimações dos parâmetros.

Neste trabalho consideramos uma abordagem puramente bayesiana que usa o Amostrador de Gibbs para gerar observações da distribuição *a posteriori* e assim aproximar as quantidades de interesse. Estas aproximações dependem apenas do tamanho da simulação, que pode ser controlado computacionalmente. Para implementar o Amostrador é necessário reescrever o modelo para distinguir as células observadas das vazias. Considere partições dos vetores $y=(y^o,y^{no})$ e $\theta = (\theta^o,\theta^{no})$ onde os supra-índices o e no referem-se as células observadas e não-observadas respectivamente. Denotando por $\tau = \sigma^{-2}$ a precisão entre experimentos e por $\beta^* = X\beta$, as distribuições condicionais completas são conseqüências direta dos resultados de Lindley e Smith (1972). A condicional completa de τ surge do fato que $\pi(\tau|\beta^*,\theta,y^O)\alpha f(y|\theta)\pi(\theta|\beta^*,\tau)\pi(\beta^*)\pi(\tau)$ assim,

$$\pi(\tau\,|\,\beta^*,\theta,y^O)\alpha^{\frac{n}{2}-1}\exp\{-\frac{\tau}{2}(\theta-\beta^*)'(\theta-\beta^*)\}.$$

$$\pi(\tau|\beta^*,\theta^o,\theta^{no},y^o)\sim Gama[n/2,\frac{(\theta-\beta^*)'(\theta-\beta^*)}{2}],\quad \text{truncada em } 0.04\le\tau\le 400$$

$$\pi(\beta^*|\tau,\theta^o,\theta^{no},y^o)\sim N[X(X'X)^{-1}X'\theta,X(X'X)^{-1}X'\tau^{-1}]$$

$$\pi(\theta^{no}|\beta^*,\theta^o,\tau,y^o)\sim N[\beta^{*no},I\tau^{-1}]$$

$$\pi(\theta^o|\beta^*,\theta^{no},\tau,y^o)\sim N[(C^{o-1}+\tau I)^{-1}(C^o y^o+\tau\beta^{*o});(C^{o-1}+\tau I)^{-1}].$$

Note que os parâmetros b e V não aparecem nas condicionais completas acima. Isto é conseqüência de considerar no Modelo (1) $V=tV_0$ com $t\to\infty$, embora também é possível obter as condicionais completas com b e V arbitrários.

Para diagnosticar convergência do amostrador usamos o critério de Gelman e Rubin (1992) baseado em cadeias múltiplas. Concretamente, simulamos 10 cadeias paralelas de tamanho $2n$ com distribuição inicial sobre dispersa e descartamos os primeiros n ciclos de cada cadeia. Supondo que desejamos aproximar $E[g(\theta, \beta^*, \tau) | Dados]$, sejam B e W os erros quadráticos médios dos valores simulados de g entre e dentro das cadeias respectivamente. Sob a hipótese de convergência, a variância *a posteriori* de g pode ser estimada sem viés por $\sigma_g^2 = [(n\text{-}1)B+W]/n$, embora se estivermos longe da convergência σ_g^2 sobrestima a verdadeira variância *a posteriori*. O critério de Gelman e Rubin diagnostica convergência quando a *redução potencial estimada da escala* $\hat{R} = \sqrt{\sigma_g^2/W}$ está próxima de 1. Para o cálculo de $\hat{R}$ e os correspondentes percentis superiores usamos um programa elaborado por Gelman e Rubin (pode ser obtido em Statlib: http: //www.lib.stat.cmu.edu).

3. Resultados

3.1 Verificação de Convergência

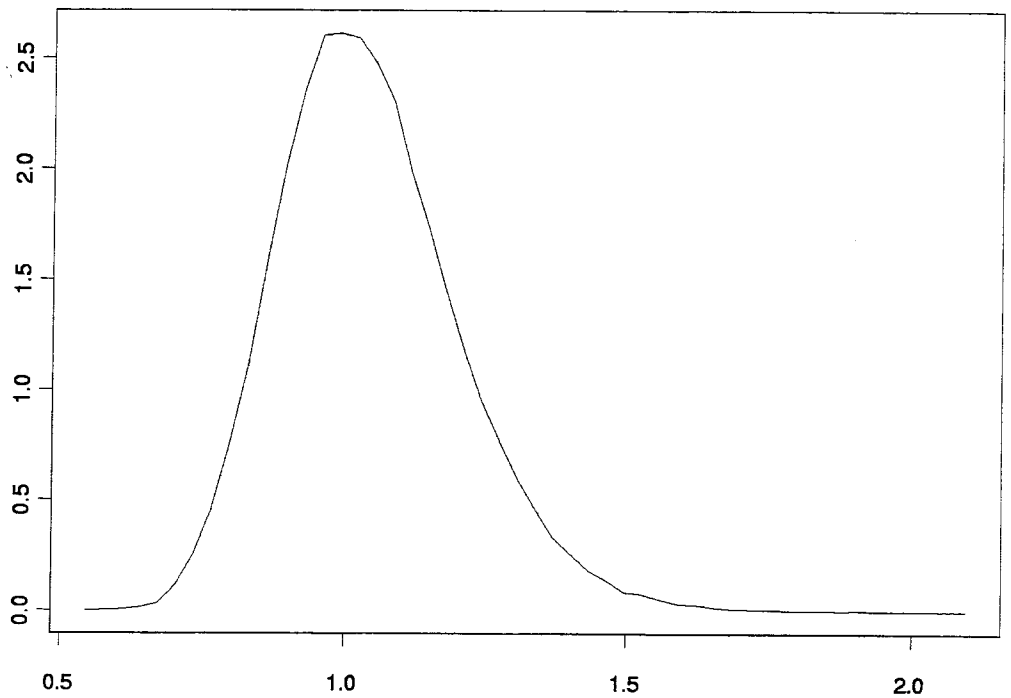
Foram simuladas 10 cadeias paralelas monitorando-se a redução potencial de escala $\hat{R}$ para as 45 componentes do vetor θ e para o erro entre experimentos $\sigma = 1/\sqrt{\tau}$. A simulação foi continuada até que o percentil 97.5% de $\hat{R}$ fosse em todos os casos inferior a 1.01. Isto significa que continuar a simulação *ad-infinitum* poderia melhorar as aproximações obtidas no máximo em 1%. O número de ciclos requeridos para cada cadeia foi de 3000. Isto permite aproximar as quantidades de interesse, após descartar os primeiros 1500 ciclos de cada cadeia, usando essencialmente uma amostra de tamanho (10)(1500)=15000. Notamos ainda que os percentis teóricos baseados na distribuição t de Student sugeridos por Gelman e Rubin são bastante próximos dos percentis empíricos baseados nos valores amostrados. Por este motivo usaremos no que resta os últimos.

3.2 Distribuição de σ

Um ponto importante a ser considerado neste problema diz respeito ao truncamento da distribuição *a priori* (e da condicional completa) de τ. Gerar valores de distribuições truncadas pode ser complicado quando o truncamento ocorre em uma região com baixas probabilidades. Felizmente isso não ocorre neste problema. Assim, podem-se amostrar valores da distribuição de τ e aceitar somente aqueles que estão no intervalo de interesse. Esse procedimento se mostrou bastante eficiente, pois não houve acréscimo significativo no tempo de simulação. A freqüência de ocorrência de dados fora do intervalo de truncamento é da ordem de dois valores fora do intervalo para cada 1000 ciclos. Isso na prática torna desnecessário o truncamento, tanto no nosso enfoque quanto no de DuMouchel e Harris. Na Figura 1, apresentamos a densidade marginal *a posteriori* aproximada de σ, obtida usando um histograma suavizado com os valores gerados por Gibbs. O erro entre experimentos é

estimado por $\hat{E}[\hat{\sigma}\,|Dados]$=1.065 com desvio-padrão de $\hat{DP}\,[\sigma\,|Dados]$=0.188. Calculando um intervalo de credibilidade, tem-se que, $P(0.77 \leq \sigma \leq 1.5)$=0.95. Portanto, pode-se concluir que a extrapolação se mostrou bastante confiável. Como uma forma alternativa de verificar a adequabilidade do modelo, repetiu-se o amostrador usando *a priori* $\log\sigma \sim N(\log 2, \log^2 2)$, também proposta por DuMouchel e Harris (1983). Dessa forma os resultados se mostraram bastante similares ao modelo com *a priori* (1). Por exemplo, as estimativas $\hat{\sigma}$=1.044 e $\hat{DP}(\sigma)$=0.158 são muito próximas daquelas do caso anterior e o mesmo acontece com o intervalo de credibilidade $P(0.78 \leq \sigma \leq 14)$=0.95. Isso nos leva a concluir pela adequabilidade do modelo.

Figura 1 - Distribuição marginal do erro entre experimentos σ



Na Tabela 2, têm-se as estimativas de θ, juntamente com seus valores observados. A avaliação destes valores estimados constitui o ponto central da análise. Primeiramente, note-se que, para um logaritmo da inclinação observada $y_{1,1}$=0.50 (piche em humanos), tem-se $\hat{E}(y_{1,1})=\hat{\theta}_{1,1}$=0.14. Essa diferença deve-se aos fatos de: i) a informação original para esta combinação espécie-agente ser bastante imprecisa-estimativa pontual de 0.50 com desvio de 1.41; e ii) está sendo utilizada, via a estrutura do modelo, informação *a priori*

relativamente forte para $\theta_{1,1}$. Note-se ainda, que em alguns casos, o intervalo de credibilidade é relativamente grande (por exemplo: diesel A, B, C, etc.). Porém, isso é de certa forma esperado, uma vez que se está trabalhando com dados bastante dispersos. Assim, se os coeficientes de variação forem altos, haverá intervalos de credibilidade grandes. Por exemplo, em forno de coque em humanos, há um baixo erro-padrão (0.34) e o intervalo de credibilidade ($0.73 \leq \hat{\theta}_2 \leq 2.04$) é pequeno quando comparado com os experimentos nos quais têm-se coeficientes de variação maiores. Das substâncias estudadas, a que se mostrou mais perigosa ao homem foi o benzopirene, para o qual o logaritmo do potencial de causar câncer de pulmão é de aproximadamente 5.62. O benzopirene também causa distúrbios genéticos em ratos com ou sem a presença de ativador metabólico, sendo que, com a utilização do ativador, o potencial é duplicado, isto é, tem-se um aumento de $\hat{\theta}_{3,4}$=6.46 para $\hat{\theta}_{4,4}$=7.2. Nas Figuras 2 e 3, mostramos as aproximações das densidades marginais *a posteriori* de $\theta_{1,2}$ e $\theta_{5,8}$ respectivamente.

Tabela 2 - Tabela de Resultados - Estimativas *a posteriori* do potencial cancerígeno das substâncias

Estudo	Emissões de piche	Emissões de forno de coque	Emissões de motores a diesel				Emissões de motores a gasolina	Benzo-pirene	Fumantes de cigarros
			A	B	C	D			
Câncer de pulmão (Humanos)	1,93 (1,64)	4,19 (4,40)	2,07	0,44	0,91	0,09	0,39	1096,05	0,03 (0,03)
	1,01(1,41)	0,34(0,34)	1,45	1,44	1,48	1,47	1,43	1,57	0,15(0,15)
	0,14(0,50)	1,38(1,48)	-0,44	-1,98	-1,23	-3,58	-2,02	5,62	-3,43(-3,46)
	{-1,81,2,12}	{0,73,2,04}	{-3,28,2,5 }	{-4,78,0,93}	{-4,15,1,71}	{-6,47,-0,67}	{-4,85,0,8}	{2,63,8,8}	{-3,73 ,-3,15}
Início de tumores de pele (Ratos)	0,53(0,54)	2,09(2,10)	0,53(0,53)	0,15(0,16)	0,36	0,02(0,01)	0,03(0,03)	85,69(85,28)	0,03(0,00)
	0,04(0,04)	0,04(0,04)	0,04(0,04)	0,22(0,22)	1,34	0,69(0,82)	0,26 (0,26)	0,03(0,03)	0,96(1,30)
	-0,63(-0,63)	0,74(0,74)	-0,64(-0,64)	-1,9(-1,86)	-1,97	-4,44(-4,51)	-3,55(-3,61)	4,45(4,45)	-4,12(-5,88)
	{-0,71,-0,55}	{0,66,0,81}	{-0,72,-0,56}	{-2,32,-1,47}	{-4,67,0,68}	{-5,8,-3,09}	{-4,06,-3,06}	{4,39,4,51}	{-6,09,-2,3}
Aumento de transf. viral (Células SHE)	2,08(2,07)	0,87(0,86)	0,65(0,65)	0,08(0,07)	0,13(0,13)	0,04 (0,04)	0,2(0,20)	539,01(540,00)	0,58(0,58)
	0,18(0,18)	0,1(0,10)	0,15(0,15)	0,32(0,33)	0,18(0,18)	0,53(0,59)	0,12(0,12)	0,04(0,04)	0,08(0,08)
	0,71(0,73)	-0,14(-0,15)	-0,44(-0,44)	-2,61(-2,70)	-2,03(-2,06)	-3,28(-3,24)	-1,59(-1,59)	6,29 (6,29)	-0,55(-0,54)
	{0,36,1,06}	{-0,33,0,05}	{-0,73,-0,14}	{-3,24,-1,98}	{-2,39,-1,68}	{-4,32,-2,23}	{-1,83,-1,36}	{6,21,6,37}	{-0,7,-0,39}
Mutagêneses - MA (Rato 5178Y)	0,42(0,31)	0,79(0,73)	1,72(1,66)	0,3(0,27)	2,49(2,55)	0,15(0,16)	0,35(0,35)	1845,91	0,58(0,59)
	0,38 (0,39)	0,21(0,21)	0,3(0,31)	0,41(0,43)	0,16(0,16)	0,23(0,24)	0,11(0,11)	1,4	0,23(0,23)
	-0,93(-1,17)	-0,26(-0,32)	0,5(0,51)	-1,29(-1,31)	0,9(0,93)	-1,91(-1,86)	-1,06(-1,06)	6,46	-0,57(-0,53)
	{-1,67,-0,19 }	{-0,67,0,15 }	{-0,1,1,09 }	{-2,08,-0,49 }	{0,58,1,21}	{-2,37,-1,45}	{-1,28,-0,85}	{3,68,9,21}	{-1,02,-0,13}
Mutagêneses + MA (Rato 5178Y)	9,53 (9,56)	9,99 (9,96)	2(1,87)	0,77(0,76)	1,05(1,01)	0,06(0,05)	0,99(0,99)	3884,78	0,46(0,45)
	0,16(0,16)	0,07(0,07)	0,26(0,26)	0,14(0,14)	0,2(0,20)	0,41(0,43)	0,1(0,10)	1,41	0,13(0,13)
	2,24(2,26)	2,3(2,30)	0,66(0,63)	-0,27(-0,27)	0,02(0,01)	-2,87(-3,02)	-0,01(-0,01)	7,2	-0,79(-0,79)
	{1,93,2,55}	{2,16,2,44}	{0,16,1,16}	{-0,54,0}	{-0,36,0,42}	{-3,68,-2,07}	{-0,21,0,18}	{4,46,10}	{-1,04,-0,53}

Cada experimento da tabela é organizado como segue: na primeira linha tem-se a inclinação estimada (potencial) e entre parênteses a inclinação observada; na linha dois, o coeficiente de variação estimado e entre parênteses o coeficiente de variação observado; na linha três o logaritmo do p timado e entre parênteses o logaritmo do potencial observado; e finalmente na quarta linha, o intervalo de credibilidade de nível 95% do logaritmo do potencial.

Figura 2 - Distribuição marginal *a posteriori* de $\theta_{1,2}$ - potencial cancerígeno das substâncias emitidas pelo forno de coque em causar câncer de [illegible]mão em humanos

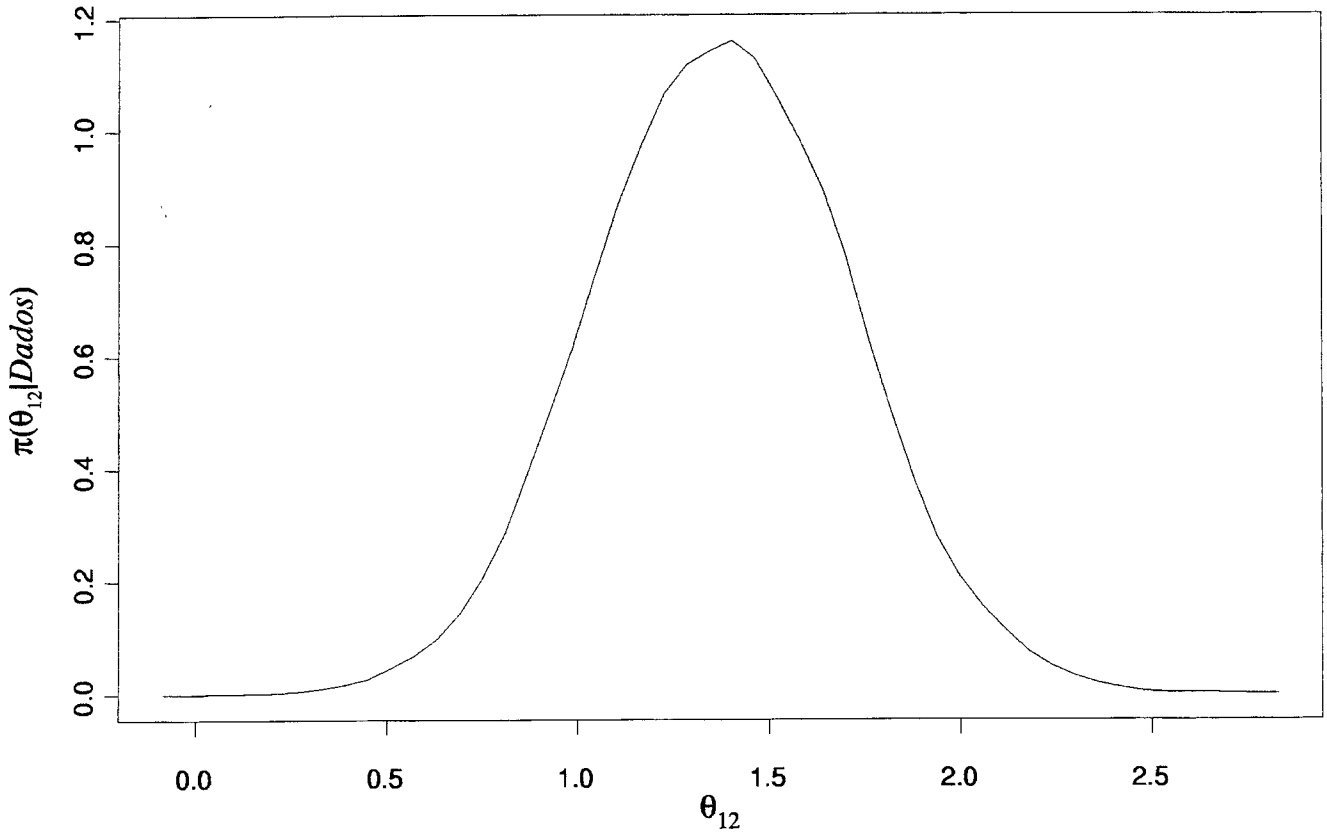
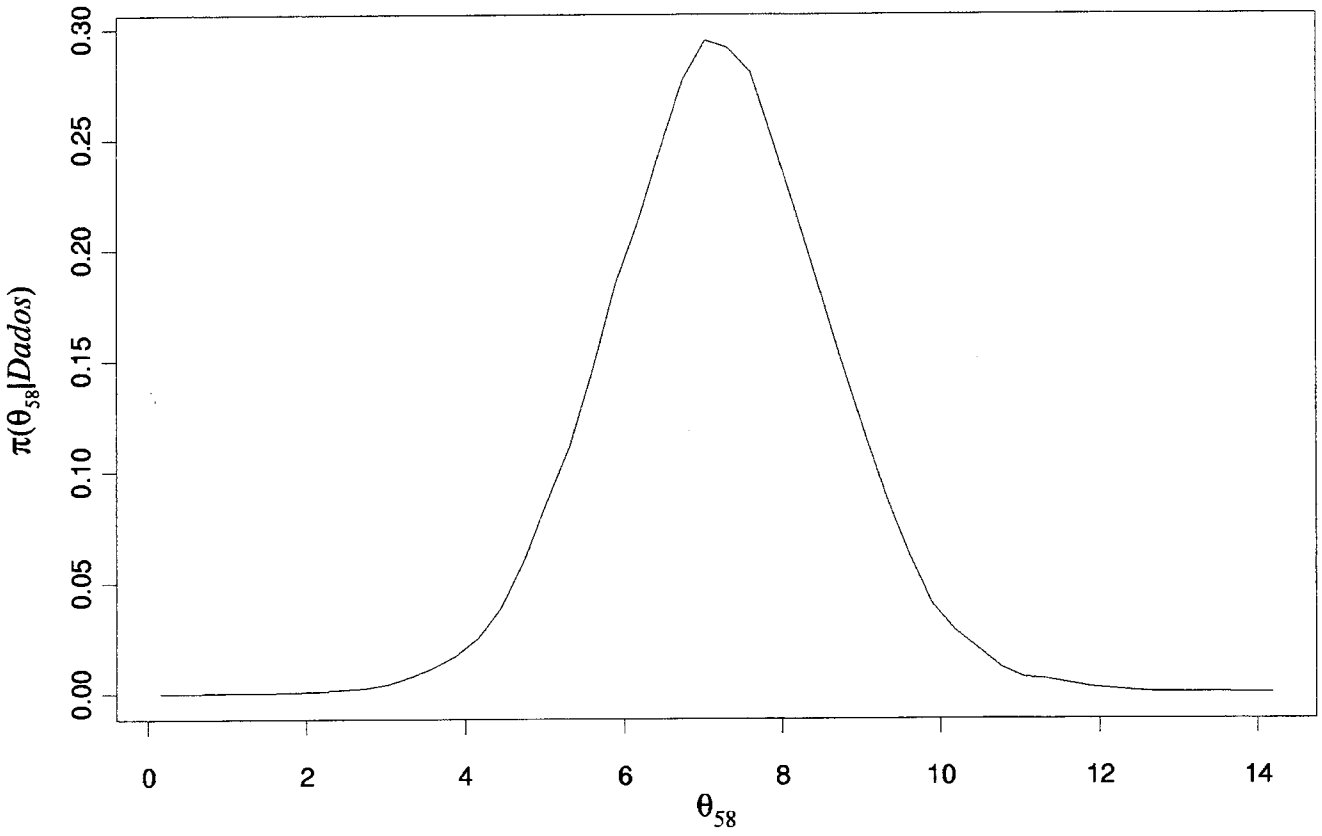


Figura 3 - Distribuição marginal *a posteriori* de $\theta_{5,8}$ - potencial cancerígeno do benzopirene em causar mutações genéticas + MA em ratos



4. Conclusões

Os resultados obtidos aqui são bastantes similares aos obtidos por Dumouchel e Harris (1983) usando métodos bayesianos empíricos. Porem isto só é conhecido após a implementação do Amostrador de Gibbs. Em geral o uso de métodos bayesianos empíricos apresenta o perigo potencial de indicar uma redução artificial da incerteza presente nas inferências finais (por exemplo, Gaver et al., 1992).

A amplitude dos intervalos de credibilidade poderia ser reduzida de duas formas. A primeira é obter informação *a priori* mais precisa sobre os hiperparâmetros β. A segunda, discutida por DuMouchel e Harris, é escolhendo um subconjunto dos experimentos descritos na Tabela 1 de forma a diminuir o erro entre experimentos σ. Em particular, DuMouchel e Harris retiraram gradativamente os estudos mutagêneses - MA, cigarros, tumor de pele e forno de coque, e obtiveram então um novo valor de $\hat{\sigma}$ =0.40. A implementação do

Amostrador para subconjuntos dos experimentos procede essencialmente da mesma forma descrita acima. Como as conclusões obtidas também são bastante similares as de DuMouchel e Harris, por motivo de espaço temos optado por não reportar os resultados aqui.

Referências bibliográficas

DUMOUCHEL, WILLIAM H. E HARRIS, JEFFREY E. (1983) Bayes Methods for Combining thet Results of Cancer Studies in Humans and Other Species, *Journal of the American Statistical Association*, 78, 382, 293-315.

DUMOUCHEL, WILLIAM H. (1981) Documentation for CATDATA., *Cambridge, Massachusetts: Massachusetts Institute of Technology, Statistics Center Technical Report.*

FREEDMAN, D. A. E ZEISEL, H. (1992) Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, 7, 4, 457-472.

GAVER, D. P., D. DRAPER, P. K. GOEL, J. B. GREEHOUSE, L. V. HEDGES,C. N. MORRIS E C. WATERNAUX (1972) On Combining Information: Statistical Issues and Opportunities for Research, *Report of a panel of the Committee on Applied and Theoretical Statistics of the National Research Council of the USA.*

GELFAND, ALAN E. E SMITH, ADRIAN F. (1990a) Sampling-Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, 85, 398-409.

GELFAND, ALAN E. E SMITH, ADRIAN F. (1990b) Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling, *Journal of the American Statistical Association*, 85, 412, 972-985.

HAMMOND, E. C., SELIKOFF, I. J., LAWTHER, P. L. E SEIDMAN, H. (1976) Inhalation of Benzpyrene and Cancer in Man, *Annals of the New York Academic of Sciences*, 271, 116-124.

HARRIS, J. E. (1981) Potential Risk of Lung Cancer from Diesel Engine Emissions, National Academic Press.

HARRIS, J. E. (1983) Diesel Emissions and Lung Cancer, Risk Analysis.

KHAN, H. A. (1966) The Dorn Study of Smoking and Mortality Among U. S. Veterans: Report on Eight and One Half Years of Observations, *Epidemiological Approaches to the Study of Cancer and other Chronic Diseases. ed. W. Haenszel. National Cancer institute Monograph*, 13, 1-41.

LLOYD, W. J. (1971) Long-Term Mortality Study of Steel-workers vs Respiratory Cancer in Coke Oven Workers, *Journal of Occupational Medicine*, 13, 53-68.

MEYN, S. P. E TWEEDIE, R. L. (1993) Markov Chains and Stochastic Stability, Springer-Verlag.

RUBINSTEIN, REUVEN Y. (1981) Simulation and the Monte Carlo Method", John Wiley and Sons.

GELMAN, A. E RUBIN, DONALD B. (1992) Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, 7, 4, 457-511.

TIERNEY, LUKE (1994) Markov Chains for Exploring Posterior Distributions, *Annals of Statistics*, 22, 4, 1701-1729.

ZELLNER, ARNOLD E MIN, CHUNG-KI (1995) Gibbs Sampler Convergence Criteria, *Journal of the American Statistical Association*, 431, 90, 921-927.

Abstract

We study a hierarquical model proposed by DuMouchel and Harris to combine the results of studies about the cancer risk posed by several substances to humans and other species. Originally the authors proposed using empirical bayes procedures to estimate the quantities of interest. This may, at least in theory, underestimate the uncertainty in the final estimates. Here we consider a fully Bayesian approach that uses the Gibbs' sampler to simulate observations from the posterior distribution of the parameters.

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 apresenta 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 devem ter conteúdo estatístico substancial. As análises devem 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; e
k) modelos e métodos econométricos.

Todos os artigos submetidos serão avaliados quanto à qualidade e à relevância por dois especialistas indicados pelo Comitê Editorial da Revista Brasileira de Estatística. 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 - 2514 4548
Fax: +55 - 21 - 2514 0039
E-mail: pedrosilva@ibge.gov.br

Para cada artigo publicado, serão fornecidas gratuitamente 20 separatas.

Instruções para preparo de originais:

Os originais entregues para publicação devem obedecer às seguintes normas:

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

Se o assunto é **Brasil**,
procure o **IBGE**

www.ibge.gov.br
wap.ibge.gov.br

atendimento
0800 218181