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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 63 número 219 janeiro/junho 2002

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 63, n. 219, p. 1-100, jan./jun. 2002

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:

Renato Martins Assunção
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.

Impressão
GráficaDigital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2004.

Capa
Renato J. Aguiar – Coordenação de Marketing/CDDI

Ilustração da Capa
Marcos Balster – Coordenação de Marketing/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

A Revista Brasileira de Estatística - RBEs – publica artigos em português que possuam uma ênfase em aplicações de métodos estatísticos. É claro que a revista não é apenas isso mas eu gostaria de aproveitar esta oportunidade para exaltar as virtudes deste tipo de artigos que a RBEs publica.

Em minha opinião, um importante propósito da RBEs é encorajar a produção de artigos que apliquem métodos estatísticos de forma correta e que possam ser lidos por um público amplo composto por acadêmicos mas também por estudantes de graduação e por estatísticos trabalhando no mercado. Não existe hoje no Brasil outra revista que atenda às necessidades de informação e de formação desse público, um público fundamental se a estatística, enquanto ciência e prática metodológica, quiser ter o reconhecimento social e profissional que merece. É esse público amplo que vai formar junto à sociedade brasileira a imagem que a estatística e os estatísticos terão e é ele que, de forma direta ou indireta, traz também benefícios para os programas de pós-graduação no país. Afinal, é a demanda por estudar mais e melhor as novas e antigas técnicas estatísticas que traz de volta ou mantém na universidade os ex-estudantes de graduação. Por outro lado, essa demanda é criada pela necessidade de uma formação sólida num mercado competitivo que, atribuindo grande importância à análise de dados, demanda um trabalho de qualidade de seus estatísticos.

Com este argumento eu gostaria de enfatizar a necessidade de que os programa de pós-graduação em estatística estimulem seus alunos e os orientadores a submeterem parte de suas dissertações de mestrado ou teses de doutorado à Revista Brasileira de Estatística. A todo instante, esses programas estão produzindo trabalhos aplicados de excelente qualidade que seriam de muito interesse para o público da RBEs. É claro que a produção de bons trabalhos estatísticos também ocorre fora da academia, e nós da RBEs estamos muito interessados nestas contribuições também. No entanto, ainda é na pós-graduação brasileira que a maior parte desses bons trabalhos são feitos e por isto esta nota editorial está focada nesta fonte. De fato, existe uma enorme quantidade de dissertações de mestrado produzidas no país que não são publicadas. Entretanto, esses trabalhos possuem textos cuidadosamente escritos e revisados pelos orientadores e banca, com análises de dados originais que poderiam ser muito úteis para aquele público mencionado acima. Assim, gostaria de encorajar os alunos de pós-graduação e os seus orientadores para que submetam artigos à Revista Brasileira de Estatística. A estatística desenvolvida e aplicada no Brasil só terá a ganhar com a divulgação desses trabalhos.

Neste número, temos quatro artigos sendo publicados mostrando a diversidade de aplicações e métodos que são de interesse para nosso público. Boa leitura de mais um número da Revista Brasileira de Estatística.

Renato Martins Assunção
Editor Responsável

Modelos lineares hierárquicos em pesquisas por amostragem - relacionando o índice de massa corporal às variáveis da pesquisa sobre padrões de vida/IBGE

Solange Trindade Corrêa*
Denise Britz do Nascimento Silva*

Resumo

Dados com estrutura hierárquica são comumente encontrados em pesquisas sociais. Quando estes dados são obtidos através de pesquisas por amostragem, surge a questão da incorporação dos pesos amostrais e do plano amostral na análise multinível, uma vez que a realização desta análise sem a incorporação do plano amostral pode acarretar estimadores viciados dos parâmetros do modelo (Pfeffermann et al., 1998).

O objetivo deste trabalho é aplicar o procedimento de ponderação desenvolvido por Pfeffermann et al. (1998), que incorpora o plano amostral na estimação dos parâmetros de modelos multiníveis, aos dados da Pesquisa sobre Padrões de Vida -PPV-, realizada pelo IBGE nos anos de 1996-1997. O desempenho do método é ilustrado através do ajuste de um modelo hierárquico de dois níveis (modelo de componentes de variância) aos dados da PPV. A análise considera o Índice de Massa Corporal -IMC- de pessoas com 20 anos ou mais de idade como variável resposta do modelo e as seguintes variáveis explicativas: sexo, idade, raça, anos de estudo, renda domiciliar mensal per capita, despesa domiciliar diária per capita com alimentação e área de localização do domicílio. A análise do modelo indica que, em média, o IMC estimado é maior para pessoas moradoras em domicílios localizados em área urbana, com maior renda domiciliar mensal per capita e maior despesa domiciliar diária per capita com alimentação. Além disso, conclui-se que, em média, o IMC de um indivíduo decresce com o aumento do seu nível de instrução e cresce de acordo com a idade até a faixa etária de 40 a 49 anos. Para homens acima desta faixa, o IMC diminui, apresentando um leve acréscimo para mulheres de 50 anos de idade ou mais.

* Endereços para correspondências: Escola Nacional de Estatísticas - ENCE/IBGE - Rio de Janeiro, Brasil e-mail: scorrea@ibge.gov.br e denisesilva@igbe.gov.br

O procedimento de estimação ponderado, implementado neste trabalho, é comparado ao método de estimação convencional que ignora os pesos amostrais e o plano amostral. Os resultados indicam a necessidade de se incorporar o plano amostral na análise multinível, visando a evitar estimativas viciadas dos parâmetros do modelo.

Palavras-chave: Índice de massa corporal, mínimos quadrados generalizados iterativo, modelos hierárquicos, não-resposta, pesquisas por amostragem.

1. Introdução

Dados com estrutura hierárquica são freqüentemente encontrados em pesquisas sociais. Um exemplo tradicional é o estudo na área de educação onde se deseja relacionar, por exemplo, as notas dos alunos com o tempo de experiência do professor e o nível socioeconômico médio da turma. Nesta situação, os estudantes agrupam-se em turmas, as turmas em escolas, as escolas em distritos escolares, e assim por diante, com variáveis descrevendo características dos alunos e das turmas. Os modelos hierárquicos, ou modelos multiníveis, são uma classe importante de modelos de regressão adequados para representar dados com tal estrutura, pois incorporam explicitamente o efeito de conglomeração das unidades de análise, permitindo medir a variabilidade entre elas.

Em pesquisas por amostragem, como as realizadas pelo Instituto Brasileiro de Geografia e Estatística - IBGE -, é muito freqüente a utilização de planos amostrais complexos com o uso de estratificação e conglomeração das unidades de análise, seleção das unidades amostrais em vários estágios e probabilidades distintas de seleção. Geralmente, estes planos amostrais são ignorados quando se ajusta modelos multiníveis a dados de pesquisas amostrais complexas.

Entretanto, o estudo da estrutura hierárquica da população, incorporando o plano amostral utilizado para a obtenção dos dados é de grande interesse para os analistas de dados de pesquisas sociais. É importante destacar, ainda, que a estrutura hierárquica é uma propriedade intrínseca à população de estudo, e que a utilização de modelos para descrevê-la é motivada por sua própria estrutura, independentemente do procedimento amostral utilizado para a obtenção das informações.

Quando consideram-se dados obtidos por amostragem complexa, a prática de ignorar o plano amostral no ajuste de modelos hierárquicos é, por vezes, justificada pelos analistas sob o argumento de que a inclusão de características do desenho amostral, tais como variáveis indicadoras de estratos e/ou conglomerados, como variáveis explicativas do modelo seria suficiente para "representar" o plano amostral. Contudo, este argumento pode não ser adequado quando as unidades populacionais são selecionadas com probabilidades desiguais, sendo necessário incorporar o plano amostral apropriadamente na análise, visando a minimizar os vícios na estimação dos parâmetros do modelo. Por exemplo, os estimadores dos parâmetros do modelo hierárquico disponíveis nos pacotes estatísticos convencionais podem ser assintoticamente viciados se as probabilidades de seleção das unidades, em qualquer nível da hierarquia, são relacionadas à variável resposta (ou mais precisamente, aos erros aleatórios) mesmo depois de condicionadas às variáveis explicativas do modelo.

O propósito deste trabalho é aplicar aos dados da Pesquisa sobre Padrões de Vida, realizada pelo IBGE no período de 1996-1997, o procedimento de *ponderação pelas probabilidades* na estimação dos parâmetros de

modelos lineares hierárquicos, desenvolvido por Pfeffermann et al. (1998). Nesta aplicação, relaciona-se o Índice de Massa Corporal -IMC- às variáveis socioeconômicas da pesquisa, comparando as estimativas obtidas por esta abordagem àquelas obtidas pelos procedimentos usuais de estimação.

2. Incorporação do plano amostral na estimação dos modelos lineares hierárquicos

O procedimento de ponderação aplicado aos modelos de regressão usuais pode ser considerado uma aplicação da abordagem de "máxima pseudo-verossimilhança" (MPV), como descrito por Skinner, et al. (Capítulo 3, 1989). A idéia básica desta abordagem é que a seleção da amostra não acarretaria vícios na estimação dos parâmetros do modelo, se os valores das variáveis de interesse fossem observados para todas as unidades da população (como em um censo). Nesta situação, seria possível definir a função de verossimilhança populacional e, através da maximização desta função, obter estimadores consistentes para os parâmetros do modelo.

Quando modelos de regressão são ajustados a um conjunto de dados, as variáveis aleatórias correspondentes às variáveis de interesse para as unidades da população finita são consideradas independentes, implicando que a função de verossimilhança pode ser representada pelo produto das funções de densidade de probabilidade daquelas variáveis aleatórias. Conseqüentemente, o logaritmo da função de verossimilhança é dado por uma soma. No caso de dados provenientes de pesquisas por amostragem, esta soma é estimada consistentemente por simples ponderação das observações amostrais, onde o peso de cada observação é, usualmente, igual ao inverso de sua probabilidade de inclusão na amostra. Do ponto de vista matemático, isto é escrito como segue.

Considere y_i o valor observado da variável de pesquisa Y_i para o elemento i da população U, onde $i = 1,...,N$. Suponha que $Y_1,...,Y_N$ são independentes e identicamente distribuídas (IID) com densidade $f(y;\theta)$, onde θ é um vetor de parâmetros desconhecidos de interesse de dimensão $k \times 1$. Supondo conhecidos os elementos da população finita, a função de verossimilhança populacional seria dada por

$$l_U(\theta) = \prod_{i=1}^{N} f(y_i;\theta)$$

e seu logaritmo por

$$L_U(\theta) = \sum_{i=1}^{N} \log[f(y_i;\theta)]$$

Calculando as derivadas parciais de $L_U(\theta)$ em relação a cada componente de θ e igualando a zero, obtemos as equações de verossimilhança populacionais dadas por

$$\sum_{i=1}^{N} \mathbf{u}_i(\theta) = 0$$

onde $\mathbf{u}_i(\theta) = \dfrac{\partial \log[f(y_i;\theta)]}{\partial \theta}$ é o vetor de dimensão $k \times 1$ dos escores do elemento i, $i = 1,...,N$.

Sob certas condições de regularidade, a solução θ_U deste sistema é o Estimador de Máxima Verossimilhança de θ no caso de um censo. Como θ_U não é calculável a menos que um censo seja realizado, θ_U passa a ser visto como um pseudo-parâmetro que é o alvo da inferência que incorpora o plano amostral. De acordo com Skinner, et al. (Capítulo 3, 1989), o estimador de Máxima Pseudo-Verossimilhança -MPV- $\hat{\theta}_{MPV}$ de θ_U (e conseqüentemente de θ) será a solução das equações de Pseudo-Verossimilhança dadas por

$$\sum_{i=1}^{n} w_i \mathbf{u}_i(\theta) = 0$$

onde w_i são pesos propriamente definidos, usualmente igual ao inverso da probabilidade de inclusão da unidade i na amostra.

Isto é, o valor do parâmetro que maximiza o logaritmo da função de verossimilhança estimada é o estimador de MPV que, sob condições gerais, é consistente para o parâmetro do modelo correspondente (Detalhes em Skinner, et al., 1989).

No modelo multinível, o procedimento de ponderação é diferente dos convencionais, pois as observações da população finita não são independentes. Por conseguinte, o logaritmo da função de verossimilhança populacional não é uma simples soma dos valores da população finita, implicando que não pode ser estimada por simples ponderação das observações amostrais.

Pfeffermann et al. (1998) desenvolveram uma forma de incorporar o plano amostral e os pesos amostrais das unidades de análise no ajuste de modelos hierárquicos para compensar diferentes probabilidades de inclusão das unidades na amostra.

2.1 Especificação do modelo e do plano amostral

Considere uma população com estrutura hierárquica de dois níveis, com N_j unidades de nível 1 aglomeradas na j-ésima unidade de nível 2, onde $j = 1,...,M$. Suponha que y_{ij}, o valor da variável resposta associado à i-ésima unidade de nível 1 da j-ésima unidade de nível 2, seja gerado pelo modelo linear hierárquico de dois níveis

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u} + z_{0ij}\varepsilon_{ij}; \qquad \mathbf{u} \sim N(\mathbf{0}, \Omega); \qquad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{2.1}$$

onde

$\mathbf{x}_{ij} = [x_{ij1} \quad \ldots \quad x_{ijp}]$ é um vetor linha de variáveis explicativas associado aos efeitos fixos, de dimensão p;

$\boldsymbol{\beta}$ é o vetor de efeitos fixos de dimensão p;

$\mathbf{z}_{ij} = \lfloor z_{ij1} \quad \ldots \quad z_{ijq} \rfloor$ é um vetor linha de variáveis explicativas associado aos efeitos aleatórios de nível 2, de dimensão q;

$\mathbf{u}$ é o vetor de efeitos aleatórios de nível 2 de dimensão q;

z_{0ij} é um escalar.

ε_{ij} é o efeito aleatório associado à i-ésima unidade de nível 1 da j-ésima unidade de nível 2;

$\mathbf{0}$ é um vetor de zeros de dimensão q;

$\Omega = VAR(\mathbf{u})$ é a matriz de variâncias-covariâncias dos efeitos aleatórios de nível 2, de dimensão $q \times q$;

σ^2 é a variância do efeito aleatório de nível 1.

Na maioria das aplicações, inclusive neste trabalho, z_{0ij} é igual a 1 para todo i e j. Porém, a utilização de z_{0ij} desiguais permite a representação de padrões conhecidos de heteroscedasticidade dentro dos conglomerados.

No procedimento descrito na próxima seção, é suposto um plano amostral de dois estágios. No primeiro estágio, são selecionadas m unidades de nível 2 com probabilidades de inclusão na amostra denotadas por π_j ($j = 1,\ldots,M$). Em cada unidade de nível 2 pertencente à amostra, são selecionadas n_j unidades de nível 1 com probabilidades de inclusão na amostra denotadas por $\pi_{i|j}$. Portanto, a probabilidade não condicional da i-ésima unidade de nível 1, pertencente à j-ésima unidade de nível 2, ser incluída na amostra é $\pi_{ij} = \pi_{i|j}\pi_j$.

2.2 Estimação

Para aplicação imediata do método de estimação de máxima pseudo-verossimilhança, seria necessário escrever uma expressão para a função de verossimilhança das unidades populacionais, estimar o logaritmo natural desta função e maximizá-lo numericamente. Contudo, por questões de eficiência computacional e simplicidade de estimação, Pfeffermann et al. (1998) adotaram uma adaptação do método de estimação dos Mínimos Quadrados Generalizados Iterativo -MQGI-, por analogia ao método de Máxima Pseudo-Verossimilhança.

O método de MQGI alterna, iterativamente, entre a estimação dos efeitos fixos ($\boldsymbol{\beta}$) e a estimação das variâncias e covariâncias dos efeitos aleatórios (elementos distintos de Ω e σ^2). Na adaptação deste método, escrevem-se expressões para os estimadores de $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$ usados no algoritmo de MQGI para o caso onde são observados os valores de todas as unidades da população, sendo $\boldsymbol{\theta}$ um vetor coluna formado pelos elementos distintos de Ω e σ^2. Na literatura, esses estimadores são referidos por "estimadores do censo", por analogia a

uma operação censitária onde todas as unidades da população são investigadas. Esses "estimadores do censo" são, então, substituídos por estimadores amostrais ponderados.

Considere o algoritmo de MQGI para o caso hipotético onde os valores y_{ij}, $\mathbf{x}_{ij}$, $\mathbf{z}_{ij}$ e z_{0ij} são observados para todas as unidades da população, isto é, $i = 1,...,N_j$ e $j = 1,...,M$. Sejam

$$\mathbf{Y}_j = \begin{bmatrix} y_{1j} \\ \vdots \\ y_{N_j j} \end{bmatrix}, \quad X_j = \begin{bmatrix} \mathbf{x}_{1j} \\ \vdots \\ \mathbf{x}_{N_j j} \end{bmatrix} \quad \text{e} \quad \mathbf{r}_j = \begin{bmatrix} r_{1j} \\ \vdots \\ r_{N_j j} \end{bmatrix}, \quad \text{onde} \quad r_{ij} = \mathbf{z}_{ij}\mathbf{u} + z_{0ij}\varepsilon_{ij}.$$

Então, o modelo definido na equação (2.1) pode ser expresso, em forma matricial, como:

$$\mathbf{Y}_j = X_j\boldsymbol{\beta} + \mathbf{r}_j, \qquad \mathbf{r}_j \sim N(\mathbf{0}, V_j) \tag{2.2}$$

onde $V_j = Z_j \Omega Z_j^t + \sigma^2 D_j$, $Z_j = \begin{bmatrix} \mathbf{z}_{1j} \\ \vdots \\ \mathbf{z}_{N_j j} \end{bmatrix}$ e $D_j = diag\{z_{0ij}^2,...,z_{0N_j j}^2\}$.

Seja $\boldsymbol{\theta}^t = [\theta_1 \quad ... \quad \theta_b]$ o transposto do vetor $\boldsymbol{\theta}$ de dimensão b, sendo $\theta_1,...,\theta_{b-1}$ os elementos distintos de Ω e $\theta_b = \sigma^2$, onde $b = \frac{q(q+1)}{2} + 1$ (q variâncias de efeitos aleatórios de nível 2, $\frac{q(q-1)}{2}$ covariâncias destes efeitos aleatórios e a variância do efeito aleatório de nível 1). Expressando-se, então, V_j como uma função linear de $\boldsymbol{\theta}$, tem-se:

$$V_j(\boldsymbol{\theta}) = \sum_{k=1}^{b} \theta_k G_{kj}$$

onde $G_{kj} = Z_j H_{kj} Z_j^t + \delta_{kb} D_j$, H_{kj} é uma matriz conhecida de dimensão $q \times q$, cujos elementos assumem valores 0 ou 1, e $\delta_{kb} = \begin{cases} 0, & \text{se k} \neq \text{b} \\ 1, & \text{se k} = \text{b} \end{cases}$.

Escrevendo $\mathbf{r}_j(\boldsymbol{\beta}) = (\mathbf{Y}_j - X_j\boldsymbol{\beta})(\mathbf{Y}_j - X_j\boldsymbol{\beta})^t$, nota-se que este é o vetor de resíduos globais para a j-ésima unidade de nível 2, com dimensão N_j, como especificado na equação (2.2). o algoritmo de MQGI envolve o cálculo de uma seqüência de "estimadores do censo" de $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$, denotados por $\boldsymbol{\beta}_C^{(r)}$ e $\boldsymbol{\theta}_C^{(r)}$, $r = 1, 2,...$, como segue:

Estágio 1: Faça $\boldsymbol{\beta}_C^{(r)} = P^{(r)^{-1}} Q^{(r)}$ (2.3)

onde $P^{(r)} = \sum_j X_j^t V_{jr}^{-1} X_j$, $Q^{(r)} = \sum_j X_j^t V_{jr}^{-1} \mathbf{Y}_j$, $V_{jr} = V_j\left(\boldsymbol{\theta}_C^{(r-1)}\right)$ e $\sum_j$ denota a soma para todo $j = 1,...,M$.

Estágio 2: Faça $\boldsymbol{\theta}_C^{(r)} = R^{(r)^{-1}} S^{(r)}$ (2.4)

onde o kl -ésimo elemento da matriz $R^{(r)}$ de dimensões $b \times b$ é $\sum_j tr\left(V_{jr}^{-1} G_{kj} V_{jr}^{-1} G_{lj}\right)$ e o k –ésimo elemento do vetor $S^{(r)}$ de dimensão b é $\sum_j tr\left\{V_{jr}^{-1} G_{kj} V_{jr}^{-1} \mathbf{r}_j\left(\boldsymbol{\beta}_C^{(r)}\right)\right\}$.

O processo seria iniciado em um valor arbitrário $\boldsymbol{\theta}_C^{(0)}$, como por exemplo o "estimador de Mínimos Quadrados Ordinários do censo" no caso de σ^2 e ω^2 igual a zero, com $\boldsymbol{\beta}_C^{(r)}$ e $\boldsymbol{\theta}_C^{(r)}$ convergindo para os "estimadores de MQGI do censo", denotados por $\boldsymbol{\beta}_C$ e $\boldsymbol{\theta}_C$, quando r tende a infinito (Pfeffermann, 1993).

Os "estimadores do censo" são funções dos valores da população e, portanto, suas estimativas não podem ser calculadas se amostragem é aplicada para obtenção dos dados da análise. Esses "estimadores do censo" ou quantidades descritivas populacionais correspondentes são, então, substituídos por estimadores baseados apenas nos dados amostrais ($\hat{P}^{(r)}$, $\hat{Q}^{(r)}$, $\hat{R}^{(r)}$ e $\hat{S}^{(r)}$), isto é, substituindo-se $j = 1,...,M$ por $j = 1,...,m$ nas expressões (2.3) e (2.4) sem considerar os pesos amostrais. Neste caso, $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$ são estimados por $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\theta}}$, sendo

$$\hat{\boldsymbol{\beta}} = \lim_{r \to \infty} \hat{\boldsymbol{\beta}}^{(r)} \quad \text{e} \quad \hat{\boldsymbol{\theta}} = \lim_{r \to \infty} \hat{\boldsymbol{\theta}}^{(r)}$$

onde

$$\hat{\boldsymbol{\beta}}^{(r)} = \hat{P}^{(r)^{-1}} \hat{Q}^{(r)} \quad \text{e} \quad \hat{\boldsymbol{\theta}}^{(r)} = \hat{R}^{(r)^{-1}} \hat{S}^{(r)} \tag{2.5}$$

Se $\hat{P}^{(r)}$, $\hat{Q}^{(r)}$, $\hat{R}^{(r)}$ e $\hat{S}^{(r)}$ são versões amostrais de $P^{(r)}$, $Q^{(r)}$, $R^{(r)}$ e $S^{(r)}$, ainda não corrigidas pelas probabilidade de seleção, então $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\theta}}$ são os *estimadores usuais de MQGI não-ponderados*. Como esses estimadores não incorporam o plano amostral, estes precisam ser adaptados para estimar consistentemente as quantidades da população finita $P^{(r)}$, $Q^{(r)}$, $R^{(r)}$ e $S^{(r)}$.

O procedimento de *ponderação pelas probabilidades* proposto por Pfeffermann et al. (1998) e aplicado em Corrêa (2001) consiste em expressar $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\theta}}$ como funções de somas em i e j, isoladamente, onde i representa as unidades de nível 1 e j as unidades de nível 2. A idéia, então, é substituir cada soma em j por uma soma das unidades da amostra ponderada por $w_j = \pi_j^{-1}$, e cada soma em i por uma soma das unidades amostrais ponderada por $w_{i|j} = \pi_{i|j}^{-1}$. Os estimadores resultantes são referidos por *estimadores de Mínimos Quadrados Generalizados Iterativo Ponderados pelas Probabilidades* -MQGIPP- (Pfeffermann et al., 1998) e dados por:

$$\hat{P}^{(r)} = \sum_j w_j \left(\hat{T}_{1j} - \hat{a}_j \hat{T}_{2j} \hat{T}_{2j}^t \right) \quad \text{e} \quad \hat{Q}^{(r)} = \sum_j w_j \left(\hat{T}_{3j} - \hat{a}_j \hat{T}_{2j} \hat{T}_{4j} \right)$$

onde $\hat{T}_{1j} = \sum_i w_{i|j} \frac{\mathbf{x}_{ij}^t \mathbf{x}_{ij}}{z_{0ij}^2}$, $\hat{T}_{2j} = \sum_i w_{i|j} \frac{\mathbf{x}_{ij}^t \mathbf{z}_{ij}}{z_{0ij}^2}$, $\hat{T}_{3j} = \sum_i w_{i|j} \frac{\mathbf{x}_{ij}^t y_{ij}}{z_{0ij}^2}$, $\hat{T}_{4j} = \sum_i w_{i|j} \frac{y_{ij} \mathbf{z}_{ij}}{z_{0ij}^2}$,

$$\hat{a}_j = \left(\hat{T}_{5j} + \frac{\hat{\sigma}^2}{\hat{\omega}_0^2} \right)^{-1} \text{ e } \hat{T}_{5j} = \sum_i w_{i|j} \frac{\mathbf{z}_{ij}^2}{z_{0ij}^2}.$$

e

$$\hat{R}^{(r)} = \begin{bmatrix} \sum_j w_j \hat{b}_j^2 & \sum_j w_j \frac{\hat{b}_j^2}{\hat{T}_{5j}} \\ \sum_j w_j \frac{\hat{b}_j^2}{\hat{T}_{5j}} & \sum_j w_j \left\{ \hat{\sigma}^{-4} \left(\hat{N}_j - 1 \right) + \frac{\hat{b}_j^2}{\hat{T}_{5j}^2} \right\} \end{bmatrix} \quad \text{e} \quad \hat{S}^{(r)} = \begin{bmatrix} \sum_j w_j \hat{b}_j^2 \hat{\mathbf{u}}^2 \\ \sum_j w_j \left(\hat{\sigma}^{-4} \hat{T}_{6j} + \frac{\hat{b}_j^2 \hat{\mathbf{u}}^2}{\hat{T}_{5j}} \right) \end{bmatrix}$$

onde $\hat{b}_j = \left(\hat{\omega}_0^2 + \frac{\hat{\sigma}^2}{\hat{T}_{5j}} \right)^{-1}$, $\hat{T}_{6j} = \sum_i w_{i|j} \hat{\varepsilon}_{ij}^2$, $\hat{\mathbf{u}} = \frac{\left(\sum_i w_{i|j} \hat{r}_{ij} \mathbf{z}_{ij} / z_{0ij}^2 \right)}{\hat{T}_{5j}}$, $\hat{N}_j = \sum_i^s w_{i|j}$, $\hat{\varepsilon}_{ij} = \frac{\left(\hat{r}_{ij} - \mathbf{z}_{ij} \hat{\mathbf{u}} \right)}{z_{0ij}}$ e

$\hat{r}_{ij} = y_{ij} - \mathbf{x}_{ij} \hat{\boldsymbol{\beta}}^{(r)}$.

A fim de minimizar vícios de pequenas amostras na estimação dos parâmetros $\boldsymbol{\beta}$ e θ, Pfeffermann et al. (1998) sugerem padronizar os pesos amostrais. Como $\hat{\boldsymbol{\beta}}$ e $\hat{\theta}$ são invariantes por mudança de escala em w_j, a padronização foi aplicada apenas a $w_{i|j}$ de forma que a soma destes pesos fosse igual a n_j. Então, $w_{i|j}$ foi substituído por $w_{i|j}^* = \lambda_j w_{i|j}$ nas expressões dos estimadores de MQGIPP utilizados na implementação do método (seção 4), com $\lambda_j = \left(\frac{\sum_i^s w_{i|j}}{n_j} \right)^{-1}$ (Corrêa, 2001, seção 3.4).

Com base em estudos de simulação, envolvendo um modelo hierárquico de dois níveis de componentes de variância sem variáveis explicativas, Pfeffermann et al. (1998) concluíram que os estimadores de MQGI e de MQGIPP têm comportamentos similares no caso de planos amostrais ignoráveis[1], com estimativas praticamente iguais para todos os parâmetros do modelo testado, apresentando vício não desprezível para os estimadores de ω^2. Os autores afirmam, entretanto, que há pouca desvantagem, em termos de vício ou precisão, na utilização de estimadores de MQGIPP padronizados no caso de planos amostrais ignoráveis.Para planos amostrais informativos[2] em algum nível do modelo hierárquico, os autores concluíram que os estimadores de MQGIPP

[1] Um plano amostral é dito *ignorável* quando o modelo probabilístico da variável resposta (ou seja, a distribuição de probabilidade da variável de interesse) na amostra é o mesmo que o da população (Pfeffermann, 1993).
[2] Um plano amostral é dito *informativo* quando o esquema de seleção da amostra depende também da variável de interesse (Pfeffermann, 1993).

não-padronizados têm menor vício que os estimadores de MQGI quando os tamanhos amostrais n_j são grandes. Porém, $\hat{\theta}$ permanece com vício não desprezível no caso de tamanhos n_j pequenos. Já os estimadores de MQGIPP padronizados têm menor vício que os estimadores de MQGI e que os estimadores de MQGIPP não-padronizados no caso de tamanhos n_j pequenos, embora ainda permaneçam com vício não desprezível. Ao contrário dos métodos de estimação disponíveis nos pacotes computacionais convencionais, o procedimento de estimação das variâncias dos estimadores de MQGIPP não considera apenas a distribuição induzida pelo modelo, mas sim a distribuição combinada: do modelo (ξ) e do plano amostral (P). A variância do estimador de MQGIPP $\hat{\boldsymbol{\beta}}$ sob esta distribuição combinada pode ser decomposta como (idem para $\hat{\theta}$):

$$VAR_{P\xi}(\hat{\boldsymbol{\beta}}) = E_{\xi}[VAR_P(\hat{\boldsymbol{\beta}})] + VAR_{\xi}[E_P(\hat{\boldsymbol{\beta}})] = E_{\xi}[VAR_P(\hat{\boldsymbol{\beta}})] + O(N^{-1}) \tag{2.6}$$

isto é, para o caso usual onde a população é muito maior que a amostra, a variância de $\hat{\boldsymbol{\beta}}$ sob a distribuição conjunta $P\xi$ é aproximadamente o valor esperado, sob o modelo, da variância de $\hat{\boldsymbol{\beta}}$ sob a distribuição induzida pelo plano amostral.

Para frações amostrais suficientemente pequenas em ambos os níveis da hierarquia, o lado direito da equação (2.6) pode ser estimado consistentemente através da estimativa da variância de $\hat{\boldsymbol{\beta}}$ sob a distribuição de aleatorização.

Esta estimativa pode ser obtida através do método de Conglomerado Primário (do inglês "Ultimate Cluster") de estimação de variâncias para estimadores de totais e médias em planos amostrais complexos (Hansen, Hurwitz e Madow, 1953). A idéia deste procedimento é considerar apenas a variação entre as unidades de nível 2 (conglomerados) no cálculo dos estimadores de variância, admitindo que elas teriam sido selecionadas por amostragem com reposição. Maiores detalhes sobre os estimadores de variância podem ser encontrados em Corrêa (2001).

A aplicação da metodologia de incorporação do plano amostral na estimação de modelos lineares hierárquicos é estudada na próxima seção, onde utilizam-se os estimadores de MQGIPP padronizados para ajustar um modelo hierárquico ao Índice de Massa Corporal, relacionando-o com variáveis pertinentes ao estudo de interesse. Esta análise foi desenvolvida com base nos dados da Pesquisa sobre Padrões de Vida do IBGE 1996-1997.

3. O Índice de massa corporal como um indicador do estado nutricional de adultos e a pesquisa sobre padrões de vida do IBGE

O Índice de Massa Corporal -IMC-, definido pela razão peso/altura^2, com o peso medido em quilograma e a altura medida em metro, foi escolhido como foco principal da análise desenvolvida neste trabalho por sua comprovada relevância na avaliação do estado nutricional de adultos em estudos epidemiológicos, uma vez que

este índice apresenta alta correlação com o peso corporal, baixa correlação com estatura e, principalmente, simplicidade e ampla aplicabilidade.

Uma motivação adicional, porém não menos importante, para estudar este índice é que análises envolvendo o Índice de Massa Corporal como a apresentada neste trabalho são raras e não triviais, pois, nesta aplicação, considera-se uma base de dados relativamente atual e técnicas estatísticas modernas, onde são enfatizadas duas importantes características dos dados: o plano amostral utilizado na obtenção dos mesmos e a estrutura hierárquica da população de onde foram extraídos.

O Índice de Massa Corporal já vinha sendo utilizado na avaliação da obesidade em adultos, quando James et al. (1988) propuseram sua utilização como indicador de Deficiência Energética Crônica -DEC- em adultos. Valores de IMC inferiores a 16 indicariam DEC de terceiro grau, valores entre 16 e 18,4 poderiam indicar DEC de graus 1 e 2, dependendo do nível de atividade física do indivíduo e valores acima de 18,4 indicariam normalidade. Mais tarde, Ferro-Luzzi et al. (1992) sugeriram uma simplificação na avaliação da DEC, baseando-a apenas no IMC. A Tabela 3.1 resume essas informações, incluindo os limites adotados pela Organização Mundial da Saúde para normalidade, sobrepeso e obesidade.

Tabela 3.1 - Limites para o Índice de Massa Corporal

IMC	<16	16-16,9	17-18,4	18,5-25	25,1-29,9	>30
Classificação	DEC III	DEC II	DEC I	normal	pesado	obeso

Fontes: Ferro-Luzzi *et al.* (1992) e WORLD HEALTH ORGANIZATION. World Health Report 1998 [online]. Disponível: http://www.who.int/whr/1998/fig18e.jpg [capturado em 4 out. 2000].

A Pesquisa sobre Padrões de Vida do IBGE - PPV/IBGE - foi escolhida para este estudo por sua abrangência temática, o que permite realizar análises complexas das interações entre os vários aspectos do bem-estar social, e pela hierarquia existente na estrutura dos dados (seção 4).

A PPV/IBGE tem um plano amostral conglomerado em dois estágios. A Unidade Primária de Amostragem -UPA- é o setor censitário da base geográfica do Censo Demográfico 1991 e a Unidade Secundária de Amostragem -USA- é o domicílio. Os setores censitários foram estratificados por divisão geográfica e de acordo com a média da renda mensal dos chefes dos domicílios de cada setor. A seleção dos setores em cada estrato geográfico foi realizada com probabilidade proporcional ao número de domicílios particulares permanentes ocupados no estrato, e a seleção de domicílios em cada setor foi por amostragem aleatória simples. A amostra é composta por 4 940 domicílios e 554 setores.

Uma etapa importante que antecede a modelagem multinível do Índice de Massa Corporal desenvolvida na próxima seção é a aplicação de tratamentos de não-resposta aos dados da PPV/IBGE, devido a ocorrência de um número significativo de não-respondentes em duas variáveis fundamentais para o estudo de interesse: o Índice de Massa Corporal e a renda domiciliar per capita. Estes tratamentos baseiam-se na reponderação dos pesos amostrais dos respondentes segundo um modelo de regressão logística que incorpora o plano amostral, onde a variável resposta é a probabilidade de não responder à variável de interesse (detalhes em Corrêa, 2001).

4. Índice de massa corporal e sua relação com variáveis da pesquisa sobre padrões de vida do IBGE

O índice de massa corporal, por ser um indicador do estado nutricional de adultos, fornece uma medida do bem-estar e do nível socioeconômico do indivíduo. A relação do índice de massa corporal com variáveis nutricionais e socioeconômicas da PPV/IBGE é estudada nesta seção, visando a identificar fatores que estejam associados à desnutrição e ao sobrepeso de adultos das Regiões Nordeste e Sudeste do Brasil.

É razoável imaginar que moradores de um mesmo domicílio tenham padrões de vida semelhantes, com condições similares de acesso à educação, à boa alimentação e à prática de atividades físicas. A relação entre o IMC e esses fatores deve, portanto, ser afetada pelo efeito de conglomeração de pessoas em domicílios, já que este índice é influenciado pelos elementos acima citados.

Para avaliar a existência e o grau de associação entre o Índice de Massa Corporal e as demais variáveis pertinentes ao estudo, foram ajustados modelos lineares hierárquicos com dois níveis, sendo pessoa o nível 1 e domicílio o nível 2, com efeito aleatório de nível 2 (domicílio) apenas no intercepto do modelo. Isto significa que apenas o IMC médio das pessoas (intercepto) varia entre domicílios, mas não a intensidade com que cada variável preditora influencia o índice. Não foram testados modelos com mais de um efeito aleatório de nível 2 por envolver extensões não triviais da metodologia descrita em Pfeffermann (1998).

Corrêa (2001) desenvolveu programas utilizando a linguagem matricial do SAS. PROC IML (SAS Institute INC., 1990) para realizar a análise de interesse, implementando o procedimento descrito na seção 2 que incorpora o plano amostral na estimação dos parâmetros do modelo hierárquico e de seus respectivos desvios padrões. As expressões dos estimadores de MQGIPP e de suas respectivas variâncias foram implementadas considerando-se os pesos condicionais das pessoas padronizados, como citado na seção 2.

A variável resposta do modelo considerado é o logaritmo do Índice de Massa Corporal de moradores com 20 anos ou mais de idade, uma vez que a transformação logarítmica aproxima a distribuição do IMC da distribuição normal. As variáveis preditoras disponíveis para o estudo de interesse são:

Nível 2 (domicílio) - quintos de renda domiciliar mensal per capita (QRD), logaritmo da despesa domiciliar per capita com alimentação diária (LDE), consumo calórico domiciliar per capita diário (LCA), área de localização do domicílio - urbana ou rural - (ARE), região do domicílio - Nordeste ou Sudeste - (REG), número de moradores do domicílio (NUM), anos de estudo do chefe do domicílio (AEC) e nível de adequação do domicílio para habitação (HAB).

Nível 1 (pessoa) - sexo (SEX), faixa etária (FXE), raça (RAC), renda individual mensal (REN), anos de estudo (AES), condição de ocupação (OCU) e condição no domicílio (CON).

As variáveis de nível 1 que apresentaram efeitos significativos no modelo foram: faixa etária (de 20 a 29 anos; de 30 a 39 anos; de 40 a 49 anos e 50 anos ou mais), sexo (masculino; feminino), anos de estudo (menos de 1 ano; de 1 a 7 anos; de 8 a 10 anos; 11 anos e 12 anos ou mais) e raça (não-branca; branca). Para explicar parte da variabilidade entre domicílios, foram incluídas no modelo as seguintes variáveis de nível 2: área de localização do domicílio (urbana; rural), logaritmo da despesa domiciliar per capita diária com alimentação e quintos de renda domiciliar mensal per capita.

A expressão do modelo escolhido está explicitada na equação (4.1) e as estimativas dos parâmetros e seus respectivos desvios padrões estão apresentados na Tabela (4.1), em que o símbolo (*) indica que o efeito é significativo a um nível de significância de 5%. Na notação adotada para representar o modelo, as letras maiúsculas abreviam o nome da variável, os sobrescritos denotam os efeitos dos níveis de cada fator e os subscritos ij ou j indicam se a variável é definida no nível de pessoa ou domicílio, respectivamente. As variáveis sem sobrescrito são contínuas.

$$\begin{cases} \ln(IMC)_{ij} = \beta_{0j} + FXE_{ij}^{r} + SEX_{ij}^{s} + AES_{ij}^{t} + RAC_{ij}^{u} + FXE_{ij}^{r} * SEX_{ij}^{s} + SEX_{ij}^{s} * RAC_{ij}^{u} + \varepsilon_{ij} \\ \beta_{0j} = \gamma_{00} + \gamma_{01} LDE_{j} + QRD_{j}^{v} + ARE_{j}^{z} + u_{0j} \end{cases} \tag{4.1}$$

Observando a relação entre a estimativa de ω^2 e seu desvio padrão estimado (Tabela 4.1), e considerando a normalidade assintótica de ω^2, conclui-se que o efeito de domicílio é significativo. Isto é, a relação entre o IMC e as demais variáveis envolvidas no estudo é afetada pela conglomeração de pessoas em domicílios.

O Índice de Massa Corporal médio estimado quando todas as variáveis estão em sua categoria base (primeira categoria de cada variável) é igual a 22,14 (valor obtido tomando-se o exponencial do intercepto), o que indica que a população adulta das Regiões Nordeste e Sudeste encontra-se na faixa normal de peso. Apesar deste estimador ser viciado, seu vício é desprezível considerando o tamanho da amostra em estudo.

Observa-se, ainda, que o IMC médio estimado é maior para pessoas moradoras em domicílios localizados em área urbana, com maior despesa domiciliar per capita diária com alimentação e maior renda domiciliar mensal per capita. Adicionalmente, as mulheres da raça não-branca têm, em média, maior IMC que as mulheres da raça branca. Já para o sexo masculino, o comportamento encontrado é inverso (Figura 4.1).

Conclui-se, ainda, que, em média, o IMC do indivíduo reduz com o aumento do seu nível de instrução e cresce de acordo com a idade até a faixa de 40 a 49 anos, a partir da qual o IMC dos homens apresenta uma redução significativa e o das mulheres apresenta um leve acréscimo. Verifica-se, também, que em geral, as mulheres têm níveis de IMC iguais ou maiores que os dos homens para o caso da raça não-branca, enquanto para a raça branca, o comportamento é oposto (Figura 4.2).

Tabela 4.1 - Estimativas dos parâmetros e de seus respectivos desvios padrões para o modelo hierárquico especificado na equação (4.1), considerando o método de estimação MQGIPP

Parâmetro	Estimativa	Desvio padrão
γ_{00}	3,0974(*)	0,0097
γ_{01}	0,0181(*)	0,0049
Faixa etária		
De 20 a 29 anos (FX1)	0,0000	.
De 30 a 39 anos (FX2)	1,0521(*)	0,0074
De 40 a 49 anos (FX3)	1,0614(*)	0,0087
50 anos ou mais (FX4)	0,0281(*)	0,0076
Sexo		
Masculino (S1)	0,0000	.
Feminino (S2)	-0,0025	0,0084
Anos de estudo		
Menos de 1 ano	0,0000	.
De 1 a 7 anos	0,0192(*)	0,0068
De 8 a 10 anos	0,0033	0,0092
11 anos	-0,0140	0,0092
12 anos ou mais	-0,0166	0,0111
Raça		
Não branca (R1)	0,0000	.
Branca (R2)	0,0224(*)	0,0062
Faixa etária * Sexo		
FX2.S2	0,0022	0,0111
FX3.S2	0,0395(*)	0,0127
FX4.S2	0,0787(*)	0,0110
Sexo * Raça		
S2.R2	-0,0392(*)	0,0081
Quintos de renda domiciliar mensal per capita		
1° quinto	0,0000	.
2° quinto	0,0168(*)	0,0089
3° quinto	0,0281(*)	0,0094
4° quinto	0,0388(*)	0,0104
5° quinto	0,0361(*)	0,0113
Área de localização do domicílio		
Urbana	0,0000	.
Rural	-0,0350(*)	0,0060
Variâncias dos erros aleatórios		
$var(u_{0j}) = \omega^2$	0,0046(*)	0,0005
$var(\varepsilon_{ij}) = \sigma^2$	0,0219(*)	0,0007

(*) efeito significativamente diferente de zero a um nível de confiança de 5%.

Figura 4.1 - Gráfico da interação entre as variáveis sexo e raça para o modelo do Índice de Massa Corporal -IMC- explicitado na equação (4.1)

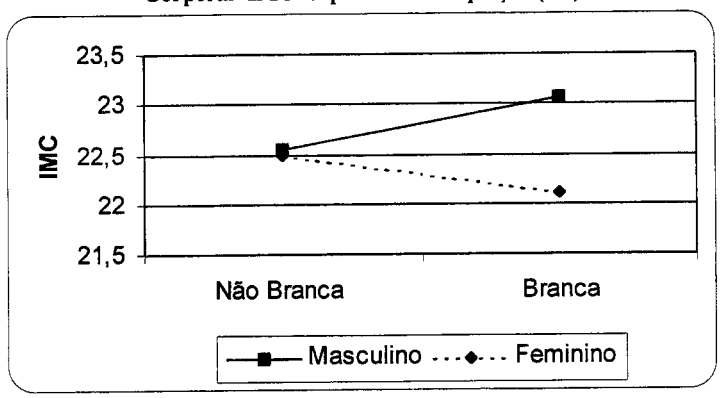
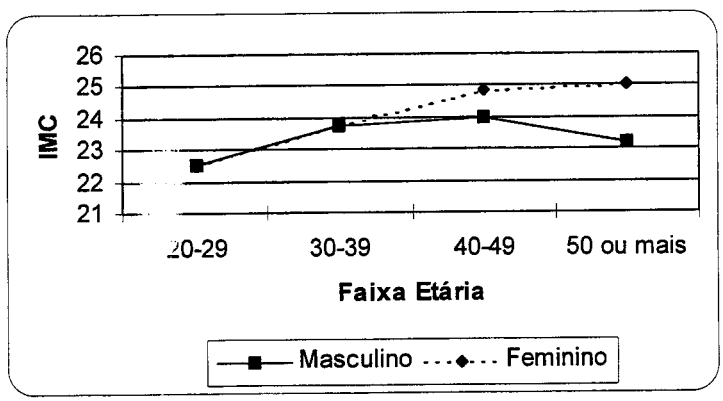
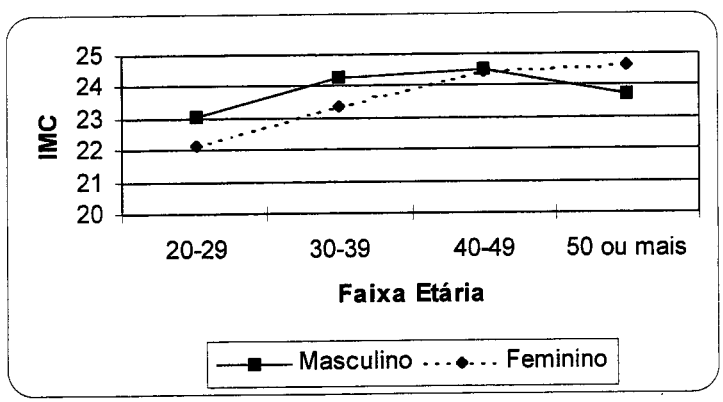


Figura 4.2 - Gráficos da interação entre as variáveis sexo e faixa etária, segundo a variável raça para o modelo do Índice de Massa Corporal -IMC- explicitado na equação (4.1)

RAÇA NÃO-BRANCA



RAÇA BRANCA



O fato de que renda domiciliar e anos de estudo influenciam o IMC em direções diferentes parece contraditório, uma vez que renda e nível de instrução, em geral, têm uma associação positiva. Uma possível explicação para tal comportamento é que os quintos de renda foram construídos com base nos valores de renda dos indivíduos de toda a amostra (Nordeste e Sudeste). Entretanto, as distribuições de renda destas duas Regiões são diferentes, indicando que, para a Região Nordeste, os indivíduos classificados nos níveis 1 e 2 da variável anos de estudo (menos de 1 ano de estudo e de 1 a 7 anos de estudo, respectivamente) estão concentrados nos quintos de renda 1 e 2. Já para a Região Sudeste, os indivíduos classificados naqueles mesmos níveis de instrução estão concentrados nos quintos de renda 3, 4 e 5.

Logo, para a base de dados aqui considerada, a relação positiva esperada entre renda e anos de estudo não se aplica no Sudeste, influenciando, assim, a distribuição de renda para as duas Regiões em conjunto (Nordeste e Sudeste) e, conseqüentemente, as estimativas dos coeficientes do modelo estudado.

As Tabelas 4.2 (a), 4.2 (b) e 4.2 (c) apresentam as distribuições de renda para as duas regiões em conjunto e para cada uma delas isoladamente. Os cálculos foram efetuados na PROC CROSSTAB do SUDAAN (Shah et al., 1992) e resumem o que foi discutido acima.

Tabela 4.2 - Distribuição dos indivíduos da Pesquisa sobre Padrões de Vida 1996-1997, segundo anos de estudo e quintos de renda domiciliar mensal per capita

(a) Regiões Nordeste e Sudeste

Anos de estudo do indivíduo	Quintos de renda domiciliar mensal per capita					Total
	1	2	3	4	5	(%)
Menos de 1 ano	**32,89**	**26,90**	**23,40**	**12,43**	**4,46**	100,00
1 a 7 anos	**15,07**	**20,49**	**24,80**	**23,29**	**16,36**	100,00
8 a 11 anos	5,41	15,19	19,70	31,89	27,80	100,00
11 anos	1,47	5,71	14,68	31,37	46,78	100,00
12 anos ou mais	0,00	0,18	1,73	14,41	83,68	100,00
Total (%)	14,32	17,36	20,49	22,41	25,43	100,00

(b) Região Nordeste

Anos de estudo do indivíduo	Quintos de renda domiciliar mensal per capita					Total
	1	2	3	4	5	(%)
Menos de 1 ano	**48,06**	**27,26**	18,64	5,03	1,01	100,00
1 a 7 anos	**31,41**	**30,17**	20,26	12,46	5,69	100,00
8 a 11 anos	10,63	28,16	24,50	24,72	11,99	100,00
11 anos	2,80	10,90	22,45	32,55	31,30	100,00
12 anos ou mais	0,00	0,12	4,60	22,19	73,10	100,00
Total (%)	30,25	25,18	19,58	13,92	11,07	100,00

(c) Região Sudeste

Anos de estudo do indivíduo	Quintos de renda domiciliar mensal per capita					Total
	1	2	3	4	5	(%)
Menos de 1 ano	9,31	26,35	**30,82**	**23,69**	**9,83**	100,00
1 a 7 anos	7,80	16,18	**26,82**	**28,09**	**21,10**	100,00
8 a 11 anos	3,54	10,53	17,98	34,46	53,49	100,00
11 anos	0,83	3,22	10,96	30,81	54,19	100,00
12 anos ou mais	0,00	0,19	1,04	12,52	86,26	100,00
Total (%)	5,55	13,05	20,99	27,08	33,32	100,00

Os gráficos de percentis da distribuição normal para os resíduos padronizados de nível 2 e de nível 1 do modelo (4.1) indicam que os erros podem ser considerados aproximadamente normalmente distribuídos no centro da distribuição, apresentando desvio de normalidade em ambas as caudas no caso dos resíduos de nível 2, e apenas na cauda direita no caso dos resíduos de nível 1, como indicado pelas Figuras 4.3 e 4.4, respectivamente.

Figura 4.3 -Gráfico de percentis da distribuição normal para os resíduos de nível 2 padronizados do modelo explicitado na equação (4.1)

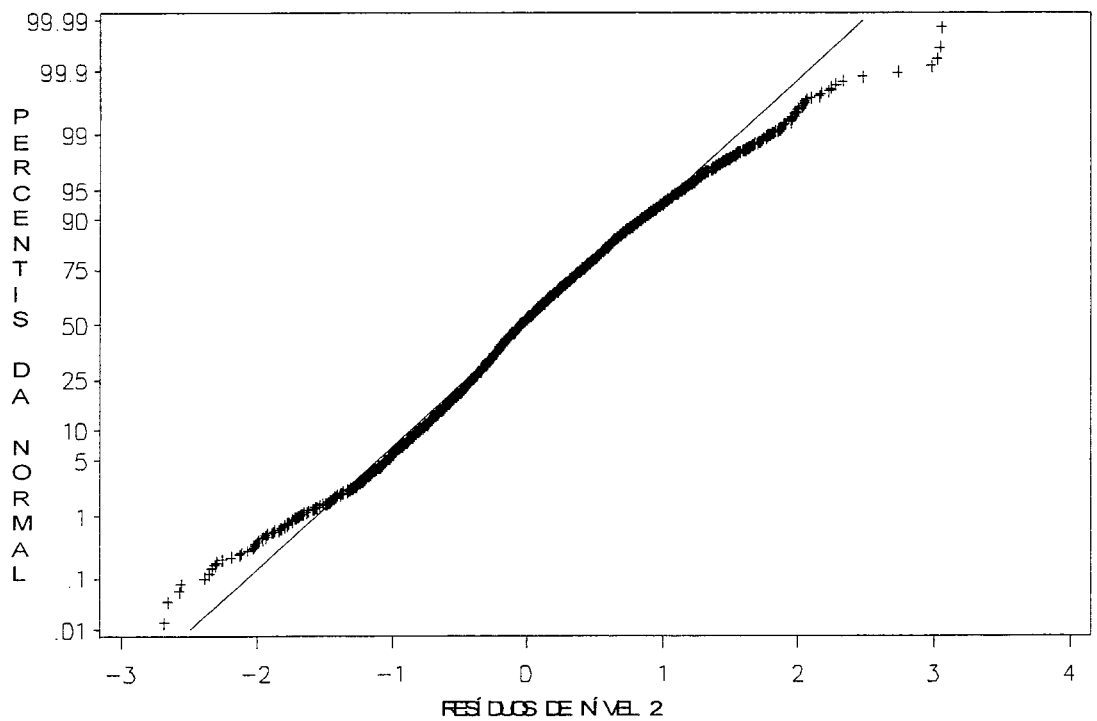
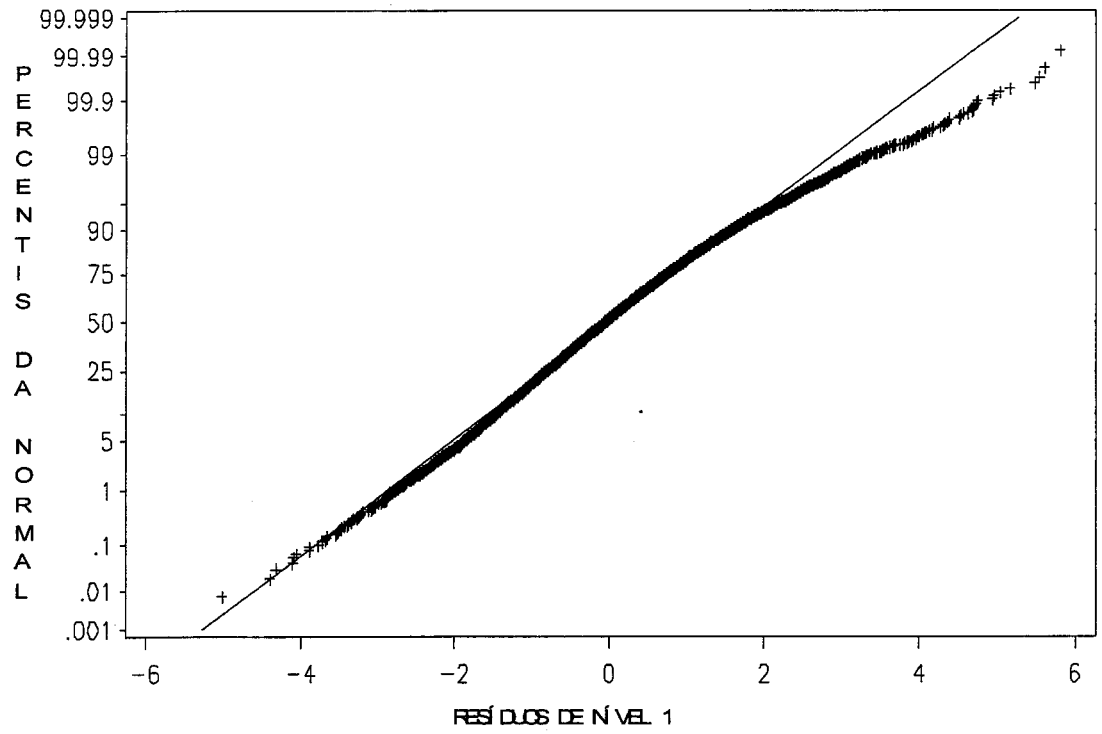


Figura 4.4 - Gráfico de percentis da distribuição normal para os resíduos de nível 1 padronizados do modelo explicitado na equação (4.1)



5. Método dos mínimos quadrados generalizados iterativo ponderados pelas probabilidades *versus* utilização de pacotes estatísticos convencionais

Diante da necessidade de se implementar o procedimento de Mínimos Quadrados Generalizados Iterativo Ponderado pelas Probabilidades, uma vez que este não se encontra disponível em pacotes computacionais, é importante avaliar como o modelo analisado na seção anterior se comporta quando ajustado em um pacote convencional que não incorpora o plano amostral de maneira apropriada. O pacote escolhido para efeito de comparação foi o MLWIN 1.10 (Rasbash et al., 1999).

O modelo analisado na seção anterior foi ajustado no MLWIN 1.10 de três diferentes maneiras. A primeira delas é o ajuste do modelo utilizando o comando WEIGHTS com a opção "pesos padronizados" (MPP), que permite que os pesos amostrais w_j e $w_{i|j}$ sejam incorporados na análise. A segunda maneira é a utilização do comando WEIGHTS com a opção "pesos não-padronizados" (MPNP). A terceira forma corresponde ao ajuste do modelo sem pesos amostrais (MSP). Para cada uma das maneiras, foi calculada a diferença relativa entre a estimativa obtida no MLWIN 1.10 e a obtida pelo método MQGIPP (*estimativa* MLWIN - MQGIPP)/MQGIPP), originando a distribuição destas diferenças relativas para cada um dos métodos testados no MLWIN 1.10 (Tabelas 5.1 a 5.4). A Tabela 5.5 apresenta as estimativas dos parâmetros

e de seus desvios padrões para o modelo explicitado em (4.1) obtidas pelos quatro procedimentos (MQGIPP, MPP, MPNP, MSP).

Cabe ressaltar que a única informação requerida pelo MLWIN 1.10 a respeito do plano amostral é o peso das observações, sem especificação do plano amostral utilizado na seleção das unidades populacionais.

Tabela 5.1 - Distribuição das diferenças relativas nas estimativas dos betas

Métodos comparados	Mínimo	1ºQuartil	Mediana	3º Quartil	Máximo
MSP x MQGIPP	-0,9153	-0,12111	-0,0008	0,17384	7,3484
MPNP x MQGIPP	-0,3161	-0,10176	-0,0219	0,17064	6,3466
MPP x MQGIPP	-0,0763	-0,00604	0,0010	0,00800	0,0374

Tabela 5.2 - Diferenças relativas nas estimativas de $\text{var}(\varepsilon_{ij}) = \hat{\sigma}^2$ **e** $\text{var}(u_{0j}) = \hat{\omega}^2$

Métodos comparados	$\hat{\sigma}^2$	$\hat{\omega}^2$
MSP x MQGIPP	0,05556	0,01733
MPNP x MQGIPP	1622,64954	0,19745
MPP x MQGIPP	0,02547	-0,00319

Tabela 5.3 - Distribuição das diferenças relativas nas estimativas dos desvios padrões dos betas

Métodos comparados	Mínimo	1ºQuartil	Mediana	3º Quartil	Máximo
MSP x MQGIPP	-0,2677	-0,23438	-0,2054	-0,15989	6,6019
MPNP x MQGIPP	-0,3064	-0,27344	-0,2088	-0,18242	6,6809
MPP x MQGIPP	-0,2705	-0,24089	-0,2172	-0,17200	-0,0991

Tabela 5.4 - Diferenças relativas nas estimativas dos desvios padrões de $\text{var}(\varepsilon_{ij}) = \hat{\sigma}^2$ **e** $\text{var}(u_{0j}) = \hat{\omega}^2$

Métodos comparados	$dp(\hat{\sigma}^2)$	$dp(\hat{\omega}^2)$
MSP x MQGIPP	-0,25445	-0,33936
MPNP x MQGIPP	1575,77087	-0,30487
MPP x MQGIPP	-0,27523	-0,35270

Tabela 5.5 - Comparação das estimativas dos parâmetros e de seus respectivos desvios padrões para o modelo hierárquico especificado na equação (4.1) obtidas pelo método MQGIPP e pelo MLWIN

Parâmetro	Estimativa			
	MQGIPP	**MLWin – PP[1]**	**MLWin – PNP[2]**	**MLWin – SP[3]**
Constante	3,0978 (0,0098)	3,098 (0,008)	3,102 (0,008)	3,102 (0,008)
Faixa etária				
20 a 29 anos	0,0000 (.)	0,0000 (.)	0,0000 (.)	0,0000 (.)
30 a 39 anos (FX2)	**0,0521 (0,0074)**	**0,052 (0,007)**	**0,046 (0,007)**	**0,046 (0,007)**
40 a 49 anos (FX3)	0,0614 (0,0087)	0,062 (0,007)	0,058 (0,007)	0,058 (0,007)
50 anos ou mais (FX4)	**0,0281 (0,0076)**	**0,028 (0,007)**	**0,033 (0,007)**	**0,033 (0,007)**
Sexo				
Masculino (S1)	0,0000 (.)	0,0000 (.)	0,0000 (.)	0,0000 (.)
Feminino (S2)	**-0,0025 (0,0084)**	**-0,002 (0,007)**	**-0,009 (0,007)**	**-0,011 (0,007)**
Anos de estudo				
Menos de 1 ano	0,0000 (.)	0,0000 (.)	0,0000 (.)	0,0000 (.)
De 1 a 7 anos	0,0192 (0,0068)	0,019 (0,005)	0,018 (0,005)	0,017 (0,005)
De 8 a 10 anos	**0,0033 (0,0092)**	**0,003 (0,007)**	**0,002 (0,007)**	**0,000 (0,007)**
11 anos	**-0,0140 (0,0092)**	**-0,014 (0,007)**	**-0,015 (0,007)**	**-0,016 (0,007)**
12 anos ou mais	**-0,0166 (0,0111)**	**-0,017 (0,009)**	**-0,025 (0,009)**	**-0,026 (0,009)**
Raça				
Não Branca (R1)	0,0000 (.)	0,0000 (.)	0,0000 (.)	0,0000 (.)
Branca (R2)	**0,0224 (0,0062)**	**-0,022 (0,005)**	**0,016 (0,005)**	**0,016 (0,005)**
Faixa etária * Sexo				
FX2.S2	**0,0022 (0,0111)**	**0,002 (0,009)**	**0,016 (0,009)**	**0,019 (0,009)**
FX3.S2	**0,0395 (0,0127)**	**0,039 (0,009)**	**0,048 (0,010)**	**0,051 (0,010)**
FX4.S2	0,0787 (0,0110)	0,079 (0,009)	0,077 (0,009)	0,079 (0,009)
Sexo * Raça				
S2.R2	-0,0392 (0,0081)	-0,039 (0,006)	-0,035 (0,007)	-0,034 (0,006)
Quintos de renda domiciliar mensal per capita				
1	0,0000 (.)	0,0000 (.)	0,0000 (.)	0,0000 (.)
2	**0,0168 (0,0089)**	**0,017 (0,007)**	**0,016 (0,006)**	**0,018 (0,007)**
3	**0,0281 (0,0094)**	**0,028 (0,007)**	**0,022 (0,007)**	**0,023 (0,007)**
4	0,0388 (0,0104)	0,039 (0,008)	0,039 (0,007)	0,041 (0,008)
5	0,0361 (0,0113)	0,036 (0,009)	0,039 (0,008)	0,038 (0,009)
Área de localização do domicílio				
Urbana	0,0000 (.)	0,0000 (.)	0,0000 (.)	0,0000 (.)
Rural	-0,0350 (0,0060)	-0,035 (0,005)	-0,032 (0,005)	-0,031 (0,005)
Despesa domiciliar mensal *per capita*				
Coeficiente	0,0181 (0,0050)	0,018 (0,004)	0,017 (0,004)	0,018 (0,004)
Variâncias dos erros aleatórios				
$var(u_{0j})$	0,0046 (0,0005)	0,005 (0,000)	7,390 (0,796)	0,005 (0,000)
$var(\varepsilon_{ij})$	0,0219 (0,0007)	0,022 (0,000)	0,026 (0,000)	0,022 (0,000)

Nota: [1] Pesos Padronizados; [2] Pesos Não-Padronizados; [3] Sem Pesos.

Para o ajuste no MLWIN 1.10 sem pesos amostrais e com pesos não-padronizados, é possível constatar discrepâncias na maioria das estimativas dos efeitos (linhas 1 e 2 das Tabelas 5.1 e 5.2) e de seus respectivos desvios padrões (linhas 1 e 2 das Tabelas 5.3 e 5.4), principalmente nas estimativas das componentes $\hat{\sigma}^2$ e $\hat{\omega}^2$. Quando avaliam-se os valores obtidos no MLWIN 1.10 com a incorporação dos pesos amostrais padronizados, as estimativas dos efeitos são praticamente idênticas àquelas obtidas pelo método MQGIPP (linha 3 das Tabelas 5.1 e 5.2). Entretanto, seus desvios padrões ainda permanecem distintos, sendo, em sua maioria, subestimados pelo MLWIN 1.10. A razão para tal diferença é que as expressões para o cálculo das variâncias de $\hat{\beta}$ e $\hat{\theta}$ utilizadas pelo MLWIN 1.10 não são as mesmas que aquelas referidas na seção 2, e explicitadas em Pfeffermann et al. (1998), acarretando estimativas viciadas dos parâmetros do modelo.

6. Conclusões

Neste trabalho, aplicou-se a metodologia que incorpora o plano amostral na estimação dos parâmetros de modelos lineares hierárquicos (MQGIPP), desenvolvida por Pfeffermann et al. (1998), considerando-se a existência de efeito aleatório de nível 2 (domicílio) apenas no intercepto. O modelo encontrado relacionou o índice de massa corporal de pessoas com 20 ou mais anos de idade com as variáveis sexo, idade, raça, anos de estudo, renda domiciliar per capita mensal, despesa domiciliar per capita diária com alimentação e área de localização do domicílio. Esta análise revelou que o IMC é influenciado pelo fato de pessoas estarem aglomeradas em domicílios, indicando que a utilização de modelos hierárquicos neste estudo foi adequada.

É reconhecida, entretanto, a dificuldade em se aplicar a metodologia citada acima, uma vez que não há pacotes estatísticos que considere o planejamento amostral na estimação dos parâmetros daqueles modelos, exigindo um esforço adicional de implementação do procedimento. Por outro lado, a não incorporação do plano amostral na análise multinível pode acarretar estimativas viciadas para os parâmetros do modelo de interesse, como verificado na seção 5.

A extensão desse procedimento para o caso de mais de um efeito aleatório de nível 2 não foi estudada neste trabalho por envolver maiores complicações computacionais. Modelos mais completos podem ser estudados mediante generalização do programa utilizado neste estudo.

Cabe ressaltar que o modelo linear hierárquico ajustado para o índice de massa corporal com base nos dados da PPV/IBGE não considerou variáveis nutricionais relevantes para o estudo de interesse, tais como: medidas de acúmulo de massa, consumo calórico individual, etc., por esta não ser uma pesquisa que tenha como objetivo principal o perfil nutricional dos indivíduos.

A idéia de repetir esta análise utilizando uma pesquisa direcionada a estudos nutricionais é bastante pertinente, podendo revelar informações relevantes para se compreender como variáveis que caracterizam o bem-estar social, o nível socioeconômico, a propensão à obesidade e os hábitos alimentares dos indivíduos influenciam a desnutrição e o sobrepeso dos mesmos.

Referências bibliográficas

Corrêa A, S. T. *Modelos lineares hierárquicos em pesquisas por amostragem – relacionando o índice de massa corporal às variáveis da Pesquisa sobre Padrões de Vida/IBGE.* Rio de Janeiro, 2001. 110 p. Dissertação (mestrado) – Escola Nacional de Ciências Estatísticas/IBGE.

Ferro-Luzzi, A.; Sette, S.; Franklin, M.; James, W.P.T. *A simplified approach to assessing adult chronic energy deficiency.* European Journal of Clinical Nutrition, v. 46, sup. 1, p. 173-186. 1992.

Hansen, M.H.; Hurwitz, W.N.; Madow, W.G. *Sample survey methods and theory.* Nova Iorque: John Wiley, 1953.

James, W.P.T.; Ferro-Luzzi, A.; Waterlow, J.C. *Definition of chronic energy deficiency in adults.* European Journal of Clinical Nutrition, v. 42, sup. 2, p. 969-981. 1988.

Pfeffermann, Danny; Skinner, C.J.; Holmes, D.J.; et al. Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society B, v. 60, parte 1, p. 23-40. 1998.

Pfeffermann, Danny. *The role of sampling weights when modeling survey data.* International Statistical Review, v. 61, n. 2, august, p. 317-337. 1993.

Rasbash, J.; Browne, W.; Goldstein, H.; et al. *A user's guide to MLwiN.* London: Institute of Education, 1999. 142 p.

SAS INSTITUTE INC. SAS® *Language: reference version 6.* 1.ed. NC: SAS Institute Inc., 1990. 1042 p.

Shah, Babubhaiv V.; Barnwell; B.G.; Hunt, P.N.; et. al. *SUDAAN User's Manual - Professional Software for SUrvey DAta ANalysis for multi-stage sample designs - release 6.0.* NC: Research Triangle Institute, 1992. 592 p.

Skinner, C.J.; Holt, D.; Smith, T.M.F. (eds). *Analysis of complex surveys.* Chichester: Wiley, 1989. 406 p.

Agradecimentos

Especiais agradecimentos aos professores Djalma Galvão Carneiro Pessoa, Pedro Luis do Nascimento Silva e Danny Pfeffermann pela supervisão e incentivo de fundamental importância para a produção deste trabalho. As autoras agradecem ainda ao Instituto Brasileiro de Geografia e Estatística e aos colegas da Coordenação de Métodos e Qualidade, pelo apoio e assistência durante a realização deste estudo. E, finalmente, agradecemos aos revisores, pelas sugestões e comentários de grande relevância.

Abstract

Hierarchically structured data are customary in Social Science. When data are obtained from sample surveys, the problem of how to incorporate the sampling weights in multilevel analysis arises. Unweighted multilevel analysis of survey data may lead to biased estimators of model parameters (see e.g. Pfeffermann et al., 1998).

Therefore, the main objective of this work is to implement the weighting procedure, developed in Pfeffermann et al. (1998), that incorporates the sampling design in the estimation of multilevel models. The properties and usefulness of the method are illustrated fitting a two level (variance components)

model using data from the Brazilian Living Standards Measurement Survey carried out in 1996-1997. The analysis considers the Body Mass Index (BMI) from people aged 20 years and over as response variable of the model and the following explanatory variables: sex, age, race, education, per capita household income, per capita food diary expenditure and household location area.

The model analysis indicates that, on average, the estimated BMI is larger for people living in urban households, with higher per capita household income and per capita diary food expenditure. In addition, the BMI decreases according to the individual's education and increases up to the age range 40 to 49. For men over this age range decreases, and it increases for females aged 50 and over.

The weighting method implemented in this work is compared to the conventional modeling procedure that ignores weights and sampling design. The results indicate the need to take into account the sampling design in multilevel analysis to protect against biased parameter estimates.

Key Words: Body mass index, hierarchical models, iterative generalized least squares, nonresponse, sample survey.

Inferência ecológica para recuperação de dados desagregados

Rogério Silva de Mattos*
Álvaro Veiga Filho**

Resumo

A escassez de dados desagregados representa séria restrição para estudos sociais em uma perspectiva espacial. O problema é agravado no Brasil pelo enxugamento do Sistema Nacional de Estatística do IBGE, ocorrido nos anos de 1990, e pela crise financeira de estados e municípios, que lhes dificulta fazerem dispendiosos levantamentos de dados. Técnicas de inferência ecológica (IE) são úteis nessa situação. Aplicações de IE incluem estudos de padrões de migração em demografia, estimação de tráfego em planejamento de transportes e atualização de matrizes de insumo–produto em economia, dentre várias outras. O artigo apresenta, discute e exemplifica os principais e mais recentes métodos para IE propostos na literatura, visando a salientar sua aplicabilidade para diversos problemas que aparecem em estudos sociais empíricos, bem como apresentar o estado da arte na área com indicação de softwares disponíveis.

Palavras-chave: inferência ecológica; desagregação de dados; tabelas de contingência; simulação.

1. Introdução

No Brasil, a demanda por informações socioeconômicas desagregadas, sobretudo em termos espaciais, vem se fazendo crescente. Essa demanda se faz tanto da parte de estudiosos acadêmicos quanto de formuladores e gestores de políticas públicas. No caso dos últimos, o processo de municipalização deslanchado pela Constituição de 1988, com a transferência de responsabilidades das esferas federal e estadual de governo para a esfera municipal, levou-os a se ressentir, cada vez mais, da escassez de informações socioeconômicas desagregadas ao nível das localidades e regiões em que atuam. Agravando essa situação, a Fundação Instituto Brasileiro de Geografia e Estatística -FIBGE- veio reformulando desde o início dos anos de 1990 seu sistema de

* Endereços para correspondências: Departamento de Análise Econômica – Univ. Federal de Juiz de Fora - Campus Universitário, Martelos, 36036-330, Juiz de Fora, MG. e-mail: rmattos@fea.ufjf.br
** Departamento de Engenharia Elétrica – PUC – Rua Marquês de São Vicente, 225 – Gávea – 22453-900 - Rio de Janeiro - RJ- e-mail: alvf@ele.puc-rio.br

dados socioeconômicos na direção de uma estrutura mais enxuta, com menor número de variáveis levantadas e a produção de estatísticas com um maior grau de agregação (Góes, 1996).

Com o novo sistema, recai sobre os estados e até sobre os municípios a carga de terem de levantar muitos dos dados desagregados ao nível regional/local de seu interesse. Entretanto, os levantamentos de dados estatísticos (*surveys*) são em geral custosos e as dificuldades financeiras de muitos estados e municípios lhes restringe de implementá-los. Uma forma de enfrentar essa situação é o apelo a "métodos não-*survey*" (tal como em inglês, *nonsurvey methods*), que se referem a quaisquer procedimentos alternativos aos *surveys*, *ad-hoc* ou formais, que possam ser usados para se aproximar os dados desagregados não disponíveis. Por exemplo, em estudos econômicos setoriais sob uma ótica regional, é usual desagregar-se espacialmente matrizes de insumo-produto através de um procedimento não-*survey* baseado em coeficientes locacionais (Round, 1978 e 1983). Em geografia humana e planejamento urbano e de transportes, vários problemas de estimação de dados socioeconômicos desagregados são tratados por métodos não–*survey* baseados em otimização de entropia (Novaes, 1981).

Uma área de pesquisa em métodos não-*survey* que vem se destacando recentemente é a de *inferência ecológica* (IE). A pesquisa em IE volta-se para problemas de recuperação ou estimação de dados desagregados (não-disponíveis) a partir de dados agregados (disponíveis)[1]. Especificamente, ela reúne o conjunto de procedimentos para se aproximar o conteúdo desconhecido de células em tabelas de contingência ou valores quando só se conhecem os totais das linhas e das colunas das tabelas. Como vários problemas em diferentes áreas, sobretudo das ciências sociais, se enquadram nessa caracterização, as técnicas de IE encontram diversas aplicações. A literatura sobre IE avançou devagar até meados da década de 1990, quando então sofisticados métodos estatísticos começaram a ser propostos e assim reativaram as pesquisas na área.

O objetivo deste artigo é apresentar alguns métodos estatísticos para se fazer IE, selecionados dentre os mais usados e os propostos recentemente. Não existe, aqui, a pretensão de se fazer um inventário extenso da literatura, mas apenas apresentar e exemplificar métodos para IE dentro das principais linhas de pesquisa em IE existentes no momento. Uma significativa relação de trabalhos é apresentada na seção de referências. Com isso, os autores pretendem difundir e estimular as aplicações de técnicas de IE por estatísticos e outros pesquisadores no contexto brasileiro.

O artigo está organizado da seguinte forma. Na seção 2, é feita uma breve revisão da literatura sobre IE, com ênfase nos desenvolvimentos recentes. Na seção 3, o problema da IE é descrito formalmente. Na seção 4, são apresentados cinco enfoques diferentes para tratar o problema. Na seção 6, é feita uma breve avaliação de softwares disponíveis para se implementar os métodos para IE apresentados. Na seção 7, são tecidos alguns comentários conclusivos.

[1] O termo "inferência ecológica" não se refere a procedimentos de inferência aplicados em ecologia, mas ao uso particular de dados agregados, também chamados de *dados ecológicos*, relativos a uma certa população para se estimar características de subgrupos da mesma

2. Literatura sobre IE

A maioria dos métodos para IE propostos na literatura foram motivados por pesquisas em ciência política, geografia regional e planejamento urbano e de transportes. Embora a preocupação de se fazer IE provavelmente exista há muito tempo, segundo King (1997) as primeiras incursões na literatura aparecem em estudos americanos de comportamento de voto na década de 1910. Uma posição cética quanto à real possibilidade de se fazer IE, posta por Robinson (1950), levou a pesquisa em IE a ficar praticamente adormecida por algumas décadas. De fato, do início da década de 1950 até meados da década de 1990, poucos métodos foram sugeridos na literatura. Em ciência política, destacaram-se nesse período o método dos limites de Duncan e Davis (1953), a chamada "Regressão de Goodman" (Goodman, 1953 e 1959), um método de otimização de entropia proposto por Johnston e Hay (1983) e o modelo agregado multinomial composto de Brown e Payne (1986). Em planejamento urbano e de transportes, métodos também baseados em otimização da entropia foram propostos por Wilson (1970a, 1970b; ver também Chilton e Poet, 1973; e Novaes, 1981). Ainda nesse período, o único estudo comparativo de métodos para IE foi feito por Cleave (1992) e reapresentado em Cleave, Brown e Payne (1995). Dentre esses métodos mais antigos, apenas o de Goodman (1953 e 1959) e o de Brown e Payne (1985) são baseados em formulação estatística. Os demais, inclusive os baseados em otimização da entropia, apresentam uma natureza *ad-hoc* ou determinística.

A pesquisa em IE ressurge no final da década de 1990 com o método proposto por King (1997), baseado na distribuição normal truncada. Este método foi algo revolucionário por usar um sofisticado modelo estatístico que também integra todas as informações determinísticas disponíveis sobre o problema, o que não fora feito por nenhum dos métodos propostos anteriormente. O método de King também inaugurou na literatura sobre IE o uso de modernos recursos de simulação estocástica para inferências estatísticas complexas (em geral, Tanner ,1996). Apesar disso, o método de King tem sido objeto de várias controvérsias (Cho, 1998; Friedman et al. 1999 e 2000; King, 1999a e 1999b; McCue, 2001; Anselin e Cho, 2002). Uma leitura cuidadosa dessas controvérsias, no entanto, indica que elas se devem mais à complexidade inerente ao processo de se fazer IE do que a eventuais limitações do método de King.

Pouco depois, King, Rosen e Tanner (1999) introduzem novo método, baseado em um modelo hierárquico binomial-beta de formulação bayesiana, que segundo os autores é mais versátil do que o método normal-truncada de King (1997) para se fazer IE. Tanto o método de King (1997) quanto o de King, Rosen e Tanner são restritos a problemas de IE em tabelas 2x2. Posteriormente, Rosen et al. (2000) generalizaram o último método para aplicações envolvendo tabelas de qualquer tamanho. A inferência com o modelo hierárquico binomial-beta (e sua versão generalizada) também é implementada com métodos de simulação estocástica, em particular os algoritmos de simulação de Monte Carlo por Cadeia de Markov (*Markov Chain Monte Carlo* - MCMC -; em geral, Tanner, 1996; Gelman et al. 1995), que são baseados em computação intensiva[2]. Antes de

[2] Mattos e Veiga (2002b) desenvolveram um método mais rápido para implementar uma versão ligeiramente modificada do modelo hierárquico binomial-beta baseado no algoritmo ECM (Meng e Rubin, 1993).

apresentarmos alguns desses métodos, faremos na próxima seção uma descrição formal do problema da inferência ecológica.

3. O problema

Tecnicamente, o problema da inferência ecológica, daqui por diante simplesmente *problema IE*, refere-se a como determinar o conteúdo das células em tabelas de contingência ou valores quando só são conhecidos os totais de linhas e colunas das tabelas. A Tabela 3.1 ilustra essa situação em um problema de determinação do comportamento de voto para uma região hipotética dos Estados Unidos da América.

Tabela 3.1 - Ilustração do problema da inferência ecológica

	Republicanos	Democratas	Não-votantes	
Brancos	?	?	?	3.246
Negros	?	?	?	1.173
	1.543	1.979	893	4.419

Nessa tabela, os números de brancos e negros em idade de votar (totais das linhas) bem como os números de votos para os candidatos Republicano e Democrata mais o total de não-votantes (totais das colunas) são *conhecidos*. Entretanto, não se sabem, por exemplo, os números de brancos que votaram no candidato republicano ou o de negros que votaram no candidato democrata. Ou seja, são *desconhecidos* os conteúdos das células, que por isso são representadas com um ponto de interrogação "?". O objetivo no problema IE é inferir os valores desagregados das células. Embora a Tabela 3.1 seja de ordem 2x3, o problema pode ser descrito em termos gerais para tabelas de ordem $R \times C$, onde R é um número qualquer de linhas e C um número qualquer de colunas.

3.1 O caso 2x2 com variáveis em proporções

Na maior parte deste artigo, estaremos trabalhando com a situação mais simples para o problema IE em que as tabelas são de tamanho 2×2 e, ao mesmo tempo, as variáveis (agregadas e desagregadas) são representadas como proporções. Este caso está apresentado formalmente na Tabela 3.2.

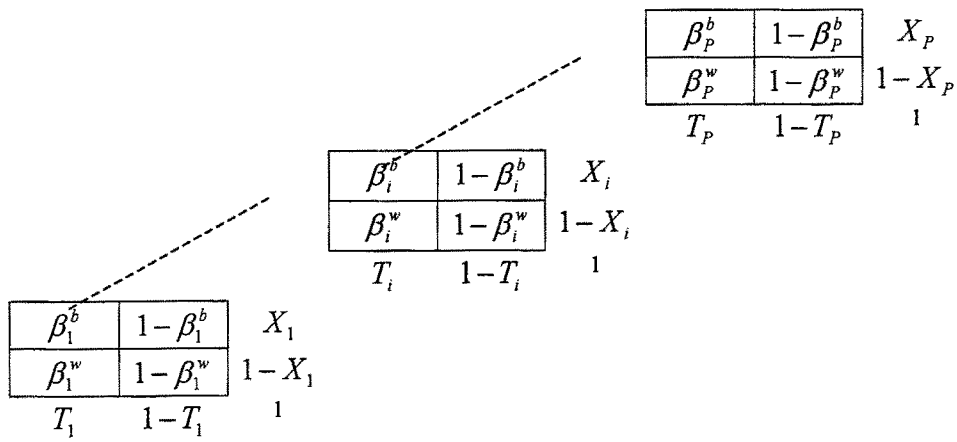
Tabela 3.2. Representação do problema IE para tabelas 2×2 e com variáveis medidas em proporções

		Variável II		
		1	2	*Totais*
Variável I	1	β_i^b	$1-\beta_i^b$	
	2	β_i^w	$1-\beta_i^w$	$1-X_i$
	Totais	T_i	$1-T_i$	1

As variáveis são representadas em proporções porque isso às vezes provê uma interpretação mais direta dos resultados. β_i^b e β_i^w representam as proporções *desagregadas* da primeira categoria da variável II no total das linhas correspondentes. X_i representa a proporção *agregada* da primeira categoria da variável I no total das suas duas categorias e T_i, por sua vez, representa a proporção *agregada* da primeira categoria da variável II no total das suas duas categorias. Apenas X_i e T_i *são observáveis* e, uma vez observado um conjunto de dados para as mesmas, o objetivo é recuperar ou prever os valores das proporções β_i^b e β_i^w, para $i = 1,...,p$. Estas proporções serão chamadas aqui de *quantidades de interesse* do problema IE. O subscrito i na Tabela 3.2 indica a i-ésima tabela ou unidade amostral, dentre um total de p tabelas que são consideradas na análise.

A notação da Tabela. 3.2 é geral e aplicável a vários contextos. Por exemplo, em economia, a variável I pode representar níveis de renda familiar e a variável II gastos em diferentes tipos de bens de consumo; em sociologia, a variável I pode representar o número de crimes, segundo diferentes regiões da cidade e a variável II o número de crimes por tipo; em planejamento de transportes, a variável I pode representar o número de residentes por área residencial e a variável II o número de empregos por área comercial. A flexibilidade da representação do problema IE faz com que as técnicas desenvolvidas para tratá-lo possam ser aplicadas em problemas de diversas áreas de pesquisa.

Figura 3.1 -O uso de várias tabelas como unidades amostrais



A idéia, seguida por muitos autores, de se trabalhar com várias tabelas ao mesmo tempo é, de um lado, usar um maior número de observações agregadas e, de outro, explorar aspectos comuns do processo gerador dos dados em cada tabela e com isso obter-se aproximações mais eficientes. Na prática, nem sempre existe tal similaridade pelo menos entre todas as p tabelas, e por isso alguns dos métodos a serem apresentados admitem extensões que permitem incluir variáveis explicativas.

3.2 Aspectos determinísticos

Há dois fatos determinísticos importantes do problema descrito na Tabela 3.2 que são usualmente considerados em métodos para IE. O primeiro é a *identidade contábil*, que formalmente é representada como:

$$T_i = \beta_i^b X_i + \beta_i^w (1 - X_i) \tag{1}$$

A equação (1) retrata uma relação exata entre as proporções agregadas e desagregadas. Ela é a contrapartida no espaço das proporções do fato de que, para uma tabela de valores, *a soma das células em uma certa coluna deve ser igual ao total da coluna.* Quando T_i e X_i são dados, a expressão (1) passa a caracterizar uma relação linear entre os valores possíveis para β_i^b e β_i^w. De fato, isso fica claro quando a reescrevemos conforme:

$$\beta_i^w = \frac{T_i}{1 - X_i} - \frac{X_i}{1 - X_i} \beta_i^b \tag{2}$$

A identidade contábil também é importante porque permite estabelecer intervalos admissíveis para as quantidades de interesse. Enquanto proporções, β_i^b e β_i^w estão restritas a assumirem valores no intervalo unitário [0,1], mas é possível mostrar a partir de (1) ou (2) que, dependendo dos valores observados de T_i e X_i, os intervalos de valores admissíveis podem ser mais estreitos. Isto significa que $\beta_i^b \in [L_i^b, U_i^b] \subseteq [0,1]$ e $\beta_i^w \in [L_i^w, U_i^w] \subseteq [0,1]$, onde:

$$L_i^b = \max(0, (T_i - 1 + X_i)/X_i) \tag{3}$$

$$U_i^b = \min(T_i / X_i, 1) \tag{4}$$

$$L_i^w = \max(0, (T_i - X_i)/(1 - X_i)) \tag{5}$$

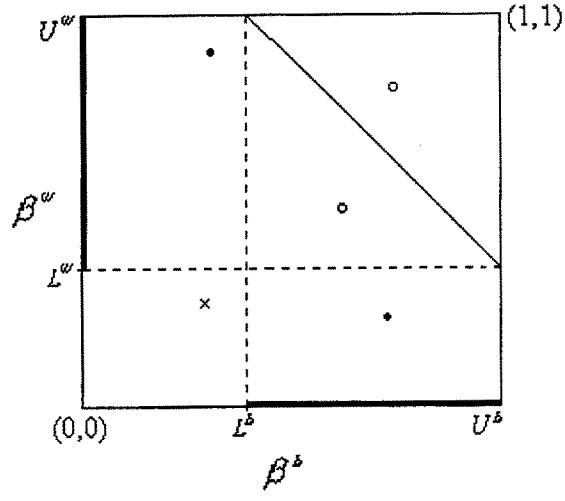
$$U_i^w = \min(T_i/(1 - X_i), 1) \tag{6}$$

(para uma prova, ver King(1997: 302-303)), onde L e U indicam limite *inferior* e limite *superior*, respectivamente. As expressões (3)-(6) foram estabelecidas na literatura sobre IE por Duncan e Davis (1953). Uma generalização das mesmas para tabelas $R \times C$ é apresentada em King(1997; 269-271).

3.3 Características das soluções

A Figura 3.2 ilustra a identidade contábil, os limites determinísticos e a propriedade de consistência na agregação[3]. Ela mostra o espaço-produto *a priori* de $\beta_i^b \times \beta_i^w$, formado pelo quadrado unitário [0,1]x[0,1], com uma linha negativamente inclinada. Para um dado par de proporções agregadas (X_i, T_i), essa linha representa a identidade contábil conforme a expressão (2). Diferentes pares (X_i, T_i) determinam diferentes linhas negativamente inclinadas cruzando o quadrado unitário. Antes de um par (X_i, T_i) ser observado, a linha correspondente à identidade contábil ainda não está determinada e o par de proporções desagregadas (β_i^b, β_i^w) pode ser qualquer ponto sobre o quadrado unitário. Porém, quando um par (X_i, T_i) é dado ou observado, (β_i^b, β_i^w) tem de ser necessariamente um dos pontos situados sobre a linha. A informação contida nos dados agregados, portanto, pode trazer substancial de redução de incerteza (de todo o quadrado unitário para uma linha) quanto aos valores possíveis para (β_i^b, β_i^w).

Figura 3.2 - Identidade contábil, limites e consistência na agregação



Além disso, note-se que as projeções dessa linha sobre os eixos caracterizam os intervalos admissíveis $[L_i^b, U_i^b]$ e $[L_i^w, U_i^w]$, respectivamente. Se um determinado método para IE respeita a identidade contábil, isto significa que os valores $\hat{\beta}_i^b$ e $\hat{\beta}_i^w$ aproximados ou previstos, segundo ele, necessariamente respeitam os limites (ou os intervalos admissíveis) e apresentam consistência na agregação, porque o ponto representado por $(\hat{\beta}_i^b, \hat{\beta}_i^w)$ iria se situar sobre a linha negativamente inclinada na Figura 3.2. Embora os pontos *fora da linha*

[3] Consistência na agregação é uma propriedade desejável de um método para IE. As previsões $\hat{\beta}_i^b$ e $\hat{\beta}_i^w$ geradas por este método têm de ser tais que, quando substituídas em (1), façam com que a identidade continue válida. Se nessa substituição, ao contrário, o lado esquerdo diferir do lado direito, então dizemos que $\hat{\beta}_i^b$ e $\hat{\beta}_i^w$ são previsões inconsistentes na agregação.

representem previsões que não apresentam consistência na agregação, eles podem, no entanto, respeitar ou não os limites determinísticos. Por exemplo, os círculos respeitam ambos os intervalos; os pontos pretos apenas um dos intervalos; e o "x" nenhum dos intervalos.

4. Enfoques

Alguns dos métodos propostos para solucionar o problema IE descrito na seção 3 serão agora apresentados, discutidos e exemplificados, a saber: o método dos limites de Duncan e Davis (1953), a Regressão de Goodman (1953, 1959), o modelo normal truncada de King (1997), o modelo hierárquico binomial-beta de King, Rosen e Tanner (1999), e o método de otimização de entropia de Wilson (1970a e 1970b). Esses métodos correspondem aos principais enfoques para IE disponíveis no momento. Para quase todos, existem softwares disponíveis para implementá-los, sobre o que falaremos na seção 6.

4.1 Método dos Limites

Um dos primeiros métodos de IE foi proposto por Duncan e Davis (1953) e envolve se trabalhar apenas com as informações determinísticas presentes nos dados agregados das tabelas. Os autores sugerem fazer as previsões das proporções desagregadas usando o ponto médio dos intervalos admissíveis, isto é $\hat{\beta}_i^b = (L_i^b + U_i^b)/2$ e $\hat{\beta}_i^w = (L_i^w + U_i^w)/2$, onde L_i^b, U_i^b, L_i^w e U_i^w são dados segundo (3)-(6). Este procedimento é conhecido na literatura sobre IE como método dos limites. Ele apresenta as vantagens de ser simples de ser aplicado, de gerar previsões que respeitam a identidade contábil e de funcionar bem quando os intervalos admissíveis são estreitos. Por outro lado, ele possui uma natureza *ad hoc* e logo só permite produzir previsões pontuais.

4.2 Regressão de Goodman

Goodman (1953) também propôs um dos primeiros métodos para se resolver o problema IE. Por se basear em um modelo clássico de regressão linear, o método ficou conhecido como "Regressão de Goodman" e também pelo termo "regressão ecológica". Em seu desenvolvimento, Goodman faz uma modificação na identidade contábil em (1), assumindo que as proporções desagregadas são as mesmas por tabela, ou seja, que $\beta_i^b = \mu^b$ e $\beta_i^w = \mu^w$, onde μ^b e μ^w são constantes para $i = 1,...p$. As diferenças decorrentes entre o lado esquerdo e o direito da identidade contábil em (1) seriam devidas a um erro aleatório ε_i, o que permitiria reescrevê-la como:

$$T_i = \mu^b X_i + \mu^w (1 - X_i) + \varepsilon_i \qquad i = 1,...,p \qquad (7)$$

Nesta forma, a identidade contábil vira um modelo de regressão linear sem constante. Para cada $i = 1,...,p$ podem ser determinadas estimativas de mínimos quadrados (constantes) para as quantidades de interesse. É imediato perceber que este método pode ser generalizado para tabelas de ordem $R \times C$.

Embora o método de Goodman seja baseado num procedimento estatístico, ele é limitado por duas razões. Primeiro, embora seja a hipótese de constância ao longo das diferentes tabelas que permite o uso de mínimos quadrados ordinários para se estimar os conteúdos das células, a evidência empírica em geral vai contra essa hipótese (Cho, 1998). Segundo, ao modificar a identidade contábil, o método de Goodman deixa de apresentar as boas propriedades, discutidas na seção 3.2, de consistência na agregação e respeito aos limites determinísticos. Não existe restrição para o valor assumido pelos parâmetros, quando estes deveriam se situar dentro dos intervalos admissíveis conforme os limites de Duncan e Davis, ou pelo menos dentro do intervalo [0,1], pois são proporções. Na prática, alguém pode obter proporções estimadas maiores do que 100% e até negativas (para um exemplo, ver King, 1997; Tabela 1.3, p. 16).

4.3 Modelo Normal-Truncada de King

O modelo de King(1997), que hoje é um marco na literatura sobre IE, consegue superar de forma consistente as limitações mencionadas da Regressão de Goodman. Essencialmente, este modelo admite que as quantidades de interesse possam variar de tabela para tabela (ao contrário da Regressão de Goodman) e faz isso de forma combinada com um modelo probabilístico para as quantidades de interesse. No final, é possível se fazer previsões estatísticas das mesmas com intervalos de confiança ainda mais estreitos que os definidos pelo limites determinísticos de Duncan e Davis. King (1997) apresenta duas formulações: o *modelo básico*, que usa só os dados agregados, e o *modelo estendido*, que também incorpora efeitos de variáveis explicativas.

Versão básica

O modelo básico de King (1997) usa a identidade contábil $T_i = \beta_i^b X_i + \beta_i^w (1 - X_i)$ de forma estrita, sem impor constância para as variáveis de interesse. Consequentemente, as previsões das proporções desagregadas produzidas por seu método apresentam boas propriedades, isto é, respeitam os limites determinísticos e apresentam consistência na agregação, como visto na seção 3.2. O modelo também apresenta as seguintes hipóteses:

a) as proporções desagregadas seguem *a priori* uma normal bivariada truncada sobre o quadrado unitário $A=[0,1]\times[0,1] \in R^2$, condicionada em X_i como abaixo:

$$(\beta_i^b, \beta_i^w \mid X_i) \sim TN_A(\breve{\psi}) \qquad i = 1,...,p \tag{8}$$

onde: $\breve{\psi}^{\mathrm{T}} = [\breve{\mu}^b, \breve{\mu}^w, \breve{\sigma}_b^2, \breve{\sigma}_w^2, \breve{\rho}]$ é o vetor de parâmetros da normal original (não-truncada) que dá origem à truncada;

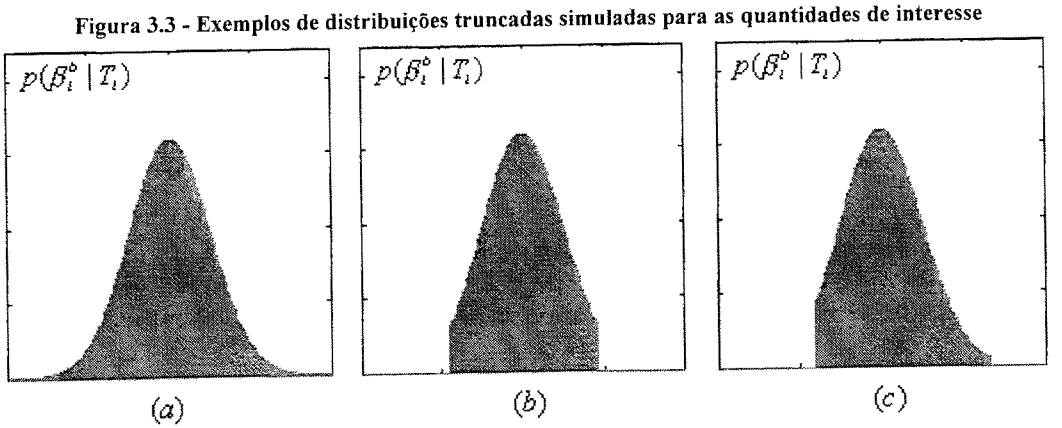
b) as médias de β_i^b e β_i^w independem de X_i ; e

c) $T_i \mid X_i$ é independente de $T_j \mid X_j$ para $i \neq j$.

Note-se na hipótese (a) que o modelo é condicional em X_i, isto é, essa variável é assumida como dada ou determinística. A distribuição truncada sobre [0,1]×[0,1] se deve ao fato de que, *a priori* ou antes de serem observados os dados agregados para T_i (uma vez que X_i é tomada como dada), as quantidades de interesse β_i^b e β_i^w são proporções e, portanto, assumem cada uma valores no intervalo [0,1]. A hipótese (b), por sua vez, significa que, apesar de X_i ser dada, as quantidades β_i^b e β_i^w são variáveis aleatórias cujas médias não dependem de X_i. E a hipótese (c) significa que o comportamento dos agregados em uma tabela é independente do comportamento dos agregados em outras tabelas.

A partir das hipóteses (a), (b) e (c) e da identidade contábil é possível derivar-se a distribuição marginal $p(T_i \mid \tilde{\psi})$, a função de verossimilhança $L(\tilde{\psi}) = \Pi_{i=1}^{P} p(T_i \mid \tilde{\psi})$, bem como as distribuições preditivas $p(\beta_i^b \mid T_i, \tilde{\psi})$ e $p(\beta_i^w \mid T_i, \tilde{\psi})$ para as quantidades de interesse. Embora King (1997) saliente que seu método possa ser usado sob um enfoque clássico ou bayesiano de estatística, ele no final acaba usando a segunda abordagem devido à necessidade de usar alguma distribuição *a priori* para $\tilde{\psi}$. Portanto, na prática seu método se baseia na distribuição *a posteriori* $p(\tilde{\psi} \mid T) = p(\tilde{\psi}) L(\tilde{\psi})$ e nas distribuições preditivas *a posteriori* $p(\beta_i^b \mid T_i)$ e $p(\beta_i^w \mid T_i)$. A implementação do modelo é feita em dois estágios: no primeiro, obtém-se a moda $\hat{\tilde{\psi}}$ da *posteriori* $p(\tilde{\psi} \mid T)$. No segundo, assume-se uma aproximação normal da *posteriori* em torno da moda e são obtidas por simulação as distribuições preditivas $p(\beta_i^b \mid T_i)$ e $p(\beta_i^w \mid T_i)$. Esse procedimento é usado porque a determinação dessas distribuições por meios analíticos é complexa e King o faz usando amostragem de importância (para detalhes, ver King, 1997; Capítulo 8).

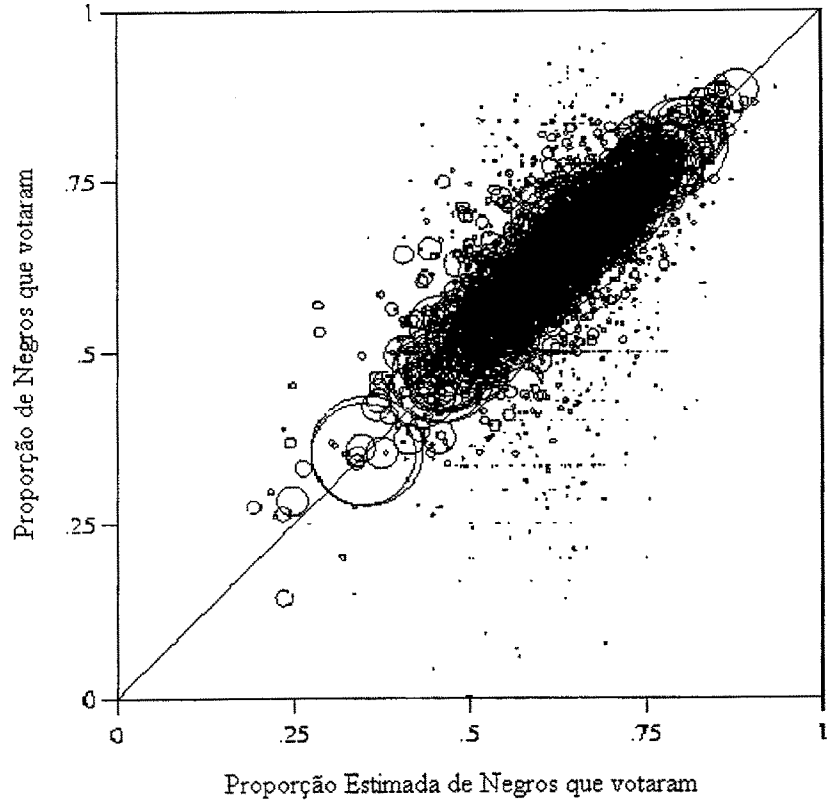
Figura 3.3 - Exemplos de distribuições truncadas simuladas para as quantidades de interesse



A Figura 3.3 ilustra os tipos de distribuições preditivas para as quantidades de interesse que podem resultar de se aplicar o método descrito acima na *i*-ésima tabela. No Gráfico 3.3(*a*), em que os truncamentos inferior e superior não são operantes (ocorrem em regiões de baixa probabilidade para a distribuição simulada), a curva resultante se assemelha a uma distribuição normal. Nos outros dois gráficos, isso não ocorre. No Gráfico 3.3(*b*), o limite inferior é operante, ao passo que o superior não. Isto gera uma assimetria da distribuição com corte abrupto em L_i^b. No Gráfico 3.3(*c*), ambos os limites são operantes e produzem alguma assimetria da distribuição resultante. Assim, em (*b*) e (*c*), as médias e variâncias das distribuições simuladas são significativamente diferentes das correspondentes à versão não-truncada. Situações similares aos três gráficos da Figura 3.3 podem ainda ocorrer quando $L_i^b = 0$ e $U_i^b = 1$, isto é, quando os limites de truncamento são iguais aos limites *a priori*.

O processo de implementação do método de King termina quando são produzidas previsões pontuais e intervalares para as quantidades de interesse, o que é feito computando-se as médias e os desvios padrão das versões simuladas de $p(\beta_i^b \mid T_i)$ e $p(\beta_i^w \mid T_i)$. A Figura 3.4 ilustra uma aplicação do modelo básico. Ela compara as proporções *verdadeiras* de negros que votaram com as correspondentes proporções *previstas* pelo método de King, usando-se apenas dados agregados (sem variáveis explicativas), para p = 3 262 zonas eleitorais (tabelas) na Louisiana, Estados Unidos (King, 1997, Figura 1.1, p.23). O tamanho de cada círculo que aparece na Figura 3.4 é proporcional ao número de negros em idade de votar na zona eleitoral a que o círculo se refere. Segundo King: "Que a vasta maioria de círculos recai próxima à linha diagonal é uma forte confirmação do método" (King, 1997: p. 23).

Figura 3.4 - Aplicação do modelo básico de King para 3 262 zonas eleitorais da Louisiana, Estados Unidos da América



Fonte: Reproduzido de King (1997, Figura 1.1, p. 23)

Versão estendida

O modelo estendido de King(1997) incorpora o efeito de variáveis explicativas, através de uma ligeira modificação do modelo básico. Para tanto, são considerados dois vetores de variáveis aleatórias: Z_i^b e Z_i^w, de ordens $(m\times 1)$ e $(n\times 1)$, respectivamente. Cada um desses vetores contém variáveis explicativas que afetam, respectivamente, o comportamento de β_i^b e β_i^w. Os vetores Z_i^b e Z_i^w podem conter uma única variável explicativa cada um, números iguais ou diferentes de variáveis, as mesmas variáveis ou algumas variáveis em comum e outras não. A incorporação dos efeitos dessas variáveis explicativas é feita assumindo-se que:

$$\breve{\mu}_i^b = a^b + \left(Z_i^b - \bar{Z}\right)^{\mathrm{T}} \alpha^b \tag{9}$$

$$\breve{\mu}_i^w = a^w + \left(Z_i^w - \bar{Z}\right)^{\mathrm{T}} \alpha^w \tag{10}$$

onde a^b e a^w são constantes que dependem dos parâmetros ψ da distribuição normal truncada para $(\beta_i^b, \beta_i^W \mid X_i)$. O vetor α^b contém os parâmetros desconhecidos que multiplicam as variáveis em Z_i^b e o vetor α^w os parâmetros desconhecidos que multiplicam as variáveis Z_i^w. O vetor de parâmetros para a ser dado por $\gamma^T = [a^b, a^w \alpha^{bT}, \alpha^{wT}, \breve{\sigma}_b^2, \breve{\sigma}_w^2, \breve{\rho}]$ e a distribuição *a posteriori* fica $p(\gamma \mid T) = p(\gamma)L(\gamma)$

A implementação do modelo estendido também é feita em dois estágios: no primeiro, obtém-se a moda $\hat{\gamma}$ da posteriori $p(\gamma \mid T)$. No segundo, assume-se uma aproximação normal da posteriori em torno da moda e são obtidas por simulação as distribuições preditivas $p(\beta_i^b \mid T_i)$ e $p(\beta_i^w \mid T_i)$. O procedimento termina quanto previsões pontuais e intervalares baseadas nas médias e desvios padrão dessas distribuições são calculados.

Uma importante limitação para se aplicar o método de King, tanto na versão básica como na estendida, é que ele só está implementado para tabelas de ordem 2×2. No Capítulo 15 de seu livro, King (1997) discute uma generalização de seu modelo para se implementar IE em tabelas de ordem maior (R×C), mas que, devido à falta de algoritmos computacionais rápidos para cômputo do fator de truncamento da distribuição normal multivariada truncada, não foi implementada pelo autor.

4.4 Modelo binomial-Beta de King, Rosen e Tanner

King, Rosen e Tanner(1999), doravante simplesmente KRT, introduzem novo método para IE baseado num modelo hierárquico binomial-beta. Este método segue uma abordagem bayesiana e também é voltado para tabelas 2x2. Embora seja diferente do método de King (1997) em vários aspectos, o método de KRT também usa uma formulação estatística rigorosa e é implementado por meio de algoritmos de simulação estocástica baseados em MCMC. Segundo KRT, o modelo hierárquico binomial-beta é mais flexível que o modelo normal-truncada de King para se fazer IE em problemas mais complexos: por exemplo, quando houver mais de uma moda nas distribuições preditivas para as quantidades de interesse. O modelo hierárquico binomial-beta também foi desenvolvido de modo a permitir o uso de variáveis explicativas, isto é, também apresenta uma versão básica e uma versão estendida.

Versão básica

Na versão básica, os autores constroem o modelo hierarquicamente em três estágios. No primeiro, admitem que a variável agregada T_i^* é uma variável inteira (isto é, não é uma proporção) que segue uma distribuição binomial com parâmetros de contagem N_i (= tamanho da população na unidade i) e parâmetro de probabilidade $\beta_i^b X_i + \beta_i^w (1 - X_i)$, isto é:

$$T_i^* \sim Bin\left(N_i, \beta_i^b X_i + \beta_i^w (1 - X_i)\right) \tag{11}$$

Aqui, os autores mantêm a hipótese usada por King (1997) de que as X_i s são dadas ou determinísticas. No segundo estágio, assumem que as quantidades de interesse β_i^b e β_i^w são variáveis aleatórias independentes, seguindo distribuições beta:

$$\beta_i^b \sim Beta(c_b, d_b) \tag{12}$$

$$\beta_i^w \sim Beta(c_w, d_w) \tag{13}$$

onde c_b, d_b, c_w e d_w são os parâmetros das betas em (12) e (13). No terceiro e último estágio da hierarquia, os autores assumem que os parâmetros das betas seguem, cada um, distribuições *a priori* independentes do tipo exponencial:

$$c_b \sim Expo(\lambda) \qquad d_b \sim Expo(\lambda) \tag{14}$$

$$c_w \sim Expo(\lambda) \qquad d_w \sim Expo(\lambda) \tag{15}$$

Note–se que o parâmetro das quatro exponenciais é o mesmo e igual a λ. Os autores assumem que a média dessas exponenciais é alta e correspondente a $1/\lambda = 2$, o que implica $\lambda = 2$ e que as distribuições *a priori* para c_b, d_b, c_w e d_w são não informativas.

Assim, dadas as hipóteses distribucionais nos três estágios, KRT constroem a distribuição *a posteriori* como segue:

$$P(Q \mid T) \propto \prod_{i=1}^{p} Bin\left(T_i \mid N_i, \beta_i^b X_i + \beta_i^w (1 - X_i)\right) \prod_{i=1}^{p} Beta\left(\beta_i^b \mid c_b, d_b\right) Beta\left(\beta_i^w \mid c_w, d_w\right) \times Expo\left(c_b \mid 1/2\right) Expo\left(d_b \mid 1/2\right) Expo\left(c_w \mid 1/2\right) Expo\left(d_w \mid 1/2\right) \tag{16}$$

onde $Q^{\mathrm{T}} = \left[\beta_1^b, ..., \beta_p^b, \beta_1^w, ..., \beta_p^w, c_b, d_b, c_w, d_w\right]$ é um vetor de ordem ((2*p*+4)×1) que contém todas as quantidades de interesse (nas *p* tabelas) mais os parâmetros das betas, e $T^{\mathrm{T}} = [T_1, ..., T_p]$ é um vetor (2*p*×1) contendo os *p* dados agregados para a variável T_i.

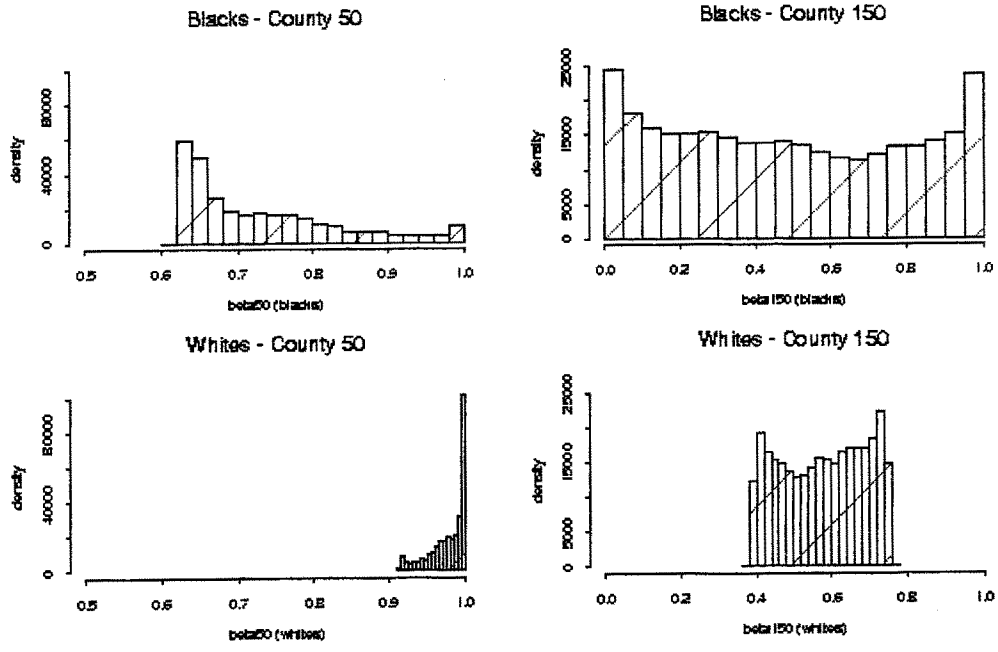
A fim de implementar a inferência bayesiana de forma rigorosa, isto é, recuperando completamente a distribuição *a posteriori* $P(Q|T)$ dada em (16) e suas marginais $P(Q_i|T)$, é preciso usar algum procedimento de simulação porque é difícil determinar $P(Q|T)$ por meios analíticos. KRT propõem o uso de algoritmos de MCMC. Como resultado, são obtidas as distribuições marginais *a posteriori* para os elementos de Q, isto é, para as quantidades de interesse β_i^b e β_i^w, $i = 1,...,p$, e os parâmetros c_b, d_b, c_w e d_w.

KRT apresentam, como exemplo, uma aplicação do modelo básico sobre os dados usados por King (1997; Capítulo 10). Esses dados referem-se a pessoas registradas e não-registradas, segundo origem racial em

275 condados de quatro estados do sudeste americano: Flórida, Louisiana, Carolina do Norte e Carolina do Sul. O objetivo é determinar, para cada condado, os percentuais de negros registrados β_i^b e não-registrados $(1-\beta_i^b)$, assim como de brancos registrados β_i^w e não-registrados $(1-\beta_i^w)$, usando-se informações agregadas sobre o percentual de negros na população X_i e sobre o total de pessoas registradas T_i^*.

Na Figura 3.5, estão apresentadas as distribuições marginais *a posteriori* obtidas para as quantidades de interesse em dois condados: os de números 50 (Figura 3.5(a)) e 150 (Figura 3.5(b)). O importante a salientar é a riqueza de formas distribucionais que o modelo hierárquico binomial-beta é capaz de captar, em contraposição ao modelo normal de King (1997). Na Figura 3.5(*a*), os gráficos para as posteriores de β_{50}^b e de β_{50}^w captam distribuições unimodais assimétricas com altas concentrações de probabilidades à esquerda e à direita, respectivamente. Já os gráficos para as posteriores de β_{150}^b e β_{150}^w apresentam formas diferentes das outras duas, apresentando duas modas salientes cada uma. Note-se também que as duas modas ocorrem em lados opostos nessas duas densidades, isto é, em seus extremos (caso de β_{150}^b) ou próximo a seus extremos (caso de β_{150}^w). Essa ocorrência de bimodalidade serve para representar a incerteza acerca do padrão de registro dos indivíduos em um dado condado quando há fortes expectativas tanto de que a taxa de registro seja alta como de que ela seja baixa.

Figura 3.5. - Aplicação do modelo binomial-beta sem variáveis explicativas: condados selecionados



(*a*) Distribuições marginais *a posteriori* para β_{50}^b e β_{50}^w

(*b*) Distribuições marginais *a posteriori* para β_{150}^b e β_{50}^w

Fonte: Adaptado de King, Rosen e Tanner (1998; Figuras 3 e 4).

Versão Estendida

O modelo binomial-beta apresentado na subseção anterior pode ser ligeiramente modificado para permitir a inclusão de variáveis explicativas, assim como no caso do modelo normal-truncada. Para tanto, KRT consideraram por simplificação uma única variável explicativa, denotada por Z_i e que é a mesma para ambas as quantidades de interesse β_i^b e β_i^w. Além disso, substituíram os parâmetros c_b e c_w nas distribuições beta do segundo estágio, redefinindo-as da seguinte forma:

$$\beta_i^b \sim Beta\left(d_b \exp(\alpha + \beta Z_i), d_b\right) \tag{17}$$

$$\beta_i^w \sim Beta\left(d_w \exp(\gamma + \delta Z_i), d_w\right) \tag{18}$$

No terceiro e último estágio da versão estendida, os autores continuam assumindo:

$$d_b \sim Expo(\lambda) \qquad\qquad d_w \sim Expo(\lambda) \tag{19}$$

onde as distribuições exponenciais em (19) apresentam médias $1/\lambda = 2$, de modo a serem priores não-informativas. Com relação aos parâmetros α, β, γ e δ, os autores seguem a filosofia usada em modelos bayesianos de regressão, isto é, assumem que são independentes e seguindo também priores não-informativas. A distribuição *a posteriori* para esta nova situação fica escrita como:

$$\begin{aligned} P(R \mid T) = & \prod_{i=1}^{p} Bin\left(T_i \mid N_i, \beta_i^b X_i + \beta_i^w (1 - X_i)\right) \\ & \times \prod_{i=1}^{p} Beta\left(\beta_i^b \mid d_b \exp(\alpha + \beta Z_i), d_b\right) \times \prod_{i=1}^{p} Beta\left(\beta_i^w \mid d_w \exp(\gamma + \delta Z_i), d_w\right) \\ & \times Expo\left(d_b \mid 1/2\right) Expo\left(d_w \mid 1/2\right) \end{aligned} \tag{20}$$

onde $R^T = \left[\beta_1^b, ..., \beta_p^b, \beta_1^w, ..., \beta_p^w, d_b, d_w, \alpha, \beta, \gamma, \delta\right]$ é um vetor de ordem $((2p + 6)\times 1)$ que contém todas as quantidades de interesse (nas p tabelas) mais os parâmetros das betas e os coeficientes de regressão. Da mesma forma que antes, é preciso aproximar por simulação $P(R|T)$ e as marginais $P(R_i|T)$ e, para isso, os autores propõem usar os mesmos algoritmos de MCMC de antes (ver KRT).

Mattos e Veiga (2002c) apresentam a primeira comparação extensiva entre a Regressão de Goodman, o modelo normal truncada e o modelo hierárquico binomial-beta. A comparação é feita com base em um experimento de Monte Carlo e mostra que o modelo normal truncada é o que tende a apresentar maior eficiência preditiva (em termos do erro quadrático médio) para recuperação das proporções desagregadas, vindo em

seguida o modelo hierárquico binomial-beta e por último a Regressão de Goodman.

4.5. Otimização da Entropia

Métodos de IE baseados em otimização da entropia foram propostos, de forma independente e em diferentes variações na literatura sobre IE. As técnicas pioneiras surgiram em planejamento urbano e de transportes, com os trabalhos de Wilson (1970a e 1970b). Posteriormente, em ciência política foi proposto o método de Jonhston e Hay (1983). A principal vantagem desses métodos entrópicos reside na facilidade de implementação, mesmo para tabelas de ordem $R \times C$. Por outro lado, suas desvantagens associam-se a serem métodos determinísticos e por não incorporarem variáveis explicativas. Recentemente, Judge, Miller e Cho (2002) apresentaram um novo método para IE, baseado em otimização de entropia que se estrutura dentro de uma abordagem estatística

Para falar desses métodos, iremos mudar ligeiramente a notação do problema IE. Assumimos, aqui, que este problema pode ser representado tal com na Tabela 3.3.

Tabela 3.3 - Representação alternativa do problema IE com tabelas *RxC*

		Variável II			
		1	...	C	**Totais**
Variável I	1	p_{11}	...	p_{1C}	$p_{1\bullet}$
	:	:	⋱	:	:
	R	p_{R1}	...	p_{RC}	$p_{R\bullet}$
	Totais	$p_{\bullet 1}$	...	$p_{\bullet C}$	1

Como antes, consideram-se conhecidos os totais das linhas $p_{i\bullet} = \Sigma_{j=1}^{C} p_{ij}$, $i = 1,...,R$, e das colunas $p_{\bullet j} = \Sigma_{i=1}^{R} p_{ij}$, $j = 1,...,C$. O objetivo é determinar as proporções p_{ij} referentes às células. Este formalismo acomoda vários problemas IE relacionados a tabelas de contingência ou de valores, porque os valores *absolutos* de células e totais de linhas e colunas podem ser representados em termos de proporções exatamente como na Tabela 3.3. Uma diferença em relação à notação da Tabela 3.2 é que agora as p_{ij}s são proporções das células no total geral da tabela e não simplesmente no total da i-ésima linha a que a célula pertence. Outra diferença é que por enquanto iremos considerar uma única tabela na análise, ao invés de várias, como fizemos anteriormente.

Os métodos de otimização da entropia operam sobre medidas de entropia em teoria da informação, como as de Shannon (1948) ou de Kullback (1959). Seja $\mathbf{P} = \{p_{ij}\}$ a matriz de ordem $R \times C$ formada pelos conteúdos das células na Tabela 3.3 e $F(\mathbf{P})$ uma função real que representa uma medida de entropia. Então, um método de IE baseado em otimização da entropia poderia ser genericamente descrito da seguinte forma:

$$\underset{\mathbf{P}}{Max} \quad F(\mathbf{P})$$

$$s.a. \quad \begin{cases} \sum_{i=1}^{R}\sum_{j=1}^{C} p_{ij} = 1 \\ \sum_{j=1}^{C} p_{ij} = p_{i*} \quad i = 1,...,R \\ \sum_{i=1}^{R} p_{ij} = p_{*j} \quad j = 1,...,C \end{cases} \tag{21}$$

onde s.a. significa "sujeito a". Quando trabalhamos com a medida de entropia de Shannon, aqui denotada por $S(\mathbf{P})$ e dada por:

$$S(\mathbf{P}) = -\sum_{i=1}^{R}\sum_{j=1}^{C} p_{ij} \ln p_{ij} \tag{22}$$

fazemos $F(\mathbf{P}) = S(\mathbf{P})$. Neste caso, estamos aplicando o conhecido princípio Maxent proposto por Jaynes (1957a e 1957b). Se, ao invés, trabalhamos com a medida de entropia-cruzada de Kullback, aqui denotada por $K(\mathbf{P})$ e dada por:

$$K(\mathbf{P}) = \sum_{i=1}^{R}\sum_{j=1}^{C} p_{ij} \ln\left(\frac{p_{ij}}{q_{ij}}\right) \tag{23}$$

fazemos $F(\mathbf{P}) = -K(\mathbf{P})$. Nesta segunda situação, estamos aplicando o também conhecido princípio MinxEnt proposto por Kullback (1959). Esse princípio, na verdade, se refere à minimização de K(**P**), e não à sua maximização, por isso que $F(\mathbf{P})$ tem de ser definida como o negativo de $K(\mathbf{P})$ para preservarmos a representação em (21).

As restrições que aparecem na representação em (21) provêm informações determinísticas sobre o problema IE, como o fato de que a soma de todas as células é igual a um, e de que as somas das células desconhecidas ao longo das linhas e colunas são iguais a totais conhecidos. A primeira restrição, de soma um, é em geral denominada de *restrição natural* e as demais de *restrições de consistência.* Elas provêm a mesma informação que a identidade contábil (1) e, portanto, as predições das quantidades de interesse são consistentes na agregação e respeitam os limites de Duncan e Davis.

A diferença entre os métodos MaxEnt e MinxEnt é que, basicamente, o primeiro procura achar a distribuição **P** que é mais próxima da uniforme, diferindo desta última em função apenas do conjunto de restrições. O princípio MinxEnt, por sua vez, tenta achar a distribuição dentro da matriz **P** que é mais próxima de uma distribuição qualquer conhecida *a priori* $\mathbf{Q} = \{q_{ij}\}$, também sujeita a restrições. Quando **Q** é a distribuição uniforme, MaxEnt e MinxEnt são equivalentes, isto é, geram a mesma solução ótima $\hat{\mathbf{P}}$, mas em caso contrário

não (para detalhes, ver Mattos e Veiga, 2002). Em geral, o método MinxEnt tende a ser de mais utilidade pois podemos introduzir mais informação na resolução do problema IE, como por exemplo, informações sobre os dados desagregados disponíveis para um período de tempo anterior (vide exemplo logo a seguir).

Exemplo

Para ilustrar a aplicação dos princípios de otimização da entropia MaxEnt e MinxEnt, esta seção apresenta um pequeno exemplo onde eles são usados para resolver um problema IE típico de planejamento de transportes. Este problema refere-se à determinação do número de viagens entre regiões dentro de uma localidade hipotética (em geral, Novaes, 1981).

O dados disponíveis foram criados artificialmente e estão apresentados na Tabela 3.4. O lado esquerdo da tabela apresenta as freqüências absolutas de viagens originadas das regiões O_1, O_2 e O_3, com destino às regiões D_1, D_2 e D_3. Por sua vez, o lado direito apresenta esses dados como proporções do total de viagens. Por exemplo, o número de viagens de O_1 para D_1 é 30 na matriz de fluxos e está associado à proporção 0,0952 (=30/315) na matriz de proporções.

Tabela 3.4 - Dados artificiais de número de viagens

	Matriz de Fluxos				Matriz de Proporções			
	D_1	D_2	D_3					
O_1	30	52	10	92	0,0952	0,1651	0,0317	0,2921
O_2	45	65	35	145	0,1429	0,2063	0,1111	0,4603
O_3	9	42	27	78	0,0286	0,1333	0,0857	0,2476
	84	159	72	315	0,2667	0,5048	0,2286	1

Em geral, no problema de determinação do número de viagens, são conhecidos os valores totais das colunas e das linhas das matrizes, mas não o conteúdo das células. O objetivo é determinar as proporções de viagens de cada região para as demais, que se assume sejam desconhecidas (e, determinando-se as proporções, é fácil calcular as freqüências correspondentes). Em outras palavras, o objetivo é determinar, ou "estimar", a matriz de proporções $\mathbf{P} = \{p_{ij}\}$ apresentada na Tabela 3.4. O fato de que esta seja conhecida no âmbito do exemplo é útil porque permite avaliar a capacidade de recuperação das proporções desagregadas com os métodos MaxEnt e MinxEnt, bem como compará-los entre si.

Os dados da Tabela 3.4 embutem as informações do conjunto de restrições do problema. A primeira é a restrição de soma 1 para o conteúdo das células. A tabela apresenta também o total das três linhas e das três colunas da matriz de proporções, o que permite especificar as restrições de consistência.

Aplicando o princípio MaxEnt (os cálculos foram feitos usando-se uma rotina desenvolvida por Mattos e Veiga (2002a)), a estimativa da matriz **P** obtida seria:

$$\hat{\mathbf{P}} = \begin{bmatrix} 0,0779 & 0,1474 & 0,0668 \\ 0,1228 & 0,2324 & 0,1052 \\ 0,0660 & 0,1250 & 0,0566 \end{bmatrix}$$

e o grau de aderência de $\hat{\mathbf{P}}$ em relação a $\mathbf{P}$ (que é conhecida hipoteticamente) poderia ser computado como:

$$\hat{s} = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}(p_{ij} - \hat{p}_{ij})^2} = 7,262 \times 10^{-2}.$$

A estimativa de $\mathbf{P}$ ainda poderia ser melhorada através do método MinxEnt. Suponhamos que houvesse uma matriz de proporções disponível, correspondente a um período anterior, e que tivesse sido computada a partir de um *survey* ou pesquisa de campo. Por exemplo:

$$\mathbf{Q} = \begin{bmatrix} 0,0975 & 0,1656 & 0,0290 \\ 0,1430 & 0,1967 & 0,1104 \\ 0,0299 & 0,1397 & 0,0883 \end{bmatrix}.$$

A matriz $\mathbf{Q}$ também foi produzida artificialmente, a partir da multiplicação dos valores da matriz de fluxo da Tabela 3.4 por variáveis geradas aleatoriamente segundo uma distribuição normal com média um e desvio padrão 0,05. Obtemos, então, uma nova estimativa da matriz $\mathbf{P}$:

$$\tilde{\mathbf{P}} = \begin{bmatrix} 0,0955 & 0.[illegible]72 & 0,0293 \\ 0,1432 & 0[illegible]0 & 0,1141 \\ 0,0280 & 0,[illegible]6 & 0,0852 \end{bmatrix}.$$

Como seria de se esperar, a matriz $\tilde{\mathbf{P}}$ aproxima melhor a matriz $\mathbf{P}$ devido ao uso das informações *a priori*, o que se reflete no seu grau de aderência:

$$\tilde{s} = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}(p_{ij} - \tilde{p}_{ij})^2} = 1,819 \times 10^{-2}$$

que é cerca de quatro vezes menor do que $\hat{s}$.

Jonsthon e Hay (1983) propuseram uma extensão desse enfoque de otimização da entropia onde se trabalha com várias tabelas (unidades amostrais) ao mesmo tempo, isto é, com um "cubo de dados" onde só os totais de linhas e colunas de cada seção (tabela) do cubo e a soma das células internas ao longo das tabelas são

conhecidas. Ou seja, é uma situação similar à representada na Figura 3.1. Neste enfoque, as proporções p_{ij} ficam redefinidas como p_{ijm}, onde o subscrito m se refere à m-ésima tabela, e as medidas de entropia S e K passam a incorporar mais um somatório, ao longo das M tabelas. Como mencionado na seção 3, essa forma de trabalhar permite "buscar força" do que há de comum entre diferentes tabelas (ou unidades amostrais). Embora seja mais geral que o método aqui apresentado para uma tabela, este enfoque preserva a limitação de ser determinístico.

Judge, Miller e Cho (2002) apresentaram recentemente um novo método de otimização da entropia para IE que é baseado em fundamentos estatísticos. Este método incorpora desenvolvimentos na área de inferência baseada em entropia de teoria da informação (*information theoretic-entropy inference*). Os principais resultados das pesquisas nessa área estão compilados em detalhe no recente livro de Mittelhammer, Judge e Miller (2000, capítulo 13). A abordagem para IE de Judge, Miller e Cho permite resolver uma versão ligeiramente modificada[4] do problema apresentado em (21), mas com uma diferença fundamental: cada restrição do problema é adicionada de um termo aleatório. Com isso, a solução de máxima entropia é tratada como um estimador ao qual está associada uma matriz de variância covariância. A partir dessa matriz, é possível construir-se margens de erro para as estimativas do conteúdo das células. Mittelhammer, Judge e Miller (2000) mostram que esse tipo de estimador apresenta boas propriedades assintóticas.

5. Softwares para IE

A disponibilidade de softwares de uso geral para se implementar métodos de IE ainda é restrita. Em geral, os autores desenvolveram rotinas específicas e nem todos as disponibilizaram para um público mais amplo. Faremos aqui, no entanto, uma indicação de softwares que viabilizam aos interessados fazerem implementações dos métodos de IE discutidos neste artigo.

O método dos limites de Duncan e Davis é muito simples e pode ser facilmente implementado mesmo em softwares do tipo planilha. A Regressão de Goodman pode ser implementada diretamente em softwares estatísticos ou econométricos que possuam rotinas especializadas para análise de regressão, como SPSS, Eviews e Shazam.

Para o método normal-truncada de King, existem duas opções restritas ao modelo para o caso 2x2 (com uma adaptação para o caso 2x3). A primeira é usar o pacote EI, que é um grupo de rotinas em linguagem GAUSS desenvolvidas por Benoit e King (1996); nesta opção, é necessário o sistema GAUSS e a aplicação *Constrained Maximum Likelihood* - CML/GAUSS, desenvolvida por Schoenberg (1996 e 1997) para maximização de funções de verossimilhanças ou posteriores com restrições. A segunda opção, mais acessível, é usar o software EzI, que é uma versão independente e amigável do software EI. Ambos embutem também o método dos limites e a Regressão de Goodman. Tanto EI quanto EzI podem ser obtidos livremente na *homepage* de Gary King em *http://gking.harvard.edu*. Para o método baseado no modelo hierárquico binomial-beta de

[4] As pequenas modificações são o fato de que são consideradas M tabelas, como no método de Johnston e Hay (1983), e as proporções p_{ijm} se referem ao total da i-ésima linha e não ao total da m-ésima tabela.

KRT, não existe software disponível para implementá-lo. Nem mesmo Rosen et al. (2000), quando lançaram a versão generalizada, disponibilizaram uma rotina específica. No entanto, para o caso 2x2, Mattos e Veiga (2002b) desenvolveram um grupo de rotinas em Matlab que implementa uma versão ligeiramente modificada desse método por um procedimento alternativo e bem mais rápido. A limitação é que são geradas previsões das quantidades de interesse apenas em termos pontuais e intervalares, isto é, sem a recuperação completa das distribuições marginais *a posteriori*. Essas rotinas podem ser obtidas em *http://www.rsmattos.ufjf.br* e, para serem usadas, é necessário o sistema Matlab (versão 5.0 ou superior) mais os *toolboxes* de estatística e de otimização.

Para o método de otimização da entropia, Mattos e Veiga (2002a) desenvolveram três rotinas em linguagem Matlab que implementam os métodos MaxEnt e MinxEnt para um única tabela. Essas rotinas também podem ser obtidas em *http://www.rsmattos.ufjf.br*. Para o método de Judge, Miller e Cho (2002), existem rotinas na linguagem GAUSS que são disponibilizadas em um CD-Rom que acompanha o livro de Mitthelhammer, Judge e Miller (2002).

6. Comentários Finais

Este artigo apresentou algumas metodologias para IE selecionadas dentre as mais usadas e as mais recentes propostas na literatura. Foram destacadas as potencialidades de tais metodologias para serem aplicadas em vários problemas de estudos sociais e, também, a sua relevância no atual contexto brasileiro decorrente da escassez relativa de dados e informações desagregados, sobretudo em termos espaciais.

Foram também apontadas vantagens e desvantagens de cada método, embora não tenha sido aqui apresentada uma comparação empírica dos mesmos usando dados reais, uma vez que o artigo teve como objetivo difundir o estado da arte e indicar ferramentas de implementação das diferentes, mas sobretudo as novas, metodologias. Um problema que existe em comparações de métodos de IE é que, na prática, os dados desagregados são desconhecidos e isto dificulta a avaliação geral de performance de diferentes métodos. Além disso, as expressões analíticas para as distribuições preditivas não estão disponíveis, como no caso dos modelos de King e hierárquico binomial-beta. Este é um dos aspectos em desenvolvimento na pesquisa sobre IE, pois até o momento só através de situações hipotéticas, isto.é, quando os dados desagregados são conhecidos, em geral, por simulação, é que é possível comparar-se o desempenho de métodos para IE. Este aspecto é discutido em Mattos e Veiga (2002c), que através de um experimento de Monte Carlo apresentam evidências de que o método normal truncada de King é mais eficiente que a regressão de Goodman e o método hierárquico binomial-beta para se gerar previsões pontuais.

Finalmente, é importante atentar para que, assim como qualquer método não-*survey*, em geral os métodos para IE não são substitutos para os tradicionais *surveys* quando houver a possibilidade e os recursos de se realizar os últimos. Dificilmente, os métodos para IE poderão proporcionar estimativas com menor grau de incerteza acerca dos dados desagregados. Porém, tanto os métodos disponíveis como as pesquisas ora em andamento de novos procedimentos para IE são desenvolvimentos promissores. São muitos os problemas de interesse em várias áreas das ciências sociais e no âmbito da formulação e gestão de políticas públicas para cujas

soluções podem contribuir os métodos para IE. Esta é um área que merece ser explorada por estatísticos e outros pesquisadores que com freqüência se deparam com a falta de dados desagregados.

Referências bibliográficas

Anselin, L. e W.K.T. Cho (2002). Spatial effects and ecological inference. Political Analysis 10, 3, 276-297.

Achen, C. H. e W. P. Shively (1995). Cross–level inference. Chicago: University of Chicago Press.

Benoith, K. e G. King (1996). EzI: An easy program for ecological inference. Manuscrito. (disponível em http://gking.harvard.edu)

Brown, P.J., and C.D. Payne (1986). Aggregate data, Ecological Regression, and Voting Transitions. Journal of the American Statistical Association, 81, 452-460.

Chilton, R. e R. W. W. Poet (1973). An entropy maximising approach to the recovery of detailed migration patterns from aggregate census data. Environment and Planning, 5, 135–146.

Cho, W. K. T. (1998) Iff the assumption fits... A comment on the King ecological inference solution. Political Analysis, 7, 143-163.

Cleave, N. (1992). Ecological inference. PhD. Dissertation, University of Liverpoool.

Cleave, N., P. J. Brown e C.D. Payne (1995). Evaluation of methods for ecological inference. Journal of The Royal Statistical Society, A, 158, Part 1, 55–72.

Ducan, O. D. e B. Davis (1953). An alternative to ecological correlation. American Sociological Review,18, 665–666.

Freedman, D.A., S.P. Klein, M. Ostland and M.R. Roberts (1998). A solution to the ecological inference problem (Book Review). Journal of the American Statistical Association, 93, 1518–120.

Freedman, D.A., M. Ostland, M.R. Roberts and S.P. Klein, (1999).Response to King. Letter to the Editor in Journal of the American Statistical Association, 94,355–357.

Gelman, A. B., J. S. Carlin, H. S. Stern and D. B.Rubin (1995). Bayesian data analysis. New York: Chapman & Hall/CRC.

Góes, M. C. (1996) A modernização das estatísticas econômicas. Trabalho apresentado no I Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais. Rio de Janeiro: FIBGE.

Goodman, L. (1953). Ecological regression and the behavior of individuals. American Sociological Review, 18, 663–664

Goodman, L. (1959). Some alternatives to ecological correlation. American Journal of Sociology, 64, 610–25.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. Physics Review, 106, 620–630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics II. Physics Review, 108, 171–190.

Johnston, R.J. e A. M. Hay (1983). Voter transition probability estimates: An entropy–maximizing approach. European Journal of Political Research, 11, 93–98.

Judge, G., D. Miller e W.K.T.Cho (2002). An information theoretic approach to ecological estimation and inference. in King, G., O. Rosen e M. Tanner (eds.). Ecological Inference: New Methodological Strategies. New York: Cambridge University Press (Capítulo 7). A sair.

Kullback, S. (1959). Information theory and statistics. New York: Wiley

King, G. (1997). A solution to the ecological inference problem: reconstructing individual behavior from aggregate data. Princeton: Princeton Universisy Press.

King, G. (1999a). The future of ecological inference research: a reply to Freedman et al. Journal of the American Statistical Association, 94, 352–355.

King, G. (1999b). Geography, statistics, and ecological inference. Manuscrito (disponível em http://gking.harvard.edu).

King, G., O. Rosen e M. A. Tanner (1999). Binomial–beta hierarquical models for ecological inference. Sociological Methods and Research, 28, 61–90.

Mattos, Rogério S. e Álvaro Veiga (2002a). Otimização de Entropia: Implementação Computacional dos Princípios MaxEnt e MinxEnt. Revista Pesquisa Operacional, 22, 1, 37-59.

_____ (2002b) The binomial–beta hierarchical model for ecological inference revisited and implemented via the ECM algorithm. (disponível no Paper Archive da Society for Political Methodology: http://web.polmeth.ufl.edu/)

_____ (2002c). A structured comparison of the Goodman Regression, the truncated normal, and the binomial-beta hierarquical methods for ecological inference. in King, G., O. Rosen e M. Tanner (eds.). Ecological Inference: New Methodological Strategies (Capítulo 15). New York: Cambridge University Press. A sair.

McCue, K. F. (2001). The statistical foundations of the EI method. The American Statistician, 55, 2, 106-110.

Meng, X-L. e D.B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika, 80, 267-278.

Mittelhammer, G. Judge e D. Miller (2000). Econometric Foundations. New York: Cambridge University Press.

Novaes, A. G. (1981). Modelos em planejamento urbano, regional e de transportes. São Paulo: Edgar Blücher.

Rosen, O., W. Jiang, G. King, and M. A. Tanner (2000). Bayesian and frequentist inference for ecological inference: the RxC case. Manuscrito. (disponível em http://gking.havard.edu)

Round, Jeffrey I. (1978). An interregional input–output approach to the evaluation of nonsurvey methods. Journal of Regional Science 18, 2, 179–194.

Round, Jeffrey I. (1983). Nonsurvey techniques: a critical review of the theory and the evidence. International Regional Science Review 8, 189–212.

Schoenberg, Ronald (1996). Constrained maximum likelihood. Manuscrito.

Schoenberg, Ronald (1997). Simulation of bayesian posterior distributions of parameters of constrained models. Manuscrito.

Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379-423.

Tanner, M. A. (1996). Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions. 3rd Edition. New York: Springer.

Wilson, A.G. (1970a). The use of the concept of entropy in system modelling. Operational Research Quarterly, 21, 247-265.

Wilson, A. G. (1970b). Entropy in urban and regional modelling. Monographs in spatial and environmental systems analysis. London: Pion.

Abstract

The shortage of disaggregate data is a severe restriction to the development of social studies under spatial perspectives. The problem is of major importance in Brazil because of the shrinkage of IBGE's Official Statistical System during the 1990's along with the financial crises of states and municipalities, what prevent them to implement expensive surveys. Techniques of ecological inference (EI) are useful in such instances. Applications of EI include assessment of migration patterns in demography, estimation of traffic flows in transportation planning, and updating of input–output matrixes in economics, among a range of others. The paper presents, discusses and exemplifies the major and most recent methods for EI proposed in the literature, aiming to highlight their applicability to a number of problems that arise in empirical social studies, as well as to present the state of the art in the area with indication of available softwares.

Key words: ecological inference; disaggregation of data; contingency tables, simulation.

O desenho amostral da pesquisa industrial - inovação tecnológica

Ana Maria Lima de Farias*
Wasmália Socorro Barata Bivar**
Aline Visconti Rodrigues**

Resumo

A hipótese central na qual se baseia o desenho amostral da Pesquisa Industrial - Inovação Tecnológica - PINTEC - é a de que a inovação é um fenômeno relativamente raro. Neste sentido, a adoção de desenhos tradicionais (geralmente é utilizada a amostragem aleatória estratificada por localização geográfica, atividade e porte da empresa) pode resultar em amostras com baixa freqüência de empresas inovadoras. Esta constatação indica a necessidade de identificar previamente, no cadastro de seleção, as empresas que têm maior probabilidade de serem inovadoras e aumentar a fração amostral para este subconjunto de empresas. Existe uma série de indicadores derivados da Pesquisa Industrial Anual e de informações de outras instituições que podem ser utilizados com este propósito. Este artigo descreve a construção de indicadores de inovação que permitiram a estratificação da população-alvo da PINTEC de acordo com as chances da empresa ser ou não inovadora, bem como o desenho amostral final.

Palavras-chave: evento raro, inovação tecnológica e alocação não-proporcional.

1. Introdução

É crescente o reconhecimento da utilidade da informação estatística e da necessidade de empregá-la na tomada de decisão, visando a reduzir sua incerteza e complexidade. Neste contexto, têm sido ampliados os temas sobre os quais a sociedade requer informação. Diante do intenso e rápido processo de mudança técnica, torna-se urgente a criação de um sistema de informações sobre as atividades de inovação tecnológica das empresas industriais no Brasil. Em vários países, sobretudo os europeus, este tipo de informação já vem sendo coletado há

* Endereços para correspondências: UFF – Departamento de Estatística, Rua Mario Santos Braga s/nº - Campus do Valonguinho, Instituto de Matemática - 7º andar, 24020-110 - Niterói - RJ, amlima.ntg@terra.com.br.
** IBGE - Departamento de Indústria, Av. República do Chile 500, 4º andar, 20031-170 - Rio de Janeiro - RJ, bivar@ibge.gov.br e alinevisconti@ibge.gov.br.

alguns anos e já existem recomendações internacionais em termos conceituais e metodológicos para o seu levantamento.

O Instituto Brasileiro de Geografia e Estatística - IBGE - realizou, com apoio da Financiadora de Estudos e Projetos - FINEP - e do Ministério da Ciência e Tecnologia - MCT -, a Pesquisa Industrial - Inovação Tecnológica - PINTEC -, que tem por objetivo geral a construção de indicadores das atividades de inovação tecnológica das empresas industriais brasileiras, compatíveis com as recomendações internacionais e que permitam às empresas avaliarem o seu desempenho em relação às médias setoriais, às entidades de classe analisarem a conduta tecnológica dos setores e aos governos desenvolverem e avaliarem políticas públicas.

Este artigo descreve as decisões que levaram ao desenho da amostra da PINTEC. Na seção 2 apresentam-se as definições metodológicas que levaram ao desenho final da amostra; na seção 3 detalha-se o desenho da amostra propriamente dito; e na seção 4 são descritos os procedimentos de estimação das variáveis da pesquisa, apresentando-se, no Apêndice 1, as fórmulas detalhadas dos estimadores e de suas variâncias estimadas. No Apêndice 2, apresenta-se a relação das atividades econômicas consideradas, bem como a distribuição da população e da amostra nos estratos finais.

2. Definições metodológicas

As principais recomendações metodológicas para a produção de estatísticas na área de ciência e tecnologia estão consolidadas nos manuais da Organisation for Economic Co-operation and Development - OECD -, que têm por objetivo a sistematização de critérios e procedimentos que permitam a construção de indicadores comparáveis entre países e a difusão das boas práticas. Dentre eles, o Oslo Manual - Proposed Guidelines for Collecting and Interpreting Technological Innovation Data (1997) trata da mensuração direta da inovação tecnológica, recomendando que seja adotado como objeto de análise o comportamento e as atividades inovativas desenvolvidas pelas empresas, concentrando-se sobre as estratégias, os incentivos e os obstáculos das empresas nesta área, como também sobre os resultados e os efeitos da inovação tecnológica.

O Manual de Oslo pretende ser um guia para a realização e normatização das pesquisas de inovação tecnológica e busca ser exaustivo, abordando todos os procedimentos e critérios metodológicos que permitam obter indicadores de qualidade e comparáveis internacionalmente.

As escolhas acerca das características da PINTEC, além de considerar as recomendações internacionais, foram muito influenciadas pelo fato de esta ser a primeira pesquisa de tal natureza realizada no Brasil, apresentando pela primeira vez às empresas industriais a necessidade de mensurar um conjunto de conceitos e definições complexos relacionados à inovação tecnológica. O próprio termo inovação tem múltiplas significações, as instruções escritas (que possam constar de um manual de instruções de preenchimento) têm pouca influência sobre as idéias preconcebidas com que os respondentes iniciam a compilação de questionários e o entendimento de instruções escritas pode variar de acordo com a familiaridade com o tema. Deste modo, optou-se por realizar entrevistas diretas em todas as empresas da amostra, como forma de obter uniformidade conceitual das informações obtidas nas empresas.

Esta decisão condicionou fortemente as escolhas sucessivas, em especial o âmbito e o desenho amostral, uma vez que o tamanho da amostra deveria ser tal que tornasse viável a realização das entrevistas.

No Capítulo 7 do Manual de Oslo são discutidos os procedimentos de amostragem e estimação, no qual sugere-se, por razões práticas, que apenas as empresas com pelo menos 10 pessoas ocupadas sejam investigadas, embora as atividades inovativas sejam realizadas por empresas de todos os portes. A população-alvo da PINTEC foi, então, definida como o conjunto de empresas com 10 ou mais pessoas ocupadas atuando nas atividades das indústrias extrativas e de transformação.

Neste mesmo capítulo, a técnica de amostragem probabilística estratificada é apresentada como aquela que apresenta resultados mais confiáveis, recomendando-se a estratificação da população por tamanho da empresa, medido pelo número de trabalhadores, e pela principal atividade econômica, sendo sugerido o detalhamento mínimo da atividade equivalente à divisão (2 dígitos) da Classificação Nacional de Atividades Econômicas - CNAE.

É interessante observar que a técnica de amostragem estratificada é amplamente utilizada nas pesquisas econômicas oficiais brasileiras, cujas estimativas se baseiam em conceitos contábeis, a partir dos quais é possível mensurar fenômenos econômicos que são, em geral, comuns a todas as empresas ativas durante um certo período de tempo (todas as empresas empregam, pagam salários, vendem, etc.). Por outro lado, as estatísticas econômicas oficiais utilizam como cadastro básico de seleção das amostras o Cadastro Central de Empresas - CEMPRE -, no qual são identificadas as unidades que formam a população de referência e que contém as variáveis necessárias para a estratificação (porte, atividade e localização) e a variável que é utilizada para controlar o coeficiente de variação das estimativas de cada subconjunto da população, para as quais são divulgados os resultados das pesquisas. No caso destas pesquisas, esta variável é o número de pessoas ocupadas (PO), informação que está disponível para todas as unidades que compõem o CEMPRE. Utilizar o controle da precisão da estimativa do total do pessoal ocupado na definição do tamanho da amostra é justificado pelo fato de esta ser uma das principais variáveis das pesquisas econômicas, que, por sua vez, é fortemente correlacionada com as demais variáveis para as quais se deseja obter estimativas.

A primeira diferença entre a PINTEC e as demais pesquisas econômicas é que nem todas as empresas estão envolvidas em atividades inovativas durante o período analisado pela pesquisa (1998/2000) e, em relação aos fenômenos mensurados nestas pesquisas (vendas, emprego, etc.), a inovação é um fenômeno relativamente raro. Neste sentido, a adoção de desenhos tradicionais pode resultar em amostras com baixa freqüência de empresas inovadoras. Além disso, não existe no cadastro de seleção nenhuma variável diretamente relacionada às variáveis que a PINTEC pretende estimar.

Por outro lado, o tamanho da amostra estava condicionado pela capacidade operacional de realização de entrevistas em todas as empresas da amostra, que é uma decisão gerencial orientada pelos recursos e tempo disponíveis, não sendo objeto de uma escolha técnica. A aplicação de entrevistas presenciais para toda a amostra da PINTEC implicaria em custos impraticáveis. Visando a tornar factível a realização de entrevistas com todas as empresas, optou-se por obter as informações da pesquisa através de entrevistas por telefone para a maior parte das empresas da amostra (utilizando um sistema de entrevistas telefônicas assistidas por computador,

desenvolvido especialmente para este fim), combinadas com a realização de entrevistas presenciais nas empresas de maior porte, que têm estruturas organizacionais muito mais complexas e nas quais, de acordo com a literatura econômica, se concentra a atividade de inovação tecnológica.

Avaliou-se que a adoção desta estratégia de captura permitiria a realização de cerca de 10 mil entrevistas. Como a população de empresas industriais com 10 ou mais pessoas ocupadas no cadastro básico de seleção é de aproximadamente 75 mil unidades, uma amostra de 10 mil empresas resultaria em uma fração amostral da ordem de 13%.

3. O desenho amostral

A hipótese central na qual se baseia o desenho amostral da PINTEC é a de que a inovação é um fenômeno relativamente raro. Neste sentido, a adoção de desenhos tradicionais (geralmente é utilizada a amostragem aleatória estratificada por localização, atividade e porte da empresa) pode resultar em amostras com baixa freqüência de empresas inovadoras. Esta constatação indica a necessidade de identificar previamente, no cadastro de seleção, as empresas que têm maior probabilidade de serem inovadoras, e aumentar a fração amostral para este subconjunto de empresas. Existe uma série de indicadores derivados da Pesquisa Industrial Anual e de informações de outras instituições que podem ser utilizados com este propósito.

Diante da impossibilidade de uma operação prévia de *screening* (uma inspeção exaustiva das empresas do cadastro, de modo a identificar as empresas inovadoras), foram utilizadas informações oriundas das seguintes fontes para gerar indicadores capazes de identificar este subconjunto:

- Cadastro do Ministério da Ciência e Tecnologia, contendo a relação das empresas industriais que se beneficiaram de incentivos fiscais concedidos pelas Leis de nºs 8.661 e 8.248. A variável FINANC indica que a empresa consta desta relação;
- Banco de Dados de Patentes e de Contratos de Transferência de Tecnologia do Instituto Nacional de Propriedade Industrial - INPI. Com este banco, foram construídas as seguintes variáveis indicadoras: CLPAT1 indica as empresas que possuíam entre 1 e 19 patentes registradas; CLPAT2, as que tinham 20 ou mais patentes registradas; CLPAT0, as que não possuíam registro de patente; CTECALL, número de contratos de transferência de tecnologia registrados nos anos de 1998 a 2000;
- Cadastro da pesquisa realizada pela Associação Nacional de Pesquisa, Desenvolvimento e Engenharia das Empresas Inovadoras - ANPEI - nos anos de 1998 e 1999, que informam anualmente seus dispêndios nas atividades de pesquisa, desenvolvimento e engenharia não rotineira (P&D&E). As empresas contidas neste cadastro foram identificadas com as seguintes variáveis indicadoras: ANPEI98 e ANPEI99;
- Cadastro da Fundação Sistema Estadual de Análise de Dados de São Paulo, contendo as empresas que declararam ser inovadoras na Pesquisa da Atividade Econômica Paulista de 1996 (variável indicadora PAEP) e na Pesquisa da Atividade Econômica Regional (variável PAER);

- Cadastro do Censo de Capitais Estrangeiros do Banco Central, contendo a participação do capital estrangeiro, no ano de 1995, nas empresas atuantes no Brasil. A variável CAPEST indica as empresas com 10% ou mais de participação de capital estrangeiro;
- Das informações das empresas que participaram das amostras de 1998, 1999 e 2000 da Pesquisa Industrial Anual, foram identificadas aquelas que declararam ter realizado aquisições incorporadas ao ativo imobilizado e, para aquelas que possuem 30 ou mais pessoas ocupadas, a aquisição de máquinas e equipamentos (variáveis AQUMAQ98, AQUMAQ99 e AQUMAQ00); as empresas que realizaram dispêndios para o pagamento de *royalties* e assistência técnica, tendo sido construídas as variáveis ROYALT98, ROYALT99, ROYALT00; e aquelas que, em 1998 e 2000, declararam ter realizado dispêndios com capacitação tecnológica, sendo INOVA98, INOVA00 as respectivas variáveis indicadoras;
- Cadastro de empresas graduadas em incubadoras, obtido por meio de pesquisa realizada pela COPPE/UFRJ com o apoio do Instituto Euvaldo Lodi e do Ministério da Ciência e Tecnologia, cuja variável indicadora é INCUBA; e
- Cadastro de empresas de base tecnológica no Estado de São Paulo, obtido do estudo de Fernandes e Côrtes (1998), com variável indicadora EBT.

3.1 - Indicadores de inovação

A partir das variáveis indicadoras acima, foram criados outros indicadores, que foram, por sua vez, divididos em dois grupos: no primeiro grupo estão os indicadores principais, que fornecem fortes indícios de que a empresa desenvolveu alguma atividade de inovação tecnológica, e no segundo grupo estão os indicadores secundários, que fornecem evidências moderadas. Os dois grupos de indicadores foram assim formados:

- Indicadores principais: as empresas com mais de um contrato de tecnologia nos anos de 1998, 1999 e 2000 (CTECALL); as empresas de base tecnológica (EBT), as empresas que se beneficiaram de incentivos concedidos pelas Leis de nºs 8.661 e 8.248 (FINANC), as empresas com 20 ou mais patentes registradas (CLPAT2), as empresas que realizaram pagamentos de *royalties* em três anos consecutivos (1998 a 2000 - ROYALT) e as empresas graduadas em incubadora (INCUBA); e
- Indicadores secundários: todas as demais empresas que constavam em pelo menos um dos cadastros definidos acima.

3.2 - Composição dos estratos

Seguindo as recomendações internacionais, a PINTEC deve fornecer estimativas por atividade econômica. Além disso, é fundamental que a pesquisa seja capaz de fornecer estimativas também para as diversas regiões geográficas do País. Considerando o tamanho pré-fixado da amostra e levando em conta o fato de estarmos lidando com um fenômeno relativamente raro, a estratificação da população-alvo da PINTEC foi

definida pelos cruzamentos de atividade econômica, conforme os agrupamentos da CNAE dados na Tabela A1 do Apêndice 2, e indicadores de inovação, definidos a seguir.

O objetivo da estratificação da população-alvo da PINTEC por indicadores de inovação é identificar e separar as empresas de acordo com as chances de serem ou não inovadoras. Assim, foram criados três estratos: um estrato certo, onde todas as empresas serão incluídas com probabilidade um na amostra e dois estratos amostrados, diferenciados pelo grau de incerteza com relação à presença do fenômeno em estudo.

Uma empresa, com qualquer indicador principal ativo, foi alocada no estrato certo. Como, de acordo com a literatura, as grandes empresas têm maior probabilidade de serem inovadoras, o estrato certo também incluiu todas as empresas com PO ≥ 500. Essa definição representa também um compromisso com as recomendações internacionais e as práticas de outros países.

Para as empresas restantes, a partir dos indicadores secundários, definiu-se uma nova variável, TOTINOV, que indica o número total desses indicadores ativos:

$$\begin{aligned} TOTINOV = {} & ANPEI98 + ANPEI99 + CAPEST + PAEP + PAER + \\ & AQUMAQ98 + AQUMAQ99 + AQUMAQ00 + \\ & INOVA98 + INOVA00 + CLPAT1 + CTECALL(1) + \\ & ROYALT98 + ROYALT99 + ROYALT00 \end{aligned}$$

As empresas com valor de TOTINOV igual ou superior a nove também foram alocadas no estrato certo. As empresas sem nenhum indicador foram alocadas no estrato amostrado não-elegível. As demais empresas (0 < TOTINOV < 9) foram alocadas no estrato amostrado elegível.

Além desses indicadores, foram considerados também alguns setores industriais, que se caracterizam por maiores oportunidades tecnológicas, a saber:

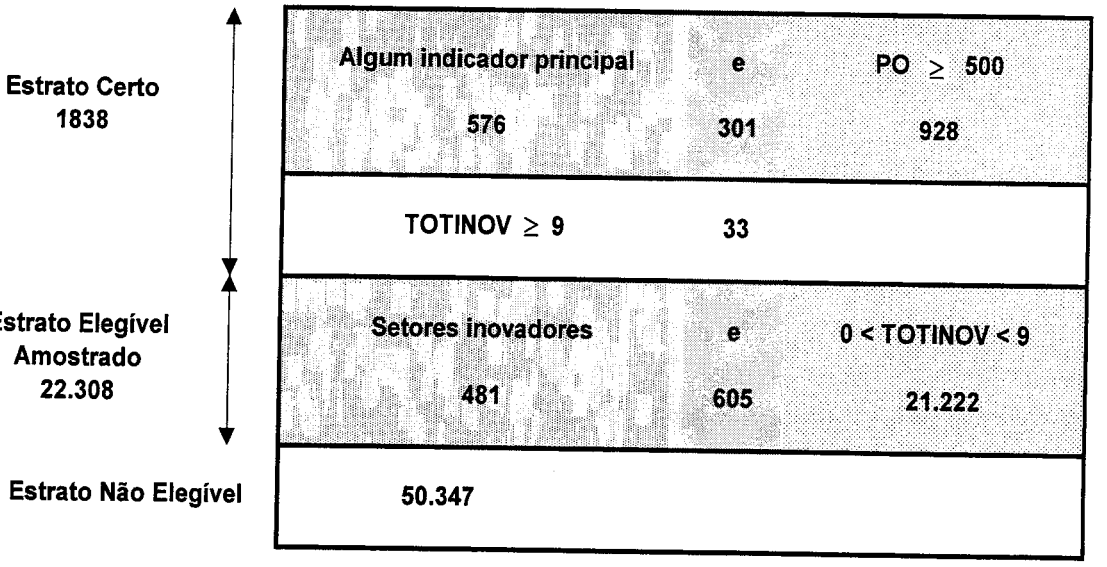
- Farmacêutica (CNAE3 = 245)
- Defensivos (CNAE3 = 246)
- Máquinas para escritório e equipamentos de informática (CNAE = 301 ou 302)
- Fabricação de material eletrônico básico (CNAE3 = 321)
- Fabricação de aparelhos de telefonia e TV (CNAE3 = 322 ou 323)
- Fabricação de aeronaves (CNAE3 = 353)

As empresas desses setores foram incluídas no estrato amostrado elegível.

Como já dito, o cadastro de seleção da amostra da PINTEC foi o CEMPRE, tomando-se como população de referência o conjunto de empresas industriais com 10 ou mais pessoas ocupadas. Na Figura 1, ilustra-se a distribuição da população de referência da PINTEC nos estratos definidos pelos indicadores de inovação.[1]

[1] Depois de feitos os acertos finais no cadastro, a população de referência totalizava 72 003 empresas.

Figura 1- Distribuição da População de Referência da PINTEC nos Estratos de Inovação
74 493 empresas



3.3 - A amostra da PINTEC

Por restrições orçamentárias, o tamanho total da amostra da PINTEC foi fixado em 10 000 empresas. Por estarmos em um contexto de população rara, vamos trabalhar com amostra estratificada desproporcional, aumentando a fração amostral entre as empresas do estrato amostrado elegível. Essa distribuição foi feita de modo que 80% da amostra virão dos estratos elegíveis e 20% dos estratos não-elegíveis. Descontadas as 2 mil empresas do estrato amostrado não-elegível, como há 1 838 empresas no estrato certo, 6 162 serão retiradas do estrato amostrado elegível. A alocação nos estratos finais (CNAE) foi proporcional à raiz quadrada do PO; com esse procedimento, incorpora-se implicitamente na estratificação o fato de as grandes empresas serem prováveis inovadoras, mas suavizando este efeito.

Com o objetivo de assegurar a obtenção de informações para as 10 mil empresas selecionadas e minimizar os problemas decorrentes de situações especiais de coleta que implicam perdas, ou seja, na falta de informações, a amostra original foi aumentada. Estimou-se, a partir da amostra original da PIA-Empresa de 1999, em 10% a taxa de perda, e o tamanho da amostra foi aumentado nesta proporção, o que significou um acréscimo de 1000 empresas. Como não era possível aumentar o tamanho da amostra no estrato certo, a distribuição destas empresas foi feita proporcionalmente ao tamanho das amostras dos estratos amostrado não-ilegível e ilegível. Sendo assim, no estrato não-ilegível foram alocadas $\frac{2000}{2000+6162} \times 1000 \approx 250$ empresas e as 750 empresas restantes foram alocadas no estrato amostrado elegível. A distribuição dessas empresas nos agrupamentos de atividade levou em consideração as taxas de perda específicas ocorridas na amostra da PIA-Empresa de 1999 nestas agregações.

Definido, assim, o tamanho da amostra em cada estrato final, o sorteio foi feito independentemente em cada um deles com probabilidade de seleção proporcional ao PO, tendo-se o cuidado de ordenar as empresas por região de interesse (Norte, Nordeste, Sudeste, São Paulo, Sul e Centro-Oeste). Esse procedimento de estratificação implícita tenta garantir a representatividade de todas as regiões geográficas na amostra. Por questões de arredondamento, o tamanho final da amostra ficou em 11 044 empresas, cuja distribuição pelos estratos finais é dada na Tabela A2 do Apêndice 2.

3.4 - Algoritmo de seleção da amostra

Em cada estrato final, definido pelos cruzamentos ELEGÍVEL x CNAE, foram executados os seguintes passos:

1) Ordenação das empresas por REGIÃO e CNPJ

- A ordenação pela variável REGIÃO, que define a região geográfica, tenta garantir a representatividade de todas as regiões na amostra; e
- Com a ordenação pelo CNPJ da empresa, espera-se que não haja tendências e estratificações implícitas ou mesmo correlação entre observações vizinhas. Sendo assim, a amostragem sistemática é essencialmente equivalente à amostragem aleatória simples.

2) Acumulação do Pessoal Ocupado empresa a empresa.

- A variável PO define a probabilidade de seleção de cada empresa; e
- Para cada empresa, calcula-se o valor da variável POAC, que é igual ao PO da empresa mais o PO de todas as empresas anteriores, ou seja, POAC é o PO acumulado.

3) Cálculo do tamanho do salto da amostragem sistemática.

- O salto é definido como

$$SALTO = \frac{PO_h}{n_h}$$

onde

PO_h = total do PO no estrato h; e

n_h = tamanho da amostra no estrato h.

4) Sorteio de um começo aleatório a_0.

- O começo a_0 é tal que $0 < a_0 < SALTO$; e
- Os sorteios para os diferentes estratos são feitos de forma independente, o que garante que as amostras sejam independentes.

5) Seleção sistemática das empresas da amostra.

- A j-ésima empresa da amostra é a primeira empresa tal que
  $POAC > a_0 + (j-1) \times SALTO$.

4. Estimação

Nesta seção descrevemos os procedimentos gerais de estimação das variáveis da PINTEC, apresentando no Apêndice 1, as fórmulas detalhadas dos estimadores e suas variâncias.

4.1 Estimação das variáveis da PINTEC

As variáveis da PINTEC estão divididas em dois grupos: variáveis quantitativas e variáveis qualitativas.[2]

As variáveis quantitativas da PINTEC referem-se basicamente aos dispêndios e ao pessoal ocupado, para os quais serão dadas estimativas de totais. Nas equações (8) e (10) do Apêndice 1, temos as fórmulas do estimador e de sua variância estimada, respectivamente.

Com relação às fontes de financiamento, a partir dos percentuais e respectivos valores informados por cada empresa, foram calculados os valores obtidos de cada fonte de financiamento. Os percentuais dos dispêndios, de acordo com a fonte de financiamento, foram então estimados através do estimador de uma razão (ver seção A.2 do Apêndice 1).

Para a variável vendas, embora as informações dadas pelas empresas sejam de natureza quantitativa, como não se dispunha do valor total das vendas, elas foram tratadas como variáveis qualitativas, dividindo-se as empresas em categorias de acordo com o percentual de vendas.

Para as variáveis qualitativas (por exemplo, se a empresa implementou inovação de produto e processo, importância das atividades de P&D, etc.), deseja-se estimar o número de empresas em cada categoria, assim como o respectivo percentual. Para isso, define-se uma variável indicadora para cada uma das categorias de uma variável qualitativa como

$$\delta_{hi}(c) = \begin{cases} 1 & \text{se unidade } i \text{ do estrato } h \text{ pertence à categoria de interesse } c \\ 0 & \text{caso contrário} \end{cases}$$

e o número de empresas $\hat{N}(c)$ na categoria c é estimado como o total da variável indicadora acima através da fórmula (8) e a variância dessa estimativa é obtida através da fórmula (10), ambas no Apêndice 1.

[2] O questionário da PINTEC encontra-se disponível no endereço www.pintec.ibge.gov.br.

Os percentuais de interesse podem ser relativos ao tota do universo (por exemplo, percentual de empresas inovadoras) ou ao total de um subgrupo (por exemplo, tre as empresas inovadoras). No primeiro caso, as fórmulas para o estimador de uma razão se aplicam (ver seção A.2 do Apêndice 1), definindo a variável do denominador como $x_{hi} = 1$ para todas as unidades da amostra. No segundo caso, as mesmas fórmulas se aplicam, também, levando em conta que o número de empresas $\hat{N}(g)$ no subgrupo g de interesse é estimado através da variável indicadora apropriada, como explicado acima.

A estimação de todas as variáveis da PINTEC foi feita com o programa SUDAAN [Research Triangle Institute (2001)]. Para maiores detalhes dos procedimentos, consulte também IBGE (2002).

4.2 Calibração dos pesos amostrais

O número de empresas estimado a partir da amostra da PINTEC é uma informação muito importante, principalmente porque várias das informações geradas pela pesquisa referem-se a percentuais. Para evitar uma discrepância entre o número de empresas do cadastro (população de referência) e o número de empresas estimado a partir da amostra, foi feita uma calibração dos pesos amostrais de modo a garantir compatibilidade entre essas fontes de informação.

Uma correção inicial foi feita nos pesos originais do desenho para corrigir algumas inconsistências do cadastro, tais como: mortes de empresas, mudança de estrato, inclusão de novas empresas, etc. Esse conjunto de pesos foi passado para o programa de calibração como sendo o conjunto de pesos do desenho.

A calibração final foi feita com o programa GES [ver Statistics Canada (2002)], usando como totais de calibração o número de empresas por atividade econômica e por classe de pessoal ocupado no cadastro e impondo a condição dos pesos de calibração serem todos maiores ou iguais a um. A partir dos pesos amostrais do desenho, o programa calcula os pesos de calibração que garantem a coincidência do número de empresas por classe de PO e por atividade econômica, estimado a partir da amostra com o respectivo número de empresas do cadastro. Esses pesos de calibração são, então, usados pelo programa SUDAAN no cálculo dos estimadores e suas variâncias (eles correspondem aos w_{hi} utilizados nas fórmulas do Apêndice 1). Para maiores detalhes dos procedimentos, consulte IBGE (2002).

Na Tabela 1, a seguir, temos as estatísticas-resumo dos três conjuntos de pesos, onde PesoOrig se refere ao peso inicial do desenho, PesoFin se refere aos pesos corrigidos para ajustes no cadastro e PesoCal se refere ao peso final, calibrado, que foi usado no cálculo das estimativas. Pode-se ver que os números de empresas estimados a partir de cada um dos conjuntos de peso são diferentes; com o peso original, houve uma subestimação e com o peso do desenho, uma superestimação. Os pesos de calibração corrigem essas diferenças.

Tabela 1 - Estatísticas-resumo dos pesos amostrais originais, corrigidos e calibrados

Estatística	PesoOrig.	PesoFin.	PesoCal.
Número de Observações	10.328	10.328	10.328
Média	6,63	7,33	6,97
Moda	1,00	1,08	1,26
Soma	**68.519**	**75.574**	**72.005**
Desvio Padrão	9,77	11,18	10,51
Máximo	50,25	59,70	60,26
Percentil 90%	20,83	23,39	21,99
3º quartil	6,79	7,15	6,38
2º quartil	2,36	2,47	2,48
1° quartil	1,00	1,07	1,23
Percentil 10%	0,93	0,97	1,03
Mínimo	0,40	0,43	1,00

5. Conclusão

Neste artigo descreveu-se o desenho amostral da Pesquisa Industrial – Inovação Tecnológica - PINTEC -, cuja característica principal foi a utilização de indicadores auxiliares, derivados de diversas fontes, que permitiram estratificar a população de referência de acordo com as chances de a empresa ser ou não inovadora. Essa característica, aliada com uma alocação desproporcional, tenta superar o fato de inovação tecnológica ser um evento raro, buscando melhorar a representatividade do segmento de interesse das empresas inovadoras na amostra. A estratificação final da população levou em conta os diferentes graus de certeza sobre a inovação, bem como a principal atividade econômica da empresa.

A estimação das variáveis foi feita com o programa SUDAAN, usando-se pesos amostrais calibrados pelo programa GES, de modo a garantir a coincidência do número total de empresas em cada atividade econômica e classe de tamanho.

Esse desenho amostral, bem como um sistema eficiente de captura das informações, levou a resultados bastante coerentes e de boa qualidade, já disponibilizados em CD-ROM pelo IBGE.

Referências bibliográficas

Cochran, W.G. (1977). ***Sampling Techniques*****, 3ª edição. Nova York: Wiley.**

Fernandes, A. C. e Côrtes, M. R. (1998) ***Caracterização do Perfil da Pequena Empresa de Base Tecnológica no Estado de São Paulo: uma análise preliminar.*** **Relatório de pesquisa. São Carlos: UFSCar, mimeo.**

IBGE (2002) ***Estimação da PINTEC: Uma aplicação dos programas GES e SUDAAN.*** **Relatório Interno. Rio de Janeiro: IBGE, Departamento de Indústria.**

Research Triangle Institute (2001). *SUDAAN User's Manual, Release 8.0.* Research Triangle Park, NC: Research Triangle Institute.

Särndal, C.E., Swensson, B. e Wretman, J. (1992). *Model Assisted Survey Sampling.* Nova York: Springer.

Shah, B.V., Folson, R.E., LaVange, L.M., Wheeless, S.C., Boyle, K.E., Williams, R.L. (1995) *Statistical Methods and Mathematical Algorithms Used in SUDAAN.* Research Triangle Park, NC: Research Triangle Institute.

Statistics Canada (2002) *Generalized Estimation System (GES)*, v. 4.2.

Apêndice 1

Estimadores e suas variâncias

Como já dito, o sorteio da amostra da PNTEC foi feito de tal forma que o desenho sistemático pode ser aproximado por um desenho de amostragem aleatória simples. Sendo assim, apresentamos, a seguir, as fórmulas dos estimadores e de suas variâncias relativas ao desenho de amostragem estratificada com reposição. Para isso, faremos uso da seguinte notação:

N_h	tamanho da população no estrato h
n_h	tamanho da amostra no estrato h
y_{hi}	valor da variável y para a unidade i do estrato h

Suponha que a probabilidade de seleção de cada elemento da amostra seja proporcional à variável P. Denotando por P_{hi} o valor dessa variável para a unidade i do estrato h, definimos, então, as seguintes quantidades, onde $h = 1,\ldots,H; \quad i = 1,\ldots,N_H$:

$$p_{hi} = \frac{P_{hi}}{\sum_{i=1}^{N_h} P_{hi}}$$

$$\pi_{hi} = n_h p_{hi}$$

$$w_{hi} = \frac{1}{\pi_{hi}}$$

A.1 Estimação de totais

O estimador do **total** populacional de uma variável y é dado por

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{1}{n_h} \frac{y_{hi}}{p_{hi}} = \qquad (1)$$

$$= \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{y_{hi}}{\pi_{hi}} = \tag{2}$$

$$= \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} y_{hi} \tag{3}$$

$$\doteq \sum_{h=1}^{H} \hat{Y}_h \tag{4}$$

e sua variância é estimada por

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{1}{n_h(n_h-1)} \left(\frac{y_{hi}}{p_{hi}} - \hat{Y}_h \right)^2 \tag{5}$$

Definindo

$$Z_{hi} = \frac{y_{hi}}{\pi_{hi}} = \frac{y_{hi}}{n_h p_{hi}} = w_{hi} y_{hi} \tag{6}$$

podemos escrever

$$\hat{Y}_h = \sum_{i=1}^{n_h} w_{hi} y_{hi} = \sum_{i=1}^{n_h} Z_{hi} = n_h \bar{Z}_h \tag{7}$$

e

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} Z_{hi} \tag{8}$$

Substituindo (6), (7) e (8) em (5), obtemos que

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{1}{n_h(n_h-1)} \left(n_h Z_{hi} - n_h \bar{Z}_h \right)^2 \tag{9}$$

ou seja

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} n_h \left[\frac{1}{n_h - 1} \left(Z_{hi} - \bar{Z}_h \right)^2 \right] \doteq \sum_{h=1}^{H} n_h S_{Z,h}^2 \tag{10}$$

A.2 Estimação de Razão

Sejam X e Y totais populacionais para os quais se deseja estimar a razão

$$R = \frac{Y}{X} \tag{11}$$

O estimador de R é

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi} y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi} x_{hi}} \tag{12}$$

A variância de tal estimador é estimada fazendo uso do método da aproximação linear pela série de Taylor, que descrevemos, a seguir, em um contexto mais geral.[3]

Seja $\theta = F(X,Y)$ uma função não-linear dos totais populacionais X e Y. O estimador de θ é $\hat{\theta} = F(\hat{X},\hat{Y})$. Usando a expansão de Taylor, podemos escrever:

$$\hat{\theta} = F(\hat{X},\hat{Y}) = F(X,Y) + \partial F_X(X,Y)(\hat{X} - X) + \partial F_Y(X,Y)(\hat{Y} - Y) + \text{termos de ordem superior}$$

Assumindo que os termos de ordem superior possam ser desprezados, resulta:

$$\hat{\theta} = F(\hat{X},\hat{Y}) = F(X,Y) + \partial F_X(X,Y)(\hat{X} - X) + \partial F_Y(X,Y)(\hat{Y} - Y) \tag{13}$$

A variância de $\hat{\theta}$ pode, então, ser estimada da seguinte forma:

$$\hat{V}(\hat{\theta}) \cong E(\hat{\theta} - \theta)^2$$
$$\cong [\partial F_X(X,Y)]^2 E(\hat{X} - X)^2 + [\partial F_Y(X,Y)]^2 E(\hat{Y} - Y)^2 + 2\partial F_X(X,Y)\partial F_Y(X,Y)E(\hat{X} - X)(\hat{Y} - Y) =$$
$$= [\partial F_X(X,Y)]^2 V(\hat{X}) + [\partial F_Y(X,Y)]^2 V(\hat{Y}) + 2\partial F_X(X,Y)\partial F_Y(X,Y)Cov(\hat{X},\hat{Y})$$

Como os totais populacionais são desconhecidos, eles são substituídos por suas estimativas, o que nos leva à seguinte expressão para a estimativa da variância de $\hat{\theta}$:

[3] O procedimento apresentado se generaliza para o caso de mais de dois totais populacionais.

$$\hat{V}(\hat{\theta}) \cong [\partial F_X(\hat{X},\hat{Y})]^2 V(\hat{X}) + [\partial F_Y(\hat{X},\hat{Y})]^2 V(\hat{Y}) + 2\partial F_X(\hat{X},\hat{Y})\partial F_Y(\hat{X},\hat{Y}) Cov(\hat{X},\hat{Y}) \quad (14)$$

Note, entretanto, que, se definirmos

$$\hat{Z} = \partial F_X(\hat{X},\hat{Y})\hat{X} + \partial F_Y(\hat{X},\hat{Y})\hat{Y} \quad (15)$$

então $\hat{V}(\hat{\theta}) = V(\hat{Z})$. Mas $\hat{Z}$ é uma função linear das variáveis X e Y medidas para cada elemento da amostra; mais precisamente,

$$\begin{aligned} \hat{Z} &= \sum_{h=1}^{H}\sum_{i=1}^{n_h} \partial F_X(\hat{X},\hat{Y}) w_{hi} x_{hi} + \sum_{h=1}^{H}\sum_{i=1}^{n_h} \partial F_Y(\hat{X},\hat{Y}) w_{hi} y_{hi} = \\ &= \sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}[\partial F_X(\hat{X},\hat{Y}) x_{hi} + \partial F_Y(\hat{X},\hat{Y}) y_{hi}] \end{aligned} \quad (16)$$

Sendo assim, a variância de $\hat{Z}$, ou melhor, a variância estimada de $\hat{\theta}$ é obtida substituindo-se a variável linearizada

$$Z_{hi} = w_{hi}[\partial F_X(\hat{X},\hat{Y}) x_{hi} + \partial F_Y(\hat{X},\hat{Y}) y_{hi}] \quad (17)$$

na fórmula da variância (10).

Quando $\theta = \frac{Y}{X}$, temos que $\partial F_X(X,Y) = -\frac{Y}{X^2}$ e $\partial F_Y(X,Y) = \frac{1}{X}$. Então, a variável linearizada para estimar a variância de $\hat{R} = \frac{\hat{Y}}{\hat{X}}$ é

$$Z_{hi} = w_{hi}\left(-\frac{\hat{Y}}{\hat{X}^2} x_{hi} + \frac{1}{\hat{X}} y_{hi}\right)$$

ou

$$Z_{hi} = w_{hi}\left(\frac{y_{hi} - \hat{R}x_{hi}}{\hat{X}}\right) \quad (18)$$

A.3 Estimação em domínios de análise

A estimação em domínios de análise é feita definindo-se uma variável indicadora para cada elemento da amostra como

$$\delta_{hi}(d) = \begin{cases} 1 & \text{se unidade } i \text{ do estrato } h \text{ pertence ao domínio de interesse } d \\ 0 & \text{caso contrário} \end{cases} \tag{19}$$

O estimador $\hat{Y}(d)$ do total e de sua variância para o domínio d são obtidos substituindo Z_{hi} por

$$Z_{hi}(d) = \delta_{hi}(d) w_{hi} y_{hi} \tag{20}$$

nas fórmulas (8) e (10).

O estimador $\hat{R}(d)$ de uma razão em um domínio de análise d é

$$\hat{R}(d) = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} \delta_{hi}(d) w_{hi} y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} \delta_{hi}(d) w_{hi} x_{hi}} = \frac{\hat{Y}(d)}{\hat{X}(d)} \tag{21}$$

e a variável linearizada para a estimação da variância é

$$Z_{hi}(d) = \frac{\delta_{hi}(d) w_{hi} [y_{hi} - \hat{R}(d) x_{hi}]}{\hat{X}(d)} \tag{22}$$

Para maiores detalhes sobre os estimadores e suas variâncias, sugere-se a leitura de Shah et al (1995), Särndal et al (1992) e Cochran (1977).

Apêndice 2

Tabela A1 - Agrupamentos da CNAE para a PINTEC

CNAE	Código	Descrição
10 + 11 + 13 + 14	100	Indústrias Extrativas
15 (excl. 15.9)	150	Fab. de produtos alimentícios
15.9	159	Fab. de bebidas
16	160	Fab. de Fumo
17	170	Fab. de produtos têxteis
18	180	Confecção de artigos de vestuário
19	190	Preparação de couro e fab. de artefatos de couro e artigos de viagem e calçados
20	200	Fab. de produtos de madeira
21.1	211	Fab. de celulose e outras pastas para fab. de papel e papelão liso
21.2 + 21.3 + 21.4	210	Fab. de embalagens de papel ou papelão e fab. de artefatos diversos de papel
22	220	Edição, impressão
23 excl. 23.2	230	Fab. de coque, elaboração de combustíveis, excl. refino de petróleo
23.2	232	Refino de petróleo
24 excl. 24.5 e 24.6	240	Química, excl. farmacêutica e defensivos
24.5	245	Farmacêutica
24.6	246	Defensivos
25	250	Fab. de artigos de borracha e plásticos
26	260	Minerais não-metálicos
27.1 + 27.2 + 27.3	271	Metalurgia básica e ferrosos
27.4 + 27.5	274	Metalurgia básica e não-ferrosos
28	280	Fab. de produtos de metal, excl. máquinas e equipamentos
29	290	Fab. de máquinas e equipamentos
30	300	Fab. de máquinas para escritório e equipamentos de informática
31	310	Fab. de máquinas, aparelhos e materiais elétricos
32.1	321	Fab. de material eletrônico básico
32.2 + 32.3	322	Fab. de aparelhos de telefonia e TV
33	330	Fab. de equipamentos de instrumentação médico-hospitalares
34.1 + 34.2	341	Fab. de automóveis e caminhões
34.4	344	Fab. de peças e acessórios pra veículos
34.3 + 34.5	343	Fab. de cabines e recondicionamento de motores
35.1 + 35.2 + 35.9	351	Construção e reparação de embarcações e outros
35.3	353	Aeronaves
36.1	361	Fab. de artigos de mobiliário
36.9	369	Fab. de produtos diversos
37	370	Reciclagem

Tabela A2 - Distribuição da população de referência e da amostra da PINTEC nos estratos finais

CNAE	Estrato					TOTAL	
	Não-elegível		Elegível Amostrado		Certo		
	População	Amostra	População	Amostra	População	População	Amostra
100	1 195	63	538	181	38	1 771	282
150	7 206	313	2 190	777	263	9 659	1 353
159	474	23	333	115	34	841	172
160	20	5	25	10	8	53	23
170	1 729	86	1 006	341	112	2 847	539
180	7 767	341	1 754	552	76	9 597	969
190	2 113	105	1 107	373	78	3 298	556
200	3 948	173	998	333	33	4 979	539
211	10	5	9	5	7	26	17
212	665	36	607	198	48	1 320	282
220	2 573	104	826	246	56	3 455	406
230	17	5	102	54	35	154	94
232	0	0	27	10	5	32	15
240	1 216	59	1 052	319	111	2 379	489
245	0	0	490	127	64	554	191
246	0	0	36	12	11	47	23
250	2 490	112	1 608	478	81	4 179	671
260	4 799	206	1 408	435	61	6 268	702
271	184	12	212	81	26	422	119
274	823	37	365	101	29	1 217	167
280	4 203	175	1 787	512	88	6.078	775
290	2 122	97	1 680	473	169	3 971	739
300	52	5	66	22	42	160	69
310	736	37	673	194	71	1 480	302
321	0	0	247	54	21	268	75
322	0	0	257	64	40	297	104
330	306	16	334	92	51	691	159
341	4	4	14	6	16	34	26
343	551	25	256	67	11	818	103
344	365	19	450	151	62	877	232
351	167	10	157	53	11	335	74
353	0	0	56	13	5	61	18
361	3 253	139	1 163	356	43	4 459	538
369	1 249	52	419	116	31	1 699	199
370	110	6	56	15	1	167	22
TOTAL	50 347	2 270	22 308	6 936	1 838	74 493	11 044

Abstract

The sampling design of the Industrial Survey - Technological Innovation (PINTEC) is based on the hypothesis that innovation is a relatively rare phenomenon. Thus, the use of traditional sampling designs, such as stratified random sample (in general, the strata are defined by location, activity and size), may result in samples with very low frequencies of innovative businesses. It is important to identify previously, in the sampling frame, the businesses with higher probabilities of being innovative and increase the sampling fraction for this subset. There is a collection of variables derived from the Annual Industrial Survey and other sources that can be used for this purpose. This article describes the innovation indicators used to stratify the target population of the PINTEC, which take into account the chances of the business being innovative or not. The complete sampling design is also presented.

Key words: rare event; technological innovation; disproportionate allocation

Agradecimentos

As autoras agradecem a Pedro Luis do Nascimento Silva, Sonia Albieri, Zélia Magalhães Bianchini e Magdalena Góes pelas sugestões recebidas durante a realização deste trabalho.

Associação entre séries de mortalidade/morbidade e concentrações de poluentes atmosféricos: uma estratégia para análise estatística

Gleice M. S. Conceição *
Paulo H. N. Saldiva *
Julio M. Singer **

Resumo

Comparamos algumas técnicas estatísticas para avaliar a associação entre poluição atmosférica e indicadores de morbidade e mortalidade em estudos ecológicos longitudinais, enfatizando alguns aspectos críticos na modelagem como a presença de tendência, de sazonalidade, de sobredispersão e a possível existência de autocorrelação entre observações medidas ao longo do tempo. Como exemplo de aplicação, avaliamos a associação entre o número diário de óbitos e as concentrações diárias de poluentes (PM_{10}, SO_2, CO e O_3) na cidade de São Paulo de 1994 a 1997. Também investigamos a existência do efeito de um rodízio de veículos (adotado a partir de 1996) nessa associação. Foram observadas associações dose-dependentes entre mortalidade e os níveis de SO_2 e CO. Os diferentes modelos produziram resultados coerentes, mas aqueles mais sofisticados foram mais sensíveis para detectar efeitos significativos, que, como era de se esperar, foram de pequena magnitude. Nenhum efeito da adoção do rodízio na mortalidade foi identificado.

Palavras-chave: modelos aditivos generalizados, regressão de Poisson, análise estatística, morbidade, mortalidade, poluição atmosférica.

Este estudo foi parcialmente financiado pela FAPESP, CNPq e LIM05-HCFMUSP.

* Endereços para correspondências: Laboratório de Poluição Atmosférica Experimental, Departamento de Patologia, Faculdade de Medicina da Universidade de São Paulo. Av. Dr. Arnaldo 455, São Paulo, SP, 01246-903, Brasil.
** Departamento de Estatística, Instituto de Matemática e Estatística da Universidade de São Paulo. Caixa Postal 66281, São Paulo, SP, 05311-970, Brasil.

1. Introdução

Os piores desastres ocasionados por episódios de altas concentrações de poluentes na atmosfera ocorreram no Vale Meuse em 1930 (Firket, 1931), em Donora, em 1948 (Schrenk et al., 1948) e em Londres, em 1952 (Logan, 1953). Nestes episódios, as concentrações de poluentes atingiram níveis extremamente altos e boa parte das conseqüências à saúde das populações afetadas foi imediata e bastante evidente. Por exemplo, em Donora, quando o episódio de inversão térmica impediu a dispersão de poluentes em outubro de 1948, foram observadas 20 mortes em lugar das duas normalmente esperadas em uma comunidade de apenas 14 000 pessoas. Quase metade da população necessitou de atendimento médico. Em dezembro de 1952 em Londres, no mais clássico dos episódios, uma nuvem de enxofre permaneceu estacionada sobre a cidade por três dias, e num período de poucas semanas registrou-se um excesso de aproximadamente 4 000 mortes em relação ao número médio de óbitos em períodos semelhantes.

Dada a clareza dos fatos, avaliar o efeito da poluição sobre a saúde em situações como as descritas acima não requer ferramentas estatísticas muito sofisticadas e é difícil incorrer em maiores vícios ou erros metodológicos. Os níveis de poluição comumente encontrados nos grandes centros urbanos são muito menores do que aqueles registrados nas grandes catástrofes e, portanto, apresentam maiores dificuldades de percepção e medição. Além disso, fatores como variações sazonais, variações climáticas ou epidemias são os maiores responsáveis pelas mudanças na morbidade e na mortalidade ao longo do tempo (Thurston E Kinney, 1995). Conseqüentemente, os efeitos da poluição são dificilmente evidenciados sem um esforço analítico adequado. Por isso, sob os níveis de poluição usualmente encontrados nas grandes cidades, uma caracterização mais precisa dos seus efeitos sobre a saúde humana exige uma modelagem estatística cuidadosa.

Estudos ecológicos envolvendo coleta cronológica de dados e o ajuste de modelos de regressão têm sido largamente utilizados para avaliar as implicações da exposição à poluição ambiente (Schwartz et al., 1996a, Saldiva et al., 1994, por exemplo). Nesses estudos, uma mesma população é observada ao longo do tempo, de modo que ela atua como seu próprio controle, eliminando assim algumas preocupações metodológicas. As variáveis que caracterizam diferenças entre os indivíduos dentro da população, mas que não sofrem grandes variações ao longo do tempo, deixam de constituir fatores de confundimento e não necessitam ser levadas em conta no estudo; o hábito de fumar ou a presença de alguma doença pulmonar são exemplos típicos. Da mesma forma, variáveis que sofrem variações ao longo do tempo, mas que não estão associadas com a exposição também não precisam ser consideradas como variáveis de confundimento; o índice de massa corpórea é um exemplo. Esse tipo de estudo oferece seus próprios desafios, pois envolve longos períodos de coleta de dados e exige ferramentas estatísticas mais sofisticadas para sua análise. Além disso, fatores ambientais, e mesmo a própria natureza sazonal dos dados podem mascarar seriamente os resultados.

Com os avanços metodológicos e tecnológicos ocorridos nos últimos 10 anos, foi possível desenvolver e implementar computacionalmente uma larga variedade de modelos e técnicas estatísticas, cada uma com suas próprias vantagens e desvantagens. Entretanto, de maneira aparentemente contraditória, a possibilidade de utilização de tantas abordagens estatísticas e as interfaces cada vez mais amigáveis dos pacotes computacionais

armadilha. O ajuste de diferentes modelos pode levar a conclusões discordantes, especialmente se algum modelo é inapropriado para um conjunto de dados específico.

A dificuldade de interpretação dos resultado de alguns modelos também pode acarretar conclusões carregadas de vícios. Cada modelo apresenta um conjunto próprio de suposições que deve ser levado em conta. Por exemplo, o modelo de regressão linear serve para investigar se uma variável reposta Y depende de variáveis explicativas X_1, X_2,..., X_N, mas este tipo de modelo avalia esta dependência apenas sob o aspecto linear, e não leva em conta outras formas de associação.

Mesmo em situações bem controladas, os modelos estatísticos, por mais sofisticados que sejam, requerem suposições que dificilmente correspondem à realidade de modo exato; apesar disso são ferramentas extremamente úteis para resumir e interpretar os dados.

Para obter resultados plausíveis é importante lembrar que a associação que desejamos caracterizar é possivelmente influenciada por uma série de fatores que podem mascará-la ou invalidar as conclusões. Estes fatores podem incluir: a presença de tendência e sazonalidade, variáveis de confusão, defasagem entre causa e efeito, forma da distribuição da variável resposta, multicolinearidade, sobredispersão e a possível existência de autocorrelação entre observações medidas ao longo do tempo. O objetivo deste trabalho é tecer considerações gerais sobre cada um desses fatores e ilustrar possíveis tratamentos estatísticos por intermédio de um exemplo.

Na seção 2, apresentamos pormenores sobre cada um dos fatores identificados acima, sugerindo técnicas estatísticas que podem ser empregadas para sua incorporação na análise. Na seção 3, descrevemos os dados que servirão para ilustrar a análise estatística. Na seção 4, delineamos uma estratégia de análise, cuja concretização é descrita na seção 5. Detalhes metodológicos sobre os modelos estatísticos empregados podem ser encontrados em Hastie e Tibshirani (1990), Conceição et al. (no prelo), Lima et al. (submetido), por exemplo.

2. Fatores a serem considerados nos modelos estatísticos

Tendência e sazonalidade

Uma importante tarefa na análise de longas séries de observações é avaliar a existência de possíveis tendências e/ou padrões sazonais. Em geral, quaisquer duas variáveis que apresentem um comportamento sistemático ao longo do tempo podem estar relacionadas sem que isto implique uma estrutura causal entre elas. Por exemplo, o número de casos de AIDS e a temperatura média no planeta têm aumentado ao longo do últimos anos. Ambos apresentam uma tendência crescente e uma análise estatística simplista poderia mostrar uma associação altamente significante. É prudente remover tendências de longo prazo antes de tentar estabelecer relações causais.

Muitas variáveis costumam apresentar variações sistemáticas ao longo do ano, além de algum tipo de tendência. Estas variações também podem nos levar a identificar uma associação entre elas, mesmo quando essa associação não existe, fazendo-nos superestimar (ou mesmo subestimar) o efeito de algum poluente na saúde.

Por exemplo, sabemos que a ocorrência de doenças respiratórias é mais freqüente no inverno e que em alguns lugares nos Estados Unidos as concentrações de ozônio costumam ser menores nessa época do ano, o que poderia sugerir, erroneamente, que o ozônio tem um efeito protetor em relação às doenças respiratórias. Esse fato mostra que também é necessário remover padrões sazonais antes de investigar a existência de associações causais.

Indicadores de saúde, poluição e variáveis meteorológicas, em geral, apresentam tanto tendência quanto sazonalidade, principalmente quando o período de observação é muito grande (de alguns anos). Para controlar esses fatores, diversas alternativas podem ser utilizadas. Uma abordagem clássica consiste na utilização de variáveis indicadoras para estações do ano (Schwartz e Dockery, 1992; Saldiva et al., 1994, por exemplo), meses do ano ou meses de estudo (Schwartz, 1994a; Saldiva et al., 1995; Pereira et al., 1998; Braga et al., 1999 ; Lin et al., 1999, por exemplo). Na verdade, trata-se de transformar uma variável contínua (o tempo desde o início do estudo) em categorias de amplitude arbitrária (um mês ou três meses, no caso de estações do ano). Uma outra alternativa é incluir, como variáveis explicativas no modelo, não só o número de dias transcorridos (como em Pope et al., 1992), como também o quadrado deste número, o que permite modelar possíveis não linearidades na mortalidade (esta abordagem é válida quando o tempo de seguimento não é muito grande). Outras abordagens envolvem a utilização de termos de seno e co-seno de diferentes freqüências (Katsouyanni et al., 1996) ou o ajuste de curvas alisadas em função do número de dias transcorridos, como em Schwartz et al. (1996a).

É comum observar ainda a presença de sazonalidade de curto prazo, como a de dias da semana. A concentração de poluentes é mais alta durante os dias úteis, quando o número de veículos circulando é maior, do que nos fins de semana. Alguns estudos (Schwartz et al., 1993; Pereira et al., 1998; BRAGA et al., 1999) relatam que o número de atendimentos de emergência, óbitos e internações também é maior durante os dias úteis. Neste caso, o recurso a variáveis indicadoras para dias úteis (ou de uma variável indicadora para fins de semana) pode ser bastante eficiente.

Variáveis de confundimento

Quando a associação entre duas variáveis pode ser afetada por um terceiro fator, dizemos que este é um fator de confundimento. Em nosso caso, estamos interessados em avaliar a associação entre mortalidade e poluição, mas sabemos que ambos podem ser influenciados por fatores meteorológicos (temperatura e umidade). Em geral, a mortalidade é maior na presença de baixas temperaturas, principalmente em se tratando de idosos. O mesmo ocorre com a concentração de poluentes, já que baixas temperaturas desfavorecem sua dispersão. As variáveis meteorológicas devem, portanto, ser consideradas também como variáveis explicativas nos modelos sob investigação.

Como a relação entre a mortalidade e as variáveis meteorológicas não é necessariamente linear, a análise pode envolver a utilização de variáveis indicadoras para categorias de temperatura e umidade (como em Pope et al., 1992), o ajuste de polinômios (Sunyer et al., 1996) e curvas de alisamento (Schwartz, 1994b; Schwartz et al.,

1996a, Lima et al., 2001). Também é conveniente investigar a existência de uma possível interação entre temperatura e umidade.

Defasagem entre causa e efeito

É razoável conjeturar que o efeito das variáveis explicativas (temperatura, umidade, concentração de poluentes, etc.) nos indicadores de saúde não se dá necessariamente no mesmo dia em que ocorre o evento de interesse (óbito ou internação). Ou seja, o número de óbitos e internações ocorridos no dia de hoje pode ser uma conseqüência das condições meteorológicas e da poluição não apenas de hoje, mas também de alguns dias anteriores. Uma ferramenta útil para examinar a contribuição desses fatores em vários dias é a utilização de modelos com defasagens ou médias móveis das variáveis meteorológicas e da concentração de poluentes. Uma questão importante é avaliar quantos dias devem ser levados em conta na modelagem. Médias móveis de variáveis explicativas envolvendo mais de uma semana podem detectar associações espúrias e não são muito razoáveis em epidemiologia (Schwartz et al., 1996b).

Distribuição da variável resposta

Estudos ecológicos realizados com o objetivo de avaliar a associação entre poluição e morbi-mortalidade freqüentemente se valem de contadores diários destes eventos, isto é, utilizam como representantes do prejuízo à saúde o número diário de casos observados de um determinado evento (internações por causas respiratórias, por exemplo). Contagens, em geral, sugerem a utilização de distribuições de Poisson e tais modelos são candidatos naturais para sua análise. Se o número de eventos for grande, modelos Gaussianos podem servir como aproximações. Neste caso, é aconselhável utilizar a transformação logarítmica para a variável resposta, com o objetivo de tentar estabilizar a variância das observações (Sen and Singer, 1993, por exemplo). Em geral, esses modelos convenientemente parametrizados produzem estimativas comparáveis dos parâmetros de interesse. Em ambos os casos, os coeficientes estão relacionados com o logaritmo do risco relativo associado a cada variável explicativa.

Alternativamente, como veremos a seguir, é possível considerar modelos em que a forma da distribuição da variável resposta não é especificada.

Sobredispersão

Uma característica bastante comum de estudos epidemiológicos, onde a população sob investigação apresenta-se dividida em subgrupos (definidos por fatores como classe social, idade, presença de alguma doença de base, nível de exposição, etc.) com diferentes riscos de ocorrência do evento de interesse é ter a variância da variável de interesse maior que sua média. Esse fenômeno chama-se sobredispersão, e nesse contexto, as suposições dos modelos de regressão Poisson normalmente utilizados não são satisfeitas. Uma possível abordagem para contornar este problema é a utilização de métodos de "quasi-verossimilhança" (McCullagh e

Nelder, 1989), que permitem estimar a magnitude da sobredispersão e incorporá-la no modelo. Para isso, pode-se recorrer à metodologia de Equações de Estimação Generalizadas (Zeger e Liang, 1986; Koyama, 1997) segundo a qual podem-se ajustar modelos com parâmetro de sobredispersão, além daqueles que incluem algumas estruturas de dependência entre as observações.

Inter-relação entre as variáveis explicativas

A existência de relações lineares entre as variáveis explicativas é chamada de multicolinearidade, que pode ser tanto inerente às mesmas, quanto fruto das especificações do modelo.

No primeiro caso podemos citar, como exemplo, a relação existente entre as concentrações dos diversos poluentes na atmosfera. A concentração de cada poluente é medida por um aparelho ou técnica específica, mas não é razoável supor que cada um esteja isolado na atmosfera, ou que seus efeitos sobre a mortalidade sejam independentes. Além disso, poluentes gerados basicamente por fontes automotivas, como CO, NO_2 e SO_2 (e mesmo o PM_{10}, pois sua concentração, em grande parte, é resultado da emissão direta ou da ressuspensão ocasionada pelo tráfego de veículos) vão estar correlacionados porque suas concentrações dependem, ao menos em parte, do volume de gases emitidos. A inclusão simultânea de dois ou mais poluentes como variáveis explicativas da mortalidade no modelo de regressão pode gerar multicolinearidade.

No segundo caso, se especificarmos um modelo contendo, como variáveis explicativas, X e X^2 (onde X pode ser a temperatura, por exemplo) e o domínio dos valores de X for pequeno, X e X^2 podem ser altamente correlacionadas, o que também pode resultar em multicolinearidade.

Na presença de multicolinearidade, os coeficientes se tornam muito sensíveis a pequenas mudanças no conjunto de dados e a precisão das estimativas diminui. Isto significa que as estimativas dos coeficientes das variáveis em questão podem ter erros padrão muito grandes, de modo que os coeficientes correspondentes não possam ser considerados significativamente diferentes de zero, induzindo à conclusão (errada) de que essas variáveis não são importantes para explicar a mortalidade.

Autocorrelação

Observações realizadas na mesma população em dias consecutivos podem estar correlacionadas, o que vai contra as suposições da maioria dos modelos clássicos de regressão. Esse tipo de correlação recebe o nome de autocorrelação. Se a correlação entre observações mais próximas no tempo é maior do que entre observações mais distantes, dizemos que existe autocorrelação serial. Se a variável resposta apresentar autocorrelação serial e esta não puder ser explicada ou devidamente controlada no modelo, então os resíduos poderão estar serialmente autocorrelacionados (Schwartz et al., 1996b). Isto pode não acarretar viés nos coeficientes de regressão estimados, mas os erros padrão poderão estar sub ou superestimados.

Alternativas para lidar com este problema incluem o uso de modelos autorregresivos (Wei, 1990) e a utilização de médias móveis. Gráficos de autocorrelação parcial dos resíduos são bons auxiliares no diagnóstico da autocorrelação (Wei, 1990).

Não parece razoável supor que o óbito de um indivíduo ocorrido no dia de hoje interfira diretamente no óbito de outros indivíduos em dias subsequentes. Assim, a autocorrelação no número de óbitos medidos ao longo do tempo deve ser explicada por outros fatores, também ordenados no tempo e que são, de fato, os responsáveis pelas alterações no risco de óbito. Se esses fatores pudessem ser medidos com precisão e colocados no modelo, não haveria autocorrelação nos resíduos. O modelo de Zeger para séries de contagens (Zeger, 1988; Koyama, 1997) incorpora essas características, e tanto a autocorrelação quanto a sobredispersão são incluídas no modelo por intermédio de uma variável não observada; os parâmetros desse modelo são estimados através de Equações de Estimação Generalizadas (Zeger e Liang, 1986). Um elucidativo exemplo envolvendo esta abordagem num outro contexto está apresentado em Campbell (1994).

Alguns autores, como Schwartz et al. (1996b), afirmam que após controlar adequadamente a sazonalidade e a tendência, a magnitude da autocorrelação serial em dados de mortalidade e admissões hospitalares é pequena e as estimativas obtidas não sofrerão grandes alterações com a incorporação de controladores de autocorrelação no modelo.

Nosso objetivo é apresentar um exemplo concreto de avaliação da magnitude da associação entre séries de concentrações de poluentes atmosféricos e de mortalidade, identificando uma estratégia de análise que permita contemplar as características discutidas acima.

Com essa finalidade procuramos quantificar a associação entre a concentração de poluentes atmosféricos e mortalidade na cidade de São Paulo no período de 1994 a 1997, identificando possíveis efeitos da adoção de medidas de restrição parcial de circulação de veículos automotores no município (rodízio).

3. Descrição dos dados

Diferentes subconjuntos dos dados utilizados para este exemplo foram analisados por Conceição et al. (2001) e Conceição et al. (submetido). As séries correspondentes estão disponíveis em www.ime.usp.br/~jmsinger.

Dados de mortalidade

Foram obtidos do Programa de Aprimoramento das Informações de Mortalidade do Município de São Paulo (PRO-AIM) dados referentes a todos os óbitos de residentes na cidade de São Paulo, nos anos de 1994 a 1997. A partir dessas informações, foi calculado o número diário de óbitos por todas as causas durante o período de estudo.

Dados de poluição atmosférica

A Companhia de Tecnologia de Saneamento Ambiental - CETESB - possui uma rede telemétrica com 13 estações medidoras dos níveis de poluição espalhadas pela cidade de São Paulo, que forneceram, para o mesmo período, informações sobre as concentrações de PM_{10} ($\mu g/m^3$), SO_2 ($\mu g/m^3$), CO (ppm), Ozônio ($\mu g/m^3$) e NO_2

($\mu g/m^3$). Nem todas as estações medem todos os poluentes. Os dados relativos à concentração de poluentes são gerados continuamente e inicialmente são calculadas, para cada hora, a média horária e a média para o período das últimas 8 horas. Posteriormente são obtidas as médias diárias para o período de 24 horas (com início às 16 horas do dia anterior e término às 16 horas do dia corrente), definidas segundo os critérios da "Environmental Protection Agency - EPA" da seguinte forma (CETESB, 1997):

- PM_{10} e SO_2: média do período de 24h;
- NO_2 e O_3: maior média horária observada no período de 24h; e
- CO: maior média de 8h observada no período de 24h.

Neste estudo, a medida de concentração diária para cada poluente foi definida como a média das médias diárias correspondentes às estações onde o poluente foi medido. É importante ressaltar que, embora as concentrações de poluentes sejam, em média, diferentes de uma região para outra na cidade, o padrão diário de variação ao longo do tempo é semelhante em todas as regiões. Dessa forma poderemos, sem maiores problemas, utilizar a média de todas as estações como representativa da concentração do poluente na cidade.

O NO_2 não foi medido em mais da metade do período e não será levado em conta na análise.

Informações sobre rodízio

As informações sobre a realização do rodízio estadual foram obtidas da CETESB. Durante o estudo, a operação rodízio foi realizada, de segunda a sexta-feira, nos períodos de 5/8/96 a 30/8/96 e de 23/6/97 a 30/9/97.

Variáveis meteorológicas

Foram obtidas do Instituto Astronômico e Geofísico da Universidade de São Paulo -IAG- USP a temperatura mínima e máxima (°C) e a umidade relativa do ar média e mínima (%) por dia.

Análise descritiva dos dados

Para descrever o comportamento do número de óbitos, apresentamos uma série de gráficos.

A Figura 1 mostra o número de óbitos em função do tempo. A sazonalidade é bastante evidente e os picos de mortalidade correspondem aos meses de inverno. A Figura 2 contém gráficos do tipo "box-plot" para o número de óbitos a cada ano. Na Figura 3 apresentamos o histograma do número de óbitos no período de estudo. A Figura 4 mostra o diagrama de dispersão da média mensal em função da variância mensal do número de óbitos; cada ponto deste gráfico corresponde a um dos 48 meses do período de estudo. O objetivo desses dois gráficos é tentar caracterizar a distribuição da variável de interesse e utilizar esta informação na análise.

Figura 1 - Número de óbitos ao longo do tempo, para o período de janeiro de 1994 a dezembro de 1997

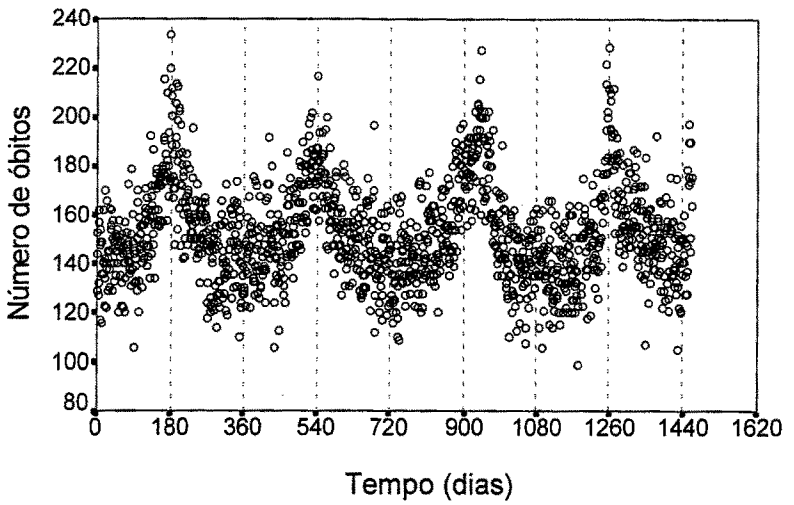
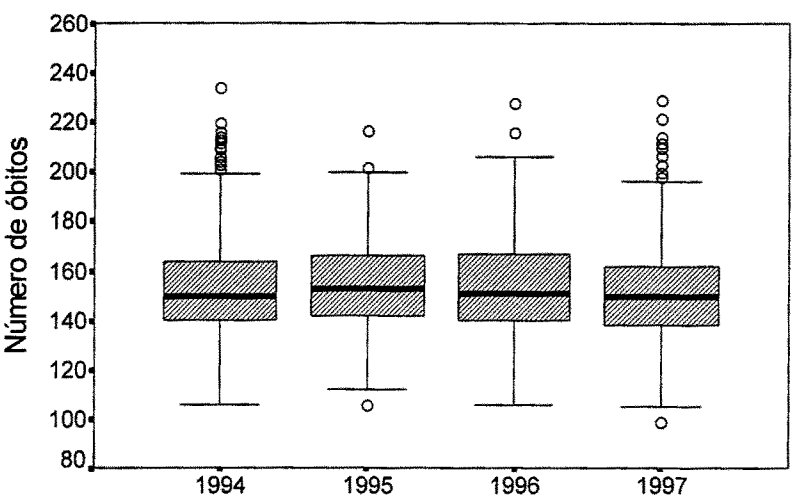


Figura 2 - "Box-plot" do número de óbitos para cada ano



Neste último gráfico, uma curva alisada facilita a visualização da tendência presente nesses dados. Para ajustá-la foi utilizado um algoritmo de regressão ponderada local (*locally weighted running line smoother - loess)*, desenvolvido por Cleveland (1979) e posteriormente descrito por Hastie e Tibshirani (1990) e implementado no software S-plus. Basicamente, para cada ponto (x_i,y_i) do diagrama de dispersão foi definida uma vizinhança contendo aproximadamente 50% do número total de pontos (isto é, 25 pontos: os 12 pontos mais próximos à direita de x_i , os 12 pontos mais próximos à esquerda de x_i , quando possível, além do próprio x_i) e ajustada uma regressão ponderada de y em x ($y_i = \alpha + \beta x_i + e_i$) nessa vizinhança. O valor da curva alisada no ponto x_i é o valor ajustado pela reta de regressão em x_i. Os pesos w_j associados ao j-ésimo ponto da vizinhança

de x_i são atribuídos pela função tri-cúbica $W\left(\frac{|x_i - x_j|}{\Delta(x_i)}\right)$, onde $W(u) = \begin{cases} (1-u^3)^3, \text{para } 0 \le u < 1; \\ 0 \quad \text{caso contrário} \end{cases}$ e

$\Delta(x_i) = \max|x_i - x_j|$, e decrescem à medida que x_j se afasta de x_i.

Na Figura 3, o histograma apresenta uma leve assimetria, com uma cauda mais extensa à direita, sugerindo uma distribuição do tipo Poisson como possível modelo para o processo de geração dos dados. Na Figura 4, pode-se notar que a variância aumenta à medida que a média aumenta, o que reforça a proposta de uma distribuição do tipo Poisson. Na maior parte dos casos, a variância é maior do que a média, caracterizando a existência de uma possível sobredispersão.

Figura 3 - Histograma do número de óbitos, para o período de janeiro de 1994 a dezembro de 1997

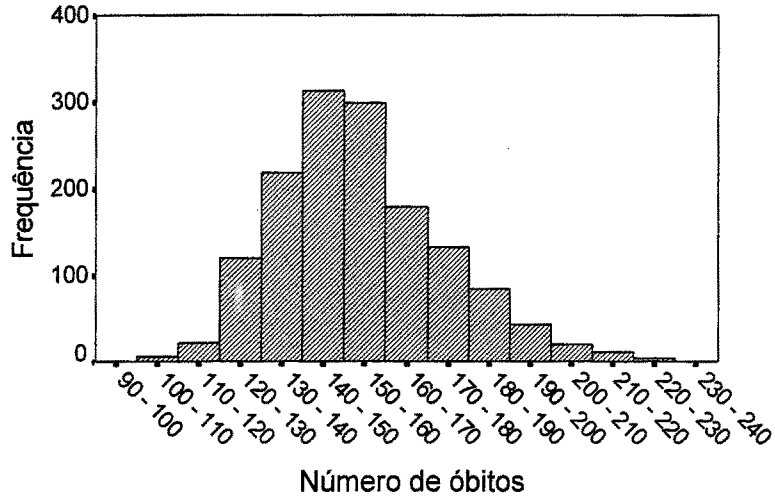
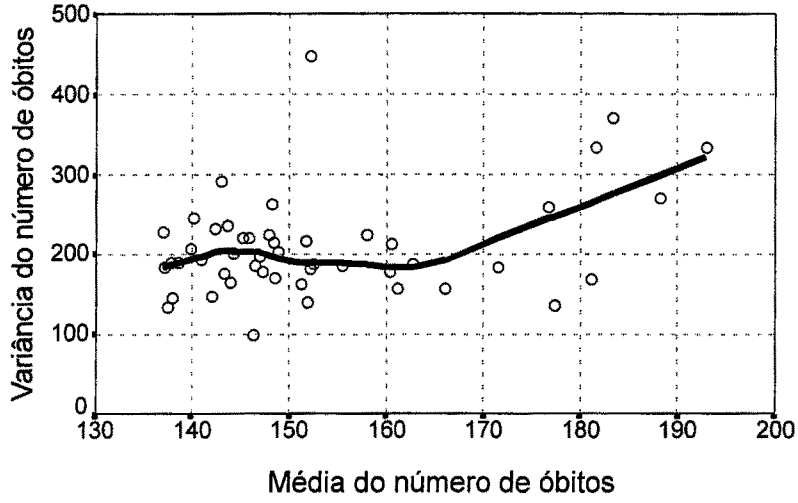


Figura 4 - Diagrama de dispersão da variância mensal em função da média mensal do número de óbitos para o período de janeiro de 1994 a dezembro de 1997



Na Figura 5, apresentamos a concentração diária de cada um dos quatro poluentes medidos ao longo do tempo e, na Figura 6, a temperatura mínima diária e a umidade relativa média diária ao longo do tempo. A curva alisada, em cada gráfico, foi construída de forma semelhante àquela descrita para a Figura 3, utilizando aproximadamente 50% dos pontos em cada vizinhança, e descreve a tendência e a sazonalidade presente nos dados.

Figura 5 - Concentração de poluentes ao longo do tempo, para o período de janeiro de 1994 a dezembro de 1997

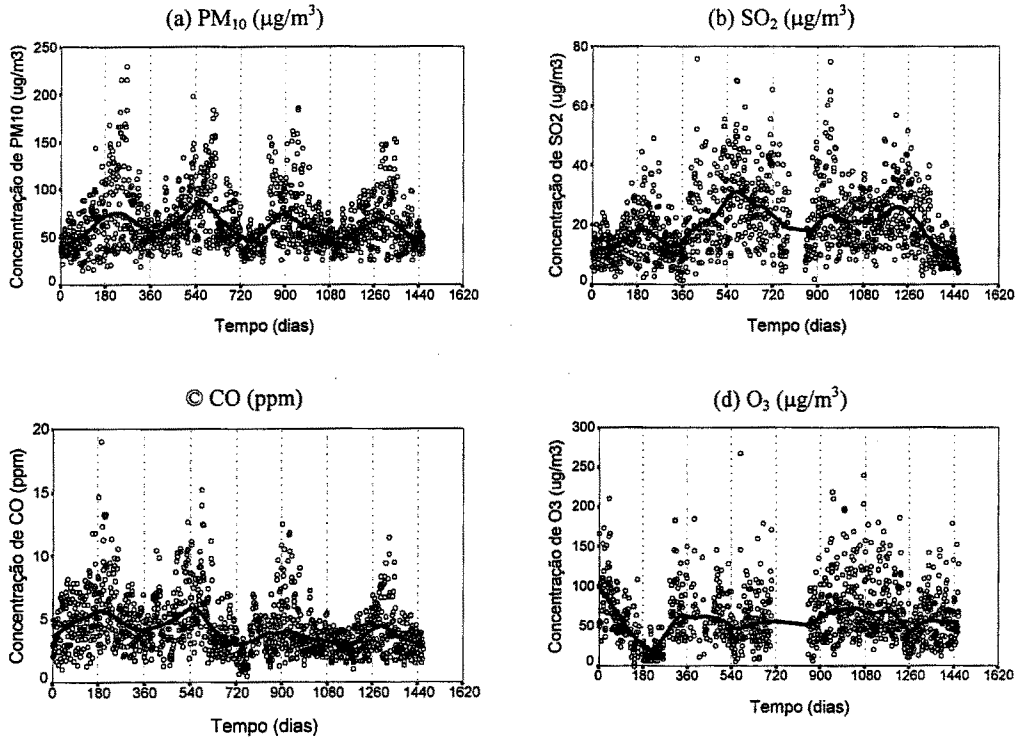
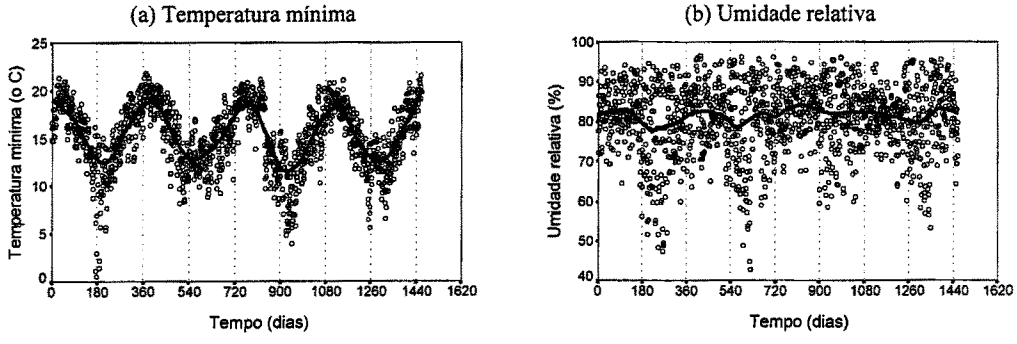


Figura 6 - Variáveis meteorológicas ao longo do tempo, para o período de janeiro de 1994 a dezembro de 1997



De forma semelhante àquela observada para o número de óbitos, existe uma forte sazonalidade na concentração de poluentes e os picos de concentração ocorrem durante o inverno, uma vez que baixas temperaturas desfavorecem a dispersão dos poluentes. A única exceção é o Ozônio, pois a sua concentração está

relacionada à incidência de luz solar, que é mais baixa durante o inverno. Nota-se uma diminuição nas concentrações de PM_{10} e CO nos períodos de inverno de um ano para outro, o que poderia ser uma conseqüência do rodízio.

A Tabela 1 contém medidas descritivas (média, desvio padrão, valores mínimo e máximo) para o número de óbitos, para as concentrações de poluentes e para as variáveis meteorológicas, respectivamente.

Com exceção do ozônio, existe uma alta correlação entre as concentrações dos poluentes, como se pode deduzir dos coeficientes de correlação de Pearson apresentados na Tabela 2.

Tabela 1 - Medidas descritivas para as variáveis em estudo

	1994	1995	1996	1997	total
Número de óbitos					
média	153	154	154	152	153
desvio padrão	21	19	22	21	20
mínimo	106	106	106	99	99
máximo	234	217	228	229	234
Concentração de poluentes					
PM_{10} ($\mu g/m^3$)					
média	67	74	64	60	66
desvio padrão	36	31	30	25	31
mínimo	16	31	24	26	16
máximo	230	198	186	153	230
SO_2 ($\mu g/m^3$)					
média	15	27	23	20	21
desvio padrão	7	13	11	10	12
mínimo	2	4	2	3	2
máximo	49	76	75	57	76
CO (ppm)					
média	5	5	4	4	4
desvio padrão	2	2	2	2	2
mínimo	1	1	1	1	1
máximo	19	15	13	11	19
O_3 ($\mu g/m^3$)					
média	57	61	76	63	64
desvio padrão	39	35	42	33	38
mínimo	0	6	8	12	0
máximo	212	269	240	188	269
Variáveis meteorológicas					
Temperatura mínima (°C)					
média	15	15	15	15	15
desvio padrão	4	3	4	3	3
mínimo	1	8	4	6	1
máximo	21	22	21	22	22
Umidade relativa (%)					
média	80	80	82	81	81
desvio padrão	10	10	8	8	9
mínimo	48	43	59	53	43
máximo	95	97	96	96	97

Tabela 2 - Coeficientes de correlação de Pearson para as concentrações de poluentes

	PM_{10} ($\mu g/m^3$)	SO_2 ($\mu g/m^3$)	CO (ppm)	O_3 ($\mu g/m^3$)
SO_2 ($\mu g/m^3$)	.60*			
CO ($\mu g/m^3$)	.62*	.44*		
O_3 ($\mu g/m^3$)	.13*	.16*	-.12*	
NO_2 ($\mu g/m^3$)	.82*	.64*	.61*	.47*

4. $p<0.01$

4. Estratégia de análise

No processo de modelagem, procuramos identificar os controles mais adequados para tendência e sazonalidade, a estrutura de defasagem que melhor representa uma possível relação entre causa e efeito, o tipo de relação existente entre as variáveis meteorológicas e a mortalidade, além das outras características comentadas na seção 2. A opção entre as diversas alternativas disponíveis para a solução destes problemas e a ordem em que devem ser consideradas no processo de modelagem são questões difíceis. Fundamentando-nos em trabalhos de SCHWARTZ (1994b), POPE e SCHWARTZ (1996), KATSOUYANNI et al. (1996), SCHWARTZ et al. (1996b), definimos uma estratégia para o ajuste dos modelos de regressão cujo objetivo básico é tentar explicar ao máximo a variabilidade na mortalidade por intermédio de controles para tendência, sazonalidade, clima, etc., antes de adicionar as concentrações de poluentes. Os passos para a concretização dessa estratégia são:

1. Incluir no modelo variáveis para controlar a sazonalidade (de curto e longo prazo) e a tendência, quais sejam:
   - variáveis indicadoras para meses e anos ou curvas de alisamento; e
   - variáveis indicadoras para dias da semana

2. Incluir as variáveis meteorológicas (temperatura mínima e umidade relativa média), considerando:
   - termos lineares e/ou quadráticos;
   - curvas de alisamento;
   - possível interação entre temperatura e umidade; e
   - medidas com diferentes defasagens (por exemplo, dia corrente, dois dias anteriores ou média móvel de dois dias).

3. Incluir concentrações de poluentes, considerando:
   - medidas do dia corrente, defasagens de até três dias e médias móveis de dois a sete dias.

4. Avaliar a existência de autocorrelação por meio de:
   - gráficos de resíduos; e

- modelos que levem em conta termos para autocorrelação.

5. Avaliar se o coeficiente estimado da concentração de poluentes é robusto, isto é, se ele é consideravelmente afetado por perturbações nas especificações do modelo. Para isso, ajustar modelos com diferentes conjuntos de variáveis explicativas, além da concentração do poluente em questão.

6. Avaliar se a associação observada entre mortalidade e poluição é dependente da dose e se esta associação é realmente linear ou se, na verdade, existe uma concentração abaixo da qual o efeito da poluição é inócuo. Com essa finalidade, ajustar modelos de regressão contendo categorias da concentração de poluentes em lugar da variável original.

7. Avaliar a existência de um possível efeito de rodízio, através da inclusão de uma variável indicadora para rodízio no modelo. Esta análise deverá ser repetida, considerando-se apenas os meses de inverno dos quatro anos, para garantir que o efeito de rodízio não será confundido com um efeito de inverno.

Nos modelos envolvendo curvas de alisamento, o parâmetro de alisamento deverá ser fixado após a investigação de gráficos de alisamento do número de óbitos com diferentes parâmetros. Lembramos que quando o objetivo é modelar a sazonalidade e uma possível tendência, devemos definir um parâmetro que apreenda apenas o padrão sazonal e a tendência. É importante que o alisador não capture a variabilidade diária do número de óbitos, porque esta pode ser explicada pelas demais variáveis, inclusive por aquelas relacionadas com a poluição.

Para fins comparativos, serão ajustados três tipos de modelo (seguindo os passos descritos anteriormente):

I. Regressão log-linear gaussiana, onde o logaritmo do número de óbitos será modelado como uma função das seguintes variáveis explicativas: indicadores para meses do ano, anos, dias da semana e funções de temperatura e umidade. Os correspondentes modelos log-lineares serão identificados pela sigla MLL.

II. Regressão log-linear de Poisson contendo as mesmas variáveis explicativas do modelo log-linear gaussiano. Esses modelos serão identificados pela sigla MLG.

III. Regressão log-linear de Poisson contendo como variáveis explicativas um alisador do número de óbitos em função do tempo, dias da semana, e alisadores do número de óbitos em função da temperatura e da umidade. Esses modelos serão identificados pela sigla MAG.

A variável resposta nos modelos de regressão será o número de óbitos.

Ao final de cada passo, serão utilizadas medidas de diagnóstico para verificar se o objetivo proposto foi atingido. Assim, após cada passo, deverão ser construídos gráficos de dispersão de resíduos *versus* tempo. No passo 2, modelos com diferentes especificações para temperatura e umidade serão ajustados e comparados. O

critério para escolha do "melhor" conjunto de variáveis meteorológicas e da estrutura de defasagem mais adequada neste caso (dia corrente, anterior ou dois dias) será baseado:

a) no coeficiente de determinação ajustado (R^2 ajustado) que, no caso dos modelos log-lineares gaussianos, dá uma idéia da porcentagem da variabilidade do logaritmo da mortalidade explicada pelas variáveis independentes e corrigida de acordo com o número de variáveis (Neter et al., 1996);
b) no critério de informação de Akaike (Hastie e Tibshirani, 1990), para os modelos de regressão log-linear de Poisson;
c) na significância dos coeficientes das variáveis meteorológicas do modelo em questão; e
d) no nível descritivo do teste que compara o modelo em questão com o modelo sem as variáveis meteorológicas (teste F para modelos log-lineares gaussianos e teste da razão de verossimilhanças para modelos de regressão log-linear de Poisson).

A análise será realizada com o auxílio do software S-Plus 4.0 (Mathsoft, 1997).

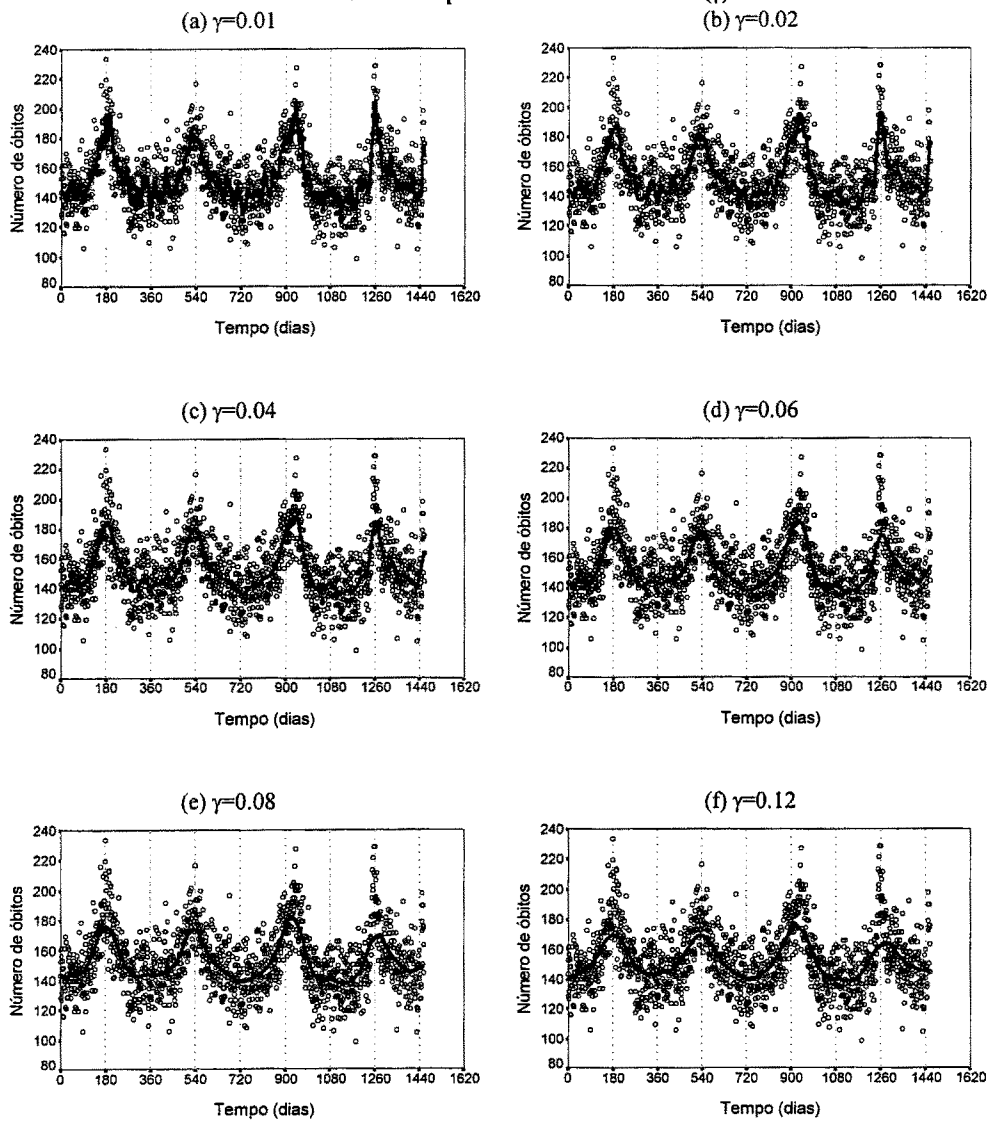
5. Resultados

Calibração do modelo

Diversos parâmetros de alisamento foram testados e com base numa inspeção (subjetiva) dos gráficos apresentados na Figura 7, escolhemos o valor 0.06 (Figura 7-d) para o alisador do número de óbitos em função do tempo. Isto significa que a vizinhança em cada ponto deverá conter 6% das observações, o que corresponde a 90 pontos (aproximadamente 3 meses). O alisador com parâmetro igual a 0.12 (Figura 2-f) não captura a variabilidade de forma satisfatória e aquele com parâmetro igual a 0.01 (Figura 2-a) o faz de forma exagerada, não reproduzindo a sazonalidade adequadamente. Os alisadores com parâmetros 0.04 e 0.06 parecem satisfatórios, apreendem a sazonalidade e a tendência sem capturar a variabilidade diária da mortalidade, que desejamos explicar posteriormente com a incorporação das variáveis meteorológicas e da poluição.

Os parâmetros de alisamento para a temperatura e para a umidade foram fixados em 0.4, porque este valor produz curvas em formato de "U" ou "S", que não evidenciam muita variabilidade, mas são suficientes para representar a associação entre essas variáveis e a mortalidade, sem apresentar a forma rígida de uma função quadrática.

Figura 7 - Curva de alisamento ajustada para o número de óbitos em função do tempo, para diferentes parâmetros de alisamento (γ)



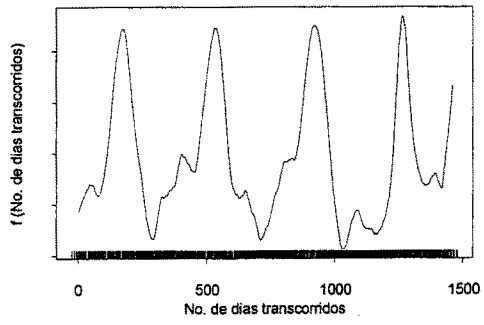
Após o ajuste de uma série de modelos, foram definidos os "melhores" modelos (sem a inclusão de poluentes) segundo os critérios apresentados anteriormente. O "melhor" modelo para cada uma das abordagens consideradas (MLL, MLG e MAG) está apresentado na Tabela 3. Os modelos envolvendo a temperatura do dia anterior e a umidade do dia corrente apresentaram melhor desempenho do que os demais, isto é, tiveram os maiores coeficientes de determinação ajustados (no caso dos MLLs), os menores valores para o critério de informação de Akaike (no caso dos MLGs e dos MAGs) e os menores níveis descritivos para o teste que compara o modelo em questão com o modelo sem as variáveis meteorológicas. Os modelos envolvendo um

termo linear para umidade e uma função quadrática para temperatura (no casos dos MLLs e dos MLGs) ou uma curva de alisamento para a temperatura (no caso dos MAGs) apresentaram melhor desempenho do que os demais.

Como ilustração, apresentamos na Figura 8 as curvas de alisamento ajustadas para o número de óbitos em função do tempo e da temperatura, sob o modelo MAG, cujos parâmetros estimados estão dispostos na Tabela 3.

Figura 8 - Curvas de alisamento estimadas através do modelo MAG da Tabela 3 para modelar a sazonalidade e a relação entre o número de óbitos e a temperatura

(a) tempo



(b) temperatura

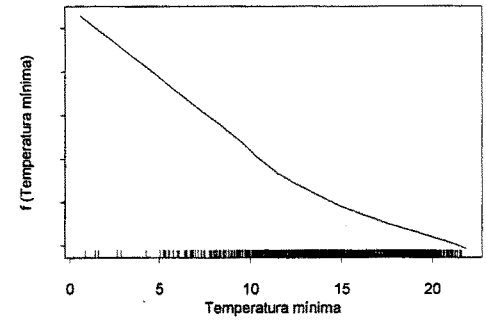


Tabela 3 – Estimativas para os parâmetros dos três modelos, antes da inclusão dos poluentes.

variáveis explicativas	MLL coef	MLL ep	MLG coef	MLG ep	MAG coef	MAG ep
intercepto	5.392**	0.042	5.389**	0.039	5.198**	0.023
meses						
janeiro						
fevereiro	0.021	0.012	0.020	0.013		
março	0.018	0.012	0.018	0.013		
abril	0.039**	0.013	0.039**	0.013		
maio	0.102**	0.014	0.102**	0.014		
junho	0.214**	0.015	0.213**	0.015		
julho	0.161**	0.015	0.162**	0.015		
agosto	0.085**	0.016	0.086**	0.016		
setembro	0.036*	0.015	0.035*	0.015		
outubro	0.004	0.014	0.005	[illegible]		
novembro	-0.003	0.013	-0.002	[illegible]		
dezembro	0.015	0.012	0.016	0.013		
anos						
1994						
1995	0.010	0.007	0.009	0.007		
1996	0.006	0.007	0.008	0.007		
1997	-0.005	0.007	-0.004	0.007		
dias da semana						
domingo						
segunda	0.023*	0.009	0.021*	0.009	0.020*	0.009
terça	0.008	0.009	0.007	0.009	0.005	0.009
quarta	-0.003	0.009	-0.004	0.009	-0.007	0.009
quinta	0.002	0.009	0.000	0.009	-0.001	0.009
sexta	0.008	0.009	0.007	0.009	0.005	0.009
sábado	0.002	0.009	0.001	0.009	0.000	0.009
variáveis meteorológicas						
temperatura - dia anterior	-0.034**	0.005	-0.034**	0.004		
(temperatura - dia anterior)^2	0.001**	0.000	0.001**	0.000		
umidade - dia corrente	-0.002**	0.000	-0.002**	0.000		
f (No. de dias)					-0.126**	0.091
f (temperatura – dia anterior)					-1.391**	0.090
umidade - média móvel 2 dias					-0.002**	0.000
R^2 adj ou AIC	0.48		2014.01		1871.73	
nível descritivo (p) #	0.00		0.00		0.00	

* p<.10; ** p<.05; ***p<.01

\# teste que compara o modelo em questão com o modelo sem as variáveis meteorológicas

Tabela 3 - Estimativas para os parâmetros dos três modelos, antes da inclusão dos poluentes

Resultados obtidos com a inclusão de poluentes

Com base nos modelos obtidos após a calibração, foram ajustados modelos que incluem a concentração de cada um dos quatro poluentes separadamente. Utilizamos a concentração do poluente no dia corrente (ou seja, no dia do óbito), a concentração defasada em até três dias anteriores, além de médias móveis de dois até sete dias.

Na Tabela 4, apresentamos os coeficientes e respectivos erros padrão para cada poluente, quando incluídos separadamente ou em conjunto com os demais, em modelos controlados para sazonalidade e clima. Para SO_2 e CO, utilizamos a média móvel das concentrações de 2 dias. Para uniformizar a análise, utilizaremos esta medida nos passos subseqüentes. Para PM_{10} e o ozônio, que não se mostraram positivamente associados à mortalidade, utilizamos a concentração do dia corrente como variável explicativa.

Tabela 4 - Coeficiente de regressão e respectivo erro padrão para cada poluente, em modelos controlados para sazonalidade e clima, quando incluídos separadamente ou juntos no modelo

	Separados			Juntos		
Poluentes	MLL	MLG	MAG	MLL	MLG	MAG
PM_{10}	<.0001 (.0001)	<.0001 (.0001)	.0001 (.0001)	-.0002 (.0002)	-.0002 (.0002)	<.0001 (.0002)
SO_2	-.0001 (.0003)	<.0001 (.0003)	.0011** (.0003)	.0002 (.0005)	.0003 (.0005)	.0009* (.0004)
CO	.0029 (.0017)	.0026 (.0016)	.0043** (.0014)	.0018 (.0022)	.0017 (.0021)	.0017 (.0019)
O_3	-.0001 (.0001)	-.0001 (.0001)	-.0001 (.0001)	-.0001 (.0001)	-.0001 (.0001)	-.0002 (.0001)

*** p< 0.05 ** p < 0.01.**

De um modo geral, os coeficientes estimados da concentração de poluentes apresentaram maiores valores sob os MAGs; além disso os correspondentes erros padrão foram menores ou iguais àqueles obtidos dos MLLs e dos MLGs.

Sob os MLLs e os MLGs, com todos os poluentes incluídos simultaneamente, nenhum dos poluentes apresentou coeficientes estatisticamente significativos. Sob os MAGs, quando os poluentes são incluídos simultaneamente, apenas o efeito do SO_2 permanece significativo, sugerindo que o efeito correspondente independe da concentração dos demais poluentes.

Na Tabela 5, apresentamos os coeficientes e respectivos erros padrão (entre parênteses) associados ao SO_2 e CO em modelos MAGs com diferentes conjuntos de variáveis explicativas. No último modelo, foram adicionados sete parâmetros auto-regressivos e o modelo foi re-estimado segundo a metodologia de Zeger para séries de contagens (Zeger, 1988). Neste caso, estamos considerando que os resíduos do modelo em um determinado dia podem estar correlacionados com os resíduos de até sete dias anteriores. Em geral, o coeficiente

diminui à medida que novas variáveis são incluídas no modelo, atingindo o menor valor sob o modelo 6, após o controle da autocorrelação e da sobredispersão. Isto sugere que grande parte dos efeitos desses poluentes detectados no modelo 1 deve-se, na verdade, a fatores sazonais e climáticos, e à existência de autocorrelação. Entretanto, os coeficientes continuam significativos mesmo com a inclusão de todas as variáveis. Essa abordagem foi empregada apenas para os MAGs, onde as associações se mostraram mais evidentes.

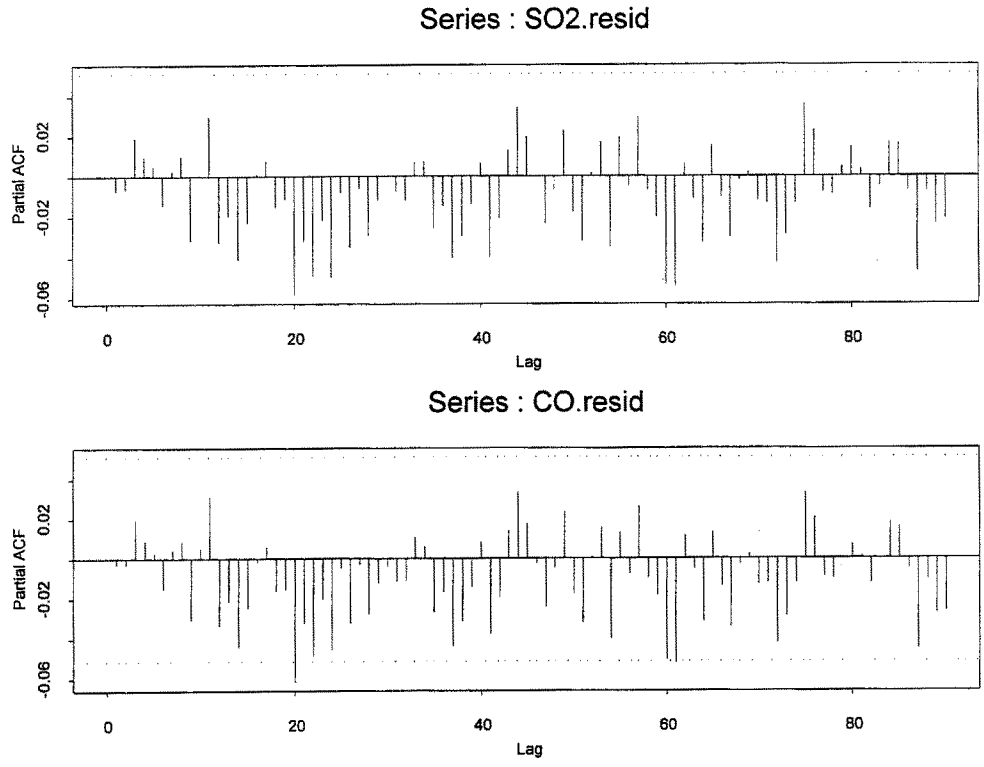
Tabela 5 - Coeficiente e erro padrão para cada poluente em modelos MAGs com diferentes conjuntos de variáveis explicativas

Modelo: variáveis explicativas	SO_2	CO
1: poluente	.0032** (.0003)	.0786** (.0016)
2: 1+ f(tempo)	.0021** (.0002)	.0229** (.0013)
3: 2 + dias da semana	.0025** (.0002)	.0074** (.0013)
4: 3 + temperatura	.0019** (.0003)	.0071** (.0014)
5: 4 + umidade	.0011** (.0003)	.0143** (.0014)
6: 5 + parâmetros autorregressivos	.0010** (0.0003)	.0035* (0.0014)

*** p < 0.05 ** p < 0.01.**

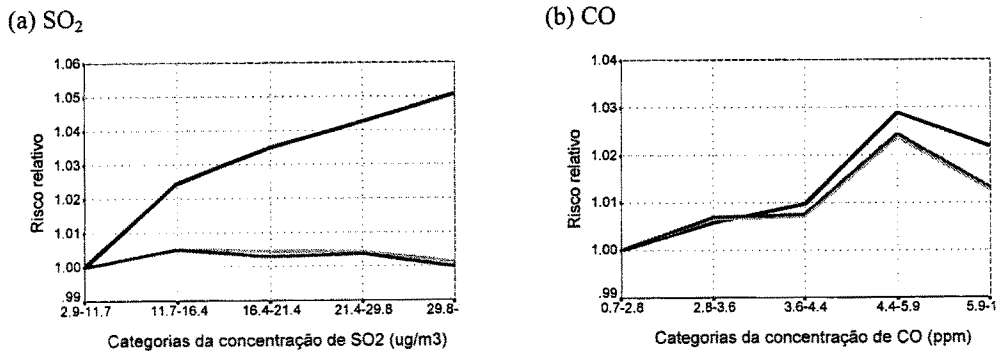
Os parâmetros auto-regressivos com defasagens de até quatro dias foram significativos em todos os modelos ajustados. A Figura 9 apresenta a autocorrelação parcial dos resíduos dos MAGs após o controle da autocorrelação; nota-se que os modelos auto-regressivos foram eficientes nesta tarefa.

Figura 9 - Autocorrelação parcial dos resíduos dos MAGs após a inclusão dos parâmetros auto-regressivos



A Figura 10 apresenta estimativas do risco relativo de mortalidade nas diferentes categorias da concentração de SO_2 e CO (média móvel de dois dias), quando comparadas com a primeira categoria, que representa os dias menos poluídos.

Figura 10 - Risco relativo de mortalidade segundo as concentrações de cada poluente



Os riscos associados aos dois poluentes apresentaram um comportamento tipo dose-resposta sob os MAGs. No caso do CO, este comportamento foi observado também nas duas outras classes de modelos. Entretanto, os riscos obtidos através dos MLLs e MLGs não foram significativos, com raras exceções.

As curvas de riscos relativos estimadas sob os MAGs não apontam para um nível de segurança na concentração destes poluentes. O risco relativo estimado de mortalidade para o SO_2, quando comparamos os dias mais poluídos com os menos poluídos, é aproximadamente 1.05 (IC: 1.03 a 1.07). No caso do CO, o risco está em torno de 1.02 (IC: 1.00 a 1.04)

Utilizando as estimativas obtidas sob os MAGs da Tabela 4 (quando os poluentes são incluídos separadamente no modelo), e considerando os níveis médios de SO_2 e CO (Tabela 1), podemos estimar a porcentagem de mortalidade associada à poluição no período de estudo. Esse valor é de 2% $((e^{0.0011*21} - 1)*100)$ do total de óbitos para o SO_2 e 1.7% para o CO.

Para avaliar seu possível efeito, uma variável indicadora para os períodos de rodízio foi incluída nos modelos da Tabela 3. Os coeficientes estimados (e erros padrão entre parênteses) nos MLLs, MLGs e MAGs foram, respectivamente, -0.0038 (0.0125), -0.0047 (0.0121) e -0.0127 (0.0104). Estes coeficientes não foram significativos sob nenhum dos modelos ajustados. O mesmo ocorreu nos modelos que consideraram apenas os meses de inverno, sugerindo que o rodízio não influenciou o número de óbitos de forma significativa durante o período de estudo.

6. Conclusões

Os modelos apresentaram resultados coerentes, mas a única classe de modelos que permitiu detectar associações significativas entre a concentração de poluentes atmosféricos e mortalidade foi a dos MAGs. Neste caso, a existência de associação entre poluição e mortalidade mostrou-se dependente do modelo empregado na análise estatística. Os resultados sugerem que a utilização de modelos mais sofisticados é necessária para detectar os efeitos dos poluentes, que como era de se esperar, são de pequena magnitude. Observaram-se resíduos autocorrelacionados, mesmo após a inclusão das variáveis para controle da sazonalidade e clima. O controle da autocorrelação e da sobredispersão nos modelos que detectaram associações acarretou uma diminuição dos coeficientes ajustados, mas não mudou drasticamente as conclusões do estudo.

Referências bibliográficas

Braga, A.L.F.; Conceição, G.M.S.; Pereira, L.A.A.; Kishi, H.; Pereira, J.C.R.; Andrade, M.F.; Gonçalves, F.L.T.; Saldiva, P.H.N.; Latorre, M.R.D.O. Air pollution and pediatric respiratory admissions in São Paulo, Brazil. *J. Environ. Med.*, v.1, p. 95-102, 1999.

Campbell, M.J. Times series regression for counts: an investigation into the relationship between sudden infant death syndrome and environmental temperature. *J. R. Statist. Soc.*, v. A157, p. 191-208, 1994.

CETESB. *Relatório anual de qualidade do ar - 1996.* CETESB, São Paulo, 1997.

Cleveland, W.S. Robust locally-weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, v.74, p.829-836, 1979.

Conceição, G.M.S.; Miraglia, S.G.E.K; Kishi, H.S.; Saldiva, P.H.N.; Singer, J.M. Air pollution and children mortality: a time series study in São Paulo, Brazil. *Environ. Health Perspect.*, v.109(suppl 3), p. 347-350, 2001.

Conceição G.M.S, Saldiva P.H.N e Singer J.M. Modelos G.L.M e G.'M para análise de associação entre poluição atmosférica e mortalidade: uma introdução baseada em dados da cidade de São Paulo. *Brazilian J. Epidemiol.*, v.4(3), p.206-219.

Firket, J. Sur les causes des accidents survenus dans la vallée de la Meuse, lors des brouillards de décembre 1930. *Bull. Mem. Acad. R. Med. Belg.*, v. 11, p. 683-741, 1931.

Hastie, T.; Tibshirani, R. *Generalized additive models.* Chapman & Hall, London, 1990.

Katsouyanni, K.; Schwartz, J.; Spix, C.; Touloumi, G.; Zmirou, D.; Zanobetti, A.; Wojtyniak, B.; Vonk, J.M.; Tobias, A.; Pönka, A.; Medina, S.; Bachárová, L.; Anderson, H.R. Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. *J. Epidemiol. Community Health.*, v. 50 (Supl 1), p. S12-S18, 1996.

Koyama, M.A.H. *Modelo de Zeger para análise de séries de contagens.* São Paulo, 1997. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo.

Lima L.P, André CDS e Singer JM. Modelos aditivos generalizados: metodologia e prática. R. Bras. Estat., v.62(217), p.37-69, 2001.

Lin, C.A.; Martins, M.A.; Farhat, S.C.L.; Pope III, C.A.; Conceição, G.M.S.; Anastácio, V.M.; Hatanaka, M.; Andrade, W.C.; Hamaue, W.R.; Böhm, G.M.; Saldiva, P.H.N. Air pollution and respiratory illness of children in São Paulo, Brazil. *Pediatr. Perinat. Epidemiol.*, v.13, p. 475-488, 1999.

Logan, W.P.D. Mortality in London fog incident, 1952. *Lancet*, v. 1, p. 336-338, 1953.

MathSoft. *S-PLUS 4 User's Guide.* Data Analysis Products Division, MathSofth, Inc, Seattle, Washington, 1997.

McCullagh, P.; Nelder, J.A. *Generalized linear models.* 2.ed. Chapman & Hall, London, 1989.

Neter, J.; Kutner, M.H.; Nachtsheim, C.J.; Wasserman, W. *Applied Linear Statistical Models.* 4.ed. Times Mirror Higher Education Group, Inc., Chicago, 1996.

Pereira, L.A.A.; Loomis, D.; Conceição, G.M.S.; Braga, A.L.F.; Arcas, R.M.; Kishi, H.; Singer, J.M.; Böhm, G.M.; Saldiva, P.H.N. Association between air pollution and intrauterine mortality in São Paulo, Brazil. *Environ. Health Perspect.*, v. 106, p. 325-329, 1998.

Pope III, C.A.; Schwartz, J.; Ranson, M.R. Daily mortality and PM_{10} pollution in Utah Valley. *Arch. Environ. Health*, v. 47, p. 211-217, 1992.

Pope III, C.A.; Schwartz, J. Time series for the analysis of pulmonary health data. *Am. J. Respir. Crit. Care Med.*, v. 154, p. S229-S233, 1996.

Saldiva, P.H.N.; Lichtenfels, A.J.F.C.; Paiva, P.S.O.; Barone, I.A.; Martins, M.A.; Massad, E.; Pereira, J.C.R.; Xavier, V.P.; Singer, J.M.; Böhm, G.M. Association between air pollution and mortality due to respiratory diseases in children in São Paulo, Brazil: a preliminar report. *Environ. Res.*, v. 65, p. 218-25, 1994.

Saldiva, P.H.N.; Pope III, C.A.; Schwartz, J.; Dockery, D.W.; Lichtenfels, A.J.F.C.; Salge, J.M.; Barone, I.A.; Böhm, G.M. Air pollution and mortality in elderly people: a time-series study in São Paulo, Brazil. *Arch. Environ. Health,* v. 50, p. 159-63, 1995.

Schrenk, H.H.; Heimamm, H.; Clayton, G.D.; Gafafer, W.M.; Wexler, H. Air pollution in Donora, PA: epidemiology of an unusual smog episode of October 1948. Washington, DC, 1949. *Federal Security Agency, Public Health Bulletin n.306* apud Bascon, R.; Bromberg, P.A.; Costa, D.A.; Devlin, R.; Dockery, D.W.; Frampton, M.W.; Lambert, W.; Samet, J.M.; Speizer, F.E.; Utell, M. Health effects of outdoor pollution. *Am. J. Respir. Crit. Care Med.*, v. 153, p. 3-50, 1996.

Schwartz, J.; Dockery, D.W. Particulate air pollution and daily mortality in Steubenville, Ohio. *Am. J. Epidemiol.*, v. 135, p. 12-19, 1992.

Schwartz, J.; Slater, D.; Larson, T.V.; pierson, w.e.; Koenig, J. Particulate air pollution and hospital emergency room visits for asthma in Seattle. *Am. Rev. Respir. Dis.*, v. 147(4), p. 826-831, 1993.

Schwartz, J. Air pollution and hopital admissions for the elderly in Detroit, Michigan. *Am. J. Respir. Crit. Care Med.*, v. 150(3), p. 648-655, 1994a.

Schwartz, J. Nonparametric smoothing in the analysis of air pollution and respiratory illness. *Can. J. Stat.*, v. 22, p. 471-487, 1994b.

Schwartz, J.; Dockery, D.W.; Neas, L.M. Is daily mortality associated specifically with fine particles? *J. Air Waste Manage. Assoc.*, v. 46, p. 927-939, 1996a.

Schwartz, J.; Spix, C.; Touloumi, G.; Bachárová, L.; Barumamdzadeh, T.; Tertre, A. ; Piekarksi, T.; Leon, A.P.; Pönka, A.; Rossi, G.; Saez, M.; Schouten, J.P. Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *J. Epidemiol. Community Health*, v. 50 (Supl 1), p. S3-S11, 1996b.

Sen, P.K. e Singer, J.M. *Large sample methods in statistics – an introduction with applications.* Chapman & Hall, London, 1993.

Sunyer, J.; Castellsagué, J.; Sáez, M.; Tobias, A.; Antó, J.M. Air pollution and mortality in Barcelona. *J. Epidemiol. Community Health*, v. 50 (Supl 1), p. S76-S80, 1996.

Thurston, G.D.; Kinney, P.L. Air pollution epidemiology: considerations in time-series modeling. *Inhal. Toxicol.*, v. 7, p. 71-83, 1995.

Wei, W.W.S. Time series analysis. *Univariate and Multivariate Methods.* Addison-Wesley Publishing Company, Inc, Redwood City, California, 1990.

Zeger, S.L.; Liang, K.Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, v. 42, 121-130, 1986.

Zeger, S.L. A regression model for times series of counts. *Biometrika*, v.75(4), p.621-629, 1988.

Abstract

We compare some statistical techniques to evaluate the association between atmospheric pollution and morbidity and mortality in longitudinal ecological studies, emphasizing some critical aspects in the modeling strategy, such as long term trends and seasonality, overdispersion and the existence of autocorrelation in measurements over time. As an example, we evaluate the association between the daily number of deaths and the daily pollutant concentrations (PM_{10}, SO_2, CO e O_3) observed in the city of São Paulo during the years of 1994 to 1997. We also investigate the possibility of an effect of the car scheduling program (adopted since 1996) on such association. Dose-dependent associations between mortality and the levels of SO_2 and CO were observed. The different models produced coherent results, but the most sophisticated ones were more sensitive to detect significant effects. No effect of the adoption of the car scheduling program in the mortality was identified.

Key words: generalized additive models, Poisson regression, statistical analysis, morbidity, mortality, air pollution.

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:

planejamento de pesquisas;

avaliação e mensuração de erros em pesquisas;

uso e combinação de fontes alternativas de informação; integração de dados;

novos desenvolvimentos em metodologia de pesquisa;

crítica e imputação de dados;

amostragem e estimação;

disseminação e confiabilidade de dados;

análise de dados;

análise de séries temporais;

modelos e métodos demográficos; e

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 três vias (que não serão devolvidas) para:

Renato Assunção
Editor Responsável
Revista Brasileira de Estatística - RBEs
Av. República do Chile 500, 10º andar
Rio de Janeiro – RJ – 20031-170
Tel.: +55 - 21 - 2142 0472
Fax: +55 - 21 - 2142 0039
E-mail: assuncao@est.ufmg.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 precedidos 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, com 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