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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 69 número 231 julho/dezembro 2008

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 69, n. 231, p. 1-119, jul./dez. 2008

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

Revista Brasileira de Estatística, ISSN 0034-7175

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

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

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

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

Francisco Louzada-Neto
Editor responsável - RBEs - IBGE.
Rua André Cavalcanti, 106 - Santa Teresa
20231-050 - Rio de Janeiro, RJ.

Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.
A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Editoração
Ismenia Blavatsky de Magalhães - Escola Nacional de Ciências Estatísticas - ENCE/IBGE

Impressão
Gráfica Digital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2009.

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 do Editor *5*

Artigos

Uma aplicação de influência para a distribuição Dirichlet.....*7*
Gecynalda S. S. Gomes
Francisco Cribari–Neto
Klaus L.P. Vasconcellos

Modelagem de séries representativas do setor energético Brasileiro.....*33*
Luciene Resende Gonçalves
Thelma Sáfadi

Modelo de previsão combinada: Uma aplicação à série mensal de passageiros transportados do Sistema de Transporte Público da Região Metropolitana do Recife – PE *69*

Dirac M. Cordeiro
Gauss M. Cordeiro

Avaliação do uso de redes bayesianas discretas para imputação de dados.....*89*
Ismenia Blavatsky
Fabio G. Cozman

Política editorial *117*

Nota do Editor

Este segundo volume da RBEs do ano de 2008 é composto por quatro artigos. O primeiro artigo, de autoria de Gecynalda S.S. Gomes, Francisco Cribari-Neto e Klaus L.P. Vasconcelos, que avalia a influência local das observações obtidas de uma distribuição Dirichlet, que é uma generalização multivariada da distribuição beta. O segundo artigo, de autoria de Luciene R. Gonçalves e Thelma Sáfadi, que apresenta uma modelagem de séries representativas do setor energético brasileiro. O terceiro artigo, de autoria de Dirac M. Cordeiro e Gauss M. Cordeiro, que desenvolve uma modelagem de previsão combinada aplicada à série mensal de passageiros transportados pelo sistema de transporte público da Região Metropolitana do Recife-PE. O quarto artigo, de autoria de Ismenia Blavatsky e Fabio G. Cozman avalia a utilização de redes Bayesianas discretas para imputação de dados.

Aproveito a oportunidade para me despedir da editoração do periódico. Agradeço à direção da ABE por ter me confiado a tarefa de editorar a RBEs, e informo que foi com muito prazer que atuei como Editor-Responsável, juntamente com todo o Corpo Editorial. Nossa atuação se deu no período de agosto de 2004 até o momento, durante o qual foram publicados dez volumes do periódico, do volume 222 ao 231.

Também, agradeço a colaboração de todos os Editores Associados, revisores do periódico, autores e à equipe do IBGE. Em particular, gostaria de agradecer a Ismenia Blavastsky, que com muita garra foi responsável da editoração executiva da revista, aos alunos de graduação, monitores da Escola Nacional de Ciências Estatísticas, Juliana F. C. Macedo e Jéferson M. Ramos que colaboraram na editoração deste volume.

Finalmente gostaria de desejar, em nome de todo o Corpo Editorial atual da RBEs, uma excelente gestão para o novo Corpo Editorial que deverá tomar posse em breve.

Uma excelente leitura.

Francisco Louzada-Neto
Editor Responsável

Uma aplicação de influência para a distribuição Dirichlet

Gecynalda S.S. Gomes[*]
Francisco Cribari–Neto[**]
Klaus L.P. Vasconcellos[***]

Resumo

Este artigo avalia a influência local das observações obtidas de uma distribuição Dirichlet, que é uma generalização multivariada da distribuição beta. Com esta finalidade, nós empregamos a medida proposta por Poon & Poon (1999). Diferentes esquemas de perturbação são considerados. A metodologia é aplicada a dados simulados e reais.

Palavras-chave: Curvatura normal conforme, distribuição Dirichlet, esquema de perturbação.

1. Introdução

Modelos estatísticos são extremamente úteis para descrever características essenciais de um conjunto de dados. Todavia, esses modelos são descrições aproximadas de algum processo mais complexo. O estudo da variação dos resultados de uma análise sob modestas modificações torna-se, assim, importante. Caso exista em uma descrição aproximada uma pequena modificação que influencie severamente resultados fundamentais de uma análise, seguramente há motivos para preocupação. Por outro lado, se tais modificações não têm importância, dizemos que a amostra é insensível com respeito às perturbações induzidas.

∗ Endereço para correspondência: Departamento de Estatística, CCEN - Universidade Federal de Pernambuco, Av. Prof. Luiz Freire, s/n^0 - Cidade Universitária, CEP: 50740-540, Recife-PE-BRASIL.-- *e-mail*: gssg@cin.ufpe.br.
∗∗ *e-mail*: cribari@de.ufpe.br.
∗∗∗ *e-mail*: klaus@de.ufpe.br.

Um dos métodos mais modernos de diagnóstico foi introduzido por Cook (1986). Um princípio fundamental consiste em estudar o comportamento de alguma medida particular de influência que mede pequenas perturbações nos dados ou no modelo e, assim, verificar a existência de pontos que sob modestas modificações no modelo causam variações significativas nos resultados (PAULA, 2004).

O método proposto por Cook (1986) utiliza conceitos de geometria diferencial para avaliar o comportamento da função de deslocamento da verossimilhança. O método é de grande utilidade, porém várias questões foram levantadas pelo autor, tais como o fato de que a curvatura normal depende de como o modelo será perturbado, de que a curvatura normal não é invariante sob reparametrização, entre outras. Poon & Poon (1999) propuseram métodos que solucionam esses e outros problemas. Eles construíram uma medida, denominada curvatura normal conforme, que é função biunívoca da curvatura normal e restrita ao intervalo [0, 1].

Pesquisadores frequentemente precisam modelar dados restritos ao intervalo (0, 1), como, por exemplo, a taxa de desemprego, alguma medida de concentração de renda, proporções de pacientes infectados por alguma doença, etc. A distribuição beta fornece um aparato flexível para a modelagem de dados dessa natureza. Quando o interesse é trabalhar com modelos multivariados e estudar a distribuição conjunta de variáveis que estão compreendidas no intervalo (0, 1), podemos fazer uso da distribuição Dirichlet, que é uma generalização da distribuição beta em situações em que a soma destas variáveis é sempre menor ou igual a um.

A principal contribuição deste artigo é avaliar a influência local através da curvatura normal conforme para dados correspondentes a variáveis independentes e identicamente distribuídas (*i.i.d.*) segundo a distribuição Dirichlet. A distribuição a ser estudada é multivariada com *p* componentes e, segundo a sugestão de Cook (1986) de perturbar variáveis respostas (ou covariáveis), perturbaremos uma das $p-1$ variáveis aleatórias, Y_j. O esquema de perturbação considerado é aquele em que uma das componentes é modificada para permitir perturbações convenientes no modelo, essas perturbações sendo dos tipos aditivo e multiplicativo; outro esquema considerado foi o de perturbar a log-verossimilhança multiplicativamente.

Um outro objetivo deste trabalho é comparar as estimativas dos parâmetros da distribuição Dirichlet sem a presença de determinada observação influente com as estimativas dos parâmetros na presença de todas as observações, com a finalidade de verificar o quanto as estimativas se modificam com a retirada de qualquer observação influente, ou seja, de

qualquer observação que produza mudanças substanciais nas estimativas dos parâmetros da distribuição.

O presente artigo está organizado da seguinte forma. A Seção 2 apresenta a distribuição Dirichlet, a Seção 3 mostra a construção da curvatura normal conforme e as vantagens de utilizá-la para avaliar a influência local, na Seção 4 são apresentados os esquemas de perturbação e as funções de log-verossimilhança, na Seção 5 apresentamos resultados de uma aplicação empírica cujos dados são reproduzidos em Aitchison (2003) e na Seção 6 apresentamos as considerações finais.

2. Distribuição Dirichlet

A distribuição Dirichlet é uma generalização da distribuição beta e pode ser empregada no estudo da distribuição conjunta de variáveis que estejam compreendidas no intervalo (0, 1) cuja soma é menor ou igual a um. Para maiores detalhes desta distribuição ver em Schervish (1995), Gelman *et. al*. (1995), Minka (2003), entre outros. A seguir apresentaremos definições e algumas propriedades dessa distribuição.

Sejam $Y_1, Y_2, \ldots, Y_{p-1}$ variáveis aleatórias. Elas são distribuídas de acordo com a distribuição Dirichlet quando sua função densidade de probabilidade é

$$f\left(\underset{\sim}{y}\right) = \begin{cases} \dfrac{\Gamma\left(\sum_{j=1}^{p} \alpha_j\right)}{\prod_{j=1}^{p} \Gamma(\alpha_j)} \left(\prod_{j=1}^{p-1} y_j^{\alpha_j - 1}\right)\left(1 - \sum_{j=1}^{p-1} y_j\right)^{\alpha_p - 1}, & \underset{\sim}{y} \in \Re, \\ 0, & \underset{\sim}{y} \notin \Re, \end{cases} \tag{1}$$

em que $\underset{\sim}{y} = (y_1, y_2, \ldots, y_{p-1})$, os parâmetros α's são estritamente positivos e a região $\Re$ é dada por:

$$\Re = \left\{(y_1, y_2, \ldots, y_{p-1}) : y_j > 0; \sum_{j=1}^{p-1} y_j < 1\right\}.$$

A Figura 1 apresenta diferentes densidades Dirichlet bivariadas com parâmetros α_1, α_2 e α_3, onde fixamos $\alpha_1 = \alpha_2 = 2$ e variamos α_3. Observamos que a forma da densidade muda de acordo com as diferentes escolhas de α_3. Podemos observar ainda que a variabilidade da distribuição diminui à medida que aumentamos o valor de α_3. (Note que

fixamos α_1 e α_2 e variamos α_3 sem perda de generalidade, uma vez que a forma da distribuição é invariante por permutação de qualquer coordenada.)

Seja $\phi = \sum_{j=1}^{p} \alpha_j$. O valor esperado, a variância e a estrutura de covariância da distribuição Dirichlet são, respectivamente,

$$\mathrm{E}(Y_j) = \frac{\alpha_j}{\phi}, \quad j = 1, \ldots, p-1,$$

$$\mathrm{Var}(Y_j) = \frac{\alpha_j(\varphi - \alpha_j)}{\phi^2(\phi+1)}, \quad j = 1, \ldots, p-1,$$

$$\mathrm{Cov}(Y_j) = -\frac{\alpha_j \alpha_h)}{\phi^2(\phi+1)}, \quad j \neq h,\ j, h = 1, \ldots, p-1.$$

Sejam amostras *i.i.d.* de tamanho n das variáveis aleatórias $Y_1, Y_2, \ldots, Y_{p-1}$, as quais são distribuídas conjuntamente de acordo com a distribuição Dirichlet com função densidade de probabilidade (1). A função de verossimilhança é dada por

$$L = \prod_{i=1}^{n} \left\{ \frac{\Gamma\left(\sum_{j=1}^{p} \alpha_j\right)}{\prod_{j=1}^{p} \Gamma(\alpha_j)} \prod_{j=1}^{p-1} Y_{ij}^{\alpha_j - 1} \left(1 - \sum_{j=1}^{p-1} Y_{ij}\right)^{\alpha_p - 1} \right\}. \tag{2}$$

As estatísticas suficientes para os parâmetros $\alpha_1, \alpha_2, \ldots, \alpha_p$ são as médias geométricas de Y_{ij} (Barbosa, 1977):

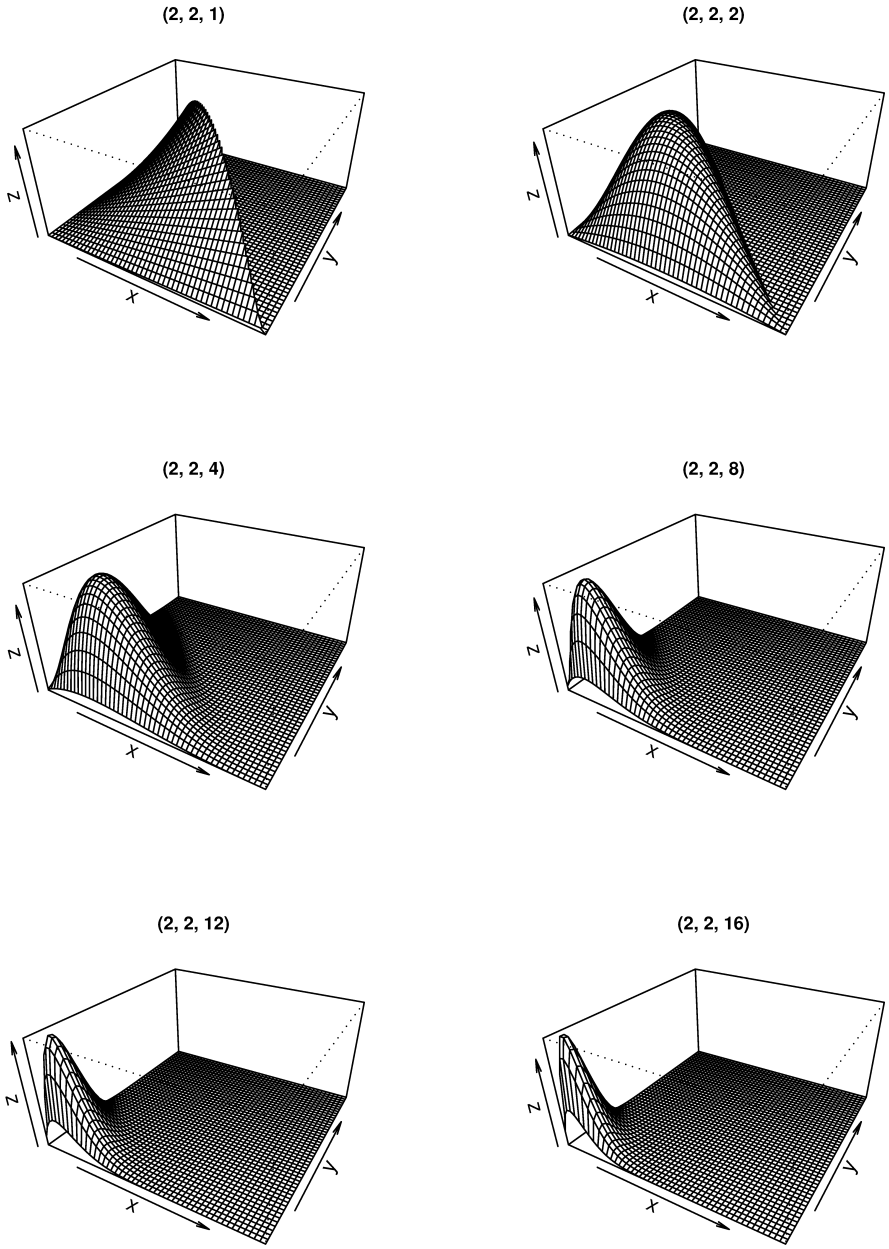
$$G_j = \left\{ \prod_{i=1}^{n} Y_{ij} \right\}^{1/n}, \ j = 1, \ldots, p-1,$$

$$G_p = \left\{ \prod_{i=1}^{n} Y^{ij} \left(1 - \sum_{j=1}^{p-1} Y_{ij}\right) \right\}^{1/n}.$$

Portanto, estas quantidades sintetizam toda a informação da amostra. O logaritmo (base natural) da função de verossimilhança (2) pode ser escrito como:

$$\ell = \log L = \sum_{i=1}^{n} \left\{ \log\Gamma\left(\sum_{j=1}^{p} \alpha_j\right) - \sum_{j=1}^{p} \log\Gamma(\alpha_j) + \sum_{j=1}^{p} (\alpha_j - 1) \log G_j \right\}.$$

Figura 1 - Gráfico da densidade Dirichlet bivariada no quadrado unitário para diferentes valores de α_3 fixando $\alpha_1 = 2$ e $\alpha_2 = 2$



Os estimadores de máxima verossimilhança dos parâmetros da distribuição Dirichlet são obtidos maximizando (2) com respeito aos parâmetros α's. A derivada parcial de (2) em relação a α_j é

$$\frac{\partial \log L}{\partial \alpha_j} = \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \alpha_j} \log \Gamma(\phi) - \frac{\partial}{\partial \alpha_j} \log \prod_{j=1}^{p} \Gamma(\alpha_j) + \sum_{j=1}^{p} (\alpha_j - 1) \log G_j \right\}. \tag{3}$$

O valor de cada α_j, $j = 1, \ldots, p-1$, que torna (3) igual a zero é a solução da equação

$$\log G_j + \psi(\hat{\alpha}_j) - \psi(\hat{\phi}), \tag{4}$$

onde $\psi(\phi) = \frac{\partial}{\partial \alpha_j} \log \Gamma(\phi)$, ψ representando a função digama.

O sistema de equações em (4) é não-linear nos parâmetros α's e não pode ser solucionado analiticamente, uma vez que a solução não possui forma fechada. É necessário, portanto, maximizar a função de log-verossimilhança numericamente através de um algoritmo iterativo (*e.g.*, Newton-Raphson, BFGS, etc.).

3. Curvatura normal conforme e influência local

Poon & Poon (1999) desenvolveram uma metodologia através de uma geometria simples, denominada curvatura normal conforme, que é função biunívoca da curvatura normal e assume valores no intervalo [0, 1]. Segundo Kobayashi (1972), uma das várias propriedades deste método é a invariância conforme (KOBAYASHI, 1972), ou seja, invariância sob uma transformação conforme. Com base nesta e em outras propriedades, valores críticos podem ser obtidos para a influência local de observações. Adicionalmente, através da curvatura normal conforme, é possível construir uma medida agregada para cada vetor básico do espaço de perturbação e diferentes autovetores podem ser usados na agregação. Poon & Poon (1999) afirmam que quando todos os autovetores são considerados, a contribuição agregada para cada vetor de perturbação básico é equivalente à sua curvatura normal conforme. Nesta seção serão apresentadas as definições dessa metodologia.

3.1. Curvatura normal conforme

Seja $\ell(\theta)$ a log-verossimilhança para o modelo sem perturbação, em que θ é um vetor $p \times 1$ de parâmetros desconhecidos. Adicionalmente, seja $\ell(\theta \mid \omega)$ a log-verossimilhança

correspondente ao modelo perturbado para um dado ω, onde $\omega^{\mathsf{T}} = (\omega_1, \ldots, \omega_n)$ é um vetor $n \times 1$ em $\Omega \subset \mathbb{R}^n$, em que Ω representa o conjunto de perturbações relevantes. Supõe-se que há ω_0 tal que $\ell(\theta \mid \omega_0) = \ell(\theta)$ para todo θ. Sejam $\hat{\theta}$ e $\hat{\theta}_\omega$ estimadores de máxima verossimilhança de θ sob $\ell(\theta \mid \omega_0)$ e $\ell(\theta \mid \omega)$, respectivamente. A função de deslocamento da verossimilhança (Cook, 1986) é dada por

$$f(\omega) = 2\{\ell(\hat{\theta} \mid \omega_0) - \ell(\hat{\theta}_\omega \mid \omega_0)\}$$

e pode ser usada para avaliar a influência de perturbações pequenas. Seja

$$\alpha(\omega) = (\omega^{\mathsf{T}}, f(\omega))^{\mathsf{T}}$$

o gráfico da influência, formado pelos valores de um vetor $(n+1) \times 1$. Cook (1986) propôs a curvatura normal para avaliar a influência local da perturbação A curvatura normal do gráfico α na direção de l no ponto ω_0 é

$$C_l = C(l,l) = -2\{l^{\mathsf{T}} \Delta^{\mathsf{T}} (\ddot{L})^{-1} \Delta l\}\Big|_{\theta=\hat{\theta}, \omega=\omega_0},$$

onde Δ é uma matriz $p \times n$ e $\ddot{L}$ é uma matriz $p \times p$ cujos elementos são dados, respectivamente, por

$$\Delta_{it} = \frac{\partial^2 \ell(\theta \mid \omega)}{\partial \theta_i \partial \omega_t} \text{ e } \ddot{L}_{it} = \frac{\partial^2 \ell(\theta \mid \omega)}{\partial \theta_i \partial \theta_t},$$

avaliados em $\theta = \hat{\theta}$ e $\omega = \omega_0$.

A matriz $-\Delta^{\mathsf{T}} (\ddot{L})^{-1} \Delta$ é positiva semidefinida em ω_0, uma vez que a função de deslocamento da verossimilhança atinge mínimo neste ponto. Quando l tem a direção do autovetor correspondente ao maior autovalor de $-\Delta^{\mathsf{T}} (\ddot{L})^{-1} \Delta$, a curvatura normal C_l atinge seu máximo.

A curvatura normal conforme no ponto ω_0 do gráfico α na direção de l é

$$B_l = -\frac{l^{\mathsf{T}} \Delta^{\mathsf{T}} (\ddot{L})^{-1} \Delta l}{\sqrt{\operatorname{tr}\{\left(-\Delta^{\mathsf{T}} (\ddot{L})^{-1} \Delta\right)^2\}}}\Bigg|_{\theta=\hat{\theta}, \omega=\omega_0}.$$

A curvatura normal conforme possui muitas propriedades que estão resumidas em alguns teoremas em Poon & Poon (1999). O interesse inicial é na transformação conforme, que é uma transformação múltipla de uma transformação ortogonal. Uma matriz M de dimensão $n \times n$ é dita conforme se houver um número positivo τ tal que $M^{\mathsf{T}} M = \tau I_n$. A transformação aqui foi considerada como “modificação do esquema de perturbação” (COOK,

1986). Um exemplo de reparametrização conforme visto em Poon & Poon (1999) é $\phi(\omega) = M\omega + c$, tal que M é a matriz conforme. Por exemplo, na análise de regressão linear, quando $\omega^{\mathrm{T}} = (\omega_1, \ldots, \omega_n)$ é o vetor de perturbação para uma amostra n (COOK, 1986, Seção 4), a curvatura normal conforme é invariante com respeito à reparametrização $(1+\omega_i)/2$, isso foi estudado por Loynes (1986). Neste caso, $c = (\frac{1}{2}, \ldots, \frac{1}{2})$ é um vetor $n \times 1$ e $M = \mathrm{diag}(\frac{1}{2}, \ldots, \frac{1}{2})$ é uma matriz $n \times n$.

Seja Ω o conjunto das perturbações. Se a transformação de Ω é conforme no ponto crítico ω_0 no gráfico de f sobre Ω, então a curvatura normal conforme em qualquer direção é invariante sob transformação.

Para qualquer direção l, B_l satisfaz a condição de que $0 \leq |B_l| \leq 1$, ou seja, B_l é uma medida normalizada e, assim, torna-se mais fácil a interpretação do seu valor. Além disso, se $e_1, \ldots, e_n$ são autovalores de $-\ddot{F}$, B_{e_i} corresponde ao autovalor normalizado $\hat{\lambda}_i$, que é dado por:

$$\hat{\lambda}_i = \frac{\lambda_i}{\sqrt{\sum_{i=1}^{n} \lambda_i^2}}. \tag{5}$$

Note que $\sum_i B_{e_i}^2 = 1$ e, assim, se as curvaturas normais conformes para todos os autovetores forem idênticas, então todas serão iguais a $1/\sqrt{n}$. Com isso, a avaliação da influência local pode ser feita de uma forma mais objetiva e sistemática.

3.2. Avaliação da influência local

Foi visto que a curvatura normal C_l e a direção associada l são usadas para avaliar a influência local. Assim, Cook (1986) sugere examinar o autovetor $e_{\max}$ com a curvatura normal máxima $C_{\max}$ independentemente do seu tamanho.

Quando l tiver a direção do autovetor $e_{\max}$ que corresponde ao maior autovalor, a curvatura normal conforme B_l assumirá o valor máximo. A curvatura normal e a curvatura normal conforme são medidas de diagnóstico equivalentes, diferindo apenas por um fator positivo. Entretanto, é preferível utilizar a curvatura normal conforme devido à sua propriedade de invariância e à sua natureza normalizada, que facilitam sua interpretação. Adicionalmente,

Poon & Poon (1999) sugerem que um valor de referência seja calculado para comparar os efeitos de B_{e_i} e B_l em vários níveis. Portanto, faz-se necessária a seguinte definição: um autovetor e é q-influente se $|B_e| \geq q/\sqrt{n}$.

Nós podemos verificar de forma mais abrangente se um autovetor é influente. Seja E_t o t-ésimo vetor de perturbação básico do espaço de perturbação cujo t-ésimo elemento é 1 e todos os restantes são iguais a 0. Seja $\{e_i : 1 \leq i \leq n\}$ uma sequência de autovetores ortonormais com autovalor normalizado correspondente $\hat{\lambda}_i$ dado na equação (5).

Quando $e_i = \sum_{t=1}^{n} a_{it} E_t$, temos $\sum_{t=1}^{n} a_{it}^2 = 1$. Isto significa que, para qualquer i fixo, se a contribuição de todos for uniforme, então $|a_{it}| = 1/\sqrt{n}$. Com isso, pode-se construir, inclusive, o valor de referência. Este método pode ser aplicado para estudar $e_{\max}$ ou quaisquer individuais autovetores influentes, de acordo com Poon & Poon (1999).

Os vetores de perturbação básicos próximos de e são encontrados para analisar a contribuição do vetor de perturbação básico para a influência do autovetor influente e.

De forma mais geral, a análise dos vetores de perturbação básicos pode ser feita para todos os autovetores influentes, definindo $\mu_i = |\hat{\lambda}_i|$ e usando a_{it} para denotar o t-ésimo elemento do autovetor normalizado correspondente a μ_i. A contribuição agregada do t-ésimo vetor de perturbação básico para todos os autovetores q-influentes é

$$m(q)_t = \sqrt{\sum_{i=1}^{k} \mu_i a_{it}^2},$$

onde $1 \leq t \leq n$. Poon & Poon (1999) sugerem comparar a contribuição agregada $m(q)_t$ com

$$\overline{m}(q) = \sqrt{\frac{1}{n} \sum_{i=1}^{k} \mu_i},$$

ou seja, se a contribuição de todos os vetores de perturbação básicos é uniforme, então cada contribuição deve ser igual a $\overline{m}(q)$. Isto significa que, ao analisarmos a contribuição $\overline{m}(q)$, determinamos a significância da contribuição individual dos vetores de perturbação básicos.

Poon & Poon (1999) consideram duas possibilidades. Uma é fazer com que q seja suficientemente grande para que a contribuição individual dos vetores de perturbação básicos seja considerada apenas para $e_{\max}$. Outra é fazer $q = 0$ de modo que todos os autovalores sejam incluídos na análise.

Quando $q = 0$, a contribuição $m(q)_t$ é chamada de contribuição total e é dada por

$$m_t = m(0)_t = \sqrt{\sum_{i=1}^{n} \mu_i a_{it}^2} .$$

Poon & Poon (1999) sugerem comparar a contribuição total $m(0)_t$ com

$$\overline{m} = \overline{m}(0) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mu_i} = \sqrt{\sum_{i=1}^{n} |\lambda_i| \Big/ n\sqrt{\sum_{i=1}^{n} \lambda_i^2}} ,$$

ou seja, se a contribuição de todos os parâmetros básicos de perturbação for uniforme, então cada contribuição total deve ser igual a $\overline{m}(0)$.

A contribuição total m_t e a curvatura normal conforme B_{E_t} do vetor de perturbação básico são altamente relacionadas. Se todos os autovalores forem não-negativos, então B_{E_t} será igual ao quadrado da contribuição total do t-ésimo vetor de perturbação básico. Resumindo, se a matriz hessiana H_f é positiva semi-definida e todos os autovalores são não-negativos, então $m_t^2 = B_{E_t}, \forall t$.

Se a contribuição de todos B_{E_t} 's for uniforme, então cada contribuição deve ser igual a

$$b = \overline{m}^2 .$$

Logo, b pode ser utilizado como valor crítico para as curvaturas dos vetores de perturbação básicos.

4. Esquemas de perturbação e funções de log-verossimilhança

Para o nosso estudo utilizamos dois esquemas de perturbação, um em que uma das componentes é modificada através da forma e outro em que perturbamos a log-verossimilhança de forma multiplicativa. Apresentaremos nesta seção a forma como cada esquema foi trabalhado. Seja Y a matriz $n \times p$ cuja i-ésima linha representa a i-ésima observação na amostra. Para o esquema em que perturbamos uma das componentes da matriz Y, apresentaremos as matrizes para cada tipo de perturbação aditiva 1 e aditiva 2. Para o esquema em que perturbamos a função de log-verossimilhança será apresentada apenas a função com perturbação multiplicativa. O modelo é perturbado para um determinado ω, onde $\omega^{\mathsf{T}} = (\omega_1, \ldots, \omega_n)$. Nesta seção, apresentaremos a forma da matriz Y para o modelo sem perturbação e para os esquemas de perturbação aqui estudados, assim como as funções de log-verossimilhança para esses esquemas.

4.1. Formas das matrizes

- Matriz Y para o modelo sem perturbação:

$$\begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1p-1} & Y_{1p} \\ Y_{21} & Y_{22} & \dots & Y_{2p-1} & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Y_{n1} & Y_{n1} & \dots & Y_{np-1} & Y_{np} \end{pmatrix}.$$

- Matriz Y para o esquema de perturbação aditiva 1:

$$\begin{pmatrix} Y_{11}+\omega_1 & Y_{12}-\frac{\omega_1}{p-2} & \dots & Y_{1p-1}-\frac{\omega_1}{p-2} & Y_{1p} \\ Y_{21}+\omega_2 & Y_{22}-\frac{\omega_2}{p-2} & \dots & Y_{2p-1}-\frac{\omega_2}{p-2} & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Y_{n1}+\omega_n & Y_{n1}-\frac{\omega_n}{p-2} & \dots & Y_{np-1}-\frac{\omega_n}{p-2} & Y_{np} \end{pmatrix}.$$

- Matriz Y para o esquema de perturbação aditiva 2:

$$\begin{pmatrix} Y_{11}+\omega_1 & Y_{12}-\frac{\omega_1}{p-1} & \dots & Y_{1p-1}-\frac{\omega_1}{p-1} & Y_{1p} \\ Y_{21}+\omega_2 & Y_{22}-\frac{\omega_2}{p-1} & \dots & Y_{2p-1}-\frac{\omega_2}{p-1} & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Y_{n1}+\omega_n & Y_{n1}-\frac{\omega_n}{p-1} & \dots & Y_{np-1}-\frac{\omega_n}{p-1} & Y_{np} \end{pmatrix}.$$

- Para os esquemas de perturbação aditiva, $\omega^{\mathsf{T}} = \omega_0^{\mathsf{T}}$ (0, 0, ..., 0).

  - Esquema de perturbação multiplicativa na log-verossimilhança:

$$\ell = (\alpha \mid \omega) = \sum_{i=1}^{n} \omega_i \ell_i(\alpha).$$

  - Para os esquemas de perturbação multiplicativa, $\omega^{\mathsf{T}} = \omega_0^{\mathsf{T}}$ (1, 1, ..., 1).

4.2. Funções de log-verossimilhança

Dadas n observações independentes com mesma distribuição Dirichlet com parâmetros $\alpha_1, \alpha_2, \ldots, \alpha_p$, temos que a função de log-verossimilhança para o modelo sem perturbação é dada por:

$$\ell(\alpha_j) = n\left\{\log\Gamma(\phi) - \sum_{j=1}^{p}\log\Gamma(\alpha_j)\right\} + \sum_{i=1}^{n}\left\{\sum_{j=1}^{p-1}(\alpha_j - 1)\log y_{ij} + (\alpha_p - 1)\log\left(1 - \sum_{j=1}^{p-1} y_{ij}\right)\right\}.$$

4.2.1. Log-verossimilhança para o esquema de perturbação aditiva 1

Para a função de log-verossimilhança para o modelo sem perturbação, temos $\ell(\alpha_j \mid \omega_0) = \ell(\alpha_j)$. Para a função de log-verossimilhança para o esquema de perturbação aditiva 1 considerada, $\ell(\alpha_j \mid \omega)$, é dada por

$$\begin{aligned}\ell(\alpha_j \mid \omega) = n\left\{\log\Gamma(\phi) - \sum_{j=1}^{p}\log\Gamma(\alpha_j)\right\} + \sum_{i=1}^{n}\{(\alpha_1 - 1)\log(y_{i1} + \omega_i) \\ + \sum_{j=2}^{p-1}(\alpha_j - 1)\log\left(y_{ij} - \frac{\omega_i}{p-2}\right) + (\alpha_p - 1)\log\left(1 - (y_{i1} + \omega_i) - \right. \\ \left. - \sum_{j=2}^{p-1}\left(y_{ij} - \frac{\omega_i}{p-2}\right)\right)\Bigg\}.\end{aligned}$$

4.2.2. Log-verossimilhança para o esquema de perturbação aditiva 2

A função de log-verossimilhança para o esquema de perturbação aditiva 2 considerada, $\ell(\alpha_j \mid \omega)$, é dada por

$$\ell(\alpha_j \mid \omega) = n\left\{\log\Gamma(\phi) - \sum_{j=1}^{p}\log\Gamma(\alpha_j)\right\} + \sum_{i=1}^{n}\{(\alpha_1 - 1)\log(y_{i1} + \omega_i) + \sum_{j=2}^{p}(\alpha_j - 1)\log\left(y_{ij} - \frac{\omega_i}{p-1}\right).$$

4.2.3. Função para o esquema de perturbação multiplicativa log-verossimilhança

A função de verossimilhança para o esquema de perturbação multiplicativa na log-verossimilhança considerada, $\ell(\alpha_j \mid \omega)$, é dada por:

$$\ell(\alpha_j \mid \omega) = \sum_{i=1}^{n}\omega_i l_i(\alpha_j) = \sum_{i=1}^{n}\omega_i\left\{\log\Gamma(\phi) - \sum_{j=1}^{p}\log\Gamma(\alpha_j) + \sum_{j=1}^{p-1}(\alpha_j - 1)\log y_{ij} + (\alpha_p - 1)\log\left(1 - \sum_{j=1}^{p-1}y_{ij}\right)\right\}.$$

5. Aplicação empírica

5.1. Introdução

Nesta seção utilizaremos, como exemplo, um conjunto de dados retirado de Aitchison (2003) sobre composição de rochas do tipo *kongite*.

Os dados $Y_{tj}, t = 1,\ldots,n$ e $i = 1,\ldots,p$ estão disponíveis como p amostras *i.i.d.* de tamanho n tais que $(Y_{t1},\ldots,Y_{tp-1})$ tem distribuição Dirichlet e $\sum_{j=1}^{p} y_j = 1$, para todo t fixo. Temos, porém, que trabalhar numericamente com $p-1$ variáveis, por isso é necessário retirar uma das p variáveis na análise numérica. Optamos por fazer uma permutação nessas retiradas, inicialmente fixando a primeira, depois a segunda, e assim por diante, para poder analisar a influência das observações. Com isso, estaremos analisando a influência em várias direções.

As contribuições agregadas $(m(q)_t)$ do t-ésimo componente de perturbação básico para todos os autovetores q influentes, com q variando de 0 a 3, assim como a curvatura

normal conforme do vetor de perturbação básico (B_{E_t}), foram calculadas para os esquemas com os dois tipos de perturbação.

Nas tabelas, são apresentados os resultados referentes às observações identificadas como influentes para um esquema de determinado tipo de perturbação, assim como os resultados das observações que foram identificadas como influentes por outro esquema, sendo que esses resultados serão apenas ilustrativos.

A contribuição total $m(0)_t$ não é apresentada porque é equivalente a B_{E_t}. Segundo a prática comum, se a contribuição individual agregada é maior que duas vezes a média, então a observação correspondente é considerada influente. Utilizamos $2b$ e $\sqrt{2}\overline{m}(q)$ como valores críticos para B_{E_t} e $m(q)_t$, respectivamente, onde b e $\overline{m}(q)$ são definidos na Seção 3.2. Nós usamos $\sqrt{2}$ em vez de 2 nos valores críticos para $m(q)_t$ devido à relação quadrática entre $m(q)_t$ e $\overline{m}(q)$. A partir das observações verificadas como influentes, faremos uma retirada de tais observações uma a uma, com a finalidade de comparar as novas estimativas dos parâmetros com as estimativas iniciais e verificar o impacto que cada observação causa nessas estimativas.

A avaliação da influência será definida da seguinte forma:

- se $m(1)_t$ for maior que seu valor crítico, diremos que a observação é influente (ou levemente influente);
- se $m(2)_t$ for maior que seu valor crítico, diremos que a observação é bem influente (ou moderadamente influente);
- se $m(3)_t$ for maior que seu valor crítico, diremos que a observação é extremamente influente (ou fortemente influente); e
- se B_{E_t} for maior que seu valor crítico, diremos que o nível de influência da observação é muito grande, ou seja, que a observação é mais influente que a média.

Os esquemas de perturbação que serão apresentados são os esquemas de perturbação aditiva nas variáveis. O esquema de perturbação multiplicativa na verossimilhança não será mostrado porque não foi identificada nenhuma observação influente segundo esse esquema; isso significa que, de fato, as observações têm características da distribuição Dirichlet.

5.2. Dados de composição de *kongite*

Considere os dados reproduzidos em Aitchison (2003, p. 356) sobre composição de 25 espécimes de rochas do tipo *kongite*. Cada composição consiste do percentual por peso de cinco minerais, que são *albite* (Y_{t1}), *blandite* (Y_{t2}), *cornite* (Y_{t3}), *daubite* (Y_{t4}) e *endite* (Y_{t5}), $t = 1,..., n$, onde n representa o número de observações.

Neste caso, há cinco componentes ($p = 5$) de uma distribuição que modelamos como Dirichlet, porém precisamos trabalhar com apenas $p-1$ variáveis aleatórias, por isso temos que escolher uma das variáveis para ser retirada da análise. O procedimento que seguimos foi o seguinte: a variável Y_{t1}, que representa a primeira componente, será perturbada segundo os esquemas definidos anteriormente e essa será a variável fixa; a variável a ser retirada pode ser qualquer uma das que restaram; optamos por fazer uma permutação entre essas variáveis com o intuito de verificar a influência em várias direções.

5.2.1. Perturbando a variável *albite* (componente 1)

Na Tabela 1, apresentamos os autovalores normalizados não-nulos para os esquemas de perturbação trabalhados com as diferentes direções utilizadas. Podemos observar que, no caso da perturbação aditiva 1, quando retiramos o mineral *blandite* obtivemos o maior autovalor entre os autovalores máximos com a retirada de cada mineral; com isso concluímos que a maior direção ocorre com a retirada de *blandite*. Entretanto, vamos apresentar a influência em várias direções. Nota-se que, através dos esquemas de perturbações aditiva 2 e multiplicativa na log-verossimilhança, os autovalores normalizados são iguais com a retirada de qualquer mineral, ou seja, a inferência é invariante em relação à escolha de variáveis a serem retiradas do modelo Dirichlet.

Tabela 1 - Autovalores normalizados

Perturbação	Valores com a retirada desses minerais			
	Endite	***Daubite***	***Cornite***	***Blandite***
Aditiva 1	0.98750	0.9862	0.98769	0.98945
	0.15458	0.16257	0.15365	0.14482
	0.03071	0.03121	0.02568	0.00443
	0.00203	0.00325	0.01419	0.00203
Aditiva 2	0.97704	0.97704	0.97704	0.97704
	0.21118	0.21118	0.21118	0.21118
	0.02811	0.02811	0.02811	0.02811
	0.00325	0.00325	0.00325	0.00325
	0.00106	0.00106	0.00106	0.00106

Nas Tabelas 2 e 3, apresentamos os resultados para o esquema de perturbação aditiva 1 e para o esquema de perturbação aditiva 2, respectivamente. Para este conjunto de

dados foi encontrado apenas um autovetor influente para as medidas $m(q)_t$, por isso os resultados das contribuições agregadas são iguais, ou seja, $m(1)_t = m(2)_t = m(3)_t$.

Tabela 2 - Medidas de observações influentes para esquema de perturbação aditiva 1

Sem o mineral	Medidas	Q_e	Média	Valores críticos	Valores para as observações			
					5	9	16	24
Endite	$m(1)_t$	1	0.1987	0.2811	0.8588	0.1120	0.3113	0.0522
	$m(2)_t$	1	0.1987	0.2811	0.8588	0.1120	0.3113	0.0522
	$m(3)_t$	1	0.1987	0.2811	0.8588	0.1120	0.3113	0.0522
	B_{E_t}	5	0.0470	0.0940	0.5597	0.0056	0.0095	0.0004
Daubite	$m(1)_t$	1	0.1986	0.2809	0.8550	0.1018	0.3074	0.0606
	$m(2)_t$	1	0.1986	0.2809	0.8550	0.1018	0.3074	0.0606
	$m(3)_t$	1	0.1986	0.2809	0.8550	0.1018	0.3074	0.0606
	B_{E_t}	5	0.0473	0.0947	0.5516	0.0054	0.0091	0.0005
Cornite	$m(1)_t$	1	0.1988	0.2811	0.2184	0.5323	0.1677	0.3064
	$m(2)_t$	1	0.1988	0.2811	0.2184	0.5323	0.1677	0.3064
	$m(3)_t$	1	0.1988	0.2811	0.2184	0.5323	0.1677	0.3064
	B_{E_t}	5	0.0472	0.0945	0.0045	0.1219	0.0014	0.0109
Blandite	$m(1)_t$	1	0.1989	0.2813	0.8576	0.0549	0.3193	0.0187
	$m(2)_t$	1	0.1989	0.2813	0.8576	0.0549	0.3193	0.0187
	$m(3)_t$	1	0.1989	0.2813	0.8576	0.0549	0.3193	0.0187
	B_{E_t}	5	0.0456	0.0913	0.5472	0.0012	0.0106	0.0001

Q_e representa a quantidade de autovetores influentes.

Tabela 3 - Medidas de observações influentes para esquema de perturbação aditiva 2

Sem o mineral	Medidas	Q_e	Média	Valores críticos	Valores para as observações			
					5	9	16	24
Endite	$m(1)_t$	2	0.2180	0.3083	0.8485	0.2521	0.3115	0.1495
	$m(2)_t$	1	0.1977	0.2796	0.8376	0.1067	0.3114	0.0611
	$m(3)_t$	1	0.1977	0.2796	0.8376	0.1067	0.3114	0.0611
	B_{E_t}	5	0.0488	0.0977	0.5193	0.0059	0.0095	0.0006
Daubite	$m(1)_t$	2	0.2180	0.3083	0.8485	0.2521	0.3115	0.1495
	$m(2)_t$	1	0.1977	0.2796	0.8376	0.1067	0.3114	0.0611
	$m(3)_t$	1	0.1977	0.2796	0.8376	0.1067	0.3114	0.0611
	B_{E_t}	5	0.0488	0.0977	0.5193	0.0059	0.0095	0.0006
Cornite	$m(1)_t$	2	0.2180	0.3083	0.8485	0.2521	0.3115	0.1495
	$m(2)_t$	1	0.1977	0.2796	0.8376	0.1067	0.3114	0.0611
	$m(3)_t$	1	0.1977	0.2796	0.8376	0.1067	0.3114	0.0611
	B_{E_t}	5	0.0488	0.0977	0.5193	0.0059	0.0095	0.0006
Blandite	$m(1)_t$	2	0.2180	0.3083	0.8485	0.2521	0.3115	0.1495
	$m(2)_t$	1	0.1977	0.2796	0.8376	0.1067	0.3114	0.0611
	$m(3)_t$	1	0.1977	0.2796	0.8376	0.1067	0.3114	0.0611
	B_{E_t}	5	0.0488	0.0977	0.5193	0.0059	0.0095	0.0006

Q_e representa a quantidade de autovetores influentes.

Para o esquema de perturbação aditiva 1, notamos que com a retirada dos minerais *endite*, *daubite* e *blandite*, as observações 5 e 16 são fortemente influentes, sendo a observação 5 destacada como influente pela medida B_{E_t}, ou seja, seu nível de influência está aci-

ma da média; com a retirada do mineral *cornite,* destacamos as observações 9 e 24 como fortemente influentes e a observação 9 é apontada como influente pela medida B_{E_t} (Tabela 2).

Para o esquema de perturbação aditiva 2, observamos que, com a retirada de qualquer mineral, as observações 5 e 16 são extremamente influentes, entretanto apenas a observação 5 tem nível de influência acima da média (Tabela 3).

5.2.2. Perturbando a variável *blandite* (componente 2)

Tabela 4 - Autovalores normalizados

Perturbação	Valores com a retirada desses minerais			
	Endite	***Daubite***	***Cornite***	***Blandite***
Aditiva 1	0.91784	0.91923	0.99266	0.92146
	0.39686	0.39351	0.12090	0.38827
	0.00872	0.01285	0.00298	0.01286
	0.00020	0.00020	0.00020	0.00156
Aditiva 2	0.94072	0.94072	0.94072	0.94072
	0.33897	0.33897	0.33897	0.33897
	0.01189	0.01189	0.01189	0.01189
	0.00110	0.00110	0.00110	0.00110
	0.00003	0.00003	0.00003	0.00003

Tabela 5 - Medidas de observações influentes para esquema de perturbação aditiva 1

Sem o mineral	Medidas	Q_e	Média	Valores críticos	Valores para as observações			
					5	**9**	**16**	**24**
Endite	$m(1)_t$	2	0.2293	0.3243	0.6595	0.6680	0.2618	0.3336
	$m(2)_t$	1	0.1916	0.2710	0.5354	0.6078	0.2050	0.2997
	$m(3)_t$	1	0.1916	0.2710	0.5354	0.6078	0.2050	0.2997
	B_{E_t}	5	0.0529	0.1059	0.1904	0.2000	0.0047	0.0124
Daubite	$m(1)_t$	2	0.2291	0.3241	0.6464	0.6762	0.2548	0.3308
	$m(2)_t$	1	0.1918	0.2712	0.5152	0.6224	0.1951	0.2976
	$m(3)_t$	1	0.1918	0.2712	0.5152	0.6224	0.1951	0.2976
	B_{E_t}	5	0.0530	0.1061	0.1764	0.2102	0.0042	0.0120
Cornite	$m(1)_t$	2	0.1993	0.2818	0.0404	0.7939	0.0190	0.3904
	$m(2)_t$	1	0.1993	0.2818	0.0404	0.7939	0.0190	0.3904
	$m(3)_t$	1	0.1993	0.2818	0.0404	0.7939	0.0190	0.3904
	B_{E_t}	5	0.0447	0.0893	0.0005	0.4030	0.0002	0.0233
Blandite	$m(1)_t$	2	0.2289	0.3237	0.6514	0.6665	0.2604	0.3286
	$m(2)_t$	1	0.1920	0.2715	0.5348	0.6084	0.2053	0.2936
	$m(3)_t$	1	0.1920	0.2715	0.5348	0.6084	0.2053	0.2936
	B_{E_t}	5	0.0530	0.1059	0.1822	0.1987	0.0046	0.0117

Q_e representa a quantidade de autovetores influentes.

Outro procedimento seguido foi o seguinte: a variável Y_{t2}, que representa a segunda componente, foi perturbada segundo os esquemas de perturbação anteriormente definidos; a variável a ser retirada pode ser qualquer uma das que restaram, da mesma forma como foi feito ao perturbar a primeira componente.

Na Tabela 4, apresentamos os autovalores normalizados não-nulos. Percebemos que, para o esquema de perturbação aditiva 1, o maior autovalor, entre os máximos autovalores, foi encontrado com a retirada da variável *cornite*, ou seja, a maior direção ocorre com a retirada deste mineral.

Através dos resultados para a perturbação aditiva 1, com a retirada de qualquer mineral, observamos que as observações 9 e 24 são extremamente influentes e a observação 9 está com nível de influência acima da média. Além dessas duas observações identificadas como influentes, verificamos que a observação 5 também é extremamente influente, com seu nível de influência muito grande, porém isso ocorre apenas quando retiramos os minerais *endite*, *daubite* e *albite* (Tabela 5).

Na Tabela 6, apresentamos os resultados para o esquema de perturbação aditiva 2. Verificamos que, com a retirada de qualquer mineral, as observações identificadas como extremamente influentes foram 5, 9 e 24, sendo que a observação 9 foi a única com nível de influência acima da média.

Tabela 6 - Medidas de observações influentes para esquema de perturbação aditiva 2

Sem o mineral	Medidas	Q_e	Média	Valores críticos	Valores para as observações			
					5	9	16	24
Endite	$m(1)_t$	2	0.2262	0.3200	0.5540	0.7245	0.2280	0.3568
	$m(2)_t$	1	0.1940	0.2743	0.3764	0.7080	0.1450	0.3452
	$m(3)_t$	1	0.1940	0.2743	0.3764	0.7080	0.1450	0.3452
	B_{E_t}	5	0.0517	0.1034	0.0958	0.2770	0.0027	0.0162
Daubite	$m(1)_t$	2	0.2262	0.3200	0.5540	0.7245	0.2280	0.3568
	$m(2)_t$	1	0.1940	0.2743	0.3764	0.7080	0.1450	0.3452
	$m(3)_t$	1	0.1940	0.2743	0.3764	0.7080	0.1450	0.3452
	B_{E_t}	5	0.0517	0.1034	0.0958	0.2770	0.0027	0.0162
Cornite	$m(1)_t$	2	0.2262	0.3200	0.5540	0.7245	0.2280	0.3568
	$m(2)_t$	1	0.1940	0.2743	0.3764	0.7080	0.1450	0.3452
	$m(3)_t$	1	0.1940	0.2743	0.3764	0.7080	0.1450	0.3452
	B_{E_t}	5	0.0517	0.1034	0.0958	0.2770	0.0027	0.0162
Blandite	$m(1)_t$	2	0.2262	0.3200	0.5540	0.7245	0.2280	0.3568
	$m(2)_t$	1	0.1940	0.2743	0.3764	0.7080	0.1450	0.3452
	$m(3)_t$	1	0.1940	0.2743	0.3764	0.7080	0.1450	0.3452
	B_{E_t}	5	0.0517	0.1034	0.0958	0.2770	0.0027	0.0162

Q_e representa a quantidade de autovetores influentes.

5.2.3. Perturbando a variável *cornite* (componente 3)

Aqui, a variável Y_{t3}, que representa a terceira componente, foi perturbada; essa é a variável fixa. Para o esquema de perturbação aditiva 1, percebemos que a maior direção ocorre com a retirada do mineral *blandite*. Na Tabela 7, apresentamos os autovalores normalizados não-nulos.

Tabela 7 - Autovalores normalizados

Perturbação	Valores com a retirada desses minerais			
	Endite	***Daubite***	***Cornite***	***Blandite***
Aditiva 1	0.99907	0.99911	0.99971	0.99905
	0.04319	0.04213	0.02418	0.04353
	0.00109	0.00164	0.00047	0.00183
	0.00004	0.00004	0.00004	0.00039
Aditiva 2	0.99942	0.99942	0.99942	0.99942
	0.03409	0.03409	0.03409	0.03409
	0.00108	0.00108	0.00108	0.00108
	0.00022	0.00022	0.00022	0.00022
	0.00001	0.00001	0.00001	0.00001

Nas Tabelas 8 e 9, apresentamos os resultados para os esquemas de perturbação aditiva 1 e aditiva 2, respectivamente. Verificamos que, para os dois esquemas, as observações 5 e 16 são extremamente influentes e que o nível de influência da observação 5 está acima da média, o que ocorre independentemente do mineral retirado.

Tabela 8 - Medidas de observações influentes para esquema de perturbação aditiva 1

Sem o mineral	Medidas	Q_e	Média	Valores críticos	Valores para as observações				
					5	**9**	**11**	**16**	**24**
Endite	$m(1)_t$	1	0.1999	0.2827	0.8768	0.0840	0.0312	0.3160	0.0382
	$m(2)_t$	1	0.1999	0.2827	0.8768	0.0840	0.0312	0.3160	0.0382
	$m(3)_t$	1	0.1999	0.2827	0.8768	0.0840	0.0312	0.3160	0.0382
	B_{E_t}	5	0.0417	0.0835	0.5912	0.0006	2,0E-05	0.0100	4,0E-05
Daubite	$m(1)_t$	1	0.1999	0.2827	0.8781	0.0816	0.0249	0.3172	0.0405
	$m(2)_t$	1	0.1999	0.2827	0.8781	0.0816	0.0249	0.3172	0.0405
	$m(3)_t$	1	0.1999	0.2827	0.8781	0.0816	0.0249	0.3172	0.0405
	B_{E_t}	5	0.0417	0.0834	0.5949	0.0005	1,0E-05	0.0102	5,0E-05
Cornite	$m(1)_t$	1	0.2000	0.2828	0.8829	0.0301	0.0137	0.3161	0.0146
	$m(2)_t$	1	0.2000	0.2828	0.8829	0.0301	0.0137	0.3161	0.0146
	$m(3)_t$	1	0.2000	0.2828	0.8829	0.0301	0.0137	0.3161	0.0146
	B_{E_t}	5	0.0410	0.0820	0.6091	2,0E-05	4,0E-06	0.0100	4,0E-06
Blandite	$m(1)_t$	1	0.1999	0.2827	0.8785	0.0750	0.0243	0.3170	0.0358
	$m(2)_t$	1	0.1999	0.2827	0.8785	0.0750	0.0243	0.3170	0.0358
	$m(3)_t$	1	0.1999	0.2827	0.8785	0.0750	0.0243	0.3170	0.0358
	B_{E_t}	5	0.0418	0.0836	0.5962	0.0005	1,0E-05	0.0101	4,0E-05

Q_e representa a quantidade de autovetores influentes.

5.2.4. Perturbando a variável *daubite* (componente 4)

Neste caso, a variável Y_{t4}, que representa a quarta componente, será perturbada; essa é a variável fixa.

Na Tabela 10, apresentamos os autovalores normalizados não-nulos. Observamos que as maiores direções ocorrem com a retirada dos minerais *endite*.

Com os resultados obtidos para o esquema de perturbação aditiva 1 (Tabela 11), percebemos que a observação 5 foi identificada como extremamente influente, com a retirada de qualquer mineral, porém seu nível de influência está acima da média apenas com a retirada dos minerais *endite*, *blandite* e *albite*. A observação 9 foi identificada como fortemente influente apenas com a retirada do mineral *cornite* e a observação 16 só não foi identificada como influente com a retirada deste mineral. Com a retirada do mineral *albite*, a observação 16 contribui apenas nas direções de maior influência.

Tabela 9 - Medidas de observações influentes para esquema de perturbação aditiva 2

Sem o mineral	Medidas	Q_e	Média	Valores críticos	Valores para as observações				
					5	9	11	16	24
Endite	$m(1)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	$m(2)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	$m(3)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	B_{E_t}	5	0.0414	0.0828	0.5993	0.0003	9,8E-06	0.0101	2,4E-05
Daubite	$m(1)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	$m(2)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	$m(3)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	B_{E_t}	5	0.0414	0.0828	0.5993	0.0003	9,8E-06	0.0101	2,4E-05
Cornite	$m(1)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	$m(2)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	$m(3)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	B_{E_t}	5	0.0414	0.0828	0.5993	0.0003	9,8E-06	0.0101	2,4E-05
Blandite	$m(1)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	$m(2)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	$m(3)_t$	1	0.1999	0.2828	0.8797	0.0679	0.0238	0.3166	0.0325
	B_{E_t}	5	0.0414	0.0828	0.5993	0.0003	9,8E-06	0.0101	2,4E-05

Q_e representa a quantidade de autovetores influentes.

Tabela 10 - Autovalores normalizados

Perturbação	Valores com a retirada desses minerais			
	Endite	***Daubite***	***Cornite***	***Blandite***
Aditiva 1	0.97664	0.95786	0.9735	0.97579
	0.20219	0.28663	0.22824	0.20943
	0.07279	0.01844	0.01412	0.06249
	0.00058	0.00045	0.00046	0.00795
Aditiva 2	0.97674	0.97674	0.97674	0.97674
	0.20426	0.20426	0.20426	0.20426
	0.06489	0.06489	0.06489	0.06489
	0.00723	0.00723	0.00723	0.00723

	0.00007	0.00007	0.00007	0.00007

Na Tabela 12, apresentamos os resultados para o esquema de perturbação aditiva 2. Observamos que apenas a observação 5 foi identificada como influente (extremamente influente), apresentando nível de influência muito elevado.

5.2.5. Perturbando a variável *endite* (componente 5)

Aqui, a variável Y_{t5}, que representa a quinta componente, foi perturbada segundo os esquemas de perturbação; essa é a variável fixa.

Tabela 11 - Medidas de observações influentes para esquema de perturbação aditiva 1

Sem o mineral	Medidas	Q_e	Média	Valores críticos	Valores para as observações		
					5	9	16
Endite	*m*(1)*t*	2	0.2171	0.3071	0.7787	0.2809	0.3088
	m(2)t	1	0.1976	0.2795	0.7432	0.2186	0.3084
	m(3)t	1	0.1976	0.2795	0.7432	0.2186	0.3084
	B_{E_t}	5	0.0501	0.1002	0.3686	0.0128	0.0095
Cornite	*m(1)t*	2	0.2231	0.3155	0.4113	0.4948	0.2770
	m(2)t	1	0.1957	0.2768	0.3413	0.3578	0.2436
	m(3)t	1	0.1957	0.2768	0.3413	0.3578	0.2436
	B_{E_t}	5	0.0505	0.1011	0.0288	0.0610	0.0059
Blandite	*m(1)t*	2	0.2192	0.3101	0.7951	0.1959	0.3225
	m(2)t	1	0.1973	0.2791	0.7667	0.0931	0.3224
	m(3)t	1	0.1973	0.2791	0.7667	0.0931	0.3224
	B_{E_t}	5	0.0487	0.0973	0.4001	0.0015	0.0108
Albite	*m(1)t*	2	0.2177	0.3079	0.7755	0.2634	0.3086
	m(2)t	1	0.1976	0.2794	0.7389	0.2031	0.3084
	m(3)t	1	0.1976	0.2794	0.7389	0.2031	0.3084
	B_{E_t}	5	0.0502	0.1005	0.3624	0.0098	0.0094

Q_e representa a quantidade de autovetores influentes.

Com os resultados da Tabela 13, podemos observar que a maior direção é dada pela retirada do mineral *albite*, independentemente do tipo de perturbação estudado. Através dos resultados para o esquema de perturbação aditiva 1 (Tabela 14), percebemos que as observações 5 e 16 são extremamente influentes, sendo que o nível de influência da observação 5 está acima da média, o que é detectado apenas com a retirada dos minerais *daubite*, *blandite* e *albite*. Com a retirada do mineral *cornite*, verificamos que a observação 9 é influente (fortemente influente), porém seu nível de influência não está acima da média.

Tabela 12 - Medidas de observações influentes para esquema de perturbação aditiva 2

Sem o mineral	Medidas	Q_e	Média	Valores críticos	Valores para as observações		
					5	9	16
Endite	$m(1)_t$	2	0.2173	0.3074	0.7064	0.2756	0.3051
	$m(2)_t$	1	0.1977	0.2795	0.6466	0.2433	0.3038
	$m(3)_t$	1	0.1977	0.2795	0.6466	0.2433	0.3038
	B_{E_t}	5	0.0501	0.1003	0.2493	0.0113	0.0090
Cornite	$m(1)_t$	2	0.2173	0.3074	0.7064	0.2756	0.3051
	$m(2)_t$	1	0.1977	0.2795	0.6466	0.2433	0.3038
	$m(3)_t$	1	0.1977	0.2795	0.6466	0.2433	0.3038
	B_{E_t}	5	0.0501	0.1003	0.2493	0.0113	0.0090
Blandite	$m(1)_t$	2	0.2173	0.3074	0.7064	0.2756	0.3051
	$m(2)_t$	1	0.1977	0.2795	0.6466	0.2433	0.3038
	$m(3)_t$	1	0.1977	0.2795	0.6466	0.2433	0.3038
	B_{E_t}	5	0.0501	0.1003	0.2493	0.0113	0.0090
Albite	$m(1)_t$	2	0.2173	0.3074	0.7064	0.2756	0.3051
	$m(2)_t$	1	0.1977	0.2795	0.6466	0.2433	0.3038
	$m(3)_t$	1	0.1977	0.2795	0.6466	0.2433	0.3038
	B_{E_t}	5	0.0501	0.1003	0.2493	0.0113	0.0090

Q_e representa a quantidade de autovetores influentes.

Na Tabela 15, apresentamos os resultados para o esquema de perturbação aditiva 2. Observamos que as observações 5 e 16 foram identificadas como extremamente influentes, a observação 5 tem nível de influência acima da média e a observação 16 contribui apenas nas direções de maior influência.

Tabela 13 - Autovalores normalizados

Perturbação	Valores com a retirada desses minerais			
	Daubite	***Cornite***	***Blandite***	***Albite***
Aditiva 1	0.96298	0.94814	0.96119	0.96367
	0.25093	0.31749	0.27556	0.25326
	0.09854	0.01522	0.01306	0.08479
	0.00046	0.00048	0.00034	0.00395
Aditiva 2	0.95473	0.95473	0.95473	0.95473
	0.28437	0.28437	0.28437	0.28437
	0.08721	0.08721	0.08721	0.08721
	0.00364	0.00364	0.00364	0.00364
	0.00007	0.00007	0.00007	0.00007

Tabela 14 - Medidas de observações influentes para esquema de perturbação aditiva 1

Sem o mineral	Medidas	Q_e	Média	Valores críticos	Valores para as observações		
					5	9	16
Daubite	$m(1)_t$	2	0.2204	0.3116	0.7903	0.2388	0.3035
	$m(2)_t$	1	0.1963	0.2776	0.7471	0.2191	0.3028
	$m(3)_t$	1	0.1963	0.2776	0.7471	0.2191	0.3028
	B_{E_t}	5	0.0525	0.1050	0.3903	0.0131	0.0087
Cornite	$m(1)_t$	2	0.2250	0.3182	0.2827	0.5214	0.2184
	$m(2)_t$	1	0.1947	0.2754	0.2626	0.3464	0.2054
	$m(3)_t$	1	0.1947	0.2754	0.2626	0.3464	0.2054
	B_{E_t}	5	0.0513	0.1025	0.0068	0.0743	0.0024
Blandite	$m(1)_t$	2	0.2224	0.3145	0.8146	0.1528	0.3230
	$m(2)_t$	1	0.1961	0.2773	0.7826	0.0866	0.3228
	$m(3)_t$	1	0.1961	0.2773	0.7826	0.0866	0.3228
	B_{E_t}	5	0.0500	0.1000	0.4404	0.0006	0.0109
Albite	$m(1)_t$	2	0.2206	0.3120	0.7930	0.2293	0.3077
	$m(2)_t$	1	0.1963	0.2777	0.7532	0.2089	0.3072
	$m(3)_t$	1	0.1963	0.2777	0.7532	0.2089	0.3072
	B_{E_t}	5	0.0522	0.1045	0.3961	0.0101	0.0091

Q_e representa a quantidade de autovetores influentes.

Tabela 15 - Medidas de observações influentes para esquema de perturbação aditiva 2

Sem o mineral	Medidas	Q_e	Média	Valores críticos	Valores para as observações		
					5	9	16
Daubite	$m(1)_t$	2	0.2226	0.3148	0.7232	0.2436	0.2980
	$m(2)_t$	1	0.1954	0.2764	0.6440	0.2406	0.2945
	$m(3)_t$	1	0.1954	0.2764	0.6440	0.2406	0.2945
	B_{E_t}	5	0.0532	0.1064	0.2754	0.0117	0.0081
Cornite	$m(1)_t$	2	0.2226	0.3148	0.7232	0.2436	0.2980
	$m(2)_t$	1	0.1954	0.2764	0.6440	0.2406	0.2945
	$m(3)_t$	1	0.1954	0.2764	0.6440	0.2406	0.2945
	B_{E_t}	5	0.0532	0.1064	0.2754	0.0117	0.0081
Blandite	$m(1)_t$	2	0.2226	0.3148	0.7232	0.2436	0.2980
	$m(2)_t$	1	0.1954	0.2764	0.6440	0.2406	0.2945
	$m(3)_t$	1	0.1954	0.2764	0.6440	0.2406	0.2945
	B_{E_t}	5	0.0532	0.1064	0.2754	0.0117	0.0081
Albite	$m(1)_t$	2	0.2226	0.3148	0.7232	0.2436	0.2980
	$m(2)_t$	1	0.1954	0.2764	0.6440	0.2406	0.2945
	$m(3)_t$	1	0.1954	0.2764	0.6440	0.2406	0.2945
	B_{E_t}	5	0.0532	0.1064	0.2754	0.0117	0.0081

Q_e representa a quantidade de autovetores influentes.

5.2.6. Resultados da comparação das estimativas

A Tabela 16 apresenta os resultados de comparação das estimativas dos parâmetros da distribuição Dirichlet, usando todas as observações com as estimativas após a retirada de algumas observações influentes. Como já foi visto, a observação 5 foi a mais influente quando retiramos os minerais *endite*, *daubite* e *blandite*, neste sentido notamos que essa observação foi a que causou maior impacto nas estimativas de todos os parâmetros. Quando retiramos o mineral *cornite*, as observações 9 e 24 causaram bastante impacto nas estimativas dos

parâmetros, sendo que a retirada da observação 9 causou impacto maior ainda. A retirada da observação 11 não causa uma mudança muito grande nas estimativas dos parâmetros.

6. Considerações finais

Neste artigo avaliamos a influência local através da curvatura normal conforme, metodologia proposta por Poon & Poon (1999), de observações correspondentes a variáveis independentes e identicamente distribuídas segundo a distribuição Dirichlet.

Em relação aos esquemas de perturbação, podemos dizer que o tipo de perturbação aditivo 2 é o mais recomendado, pois todas as p componentes da distribuição serão alteradas e não apenas $p-1$ componentes, como no esquema de perturbação aditiva 1. Para um conjunto de dados reais, não há dificuldades de escolher qual variável será a p-ésima componente, já que o resultado é invariante a essa escolha. O outro esquema considerado foi o de multiplicar o vetor de perturbação com a função de log-verossimilhança, porém para o esquema deste tipo de perturbação não foi identificada nenhuma observação influente em nenhum dos conjuntos de dados reais estudados, o que se deve ao fato de que as observações realmente têm características Dirichlet, ou seja, não encontramos nenhuma observação com padrão distinto do esperado para esta distribuição.

A variável $Y_j, j=1,\ldots,p$ foi escolhida para ser perturbada segundo os diversos esquemas de perturbação, ou seja, trabalhamos de forma que pudéssemos perturbar qualquer componente do modelo. Como $\sum_{j=1}^{p} Y_j = 1$, é necessário retirar uma das p variáveis, e trabalhar com apenas $p-1$ variáveis aleatórias. Optamos por fazer uma permutação entre essas variáveis e verificar qual delas tem a direção de maior influência, ou seja, a direção obtida através do maior autovalor normalizado entre os autovalores máximos.

Tabela 16 - Comparação (e variação percentual) das estimativas dos parâmetros com a retirada de algumas observações influentes

Parâmetro	Estimativas	Estimativas com a retirada das seguintes observações		
		Obs. 5	Obs. 9	Obs. 11
α_1	7.4833	8.2206 (9.85%)	9.8127 (31.13%)	8.0808 (7.98%)
α_2	3.3723	3.5468 (5.17%)	4.5422 (34.69%)	3.7042 (9.84%)
α_3	2.1194	2.5228 (19.03%)	2.4396 (15.11%)	2.1319 (0.59%)
α_4	1.9798	2.1981 (11.03%)	2.4599 (24.25%)	2.1136 (6.76%)
α_5	1.8285	2.0011 (9.44%)	2.2427 (22.65%)	1.8966 (3.72%)
Parâmetro	**Estimativas**	**Estimativas com a retirada das seguintes observações**		
		Obs. 16	**Obs. 24**	**-**
α_1	7.4833	7.7719 (3.86%)	8.2776 (10.61%)	-
α_2	3.3723	3.3799 (0.23%)	3.8460 (14.05%)	-
α_3	2.1194	2.3252 (9.71%)	2.1704 (2.41%)	-
α_4	1.9798	2.0721 (4.66%)	2.1242 (7.29%)	-
α_5	1.8285	1.8961 (3.70%)	1.9753 (8.03%)	-

Para diferentes esquemas de perturbação e/ou para perturbações em diferentes componentes, detectamos observações influentes de formas diferenciadas. E mais, através das medidas $m(q)_t$, $q = 1, 2, 3$, vimos que nem sempre determinada observação identificada como influente segundo $m(3)_t$ também será identificada pelas outras medidas, $m(1)_t$ ou $m(2)_t$, significando que essa observação contribui apenas na direção de maior influência.

Por fim, ao analisarmos as observações influentes através das medidas estudadas, retiramos essas observações uma a uma e calculamos novas estimativas dos parâmetros da distribuição Dirichlet, comparando-as com as estimativas na presença de todas as observações. Através dos esquemas estudados, verificamos que a retirada de observações influentes teve impacto nas estimativas, este impacto sendo maior após a retirada da observação identificada como a mais influente. Entretanto, não podemos afirmar que a retirada de tais observações melhora as estimativas dos parâmetros.

Referências bibliográficas

AITCHISON, J. (2003). *The Statistical Analysis of Compositional Data*. New Jersey: Blackburn Press.

BARBOSA, F. B. (1977). Estimadores de máxima verossimilhança dos parâmetros da distribuição de Dirichlet. *Revista Brasileira de Estatística*, 38, 203–219.

COOK, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society B*, 48, 133–169.

GELMAN, A., CARLIN, J. B., STERN, H. S., RUBIN D. B. (1995). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall/CRC.

KOBAYASHI, S. (1972). *Transformation Goups in Differential Geometry*. New York: Springer-Verlag.

LOYNES, R. M. (1986). Discussion on Assessment of local influence (by R. D. Cook). *Journal of the Royal Statistical Society B*, 48, 156–157.

MINKA, T. (2003). Estimating a Dirichlet distribution. www.stat.cmu.edu/˜minka/papers/dirichlet.

PAULA, G. A. (2004). *Modelos de Regressão com Apoio Computacional*. São Paulo: Instituto de Matemática e Estatística - USP.

POON, W. Y., POON, Y. S. (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society B*, 61, 51–61.

SCHARVISH, M. J. (1995). *Theory of Statistics*. New York: Springer.

Abstract

This paper considers the issue of assessing local influence of observations obtained from a Dirichlet distribution, which is a useful multivariate generalization of the beta distribution. To that end, we employ the measure proposed by Poon & Poon (1999). Different perturbation schemes are considered. The methodology is applied to both simulated and real data.

Keywords: Conformal normal curvature, Dirichlet distribution, perturbation schemes.

Agradecimentos

Os autores agradecem à CAPES e ao CNPq pelo apoio financeiro.

Modelagem de Séries Representativas do Setor Energético Brasileiro

Luciene Resende Gonçalves[*]
Thelma Sáfadi[*]

Resumo

O setor energético brasileiro é constituído por fontes renováveis e não renováveis. As energias não renováveis como os derivados do petróleo, óleo diesel e gasolina, contribuíram para um grande avanço tecnológico. No entanto, o uso desses fósseis vem provocando altos índices de poluição que levam a alterações climáticas. As energias provenientes da cana-de-açúcar e do álcool entram em cena como alternativa às já existentes. As técnicas de séries temporais foram utilizadas neste trabalho para modelar a produção de cana, álcool, óleo diesel e gasolina. Foram obtidos modelos univariados para as séries de produção de cana e álcool em que as melhores previsões foram obtidas por meio do modelo com intervenção. As séries de produção dos combustíveis: álcool, óleo diesel e gasolina foram utilizadas na modelagem multivariada sendo ajustado um modelo VAR (2). Verificou-se que as produções de óleo e gasolina estão interligadas e a produção de álcool é independente das duas primeiras.

Palavras-chave: fontes de energia, séries temporais, modelo SARIMA, intervenção, modelo VAR.

* Endereço para correspondência: Departamento de Ciências Exatas- Universidade Federal de Lavras -- e-mail:safadi@ufla.br

1. Introdução

A energia faz parte da vida do homem desde os tempos remotos, mas foi o progresso trazido pela Revolução Industrial e pela descoberta do petróleo que selou, de uma vez por todas, a dependência energética do mundo moderno (LEITE, 1997).

O atual setor energético brasileiro é constituído por fontes renováveis e não renováveis de energia. A fonte renovável biomassa é uma das mais exploradas atualmente. Dela provém a cana-de-açúcar e desta o álcool, conhecido como o combustível do futuro por ser, até agora, o mais adequado para substituir a gasolina. A gasolina e o óleo diesel são derivados da maior fonte energética de todos os tempos: o petróleo. Porém, os altos índices de poluição gerados com a queima de derivados fósseis vêm causando sérios problemas ao meio ambiente provocando catástrofes ambientais em diversos lugares do mundo.

Trabalhos na literatura relativos à produção dos combustíveis: álcool, óleo diesel e gasolina são escassos. Existem diversos estudos referentes ao consumo dos derivados de petróleo relacionados a variáveis macroeconômicas.

Burnquist e Bacchi (2002) apresentaram estimativas de elasticidades preço e renda da demanda por gasolina no Brasil, referentes ao período de 1973 a 1998. Os resultados obtidos indicam que, no curto prazo, a demanda por gasolina no País é inelástica a mudanças na renda real, dado que um aumento da ordem de 1% nessa variável resulta em incremento pouco expressivo, da ordem de 0,6% no consumo de gasolina. No longo prazo, conforme esperado, a elasticidade renda obtida apresentou-se relativamente mais elevada, embora pouco inferior à unidade, da ordem de 0,959. No que se refere à elasticidade preço da demanda, os resultados mostraram que o consumo de gasolina, no contexto da economia brasileira, é aparentemente pouco sensível a mudanças nos preços desse combustível, tanto no curto como no longo prazo.

Maciel (1991) analisou os aspectos relacionados à oferta e à demanda dos produtos envolvidos no Proálcool. Desenvolveu-se um modelo constituído por um conjunto de equações de comportamento e de identidades para cada produto: álcool, gasolina e açúcar. As equações foram estimadas pelo método de mínimos quadrados ordinários. A partir deste modelo, fizeram-se simulações econométricas. Após a validação do modelo, passou-se à etapa de simulações. Os resultados obtidos mostraram que, a persistirem as tendências dos últimos anos, verificar-se-iam um déficit cada vez mais acentuado no abastecimento de álcool, evidenciando a divergência existente na estrutura que vem determinando a oferta e a demanda do produto.

Assis e Lopes (1980) avaliaram o comportamento do consumo de gasolina e óleo diesel entre 1970 e 1977, obtendo estimativas de preço de curto e longo prazos, para o Brasil. Utilizaram a técnica *pooled regressions*, juntamente com observações de séries temporais e estimaram a equação de consumo de gasolina e óleo diesel por mínimos quadrados ordinários, com variáveis *dummies*. Os resultados obtidos indicaram tanto baixas elasticidades-preço da demanda de gasolina e óleo diesel como baixa sensibilidade a variações na renda para o consumo de gasolina. O consumo mostrou-se sensível à variação na renda para o óleo diesel, com exceção da Região Norte.

Moreira (1996) enfocou o consumo de óleo diesel no Brasil, empregando dois modelos de projeção de longo prazo do consumo, sendo um para o nível nacional e outro para análise regional. O autor utilizou um modelo para a realização de previsões condicionadas ao preço do combustível e ao nível de atividade (PIB) para a análise de abrangência nacional. A metodologia empregada foi a de análise de autorregressão vetorial com correção de erros (VEC). Essa escolha foi feita por se tratar de um procedimento de estimação que permite testar a existência e avaliar os parâmetros das relações de longo prazo. Foram testadas algumas especificações alternativas do modelo, para a análise regional, não tendo sido avaliada a possibilidade de existirem relações de equilíbrio de longo prazo.

Brown (1980) analisou a utilização de álcool como substituto para os derivados de petróleo. Ele partiu de hipóteses, como crescimento econômico, produção de etanol, produção nacional de petróleo e preço por barril de petróleo. Investigou as diversas formas de suprir a demanda por derivados do petróleo com o uso do álcool. Concluiu que o álcool pode servir para diminuir a importação de petróleo, desde que combinado com uma política de substituição adequada, sendo talvez necessário alterar a composição da frota de veículos.

O objetivo deste trabalho foi fazer um estudo dessas fontes de energia por meio das técnicas de séries temporais. A análise univariada foi considerada para as séries de produção de álcool e cana-de-açúcar com o objetivo de compreender o comportamento da produção e obter previsões. As previsões constituem um meio para uma consequente tomada de decisões e este tipo de informação é crucial no mercado da cana e do álcool já que a energia proveniente dessas fontes é vista como solução energética de um futuro não muito distante. Para esta análise também se verificou a presença de possíveis intervenções no período considerado. A análise multivariada foi considerada para as séries de produção dos combustíveis: álcool, gasolina e óleo diesel. Neste caso, foi ajustado o modelo vetorial auto-regressivo. Este tipo de modelagem permite a obtenção de relações de dependência entre um conjunto de séries. O objetivo aqui foi somente identificar se há alguma relação de dependência entre tais produções, pois é sabido que são muitos os condicionantes de cada

produção de combustível, por exemplo, a produção de óleo e gasolina é quase totalmente dependente da produção de petróleo – que depende de fatores geológicos relacionados à expansão das reservas e ao ritmo de exploração, condicionado por fatores específicos a cada jazida.

2. Referencial Teórico

2.1. Séries temporais

Uma série temporal é um conjunto de observações coletadas de forma sequencial, ao longo do tempo. A dependência entre as observações é o que caracteriza as aplicações das técnicas de séries temporais.

Um dos principais objetivos de análise desse tipo de dados é a construção de modelos com propósitos determinados, tais como a realização de previsão, que é a determinação de valores futuros a partir de valores presentes e passados da série, bem como a identificação de relação de dependência entre um conjunto de séries.

2.1.1. Modelos para séries temporais

Conforme Morettin e Toloi (2006), os modelos utilizados para descrever séries temporais são processos estocásticos, controlados por leis probabilísticas. Um processo estocástico é definido como sendo uma coleção de variáveis aleatórias sequenciadas no tempo e definidas em um conjunto de pontos *T*, que pode ser contínuo ou discreto. A variável aleatória no tempo t é denotada por Z_t, $t = 1, \ldots, N \in T$.

2.1.2. Estacionaridade

Uma das suposições mais frequentes que diz respeito a uma série temporal é a de que ela é estacionária, ou seja, ela se desenvolve no tempo aleatoriamente ao redor de uma média constante, refletindo alguma forma de equilíbrio estável. O conceito de estacionaridade é baseado nos momentos de uma série temporal, mais precisamente nos momentos de primeira e segunda ordem. Isto é dado pela estacionaridade fraca.

Morettin e Toloi (2006) definem um processo estocástico como sendo fracamente estacionário ou estacionário de segunda ordem se:

(i) $E\{Z(t)\} = \mu(t),\ \forall\, t \in T$;

(ii) $E\{Z^2(t)\} < \infty,\ \forall\, t \in T$;

(iii) $\gamma(t_1, t_2) = \text{cov}\{Z(t_1), Z(t_2)\}$ é uma função de $|t_1 - t_2|$.

Esta definição confirma que um processo estocástico fica bem descrito por meio das funções média, variância e autocovariância.

2.1.3. Função de autocovariância e autocorrelação

Conforme Morettin e Toloi (2006), a função de autocovariância é definida por $\gamma_r = E\{Z_t Z_{t+\tau}\}$, em que Z_t é um processo estacionário real discreto de média zero e τ é chamado de *lag* e representa a defasagem no tempo.

A função de autocorrelação (fac) de um processo estacionário é definida por $\rho_\tau = \frac{\gamma_\tau}{\gamma_0}$, em que ρ_τ é a autocorrelação no *lag* τ pertencente aos inteiros, γ_τ é a autocovariância no *lag* τ e γ_0 é a autocovariância em zero.

O estimador da função de autocovariância γ_τ é $c_\tau = \frac{1}{N}\sum_{t=1}^{N-\tau}(Z_t - \overline{Z})(Z_{t+\tau} - \overline{Z})$, em que c_τ é a autocovariância estimada no *lag* $\tau = 1, \ldots, N-1$, N é o número de observações da série e $\overline{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$ é a média amostral. A estimativa da função de autocorrelação ρ_τ é dada por $r_\tau = \frac{c_\tau}{c_0}$, sendo c_τ a função de autocovariância no *lag* τ e c_0 a variância.

2.1.4. Decomposição Clássica

Cada observação temporal Z_t pode ser escrita como uma soma de três componentes não-observáveis:

$$Z_t = T_t + S_t + a_t, \qquad (1)$$

em que T_t e S_t representam, respectivamente, a tendência e a sazonalidade, enquanto a_t é uma componente aleatória, de média zero e variância constante σ_a^2.

A tendência pode ser entendida como um aumento ou diminuição gradual das observações ao longo do tempo; a sazonalidade indica possíveis flutuações ocorridas em

períodos menores ou iguais a 12 meses e a componente aleatória são oscilações aleatórias irregulares. Quando as componentes tendência e sazonalidade são retiradas da série o que resta é a_t (erro aleatório). Para Morettin e Toloi (2006), existindo uma sequência $\{a_t, t \in T\}$ de variáveis *i.i.d.*, com média zero e variância constante, tais variáveis são chamadas choques aleatórios e a sequência é denominada ruído branco.

Ainda de acordo com Morettin e Toloi (2006), o principal interesse em considerar um modelo do tipo (1) é estimar a sazonalidade S_t e a tendência T_t, pois estas duas componentes estão intrinsecamente ligadas e, conforme Pierce (1979), a influência da tendência sobre a componente sazonal pode ser muito forte, por duas razões:

(i) se a tendência não for levada em conta, métodos de estimação de S_t podem ser bastante afetados;

(ii) a especificação de S_t depende da especificação de T_t.

Estas componentes fazem com que a série não atinja seu estágio estacionário, condição exigida na metodologia de ajuste dos modelos de Box e Jenkins usada neste artigo. Uma vez estimadas T_t e S_t, elas são subtraídas de Z_t, restando apenas uma estimativa da componente aleatória a_t.

2.1.5. Tendência

Supondo-se a ausência de sazonalidade, tem-se o modelo:

$$Z_t = T_t + a_t, \quad t = 1, 2, \ldots, N$$

em que a_t é um ruído branco, com variância σ_a^2.

Um procedimento bastante utilizado para eliminar a tendência consiste em tomar diferenças sucessivas da série original. A primeira diferença dada por $\Delta Z_t = Z_t - Z_{t-1}$ geralmente é suficiente para deixar a série estacionária.

A construção de gráfico na análise de séries temporais é uma ferramenta importante. Por meio dele é possível identificar características inerentes aos dados, como variabilidade, observações atípicas, sazonalidade e tendência, dentre outras. Mas, como os procedimentos visuais nem sempre são confiáveis, existem testes para confirmar a presença das componentes tendência e sazonalidade.

2.1.5.1. Teste do sinal para identificação de tendência

O teste consiste em dividir a série em dois grupos, nos quais são comparadas as observações $(Z_1, Z_{1+c}), (Z_2, Z_{2+c}), \ldots, (Z_N, Z_{N+c})$, em que $c = \frac{N}{2}$ para as situações em que N é um número par e $c = \frac{N+1}{2}$, quando N for um número ímpar. É atribuído sinal "+" sempre que $Z_i < Z_{i+c}$ e sinal "−" para $Z_i > Z_{i+c}$, os empates são eliminados e n é o número de pares em que $Z_i \neq Z_{i+c}$, ou seja, é a soma de sinais "+" com os sinais "−". É testada a hipótese bilateral:

H_0: $P(Z_i < Z_{i+c}) = P(Z_i > Z_{i+c})$, $\forall i$: não existe tendência;

H_1: $P(Z_i < Z_{i+c}) \neq P(Z_i > Z_{i+c})$, $\forall i$: existe tendência.

Definindo T_2 como o número de sinais positivos, rejeita-se H_0, ou seja, a série apresenta tendência se $T_2 \geq n \cdot t$, em que t é obtido por meio de uma distribuição binomial com parâmetros $\left(n, p = \frac{1}{2}\right)$ e α um dado nível de significância. Para n > 20, utiliza-se a aproximação normal.

2.1.6. Sazonalidade

A sazonalidade (ou periodicidade) constitui outra forma de não-estacionaridade e deve ser estimada e retirada da série. Sendo $\hat{S}_t$ a estimativa de S_t, a série sazonalmente ajustada é $Z_t^{SA} = Z_t - \hat{S}_t$.

Priestley (1989) propõe o teste de Fisher para testar a presença de sazonalidade determinística (que pode ser prevista a partir de meses anteriores), baseado na análise de uma quantidade chamada de periodograma, a qual é dependente das funções seno e co-seno.

O periodograma é uma descrição dos valores observados numa realização de uma série através da sobreposição de ondas sinusoidais com várias frequências. A aplicação prática mais óbvia desta decomposição é a de servir de instrumento à identificação de componentes cíclicas ou periódicas.

Conforme Priestley (1989), a função periódica é dada por:

$$I_p(f_i) = \frac{2}{n}\left[\left(\sum_{t=1}^{n} a_t \cos\frac{2\pi i}{n}t\right)\left(\sum_{t=1}^{n} a_t \operatorname{sen}\frac{2\pi i}{n}t\right)^2\right] \quad (2)$$

em que $0 < f_i < \frac{1}{2}$ e $t = 1, 2, \ldots, n$. $I_p(f_i)$ é a intensidade da frequência f_i. A periodicidade de período $\frac{1}{f_i}$ pode ser observada pela existência de picos na frequência $f_i = \frac{1}{n}$.

2.1.6.1. Teste de Fisher

O teste de Fisher foi proposto, inicialmente, para testar o maior período. As hipóteses a serem testadas são:

H_0: não existe sazonalidade;

H_1: existe sazonalidade.

A estatística do teste é dada por $g = \frac{\max I_p}{\sum_{p=1}^{N/2} I_p}$, em que I_p é o valor do periodograma (dado pela equação 2) no período p e N é o número de observações da série. A estatística do teste de Fisher, Z_α, é $Z_\alpha = 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{n-1}}$ sendo $n = \frac{N}{2}$ e α é o nível de significância do teste. Se $g > z$, rejeita-se H_0, ou seja, a série apresenta periodicidade p.

2.1.6.2. Estimação da sazonalidade

A sazonalidade encontrada na série pode ser determinística ou estocástica. A sazonalidade determinística pode ser prevista a partir de meses anteriores. A sua eliminação da série pode ser feita pela utilização de diferenças, cuja ordem corresponde ao valor do período *p* encontrado pelo periodograma.

A outra forma de eliminação é pela estimação dada pelo método de regressão. Este método consiste em considerar o modelo $Z_t = T_t + S_t + a_t$, em que

$$T_t = \sum_{j=0}^{m} \beta_j t^j \quad \text{e} \quad S_t = \sum_{j=1}^{12} \alpha_j d_{jt},$$

d_{jt} são variáveis periódicas (senos, cossenos ou variáveis sazonais *dummies*) e a_t é ruído branco, com média zero e variância σ_a^2.

2.1.7. Modelos de Box e Jenkins

Dentre os diversos métodos e modelos de previsão existentes, destacam-se os modelos de Box e Jenkins, cuja metodologia consiste em ajustar modelos auto-regressivos integrados de médias móveis, ARIMA (p, d, q), a um conjunto de dados. Tais modelos se caracterizam, ainda, por serem simples e parcimoniosos; as previsões são bastante precisas, comparando-se favoravelmente com os demais métodos de previsão. Classificam-se em modelos lineares estacionários e não estacionários.

Segundo Morettin e Toloi (2006), a estratégia para a construção desses modelos é baseada em um ciclo iterativo. Os estágios deste ciclo são: especificação, identificação, estimação e verificação. Esses autores ainda salientam que a fase mais crítica desta iteração é a identificação, pois, se o modelo não for adequado, o ciclo é repetido voltando-se a este estágio. Um procedimento prático é identificar vários modelos e, dentre os que se ajustam, escolhe-se o que fornece o menor erro quadrático médio de previsão (EQMP), se este for o objetivo do ajuste.

A notação de operadores é bastante utilizada nestes modelos. Estes operadores são:

(i) B - operador translação para o passado: este operador ocasiona uma defasagem de um período m para trás, cada vez que é utilizado. É definido por $B^m Z_t = Z_{t-m}$;

(ii) Δ - operador diferença, definido por $\Delta Z_t = Z_t - Z_{t-1} = (1-B)Z_t \Rightarrow \Delta = 1-B$. Tem-se que $\Delta^n Z_t = (1-B)^n Z_t \Rightarrow \Delta_n = (1-B)^n$.

2.1.7.1. Modelos lineares estacionários

A metodologia de Box e Jenkins supõe que a série temporal Z_t é o resultado da passagem de um processo aleatório (ruído branco) a_t por um filtro ou sistema linear $\Psi(B)$, ou seja, $Z_t = \mu + a_t + \psi_1 a_{t-1} + \ldots = \mu + \psi(B)a_t$, em que $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \ldots$ é a função de transferência do filtro e μ é um parâmetro que determina o nível da série.

Chamando $\tilde{Z}_t = Z_t - \mu$, tem-se $\tilde{Z}_t = \psi(B)a_t$ e de forma alternativa $\tilde{Z}_t$ pode ser escrito como uma soma ponderada de valores passados mais um ruído como

$$\tilde{Z}_t = \pi_1 \tilde{Z}_{t-1} + \pi_2 \tilde{Z}_{t-2} + ... + a_t = \pi(B)\tilde{Z}_t + a_t,$$

sendo $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \ldots$.

2.1.7.1.1. Processos auto-regressivos e de médias móveis de ordens p e q – ARIMA (p,q)

Os modelos auto-regressivos e de médias móveis são uma combinação linear dos modelos auto-regressivos com médias móveis, podendo, então, serem escritos da forma

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \phi_2 \tilde{Z}_{t-2} + ... + \phi_p \tilde{Z}_{t-p} + a_t - \theta_1 a_{t-1} - ... - \theta_q a_{t-q},$$

ou ainda de forma compacta $\phi(B)\tilde{Z}_t = \theta(B)a_t$, sendo $\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$, o polinômio auto-regressivo de ordem p e $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$, o polinômio de médias móveis de ordem q.

2.1.7.2. Modelos não-estacionários

Grande parte das séries encontradas na prática apresenta alguma forma de não estacionaridade e, como a maioria dos procedimentos utilizados em séries temporais é para séries estacionárias, é necessário tomar um número d de diferenças finitas para atingir este estágio. Quando isso é possível, tem-se um caso de séries não-estacionárias homogêneas ou séries portadoras de raízes unitárias.

Os modelos usados para séries com este comportamento são os modelos ARIMA e SARIMA.

2.1.7.2.1. Modelo auto-regressivo integrado de médias móveis - ARIMA (p,d,q)

Este modelo é o caso geral dos modelos de Box e Jenkins, sendo dado por $\phi(B)\Delta^d Z_t = \theta(B)a_t$, em que Δ é o operador diferença e *d* o número de diferenças para tornar a série estacionária.

2.1.7.2.2. Modelo sazonal auto-regressivo integrado de médias móveis – SARIMA (p,d,q) × (P,D,Q)

Box e Jenkins (1976) generalizaram o modelo ARIMA para lidar com sazonalidade e definiram um modelo ARIMA sazonal multiplicativo, denominado SARIMA, representado por $\phi(B)\Phi(B^s)\Delta^d \Delta_s{}^D Z_t = \theta(B)\Theta(B^s)a_t$, em que:

- $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ é o polinômio auto-regressivo de ordem p;
- $\Phi(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_p B^{P_s}$ é o polinômio auto-regressivo sazonal de ordem P;

- $\Delta^d = (1-B)^d$ é o operador diferença e d é o número de diferenças necessárias para retirar a tendência da série;
- $\Delta_s{}^D = (1-B^s)^D$ é o operador diferença generalizado, quando duas observações estão distantes entre si de s intervalos de tempos que apresentam alguma semelhança e D é o número de diferenças de *lags* s necessárias para retirar a sazonalidade da série;
- $\theta(B) = 1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q$ é o polinômio de médias móveis de ordem q;
- $\Theta(B) = 1-\Theta_s B^s-\ldots-\Theta_p B^{Q_s}$ é o polinômio de médias móveis sazonal de ordem Q.

2.1.7.3. Estágios do ciclo iterativo

Os estágios do ciclo iterativo do método de Box e Jenkins são a identificação, a estimação e a verificação, dado que a classe geral de modelos ARIMA foi especificada.

– **Identificação**

O objetivo da identificação é determinar os valores de p,d,q, do modelo ARIMA (p,d,q) e (P,D,Q), quando a série apresentar sazonalidade estocástica. A escolha das ordens é feita com base nas autocorrelações e autocorrelações parciais (mede o grau de associação entre as observações Z_t e Z_{t-k}) estimadas, que se espera representar adequadamente as respectivas quantidades teóricas, as quais são desconhecidas.

.

– **Estimação**

Ao se identificar um modelo provisório para a série temporal, o próximo passo é estimar seus parâmetros e o método usado é o método da máxima verossimilhança.

– **Verificação ou diagnóstico**

Após identificar a ordem e estimarem-se eficientemente os parâmetros de um modelo, é necessário verificar se ele representa os dados de maneira adequada. Esta verificação é feita pelo teste de Box-Pierce.

Teste de Box-Pierce

Se o modelo for apropriado, a estatística:

$$Q(k) = n(n+2)\sum_{j=1}^{K}\frac{\hat{r}_j^2}{(n-j)} \sim \chi_{K-p-q},$$

em que $k = \{1, 2, \ldots, K\}$ são as primeiras k autocorrelações de $\hat{a}_t$, $\hat{r}_j$ é a autocorrelação estimada em j e $n = N - d$, sendo N o número total de observações e d o número de diferenças necessárias para obter estacionaridade, p é a ordem auto-regressiva do modelo e q é a ordem de médias móveis. A hipótese de ruído branco é rejeitada para $Q(K) > \chi^2_{K-p-q}$.

2.1.7.4. Critérios para escolha do modelo

Vários modelos podem ser identificados para descrever uma série, mas existem critérios para escolha do melhor modelo de acordo com o objetivo do ajuste. Dentre diversos critérios, tem-se o critério de informação de Akaike (AIC) e o critério do erro quadrático médio de previsão (EQMP).

2.1.7.4.1. Critério de informação de Akaike (AIC)

Akaike (1973) sugere escolher o modelo cujas ordens p e q minimizem o critério. O critério de Akaike, na comparação de diversos modelos, com N fixo, pode ser expresso por $AIC = N \log \hat{\sigma}_a^2 + 2(k + l + 2)$, em que $\hat{\sigma}_a^2$ é o estimador de máxima verossimilhança de $\sigma_a{}^2$, $0 \le k \le p$ e $0 \le l \le q$. O melhor modelo é aquele que apresenta menor AIC.

2.1.7.4.2. Critério do erro quadrático médio de previsão (EQMP)

Sendo o objetivo do ajuste a realização de previsão, o melhor modelo será o que apresentar o menor erro quadrático médio de previsão (EQMP). Mesmo que não tenha o menor valor de AIC. As estimativas EQMP são dadas pela média dos quadrados das diferenças entre valores observados e valores preditos. Logo, o EQMP com origem em t é dado por: $EQMP_t = \frac{1}{N} \sum_{h=1}^{n} [Z_{t+h} - \hat{Z}_t(h)]^2$ em que Z_{t+h} é o valor real, h é o número de previsões e $\hat{Z}_t(h)$ é o valor predito.

2.1.7.4.3. EPM

O erro percentual médio (EPM), dado por $EPM = \frac{1}{h}\sum_{1}^{h}\left|\frac{e_t(h)}{Z_{t+h}}\right| \times 100$ em que $e_t(h)$ é o erro de previsão, é utilizado para verificar o bom desempenho de ajuste de modelos.

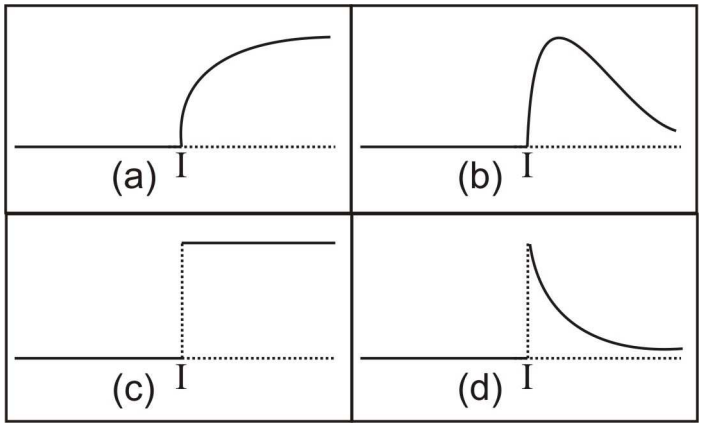
2.1.8. Análise de intervenção

O comportamento de uma série temporal pode vir a ser afetado pela presença de algum evento que pode alterar a real trajetória da série em um dado instante de tempo previamente conhecido. A esse evento dá-se o nome de intervenção e sua incidência pode ser momentânea ou perdurar por algum intervalo de tempo. O objetivo da análise de intervenção é avaliar o impacto de tais eventos no comportamento da série (PRIESTLEY, 1989).

2.1.8.1. Efeitos da intervenção

Uma intervenção pode afetar uma série temporal de várias maneiras. Na sua manifestação, ela pode ser abrupta ou gradual; na sua duração, pode ser permanente ou temporária. Os tipos mais comuns de efeitos de uma intervenção estão resumidos na Figura 1.

FIGURA 1- Efeito gradual permanente (a), gradual temporário (b), abrupto permanente (c) e abrupto temporário (d), sobre uma série temporal Z_t

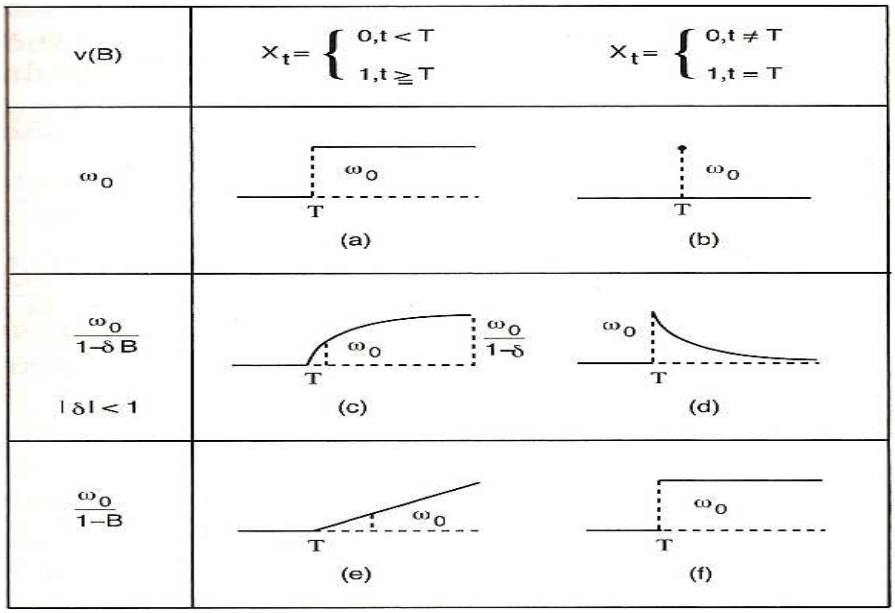


Uma classe geral de modelos, que leva em conta a ocorrência de múltiplas intervenções, é dado por $Z_t = \sum_{j=1}^{k} v_j(B) X_{j,t} + N_t$, em que:

- $X_{j,t}$, $j = 1, 2, \ldots, k$, são variáveis de intervenção do tipo $X_{j,t} = S_t^{(T)} = \begin{cases} 0, t < T \\ 1, t \geq T \end{cases}$ e $X_{j,t} = I_t^{(T)} = \begin{cases} 0, t \neq T \\ 1, t = T \end{cases}$;
- $v_j(B)$, $j = 1, \ldots, k$, são funções racionais da forma $\frac{\omega_j(B)B^{b_j}}{\delta_j(B)}$, em que $\omega_j(B) = \omega_{j,0} - \omega_{j,1}B - \ldots - \omega_{j,S}B^S$ e $\delta(B) = 1 - \delta_{j,1}B - \ldots - \delta_{j,r}B^r$ são polinômios em B, b_j é a defasagem no tempo para o início do efeito da j-ésima intervenção;
- N_t é a série temporal livre do efeito das intervenções, denominada série residual.

Para cada efeito de intervenção, tem-se uma forma apropriada para a função de transferência $v_j(B)$. A Figura 2 exibe algumas formas que $v(B)$ pode assumir, considerando, por simplicidade, o caso de uma única função de transferência $Z_t = v(B)X_t + N_t$, em que $v(B) = \frac{\omega(B)}{\delta(B)}$.

FIGURA 2 - Estrutura da função de transferência $v_j(B)$ de uma série temporal



2.2. Modelos lineares multivariados

O modelo linear multivariado é o modelo estabelecido para obter relações dinâmicas entre as séries de uma série vetorial Z_t constituída por n componentes, com denotação Z_{it} em que $i = 1, \cdots n$ é a i-ésima componente desse vetor e t é o tempo.

O vetor de médias de $\mathbf{Z}_t$ é denotado por $\boldsymbol{\mu}_t = E(\mathbf{Z}_t) = (\mu_{1t}, \mu_{2t}, \cdots, \mu_{nt})'$ e a matriz de covariâncias de $\mathbf{Z}_t$ é definida por $\boldsymbol{\Gamma}(t+\tau, t) = \mathrm{E}\{(\mathbf{Z}_{t+\tau} - \boldsymbol{\mu}_{t+\tau})(\mathbf{Z}_t - \boldsymbol{\mu}_t)'\}$, sendo τ o número de intervalos de tempo defasados.

A série n-variada $\mathbf{Z}_t$ é estacionária se a média $\boldsymbol{\mu}_t$ e a matriz de covariâncias $\boldsymbol{\Gamma}(t+\tau, t)$, t e τ inteiros, não dependem do tempo t.

2.2.1. Modelos auto-regressivos vetoriais – VAR (p)

O processo $\mathbf{Z}_t$ de ordem $n \times 1$, segue um modelo VAR de ordem p, VAR (p), se

$$\boldsymbol{Z}_t = \boldsymbol{\Phi}_0 + \boldsymbol{\Phi}_1 \boldsymbol{Z}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{Z}_{t-p} + \boldsymbol{a}_t, \tag{3}$$

em que $\boldsymbol{a}_t$ é ruído branco com vetor de médias $\mathbf{0}$ e matriz de covariâncias $\boldsymbol{\Sigma}$, $\boldsymbol{\Phi}_0 = (\phi_{10}, \ldots, \phi_{n0})'$ é um vetor $n \times 1$ de constantes e $\boldsymbol{\Phi}_k$ são matrizes de ordens $n \times n$ constantes, com elementos $\phi_{ij}^{\ k}, i, j = 1, \ldots, n$ e $k = 1, \ldots, p$. Sendo $\boldsymbol{I}_n$ a matriz identidade de ordem n, a equação (3) pode ser reescrita na forma $\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{\Phi}_0 + \boldsymbol{a}_t$, em que $\boldsymbol{\Phi}(B) = \boldsymbol{I}_n - \boldsymbol{\Phi}_1 B - \ldots - \boldsymbol{\Phi}_p B^p$ é o operador auto-regressivo vetorial de ordem p ou, ainda, um polinômio matricial $n \times n$ em B. O elemento genérico de $\boldsymbol{\Phi}(B)$ é $[\delta_{ij} - \phi_{ij}^{\ (1)} B - \ldots - \phi_{ij}^{\ (p)} B^p]$, para $i, j = 1, \ldots, n$ e $\delta_{ij} = 1$, se $i = j$ e igual a zero, caso contrário.

2.2.2. Construção de modelos VAR

A construção de modelos VAR segue o mesmo ciclo de identificação, estimação e diagnóstico usado para modelos univariados da classe ARIMA.

De acordo com Morettin (2006), uma maneira de identificar a ordem p de um modelo VAR (p) consiste em ajustar sequencialmente modelos auto-regressivos vetoriais de ordens

$1,2,\ldots k$ e testar a significância dos coeficientes (matrizes). Outra maneira de identificar a ordem de um VAR é usar algum critério de informação, como o de Akaike (AIC) ou Schwarz (BIC), dados por:

$$AIC(k) = \ln\left(\left|\hat{\Sigma}_k\right|\right) + 2kn^2 / T$$

e $$BIC(k) = \ln\left(\left|\hat{\Sigma}_k\right|\right) + kn^2 \ln(T) / T \ ,$$

em que $\hat{\mathbf{\Sigma}}_{\mathbf{k}}$ é a matriz de covariância dos resíduos, estimador de $\mathbf{\Sigma}$, e T é o tamanho da série.

Identificado o valor de p e supondo-se $\mathbf{a}_t \sim N(0, \mathbf{\Sigma})$, é possível estimar os coeficientes por máxima verossimilhança. Neste caso, os estimadores de mínimos quadrados são equivalentes aos estimadores de máxima verossimilhança condicionais que são obtidos por métodos de maximização numérica.

Os resíduos do modelo estimado são utilizados para construir a versão multivariada da estatística de Box-Ljung-Pierce para testar se o modelo é adequado. Esta estatística é dada por:

$$Q(m) = T^2 \sum_{\tau=1}^{m} \frac{1}{T-\tau} tr(\mathbf{\Gamma}(\tau)' \mathbf{\Gamma}(0)^{-1} \mathbf{\Gamma}(\tau) \mathbf{\Gamma}(0)^{-1}) \ ,$$

que sob a hipótese nula, de que a série $\mathbf{a}_t$ é ruído branco, tem distribuição $\chi^2(n^2(m-p))$. Para que o número de graus de liberdade seja positivo, m deve ser maior do que p.

3. Material e Métodos

3.1. Material

- Série de produção de cana-de-açúcar, em milhões de toneladas, coletada no período anual de 1947 a 1998, num total de 52 observações. As observações referentes aos anos de 1999 até 2004 foram reservadas para efeito de comparação com as previsões. Esses dados foram obtidos junto ao banco de dados do Instituto de Pesquisas Econômicas Aplicadas – IPEA (IPEA, 2005);
- Série de produção mensal de álcool de cana-de-açúcar - índice (média 2002 = 100), coletada no período de janeiro de 1991 a dezembro de 2005, num total de 180 observações. As observações do período de janeiro de 2006 a setembro do mesmo ano foram reservadas para serem comparadas com as previsões. Esses dados também foram obtidos junto ao banco de dados do IPEA; e

• Séries de produção de óleo diesel e gasolina, em metros cúbicos (m^3), com 81 observações cada uma, coletadas mensalmente no período de janeiro de 2000 a setembro de 2006. Essas séries foram obtidas pelo banco de dados da Agência Nacional do Petróleo - ANP (ANP, 2006).

Na análise multivariada, a série de produção de álcool foi utilizada no período de janeiro de 2000 a setembro de 2006.

3.2. Metodologia

A análise foi dividida em duas etapas, na primeira foi considerada a análise de cada série separadamente e posteriormente considerou-se a análise conjunta das séries de óleo diesel, gasolina e álcool.

No ajuste univariado, primeiramente, foi construído o gráfico das séries para verificar visualmente indícios de componentes de tendência, sazonalidade, variações atípicas e dados discrepantes. Em seguida, foram observadas as funções de autocorrelação e autocorrelação parcial dos dados originais que também auxiliam na identificação de componentes não estacionárias. Foram aplicados os testes de Cox-Stuart (MORETTIN; TOLOI, 2006) para tendência e teste de Fisher (PRIESTLEY, 1989) para sazonalidade. Confirmadas as componentes elas foram eliminadas por diferenciação e então se utilizou as funções de autocorrelação e autocorrelação parcial das séries diferenciadas para obter as possíveis ordens dos modelos a serem ajustados. Essas ordens levaram ao ajuste dos modelos cujos parâmetros foram estimados pelo método da máxima verossimilhança. O teste de Box-Pierce foi utilizado para diagnosticar se o resíduo do modelo ajustado constitui um ruído branco. Entre os modelos ajustados foram considerados modelos com intervenções. A escolha do melhor modelo foi feita considerando-se os critérios de Akaike (AIC) e erro quadrático médio de previsão (EQMP).

Para o ajuste multivariado, o primeiro passo foi colocar as séries em estágio estacionário. Para isso foram aplicados os testes de Fisher para testar sazonalidade e teste de Cox-Stuart para testar tendência. Uma vez confirmada a presença dessas componentes, elas foram estimadas pelo método de regressão e em seguida subtraídas das séries originais, obtendo assim séries estacionárias. Finalmente, foram ajustados modelos auto-regressivos multivariados de ordem *p* para $p = 1,\ldots,8$ e escolhido o modelo que apresentou menor AIC e BIC.

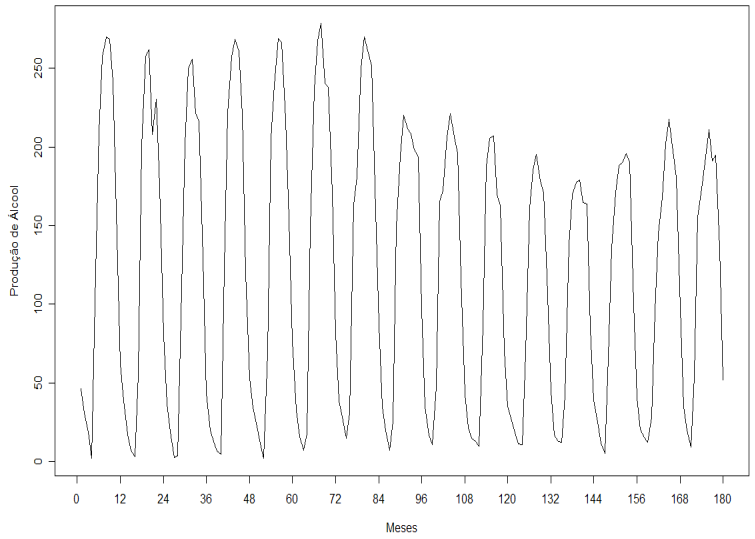
4. Resultados e Discussão

4.1. Modelagem Univariada

4.1.1. Série de produção de álcool

Na Figura 3, é apresentado o gráfico da série de produção nacional de álcool. Visualmente, é possível intuir a presença de sazonalidade e ausência de tendência.

FIGURA 3 - Série de produção brasileira de álcool - índice (média 2002=100), coletada no período mensal de janeiro de 1991 a dezembro de 2005

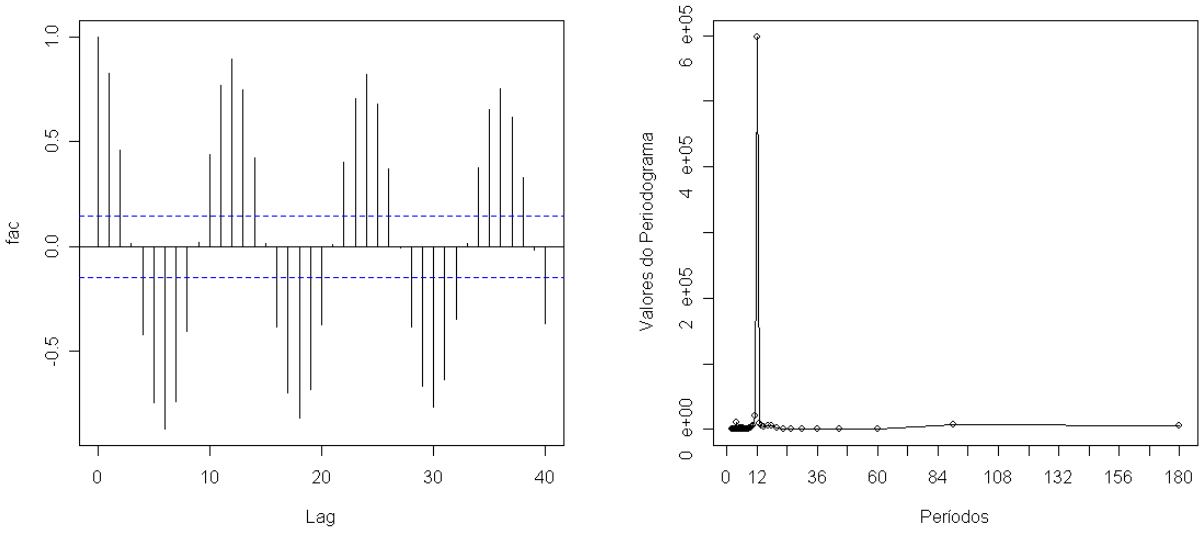


Fonte: IPEA.

Na Figura 4, estão representadas a função de autocorrelação da série original de produção mensal de álcool e o periodograma. A estrutura da função de autocorrelação é mais um indício de que a série apresenta componente sazonal, e o periodograma apresenta um pico no período de 12 meses. Aplicando-se o teste de Fisher para testar a existência do efeito sazonal nesse período, obteve-se g = 0,90 e $Z_{0,05}$ = 0,08. Como $g > z$, a série apresenta sazonalidade de 12 meses, a 5% de significância.

Segundo Morettin e Toloi (2006), havendo outra componente na série ela deve ser eliminada antes de se aplicar o teste para tendência. Para a série sazonalmente ajustada (sem sazonalidade) foi aplicado o teste de Cox-Stuart, a 5% de significância. O resultado do teste foi: $c = \frac{168}{2} = 84$ e $n = 84$ ($Z_t \neq Z_{t+c}\ \forall t$). O número de sinais positivos T_2 é igual a 30 e $t = 34,47$. Como $T_2 < 84 - 34,47$, a série não apresenta tendência.

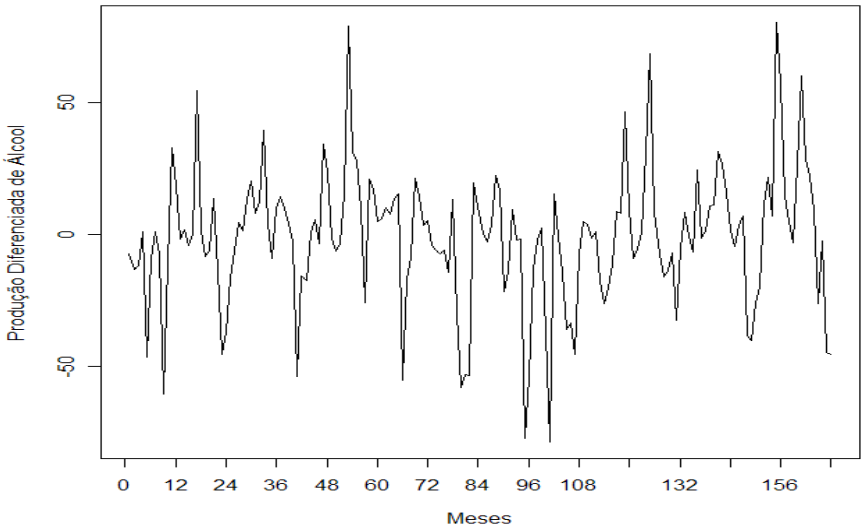
FIGURA 4 - Função de autocorrelação e periodograma da série de produção brasileira de álcool, coletada no período mensal de janeiro de 1991 a dezembro de 2005



A Figura 5 é a representação da série sazonalmente ajustada. Embora não seja fácil observar nessa figura, possíveis intervenções, nos períodos correspondentes a maio de 1996 (referente à observação 65), a junho de 1997 (observação 78) e a dezembro de 2000 (observação 120), serão consideradas na análise. O efeito abrupto no aumento da produção refletido em maio de 1996 está associado ao período de estabilização da economia, em que foi estabelecido um novo ciclo de desenvolvimento elevando os índices de expansão da

economia e do consumo de energia (BRASIL, 2007). Outro fato associado foi a criação, nesta época, da chamada "frota verde" que foi o estímulo e a determinação do uso do álcool hidratado em determinadas classes de veículos leves, como os carros oficiais e táxis. Houve também o aumento de 22% para 24% do percentual de adição de álcool etílico anidro à gasolina (Evolução do Programa Nacional do álcool – Proálcool, 2008). Os efeitos refletidos em junho de 1997 e dezembro de 2000, quando ocorreram quedas na produção, são explicados, em parte, por sucessivas crises externas, principalmente a crise cambial nos países asiáticos, que acabaram contaminando a economia nacional, obrigando o governo a tomar medidas que levaram a uma forte retração no crescimento econômico, tendo o PIB apresentado um crescimento de apenas 0,9% no ano de 1998 e de 1,6% em 1999. O baixo desempenho da economia teve reflexos no consumo de energia de 1999, notadamente quanto às energias associadas ao uso individual, como o álcool hidratado com queda de 8,6% no consumo, a gasolina automotiva com queda de 6,3%, o querosene de aviação com queda de 6,3% e energia elétrica residencial, com apenas 2,4% de crescimento (BRASIL, 2007).

FIGURA 5 - Série de produção brasileira de álcool, em meses, sazonalmente ajustada, no período de janeiro de 1991 a dezembro de 2005



As funções de autocorrelação e autocorrelação parcial da série sazonalmente ajustada de produção mensal de álcool que possuem a característica de identificar as ordens do modelo são apresentadas na Figura 6.

Pela observação dos gráficos da Figura 6, com base nos *lags* significativos 1 e 12, os modelos sugeridos foram:

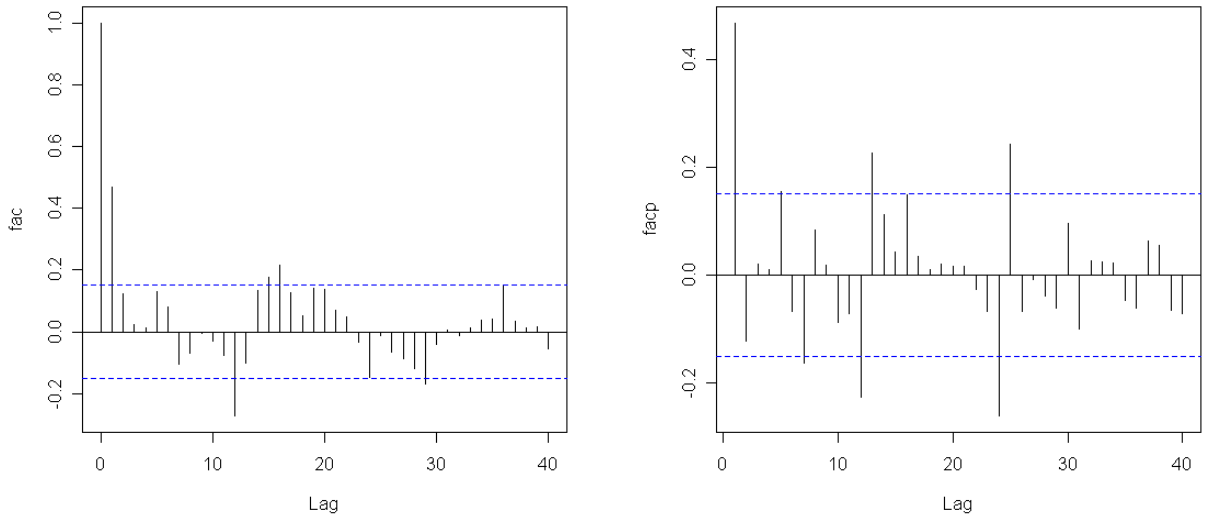
- SARIMA $(1,0,0)(0,1,1)_{12}$ descrito por:

$$Z_t = \frac{1-\Theta_1 B^{12}}{(1-B^{12})(1-\phi_1 B)} a_t \,, \qquad (4)$$

em que Z_t é a observação temporal no instante t, Θ_1 é o parâmetro de médias móveis sazonal de ordem 1, B é o operador translação para o passado, definido por $B^m Z_t = Z_{t-m}$, sendo m o período de tempo defasado, a_t é a componente aleatória e ϕ_1 é o parâmetro do coeficiente auto-regressivo de ordem 1.

- SARIMA $(1,0,0)(0,1,1)_{12}$ com intervenções abruptas temporárias em maio/1996 (observação 65) e junho/1997 (observação 78), intervenção abrupta permanente em dezembro/2000 (observação 120).

FIGURA 6 - Função de autocorrelação (*fac*) e autocorrelação parcial (*facp*) da série de produção brasileira de álcool, em meses, sazonalmente ajustada, no período de janeiro de 1991 a dezembro de 2005



O modelo adequado para esta situação é dado por:

$$Z_t = \omega_1 x_{1,t} + \omega_2 x_{2,t} + \frac{\omega_3}{1-\delta_3 B} x_{3,t} + \frac{1-\Theta_1 B^{12}}{(1-B^{12})(1-\phi_1 B)} a_t \,, \qquad (5)$$

em que os efeitos de intervenção representados por $x_{1,t}$, $x_{2,t}$ e $x_{3,t}$ são:

$$x_{1,t} = \begin{cases} 0, t < 65; \\ 1, t \geq 65; \end{cases}, \quad x_{2,t} = \begin{cases} 0, t < 78; \\ 1, t \geq 78; \end{cases}, \quad x_{3,t} = \begin{cases} 0, t \neq 120; \\ 1, t = 120. \end{cases}$$

As estimativas dos parâmetros dos modelos propostos são apresentadas na Tabela 1.

Na Tabela 2, são exibidas as estatísticas para o teste de Box-Pierce, o qual verifica que os resíduos do modelo são ruído branco (independente e identicamente distribuído).

Os valores reais, as estimativas dos valores preditos para a produção de álcool e o erro de previsão, para o período de janeiro de 2006 a setembro do mesmo ano, encontram-se na Tabela A do apêndice.

A escolha do melhor modelo foi feita com base no critério de informação de Akaike (AIC) e erro quadrático médio de previsão (EQMP). Esses dois critérios e o EPM, o qual avalia o desempenho do ajuste, são mostrados na Tabela 3.

TABELA 1 - Estimativas dos parâmetros dos modelos SARIMA com e sem intervenções para a série de produção brasileira de álcool, em meses, no período de janeiro de 1991 a dezembro de 2005

Modelo	Parâmetro	Estimativa	Erro padrão	p-valor
SARIMA $(1,0,0)(0,1,1)_{12}$	ϕ_1	0,5843	0,0658	0,0000
	Θ_1	0,6428	0,0683	0,0000
SARIMA $(1,0,0)(0,1,1)_{12}$*	ϕ_1	0,5401	0,0696	0,0000
	Θ_1	0,6658	0,0599	0,0000
	ω_1	24,4040	10,5352	0,0218
	ω_2	-29,6270	11,0087	0,0077
	ω_3	-30,7567	12,6068	0,0158
	δ_3	0,9665	0,0332	0,0000

Nota:*Modelo com intervenção.

TABELA 2 - Estatísticas para o teste de Box-Pierce de modelos SARIMA com e sem intervenções para a série de produção brasileira de álcool, em meses, no período de janeiro de 1991 a dezembro de 2005

Modelo	GL	Q40	χ^2
SARIMA $(1,0,0)(0,1,1)_{12}$	38	39,28	53,38
SARIMA $(1,0,0)(0,1,1)_{12}$*	34	37,52	49,80

Nota:*Modelo com intervenção.

TABELA 3 - Estimativas dos critérios de AIC, EQMP e EPM de modelos SARIMA com e sem intervenção da série de produção brasileira de álcool, em meses, no período de janeiro de 1991 a dezembro de 2005

Modelo	AIC	EQMP	EPM
SARIMA $(1,0,0)(0,1,1)_{12}$	1014,58	519,47	17,52%
SARIMA$(1,0,0)(0,1,1)_{12}$*	1007,37	382,41	25,97%

Nota:*Modelo com intervenção.

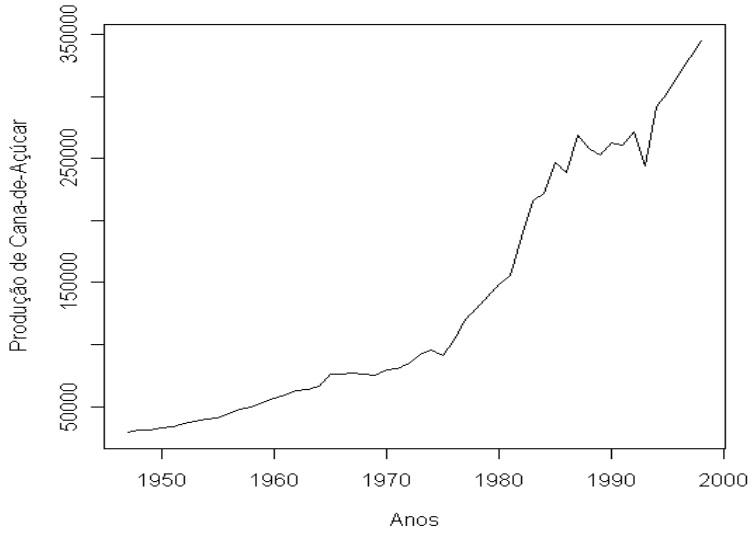
O modelo escolhido para representar a série de produção mensal de álcool foi o SARIMA$(1,0,0)(0,1,1)_{12}$ com intervenção, equação (5). Este modelo, apesar de possuir um maior erro percentual de previsão, possui menor AIC e menor EQMP. O modelo (5) substituído por suas respectivas estimativas é dado por:

$$Z_t = 24,404x_{1,t} - 29,627x_{2,t} - \frac{30,7567}{1-0,966B}x_{3,t} + \frac{1-0,666B^{12}}{(1-B^{12})(1-0,540B)}a_t .$$

4.1.2. Série de produção de cana-de-açúcar

O gráfico da série de produção anual de cana-de-açúcar pode ser visto na Figura 7. Visualmente, pode-se intuir a presença da componente tendência e ausência de ciclo, o qual corresponde a períodos maiores que 12 meses.

FIGURA 7 - Série de produção brasileira de cana-de-açúcar, em milhões de toneladas, no período anual de 1947 a 1998



Fonte: IPEA

A função de autocorrelação e o periodograma da série de produção anual de cana-de-açúcar são exibidos na Figura 8. No gráfico da função de autocorrelação é possível perceber que a série não decai rapidamente para zero, indicando a sua não estacionaridade. O teste do sinal (Cox-Stuart) foi aplicado para confirmar a existência da tendência. Considerando-se um nível de significância $\alpha = 0,05$ e tomando-se as 52 observações, tem-se que: $c = \frac{52}{2} = 26$ e $n = 26(Z_t \neq Z_{t+c}\ \forall t)$. O número de sinais positivos T_2 é igual a 26 e $t = 17,19$. Como $T_2 > 26 - 17,19$, a série apresenta tendência.

A série sem tendência é apresentada na Figura 9. É possível observar várias oscilações na produção de cana, especialmente a partir de 1975. Essas oscilações podem ser efeito de intervenções que são bem explicadas pela fase de implantação do Proálcool - medida tomada pelo governo para solucionar a crise energética ocasionada pelas crises de abastecimento do petróleo. Iniciou-se um deslocamento do consumo dos combustíveis derivados dos fósseis para o álcool derivado da cana. Dentre as várias medidas tomadas pelo governo, pode-se citar as políticas de incentivo ao crédito para a produção de cana (MACIEL, 1991). Inicialmente, existem indícios de intervenções nas observações 37 e 48, referentes aos anos de 1983 e 1994, respectivamente. Apesar disso, é considerada no ajuste como possível intervenção a referente ao ano de 1983.

FIGURA 8 - Função de autocorrelação e periodograma da série de produção brasileira de cana-de-açúcar, em milhões de toneladas, no período anual de 1947 a 1998

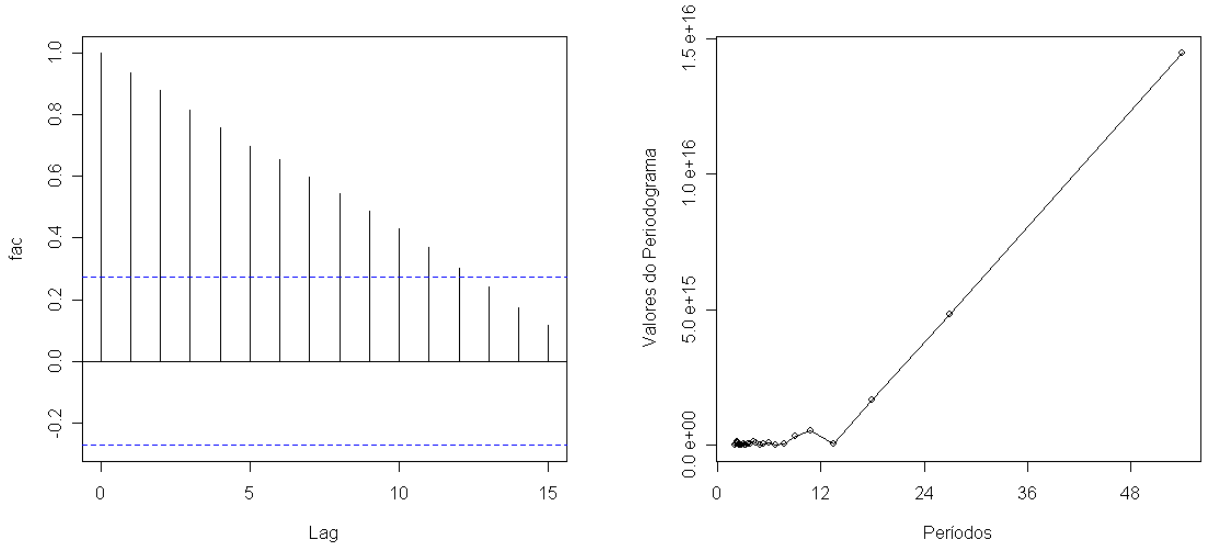
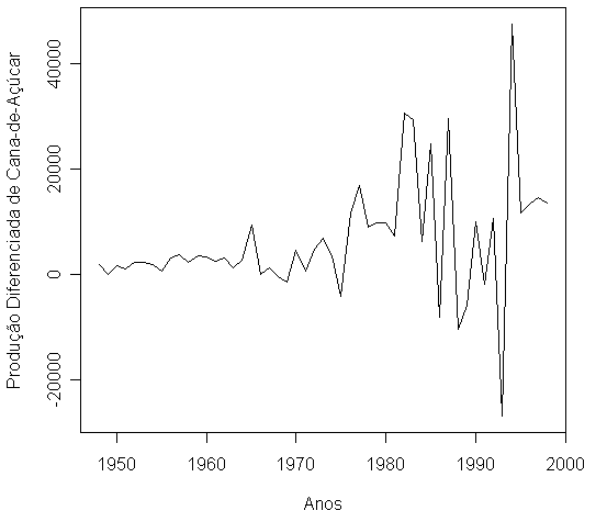
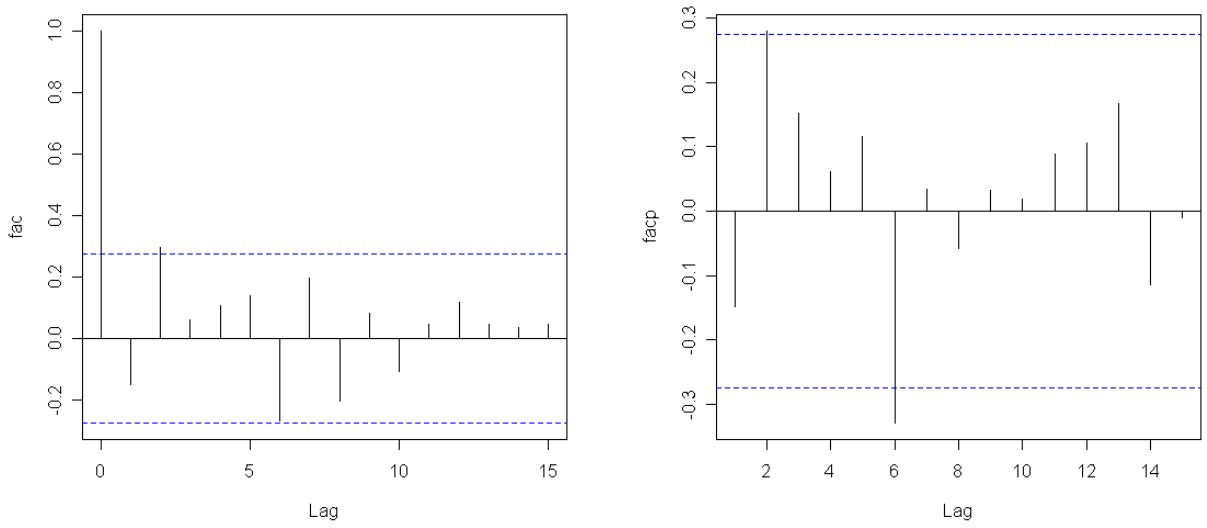


FIGURA 9 - Série diferenciada da produção brasileira de cana-de-açúcar, em milhões de toneladas, no período anual de 1947 a 1998



As funções de autocorrelação e autocorrelação parcial da série sem tendência, usadas na identificação das ordens do modelo, são exibidas na Figura 10.

FIGURA 10 - Funções de autocorrelação (fac) e autocorrelação parcial (facp) da série diferenciada da produção brasileira de cana-de-açúcar, em milhões de toneladas, no período anual de 1947 a 1998



Os modelos propostos a partir do *lag* significativo 2 são:

- ARIMA (2,1,1) dado por:

$$Z_t = \frac{(1-\theta_1 B)}{(1-\phi_1 B-\phi_2 B^2)(1-B)} a_t \ , \qquad (6)$$

em que Z_t é a observação temporal no instante t, θ_1 é o parâmetro de médias móveis de ordem 1, B é o operador translação para o passado, definido por $B^m Z_t = Z_{t-m}$, sendo m o período de tempo defasado, a_t a componente aleatória, ϕ_1 o parâmetro auto-regressivo de ordem 1 e ϕ_2 o parâmetro auto-regressivo de ordem 2.

- ARIMA (2,1,1) com intervenção abrupta permanente no ano de 1983 (observação 37) . É descrito por:

$$Z_t = \frac{\omega_1}{1-\delta_1 B} x_{1,t} + \frac{(1-\theta_1 B)}{(1-\phi_1 B-\phi_2 B^2)(1-B)} a_t \ , \qquad (7)$$

em que ω_1 e δ_1 são os parâmetros de intervenção e a variável *dummy* $x_{1,t}$ é:

$$x_{1,t} = \begin{cases} 0, t \neq 1983; \\ 1, t = 1983. \end{cases}$$

Na Tabela 4, são apresentadas as estimativas dos parâmetros dos modelos propostos.

TABELA 4 - Estimativas dos parâmetros do modelo ARIMA com e sem intervenção da série de produção brasileira de cana-de-açúcar, em milhões de toneladas, no período anual de 1947 a 1998

Modelo	Parâmetro	Estimativa	Erro padrão	p-valor
ARIMA(2,1,1)	ϕ_1	0,5214	0,2218	0,023
	ϕ_2	0,420	0,161	0,012
	θ_1	0,619	0,237	0,012
ARIMA(2,1,1)*	ϕ_1	0,477	0,000	0,000
	ϕ_2	0,523	0,000	0,000
	θ_1	0,858	0,068	0,000
	ω_1	30.722,0	10.978,0	0,006
	δ_1	0,777	0,206	0,000

Nota:*Modelo com intervenção.

Na Tabela 5, são exibidas as estatísticas para o teste de Box-Pierce, o qual identifica se o resíduo é ruído branco.

TABELA 5 - Estatísticas para o teste de Box-Pierce do modelo ARIMA com e sem intervenção da série de produção brasileira de cana-de-açúcar, em milhões de toneladas, no período anual de 1947 a 1998

Modelo	GL	Q15	χ^2
ARIMA (2,1,1)	12	13,98	21,03
ARIMA (2,1,1) *	10	13,17	18,31

Nota:*Modelo com intervenção.

Pela análise da Tabela 5, observa-se que os dois resíduos constituem um ruído branco. Os valores reais, as estimativas dos valores preditos para a produção de cana-de-açúcar e o erro de previsão, para o período de 1999 a 2004, encontram-se na Tabela B do apêndice.

A escolha do melhor modelo foi feita com base no critério de informação de Akaike (AIC) e erro quadrático médio de previsão (EQMP). Esses dois critérios e o MAPE, que avalia o desempenho dos ajustes, podem ser vistos na Tabela 6.

TABELA 6 - Estimativas dos critérios de AIC, EQMP e EPM do modelo ARIMA com e sem intervenção, da produção brasileira de cana-de-açúcar, em milhões de toneladas, no período anual de 1947 a 1998

Modelo	AIC	EQMP	EPM
ARIMA (2,1,1)	957,68	714.168.780	6,57%
ARIMA (2,1,1)*	955,55	644.631.274	6,19%

Notas:*Modelo com intervenção.

O modelo escolhido para representar a série de produção anual de cana-de-açúcar, em milhões de toneladas, é o ARIMA (2,1,1) com intervenção na observação 37 referente ao ano de 1983, por apresentar menor AIC e menor EQMP. Este modelo apresentou também um bom desempenho de ajuste com EPM pequeno. O modelo (7) substituído por suas respectivas estimativas é dado por:

$$Z_t = \frac{30722,0}{1-0,777_1 B} x_{1,t} + \frac{(1-0,858B)}{(1-0,477B-0,523B^2)(1-B)} a_t \quad ,$$

em que em que Z_t é a observação temporal no instante t, B é o operador translação para o passado e a_t é a componente aleatória.

4.2. Modelagem multivariada

4.2.1. Estágio estacionário

As séries usadas nesta análise foram as séries de produção dos combustíveis: álcool, gasolina e óleo diesel no período de janeiro de 2000 a setembro de 2006. Para o ajuste do modelo VAR, as séries devem se encontrar em estágio estacionário. As séries de óleo diesel e gasolina apresentam tendência e a série de álcool apresenta sazonalidade.

Na Tabela 7, são apresentados os testes de Cox-Stuart para tendência e de Fisher para sazonalidade.

TABELA 7 - Teste de Cox-Stuart para tendência e teste de Fisher para sazonalidade das séries de produção mensal dos combustíveis álcool, gasolina e óleo diesel no período de janeiro de 2000 a setembro de 2006

Variável	g	$Z_{0,05}$	T_2	n	t	Conclusão
Álcool	0,62	0,16	-	-	-	$g > Z$: tem sazonalidade
Gasolina	-	-	22	40	25,19	$T_2 > n-t$: tem tendência
Óleo diesel	-	-	38	40	25,19	$T_2 > n-t$: tem tendência

Para eliminar a tendência das séries de gasolina e óleo diesel e a sazonalidade da série de álcool, a tendência e os componentes sazonais foram estimados pelo método de regressão e são apresentados nas Tabelas 8 e 9, respectivamente.

TABELA 8 - Coeficientes de regressão e erro padrão ajustados às séries de produção brasileira de óleo diesel e gasolina, em metros cúbicos (m^3), no período mensal de janeiro de 2000 a setembro de 2006

Variável	Coeficientes	Estimativas	Erro Padrão
Gasolina	b_0	1.482.239,18	42.263,54
	b_1	14.274,64	4.436,26
	b_2	-540,02	125,30
	b_3	5,09	1,00
Óleo diesel	b_0	2.498.614	44.948,80
	b_1	10.327	952,34

Na Tabela 9, são apresentados os fatores sazonais para cada mês do ano utilizados para encontrar a série sazonalmente ajustada da produção de álcool.

TABELA 9 - Fatores sazonais da série de produção brasileira de álcool, em meses, no período de janeiro de 1991 a dezembro de 2005

Fator Sazonal	Valor	Fator Sazonal	Valor
α_1	-83,90	α_7	79,04
α_2	-94,09	α_8	90,81
α_3	-99,26	α_9	79,76
α_4	-64,98	α_{10}	73,39
α_5	11,99	α_{11}	7,69
α_6	56,25	α_{12}	-56,70

4.2.2. Ajuste do modelo VAR

O vetor de séries $\boldsymbol{Z}_t$, constituído pelas séries de produção de álcool, óleo diesel e gasolina, representado respectivamente por A_t, G_t e O_t, é dado por $\boldsymbol{Z}_t = \begin{bmatrix} A_t \\ G_t \\ O_t \end{bmatrix}$.

Na Tabela 10, são apresentados os critérios de informação, AIC e BIC, do ajuste do VAR, para ordens de 1 a 8. A ordem selecionada foi $p = 2$, por apresentar os menores valores de AIC e BIC.

TABELA 10 - Estatísticas resultantes de ajustes de modelos VAR (*p*), *p* = 1,..., 8, para as séries de produção brasileira de álcool, óleo diesel e gasolina, em metros cúbicos (m^3), no período mensal de janeiro de 2000 a setembro de 2006

Ordem	1	2	3	4	5	6	7	8
AIC	61,79	61,44	61,47	61,57	52,78	61,69	61,87	62,02
BIC	62,08	62,01	62,33	62,61	62,99	63,40	63,86	64,30

A equação do modelo VAR (2) pode ser escrita como:

$$\boldsymbol{Z}_t = \boldsymbol{\Phi}_0 + \boldsymbol{\Phi}_1 \boldsymbol{Z}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{Z}_{t-2} + a_t ,$$

em que $\boldsymbol{Z}_{t-j}$ é o vetor das variáveis defasadas de ordem j, $j = 1,2$ e $\boldsymbol{a}_t = \begin{bmatrix} \varepsilon_{A_t} \\ \varepsilon_{Ot} \\ \varepsilon_{Gt} \end{bmatrix}$ é o vetor de resíduos sendo $a_t \sim RB(0, \boldsymbol{\Sigma})$.

As estimativas dos parâmetros e seus erros padrão são dados na Tabela 11.

TABELA 11 - Ajuste de um modelo VAR (2) para as séries de produção de álcool, óleo diesel e gasolina, em metros cúbicos (m3), no período mensal de janeiro de 2000 a setembro de 2006

Parâmetros	Estimativa			Erro padrão		
$\boldsymbol{\Phi}_1$	$0,739^*$	$-0,000$	$0,000$	$0,113$	$0,000$	$0,000$
	$1375,860$	$-0,389^*$	$-0,755^*$	$1489,010$	$0,118$	$0,229$
	$-190,066$	$-0,075$	$-0,698^*$	$670,396$	$0,053$	$0,103$
$\boldsymbol{\Phi}_2$	$-0,190^*$	$0,000$	$0,000$	$0,114$	$0,000$	$0,000$
	$-2173,030$	$-0,049$	$-0,479^*$	$1490,840$	$0,114$	$0,228$
	$775,582$	$-0,106^*$	$-0,590^*$	$671,217$	$0,051$	$0,102$

Nota:*Estimativas significativas a 5%.

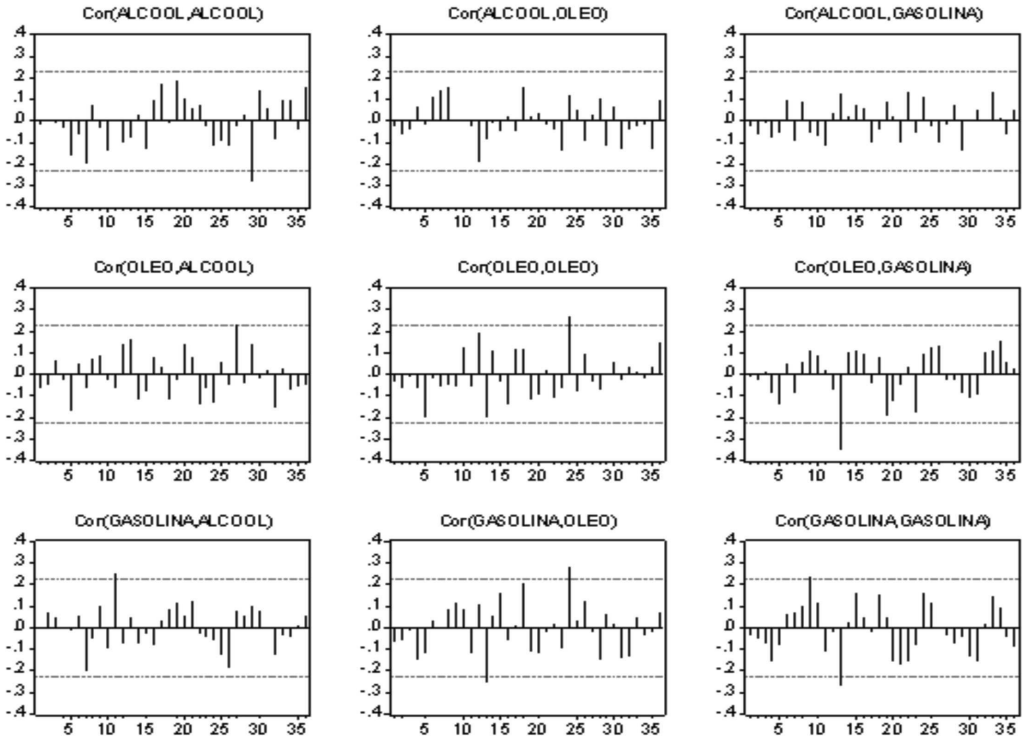
Levando-se em conta apenas os coeficientes significativos, o modelo ajustado pode ser representado aproximadamente por:

$$\begin{aligned} A_t &= 0,739 A_{t-1} - 0,190 A_{t-2} + \varepsilon_{A_t} \\ O_t &= -0,389 O_{t-1} - 0,755 G_{t-1} - 0,479 G_{t-2} + \varepsilon_{Ot}, \\ G_t &= 0,106 O_{t-2} - 0,698 G_{t-1} - 0,590 G_{t-2} + \varepsilon_{Gt}. \end{aligned}$$

A estatística multivariada de Box-Ljung-Pierce, para verificar se o resíduo é ruído branco, resultou em: $Q(36)$=243,786 e $\chi^2(306) = 347,79$. Como $Q(36) < \chi^2(306)$, o ajuste é adequado.

Na Figura 11, é apresentada a função de correlação para os resíduos do modelo VAR(2), confirmando-se a adequação do modelo.

FIGURA 11 - Representação gráfica da função de correlação dos resíduos das produções brasileiras de álcool, óleo diesel e gasolina, em metros cúbicos (m^3), no período mensal de janeiro de 2000 a setembro de 2006



Por estas estimativas, conclui-se que os valores atuais da produção de álcool não são influenciados pelos valores defasados do óleo diesel e da gasolina. Esta conclusão estatística contraria os fatos histórico e econômico referentes ao mercado de combustíveis. Esperava-se uma dependência do álcool em relação à gasolina e ao óleo diesel porque o uso do álcool como combustível reduz a participação da gasolina na estrutura de refino do petróleo e isso amplia a produção de óleo diesel.

Os valores presentes de produção de óleo diesel são influenciados por seus valores passados e também pelos valores passados da produção de gasolina.

Os valores atuais da gasolina são influenciados por seus valores passados e também pelos valores passados de óleo diesel.

5. Conclusões

De modo geral, os modelos univariados, propostos por Box e Jenkins, ajustaram-se bem aos dados. Para a série de produção de cana-de-açúcar, o modelo que melhor se ajustou aos dados foi o ARIMA (2,1,1) com intervenção no ano de 1983 (observação 37). Neste ano, observou-se um aumento de 307.220 milhões de toneladas na produção de cana em razão da alta demanda desta cultura para a produção de álcool combustível.

A série de produção de álcool apresentou correlações sazonais, sendo o modelo SARIMA $(1,0,0)(0,1,1)_{12}$ com intervenções em maio de 1996 (referente à observação 65), junho de 1997 (observação 78) e dezembro de 2000 (observação 120), o modelo de melhor ajuste. O aumento de 24,40 da produção, refletido em maio de 1996, está associado ao período de estabilização da economia, em que foi estabelecido um novo ciclo de desenvolvimento elevando os índices de expansão da economia e do consumo de energia (Brasil, 2007). Outro fato associado foi a criação, nesta época, da chamada "frota verde" que foi o estímulo e a determinação do uso do álcool hidratado em determinadas classes de veículos leves, como os carros oficiais e táxis. Houve também o aumento de 22% para 24% do percentual de adição de álcool etílico anidro à gasolina (Evolução do Programa Nacional do álcool – Proálcool, 2008). Os efeitos refletidos em junho de 1997 e dezembro de 2000, quando ocorreram quedas na produção, são explicados, em parte, por sucessivas crises externas, principalmente a crise cambial nos países asiáticos, que acabaram contaminando a economia nacional, obrigando o governo a tomar medidas que levaram a uma forte retração no crescimento econômico, tendo o PIB apresentado um crescimento de apenas 0,9% no ano de 1998 e de 1,6% em 1999. O baixo desempenho da economia teve reflexos no consumo de energia de 1999, notadamente quanto às energias associadas ao uso individual, como o álcool hidratado com queda de 8,6% no consumo, a gasolina automotiva com queda de 6,3%, o querosene de aviação com queda de 6,3% e energia elétrica residencial, com apenas 2,4% de crescimento (Brasil, 2007). Estas quedas foram de 29,63 e 30,76, respectivamente.

No ajuste multivariado, foram obtidas relações de dependência entre as séries de produção de gasolina e óleo diesel. O melhor ajuste foi dado pelo modelo VAR (2), indicando uma dependência entre a produção de óleo e gasolina de até dois períodos.

Referências bibliográficas

AKAIKE, H. (1973). Maximum likelehood identification of gaussian autoregressive moving average models. Biometrika, London, v.60, n.21, p. 255-265.

ANP. AGÊNCIA NACIONAL DO PETRÓLEO. Disponível em <www.anp.gov.br/doc/dados−estatisticos.> Acesso em: 14 set. 2006.

ASSIS, A. N.; LOPES, L. B. R. (1980). A ineficiência da política de preços para conter o consumo de derivados de petróleo. *Revista Brasileira de Economia*, Rio de Janeiro, v. 34, n.3, p.417-428, jul./set.

BRASIL. (2007). Ministério de Minas e Energia. Empresa de Pesquisa Energética Balanço Energético Nacional 2007: Ano base 2006. Relatório final/Ministério de Minas e Energia. Rio de Janeiro: Empresa de Pesquisa Energética. 192 p. Disponível em < http://www.mme.gov.br/site/menu/select_main_menu_item.do?channelId=1432> Acesso em: 07 de julho 2008.

BOX, G. E. P.; JENKINS, G. M. (1976). *Time Series Analysis: forecasting and control*. San Francisco: Holden Day (Revised edition).

BROWN, R. I. (1980). Um esquema para avaliar os impactos de estratégias diversas para etanol, outras substituições e a racionalização na demanda de derivados de petróleo em 1985. *Ciência e Cultura*, Campinas: v. 32, n. 8, p. 1032-1040.

BURNQUIST, H. L.; BACCHI, M. R. P. (1980). A demanda por gasolina no Brasil: uma análise utilizando técnicas de co-integração. In: *CONGRESSO BRASILEIRO DE ECONOMIA E SOCIOLOGIA RURAL*, 40, 002, Passo Fundo, RS. Equidade e Eficiência na Agricultura Brasileira, 2002. Ciência e Cultura, Campinas: v. 32, n. 8, p. 1032-1040.

EVOLUÇÃO DO PROGRAMA NACIONAL DO ÁLCOOL – Proálcool. Disponível em <http://www.biodieselbr.com/proalcool/pro-alcool.htm>. Acesso em: 08 jul. 2008.

IPEA. Instituto de Pesquisas Econômicas Aplicadas. Disponível em <www.ipeadata.gov.br/ipeaweb.dll/ipeadata? 406640843>. Acesso em: 10 dez. 2005.

LEITE, A. D. (1997). *A energia do Brasil*. Rio de Janeiro: Nova Fronteira, 598 p.

MACIEL, T. J. L. (1991). *Análise do setor alcooleiro do Brasil: perspectivas do Proálcool*. Tese (Doutorado em Economia Rural) – Viçosa: Universidade Federal de Viçosa, MG, 83 p.

MOREIRA, R. B. (1996). Modelos para a projeção do consumo nacional e regional de óleo diesel. *Texto para Discussão n. 443*, Rio de Janeiro: IPEA.

MORETTIN, P. A. (2008). *Econometria financeira - um curso em séries temporais financeiras.* São Paulo: Editora Blucher, 319 p.

MORETTIN, P. A.; TOLOI, C. M. C. (2006). Análise de séries temporais. São Paulo: Edgard Blücher. 439 p.

PIERCE, D. A. (1979). Some recent developments in seasonal adjustment. In: BRILLINGER, D. R.; TIAO, G. C. (Ed.). *Reports on Directions in Time Series*. Institute of Mathematical Statistics. p.123-140.

PRIESTLEY, M. B. (1989). *Spectral Analysis and Time Series*. London: Academic Press. 407p.

APÊNDICE

TABELA A - Valores reais, preditos e erro de previsão para a série de produção brasileira de álcool, em meses, no período de janeiro de 1991 a dezembro de 2005

Modelo	Meses	Z_{t+h}	$\hat{Z}_t(h)$	$e_t(h)$
SARIMA(1,0,0)(1,0,1)$_{12}$	Jan./06	24,78	20,25	4,53
	Fev./06	9,43	12,50	-3,07
	Mar./06	7,48	7,53	-0,06
	Abr./06	65,66	41,33	24,34
	Mai./06	169,60	127,93	41,68
	Jun./06	197,85	166,86	30,99
	Jul./06	213,14	186,25	26,89
	Ago./06	223,15	203,57	19,58
	Set./06	212,04	196,23	15,81
SARIMA(1,0,0)(1,0,1)$_{12}$*	Jan./06	24,78	23,35	1,43
	Fev./06	9,43	16,68	-7,25
	Mar./06	7,48	12,12	-4,64
	Abr./06	65,66	45,15	20,50
	Mai./06	169,6	131,54	38,06
	Jun./06	197,85	171,94	25,91
	Jul./06	213,14	191,51	21,63
	Ago./06	223,15	208,18	14,97
	Set./06	212,04	200,49	11,55

Notas:*Modelo com intervenção. Z_{t+h} = valor real; $\hat{Z}_t(h)$ = valor predito; $e_t(h)$ = erro de previsão

TABELA B - Valores reais e preditos da produção brasileira de cana-de-açúcar, em milhões de toneladas, no período anual de 1947 a 1998

Modelo	Meses	Z_{t+h}	$\hat{Z}_t(h)$	$e_t(h)$
ARIMA(2,1,1)	1999	333.847	356.816,1	-22.969,1
	2000	326.121	368.572,3	-42.451,3
	2001	344.292	379.556,5	-35.264,5
	2002	364.389	390.219,7	-25.830,7
	2003	396.012	400.391,4	-4.379,4
	2004	415.205	410.172,0	5.033,0
ARIMA(2,1,1)*	1999	333.847	355.030,3	-21.183,3
	2000	326.121	366.885,1	-40.764,1
	2001	344.292	377.696,8	-33.404,8
	2002	364.389	389.088,0	-24.699,0
	2003	396.012	400.202,3	-4.190,30
	2004	415.205	411.481,4	3.723,60

Notas:*Modelo com intervenção. Z_{t+h} = valor real; $\hat{Z}_t(h)$ = valor predito; $e_t(h)$ = erro de previsão

Abstract

The Brazilian energy sector is consisted of renewable and non-renewable sources. Non-renewable energy sources such as those derived from oil, diesel and gasoline, have contributed to a technological progress never seen before. However, using such fossils has increasing the pollution indices and changing the weather. Nowadays, sugar cane and the alcohol appear as alternative sources of energy. Time series techniques were used in this study to model the production of sugar cane, alcohol, diesel oil and oil. Univariate models were obtained for the series of production of cane and alcohol, for which the best estimates were obtained using the model with intervention. The series of production of alcohol, diesel oil and gasoline were used in the multivariate modeling, where the VAR (2) model was fitted. It was observed that the production of diesel oil and gasoline are related, and the production of alcohol is independent of them.

Keywords: energy sources, time series, SARIMA model, intervention, VAR model.

Modelo de Previsão Combinada: Uma Aplicação à Série Mensal de Passageiros Transportados do Sistema de Transporte Público da Região Metropolitana do Recife-PE

*Dirac M. Cordeiro**
*Gauss M. Cordeiro***

Resumo

Neste artigo desenvolve-se um modelo de previsão combinada para explicar o comportamento da série temporal "passageiros transportados" do Sistema de Transporte Público de Passageiros da Região Metropolitana do Recife-PE (STPP/RMR). Nesse modelo, aplica-se a proposta de Bunn (1985). Duas previsões pontuais, as melhores em termos da medida de eficiência, fundamentada no erro quadrático médio (EQM), foram combinadas linearmente para a obtenção de melhores previsões, através de métodos estatísticos, análise dos resíduos e no percentual de redução da variância não explicada da modelagem combinada, em relação às individualizadas. Constata-se que em todas as abordagens atribuídas à série de interesse, as previsões cresceram, quando comparadas com as observadas antes de 2003 – período em que houve uma queda acentuada na demanda de passageiros, em virtude da grande incidência do transporte "clandestino".

Palavras-chave: Erro Quadrático Médio, Erro de Monte Carlo, Metodologia de Box-Jenkins, Modelo de Holt-Winters, Modelo de Previsão Combinada.

* Endereço para correspondência: Universidade de Pernambuco – UPE, Rua Benfica, nº. 455, Madalena, CEP 50751-460, Recife – PE, Brasil. *E-mail*: dmc@upe.poli.br.

** Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco – UFRPE, Rua Dom Manoel de Medeiros, s/nº, Dois Irmaõs, CEP 50171-900, Recife, PE, Brasil. E-mail: gauss@deinfo.ufrpe.br.

1. INTRODUÇÃO

O objetivo deste trabalho é propor uma modelagem estatística que substitua de forma adequada os modelos de Box-Jenkins (1976) e Holt-Winters (MAKRIDAKIS et al., 1998), atualmente adotados, para explicar o comportamento da série do número médio de passageiros transportados do Sistema de Transporte Público de Passageiros da Região Metropolitana do Recife-PE (STPP/RMR). O modelo a ser proposto é obtido por meio de uma combinação linear entre os modelos anteriores, cuja equação de previsão é representada por uma ponderação, utilizando inferência bayesiana para seleção sequencial dos pesos. Nestes termos, busca-se um modelo com melhor desempenho, por meio da análise dos resíduos e minimização do Erro Quadrático Médio (EQM). Para se ter uma ideia da importância do sistema no cenário econômico da região, deve-se considerar algumas informações relevantes, dos pontos de vista operacional e financeiro, tais como: investimento em frota de ônibus – U$ 76 milhões; faturamento mensal do Sistema – U$ 23 milhões; passageiro médio transportado por dia útil – 1,3 milhão; pessoal alocado para operação – 15 mil; número médio de viagens por dia útil – 29 mil e arrecadação de impostos por mês – U$ 1,2 milhão.

O modelo de remuneração das empresas operadoras apoia-se num orçamento de arrecadação, representado por metas operacionais de serviço e de passageiros, propostas pelas empresas operadoras, com projeções orçamentárias que serão a base da arrecadação e da remuneração do Sistema. O atual cenário do “Sistema de Transporte Público de Passageiros da Região Metropolitana do Recife” (STPP/RMR), no que se refere à demanda, é hoje completamente diferente daquele que se verificava em junho de 2003, quando esse sistema passou por uma maciça intervenção, viabilizada pelos poderes executivos dos municípios da Região Metropolitana do Recife e do Estado de Pernambuco. Uma decisão conjunta desses poderes desencadeou uma ação integrada entre os órgãos gestores de transporte e trânsito, e polícia militar, no sentido de eliminar a concorrência danosa provocada pelo transporte “clandestino” que se proliferou na região, sem quaisquer tipos de obrigação ou responsabilidade social (CORDEIRO et al., 2006). O sucesso dessa ação resultou na devolução de grande fatia de demanda, que passou a usar o transporte formal e regular, operado pelo STPP/RMR. Essa nova realidade propiciou a implantação de uma nova metodologia de remuneração, em que as variáveis operacionais são pré-estabelecidas através de metas que permitem um orçamento prévio, com base na otimização do serviço ofertado. Na Seção 2, apresentam-se os testes estatísticos para avaliar as componentes da série temporal. A Seção 3 introduz a metodologia de um modelo de previsão combinada com base nos modelos de Holt-Winters e Box-Jenkins. A Seção 4 trata do desenvolvimento do modelo especificado na Seção 3. Na Seção 5, aborda-se as estatísticas para as autocorrelações

residuais e faz-se uma análise detalhada dos resultados. Na Seção 6, compara-se os resultados das previsões dos modelos individualizados com o modelo combinado.

2. ANÁLISE EXPLORATÓRIA DOS DADOS DA SÉRIE TEMPORAL

A série dos passageiros transportados representada pela média mensal dos dias úteis revela, entre outros atributos:

Assimetria positiva da distribuição da série, ou seja, a amostra é mais populosa em valores baixos do que em valores altos – somente 27% dos passageiros transportados são acima de 1,24 milhão, dos quais 46% são acima de 1,34 milhão;

Abaixo de 940 mil passageiros não se registra resultados para essa demanda. Especificamente, em situações não-adversas, esse resultado seria sempre mantido. Portanto, valores de demanda abaixo de 940 mil passageiros teriam probabilidade de ocorrência próxima de zero.

Além da inspeção gráfica, utilizamos alguns testes de hipóteses com o intuito de verificar de forma excludente se realmente existem as principais componentes: tendência e sazonalidade. Logo, as hipóteses a serem testadas são: H_0: não existe componente e H_1: existe componente.

O teste para avaliação da tendência antes da sua estimação é baseado na série temporal $A[X_t]$ livre da componente sazonal e obtida por meio da aplicação de uma operação linear que transforma $A[X_t] = T_t -$ componente tendência para $t = 1,\ldots,N$. A transformação aplicada a X_t para obtenção de T_t é a de um filtro linear representado por uma média móvel centrada e de tamanho igual ao comprimento sazonal $L = 12$. Esse teste não-paramétrico é baseado no coeficiente de correlação de Spearman ρ (MILONE, 2003):

$$\rho = 1 - \frac{6\sum_{t=1}^{n}(R_t - t)^2}{N\left[\left(N^2 - 1\right)\right]}, \qquad (1)$$

em que $R_t = r(T_t)$ é o posto de T_t com N observações e aqui $N = 48$. O valor desse coeficiente indica a interpretação da correlação; no caso do estudo da série X_t a interpretação da influência da componente tendência é relevante, pois $\rho = 0{,}65$. Desta forma, admite-se a hipótese não-nula para T_t. Quanto ao teste para avaliação da sazonalidade antes da estimação, utiliza-se o teste paramétrico com base no cálculo da estatística ES (MORETTIN; TOLOI, 1986) dada pela equação:

$$ES = \left(\frac{N-L}{L-1}\right)\left[\frac{m\sum_{j=1}^{L}\left(\overline{X_j^*} - \overline{X^*}\right)^2}{\sum_{j=1}^{L}\sum_{i=1}^{m}\left(X_{ij}^* - \overline{X_j^*}\right)^2}\right], \quad (2)$$

em que $m = 12$ representa o número constante de estações sazonais para o mês $j = 1,\ldots,12, X_{ij}^*$ é a série livre da componente tendência para o mês j da estação i, $\overline{X_j^*}$ e $\overline{X^*}$ representam as médias para o mês j e segundo todos os meses, respectivamente. O valor dessa estatística é igual a 10,46 que comparado com o ponto crítico da distribuição F (11, 36, 5%) = 2,06 implica na rejeição da hipótese nula de existência de fatores sazonais. Mais detalhes sobre o assunto, *vide* Cordeiro (2002).

Na Figura 1, observa-se, claramente, que as amplitudes da série variam com a média, sendo isto um forte indicativo para a modelagem multiplicativa da componente sazonal. Bowerman (1987) sugere que o melhor critério de escolha entre os fatores multiplicativos ou aditivos é calcular o valor do Erro Absoluto Médio (EAM); e, para ambas as modelagens, a opção escolhida corresponde ao menor erro. Com efeito, o valor do EAM para as modelagens aditiva e multiplicativa diferiu muito pouco; porém, com menor valor, quando se atribui um ajuste multiplicativo à componente sazonal.

A série livre do efeito sazonal, na Figura 2, é discernida por meio de um amortecimento usando uma média móvel centrada em $C = 7$ e de tamanho $q = 12$. Observe-se que essa série inicialmente tem uma inclinação quase nula e, somente a partir do trigésimo mês, existe uma forte inclinação positiva. Ressalta-se que, o período compreendido até $t = 30$, o sistema de transporte sofria com uma concorrência bastante danosa, proveniente do transporte não regulamentado (“clandestino”). Como agravante, o transporte não regulamentado chegou a transportar em torno de 25% da demanda de passageiros do sistema, representando cerca de 300.000 passageiros por dia, com uma perda de receita estimada em US$ 10 milhões de dólares a cada mês (EMTU, 2004).

A Figura 3 elucida a comparação gráfica entre a série atual e a série sazonalmente ajustada, obtida em função das estimativas dos fatores sazonais. Além disso, está bastante visível, que em $t = 30$, a magnitude do efeito tendência influencia no comportamento da série. Com efeito, a não-estacionariedade é provocada pela componente tendência, visto que os valores da série livre deste efeito pouco variam em relação à média. Esse fato é reforçado pelo correlograma da série livre da componente tendência, em que se aceita a hipótese nula para as autocorrelações rk ao longo dos *lags* $k \leq 12$. Assim, podemos extrapolar a série para um regime não-estacionário, quanto ao nível e inclinação – estacionariedade homogênea.

Figura 1 – Série dos Passageiros Transportados – Média dos Dias Úteis

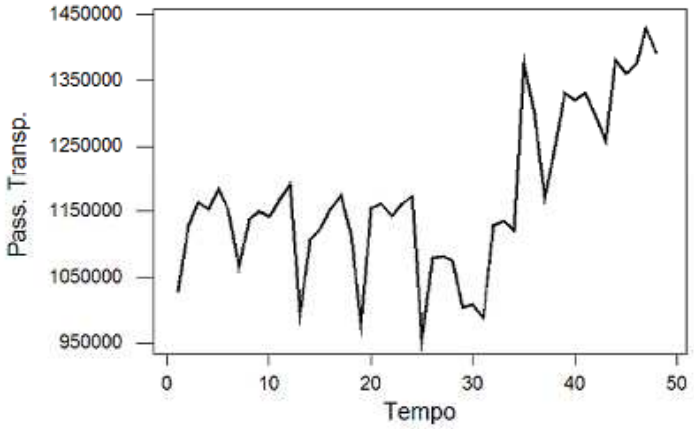


Figura 2 – Série Original – Livre do Efeito Tendência

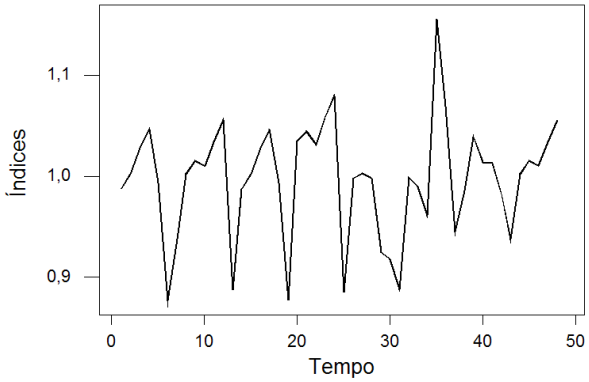
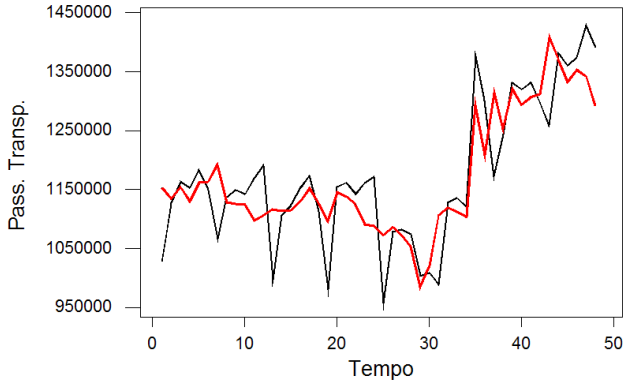


Figura 3 – Série Original e Série Sazonalmente Ajustada



2.1. Modelos individualizados para a série de passageiros transportados

Após as análises, tanto quantitativas quanto qualitativas, destacamos para a série X_t o Amortecimento Exponencial Sazonal Multiplicativo de Holt-Winters (MAKRIDAKIS et al., 1998; CORDEIRO, 2002), com os seguintes parâmetros: $\alpha = 0.4$ (nível), $\beta = 0.7$ (tendência) e $\gamma = 0.3$ (sazonalidade). O modelo teve um bom desempenho com base no EQM e na análise dos resíduos não constatamos nenhum comportamento sistemático. Em suma, com base na estatística de Box-Pierce aceita-se estatisticamente, ao nível de significancia de 5%, a hipótese nula para as autocorrelações residuais: $r_k(\varepsilon) = 0$ em todos os *lags* $k \leq 12$. Conforme fora exibido na Figura 1, apresentamos na Tabela 1, os quantitativos de passageiros transportados, média dos dias úteis, no período de janeiro de 2001 a dezembro de 2004. Podemos concluir que o crescimento acentuado dos valores da série a partir de junho de 2003 foi decorrente basicamente de ações coercitivas do Poder Público no combate ao transporte não-regulamentado (este argumento explica o processo heterocedástico – a variabilidade da série no ano de 2003).

Tabela 1 – Série dos Passageiros Transportados – Média dos Dias Úteis 01/2001 a 12/2004

MÊS	2001	2002	2003	2004
1	1.027.719	995.720	956.412	1.172.024
2	1.126.509	1.106.165	1.078.185	1.243.680
3	1.163.602	1.123.667	1.081.768	1.331.795
4	1.152.961	1.153.189	1.074.188	1.320.557
5	1.183.566	1.173.932	1.003.120	1.332.021
6	1.149.484	1.113.219	1.008.384	1.297.243
7	1.065.541	979.731	988.432	1.258.936
8	1.136.829	1.154.485	1.127.907	1.381.765
9	1.149.577	1.161.803	1.135.227	1.361.201
10	1.141.961	1.142.222	1.121.622	1.375.018
11	1.168.052	1.161.470	1.377.184	1.429.475
12	1.191.125	1.172.110	1.301.661	1.391.442

Fonte – Anuário Estatístico da EMTU/Recife.

Como a maioria dos procedimentos de análise estatística de séries temporais supõe que estas sejam estacionárias, faz-se necessária à adoção de transformações, principalmente para corrigir problemas provenientes da hetorocedasticidade e, consequentemente, a não-normalidade dos dados. Dessa maneira, elimina-se o comportamento instável dos dados em torno da média, melhorando significativamente as estimativas dos parâmetros do modelo. A transformação mais comum consiste em aplicar o operador de diferenças sucessivas Δ à série original X_t até se obter uma série estacionária.

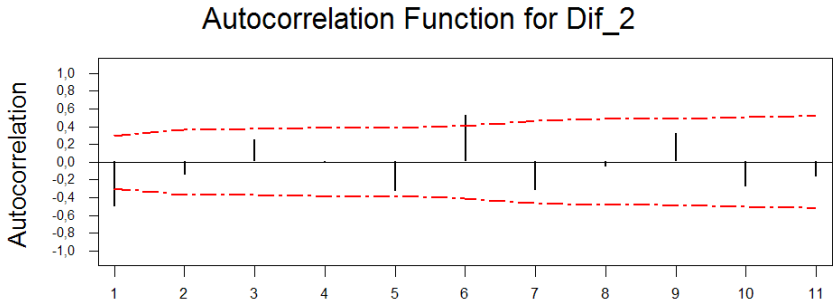
De um modo geral, a n-ésima diferença de X_t é dada por $\Delta^n X_t = \Delta\left[\Delta^{n-1} X_t\right]$. Sendo assim, aplicando esta transformação à serie original X_t consegue-se, senão, corrigir totalmente essa assimetria ou pelo menos minimizá-la, tornando-a mais próxima de uma distribuição Gaussiana. Com efeito, aceita-se, pelo teste não-paramétrico de Kolmogorov-Smirnov, a hipótese de normalidade para as duas diferenças. A escolha de $n = 2$ é decorrente unicamente do fator de redução da variância de 0, 8631. Após a transformação X_t em Y_t, determinamos o correlograma da série Y_t que está exibido na Figura 4. Após a análise do correlograma da série Y_t, nota-se claramente dois eventos:

Existência de uma componente periódica de período ou comprimento sazonal $L = 12$;

As autocorrelações estimadas apresentam picos decrescentes – exponenciais amortecidas por senóides nos *lags* múltiplos de L.

Isto indica, entre outras coisas, que a série transformada Y_t não precisa ser submetida a um filtro linear instável $\left(1 - B^{12}\right) = \Delta_{12}$. Torna-se, então, necessário a utilização desse filtro para eliminar o efeito da não-estacionariedade sazonal. Entretanto, como os fatores sazonais seguem um regime estacionário, reforça-se a tese de não ser necessário a utilização do operador de diferenças sucessivas de *lag12* e, dessa maneira, evita-se reduzir em 26% o tamanho da série.

Figura 4 – Correlograma da Série Transformada Y_t



Lag	Corr	T	Lag	Corr	T
1	-0,51	-3,44	8	-0,06	-0,24
2	-0,15	-0,83	9	0,32	1,33
3	0,25	1,35	10	-0,29	-1,14
4	-0,01	-0,07	11	-0,17	-0,66
5	-0,33	-1,73			
6	0,53	2,59			
7	-0,32	-1,40			

Com efeito, está evidente o decréscimo exponencial das autocorrelações, o que é um forte indício de que as componentes ARs do modelo são relevantes. A primeira modelagem para a série X_t foi um modelo ARIMA(1, 2, 0)$\times$(1, 0, 0)$_{12}$. O modelo funcionou satisfatoriamente, com relação à condição de estacionariedade, rejeitando-se a hipótese nula dos parâmetros, com base na estatística t. Através da estatística de Portmanteau – análise de adequação, aceita-se a hipótese nula para as autocorrelações dos resíduos; isto é, $r_k(\varepsilon)=0$ para todos os *lags* $k \leq 11$. Torna-se importante salientar que a inspeção gráfica da série é uma ferramenta poderosa para se fazer inferência e para uma melhor prospecção do modelo (CHATFIELD, 1989). Nestes termos, estabeleceu-se para o modelo inicial um teste sobre fixação, cujo objetivo é verificar se o número de parâmetros é adequado. O modelo melhora o ARIMA(1, 2, 0)$\times$(1, 0, 0)$_{12}$ pela inclusão do parâmetro referente à média móvel, passando para um ARIMA(1, 2, 1)$\times$(1, 0, 1)$_{12}$, cuja estrutura da equação mais parcimoniosa do modelo Box-Jenkins é apresentada pela equação (3):

$$(1+0{,}4851B)(1-0{,}99510B^{12})\Delta^2 X_t = (1-0{,}7143B)(1-0{,}6528B^{12})\varepsilon_t, \qquad (3)$$

Com efeito, esse modelo além de atender às condições de estacionariedade e invertibilidade apresenta todas as qualidades estatísticas do modelo anterior, porém com melhor desempenho, com base na análise dos resíduos, bem como na variância residual minimizada em torno de 30%.

3. COMBINAÇÃO DOS MODELOS DE PREVISÃO

Essa formulação foi proposta por Granger (1980) através da combinação linear de previsões pontuais, calculadas por meio de modelos com características distintas. Essa metodologia, embora tenha cerca de 27 anos, continua sendo bastante utilizada nos dias de hoje com base no seguinte argumento científico: “Se duas previsões pontuais, as melhores em termos de uma medida de eficiência (Erro Médio Absoluto ou Erro Médio Quadrático), são obtidas em modelagens distintas, então, se as previsões são combinadas linearmente, o resultado da previsão final será sempre melhor que as previsões não-combinadas ou individualizadas”.

No caso particular de dois modelos produzindo previsões pontuais $\widehat{X}_{t+1}^{(1)}$ e $\widehat{X}_{t+1}^{(2)}$, respectivamente, a um passo à frente, a previsão combinada $\widehat{X}_{t+1}^{(c)}$ é obtida pela seguinte combinação linear:

$$\widehat{X}_{t+1}^{(c)} = \widehat{p}\widehat{X}_{t+1}^{(1)} + (1-\widehat{p})\widehat{X}_{t+1}^{(2)} . \quad (4)$$

Aqui p é o peso e, no caso clássico, o seu cálculo é feito de modo a minimizar a variância do erro de previsão combinada. Logo, em virtude da equação (4) estar vinculada à minimização da variância residual, é interessante mencionar a abordagem bayesiana para o referido problema. Em Granger (1980), encontra-se uma descrição detalhada do enfoque bayesiano para a combinação de previsões calculadas pelos modelos especializados. Uma outra análise de natureza Bayesiana para a seleção sequencial do peso p referente à equação (4) é proposta por Bunn (1985), na qual o peso pertencente ao intervalo (0,1) é considerado como uma variável aleatória com distribuição beta, sequencialmente atualizada no tempo. O interesse maior na adoção dessa metodologia está na minimização da variância residual, vinculada à redução do número de observações não-explicadas. Souza (1982) apresenta, de forma bastante detalhada, o processo de combinação para a previsão, utilizando as distribuições beta e binomial. Em suma, estamos interessados em estimar p; o seu valor baseado numa amostra aleatória X_t para $t = 1,\ldots,N$ é aquele que minimiza o desvio padrão. Naturalmente, o *p* será uma estatística e, por isso, uma variável aleatória, pois seu valor dependerá da amostra ($X_t, t = 1,\ldots,N$).

4. DESENVOLVIMENTO DE UM MODELO DE PREVISÃO COMBINADA – SÉRIE DOS PASSAGEIROS TRANSPORTADOS

4.1. Descrição do processo de modelagem

O princípio do método de previsão através de combinação linear estabelece que uma combinação linear de modelos lineares, também, representa um modelo linear. Embora simples, o método é importante para construção de modelos mais adequados ao comportamento dos dados. O cerne do método é que um grande modelo linear poderá ser considerado como uma combinação linear de modelos lineares. Suponha que dois modelos distintos ($j = 1, 2$) geram os dados; como, por exemplo, o efeito tendência sobre imposto ao crescimento ou decrescimento da série de passageiros transportados. Logo, podemos construir as equações de estado (KALMAN; BUCK, 1961) para ambos os modelos, ou seja, pelas equações (5):

$$\begin{aligned} \underset{\sim t}{\widehat{X}}^{(j)} &= Z_t^{(j)} \underset{\sim t}{\widehat{b}}^{(j)} \quad \text{e} \\ \underset{\sim t}{b}^{(j)} &= G_t \underset{\sim t-1}{b}^{(j)} + \underset{\sim t}{w}^{(j)}, \end{aligned} \tag{5}$$

em que:

$\underset{\sim t}{X}^{(j)}$: vetor de observações do processo j no instante t de dimensão $N \times 1$;

$Z_t^{(j)}$: matriz de variáveis independentes conhecidas do processo j no instante t de dimensão $N \times k$;

$\underset{\sim t}{b}^{(j)}$: vetor de parâmetros do processo j no instante t de dimensão $k \times 1$;

G_t: matriz de transição do sistema não-conhecida no instante t de dimensão $k \times k$;

$\underset{\sim t}{w}^{(j)}$: vetor aleatório normal $(k \times k)$ de média nula e matriz de variância-covariância calculada por $E\left(\underset{\sim t}{w}\, \underset{\sim t}{w'}\right) = \sigma_w^2 I_k$ e conhecida no instante t.

Com base nos dois processos (j = 1, 2), observe-se que o ruído branco das observações não é especificado, pois $\widehat{X}_t^{(1)}$ e $\widehat{X}_t^{(2)}$ são valores idealizados e são observáveis somente em combinação por $\widehat{X}_t^{(c)}$, sendo que $\underset{\sim t}{\varepsilon}$ representa um vetor ruído branco que tem distribuição $N\left(\underset{\sim t}{0}, \sigma_\varepsilon^2 I_N\right)$. Na equação (6) tem-se uma modelagem combinada:

$$\underset{\sim t}{X}^{(c)} = \underset{\sim t}{\widehat{X}}^{(1)} + \underset{\sim t}{\widehat{X}}^{(2)} + \underset{\sim t}{\varepsilon}. \tag{6}$$

O sistema (5) pode ser descrito por meio de uma combinação linear conforme apresentado na equação (7). Dessa maneira, tem-se as equações do modelo numa formulação parametrizada, em que as observações da série temporal X_t se relacionam linearmente com b_t que, por sua vez, evoluem por meio de dinâmica de estado, ou seja, a melhor estimativa atual para o vetor b_t é com base no estado prévio do vetor b_{t-1}. Os modelos que seguem essa formulação denominam-se de Modelo Linear Dinâmico. Uma ampla abordagem desses modelos encontra-se devidamente descrita em Harrison e Stevens (1976), Otter (1984), Souza e Baratojo (1998) e Harvey (1989). Tem-se,

$$
\begin{aligned}
\underset{\sim t}{X}^{(c)} &= \begin{pmatrix} Z_t^{(1)} & Z_t^{(2)} \end{pmatrix} \begin{pmatrix} \hat{\underset{\sim t}{b}}^{(1)} \\ \hat{\underset{\sim t}{b}}^{(2)} \end{pmatrix} + \underset{\sim t}{\varepsilon}, \\
\begin{pmatrix} \underset{\sim t}{b}^{(1)} \\ \underset{\sim t}{b}^{(2)} \end{pmatrix} &= \begin{pmatrix} G_{11} & 0 \\ 0 & G_{22} \end{pmatrix} \begin{pmatrix} \underset{\sim t-1}{b}^{(1)} \\ \underset{\sim t-1}{b}^{(2)} \end{pmatrix} + \begin{pmatrix} \underset{\sim t}{w}^{(1)} \\ \underset{\sim t}{w}^{(2)} \end{pmatrix}.
\end{aligned}
\tag{7}
$$

Vale salientar que sua extensão para três ou mais modelos é imediata. Assim, no intuito de melhorar as previsões pontuais, atribuímos à série temporal de interesse uma subjetividade para minorar o efeito explosivo dos dados, decorrente dos fatores tendência e sazonalidade. Dessa maneira, deve-se reduzir a capacidade de influência das observações nos períodos onde os efeitos sobrepõem-se às observações.

Apresenta-se, na Seção 5.1, a estrutura canônica do modelo representado pelos pesos $p_i s$, bem como pelas estimativas dos modelos de melhor desempenho, até então analisados com base em estatísticas e medidas de qualidade. Algumas considerações para elaboração do processo de combinação foram introduzidas, objetivando o melhor entendimento da estrutura das equações do modelo e, ao mesmo tempo, facilitando os procedimentos computacionais inerentes à convergência das estimativas dos pesos $p_i s$. A seguir, apresentamos as premissas que nortearam a concepção metodológica bayesiana para a modelagem da série dos passageiros transportados. Mais detalhes sobre o assunto em Cordeiro & Cordeiro (2007).

O número médio esperado para a demanda r_i para cada ano i segue uma distribuição uniforme com estimativa otimista para os pesos p_i, isto é, $r_i \sim U(n_i \quad p_i)$, sendo n_i o número médio de passageiros transportados por dia útil em cada mês i;

As variáveis aleatórias $p_i s$ são independentes e podem seguir uma distribuição não-informativa do tipo beta $B(1, 1)$;

Os pesos p_i do modelo são estimados a partir de a) e essas estimativas são reavaliadas em cada instante de tempo $i = 1, 2, 3, 4$;

A distribuição *a posteriori* de p_i é calculada a partir da distribuição *a priori* utilizando o Teorema de Bayes;

Obtém-se o estimador de Bayes – máxima probabilidade e menor EQM – a partir da distribuição *a posteriori*, que pode ser qualquer medida de posição dessa distribuição, como, por exemplo, moda, média ou mediana.

Sendo as três medidas coincidentes, a função densidade de probabilidade da distribuição é simétrica. Logo, com a distribuição *a posteriori* do vetor de pesos p_i, em cada instante de tempo i, utiliza-se esta informação para inferir sobre as distribuições das futuras previsões. Dessa forma, as respostas das estimativas dos $p_i s$ aparentemente podem ser distribuições, seguindo delineamento da distribuição beta.

Com efeito, faz-se necessário testar se a distribuição *a posteriori* dos $p_i s$ pode ser representada por uma distribuição beta. O teste para avaliar a igualdade entre as médias das distribuições – hipótese H_0 – foi o teste não-paramétrico de Mann-Whitney (Teste U). O teste consiste em testar as médias de dois grupos não-pareados com grandes amostras (ROSNER, 1995). Verifica-se para o nível de significância de 5%, que os valores dos $Z_U s$ ficaram muito aquém de $z_c = 1{,}96$. Com efeito, aceita-se, estatisticamente, ao nível de significância de 5%, que a distribuição *a posteriori* para os $p_i s$ aparentemente tem o mesmo delineamento da distribuição da hipótese H_0 – distribuição B (1, 1).

5. DISCUSSÃO E ANÁLISE DOS RESULTADOS

Apresentamos na Tabela 2 um resumo das estatísticas referentes à distribuição *a posteriori* dos pesos p_i, onde se definem suas estimativas médias ao longo dos 48 meses. Temos que $\hat{p}_1$ é o valor da estimativa de p para o ano de 2001; $\hat{p}_2$ a estimativa de p para o ano de 2002; e, assim, sucessivamente, até o último ano, ou seja, 2004. Para o cálculo de $\hat{p}_1$, faz-se necessário usar o algoritmo de convergência com a finalidade de minimizar o erro-padrão e, consequentemente, o erro de Monte Carlo (MC erro). Uma forma de se avaliar adequadamente o peso de interesse da distribuição *a posteriori* é calcular o quociente entre o erro de Monte Carlo e o seu respectivo desvio padrão. Caso essa relação seja menor que 5%, então, pode-se afirmar que a estimativa $\hat{p}_1$ representa a verdadeira média da distribuição *a posteriori.* No caso, tem-se que este valor ficou aquém, algo em torno 2,0%. Ainda em relação à Tabela 2, nota-se a plena caracterização do estimador de Bayes, através da igualdade entre a média e a mediana, bem como o menor desvio padrão.

Tabela 2 – Sumário das Estatísticas Referentes à Distribuição *a posteriori* – Beta

peso	média	sd	MC Erro	2,5%	Mediana	97,5%
p_1	0,8787	3,044E-4	1,153E-5	0,8781	0,8787	0,8793
p_2	0,8930	3,067E-4	9,437E-6	0,8924	0,8930	0,8936
p_3	0,9054	2,705E-4	9,765E-6	0,9049	0,9054	0,9059
p_4	0,7549	3,801E-4	1,388E-5	0,7542	0,7550	0,7557

A convergência foi obtida em função de 1.000 iterações, evidenciando, aparentemente, que não existiu descontinuidade no processo de convergência, ou seja, não ocorreu *burn-in* no período de formação da cadeia dos $p_i s$, conforme revelado numa análise gráfica exaustiva realizada. Essa análise consistiu em gráficos de p_i *versus* o número de iterações que evidenciam que 1.000 iterações são suficientes para convergência da cadeia de formação de p_i. Quanto às autocorrelações, os seus valores seguem um regime estacionário, ou seja, são estatisticamente não-significativos em todas as iterações.

Os gráficos das distribuições dos pesos (distribuição *a posteriori*) são ilustrados na Figura 5 e representam, conforme o teste de Mann-Whitney, distribuições betas com os seguintes delineamentos:

$$p_1 \sim B(1011209, 139546),\ p_2 \sim B(911155, 109156),$$

$$p_3 \sim B(3346828, 349676) \text{ e } p_4 \sim B(966868, 314232).$$

Nota-se, também, na Figura 6, que os estimadores de p_1 decrescem com o crescimento dos dados observados mensalmente, ano a ano. Nesse caso, concluímos que os valores estimados, de certa forma, funcionaram como um filtro de amortecimento das componentes, de modo a minimizar os efeitos decorrentes de fatores externos, a exemplo do transporte clandestino, que imputou uma drástica redução na demanda de passageiros no período de 1998 a 2002. Com efeito, tendo em vista os intervalos de credibilidade determinados para cada um dos $p_i s$, fica totalmente excluída a possibilidade de nulidade para os referidos pesos. Os intervalos de credibilidade dos p_i na sequência $i = 1,2,\ldots$, são: (0, 878498, 0, 8788902); (0, 892696, 0, 893204); (0, 905220, 0, 905579); (0, 754648, 0, 755152).

Figura 5 – Distribuição *a posteriori* de *p*

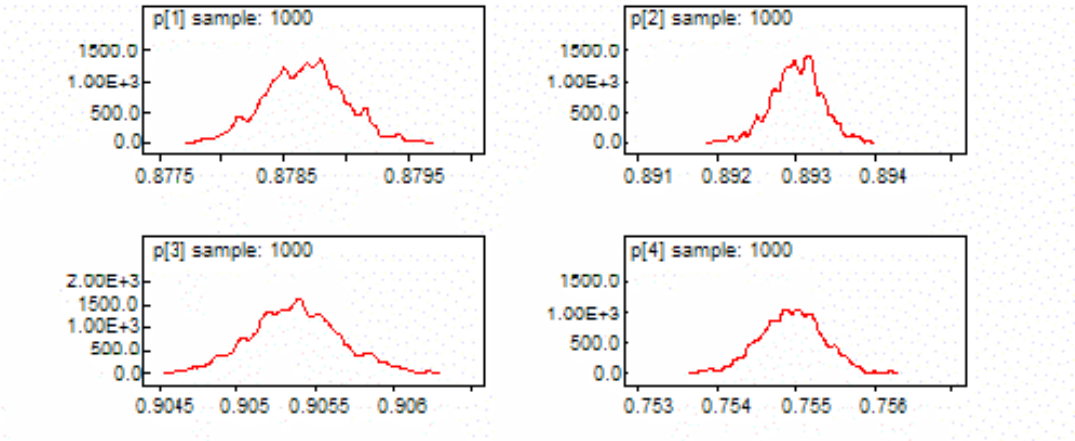
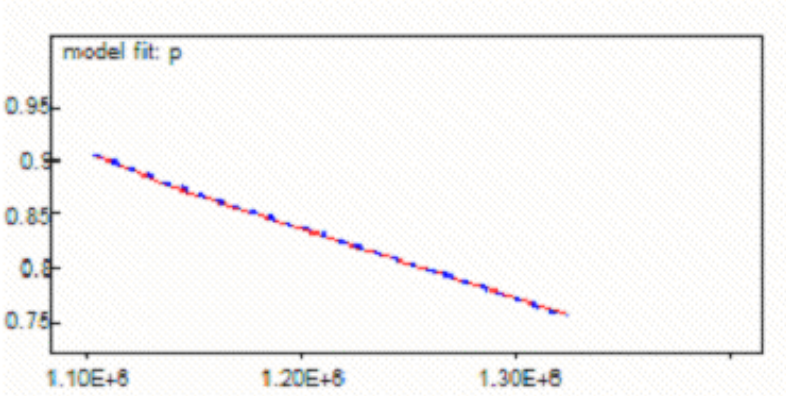


Figura 6 – Valores Ajustados p_i *versus* Valores Médios da Série Mensal dos Passageiros 2001-2004



5.1. Combinação das Previsões

Considere a equação de combinação linear, na qual o peso $p_i \in$ (0, 1), é uma variável aleatória com distribuição beta – sequencialmente atualizada no tempo. A equação de previ-são apresenta-se na estrutura matricial dada pela equação (8):

$$\widehat{X}_t^{(c)} = \left(\underset{\sim t}{\widehat{p}} : 1 - \underset{\sim}{} \underset{\sim t}{\widehat{p}} \right) \begin{pmatrix} \underset{\sim t}{\widehat{X}}^{(1)} \\ \underset{\sim t}{\widehat{X}}^{(2)} \end{pmatrix}, \qquad (8)$$

em que:

$\underset{\sim t}{\widehat{p}}$: representa o vetor das estimativas dos parâmetros p_i, para $i = 1,2,\ldots,t$, cuja dimensão é 1×48;

$\underset{\sim}{1}$: vetor unitário de dimensão 1×48;

$\underset{\sim t}{\widehat{X}}^{(1)}$: vetor das estimativas da série passageiros transportados calculado pelo processo 1;

$\underset{\sim t}{\widehat{X}}^{(2)}$: vetor das estimativas da série passageiros transportados calculado pelo processo 2.

A seguir, apresenta-se na Tabela 3, os resultados das previsões combinadas, usando os modelos sazonais, ou seja, SARIMA e Holt-Winters – Amortecimento Exponencial Sazonal Multiplicativo.

Tabela 3 – Previsões Pontuais com Modelagem Combinada – 3 Passos à Frente

Mês	SARIMA	Winters	Combinação
1	1.198.900	1.160.666	1.189.590
2	1.274.771	1.237.302	1.265.587
3	1.297.783	1.246.072	1.285.109

Na Tabela 4, além das análises dos resíduos dos três modelos, têm-se os resultados dos desempenhos com base no percentual de redução da variância não-explicada do modelo combinado em relação às variâncias individualizadas. Logo, em função do redutor da variância adicionado ao número de *outliers* – (observações não-explicadas), obtiveram-se o percentual de desempenho da modelagem combinada *versus* os processos individualizados – modelo de amortecimento exponencial sazonal – M_2, SARIMA – M_1 e combinação – $(M_1 \times M_2)^c$.

Tabela 4 – Análise Comparativa: Modelagem Combinada × Individualizada

Modelo	Análise dos Resíduos	Análise da Hipótese $r_k(\varepsilon)=0$	EQM	*Outliers*	Desempenho (%)
M_1	$Q = 3{,}1404$	Aceita-se	1, 87E + 9	3	–
M_2	$Q = 3{,}3702$	Aceita-se	2, 62E + 9	3	–
$(M_1 \times M_2)^c \times M_1$	$Q = 3{,}7522$	Aceita-se	2, 21E + 8	2	92
$(M_1 \times M_2)^c \times M_2$	$Q = 3{,}7522$	Aceita-se	2, 21E + 8	2	95

Notas: M_1: SARIMA M_2: WINTERS $(M_1 \times M_2)^c$: COMBINAÇÃO.

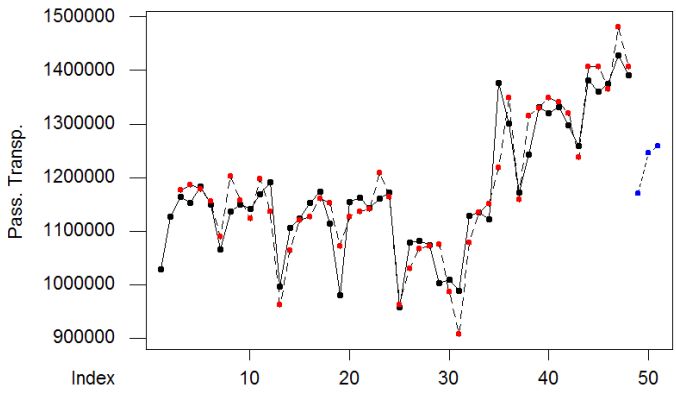
A Figura 7 mostra as estimativas e previsões – três passos à frente da modelagem combinada. Pelo correlograma, observa-se que é aceita a hipótese nula para as autocorrelações residuais; ou seja, estatisticamente, os resíduos não são correlacionados e, sendo assim, $r_k(\varepsilon)=0$. Verifica-se, também, que mesmo usando uma metodologia combinada, o correlograma é constituído por exponencial amortecida através de senóides. Portanto, a componente autorregressiva continua sendo preponderante no processo combinado.

6. CONCLUSÃO

A proposta de usar a combinação linear para previsões pontuais é factível e mais robusta em relação às previsões individuais, adotadas até o ano de 2004. Logo, a afirmação de Granger (1980) é de fato verossímil, ou seja, a previsão pontual combinada é sempre melhor que as previsões individuais. Em síntese, constata-se que todas as abordagens metodológicas atribuídas à série de interesse – Modelo Exponencial e Modelo de Box-Jenkins – revelaram que as estimativas dos passageiros transportados, média dos dias úteis, cresceram, quando comparadas com 2001 a 2003. Nos três modelos utilizados, tanto para as previsões pontuais como combinadas, os resultados das previsões são factíveis e, nesse caso, pode-se extrapolar o comportamento da série de passageiros transportados, média dos dias úteis – períodos típicos e atípicos – para um processo com poucas oscilações.

A série tende a se estabilizar em relação às componentes sazonais e tendência. Com efeito, essa estabilização deixaria de existir em casos extremos, como ocorreu no período anterior a julho de 2003 – período em que houve uma queda acentuada na demanda de passageiros, em virtude da grande concorrência do transporte “clandestino”. Logo, a reincidência de eventos desse porte, em qualquer ano, posterior a 2004, sem nenhuma dúvida, propiciará um desequilíbrio entre as componentes da série e, neste caso, uma queda brutal da demanda nos meses posteriores é muito provável de acontecer. Convém enfatizar que as séries que sofrem bruscas interferências, a componente sazonal contribui de forma importante na formação da série temporal, pois os valores das suas observações ficam muito sensíveis às variações em torno da sua média. Nesse caso, o efeito sazonal passa a funcionar como vetor contaminador redutor de longo prazo, notadamente, em torno da média da série temporal.

Figura 7 – Valores Ajustados e Previstos da Série dos Passageiros Transportados



Referências bibliográficas

BOWERMAN, B.L. (1987). ***Times Series Forecasting: unified concepts and computer implementation*****. Boston: Duxbury.**

BOX, G.E.P. E JENKINS, G.M. (1976). ***Time Series Analysis: Forecasting and Control*****. San Francisco: Holden Day.**

BUNN, D.W. (1985). Statistical Efficiency in the Linear Combination of Forecasting, ***Int. J. Forecast*****., 1, 151-163.**

CHATFIELD, C. (1989). ***The Analysis of Time Series: An Introducion.*** **London: Chapman and Hall.**

CORDEIRO, D.M. (2002). ***Séries temporais, análise quantitativa: teoria e aplicações*****, Série Ciência e Tecnologia, EDUPE.**

CORDEIRO, D.C., BARRETO, A.C., LIMA, D.C. (2006). Aplicação da Função Transação Demanda "versus" Tarifa como Modelo para Comprovação do Princípio da Elasticidade da Demanda. Comunicação Técnica. ***Segundo Congresso Luso Brasileiro para o Planejamento, Urbano, Regional, Sustentável, P LU RI S 2006*****, Escola de Engenharia, Universidade do Minho, Braga, Portugal.**

CORDEIRO, D.M. E CORDEIRO, G.M. (2007). Model of Combined Prevision: An Application of the Monthly Series of Dengue Notifications in the State of Pernambuco, ***Comm. Statist. Simul. and Comput*****., 36, 719-740.**

EMTU. (2004). Anuário Estatístico do Sistema de Transporte Público de Passageiros da RMR-PE.

GRANGER, C.W.J. (1980). ***Forecasting in business and economics*****. New York: Academic Press.**

HARISON, P.J., STEVENS, C.F. (1976). Structural Times Series Models and the Kalman Filter, ***J. R. Statist. Soc. B*****, 38, 205-207.**

HARVEY, A.C., Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge: Cambridge University Press, 1989.

KALMAN, R.E. E BUCK, R.S. (1961). New results in linear filtering and prediction theory, ***J. Basic Eng*****., 82D, 95-108.**

MAKRIDAKIS, S.G., WHELL, S.C., HYNDMAN, R.J. (1998). ***Forecasting: methods and applications*****. New York: John Wiley.**

MILONE, G. (2003). ***Estatística Geral e Aplicada*****. São Paulo: Thompson Learning.**

MORETTIN, P.A., TOLOI, C., (1986). ***Séries Temporais*****. São Paulo: Atual.**

OTTER, P.W. (1984). The discrete Kalman filter applied to linear regression models: statistical consideration and application. ***Statist. Neerlandica*****, 31, 41-56.**

ROSNER, B. (1995). ***Fundamentals of Biostatistics*****. Belmont: Duxbury Press.**

SOUZA, R.C. (1982). ***A Bayesian entropy approach to forecasting: the binomial – beta model*****, em Time Series Analysis: Theory and Practice (Editor O. D. Anderson), Amsterdam, 1, 475-486.**

SOUZA, R.C. E BARATOJO, S.C. (1988). Combinação Bayesiana de Previsões: Aplicação ao IGP-DI, ***Comunicação GSC-36/88*****, Grupo de Sistemas, DEE, PUC/RJ.**

Abstract

In this article we developed a combined prevision model to explain the behavior of the passengers transported by the Passenger System for Public Transportation of Passengers of the Metropolitan Region of Recife - PE (STPP/RMR). In this model we applied the Bunn's proposal (1985) where we combined linearly two punctual forecasts, the best in efficiency - average quadratic error (AQE) - to obtain the best previsions of the statistical models, analysis of residuals and the percentage of reduction of the non-explained variance for the modeling combined, in relation to the individual estimates. It is evidenced that in all the boardings imputed to the interest series, the forecasts had grown, when comparative to the observed ones before 2003 - period where it had a fall accented in the demand of passengers in virtue of the great incidence of the illegal transport.

Keywords: Average quadratic error, Box-Jenkins model, Combined prevision model, Holt-Winters model, Monte Carlo error.

Avaliação do Uso de Redes Bayesianas Discretas para Imputação de Dados

Ismenia Blavatsky [*]
Fabio Gagliardi Cozman[**]

Resumo

Redes bayesianas são estruturas que combinam distribuições de probabilidade e grafos. Apesar de as redes bayesianas terem surgido na década de 1980 e as primeiras tentativas em solucionar os problemas em bases de dados gerados a partir da não resposta datarem das décadas de 1930 e 1940, a utilização de estruturas deste tipo especificamente para imputação é bem recente: surgiram, em 2002, em institutos oficiais de estatística e, em 2003, no contexto de mineração de dados. Como existem questões teóricas e propriedades ainda não abordadas na literatura, além de poucas referências ao assunto, pretende-se tecer maiores considerações a respeito da aplicação de redes bayesianas como um método de imputação. Apresenta-se neste artigo resultados da aplicação de um algoritmo para a imputação de dados baseado em redes bayesianas discretas. Utilizam-se as avaliações sugeridas por Di Zio *et al*. (2004) e propõe-se um novo tipo de consistência, a consistência estrutural, que se relaciona à manutenção da estrutura da rede bayesiana em sua classe de equivalência após a imputação. Aplicações são feitas com o uso dos dados do Censo Demográfico Brasileiro em estrutura de redes simples. Redes construídas com estruturas simples mantêm-se totalmente em sua classe de equivalência mesmo após a inclusão de percentuais altos de dados imputados.

1. Introdução

A imputação de dados já é uma forma bem conhecida de tratamento da não resposta que tem por objetivo “completar" os espaços vazios, de maneira que se possa utilizar de ferramentas para dados completos na busca por características de uma dada população de interesse.

[*] Endereço para correspondência: IBGE–ENCE, Rua André Cavalcanti, nº. 106, sala 403, Santa Teresa, Rio de Janeiro – RJ, CEP 20231-050. *E-mail*: ismenia.magalhaes@ibge.gov.br.

[**] *E-mail*: fgcozman@usp.br.

Em geral, as avaliações sobre os métodos de imputação contemplam os resultados em apenas uma variável, referenciando–se a apenas um parâmetro de interesse. Textos publicados pelo projeto EUREDIT[1] relatam a experiência de que o melhor método para imputação varia de acordo com a aplicação e, dependendo do método, existem bons resultados garantidos para o caso univariado.

Para realizar imputação no contexto multivariado, em 2002 surgem as primeiras aplicações de redes bayesianas em estatísticas oficiais (THIBAUDEAU; WINKLER, 2002) (DI ZIO *et al*., 2004) e em 2003, em mineração de dados (HRUSCHKA–JR., 2003). Observa-se a necessidade de aprimoramento da teoria e discussão dos resultados até então obtidos.

As redes bayesianas surgiram na década de 1980 e em pouco mais de duas décadas de existência, tornaram–se populares e difundiram–se em aplicações nas mais diferentes áreas, como por exemplo, aplicações na área médica (HECKERMAN, 1988) (SAHEKI, 2005), mineração de dados (HRUSCHKA–JR, 2003), reconhecimento de padrões (FREY, 1998), em finanças (BINNER; KENDALL; CHEN, 2005). Este texto trata de grafos e redes bayesianas discretas e suas principais características que venham a compor o cenário para a sua utilização em imputação.

Como uma primeira avaliação do uso de redes bayesianas para imputação no contexto de estatísticas oficiais, o projeto EUREDIT apresenta um apêndice (DI ZIO; SCANU, 2003) ressaltando que não haveria comparação com os demais métodos estudados pelo projeto, por serem estes apenas avaliações preliminares e não terem sido conduzidos sob os protocolos de experimentos do conjunto de estudos. Scanu, Di Zio e Vicard (2003) apresentam, ainda com o trabalho em progresso, alguns aspectos computacionais de execução da imputação utilizando redes bayesianas e Di Zio, Scanu e Vicard (2003) relacionam os problemas em aberto e novas perspectivas do uso de redes bayesianas para imputação.

Posteriormente, Di Zio *et al*. (2004) publicam o primeiro artigo com alguns resultados do método, aplicado em parte de uma base de dados não identificados do Censo Demográfico 1991 da Inglaterra. Estes dados foram os mesmos utilizados no projeto EUREDIT, sendo simuladas não respostas do tipo MAR e MCAR[2] em 5% e 10% do total e os resultados foram comparados com procedimentos *hot–deck*.

Já na área de mineração de dados, Hruschka–Jr. (2003) propõe dois algoritmos para imputação bayesiana, dos quais um deles executa um teste baseado na estatística qui–quadrado para a ordenação de variáveis na construção da estrutura. Os resultados são

[1] Um projeto realizado para o desenvolvimento e avaliação de novos métodos para edição e imputação. Ver em: www.cs.york.ac.uk\euredit\

[2] MAR é a sigla para o mecanismo de perda aleatória denominado *Missing At Random* e MCAR é a sigla para o mecanismo de perda aleatória denominado *Missing Completely At Random* (Nota dos autores; maiores detalhes em Magalhães, 2007).

apresentados para problemas de classificação, agrupamento e seleção de atributos e todas as simulações foram realizadas em sete bancos extraídos de repositórios de dados e dois bancos de dados gerados aleatoriamente com caráter didático.

Apesar dos contextos diferentes para o surgimento dos métodos de imputação usando redes bayesianas, nenhum dos textos apresenta detalhamento teórico da metodologia em questão. Em todos eles sugere–se que estudos mais aprofundados são necessários.

O objetivo deste trabalho é, a partir de um algoritmo para imputação usando redes bayesianas, avaliar este método utilizando-se das medidas de consistência propostas por Di Zio *et al.* (2004). Além disso, é proposta uma nova possibilidade de consistência, a consistência estrutural, que avalia a propriedade de manutenção da estrutura da rede após a imputação. São apresentadas simulações e aplicações com dados do Censo Demográfico Brasileiro do ano de 2000.

Este texto divide-se como segue. Na Seção 2, tem–se uma rápida descrição da metodologia, o que inclui grafos e redes bayesianas discretas e também são apresentadas as redes bayesianas como ferramenta para imputação. Na Seção 3, são apresentados alguns resultados obtidos com simulações nos dados do Censo Demográfico Brasileiro 2000. Na Seção 4, é apresentado um relato das conclusões e futuros direcionamentos para pesquisas relacionadas.

2. Metodologia

2.1. Grafos e redes bayesianas

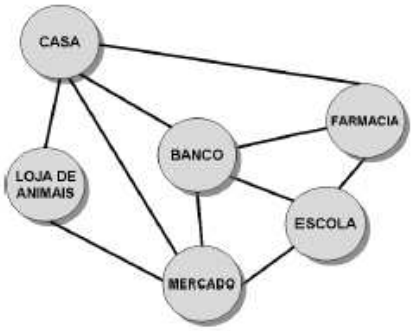
Das definições para redes bayesianas existentes na literatura, a que usamos neste texto considera dois elementos como seus componentes: o grafo e a distribuição de probabilidade associada a um conjunto de variáveis de interesse. A grande vantagem no uso deste tipo de estrutura está em conseguir representar a incerteza de forma graficamente compacta (PEARL, 1988) (CHARNIAK, 1991) através do uso de grafos.

Como componente fundamental da rede bayesiana, o grafo será de destaque quando se tratar mais adiante da consistência estrutural. Existem muitas aplicações de grafos na literatura e inicia-se esta seção com um exemplo dado na Figura 1. Esta figura representa a seguinte situação: uma dona de casa resolve mapear os locais por onde ela deve passar para realizar as suas tarefas em um determinado período do dia (ver Quadro 1 a seguir).

Quadro 1 - Tarefas e seus respectivos locais de realização para a situação hipotética dos possíveis caminhos de uma dona de casa

Tarefa	Local
1. limpar a casa	Casa
2. levar o cachorro para tomar vacina	Loja de animais
3. pagar as contas	Banco
4. fazer compras	Mercado
5. buscar as crianças na escola	Escola
6. comprar o remédio do marido	Farmácia

Figura 1 - Exemplo de grafo para situação hipotética dos possíveis caminhos de uma dona de casa



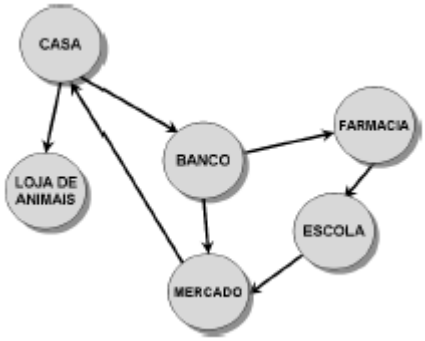
Podemos pensar em minimizar o percurso que a senhora percorre ou então em otimizar o tempo de realização das suas tarefas. Assim, como estes dois, diversos outros objetivos podem ser modelados com o uso de um grafo. Exemplos podem ser encontrados em variadas áreas, como em informática, com a transmissão de informações; na área de logística, com a modelagem da localização do armazenamento de mercadorias; em segurança pública, no fluxo de informações entre os sistemas de justiça; na arquitetura, com a definição de instalações elétricas domiciliares, etc. Principalmente para profissionais da matemática e da ciência da computação, a teoria dos grafos apresenta vários tipos de aplicações.

Definição 1 – O grafo G é um composto de um conjunto de vértices ou nós (V) conectados por um conjunto de arcos (A), que representam as ligações entre os nós.

Para construir uma rede bayesiana, necessitamos de um grafo orientado (ou dígrafo) que estabelecerá uma relação de dependência direta entre os vértices. No grafo orientado, as arestas são chamadas de arcos e a relação definida pelo conjunto A não é simétrica, existindo uma orientação na relação entre os nós. Imagine, por exemplo, que alguém que conheça a rotina da dona de casa do Exemplo 1 resolvesse prever a realização das suas tarefas. Sabe–se que, se a senhora estiver em casa, ela poderá ir ao banco ou à loja de animais com alguma probabilidade. Se ela estiver no banco, a tendência será de ela ir à farmácia comprar os remédios do marido ou ao mercado para realizar suas compras.

A sequência dada na Figura 2 define um caminho (ou cadeia), que é aquele realizado do nó casa até o nó mercado, da forma [casa, banco, farmácia, escola, mercado]. Um subcaminho é qualquer conjunto de nós e arcos que esteja inserido no caminho, por exemplo, o trajeto [banco, escola]. Um nó pai é aquele em que existe um arco partindo dele a qualquer outro nó no grafo e todo nó que recebe um arco a partir de um nó pai é chamado seu filho ou descendente. Nota–se na Figura 2 que o retorno da senhora à sua casa permite que haja um ciclo que se define como sendo o caminho de um vértice até si mesmo (NEAPOLITAN, 2004).

Figura 2 - Grafo que identifica um possível caminho percorrido pela dona de casa no exemplo hipotético



Definição 2 – Um grafo direcionado acíclico (GDA) é um grafo composto por nós e arcos no qual não existe a ocorrência de ciclos.

Para compor uma rede bayesiana devemos ter um grafo direcionado acíclico. Tendo um conjunto de variáveis aleatórias $\mathbf{X} = X_1,\ldots,X_k$ associadas aos nós do grafo e uma distribuição de probabilidade conjunta definida em $\mathbf{X}$, podemos construir a seguinte definição:

Definição 3 – Uma rede bayesiana é um par $B = (G,\theta)$ definido sobre um conjunto de variáveis aleatórias $\mathbf{X} = \{X_1, X_2,\ldots,X_k\}$, onde cada X_i corresponde a um nó, G é um grafo direcionado acíclico que será chamado de estrutura e θ é um conjunto de parâmetros que especificam distribuições de probabilidades condicionais que satisfaçam a condição de Markov:

$$P_\theta\left[X_i \middle| X_j, pa(X_i)\right] = P_\theta\left[X_i \middle| pa(X_i)\right], \quad (1)$$

onde $pa(X_i)$ é o conjunto de nós que são pais da variável X_i.

Em outras palavras, a condição de Markov para uma rede bayesiana diz que qualquer nó na rede é condicionalmente independente de seus não descendentes (X_j) condicionado a seus pais.

As redes bayesianas discretas são aquelas construídas de tal forma que a cada nó esteja associada uma variável aleatória do tipo discreto. Neste caso, a distribuição de probabilidade conjunta das variáveis aleatórias $\mathbf{X}$ é obtida pela fatoração:

$$P(\mathbf{X}) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \prod_{i=1}^{k} P(X_i = x_i \,|\, pa(X_i)), \qquad (2)$$

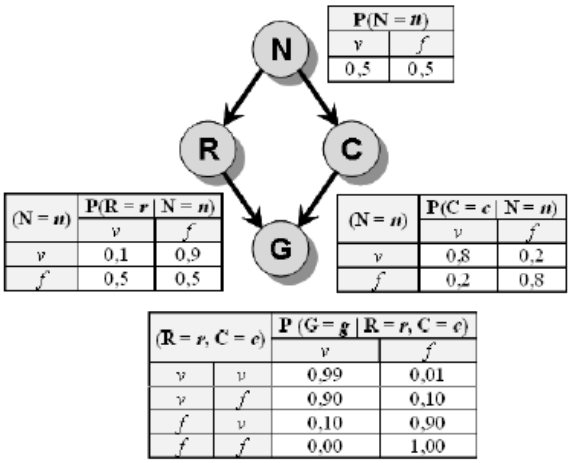
e, de acordo com a Definição 3, a possível falta de arcos entre as variáveis X_i e X_j no grafo G significa a independência entre as mesmas.

No Exemplo 2, ilustrado pela Figura 3 (MURPHY, 1998), observa–se uma rede bayesiana discreta. A rede para este exemplo é formada pela estrutura gráfica (G) e pelo conjunto de probabilidades associadas (θ) explicitadas nas tabelas. Esta rede especifica o relacionamento entre quatro variáveis aleatórias discretas, cada uma contendo dois valores possíveis (v = verdadeiro e f = falso). As variáveis são: a existência de nuvens no céu (N), o regador ligado (R), a ocorrência de chuva (C) e a grama molhada (G). Neste caso, podemos quantificar incertezas a partir da dependência entre a grama estar molhada e o regador estar ligado ou à ocorrência de chuva.

Observamos na Figura 3, uma vez condicionado em R e C, que as variáveis N e G são independentes, ou seja, a grama estar molhada não depende de haver nuvens no céu condicionado às variáveis ocorrência de chuva e o regador estar ligado. Os relacionamentos de causa e efeito entre as variáveis podem gerar as chamadas classes de equivalência que representam as mesmas relações de independência condicional em redes bayesianas. Neste mesmo exemplo, temos que {N} é o pai de R e C e pa (N) = $\varnothing$. Na mesma rede, pa (G) = {R, C} e P [G, N| pa (G)] = P [G| pa (G)], pois G e N são independentes se condicionado aos pais de G (neste caso, R e C).

A Definição 3 leva a uma propriedade do GDA chamada d–separação. Esta propriedade será importante quando se tratar mais adiante da consistência estrutural, proposta neste texto para a avaliação do método de imputação. Se $B = (G, \theta)$ é uma rede bayesiana que satisfaz a condição de Markov, então é possível identificar os nós condicionalmente independentes a partir da d–separação. Como exemplo, na rede da Figura 3, N e G são d–separados pelo conjunto de nós {R, C}.

Figura 3 - Rede para a grama molhada citada no Exemplo 2 (MURPHY, 1998)



Para definir formalmente a d–separação, será necessária a utilização de alguns conceitos a mais sobre o relacionamento entre os nós em um grafo. Sejam X, Y e Z três nós que formam um subcaminho no grafo. Diz–se que (NEAPOLITAN, 2004):

(a). Se X $\rightarrow$ Z $\rightarrow$ Y, temos um relacionamento do tipo topo–para–final;

(b). Se X $\leftarrow$ Z $\rightarrow$ Y, temos um relacionamento do tipo final–para–final;

(c). Se X $\rightarrow$ Z $\leftarrow$ Y, temos um relacionamento topo–para–topo[3].

Nos casos (a) e (b), ao se observar o nó Z, vemos que ele bloqueia o caminho entre os nós X e Y. Em (c), se o nó Z não for observado e nenhum dos seus descendentes for observado, ele bloqueará todos os caminhos na rede que pudessem seguir a partir de Z. Maiores detalhes em Neapolitan (2004). Com estas informações podemos definir formalmente d–separação conforme a Definição 4.

Definição 4 – Seja G =(V,A) um grafo direcionado acíclico (GDA), e E um subconjunto de V, com X e Y nós distintos em V–E. Dizemos que X e Y são d–separados pelo conjunto E em G se todo caminho entre X e Y é bloqueado por E.

Pode ser construído para cada GDA um conjunto de equivalência, chamado de conjunto de equivalência de Markov, no sentido de que as redes preservariam as mesmas independências condicionais a partir das *d*–separações identificadas.

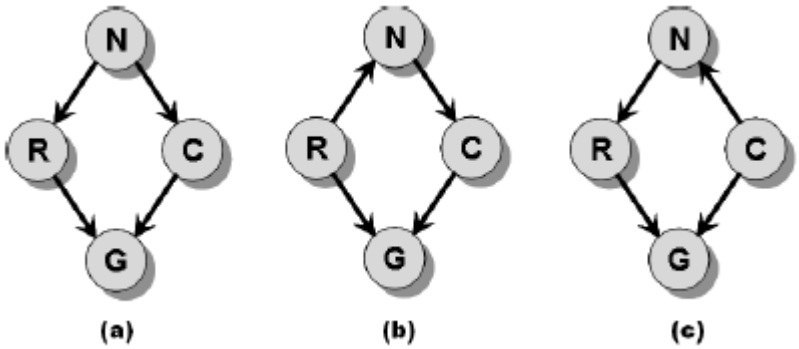
Segundo Neapolitan (2004), dois grafos são equivalentes se e somente se, baseado na condição de Markov, eles representam as mesmas independências condicionais. Em outras palavras, dois grafos são Markov–equivalentes se tiverem as mesmas ligações entre os nós

[3] Esta tradução foi realizada de forma livre pelos autores e se deu a partir de Neapolitan (2004).

sem considerar suas direções, além de preservar os mesmos relacionamentos topo–para–topo no grafo. Outras condições existem, mas para o desenvolvimento deste trabalho estas serão suficientes. O leitor que tiver necessidade de maior detalhamento teórico pode consultar Neapolitan (2004).

Do Exemplo 2, as Figuras 4(b) e 4(c), a seguir, mostram a classe de equivalência do GDA da rede original (reproduzida na Figura 4(a)). Nenhum outro grafo é equivalente a eles.

Figura 4 - (a) Rede original (de Murphy, 1998) e (b) e (c) Elementos da classe de equivalência da rede original para o Exemplo 2



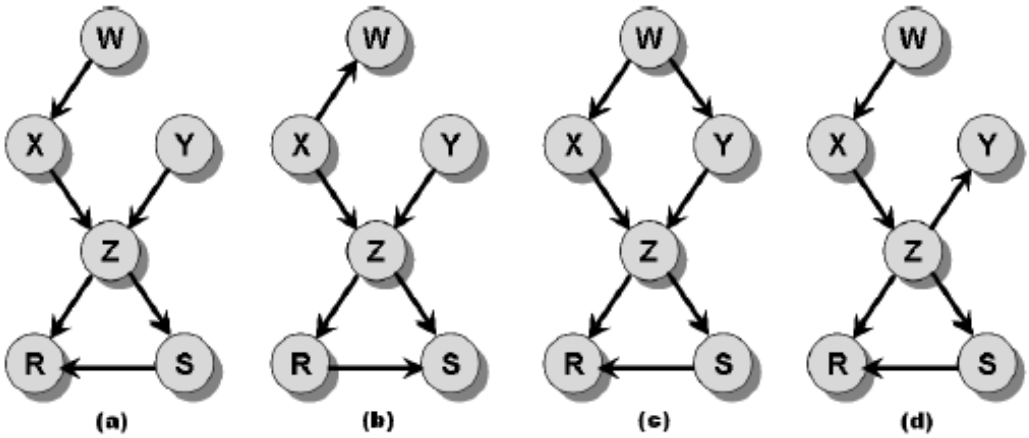
Definição 5 – Sejam $G_1 = (V, A_1)$ e $G_2 = (V, A_2)$ dois grafos direcionados acíclicos contendo o mesmo conjunto de nós V e A_1, A_2 seus respectivos conjuntos de arcos. G_1 e G_2 são ditos da mesma classe de equivalência se, para todos os subconjuntos mutuamente disjuntos B, C, $D \subseteq V$, B e C são d–separados por D em G_1 e B e C são d–separados por D em G_2. Isto implica que as mesmas independências condicionais entre B e C são observadas nos grafos G_1 e G_2.

O exemplo na Figura 5, a seguir, é reproduzido de Neapolitan (2004) e exemplifica a identificação de elementos da classe de equivalência do grafo da Figura 5(a). Os grafos na Figura 5(a) e na Figura 5(b) são equivalentes, pois apresentam o mesmo conjunto de arestas, sem considerar suas direções e o mesmo relacionamento topo–para–topo ($X \rightarrow Z \leftarrow Y$) no nó de bloqueio Z está presente nos dois grafos. Considerando agora o conjunto de Figuras 5(a) e 5(b), denotado por {5(a), 5(b)}, e avaliando as Figuras 5(c) e 5(d), temos que:

- As Figuras 5(c) e {5(a), 5(b)} não são equivalentes por causa da presença do arco em $W \rightarrow Y$ no grafo em 5(c), o que implica em conjuntos diferentes de arestas entre os dois grafos;

– As Figuras 5(d) e {5(a), 5(b)} não são equivalentes por não manterem o relacionamento topo–para–topo no nó de bloqueio Z no grafo em 5(d).

Figura 5 - Exemplo de identificação de elementos de uma mesma classe de equivalência (reproduzido de Neapolitan (2004))



Comparando agora as Figuras 5(c) e 5(d), elas não são equivalentes entre si por causa da existência do arco W $\rightarrow$ Y em 5(c) e pelo não relacionamento topo–para–topo em 5(d).

Uma vez definida a estrutura da rede e seu conjunto de parâmetros θ, pode–se calcular uma função que atribui um valor para cada GDA (ou a um determinado padrão de GDAs) baseado nos dados. Esta quantidade atribuída ao grafo é denominada de função escore e seu cálculo depende da distribuição de probabilidade associada às variáveis aleatórias do problema em questão. A função escore é uma outra possibilidade de se encontrar redes bayesianas que estejam na mesma classe de equivalência de uma rede de interesse. Maiores informações podem ser encontradas no texto Magalhães (2007).

2.2. Algoritmo para o uso de redes bayesianas como ferramenta para imputação

Considere que já exista uma rede bayesiana $B = (G, \theta)$ ajustada de pesquisas anteriores ou construída a partir de uma base de dados contendo não resposta em pelo menos uma das variáveis $X_1, X_2, \ldots, X_k$ associadas aos nós da rede e onde se conhecem os parâmetros θ da sua distribuição. É importante observar que, para justificar o uso da rede bayesiana para imputação, esta deve manter as relações de causa e efeito das variáveis consideradas, ou seja, os critérios para construir as relações de independência condicional devem estar identificados na rede. Registramos que o objetivo de usar uma rede bayesiana para imputação está em preservar o relacionamento existente entre as variáveis, então o

ajuste inicial (ou rede original) será fundamental para esta verificação. Se esta rede contiver algum problema de ajuste, a imputação ou qualquer procedimento inferencial a partir dela conterá problemas também[4].

Dispomos de n observações das quais n^*, $n^* \leq n$, contêm pelo menos um item faltante. Podem ser divididas as $X_i, i = 1, 2, \ldots, k$, variáveis em grupos disjuntos conforme identificados na estrutura G ajustada (a exemplo do que é conduzido por Di Zio *et al*. (2004), só que em sua proposta a identificação é feita antes do ajuste da rede): o subconjunto P_0 contém as variáveis órfãs (sem pais) em G; o subconjunto P_1 contém aquelas com pais somente em P_0; já o subconjunto P_2 contém pais em $P_0 \bigcup P_1$, e assim sucessivamente até o último subconjunto $j = v - 1$, que contém os pais das variáveis em v.

Se a variável estiver em P_0, pode ser imputado um valor (ou classe) de acordo com a distribuição marginal da variável em questão. Se a variável estiver em P_1, a imputação será dada em função da distribuição condicionada em P_0, ou seja, é gerado um valor da distribuição $P(X_i \mid pa(X_i))$, sabendo que $X_i \subseteq P_1$ e $pa(X_i) \subseteq P_0$. Este procedimento prossegue até que a imputação nas variáveis do subconjunto v, que têm pais no conjunto $\bigcup_{i=1}^{j+1} P_i$, sejam imputadas. O Quadro 2 resume este algoritmo.

Quadro 2 – Algoritmo proposto para uso de redes bayesianas discretas em imputação

Algoritmo proposto
Entrada: Rede bayesiana ajustada e Base de dados com valores faltantes. Saída: Base de dados imputada.
1. Identifique os subconjuntos $P_0, P_1, \ldots, P_v$ na rede bayesiana de entrada; 2. Defina uma ordem de imputação em cada subconjunto de acordo com algum critério; 3. Para cada subconjunto P_j, j =1, 2, ..., v, faça: 3.1. Se a variável pertencer ao primeiro conjunto (ou conjunto das variáveis sem pais), gere aleatoriamente um dado a ser imputado de acordo com a distribuição marginal da mesma; 3.2. Se não, gere um dado a ser imputado de acordo com a estrutura da rede ajustada em 2.; 4. Retorne à base imputada.

[4] Em algum momento seria interessante fazer uma avaliação da rede inicialmente ajustada, ou avaliar se esta descreve bem o fenômeno. Este não foi o objetivo deste artigo.

A ideia de se particionar as variáveis em subconjuntos de acordo com a estrutura G da rede bayesiana para imputação tem como principal objetivo considerar que, o que acontecer para um determinado subconjunto possa influenciar na imputação das variáveis do subconjunto seguinte. As imputações que forem conduzidas por este método nas variáveis em P_0 serão equivalentes àquelas produzidas pela imputação geral aleatória, pois não estarão condicionadas a nenhuma variável aleatória na rede. Já as imputações realizadas em $P_1, P_2, \ldots$, serão equivalentes às imputações aleatórias dentro de classes, pois são obtidas a partir das distribuições de probabilidade condicionadas em seus pais. As classes de imputação, porém, são delimitadas pela rede $B = (G, \theta)$.

A avaliação do método também se torna facilitada, pois uma imputação em P_1 depende de menor número de parâmetros que aquelas em P_2, e esse pode ser um fator importante para se considerar as consistências em subconjuntos distintos de variáveis.

Um ponto limitante do algoritmo proposto no Quadro 2 está em não permitir alterar os valores dos parâmetros da rede, o que garante que todos os itens imputados em uma mesma variável apresentam a mesma distribuição descrita inicialmente. Isso significa que não se considera uma atualização de θ. No caso de a rede de entrada no algoritmo ser ajustada com a presença de itens faltantes em alguma variável, os parâmetros em cada nó seriam atualizados antes da fase de imputação.

Este método de imputação permite o tratamento de forma simples dos chamados zeros estruturais. Os zeros estruturais são aqueles casos em que existem blocos de perguntas que não se aplicam a determinados respondentes no questionário (como, por exemplo, questões relacionadas ao trabalho para pessoas não inseridas no mercado). A rede bayesiana funciona aqui como um dispositivo que, com probabilidades conhecidas em θ, permite gerar ocorrências para variáveis a partir de θ. Se um zero estrutural for considerado como uma classe da variável aleatória que a contém, então associa–se a este uma probabilidade nula (ou quase nula) se condicionada a uma classe que a determine (no caso do exemplo, se a pessoa não estiver no mercado, então a probabilidade de pertencer a uma classe zero estrutural às questões do trabalho é unitária).

Di Zio *et al*. (2004) defendem a aplicação das redes bayesianas para o processo de identificação dos zeros estruturais antes de considerar a imputação. Juntamente com este passo, os autores sugerem que seja incorporado o processo de crítica dos dados, que identificaria algum valor incoerente que passaria a ser um dado faltante. Uma vez realizado o processo de crítica, os itens faltantes seriam previamente classificados em válidos, de acordo

com suas características, ou em não respostas a serem imputadas, evitando que imputações fossem realizadas em locais indevidos.

O algoritmo proposto no Quadro 2 permitiria manter a distribuição multivariada dos dados representada pela rede bayesiana de entrada. Só que no algoritmo proposto nesta seção, as relações existentes entre as variáveis são aquelas que determinam sua probabilidade de imputação e tornam estas informações mais fidedignas à realidade dos dados. O ponto limitante está na necessidade de algum conhecimento prévio do comportamento dos dados. Isso implica que a utilização da experiência de especialistas na formação das redes e obtenção dos parâmetros é imprescindível para a funcionalidade do método.

2.3. Avaliação do uso da rede discreta para imputação

Os estudos atuais sobre a imputação a partir de redes bayesianas levam à conclusão de que estas preservam a distribuição multivariada dos dados por causa da utilização da estrutura de independência na obtenção de um dado a ser imputado, embora esta afirmação ainda esteja sem confirmação teórica e existam poucos experimentos neste sentido. Devido à especificidade de cada base de dados e de cada problema, os resultados de uma avaliação comparativa a outros métodos de imputação dependeriam das variáveis aleatórias e percentuais de não resposta em cada uma delas, número de observações, grau de dependência entre as variáveis e, especificamente no caso das redes bayesianas, da ordenação em que as variáveis se encontram na base de dados.

Di Zio *et al*. (2004) sugerem três parâmetros para a avaliação do uso de redes bayesianas discretas para imputação em aspectos distintos:

- A consistência da base de dados, que avalia a preservação dos microdados;
- A consistência lógica, para avaliar a preservação das restrições lógicas;
- A consistência estatística, com a preservação dos parâmetros e quantidades associadas à base de dados.

A primeira forma de avaliação diz respeito à preservação dos microdados que seria a propriedade de recuperar exatamente a informação perdida. A preservação das restrições lógicas está na ideia de que o método resgata a plausibilidade dos valores imputados com respeito às restrições lógicas. Isso pode ser feito de forma direta para variáveis discretas e um bom exemplo está na variável faixa etária. Não seria muito razoável que uma faixa etária imputada de “menos de dez anos de idade” fosse feita em um estado civil observado “casado”. A consistência estatística analisa a preservação dos parâmetros da distribuição conjunta a

partir de índices descritivos simples. Neste texto estendemos a consistência estatística para parâmetros que possam mensurar o relacionamento entre as variáveis aleatórias, como é o caso da medida de associação entre estas.

São avaliadas neste trabalho três medidas de consistência sugeridas em Di Zio *et al*. (2004): da base de dados, lógica e estatística. É proposta mais um tipo de consistência, esta com relação à preservação da estrutura da rede bayesiana após a imputação. A ideia destas quatro maneiras de verificação está em identificar se o conhecimento *a priori* com respeito à base de dados se mantém e se não existem alterações bruscas, principalmente nas consistências estrutural e estatística, que resumiriam os principais interesses dos mantenedores da base de dados e dos usuários, respectivamente. Serão descritas, a seguir, as consistências estrutural e estatística. Maior detalhamento sobre estas medidas e as demais podem ser encontradas em Magalhães (2007).

2.3.1. Consistência estrutural

A medida de consistência estrutural proposta neste artigo é a propriedade que o método tem de manter a estrutura construída a partir dos dados imputados na mesma classe de equivalência da rede original. A estrutura da rede é necessária para a representação do relacionamento de independência condicional entre as variáveis, portanto, a verificação de sua manutenção após a imputação é de fundamental importância para a confiabilidade do método. A consistência estrutural pode ser avaliada de duas formas: a partir do grafo (ou da estrutura propriamente dita) ou a partir de uma medida resumo do ajuste da estrutura, por exemplo, a função escore calculada para a rede. No caso da consistência estrutural pela avaliação do grafo, partimos da observação após a imputação, dos mesmos nós e mesma sequência de arestas (independentes da direção) e relações topo–para–topo entre a rede original $B = (G, \theta)$ e a rede ajustada após a imputação $\tilde{B} = (\tilde{G}, \theta)$, conforme a equivalência dos grafos descrita anteriormente.

Nesta forma de verificação temos como hipótese que uma imputação que mantém o relacionamento de independência nos dados é aquela que consegue reproduzir a estrutura G original dos dados após a construção da estrutura $\tilde{G}$ a partir dos dados imputados. Como parâmetros, perseguimos alguns aspectos que permitem identificar se o grafo gerado a partir da base imputada pertence à mesma classe de equivalência da rede original.

Nesse tipo de consistência existem algumas considerações a serem feitas sobre a forma de ajuste e as características dos dados que podem influenciar nos resultados. A primeira a ser considerada refere–se ao número de nós, ou variáveis aleatórias. Quanto maior

o número de nós, maior a possibilidade de existência de arestas entre eles, o que pode levar a um maior número de redes na classe de equivalência da rede original. Certamente, uma classe de equivalência de uma rede para uma base que possui três variáveis aleatórias, e que pode apresentar no máximo três arestas em uma rede construída a partir delas, é muito menos sensível do que uma classe de equivalência para uma estrutura que contenha quatro variáveis e que pode conter no máximo seis arestas.

O mesmo ocorre com a quantidade de observações disponíveis e faltantes na base de dados. Parece intuitivo que, quanto maior o número de observações da base de dados, mais estável é a rede bayesiana $B = (G, \theta)$ com relação à estrutura e aos parâmetros associados. Já quanto maior o percentual de dados faltantes, maior a incerteza em se manter a estrutura dos dados originais. Isso porque, mesmo em se obtendo os mesmos parâmetros para as distribuições univariadas, as imputações em registros individuais podem ocasionar alterações nas relações de independência condicional entre as variáveis.

Para se construir uma medida de avaliação da consistência estrutural, consideramos o caso em que o número de nós da estrutura G de entrada é o mesmo a ser trabalhado na construção da rede após a imputação. Mesmo que não existam itens faltantes em alguma variável, é imprescindível que a estrutura $\tilde{G}$ tenha as mesmas variáveis da estrutura original para que esteja em sua classe de equivalência. Além disso,

- para que a estrutura $\tilde{G}$ esteja na mesma classe de equivalência de G é necessário que exista o mesmo número de arestas entre a rede original e a rede ajustada a partir dos dados imputados e;
- deve ser considerado que a estrutura construída a partir dos dados imputados mantenha os mesmos relacionamentos topo–para–topo em nós de bloqueio.

Um número de arestas menor que o da estrutura original implica que o método modifica o relacionamento inicial na direção da independência entre as variáveis e um número de arestas maior que o da estrutura original implica em modificar o relacionamento inicial na direção de maior dependência entre as variáveis.

Para se obter uma indicação do que ocorre neste sentido, define–se um conjunto de informações que expresse estas quantidades. Seja B uma variável aleatória que especifica os eventos b_1, se o número de arestas do grafo após a imputação é menor que o número de arestas no grafo original; b_2, se o número de arestas do grafo após a imputação é igual ao número de arestas do grafo original; e b_3, se o número de arestas do grafo após a imputação é maior que o número de arestas do grafo original.

A variável B está então definida como

$$B = \begin{cases} b_1, & \text{se} \quad n(\tilde{G}) < n(G) \\ b_2, & \text{se} \quad n(\tilde{G}) = n(G) \\ b_3, & \text{se} \quad n(\tilde{G}) > n(G) \end{cases}, \tag{3}$$

onde $n(G)$ representa o número de arestas do grafo G originalmente ajustado e $n(\tilde{G})$ representa o número de arestas em $\tilde{G}$.

As redes pertencentes à mesma classe de equivalência de G estão contidas no conjunto $\{B = b_2\}$, pois esta é uma condição para se delimitar os elementos de uma mesma classe de equivalência de um grafo. O procedimento seguinte consiste em verificar as posições dos arcos independente de suas direções. Para cada estrutura pertencente ao conjunto de possíveis redes equivalentes, defina a variável aleatória C, que assume os valores c_1, se o sentido da aresta i na estrutura original é o mesmo da aresta i na estrutura após a base imputada e c_2, se o sentido da aresta i na estrutura original é diferente da aresta i na estrutura após a base imputada, ou seja,

$$C = \begin{cases} c_1, & \text{se} \quad l_i(\tilde{G}) = l_i(G) \\ c_2, & \text{se} \quad l_i(\tilde{G}) \neq l_i(G) \end{cases}, \qquad i = 1, \ldots, n(\tilde{G}), \tag{4}$$

onde $l_i(G)$ é o sentido da aresta i na estrutura construída a partir da base original e $l_i(\tilde{G})$ é o sentido da aresta i na estrutura construída a partir da base imputada.

Se for válida a suposição de que a imputação a partir de redes bayesianas mantém a distribuição multivariada dos dados, então a consistência estrutural a partir do grafo deve ser mantida após a construção da rede com a base imputada. Esta suposição será verificada neste texto através de simulação. Sugerimos que seja conduzido um estudo futuro sobre as propriedades teóricas destas variáveis.

A segunda forma de verificação da consistência estrutural é desenvolvida com base em uma quantidade calculada sob a rede original e o equivalente com a rede construída após a imputação. Isso pode ser feito a partir do valor de qualquer função, por exemplo o escore da rede, que é usado como métrica para o aprendizado da estrutura G e compõe a base de uma estratégia de busca heurística no seu ajuste. Segundo Bottcher e Dethlefsen (2003), esta quantidade representa as independências condicionais entre as variáveis aleatórias a partir da probabilidade relativa:

$$S(G) = P(G, d) = P(d|G)P(G), \tag{5}$$

onde d representa um conjunto de dados específico para o ajuste da rede bayesiana e $S(G)$ é denominada de função escore para a rede construída.

Neapolitan (2004) mostra que a função escore é consistente para a classe de modelos de uma distribuição representada pela rede bayesiana que identifica o mapa de independências ótimo dos parâmetros de uma base de dados.

A hipótese levantada neste trabalho é de que, se a rede bayesiana para a imputação mantém a relação entre as variáveis observada na rede original, então os valores da função escore devem se manter os mesmos com a rede ajustada após a imputação. Para avaliar esta afirmação deve–se construir uma medida simples de forma que:

$$\tau = S(G) - S(\tilde{G}), \tag{6}$$

onde $S(G)$ é o escore da rede original e $S(\tilde{G})$ é o escore da rede calculado após a imputação. O que se espera é que, para duas redes da mesma classe de equivalência, esta diferença não seja muito afastada de um determinado ε que varia conforme características próprias da rede como número de nós, de arestas, de classes em cada variável, de observações, etc.

Torna–se interessante aqui a obtenção das propriedades de τ para verificar a sua composição no contexto de imputação, ou seja, avaliar os valores de

$$E[\tau|d] = E\{[S(G) - S(\tilde{G})]|d\} \quad \text{e} \quad Var[\tau|d] = Var\{[S(G) - S(\tilde{G})]|d\}. \tag{7}$$

Não se sabe ao certo qual a variação na quantidade de τ permite afirmar o quanto duas redes estão em uma mesma classe de equivalência a partir dos valores da função escore. Este é um trabalho que ainda se encontra em aberto e não será abordado neste texto.

2.3.2. Consistência estatística

Com relação à consistência estatística, existem diversas formas de fazê–la. Di Zio *et al*. (2004) avaliam este tipo de consistência do ponto de vista dos parâmetros da rede em cada nó, mas aqui esta consistência é tratada de forma mais abrangente. Além dos parâmetros da rede, também são consideradas medidas importantes que decorrem diretamente do relacionamento entre as variáveis, como medidas de associação e parâmetros de modelos ajustados após os dados imputados.

De uma maneira geral, a consistência estatística após a imputação é a propriedade que o método tem de manter os parâmetros da rede, características das distribuições univariadas e multivariadas e quantidades de interesse.

Para avaliar os parâmetros da rede, uma medida simples proposta no texto de Di Zio *et al*. (2004) está em considerar uma distância entre as frequências relativas da classe z antes e depois da imputação. Seja

$$\Delta = \frac{1}{2}\sum_{z}\left| f_z - \tilde{f}_z \right|, \tag{8}$$

onde f_z denota a frequência relativa da categoria z de X no conjunto de dados reais dos n^* itens faltantes. Da mesma forma, $\tilde{f}_z$ representa a frequência relativa depois da imputação. O valor de Δ será um resultado entre 0 (que significa a igualdade nas duas distribuições) e 1, e pode ser extendido facilmente para o caso multivariado onde, neste caso, os autores observam que se podem avaliar diferentes estratégias de imputação para o mesmo número de variáveis e categorizações.

Especificamente neste texto tratamos, além dos valores de Δ, de uma medida de associação entre duas variáveis, aqui denotada por $V(X_i, X_j)$. Existe uma ampla variedade de medidas apropriadas para avaliar a associação entre variáveis nominais, tais como: o teste qui quadrado, o qui quadrado da razão de verossimilhança, o qui–quadrado de Mantel–Haenszel, o valor Phi, a medida V de Cramer, o valor Tau de Goodman e Kruskal e o coeficiente Kappa (ver, por exemplo (GOODMAN; KRUSKAL, 1954) (HINKLE; WIERSMA; JURS, 1994) (AGRESTI, 2002)). Será considerada neste texto a medida V de Cramer, que é uma medida conceitualmente similar ao coeficiente de correlação, que se refere ao grau de associação linear entre duas variáveis do tipo contínuo. Os valores possíveis para o coeficiente de correlação estão no intervalo entre -1 e 1. Já para as medidas de associação, estas variam, mas a maior parte delas resulta em um valor entre 0 e 1. Cabe observar que estas medidas de associação não assumem valores negativos devido a não ordenação das classes de uma variável nominal. Para variáveis ordinais, por exemplo, as medidas de associação podem ser negativas.

A ideia de aplicação da consistência estatística neste caso é a de avaliar se a existência de associação entre duas variáveis após a imputação permanece a mesma ou é influenciada por algum fator inerente ao método que usa a rede bayesiana.

Apesar de os testes de associação entre duas variáveis apresentarem significância com base na distribuição qui–quadrado, o valor da estatística é difícil de ser interpretado por ser uma função do tamanho da amostra, da independência entre as variáveis e dos graus de liberdade. O valor V de Cramer, assim como o de outras medidas de associação, mostra–se como solução para esta dificuldade de interpretação, embora também seja construído com base no qui–quadrado (AGRESTI, 2002). A fórmula do teste qui–quadrado é:

$$\chi^2(X_i, X_j) = \sum_{k=1}^{c} \frac{(o_k - e_k)^2}{e_k},$$

onde o_k é a frequência observada e e_k é a frequência esperada nas c combinações de classes das variáveis X_i e X_j consideradas em uma tabela de contingência. O valor V de Cramer é calculado por (CRAMER, 1999):

$$V(X_i, X_j) = \sqrt{\frac{\chi^2(X_i, X_j)}{n(l-1)}}, \tag{9}$$

onde n é o número total de casos na tabela de contingência e l é o menor entre os números de linhas e colunas da tabela. Esta medida é a mais popular das medidas de associação para variáveis nominais e é apropriada para tabelas que são maiores que as do tipo 2×2 (ou seja, tabelas para variáveis do tipo dicotômicas).

Tão importante quanto a obtenção do valor da associação após a imputação por redes bayesianas, está a verificação de alguma possível alteração na conclusão a partir do seu valor calculado. Para avaliar os valores de $V(X_i, X_j)$ tomaremos por base a Tabela 2 que identifica a interpretação referente a cada intervalo do valor do coeficiente de correlação de Pearson – $\rho(X_i, X_j)$ (SHIMAKURA, 2006). Uma associação que estiver, por exemplo, entre 0 e 0,19 é interpretada como bem fraca e uma mudança neste patamar poderia implicar numa alteração brusca na interpretação do valor de $V(X_i, X_j)$. O valor nulo implica em não associação entre as variáveis.

Para avaliar o comportamento do coeficiente de correlação após a imputação a partir de redes bayesianas é feita a suposição de que os relacionamentos entre as variáveis mapeadas pela estrutura G permanecem inalterados quando obtido $\tilde{G}$, e que a variação que ocorrer de $V(X_i, X_j)$ para o valor de $\tilde{V}(X_i, X_j)$ não altera a interpretação da associação entre X_i e X_j.

Tabela 2 – Interpretação dos valores do coeficiente de correlação de Pearson ($\rho(X_i, X_j)$) entre duas variáveis que servem de base para a interpretação do valor $V(X_i, X_j)$ de Cramer

Valor de $\lvert\rho(X_i, X_j)\rvert$	Interpretação da correlação
0,00 a 0,19	Bem fraca
0,20 a 0,39	Fraca
0,40 a 0,69	Moderada
0,70 a 0,89	Forte
0,90 a 1,00	Muito forte

A próxima seção traz resultados destas avaliações em rede obtida a partir dos dados do Censo Demográfico Brasileiro.

3. Aplicação aos dados do Censo Demográfico Brasileiro

Como aplicação, foram utilizadas redes bayesianas para imputação com configurações de redes de três, quatro e cinco nós, nas características do domicílio do questionário básico CD–01 do Censo Demográfico Brasileiro. Apresentamos, nesta seção, os resultados para uma rede com cinco nós e os demais podem ser consultados em Magalhães (2007).

O Censo Demográfico no Brasil compreende um grande conjunto de operações de coleta, processamento, análise e disseminação de dados populacionais que ocorrem a cada e durante dez anos. Os Censos Demográficos produzem informações imprescindíveis para a definição e acompanhamento de políticas públicas e tomada de decisões de investimento, sejam elas de caráter público ou privado. Todas as etapas de planejamento, treinamento, coleta e outras relacionadas ao Censo Demográfico Brasileiro podem ser consultadas no documento Metodologia do Censo Demográfico 2000 (IBGE, 2003), sendo importante destacar aqui apenas alguns pontos.

O Censo Demográfico é a pesquisa responsável pela atualização das estatísticas demográficas oficiais do País. É realizado pelo IBGE em todo o Território Nacional e é coletado em dois questionários: o questionário básico (CD–01), que possui itens pesquisados para todos os habitantes e o questionário da amostra (CD–02), que é aplicado a amostras de 10% ou 20% dos domicílios de cada um dos municípios brasileiros. O questionário básico contém duas partes: uma com dez perguntas sobre características do domicílio e uma segunda contendo nove perguntas sobre as características dos moradores a ser preenchido para cada morador. Já o questionário da amostra é aplicado a parte dos domicílios e é mais extenso que o CD–01. Apresenta maior variedade de perguntas, e em seu conteúdo também contém as questões encontradas no questionário básico. No CD–02, a cada observação é associado um peso de expansão da amostra.

Nesta seção avaliamos o uso de redes bayesianas discretas para imputação conforme descrito na seção anterior. São construídos os indicadores de consistência estrutural e consistência estatística para redes de domicílios do Município de Natal com renda domiciliar (informada ou imputada) nula na base de dados original. Isso totaliza 15.225 de um total de 179.822 domicílios.

As variáveis consideradas para a construção das redes discretas estão listadas na Tabela 3. Essas variáveis foram escolhidas dentre as dez pelo relacionamento existente entre elas e pela existência de zeros estruturais, ou seja, em algum momento uma dada resposta leva o respondente a não responder uma ou mais variáveis do questionário, sem que isso seja caracterizado como não resposta. Todas as variáveis consideradas referem–se a características do domicílio, por exemplo, tipo do domícilio, identificada por TIPODOM, que

registra se o mesmo é uma casa, apartamento ou cômodo. Ainda na Tabela 3 são descritas as classes de cada variável, bem como o percentual de cada uma delas do total de 15.225 domicílios.

Tabela 3 – Características de domicílios consideradas no ajuste das redes discretas para imputação de dados do Censo Demográfico, Natal, 2000

Nome da variável	Descrição da variável	Classes	(%) do total
TIPODOM	Tipo do domicílio	1 – casa 2 – apartamento 3 – cômodo	0,964 0,024 0,012
CONDDOM	Condição do domicílio	1 – próprio (já pago) 2 – próprio (ainda pagando) 3 – alugado 4 – cedido por empregador 5 – cedido de outra forma 6 – outra condição	0,634 0,106 0,160 0,004 0,066 0,030
CONDTER	Condição do terreno	1 – próprio 2 – cedido 3 – outra condição Z – zero estrutural	0,703 0,020 0,017 0,259
ABASTEC	Forma de abastecimento de água	1 – rede geral 2 – poço ou nascente 3 – outra	0,940 0,020 0,040
TIPOCAN	Como a água chega no domicílio	1 – canalizada 2 – canalizada no terreno 3 – não canalizada	0,853 0,097 0,050

Fonte: IBGE, Censo Demográfico, 2000.

A título de informação, na Tabela 4, a seguir, são descritos os percentuais de não resposta em cada uma destas variáveis na base original. Podemos perceber que o percentual de não resposta nestas variáveis é baixo, por exemplo, 0,011% de não resposta para tipo de domicílio (TIPODOM). Por este motivo, para todas as aplicações construídas foram geradas várias situações de não resposta do tipo *Missing Completely at Random* (MCAR) em percentuais variados (1%, 3%, 5%, 7%, 10%, 20%, 30%, 40% e 50%) para cada variável. Foram avaliadas as consistências da base de dados, lógica, estrutural e estatística em cada caso após 500 processos de imputação em cada uma delas.

Tabela 4 – Número e percentual de não resposta das variáveis na base de dados original, Natal, 2000

Nome da variável	Número de imputados	(%) de imputados
TIPODOM	165	0,011
CONDDOM	222	0,015
CONDTER	209	0,014
ABASTEC	171	0,011
TIPOCAN	392	0,026

Fonte: IBGE, Censo Demográfico, 2000.

Com os dados do Censo Demográfico foram avaliadas a imputação em redes discretas de cinco nós, que foram construídas agregando o conhecimento de especialistas, impedindo que relações de causa e efeito pouco coerentes com a realidade fossem obtidas no ajuste da rede, por exemplo, a condição do domicílio (CONDDOM) ser influenciada pela condição do terreno (CONDTER). Neste caso, a determinante para se responder à condição do terreno no questionário é dada a partir da variável condição do domicílio. Se fossem registradas como resposta algumas das classes (alugado, cedido por empregador, cedido de outra forma ou outra condição) em CONDDOM, então CONDTER teria obrigatoriamente uma não resposta do tipo zero estrutural por não se aplicar esta pergunta a estes casos. Dessa forma, não seria coerente considerar um arco de CONDTER para CONDDOM.

Funcionaram como especialistas na definição destas redes os técnicos que trabalharam diretamente com a metodologia e análise dos dados do Censo Demográfico 2000, e os técnicos que trabalharam na definição do método para imputação da variável de renda e a bibliografia disponível na área demográfica.

3.1. Rede com cinco nós

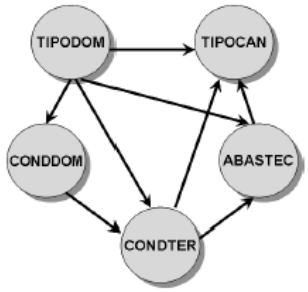
Para os estudos de simulação conduzidos com a rede de cinco nós, consideramos as variáveis listadas na Tabela 3. A escolha destas variáveis foi devido ao relacionamento que elas possuem entre si, pela ordenação no questionário e pela ocorrência de zero estrutural na variável CONDTER, onde o zero estrutural a essa questão decorre da resposta das classes (alugado, cedido por empregador, cedido de outra forma ou outra condição) em CONDDOM. Foram mantidas as seis categorias da variável CONDDOM para a avaliação da imputação naquelas que contivessem muitas classes. A Figura 6 traz o grafo da rede para as cinco variáveis, ajustada pelo pacote *deal* do R (BOTTCHER; DETHLEFSEN, 2003).

A Tabela 3 mostra que a variável TIPOCAN possui características semelhantes às variáveis TIPODOM e ABASTEC, com poucas classes e uma delas com grande percentual de ocorrência nos dados.

Ao identificar os conjuntos de variáveis no grafo da Figura 6, temos que a variável TIPODOM pertence ao conjunto P_0 das variáveis sem pais, CONDDOM ao conjunto P_1 tendo como pai a variável em P_0, e em P_2 está CONDTER cujos pais estão no conjunto $P_0 \bigcup P_1$. Ao se incluir a variável TIPOCAN, define–se o conjunto P_4, cujos pais estão em $\bigcup_{i=0}^{3} P_i$. A sequência de imputação na rede é dada pela distribuição condicional das variáveis de acordo com a relação descrita no grafo. É imputado TIPODOM de acordo com a sua distribuição marginal, depois é imputado CONDDOM de acordo com P(CONDDOM | TIPODOM) e

finalmente, CONDTER segundo P(CONDTER | CONDDOM, TIPODOM). A imputação em TIPOCAN será dada pela relação P(TIPOCAN | TIPODOM, CONDTER, ABASTEC) identificada pelo grafo na Figura 6.

Figura 6 – Grafo da rede ajustada para cinco variáveis discretas



Fonte: IBGE, Censo Demográfico, 2000.

Com uma estrutura complexa, a avaliação da consistência estrutural apresentou resultados interessantes. Para baixos percentuais de não resposta a estrutura está mantida preservada independente da combinação de variáveis que será imputada. O mesmo não ocorreu para percentuais mais altos de dados faltantes. Este foi o caso, por exemplo, das imputações conduzidas em percentuais acima dos 10% de não resposta, aonde se chegou a até 40% de alteração da estrutura da rede após o seu ajuste com os dados imputados.

Partindo para a avaliação da consistência estatística percebemos, a partir dos resultados na Figura 7, que os resultados na consistência dos dados e estrutural não parecem influenciar a consistência estatística. Com relação aos valores de Δ, observamos que estes melhoram à medida que se aumenta o percentual de não resposta simulado. A Figura 7 explicita os valores de Δ para os diferentes percentuais de não resposta estudados e em diferentes combinações de variáveis. Além da melhoria na consistência estatística associada aos valores dos parâmetros da rede, observamos uma diminuição dos correspondentes desvios à medida que se aumenta o percentual de não resposta simulado.

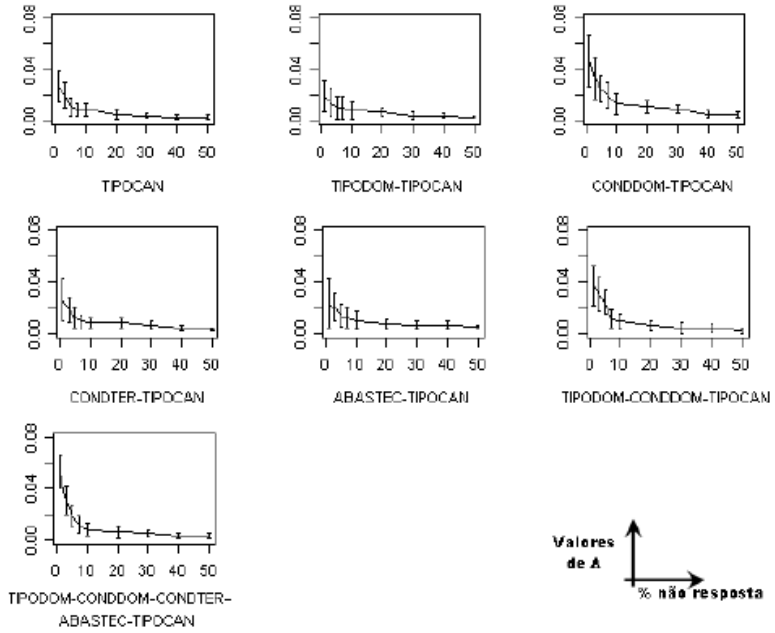
Ainda na consistência estatística, mas observando agora o valor V de Cramer entre duas variáeis após a imputação, observamos que, à medida que se aumenta o percentual de não resposta, tende–se a diminuir a associação entre as variáveis consideradas. A Figura 8 traz o resultado dos cálculos das medidas de associação após as 500 simulações nos diferentes percentuais de não resposta estudados. Para a comparação, utilizamos os valores calculados a partir da base original que podem ser observados na Tabela 7.

Tabela 7 – Medidas de associação (V de Cramer) entre duas variáveis na rede original com cinco nós, dados de Natal, 2000

	TIPODOM	CONDDOM	CONDTER	ABASTEC	TIPOCAN
TIPODOM	1	–	–	–	–
CONDDOM	0,099	1	–	–	–
CONDTER	0,077	0,708	1	–	–
ABASTEC	0,052	0,325	0,118	1	–
TIPOCAN	0,094	0,298	0,124	0,638	1

Fonte: IBGE, Censo Demográfico, 2000.

Figura 7 – Resultado da avaliação da consistência estatística – valor de Δ após 500 imputações a partir de rede com cinco nós e em diferentes percentuais de não resposta, dados de Natal, 2000



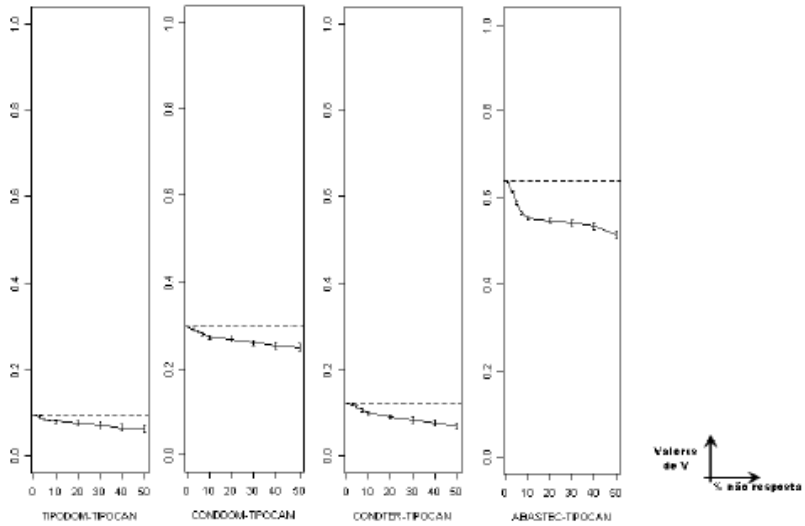
Fonte: IBGE, Censo Demográfico, 2000.

4. Conclusões e futuros direcionamentos

Neste artigo foi proposto um algoritmo para o uso de redes bayesianas para imputação em variáveis aleatórias discretas que tem como finalidade agregar o conhecimento de especialistas sobre as variáveis e seus relacionamentos de independência condicional na construção da rede. Essa informação evita uma ordenação prévia dos atributos que implique em uma alteração na distribuição multivariada das variáveis.

Na avaliação, foi conduzida uma medida de consistência proposta por Di Zio *et al*. (2004), a consistência estatística. Além disso, foi proposto neste texto a consistência estrutural, com o objetivo de auxiliar na verificação da hipótese de que a imputação conduzida sob redes bayesianas preservaria o relacionamento multivariado entre as variáveis da rede.

Figura 8 – Gráficos do valor V de Cramer calculado após 500 imputações a partir de rede com cinco nós e em diferentes percentuais de não resposta (a linha tracejada refere–se à medida de associação calculada a partir da base original), dados de Natal, 2000



Fonte: IBGE, Censo Demográfico, 2000.

Considera–se fundamental que outros aspectos sejam observados ao se examinar a consistência estrutural, como os parâmetros da rede e a manutenção dos registros individuais. Por isso, a consistência estatística foi estendida para além dos parâmetros da rede com o cálculo de uma medida de associação entre as variáveis. A rede estudada com os dados do Censo Demográfico foi pequena (cinco nós) para se equiparar ao estudo realizado por Di Zio *et al*. (2004), embora o objetivo neste trabalho não fosse a comparação em si.

Em suma, os resultados obtidos com as redes estudadas serviram para a percepção de aspectos conjuntos que se buscam ao se definir um método de imputação e para balizar futuros direcionamentos no tema. Por ser esta uma aplicação relativamente recente das redes bayesianas, ainda muito existe para ser estudado e as avaliações que se seguem abaixo indicam que o estudo e aplicação deste método para imputação mostraram-se promissores:

- a preservação da estrutura, em geral, não depende do percentual de não resposta quando o número de observações da base de dados é relativamente alto;

Conforme em Neapolitan (2004), a classe de equivalência de uma dada estrutura pode ser obtida para $n \to \infty$ em determinadas classes de modelos. As simulações conduzidas com redes discretas consideraram estruturas simples, com poucas variáveis e nas situações de grande número de unidades na base (o caso do Censo Demográfico) e um número mais restrito. Para a base de dados maior, mais de 90% das redes bayesianas aprendidas sob os dados imputados em uma rede com quatro nós e sob 50% de não resposta, por exemplo,

pertenciam à classe de equivalência da rede original. Um fator que exerce influência na medida de consistência estrutural é o número de nós, o que implica também na complexidade da rede. No caso dos dados do Censo Demográfico na rede construída com cinco nós, em altos percentuais de não resposta a estrutura construída após a imputação encontrou–se na mesma classe de equivalência da estrutura original em uma ordem de 60% das vezes.

- com relação aos parâmetros da rede, estes se mostram mais próximos dos parâmetros originais à medida que se aumenta o percentual de não resposta na base, independentemente da estrutura;

Este resultado entra em contradição com o anterior para bases de dados pequenas. Nos dados do Censo Demográfico, em todos os percentuais de não resposta simulados, os valores de Δ mostraram–se próximos de zero, ou seja, os parâmetros da rede ajustada após a imputação não sofreram alterações bruscas, salvo para aquelas variáveis com muitas classes, que apresentaram desempenho menor para esta consistência.

- altos valores de medidas de associação entre as variáveis são afetadas pelos maiores percentuais de não resposta após a imputação via rede bayesiana;

Nas situações avaliadas, os baixos valores calculados da medida V de Cramer não sofreram alterações após a imputação. Este foi o caso, por exemplo, da associação entre TIPODOM e TIPOCAN. Já as medidas de associação mais altas (em comparação com as demais calculadas na mesma rede), são mais sensíveis ao aumento do percentual de não resposta. Possivelmente, esta mudança no nível da medida de associação pode estar relacionada com a mudança na relação observada entre as variáveis que são explicitadas na estrutura da rede. As relações de dependência entre os nós não são necessariamente as mesmas identificadas pela medida de associação V de Cramer calculada entre as variáveis. Neste caso, a imputação a partir de redes bayesianas como método tende a diminuir a associação existente entre as variáveis à medida que se aumenta o percentual de não resposta.

Estes resultados foram algumas observações que se tornaram recorrentes para as bases de dados, estruturas e parâmetros diferentes, portanto percebeu–se a necessidade de registrá–los. Outros aspectos, como o decaimento não-linear do coeficiente de determinação em função do percentual de não resposta ou a diminuição do desvio das estimativas também em função do percentual de não resposta para os valores de Δ, são itens específicos e dependentes do modelo real e suposto aos dados. Estes aspectos carecem de maiores estudos e apresentam–se como futuros direcionamentos deste trabalho.

Referências bibliográficas

AGRESTI, A. Categorical Data Analysis, New Jersey: John Wiley & Sons, 2002.

BINNER, J. M., KENDALL, G., CHEN, S. H. Applications of Artificial Intelligence in Finance and Economics. Hardbound: Elsevier, 2005.

BOTTCHER, S. G., DETHLEFSEN, C. deal: A package for learning Bayesian networks. Journal of Statistical Software, 8(20), 2003.

CHAMBERS, R. Evaluation criteria for statistical editing and imputation. Technical Report #28, National Statistics Methodology Series, 2000.

CHARNIAK, E. Bayesian networks without tears. AI Magazine, p. 50–630, 1991.

CRAMER, H. Mathematical Methods of Statistics. New Jersey: Princeton University Press, 1999.

DI ZIO, M., SCANU, M. Bayesian networks. Methods and experimental results from the Euredit project. Technical Appendices E, Volume 2, Euredit deliverableD6.1, on CD with Volume 1, 2003.

DI ZIO, M., SCANU, M., COPPOLA, L., LUZI, O., PONTI, A. Bayesian networks for imputation. Journal of the Royal Statistical Society A, 167, Part 2, p. 309–322, 2004.

DI ZIO, M., SCANU, M., VICARD, P. Open problems and new perspectives for imputation using Bayesian networks. Modelli Complesi e Metodi Computazionali Intensivi per la Stima e la Previsione. Treviso: 2003.
disponível em: www.dst.unive.it\sco2003\.

FREY, B. J. Graphical Models for Machine Learning and Digital Communication. Cambridge: Bradford Book, 1998.

GOODMAN, L. A., KRUSKAL, W. H. Measures of association for cross–classification. Journal of the American Statistical Association, 49, 732–764, 1954.

HECKERMAN, D. E. An axiomatic framework for belief updates. Uncertainty in Artificial Intelligence 2, New York: 1988.

HINKLE, D. E., WIERSMA, W., JURS, S. G. Applied Statistics for the Behavioral Sciences. Boston: Houghton Mifflin, 1994.

HRUSCHKA--JR., E. R. Imputação Bayesiana no Contexto da Mineração de Dados. Tese (Doutorado). Rio de Janeiro: COPPE--UFRJ, 2003.

IBGE. Metodologia do Censo Demográfico 2000. Rio de Janeiro: IBGE, 2003.

MAGALHÃES, I. Avaliação de redes bayesianas para imputação em variáveis qualitativas e quantitativas. Tese (Doutorado). Escola Politécnica, São Paulo: USP, 2007.

MURPHY, K. P. Inference and learning in hybrid Bayesian networks. Technical Report No. UCB/CSD-98-990, Berkeley: Computer Science Division, University of California, 1998.

NEAPOLITAN, R. E. Learning Bayesian Networks. New Jersey: Pearson Prentice Hall, 2004.

PEARL, J. Probabilistic Reasoning in Intelligent Systems. California: Morgan Kaufmann, 1988.

SAHEKI, A. H. Construção de uma Rede Bayesiana Aplicada ao Diagnóstico de Doenças Cardíacas. Tese (Mestrado). São Paulo: Escola Politécnica, Universidade de São Paulo, 2005.

SCANU, M., DI ZIO, M., VICARD, P. Computational aspects of imputation with Bayesian networks. A MaPhySto Workshop on Computational Aspects of Graphical Models. Aalborg, Dinamarca: Aalborg University, 2003. disponível em www.math.aau.dk\gr\material\scanu.pdf.

SHIMAKURA, S. E. Interpretação do coeficiente de correlação. Notas de Aula Online. Paraná: UFPR, 2006. disponível em http:\\leg.ufpr.br\~silvia\CE003.

THIBAUDEAU, Y., WINKLER, W. E. Bayesian networks representation, generalized imputation, and synthetic micro--data satisfying analytic constraints. Technical Report #2002-09, Washington: U.S. Bureau of the Census, 2002.

Abstract

Bayesian networks are structures that combine probability distributions with graphs. Although Bayesian networks initially appeared in the 1980s and the first attempts to solve problems generated from the non-response date back to the 1930s and 1940s, the use of structures of this kind specifically for imputation is rather recent: in 2002 by official statistical institutes, and in 2003 in the context of data mining. As it exists theoretical questions and still not boarded properties in literature, beyond few references to the subject, we intended to advance in the properties and to weave greaters considerations regarding the application of Bayesians networks as an imputation method. We present results from the application of a new algorithm for data imputation based on discrete Bayesian networks built combining knowledge obtained from experts. To evaluate Bayesian networks in this context, we use three types of consistence suggested by Di Zio *et al*. (2004) and we propose a new one: the structural consistence, which can be defined as the ability of a network to maintain its structure in the equivalence class of the original network when built from the data after imputation. The importance of this consistency measure is to evaluate if the relationship of conditional independence between variables are preserved after the imputation process. Applications are conducted using data from 2000 Demographic Census based on different structures of simple networks.

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 via e-mail para: rbe@ibge.gov.br.

Após a submissão, o autor receberá um código para acompanhar o processo de avaliação do artigo. Caso não receba um aviso com este código no prazo de uma semana, fazer contato com a secretaria da revista no endereço:

Revista Brasileira de Estatística – RBEs
IBGE – ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS
Rua André Cavalcanti, 106, sala 111
Centro, Rio de Janeiro – RJ
CEP: 20031-170
Tels.: 55 21 2142-4682 (Sandra Cavalcanti Barros – Secretária)
55 21 2142-4686 (Ismenia Blavatsky – 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 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 do 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.