REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034-7175

volume 73
número 237
julho / dezembro 2012

IBGE
Instituto Brasileiro de Geografia e Estatística

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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 73 número 237 julho/dezembro 2012

ISSN 0034-7175

R. Bras. Estat., Rio de Janeiro, v. 73, n. 237, p. 1-145, jul./dez. 2012

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 através de divulgação de artigos inéditos tratando de aplicações da Estatística nas mais diversas áreas do conhecimento. Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para a produção e uso de estatísticas públicas.

Os originais para publicação deverão ser submetidos para o site http://rbes.submitcentral.com.br/login.php

Os artigos submetidos à 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.

Editor Responsável
Lúcia Pereira Barroso (IME-USP)

Editoração
Marilene Pereira Piau Câmara - ENCE/IBGE
Dyana Cristina da Silva Braga – ENCE/IBGE

Impressão
Gráfica Digital / Centro de Documentação e Disseminação de Informações - CDDI/IBGE

Capa
Renato J. Aguiar - Coordenação de *Marketing*/CDDI/IBGE

Ilustração da Capa
Marcos Balster - Coordenação de *Marketing*/CDDI/IBGE

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.

Gerência de Biblioteca e Acervos Especiais CDU 31(05)
RJ-IBGE/88-05 (rev.2009) PERIÓDICO

Impresso no Brasil/*Printed in Brazil*

Sumário

Nota da Editora

Este número da RBE de 2012 reúne cinco artigos envolvendo aplicações diversas. O artigo de autoria de Roberto Luis da Silva Carvalho, Lavínia Davis Rangel Pessanha e Eduardo Lima Campos apresenta um modelo multivariado para os gastos familiares mensais com animais de estimação, coletados na Pesquisa Domiciliar sobre Cães e Gatos: Humanização e Padrões de Consumo. No artigo de Wesley Rocha Gripa, Valdério A. Reisen, Fabio A. Fajardo e Neyval Costa Reis Junior modelos de Séries Temporais e de Regressão Linear Múltipla são aplicados na previsão da concentração média de material Particulado Inalável, na cidade de Carapina, Es. O artigo de autoria de Silvia Nagib Elian e Koki Fernando Oikawa apresenta e discute medidas de diagnóstico e análise de influência para o procedimento de Regressão em Cristas, que é geralmente utilizado para contornar problemas de multicolinearidade. Abordam ainda medidas de influência local. No artigo de Andrés Gutiérrez e Jorge Ortiz, escrito em espanhol, os autores implementam três funções em linguagem R sobre a aplicação de critérios de emparelhamento definidos anteriormente por outros autores, facilitando assim o acesso aos dados da Pesquisa Mensal de Emprego a um público mais amplo. O artigo de Natália Santana Paiva e Leonardo Soares Bastos faz uso da inferência bayesiana e implementa o método INLA no ambiente R, demonstrando a utilização de alguns modelos de regressão de Poisson com efeitos aleatórios da detecção de padrões de variação do risco de morbidade hospitalar do SUS e mortalidade para doenças do aparelho respiratório no estado do Rio de Janeiro.

Aproveito a oportunidade para agradecer a colaboração dos Editores Executivos Pedro Luis do Nascimento Silva (ENCE/IBGE) e Mário de Castro Andrade Filho (ICMC-USP), o Editor de estatísticas Oficiais José André de Moura Brito (ENCE/IBGE) e o Editor de Metodologias Fernando Antonio da Silva Moura (UFRJ). Agradeço também aos Editores Associados, aos autores, IBGE, ABE, aos revisores, que anonimamente contribuíram para mais este número da Revista Brasileira de Estatística e a Marilene Pereira Piau Câmara pela editoração.

Tenham uma excelente leitura.

Lúcia Pereira Barroso
Editora Responsável

Análise dos gastos familiares com animais de estimação: aplicação de um modelo de regressão múltipla com resposta univariada

Roberto Luís da Silva Carvalho [1]
Lavínia Davis Rangel Pessanha [2]
Eduardo Lima Campos[3]

Resumo

O presente estudo teve como objetivo investigar os gastos familiares nos setores de higiene, beleza, saúde, alimentação e lazer, destinados aos animais de estimação, nos domicílios particulares permanentes do Grande Méier, de Todos os Santos, do Engenho Novo e de Lins de Vasconcelos, no município do Rio de Janeiro, em 2007. Os dados foram obtidos na Pesquisa domiciliar sobre cães e gatos: humanização e padrões de consumo, do Instituto Brasileiro de Geografia e Estatística. A metodologia foi iniciada com um levantamento bibliográfico, seguido de uma análise exploratória dos dados. Posteriormente, foi ajustado um modelo estatístico para o gasto domiciliar total mensal com animais de estimação. Dentre os resultados, foi verificado que os moradores gastaram em média R$ 149,47 (s = 11,33) por mês com animais de estimação. No modelo ajustado, o vínculo antropomórfico entre proprietários e animais foi identificado através do gasto familiar com animais de estimação.

Palavras-chave: animais de estimação, gastos familiares, regressão linear múltipla.

[1] Universidade Federal do Amazonas - Instituto de Ciências Sociais, Educação e Zootécnica. E-mail: robertoluis.carvalho@gmail.com.
[2] Escola Nacional de Ciências Estatísticas. E-mail: laviniap.pessanha@ibge.gov.br
[3] Escola Nacional de Ciências Estatísticas. E-mail: Eduardo.campos@ibge.gov.br

1. Introdução

Os animais estão presentes no cotidiano das famílias, por serem considerados como companheiros e membros da casa, ou por serem identificados como um recurso utilitário ou econômico, em ações voltadas segurança, controle de pragas ou roedores, ou outras. O consumo de produtos para animais de estimação crescendo na sociedade (OLIVEIRA, 2006). Assim, entender o comportamento e as mudanças nos padrões de consumo das famílias proprietárias de animais de estimação é uma forma de contribuir para o entendimento da vida social contemporânea.

Diversos pesquisadores buscam compreender como as mudanças sociais foram marcadas por influências do consumo. Barbosa e Campbell (2006, pág. 26) afirmam que o consumo é definido como um processo de aquisição de bens e serviços por distintos meios e formas de acesso, e também como uma forma de produzir sentidos e identidades, sendo uma "estratégia" essencial dos grupos sociais para definir estilos de vida.

Diversos fatores são apresentados para a identificação das motivações para o consumo. Para Campbell (2006, pág. 48) o consumo é delimitado por dois aspectos, o primeiro "é o lugar central ocupado pela emoção e pelo desejo" e o segundo pela "individualidade", de modo que ao consumir os indivíduos também constituem suas identidades. Para Miller (2002), os indivíduos não se orientam somente pela relação custo benefício, mas também pelos meios de expressar afeto e construir relações de amor e carinho. Nas pequenas compras cotidianas são levados em consideração sentimentos, compromissos e a responsabilidade com a pessoa e os demais membros da família, de tal modo que mães buscam comprar o que é de melhor para seus filhos em termos de dar qualidade devida ou para satisfazer o desejo dos mesmos; avós buscam satisfazer os desejos dos netos; entre outros.

Diante dos aspectos apresentados algumas hipóteses foram construídas, com base na literatura acadêmica sobre comportamento, consumo e afetividade. Assim, se poderia verificar um consumo por afetividade, como citado por Miller (2002), de tal modo que a partir do consumo e da escolha de produtos o proprietário pode expressar a relação afetiva com seu animal de estimação. Por outro lado, de acordo com Campbell (2006), sendo ato de consumir um meio de expressar necessidades e desejos, é possível propor que o consumo de produtos para animais de estimação seja um modo expressão e realização de desejos e necessidades do proprietário e não aquelas dos animais.

Neste sentido, o presente artigo tem como objetivo investigar os gastos familiares nos setores de higiene, beleza, saúde, alimentação e lazer, destinados aos animais de estimação, nos domicílios particulares permanentes do Grande Méier, de Todos os Santos, do Engenho Novo e de Lins de Vasconcelos, no Município do Rio de Janeiro, no ano de 2007. Inicialmente, foram analisados os gastos familiares com os animais de estimação com higiene, beleza, saúde, alimentação e lazer. Em seguida foram identificados os principais fatores preditores do gasto mensal total domiciliar com os animais por meio de ajuste de um modelo de regressão múltipla.

2. Metodologia

2.1. Participantes do estudo e técnica de amostragem utilizada

Para análise dos gastos familiares foram utilizados os microdados da pesquisa realizada pela Escola Nacional de Ciências Estatísticas do Instituto Brasileiro de Geografia e Estatística (IBGE/ENCE, 2007), intitulada "Pesquisa domiciliar sobre cães e gatos: humanização e padrão de consumo". A pesquisa coletou dados em domicílios particulares permanentes na área do Grande Méier, que corresponde aos bairros do Méier, de Todos os Santos, do Engenho Novo e de Lins de Vasconcelos, no município do Rio de Janeiro, onde as pessoas residentes declararam possuir cães e gatos, no período de 6 e 14 de outubro de 2007.

Os dados da referida pesquisa foram obtidos através de amostragem probabilística com conglomerados em dois estágios. Os desenvolvedores da pesquisa utilizaram a Base Operacional Geográfica do Censo Demográfico de 2000, organizada em setores censitários. Os aglomerados subnormais e o entorno, áreas de difícil acesso e setores de domicílio coletivo foram excluídos, sendo abrangida uma área de 25 setores censitários. As unidades da amostra foram selecionadas em dois estágios. A unidade primária de amostragem - UPA foi o setor censitário e a unidade secundária de amostragem - USA foi domicílio, sendo tomado como informante a pessoa com pelo menos 16 anos de idade, responsável pelo cuidado e/ou gasto com o animal.

No primeiro estágio foram selecionados 25 setores censitários através de seleção sistemática com probabilidade proporcional ao tamanho – PPTSis, sendo que cada setor correspondia a um conglomerado.

No segundo estágio foi utilizada a técnica de amostragem inversa. Esta técnica foi escolhida pelo "não conhecimento de todos os elementos da população, da escassez de tempo e recursos financeiros e, ao mesmo tempo, por possibilitar o cálculo da precisão das estimativas inferidas para a população-alvo" (IBGE/ENCE, 2007).

O dimensionamento da amostra considerou as seguintes premissas (IBGE/ENCE, 2007, p. 24): onde se desejou estimar uma proporção de 6% de uma característica rara da população-alvo com efeito de conglomeração de 1,5; com uma taxa de não entrevista de 0% (devido à amostragem inversa). O total de domicílios na área de interesse foi de 18.313, com 25 setores censitários na amostra. Um intervalo de confiança de 95% com um valor de z padronizado sobre a curva normal de 1,96 foi considerado. Foi selecionada uma amostra de 600 domicílios, sendo 24 por setor censitário, obtendo para tal característica rara um coeficiente de variação de 20%.

2.2. Variáveis do Estudo

As variáveis do estudo foram classificadas em 4 grupos distintos: (1) características domiciliares; (2) características dos animais de estimação; (3) comportamento em relação a cães e gatos e (4) caracterização do padrão de consumo (gastos domiciliares com animais de estimação), e são listadas no Anexo I.

2.3. Estimadores de total, média e razão utilizados

Para estimar o total Y em um plano amostral probabilístico foi utilizado o estimador de Horvitz-Thompson (HORVITZ e THOMPSON, 1952), sendo este um estimador ponderado (π- ponderado). Para calcular este estimador, primeiramente é necessário identificar as probabilidades de inclusão π_i e π_{ij}, que são definidas por:

$$\pi_i = \sum_{i \in s} p(s); \quad i = 1, 2, \ldots, N ,$$

e

$$\pi_{ij} = \sum_{i,j \in s} p(s); \quad i \neq j = 1, 2, \ldots, N$$

sendo π_{ij} a probabilidade de inclusão simultânea, conjunta ou de segunda ordem, associada às unidades i e j, dada por $\pi_{ij} = P[(i \in s) \cap (j \in s)]$ e onde $p(s)$ é definida com a probabilidade de seleção da amostra s no conjunto S de todas as amostras possíveis. Assim, o estimador de total (HORVITZ e THOMPSON, 1952, p. 667; PESSOA e SILVA, 1998) será definido por:

$$\hat{Y}_\pi = \sum_{i \in s} \frac{y_i}{\pi_i},$$

sendo este um estimador linear, podendo ser reescrito da forma

$$\hat{Y}_\pi = \sum_{i \in s} w_i y_i ,$$

onde w_i, com $i = 1, \cdots, n$ é o peso amostral do elemento i selecionado na amostra, e y_i é o valor da variável de interesse para o elemento i. Mais especificamente, w_i é o inverso da probabilidade de seleção, isto é, $w_i = \frac{1}{\pi_i}$.

O estimador de média (HORVITZ e THOMPSON, 1952, p. 670) é definido da mesma forma, mas dividindo o estimador de total por N:

$$\hat{\bar{Y}}_{\pi} = \frac{\hat{Y}_{\pi}}{N} = \sum_{i \in s} \frac{y_i}{N\pi_i} \qquad \text{assim } w_i = \frac{1}{N\pi_i}.$$

O estimador de razão para duas variáveis observadas x e y é dado por:

$$\hat{R} = \frac{\sum_{e \in s} w_i y_i}{\sum_{e \in s} w_i x_i}$$

A probabilidade final de seleção foi definida pelo produto das probabilidades de seleção em cada estágio, aplicadas as correções sobre as recusas devido aos domicílios fechados e não respostas dentro e fora do âmbito da pesquisa. Assim, o peso amostral foi definido por (IBGE/ENCE, 2007, p. 26):

$$w_i = \frac{N}{m \times N_i^*} \times \frac{N_i}{er_i} \times \frac{er_i - 1}{n_1 - 1} \times \frac{n_i}{v_i} \times \frac{d_i}{er_i}$$

sendo $\frac{N}{m \times N_i^*}$ referente ao primeiro estágio e $\frac{N_i}{er_i} \times \frac{er_i - 1}{n_1 - 1} \times \frac{n_i}{v_i} \times \frac{d_i}{er_i}$ referente ao segundo estágio. Onde:

N = total de domicílios da área (18.254 domicílios);

m = número de setores na amostra (25 setores censitários);

N_i^* = número de domicílios no i-ésimo setor (Censo 2000);

N_i = número de domicílios no i- ésimo setor (listagem);

er_i = número de entrevistas realizadas no i- ésimo setor;

n_i = número de domicílios visitados no i- ésimo o setor;

v_i = número de domicílios exceto "recusa" e "fechado" no setor i;

d_i = número de domicílios com cão ou gato no setor i.

Os pesos amostrais para os domicílios variaram de 3,8 a 13,7.

2.4. Estimadores de variância dos estimadores de dados amostrais

A variância do estimador de total é definida por (PESSOA e SILVA, 1998):

$$V(\hat{Y}_\pi) = \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{y_i}{\pi_i} \frac{y_j'}{\pi_j},$$

sendo a probabilidade de inclusão positiva, isto é, $\pi_i > 0, \ \forall \ i$.

O estimador da variância do estimador de total é não viciado (PESSOA e SILVA, 1998) desde que $\pi_{ij} > 0$, $\forall i, j \in U$, sendo definido por

$$\hat{V}(\hat{Y}_\pi) = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j'}{\pi_j}$$

Da mesma forma, para se obter a variância da média populacional, basta dividir o estimador de total pelo fator $\frac{1}{N^2}$, assim:

$$V(\hat{\bar{Y}}_\pi) = \frac{V(\hat{Y}_\pi)}{N^2}$$

Obtendo o seguinte estimador da variância do estimador de total:

$$\hat{V}(\hat{\bar{Y}}_\pi) = \frac{\hat{V}(\hat{Y}_\pi)}{N^2}.$$

A variância do estimador de razão será calculada utilizando a linearização de Taylor. A estimativa pode obtida por

$$\hat{V}(\hat{R}) = \frac{M-m}{Mm(m-1)} \sum_{i=1}^{N} \left(\frac{N_i^{'}}{\bar{x}_i} \right)^2 (\bar{y}_i - \hat{R})^2 + \frac{1}{Mm} \sum_{i=1}^{N} \left(\frac{N_i^{'}}{\bar{x}_i} \right)^2 \frac{N_i^{'} - n_i^{i}}{N_i^{'}} \frac{S_i^{'2}}{n_i^{'}}$$

Onde $N_i^{'}$ é o número de unidades secundárias na $UP_i^{'}$; $n_i^{'}$ é o número de unidades secundárias selecionadas; m = unidades primárias selecionadas e M = quantidade de unidades primárias.

2.5. Intervalos de confiança

Os intervalos de confiança (IC) para os estimadores descritos na Seção 2.3 são baseados nas aproximações assintóticas da distribuição Normal e são dados por (PESSOA e SILVA, 1998):

$$IC\left[\hat{\theta}; \hat{V}_{\pi}(\hat{\theta})\right] = \left[\hat{\theta} \pm z_{\alpha/2} \sqrt{\hat{V}_{\pi}(\hat{\theta})} \right]$$

onde $\hat{\theta}$ é o estimador da estatística (média, total, etc.) da característica da população de interesse e $\hat{V}_{\pi}(\hat{\theta})$ é o estimador de variância correspondente.

2.6. Regressão linear múltipla

O modelo de regressão linear múltipla busca identificar a relação linear da variável dependente *Y* (variável resposta) com as k variáveis x independentes ou regressoras (MONTGOMERY e RUNGER, 2009) disponíveis para análise. Assim o modelo é definido por:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

onde β_j, com j = 0, 1, ..., *k*, são os parâmetros (coeficientes) do modelo e ε é o erro aleatório (ruído).

Assim, para modelar o fenômeno de interesse estimam-se os parâmetros do modelo, com base em uma amostra aleatória de n observações, com a seguinte expressão:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i$$

onde ε_i é o erro aleatório (ruído), com $E(\varepsilon_i) = 0$ e $Var(\varepsilon_i) = \sigma^2$; e $\forall i \neq j$ os erros ε_i e ε_j não são correlacionados, isto é, $Cov(\varepsilon_i, \varepsilon_j) = 0$. A seguir, é apresentado o método de estimação dos parâmetros.

2.7. Estimadores de máxima pseudo-verossimilhança dos parâmetros do modelo

O método de máxima pseudo-verossimilhança (PESSOA e SILVA, 1998) consiste basicamente em incluir os pesos amostrais no processo de inferência, pois a não inclusão destes pode introduzir ou causar vício nas estimativas, ou até mesmo, ocasionar uma má especificação do modelo. Sejam y_i vetores observados das variáveis de pesquisa para o elemento i gerados por vetores aleatórios Y_i, para $i \in U$. Suponha que $Y_1, Y_2, \cdots, Y_n$ são independentes e identicamente distribuídos (IID) com densidade $f(y_i, \theta)$. Com isso, têm-se as funções de verossimilhança e de log-verossimilhança populacionais:

$$l_U(\theta) = f(y_1, \theta) \times f(y_2, \theta) \times \cdots \times f(y_n, \theta) = \prod_{i \in U} f(y_i, \theta)$$

e

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

Da mesma forma, as equações de verossimilhança populacionais correspondentes

$$\sum_{i \in U} u_i(\theta) = 0$$

onde $u_i(\theta) = \dfrac{\partial L(\theta)[f(y_i; \theta)]}{\partial \theta}$ é o vetor $K \times 1$ dos escores do elemento i, para $i \in U$.

Seja $T = \sum_{i \in U} u_i(\theta)$ a soma dos vetores dos escores populacionais, ou seja, um vetor de totais populacionais. O estimador de máxima pseudo-verossimilhança (MPV) é definido pela solução das equações de pseudo-verossimilhança dada por

$$\hat{T} = \sum_{i \in s} w_i u_i(\theta) = 0,$$ onde w_i são os pesos amostrais.

A variância estimada de $\hat{\theta}_\pi$ é calculada por

$$\hat{V}_p(\hat{\theta}_\pi) = \left[\hat{J}(\hat{\theta}_\pi)\right]^{-1} \hat{V}_p\left[\sum_{i\in U} \pi_i^{-1} u_i(\hat{\theta}_\pi)\right]\left[\hat{J}(\hat{\theta}_\pi)\right]^{-1}$$

onde

$$\hat{V}_p\left[\sum_{i\in s} \pi_i^{-1} u_i(\hat{\theta}_\pi)\right] = \sum_{i\in s}\sum_{i\in s} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\left[u_i(\hat{\theta}_\pi)\right]\left[u_i(\hat{\theta}_\pi)\right]'$$

e

$$\hat{J}(\hat{\theta}_\pi) = \left.\frac{\partial \hat{T}(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_\pi} = \sum_{i\in s} \pi_i^{-1} \left.\frac{\partial u_i(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_\pi}$$

Encontrados os estimadores para $\hat{\theta}_\pi$ e $\hat{V}_p(\hat{\theta}_\pi)$ será possível encontrar os intervalos de confiança para os parâmetros citados através da distribuição normal (PESSOA e SILVA, 1998).

2.8. Testes de ajuste do modelo

O modelo para ser considerado adequado deve atender algumas suposições básicas (HAIR JR. et al., 2009): (1) linearidade do fenômeno estudado; (2) variância constante dos termos de erro (homocedasticidade); (3) independência dos termos de erro e (4) normalidade da distribuição dos termos de erro. Para esta análise, foi utilizado o erro de previsão (resíduo) para a variável estatística, definido pela diferença entre os valores observados e os previstos pelo modelo para a variável dependente, para fins de comparação. Geralmente, a utilização desta medida padronizada evita distorções nos resultados.

A análise dos resíduos consiste em verificar o comportamento da distribuição dos mesmos em relação aos valores observados e valores preditos da variável independente. Um método comumente utilizado é a plotagem dos resíduos versus os valores preditos.

Segundo Graubard e Korn (2009), as análises gráficas em pesquisas por amostragem complexa não devem ser realizadas sem a inclusão dos pesos amostrais, pois somente com a inserção destes será possível identificar os pontos de influência na dispersão. Com isso, os pesos amostrais foram incluídos na análise gráfica, de forma que as bolhas fossem proporcionais aos pesos.

Tão importante quanto a análise dos resíduos, os parâmetros do modelo devem ser significativos, isto é, devemos analisar as hipóteses de inclusão desses parâmetros no modelo. Os testes mais utilizados são o teste t de *Student* ou teste F, mas segundo Pessoa e Silva (1998) estes testes utilizam a razão de máxima verossimilhança. Os autores sugerem a utilização da estatística *Wald,* que mede a distância entre a estimativa pontual e o valor hipotético do parâmetro numa métrica definida pela matriz de covariância do estimador.

A estatística *Wald*, para testar a hipótese nula H_0 de um problema linear geral ($H_0 = C\beta = c$), onde C é uma matriz de dimensão $R \times P$ (onde *P* é a quantidade de parâmetros e *R* é a quantidade de parâmetros ajustados no modelo) de posto pleno igual a $R = P - Q$ e c é um vetor $R \times 1$, é definida por:

$$X_W^2 = (C\hat{\beta} - c)'(C\hat{V}(\hat{\beta})C')^{-1}(C\hat{\beta} - c)$$

onde $\hat{\beta}$ e $\hat{V}(\hat{\beta})$ são estimadores de mínimos quadrados ordinários. Com isso, sob a hipótese nula H_0, a distribuição assintótica da estatística X_W^2 é uma distribuição $\chi^2(R)$. No entanto, para utilização desta estatística em planos amostrais complexos deve ser utilizado o estimador de maxíssima pseudo-verossimilhança (MPV) de β, isto é, $\hat{\beta}_\pi$, bem como a matriz de covariância $\hat{V}_P(\hat{\beta}_\pi)$ correspondente ao invés dos estimadores de MQO de β e $V(\beta)$. Assim,

$$X_W^2 / R \sim F(R, v)$$

onde $v = m - H$ **,** que é definido pelo número de unidades primárias de amostragem (UPAs), m, menos o número de estratos considerados no plano amostral para seleção das UPAs, H.

2.9. Procedimento para análise de dados amostrais complexos

Para análise de dados foi utilizado o software SPSS. A análise de dados seguiu os passos abaixo:

(1) Primeiramente, se fez a inclusão do plano e dos pesos amostrais (ENCE/IBGE, 2007) na análise de dados;

(2) Estatísticas descritivas foram construídas para o entendimento das variáveis do estudo e a identificação de *outliers* ou possíveis erros;

(3) Por último, um modelo estatístico foi proposto através da técnica de dependência de análise de regressão múltipla (HAIR JR. et al., 2009) para o gasto domiciliar total com animal de estimação. Na construção do modelo, foram incluídos o plano amostral e os pesos amostrais, como sugerido por Pessoa e Silva (1999), utilizando o método de máxima pseudo-verossimilhança.

Para atender o objetivo do presente estudo de identificar os principais fatores preditores do gasto domiciliar com os animais de estimação foi construído um modelo para o logaritmo neperiano (Ln) do gasto mensal total com animais de estimação igual ao somatório de todos os gastos em função de variáveis explicativas que representam as características domiciliares, as características dos animais de estimação e o comportamento dos proprietários em relação a seus animais de estimação. Os dados originais foram transformados com a aplicação da função Ln, pois a distribuição dos dados não seguia a distribuição normal. As variáveis estudadas estão listadas a seguir nas tabelas 1, 2 e 3.

Tabela 1 – Supostas variáveis preditoras referentes às características domiciliares

Código	Descrição	Tipo de variável
X_1	Tipo de domicílio	Dummy (0 para casa e 1 para apartamento)
X_2	Total de homens no domicílio	Quantitativa discreta
X_3	Total de mulheres no domicílio	Quantitativa discreta
X_4	Ln do rendimento domiciliar	Quantitativa intervalar
X_5	Existe pelo menos uma criança no domicílio	Dummy (1 se o domicílio possuir pelo menos uma criança e 0 caso contrário)
X_6	Existe pelo menos um idoso no domicílio	Dummy (1 se o domicílio possuir pelo menos um idoso no domicílio e 0 caso contrário)
X_7	Tipo de arranjo familiar - unipessoal	Dummy (1 se o domicílio se enquadrar no tipo específico de arranjo familiar e 0 caso contrário)
X_8	Tipo de arranjo familiar – nuclear / monoparental	Dummy (1 se o domicílio se enquadrar no arranjo familiar e 0 caso contrário)
X_9	Tipo de arranjo familiar - nuclear / biparental / sem filhos	Dummy (1 se o domicílio se enquadrar no arranjo familiar e 0 caso contrário)
X_{10}	Tipo de arranjo familiar - nuclear / biparental / com filhos	Dummy (1 se o domicílio se enquadrar no arranjo familiar e 0 caso contrário)
X_{11}	Tipo de arranjo familiar – estendido	Dummy (1 se o domicílio se enquadrar no arranjo familiar e 0 caso contrário)
X_{12}	Tipo de arranjo familiar – composto	Dummy (1 se o domicílio se enquadrar no arranjo familiar e 0 caso contrário)
X_{13}	Sexo do chefe	Dummy (1 se o sexo do chefe é masculino e 0 caso contrário)

As variáveis apresentadas na Tabela 1 foram incluídas na análise, devido à hipótese de que os novos padrões familiares estejam influenciando os gastos com animais de estimação. Esta relação está sendo baseada nos benefícios resultantes da relação homem e animal: benefícios médicos (ALLEN, 2003; BUSSOTTI et al., 2005; CUTT et al., 2007) e benefícios psicológicos e psicoterápicos (ORMEROD, 2005; HARA, 2007; GREGHI et al., 2008). Foi utilizada a composição de arranjos familiares proposta por Arriagada (2001).

A seguir, são apresentadas as supostas variáveis preditoras referentes às caraterísticas dos animais de estimação (Tabela 2).

Tabela 2 – Supostas variáveis preditoras referentes às características dos animais

Código	Descrição	Tipo de variável
X_{14}	Total de cães no domicílio	Quantitativa discreta
X_{15}	Total de gatos no domicílio	Quantitativa discreta
X_{16}	Existência de pelo menos um animal de raça no domicílio	Dummy (1 se possui raça e 0 caso contrário)
X_{17}	Existência de pelo menos um animal de raça com pedigree no domicílio	Dummy (1 se possui pedigree e 0 caso contrário)
X_{18}	Existe no domicílio pelo menos um animal de pequeno porte	Dummy (1 se o animal for pequeno e 0 caso contrário)
X_{19}	Existe no domicílio pelo menos um animal de médio porte	Dummy (1 se o animal for médio e 0 caso contrário)
X_{20}	Existe no domicílio pelo menos um animal de grande porte	Dummy (1 se o animal for grande e 0 caso contrário)
X_{21}	Principal motivo de aquisição do cão e/ou gato – companhia	Dummy (1 se motivo companhia e 0 caso contrário)
X_{22}	Principal motivo de aquisição do cão e/ou gato - diversão / afetividade	Dummy (1 se motivo diversão / afetividade e 0 caso contrário)
X_{23}	Principal motivo de aquisição do cão e/ou gato - status / moda / distinção social	Dummy (1 se motivo status / moda / distinção social e 0 caso contrário)
X_{24}	Principal motivo de aquisição do cão e/ou gato - recomendação médica / terapia / guia	Dummy (1 se motivo recomendação médica / terapia / guia e 0 caso contrário)
X_{25}	Principal motivo de aquisição do cão e/ou gato - reprodução / negócios	Dummy (1 se motivo reprodução / negócios e 0 caso contrário)
X_{26}	Principal motivo de aquisição do cão e/ou gato – segurança ou controle de roedores	Dummy (1 se motivo segurança ou controle de roedores e 0 caso contrário)

As variáveis apresentadas na Tabela 2 foram incluídas na análise devido à hipótese de que características referentes à raça e ao pedigree (CLARK e PAGE, 2009), ao porte do animal e ao motivo de aquisição do animal (CAVANAUGH, LEONARD e SCAMMON, 2008) influenciem os gastos com animais de estimação.

Na Tabela 3 são discriminadas as variáveis referentes ao comportamento dos proprietários de animais.

Tabela 3 – Supostas variáveis preditoras referentes ao comportamento dos proprietários em relação a seus animais de estimação

Código	Descrição	Tipo de variável
X_{27}	Circulação irrestrita no domicílio de pelo menos um cão ou gato - Permissão de circulação de animais	Dummy (1 se o animal possui circulação irrestrita no domicílio e 0 caso contrário)
X_{28}	Existência no domicílio de pelo menos um animal de estimação que utilizou roupas e/ou adornos - uso de vestuário	Dummy (1 se o animal utilizou roupas e/ou adornos e 0 caso contrário)
X_{29}	Existência no domicílio de pelo menos um animal de estimação que utilizou acessórios e/ou brinquedos - uso de acessórios	Dummy (1 se o animal utilizou acessórios e/ou brinquedos e 0 caso contrário)
X_{30}	Existência no domicílio de oferta habitual de guloseimas próprias para animais a pelo menos um animal de estimação – consumo de guloseimas	Dummy (1 se o animal possui guloseimas próprias e 0 caso contrário)

Por fim, as variáveis apresentadas na Tabela 3 foram incluídas na análise, devido à hipótese de que existe uma relação direta do comportamento do proprietário com seu animal de estimação. Esta relação está sendo baseada no princípio de antropomorfismo dos animais e na importância que os membros familiares estão dando aos animais de estimação (ECKSTEIN, 2000, SERPEL 2003; KONECKI 2007, RIDGWAY et al., 2008) que acabam por influenciar os gastos com animais de estimação.

A construção do modelo seguiu a técnica *Backward* de seleção de variável, onde todas as variáveis foram incluídas no modelo e em seguida foram retiradas de acordo com o nível de significância e a estatística *Wald (PESSOA e SILVA*, 1998). No entanto, buscou-se a parcimônia do modelo.

3. Resultados e discussão

Inicialmente, serão apresentadas as médias e desvios padrão dos gastos com lazer, beleza, higiene, saúde, alimentação e gasto total domiciliar com animais de estimação. Os valores estão apresentados na Tabela 4.

Tabela 4 – Valores gastos em média por domicílio, com higiene, beleza, lazer, saúde, alimentação e gasto total com animais de estimação, na área do Grande Méier, no ano de 2007, valores em reais

Variável	$\hat{\bar{X}}$	$s(X)$	$IC(\mu)_{95\%}$		$\hat{N}_i$
			Limite inferior	Limite superior	
Lazer	6,09	0,77	4,50	7,68	4733
Beleza	6,49	0,93	4,57	8,42	4729
Higiene	32,99	3,16	26,46	39,51	4741
Saúde	41,92	4,75	32,12	51,72	4702
Alimentação	63,98	6,49	50,58	77,37	4771
Gasto Total	149,47	11,33	126,09	172,84	4805
Rendimento	3439,89	192,26	3043,08	3836,69	4280

Fonte: Microdados da Pesquisa domiciliar sobre cães e gatos: humanização e padrões de consumo – IBGE / ENCE (2007)

A Tabela 4 mostra que os moradores dos domicílios da região do Grande Méier, em 2007, gastaram em média 149,47 reais com os animais de estimação. Mais especificamente, o gasto médio com alimentação foi de 63,98 reais, sendo este considerado o principal gasto com animais de estimação; com saúde o gasto médio obtido foi de 41,92 reais; com higiene o gasto médio observado foi de 32,99 reais; com beleza 6,49 reais e com lazer 6,09 reais. Para esta análise foram desconsiderados os dados de três domicílios.

Com o intuito de identificar a proporção dos gastos com cada item em relação ao gasto total foi construído o estimador de razão para cada gasto médio com animal de estimação em relação ao gasto total. Para esta análise, foi considerado um total estimado de 4548 domicílios. Assim a seguir na Tabela 5, são apresentadas estas estimativas:

Tabela 5 – Estimadores de razão dos valores dos gastos domiciliares com higiene, beleza, lazer, saúde, alimentação em relação ao gasto total com animais de estimação, na área do Grande Méier, em 2007

Modalidade de gasto	$\hat{r}_i$	$s(\hat{r}_i)$	$IC(\hat{r}_i)_{95\%}$	
			Limite inferior	Limite superior
Higiene	0,215	0,017	0,179	0,251
Beleza	0,042	0,005	0,033	0,52
Lazer	0,040	0,004	0,032	0,048
Saúde	0,275	0,024	0,226	0,324
Alimentação	0,427	0,025	0,377	0,478

Fonte: Microdados da Pesquisa domiciliar sobre cães e gatos: humanização e padrões de consumo – IBGE / ENCE (2007)

De acordo com os estimadores de razão, os gastos com alimentação representam 42,7% do valor gasto, os gastos com saúde representam 27,5%, os gastos com higiene representam 21,5%, os gastos com beleza representam 4,2% e lazer, 4,0%.

3.1. Modelo de regressão múltipla para o gasto total mensal domiciliar com animais de estimação

A ordem de retirada das variáveis é apresentada no ANEXO 2. A seguir, são apresentados os valores estimados dos parâmetros (Tabela 6) do modelo para Ln do gasto total com animais de estimação, seus respectivos intervalos de confiança, as estatísticas Wald e os p-valores do modelo ajustado (Wald F(9, 16) = 42,696; p = 0,000). O R2 encontrado foi de 37,3%.

Tabela 6 – Estimativas para o modelo Ln do gasto total mensal domiciliar com animais de estimação

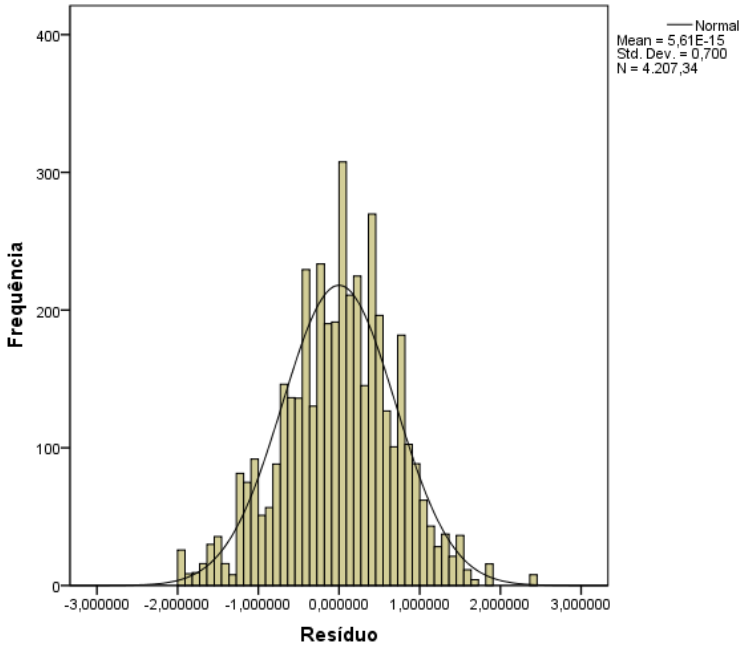
Parâmetros	Estimativas	Desvio padrão	$IC_{95\%}(\beta_i)$		Wald	p
			Mínimo	Máximo		
Intercepto	1,149	0,471	0,176	2,121	5,941	0,023
Total de cães no domicílio	0,257	0,020	0,216	0,298	166,262	0,000
Total de gatos no domicílio	0,167	0,033	0,098	0,237	25,019	0,000
O pet usou um acessório pelo menos uma vez	0,358	0,093	0,165	0,550	14,743	0,001
O pet ganhou guloseimas pelo menos uma vez	0,233	0,071	0,086	0,380	10,684	0,003
Há pelo menos um animal com pedigree no domicílio	0,347	0,106	0,129	0,565	10,785	0,003
Arranjo familiar: unipessoal	0,321	0,146	0,020	0,621	4,849	0,038
Arranjo familiar: monoparental	0,238	0,074	0,085	0,391	10,344	0,004
Ln da renda mensal domiciliar	0,328	0,053	0,218	0,438	37,957	0,000
Motivo de aquisição do pet: reprodução	-0,616	0,203	-1,036	-0,197	9,182	0,006

Assim, temos o seguinte modelo ajustado:

Ln do gasto total mensal domiciliar com animais de estimação = 1,149 + 0,257 (total de cães no domicílio) + 0,167 (total de gatos no domicílio) + 0,358 (o *pet* já usou acessórios) + 0,233 (o *pet* já ganhou guloseimas) + 0,347 (há pelo menos um animal com pedigree no domicílio) + 0,321 (arranjo familiar: unipessoal) + 0,238 (arranjo familiar: monoparental) + 0,328 (Ln da renda mensal domiciliar) – 0,616 (motivo de aquisição do *pet*: reprodução).

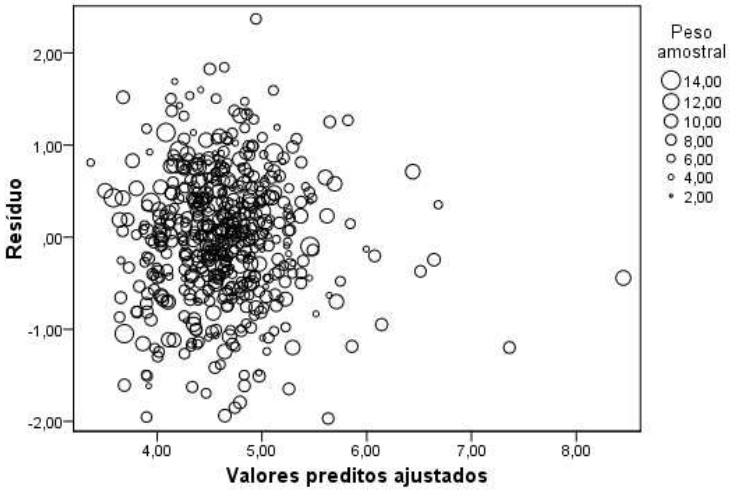
Para verificar o ajuste do modelo proposto, a seguir serão apresentados os gráficos: histograma dos resíduos (Gráfico 1) e o diagrama de dispersão dos resíduos padronizados em relação aos valores preditos ajustados (Gráfico 2).

Gráfico 1 – Histograma dos resíduos do modelo ajustado do Ln do gasto total mensal domiciliar com animais de estimação



Fonte: Microdados da Pesquisa domiciliar sobre cães e gatos: humanização e padrões de consumo – IBGE / ENCE / CDHP (2007)

Gráfico 2 - Diagrama de dispersão dos resíduos em relação aos valores preditos ajustados do modelo Ln do gasto total mensal domiciliar com animais de estimação



Fonte: Microdados da Pesquisa domiciliar sobre cães e gatos: humanização e padrões de consumo – IBGE / ENCE / CDHP (2007)

De acordo com a análise dos gráficos 1 e 2, o modelo em questão atende as premissas básicas de um modelo de regressão linear múltipla, isto é, os erros seguem a distribuição normal e estão aleatoriamente distribuídos.

A seguir, são apresentados os resultados analisados por categorias de variáveis supostamente preditoras.

O 'Ln do gasto mensal total domiciliar com animais de estimação' está relacionado positivamente com as variáveis referentes às características domiciliares, como 'Ln da renda mensal domiciliar' (+ 0,328) e com os arranjos familiares: (1) 'unipessoal' (+ 0,321) e (2) 'nuclear / monoparental' (+ 0,238).

Da mesma forma, o 'Ln do gasto mensal total domiciliar com animais de estimação' recebe maior influência do arranjo familiar 'unipessoal' do que o arranjo 'nuclear / monoparental', pois seus coeficientes acrescentam 0,321 e 0,238, respectivamente, e consequentemente, estes são mais influentes que os demais arranjos familiares. Assim, é possível que as variáveis afetivas, inerentes à solidão, venham a influenciar esta relação, pois como sugerido por Dotson e Hyatt (2008) os animais de estimação acabam por satisfazer as necessidades humanas de companhia, amizade, amor incondicional e afeto. E da mesma forma, como sugere Serpell (2003), os

proprietários vêm os animais como fontes alternativas de apoio social e, de um modo geral, os animais se apresentam como um meio de obtenção de benefícios emocionais e físicos. Outro ponto é que essas relações aumentam de intensidade quando os proprietários estão vulneráveis emocionalmente (EL-ALAYLI et al., 2006) e pode de ser ampliada ainda mais quando o integrante da família tratar seu animal com um "filho" ou "membro da família", como citado por Cohen (2002).

A relação existente com a 'renda mensal domiciliar' vai ao encontro dos resultados de diversos estudos de consumo, que sugerem uma relação do aumento do poder econômico com a ampliação do acesso a bens e serviços (SLATER, 2002).

O 'gasto total' está relacionado com as variáveis referentes às características dos animais, como 'total de cães no domicílio' (+ 0,257), 'total de gatos no domicílio' (+ 0,167) e 'existência de pelo menos um animal com pedigree no domicílio' (+ 0.347). De certo modo, a relação de crescimento proporcional do 'gasto total' com o número de cães e gatos já era esperada, devido ser uma característica da criação dos animais. Mas, este modelo corrobora a afirmação de Clark e Page (2009) que existe relação entre o gasto com animais de estimação e existência de pedigree. Segundo Oliveira (2006, p.85) "as raças puras e o pedigree, não apenas caracterizam um cão, mas simbolizam as características de seus proprietários, através de sua beleza, qualidade, afeto, indicando traços de sua personalidade social". Neste sentido, a relação existente com o fato de no 'domicílio possuir um animal com pedigree' e o aumento no 'gasto domiciliar', reforça a definição do status em função de produtos e imagens com cacterísticas e significações relevantes.

Seguindo a análise do modelo, foi observado que o 'Ln gasto total' está relacionado positivamente com as variáveis que representam o comportamento em relação a cães e gatos: '*O pet* já usou acessórios pelo menos uma vez' (+ 0,358) e o '*pet* ganhou pelo menos uma vez guloseimas' (+ 0,233). A inclusão destas variáveis confirma a relação existente entre o comportamento do proprietário e o gasto com o animal de estimação sugerido por diversos autores (ECKSTEIN, 2000, SERPELL 2003; KONECKI 2007, RIDGWAY et al., 2008). Nos termos, da relação do consumo por afetividade proposta por Miller (2002), é possível que os proprietários busquem dar o que acreditam que é o melhor para seus animais de estimação.

A última variável ‘adquiriu o animal por motivo de reprodução’ apresentou uma relação negativa com o Ln do gasto total (– 0,616). Este fato está associado à relação o animal como um recurso econômico ou utilitário (KONECKI, 2007) e reforça a relação existente entre as variáveis afetivas e o consumo de produtos para animais de estimação, reforçando a influência do antropomorfismo no valor gasto com os produtos para animais de estimação.

4. Conclusão

De um modo geral, nos estudos sobre consumo, se busca compreender as particularidades dos agentes envolvidos no processo, bem como os fatores inerentes para tal ação. O estudo é uma fonte de informação para os estudos sobre o comportamento padrão de gastos dos proprietários de animais de estimação, dada a concepção da amostra e construção da pesquisa de caráter domiciliar, apesar de ter sido realizado com dados de uma área delimitada ao Grande Méier, no Rio de Janeiro.

Os resultados indicam o consumo por afetividade através das relações identificadas nos gastos domiciliares, que diferenciam aqueles proprietários que oferecem guloseimas ou compram acessórios para seus animais.

A existência de raça sem pedigree é um fator não preponderante para ampliar o gasto com animal de estimação, mas nos casos em que os proprietários registraram o pedigree o valor do gasto total se amplia. Este resultado converge com as conclusões do estudo realizado por Clark e Page (2009). Estes proprietários estão inclusos do num grupo diferenciado por status social, pois ao possuem um animal com pedigree certificado e gastam mais com os animais do que os demais.

Os valores gastos com os animais de estimação são maiores nos arranjos domiciliares de tipo unipessoal e nuclear/monoparental, representando assim uma ordem inversa com o tamanho do domicílio, isto é, um domicílio com número menor de indivíduos gasta mais com animais. A relação entre o tipo de arranjo domiciliar e o gasto com animais de estimação sugere a busca dos benefícios médicos (ALLEN, 2003; BUSSOTTI et al., 2005; CUTT et al., 2007) e benefícios psicológicos e psicoterápicos (ORMEROD, 2005; HARA, 2007; GREGHI et al., 2008) pelos proprietários.

Por outro lado, foi observada a relação positiva da renda domiciliar com os gastos com animais de estimação, evidenciando a relação do poder econômico domiciliar com a ampliação do acesso a bens e serviços.

Dentre as limitações do estudo, o coeficiente de determinação do modelo estatístico ajustado (37,3%) foi considerado baixo, indicando que outras variáveis devem ser incluídas em análises futuras.

Outra limitação observada foi a aferição das variáveis afetivas. Sendo assim, é recomendável que os próximos estudos a incluam de medidas de atitude em relação aos animais, entre outras variáveis.

A diversificação da amostra da pesquisa, no que tange a localidade e as características urbana e rural do domicílio é recomendada com o intuito de ampliar a avaliação da influência do antropomorfismo no cuidado e gasto com animais.

O estudo permitiu conhecer um pouco mais do perfil dos proprietários/consumidores de animais de estimação, corroborando a importância da inclusão de variáveis adequadas ao estudo do perfil do consumidor.

Referências bibliográficas

ALLEN, K. Are Pets a Healthy Pleasure? The Influence of Pets on Blood Pressure. American Psychological Society, v. 12, n. 6, p.236-239, 2003.

ARRIAGADA, I. Familias latinoamericanas. Diagnóstico y políticas públicas en los inicios del nuevo siglo. Naciones unidas / División de Desarrollo Social / CEPAL - SERIE Políticas sociales, n. 57, p.1-55, 2001.

BARBOSA, L.; CAMPBELL, C. O estudo do consumo nas ciências contemporâneas. In: Cultura, consumo e identidade. Organizadores Lívia Barbosa, Colin Campbell, Rio de Janeiro: Editora FGV, 2006.

BUSSOTTI, E. A.; LEÃO, E. R.; CHIMENTÃO, D. M. N.; SILVA, C. P. R. Assistência individualizada: "Posso trazer meu cachorro?" Revista Escola de Enfermagem – USP, v. 39, n. 2, p.195-201, 2005.

CAMPBELL, C. Eu compro, logo sei que existo: as bases metafísicas do consumo moderno. In: Cultura, consumo e identidade. Organizadores Lívia Barbosa, Colin Campbell, Rio de Janeiro: Editora FGV, 2006.

CAVANAUGH, L. A.; LEONARD, H. A.; SCAMMON, D. L.. A tail of two personalities: How canine companions shape relationships and well-being. Journal of Business Research, v. 61, n.5, p. 469–479, 2008.

CLARK, P. W.; PAGE, J. B. Examining Role Model and Information Source Influence on Breed Loyalty: Implications in Four Important Product Categories. Journal of Management and Marketing Research, v. 2, n. 1, p. 1-14, 2009.

COHEN, S. P. Can Pets Function as Family Members? Western Journal of Nursing Research, v. 24, n. 6, p. 621-638, 2002.

CUTT, H.; GILES-CORTI, B.; KNUIMAN, M.; BURKE, V. Dog ownership, health and physical activity: A critical review of the literature. Health & Place, v. 13, n. 1, p. 261–272, 2007.

DOTSON, M. J.; HYATT, E. M. Understanding dog–human companionship. Journal of Business Research, v. 61, n. 5, p. 457–466, 2008.

ECKSTEIN, D. The Pet Relationship Impact Inventory. The family journal: counseling and therapy for Couples and families, v. 8, n. 2, p. 192-198, 2000.

EL-ALAYLI, A.; LYSTAD, A. L.; WEBB, S. R., HOLLINGSWORTH, S. L.; CIOLLI, J. L. Reigning Cats and Dogs: A Pet-Enhancement Bias and Its Link to Pet Attachment, Pet–Self Similarity, Self-Enhancement, and Well-Being. Basic and Applied Social Psychology, v. 28, n. 2, p. 131–143, 2006.

GRAUBARD, B. I.; KORN, E. L. Scatterplots with Survey Data. In. Sample Surveys: Inference and Analysis (Ed. PFEFFERMANN, D.; RAO, C. R.). Handbook of statistics, v. 29B, p. 397-422, 2009.

GREGHI, G. F.; MARTINS, M. F.; SILVA, M. R.; SANCHES, Y. C.; POZZOBOM, N. M. Estudo da percepção da auto qualidade de vida e bem-estar em idosos proprietários de animais. In.: 35º Congresso Brasileiro de Medicina Veterinária, Gramado / Rio Grande do Sul, 2008. Anais..., Gramado / Rio Grande do Sul, 2008. p.1-6. Disponível em <http://www.sovergs.com.br/conbravet2008/anais/cd/lista_area_23.htm>. Acesso em 26/07/2010.

HAIR JR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L.. Análise multivariada de dados. Tradução Adonai Schlup Sant'ana e Anselmo Chaves Neto, 6a. edição. – Porto Alegre: Bookmam, 688p, 2009.

HARA, S. Managing the dyad between independence and dependence: case studies of the american elderly and their lives with pets. International Journal of Japanese Sociology, 2007, v. 16, n. 1, p. 100-114.

HORVITZ, D. G.; THOMPSON, D. J. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, v. 47, n. 260, p. 663- 685, 1952. Disponível em <http://www.jstor.org/stable/2280784>. Acesso em 04/09/2010.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍTICA / ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS. Pesquisa domiciliar sobre cães e gatos: humanização e padrões de consumo. Relatório de pesquisa. Rio de Janeiro, IBGE/ENCE/CDHP, 2007.

KONECKI, K. T. Pets of Konrad Lorenz. Theorizing in the social world of pet owners. Qualitative Sociology Review, Volume 3, n. 1, p. 110-127, 2007.

MILLER, D. Teoria das compras: o que orienta as compras dos consumidores. São Paulo: Nobel, 2002.

MONTGOMERY, D. C.; RUNGER, G. C. Estatística aplicada e probabilidade para engenheiros. Tradução: Verônica Calado. 4ª. Edição – Rio de Janeiro: LTC, 2009.

OLIVEIRA, S. B. C. Sobre homens e cães: um estudo antropológico sobre afetividade, consumo e distinção. Rio de Janeiro, 2006. Dissertação (Mestrado em Sociologia e Antropologia). IFCS/PPGSA, Universidade Federal do Rio de Janeiro, 2006.

ORMEROD, E. Companion animals. Working with Older People, v. 9, n. 3, p. 23-27, 2005.

PESSOA, D. G. C.; SILVA, P. L. N. Análise de dados amostrais complexos. In.: Simpósio Nacional de Análise de dados, Probabilidade e Estatística. Anais..., Associação Brasileira de Estatística, Caxambu, 1998, p.170.

RIDGWAY, N. M.; KUKAR-KINNEY, M.; MONROE, K. B.; CHAMBERLIN, E.. Does excessive buying for self relate to spending on pets? Journal of Business Research, v. 61, n. 5, p. 392–396, 2008.

SERPELL, J. A. Anthropomorphism and Anthropomorphic Selection—Beyond the "Cute Response". Society & Animals, v. 11, n. 1, p. 83-100, 2003.

SLATER, D. Cultura do consumo e modernidade. Tradução Dinah de Abreu Azevedo. São Paulo: Nobel, 2002.

ANEXO I - Variáveis do Estudo

Variáveis referentes às características domiciliares:

- Tipo de domicílio: casa, apartamento e outros;
- Total de homens no domicílio;
- Total de mulheres no domicílio;
- Rendimento domiciliar;
- O domicílio possui pelo menos uma criança;
- O domicílio possui pelo menos um idoso;
- Sexo do chefe;
- Tipo de arranjo familiar: unipessoal, famílias nucleares, famílias estendidas e famílias compostas.

Variáveis referentes às características dos animais:

- Total de cães no domicílio;
- Total de gatos no domicílio;
- Existência de pelo menos um animal com raça no domicílio;
- Existência de pelo menos um animal com pedigree no domicílio;
- Porte do animal de estimação: pequeno, médio e grande;
- Principal motivo de aquisição do cão e/ou gato: companhia, diversão/afetividade, status/moda /distinção social, recomendação médica/terapia/guia, segurança/controle de roedores e reprodução/negócio;
- Forma de aquisição do cão e/ou gato: doação, adoção, compra em pet shop, compra em criadores profissionais (canis/gatis), compra em mercado informal, cria da casa e outra forma.

Variáveis referentes à caracterização do padrão de consumo:

- Gasto médio mensal domiciliar com saúde do(s) animal(is);
- Gasto médio mensal domiciliar com alimentação do(s) animal(is);
- Gasto médio mensal domiciliar com beleza/adornos/roupas do(s) animal(is);
- Gasto médio mensal domiciliar com lazer/acessórios e brinquedos do(s) animal(is);
- Gasto médio mensal domiciliar com higiene do(s) animal(is);
- Gasto total domiciliar com animais de estimação = somatório de todos os gastos.

Variáveis referentes ao comportamento dos proprietários em relação a cães e gatos – Antropomorfismo:

- Circulação irrestrita no domicílio de pelo menos um cão ou gato — permissão de circulação de animais — variável referente ao "pet love";
- Existência no domicílio de pelo menos um animal de estimação que utilizou roupas e/ou adornos — uso de vestuário — variável referente ao "consumo pet";
- Existência no domicílio de pelo menos um animal de estimação que utilizou acessórios e/ou brinquedos — uso de acessórios — variável referente ao "consumo pet";
- Existência no domicílio de oferta habitual de guloseimas próprias para animais a pelo menos um animal de estimação — consumo de guloseimas — variável referente ao "consumo pet".

ANEXO 2 - Ordem de retirada das variáveis, R2, valores das estatísticas Wald´s e os respectivos p-valores do modelo para o Logaritmo neperiano do gasto total

R2	Variável retirada	Wald	p
0,409	Tipo domicilio X motivo de aquisição por diversão	0,000	0,987
0,409	Tipo de arranjo familiar: nuclear / biparental / sem filhos	0,002	0,964
0,409	Tipo de domicílio X possui um animal de médio porte	0,005	0,943
0,409	Tipo de arranjo familiar: composto	0,007	0,934
0,409	Total de homens	0,008	0,930
0,409	Forma de aquisição: cria de casa	0,056	0,814
0,409	Tipo domicilio X possui um animal de pequeno porte	0,111	0,742
0,409	Total de mulheres	0,125	0,727
0,409	Forma de aquisição: adoção	0,100	0,755
0,409	Forma de aquisição: doação	0,060	0,809
0,409	Forma de aquisição: comprou criador	0,079	0,781
0,408	Motivo de aquisição: companhia	0,155	0,697
0,408	Tipo domicílio	0,145	0,707
0,408	O domicilio possui um idoso	0,225	0,640
0,408	Tipo de arranjo familiar: nuclear / biparental / com filhos	0,370	0,549
0,407	Tem um animal de raça X possui um animal de médio porte	0,366	0,551
0,407	Tem um animal de raça X possui um animal de grande porte	0,377	0,545
0,406	Tem um animal de raça	0,053	0,819
0,406	Possui um animal de grande porte	0,427	0,520
0,406	Tipo domicilio X motivo de aquisição: reprodução	0,347	0,561
0,406	Tem uma criança no domicílio	0,815	0,376
0,404	Tipo domicílio X tem uma criança no domicílio	0,272	0,607
0,404	Tem jovem no domicílio	0,166	0,688
0,404	Tipo domicilio X tem um menor de idade no domicílio	0,707	0,409
0,403	Possui um animal de médio porte	0,720	0,405
0,402	Tipo domicílio X tem um idoso no domicílio	1,195	0,285
0,402	Tem um animal de raça X possui um animal de pequeno porte	1,451	0,240
0,400	Possui um animal de pequeno porte	0,566	0,459
0,399	O animal tem permissão de circular no domicílio	1,539	0,227
0,397	O animal usou roupas pelo menos uma vez	1,700	0,205
0,395	Tipo domicilio X motivo de aquisição: companhia	1,763	0,197
0,393	Motivo de aquisição: diversão	0,515	0,480
0,392	Motivo de aquisição: segurança	0,793	0,382
0,392	Tipo domicílio X motivo de aquisição: segurança	0,905	0,351
0,391	Tipo domicílio X motivo de aquisição: terapia	1,222	0,280
0,391	Motivo de aquisição: terapia	1,465	0,238
0,391	Forma de aquisição do pet: comprou pet shop	2,475	0,129
0,388	Sexo do chefe do domicilio	3,169	0,088
0,383	Forma de aquisição do pet: comprou no mercado	3,379	0,078
0,381	Tipo domicilio X possui um animal de grande porte	3,657	0,068
0,378	Tipo de arranjo familiar: estendido	3,835	0,062

ABSTRACT

This study aimed to investigate household expenditure in the sectors of hygiene, beauty, health, food and entertainment, intended to pets in permanent private households in Grande Méier, Todos os Santos, Engenho Novo, and Lins de Vasconcelos, in the city of Rio de Janeiro, in 2007. The data were obtained from the study "Pesquisa domiciliar sobre cães e gatos: humanização e padrões de consumo", of the Instituto Brasileiro de Geografia e Estatística. The methodology began with a literature review, following, an exploratory data analysis was performed. Subsequently, statistical model were adjusted for the total household monthly expenditure with pets. Among the results, verified that the residents spend on average R$ 149.47 (s = 11.33) per month with pets. In adjusted model, was observed consumption, with anthropomorphic bond between owners and their animals, through the relationships identified in spending on animals.

Keywords: Estimate animals, family spending, multiple linear regression.

Análise de predição e previsão das concentrações de material particulado inalável (PM_{10}) na cidade de Carapina, ES

Wesley R. Gripa[1]
Valdério A. Reisen[2]
Fabio A. Fajardo[3]
Neyval C. Reis Jr.[4]

Resumo

Material Particulado Constitui um grande problema para a qualidade do ar em regiões metropolitanas. Neste artigo são discutidas técnicas de previsão da qualidade do ar para concentrações médias de Material Particulado Inalável (PM_{10}) com a consideração de fatores metereológicos. Os modelos de séries temporais e de regressão linear múltipla são as metodologias usadas para o ajuste da dinâmica da concentração média diária de PM_{10} na cidade de Carapina, Região da Grande Vitória, ES, Brasil. Na primeira metodologia considera-se um cenário com erros não-correlacionados no modelo ajustado e na segunda metodologia avalia-se a presença de erros correlacionados em um modelo de regressão linear múltipla. Os modelos ajustados foram considerados para predição e previsão do conjunto de observações. Ambos modelos evidenciaram resultados semelhantes, no entanto o modelo de regressão apresentou medidas de previsão das concentrações médias de PM_{10} um pouco melhores que as do modelo de séries temporais.

Palavras-chave: Modelagem estocástica, regressão múltipla, SARIMA, qualidade do ar, dados faltantes.

[1]Programa de Pós-Graduação em Engenharia Ambiental; Departamento de Matemática Aplicada - CEUNE/UFES, São Mateus, ES
[2] Programa de Pós-Graduação em Engenharia Ambiental; Departamento de Estatística – UFES, Vitória, ES.
[3] Departamento de Estatística – UFES, Vitória, ES.
[4] Programa de Pós-Graduação em Engenharia Ambiental – UFES,Vitória, ES.

1. Introdução

Topografia, densidade populacional, frota veicular, atividades industriais e condições meteorológicas, entre outros fatores, contribuem para o aumento da poluição atmosférica (Perz & Reyes 2002). Um poluente comum nas regiões urbanas é o material particulado inalável. O PM_{10} é constituído pelas partículas com um diâmetro aerodinâmico menos ou igual a 10μm que possuem como fontes principais a queima de combustíveis fósseis e processos industriais. De acordo com Souza (2002), a caracterização do PM_{10} na Região da Grande Vitória, ES , Brasil, é dada principalmente pelas contribuições industriais, as atividades humanas (emissão veicular, queimadas, construção civil, entre outras) e as emissões naturais. Do ano de 1995 a 1998, esses fatores representaram 34.6%, 54.6% e cerca de 10.8%, respectivamente, do material particulado coletado nessa região (Souza(2002)).

Na atualidade, o estudo das concentrações do material particulado em regiões metropolitanas é um tópico de grande relevância para prevenção de doenças respiratórias e cardiovasculares. Estudos realizados por Ostro et al. (1996) mostraram uma associação significativa entre PM_{10} e problemas respiratórios em grupos de crianças na cidade de Santiago (Chile). Através da análise de séries temporais, os autores detectaram um aumento de 7% no número de internações em crianças com idade de 2 a 15 anos para um aumento de 50$\mu g/m^3$ na concentração de PM_{10}, com defasagem de 5 dias. Estudos similares conduzidos na Califórnia (Estados Unidos) e Bancoque (Tailândia) indicaram uma associação estatística significativa entre PM_{10} e mortalidade, em que um incremento de 10$\mu g/m^3$ na concentração diária de PM_{10} é associado com cerca de 1% de aumento da mortalidade total e cerca de 3% para doenças respiratórias e cardiovasculares (para detalhes ver, e.g., Ostro, Hurley & Lipsett (1999) e Ostro, Eskeland, Sanchez & Feyzioglu (1999)).

Maddison (2005) avaliou a relação entre o número de admissões hospitalares e poluentes atmosféricos na cidade de Londres, por meio de modelos de séries temporais com variáveis exógenas. Os resultados mostraram que uma redução de 1% nos níveis de PM_{10} resultaria, a longo prazo, numa redução de 0.14% no número de admissões hospitalares por causas respiratórias. Em contrapartida, o autor afirma que o aumento das internações por causas cardiovasculares não pode ser explicado somente pela poluição atmosférica. Neuberger et al. (2007) verificaram a relação entre indicadores de poluição do ar no perímetro urbano com a mortalidade diária utilizando séries diárias de diversos poluentes para a cidade de Viena (Áustria) no período de 2000 a 2004. Resultados evidenciaram que um incremento de $10\mu g/m^3$ na concentração dos poluentes $PM_{2.5}$ e NO_2 indicaram, respectivamente, acréscimos de 2.6% e 2.9% na mortalidade por causas respiratórias para uma defasagem de $0 - 14$ dias.

Goyal et al. (2006) estudaram a dinâmica do PM_{10} para as cidades de Delhi e Hong Kong através de modelos estatísticos de regressão linear múltipla com erros correlacionados e um modelo de séries temporais. Os autores verificaram o desempenho dos modelos e mostraram que a regressão linear múltipla com erros correlacionados captou melhor a variabilidade dos dados para ambas as cidades. Os resultados obtidos pelos autores motivaram a utilização desses modelos na análise da dinâmica das concentrações do material particulado na cidade de Carapina.

O principal objetivo deste trabalho é comparar, por meio de modelos de séries temporais e de regressão linear múltipla, a qualidade do ajuste e a capacidade preditiva desses modelos para a variável do material particulado inalável, observada na cidade de Carapina, Região da Grande Vitória (RGV), no período 01 de janeiro a 31 de dezembro de 2006. No modelo de séries temporais assume-se a independência no erro, e no modelo regressão linear múltipla, com erros correlacionados, considerou-se uma variável metereológica para explicar a dinâmica das observações. Os modelos ajustados apresentam, como parte relevante do ajuste, uma componente adicional que permite modelar a sazonalidade dos dados.

Séries temporais provenientes de monitoramentos da qualidade do ar frequentemente sofrem com o problema de observações faltantes (ver, e.g., Palma & del Pino (1999) e Iglesias et al. (2006)). Portanto, o processo de modelagem é precedido pela imputação das observações faltantes através de uma modificação do filtro de Kalman e a aplicação do algoritmo EM ("Expectation Maximization"), sugerida por Palma & Chan (1997) e Shumway & Stoffer (1982). Para seleção do modelo que melhor represente a dinâmica dos dados sob estudo, medidas de erros foram calculadas para tal propósito. Os resultados indicam que o modelo de regressão com erros correlacionados, e com co-variável a *velocidade do vento*, representa adequadamente o comportamento do PM_{10} na cidade de Carapina e apresenta melhor desempenho no cálculo de previsões.

Este artigo está dividido como se segue. A Seção 2 descreve as técnicas de modelagem utilizadas para o ajuste das observações. A Seção 3 apresenta as análises e os resultados obtidos. Finalmente, a Seção 4 apresenta as conclusões e alguns comentários finais sobre a análise dos dados.

2. Metodologia

2.1. Modelo SARIMA $(p,d,q) \times (P.D,Q)_s$. Seja $Z_t \equiv \{Z_t; t \in Z\}$ um processo linear com representação dada por

$$\Phi(B^s)\phi(B)\nabla^d Z_t = \Theta(B^s)\theta(B)\in_t \qquad (1)$$

onde s é chamado período sazonal do processo, $\{\in_t\}$ é um processo de ruído branco com $E[\in_t] = 0$ e $Var[\in_t] = \sigma_\in^2$. O operador ∇^d , onde $d = (d, D)$ e d, D são números inteiros não negativos, é definido por:

$$\nabla^d = (1-B)^d (1-B^s)^D \qquad (2)$$

O operador de defasagem B é definido como $B^k Z_t = Z_{t-k}, .\kappa \in \mathrm{N}, \Phi(z^s) = 1 - \Sigma_{i=1}^{P} \Phi_i z^{is}, \phi(z) = 1 - \Sigma_{j=1}^{p} \phi_j z^j, \Theta(z^s) = 1 - \Sigma_{k=1}^{Q} \Theta_k z^{ks}$ e $\theta(z) = 1 - \Sigma_{l=1}^{q} \theta_l z^l$ são polinômios de ordem P, p, Q, q $\in$ $\mathbb{N}$, respectivamente, com $z \in C$ e $\{\Phi_i\}$, $\{\phi_j\}$, $\{\Theta_k\}$, $\{\theta_l\}$ são sequências de números reais. O processo Z_t com representação dada pela Eq. 1 é chamado de SARIMA $(p.d,q)$ x $(P,D,Q)_s$. O processo Z_t é estacionário e invertível se $d = D = 0$ e as raízes de $\Phi(z^s)\phi(z)$ e $\Theta(z^s)\theta(z)$ são não comuns e encontram-se fora do círculo unitário (para detalhes ver, e.g., *Shumway & Stoffer (2010)* e *Wei (2005)*). Extensão do modelo acima é o processo SARFIMA $(p.d,q)$ x $(P,D,Q)_s$ onde $d, D \in$ R(veja os recentes trabalhos de *Reisen et al. (2006), Fajardo et al. (2009), Reisen et al. (2010), Reisen & Fajardo (2012), Reisen & Fajardo (2008),* entre outros).

2.2. Modelos de regressão com erros correlacionados.

Análise de regressão é uma técnica estatística bastante utilizada em pesquisas ambientais. A idéia básica do modelo de regressão consiste em descrever a relação existente entre uma variável dependente e variáveis denominadas independentes. O modelo de regressão pode ser representado por

$$Z_t = \mu(\mathrm{x}_t) + u_t, \quad t = 1, 2, ..., \mathrm{n} \qquad (3)$$

onde Z_t é a variável resposta, u_t é o termo de erro e *n* é o tamanho de amostra. A função $\mu(\mathrm{x}_t)$ deve ser definida pelo pesquisador em função do grau de conhecimento do fenômeno sob estudo. Na função $\mu(.)$ são definidos o vetor das variáveis explicativas x_t e as componentes sazonais que possam explicar a dinâmica do conjunto de dados. Em diferentes cenários as componentes sazonais podem ser representadas em forma trigonométrica como:

$$S_t = \sum_{j=1}^{\lfloor s/2 \rfloor} \left[a_j \operatorname{sen}\left(\frac{2\pi jt}{s}\right) + b_j \cos\left(\frac{2\pi jt}{s}\right) \right], \qquad (4)$$

onde $\lfloor . \rfloor$ denota a função parte inteira e $\{a_j\}$, $\{b_j\}$ representam sequências de constantes reais.

Assume-se que o termo de erro u_t segue uma representação Autorregressiva e de Médias Móveis (ARMA) estacionária, i.e. a representação na Eq. 1 com período sazonal $\mathrm{s} = 0$ e $\mathrm{d} = \mathrm{D} = 0$, dado por

$$\phi(B)u_t = \theta(B)\in_t, \qquad (5)$$

Onde $\{\in_t\}$é um processo de ruído branco com média zero e variância constante. Um procedimento geral para estimação dos parâmetros do modelo na *Eq. 3* é desenvolvido por *Cochrane & Orcutt (1949).*

2.3. Estimação de dados faltantes

A presença de observações faltantes em conjuntos de dados é um problema comumente encontrado em aplicações práticas. O tratamento desse tipo de dados tem sido estudado para diferentes contextos nos trabalhos de *Palma & Cham (1997)* e *Shumway & Stoffer (1982),* entre outros. Neste trabalho aplica-se a metodologia proposta por *Shumnway & Stoffer (1982)* baseada representação de espaço de estados do modelo e uma combinação entre o filtro de *Kalman e o algoritmo* EM. A metodologia proposta pelos autores assume que a dinâmica do conjunto de observações pode ser representada pela equação de estados

$$z_t = \mathrm{A}_t x_t + v_t, \quad t = 1, 2, ..., n$$

onde z_t são vetores q x 1 de observações, A_t é uma matriz q x p de constantes, são vetores de estados p x 1 com representação $x_t = \Phi x_{t-1} + w_t$ *e* w_t são vetores p x 1 de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) com distribuição normal com média **0** e matriz de covariâncias *Q*. $\{v_t\}$ é um processo de ruído branco gaussiano não-correlacionado com $\{w_t\}$.

Para um tempo *t*, define-se uma partição do vetor $z_t = \left(z_t^{(1)'}, z_t^{(2)'}\right)$, onde a primeira componente $z_t^{(1)'}$, de tamanho $q_{1t} \times 1$, representa o vetor de valores observados e a segunda componente $z_t^{(2)'}$, de tamanho $q_{2t} \times 1$, representa o vetor de valores faltantes, tal que $q_{1t} + q_{2t} = q$. Então,

$$\begin{pmatrix} z_t^{(1)} \\ z_t^{(2)} \end{pmatrix} = \begin{bmatrix} A_t^{(1)} \\ A_t^{(2)} \end{bmatrix} x_t + \begin{pmatrix} v_t^{(1)} \\ v_t^{(2)} \end{pmatrix},$$

onde $A_t^{(1)}$ *e* $A_t^{(2)}$ são as partições da matriz A_t com tamanhos $q_{1t} \times p$ *e* $q_{2t} \times p$, respectivamente, e

$$\operatorname{cov}\begin{pmatrix} v_t^{(1)} \\ v_t^{(2)} \end{pmatrix} = \begin{bmatrix} R_{11t} & R_{12t} \\ R_{21t} & R_{22t} \end{bmatrix}. \qquad (6)$$

Shumway & Stoffer (1982) estabelecem as equações de filtragem para o caso de observações faltantes se, no instante *t*, utilizam-se às substituições

$$z_{(t)} = \begin{pmatrix} z_t^{(1)} \\ 0 \end{pmatrix}, \qquad A_{(t)} = \begin{bmatrix} A_t^{(1)} \\ 0 \end{bmatrix}, \qquad R_{(t)} = \begin{bmatrix} R_{11t} & 0 \\ 0 & R_{22t} \end{bmatrix}.$$

Baseado nessa substituição, os estimadores dos estados são dados por

$$x_t^{(s)} = E\left[x_t \middle| y_1^{(1)}, \ldots, y_s^{(1)}\right],$$

com matriz de variâncias dos erros $P_t^{(s)} = E\left[\left(x_t - x_t^{(s)}\right)\left(x_t - x_t^{(s)}\right)'\right]$. Os estimadores de máxima verossimilhança, como calculados no algoritmo EM, sofrem uma pequena mudança devido à presença de observações faltantes. Para implementar o passo E, na iteração *j*, deve-se calcular

$$E\left[-21nL_{X,Y}(\Theta)\middle|Y_n^{(1)}, \Theta^{(j-1)}\right] = E_*\left[1n\left|\Sigma_0\right| + tr\Sigma_0^{-1}(x_0 - \mu_0)(x_0 - \mu_0)'\middle|Y_n^{(1)}\right]$$

$$+E_*\left[n\ln\left|Q\right| + \sum_{t=1}^{n} tr\left[Q^{-1}\left(x_t - \phi x_{t-1}\right)\left(x_t - \phi x_{t-1}\right)'\middle|Y_n^{(1)}\right]\right]$$

$$+E_*\left[n\ln\left|R\right| + \sum_{t=1}^{n} tr\left[R^{-1}\left(z_t - A_t x_t\right)\left(z_t - A_t x_t\right)'\middle|Y_n^{(1)}\right]\right],$$

onde E_* denota o valor esperado condicional sob $\Theta^{(j-1)}$, $Y_n^{(1)} = \{y_1^{(1)}, ..., y_n^{(1)}\}$, $\Theta = \{\mu_0, \Sigma_0, \phi, Q, R\}$ o vetor de parâmetros, com μ_0 a média inicial, Σ_0 a matriz de covariâncias, $L_{X,Y}(.)$ a função de verossimilhança (para detalhes ver Shumway & Stoffer (1982)).

3. Análises e resultados

Nesta seção são analisadas as relações existentes entre as concentrações médias diárias de PM_{10} ($\mu g/m^3$), de radiação ($Watts/m^2$), de temperatura (^oC), da umidade (%) e da velocidade do vento (m/s) medidos na cidade de Carapina, Espírito Santo, período de observação de 01 de janeiro a 31 de dezembro de 2006. Os dados em estudo foram cedidos pelo Instituto Estadual de Meio Ambiente e Recursos Hídricos - IEMA.

A presença de dados faltantes na série PM_{10} motiva o uso da metodologia baseada na aplicação do filtro de Kalman e algoritmo EM nos dados, proposta por Shumway & Stoffer (1982) e disponível no site *http://www.stat.pitt.edu/stoffer/tsa3/tsa3.rda*. A série com dados imputados das concentrações médias diárias do PM_{10} será denotada por PM_{10}^{*}.

A Tabela 1 mostra as correlações calculadas entre as variáveis sob estudo. Observa-se que a variável velocidade do vento (Vel. do Vento) apresenta uma relação linear mais forte com o PM_{10}^{*} coletado na estação de Carapina, indicando que os níveis de material particulado estão associados ás mudanças na velocidade do vento.

Tabela 1. Matriz de correlação entre as variáveis sob estudo.

	PM_{10}	Radiação	Temperatura	Umidade	Vel. do Vento
PM_{10}	1.000				
Radiação	-0.018	1.000			
Temperatura	0.049	0.694	1.000		
Umidade	-0.112	-0.563	-0.288	1.000	
Vel. do Vento	-0.217	0.252	0.241	-0.379	1.000

A Figura 1 apresenta a matriz de dispersão entre o PM_{10} e as variáveis metereológicas de interesse. O gráfico evidencia a associação entre as variáveis sob estudo. Os gráficos de dispersão corroboram com os resultados apresentados na Tabela 1. Com base na análise de correlação (1), optou-se por considerar a velocidade do vento como a variável que apresenta maior associação linear com as concentrações de PM_{10}^{*}. As Figuras 2 e 3 apresentam a evolução temporal das variáveis meteorológicas e da série PM_{10}^{*}, respectivamente.

O aumento na variabilidade do PM_{10} pode ser também explicado pela presença de valores atípicos nas observações. Uma análise detalhada para identificar os possíveis valores atípicos nas observações do PM_{10} mostra que 3 valores do conjunto encontram-se fora do padrão dos dados (observações número 53, 54 e 156), porém o tratamento desse tipo de dados não é o objetivo desta pesquisa. Nesse contexto, as metodologias apresentadas por Fajardo et al. (2009) e Reisen & Fajardo (2012) serão consideradas para futuros trabalhos na análise de dados de poluição do ar coletados na RGV.

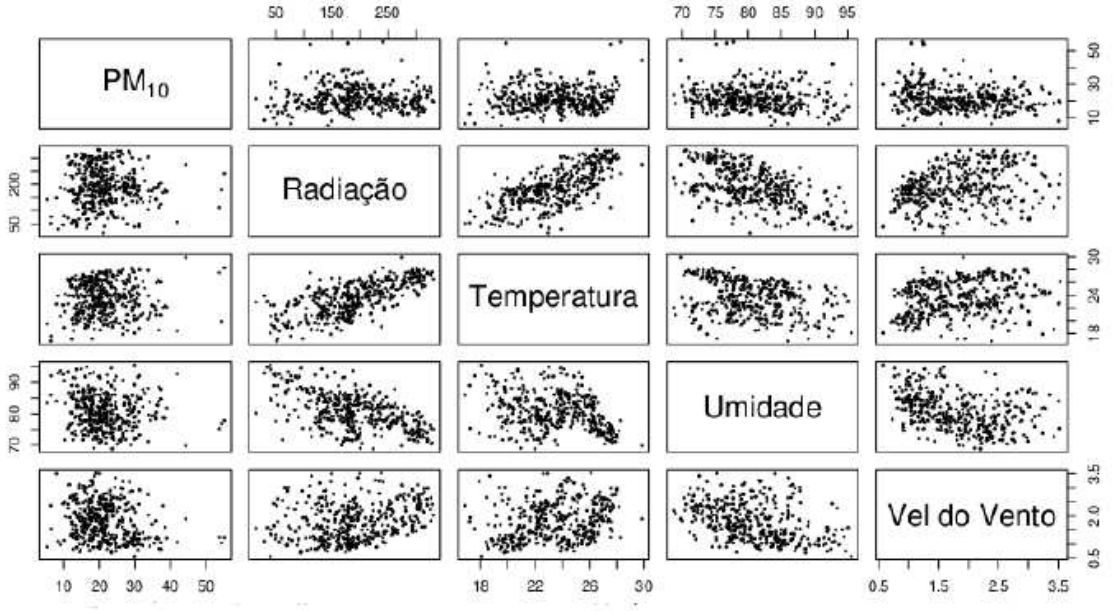


Figura 1. Matriz de dispersão entre PM10 e as variáveis metereológicas Radiação, Temperatura, Umidade e Velocidade do Vento.

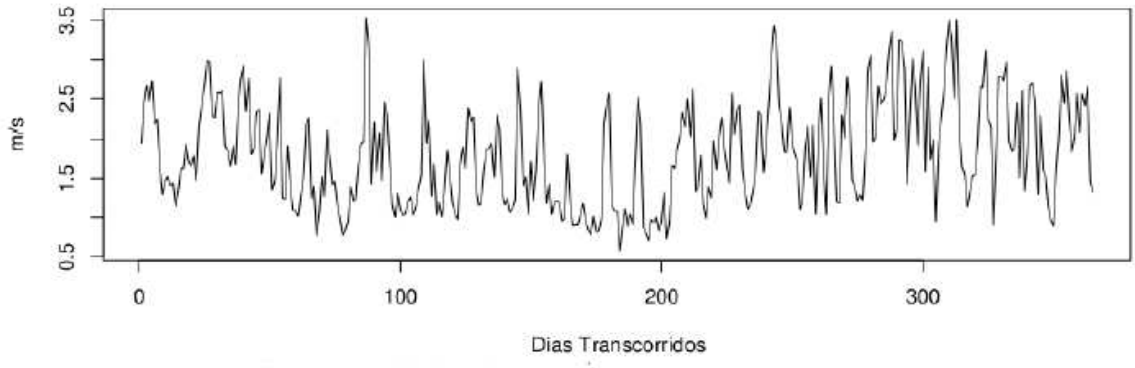


Figura 2. Dinâmica da variável que representa a Velocidade do vento da cidade de Carapina período 01 jan - 31 dez. 2006.

3.1. Modelos Ajustados

Com o objetivo de estabilizar a variância, o ajuste do modelo foi feito na série $\ln PM_{10}^{*}$ e as análises da qualidade dos modelos estimados estão descritos a seguir.

Modelo Tempora l - Modelo I. Na Figura 4 observam-se as Funções de Autocorrelação Amostral (FAC) e Autocorrelação Parcial Amostral (FACP) da série transformada. A FAC (Figura 4(a)) apresenta um decaimento exponencial e correlações estatisticamente significativas para defasagens múltiplas de 7, o que sugere a presença de sazonalidade com período $s = 7$. As evidências empíricas da FAC recomendam o ajuste de um modelo da classe SARMA (p,q) $\times$ $(P,Q)_7$, onde p e q representam as ordens autorregressiva e de médias móveis, respectivamente. Os valores P e Q representam as ordens autoregressiva e de médias móveis, sazonal respectivamente. Para identificação das ordens do modelo utilizou-se o critério de informação de Akaike (AIC) e as funções FAC e FACP. A análise dessas medidas sugere que um modelo adequado para explicar a dinâmica do logaritmo do PM_{10}^{*} e um modelo SARMA(1, 0) $\times$ (2, 0)7 (Modelo I). As estimativas dos parâmetros do modelo ajustado são apresentadas na Tabela 2.

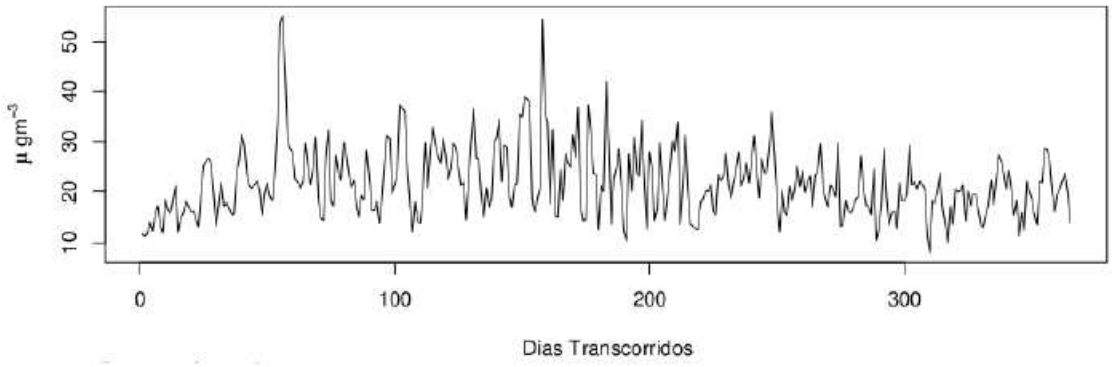


Figura 3. Série das concentrações PM*10 na cidade de Carapina período 01 jan - 31 dez de 2006.

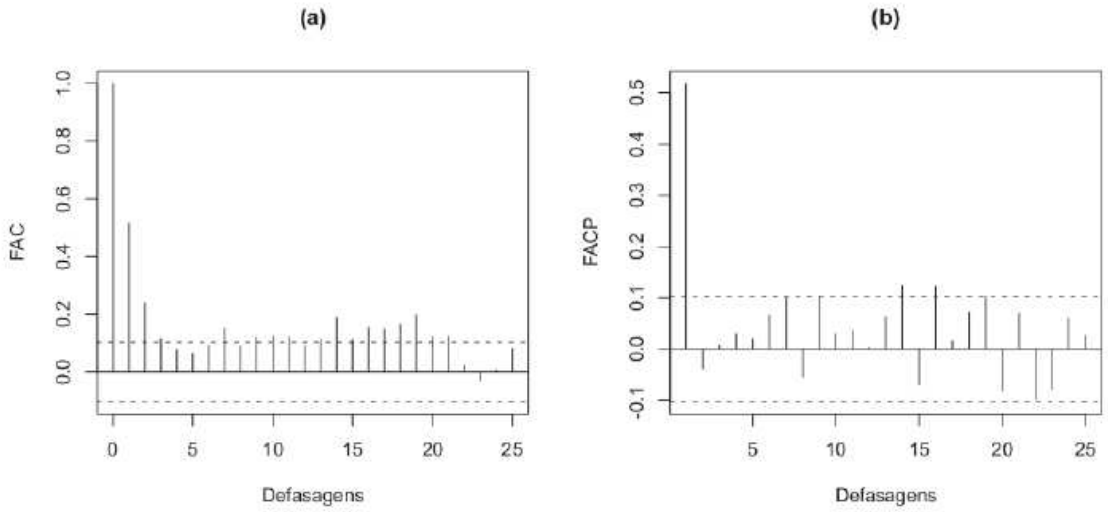


Figura 4. Funções de Autocorrelação Amostral do logaritmo do PM*10; (a) FAC e (b) FACP.

Tabela 2. Estimativas dos parâmetros do Modelo I.

Parâmetro	Estimativa	Teste t
Constante	3.030	81.676
ϕ	0.514	11.403
Φ_1	0.117	2.270
Φ_2	0.157	3.023

Modelo de Regressão - Modelo II. Para efeitos de comparação das técnicas de modelagem, optou-se pelo ajuste de um modelo de regressão com erros correlacionados aos dados de PM_{10}^{*}, com a consideração da variável exógena velocidade do vento (Figura 2). Como observado anteriormente, a série PM_{10}^{*} apresenta sazonalidade com período $s = 7$. Para controlar essa propriedade, foi incluído no modelo de regressão a componente S_t (Eq. 4). O modelo de regressão inicialmente ajustado é da forma

$$\ln PM_t = \beta_0 + \beta_1 \ln VENT_t + \sum_{j=1}^{3}\left[\alpha_j \operatorname{sen}\left(\frac{2\pi jt}{7}\right) + \gamma_j \cos\left(\frac{2\pi jt}{7}\right)\right] + ut, \qquad (7)$$

onde ut é o erro com variância Γ_u. Adicionalmente, verificou-se que não existe associação linear entre a concentração de PM_{10} e a variável velocidade do vento defasada p períodos, não sendo nenhuma defasagem estatisticamente significativa. Tal resultado permite o ajuste de um modelo parcimonioso para as variáveis sob estudo.

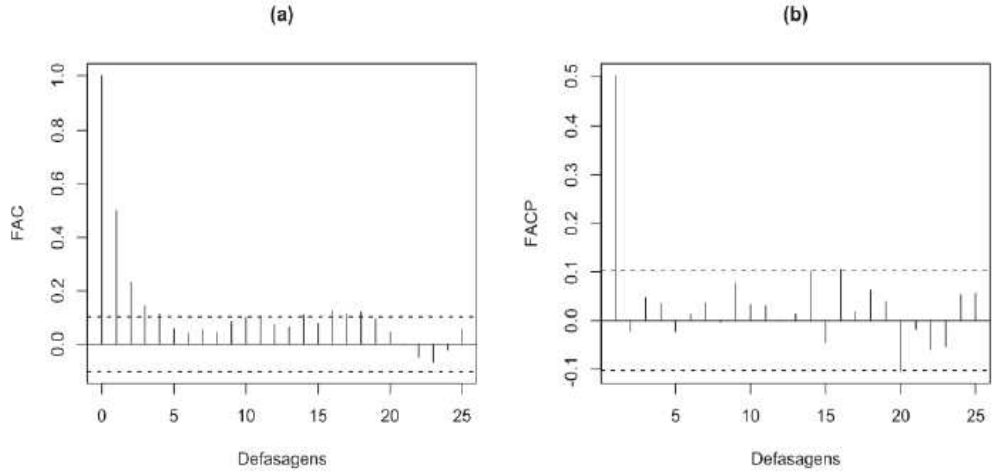


Figura 5. Funções de Autocorrelação Amostral do ût; (a) FAC e (b) FACP.

A Figura 5 mostra as funções FAC e FACP amostrais para a componente *ut*. A identificação da ordem é realizada através do AIC e as funções FAC e FACP. Portanto, o modelo selecionado para $\hat{u}t$ e o $\mathrm{AR}(1)$. Por isso, o modelo II ajustado é uma regressão linear múltipla com erros $\mathrm{AR}(1)$.

A Tabela 3 apresenta as estimativas dos parâmetros incluídos no modelo II. O coeficiente $\hat{\beta}_1 = -0.133$ indica a existência de uma relação inversa entre a concentração de PM_{10} e a velocidade do vento. Fisicamente, essa relação é verificada, pois com o aumento da velocidade do vento ocorre um aumento no espalhamento da pluma das partículas PM_{10} e, consequentemente, a concentração pontual do poluente se reduz.

Tabela 3. Estimativas dos parâmetros do Modelo II.

Parâmetro	Estimativa	Teste t
β_0	0.001	0.085
β_1	-0.133	-2.900
α_1	-0.066	-3.615
γ_1	-0.039	-2.152
α_2	0.007	0.405
γ_2	0.055	3.027
α_3	0.034	1.879
γ_3	-0.048	-2.628
ϕ	0.508	11.202

Adequação dos modelos. Para verificar a adequação dos modelos a análise dos resíduos se torna uma ferramenta essencial para o sucesso da modelagem. A Tabela 4 apresenta os valores *p* do teste de normalidade Jarque & Bera (1981), de não correlação residual Box & Pierce (1970) e de homoscedasticidade (Multiplicadores de Lagrange). Todos os testes não rejeitaram as hipóteses nulas de normalidade, de não-correlação residual e de homoscedasticidade, respectivamente. As Figuras 6 e 7 mostram as análises gráficas dos resíduos para os modelos I e II, respectivamente, e corroboram os resultados apresentados na Tabela 4.

Tabela 4. Testes estatísticos de normalidade*, não-correlação e homoscedasticidade*** residual**

	Modelo I	Modelo II
Teste	valor p	valor p
Jarque Bera*	0.257	0.918
Box Pierce**	0.471	0.641
Multiplicador de Lagrange***	0.972	0.857

Na Figura 8 é visualizado o ajuste dos modelos I e II. A análise gráfica mostra um comportamento semelhante para ambos modelos. Essa evidência ´e confirmada pelas medidas do Erro Quadrático Médio (EQM), do Erro Absoluto Médio (EAM), do Erro Absoluto Médio Percentual (EAMP) e da Raiz do Erro Quadrático Médio (REQM), apresentados na Tabela 5. Os resultados da Tabela 5 evidenciam um melhor desempenho do modelo de séries temporais, com medidas de erros inferiores às do modelo de regressão.

Tabela 5. Avaliação de desempenho dos modelos

Medidas de erro	Modelo I	Modelo II
EQM	15.121	16.874
EAM	2.745	2.894
EAMP	0.123	0.132
REQM	3.888	4.107

3.2. Estudo de Previsão

Nesta seção é apresentado o estudo de previsão de um passo à frente, período de 01 de janeiro a 28 de fevereiro de 2007, para comparar o desempenho dos modelos ajustados. As medidas dos erros de previsão são apresentadas na Tabela 6 e indicam que o modelo de regressão (Modelo II) obteve previsões um pouco melhores, menores EQM, EAM, EAMP e REQM, quando comparado aos valores obtidos através de séries temporais (Modelo I), o que não é um resultado surpreendente. Como é conhecido na literatura, modelos que apresentam , em geral, melhores ajustes não necessariamente são os mais indicados para previsão. Nesse contexto, pode ser considerado estudo de previsão com combinações de modelos, metodologia já explorada por trabalhos de pesquisa na área de previsão, ideia que pode ser considerada em futuras análises das variáveis em questão.

Tabela 6. Avaliação de desempenho das previsões dos modelos

Medidas de erro	Modelo I	Modelo II
EQM	11.240	10.127
EAM	2.324	2.242
EAMP	0.092	0.091
REQM	3.352	3.182

Para uma análise visual, a Figura 9 mostra as previsões dos Modelos I e II, juntamente com os valores observados para a concentração diária de PM_{10}. Nota-se que as previsões capturam bem a variabilidade dos dados com valores próximos aos medidos.

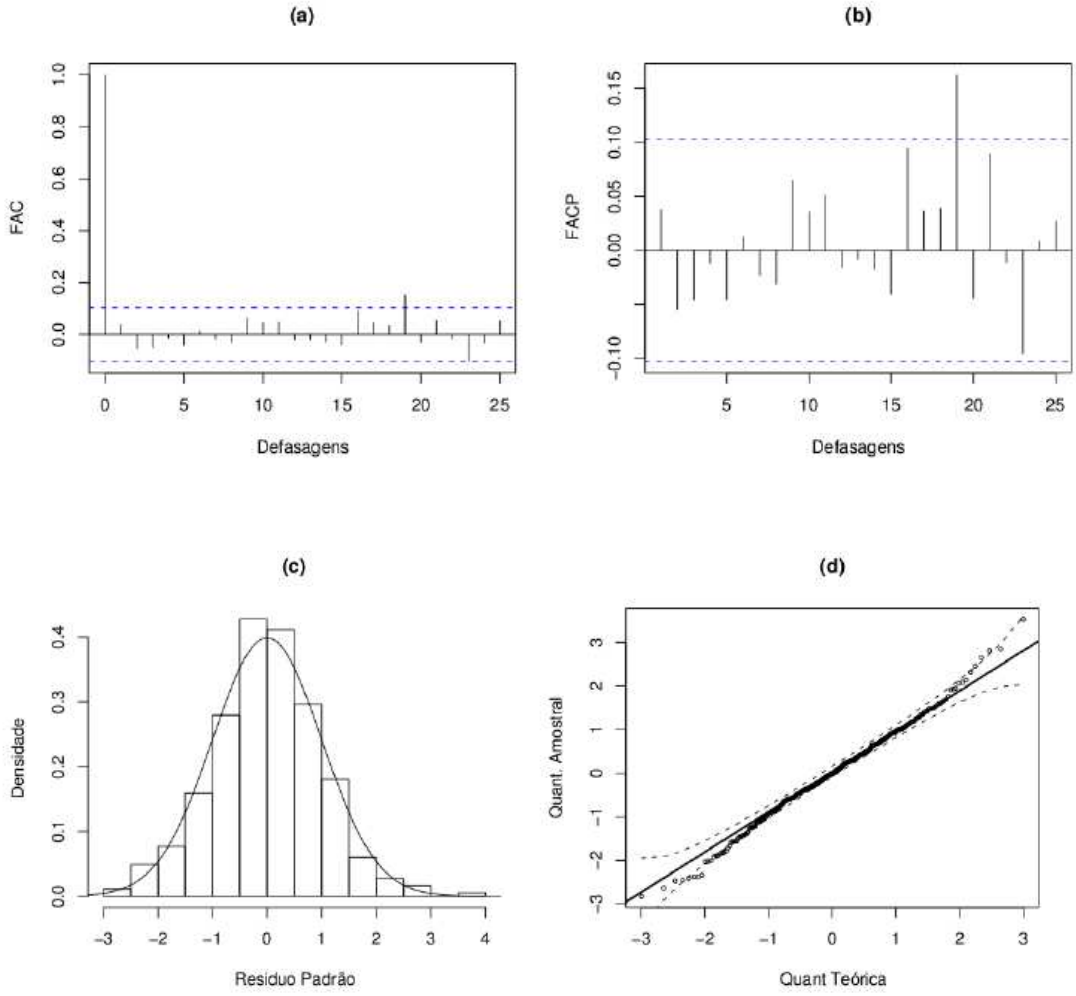


Figura 6. Gráficos para a análise residual do Modelo I: FAC, FACP, histograma e gráfico de quartis.

4. Conclusões

Neste artigo avaliaram-se a qualidade do ajuste e a capacidade preditiva de dois modelos que representam a dinâmica das concentrações de PM_{10} na cidade de Carapina, ES. Os modelos ajustados foram baseados em duas técnicas de modelagem: série temporal e regressão com erros correlacionados. Um termo adicional é considerado nas equações matemáticas que representam os modelos para avaliar a presença de uma componente sazonal nos dados. O uso da média diária da velocidade do vento como variável explicativa das concentrações de PM_{10} no modelo de regressão permitiu melhorar a capacidade preditiva do mesmo, fornecendo previsões mais próximas aos valores observados e, consequentemente, menores valores para as medidas de erro de previsão. Como parte de um estudo posterior, sugere-se aprimorar os modelos com a inclusão de variáveis metereológicas adicionais, que permitam melhorar as análises e dessa forma explicar razoavelmente o comportamento do material particulado na cidade de Carapina. O ajuste de modelos vetoriais autorregressivos pode enriquecer a capacidade preditiva e considerar as relações de longo prazo existentes entre as variáveis metereológicas. Os modelos de função de transferência podem ser considerados como parte da metodologia de modelagem para melhorar a qualidade do ajuste dos modelos sugeridos, assim como os modelos sazonais fracionários (veja referências sugeridas na introdução).

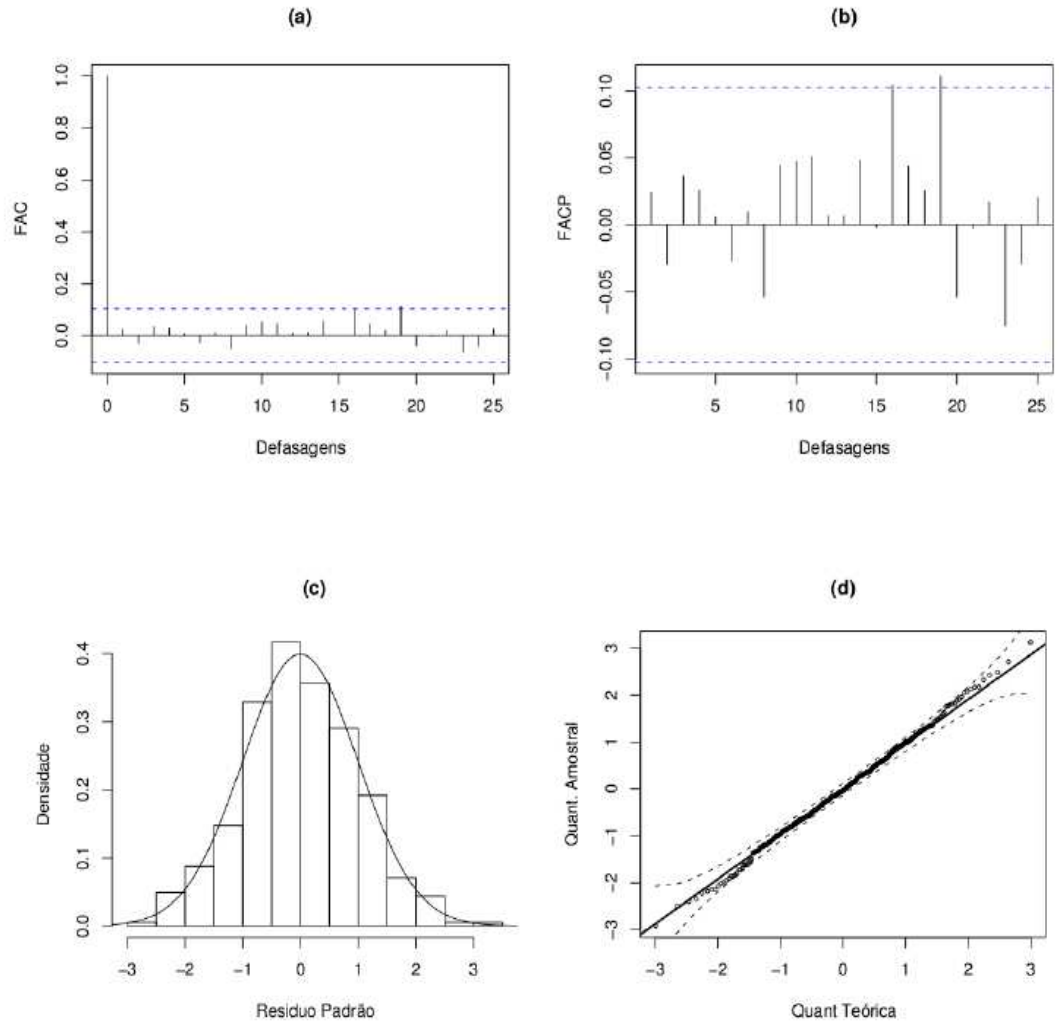


Figura 7. Gráficos para a análise residual do Modelo II: FAC, FACP, histograma e gráfico de quartis.

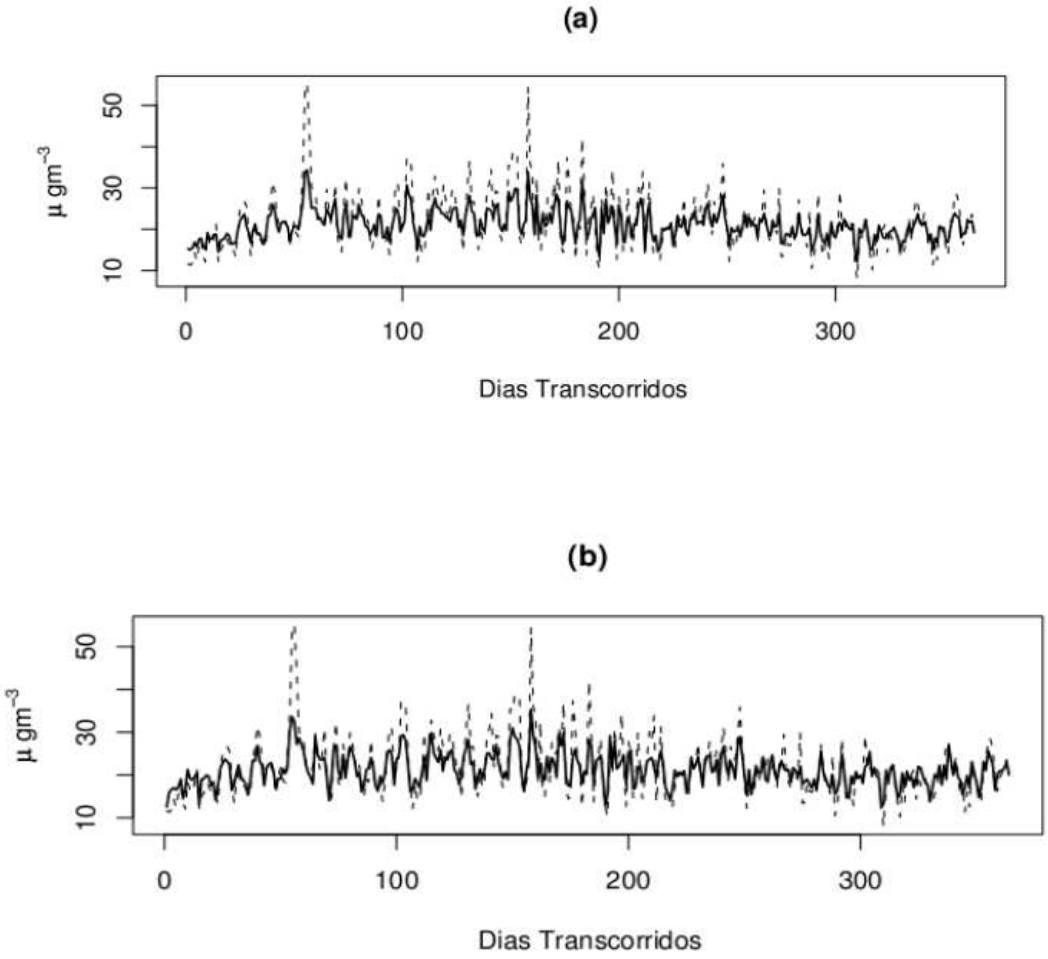


Figura 8. - - - **Valor Observado** | **Valor Ajustado: (a) Modelo I (b) Modelo II.**

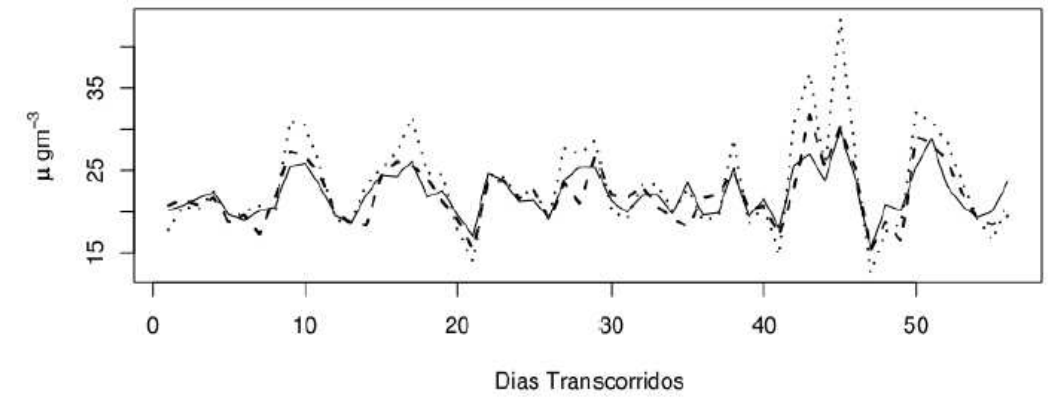


Figura 9 ... Valor Observado, — Valor previsto modelo I e - - - Valor previsto modelo II.

Referências bibliográficas

Box, G. E. & Pierce, D. A. (1970), 'Distribution of residual correlations in autorregressive integrated moving average time series models', Journal of the American Statistical Association 65, 1509–1526.

Cochrane, D. & Orcutt, G. H. (1949), 'Applications of least squares regression to relationships containing autocorrelated errors', Journal of the American Statistical Association 44, 32–61.

Fajardo, F., Reisen, V. A. & Cribari-Neto, F. (2009), 'Robust estimation in long-memory processes under additive outliers', Journal of Statistical Planning and Inference 139, 2511–2525.

Goyal, P., Chan, A. T. & Jaiswal, N. (2006), 'Statistical models for the prediction of respirable suspended particulate matter in urban cities', Atmospheric Environment 40, 2068–2077.

Iglesias, P., Jorquera, H. & Palma, W. (2006), 'Data analysis using regression models with missing observations and long-memory: an application study', Computational Statistics & Data Analysis 50, 2028–2043.

Jarque, C. M. & Bera, A. K. (1981), 'Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte carlo evidence', Economics Letters 7, 313–318.

Maddison, D. (2005), 'Air pollution and hospital admissions: an ARMAX modelling approach', Journal of Environmental Economics and Management 49, 116–131.

Neuberger, M., Rabczenko, D. & Moshammer, H. (2007), 'Extended effects of air pollution on cardiopulmonary mortality in Vienna', Atmospheric Environment 47, 8549–8556.

Ostro, B. D., Eskeland, G. S., Sanchez, J. M. & Feyzioglu, T. (1999), 'Air pollution and health effects: A study of medical visits among children in Santiago, Chile', Environmental Health Perspectives 107, 69–73.

Ostro, B. D., Hurley, S. & Lipsett, M. J. (1999), 'Air pollution and daily mortality in the coachella valley, california: A study of PM_{10} dominated by coarse particles', Environmental Research 81, 231–238.

Ostro, B., Sanchez, J. M., Aranda, C. & Eskeland, G. S. (1996), 'Air pollution and mortality: results from a study of santiago, chile', Journal of Exposure Analysis and Environmental Epidemiology 6, 97–114.

Palma, W. & Chan, N. H. (1997), 'Estimation and forecasting of long-memory processes with missing values', Journal of Forecasting 16, 395–410.

Palma, W. & del Pino, G. (1999), 'Statistical analysis of incomplete long-range dependent data', Biometrika 86, 965–972.

Perez, P. & Reyes, J. (2002), 'Prediction of maximum of 24-h average of PM10 concentrations 30 h in advance in Santiago, Chile', Atmospheric Environment 36, 4555–4561.

Reisen, V. A. & Fajardo, F. (2008), Robust estimation in seasonal long-memory processes with outliers, in 'Annals of Latin American Meeting of the Econometric Society', http://www.webmeets.com/files/papers/LACEA-LAMES/2008/707/SeasonalOutlierV2.pdf.

Reisen, V. A. & Fajardo, F. (2012), 'Robust estimation in time series with long and short memory properties', Annales Mathematicae et Informaticae 39, 20–36.

Reisen, V. A., Moulines, E., Soulier, P. & Franco, G. C. (2010), 'On the properties of the periodogram of a stationary long-memory process over different epochs with applications', Journal of Time Series Analysis 31, 20–36.

Reisen, V. A., Rodrigues, A. & Palma, W. (2006), 'Estimation of seasonal fractionally integrated processes', Computational Statistics & Data Analysis 50, 568–582.

Shumway, R. H. & Stoffer, D. S. (2010), Time Series Analysis and Its Applications: With R Examples, 3rd edn, New York: Springer.

Shumway, R. & Stoffer, D. (1982), 'An approach to time series smoothing and forecasting using the EM algorithm', Journal of time series analysis 3, 253–264.

Souza, P. A. (2002), Intelligent receptor modelling, in 'I International Workshop on Industrialized Urban Centers', Vitória, ES: Icon Graphics, pp. 1–18.

Wei, W. (2005), Time Series Analysis: Univariate and Multivariate Methods, 2nd edn, Addison Wesley.

Agradecimentos

Os autores agradecem os comentários e sugestões dos avaliadores anônimos, assim como o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e da FAPES (ES)- Brasil.

Abstract

Particulate matter is a type of contaminant that makes a large impact in the air quality of a metropolitan area. This paper discusses the mean level estimation and forecasting properties of a model to explain inhalant particulate matter (PM10) with meteorological factors. A time series model and multiple linear regression are the tools considered to model the level of the PM10 data measured in the Carapina area, in the Great Vitória Region (RGV), ES, Brazil. The former approach uses as an adjusted model a time series model with uncorrelated errors whereas the latter one deals with a multiple linear regression model with correlated errors. The prediction and forecast model properties of the PM10 data were also analyzed and compared. Both models displayed similar results in terms of modeling the level of the contaminant, nevertheless the forecasting issues indicated that the regression approach gave a slightly better result than the time series model.

Análise de influência na regressão em cristas

Silvia Nagib Elian[1]
Koki Fernando Oikawa [2]

Resumo

Os Modelos de Regressão em Cristas apresentam características próprias e problemas específicos. São geralmente utilizados para contornar o problema da multicolinearidade, consequência da existência de relações lineares entre as variáveis explicativas. O objetivo do presente trabalho é apresentar e discutir as medidas de diagnóstico e a correspondente análise de influência quando é utilizado o procedimento de regressão em cristas. Apresentaremos inicialmente medidas de influência específicas para a regressão em cristas. Neste mesmo contexto, serão ainda abordadas medidas de influência local. Finalmente, os procedimentos descritos serão aplicados a um conjunto de dados reais.

[1] Departamento de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo. C.P:66281- São Paulo, Brasil - E-mail: selian@ime.usp.br.
[2] Faculdade Capital São Paulo, Brasil - E-mail: kfoikawa@gmail.com.

1. Medidas de influência na regressão em cristas

Consideremos o modelo de regressão linear

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$$

onde $\mathbf{y}$ é um vetor de variáveis aleatórias observáveis, $\mathbf{1}$ é um vetor contendo o valor 1 em todas as posições, β_0 é um parâmetro desconhecido, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ é uma matriz $n\ x\ p$ centralizada e padronizada de constantes conhecidas $(\mathbf{1}'\mathbf{x}_i = 0,\ \mathbf{x}_i'\mathbf{x}_i = 1,\ i = 1, ..., p)$, $\boldsymbol{\beta}_1$ é um vetor de parâmetros desconhecidos e ε é um vetor de erros não observáveis com $E(\boldsymbol{\varepsilon}) = 0$ e $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

Se $\mathbf{Z} = (\mathbf{1}, \mathbf{X})$, o estimador de mínimos quadrados de $\boldsymbol{\beta}$ $[\boldsymbol{\beta}' = (\beta_0, \boldsymbol{\beta}'_1)]$ é $\mathbf{b} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$, o vetor de respostas ajustadas fica $\hat{\mathbf{y}} = \mathbf{Z}\mathbf{b}$ e o estimador de σ^2 é $s^2 = \mathbf{e}'\mathbf{e}/(n - p - 1)$, sendo que $\mathbf{e}$ é o vetor de resíduos $(\mathbf{y} - \hat{\mathbf{y}})$.

O estimador em cristas, proposto por Hoerl e Kennard (1970), surgiu como uma forma de contornar o problema de multicolinearidade que pode ocorrer nas variáveis explicativas dos modelos de regressão.

O problema principal quando se utiliza o estimador de mínimos quadrados na presença de multicolinearidade é que, embora este seja não viciado, sua variância é grande.

Por outro lado, o estimador em cristas é viciado, mas seu erro quadrático médio pode ser menor que o do estimador não viciado $\hat{\boldsymbol{\beta}}$ de mínimos quadrados, devido ao decréscimo na variância .

O estimador em cristas é definido como

$$\mathbf{b}^* = (\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*)^{-1}\mathbf{Z}'\mathbf{y}$$

onde $\mathbf{I}^* = diag(0, 1, ..., 1)$ de dimensão $p + 1$.

Existem inúmeros critérios para a determinação do valor de k. Nesse trabalho será utilizado o critério do traço da crista, proposto por Hoerl e Kennard (1970).

A análise de diagnóstico em modelos de regressão quando os parâmetros são estimados pelo procedimento de mínimos quadrados é bastante conhecida. No entanto, para o procedimento de regressão em cristas, a literatura não se mostra tão rica. Apresentaremos a seguir algumas medidas de diagnóstico para esse caso, extraídas do trabalho de Walker e Birch (1988).

Ao utilizarmos o estimador em cristas, o vetor de valores ajustados será

$$\begin{aligned}\hat{\mathbf{y}}^{*} &= \mathbf{Z}\mathbf{b}^{*} \\ &= \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^{*}\right)^{-1}\mathbf{Z}'\mathbf{y}.\end{aligned}$$

Portanto, a matriz $\mathbf{H}^{*} = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^{*}\right)^{-1}\mathbf{Z}'$ assume uma função similar à da matriz "hat" na estimação por mínimos quadrados. O *i-ésimo* valor previsto pode ser escrito em termos dos elementos de $\mathbf{H}^{*}$ como

$$\hat{y}_i^{*} = \sum_{j=1}^{n} h_{ij}^{*} y_j$$

Consequentemente, $\partial \hat{y}_i^{*} / \partial y_i = h_{ii}^{*} \equiv h_i^{*}$ e com isso, os elementos da diagonal da matriz "hat" do estimador em cristas podem ser interpretados, assim como no caso de mínimos quadrados, como um valor de alavancagem

Uma versão alternativa para a distância de Cook adaptada também ao contexto de regressão em cristas é dada pela expressão

$$D_i^{*} = \left(1/\left((p+1)s^2\right)\right)\left(\mathbf{b}^{*} - \mathbf{b}^{*}(i)\right)'\,\mathbf{Z}'\mathbf{Z}\,\left(\mathbf{b}^{*} - \mathbf{b}^{*}(i)\right),$$

em que $\mathbf{b}^{*}(i)$ é o estimador em cristas calculado sem a *i-ésima* observação.

A medida D_i^{*} também pode ser escrita como

$$D_i^{*} = \left(1/(p+1)s^2\right)\left(\hat{\mathbf{y}}^{*} - \hat{\mathbf{y}}^{*}(i)\right)'\left(\hat{\mathbf{y}}^{*} - \hat{\mathbf{y}}^{*}(i)\right),$$

sendo que $\hat{\mathbf{y}}^{*} = \mathbf{Z}\mathbf{b}^{*}(i)$.

2. Análise de influência local na regressão em cristas

O método da influência local foi desenvolvido por Cook (1986) e é aplicável apenas em procedimentos de estimação por máxima verossimilhança

Seja $L(\boldsymbol{\theta})$ o logaritmo da função de verossimilhança para um modelo inicial, sendo $\boldsymbol{\theta}$ um vetor $p\times 1$ de parâmetros desconhecidos com estimador de máxima verossimilhança dado por $\hat{\boldsymbol{\theta}}$.

São introduzidos distúrbios no modelo através do vetor $\mathbf{w}$, $m\times 1$, $\mathbf{w}\subset\boldsymbol{\Omega}$, $\boldsymbol{\Omega}\subset\Re^m$, onde $\boldsymbol{\Omega}$ representa um conjunto aberto de possíveis pequenos distúrbios. Do ponto de vista prático, $\mathbf{w}$ refletiria qualquer esquema de perturbação, por exemplo, nas variáveis explicativas ou na matriz de covariâncias da variável resposta do modelo de regressão.

Seja $L(\boldsymbol{\theta}\,|\,\mathbf{w})$ o logaritmo da função de verossimilhança que corresponde ao modelo que sofreu perturbação e $\hat{\boldsymbol{\theta}}_w$ o estimador de máxima verossimilhança correspondente a esse modelo. Supondo que exista um ponto $\mathbf{w}_0$ em $\boldsymbol{\Omega}$ que representa a ausência de perturbação nos dados, de modo que $L(\boldsymbol{\theta})=L(\boldsymbol{\theta}\,|\,\mathbf{w}_0)$, e assumindo que $L(\boldsymbol{\theta}\,|\,\mathbf{w})$ seja duplamente diferenciável e contínua em uma vizinhança de $(\hat{\boldsymbol{\theta}}',\mathbf{w}'_0)$, o deslocamento de verossimilhanças de Cook é definido como

$$LD(\mathbf{w})=2\left[L(\hat{\boldsymbol{\theta}})-L(\hat{\boldsymbol{\theta}}_w)\right],$$

e compara as estimativas $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_w$, podendo, assim, avaliar a influência dos distúrbios $\mathbf{w}$. Grandes valores de $LD(\mathbf{w})$ indicam que $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_w$ diferem consideravelmente em relação ao contorno da função de verossimilhança sem perturbação $L(\boldsymbol{\theta})$.

Esse método é baseado no estudo do comportamento local de um gráfico de influência $\alpha(\mathbf{w})=\begin{pmatrix}\mathbf{w}\\ LD(\mathbf{w})\end{pmatrix}$ ao redor de $\mathbf{w}_0$. O procedimento consiste em considerar $\mathbf{w}$ como $\mathbf{w}(a)=\mathbf{w}_0+a\mathbf{d}$, $a\in\Re$ e $\mathbf{d}$ um vetor direção de comprimento unitário.

Cook (1986) sugere investigar a direção na qual a medida de influência $LD(\mathbf{w})$ muda localmente mais rapidamente, que é a curvatura máxima de LD, dada por

$$C_{\max}=\max_{\|d\|=1} 2\left|\mathbf{d}'\mathbf{F}\mathbf{d}\right|$$

em que $\mathbf{F}$ é uma matriz $m\times m$ definida por

$$\mathbf{F}=\mathbf{\Delta}'\mathbf{Q}^{-1}\mathbf{\Delta}$$

$\mathbf{\Delta}$ é a matriz $p\times m$ $(p=\dim(\boldsymbol{\theta}), m=\dim(\mathbf{w}))$ com elementos

$$\mathbf{\Delta} \qquad \mathbf{\Delta}_{ij}=\frac{\partial^2 L(\boldsymbol{\theta}\mid\mathbf{w})}{\partial\boldsymbol{\theta}_i\partial\mathbf{w}_j}$$

avaliados em $\hat{\boldsymbol{\theta}}$ e $\mathbf{w}_0$, e $-\mathbf{Q}$ representa a matriz de informação observada do modelo sem distúrbios $\mathbf{Q}=\left[\partial^2 L(\boldsymbol{\theta})\mid\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\right]$, avaliada em $\hat{\boldsymbol{\theta}}$. Verifica-se que a maximização de $\left|\mathbf{d}'\mathbf{F}\mathbf{d}\right|$, sujeita à restrição que $\mathbf{d}'\mathbf{d}=1$, resulta em $\mathbf{d}_{\max}$, que representa o autovetor correspondente ao maior autovalor absoluto $C_{\max}$ de $\mathbf{F}$. A direção do vetor $\mathbf{d}_{\max}$ seria aquela que produziria a maior mudança local nas estimativas dos parâmetros.

Cook (1986) sugere como referência geral uma curvatura igual a 2, sendo que curvaturas maiores que esse valor indicariam notável sensibilidade local.

Billor e Loynes (1999) propuseram ainda uma medida alternativa, descrita por

$$LD^*(\mathbf{w})=-2\left[L(\hat{\boldsymbol{\theta}})-L(\hat{\boldsymbol{\theta}}_w\mid\mathbf{w})\right]$$

A quantidade $LD^*(\mathbf{w})$ compararia então as funções de verossimilhança das duas situações consideradas, com e sem perturbação. Para $m \geq 2$, os autores sugerem o uso da medida

$$l_{\max} = \left| \nabla LD^*(\mathbf{w}_0) \right|$$

em que $\nabla LD^*(\mathbf{w}_0)$ é o vetor gradiente da função LD^* em $\mathbf{w}_0$

Para o cálculo das medidas de influência, de acordo com essa abordagem, os autores escrevem o estimador em cristas como um estimador de pseudo-máxima verossimilhança.

Para tal fim, consideraram um modelo de regressão linear múltipla

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (2.1)$$

em que $\mathbf{X}$ é uma matriz conhecida $n \times p$ padronizada, $\boldsymbol{\beta}$ é um vetor $p \times 1$ de parâmetros conhecidos, $\boldsymbol{\varepsilon}$ é o vetor de erros $p \times 1$ independentes e com distribuição normal com média zero e variância desconhecida σ^2. Admitiu-se adicionalmente que, nesse modelo, o termo constante não foi incluído.

Marquardt (1970) demonstrou que o estimador em cristas é equivalente ao estimador de mínimos quadrados quando os dados são suplementados por um conjunto de dados fictícios tomados de acordo com a matriz de planejamento ortogonal Hk e a variável resposta $\mathbf{Y}$ sendo zero em cada ponto fictício adicionado.

O modelo aumentado com matriz de planejamento $(n+p) \times p$

$$\mathbf{X}_a = \begin{pmatrix} \mathbf{X} \\ (k\mathbf{I})^{1/2} \end{pmatrix}$$

e o vetor $(n+p) \times 1$ de variáveis resposta $\mathbf{Y}_a' = (\mathbf{Y}'\ 0')$ pode ser escrito como

$$\mathbf{Y}_a = \mathbf{X}_a\boldsymbol{\beta} + \boldsymbol{\varepsilon}_a,$$

em que $\boldsymbol{\varepsilon}_a$ representa um vetor aleatório cujas componentes são variáveis aleatórias independentes e normalmente distribuídas com média zero e variância σ^2. A função densidade de $\mathbf{Y}_a$ será denominada função pseudo-densidade e a correspondente função de pseudo-verossimilhança será descrita por

$$L_p(\boldsymbol{\beta}) = \frac{n+p}{2}\log 2\pi - \frac{n+p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}'_i\,\boldsymbol{\beta})^2 + k\boldsymbol{\beta}'\boldsymbol{\beta}\right]$$

O estimador de máxima pseudo-verossimilhança é resultante de $\frac{\partial L_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$.

Como

$$\sum_{i=1}^{n}\left(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 + k\boldsymbol{\beta}'\boldsymbol{\beta}$$

pode ser escrito na forma

$$\begin{aligned}&(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\cdot(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + k\boldsymbol{\beta}'\boldsymbol{\beta}\\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'[\mathbf{X}'\mathbf{X} + k\mathbf{I}]\boldsymbol{\beta},\end{aligned}$$

derivando-se essa expressão com relação a $\boldsymbol{\beta}$ e igualando a zero, obtem-se

$$2[\mathbf{X}'\mathbf{X} + k\mathbf{I}]\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{y} = 0$$

Resolvendo essa equação, em decorrência da utilização da matriz aumentada, obtêm-se a solução $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$, que é o estimador em cristas. Uma vez que o estimador em cristas é o estimador de máxima pseudo-verossimilhança para o modelo considerado, a medida de influência local de Cook pode ser aplicada na regressão em cristas.

Considerando o modelo originalmente descrito em (2.1), que supõe homogeneidade na variância do erro, ou seja, $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$ e que pequenos distúrbios são introduzidos na variância de $\boldsymbol{\varepsilon}_i$ por meio de um vetor de distúrbios $\mathbf{w}$, $n \times 1$, onde σ^2 é suposto conhecido, a função de pseudo-verossimilhança com distúrbios para o modelo aumentado é

$$L_p(\boldsymbol{\beta}) = cons\tan te - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}'_i\,\boldsymbol{\beta})^2 \mathbf{w}_i^* + k\boldsymbol{\beta}'\boldsymbol{\beta}\right] + \frac{1}{2}\sum_{i=1}^{n}\log \mathbf{w}_i^*$$

com $\mathbf{w}_i^* = 1 + \mathbf{w}_i$, $\mathbf{w}_i$ sendo o *i-ésimo* componente do vetor $n \times 1$ de distúrbios $\mathbf{w}$.

Para o cálculo da curvatura máxima $C_{\max}$ é necessária a obtenção dos componentes individuais da matriz da informação observada $-\mathbf{Q}$ e da matriz

$$\boldsymbol{\Delta} = \left[\frac{\partial^2 L_p(\boldsymbol{\beta} \mid \mathbf{w})}{\partial \boldsymbol{\beta}_i \partial \mathbf{w}_j}\right]$$

avaliados em $\hat{\boldsymbol{\beta}}$ e $\mathbf{w}_0$.

Nessa situação, as matrizes $\mathbf{Q}$ e $\boldsymbol{\Delta}$ são dadas por

$$-\mathbf{Q} = \frac{(\mathbf{X}'\mathbf{X} + k\mathbf{I})}{\sigma^2},$$

$$\boldsymbol{\Delta} = \frac{\mathbf{X}'\mathbf{D}(\mathbf{e}^*)}{\sigma^2},$$

onde $\mathbf{e}^*$ é o vetor de resíduos em cristas, isto é, $\mathbf{e}^* = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*$, $\hat{\boldsymbol{\beta}}^*$ é o estimador em cristas e $\mathbf{D}(\mathbf{e}^*) = diag(\mathbf{e}_1^*, ..., \mathbf{e}_n^*)$. A curvatura é obtida como

$$\begin{aligned} C_d &= 2\left|\mathbf{d}'\mathbf{F}\mathbf{d}\right| \\ &= 2\left|\mathbf{d}'\Delta'\mathbf{Q}^{-1}\Delta\mathbf{d}\right| \\ &= \frac{2\left|\mathbf{d}'\mathbf{D}(\mathbf{e}^*)\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{D}(\mathbf{e}^*)\mathbf{d}\right|}{\sigma^2}. \end{aligned}$$

Após cálculos, verifica-se que a curvatura máxima de LD é dada por

$$C_{\max} = \frac{2\lambda^*_{\max}}{\sigma^2},$$

onde $\lambda^*_{\max}$ é o maior autovalor de

$$\mathbf{D}(\mathbf{e}^*)\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{D}(\mathbf{e}^*)$$

Já, para LD*, a máxima inclinação será dada por

$$\left|\nabla LD^*_{\Re}\right| = \left[\sum_{i=1}^{n}\left(1 - \frac{(\mathbf{e}^*_i)^2}{\sigma^2}\right)^2\right]^{1/2}.$$

Com relação a uma interpretação adequada dessas medidas, Cook (1986) sugere que $C_{\max} = 2$ pode ser usado como um valor limite. Contudo, Billor e Loynes (1993) apontaram o valor $\sqrt{2n + 4(14n)^{1/2}}$ como relevante na determinação de influência local.

3. Aplicação

As técnicas de análise apresentadas foram aplicadas por Oikawa (2008) ao conjunto de dados do projeto desenvolvido em André, Elian e Bruscato (1997). O projeto é da área farmacológica e investiga o efeito de diversos tipos de anestésicos locais sobre o coração de ratos. O interesse desse estudo consistia em verificar quais características físico-químicas da molécula de determinada droga influenciam mais em sua potência tóxica, definida como a dose de droga necessária para ocorrer uma redução de 30% na frequência do átrio. Para tal, foram utilizados setenta e dois ratos, homogêneos entre si, divididos em quatorze grupos, contendo de três a oito ratos. Cada grupo foi submetido a uma droga diferente e a potência tóxica foi calculada após a realização de um experimento, descrito no referido trabalho.

Foram consideradas as variáveis explicativas:
largura do comprimento substituinte a partir do eixo da ligação, perpendiculares a ele (medida em Ângstron).

- **F**: componente de campo (adimensional);
- **R**: componente de ressonância (adimensional);
- **SIGMA**: constante de Hammet – combinação linear das duas anteriores (adimensional);
- **LOG PAPP**: logaritmo do coeficiente de partição óleo-água medido (adimensional);

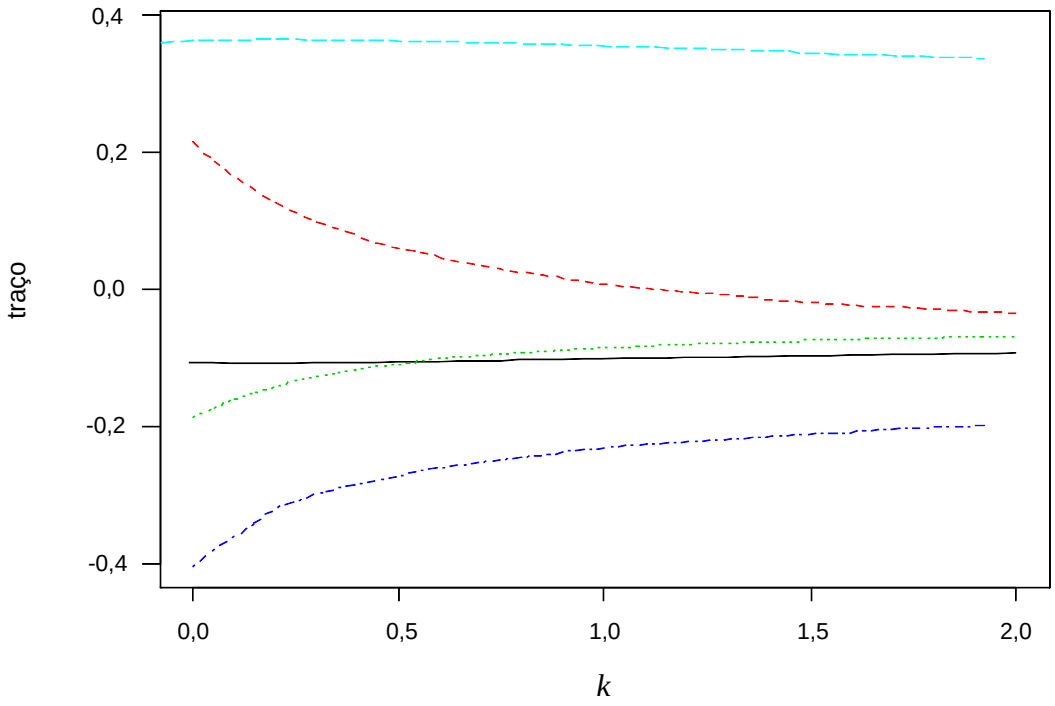
e a variável resposta é dada por:

- **POTÊNCIA**: -log (DE_{30}), onde DE_{30} é a dose de droga necessária para ocorrer uma redução de 30% na freqüência do átrio em relação ao controle (adimensional).

Na análise da relação entre a variável resposta e as variáveis explicativas, utilizou-se um modelo de regressão linear múltipla. No entanto, foi detectada a presença de multicolinearidade através do cálculo do Fator de Inflação da Variância e do número condicional, que é obtido pela razão $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$, onde $\lambda_{\max}$ é o maior autovalor da matriz $(\mathbf{X}'\mathbf{X})$, na sua forma de correlação, enquanto que $\lambda_{\min}$ é o menor autovalor dessa matriz. Os autovalores da matriz $(\mathbf{X}'\mathbf{X})$ obtidos foram: 3,1756; 1,0058; 0,6851; 0,1267 e 0,0066. $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{3{,}1756}{0{,}0066} = 481{,}15$, o que sugere existência de forte multicolinearidade nos dados.

Como forma de contornar o problema da multicolinearidade, um modelo de regressão em cristas foi ajustado. Para isso, foi utilizado o traço da crista como critério de escolha para o valor de, κ com κ variando de zero a dois.

Figura 1 – Traço das estimativas dos coeficientes de regressão das variáveis: B4, SIGMA, F, R e LOG.PAPP



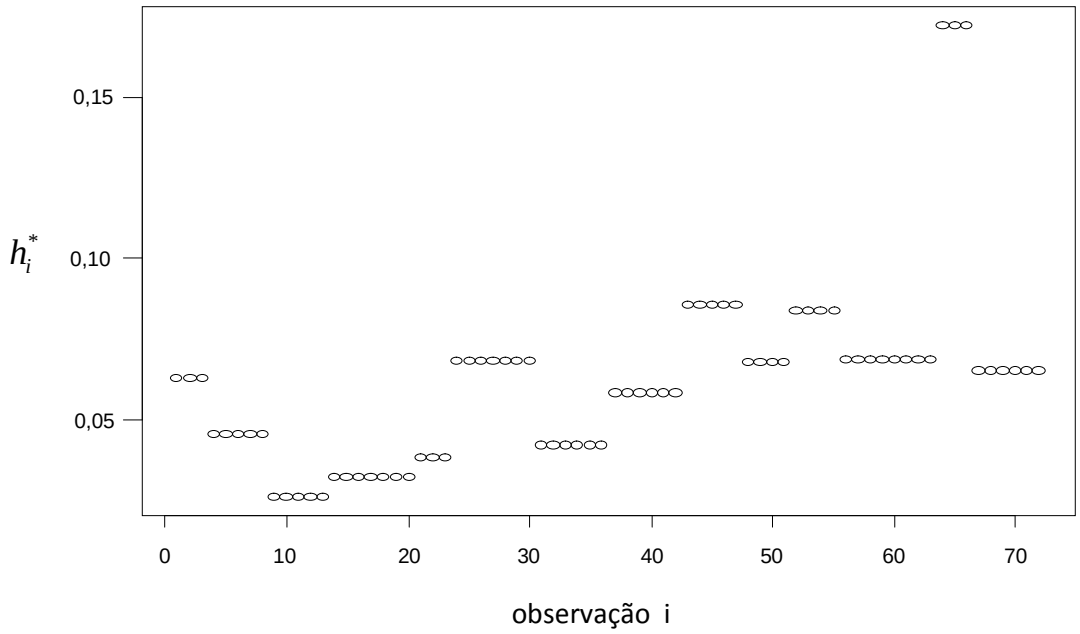
Através da Figura 1 percebe-se que, a partir de $k = 1$, os coeficientes tendem a se estabilizar. Assim, esse valor foi escolhido, obtendo-se o modelo de regressão em cristas

$$\hat{Y} = 3{,}27 - 0{,}07 \cdot \text{B4} + 0{,}02 \cdot \text{SIGMA} - 0{,}44 \cdot \text{F} - 0{,}81 \cdot \text{R} + 0{,}30 \cdot \text{LOG.PAPP} \ .$$

Posteriormente foram aplicadas algumas das técnicas de diagnóstico descritas com o auxílio de programas desenvolvidos no pacote computacional R.

Calculando-se os elementos da diagonal principal da matriz $\mathbf{H}^* = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}^*\right)^{-1}\mathbf{Z}'$ verificou-se que as observações 64, 65 e 66 eram as mais influentes, com valor $h_i^* = 0{,}172$ (Figura 2). Correspondiam a três ratos com os maiores valores da variável SIGMA e foram os únicos a apresentarem valores negativos em LOG.PAPP . Apresentavam também os maiores valores da variável F e da variável R .

Figura 2 – Valores dos elementos da diagonal principal da matriz H*

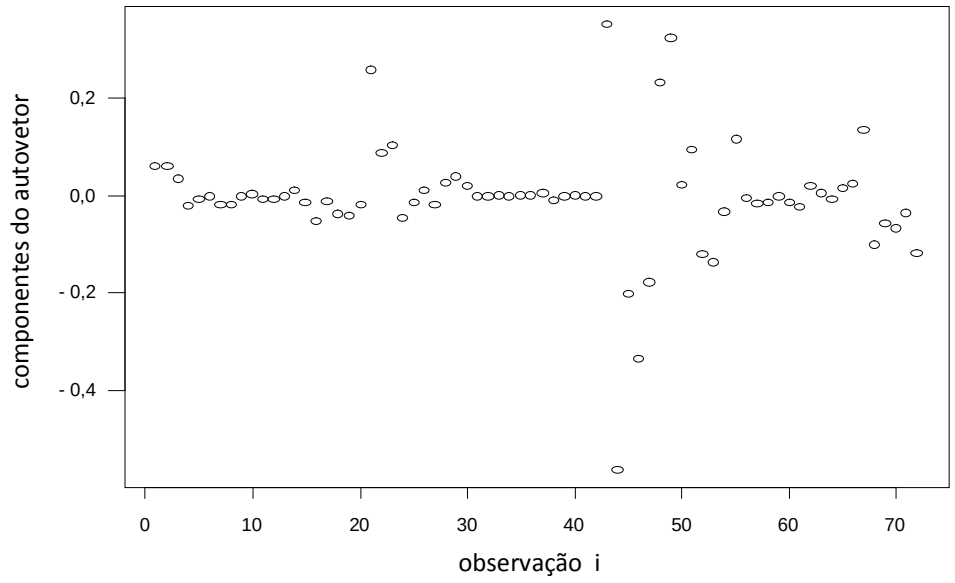


Não foram detectados pontos influentes por meio da medida D_i^*. Se o procedimento adotado fosse o de mínimos quadrados, sete seriam as observações influentes: 21, 44, 64, 65, 69, 70 e 71.

A curvatura máxima obtida foi $C_{\max} = \dfrac{2 \cdot \lambda_{\max}^*}{\hat{\sigma}^2} = \dfrac{2 \cdot 0{,}0525435}{0{,}03663602} = 2{,}87$. Dessa forma, podemos concluir por uma sensibilidade moderada nos dados, de acordo com o critério de Cook $\left(C_{\max} > 2\right)$.

O autovetor associado a $\lambda^*_{\max}$ também fornece informação sobre a influência dos pontos,de modo que as coordenadas com maiores valores correspondem aos pontos mais influentes. Segundo esse critério, foram detectados os ratos de números: 44, 43, 46, 49, 21, 48 e 45, todos com componente de $\lambda^*_{\max}$ maiores que $|0{,}2|$, como pode ser verificado na Figura 3.

Figura 3- Análise da Influência pelas componentes do autovetor associado a $\lambda^*_{\max}$



A inclinação máxima $l_{\max}$ obtida foi

$$\left|\nabla LD^*_{\Re}\right| = \left[\sum_{i=1}^{n}\left(1-\frac{\left(\mathbf{e}^*_i\right)^2}{\hat{\sigma}^2}\right)^2\right]^{1/2},$$

$$\left|\nabla LD^*_{\Re}\right| = 13{,}53.$$

Assim, como $l_{\max} = 13{,}53 < 16{,}46 \quad = \sqrt{2n + 4(14n)^{1/2}}$, esta medida não sugere sensibilidade local para os dados . Os valores absolutos individuais de l_i , em que $l_i = \left(1 - \frac{\left(\mathbf{e}_i^*\right)^2}{\hat{\sigma}^2}\right)$, encontram-se na Figura 4 e Tabela 1.

Figura 4 - Valores absolutos de l_i

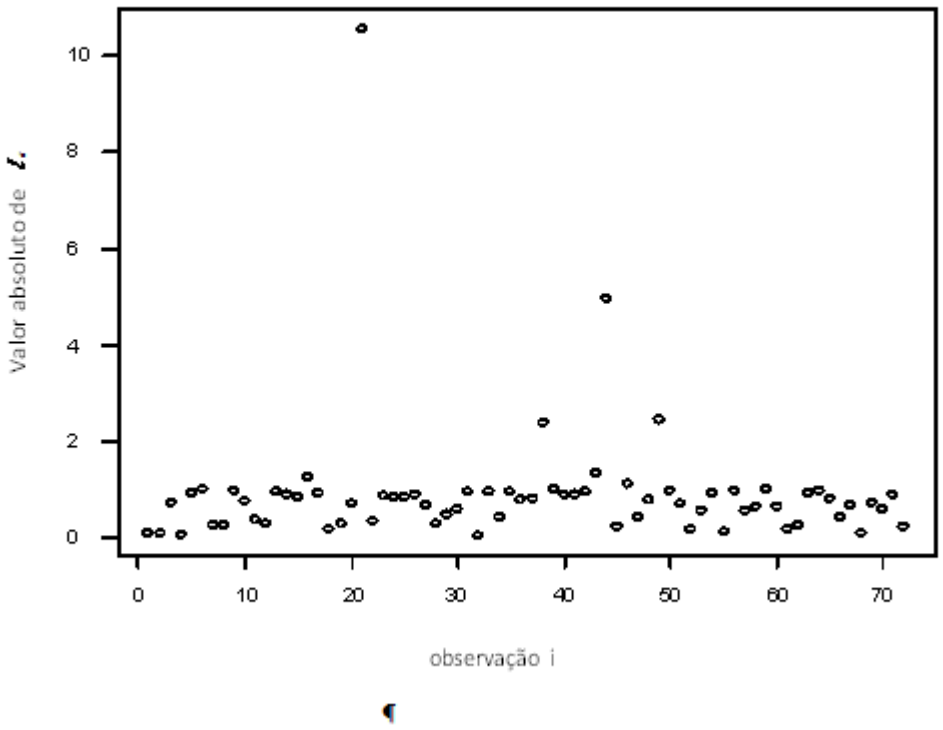


Tabela 1– Valores absolutos individuais l_i

Caso	$\|l_i\|$	Caso	$\|l_i\|$	Caso	$\|l_i\|$	Caso	$\|l_i\|$	Caso	$\|l_i\|$
1	0,10	16	1,25	31	0,94	46	1,10	61	0,16
2	0,10	17	0,91	32	0,04	47	0,41	62	0,27
3	0,72	18	0,17	33	0,95	48	0,77	63	0,92
4	0,05	19	0,29	34	0,41	**49**	**2,43**	64	0,97
5	0,91	20	0,69	35	0,95	50	0,98	65	0,80
6	0,99	**21**	**10,5**	36	0,76	51	0,70	66	0,42
7	0,24	22	0,32	37	0,81	52	0,18	67	0,68
8	0,24	23	0,85	**38**	**2,39**	53	0,54	68	0,10
9	0,97	24	0,82	39	1,00	54	0,91	69	0,72
10	0,74	25	0,83	40	0,88	55	0,12	70	0,59
11	0,37	26	0,89	41	0,89	56	0,96	71	0,90
12	0,28	27	0,68	42	0,95	57	0,57	72	0,23
13	0,95	28	0,29	43	1,34	58	0,63		
14	0,89	29	0,46	**44**	**4,96**	59	0,99		
15	0,84	30	0,60	45	0,24	60	0,63		

Com base nelas, detectamos quatro observações apresentando valores perceptivelmente maiores que as demais: 21, 38, 44 e 49, sendo que as observações 21 e 44 já haviam sido detectadas em análise prévia pelo método de mínimos quadrados e também pelas componentes de $\lambda^*_{\max}$. A observação de número 21 apresentou valor absoluto de l_i excepcionalmente maior. Verificamos que correspondia a um elemento amostral com alto valor do resíduo no modelo de regressão em cristas ajustado.

Dessa maneira, a análise de influência realizada mostrou-se plenamente satisfatória. Adicionalmente, todos os pontos diagnosticados correspondiam realmente a elementos atípicos, o que evidenciou a extrema importância das técnicas utilizadas.

Referências bibliográficas

André, C.D.S.; Elian, S.N.; Bruscato, A.(1997) Relatório de Análise Estatística sobre o projeto: "Relação Estrutura- Atividade de Anestésicos locais N,N [dimetilamina] etilBenzoatos parasubstituídos". São Paulo, IME-USP, 38p.

Billor, N. and Loynes, R. M. (1993). "Local Influence: A New Approach", Communications in Statistics – Theory and Methods, 22, 1595-1611.

Cook, R. D. (1986). "Assessment of Local Influence (wit discussion)". Journal of the Royal Statistical Society, Series B, 48, 133-169.

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems". Technometrics, 12, 55-67.

Marquardt, D. W. (1970). "Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation". Technometrics, 12, 591-612.

Oikawa, K. F. (2008). "Análise de Influência na Regressão em Cristas". Dissertação de Mestrado. IME-USP.

Walker, E. and Birch, J. B. (1988). "Influence Measures in Ridge Regression". Technometrics, 30, 221-227.

Abstract

Ridge Regression Models, even so can be considered as a particular case of the general linear regression model, they present proper characteristics and specif problems.These models are used, in general, to solve the problem of multicollinearity, which is a consequence of existence of linear relation among regressor variables. The objective of this paper is to present some influence measures in ridge regression. First, It wiil be discussed influence measures specific to ridge regression models. Also, we present local influence measures in this kind of analysis. Finally, some of the described procedures will be applied to a real data set.

Emparejamiento de paneles y clasificación de la ausencia de respuesta en la Pesquisa Mensal de Emprego usando funciones en R

Andrés Gutiérrez,[1]
Jorge Ortiz,[2]

Resumen.

La Pesquisa Mensal de Emprego es un instrumento de seguimiento de las condiciones de empleo en los principales centros metropolitanos de Brasil, de gran utilidad y uso, tanto administrativo estatal como investigativo y académico. Su estructuración en paneles rotativos la hace especialmente vulnerable, no solo a la información faltante, sino también a cambios en los datos básicos que desarticulan la identificación de los individuos generando un falso incremento en la ausencia de respuesta. En el presente artículo se proponen tres funciones en lenguaje R para implementar la aplicación de criterios de emparejamiento definidos anteriormente por otros autores con el fin de reducir el desgaste del panel ocasionado por esta desarticulación y al mismo tiempo, facilitar el acceso de los datos a un público investigador más amplio. Como resultado de las funciones es posible realizar una clasificación de la ausencia de respuesta de los individuos que permitirá un análisis más riguroso sobre parámetros de interés tales como cambios brutos en el nivel de ocupación.

[1] Profesor Facultad de Estadística - Universidad Santo Tomás. E-mail: hugogutierrez@usantotomas.edu.co,
[2] Profesor Facultad de Estadística - Universidad Santo Tomás. E-mail: jorgeortiz@usantotomas.edu.co

1. Introducción

La Pesquisa Mensal de Emprego (PME) es una encuesta que provee indicadores mensuales para la obtención de información en el mercado laboral en las áreas metropolitanas de Brasil. Su objetivo principal es producir estimaciones de la fuerza de trabajo mensual para evaluar las fluctuaciones y la tendencia del mercado laboral metropolitano en el mediano y largo plazo. Con ella se obtienen indicaciones rápidas de los efectos de las condiciones económicas en el mercado de trabajo y se satisfacen necesidades importantes para la planificación y el desarrollo socio-económico. Esta encuesta ha sido aplicada desde 1980, con algunos cambios metodológicos mayores en 1982, 1988, 1993 y 2001 (IBGE 2007) .

En el sitio web de la PME, es posible encontrar los microdatos anonimizados de las encuestas mensuales desde el año 2002, en el mayor nivel de desagregación posible y acompañados de documentación que proporciona los nombres y los códigos de las variables de interés y sus categorías. También se puede consultar la metodología de la encuesta y el instrumento de recolección de datos. Lo anterior constituye una valiosa herramienta de investigación, máxime cuando la PME realiza un seguimiento continuo de hogares a lo largo del tiempo.

Muchos investigadores se han visto beneficiados con la publicación en línea de los microdatos de la PME como base para diversos análisis estadísticos de la encuesta. Con algunos conocimientos en lenguajes de programación o de software especializado, pueden calcular estimaciones, crear sus propios cuadros y comprobar que los indicadores de empleo son consistentes con la realidad.

Las características de la encuesta hacen que la tarea de la reconstrucción de los paneles inducidos por el diseño rotativo de la PME no resulte sencilla por la ausencia en los microdatos de un código único de identificación de los individuos en el hogar en periodos de medición distintos. El emparejamiento básico de la información considera algunas características socio-demográficas reportadas por los individuos, que permiten su reconocimiento a lo largo de los periodos de medición en el hogar. Cuando la información básica no es consistente de un periodo a otro, se hace difícil encontrar a un mismo individuo a través del tiempo y asimismo calcular estimaciones apropiadas a partir de los microdatos. Por lo tanto, se hace necesario el uso de criterios adicionales de emparejamiento para identificar a las personas.

Después de un proceso exhaustivo de identificación, incluyendo los criterios adicionales de emparemiento, el investigador se enfrenta al problema de la ausencia de respuesta: para algunos individuos puede faltar información en algunas mediciones. Por ejemplo, cuando se comparan dos períodos específicos, se encuentran unas personas con datos en las dos mediciones respectivas, otras con respuesta en solo una de ellas, y además, otras que no responden en ninguna de las dos. Cada una de estas situaciones debe tenerse en cuenta cuando se calculan estimaciones comparativas de indicadores de empleo de un periodo a otro.

El objetivo de este artículo es proveer una herramienta automática, sencilla y fácil de usar para reconstruir los paneles de la PME, que al mismo tiempo clasifique la ausencia de respuesta en los periodos de medición. Además, se revisan los aspectos principales de la PME, junto con las características propias de las encuestas a hogares (de tipo transversal y longitudinal); se examina brevemente la metodología del muestreo probabilístico, y se exploran algunos criterios de emparedamiento en datos anonimizados utilizados por Perez & Dillon (2009), quienes desarrollaron un procedimiento computacional en STATA. Además de lo anterior, en este trabajo de investigación se implementan procedimientos de clasificación de la ausencia de respuesta con miras a la estimación puntual de los cambios brutos en dos periodos de referencia de la encuesta.

En este artículo, hemos escogido implementar un conjunto de funciones[1] en el software estadístico R (R Development Core Team 2012), considerando que es un software libre, de uso gratuito, disponible en distintas plataformas y sistemas operativos (Windows, Linux, Mac), aceptado por otros paquetes computacionales como SAS y IBM-SPSS, siendo además, el software estadístico más SAS y IBM-SPSS, siendo además, el software estadístico más utilizado actualmente en investigación.

Después de una breve introducción, en la sección dos se describe el proceso metodológico en encuestas de hogares como la PME, la sección tres aborda el tópico de la ausencia de respuesta y sus clasificaciones cuando se realizan comparaciones en dos periodos de tiempo. La sección cuatro trata de la inferencia puntual que se puede realizar con este tipo de encuestas repetidas, y en particular se muestra cómo estimar el tamaño del panel. En la quinta sección se analizan varias formas de emparejamiento de las personas a través del panel mediante criterios objetivos implementados paso a paso. También se aborda la clasificación de la ausencia de respuesta para el nivel de ocupación en la PME y se establecen condiciones de consistencia. La sección seis describe funciones computacionales, programadas en R, como una solución que se propone en este artículo para realizar el emparejamiento de los paneles y la clasificación de la ausencia de respuesta. La sección siete muestra, paso a paso, el proceso de emparejamiento del panel P6 seguido desde 2010 hasta 2012. De la misma forma, se establece la clasificación de la ausencia de respuesta en los ocho meses en los cuales se realizó la medición a este panel particular. En la última sección se discuten los resultados encontrados en este trabajo de investigación y se concluye acerca de los alcances de la implementación de las funciones computacionales propuestas.

[1] Resaltamos que la metodología propuesta es útil en el emparejamiento de paneles fijos y rotativos.

2. PME: una encuesta de hogares.

La PME es una encuesta con características particulares que se exponen en esta sección. En general, se describe el proceso de encuestas que brindan estadísticas oficiales, de tipo panel rotativo en donde el hogar, como unidad de muestreo, permanece en la muestra durante varios periodos y posteriormente no es considerado como respondiente.

2.1. Plan de muestreo

Gambino & Silva (2009) afirman que en una encuesta de hogares, interesan las características de algunos miembros del hogar que pueden estar relacionadas con salud, educación, ingresos/gastos, estado de empleo, usos de diferentes servicios, etc. Los diseños de muestreo utilizados para estos estudios son complejos y pueden abarcar técnicas como conglomeración, estratificación o selección de unidades con diferentes probabilidades. En una gran cantidad de encuestas de este tipo se considera la vivienda como unidad de muestreo y las personas o los hogares como unidades de observación (Agresti 2002).

Tradicionalmente, para este tipo de encuestas, se utilizan marcos de áreas y de lista. Alternativamente, pueden utilizarse otros, como los de números telefónicos o incluso internet. Si se encuentra disponible una lista de unidades poblacionales, entonces puede ser utilizada para seleccionar derectamente una muestra, posiblemente después de la estratificación de las unidades de muestreo en grupos homogéneos. Por ejemplo, se puede tener acceso a listas de hogares, de personas o listas de números telefónicos. Por lo general, es difícil mantener actualizado este tipo de marcos por características inherentes a las unidades de muestreo y por razones de mudanzas, matrimonios, divorcios, nacimientos o muertes que generan modificaciones en la población. Además, pueden presentarse problemas de duplicación y sobre-cobertura puesto que una persona puede quedar enlistada varias veces o incluso, no pertenecer a la población de interés.

Por las dificultades mencionadas, en las encuestas de hogares se acostumbra el uso de marcos de áreas, que se obtienen mediante la división de un país, provincia o estado, en muchas áreas pequeñas, mutuamente excluyentes y exhaustivas, proporcionando una cobertura completa de la población de interés. El uso de marcos de áreas conlleva naturalmente a los diseños en varias etapas, en donde se definen conglomerados a partir de la ubicación geográfica. El proceso de selección de muestras empieza en alguno de los niveles geográficos. Tanto en países desarrollados como en vía de desarrollo, los marcos de áreas son de uso frecuente, pues reducen los costos de transporte del personal de campo.

Es usual que, en la última etapa de este tipo de estudios, se seleccionen personas dentro de cada hogar. Por lo tanto, se hace necesario determinar el número óptimo de personas por seleccionar en cada uno para evitar problemas de sobre-representación o sub-representación de subgrupos poblacionales (Clark & Steel 2007). Béland, Dale, Dufour & Hamel (2005) citan un ejemplo en donde, mediante la selección aleatoria simple de una persona por hogar, se induce sobre-representación en grupos poblacionales por edades. Dentro del diseño de muestreo utilizado en encuestas de hogares, frecuentemente se encuentran conglomerados definidos como áreas geográficas compactas. Gambino & Silva (2009) afirman que en áreas urbanas, los conglomerados se forman mediante la combinación natural de bloques o manzanas contiguas, mientras que en áreas rurales, se forman a partir de información censal o se toman conglomerados naturales como veredas, entre otros. Cuando el tamaño de los conglomerados es muy heterogéneo, se puede recurrir al muestreo con probabilidades de selección proporcionales al tamaño. De manera similar, se toman como conglomerados los edificios de apartamentos en áreas metropolitanas grandes.

Otra caracterización del plan de muestreo en estas encuestas es la estratificación. Áreas geográficas, como provincias, estados o regiones forman un primer nivel de estratificación. En particular, la PME define a la persona residente en el hogar como una unidad de investigación. Esta encuesta se basa en una muestra probabilística de hogares, bietápica y estratificada para cada área metropolitana cubierta por la encuesta. Las municipalidades y pseudomunicipios (conjuntos de municipios más pequeños) se consideran como estratos independientes de selección, asegurando así la dispersión de la muestra para el área metropolitana. A su vez, dentro de cada municipio o pseudomunicipio se realiza una selección de unidades primarias de muestreo (correspondientes a secciones censales) y posteriormente, las unidades secundarias de muestreo (correspondientes a unidades domiciliares u hogares).

En IBGE (2007) se encuentra que la selección de los sectores se lleva a cabo mediante un muestreo sistemático con probabilidad proporcional al número total de hogares particulares. De la lista actualizada de los hogares en los sectores seleccionados, se extraen los hogares mediante muestreo sistemático simple. La PME tiene algunos aspectos a priori de un plan de muestreo autoponderado (Cochran 1977, pp. 91, 303) dentro de cada área metropolitana. Esto implica que, dependiendo Del crecimiento o la disminución del sector, el número de unidades de vivienda seleccionadas puede aumentar o disminuir.

2.2. Encuestas repetidas

Otra característica fundamental de la PME es que realiza un seguimiento continuo a las mismas unidades domiciliarias durante ocho meses. Algunas encuestas de hogares se repiten a través del tiempo con un contenido y metodología similares. En el mundo existen encuestas muy importantes que usan esquemas de seguimiento[2]. Perez & Dillon (2009) afirman que la PME presenta un esquema de muestreo equivalente al de la *US Current Populaton Survey* y describen otro tipo de encuestas que utilizan esquemas parecidos, como la *National Longitudinal Survey of Labor Market* y el *Panel Study of Income Dynamics*, en Estados Unidos. Alrededor del mundo también existen encuestas importantes con una metodología similar: la *Belgian Socio-economic Panel*, en Bélgica; la *Netherlands Socio-Economic Panel,* en Holanda; la *German Social Economics Panel*, en Alemania; la *British Household Panel Survey*, en Inglaterra; la *European Community Household Panel*, en los países de la comunidad europea; la *Household, Income and Labour Dynamics*, en Australia y la *Survey of Labor Income Dynamics*, en Canada; la *Encuesta Permanente de Empleo*, en Perú; la *Encuesta Panel CASEN*, en Chile; la *Encuesta Nacional sobre Niveles de Vida de los Hogares*, en México; entre muchas otras.

El objetivo del seguimiento a las unidades de muestreo es producir estimaciones de indicadores claves para la sociedad. Por ejemplo, en encuestas de fuerza laboral, interesa obtener estimaciones del número de personas empleadas o desempleadas en diferentes instantes de tiempo. Gambino & Silva (2009) hacen una clara distinción entre las encuestas transversales repetidas y las encuestas de tipo longitudinal. Las primeras se realizan mediante la recolección de datos de una población objetivo específica en ciertos intervalos utilizando la misma metodología o una comparable y no requiere el seguimiento de las mismas unidades de muestreo a través del tiempo. Sin embargo, generalmente se diseñan de tal manera que exista algún traslape de unidades de muestreo entre encuestas sucesivas. Las encuestas longitudinales, por el contrario, requieren que la misma muestra de unidades sea observada a través del tiempo, en por lo menos dos periodos sucesivos.

[2] Por ejemplo, Stasny (1987) considera la Labor Force Survey que se fundamenta en una muestra mensual de aproximadamente 56 mil hogares que son retinidos en la muestra por 6 meses, con una tasa de traslape mensual del 83 por ciento.

En las encuestas repetidas, se utiliza una estrategia de muestreo (especificación del procedimiento de selección de muestras junto con un procedimiento de estimación) que provea la precisión adecuada para la inferencia de los parámetros de interés. Algunos de los resultados que brindan este tipo de encuestas repetidas son la estimación de parámetros poblacionales específicos en cada punto del tiempo, la estimación del cambio de los parámetros de interés entre diferentes oleadas de encuestas y la estimación del valor promedio de los parámetros de interés sobre diferentes oleadas de encuestas. Las diferentes opciones de rotación en la muestra y la frecuencia de las entrevistas afectan la precisión de los estimadores. Específicamente, IBGE (2007) considera que la PME es una encuesta de hogares que se distribuye a través de las cuatro semanas del mes de referencia. Así, el resultado del mes se consigue mediante el conjunto de la información de las cuatro semanas. La recopilación de los datos sigue una metodología en la que cada hogar es seleccionado en la muestra durante cuatro meses consecutivos. Luego, se excluye de la muestra por los ocho meses siguientes y, después, vuelve a seleccionarse por otros cuatro meses. Después de esto, es eliminado de la muestra. En general, durante el periodo de observación del hogar, es posible que la familia cambie de domicilio y otra familia se traslade a ocupar esa unidad de alojamiento. En estos casos, la información se obtiene con la nueva familia para el resto del periodo de observación.

La PME se subdivide en ocho grupos de rotación. Cada mes se retira de la muestra un grupo de rotación y se incorpora uno nuevo, es decir, el 25% de la muestra de hogares se sustituye, y se conserva el 75% de la muestra, siguiendo un esquema de grupos de rotación y los paneles. Cada panel representa un número de unidades de vivienda y los grupos de rotación conforman conjuntos de sectores censales. Así, para el mismo mes, en pares de años consecutivos, se garantiza el 50% de la parte común de la muestra. Por ejemplo, Perez & Dillon (2009) consideran el grupo de rotación E1, que fue entrevistado de Febrero a Mayo del 2003 (cuatro meses), no fue entrevistado desde Junio de 2003 hasta Enero de 2004 (ocho meses) y nuevamente fue seleccionado desde Febrero hasta Mayo de 2004.

3. PME: una encuesta con ausencia de respuesta

Según Lohr (2000), la mayoría de las encuestas tienen cierta ausencia de respuesta residual, aun después de un diseño cuidadoso y un seguimiento de la ausencia de respuesta. La PME no es la excepción y en esta sección mostraremos que es posible diferenciar la ausencia de respuesta en el panel. Särndal & Lundström (2004) afirman que la ausencia de respuesta ha sido un tema de interés en las agencias de estadística que producen cifras oficiales. En las últimas décadas, la atención de la literatura hacia este tópico y sus efectos se ha incrementado considerablemente. En parte, se debe a una propensión decreciente del público para cooperar y enviar los datos solicitados por las agencias de estadística. El problema de la ausencia de respuesta es una faceta normal, aunque no deseable, en el desarrollo de una encuesta. Existe un consenso general en considerar que la ausencia de respuesta puede perjudicar severamente la calidad de las estadísticas calculadas y publicadas a partir de los datos de una encuesta.

Lohr (2000) clasifica la ausencia de respuesta en función de su relación con la característica de interés. Se define la ausencia de respuesta ignorable cuando la probabilidad de que un individuo responda no depende de la característica de interés. Por consiguiente, la ausencia de respuesta se considera no ignorable cuando la probabilidad de que un individuo responda depende de la característica de interés. Por ejemplo, si en una encuesta de fuerza laboral, se desea estimar el número de personas empleadas o desempleadas, la ausencia de respuesta es no ignorable cuando depende de la situación laboral del individuo.

Särndal & Lundström (2004) destacan la existencia de una gran cantidad de literatura acerca de la ausencia de respuesta y el interés reciente por este tema. En esta literatura se examinan dos aspectos complementarios en el ejercicio de una encuesta: la prevención de la ausencia de respuesta (antes de que ocurra) y las técnicas de estimación adecuadas para tener en cuenta la ausencia de respuesta de manera apropiada en el proceso de inferencia. La segunda actividad se conoce como ajuste para la ausencia de respuesta. Figueredo (2003) advierte que si la incidencia de este fenómeno provoca serios problemas en las encuestas realizadas en un único periodo, es decir en encuestas transversales, la situación con las encuestas repetidas se torna mucho más compleja. Además, como en las encuestas por panel, las unidades que integran la muestra son observadas repetidamente a lo largo de una serie de entrevistas, entonces se presentan distintas formas de ausencia de respuesta que, en la mayoría de los casos, es intermitente: las unidades entrevistadas responden en algunos periodos de medición, más no en todos, generando patrones complejos de respuesta.

Lumley (2010, capítulo 9) hace un análisis detallado de la ausencia de respuesta individual, en donde se tienen datos parciales para un respondiente, considerando un enfoque basado en el diseño de muestreo al ajustar los pesos muestrales. Fuller (2009, capítulo 5) cita algunas técnicas de imputación para el tratamiento de la ausencia de respuesta y conjuga modelos probabilísticas con los pesos del diseño de muestreo para mitigar los efectos de este problema. Särndal (2011) considera un enfoque asistido por modelos, en donde toma conjuntos balanceados para lograr mayor `representatividad" de las estimaciones, en el sentido de igualdad entre la media muestral y la media poblacional de una variable auxiliar disponible. De la misma forma, Särndal & Lundström (2010) proponen un conjunto de indicadores para juzgar la efectividad de la información auxiliar utilizada para controlar el sesgo generado por la ausencia de respuesta.

3.1. Diferentes clases de ausencia de respuesta

Como la PME, existen muchas encuestas que utilizan diseños de tipo panel rotativo donde los individuos son entrevistados varias veces antes de ser rotados fuera de la muestra. Estas encuestas a gran escala son utilizadas para producir estimaciones puntuales en el tiempo y realizar comparaciones entre meses y años. De esta forma, la estructura del panel rotativo resulta de la necesidad de reducir costos manteniendo los mismos entrevistadores y objetivos por más de una entrevista. Stasny (1987) consideró el problema de estimar cambios brutos entre dos periodos de tiempo utilizando datos categóricos, obtenidos de una encuesta de tipo panel, con ausencia de respuesta. En estas condiciones, algunos individuos entrevistados se clasifican adecuadamente en una tabla de contingencia; otros, sólo pueden ser clasificados parcialmente, y los demás no pueden ser clasificados. Stasny, en el artículo mencionado, utilizó un enfoque basado en modelos para ajustar la posible no respuesta, que no puede ser considerada como completamente al azar, sino que depende de la clasificación de los individuos en la tabla de contingencia.

Un enfoque para estimar el cambio bruto a través del tiempo con datos panel se basa solamente en la utilización de la información obtenida de individuos que fueron respondientes en dos periodos de tiempo. Para utilizar este enfoque, es necesario asumir que los individuos que no respondieron en uno o ambos periodos constituyen una muestra aleatoria de todos los individuos (Rubin 1976). Sin embargo, en muchas ocasiones, la ausencia de respuesta no ocurre de forma aleatoria simple, por ejemplo, los datos pueden mostrar que se relaciona con la clasificación en la fuerza laboral (Fienberg & Stasny 1983). Para producir estimaciones conábles, se necesita emparejar a los respondientes en el panel, pero también es necesario determinar y clasificar todos los posibles patrones de no respuesta.

Supóngase que, como resultado de cada entrevista, se clasifica al respondiente en una de G categorías de una variable nominal, y que se quiere estimar el cambio bruto para estas categorías utilizando registros de individuos entrevistados en dos periodos consecutivos de tiempo. La clasificación en las categorías no es clara para los individuos que fueron no respondientes en ambos periodos. De esta forma, se tiene un grupo con clasificación en ambos tiempos, otro con datos en uno de los dos, y un tercero sin respuesta en ningún periodo.

Para el primer grupo de individuos, con respuesta en los tiempos t - 1 y t, los datos de clasificación se resumen en una matriz de tamaño GxG. La información para los individuos que no respondieron la encuesta del tiempo t - 1 pero sí en el tiempo t puede resumirse en un complemento fila, mientras que la de los individuos que no respondieron en el tiempo t pero sí en el tiempo t - 1 se resume en un complemento columna. Finalmente, los que no respondieron en ningún tiempo son incluidos en una única celda de faltantes. Lo anterior se ilustra en la tabla 1, en donde N_{ij} (i,j = 1,...,G) denota el número de individuos respondientes en el universo que tienen clasificación i en el tiempo t - 1 y j en el tiempo t, R_i denota el número de individuos que fueron no respondientes en el tiempo t y tienen clasificación i en el tiempo t - 1, C_j denota el número de individuos que fueron no respondientes en el tiempo t - 1 y tuvieron clasificación j en el tiempo t, y M denota el número de individuos seleccionados que no respondieron en ningún tiempo.

Tabla 1: Cambio bruto poblacional en dos periodos consecutivos.

Tiempo $t-1$	Tiempo t						
	1	2	⋯	j	⋯	G	Complemento fila
1	N_{11}	N_{12}	⋯	N_{1j}	⋯	N_{1G}	R_1
2	N_{21}	N_{22}	⋯	N_{2j}	⋯	N_{2G}	R_2
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
i	N_{i1}	N_{i2}	⋯	N_{ij}	⋯	N_{iG}	R_i
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
G	N_{G1}	N_{G2}	⋯	N_{Gj}	⋯	N_{GG}	R_G
Complemento columna	C_1	C_2	⋯	C_j	⋯	C_G	M

4. Inferencia puntual en encuestas como la PME

En muchas encuestas, buena parte de las preguntas se presentan con opciones de respuesta categorizadas: binarias (como “sí" o “no"), multinomiales (como “empleado", ”desempleado" o “no perteneciente a la fuerza laboral") o de escalas ordinales. Se tiene entonces la necesidad de procedimientos inferenciales para los análisis univariados y multivariados de estas variables categóricas adaptados a datos provenientes de encuestas complejas. El análisis de datos categóricos incluye diversas técnicas y métodos, desde la estimación de las proporciones en las diferentes categorías hasta métodos multivariados complejos o modelos de regresión multinivel. Heeringa, West & Berglund (2010) destacan que la mayoría de los estimadores y estadísticas de prueba para datos categóricos se construyen con el método de máxima verosimilitud y asumen que los datos son independientes e idénticamente distribuidos con alguna distribución de probabilidad discreta. De acuerdo con Agresti (2002), cuando se lleva a cabo un muestreo aleatorio simple, se asume que las variables categóricas siguen alguna de las distribuciones discretas como la binomial, multinomial, Poisson e hipergeométrica. En otros casos, debido a los diferentes pesos de muestreo y a los efectos de conglomeración y de estratificación, se dificulta encontrar la función de verosimilitud de los datos muestrales. Por esta razón, el uso tradicional de métodos inferenciales de máxima verosimilitud puede no ser conveniente cuando los datos provienen de encuestas complejas. Así, los análisis estándares deben ser modificados para tener en cuenta los efectos de los diseños muestrales complejos, utilizando estimaciones ponderadas de proporciones, estimaciones de varianza basadas en el diseño muestral y correcciones generalizadas para los efectos del diseño (Pessoa & Silva 1998).

El escenario se vuelve más complejo cuando el análisis se lleva a cabo para dos o más variables categóricas. En el caso de dos variables categóricas, los datos se resumen generalmente en una tabla de contingencia. Bajo muestreo aleatorio simple, las frecuencias muestrales (sin ninguna ponderación) pueden ser utilizadas para estimar estadísticas de interés tales como las proporciones marginales por cada categoría, o para realizar pruebas, como la Ji-cuadrado de Pearson, para determinar la relación entre las variables categóricas. Sin embargo, cuando los individuos son seleccionados con un diseño muestral con probabilidades desiguales, tanto las estimaciones como las pruebas estadísticas calculadas usando frecuencias sin ninguna ponderación son sesgadas para las propiedades verdaderas de la población objetivo. Por otro lado, cuando el diseño de muestreo es complejo e induce probabilidades de inclusión desiguales (lo cual suele suceder en encuestas de hogares, de fuerza laboral, entre otras encuestas que producen estadísticas oficiales), la proporción de interés debe calcularse teniendo en cuenta expresiones teóricas que consideren las probabilidades de inclusión.

En encuestas de fuerza laboral, como la PME, es posible encontrar clasificaciones aún más complejas que dependen de los estados laborales en dos periodos consecutivos de medición. Por ejemplo, el interés puede estar centrado en la estimación del número de individuos que en un periodo pasado se encontraban laborando y continúan haciéndolo en el periodo actual, el número de individuos desempleados, tanto en el periodo pasado como en el presente, el número de individuos que en un periodo pasado se encontraban laborando y en el periodo actual están desempleados y viceversa. Skinner, Holt & Smith (1989) y Chambers & Skinner (2003, capítulo 7) proveen una óptima profundización para el análisis de datos con respuesta categórica en encuestas en donde el diseño de muestreo es informativo y complejo. Asimismo, Lumley (2010, capítulo 6) discute varias técnicas analíticas modernas para el estudio de datos binarios y categóricos.

Por otro lado, existen muchas situaciones prácticas en donde es posible abordar este tipo de estimaciones en tablas de contingencia. En términos de los totales marginales, es posible estimar los *cambios netos* mediante la comparación directa entre periodos (Kalton 2009). Por ejemplo, se determina si la tasa de desempleo subió o bajó y en qué magnitud. Se pueden realizar análisis detallados de los *cambios brutos* a partir de la descomposición de los cambios netos. De esta manera, si el desempleo subió un punto porcentual, es posible conocer si esto se debió a que el uno por ciento de los empleados en el primer periodo perdió su trabajo, o si el diez por ciento de los empleados perdió su trabajo y a la vez el nueve por ciento de los desempleados encontró un nuevo trabajo.

Con base en lo anterior, se puede concluir que los cambios brutos pueden ser estimados en dos periodos de tiempo para dos muestras diferentes, entrevistando a distintos individuos. Sin embargo, los cambios brutos sólo pueden ser estimados a partir de muestras en dos ocasiones, como se hace con la PME, en donde se entrevistan a los mismos individuos en dos periodos de tiempo consecutivos. Sin embargo, en una situación real es posible encontrar que algunos individuos no respondan la encuesta y, más aún, que esta ausencia de respuesta no sea aleatoria, sino que siga distintos patrones inuenciados por la clasificación del individuo en la tabla de contingencia. La siguiente sección abordará esta problemática.

4.1. Estimación del tamaño del panel

Si se mide la población *U* en dos instantes de tiempo t - 1 y t, los parámetros de interés se pueden representar como las entradas de la tabla 1. Nótese que fácilmente se puede concluir que el tamaño total de la población de interés, N, debe satisfacer la siguiente expresión:

$$N = \sum_i \sum_j N_{ij} + \sum_j C_j + \sum_i R_i + M$$

Donde, N_{ij} , C_j , R_i y M han sido definidos al final de la sección 3.1. Existen varios subgrupos poblaciones que inducen una partición del universo en ambos tiempos. En primer lugar, denotaremos a U_{ij} como el subgrupo de individuos respondientes en ambos periodos que tienen clasificación $_{ij}$. Además, denotaremos a U_{Cj} como el subgrupo de individuos respondientes en el tiempo t que tiene clasificación j y que no respondieron en el tiempo t - 1, U_{Ri} como el subgrupo de individuos respondientes en el tiempo t - 1 que tiene clasificación i y que no respondieron en el tiempo t. Por último se tiene que U_M es el conjunto de no respondientes en ambos tiempos. Nótese que $\#(U_{ij}) = N_{ij}, \#(U_{Cj}) = C_j, \#(U_{Ri}) = R_i y \#(U_M) = M$.

$$U = \bigcup_i \bigcup_j U_{ij} \cup \bigcup_j U_{Cj} \cup \bigcup_i U_{Ri} \cup U_M$$

Por supuesto, bajo condiciones ideales, los parámetros C_j, para $j = 1,\ldots, G$, R_i, para $i = 1,\ldots, G$ y M deberían ser siempre nulos. Lo anterior se presentaría si todos los individuos de la población fuesen respondientes en ambos periodos. De esta manera, se tendría que $\Sigma_i \Sigma_j N_{ij} = N$. Además, se tendría que $U = \bigcup_{i=1}^{G} \bigcup_{j=1}^{G} U_{ij}$, puesto que los conjuntos U_{Ri} , para $i = 1,\ldots,G, U_{Cj}$, para j = 1, ...,G, tendrán cardinalidad nula, así como el conjunto U_M. Sin embargo, dado que la realidad de las encuestas es compleja, y que no todos los entrevistados serán respondientes, entonces bajo un modelo poblacional de ausencia de respuesta, sería posible "repartir" a los no respondientes, categorizados en los complementos fila, complementos columna y no respondientes en ninguna entrevista, en las celdas de la clasificación poblacional. Si la ausencia de respuesta no fuese diferencial, entonces esa repartición debería ser equitativa y proporcional entre las celdas de interés. Sin embargo, si por alguna razón, el modelo indica que la clasificación del individuo induce la ausencia de respuesta, entonces esa repartición, no debería ser ni equitativa ni proporcional.

Mediante la definición de las siguientes características de interés, es posible definir los parámetros de interés.

$$y_{1ik} = \begin{cases} 1, \text{ si el individou k-ésimo responde en t- 1y tiene clasificación i;} \\ 0, \text{ en otro caso} \end{cases} \quad (1)$$

$$y_{2ik} = \begin{cases} 1, \text{ si el individuo } k\text{-ésimo responde en } t\text{- 1y tiene clasificación i;} \\ 0, \text{ en otro caso} \end{cases} \quad (2)$$

Por lo anterior, el producto de las anteriores cantidades, definido como FORMULA, provee una nueva característica de interés que toma el valor uno, si el individuo contestó en ambos periodos y está clasificado en la celda *ij* y cero para cualquier otro caso. De esta forma, se tiene que

$$N_{ij} = \sum_{k \varepsilon U} y_{1ik} y_{2jk} \quad (3)$$

Además se definen las siguientes características dicotómicas

$$z_{1k} = \begin{cases} 1, \text{ si el individuo } k\text{-ésimo responde en } t\text{-1;} \\ 0, \text{ en otro caso.} \end{cases} \quad (4)$$

$$z_{2k} = \begin{cases} 1, \text{ si el individuo } k\text{-ésimo responde en } t\text{;} \\ 0, \text{ en otro caso.} \end{cases} \quad (5)$$

Por lo tanto, se tiene que

$$R_i = \sum_{k \varepsilon U} y_{1ik} \left(1 - z_{2k}\right) \quad (6)$$

$$C_j = \sum_{k \varepsilon U} y_{2ik} \left(1 - z_{1k}\right) \quad (7)$$

$$M = \sum_{k \varepsilon U} \left(1 - z_{1k}\right)\left(1 - z_{2k}\right) \quad (8)$$

Sin embargo, dado que muy pocas veces es posible acceder a la población de interés, y mucho menos en dos ocasiones consecutivas, entonces se hace necesario disponer de un proceso de muestreo que permita estimar los parámetros de interés. Por lo tanto, en la mayoría de casos es imprescindible seleccionar una muestra s y esa selección induce ponderaciones que puede ser utilizadas para estimar los parámetros de interés. Si w_k es una ponderación del k-ésimo individuo inducida por una estrategia (diseño de muestreo y estimador) de muestreo, entonces las siguientes expresiones representan estimadores de los parámetros de interés

$$\hat{N}_{ij} = \sum_{k \varepsilon S} w_k y_{1ik} y_{2ik} \tag{9}$$

$$\hat{R}_i = \sum_{k \varepsilon S} w_k y_{1ik} (1 - z_{2k}) \tag{10}$$

$$\hat{C}_j = \sum_{k \varepsilon S} w_k y_{1ik} (1 - z_{1k}) \tag{11}$$

$$\hat{M} = \sum_{k \varepsilon S} w_k (1 - z_{1k})(1 - z_{2k}) \tag{12}$$

para N_{ij} , R_i, C_j y M, respectivamente. Nótese entonces que una estimación insesgada para el tamaño de la población está dada por la siguiente expresión

$$\hat{N} = \sum_i \sum_j \hat{N}_{ij} + \sum_j \hat{C}_j + \sum_i \hat{R}_i + \hat{M} = \sum_s w_k v_k \tag{13}$$

En donde

$$v_k = \sum_i y_{1ik} \sum_j y_{2ik} + \sum_j y_{2jk} (1 - z_{1k}) + \sum_i y_{1ik} (1 - z_{2k}) + (1 - z_{1k})(1 - z_{2k}) \tag{14}$$

Nótese que si w_k es el inverso de la probabilidad de inclusión de primer orden del k-ésimo individuo, entonces las anteriores expresiones se convierten en estimadores de Horvitz-Thompson cuya varianza está dada por S arndal, Swensson & Wretman (1992, resultado 2.8.1.). Más aún, si w_k corresponde a una ponderación de calibración, entonces las varianzas de los estimadores están definidas por Deville & S arndal (1992, expresión 3.1.)

5. Emparejamiento y clasificación de la respuesta en la PME

La PME es una encuesta a domicilios con periodicidad mensual y con un esquema de rotación 4-8-4, según el cual un domicilio entra en la encuesta durante cuatro meses consecutivos, sale de la muestra los siguientes ocho meses, y retorna para ser entrevistado cuatro veces consecutivas adicionales (Perez & Dillon 2009). La información se recolecta en las regiones metropolitanas de Rio de Janeiro, Sao Paulo, Porto Alegre, Belo Horizonte, Recife y Salvador.

El tema principal de la encuesta gira en torno al trabajo y la ocupación de los individuos y es posible que su estado de desempleo se relacione con la ausencia de respuesta. Por ejemplo, los individuos desempleados pueden ser más propicios a no responderla. Por lo tanto, accediendo a los microdatos provistos por el IBGE, es posible estimar de manera insesgada los tamaños N_{ij} , R_i, C_j y M.

5.1. Criterios de emparejamiento

Para reconstruir el panel, lo primero que se debe hacer es la identificación de los domicilios en los periodos de medición. Esta información está disponible en los microdatos de las encuestas; específicamente, hemos considerado las siguientes variables para la construcción del identificador unico a nivel domiciliar:

- Región metropolitana (V035): toma los valores 26 (Recife), 29 (Salvador), 31 (Belo Horizonte), 33 (Rio de Janeiro), 35 (São Paulo), (Curitiba) y 43 (Porto Alegre).
- Número de controle (V040): es una secuencia numérica de identificacióin de la entrevista.
- Número de série (V050), que identifica un domicilio seleccionado.
- Panel (V060): toma valores de A a Z e identifica un conjunto de domicilios.
- Grupo rotacional (V063), con valores de 1 al 8 y corresponde a la división de los sectores seleccionados.

Los microdatos reportados por el IBGE no incluyen un código único de identificación para los individuos que conforman los domicilios y, por lo tanto, no es posible realizar un emparejamiento rápido de personas. Siguiendo las ideas de Lopes (2003) y de Perez & Dillon (2009), adoptamos algunos criterios de emparejamiento de registros para reducir el nivel de desgaste del panel ocasionado por los errores en el diligenciamiento del cuestionario. Hemos considerado seis criterios aplicados secuencialmente.

Primer criterio: Dos registros corresponden a una misma persona en el panel si se cumplen lãs siguientes condiciones:

- Misma llave domiciliar, dada por V035, V040, V050, V060 y V063.
- Mismo sexo, dado por V203.
- Mismo día de nacimiento, dado por V204.
- Mismo mes de nacimiento, dado por V214.
- Mismo año de nacimiento, dado por V224.
- Mismo número de orden, dado por V201.

Segundo criterio: Perez & Dillon (2009, sección 4) analizan detalladamente los problemas eventuales de emparejamiento que ocasionan desgaste en el panel y definen un conjunto de criterios que denominan de emparejamiento avanzado y que se aplican a los registros que no cumplen el primer criterio. Dos registros corresponden a la misma persona si cumplen las siguientes condiciones:

- Misma llave domiciliar.
- Mismo sexo, mismo día, mismo mes y mismo año de nacimiento.

Tercer criterio: Para los registros que no cumplen alguno de los dos primeros criterios, se aplican las siguientes condiciones de emparejamiento:

- Misma llave domiciliar.
- Mismo mismo día y mismo mes de nacimiento.
- Mismo número de orden.

Cuarto criterio: Se aplica a los jefes de hogar, cónyuges e hijos con 25 años o más, cuando no se ha cumplido ninguno de los criterios anteriores. Las condiciones de emparejamiento de registros son:

- Misma llave domiciliar.
- Mismo sexo.
- Hasta cuatro días de diferencia en el día de nacimiento.
- Hasta dos meses de diferencia en el mes de nacimiento.
- Hasta dos años de diferencia en la edad presumida, si la edad presumida de la persona es menor de 25 años; o `exp(edad)/30`, si la edad presumida de la persona es mayor de 25 años. La edad presumida está dada por la variable V234.

Este último ítem representa una función de error en la edad presumida y es discutida con detalle en Perez & Dillon (2009, p. 94).

Quinto criterio: Se ejecuta sobre los registros que aún no han emparejado con alguno de los criterios anteriores, y que son jefes de hogar, cónyuges o hijos con 25 años o más y cuyo día de nacimiento no se encuentra en la base de datos:

- Misma llave domiciliar.
- Mismo sexo.

- Hasta dos meses de diferencia en el mes de nacimiento.
- Hasta dos años de diferencia en la edad presumida, si la edad presumida de la persona es menor de 25 años; o exp(edad)/30, si la edad presumida de la persona es mayor de 25 años. La edad presumida está dada por la variable V234.
- Hasta un ciclo de diferencia en el nivel de escolaridad, dado por VDAE1.

Sexto criterio: Se ejecuta sobre los registros de la base de datos que aún no han emparejado con los criterios anteriores y que corresponden a jefes de hogar, cónyuges e hijos con 25 años o más, cuando nose encuentra el mes de nascimento en la basa de datos:

- Misma llave domiciliar.
- Mismo sexo.
- Hasta cuatro días de diferança en el día de nascimento.
- Hasta dos años de diferença en la edad presumida, si la edad presumida de la persona es menor de 25 años; o exp(edad)/30, si la edad presumida de la persona es mayor de 25 años. La edad presumida está por la variable V234.
- Hasta un ciclo de diferencia en el nivel de escolaridad, dado por VDAE1.

Séptimo critério: Se aplica a los registros que aún no han emparejado con alguno de los criterios anteriores, y que son jefes de hogar, cónyuges e hijos con 25 años o más y para los cuales no se encuentra ni el día ni el mes de nacimiento en la base de datos:

- Misma llave domiciliar.
- Mismo sexo.
- Hasta dos años de diferencia en la edad presumida, si la edad presumida de la persona es menor de 25 años; o exp(edad)/30, si la edad presumida de la persona es mayor de 25 años. La edad presumida está dada por la variable V234.
- Hasta un ciclo de diferencia en el nivel de escolaridad, dado por VDAE1.

Perez & Dillon (2009, p. 93) agregan un conjunto de criterios aún más laxos para buscar emparejamientos en el resto de la base de datos. No los incluimos, teniendo en cuenta que con los descritos anteriormente se logra una reducción importante en el nivel de desgaste del panel.

5.2. Clasificación de la respuesta para el nivel de ocupación

Después de realizar el emparejamiento de acuerdo a los criterios de la sección anterior, el investigador dispone de una base de datos de individuos medidos en el tiempo. Las mediciones personales pueden ir desde una, para los registros que no pudieron emparejarse a lo largo del proceso, hasta ocho, para los que emparejaron en todos los meses de medición. Así, para realizar comparaciones entre pares de meses se tienen los siguientes conjuntos que caracterizan la respuesta:

- Respondientes (Tipo 0): definidos como las personas que respondieron en ambos periodos.
- No respondientes en la segunda ocasión (Tipo 1): se definen como las personas que respondieron en la primera ocasión, pero no en la segunda. No necesariamente deben ser personas que emparejan alguna una vez en el procedimiento, pero sí deben encontrarse en el primer periodo de interés.
- No respondientes en la primera ocasión (Tipo 2): son las personas que no respondieron en el primer periodo pero que sí respondieron en la segunda ocasión. No necesariamente emparejan alguna vez en el procedimiento, pero sí deben encontrarse en el segundo periodo de interés.
- No respondientes en ambos periodos (Tipo 3): son las personas que no respondieron en ninguno de los periodos. En este conjunto se encuentran, tanto las personas que respondieron solo una vez, como las que emparejaron alguna vez en el procedimiento, pero que no se encuentran en ninguno de los periodos de interés.

La respuesta de un individuo en una medición garantiza la clasificación del mismo en una tabla de contingencias. Por ejemplo, si la característica de interés es el nivel de ocupación – correspondiente a la variable VD1 de la PME - (Ocupado, Desocupado, Inactivo o No perteneciente a la fuerza laboral) y el individuo contesta en los dos periodos de comparación y tendrá una clasificación completa. Si responde en un periodo pero no en el otro, sólo tendrá una clasificación parcial. Por último, si el individuo no responde en ninguno de los dos periodos de referencia, no podrá ser clasificado. Lo anterior se ilustra en tabla 2, en donde, tal como se indicó en la sección 3.1, N_{ij} $(i, j=1,...,3)$ denota el número de individuos respondientes en la población que tienen clasificación i en el tiempo t - 1 y j en el tiempo t, R_i denota el número de individuos que fueron no respondientes en el tiempo t y tienen clasificación i en el tiempo t - 1, Cj denota el número de individuos que fueron no respondientes en el tiempo t - 1 y tuvieron clasificación j en el tiempo t, y M denota el número de individuos seleccionados que no respondieron en ningún tiempo.

Tabla 2: Cambio bruto poblacional en dos mediciones para el nivel de ocupación en la PME.

Tiempo $t-1$	Tiempo t				
	Ocupado	Desocupado	Inactivo	No perteneciente	Complemento fila
Ocupado	N_{11}	N_{12}	N_{13}	N_{14}	R_1
Desocupado	N_{21}	N_{22}	N_{23}	N_{24}	R_2
Inactivo	N_{31}	N_{32}	N_{33}	N_{34}	R_3
No perteneciente	N_{41}	N_{42}	N_{43}	N_{44}	R_4
Complemento columna	C_1	C_2	C_3	C_4	M

Consistencia del procedimiento

El emparejamiento y el análisis de la clasificación de las respuestas deben ser consistentes con los microdatos de la PME. Las siguientes condiciones permiten avalar los procedimientos para cualquiera de las ocho mediciones de un panel:

1. La suma de la cantidad de personas con un solo registro en el panel con la de empates debe coincidir con el número de registros en los ocho meses de medición.

2. Para cada mes, la cantidad de registros en la cuatro clasificaciones de respuesta (Tipo 0. Tipo 1, Tipo 2 y Tipo 3) debe coincidir con el número total de personas identificadas en el procedimiento de reconstrucción del panel.
3. La suma del número de respondientes en dos periodos de tiempo $t-1$ y t (Tipo 0) con el número de personas que sólo respondieron en el primer periodo $t-1$ (Tupo 1) debe ser igual al número de personas reportadas en los microdatos de la PME para el primer periodo $t-1$.
4. La suma del número de respondientes en dos periodos de tiempo $t-1$ y t (Tipo 0) con el número de personas que sólo respondieron en el segundo mes t (Tipo 1) debe ser igual al número de personas reportadas en los microdatos de la PME para el segundo periodo t.

6. Descripción de las funciones en R

En esta sección se describen brevemente las funciones programadas en R para el emparejamiento de paneles y la clasificación de la repuesta.

6.1. Recursos web

En primer lugar, el sitio web oficial del software estadístico R es www.cran.r-project.org/. Desde aquí se puede descargar, instalar y actualizar R. Además, en esta página se encuentra una gran cantidad de documentación e información sobre librerías. En particular, sobre las librerías `sqldf` (Grothendieck 2012), `car` (Fox & Weisberg 2011) y `strigr` (Wickham 2011), necesarias para ejecutar neustras funciones .

El sitio web oficial de la PME, donde se encuentra toda la información metodológica de la encuesta, desde las tablas y flujogramas de las estimaciones hasta los documentos técnicos relevantes de la encuesta, es www.ibge.gov.br/home/estatistica/indicadores/trabalhoerendimento/pme_nova. De este sitio se pueden descargar los archivos comprimidos (`.zip`) de microdatos. Los archivos de datos se encuentran en formato de texto fijo (`.txt`). La documentación completa de la PME, mes a mes, desde el año 2002, incluye el archivo INPUT.txt que permite la lectura de las bases de datos en el software SAS.

Todas nuestras funciones se encontran en el archivo PME_Matching_Nonresponse.txt, que se puede descartar desde el siguinte sitio www.gutierrezandres.com/software/matchingpme. Alternativamente, se puede ejecutar el código fuente directamente dese R, mediante el suministro de la dirección URL como argumento a la función source.Tambiém se pude descargar un archivo comprimido (en formato `.rar`) con los conjuntos de datos (en formato `.csv`), utilizando en los ejemplos de la sección 7.

6.2. Descripción de las funciones

`ReadPME`. Con esta función se leen los microdatos directamente desde R. De esta maneira , cualquier investigador puede acceder al manejo de las bases de datos de la PME, sin necesidad de adquirir un software comercial para la lectura de los microdatos. Esta función sólo tiene dos argumentos: `Format`, que corresponde al archivo `INPUT.txt`, descargado desde la página web dela encuesta, y `File`, que corresponde a los microdatos de la PME, descargados tambiém dela página web de la PME. Todos los archivos de microdatos de la encuesta tienen una identificación común a `PMEnova.mes.año.txt`. El resultado de la función es una estructura de datos (data.frame) en R con los microdatos de la PME.

`Match`. Con esta función se realiza el emparejamiento en el panel con los criterios dados en la sección 5.1. Tiene tres argumentos: B, correspondiente a una única base de datos[3] con la información de los ocho meses de medición del panel de interés; `Panel` corresponde a la letra que identifica el panel de interés; y `Group` identifica el número del grupo rotacional. El resultado de la función es una lista de bases de datos. `Match (B, Panel, Group)$Data` contiene[4] el conjunto de datos originales para todas las mediciones. `Match(B, Panel, Group)$Persons` corresponde a las personas identificadas en el panel reconstruido. La última columna indica la medición donde se identificó por primera vez a

[3] Se sugiere reducir el número de columnas en la base de datos B que sirve como argumento a la función Match. Lo anterior, puesto que el proceso se puede tornar demorado. Las columnas que debería conservar B son las variables de identificación para el emparejamiento y las otras que se considere pertinente estudiar.

[4]4 El número de filas de Match(B, Panel, Group)$Data debe estar alrededor de las cien mil personas, puesto que se realizan ocho mediciones en meses distintos y en cada mes el número de personas encuestadas en un grupo rotacional es de más de doce mil personas.

cada persona. `Match(B, Panel, Group)$Matches` contiene todos los empates encontrados en el proceso, mostrando los números de los registros que resultaron empatados, el mes en que se detectó el empate y el criterio con el que empataron. `Match(B, Panel, Group)$Loose` identifica el conjunto de personas cuyos registros no pudieron ser empatados y aparecen una sola vez en el estudio del panel.

`TypeNR`. Con esta función se clasifican las personas del panel dependiendo de su presencia en los periodos de referencia i, j. Como argumentos, requiere de un objeto `M` de tipo `Match` y de los periodos[5] i y j. El resultado de la función es una lista de bases de datos. `TypeNR(M, i, j)$Shared` contiene la información de las variables de interés de los respondientes denominados como Tipo 0. `TypeNR(M, i, j)$Only_i` muestra la información de los respondientes denominados como Tipo 1. `TypeNR(M, i, j)$Only_j` contiene la información de los respondientes en el segundo periodo identificados como Tipo 2. Finalmente, `TypeNR(M, i, j)$Neither` muestra la información de los no respondientes, denominados como Tipo 3. Si se han considerado los factores de expansión (dados por V211 en el archivo de microdatos), entonces es posible estimar el tamaño del panel, de acuerdo a las expresiones de la sección 4.1.

7. Ejemplo real: el panel P6

En esta sección se ilustra el uso de las funciones de R y se muestran los resultados obtenidos. Para ello, se utiliza el panel P6 cuyo seguimiento tuvo lugar desde Noviembre de 2010 hasta Febrero de 2011 y luego, desde Noviembre de 2011 hasta Febrero de 2012.

[5]5 Nótese que los valores de i y j no deben ser necesariamente consecutivos, pero sí diferentes. Además, deben ser enteros mayores o iguales a uno y menores o iguales a ocho.

7.1. Lectura de los microdatos

De la página web de la PME se descargan los archivos correspondientes a los ocho meses de seguimiento del panel P6, escogido para el ejemplo. De la misma página, se descarga el archivo `INPUT.txt` para la lectura de las bases de datos con SAS. En una misma carpeta se guardan los microdatos de las ocho mediciones y el archivo `INPUT.txt`. Para la sesión en R se cargan las funciones con la instrucción:

```
source("http://www.gutierrezandres.com/wpcontent/uploads/2012/08/PME_M
atching_Nonresponse.txt")
library(sqldf)
library(car)
library(stringr)
```

Si se quiere, se descarga el archivo `PME_Matching_Nonresponse.txt` en una carpeta, que puede ser la misma de los datos, y se ejecuta:

```
setwd("C:/Folder")
source("PME_Matching_Nonresponse.txt")
```

En donde `Folder` es la ubicación de los archivos en el sistema. Con la función `ReadPME`, se cargan los datos en la sesión de R. El proceso de la lectura puede ser demorado por el formato fijo en que vienen los microdatos.

```
> D1 = ReadPME(Format = "INPUT.txt", File = "PMEnova.112010.txt")
> D2 = ReadPME(Format = "INPUT.txt", File = "PMEnova.122010.txt")
> D3 = ReadPME(Format = "INPUT.txt", File = "PMEnova.012011.txt")
> D4 = ReadPME(Format = "INPUT.txt", File = "PMEnova.022011.txt")
> D5 = ReadPME(Format = "INPUT.txt", File = "PMEnova.112011.txt")
> D6 = ReadPME(Format = "INPUT.txt", File = "PMEnova.122011.txt")
> D7 = ReadPME(Format = "INPUT.txt", File = "PMEnova.012012.txt")
> D8 = ReadPME(Format = "INPUT.txt", File = "PMEnova.022012.txt")
```

Para mejorar la eficiencia computacional de los subsecuentes procesos, es conveniente conservar solamente las variables de interés en los microdatos. Para nuestro ejemplo, conservaremos las variables demográficas y de domicilio que nos permitirán realizar el proceso de emparejamiento, el nivel de ocupación, dado por VD1 y también el factor de expansión que contiene los pesos de muestreo, dado por V211.

```
> D1s = data.frame(V035 = D1$V035, V040 = D1$V040, V050=D1$V050, V060
= D1$V060,
V063 = D1$V063, V070 = D1$V070, V075 = D1$V075, V201 = D1$V201, V203 =
D1$V203,
V204 = D1$V204, V214 = D1$V214, V224 = D1$V224, V072 = D1$V072, V234 =
D1$V234,
V205 = D1$V205, V307 = D1$V307, V211 = D1$V211, VD1 = D1$VD1, VDAE1 =
D1$VDAE1)
...
> D8s = data.frame(V035 = D8$V035, V040 = D8$V040, V050 = D8$V050,
V060 = D8$V060,
V063 = D8$V063, V070 = D8$V070, V075 = D8$V075, V201 = D8$V201, V203 =
D8$V203,
V204 = D8$V204, V214 = D8$V214, V224 = D8$V224, V072 = D8$V072, V234 =
D8$V234,
V205 = D8$V205, V307 = D8$V307, V211 = D8$V211, VD8 = D8$VD8, VDAE1 =
D8$VDAE1)
```

7.2. Reconstrucción del panel

Una vez realizada la lectura de los microdatos, con la función `Match` se reconstruye el panel. Para esto, se integran las ocho bases de datos, correspondentes a los ocho meses de seguimento del panel, en una única denotada como B. Esta integración se realiza con la función `rbind` de R. Se define el panel y el grupo rotacional de interés (`Panel="P", Group=6`):

```
> B = rbind(D1s, D2s, D3s, D4s, D6s, D5s, D7s, D8s)
> M = Match(B, Panel="P", Group=6)
```

Después de ejecutar la función, se pueden guardar sus resultados:

```
> Data = M$Data
> Persons = M$Persons
> Matches = M$Matches
> Loose = M$Loose
```

En Data, se ha almacenado la información de las ocho mediciones del panel P6, en Persons, se reporta el panel reconstruido, en Matches, se muestran los criterios con los que se consiguieron los empates y la medición correspondiente, y en Loose, se encuentrn personas que nunca empataron.

Se puede verificar que el número de registro en el panel P6 durante las ocho mediciones es de 101736. El panel contiene 21374 personas distintas de las cuales, 2199 aparecen en un solo registro. Se encuentran 80362 empates. Nótese que el número de personas únicas en el panel más el número de empates coincide con el número de registros en los ocho meses de medición.

```
> nrow(Data)
[1] 101736
> nrow(Persons)
[1] 21374
> nrow(Matches)
[1] 80362
> nrow(Loose)
[1] 2199
> nrow(Persons)+nrow(Matches)==nrow(Data)
[1] TRUE
```

A continuación se muestra un breve encabezado del resultado de Data, en el cual aparecen todas las variables de interés, para los 101736 registros de la base de datos, junto con la llave de domicilio, llamado VDom, y un identificador, llamado VPer, creado con el primer criterio. Además aparece un contador del número de personas, denotado por Nro.

```
> Data

V035   V040 ...           VDom                      VPer    Nro
26 26000682 ... 26260006821166  2626000682116611919721      1
26 26000682 ... 26260006821166  2626000682116621919772      2
...
43 43603025 ... 43436030251166 4343603025116611681965 1 101736
```

Luego, en Persons se muestra la información del panel reconstruido conformado por 21374 personas diferentes. En esta base de datos aparece la información de las variables de interés además de VDom, VPer, Nro y la primera medición en donde se identificó a la persona, llamada Med1.

```
> Persons

V035    V040 ...           VDom                        VPer       Nro Med1
26  26000682 ... 26260006821166  2626000682116611919721         1    1
26  26000682 ... 26260006821166  2626000682116621919772         2    1
...
43  43600271 ... 43436002711166  43436002711166220619832  101706    8
```

La tabla 3 presenta la contribución mes a mes de las personas en la reconstrucción del panel. Llama la atención que en la quinta medición, que vuelve a realizarse ocho meses después de la medición en el cuarto mes, la contribución de personas en el panel es del 20.2%.

Tabla 3: Total y porcentaje de personas en cada medición que se identificaron por primera vez en la reconstrucción del panel.

Medición	1	2	3	4	5	6	7	8	Total
Total	12653	1030	862	777	4314	718	622	398	21374
Porcentaje	59,2 %	4,8 %	4,0 %	3,6 %	20,2 %	3,4 %	2,9 %	1,9 %	100,0 %

El siguiente resultado de la función es Matches, en el cual se presentan todos los empates encontrados. Además de VPer, esta función incluye NroA y NroB que indican el empate entre las personas; Time, que indica la medición en donde se encontró el empate y Criterium, que representa el criterio con el cual se identificó el empate.

```
> Matches
                        VPer     NroA    NroB  Time  Criterium
   2626000682116611919721         1   12654     2          1
   2626000682116621919772         2   12655     2          1
...
29290006733016619999999991    78052   90714     8          7
```

La tabla 4 presenta el número de personas empatas según los criterios de reconstrucción del panel. Nótese que el criterio más básico encuentra un 83.7% de empates y el proceso termina por encontrar los restantes 13095 empates que representan el 16.3%.

Además, es posible conocer el mes en que una persona en particular fue encontrada junto con su criterio. Por ejemplo, la persona número cinco fue empatada en siete ocasiones, desde el segundo mes de medición. Para los meses dos, tres y cuatro, el empate se identificó utilizando el criterio básico; para los meses cinco al ocho, el empate se identificó con el criterio tres.

Tabel 4 : Total y porcentaje de personas empatas por cada uno de los criterios del processo

Criterio	1	2	3	4	5	6	7	Total
Total	67267	4319	4022	4152	202	59	341	80362
Porcentaje	83,7%	5,4%	5,0%	5,2%	0,3%	0,1%	0,4%	100,0%

```
> Matches[Matches$NroA==5,]
                          VPer  NroA   NroB   Time Criterium
    5 26260006823166199991943l     5  12658      2         1
11867 26260006823166199991943l     5  25550      3         1
23962 26260006823166199991943l     5  38511      4         1
42158 26260006823166199991943l     5  51485      5         3
53960 26260006823166199991943l     5  63734      6         3
65986 26260006823166199991943l     5  76292      7         3
78300 26260006823166199991943l     5  88991      8         3
```

Nótese que es posible encontrar la información de las personas empatadas. Por ejemplo, para encontrar la información de la persona número cinco, basta con buscar los registros en la base de datos única que contiene todos los registros de las ocho mediciones. Por supuesto, esta persona aparece por primera en la primera medición. Nótese en las cuatro primeras mediciones VPer coincide plenamente. Sin embargo, desde la quinta medición la edad presumida cambió de 1943 a 1944, por lo tanto los empates se identificaron con otro criterio distinto al básico.

```
> id=c(5, Matches[Matches$NroA==5,]$NroB)
> Data[id,]
V035 ...           VDom                      Vper   Nro
  26 ... 26260006823166 26260006823166199991943l     5
  26 ... 26260006823166 26260006823166199991943l 12658
  26 ... 26260006823166 26260006823166199991943l 25550
  26 ... 26260006823166 26260006823166199991943l 38511
  26 ... 26260006823166 26260006823166199991944l 51485
  26 ... 26260006823166 26260006823166199991944l 63734
  26 ... 26260006823166 26260006823166199991944l 76292
  26 ... 26260006823166 26260006823166199991944l 88991
```

Por último, Loose presenta toda la información de las personas que nunca encontraron un empate.

7.3. Clasificación de la respuesta

Después de haber reconstruido el panel con la función Match es posible realizar comparaciones de la clasificación de la respuesta en los ocho meses de seguimiento al panel. Lo anterior se realiza definiendo los periodos de interés en la función TypeNR. Por ejemplo, si se quisieran realizar comparaciones entre el tercer y el cuarto periodo de medición, dado por los meses Enero de 2011 y Febrero de 2011, entonces la clasificación se ejecuta mediante el siguiente código:

```
> M = Match(B, Panel="P", Group=6)

> A = TypeNR(M, i=3, j=4)

> Tipo0 = A$Shared
> Tipo1 = A$Only_i
> Tipo2 = A$Only_j
> Tipo3 = A$Neither

> nrow(Tipo0)+nrow(Tipo1)+nrow(Tipo2)+nrow(Tipo3) == nrow(Persons)
[1] TRUE
```

Nótese que el número de personas en el panel reconstruido es equivalente al número de personas en los cuatro tipos de clasificación de la respuesta. Esta función reproduce todas las variables de interés de los individuos clasificados en estos dos periodos de tiempo y con esto es posible realizar análisis comparativos de las variables de investigación. Para las personas clasificadas como respondientes en ambos periodos, la función devuelve el valor de las características de interés tanto en el primero como en el segundo mes de referencia, tal como se muestra a continuación para la variable VD1 correspondiente al nivel de ocupación:

```
> Tipo0

   VDAE1i VD1j
1       2    1
2       2    3
...
11988   5    1
```

Los otros objetos resultantes de la función solo muestran la información del mes de referencia. En la tabla 5 se muestra la clasificación completa de la respuesta en las ocho mediciones del panel P6. Los valores por encima de la diagonal representan el número de personas respondientes en los dos periodos de tiempo (Tipo 0) y el número de personas que no respondieron en esos dos periodos (Tipo 3). Los valores por debajo de la diagonal representan los respondientes en el primer período de tiempo (Tipo 1) y los respondientes en el segundo periodo de tiempo (Tipo 2). Nótese que existe un decremento significativo de los respondientes Tipo 0 después del cuarto mes de medición. Lo anterior se debe a que han pasado ocho meses desde la cuarta hasta la quinta medición.

7.4. Estimación del nivel de ocupación

Una vez que hemos clasificado las respuestas, podemos establecer las debidas comparaciones en los dos periodos de referencia. Siguiendo un breve código (ver apéndice) computacional y con la ayuda de la librería `TeachingSampling` (Gutierrez 2009), calculamos la clasificación[6] en el panel P6 dada por la tabla 6.

Sin embargo, dado que el panel P6 de la PME corresponde a una muestra aleatoria de las áreas metropolitanas de Brasil, cada individuo en el panel se representa a sí mismo y a muchas más personas en la población. Por lo tanto, recurriendo al procedimiento de estimación reportado en la sección 5.1, y utilizando el factor de expansión de la encuesta dado por la variable V211, notamos que la expansión poblacional[7] del panel P6, en términos del nivel de ocupación, está dada en la tabla 7.

Tabla 5: Clasificación de la respuesta para cada par de meses en el seguimiento del panel P6. Los valores por encima de la diagonal representan el número de personas respondientes Tipo 0 y Tipo 3. Los valores por debajo de la diagonal representan los respondientes Tipo 1 y Tipo 2.

	1	2	3	4	5	6	7	8
1		11862 7691	11243 7007	10724 6467	6940 3415	6990 3152	6926 2948	6923 2892
2	791 1030		11944 7469	11373 6877	7131 3367	7209 3132	7132 2915	7150 2880
3	1410 1714	948 1013		11988 7427	7307 3478	7382 3240	7335 3053	7327 2992
4	1929 2254	1519 1605	969 990		7458 3608	7554 3391	7504 3201	7448 3092
5	5713 5306	5761 5115	5650 4939	5520 4788		11280 7849	10723 7152	10450 6826
6	5663 5569	5683 5350	5575 5177	5424 5005	966 1279		11621 7737	11237 7300
7	5727 5773	5760 5567	5622 5364	5474 5195	1523 1976	938 1978		11897 7820
8	5730 5829	5742 5602	5630 5425	5530 5304	1796 2302	1322 1515	802 855	

[6] Nótese que la suma de todas las entradas de la tabla de clasificación da como resultado el número de personas en el panel reconstruido, es decir 21374.

[7] Es importante recalcar que la suma de todas las entradas de la tabla de clasificación expandida da como resultado el número de personas para las cuales el panel P6 es representativo. Este tamaño se estima en 8465560.

Tabla 6: Cambio bruto poblacional observado en la muestra en dos mediciones para el nivel de ocupación en la PME.

Tiempo $t-1$	Tiempo t				
	Ocupado	Desocupado	Inactivo	No perteneciente	Complemento fila
Ocupado	5205	63	270	0	435
Desocupado	69	143	118	0	29
Inactivo	242	131	4297	0	364
No perteneciente	4	0	19	1427	141
Complemento columna	472	30	330	158	7427

Nótese que un paso esperado, después de haber obtenido esta tabla, consiste en la estimación de una matriz de clasificación de los estados de ocupación, mediante la incorporación de la información en los complementos fila y columna. Un enfoque similar fue seguido por Stasny (1987) sin considerar el efecto del diseño de muestreo complejo. El desarrollo de la metodología que incluye las ponderaciones del diseño de muestreo complejo en la estimación de los cambios brutos es objeto de estudio de la tesis de doctorado del primer autor.

Tabla 7: Cambio bruto poblacional estimado en dos mediciones para el nivel de ocupación en la PME.

Tiempo $t-1$	Tiempo t				
	Ocupado	Desocupado	Inactivo	No perteneciente	Complemento fila
Ocupado	2135864	26600	87748	0	147300
Desocupado	21245	55798	38679	0	10837
Inactivo	85001	54542	1717855	0	123088
No perteneciente	1295	0	5834	561208	51264
Complemento columna	169233	12360	114725	58533	2986542

8. Discusión

En este artículo se proponen varias funciones que pueden ser útiles para los investigadores: La función `ReadPME` que toma el código SAS de lectura de un archivo en formato fijo.`txt` y se ocupa de la lectura del archivo de datos en R. Así, lo dispendioso de la generación manual de las instrucciones se resuelve de manera automática y segura. La única condición del archivo de instrucciones de lectura es que cada línea contenga la información de una sola variable, así como `INPUT.txt`. La función `Match`, que aplica de manera consecutiva los criterios de empate de los registros de la base de datos utilizando esencialmente instrucciones SQL, más estándar y más fácilmente comprensibles que las de cualquier otro código secuencial generado en un lenguaje computacional de otro software. Además, la modificación de los criterios, la eliminación de algunos de ellos o la inclusión de otros nuevos es una tarea fácil de implementar, mediante la manipulación apropiada de la función. Esta función puede servir de guía para resolver problemas similares en otros estudios tipo panel. La función `TypeNR` clasifica los registros según las condiciones que determinan los tipos de no respuesta que se presentan en la PME. Igualmente, resulta sencilla su adaptación a situaciones un tanto diferentes.

Las funciones anteriores ponen al alcance no solo de quienes realizan sus análisis utilizando software libre como R, sino también de quienes trabajan con programas comerciales pues, en la actualidad la mayoría de ellos incorporan la posibilidad de utilizar el código de R.

Por otro lado, este acercamiento intuitivo genera los insumos adecuados para realizar análisis más complejos como la estimación apropiada de los cambios brutos, que debería tener en cuenta el diseño de muestreo complejo. De esta manera, fue posible estimar el alcance del panel P6, que representa a 8.465.560 brasileros en las áreas metropolitanas. Atendiendo a los resultados encontrados en el ejemplo, se advierte un comportamiento similar entre los primeros y los últimos cuatro meses de medición, en donde se encuentra un número elevado de respondientes Tipo 0 y un número bajo de respondientes Tipo 1 y Tipo 2, así como un número no despreciables de no respondientes Tipo 3. Sin embargo, cuando se realiza la comparación entre mediciones que han superado los ocho meses de espera, impuestos por el diseño del panel rotativo, la clasificación de la respuesta cambia dramáticamente, y se observa un decremento del número de respondientes Tipo 0 y del número de respondientes Tipo 3, mientras que el número de respondientes Tipo 1 y Tipo 2 sufre un incremento significativo.

Referências bibliográficas

Agresti, A. (2002), Categorical Data Analysis, John Wiley and Sons.

Béland, Y., Dale, V., Dufour, J. & Hamel, M. (2005), The Canadian Community Health Survey: Building on the Success from the Past, in A. S. Association, ed., `Proceedings of the Survey Research Methods Section, Joint Statistical Meetings', pp. 2738 { 2746.

Chambers, R. L. & Skinner, C. J., eds (2003), Analysis of Survey Data, Wiley.

Clark, R. G. & Steel, D. (2007), `Sampling within households in household surveys', Journal of the Royal Statistical Society. Series A 170, 63 { 82.

Cochran, W. (1977), Sampling Techniques, 3 edn, Wiley.

Deville, J. & S arndal, C. (1992), `Calibration estimators in survey sampling', Journal of the American Statistical Association 87, 376{382.

Fienberg, S. E. & Stasny, E. A. (1983), "Estimating monthly gross ows in labour force participation", Survey Methodology 9, 77 { 102.

Figueredo, J. S. (2003), Avalação de Desgaste de Panéis em Estudos Longitudinais: Uma Aplicação na Pesquisa Mensal de Emprego do IBGE, Dissertação de mestrado em Estudos Populacionais e Pesquisas Socias. Escola Nacional de Ciências Estatísticas.

Fox, J. &Weisberg, S. (2011), An R Companion to Applied Regression, second edn, Sage, Thousand Oaks CA.

Fuller, W. A. (2009), Sampling Statistics, Wiley.

Gambino, J. G. & Silva, P. L. (2009), Handbook of Statistics, Vol. 29A, Elsevier B.V., chapter 16: Sampling and Estimation in Household Surveys, pp. 407 { 439.

Grothendieck, G. (2012), sqldf: Perform SQL Selects on R Data Frames. R package version 0.4-6.4.

Gutierrez, H. A. (2009), TeachingSampling: Sampling designs and parameter estimation in finite population. R package version 2.0.1.

Heeringa, S. G., West, B. T. & Berglund, P. A. (2010), Applied Survey Data Analysis, CRC Press.

IBGE (2007), Pesquisa Mensal de Emprego, Vol. 23 of Série Relatórios Metodológicos, 2 edn, Instituto Brasileiro de Geografía e Estatística.

Kalton, G. (2009), Handbook of Statistics, Vol. 29A, Elsevier, chapter Designs for Surveys over Time, pp. 89 { 108.

Lohr, S. (2000), Sampling: Design and Analysis, Thompson.

Lopes, M. D. (2003), Não-resposta diferencial e tendenciosidade de grupos de rotaçãon na Pesquisa Mensal de Emprego do IBGE, Dissertação de mestrado em Estudos Populacionais e Pesquisas Socias. Escola Nacional de Ciências Estatísticas.

Lumley, T. (2010), Complex Surveys: a Guide to Analysis using R, Wiley.

Perez, R. & Dillon, S. S. (2009), "Sobre o Panel da Pesquisa Mensal de Emprego - PME do IBGE: problemas e soluções para o emparelhamento usando microdados", Revista Brasileira de Estatística 70(233), 75 {108 .

Pessoa, D. G. C. & Silva, P. L. N. (1998), Analise de Dados Amostrais Complexos.

R Development Core Team (2012), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rubin, D. (1976), `Inference and missing data', Biometrika 63, 581 { 592.

Skinner, C. J., Holt, D. & Smith, T. M. F. (1989), analysis of Complex Surveys, Wiley.

Stasny, E. A. (1987), `Some Markov-chain Models for Nonresponse in Estimating Gross Labor Force Flows', Journal of OFIcial Statistics 3, 359 { 373.

Särndal, C. E. (2011), `The 2010 Morris Hansen lecture: Dealing with survey nonresponse in data collection', Journal of Oficial Statistics 27, 1 { 21.

Särndal, C. E. & Lundstr om, S. (2004), Estimation in Surveys with Nonresponse, Wiley.

Särndal, C. E. & Lundstr om, S. (2010), `Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias', Survey Methodology 36, 131 { 144.

Särndal, C., Swensson, B. & Wretman, J. (1992), Model Assisted Survey Sampling, Springer, New York.

Wickham, H. (2011), stringr: Make it easier to work with strings. R package version 0.6.

Agradecimientos

Los autores agradecen la valiosa ayuda y disposición de Pedro Luis do Nascimento Silva y Leonardo Trujillo. También, expresan su gratitud a José Fernando Zea y Ángela Luna por sus oportunos y constructivos comentarios. Finalmente, agradecemos a Sergei Suarez Dillon Soares y Rafael Perez Ribas por compartir los códigos computacionales en STATA de su artículo del año 2009.Este artículo hace parte de la disertación doctoral de Andrés Gutiérrez en el programa de Doctorado en Ciencias - Estadística de la Universidad Nacional de Colombia bajo la dirección de Pedro do Nascimento Silva y Leonardo Trujillo. El primer autor agradece a la Universidad Santo Tomás por financiar esta investigación a través de una comisión de estudios de doctorado.

A Apéndice: código en R para la estimación del tamaño del panel

```
source("http://www.gutierrezandres.com/wp-
content/uploads/2012/08/PME_Matching_Nonresponse.txt")
library(sqldf)
library(car)
library(stringr)

M = Match(B, Panel="P", Group=6)
A = TypeNR(M, i=3, j=4)

Tipo0 = A$Shared
Tipo1 = A$Only_i
Tipo2 = A$Only_j
Tipo3 = A$Neither

z1 = rep(0, times=nrow(Persons))
z1[1:nrow(Tipo0)]=1
z1[(nrow(Tipo0)+1):(nrow(Tipo0)+nrow(Tipo1))]=1

z2 = rep(0, times=nrow(Persons))
z2[1:nrow(Tipo0)]=1
z2[(nrow(Tipo0)+nrow(Tipo1)+1):(nrow(Tipo0)+nrow(Tipo1)+nrow(Tipo2))]=
1

View(cbind(z1,z2))

Tipo0$VD1i[which(as.integer(is.na(Tipo0$VD1i))==1)]<-9
Tipo0$VD1j[which(as.integer(is.na(Tipo0$VD1j))==1)]<-9
Tipo1$VD1[which(as.integer(is.na(Tipo1$VD1))==1)]<-9
Tipo2$VD1[which(as.integer(is.na(Tipo2$VD1))==1)]<-9

y1 = matrix(0, ncol=4, nrow=nrow(Persons))
y1[(1:nrow(Tipo0)),]=Domains(Tipo0$VD1i)
y1[((nrow(Tipo0)+1):(nrow(Tipo0)+nrow(Tipo1))),]=Domains(Tipo1$VD1)

y2 = matrix(0, ncol=4, nrow=nrow(Persons))
y2[(1:nrow(Tipo0)),]=Domains(Tipo0$VD1j)
y2[((nrow(Tipo0)+nrow(Tipo1)+1):(nrow(Tipo0)+nrow(Tipo1)+nrow(Tipo2)))
,]=Domains(Tipo2$VD1)

View(cbind(y1,y2))

w = rep(0, times=nrow(Persons))
w[1:nrow(Tipo0)]=Tipo0$V211i
w[(nrow(Tipo0)+1):(nrow(Tipo0)+nrow(Tipo1))]=Tipo1$V211
w[(nrow(Tipo0)+nrow(Tipo1)+1):(nrow(Tipo0)+nrow(Tipo1)+nrow(Tipo2))]=T
ipo2$V211
w[(nrow(Tipo0)+nrow(Tipo1)+nrow(Tipo2)+1):nrow(Persons)]=Tipo3$V211
View(w)
```

```
Nij=t(y1)%*%y2
Nij
Ri=t(y1)%*%(1-z2)
Ri
Cj=t(y2)%*%(1-z1)
Cj

M=t(1-z1)%*%(1-z2)
M
N <- sum(Nij)+sum(Ri)+sum(Cj)+M
N == nrow(Persons)

Est.Nij=t(w*y1)%*%y2
Est.Nij
Est.Ri=t(w*y1)%*%(1-z2)
Est.Ri
Est.Cj=t(w*y2)%*%(1-z1)
Est.Cj
Est.M=t(w*(1-z1))%*%(1-z2)
Est.M
Est.N <- sum(Est.Nij)+sum(Est.Ri)+sum(Est.Cj)+Est.M
Est.N
```

Análise do risco de mortalidade e de morbidade hospitalar do SUS por doenças respiratórias usando modelo de regressão Poisson com efeitos aleatórios.

Natália Santana Paiva,[1]
Leonardo Soares Bastos,[2]

Resumo.

O escopo do presente trabalho e demonstrar a utilização de alguns modelos de regressão de Poisson com efeitos aleatórios na detecção de padrões de variação do risco de morbidade hospitalar do SUS e de mortalidade para doenças do aparelho respiratório no estado do Rio de Janeiro (RJ) no ano de 2003. A inferência será feita sob a ótica bayesiana usando o método INLA implementado no ambiente *R*. A aplicação aos dados do Rio de Janeiro exemplifica como modelos com efeito aleatório e estrutura espacial podem reduzir a heterogeneidade presente nos dados. No caso de dados de taxa de internação por doenças respiratórias, o modelo de regressão Poisson com efeitos taxa de internação por doenças respiratórias, o modelo de regressão Poisson com efeitos aleatórios com estrutura espacial se mostrou o mais adequado enquanto para o ajuste da taxa de mortalidade por doenças respiratórias o modelo somente com efeitos aleatórios foi o que mais se adequou. Em ambos modelos nenhuma das covariáveis disponíveis mostrou-se estatisticamente significativa.

Palavras-chave: Regressão de Poisson; Modelo com efeitos aleatórios; Doenças respiratórias; INLA; DIC.

[1] Departamento de Estatística - Universidade Federal Fluminense; Departamento de Métodos Estatísticos - Universidade Federal do Rio de Janeiro.
[2] Departamento de Estatística - Universidade Federal Fluminense; Programa de Computação Científica, - Fundação Oswaldo Cruz, E-mail: lsbastos@fiocruz.br.

1. Introdução

Efeitos da poluição do ar sobre a saúde humana têm sido constatados tanto na mortalidade geral e por doenças respiratórias e cardiovasculares como na morbidade incluindo aumento em sintomas respiratórios e diminuição nas funções pulmonares (Castro et al. 2009).

Segundo Gouveia et al. (2003), no Brasil, alguns estudos investigatórios dos efeitos da poluição do ar na saúde encontram associações estatisticamente significantes com mortalidade infantil, mortalidade em idosos, além de hospitalizações de crianças e adultos por causas respiratórias.

Evidências comprovam que fatores meteorológicos, assim com aspectos demográficos, índices de desenvolvimento humano (IDH), urbanização, padrões da industrialização, dentre outros, também afetam a qualidade do ar, com reflexos diretos sobre a saúde humana (Bueno 2008). Devidi as tais evidências, algumas covariáveis sócio-econômicas serão usadas para tentar explicar a relação entre qualidade do ar e mortalidade e/ou as internações hospitalares segundo doenças respiratórias no estado do Rio de Janeiro (RJ) no ano de 2003, tais como, IDH, taxa de urbanização do ano 2000, taxa da frota veicular, densidade demográfica entre outras.

Dados referentes ás internações hospitalares e mortalidade por doenças do aparelho respiratório (capítulo *X* da Classificação Internacional de Doenças em sua décima revisão – CID 10) foram coletados diretamente de bancos de dados informatizados, disponibilizados pelo Ministério da Saúde para os hospitais conveniados ao SUS – DATASUS. Esses bancos contém informações de todas as internações e óbitos no âmbito do SUS por intermédio das Autorizações de Internações Hospitalar (AIH) e Declarações de Óbito (DO), respectivamente .

Até meados de 1980, a poluição atmosférica provinha das emissões industriais. Com o rápido crescimento da frota veicular, verificou-se a enorme contribuição dessa fonte na degradação da qualidade do ar, principalmente nas regiões metropolitanas do país (IBGE 2008).

A estimativa de poluição usada nesse trabalho é o IPPS (abreviação do inglês para Industrial Pollution Projection System) implementado pelo Instituto Brasileiro de Geografia Estatística (IBGE). Sor et al. (2008), em um estudo inicia, não identificaram uma relação entre o IPPS e dados de doenças respiratórias do DATASUS.

Podem existir muitas covariáveis que ajudam a explicar a taxa de morbidade e mortalidade de um município. Com o objetivo de reduzir a heterogeneidade entre municípios o presente trabalho propõe a utilização de modelos de regressão com efeitos aleatórios na detecção de padrões de variação de morbidade hospitalar do SUS e de mortalidade por doenças respiratórias nos municípios do Rio de Janeiro no ano de 2003.

O escopo do presente trabalho é demonstrar a utilização do modelo de regressão com efeitos aleatórios na detecção de padrões de variação do risco de morbidade hospitalar do SUS e de mortalidade para doenças do aparelho respiratório no estado do Rio de Janeiro. Também serão considerados efeitos aleatórios com estrutura de dependência espacial, chamados efeitos espaciais. No processo de interferência, será utilizado o método INLA (abreviação do inglês para Integrated Nested Laplace Approximation), proposto por Rue et al. (2009), usado para fazer inferência bayesiana em modelos com campos aleatórios gaussianos latentes, onde o modelo de regressão com efeitos aleatórios e espaciais são casos particulares.

Este trabalho está organizado da seguinte forma, na seção 2 os dados a serem utilizados serão descritos. Na seção 3, a modelagem estatística será introduzida, onde serão descritos o modelo de regressão Poisson com efeitos aleatórios e o método de inferência. Na seção 4, será estudada a presença de possíveis fatores de risco para a morbidade hospitalar do SUS e mortalidade por doenças respiratórias no estado do Rio de Janeiro. Finalmente, na seção 5, o trabalho é concluído com uma discussão dos resultados obtidos e de possíveis extensões para a modelagem estatística.

2. Descrição dos dados

Os dados utilizados para as análises foram secundários. As variáveis dependentes, número de internações e óbitos por doenças do aparelho respiratório, na população residente do RJ, no ano de 2003, são provenientes dos Sistemas de Informação Hospitalares do SUS (SIH-SUS) e do Sistema de Mortalidade (SIM) disponibilizados no DATASUS.

O número médio de internações, por local de residência, por doenças do aparelho respiratório no estado do RJ no ano de 2003 foi 1.041,25 e o total de notificações de internações neste mesmo ano foi 95.795. O número médio de óbitos pele mesma causa e ano foi 135 e o total de registros de óbito foi 12.421.

O município que obteve o menor índice de morbidade hospitalar do SUS pela causa estuda foi o Carapebus com apenas 12 registros. Os municípios que apresentaram os maiores níveis de morbidade hospitalar do SUS e, também, de mortalidade foram Rio de Janeiro (15.212 e 5.551, respectivamente) e São Gonçalo (12.724 e 726, respectivamente). O menor número registrado de óbitos foi em Laje do Muriaé com apenas 2 casos no ano de 2003 pela causa analisada.

As covariáveis referentes aos aspectos demográficos como taxa de urbanização, dada pela da população residente da área urbana do município e a população residente total do município, e o IDH foram extraídas do censo demográfico de 2000 realizado pelo IBGE. Também do IBGE, foram coletadas informações referentes à área territorial dos municípios do RJ (km²) e à população geral estimada para o Tribunal de Contas da União (do ano de 2003) para o cálculo da densidade demográfica.

A taxa média de urbanização em 2000 foi 0,80. Sumidouro foi classificado como o município de menor taxa de urbanização (0,16) dentre todos os 92 municípios do RJ. O IDH médio do ano de 2000 foi equivalente à 0,76. O IDH mínimo e máximo ficou a cargo dos municípios de Varre-Sai (0,679) e Niterói (0,886), respectivamente. A densidade demográfica média encontrada foi 665,6 habitantes por km². O município com o maior valor registrado foi São João de Meriti com 13.110,0 habitantes por km².

Para representação do nível de poluição de cada um dos 92 municípios do RJ foram utilizadas a frota veicular total, por tipo de veículo do mês de Dezembro de 2003, obtida através do site do Departamento Estadual de Trânsito do Rio de Janeiro (Detran-RJ)[1], e estimativas IPPS a partir do estudo implementado pelo IBGE que disponibiliza a quantidade de emissão potencial IPPS (tonelada/2003) de partículas finos (PM10), partículas inaláveis, de diâmetro inferior a 10 micrómetros e de dióxido de enxofre(SO_2).

Sor et al. (2008) apontam que o estado do RJ apresenta duas áreas críticas em termos de poluição de ar: a região Metropolitana, a qual se encontra a segunda maior concentração de população, de veículos, de indústrias e de fontes emissoras de poluentes do país e a região do Médio Paraíba, cujas principais cidades são Volta Redonda, Barra do Piraí, Italiana e Resende, que é conhecida por sua atividade industrial concentrada no eixo viário que interliga as duas maiores metrópoles brasileiras, Rio de Janeiro e São Paulo.

[1] http://www.detran.rj.gov.br/_estatisticas.veiculos/index.asp, acessado em 15 de Abril de 2011.

A taxa de frota veicular média por tipo de veículo, de mês de Dezembro de 2003 é de 17,68%. O Rio de Janeiro apresenta a maior taxa dentre todos os 92 municípios com 30,83%.

Em 1990 foi criado e sancionado o Projeto de Lei da Emancipação do município de Mesquita. Sendo assim, o mesmo não entrou no censo de 2000, inviabilizando suas informações de população urbana e IDH. Para o cálculo dos dados faltantes foi construída uma média de seus vizinhos semelhantes em área geográfica: Belford Roxo (leste), Nilópolis (sul) e São João de Meriti (sudeste).

A emissão potencial IPPS média de PM10 foi de 208,3 toneladas em 2003. Os municípios que mais emitiram PM10 foram Rio de Janeiro (4.844,0 toneladas) e Volta Redonda (4.030,5 toneladas). A emissão potencial IPPS média de SO_2 encontrada foi 902,6 toneladas emitidas em 2003. Os municípios que mais emitiam SO_2 foram Rio de Janeiro (2.557,0 toneladas) e Duque de Caxias (17.959,7 toneladas).

O potencial de poluição industrial do ar também é concentrado em poucos municípios: Rio de Janeiro, Duque de Caxias e Volta Redonda, na emissão de SO_2, Rio de Janeiro, Volta Redonda e Cantagalo, em relação aos PM10, substâncias que causam danos à saúde respiratória e ao meio ambiente. Isso significa que os potenciais de poluição do ar por indústrias estão na região metropolitana do Rio de Janeiro e no médio Paraíba do Sul.

Variáveis metereológicas e medidas de poluição feitas em estações monitoradas podem explicar melhor a variação nas taxas de morbidade hospitalar do SUS e mortalidade por doenças respiratórias. No entanto, para este trabalho os autores não tiveram acesso a essa informação, e acredita-se que existia uma variabilidade signitiva entre os municípios do estado do Rio de Janeiro. A não inclusão de tais variáveis explicativas, justificam a inclusão de efeitos aleatórios para reduzir uma possível heterogeneidade presente nos dados.

Na próxima seção, modelos de regressão de Poisson com e sem efeitos aleatórios serão usados para explicar a morbidade hospitalar do SUS e mortalidade por doenças respiratórias usando as variáveis descritas nesta seção.

3. Regressão Poisson com efeitos aleatórios

Os modelos de regressão para dados de contagem já estão bem estabelecidos na literatura estatística (McCullagh AND Nelder 1989). A introdução de efeitos aleatórios independentes em modelos de regressão de Poisson foi proposta por Hougaard (1984) com o objetivo de reduzir uma possível heterogeneidade presente nos dados causada, por exemplo, devido à ausência de uma covariável importante. A Inclusão da estrutura espacial aos efeitos aleatórios foi proposta no artigo seminal de Besag, York and Mollie (1991), onde aos efeitos aleatórios foram atribuídos um processo condicional autoregressivo espacial.

Um problema para este tipo de modelagem é o custo computacional. Neste caso a inferência é baseada na distribuição a posteriori dos parâmetros do modelo que é usualmente obtida via métodos computacionalmente intensivos. Na qual os métodos de simulação de Monte Carlo via cadeias de Markov (MCMC) tem um grande destaque (Gamerman and Lopes 2006). No entanto, neste trabalho será utilizado um método altenativo ao MCMC para obtenção das distribuições marginais a posteriori dos parâmetros do modelo proposto, tal método é conhecido como INLA.

Sem perda de generalidade, os modelos descritos nessa seção serão para o número de internações por doenças respiratórias, mas a mesma proposta de modelagem estatística continua válida para o total de óbitos por doenças respiratórias. Na seção 4, os modelos propostos nesta seção serão utilizados tanto para total de internações quanto para o total de óbitos por doenças respiratórias no estado do Rio de Janeiro no de 2003.

3.1. Descrição do modelo estatístico

Seja Y_i o total de internações por doenças respiratórios do município i, $i = 1, 2 \ldots, M$, M denota o total de municípios do estado do Rio de Janeiro no ano de 2003, 92 municípios. Será assumido que o número de internações por doenças respiratórias no RJ em 2003 segue uma distribuição de Poisson com taxa $E_i\lambda_i > 0$, onde E_i é o número esperado de internações no município i. Ou seja,

$$Y_i|\lambda_i \sim Poisson(E_i\lambda_i), i = 1,2...,M \qquad (1)$$

O total esperado de internações no município i é aproximado usando a distribuição de internações do estado do Rio de Janeiro por faixa etária, informações disponíveis pelo DATA-SUS, e a população por faixa etária em cada município, informações disponíveis pelo IBGE. Ou seja,

$$E_i = \sum_k P_{ik}\pi_k^{RJ} \qquad (2)$$

onde k representa uma faixa etária {menor que 1 ano, 1 a 4 anos, 5 a 9 anos, 10 a 14 anos, 15 a 19 anos, 20 a 29 anos, 30 a 39 anos, 40 a 49 anos, 50 a 59 anos, 60 a 69 anos, 70 a 79 anos, e 80 anos ou mais}, P_{ik} é a população do município i na faixa etária k, e π_k^{RJ} é a proporção de pessoas internadas no estado do Rio de Janeiro na faixa etária k, $\sum_k \pi_k^{RJ} = 1$. Note que $E_1,\ldots, E_K$ não refletem variações espaciais da morbidade hospitalar do SUS.

O estimador de máxima verossimilhança de λ é dado por

$$\hat{\lambda}_i = \frac{Y_i}{E_i}, \qquad i = 1,2,...,\ M\ ,$$

conhecido por razão de morbidade padronizada (RMbP) ou razão de mortalidade padronizada (RMtP). A Figura 1 apresenta o mapa das razões padronizadas de morbidade hospitalar do SUS e mortalidade por doenças respiratórias no estado do Rio de Janeiro no ano de 2003. A Figura 1 (a) sugere a existência de um padrão espacial para a morbidade hospitalar do SUS. Vale destacar que os municípios do noroeste fluminense apresentaram uma morbidade hospitalar do SUS maior que a esperada, enquanto municípios da região dos lagos tiveram uma taxa de morbidade hospitalar do SUS menor que a esperada. Foi observado que 23 municípios tiveram internações abaixo do valor esperado. A Figura 1 (b) sugere que a maioria dos municípios tiveram um mortalidade por doenças respiratórias acima do que era esperado, apenas 6 em 92 municípios tiveram o total de óbitos abaixo do esperado.

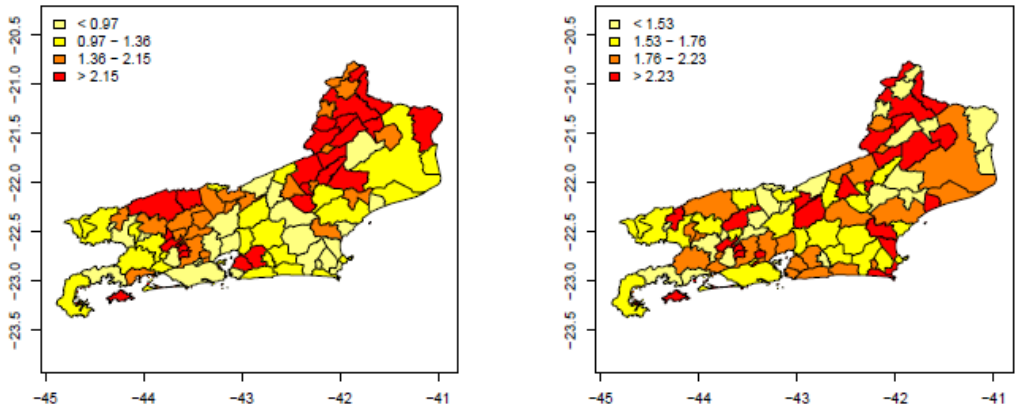


Figura 1: Mapas das razões padronizadas de morbidade hospitalar do SUS (a) e mortalidade (b) por doenças respiratórias no estado do Rio de Janeiro no ano de 2003.

Covariáveis por município, denotadas por x_i , são introduzidas no modelo através de uma função ligação, g(.), usualmente a função logarítmica. Levando ao **modelo de regressão** Poisson (MRP). Ou seja, para i = 1, 2,..., M, a função de ligação é dada por

$$g(\lambda_i) = x_i^t \beta \qquad (3)$$

Onde β é um vetor de coeficientes de regressão.

Como os municípios têm características distintas que podem não ser explicadas pelas covariáveis utilizadas, com o objetivo de reduzir a heterogeneidade dos dados, é razoável incluir ao modelo um efeito aleatório latente para cada município, ei. O **modelo de regressão Poisson com efeitos aleatórios** (MRPea), tem função de ligação para cada município dada por

$$g(\lambda_i) = x_i^t \beta + \varepsilon_i . \qquad (4)$$

onde ε_i segue uma distribuição gaussiana com média zero e variância σ_e^2 , assumindo independência entre os efeitos aleatórios. Ou seja,

$$\varepsilon_i \sim N(0; \sigma_\varepsilon^2) \qquad (5)$$

Visualmente na Figura 1 (a), as taxas de morbidade hospitalar do SUS dos municípios parecem ter uma dependência espacial, ou seja, a morbidade hospitalar do SUS do município i tende a ter um comportamento similar com a morbidade hospitalar do SUS (mortalidade) de seus municípios vizinhos. Portanto, é razoável considerar a inclusão de um efeito aleatório latente com estrutura espacial, levando ao modelo de regressão Poisson espacial (MRPesp). Para i = 1, 2,..., M, a função de ligação é dada por

$$g(\lambda_i) = x_i^t \beta + \phi_i \qquad (6)$$

onde $Ø_i$ é o efeito aleatório espacial do município i. Será assumido que a distribuição de $Ø_i$ também será gaussiana com média dada pela média de seus vizinhos, com variância t_ϕ^2. Ou seja, para cada municÍpio assume-se que

$$\phi_i|\phi_{\partial i} \sim N\left(\frac{1}{n_i}\sum_{j\varepsilon\partial i}\phi_j ; \frac{t_\phi^2}{n_i}\right) \qquad (7)$$

onde ∂i denota o conjunto de índices dos vizinhos do município ι, n_i denota o total de vizinhos do município ι,e t_ϕ^2 é a variância dos efeitos espaciais. Essa estrutura de dependência espacial é um processo auto-regressivo no espaço e é conhecida como CAR intrínseco. Besag (1974) mostra que a distribuição conjunta dos efeitos aleatórios é gaussiana com média zero e uma matriz de variância singular, implicando em uma priori conjunta imprópria para os efeitos aleatórios. O uso de prioris impróprias na inferência bayesiana é bastante estudado na literatura, para detalhes sobre prioris impróprias veja Bernardo and Smith (1994).

E um modelo que tem como caso particular os três modelos anteriores, chamado de **modelo de regressão Poisson completo** (MRPc) que tem como função de ligação

$$g(\lambda_i) = x_i^t\beta + \varepsilon_i + \phi_i \qquad (8)$$

onde a distribuição dos efeitos aleatórios independentes, $\varepsilon=\{\varepsilon_1,...,\varepsilon_M\}$, é dada por (5) e a distribuição dos efeitos aleatórios espaciais, $\phi=\{\phi_1,...,\phi_M$, é dada por (7). Note que os modelos anteriores podem ser obtidos simplesmente fixando uma ou ambas variância dos efeitos aleatórios como zero.

3.2. Inferência

A inferência para os parâmetros dos modelos de regressão apresentados na seção anterior será feita sob o paradigma bayesiano. Seja $\theta=\left(\beta,\varepsilon,\phi,\sigma_{\varepsilon}^{2},t_{\phi}^{2}\right)$ o vetor de parâmetros do modelo completo. Toda inferência é baseada somente na distribuição a posteriori de θ, que é obtida via do teorema de Bayes, ou seja,

$$p\left(\theta|y,X\right)\propto p\left(y|\theta,X\right)p(\theta) \qquad (9)$$

onde $p(\theta)$ é a distribuição a priori dos parâmetros e $p\left(y|\theta,X\right)$ é a função de verossimilhança, que sob independência condicional é dada pelo produto das densidades dadas em (1). Ou seja,

$$p\left(y|\theta,X\right)\propto\prod_{i=1}^{M}\left(E_{i}g^{-1}\left(x_{i}^{T}\beta+\varepsilon_{i}+\phi_{i}\right)\right)^{yi}e^{-Eig^{-1}\left(x_{i}^{T}\beta+\varepsilon_{i}\phi_{i}\right)} \qquad (10)$$

A especificação do modelo se completa após a elicitação da distribuição a priori para θ. A distribuição a priori tem como principal papel a inclusão de qualquer informação subjetiva a respeito dos parâmetros do modelo em forma de probabilidade. A priori para θ é dada por

$$p(\theta)=p(\beta)p\left(\varepsilon\left|\sigma_{\varepsilon}^{2}\right.\right)p\left(\sigma_{\varepsilon}^{2}\right)p\left(\phi\left|\tau_{\phi}^{2}\right.\right)p\left(\tau_{\phi}^{2}\right) \qquad (11)$$

Para os coficientes de regressão serão utilizadas prioris normais independentes com média zero e variância 100^{2}. A distribuição dos efeitos aleatórios independentes é dada em (5), e dos efeitos aleatórios espaciais é dada em (7), e as distribuições a priori para as variâncias σ_{ε}^{2} e τ_{ϕ}^{2} serão Gamas inversas independentes, GI(a; b) com moda b/(a + 1), as seguintes prioris não informativas, com a = 1 e b = 0; 00005.

3.2.1. INLA

As distribuições a posteriori (9), não tem forma analítica tratável para integação. Usualmente aproximações para essa distribuição são obtidas numericamente através de métodos computacionalmente intensivos tais como os métodos via MCMC. Rue et al. (2009) propuseram um método de aproximação para obter as distribuições marginais a posteriori para modelos gaussianos latentes sem recorrer a métodos de Monte Carlo. O método proposto é conhecido por INLA.

O objetivo do INLA é obter aproximações para as seguintes marginais a posteriori

$$p(\theta_k|y,X) = \int p(\theta_k|y,X,\psi)\, p(\psi|y,X)d \quad (12)$$

$$p(\psi_i|y,X) = \int p(\psi_i|y,X)\, d\psi_{-i}. \quad (13)$$

onde ψ corresponde aos hiper-parâmetros do modelo e ψ_{-i} corresponde ao vetor de hiper-parâmetros ψ excluindo a *i*-ésima observação. Tais aproximações são baseadas na combinação eficiente da aproximação de Laplace para as condicionais completas para $p(\psi|y,X)$ e $p(\theta_k|y,X,\psi)$ e de rotinas de integração numérica para os hiper-parâmetros ψ.

Aproximação proposta por Rue et al. (2009) consiste em trâs passos. O primeiro é a obtenção de uma aproximação de Laplace para a densidade $p(\theta|y,X,\psi)$, o segundo passo consiste em aproximar a marginal $p(\theta_k|y,X)$ novamente usando uma aproximação de Laplace, e o terceiro passo consiste em resolver numericamente as integrais (12) e (13). Veja detalhes da aproximação e da integração numérica em Rue et al. (2009) e Schrödle and Held (2010). Note que, os autores recomendam para a integração numérica que a dimensão de ψ seja baixa. Fato que É verificado na modelagem proposta neste artigo, onde a dimensão máxima de ψ é 2, pois o modelo completo tem dois hiper-parâmetros, τ_ϕ^2 e σ_ε^2.

3.2.2. Comparação de modelos

Os modelos ajustados serão comparados usando o DIC (abreviação do inglês para *Deviance Information Criterion*), proposto por Spiegelhalter et al. (2002). Seja a função desvio, ou *deviance*, definida por

$$D(\theta) = -2 \log p(y|\theta) + C$$

onde θ é o vetor de parâmetros de interesse, $p(y|\theta)$ é a função de verossimilhança e C é uma constante que não depende de θ e que se cancela ao compararmos dois modelos.

Seja $\overline{D} = E[D(\theta)|y]$ o valor esperado a posteriori para a função desvio, que é uma medida de ajuste dos dados que quanto menor melhor. Seja $p_D = \overline{D} - D(\overline{\theta})$ o número efetivo de parâmetros do modelo, onde $\overline{\theta}$ é a média a posteriori de θ. O *DIC* é definido como o valor esperado da função desvio penalizado pelo número efetivo de parâmetros, ou seja

$$DIC = p_D + \overline{D}.$$

De uma forma geral, quanto menor o valor do DIC melhor o modelo. Para mais detalhes a respeito do DIC veja em Spiegelhalter et al. (2002).

4. Análise dos dados de doenças respiratórias no estado do Rio de Janeiro

4.1. Seleção dos modelos finais

Para cada um dos modelos propostos na seção 3, modelo de regressão de Poisson (MRP), com efeitos aleatórios (MRPea), com efeitos espaciais (MRPesp) e completo (MRPc), tanto para o risco de morbidade hospitalar do SUS quanto para mortalidade, foi realizada uma seleção não automática de covariáveis baseada no método *stepwise*

(Efroymson 1960). O modelo final para cada um dos quatro casos foi escolhido segundo o critério de comparação dos DICs dentre os possíveis modelos e estes podem ser observados na Tabela 1.

Tabela 1: Modelos finais escolhidos para morbidade hospitalar do SUS (e mortalidade) segundo um método de seleção de covariáveis não automático. Os sinais + ou - indicam o sinal do efeito e NS indica variável não significativa.

Variáveis	MRP	MRPea	MRPesp	MRPc
Densidade	NS (NS)	NS (NS)	NS (NS)	NS (NS)
SO_2	- (NS)	NS (NS)	NS (NS)	NS (NS)
PM10	+ (NS)	NS (NS)	NS (NS)	NS (NS)
Frota veicular	- (NS)	NS (NS)	NS (NS)	NS (NS)
Taxa de urbanização	+ (+)	NS (NS)	NS (NS)	NS (NS)
IDH	- (-)	- (NS)	NS (NS)	NS (NS)

Vale ressaltar que os modelos finais tanto para morbidade hospitalar do SUS quanto para mortalidade as covariáveis não foram estatisticamente significativas. No entanto, a inclusão de efeitos aleatórios, espaciais ou não, capturou a heterogeneidade presente nos dados.

As análises desta seção serão realizadas a partir dos modelos finais, segundo critério de DIC, para morbidade hospitalar do SUS e mortalidade que apresentaram o menor valor do DIC (Tabela 2).

Tabela 2: Comparação dos modelos finais através do critério DIC.

	MRP	MRPea	MRPesp	MRPc
Morbidade	13.579,96	898,36	**896,99**	897,41
Mortalidade	702,55	**640,22**	661,16	640,90

4.2. Análise

O melhor modelo que ilustra a variação do risco de morbidade hospitalar do SUS por doenças do aparelho respiratório no ano de 2003 possui uma estrutura espacial no risco de morbidade hospitalar do SUS, como apontado anteriormente na Figura 1 (a). Para a variação do risco de mortalidade por doenças do aparelho respiratório no ano de 2003 tem-se que o melhor modelo possui efeitos aleatórios no risco de mortalidade. Para ambos os desfechos analisados as covariáveis disponíveis no presente estudo mostraram-se não significativas.

A questão de nenhuma covariável ter sido significativa para os desfechos abordados no presente trabalho pode ser explicado entre outros fatos pela ausência de variáveis meteorológicas. Não se pode descartar a possibilidade de que alterações climáticas sejam as responsáveis pelo agravamento dos sintomas respiratórios. Segundo Arbex et al. (2004) fatores meteorológicos como temperatura, umidade relativa do ar e precipitação, assim como os aspectos demográficos, índices de desenvolvimento humano, urbanização, padrões de industrialização, dentre outros, afetam a qualidade do ar, com reflexos diretos sobre a saúde humana.

Embora se tenha evidências constatadas em Gouveia et al. (2003) de associações entre as internações de crianças (menor de cinco anos de idade) e idosos (maior ou igual a 65 anos) devido ás doenças respiratórias e do aparelho circulatório com o PM_{10}, CO e SO_2 nas cidades de São Paulo e do Rio de Janeiro no presente trabalho as variáveis utilizadas para representar a poluição atmosférica não foram significativas. Tal resultado pode ser explicado pelo fato das variáveis utilizadas serem estimativas das quantidades de emissão potencial poluidor industrial nos municípios. No primeiro estudo que utiliza IPPS como variável *proxy* para poluição, realizado por Sor et al. (2008), também não foi identificado uma relação entre o IPPS e dados de doenças respiratórias do DATASUS.

A Figura 2 (a) apresenta o mapa da mediana a posteriori do risco de morbidade hospitalar do SUS por doenças do aparelho respiratório no ano de 2003 nos municípios do RJ e seu padrão espacial é nítido. Percebe-se que a Região Noroeste Fluminense, assim como a Microrregião geográfica da Barra do Piraí, apresenta um alto risco de morbidade hospitalar do SUS por doenças do aparelho respiratório no ano de 2003. Já as Regiões das Baixadas. Litorâneas como a da Costa Verde apresentam riscos mais baixos. As Figuras 2 (b) e (c) apresentam os mapas dos quantis 2,5% e 97,5% da distribuição a posteriori do risco de morbidade hospitalar do SUS. Esses mapas representam um intervalo de credibilidade espacial para o risco de morbidade hospitalar do SUS. Já a Figura 3 apresenta os mapas da mediana a posteriori do risco de mortalidade por doenças do aparelho respiratório no ano de 2003 nos municípios do RJ e do intervalo de credibilidade espacial de 95%.

Temos que o município do Rio de Janeiro apresenta o risco mediano a posteriori de morbidade hospitalar do SUS igual 0,39 que pertence ao intervalo o qual os riscos são menores que 0,99. Como seu intervalo de credibilidade de 95% (IC_{95}) a posteriori (Figura 2 (b),(c)) encontra-se na mesmo intervalo do risco mediano a posteriori, isto é, IC_{95}(0; 38; 0; 41) pode-se afirmar que com 95% de credibilidade o risco de internações por doenças do aparelho respiratório foi menor que o esperado nessa região no ano de 2003. A Figura 3 mostra que o risco mediano a posteriori de mortalidade no município do Rio de Janeiro está entre 1,5 e 2 com 95% de credibilidade. Para os demais municípios esses mapas não são muito informativos, porém percebe-se que o risco mediano a posteriori de mortalidade para o estado do RJ no período de 2003 foi maior que o esperado.

Alguns destes resultados como a RMbP, RMtP, risco mediano a posteriori (λ) e seu respectivo IC_{95} podem ser vistos na Tabela 3 para os 6 municÍpios do RJ considerados no estudo desenvolvido pelo IBGE (Sor et al. 2008) como os mais poluentes em relação aos particulados finos e dióxido de enxofre e Niterói por possuir o IDH mais elevado e ter uma grande representatividade no estado do RJ.

Pode-se observar que os riscos de morbidade hospitalar do SUS a posteriori e a RMbP são homogêneos para os municípios apresentados na Tabela 3. Em geral, embora se tenha uma estrutura espacial associada ao risco de morbidade hospitalar do SUS a RMbP pode representar bem esse risco. Isso pode ser atribuído á ausência de covariáveis que ajudem a explicar tal desfecho. Já para o risco de mortalidade por doenças do aparelho respiratório no ano de 2003 tem-se que os efeitos aleatórios conseguem suavizar os valores encontrados pela RMtP. Embora esta diferença causada pela suavização seja significativa ela ainda é relativamente pequena.

Tabela 3: Razão de morbidade hospitalar do SUS e mortalidade padronizadas, risco mediano a posteriori e IC95 para alguns municÍpios do RJ.

Municípios	RMbP	$\lambda_{Morb}(IC_{95})$	RMtP	$\lambda_{Mort}(IC_{95})$
Campos dos Goytacazes	1,17	1,18 (1,12;1,24)	1,81	1,82 (1,55;2,12)
Cantagalo	2,46	2,46 (2,17;2,77)	1,99	1,89 (1,41;2,52)
Duque de Caxias	0,89	0,89 (0,85;0,93)	1,84	1,84 (1,60;2,11)
Nilópolis	0,64	0,64 (0,58;0,70)	2,36	2,22 (1,83;2,70)
Niterói	0,70	0,70 (0,66;0,75)	2,70	1,81 (1,57;2,09)
Rio de Janeiro	0,39	0,39 (0,38;0,41)	1,72	1,72 (1,56;1,91)
Volta Redonda	1,01	1,01 (0,94;1,08)	2,14	2,08 (1,74;2,48)

As análises desta seção foram implementadas no ambiente livre R (R Development Core Team 2011). Os mapas foram feitos usando o pacote "spdep" (Bivand et al. 2011) com a malha obtida através do site do Instituto Brasileiro de Geografia e Estatística, a inferência feita com o pacote "INLA" (Rue and Martino 2009) e alguns códigos de autoria própria.

5. Considerações Finais.

O objetivo desse trabalho foi de ilustrar a utilização de modelos de regressão de Poisson com efeitos aleatórios espaciais ou não usando o método INLA. A metodologia foi usada com o objetivo de detectar de padrões na variação do risco de morbidade hospitalar do SUS e de mortalidade para doenças do aparelho respiratório no estado do Rio de Janeiro no ano de 2003.

As variáveis de poluição utilizadas nesse trabalho não apresentaram relação significativa com a morbidade hospitalar do SUS e mortalidade por doenças respiratórias no estado do Rio de Janeiro. Variáveis metereológicas e medidas de poluição feitas em estações monitoradoras podem explicar melhor a variação nas taxas de morbidade hospitalar do SUS e mortalidade por doenças respiratórias.

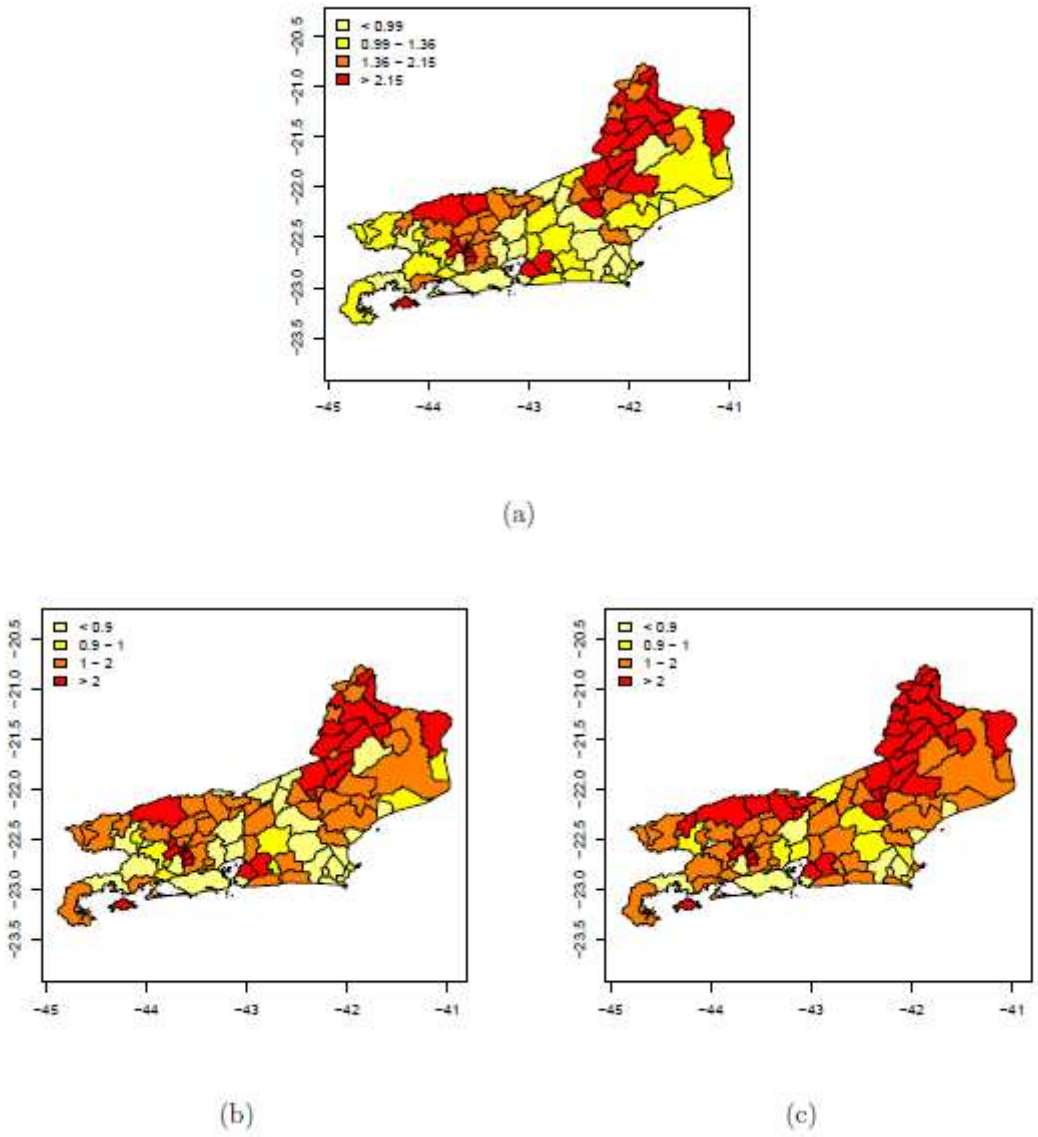


Figura 2: Mapa da mediana a posteriori dos riscos de morbidade hospitalar do SUS por doenças respiratórias no estado do Rio de Janeiro no ano de 2003 (a) e mapas de credibilidade de 95% (b) e (c).

Duas possíveis extensões para a metodologia proposta são consideradas. A primeira seria tratar conjuntamente a morbidade hospitalar do SUS e mortalidade, uma vez que é natural assumir a existência de uma correlação positiva entre tais medidas. A segunda extensão seria a modelagem espaço-temporal onde seria estudada a evolução das taxas de morbidade hospitalar do SUS e mortalidade ao longo do tempo por município.

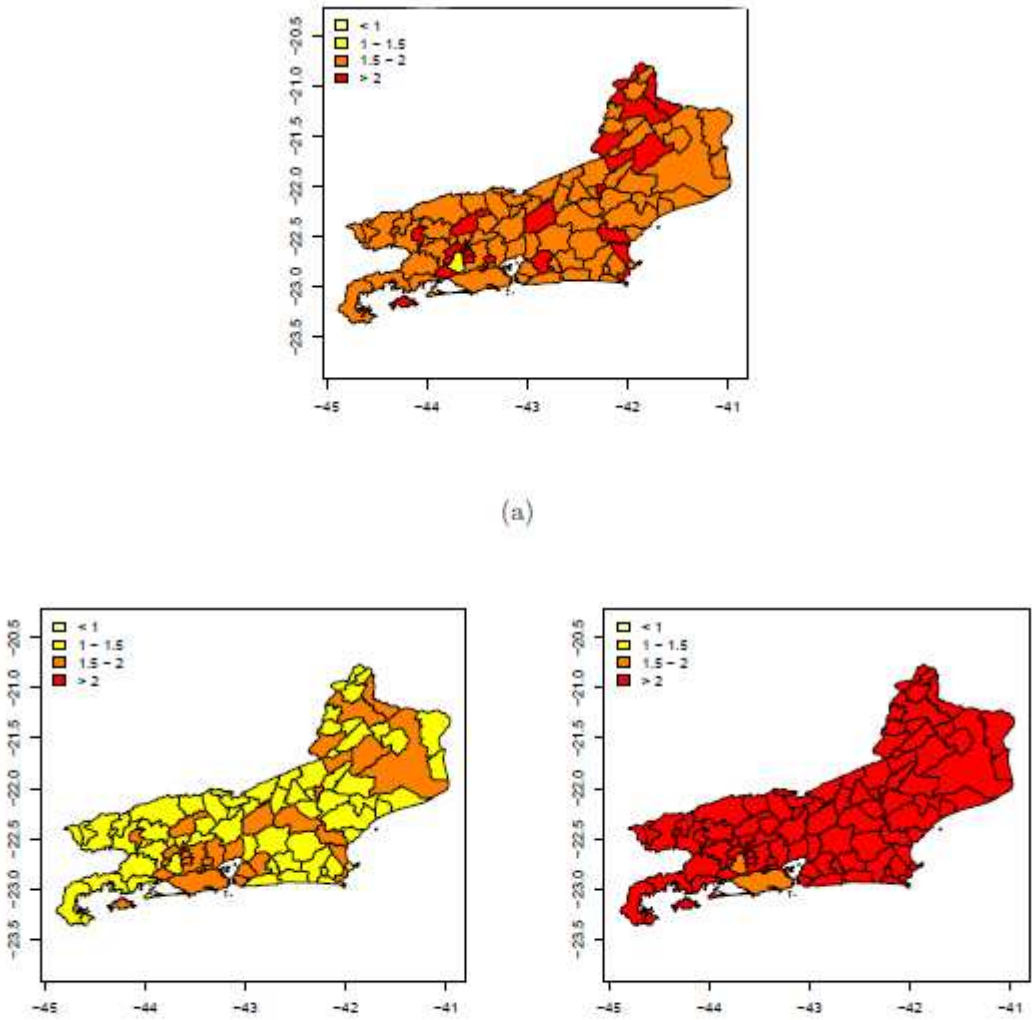


Figura 3: Mapa da Mediana a posteriori dos riscos de mortalidade por doenças respiratórias no estado do Rio de Janeiro no ano de 2003 (a) e mapas de credibilidade de 95% (b) e (c).

Referências bibliográficas

Arbex, M. A., Cançado, J.E.D., Pereira, L. A. A., Braga, A. L. F. and do Nascimento Saldiva, P. H. (2004), "Queima da biomassa e efeitos sobre a saúde", Jornal Brasileiro de Pneumologia, Vol. 30, pp.158-175.

Bernardo, J. M. and Smith, A. F. M. (1994), Bayesian Theory, Wiley.

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems", Journal of the Royal Statistical Society , Vol. 36, pp. 192-236.

Besag, J., York, J. and Mollie, A. (1991), "Bayesian image restoration, with two applications in spatial statistics", Annals of the Institute of Statistical Mathematics , Vol. 43, pp. 1-59.

Bivand, R., with contributions by Micah Altman, Anselin, L., Assunção, R., Berke, O., Bernat, A., Blanchet, G., Blankmeyer, E., Carvalho, M., Christensen, B., Chun, Y., Dormann, C., Dray, S., Halbersma, R., Krainski, E., Legendre, P., Lewin-Koh, N., Li, H., Ma, J., Millo, G., Mueller, W., Ono, H., Peres-Neto, P., Piras, G., Reder, M., Tiefelsdorf, M., and Yu., D. (2011), spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.5-33.
URL: http://CRAN.R-project.org/package=spdep

Bueno, F. (2008), Qualidade do ar e internações por doenças respiratórias em crianças, no município de Divinópolis, MG, Brasil, Master's thesis, Universidade do Estado de Minas Gerais, Fundação Educacional de Divinópolis.

Castro, H., Cunha, M., Mendonça, G., Junger, W., Cunha-Cruz, J. and Leon, A. (2009), "Efeitos da poluição do ar na função respiratória de escolares, Rio de Janeiro, RJ", Revista de Saúde Pública , Vol. 43, pp. 26-34.

Efroymson, M. (1960), Multiple regression analysis., in A. Ralston and H. Wilf, eds, 'Mathematical Methods for Digital Computers', Wiley, New York.

Gamerman, D. and Lopes, H. F. (2006), Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, second edition edn, Chapman and Hall/CRC.

Gouveia, N., Mendonça, G., Leon, A., Correia, J., Junger, W., Freitas, C., Daumas, R., Martins, L., Giussepe, L., Conceição, G., Manerich, A. and Cunha-Cruz, J. (2003), "Poluição do ar e efeitos na saúde nas populações de duas grandes metrópoles brasileiras", Epidemiologia e Serviços de Saúde , Vol. 12, pp. 29-40.

Hougaard, P. (1984), "Life table methods for heterogeneous populations: distributions describing the heterogeneity.", Biometrika , Vol. 71, pp. 75-83.

IBGE (2008), Potencial de poluição industrial do ar no estado do Rio concentra-se em 4 municípios, Technical report, Instituto Brasileiro de Geografia e Estatística.

McCullagh, P. and Nelder, J. (1989), Generalized Linear Models, second edition edn, Chapman and Hall/CRC.

R Development Core Team (2011), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: http://www.R-project.org/

Rue, H., and Martino, S. (2009), INLA: Functions which allow to perform a full Bayesian analysis of structured additive models using Integrated Nested Laplace Approximation. R package version 0.0.

Rue, H., Martino, S. and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations", Journal of the Royal Statistical Society, Series B , Vol. 71, pp. 319-392.

Schrödle, B. and Held, L. (2010), "Spatio-temporal disease mapping using inla", Envirometrics , Vol. 22, pp. 725-734.

Sor, J. L., Clevelario Junior, J., Guimarães, L. T. and de Andrade Memoria Moreno, R. (2008), Relatório piloto com aplicação da metodologia IPPS ao estado do Rio de Janeiro: Uma estimativa do potencial de poluição industrial do ar, Technical report, Diretoria de Geociências, Instituto Brasileiro de Geografia e Estatística - IBGE.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002), "Bayesian measures of model complexity and fit (with discussion)", Journal of the Royal Statistical Society, Series B , Vol. 64, pp. 583-639.

Abstract

The scope of this paper is to demonstrate the use of Poisson regression models with random effects in the detection of patterns of variation in risk of SUS hospital morbidity and mortality for respiratory diseases in the state of Rio de Janeiro (RJ) in year 2003. Inference is based on a Bayesian method using the INLA approach implemented in the R environment. For the Rio de Janeiro dataset, the Poisson regression model with spatial effects proved to be the most appropriate to adjust the rate of hospitalization for respiratory diseases. As for adjusting the rate of mortality from respiratory disease model with random effects was the most appropriate. In both models none of the covariates were statistically significant.

Keywords: Poisson regression, Random effect model; Respiratory disease; INLA; DIC.

REVISTA BRASILEIRA DE ESTATÍSTICA - RBEs

POLÍTICA EDITORIAL

A Revista Brasileira de Estatística - RBEs publica trabalhos relevantes em Estatística Aplicada, não havendo limitação no assunto ou matéria em questão. Como exemplos de áreas de aplicação, citamos as áreas de advocacia, ciências físicas e biomédicas, criminologia, demografia, economia, educação, estatísticas governamentais, finanças, indústria, medicina, meio ambiente, negócios, políticas públicas, psicologia e sociologia, entre outras. A RBEs publicará, também, artigos abordando os diversos aspectos de metodologias relevantes para usuários e produtores de estatísticas públicas, incluindo planejamento, avaliação e mensuração de erros em censos e pesquisas, novos desenvolvimentos em metodologia de pesquisa, amostragem e estimação, imputação de dados, disseminação e confiabilidade de dados, uso e combinação de fontes alternativas de informação e integração de dados, métodos e modelos demográfico e econométrico.

Os artigos submetidos devem ser inéditos e não devem ter sido submetidos simultaneamente a qualquer outro periódico.

O periódico tem como objetivo a apresentação de artigos que permitam fácil assimilação por membros da comunidade em geral. Os artigos devem incluir aplicações práticas como assunto central, com análises estatísticas exaustivas e apresentadas de forma didática. Entretanto, o emprego de métodos inovadores, apesar de ser incentivado, não é essencial para a publicação.

Artigos contendo exposição metodológica são também incentivados, desde que sejam relevantes para a área de aplicação pela qual os mesmos foram motivados, auxiliem na compreensão do problema e contenham interpretação clara das expressões algébricas apresentadas.

A RBEs tem periodicidade semestral e também publica artigos convidados e resenhas de livros, bem como incentiva a submissão de artigos voltados para a educação estatística.

Artigos em espanhol ou inglês só serão publicados caso nenhum dos autores seja brasileiro e nem resida no País.

Todos os artigos submetidos são avaliados quanto à qualidade e à relevância por dois especialistas indicados pelo Comitê Editorial da RBEs.

O processo de avaliação dos artigos submetidos é do tipo 'duplo cego', isto é, os artigos são avaliados sem a identificação de autoria e os comentários dos avaliadores também são repassados aos autores sem identificação.

INSTRUÇÃO PARA SUBMISSÃO DE ARTIGOS À RBEs

O processo editorial da RBEs é eletrônico. Os artigos devem ser submetidos para o site http://rbes.submitcentral.com.br/login.php

Secretaria da RBEs

Revista Brasileira de Estatística – RBEs

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS - IBGE

Rua André Cavalcanti, 106, sala 503-A

Centro, Rio de Janeiro – RJ

CEP: 20031-050

Tels.: 55 21 2142-3596 (Marilene Pereira Piau Câmara – Secretária)

55 21 2142-4957 (Pedro Luis do Nascimento Silva – Editor–Executivo)

Fax: 55 21 2142-0501

INSTRUÇÕES PARA PREPARO DOS ORIGINAIS

Os originais enviados para publicação devem obedecer às normas seguintes:

1. Podem ser submetidos originais processados pelo editor de texto *Word for Windows* ou originais processados em LaTeX (ou equivalente) desde que estes últimos sejam encaminhados e acompanhados de versões em pdf, conforme descrito no item 3, a seguir;

2. 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 afiliação e endereço para correspondência. Agradecimentos a colaboradores e instituições, e auxílios recebidos, se for o caso de constarem no documento, também devem figurar nesta página;

3. No caso de a submissão não ser em *Word for Windows*, três arquivos do original devem ser enviados. O primeiro deve conter os originais no processador de texto utilizado (por exemplo, LaTeX). O segundo e terceiro devem ser no formato pdf, sendo um com a primeira página, como descrito no item 2, e outro contendo apenas o título, sem a identificação do(s) autor(es) ou outros elementos que possam permitir a identificação da autoria;

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

5. 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;

6. Tratamentos algébricos exaustivos devem ser evitados ou alocados em apêndices;

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

8. As tabelas e gráficos devem ser precedidos de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas sequencialmente (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 de trabalho;

9. Gráficos e diagramas para publicação devem ser incluídos nos arquivos com os originais do artigo. Caso tenham que ser enviados em separado, devem ter nomes que facilitem a sua identificação e posicionamento correto no artigo (ex.: Gráfico 1; Figura 3; etc.). É fundamental que não existam erros, quer no desenho, quer nas legendas ou títulos;

10. Não serão permitidos itens que identifiquem os autores do artigo dentro do texto, tais como: número de projetos de órgãos de fomento, endereço, *e-mail*, etc. Caso ocorra, a responsabilidade será inteiramente dos autores; e

11. No caso de o artigo ser aceito para a publicação após a avaliação dos pareceristas, serão encaminhadas as sugestões/comentários aos autores sem a sua identificação. Uma vez nesta condição, é de responsabilidade única dos autores fazer o *download* da formatação padrão da revista (em doc ou em LaTeX) para o envio da versão corrigida.