Presidente da República
**Fernando Henrique Cardoso**

Ministro do Planejamento, Orçamento e Gestão
**Guilherme Gomes Dias**

# INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
**Sérgio Besserman Vianna**

Diretor Executivo
**Nuno Duarte da Costa Bittencourt**

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
**Maria Martha Malard Mayer**

Diretoria de Geociências
**Guido Gelli**

Diretoria de Informática
**Paulo Roberto Ribeiro da Cunha**

Centro de Documentação e Disseminação de Informações
**David Wu Tai**

Escola Nacional de Ciências Estatísticas
**Kaizô Iwakami Beltrão**

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

# REVISTA BRASILEIRA DE ESTATÍSTICA

volume 62 número 217 janeiro/junho 2001

ISSN 0034-7175

*R. bras. Estat.*, Rio de Janeiro, v. 62, n. 217 p. 1-108, jan./jun. 2001

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



**Revista Brasileira de Estatística, ISSN 0034-7175**

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

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

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

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

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

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

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



**Editoração**
Helem Ortega da Silva - Departamento de Metodologia - DPE

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

**Capa**
Renato J. Aguiar – Gerência de Criação – CDDI

**Ilustração da Capa**
Marcos Balster – Gerência de Criação – CDDI

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

v.

Trimestral (1940-1986), semestral (1987- ).

Continuação de: Revista de economia e estatística.

Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v. 50 (1980-1989).

Co-edição com a Associação Brasileira de Estatística a partir do v.58.

ISSN 0034-7175 = Revista brasileira de estatística.

I. Estatística – Periódicos. I. IBGE. II. Associação Brasileira de Estatística.

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

Impresso no Brasil/*Printed in Brazil*

# Sumário

# NOTA DO EDITOR

Mais uma vez, este número da revista apresenta uma combinação de artigos revelando contribuições de grande interesse. Cruz e Silva abrem o número apresentando uma aplicação de métodos de séries temporais para o cálculo de estimativas da precisão de séries de indicadores de desemprego sazonalmente ajustadas, baseadas em pesquisas amostrais repetidas. Esta contribuição permitirá aos usuários e analistas de tais séries conhecerem a precisão das séries e avaliar o impacto de se basearem em amostras ao fazer suas análises. Lima, André e Singer apresentam uma revisão dos modelos aditivos generalizados, seus procedimentos de ajuste e testes de hipótese, com vistas a sua aplicação a um estudo ambiental para descrever a associação entre mortalidade fetal tardia e poluição atmosférica na cidade de São Paulo. Bidurin e Milan apresentam uma revisão dos modelos semiparamétricos, destacando as formas tradicionais de estimação do componente não-paramétrico e apresentando o modelo de regressão por *spline*, juntamente com uma discussão sobre a seleção dos parâmetros de suavização. Completando o número, Sáfadi e Moretin estudam séries da mortalidade por problemas cardíacos e das temperaturas mínimas na Cidade de São Paulo, mediante uma análise Bayesiana do modelo de "open loop threshold".

Ao apresentar este número da RBEs estarei passando o posto de editor responsável ao Prof. Renato Martins Assunção, da Universidade Federal de Minas Gerais. Tendo cumprido um período de quase cinco anos como editor, chegou a hora de abrir caminho para a renovação de forças e energia no conselho editorial. Dois novos membros estão sendo adicionados, o Prof. Francisco Louzada-Neto, da Universidade Federal de São Carlos, ocupará a editoria de metodologia da revista, e o Prof. Francisco Cribari, da Universidade Federal de Pernambuco, será mais um dos editores associados. Aos novos membros do conselho editorial, as boas vindas e os votos de que sua colaboração à produção da revista seja a mais proveitosa possível. Não poderia deixar de registrar e agradecer à colaboração recebida dos membros do atual conselho editorial durante meu período como editor. Meu muito obrigado aos professores Djalma Pessoa, Hélio Migon, Renato Assunção, Wilton Bussab, Lisbeth Cordani, Gilberto Paula e Kaizo Beltrão. Um agradecimento especial vai aqui também à Helem Ortega, que tem secretariado e editorado a revista, pela competência e dedicação, e por tornar a vida do editor tão mais fácil e simples. Para terminar, um agradecimento aos muitos colegas e profissionais que colaboraram submetendo ou revisando artigos: espero que continuem contribuindo

para fazer da Revista Brasileira de Estatística uma publicação digna do nível de desenvolvimento alcançado pela comunidade estatística brasileira.

Saudações,

Pedro Luis do Nascimento Silva

Editor Responsável

# Estimação de variâncias para séries sazonalmente ajustadas pelo método X-12-ARIMA, considerando o esquema de sobreposição e rotação amostral

Marcelo Martins Cruz*

Denise Britz do Nascimento Silva**

## Resumo

**Várias séries temporais utilizadas como fonte de informação para o conhecimento da realidade e para pesquisa científica são provenientes de pesquisas por amostragem repetidas no tempo. Apesar do uso de estimativas amostrais para análise de características de interesse ser prática e amplamente difundida, nem sempre a obtenção de medidas de precisão das estimativas é incluída no processo de análise e divulgação de resultados. Quando o foco da análise é uma série temporal, a decomposição da série em componentes de tendência e sazonalidade permite uma melhor compreensão do fenômeno observado. O método X11 (Shiskin, Young e Musgrave, 1967), e variantes, são procedimentos de ajustamento sazonal muito utilizados em todo o mundo. Um problema ainda pendente com estes procedimentos é a estimação de variâncias para os componentes de tendência, sazonalidade e da série sazonalmente ajustada. Adicionalmente, no caso de séries provenientes de pesquisas amostrais, deve-se considerar o efeito que o esquema de sobreposição e rotação amostral da pesquisa exercem sobre a série observada e incorporar a variância do erro amostral no cálculo da precisão das estimativas dos componentes. O propósito deste trabalho é fornecer estimativas de desvios-padrão dos componentes produzidos pelo X-12-ARIMA (Findley et al. 1998), utilizando a metodologia proposta por Pfeffermann (1994).**

* Endereço para correspondência: IBGE, Avenida República do Chile, 500, 10º andar, CEP 20.031-170 – E-mail: mmcruz@ibge.gov.br

** Endereço para correspondência: ENCE, Rua André Cavalcanti, 106, 4º andar, CEP 20.231-050 – E-mail: denisesilva@ibge.gov.br

# 1. Introdução

O problema de estimação de variâncias para séries sazonalmente ajustadas é de longa data, conforme ilustram Wolter e Monsour (1981):

> *In the final of the President's Committee to Appraise Employment and Unemployment Statistics (the Gordon Committee) (1962), it was suggested that standard errors for an adjusted labor force series are more important than those for the original series because of the increased reliance of policy makers on the deseasonalized series, and that research be undertaken on how to estimate and publish such standard errors. Unfortunately, this suggestion did not lead to any major advances or to practical methods for estimating and publishing the standard errors. Recognizing this fact, the national Commission on Employment and Unemployment Statistics (1979) recently reaffirmed the Gordon Committee's findings. In part, their recommendations read as follows:*
>
> *"The commission reemphasizes the importance of standard errors for seasonally adjusted statistics and urges the Census Bureau to undertake research to develop them."*
>
> *Finkner and Nisselson (1978) characterized the need for variance estimates for a deseasonalized series as one of the main statistical problems with continuing cross-sectional surveys.*

Uma grande parte dos ajustamentos sazonais das agências produtoras de estatística de todo o mundo é elaborada utilizando-se métodos empíricos, notadamente o X-11, o X-11-ARIMA (Dagum, 1988) ou, ainda, o X-12-ARIMA. Estes métodos são baseados em filtros ou **médias móveis fixas**, e não em modelos de decomposição estocásticos. Adicionalmente, inúmeras séries temporais são oriundas de pesquisas repetidas com planos amostrais complexos, estando, portanto, sujeitas a erros amostrais. Sendo assim, deve-se incorporar a variabilidade do erro amostral ao cálculo da variância do componente não-observável que se deseja estimar.

Pfeffermann (1994) propôs um método para estimação de variâncias dos estimadores utilizados pelo X-11 (ou equivalentemente, X-11-ARIMA ou X-12-ARIMA), que se estrutura em duas suposições ou hipóteses fundamentais:

O método leva em consideração dois tipos diferentes de erros, a saber: o erro amostral dos estimadores em torno dos correspondentes valores populacionais, e a variabilidade dos componentes que descrevem o modelo de decomposição para os valores populacionais, ou seja, a metodologia leva em conta as contribuições dos erros amostrais e a variabilidade do componente irregular. As propriedades do **erro combinado** (*erro amostral* mais o *componente irregular*) são estimadas pelos valores do **componente irregular fornecidos pelo X-11**. Estas propriedades serão utilizadas para estimar dois tipos de variâncias.

Assume-se que os estimadores do X-11 para a tendência e sazonalidade são aproximadamente não-viciados, os erros amostrais e o componente irregular são estacionários e que **as possíveis estruturas estocásticas da tendência e da sazonalidade não necessitam ser especificadas**.

Como conseqüência das hipóteses fundamentais surgem duas importantes características:

O método leva em conta a correlação serial entre os erros amostrais presentes, por exemplo, nas pesquisas amostrais com painéis rotativos.

O método não requer estimativas externas para as covariâncias da série observada (geralmente composta de estimativas amostrais).

# 2. Modelos de decomposição

Seja $\{y_t; t = 1,...,N\}$ a série observada. As séries temporais freqüentemente produzidas são constituídas por estimativas de pesquisas amostrais e, conseqüentemente, sujeitas a erros amostrais. Na abordagem baseada no planejamento amostral, as propriedades de uma estatística ou estimador são avaliadas com respeito à distribuição de aleatorização. Denote-se por $E_P(.)$ o operador esperança referente à distribuição de probabilidade $p(s)$ induzida pelo planejamento amostral. Assim sendo, sob a teoria da amostragem, a cada tempo $t$,

$$y_t = Y_t + \varepsilon_t, \quad E_P[\varepsilon_t] = 0 \text{ e } E_P[\varepsilon_t \varepsilon_{t-k}] = \lambda_k \text{ para } k = 0,1,... \tag{2.1}$$

onde $y_t$ é a série observada, $Y_t$ é o verdadeiro valor populacional, fixo e desconhecido, e $\varepsilon_t$ é o erro amostral.

Entretanto, sob a teoria de séries temporais, pode-se definir $\{Y_t\}$ como sendo uma seqüência de valores populacionais, que evoluem no tempo segundo um processo estocástico. Observe-se que, em (2.1), os erros amostrais $\varepsilon_t$, são considerados serialmente correlacionados, e denote-se por $\lambda_k$ as autocovariâncias dos erros amostrais, admitindo-se que $y_t$ não é afetado por erros de medida ou alheios à amostragem. As correlações seriais ocorrem em pesquisas amostrais quando há alguma sobreposição das unidades amostrais ao longo do tempo, como nas pesquisas com painéis rotativos (Duncan e Kalton, 1987). Destaque-se que estimativas de variâncias dos erros amostrais, $\lambda_0$, são rotineiramente produzidas pelas agências de estatísticas mas, por outro lado, raramente são incorporadas na análise de séries temporais.

Considerando então um modelo de componentes não-observáveis com **decomposição aditiva**, os valores populacionais podem ser expressos como,

$$Y_t = T_t + S_t + I_t, \quad E_M[I_t] = 0 \text{ e } E_M[I_t I_{t-k}] = \nu_k \text{ para } k = 0,1,... \tag{2.2}$$

onde $T_t$ representa o nível do componente tendência-ciclo, $S_t$ o efeito sazonal, $I_t$ o termo irregular (erro sob o modelo) e $E_M[.]$ é o valor esperado sob o modelo de decomposição. Os termos irregulares, $I_t$, também apresentam correlação serial, sendo $\nu_k$ suas autocovariâncias. Portanto, pela definição dos dois erros (erro amostral e o erro sob o modelo), assume-se que $E[I_t \varepsilon_{t-k}] = 0$ para todo $k \geq 0$.

Substituindo-se (2.2) em (2.1), rescreve-se a decomposição para a série observada como:

$$y_t = T_t + S_t + I_t + \varepsilon_t = T_t + S_t + e_t \tag{2. 3}$$

onde o componente de erro composto, $e_t = \varepsilon_t + I_t$, é também estacionário, e sob a distribuição conjunta, designada $C$, apresenta os seguintes momentos:

$$E_C[e_t] = 0, \; E_C[e_t e_{t-k}] = V_k = \lambda_k + \nu_k \text{ para } k = 0,1,... \tag{2.4}$$

Note que para $k = 0$, $V_0 = \lambda_0 + \nu_0$. Isto é, a variância do erro composto é a soma da variância do erro amostral mais a variância do componente irregular, sob o modelo de decomposição.

A decomposição (2.3), $y_t = T_t + S_t + e_t$, sem a suposição (2.4), é a decomposição postulada como modo aditivo pelo X-11, onde $\{e_t\}$ seria o componente irregular. Ressalte-se que na decomposição de série

proveniente de uma pesquisa amostral, pode-se supor que o componente irregular incorpora a variação devida aos erros amostrais.

## 2.1 A Aproximação linear para o X-11 e variantes

A metodologia proposta para a estimação de variâncias de séries sazonalmente ajustadas **utiliza a aproximação linear do X-11 como base para a estimação das variâncias**. O procedimento X-11 consiste em uma seqüência de médias móveis ou de operações de filtros lineares aplicados aos dados observados. A rede de efeitos destas operações pode ser representada por um único conjunto de médias móveis. Assim, o Estimador de Ajustamento Sazonal, denotado por EAS, é definido como

$$= y_t - \hat{S}_t, \tag{2.5}$$

onde $\hat{N}_t$ é a estimativa da série sazonalmente ajustada, $y_t$ a série observada e $\hat{S}_t$ a estimativa do componente sazonal. O EAS, $\hat{N}_t$, é calculado pela aplicação de uma seqüência de filtros lineares e pode ser representado na forma

$$\hat{N}_t = \sum_{k=-(t-1)}^{N-t} \omega_{kt} y_{t+k} = \boldsymbol{\omega}_t' \mathbf{y}, \text{ com } t = 1,..,N, \tag{2.6}$$

onde $\mathbf{y}' = (y_1,..., y_N)$ e $\boldsymbol{\omega}_t = (\omega_{t_1},...,\omega_{t_N})'$. Os outros componentes do modelo (2.3) podem ser estimados utilizando-se uma representação similar a (2.6). Os pesos $\{\omega_{kt}\}$ não são invariantes ao longo de todo o processo, uma vez que os estimadores utilizam filtros assimétricos no começo e no final das séries.

A substituição de (2.3) em (2.6), fornece a seguinte decomposição para o EAS:

$$\hat{N}_t = \sum_{k=-(t-1)}^{N-t} \omega_{kt}(T_{t+k} + S_{t+k}) + \sum_{k=-(t-1)}^{N-t} \omega_{kt}\varepsilon_{t+k} + \sum_{k=-(t-1)}^{N-t} \omega_{kt} I_{t+k}. \tag{2.7}$$

Bell e Monsell (1992) descrevem o procedimento, baseado na abordagem de Wallis (1974), para a determinação das aproximações dos filtros lineares utilizados pelo X-11, assumindo a decomposição aditiva de uma série observada ($y_t = T_t + S_t + I_t$) e supondo que $y_t$ não é sujeita a erro amostral.

As etapas de alisamento, via médias móveis, usadas pelo X-11 para as estimativas dos componentes $\hat{T}_t$, $\hat{S}_t$ e $\hat{I}_t$ na decomposição aditiva, são descritas a seguir com o objetivo de estabelecer as expressões para o cálculo dos pesos $\omega_{kt}$ de (2.7) produzidos pelos vários filtros utilizados pelo X-11.

Etapas:

1) Extrai-se a tendência de $y_t$ por subtração de uma média móvel centrada de 12 termos (média móvel 2X12);

2) Calcula-se uma primeira média móvel sazonal (*default* = 3X3) na série resultante do passo 1, e utiliza-se como uma estimativa preliminar de $\hat{S}_t$;

3) Ajustam-se os valores de $\hat{S}_t$ para que a soma dos efeitos sazonais seja aproximadamente igual a zero em qualquer período de 12 meses, subtraindo-os de uma média móvel 2X12;

4) Subtrai-se o resultado obtido no passo 3 de $y_t$ para determinar uma série dessazonalizada preliminar;

5) Subtrai-se uma média móvel de tendência (Henderson) do resultado obtido em 4 para uma extração mais refinada da tendência;

6) Aplica-se uma segunda média móvel sazonal (*default* = 3X5) no resultado de 5;

7) Ajusta-se o resultado de 6, como na etapa 3, por subtração de uma média móvel 2X12. O resultado, $\hat{S}_t$, é uma estimava de $S_t$; e

8) A série dessazonalizada é então, $\hat{N}_t = y_t - \hat{S}_t$. A estimativa da tendência, $\hat{T}_t$, é obtida aplicando-se uma média móvel de tendência (Henderson) em $\hat{N}_t$, e a estimativa do componente irregular é obtida como $\hat{I}_t = \hat{N}_t - \hat{T}_t$.

Pode-se expressar o cálculo realizado em cada uma das etapas através de polinômios (funções do operador atraso, $B$). É possível, então, representar o filtro sazonal, o de ajustamento sazonal, o de tendência e o irregular, respectivamente $\omega_S(B)$, $\omega_N(B)$, $\omega_T(B)$, $\omega_I(B)$, por estes polinômios. Assim, a expressão ou rede de efeitos para o filtro sazonal $\omega_S(B)$ utilizado no X-11 tem a forma:

$$\omega_S(B) = [1-\mu(B)]\Gamma_2(B)[1-H(B)\{1-[1-\mu(B)]\Gamma_1(B)[1-\mu(B)]\}]. \tag{2.8}$$

Sendo $U(B) = 1 + B + ... + B^{11}$ e $F = B^{-1}$, escreve-se:

$\mu(B) = (1/24)(F^6 + F^5)U(B)$, que é uma média móvel 2X12, primeira estimativa da tendência.

Adicionalmente tem-se

$\Gamma_1(B) = (1/9)(F^{12} + 1 + B^{12})(F^{12} + 1 + B^{12})$, a primeira média móvel sazonal (o *default* utilizado no método X-11 é uma média móvel 3X3), e

$\Gamma_2(B) = (1/15)(F^{12} + 1 + B^{12})(F^{24} + F^{12} + 1 + B^{12} + B^{24})$, a segunda média móvel sazonal (o *default* utilizado no X-11 é uma média móvel 3X5). Finalmente, $H(B)$ é a média móvel de tendência (Henderson). Detalhes do desenvolvimento são apresentados em Dagum (1985, p. 634).

Os outros filtros (de ajustamento sazonal, de tendência e irregular) utilizados pelo X-11 podem ser determinados em função da expressão (2.8), como se segue:

$\omega_N(B) = 1 - \omega_S(B)$ é o filtro para a série ajustada para sazonalidade; (2.9)

$\omega_T(B) = H(B)\omega_N(B)$ é o filtro de tendência; e (2.10)

$\omega_I(B) = [1-H(B)]\omega_N(B) = \omega_N(B) - \omega_T(B)$ é o filtro dos resíduos (irregular). (2.11)

Observe-se que as expressões (2.8) a (2.11) fornecem uma maneira conveniente de calcular os pesos para os vários filtros resultantes da opção *default* do programa ou para outras escolhas de $\Gamma_1(B)$, $\Gamma_2(B)$ e $H(B)$.

Se $\omega(B) = \sum_k \omega_k B^k$ é um filtro do X-11, seja sazonal, de ajustamento sazonal, de tendência, ou de irregularidade, os pesos do filtro $\omega_k$ podem ser calculados utilizando-se uma linguagem de programação matemática que execute a rotina de multiplicação de polinômios. Os valores dos pesos $\omega_I$ são necessários na

implementação efetiva do cálculo das variâncias dos estimadores utilizados pelo X-12-ARIMA, mas como se pode observar, os valores dos pesos $\omega_{ij}$, $\omega_N$, $\omega_T$ precisam também ser determinados. Os pesos de ajustamento sazonal ($\omega_N$) e de tendência ($\omega_T$) utilizados neste trabalho foram gentilmente cedidos pelo Dr. Stuart Scott do *Bureau of Labor Statistics*, BLS/US.

### 2.2 Propriedades dos filtros lineares do X-11 e variantes

A hipótese de que os estimadores de tendência e sazonalidade são aproximadamente não-viciados é utilizada para estabelecer as expressões das variâncias dos estimadores do X-11. Esta hipótese é baseada nas propriedades dos filtros lineares do X-11 (e variantes) considerando-se que:

$$E_C[\hat{T}_t - T_t] \cong 0 \qquad E_C[\hat{S}_t - S_t] \cong 0 \tag{2.12}$$

Pfeffermann (1994, p. 88-90) justifica a suposição (2.12) utilizando vários argumentos sobre as propriedades dos filtros lineares do X-11, tanto no domínio do tempo como no domínio da freqüência. Além disso, a hipótese mencionada está sustentada por um extensivo estudo de simulação apresentado no referido artigo. Pfeffermann (1994) alerta que as aproximações em (2.12) não são definidas em termos matemáticos exatos. Assume-se que a contribuição da variância dos vícios $(\hat{T}_t - T_t)$ e $(\hat{S}_t - S_t)$ com relação ao total da variância do EAS, sobre todas as possíveis realizações dos componentes tendência e sazonal, é suficientemente pequena para ser ignorada. Para mais detalhes ver a decomposição utilizada na prova do Lema 1, no apêndice.

# 3. Cálculo das variâncias dos estimadores de ajuste sazonal

## 3.1 Definições das variâncias

O Estimador de Ajustamento Sazonal - EAS - pode ser utilizado para estimar diferentes componentes, conforme a definição empregada. Dois casos podem ser estabelecidos.

***Caso 1***: O EAS é utilizado para estimar o valor da série sazonalmente ajustada na população. Neste caso, considera-se o resíduo ou erro 1,

$$D_{1t} = \hat{N}_t - N_t^*, \tag{3.1}$$

sendo $N_t^* = Y_t - S_t$ o verdadeiro valor populacional da série sazonalmente ajustada no tempo $t$, que constitui o alvo do processo de estimação. Assim, $D_{1t} = \hat{N}_t - (Y_t - S_t)$, com $\hat{N}_t = y_t - \hat{S}_t$. Conseqüentemente,

$D_{1t} = y_t - \hat{S}_t - Y_t + S_t$. Entretanto, sob a teoria da amostragem, tem-se que $y_t = Y_t + \varepsilon_t$ e

$$D_{1t} = \varepsilon_t - (\hat{S}_t - S_t). \tag{3.2}$$

Calculando a variância de (3.2), obtém-se

$$VAR(D_{1t}) = VAR^{(1)}(\hat{N}_t) = VAR\{\varepsilon_t - (\hat{S}_t - S_t)\} = VAR\{(\hat{S}_t - S_t) - \varepsilon_t\} \tag{3.3}$$

que é a variância do estimador da série ajustada para a sazonalidade. No caso da não existência de erro amostral ( $\varepsilon_t \equiv 0$ ), a variância em (3.3) é definida como $VAR^{(1)}(\hat{N}_t) = VAR(\hat{S}_t - S_t)$.

*Caso 2*: O EAS é utilizado para estimar o componente de tendência/nível. Nesse caso, define-se o erro 2 como $D_{2t} = \hat{N}_t - T_t = y_t - \hat{S}_t - (Y_t - S_t - I_t) = Y_t + \varepsilon_t - \hat{S}_t - Y_t + S_t + I_t$ ou seja,

$$D_{2t} = e_t - (\hat{S}_t - S_t). \tag{3.4}$$

Calculando a variância de (3.4), obtém-se,

$$VAR(D_{2t}) = VAR^{(2)}(\hat{N}_t) = VAR\{(\hat{S}_t - S_t) - e_t\} \tag{3.5}$$

que é a variância do estimador do componente de tendência.

## 3.2 Expressões explícitas para as variâncias

Sob a suposição de que o estimador do componente sazonal $\hat{S}_t$ é não-viciado para $S_t$, conforme definido em (2.12), ambos os termos de erro $D_{it}$, $i = 1,2$ têm valor esperado igual a zero, isto é, $E_C[D_{1t}] = E_C[D_{2t}] = 0$. Em conseqüência resulta o seguinte lema (detalhes no apêndice):

*Lema 1*: $VAR^{(i)}(\hat{N}_t) \cong VAR_C(D_{it})$; $\quad i = 1,2$.

As expressões explícitas para as variâncias $VAR_C(D_{it})$ podem agora ser estabelecidas. Seja

$$\hat{S}_t = \sum_{k=-(t-1)}^{N-t} \tilde{\omega}_{kt} y_{t+k} = \sum_{k=-(t-1)}^{N-t} \tilde{\omega}_{kt}(T_{t+k} + S_{t+k}) + \sum_{k=-(t-1)}^{N-t} \tilde{\omega}_{kt} e_{t+k}, \tag{3.6}$$

que define a aproximação linear para o componente sazonal, onde $\tilde{\omega}_{kt} = -\omega_{kt}$ para $k \neq 0$ e $\tilde{\omega}_{0t} = 1 - \omega_{0t}$. Os coeficientes $\omega_{kt}$ definem a aproximação linear para o EAS de acordo com a equação (2.11).

Adicionalmente prova-se, a partir do *Lema 1*, que

$$VAR(\hat{S}_t - S_t) \approx VAR_C\left( \sum_{k=-(t-1)}^{N-t} \tilde{\omega}_{kt} e_{t+k} \right). \tag{3.7}$$

Utilizando as equações (3.3) e (3.7), obtém-se:

$$VAR^{(1)}(\hat{N}_t) = VAR_C\left( \sum_{k=-(t-1)}^{N-t} \tilde{\omega}_{kt} e_{t+k} \right) + \lambda_0(1 - 2\tilde{\omega}_{0t}) - 2 \sum_{\substack{k=-(t-1) \\ k \neq 0}}^{N-t} \tilde{\omega}_{kt} \lambda_k \tag{3.8}$$

que é a expressão da variância do estimador da série sazonalmente ajustada.

Já a variância do estimador da tendência $VAR^{(2)}(\hat{N}_t)$ fica estabelecida por (3.5) e (3.7), ou seja,

$$VAR^{(2)}(\hat{N}_t) = VAR\{(\hat{S}_t - S_t) - e_t\} = VAR_C\left( \sum_{k=-(t-1)}^{N-t} \omega_{kt} e_{t+k} \right). \tag{3.9}$$

Equivalentemente,

$$VAR^{(2)}(\hat{N}_t) = \sum_{k=-(t-1)}^{N-t} \omega_{kt}^2 VAR_C(e_{t+k}) + 2\sum_{k<j}^{N} \omega_{kt}\omega_{jt} COV_C(e_{t+k}, e_{t+j}). \tag{3.10}$$

É interessante observar que, de (3.2) e (3.4) escreve-se, $D_{2t} = D_{1t} + I_t$, ou ainda, $D_{1t} = D_{2t} - I_t$. Calculando-se a variância, $VAR(D_{1t}) = VAR(D_{2t}) + VAR(I_t)$, ou ainda, de acordo com (2.11) e o *Lema 1* $VAR^{(1)}(\hat{N}_t) = VAR^{(2)}(\hat{N}_t) + VAR_M(\sum_{k=-(t-1)}^{N-t} \omega_{kt} I_{t+k})$ e, conseqüentemente,

$$VAR^{(1)}(\hat{N}_t) = VAR^{(2)}(\hat{N}_t) + \nu_0(1 - 2\omega_{0t}) - 2\sum_{\substack{k=-(t-1) \\ k\neq 0}}^{N-t} \omega_{kt}\nu_k \tag{3.11}$$

onde $\nu_k = E_M[I_t I_{t-k}]$ está definido em (2.2). Conclui-se, então, de (3.8) e (3.9), que para estimar as variâncias do estimador da série sazonalmente ajustada $VAR^3$ e do estimador de tendência $VAR^2(\hat{N}_t)$ **é preciso apenas obter estimativas da variância e das autocovariâncias do erro composto** $\{e_t\}$. Tal constatação é imediata com referência à equação (3.9). A mesma conclusão é obtida para a expressão (3.8) se o último termo da equação for desconsiderado.

# 4. Estimação das autocovariâncias do erro composto

## 4.1 O Método dos Momentos

Os momentos $V_k = E_C[e_t e_{t-k}]$ podem ser estimados a partir dos **resíduos** do X-11, $R_t = y_t - \hat{T}_t - \hat{S}_t$, no centro das séries ($24 < t < N-24$). Seja

$$R_t = \sum_{k=-(t-1)}^{N-t} a_{kt} y_{t+k} = \sum_{k=-(t-1)}^{N-t} a_{kt} M_{t+k} + \sum_{k=-(t-1)}^{N-t} a_{kt} e_{t+k} \tag{4.1}$$

a aproximação linear para os resíduos do X-11, onde $M_t = T_t + S_t$ é o sinal mensal na presença de ruído. Os momentos $V_k$ são estimados sob o seguinte postulado.

***Postulado***: No centro das séries o sinal médio, $\overline{M}_t = \sum_{k=-(t-1)}^{N-t} a_{kt} M_{t+k}$ satisfaz a condição $\overline{M}_t \approx 0$.

A admissibilidade deste postulado pode ser sustentada de diferentes formas. Observe-se primeiro que sob as propriedades em (2.12) $E_C(R_t) \approx 0$, tal que calculando o valor esperado de ambos os lados de (4.1) obtém-se $E_C(\overline{M}_t) = \overline{M}_t \approx 0$, ou seja, $R_t = y_t - \hat{T}_t - \hat{S}_t$. Pode-se, então, escrever $y_t - \hat{T}_t - \hat{S}_t = \sum_{k=-(t-1)}^{N-t} a_{kt} M_{t+k} + \sum_{k=-(t-1)}^{N-t} a_{kt} e_{t+k}$, ou ainda,

tem-se a partir de (4.1) que $(T_t - \hat{T}_t) + (S_t - \hat{S}_t) + e_t = \sum_{k=-(t-1)}^{N-t} a_{kt} M_{t+k} + \sum_{k=-(t-1)}^{N-t} a_{kt} e_{t+k}$.

Calculando o valor esperado da expressão acima,

$$E_C[T_t-\hat{T}_t]+E_C[S_t-\hat{S}_t]+E_C[e_t] = \sum_{k=-(t-1)}^{N-t} a_{kt}E_C[M_{t+k}]+\sum_{k=-(t-1)}^{N-t} a_{kt}E_C[e_{t+k}].$$

Sob a hipótese (2.12), obtém-se, $\sum_{k=-(t-1)}^{N-t} a_{kt}E_C[M_{t+k}]=0$, se e somente se, $E_C[M_{t+k}]=0$.

Assim, $E_C[\overline{M}_t]=\overline{M}_t = \sum_{k=-(t-1)}^{N-t} a_{kt}E_C[M_{t+k}] = 0.$

Pfeffermann (1994, p. 93) mostra, ainda, que uma outra forma mais direta de verificar a admissibilidade do postulado é estudar as propriedades dos pesos e a função de transferência do filtro estabelecido em (4.1).

A metodologia proposta baseia-se não só na hipótese de que $\overline{M}_t \approx 0$, mas, também, na estacionariedade no centro da série ($24<t<N-24$) dos resíduos $R_t$. Neste caso, estabelece-se o segundo *Lema*.

*Lema 2*: A série dos resíduos do X-11, $\{R_t, t=25,...,N-24\}$ é estacionária.

**A hipótese de estacionariedade de $R_t=\overline{E}_t$ permite a estimação da variância do erro combinado, $e_t$, das séries $\{R_t\}$ geradas pelo X-11**. Pode-se, então, expressar a variância e as autocovariâncias das séries $\{R_t\}$ como combinações lineares da variância e autocovariâncias das séries $\{e_t\}$, substituindo os momentos desconhecidos das séries $\{R_t\}$ pelas estimativas amostrais e resolvendo-se as equações resultantes para os momentos desconhecidos das séries $\{e_t\}$.

De (2.4) obtém-se que $E_C[e_te_{t-k}]=V_k=\lambda_k+\nu_k \quad k=0,1,...$, ou seja, $V_k=COV(e_t,e_{t-k}),. \quad k=0,1,..$. Entretanto, na prática, os momentos $V_k$ decaem para zero e pode-se assumir que após um ponto de corte $C$, tem-se $V_k \cong 0$. Seja, então, $\mathbf{U}'=(U_0,U_1,...,U_C)$ um vetor $1\times(C+1)$, com os elementos $U_k=COV(R_t,R_{t-k})$, $24<t<N-24$, $k=0,...,C$, e seja $\mathbf{V}'$ um vetor $1\times(C+1)$ de variância e covariâncias $\{V_k\}$. Tendo em vista que

$$R_t=\sum_{k=-(t-1)}^{N-t} a_k e_{t+k}, \text{ então } U_k=a_{k0}V_0+a_{k1}V_1+...+a_{kc}V_C = \mathbf{a}_{kc}'\mathbf{V}, \quad k=0,1,...,C, \tag{4.2}$$

onde os coeficientes $\{a_{kc}\}$ são diretamente definidos pelos coeficientes $\{a_k\}$.

O conjunto das equações (4.2) pode ser escrito como $\mathbf{U}=\mathbf{AV}$, onde $\mathbf{A}$ é uma matriz $(C+1)\times(C+1)$, fornecendo daí o conjunto de equações $\mathbf{V}=\mathbf{A}^{-1}\mathbf{U}$. Estimando $\mathbf{U}'$ por $\hat{\mathbf{U}}'=(\hat{U}_0,...,\hat{U}_C)$ onde

$$\hat{U}_k=\frac{1}{N-48}\sum_{t=25+k}^{N-25}(R_t-\overline{R})(R_{t-k}-\overline{R}) \tag{4.3}$$

produz-se os estimadores das covariâncias

$$\hat{\mathbf{V}}=\mathbf{A}^{-1}\hat{\mathbf{U}}. \tag{4.4}$$

A matriz A correspondente ao ponto de corte $C=13$ está descrita a seguir. A matriz (simétrica) correspondente ao ponto de corte $C<13$ é obtida a partir da matriz A omitindo-se as $14-C$ últimas linhas e colunas.

Matriz A : Coeficientes $\{a_{kC}, 0 \le k; c \le C\}$ para pontos de corte $C \le 13$, da Equação (4.2)

Pesos do filtro linear do componente irregular estimado pelo X-12-ARIMA.

| C=0 | C=1 | C=2 | C=3 | C=4 | C=5 | C=6 | C=7 | C=8 | C=9 | C=10 | C=11 | C=12 | C=13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0,4432 | -0,2447 | -0,1827 | -0,1018 | -0,0232 | 0,0343 | 0,0658 | 0,0099 | 0,0080 | 0,0069 | 0,0053 | 0,0027 | -0,1763 | 0,0926 |
| -0,1901 | 0,5141 | -0,1814 | -0,1305 | -0,0664 | -0,0091 | 0,0298 | 0,0302 | 0,0030 | 0,0025 | 0,0025 | 0,0023 | 0,0768 | -0,2026 |
| -0,1106 | -0,1517 | 0,5497 | -0,1507 | -0,1079 | -0,0546 | -0,0035 | 0,0164 | 0,0104 | -0,0008 | -0,0004 | 0,0006 | 0,0454 | 0,0625 |
| -0,0445 | -0,0957 | -0,1380 | 0,5611 | -0,1422 | -0,1033 | -0,0507 | -0,0075 | 0,0122 | 0,0049 | -0,0018 | -0,0011 | 0,0179 | 0,0396 |
| 0,0017 | -0,0414 | -0,0948 | -0,1382 | 0,5591 | -0,1436 | -0,1063 | -0,0524 | -0,0074 | 0,0125 | 0,0068 | -0,0010 | -0,0012 | 0,0168 |
| 0,0197 | -0,0020 | -0,0464 | -0,0995 | -0,1431 | 0,5563 | -0,1523 | -0,1069 | -0,0510 | -0,0050 | 0,0157 | 0,0116 | -0,0077 | 0,0005 |
| 0,0137 | 0,0181 | -0,0050 | -0,0494 | -0,1025 | -0,1454 | 0,5469 | -0,1519 | -0,1059 | -0,0494 | -0,0031 | 0,0187 | 0,0075 | -0,0063 |
| 0,0027 | 0,0138 | 0,0183 | -0,0052 | -0,0501 | -0,1031 | -0,1511 | 0,5481 | -0,1508 | -0,1049 | -0,0484 | -0,0019 | 0,0156 | 0,0072 |
| 0,0053 | 0,0005 | 0,0121 | 0,0171 | -0,0061 | -0,0504 | -0,1037 | -0,1504 | 0,5485 | -0,1505 | -0,1047 | -0,0481 | -0,0043 | 0,0163 |
| 0,0112 | -0,0015 | -0,0044 | 0,0095 | 0,0164 | -0,0054 | -0,0455 | -0,1040 | -0,1510 | 0,5477 | -0,1514 | -0,1055 | -0,0488 | -0,0020 |
| 0,0166 | 0,0000 | -0,0092 | -0,0079 | 0,0097 | 0,0188 | 0,0016 | -0,0471 | -0,1056 | -0,1526 | 0,5461 | -0,1529 | -0,1032 | -0,0452 |
| 0,0177 | 0,0037 | -0,0083 | -0,0121 | -0,0061 | 0,0140 | 0,0227 | -0,0011 | -0,0492 | -0,1074 | -0,1544 | 0,5445 | -0,1467 | -0,0992 |
| -0,1320 | 0,0937 | 0,0566 | 0,0184 | -0,0087 | -0,0190 | 0,0007 | 0,0229 | 0,0024 | -0,0441 | -0,1019 | -0,1493 | 0,5508 | -0,1773 |
| 0,0743 | -0,1713 | 0,0654 | 0,0396 | 0,0122 | -0,0060 | -0,0079 | 0,0062 | 0,0214 | 0,0007 | -0,0459 | -0,1039 | -0,1666 | 0,5618 |

O procedimento descrito em (4.4) é essencialmente uma aplicação do método dos momentos e, portanto, guarda as propriedades inerentes a este método. Em particular, quando $M_t = 0$, para todo $t$, no centro das séries, e $V_k = 0$ para $k > C$, os estimadores $\hat{v}$ são consistentes, no sentido que $\underset{N\to\infty}{pLim}\hat{V}_k = V_k \quad k = 0,1,...,C$, ou seja, existe um $N^*$ tal que $N$ é maior que $N^*$ de modo que $P\{\,|\,\hat{V}_k - V\,| > \xi\,\} < \delta$.

## 4.2 Especificação do ponto de corte baseado no conhecimento do esquema de rotação da amostra

O uso prático do procedimento descrito por (4.4) requer a especificação do ponto de corte $C$. De um modo geral, quanto maior o ponto de corte $C$, menor será o vício do estimador, porém maior será sua variância. Em muitas aplicações, valores admissíveis para o ponto de corte podem ser determinados com base na informação do desenho ou plano amostral. Assim, se as amostras são independentes (sem sobreposição) de um período de tempo para outro, e se os termos irregulares (estimados pelo X-11) são também não-correlacionados, então $V_k = 0$ para $k > 0$ tal que $C = 0$. Cabe ressaltar que os resíduos do X-11 são correlacionados negativamente, mesmo quando a série de entrada é um ruído branco, tal que $T_t + S_t = 0$. Tal fato pode ser verificado examinando-se a primeira coluna da matriz $A$, os resíduos resultantes do X-11 têm correlações negativas significativas nos *lags* 1, 2 e 12, isto é,

$$Corr(R_t, R_{t-1}) = {}^{-0,190}/_{0,443} = -0,429$$

$$Corr(R_t, R_{t-2}) = {}^{-0,111}/_{0,443} = -0,251$$

$$Corr(R_t, R_{t-12}) = {}^{-0,132}/_{0,443} = -0,298 .$$

Observe-se que $VAR(R_t) = 0.443\ VAR(I_t)$.

A existência de tais relações ilustram que os resíduos do X-11 não são consistentes com a hipótese de que os termos irregulares são realizações de um processo ruído branco.

Muitas pesquisas realizadas, atualmente, empregam planos amostrais com alguns elementos coincidentes ao longo do tempo, o que implica na existência de autocorrelações entre os erros amostrais das estimativas (série observada). É o que ocorre com a Pesquisa Mensal de Emprego do IBGE (PME/IBGE), que possui um esquema de rotação da amostra. Em cada grupo de rotação, os domicílios permanecem na amostra por quatro meses consecutivos, são retirados da pesquisa por um período de oito meses e retornam à amostra para mais quatro entrevistas consecutivas (detalhes sobre a pesquisa estão descritos na seção 5.2). Com a utilização deste plano amostral espera-se encontrar autocorrelações significativas de ordem 1, 2 e 3, pelo menos, isto ao considerar apenas o primeiro período de sobreposição da amostra. Na prática, entretanto, o ponto de corte recomendado assume, em geral, um valor menor do que cinco. Isto acontece por três razões:

as correlações seriais dos erros amostrais geralmente decaem para zero ao longo do tempo;

os coeficientes $\{a_{kC}\}$ da equação (4.2) são tipicamente muito pequenos para $0 \le k \le 5$, $C \ge k+3$, exceto o coeficiente $a_{0,13}$, conforme matriz $\mathbf{A}$; e

os pesos $\{\omega_{jt}\}$ utilizados para a estimação dos *EASs* têm valores significativos somente para os *lags* sazonais $\{t \pm 12n,\ n = 0,1,...\}$.

De acordo com o exposto, e considerando o esquema de rotação da amostra da PME/IBGE, utiliza-se ponto de corte $C = 4$, para a utilização do método de estimação de variâncias dos estimadores do X-11.

Pfeffermann e Scott (1997) introduziram novos aperfeiçoamentos na metodologia de estimação de variâncias dos estimadores do X-11, descrita nas seções anteriores, objetivando a implementação de dois principais desenvolvimentos:

1) As equações de estimação em (4.2) são modificadas para permitir o uso de todos os resíduos do X-11 para a estimação das covariâncias $V_k$ e não apenas os referentes às observações centrais da série.

2) A metodologia foi modificada para permitir o uso direto de estimativas da variância e covariâncias dos erros amostrais, quando disponíveis. Sob a metodologia atualizada, as equações de estimação estão associadas somente à variância e covariâncias das irregularidades na população. Uma vez que as irregularidades são correlacionadas somente em *lags* de baixa ordem (0, 1, 2 ou 3), o número de equações é substancialmente

reduzido, requerendo o uso somente das primeiras covariâncias empíricas dos resíduos do X-11 e, portanto, fazendo o processo de estimação mais estável.

## 4.3 Uso de toda a série de resíduos do X11 para o cálculo de variâncias e covariâncias

Assumindo que os estimadores dos sinais $M_t$ são não-viciados para toda a série dos termos combinados, com $M_t$ mantido fixo, segue da equação (4.1) que os sinais médios, $\overline{M}_t$, são suficientemente próximos de zero para todo $t$ e não apenas no centro das séries, ver Figura 2 em Pfeffermann (1994, p. 94).

Portanto, a aproximação linear $R_t = \sum_{k=-(t-1)}^{N-t} a_{kt} e_{t+k}$ mantém-se ao longo de toda a série, ainda que os resíduos não sejam integralmente estacionários, devido ao uso de filtros assimétricos ou pesos que variam ao longo do tempo, tanto no começo como no fim das séries.

A incorporação da nova hipótese e o processamento das modificações necessárias nas equações de estimação (4.2), permite o uso de toda a série de resíduos do X-11.

Inicialmente, note que $E[R_t] = 0$ e $COV(R_t, R_{t+m}) = E_C[R_t R_{t+m}]$ com $t = 1,\ldots, N-m$ e, conseqüentemente, pode-se representar $\tilde{U}_m = \frac{1}{N-m}\sum_{t=1}^{N-m} R_t R_{t+m}$ como um estimador não-viciado para $U_m = \frac{1}{N-m}\sum_{t=1}^{N-m} COV(R_t, R_{t+m})$. As covariâncias $U_{tm} = COV(R_t, R_{t+m})$ podem ser estimadas sob a suposição de que $V_k = COV(e_t, e_{t+m}) = 0$ para $k > C$, utilizando a aproximação linear para $R_t$ e $R_{t+m}$ definida em (4.1).

O conjunto de equações modificadas para estimação da variância e covariâncias, $V_k$, é conseqüentemente

$$\tilde{U}(m) = \tilde{b}_{m0} V_0 + \tilde{b}_{m1} V_1 + \ldots + \tilde{b}_{mC} V_C, \quad m = 0,1,\ldots,C, \text{ onde} \tag{4.5}$$

$b_{mj}$ é o coeficiente da combinação linear das covariâncias $V_m$ quando se calcula a média das covariâncias $U_{tm}$.

## 4.4 Melhoramentos para permitir o uso das covariâncias dos erros amostrais

Suponha que as covariâncias, $\lambda_k = COV(\varepsilon_t, \varepsilon_{t-1})$, dos erros amostrais são conhecidas para $k = 0,1,\ldots$. Na prática $\lambda_k$ é substituída por $\hat{\lambda}_k$, estimada, utilizando-se os dados amostrais. Como já mencionado, as estimativas de variância e covariâncias são rotineiramente calculadas pelas agências de estatísticas, mas raramente são incorporadas na análise de séries temporais, como apresentado neste trabalho. Adicionalmente, assumindo-se que o comportamento de erro composto $(\varepsilon_t + I_t)$ é dominado pelo componente de erro amostral $(\varepsilon_t)$, supõe-se que $COV(I_t, I_{t-1}) = 0$ se $k > Q$ para algum $Q$. Nos modelos estimados para os dados amostrais, as irregularidades são geralmente representadas como ruído branco. Ainda que esta suposição possa ser restritiva, parece razoável assumir que o valor de $Q$ seja baixo, variando no intervalo de zero a três. Considerando-se a propriedade (2.4),

tem-se $COV(e_t, e_{t-1}) = COV(\varepsilon_t, \varepsilon_{t-1}) + COV(I_t, I_{t-1}) = \lambda_k + v_k$. Segue que para $k > Q$, $COV(e_t, e_{t-1}) = \lambda_k$. Com os valores conhecidos de $\lambda_k$, o conjunto de equações definido por (4.5) pode ser rescrito como:

$$\left[U(m) - \sum_{j=0}^{C} b_{mj}\lambda_j\right] = \tilde{b}_{m0}v_0 + \tilde{b}_{m1}v_1 + \ldots + \tilde{b}_{mQ}v_Q, \quad m = 0,1,\ldots,Q \tag{4.6}$$

Finalmente, somando as estimativas $\hat{v}_j$ obtidas como solução de (4.6) às estimativas $\hat{\lambda}_j$ (calculadas externamente), produz-se as estimativas $\hat{V}_k$ necessárias para a estimação das variâncias dos EAS em (3.9) e (3.11). Observe-se que, se a série em consideração não é sujeita a erros amostrais, as equações definidas em (4.5) e (4.6) são iguais, com $C = Q$.

# 5. Aplicações

## 5.1 Introdução

A metodologia descrita nas seções anteriores permite a estimação de variâncias dos componentes não-observáveis estimados pelo X-12-ARIMA e incorpora, no cálculo destas variâncias, a correlação serial dos erros amostrais (caso existente) e também a correlação devida à variabilidade do componente irregular. No caso de aplicação desta metodologia a pesquisas não-amostrais, não há que se considerar a contribuição do erro amostral, mas unicamente a contribuição da variabilidade do componente irregular.

Para ilustração da utilidade do **M**étodo Geral para **E**stimação de **V**ariâncias dos Estimadores do **X**-12-ARIMA (designado ***MEV-X***), no caso de ajustamento sazonal de séries provenientes de pesquisas por amostragem, utilizou-se a Pesquisa Mensal de Emprego do IBGE, por ser uma pesquisa por amostragem probabilística com painéis rotativos. A metodologia será aplicada à série da taxa de desemprego aberto na semana para a Região Metropolitana de São Paulo.

Na seção 5.2, descreve-se as principais características da pesquisa que deu origem à serie escolhida para análise. Note-se que ausente o erro amostral, o método apresentado para a estimação de variâncias dos estimadores do X-12-ARIMA continua válido. O que ocorre, na verdade, é uma simplificação na formulação do modelo. Nesta situação, admite-se a existência de apenas uma fonte de variabilidade, aquela proveniente do componente irregular estimado pelo X-12-ARIMA.

## 5.2 A Pesquisa Mensal de Emprego (PME/IBGE)

### 5.2.1 Aspectos Gerais

A Pesquisa Mensal de Emprego do IBGE (PME/IBGE) fornece estimativas mensais relacionadas às características da força de trabalho como, por exemplo, a taxa de desemprego aberto, que é a relação entre o

número de pessoas desocupadas (procurando emprego) e o número de pessoas economicamente ativas[1] em um determinado período de referência; a taxa de atividade e a proporção de pessoas ocupadas por setor de atividade, para seis regiões metropolitanas, a saber: Rio de Janeiro, São Paulo, Belo Horizonte, Recife, Salvador e Porto Alegre. Os indicadores nacionais são produzidos por agregação (taxa média) das seis regiões metropolitanas pesquisadas.

A definição operacional de diversos indicadores usados baseia-se na condição de atividade das pessoas em determinados períodos de tempo, denominados períodos de referência. Define-se como semana de referência a semana de domingo a sábado que precede a semana da realização da entrevista. Adicionalmente, o período de referência de 30 dias corresponde aos 30 dias que antecedem a semana fixada para a entrevista. Finalmente, o mês de referência da pesquisa é aquele que antecede ao mês de realização da pesquisa.

Os indicadores divulgados pela PME/IBGE referem-se à situação das pessoas de 15 anos e mais de idade, numa dada semana de referência, com exceção da taxa de desemprego aberto, que é divulgada, também, para o período de 30 dias, e dos rendimentos que são relat[illegible]s ao mês de referência da pesquisa.

A PME/IBGE é uma pesquisa por amostra domiciliar: em cada região metropolitana, a amostra é selecionada independentemente em cada município, em dois estágios; os municípios constituem estratos, os setores censitários são as unidades primárias de seleção da pesquisa e os domicílios são as unidades amostrais do segundo estágio de seleção. A alocação da amostra de setores é proporcional ao longo dos estratos (municípios). As unidades de primeiro estágio, os setores, são selecionadas sistematicamente com probabilidade proporcional a uma medida de tamanho (números de domicílios no setor).

A PME/IBGE é uma pesquisa com rotação de amostra: os indicadores produzidos pela Pesquisa Mensal de Emprego são calculados através de média aritmética simples de quatro réplicas da amostra. Na seção 5.2.2, descreve-se o sistema de rotação da amostra utilizado na pesquisa mensal de emprego.

## 5.2.2 O Esquema de Rotação da Amostra da PME/IBGE

Em qualquer mês de pesquisa, a amostra da PME/IBGE é composta por quatro grupos de rotação que são conjuntos mutuamente exclusivos de setores censitários (unidades primárias de amostragem). O esquema de rotação é implementado mensalmente com o uso de quatro painéis, conjuntos de domicílios (unidades secundárias de amostragem) que entram e saem da amostra nos mesmos meses.

Os domicílios de um dado painel permanecem na amostra por quatro meses consecutivos (para a primeira, segunda, terceira e quarta visitas) sendo, nos oito meses subseqüentes, retirados da amostra. Após este período, os domicílios retornam à amostra (para a quinta, sexta, sétima e oitava visitas) pela última vez, e são retirados definitivamente da pesquisa. A cada mês um painel é retirado da amostra. O painel substituto pode ser completamente novo ou um que retorna depois de oito meses de ausência. O esquema de rotação 4-8-[illegible]z a

[1] [illegible]pulação Economicamente Ativa compreende o potencial de mão-de-obra com que pode contar o setor produtivo [illegible], a p[illegible]ulação ocupada e a população desocupada.

uma estrutura complexa de correlação para os erros amostrais ao longo dos anos. A Tabela 5.1 apresenta o esquema de rotação da amostra da PME/IBGE, indicando os painéis que compõem cada amostra mensal. O esquema de rotação da amostra da PME/IBGE permite uma coincidência substancial de unidades domiciliares investigadas mês a mês, isto é, para qualquer mês, 75% dos domicílios são comuns com o mês antecessor.

Entretanto, Pfeffermann, Silva e Freitas (2000) destacam o fato de que o esquema de rotação da PME não é um esquema regular ou estacionário, no sentido de que a distribuição dos domicílios que são entrevistados nos vários meses do ano não é constante no que se refere ao número de visitas ao domicílio, isto é, ao número de vezes que os domicílios já foram visitados ou entrevistados, conforme pode ser verificado na Tabela 5.1.

A diversidade de domicílios com número diferenciado de visitas pode acarretar na não-estacionariedade da série de estimativas. Por exemplo, pode-se supor que $Corr(y_{jan|2}, y_{fev|2}) \neq Corr(y_{jan|3}, y_{fev|3})$, onde $Corr(y_{jan|2}, y_{fev|2})$ é a correlação entre as duas estimativas, sendo $y_{jan|2}$ a taxa observada no mês de janeiro do ano 2 e $y_{fev|2}$ a taxa observada no mês de fevereiro do ano 2, já que o conjunto de painéis entrevistados nos meses de janeiro e fevereiro do ano 2 diferem daqueles entrevistados nos mesmos meses do ano 3, quanto ao seu tempo de permanência na pesquisa. Isto é devido ao fato das amostras, do ano 2 em janeiro e fevereiro, serem compostas por painéis com número de visitas de 1 a 4. Já as amostras do ano 3, nos mesmos meses, são compostas por painéis com número de visitas de 5 a 8.

Adicionalmente, pelo mesmo motivo, o problema ocorre, também, dentro de um mesmo ano: $Corr(y_{jul|1}, y_{ago|1}) \neq Corr(y_{nov|1}, y_{dez|1})$ apesar de ambas serem correlações de ordem 1 (*lag* 1). Nota-se que as amostras dos meses de julho e de agosto são compostas por painéis com número de visitas de 1 a 4, enquanto a amostra de novembro é composta por painéis com número de visitas de 3 a 6, e a de dezembro, por painéis com número de visitas de 4 a 7.

Nesses casos, a variável resposta e as respectivas correlações seriais podem ter comportamentos diferenciados para os períodos relacionados violando a suposição de estacionariedade. Esclarece-se que estudos empíricos estão sendo realizados para a verificação da ocorrência de vício de rotação na amostra da PME e que tal tópico foge ao escopo deste trabalho. Entretanto, neste caso, deve-se ter cautela ao analisar as estimativas das autocorrelações dos erros amostrais.

**Tabela 5.1 - Esquema de Rotação da PME/IBGE[2]**

| Ano | Mês | Painéis | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A1 | A2 | A3 | A4 | B1 | B2 | B3 | B4 | C1 | C2 | C3 | C4 | D1 | D2 | D3 | D4 | E1 | E2 | E3 | E4 | F1 | F2 | F3 | F4 |
| 1 | Jan | 4 | 3 | 2 | 1 | | | | | | | | | | | | | | | | | | | | |
| 1 | Fev | | 4 | 3 | 2 | 1 | | | | | | | | | | | | | | | | | | | |
| 1 | Mar | | | 4 | 3 | 2 | 1 | | | | | | | | | | | | | | | | | | |
| 1 | Abr | | | | 4 | 3 | 2 | 1 | | | | | | | | | | | | | | | | | |
| 1 | Mai | | | | | 4 | 3 | 2 | 1 | | | | | | | | | | | | | | | | |
| 1 | Jun | | | | | | 4 | 3 | 2 | 1 | | | | | | | | | | | | | | | |
| 1 | Jul | | | | | | | 4 | 3 | 2 | 1 | | | | | | | | | | | | | | |
| 1 | Ago | | | | | | | | 4 | 3 | 2 | 1 | | | | | | | | | | | | | |
| 1 | Set | | | | | | | | | 4 | 3 | 2 | 1 | | | | | | | | | | | | |
| 1 | Out | 5 | | | | | | | | | 4 | 3 | 2 | | | | | | | | | | | | |
| 1 | Nov | 6 | 5 | | | | | | | | | 4 | 3 | | | | | | | | | | | | |
| 1 | Dez | 7 | 6 | 5 | | | | | | | | | 4 | | | | | | | | | | | | |
| 2 | Jan | 8 | 7 | 6 | 5 | | | | | | | | | | | | | | | | | | | | |
| 2 | Fev | | 8 | 7 | 6 | 5 | | | | | | | | | | | | | | | | | | | |
| 2 | Mar | | | 8 | 7 | 6 | 5 | | | | | | | | | | | | | | | | | | |
| 2 | Abr | | | | 8 | 7 | 6 | 5 | | | | | | | | | | | | | | | | | |
| 2 | Mai | | | | | 8 | 7 | 6 | 5 | | | | | | | | | | | | | | | | |
| 2 | Jun | | | | | | 8 | 7 | 6 | 5 | | | | | | | | | | | | | | | |
| 2 | Jul | | | | | | | 8 | 7 | 6 | 5 | | | | | | | | | | | | | | |
| 2 | Ago | | | | | | | | 8 | 7 | 6 | 5 | | | | | | | | | | | | | |
| 2 | Set | | | | | | | | | 8 | 7 | 6 | 5 | | | | | | | | | | | | |
| 2 | Out | | | | | | | | | | 8 | 7 | 6 | 1 | | | | | | | | | | | |
| 2 | Nov | | | | | | | | | | | 8 | 7 | 2 | 1 | | | | | | | | | | |
| 2 | Dez | | | | | | | | | | | | 8 | 3 | 2 | 1 | | | | | | | | | |
| 3 | Jan | | | | | | | | | | | | | 4 | 3 | 2 | 1 | | | | | | | | |
| 3 | Fev | | | | | | | | | | | | | | 4 | 3 | 2 | 1 | | | | | | | |
| 3 | Mar | | | | | | | | | | | | | | | 4 | 3 | 2 | 1 | | | | | | |
| 3 | Abr | | | | | | | | | | | | | | | | 4 | 3 | 2 | 1 | | | | | |
| 3 | Mai | | | | | | | | | | | | | | | | | 4 | 3 | 2 | 1 | | | | |
| 3 | Jun | | | | | | | | | | | | | | | | | | 4 | 3 | 2 | 1 | | | |
| 3 | Jul | | | | | | | | | | | | | | | | | | | 4 | 3 | 2 | 1 | | |
| 3 | Ago | | | | | | | | | | | | | | | | | | | | 4 | 3 | 2 | 1 | |
| 3 | Set | | | | | | | | | | | | | | | | | | | | | 4 | 3 | 2 | 1 |
| 3 | Out | | | | | | | | | | | | | 5 | | | | | | | | | 4 | 3 | 2 |
| 3 | Nov | | | | | | | | | | | | | 6 | 5 | | | | | | | | | 4 | 3 |
| 3 | Dez | | | | | | | | | | | | | 7 | 6 | 5 | | | | | | | | | 4 |
| 4 | Jan | | | | | | | | | | | | | 8 | 7 | 6 | 5 | | | | | | | | |
| 4 | Fev | | | | | | | | | | | | | | 8 | 7 | 6 | 5 | | | | | | | |
| 4 | Mar | | | | | | | | | | | | | | | 8 | 7 | 6 | 5 | | | | | | |
| 4 | Abr | | | | | | | | | | | | | | | | 8 | 7 | 6 | 5 | | | | | |
| 4 | Mai | | | | | | | | | | | | | | | | | 8 | 7 | 6 | 5 | | | | |
| 4 | Jun | | | | | | | | | | | | | | | | | | 8 | 7 | 6 | 5 | | | |
| 4 | Jul | | | | | | | | | | | | | | | | | | | 8 | 7 | 6 | 5 | | |
| 4 | Ago | | | | | | | | | | | | | | | | | | | | 8 | 7 | 6 | 5 | |
| 4 | Set | | | | | | | | | | | | | | | | | | | | | 8 | 7 | 6 | 5 |

**FONTE: Pfeffermann, Silva e Freitas (2000).**

## 5.2.3 Análise sazonal da série da taxa de desemprego na Região Metropolitana de São Paulo

É fundamental verificar se a série em questão apresenta sazonalidade, cuja existência é pressuposto básico para elaboração de ajustamentos sazonais e, portanto, para utilização da metodologia aqui apresentada. A análise foi elaborada utilizando o programa de ajuste sazonal X-12-ARIMA. Foi confirmada a presença de sazonalidade identificável e boas condições para a realização do ajustamento sazonal sob o modelo aditivo. Na seqüência são apresentados os gráficos da evolução dos fatores sazonais, da série dessazonalizada e da curva de tendência, Figuras 5.1, 5.2 e 5.3, respectivamente.

[2] Cada cela exibe o número de visitas já realizadas ao domicílio, no painel correspondente.

**Figura 5.1 – Evolução dos fatores sazonais da série da taxa de desemprego aberto**
**Região Metropolitana de São Paulo**
**Janeiro de 1991 a Dezembro de 1997**

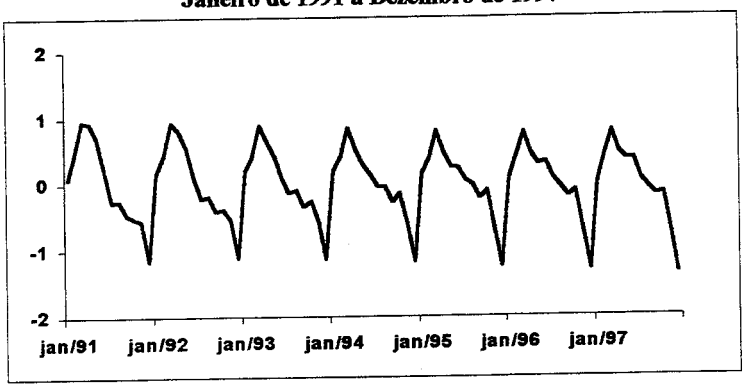



**Figura 5.2 – Série da taxa de desemprego aberto sazonalmente ajustada**
**Região Metropolitana de São Paulo**
**Janeiro de 1991 a Dezembro de 1997**

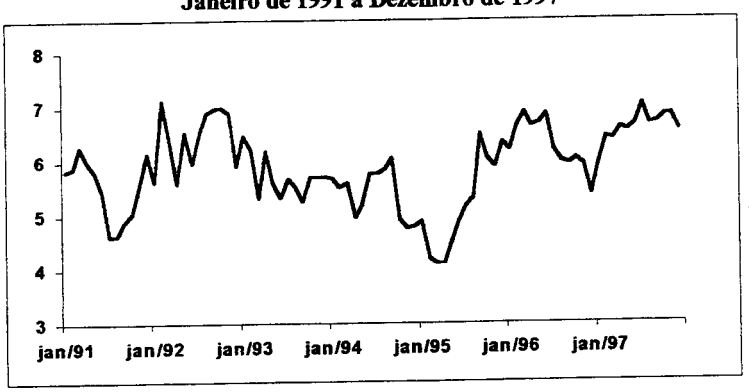
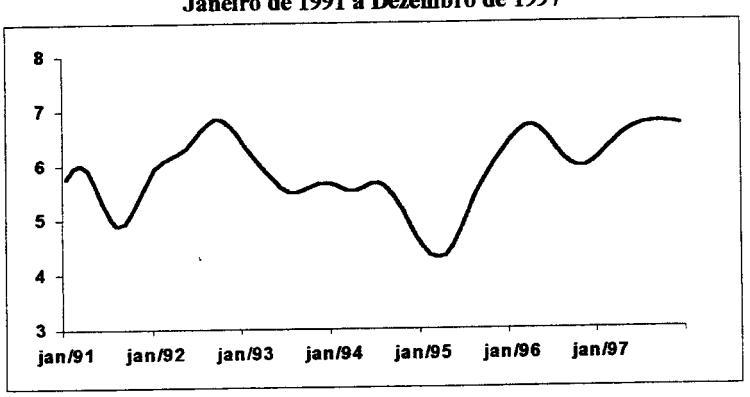



**Figura 5.3 – Série de tendência da taxa de desemprego aberto**
**para a Região Metropolitana de São Paulo**
**Janeiro de 1991 a Dezembro de 1997**




Para uma melhor compreensão do nível do desemprego deve ser considerado, necessariamente, o fenômeno sazonal. Assim, as séries dessazonalizada (Figura 5.2) e de tendência (Figura 5.3) são de enorme valia

para o estudo da evolução do desemprego. No entanto, uma indicação da precisão das estimativas destes componentes precisa ser estabelecida. Na seção 5.2.4, o método de estimação de variâncias dos estimadores do X-12-ARIMA (MEV-X) é aplicado à serie da taxa de desemprego aberto na semana para a Região Metropolitana de São Paulo. Variâncias das séries dessazonalizada e de tendência são disponibilizadas, sendo, então, possível determinar os Coeficientes de Variação - CV - das estimativas destes componentes, estabelecendo-se uma indicação da precisão das mesmas.

### 5.2.4 Estimativas de desvios-padrão para os estimadores de ajustamento sazonal e tendência na PME/IBGE

Estabeleceu-se, anteriormente, as expressões para o cálculo das variâncias (ou desvios-padrão) para o estimador de ajustamento sazonal:

$$VAR^{(1)}(\hat{N}_t) = VAR^{(2)}(\hat{N}_t) + v_0(1-2\omega_{0t}) - 2\sum_{\substack{k=-(t-1)\\k\neq 0}}^{N-t}\omega_{kt}v_k, \qquad \text{equação (3.11)}$$

e para o estimador de tendência:

$$VAR^{(2)}(\hat{N}_t) = \sum_{k=-(t-1)}^{N-t}\omega_{kt}^2 VAR_C(e_{t+k}) + 2\sum_{k<j}^{N}\omega_{kt}\omega_{jt}COV_C(e_{t+k}, e_{t+j}), \qquad \text{equação (3.10).}$$

A covariância do erro composto, $COV_C(e_{t+k}, e_{t+j}) = V_k = \lambda_k + v_k$, pode ser estimada a partir dos resíduos $R_t = y_t - \hat{T}_t - \hat{S}_t$ do X-12-ARIMA, utilizando-se a equação, $\hat{\mathbf{V}} = \mathbf{A}^{-1}\hat{\mathbf{U}}$, onde $\mathbf{A}$ é a matriz dos pesos do componente irregular e $\hat{U}_m = \frac{1}{N-m}\sum_{t=1}^{N-m}COV(R_t, R_{t+m})$, $m = 0,1,...,C$, conforme apresentado na seção 4.3.

Por sua vez, as autocovariâncias dos erros amostrais $\lambda_k$, podem ser estimadas utilizando-se as autocorrelações observadas dos pseudo-erros amostrais. Os pseudo-erros amostrais são definidos como $\varepsilon_t^j = (y_t^j - y_t)$, onde $y_t = \sum_{j=1}^{M} y_t^j \Big/ 4$ e $y_t^j$ é a estimativa amostral baseada no painel $j$, $j = 1,...,4$. Assumindo que as séries $\{\varepsilon_t^j, \quad t = 1,...,N\}$ são estacionárias, as autocorrelações dos erros amostrais podem ser obtidas, de acordo com Pfeffermann, Feder e Signorelli (1998), por:

$$\rho_k = CORR(\varepsilon_t, \varepsilon_{t-k}) = \sum_{j=1}^{4} c_k^j \Big/ \sum_{j=1}^{4} c_0^j, \quad k = 1,2,... . \qquad (5.1)$$

onde $c_k^j = \left[\sum_{t=k+1}^{N}(\varepsilon_t^j - \varepsilon^j)(\varepsilon_{t-k}^{j,t} - \varepsilon^{j,k}) \Big/ N\right]$, $\varepsilon^j = \left[\sum_{t=1}^{N}\varepsilon_t^j \Big/ N\right]$ e $\varepsilon^{j,k} = \left[\sum_{t=k+1}^{N+k}\varepsilon_{t-k}^{j,t}\right]$ tal que $c_k^j$ é a autocovariância de *lag* $k$ da série $\varepsilon^{(j)}$ de erro amostral. As autocovariâncias $\lambda_k$ podem, então, ser definidas como

$$\lambda_k = \sum_{j=1}^{4} c_k^j = \rho_k \sum_{j=1}^{4} c_0^j = \rho_k \lambda_0, \qquad (5..2)$$

sendo $\lambda_0$ a variância dos erros amostrais.

Finalmente, a variância dos erros amostrais $\lambda_0$, pode ser estimada utilizando-se o método de grupos aleatórios, Wolter (1985, p. 20). Como $y_t = 1/4\sum_{j=1}^{4} y_t^j$, então, calculando-se a variância de $y_t$, obtém-se, $VAR(\varepsilon_t) = VAR(y_t) = 1/16\sum_{j=1}^{4} VAR(y_t^j)$. Isto porque sob a teoria da amostragem, se $y_t$ é estimador não-viciado de $Y_t$, tem-se $y_t = Y_t + \varepsilon_t$ com $E(\varepsilon_t) = 0$. Então, $VAR(\varepsilon_t) = VAR(y_t - Y_t)$ ou equivalentemente, $VAR(\varepsilon_t) = VAR(y_t)$. Mas $VAR(y_t^j) = 1/3\sum_{j=1}^{4}(y_t^j - y_t)^2$. Sendo assim,

$$VAR(y_t) = (1/16)(4)\left[1/3\sum_{j=1}^{4}(y_t^j - y_t)^2\right] = \frac{1}{(4).(3)}\sum_{j=1}^{4}(y_t^j - y_t)^2 .$$ Finalmente $\lambda_0 = \frac{1}{N}\sum_{t=1}^{N}\left[\frac{1}{(4).(3)}\sum_{j=1}^{4}(y_t^j - y_t)^2\right]$.

Uma vez estimadas as autocovariâncias dos erros amostrais $\lambda_k$, as autocovariâncias dos resíduos (componente irregular) do X-12-ARIMA $\nu_k$ podem ser estimadas utilizando-se a equação $U_k - \mathbf{A}\lambda_k = \mathbf{A}\nu_k$, resultante da multiplicação da equação (4.4) pela matriz $\mathbf{A}$.

As estimativas de variâncias do erro composto são calculadas considerando-se $V_k = \lambda_k + \nu_k$, com o ponto de corte fixado em $k = 4$, conforme desenho amostral da PME/IBGE, cujo esquema de rotação é [4-8-4].

A variância do erros amostrais da taxa de desemprego aberto na semana para a Região Metropolitana de São Paulo para o período de janeiro de 1990 a junho de 1998, calculada pelo método de grupos aleatórios, é 0,1343. As autocorrelações são significativas apenas para os *lags* 4, 12 e 24, então as autocovariâncias de interesse $\lambda_k$, $k$=0,1,...,15 assumem os valores $\lambda_0 = 0.1343$, $\lambda_4 = 0{,}0313$, $\lambda_{12} = 0{,}0340$. Para os demais *lags* as autocovariâncias não são significativamente diferentes de zero. Ressalta-se que, pelo esquema de rotação da PME [4-8-4], esperava-se obter autocorrelações significativamente diferentes de zero pelo menos até o *lag* 15, considerando-se que um domicílio é entrevistado pela última vez na pesquisa 15 meses após a primeira entrevista (ver Tabela 5.1).

As medidas das variâncias são baseadas nos "filtros ou pesos combinados", isto é, nos filtros utilizados pelo X-11 (os pesos $a_{kt}$ da matriz $\mathbf{A}$) considerando-se, também, os filtros de extrapolação ARIMA[3], baseados nos modelos estimados pelo X-12-ARIMA.

A seguir, são apresentados os resultados da aplicação da metodologia estudada, designada MEV-X, na série da taxa de desemprego aberto na semana para a Região Metropolitana de São Paulo.

---

[3] Como os valores residuais extremos são excluídos do cálculo das covariâncias $\hat{U}_k$, equação (4.4), o MEV-X pode ser utilizado fixando-se os coeficientes do modelo ARIMA para os valores estimados, e modificando-se os filtros EAS empregados no final das séries, considerar o uso dos valores extrapolados.

**Figura 5.4 – Desvios-padrão para as estimativas de ajustamento sazonal da taxa de desemprego aberto para a Região Metropolitana de São Paulo, obtidas pelo MEV-X janeiro de 1991 a dezembro de 1997**

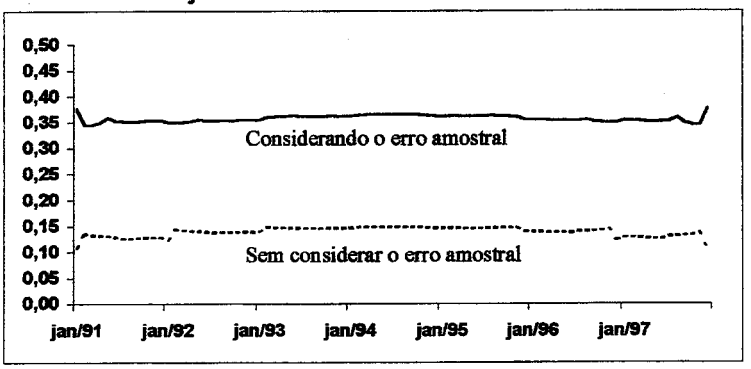



**Figura 5.5 - Série sazonalmente ajustada da taxa de desemprego aberto para a Região Metropolitana de São Paulo e correspondentes intervalos de confiança de 95%, considerando o erro amostral janeiro de 1991 a dezembro de 1997**

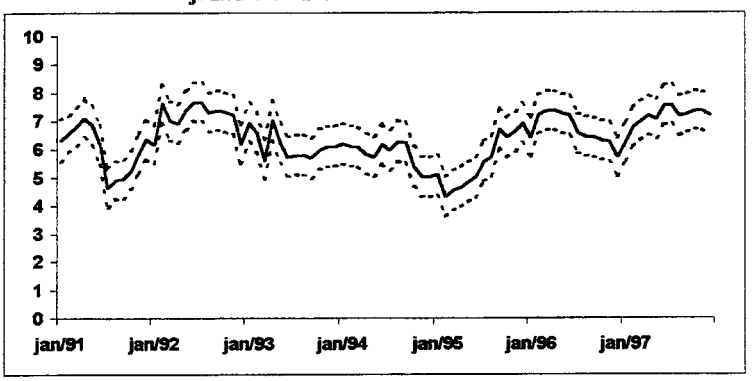



**Figura 5.6 – Desvios-padrão para as estimativas de tendência da taxa de desemprego para a Região Metropolitana de São Paulo, obtidas pelo MEV-X janeiro de 1991 a dezembro de 1997**

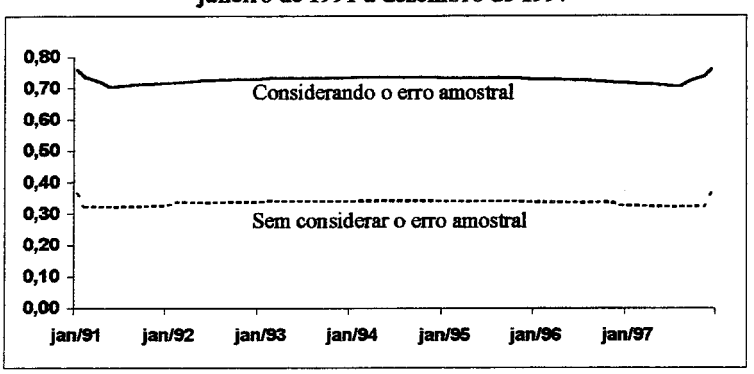
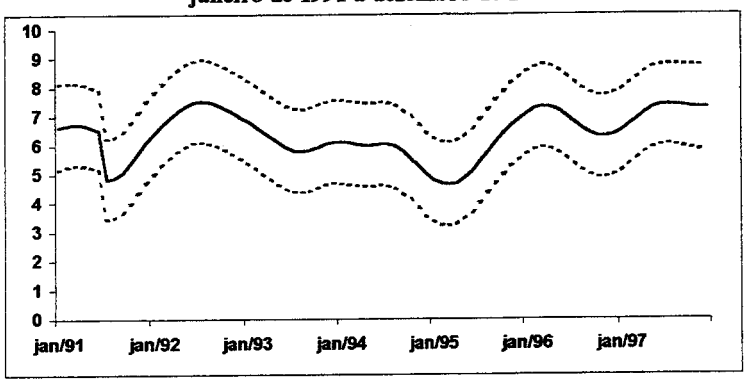


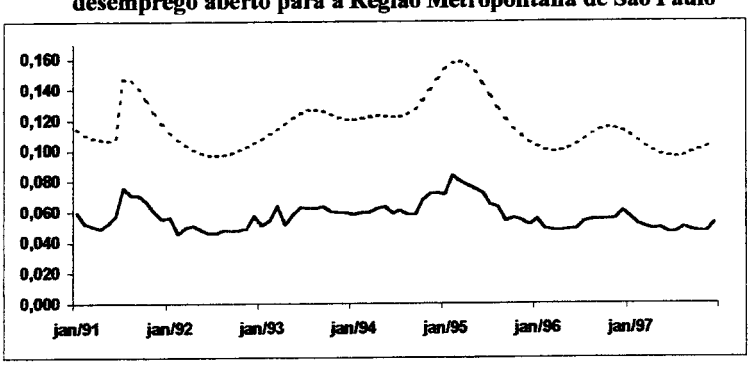
**Figura 5.7 - Série de tendência da taxa de desemprego aberto para a Região Metropolitana de São Paulo e correspondentes intervalos de confiança de 95%, considerando o erro amostral janeiro de 1991 a dezembro de 1997**




Evidências apresentadas nas Figuras 5.4 e 5.6 indicam que ignorar o erro amostral implica considerável subestimação da variabilidade dos estimadores das séries de tendência e sazonalmente ajustada. Os desvios-padrão da taxa de desemprego aberto sazonalmente ajustada flutuam em torno de 0,35 ponto percentual e da tendência da taxa em torno de 0,70 ponto percentual, quando considerado o erro amostral. É importante ressaltar que para o cálculo de variâncias para séries dessazonalizadas são consideradas duas fontes de variação, a saber, o erro amostral e a sazonalidade, enquanto para o cálculo de variâncias da tendência, três fontes de variação são incluídas: o erro amostral, os resíduos do modelo e a sazonalidade. Por este motivo, as variâncias da tendência são, em geral, maiores que as variâncias das séries ajustadas para sazonalidade.

Na Figura 5.8, estão representados os Coeficientes de Variação - CVs - das estimativas da série da taxa de desemprego aberto sazonalmente ajustada e da tendência da taxa de desemprego aberto, considerando o erro amostral. O *CV* fornece uma indicação da precisão relativa das estimativas de interesse.

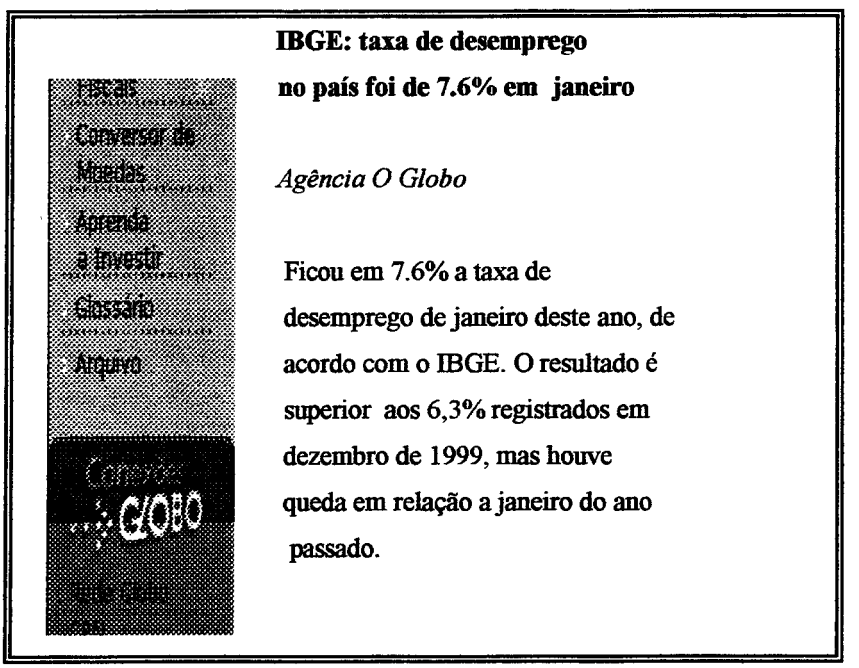
**Figura 5.8 – Coeficientes de variação para as estimativas de ajustamento sazonal e de tendência da taxa de desemprego aberto para a Região Metropolitana de São Paulo**




___ *CVs* da Tendência ___ *CVs* de Ajustamento Sazonal

Para exemplificar a importância das variâncias de ajustamento sazonal e tendência, disponibilizadas pelo método de estimação de variâncias dos estimadores do X-12-ARIMA, para análise da evolução da série taxa de desemprego aberto na semana da PME/IBGE, registre-se a notícia veiculada na imprensa, através do jornal O Globo descrita na Figura 5.9. Atualiza-se a notícia veiculada, de acordo com a introdução de informações imprescindíveis para a análise do fenômeno observado: o desemprego. A primeira contribuição é dispor de informações livres das flutuações sazonais, isto é, a série do desemprego sazonalmente ajustada. Adicionalmente, incorpora-se na análise uma medida de precisão: o coeficiente de variação. Observa-se, então, como supostamente a notícia veiculada alteraria à medida que uma nova informação fosse incorporada a análise.

**Figura 5.9 – Informe Econômico de O Globo capturado em 26 novembro de 2000**




**FONTE: www.globo.com/noticias/arquivo**

A análise apresentada na Figura 5.9 não considera que a taxa de desemprego aberto na semana para todas as áreas é caracterizada por marcantes movimentos sazonais, e por isso a comparação direta das estimativas do mês atual contra mês anterior não é apropriada. Uma forma mais adequada de analisar tal informação poderia ter sido feita considerando-se a série da taxa de desemprego aberto sazonalmente ajustada, isto é, série em que os efeitos periódicos previsíveis são removidos e que revelam as novas variações na tendência do indicador.

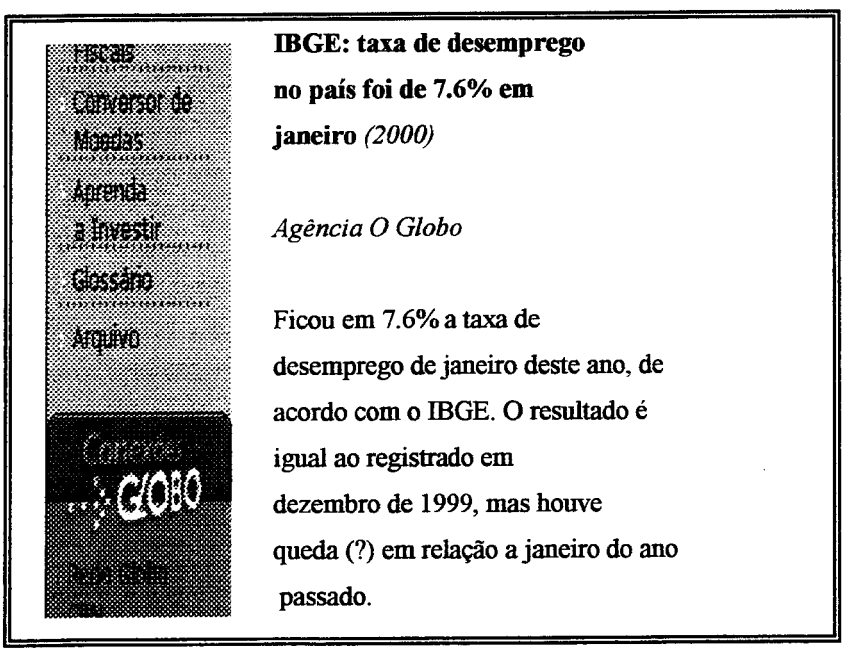
Considerando-se os valores da taxa de desemprego aberto na semana para todas as áreas, sazonalmente ajustada, Tabela 5.1, propõe-se repetir a análise apresentada na Figura 5.9, destacando-se o problema encontrado naquela abordagem.

**Tabela 5.1 – Taxa de desemprego aberto na semana para todas as áreas: valores observados e sazonalmente ajustados, para meses selecionados**

| Meses Selecionados | Série Observada | Série Sazonalmente Ajustada |
|---|---|---|
| Jan./99 | 7.7 | **7.7** |
| Nov./99 | 7.0 | **7.9** |
| Dez./99 | 6.3 | **7.6** |
| Jan./00 | 7.6 | **7.6** |

**FONTE: IBGE - Pesquisa Mensal de Emprego.**

Como seria a análise considerando a taxa de desemprego sazonalmente ajustada:




**IBGE: taxa de desemprego no país foi de 7.6% em janeiro** *(2000)*

*Agência O Globo*

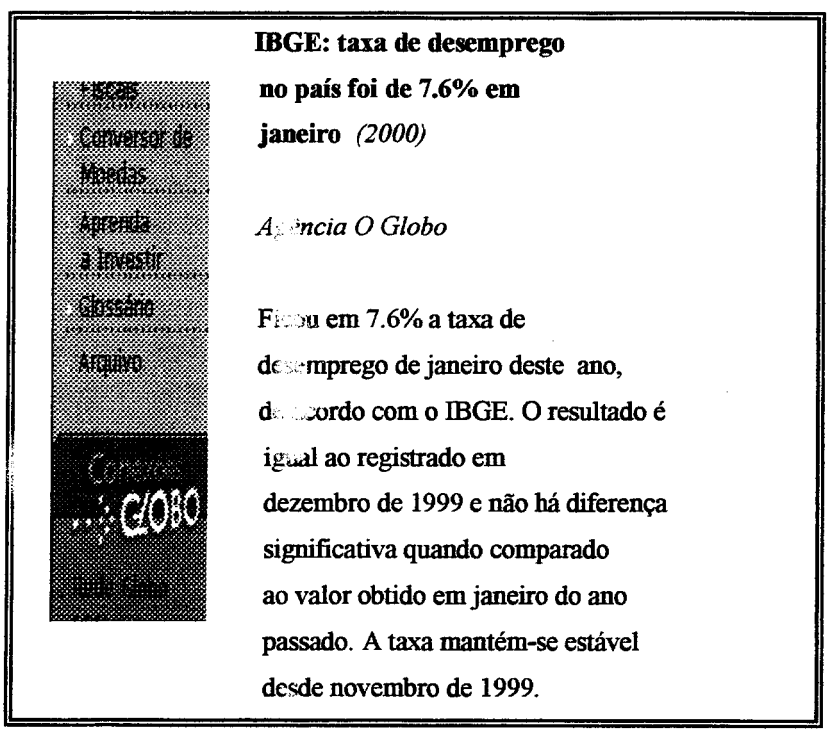
Ficou em 7.6% a taxa de desemprego de janeiro deste ano, de acordo com o IBGE. O resultado é igual ao registrado em dezembro de 1999, mas houve queda (?) em relação a janeiro do ano passado.

Para efeito de exercício, assume-se que os desvios-padrão, determinados pelo MEV-X, da série taxa de desemprego aberto na semana, sazonalmente ajustada para a Região Metropolitana de São Paulo possa representar um *proxy* para os desvios-padrão da série da taxa de desemprego aberto na semana para todas as áreas, sazonalmente ajustada (Indicador Nacional dessazonalizado).

**Tabela 5.2 – Intervalos de confiança de 95% e coeficientes de variação para a taxa de desemprego aberto na semana para todas as áreas, sazonalmente ajustada**

| Meses selecionados | Série sazonalmente ajustada | Intervalos de confiança | Coeficientes de variação (%) |
|---|---|---|---|
| Jan./99 | 7,7 | (7,0 ; 8,4) | 4,94 |
| Nov./99 | 7,9 | (7,2 ; 8,6) | 4,81 |
| Dez.99 | 7,6 | (6,9 ; 8,3) | 5,00 |
| Jan./00 | 7,6 | (6,9 ; 8,3) | 5,00 |

Então, incorporando a informação disponibilizada na Tabela 5.2 à análise em questão, pode-se rescrevê-la, da seguinte forma:




**IBGE: taxa de desemprego no país foi de 7.6% em janeiro** *(2000)*

*Agência O Globo*

Ficou em 7.6% a taxa de desemprego de janeiro deste ano, de acordo com o IBGE. O resultado é igual ao registrado em dezembro de 1999 e não há diferença significativa quando comparado ao valor obtido em janeiro do ano passado. A taxa mantém-se estável desde novembro de 1999.

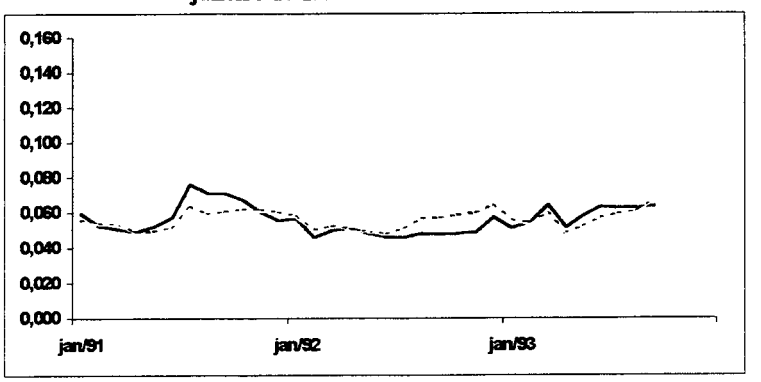
A pequena análise elaborada mostra que, o conhecimento das informações sobre a variabilidade das estimativas de ajustamento sazonal, é fundamental para a compreensão correta do fenômeno observado.

# 6. Conclusão

Duas importantes características do método proposto por Pfeffermann (1994) são, por um lado, a simplicidade que permite a direta aplicabilidade da metodologia e, por outro lado, a generalidade. Apesar da

simplicidade da abordagem utilizada, o MEV-X produz resultados satisfatórios. Por exemplo, considere-se as estimativas de coeficientes de variação de ajustamento sazonal, obtidas por Silva (1996, p. 159)[4], calculadas para a taxa de desemprego aberto na semana para a Região Metropolitana de São Paulo, utilizando um modelo estrutural básico multivariado (BSM) para o sinal, na representação de espaço de estados e modelos auto-regressivos multivariados [VAR(1) e VARMA(1,1)] para o **erro amostral**. Na Figura 6.1, são comparados os resultados obtidos pelo MEV-X e pelos modelos BSM+VAR(1) e BSM+VARMA(1,1), para o período de janeiro de 1991 a setembro de 1993.

**Figura 6.1 - Coeficientes de variação para as estimativas de ajustamento sazonal da taxa de desemprego aberto na semana para a Região Metropolitana de São Paulo janeiro de 1991 a setembro de 1993**




____MEV-X _ _ _BSM+VAR(1) ____ BSM+VARMA(1,1)

A análise da Figura 6.1 permite concluir que as estimativas de coeficientes de variação, produzidas pelo método **univariado** de estimação de variâncias dos estimadores do X-12-ARIMA (MEV-X), estão em níveis próximos aos das estimativas obtidas pela análise **multivariada**[5] de séries temporais, representada pelos modelos BSM+VAR(1) e BSM+VARMA(1,1). Isto é, abordagens diferentes chegam a resultados similares, ratificando os resultados obtidos pela aplicação do MEV-X. O procedimento de ajustamento sazonal atualmente utilizado pela PME/IBGE é o X-12-ARIMA, portanto, é imprescindível a utilização de estimativas de variâncias

[4] Silva desenvolveu um método de séries temporais para ajustar dados de composições provenientes de uma pesquisa amostral repetida no tempo. O termo composição é utilizado para definir um vetor cujos componentes assumem valores não negativos representando proporções de um dado total. Neste caso, cada um dos componentes está definido no intervalo [0,1] e sua soma é igual a um. Dados com esta características são obtidos, numa pesquisa, quando a variável de interesse admite uma resposta multinomial e a pesquisa visa a obter estimativas para a proporção de unidades classificadas em cada uma das categorias. Por exemplo, numa pesquisa sobre a força de trabalho, investiga-se a proporção de pessoas classificadas como empregadas, desempregadas e fora da força de trabalho. O método desenvolvido produz estimativas dos componentes de tendência e sazonalidade das séries multivariadas observadas, respeitando as características específicas das composições e o desenho amostral da pesquisa. O procedimento de ajuste/análise foi aplicado à Pesquisa Mensal de Emprego - PME.

[5] A abordagem multivariada considera que os vários componentes (vetor de proporções) correlacionam-se simultaneamente.

dos estimadores de ajustamento sazonal e de tendência do X-12-ARIMA para o conhecimento integral do fenômeno do desemprego no País.

Neste trabalho, tratou-se da implementação do método de estimação de variâncias dos estimadores do X-11 e variantes, **para o caso de séries com estrutura aditiva**, ($y_t = T_t + S_t + I_t$). Uma outra extensão importante do método é para o caso de decomposição multiplicativa, ($y_t = T_t \times S_t \times I_t$). A utilização de séries com estrutura multiplicativa produz resultados similares, quando aplica-se a transformação logarítmica ($\log y_t = \log T_t + \log S_t + \log I_t$). Assumindo que $\log(y_t/\hat{S}_t)$ é uma variável aleatória normal, as variâncias podem ser estimadas recorrendo-se às relações entre as variâncias das distribuições normal e log-normal. Ressalte-se que a suposição de normalidade é geralmente não-restritiva, considerando-se as trocas dos valores extremos embutidos no programa X-11. Para diferentes (mas conhecidas) distribuições do erro composto, sob a decomposição log-aditiva, as variâncias dos *EASs* podem ser estimadas por *bootstrapping* paramétrico.

A suposição de que a variância dos erros amostrais é constante ao longo do tempo pode, ocasionalmente, ser muito restritiva. Mudanças nas variâncias de aleatorização podem ser devidas às mudanças no plano amostral ou às alterações no nível das séries. Nestes casos, as covariâncias $\hat{U}_k$, dos termos do erro composto, podem ser modificadas para incorporar as mudanças nas variâncias de aleatorização (desenho amostral). Pfeffermann, Morry e Wong (1995) propõem ampliações que incluem o uso da decomposição multiplicativa e uma compensação para as mudanças na variância e covariâncias dos termos do erro composto. Adicionalmente, examinam como as variâncias dos estimadores de ajustamento sazonal são afetadas pela identificação e estimação dos modelos ARIMA utilizados para a extrapolação das séries observadas e pela identificação e gradual troca das observações extremas (procedimento padrão do X-11).

Bell e Kramer (1996) desenvolveram uma abordagem onde se assume que o alvo do ajustamento sazonal do X-11 é aquele resultante da aplicação de filtros lineares simétricos (utilizados pelo X-11), se as séries não contenham erros amostrais. O objetivo é obter variâncias de ajustamentos sazonais do X-11 considerando-se duas fontes de erro. A primeira fonte de erro é o erro amostral (aleatorização), estimado externamente, pela modelagem do erro amostral. A segunda fonte de erro resulta da necessidade de estender-se às séries temporais com previsões (para frente e para trás), através de um modelo ARIMA, antes da utilização dos filtros simétricos do X-11.

Vale, então, registrar, que uma continuidade natural deste trabalho, passa por considerar as abordagens mencionadas, implementá-las, comparar suas propriedades e descrever, se for o caso, as diferenças nos resultados obtidos.

## Referências bibliográficas

**BELL, W.R.; KRAMER, M. *Toward variances for X-11 seasonal adjustments*. Washington D.C.: Bureau of the Census, 1996. 44 p. (Statistical Research Report Series, n. RR96/07). Disponível em: http://www.census.gov/srd/www/byyear. html. Acesso em: 17 mar. 2000.**

BELL, W.R.; MONSELL, B.C. *X-11 Symmetric linear filters and their transfer functions.* Washington D.C.: Bureau of the Census, 1992. 49 p. (Statistical Research Report Series. n. RR92/15). Disponível em: http://www.census.gov/srd/ www/byyear.html. Acesso em: 17 mar. 2000.

CRUZ, M.M. *Estimação de variâncias para séries dessazonalizadas pelo método X-12-ARIMA, considerando o desenho amostral.* Rio de Janeiro: IBGE, Escola Nacional de Ciências Estatísticas, 2001. 194 p.

DAGUM, E.B. Moving averages. In: KOTZ, S.; JOHNSON, N.L. (Ed.). *Encyclopedia of statistical science.* New York: John Wiley & Sons, 1985. Vol. 5, p. 630-634.

DAGUM, E.B. *X-11-ARIMA/88 seasonal adjustment method:* foundations and users manual. Ottawa: Statistics Canada, Time Series Research and Analysis, 1988. 144 p.

DUNCAN, G.J.; KALTON, G. Issues of design and analysis of surveys across time. *International Statistical Review*, Voorburg, Holanda, v. 55, p. 1, 97-111, 1987.

ESTADOS UNIDOS. Bureau of Census. *X-12-ARIMA reference manual.* Version 0.2. Washington D.C., 1998. 165 p. Disponível em: http://www.census. gov/srd/www/x12a/. Acesso em: 21 ago. 1998.

FINDLEY, D.F. et al. New capabilities and methods of the X-12-ARIMA seasonal-adjustment program. *Journal of Business and Economic Statistics*, Alexandria, VA, v. 16, n. 2, p. 127-177, abr. 1998.

METODOLOGIA da pesquisa mensal de emprego 1980. Rio de Janeiro: IBGE, 82 p. (Relatórios metodológicos, v. 2).

PFEFFERMANN, D. A general method for estimating the variances of X-11 seasonally adjusted estimators. *Journal of Time Series Analysis*, v. 15, n. 1, p. 85-116, 1994.

PFEFFERMANN, D.; FEDER, M.; SIGNORELLI, D. Estimation of autocorrelations of survey errors with application to trend estimation small areas. *Journal of Business & Economic Statistics*, Washington D.C., v. 16, n. 3. p. 339-347. 1998.

PFEFFERMANN, D.; MORRY, M.; WONG, P. Estimation of the variances of X-11-ARIMA seasonally adjusted estimators for a multiplicative decomposition and heteroscedastic variances. *International Journal of Forecasting*, v. 11, p. 271-283, 1995.

PFEFFERMANN, D.; SCOTT, S. Variance measures for X-11 seasonally adjusted estimators: some new developments with application to labor force series. *Annual Meetings of the American Statistical Association*, Anaheim, CA, 1997.

PFEFFERMANN, D.; SILVA, L.N.S.; FREITAS, M.P.S. *Implicações do esquema de rotação da pesquisa mensal de emprego do IBGE na qualidade das estimativas publicadas.* Rio de Janeiro: IBGE, Departamento de metodologia 2000. 24 p.

SAS INSTITUTE. *SAS/IML software:* changes and enhancements through. Release 6.11. Cary, NC, 1996. 312 p.

SHISKIN, J.; YOUNG, A.H.; MUSGRAVE, J.C. *The X-11 variant of the census method II seasonal adjustment program.* Washington D.C.: Bureau of Census, Department of Commerce, 1967. 66 p. (Technical Paper n. 15).

SILVA, D.B.N. *Modelling compositional time series from repeated surveys.* 1996. 237 p. Tese (Doutorado) - University of Southampton. Faculty of Mathematical Studies.

WALLIS, K.F. Seasonal adjustment and relations between variables. *Journal of American Statistical Association*, Washington D.C., v. 69, p. 18-31, 1974.

WOLTER, K.M. *Introduction to variance estimation.* New York: Springer-Verlag., 1985. 427 p.

WOLTER, K.M.; MONSOUR, N.J. On the problem of variance estimation for a deseasonalized series. In: INTERNATIONAL SYMPOSIUM ON SURVEY SAMPLING, 1980, Carleton. *Current topics in survey sampling.* New York: Academic Press, 1981. p. 367-403.

## Abstract

**Several time series regularly used as a source of information for planning and research are obtained from repeated sample surveys. Although there has been widespread use of sample estimates from such surveys in time series analysis, it is not always the case that the precision of the time series estimates is available for assessment and publication. When the focus is on the estimation of unobservable components of a time series, such as trend and seasonals, the need for seasonal adjustment procedures arises. One very well known method is the X-11 procedure (Shiskin, Young and Musgrave, 1967). However, when using the X-11 and its successors, the problem of how to estimate variances of the unobservable components and the seasonally adjusted series appears. The ability to evaluate the precision of these estimates is crucial for data analysis and interpretation. Moreover, when dealing with time series derived from repeated surveys, variance estimation procedures must take into account the sampling error.**

**The main objective of this work is to provide standard errors for trend and seasonally adjusted regional unemployment rate series in Brazil obtained by the X-12-ARIMA procedure (Findley et al, 1998). The standard errors are estimated based on Pfeffermann (1994).**

# 7. Apêndice

- **Prova do Lema 1:**

Para verificar o *Lema 1* pode-se utilizar a seguinte decomposição para as variâncias $VAR^{(i)}(\hat{N}_t)$:

$$VAR^{(i)}(\hat{N}_t) = E_{T,S}\{VAR_C(\hat{N}_t)\} + VAR_{T,S}\{E_C(\hat{N}_t)\}, \quad i = 1,2$$

onde $E_{T,S}\{.\}$ e $VAR_{T,S}\{.\}$ definem a esperança e a variância sob todas as possíveis realizações dos componentes tendência ($T_t$) e sazonalidade ($S_t$).

Sendo o EAS dado por $\hat{N}_t = y_t - \hat{S}_t$, onde $y_t = Y_t + \varepsilon_t$ e $Y_t = T_t + S_t + I_t$, tem-se $\hat{N}_t = T_t + S_t + I_t + \varepsilon_t - \hat{S}_t$. Como $e_t = \varepsilon_t + I_t$, então,

$\hat{N}_t = T_t + e_t - (\hat{S}_t - S_t)$. Substituindo em (3.6) resulta em:

$$VAR^{(i)}(\hat{N}_t) = E_{T,S}\{VAR_C[T_t + e_t - (\hat{S}_t - S_t)]\} + VAR_{T,S}\{E_C[T_t + e_t - (\hat{S}_t - S_t)]\}$$

$VAR^{(i)}(\hat{N}_t) = E_{T,S}\{VAR_C[e_t - (\hat{S}_t - S_t)]\} + VAR_{T,S}\{T_t + E_C[e_t - (\hat{S}_t - S_t)]\}$ e considerando (3.2) e (3.4), obtém-se:

$$VAR^{(i)}(\hat{N}_t) = E_{T,S}\{VAR_C(D_{it})\} + VAR_{T,S}\{T_t + E_C[(D_{it})\}$$

Como $E_C(D_{it}) \cong 0$ conforme (2.1), (2.4) e (2.12) e $E_{T,S}\{VAR_C(D_{it})\} = VAR_C(D_{it})$ porque as variâncias $VAR_C(D_{it})$ dependem exclusivamente das distribuições dos erros amostrais e termos irregulares. Portanto,

$VAR^{(i)}(\hat{N}_t) \cong VAR_C(D_{it}), \quad i = 1,2$. Ou seja,

$VAR^{(1)}(\hat{N}_t) \cong VAR_C(D_{1t}), \quad D_{1t} = \varepsilon_t - (\hat{S}_t - S_t)$

$VAR^{(2)}(\hat{N}_t) \cong VAR_C(D_{2t}), \quad D_{2t} = e_t - (\hat{S}_t - S_t)$

- **Prova do Lema 2:**

A hipótese de estacionariedade de $R_t$ pode ser verificada considerando-se o postulado e equação (4.1).

Sendo $\overline{M}_t \approx 0$, tem-se $R_t = \overline{E}_t = \sum_{k=-(t-1)}^{N-t} a_k e_{t+k}$. Como a série $\{e_t\}$ é estacionária, ver equação (2.4) e os pesos $\{a_t\}$, no centro das séries, são invariantes no tempo, conclui-se, então, que $\{R_t\}$ é também estacionária.

# Modelos aditivos generalizados: metodologia e prática

Liliam Pereira de Lima *

Carmen Diva Saldiva de André*

Julio da Motta Singer*

**Resumo**

**Os Modelos Aditivos Generalizados - MAG - constituem uma extensão dos Modelos Lineares Generalizados - MLG -, na qual o efeito de cada variável preditora na variável resposta é modelado de forma bastante flexível por meio de uma função não especificada. Neste trabalho apresentamos uma conceituação desses modelos, bem como procedimentos de ajuste e testes de hipótese. Para ilustração, essa metodologia é aplicada a um estudo ambiental cujo objetivo é descrever a associação entre mortalidade fetal tardia e poluição atmosférica na cidade de São Paulo. Os resultados são comparados com aqueles obtidos via MLG para o mesmo conjunto de dados.**

**Palavras-chave: estudos epidemiológicos ambientais, métodos de suavização, Modelos Lineares Generalizados, regressão não-paramétrica.**

## 1. Introdução

Estudos epidemiológicos realizados em diferentes centros de pesquisa têm detectado associações significativas entre morbimortalidade por causas respiratórias e poluição atmosférica em populações urbanas (Schwartz, 1994, Saldiva et al., 1995, Braga et al., 1999 e Singer et al., 2002, por exemplo). As populações mais vulneráveis são as constituídas por crianças, idosos e pessoas que apresentam doenças respiratórias prévias (Dockery e Pope, 1994, Saldiva et al., 1995 ou Gouveia e Fletcher, 2000, por exemplo). A maior parte desses estudos tem base populacional, e um grupo, ao invés de um indivíduo, constitui a unidade de observação

* Endereço para correspondência: Departamento de Estatística. Instituto de Matemática e Estatística da Universidade de São Paulo.

(Morgenstern, 1995). Esse grupo é acompanhado ao longo do tempo. Os estudos desta classe envolvem, em geral, a observação de eventos como mortalidade, internações hospitalares ou sintomas respiratórios, além de observações de variáveis explicativas como temperatura ou concentração de algum poluente atmosférico. Essa estrutura de dados é menos suscetível a variáveis de confusão como tabagismo, pressão arterial ou fatores socio-econômicos (Rothman e Greenland, 1998), que não variam de dia para dia com a poluição atmosférica embora sejam possíveis confundidoras quando se comparam populações de localizações geográficas distintas sujeitas a diferentes níveis de poluição (Schwartz, 1994 e André et al., 2000). Em geral, variáveis temporais e climáticas são consideradas como componentes confundidores nesse contexto.

Neste tipo de estudo, é comum adotar uma estratégia de análise que consiste em, inicialmente, construir um modelo básico (incluindo somente as variáveis de controle sazonal e meteorológico) que explique ao máximo a variabilidade da resposta, e então adicionar a(s) variável(eis) relativa(s) à(s) concentrações do(s) poluente(s). O objetivo desta estratégia é evitar que um controle inadequado das variáveis confundidoras possa mascarar a associação entre a resposta e a concentração do poluente. Além disso, os efeitos das concentrações dos poluentes e das variáveis climáticas, caso existam, podem não incidir necessariamente no mesmo dia em que ocorre o evento de interesse (óbito ou internação), ou seja, o número de óbitos ou internações ocorridos no dia de hoje pode ser uma conseqüência das condições meteorológicas e da poluição não apenas de hoje, mas também de alguns dias anteriores. Por este motivo, é comum utilizarem-se modelos com defasagem ou médias móveis das variáveis meteorológicas e dos poluentes, ou ainda, modelos com defasagens distribuídas (Zanobetti, 2000).

Sob essa ótica, em um trabalho pioneiro, Pereira et al. (1998) estudaram a associação entre mortalidade fetal tardia (natimortalidade) e concentração de $NO_2$, CO, $SO_2$, $PM_{10}$ e $O_3$ sob diferentes defasagens, e entre mortalidade fetal tardia e médias móveis das concentrações desses poluentes, nos períodos de 2 até 14 dias, precedendo o registro da morte com base em dados diários referentes aos anos de 1991 e 1992. A avaliação da associação entre mortalidade fetal tardia (variável resposta) e concentrações de um determinado poluente (variável independente) foi baseada em um modelo de regressão de Poisson que pertence à classe dos Modelos Lineares Generalizados - MLG. Foram consideradas no modelo, as variáveis de controle geralmente utilizadas em pesquisas de mortalidade e morbidade por causas respiratórias (Saldiva et al., 1995 e Braga et al., 1999, por exemplo): meses do ano, dias da semana, temperatura mínima diária e umidade relativa do ar às 12 horas. O único poluente para o qual foi detectada uma associação significativa ($p = 0{,}001$) com a resposta, mostrando comportamento dose-dependente, foi o $NO_2$, cujo efeito foi avaliado por meio da média móvel de cinco dias das suas concentrações ($M5NO_2$). O modelo final adotado por Pereira et al. (1998) foi

$$\begin{aligned} log\, E[NATMOR] = {} & \alpha + \beta_1 FEV91 + \ldots + \beta_{23} DEZ92 + \\ & \beta_{24} SEG + \ldots + \beta_{29} SÁB + \beta_{30} M2TEMP + \beta_{31} M2UMID + \\ & \beta_{32} TEMP1 + \ldots + \beta_{34} TEMP3 + \beta_{35} UMID1 + \ldots + \\ & \beta_{37} UMID3 + \beta_{38} M5NO_2a + \ldots + \beta_{41} M5NO_2d \end{aligned} \qquad (1)$$

onde

• *NATMOR* é o número diário de natimortos (variável resposta),

• *FEV91,...,DEZ92* representam 23 variáveis indicadoras dos meses no período do estudo cuja finalidade é controlar a sazonalidade de longa duração; o mês de janeiro de 1991 é a categoria de referência para os demais meses,

• *SEG,...,SÁB* representam seis variáveis indicadoras dos dias da semana cuja finalidade é controlar a sazonalidade de curta duração; domingo é a categoria de referência para os demais dias da semana,

• *M2TEMP e M2UMID* representam, respectivamente, médias móveis de dois dias (média dos valores dessas variáveis no dia e dia anterior ao registro da morte) da temperatura mínima e umidade relativa do ar,

• *TEMP1,...,TEMP3 e UMID1,...,UMID3* são, respectivamente, variáveis categorizadas para controle adicional da temperatura mínima e umidade relativa, correspondentes a quatro classes delimitadas pelos seus quartis. As categorias de referência consideradas foram aquelas formadas pelos menores valores de temperatura e umidade,

• $M5NO_2a,...,M5NO_2d$ representam variáveis indicadoras dos cinco intervalos de classes delimitados pelos quintis da variável $M5NO_2$, considerando a classe composta pelas menores concentrações como referência,

• $\beta_1,...,\beta_{41}$ são os parâmetros (coeficientes no modelo de regressão) a serem estimados.

Os resultados do ajuste do modelo (1) são apresentados no Apêndice (Tabela A.1).

Nesse contexto, a utilização de variáveis categorizadas em substituição a variáveis originalmente contínuas deve-se ao fato de que o tipo de relação existente entre a resposta e o poluente pode ser não-linear e além disso, de difícil especificação. Motivo semelhante justifica a introdução de variáveis categorizadas para controle adicional da temperatura e umidade, evitando que uma relação paramétrica entre essas variáveis e a resposta precise ser introduzida no modelo. Esse procedimento de categorização das variáveis originalmente contínuas acarreta perda de informação e está sujeito a critérios de classificação subjetivos. Um procedimento alternativo consiste em adotar um modelo não-paramétrico no qual a relação entre a resposta e cada uma das variáveis é ditada pelos próprios dados. Esses são os Modelos Aditivos Generalizados - MAG - descritos por Hastie e Tibshirani (1990), cuja forma geral é

$$g[\, E(Y|X_1,...,X_p)\,] = \alpha + f_1(X_1) + ... + f_p(X_p), \qquad (2)$$

onde $Y$ é uma variável resposta, $X_1,...,X_p$ são variáveis preditoras, $g(\bullet)$ é uma função de ligação que relaciona a média da resposta com as variáveis preditoras, $\alpha$ é um parâmetro a ser estimado e $f_1,...,f_p$ são funções não especificadas a serem estimadas.

Tanto o ajuste dos MAG quanto testes de hipóteses sobre seus componentes foram desenvolvidos em analogia a procedimentos utilizados com esses objetivos nos MLG, modificando-os de forma que as funções $f_1,...,f_p$ em (2) sejam estimadas por meio de suavizadores. Por este motivo, os métodos de suavização são de extrema importância no processo de ajuste dos MAG.

Nosso objetivo é discutir aspectos técnicos e práticos dos MAG, incluindo estratégias de construção e ajuste de modelos, além de testes de hipóteses sobre seus componentes e sua aplicação a um problema prático. Com essa finalidade, discutiremos brevemente os métodos de suavização na seção 2. Na seção 3 descreveremos o processo de ajuste dos MAG. Na seção 4 repetiremos a análise feita por Pereira et al. (1998) adotando um MAG. Finalmente, conclusões e considerações finais são apresentadas na seção 5.

# 2. Métodos de suavização

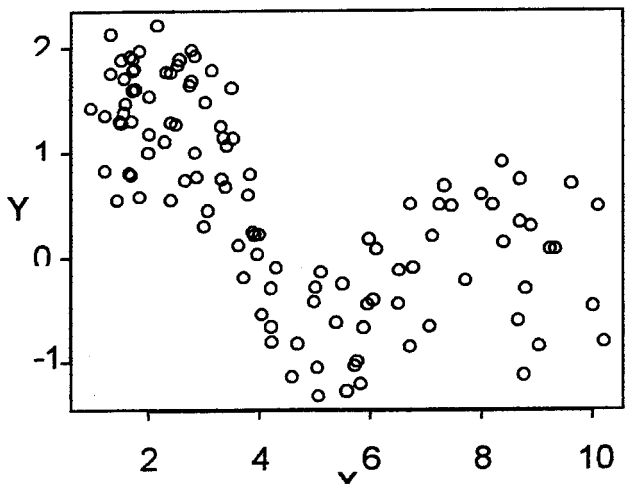
## 2.1 Introdução

Um suavizador (ou alisador) é uma ferramenta que descreve a variação da média de uma variável $Y$ como função de uma ou mais variáveis $X_1, \ldots, X_p$. Quando a variação da média de Y é descrita em função de apenas uma variável $X$, o alisador é denominado *unidimensional*. Quando p variáveis, $X_1, \ldots, X_p$, são consideradas, diz-se que o alisador é *multidimensional*.

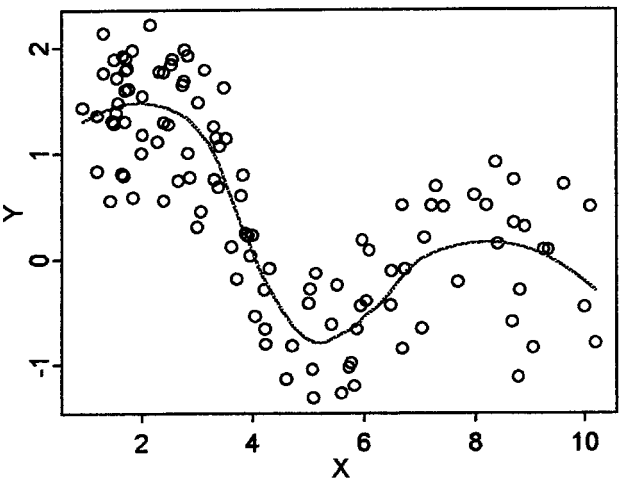
Como ilustração, considere o diagrama de dispersão apresentado na Figura 1a e sejam $Y$ uma variável aleatória e $X$ uma variável não-estocástica. A relação entre essas variáveis é melhor visualizada com o auxílio da curva representada na Figura 1b, obtida por intermédio de um procedimento de suavização. A *curva suavizada* ou *curva ajustada* é construída com base nos pontos $(x_i, \hat{y}_i)$, onde $\hat{y}_i$ é o valor previsto (pela suavização) de $Y$ para $X = x_i$. Esses valores são obtidos sem a adoção de um modelo paramétrico relacionando $Y$ e $X$.

**Figura 1 - Diagramas de dispersão de X e Y e curva suavizada pelo método *loess***

**(1a) Diagrama de dispersão de X e Y**




**(1b) Diagrama de dispersão de X e Y com curva suavizada pelo método *loess***



Formalmente, um *suavizador* é uma função de $\boldsymbol{x} = (x_1,\ldots,x_n)'$ e $\boldsymbol{y} = (y_1,\ldots,y_n)'$, digamos $s(x) = S(y|x)$, com mesmo domínio de $\boldsymbol{x}$. Para alguns suavizadores, $s(x_o)$ é definida para todo $x_o$. Outras vezes ela é definida apenas para os valores observados de $X$, e neste caso, algum tipo de interpolação é necessário para obter estimativas associadas a outros valores de $X$.

Muitas vezes, o alisador é utilizado com o objetivo de ajustar o modelo

$$y_i = f(x_i) + \varepsilon_i, \quad i=1,\ldots,n, \tag{3}$$

onde $f$ é uma função não especificada e os $\varepsilon_i$ são erros aleatórios distribuídos independentemente com média zero e variância $\sigma^2$. Este modelo pode ser considerado uma generalização do modelo de regressão linear simples que tem $Y$ como variável resposta e $X$ como variável preditora.

Na maioria das técnicas de suavização, o valor suavizado $\hat{y}_i$ é obtido com base em uma "média" de $r$ observações na vizinhança de um dado valor $x_i$. Diferentes formas de cálculo dessa média em uma vizinhança de $x_i$ definem diferentes métodos de suavização.

A escolha do *tamanho da vizinhança* é um problema importante no processo de suavização. Ele é associado a um parâmetro $(\lambda)$ denominado *parâmetro de suavização* que deve ser fixado antes do início do processo. A escolha de valores para este parâmetro depende da relação entre o *viés* e a *variância* da curva estimada, que serão definidos na seção 2.3.

Um suavizador é dito linear quando o vetor de valores previstos $\hat{\mathbf{y}} = (\hat{y}_1,\ldots,\hat{y}_n)' = (\hat{f}(x_1),\ldots,\hat{f}(x_n))' = \hat{\mathbf{f}}$ pode ser escrito como

$$\hat{\mathbf{f}} = \boldsymbol{S}\boldsymbol{y}, \tag{4}$$

onde $\boldsymbol{S} = \{s_{ij}\}$ é uma matriz de dimensão $n \times n$ chamada *matriz suavizadora*, que depende apenas de $X$ e do parâmetro de suavização $\lambda$. O valor ajustado de $Y$ em $x_i$ é

$$\hat{y}_i = \hat{f}(x_i) = s_{x_i 1}y_1 + s_{x_i 2}y_2 + \ldots + s_{x_i n}y_n, \quad i=1,\ldots,n,$$

onde $(s_{x_i 1},\ldots,s_{x_i n})$ é a $i$-ésima linha da matriz $\boldsymbol{S}$.

A literatura sobre alisadores lineares é bastante ampla. Buja et al. (1989) e Hastie e Tibshirani (1990) descrevem vários desses alisadores (como o *cubic spline* ou o *loess – locally-weighted scatterplot smoother*) e apresentam bibliografia adicional sobre o assunto. Na próxima seção descreveremos o alisador *loess* que será empregado no exemplo prático aqui analisado.

## 2.2 Suavizador *loess*

Suponha que se deseja suavizar a relação entre duas variáveis, $X$ e $Y$, com base num conjunto de pontos $(x_i,y_i)$, $i=1,\ldots,n$. O *loess*, proposto por Cleveland (1979), é um método de suavização que se baseia no ajuste sucessivo de n modelos de regressão pelo método de Mínimos Quadrados Ponderados - MQP. Cada modelo é ajustado considerando observações cujos valores de $X$ pertencem a uma vizinhança da coordenada $x_i$ de uma

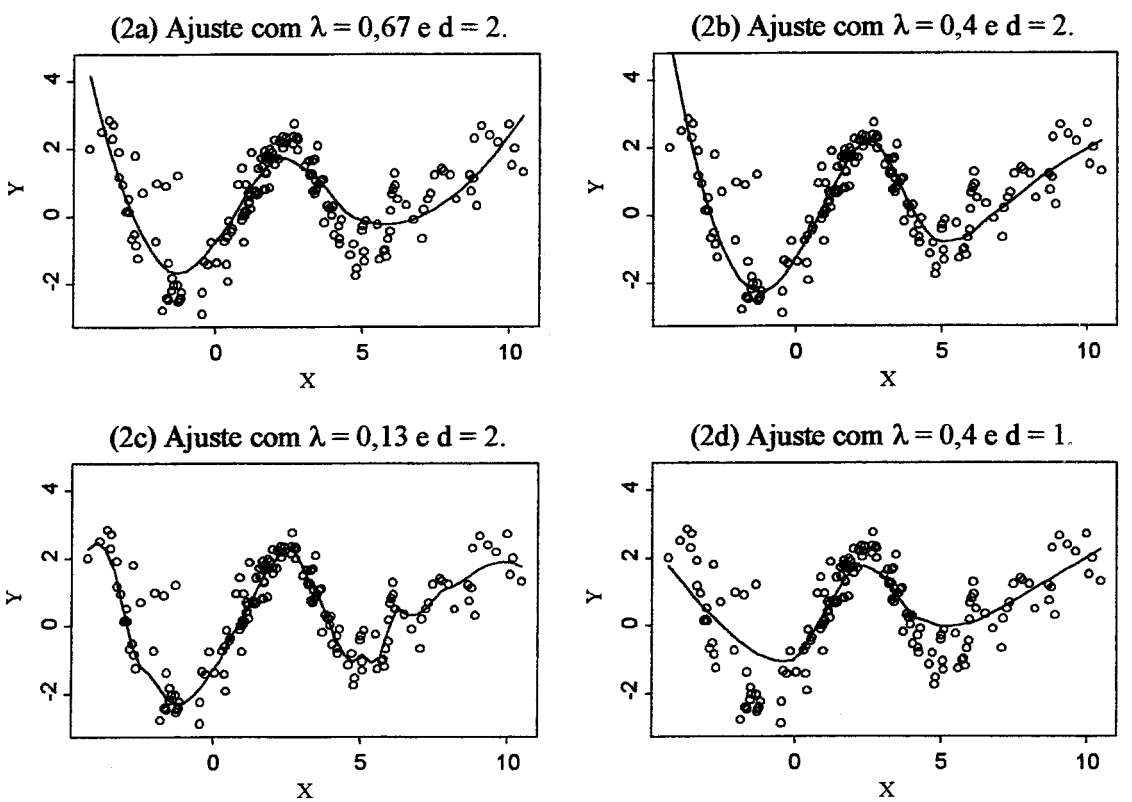
observação $(x_i,y_i)$ fixada e denominada *ponto alvo*, $i=1,\ldots,n$. O valor ajustado é $\hat{y}_i = \hat{f}(x_i)$. Portanto, considerando sucessivamente as n observações $(x_i,y_i)$ como ponto alvo, obtêm-se os pontos $(x_i, \hat{f}(x_i))$, $i=1,\ldots,n$, que geram a curva suavizada.

Para cada ponto alvo $(x_i,y_i)$ define-se uma vizinhança, e aos pontos $(x_j,y_j)$ nessa vizinhança é ajustado um polinômio de grau $d$, $y_j = \alpha + \beta_1 x_j + \ldots + \beta_d x_j^d + e_j$, $j=1,\ldots,n$, por MQP sendo os pesos associados a cada um desses ajustes locais obtidos por meio de uma função que será definida adiante.

A vizinhança de cada $(x_i,y_i)$ é constituída pelos $r$ pares de observações $(x_j,y_j)$, com as coordenadas $x_j$ mais próximas a $x_i$. O número de pontos $r$ a ser considerado é $r = \lambda n$, onde $\lambda$ $(0 < \lambda \leq 1)$ é o parâmetro de suavização que corresponde à proporção do número total de observações a ser utilizado em cada ajuste local.

Não existe um critério rígido para a escolha do valor de $\lambda$ e muitas vezes esta escolha é feita empiricamente, com a seleção de vários valores para um mesmo conjunto de dados. O parâmetro de suavização tem influência fundamental na variância e no viés da curva estimada: aumentar $\lambda$ implica aumentar a suavização da curva (diminuir a variância) e perder informação no ajuste (aumentar o viés). Este efeito é ilustrado na Figura 2, onde um conjunto de dados gerados *ad hoc* é suavizado pelo método *loess* com diferentes valores de $\lambda$.

**Figura 2 - Curvas suavizadas pelo método *loess* com diferentes valores de λ e d**



A Figura 2a apresenta uma curva obtida com $\lambda = 0,67$; ela é pouco *ondulada* e não se adapta aos *picos* e aos *vales* adequadamente. Quando $\lambda$ é reduzido para 0,40 (Figura 2b), a curva ajustada torna-se mais *ondulada* (menos suave), e consegue descrever melhor o comportamento dos dados (menor viés). Reduzindo ainda mais o valor de $\lambda$ (Figura 2c), obtém-se uma curva muito *ondulada*, que não consegue captar a tendência dos dados. Um valor adequado para $\lambda$ neste exemplo seria 0,40, que é o maior valor (dentre os considerados) capaz de minimizar a variabilidade sem distorcer a tendência dos dados.

O grau do polinômio ($d$) deve ser fixado com base no padrão apresentado pelos dados num diagrama de dispersão. De uma forma geral, se a nuvem de pontos sugere uma tendência sem máximos ou mínimos locais, então um ajuste linear ($d = 1$), é adequado. Mas se existirem regiões com máximos ou mínimos locais como nos dados da Figura 2, então um ajuste quadrático ($d = 2$) normalmente produz uma curva que melhora a descrição local do padrão dos dados.

As Figuras 2b e 2d foram suavizadas fixando-se $\lambda = 0,4$ e, respectivamente, $d = 2$ e $d = 1$. Observa-se que o ajuste linear (Figura 2d) não é capaz de acomodar os máximos e mínimos locais pois a curva permanece abaixo dos *picos* e acima dos *vales*. Para conseguir tal acomodação seria necessário diminuir muito o valor de $\lambda$. Já o ajuste local quadrático com o mesmo valor de $\lambda$ (Figura 2b) atinge os *picos* e *vales*, sugerindo uma suavização mais adequada.

A função $U$ que atribui os pesos associados a cada ajuste local do polinômio, tendo $(x_i,y_i)$ como ponto alvo, tem a forma geral

$$u_{x_i j} = U(h_i^{-1}(x_j - x_i)),\ j=1,\ldots,n,$$

onde $h_i$ é a distância entre $x_i$ e o seu $r$-ésimo vizinho mais próximo, isto é, $h_i$ é o valor que ocupa a $r$-ésima posição na seqüência crescentemente ordenada de $|x_i - x_k|$, $k = 1,\ldots,n$. Essa função deve ser especificada de forma que, para $g \in \Re$,

*i.* $U(g) > 0$ *para* $|g| < 1$;

*ii.* $U(-g) = U(g)$ ;

*iii.* $U(g)$ *é uma função decrescente para* $g \geq 0$;

*iv.* $U(g) = 0$ *para* $|g| \geq 1$.

A função tricúbica,

$$U(g) = \begin{cases} \left(1 - |g|^3\right)^3 & para\ |g| < 1 \\ 0 & para\ |g| \geq 1, \end{cases} \tag{5}$$

apresenta as propriedades descritas acima, e de acordo com Cleveland (1979), fornece uma suavização adequada na maioria dos casos.

Com base na função (5) obtém-se a matriz de pesos referente ao ponto alvo $(x_i,y_i)$, denotada por

$$U_{x_i} = diagonal\{u_{x_i,1},\ldots,u_{x_i,n}\}, \tag{6}$$

com

$$u_{x_i,j} = \begin{cases} \left(1 - \left|h_i^{-1}(x_j - x_i)\right|^3\right)^3 \text{ para } \left|h_i^{-1}(x_j - x_i)\right| < 1 \\ 0 \qquad \text{para } \left|h_i^{-1}(x_j - x_i)\right| \geq 1. \end{cases} \tag{7}$$

Assim, por (6) e (7), em um ajuste local, ao ponto alvo $(x_i, y_i)$ fica associado um peso 1; os pesos diminuem à medida que os pontos se afastam de $(x_i, y_i)$ e aos pontos fora da vizinhança de $x_i$ ficam associados a pesos nulos.

O valor ajustado é, então, calculado a partir de um modelo de regressão por MQP com pesos dados por $U_{x_i}$. Dessa forma, para cada $x_i$, $i=1,\ldots,n$, obtém-se o valor suavizado $\hat{f}(x_i)$, $i=1,\ldots,n$.

Uma ilustração numérica deste procedimento está apresentada no Apêndice.

Com finalidade computacional, note que os valores previstos de $Y$ obtidos no procedimento de suavização podem ser escritos na forma (4). A matriz suavizadora

$$S = \begin{bmatrix} s_{x_1,1} & s_{x_1,2} & \cdots & s_{x_1,n} \\ s_{x_2,1} & s_{x_2,2} & \cdots & s_{x_2,n} \\ \vdots & \vdots & & \vdots \\ s_{x_n,1} & s_{x_n,2} & \cdots & s_{x_n,n} \end{bmatrix} = \begin{bmatrix} \mathbf{s}'_{x_1} \\ \mathbf{s}'_{x_2} \\ \vdots \\ \mathbf{s}'_{x_n} \end{bmatrix}$$

tem a linha $s'_{x_i}$ correspondente à i-ésima linha da matriz

$$S_{x_i} = X(X'U_{x_i}X)^{-1}X'U_{x_i}, \tag{8}$$

construída no ajuste da regressão ponderada local que tem $(x_i, y_i)$ como ponto alvo e matriz de pesos $U_{x_i}$ definida em (6), $i=1,\ldots,n$.

O valor previsto correspondente a $x_i$ é, então, dado por

$$\hat{f}(x_i) = s_{x_i 1}y_1 + s_{x_i 2}y_2 + \ldots + s_{x_i n}y_n = \mathbf{s}'_{x_i}\mathbf{y}, \quad i=1,\ldots,n.$$

De uma forma geral, pode-se mostrar que o elemento ij da matriz suavizadora $S$ é

$$s_{x_i,j} = \frac{u_{x_i,j}\sum_{j=1}^{n} x_j^2 u_{x_i,j} - u_{x_i,j}(x_i + x_j)\sum_{j=1}^{n} x_j u_{x_i,j} + x_i x_j u_{x_i,j}\sum_{j=1}^{n} u_{x_i,j}}{\sum_{j=1}^{n} u_{x_i,j}\sum_{j=1}^{n} x_j^2 u_{x_i,j} - \left(\sum_{j=1}^{n} u_{x_i,j}x_j\right)^2}, \tag{9}$$

onde $u_{x_i,j}$ é definido de acordo com (7).

As expressões (8) e (9) mostram claramente que os elementos da matriz $S$ dependem apenas de $x_1,\ldots,x_n$ e do parâmetro de suavização $\lambda$. Portanto, o *loess* é um suavizador linear.

A matriz suavizadora $S$ desempenha papel semelhante ao da matriz chapéu (*hat matrix)* no método de estimação de mínimos quadrados e algumas de suas propriedades são demonstradas por Hoaglin e Welsch (1978), a saber

*i.* $0 \leq s_{ii} \leq 1$,

*ii.* $-1 \leq s_{ij} \leq 1$ *para* $i \neq j$,

*iii.* $s_{ii} = 1$ *se e somente se* $s_{ij} = 0$ *para todo* $i \neq j$, *e*

*iv.* $\sum_{j=1}^{n} s_{ij} = 1$.

## 2.3 Algumas propriedades dos suavizadores lineares

Quando um suavizador é linear, a *matriz de covariância* de $\hat{\mathbf{f}}$ é

$$Var(\hat{\mathbf{f}}) = Var(\boldsymbol{SY}) = \boldsymbol{S}Var(\boldsymbol{Y})\boldsymbol{S}'.$$

Sob a suposição de que os $Y_1, \ldots, Y_n$ são independentes com $Var(Y_i) = \sigma^2$, tem-se

$$Var(\hat{\mathbf{f}}) = \sigma^2 \boldsymbol{SS}'.$$

O vetor *viés* do estimador é definido como

$$\boldsymbol{b} = E(\boldsymbol{f} - \hat{\mathbf{f}}) = \boldsymbol{f} - E(\hat{\mathbf{f}}) = \boldsymbol{f} - E(\boldsymbol{SY}) = \boldsymbol{f} - \boldsymbol{Sf} = (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{f}$$

e o *erro quadrático médio* (EQM) global é dado por

$$EQM = \frac{1}{n}\sum_{i=1}^{n} E\left(f(x_i) - \hat{f}(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n} Var\left(\hat{f}(x_i)\right) + \frac{1}{n}\sum_{i=1}^{n} b_i^2 = \frac{traço(\mathrm{SS}')}{n}\sigma^2 + \frac{\mathbf{b}'\mathbf{b}}{n}. \qquad (10)$$

É interessante observar a influência de $\lambda$ nos componentes do EQM. De forma geral, aumentando $\lambda$, o traço$(\boldsymbol{SS}')$ tende a diminuir e os elementos de $\boldsymbol{b}$ tendem a aumentar, e vice-versa.

O parâmetro $\sigma^2$ em (10) geralmente é desconhecido e, assumindo que $\hat{\mathbf{f}}$ é não-viesado, um estimador é

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left\{y_i - \hat{f}(x_i)\right\}^2}{n - traço(2\mathrm{S} - \mathrm{SS}')} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n - traço(2\mathrm{S} - \mathrm{SS}')},$$

onde $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{f}} = (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{y}$. Pode-se mostrar que $E[\hat{\mathbf{e}}] = \boldsymbol{0}$ *e* $Var[\hat{\mathbf{e}}] = \sigma^2(\boldsymbol{I} - \boldsymbol{S})(\boldsymbol{I} - \boldsymbol{S})'$ (Hastie e Tibshirani, 1990).

Quando os erros têm distribuição Normal é fácil mostrar que $\hat{\sigma}^2$ é um estimador não-viesado para $\sigma^2$. Assim, $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ é uma forma quadrática em variáveis normais, e

$$\mathrm{E}[\hat{\mathbf{e}}'\hat{\mathbf{e}}] = \text{traço}[\sigma^2(\mathbf{I} - \mathrm{S})(\mathbf{I} - \mathrm{S})'] = \sigma^2[\mathrm{n} - \text{traço}(2\mathrm{S} - \mathrm{SS}')]. \qquad (11)$$

Logo,

$$\sigma^2 = \frac{\mathrm{E}[\hat{\mathbf{e}}'\hat{\mathbf{e}}]}{n - traço(2\mathrm{S} - \mathrm{SS}')}.$$

Bandas de confiança pontuais para $f(x_i)$ podem ser construídas por intermédio de

$$\hat{\mathrm{f}}(x_i) \pm 2\,\hat{ep}(\hat{\mathrm{f}}(x_i)),$$

onde $\hat{ep}(\hat{f}(x_i))$ é o erro padrão estimado de $\hat{f}(x_i)$, e corresponde à raiz quadrada do i-ésimo elemento da diagonal da matriz $\hat{\sigma}^2 SS'$. Sob as suposições de erros normais e viés desprezível, estas bandas representam intervalos de confiança pontuais para $f$. Se o viés não for desprezível (o que é difícil de ser verificado), as bandas correspondem a intervalos de confiança pontuais para os elementos do vetor $Sf$, e não para $f$.

Uma maneira de tornar diferentes procedimentos de suavização comparáveis em relação à *quantidade de suavização* que apresentam é baseada na especificação dos *graus de liberdade* de cada suavizador.

Dados um parâmetro de suavização $\lambda$ e uma matriz suavizadora $S$, o *número de graus de liberdade* (ou *número de parâmetros*) do alisador correspondente pode ser definido como

$$gl = traço(SS'), \tag{12}$$

$$gl = traço(S) \tag{13}$$

ou

$$gl = traço(2S - SS'). \tag{14}$$

Quando a matriz suavizadora $S$ é idempotente, (12), (13) e (14) coincidem.

Essas expressões foram propostas por analogia à regressão de mínimos quadrados. A motivação para (12) e (13) deve-se ao fato de que a matriz $S$ desempenha papel semelhante à matriz chapéu, $H$, para a qual

*traço(HH') = traço(H) = posto(H) = número de parâmetros do modelo.*

A motivação para (14) é a expressão (11), considerando que, na regressão de mínimos quadrados, o valor esperado da soma de quadrados do resíduo é $\sigma^2$(n – número de parâmetros no modelo). Nota-se que, em um caso extremo para o qual a suavização é mínima, isto é, em que a curva suavizadora passa por todos os pontos, tem-se traço$(SS') = n$, que é igual ao número de parâmetros em um modelo de regressão saturado, para o qual $traço(H) = n$.

De uma forma geral, quanto maior o número de graus de liberdade, menor a quantidade de suavização e, conseqüentemente, menor o valor de $\lambda$.

As expressões (12), (13) e (14) podem ser estendidas para suavizadores não-lineares, mas neste caso, os graus de liberdade dependem da distribuição de Y.

## 2.4 Suavizadores ponderados

A suposição de igualdade de variâncias dos erros assumida para o modelo (3) pode ser avaliada da mesma forma que em um modelo paramétrico de regressão. Embora os resíduos, $\hat{e}_i = y_i - \hat{f}(x_i)$, não tenham soma nula, eles podem ser examinados graficamente de forma análoga à regressão de mínimos quadrados para a verificação da hipótese de homocedasticidade (ver, por exemplo, Neter et al., 1996).

Se a análise dos resíduos indicar que um modelo heterocedástico é mais adequado, isto é, com erros aleatórios com média zero e variância $\sigma_i^2$, então, pode-se adotar um método de suavização ponderado no qual à i-ésima observação fica associado um peso $w_i = 1/\sigma_i^2$.

O método de suavização *loess* é facilmente adaptado ao caso ponderado. Para isto, basta multiplicar $U_{x_i}$, dada em (6), por $W = diagonal\{w_1,\ldots,w_n\}$ e obter uma nova matriz de pesos $A_{x_i} = U_{x_i}W$. O valor previsto correspondente a $x_i$ será

$$\hat{f}(x_i) = s_{x_i,1}y_1 + s_{x_i,2}y_2 + \ldots + s_{x_i,n}y_n = s'_{x_i}y,$$

onde $s'_{x_i}$ é a i-ésima linha da matriz

$$S_{x_i} = X(X'A_{x_i}X)^{-1}X'A_{x_i}.$$

Como raramente as variâncias são conhecidas, é necessário estimar $\sigma_i^2$. Existem várias formas de se obterem estimadores para estes parâmetros quando o método de mínimos quadrados é adotado (ver, por exemplo, Neter et al., 1996, Capítulo 10), que podem ser utilizadas aqui. Dentre elas, destacamos aquela baseada na construção do gráfico de dispersão dos resíduos versus $x_i$ e na estimação da relação entre a variância e a variável preditora. Um procedimento iterativo pode então, ser adotado para obtenção dos pesos finais.

A versão ponderada do *loess* é uma ferramenta básica para o uso deste suavizador nos MAG, como será indicado na Seção 3.

# 3. Modelos aditivos generalizados

Sejam $Y_i$, $i=1,\ldots,n$, variáveis aleatórias independentes, com função densidade de probabilidade (ou função de probabilidade) dada por

$$f(y_i;\theta_i,\phi) = \exp\left[\phi\left(y_i\theta_i - b(\theta_i)\right) + c(y_i,\phi)\right], \tag{15}$$

onde b(•) e c(•) são funções especificadas, b(•) é duas vezes diferenciável, e $\phi^{-1} > 0$ é o parâmetro de dispersão. Neste trabalho, assumiremos $\phi$ conhecido e, portanto, (15) é um modelo da família exponencial unidimensional com parâmetro canônico $\theta_i$, $i=1,\ldots,n$. Consideremos que a média de $Y_i$, digamos $\mu_i$, está relacionada com um conjunto de variáveis preditoras (não-aleatórias) $X_1,\ldots,X_p$ por meio de uma função de ligação monótona e diferenciável

$$g(\mu_i) = \eta_i$$

onde

$$\eta_i = \alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \tag{16}$$

é chamado *preditor linear*, $\alpha, \beta_1,\ldots,\beta_p$ são os parâmetros a serem estimados e $x_{i1},\ldots,x_{ip}$ são os valores observados de $X_1,\ldots,X_p$ no i-ésimo elemento da amostra. Modelos definidos desta maneira são chamados MLG. Nesse contexto, o preditor linear $\eta_i$ é uma função linear de cada uma das variáveis preditoras $X_1,\ldots,X_p$. No entanto, uma relação menos rígida pode ser adotada substituindo o termo linear correspondente a cada variável preditora por uma função não especificada dessa variável, obtendo-se o *preditor aditivo*

$$\eta_i = \alpha + f_1(x_{i1}) + \ldots + f_p(x_{ip}), \tag{17}$$

onde $\alpha$ é um parâmetro e $f_1, \ldots, f_p$ são funções não especificadas. Modelos assim obtidos são denominados MAG e podem ser vistos como generalizações dos MLG.

O preditor (17) corresponde a um modelo totalmente não-paramétrico. Modelos cujo preditor combina formas paramétricas de algumas *(r)* variáveis preditoras com termos não-paramétricos de outras *(p – r)* variáveis também fazem parte dessa classe. Neste caso, o preditor pode ser escrito como

$$\eta_i = \alpha + \beta_1 x_{i1} + \ldots + \beta_r x_{ir} + f_1(x_{i,r+1}) + \ldots + f_{p-r}(x_{ip}). \tag{18}$$

Esses modelos são denominados *modelos semiparamétricos.*

Ajuste de modelos aditivos generalizados

Considere um MAG cujo preditor aditivo é dado por (17). Seja $f_j = (f_j(x_{1j}), \ldots, f_j(x_{nj}))'$ o vetor dos valores da função $f_j$, $j=1,\ldots,p$, calculado nos n valores observados de $X_j$. O ajuste de um MAG consiste na estimação de $\alpha$ e $f_j$, $j=1,\ldots,p$.

O processo de ajuste baseia-se na combinação de dois procedimentos iterativos:

• o Procedimento de Ponderação Local (*local scoring*), doravante abreviado PPL – similar ao procedimento de Mínimos Quadrados Iterativamente Reponderados - MQIR - utilizado no ajuste dos MLGs, com o preditor aditivo (17) no lugar do preditor linear (16),

e

• o retroajuste (*backfitting*) – algoritmo responsável pela estimação de cada $f_j$ por meio da utilização de suavizadores ponderados.

O PPL corresponde a um "ciclo externo" no processo de estimação necessário para o ajuste de um modelo com estrutura semelhante a um MLG; o retroajuste é um "ciclo interno" ao PPL no qual são estimadas as funções $f_j$, $j=1,\ldots,p$.

O ajuste de um MAG pode ser efetuado nos três passos do algoritmo PPL esquematizados a seguir. O passo 2 corresponde ao retroajuste.

Fazendo $\alpha^{(0)} = g\left(\sum_{i=1}^{n} \frac{y_i}{n}\right)$ e $f_1^{(0)} = \ldots = f_p^{(0)} = 0$, o algoritmo consiste em iterar os seguintes passos para $m = 1, 2, \ldots$

***Passo 1***: Para $i=1,\ldots,n$, calcular

$$\eta_i^{(m)} = \alpha^{(m-1)} + \sum_{j=1}^{p} f_j^{(m-1)}(x_{ij}) \ ,$$

$$\mu_i^{(m)} = g^{-1}(\eta_i^{(m)}),$$

$$w_i^{(m)} = \left[\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^{(m)}\right]^2 \left[\left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{(m)}\right]^{-1}, \qquad (19)$$

$$z_i^{(m)} = \eta_i^{(m)} + (y_i - \mu_i^{(m)})\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{(m)}$$

e

$\mathbf{S}_j^{(m)}$: matriz suavizadora ponderada de dimensão *(nxn)* relativa à j-ésima covariável, com matriz de pesos $W^{(m)}=diagonal\{w_1^{(m)},\ldots,w_n^{(m)}\}$, $j=1,\ldots,p$. Esta matriz depende do método de suavização utilizado.

***Passo 2***: (Retroajuste).

Fazer $\alpha^{(m)} = \bar{z}^{(m)} = \sum_{i=1}^{n} \frac{z_i^{(m)}}{n}$, $\mathbf{f}_{j(0)}^{(m)} = \mathbf{f}_j^{(m-1)}$, $j=1,\ldots,p$ e calcular

$$f_{j(v)}^{(m)} = \mathbf{S}_j^{(m)} \mathbf{r}_{j(v)}^{(m)}, \; j=1,\ldots,p$$

onde $\mathbf{r}_{j(v)}^{(m)} = (r_{ij(v)}^{(m)},\ldots,r_{ij(v)}^{(m)})'$ é o vetor de resíduos parciais com elementos dados por

$$r_{ij(v)}^{(m)} = z_i^{(m)} - \bar{z}^{(m)} - \sum_{k=1}^{j-1} f_{k(v)}^{(m)}(x_{ik}) - \sum_{k=j+1}^{p} f_{k(v-1)}^{(m)}(x_{ik}),$$

para $v=1, 2,\ldots$, até que

$$\left\| \mathbf{f}_{j(v)}^{(m)} - \mathbf{f}_{j(v-1)}^{(m)} \right\| \leq \varepsilon, \; j=1,\ldots,p$$

para um valor $\varepsilon > 0$ preestabelecido. Nesse ponto,

$$f_j^{(m)} = f_{j(v)}^{(m)}.$$

***Passo 3***: Repetir os passos 1 e 2 até que

$$\frac{\sum_{j=1}^{p} \left\| \mathbf{f}_j^{(m)} - \mathbf{f}_j^{(m-1)} \right\|}{\sum_{j=1}^{p} \left\| \mathbf{f}_j^{(m-1)} \right\|} \leq \delta,$$

para um valor $\delta > 0$ preestabelecido.

Essencialmente, o algoritmo de retroajuste corresponde ao método de Gauss-Seidel para resolver o sistema de equações lineares

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1^{(m)} & \mathbf{S}_1^{(m)} & \cdots & \mathbf{S}_1^{(m)} \\ \mathbf{S}_2^{(m)} & \mathbf{I} & \mathbf{S}_2^{(m)} & \cdots & \mathbf{S}_2^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p^{(m)} & \mathbf{S}_p^{(m)} & \mathbf{S}_p^{(m)} & \cdots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{f}_1^{(m)} \\ \mathbf{f}_2^{(m)} \\ \vdots \\ \mathbf{f}_p^{(m)} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1^{(m)}\mathbf{z}^{(m)} \\ \mathbf{S}_2^{(m)}\mathbf{z}^{(m)} \\ \vdots \\ \mathbf{S}_p^{(m)}\mathbf{z}^{(m)} \end{bmatrix}. \qquad (20)$$

Esse sistema consiste de *np* equações que correspondem aos *np* parâmetros a serem estimados. O retroajuste é um método eficiente para resolvê-lo, principalmente quando o número de parâmetros é grande. Motivações para a utilização dessas equações na obtenção de $\mathbf{f}_1^{(m)},\ldots,\mathbf{f}_p^{(m)}$ podem ser encontradas em Hastie e Tibshirani (1990) ou Lima (2001).

As matrizes $\mathbf{S}_j^{(m)}$, $j=1,\ldots,p$, em (20) podem corresponder a diferentes métodos de suavização, mas, em geral, o mesmo suavizador é utilizado para estimar $f_1,\ldots,f_p$. No caso do suavizador *loess*, a i-ésima linha de $\mathbf{S}_j^{(m)}$ corresponde à i-ésima linha da matriz suavizadora ponderada

$$\mathbf{S}_{x_i}^{(m)} = X(X'\mathbf{A}_{x_i}^{(m)}X)^{-1}X'\mathbf{A}_{x_i}^{(m)},$$

onde $X = (1, X_j)$ e $\mathbf{A}_{x_i}^{(m)} = diagonal\{u_{x_i 1}w_1^{(m)},\ldots,u_{x_i n}w_n^{(m)}\}$ *com* $u_{x_i j}$ *e* $w_j^{(m)}$ definidos, respectivamente, em (7) e (19).

Se a variável resposta segue uma distribuição Normal, a função de ligação é a identidade, então $Z = Y$, $W = I$ e o procedimento MQIR é substituído por um método direto, ou seja, apenas o ciclo interno, correspondente ao retroajuste, é necessário.

No caso de um MAG com apenas uma função não especificada, isto é, $p = 1$ em (17), o algoritmo de retroajuste não é necessário pois $f^{(m)}$ pode ser obtido diretamente com a utilização de um alisador ponderado aplicado aos resíduos $r_i^{(m)} = z_i^{(m)} - \bar{z}^{(m)}$ em função de $x_i$, $i=1,\ldots,n$, com matriz de pesos $W^{(m)}$.

Embora o retroajuste seja um algoritmo eficiente para resolver (20), pelo menos conceitualmente, estimativas para $f_1,\ldots,f_p$ podem ser obtidas diretamente pela relação

$$\hat{\mathbf{f}} = M^{-1}Cz,$$

com

$$\hat{\mathbf{f}} = \begin{bmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_p \end{bmatrix}, \quad M = \begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \mathbf{S}_2 & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p & \mathbf{S}_p & \mathbf{S}_p & \cdots & \mathbf{I} \end{bmatrix} \quad e \quad C = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_p \end{bmatrix},$$

se a inversa de $M$ existir.

Em particular, escrevendo

$$R_j = E_j M^{-1} C, \quad j=1,\ldots,p$$

com $E_j$ denotando uma matriz de dimensão $(n \times np)$ composta por p "blocos" de dimensão $n \times n$, com todos os blocos nulos à exceção do j-ésimo, que é uma matriz identidade, temos

$$\hat{\mathbf{f}}_j = \boldsymbol{R}_j \boldsymbol{z}, \quad j=1,\ldots,p.$$

Fazendo $f = f_1 + \ldots + f_p$, segue que

$$\hat{\mathbf{f}} = \boldsymbol{R}_1 \boldsymbol{z} + \ldots + \boldsymbol{R}_p \boldsymbol{z} = \boldsymbol{R}_{NP} \boldsymbol{z}$$

onde $\boldsymbol{R}_{NP} = \boldsymbol{R}_1 + \ldots + \boldsymbol{R}_p$.

Para modelos que envolvem apenas duas matrizes suavizadoras em seu ajuste $(p = 2)$, Hastie e Tibshirani (1990) fornecem expressões mais simples para $\boldsymbol{R}_1$ e $\boldsymbol{R}_2$, dadas por

$$\mathbf{R}_1 = \mathbf{I} - (\mathbf{I} - \mathbf{S}_1\mathbf{S}_2)^{-1}(\mathbf{I} - \mathbf{S}_1)$$
$$\mathbf{R}_2 = \mathbf{I} - (\mathbf{I} - \mathbf{S}_2\mathbf{S}_1)^{-1}(\mathbf{I} - \mathbf{S}_2).$$

Neste caso,

$$\boldsymbol{R}_{NP} = (\boldsymbol{R}_1 + \boldsymbol{R}_2) = \boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{S}_2)(\boldsymbol{I} - \boldsymbol{S}_1\boldsymbol{S}_2)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1).$$

Expressões recursivas para modelos envolvendo mais de dois suavizadores foram deduzidas por Opsomer (2000). O custo computacional para obter $\boldsymbol{R}_j$, $j=1,\ldots,n$, a partir dessas expressões é, entretanto, elevado.

A apresentação da solução direta neste ponto do trabalho não tem o objetivo de propô-la como alternativa ao retroajuste na estimação de $f_1,\ldots f_p$, uma vez que este algoritmo é mais eficiente do ponto de vista computacional, mas sim de obter expressões para $\hat{\mathbf{f}}_j$ e $\hat{\boldsymbol{\eta}}$ que tornem mais simples o estudo de suas propriedades estatísticas.

A convergência do procedimento de ajuste dos MAGs está condicionada à convergência do retroajuste, uma vez que o PPL não apresenta, em geral, problemas dessa ordem (Hastie e Tibshirani, 1990). Resultados sobre a convergência desse procedimento podem ser obtidos em Buja et al. (1989) e Opsomer (2000).

Ajuste de modelos semiparamétricos

Considere o modelo semiparamétrico

$$g(\mu_i) = \alpha + \sum_{j=1}^{r} \beta_j x_{ij} + \sum_{j=r+1}^{p} f_j(x_{ij}).$$

Os parâmetros $\alpha, \beta_1,\ldots,\beta_r$ e as funções $f_{r+1},\ldots f_p$ também podem ser estimados com a utilização do PPL e do retroajuste. Dados os valores iniciais $\boldsymbol{\beta}^{(0)} = (\alpha^{(0)},\beta_1^{(0)},\ldots,\beta_r^{(0)})'$ e $f_{r+1}^{(0)},\ldots f_p^{(0)}$, estimativas para $\boldsymbol{\beta}$ e $f_{r+1},\ldots f_p$ são obtidas resolvendo-se, iterativamente, as seguintes equações:

$$\boldsymbol{\beta}^{(m)} = (\mathbf{X}'\mathbf{W}^{(m)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(m)}\left(\mathbf{z}^{(m)} - \sum_{j=r+1}^{p}\mathbf{f}_j^{(m)}\right) \tag{21}$$

e

$$\mathbf{f}_j^{(m)} = \mathbf{S}_j^{(m)}\left(\mathbf{z}^{(m)} - \mathbf{X}\boldsymbol{\beta}^{(m)} - \sum_{\substack{i=r+1 \\ i\neq j}}^{p}\mathbf{f}_i^{(m)}\right), j=r+1,\ldots,p \tag{22}$$

onde $X = (1, X_1, \ldots, X_r)$ é a matriz de especificação correspondente aos termos paramétricos do modelo com $X_j$, $j=1,\ldots,r$, denotando o vetor dos valores observados da j-ésima covariável e $\mathbf{S}_j^{(m)}$, $j=r+1,\ldots,p$, a matriz suavizadora ponderada relativa à j-ésima covariável no m-ésimo passo do PPL. Após obter as estimativas $\boldsymbol{\beta}^{(m)}$ e $\mathbf{f}_j^{(m)}$, $j=r+1,\ldots,p$, pelo retroajuste, valores de $\boldsymbol{\eta}^{(m+1)}$, $\boldsymbol{\mu}^{(m+1)}$, $\mathbf{z}^{(m+1)}$ e $\mathbf{W}^{(m+1)}$ são calculados pelo PPL delineado na seção 3.1, e o processo é repetido até a convergência.

Este procedimento é análogo àquele empregado para ajuste de um modelo não-paramétrico com $(p-r)+1$ suavizadores: um deles é o operador projeção $S_1 = X(X'WX)^{-1}X'W$, que produz o valor ajustado $X\hat{\beta}$ e os $(p-r)$ suavizadores restantes são os que geram $\hat{\mathbf{f}}_{r+1}, \ldots, \hat{\mathbf{f}}_p$.

Quando existe apenas uma função não especificada, $f$, isto é, $p = r+1$ em (18), as expressões (21) e (22) se reduzem a

$$\boldsymbol{\beta}^{(m)} = (X'W^{(m)}X)^{-1}X'W^{(m)}(z^{(m)} - f^{(m)}) \tag{23}$$

e

$$f^{(m)} = S^{(m)}(z^{(m)} - X\boldsymbol{\beta}^{(m)}). \tag{24}$$

Assim, o retroajuste pode ser evitado substituindo (24) em (23) de forma a obter

$$\boldsymbol{\beta}^{(m)} = (X'W^{(m)}(I - S^{(m)})X)^{-1}X'W^{(m)}(I - S^{(m)})z^{(m)}. \tag{25}$$

Após a obtenção de uma estimativa $\boldsymbol{\beta}^{(m)}$ segundo (25), uma estimativa $f^{(m)}$ é calculada segundo (24). Por intermédio do PPL especificado na seção 3.1, podem-se estimar, então, novos valores $\boldsymbol{\eta}^{(m+1)}$, $\boldsymbol{\mu}^{(m+1)}$, $z^{(m+1)}$ e $W^{(m+1)}$, e o processo é repetido até a convergência. As estimativas assim obtidas são idênticas às fornecidas quando o retroajuste é realizado.

Opsomer e Ruppert (1999) e Thurston et al. (2000) mostraram que no modelo envolvendo mais de um termo não-paramétrico, o retroajuste pode ser evitado considerando, no m-ésimo passo do PPL, os estimadores

$$\boldsymbol{\beta}^{(m)} = (\mathbf{X}'\mathbf{W}^{(m)}(\mathbf{I} - \mathbf{R}_{NP}^{(m)})\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(m)}(\mathbf{I} - \mathbf{R}_{NP}^{(m)})\mathbf{z}^{(m)}$$

e

$$\mathbf{f}_{NP}^{(m)} = \mathbf{R}_{NP}^{(m)}(\mathbf{z}^{(m)} - \mathbf{X}\boldsymbol{\beta}^{(m)}) = \mathbf{f}_{r+1}^{(m)} + \ldots + \mathbf{f}_p^{(m)}$$

onde

$$\mathbf{R}_{NP}^{(m)} = \sum_{j=r+1}^{p} \mathbf{R}_j^{(m)}$$

denota a matriz suavizadora ponderada generalizada correspondente aos termos não-paramétricos do modelo. As soluções obtidas por intermédio dessas equações são equivalentes às soluções do retroajuste, que, no entanto, é um procedimento computacionalmente mais eficiente.

Um fato pouco evidenciado na literatura é que, em geral, os estimadores dos modelos semiparamétricos não são identificáveis quando incluem o intercepto (Opsomer e Ruppert, 1999). Neste caso, quando a soma dos

elementos das linhas de $S$ é igual a 1 (como é o caso do *loess*), $X'W^{(m)}(I - S^{(m)})X$ e $X'W^{(m)}(I - \mathbf{R}_{NP}^{(m)})X$ são singulares e uma solução simples para esse problema é substituir as matrizes $\mathbf{S}_j^{(m)}$ por matrizes centradas, da forma $(I-11'/n)\mathbf{S}_j^{(m)}$. Este procedimento faz com que a média de $\hat{\mathbf{f}}$ seja igual a zero em cada passo e o modelo torne-se identificável.

Testes de hipóteses

Nos MLGs, o desvio para o modelo ajustado $\hat{\boldsymbol{\mu}}$, $D(y; \hat{\boldsymbol{\mu}})$, e o desvio parcial para dois modelos ajustados $\hat{\boldsymbol{\mu}}_1$ e $\hat{\boldsymbol{\mu}}_2$, $D(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2)$, são quantidades bem conhecidas, utilizadas, respectivamente, para avaliar a qualidade do ajuste e comparar dois modelos ajustados (ver, por exemplo, McCullagh e Nelder, 1989). Sem perda de generalidade, pode-se escrever $D(y; \hat{\boldsymbol{\eta}})$ no lugar de $D(y; \hat{\boldsymbol{\mu}})$, uma vez que $\hat{\boldsymbol{\mu}}$ está relacionado com $\hat{\boldsymbol{\eta}}$ por meio da função de ligação $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$. Pelo mesmo motivo, pode-se escrever $D(\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2)$ no lugar de $D(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2)$.

No caso dos MAGs, o desvio também pode ser usado como uma medida de ajuste e a comparação entre modelos pode ser feita utilizando-se o desvio parcial. Embora as distribuições assintóticas dessas estatísticas não tenham sido determinadas, Hastie e Tibshirani (1990) mostraram, por simulação, que distribuições $\chi^2$ com número de graus de liberdade determinado da maneira a seguir são boas aproximações para as distribuições dessas estatísticas.

Seja $\boldsymbol{R}$ o operador ponderado de ajuste aditivo obtido no último passo do PPL e seja $D(y; \hat{\boldsymbol{\eta}})$ o desvio correspondente a $\hat{\boldsymbol{\eta}} = \boldsymbol{Rz}$. Valores observados de $D(\mathbf{y}; \hat{\boldsymbol{\eta}})$ muito maiores que a média da distribuição $\chi^2$ com $gl = n - traço(2\boldsymbol{R} - \boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}\boldsymbol{W}^{-1})$ graus de liberdade sugerem falta de ajuste do modelo.

Suponha agora que $\boldsymbol{R}^{(1)}$ e $\boldsymbol{R}^{(2)}$ sejam operadores de ajuste aditivo ponderado e que $\hat{\boldsymbol{\eta}}_1 = \boldsymbol{R}^{(1)}\boldsymbol{z}$ e $\hat{\boldsymbol{\eta}}_2 = \boldsymbol{R}^{(2)}\boldsymbol{z}$ sejam, respectivamente, estimativas de

$$\eta_1 = \alpha + f_1(X_1) + \ldots + f_q(X_q)$$

e

$$\eta_2 = \eta_1 + f_{q+1}(X_{q+1}) + \ldots + f_p(X_p).$$

A estatística $\phi D(\hat{\boldsymbol{\eta}}_1; \hat{\boldsymbol{\eta}}_2)$ pode ser usada para avaliar a contribuição dos termos presentes apenas em $\eta_2$. A distribuição de referência é a distribuição $\chi^2$ com

$$gl(\hat{\boldsymbol{\eta}}_1) - gl(\hat{\boldsymbol{\eta}}_2) = traço(2\boldsymbol{R}^{(1)} - \boldsymbol{R}^{(1)'}\boldsymbol{W}_1\boldsymbol{R}^{(1)}\boldsymbol{W}_1^{-1}) - traço(2\boldsymbol{R}^{(2)} - \boldsymbol{R}^{(2)'}\boldsymbol{W}_2\boldsymbol{R}^{(2)}\boldsymbol{W}_2^{-1})$$

graus de liberdade. Valores grandes de $\phi D(\hat{\boldsymbol{\eta}}_1; \hat{\boldsymbol{\eta}}_2)$ sugerem uma contribuição significativa desses termos.

Os cálculos para obtenção dos graus de liberdade segundo as definições dadas acima têm custo computacional elevado por envolverem matrizes do tipo $\boldsymbol{R}$. Por este motivo, a quantidade

$$gl = n - 1 - \sum_{j=1}^{p} [\text{traço}(\mathbf{S}_j) - 1]$$

é usada como uma medida aproximada para os graus de liberdade de $D(y; \hat{\eta})$.

Medidas de precisão e bandas de confiança pontuais

Bandas de confiança pontuais para os elementos de $\eta$ podem ser obtidas usando a metodologia descrita em Hastie e Tibshirani (1990), que se baseia em procedimentos de linearização. A idéia é aproximar a variável resposta modificada $z$ por uma quantidade assintoticamente equivalente $z_0$, assumindo que o modelo é consistente. A matriz de covariância assintótica de $\hat{\eta} = \mathbf{Rz}$ é, então, estimada com base na matriz de covariância de $z_0$, nomeadamente $W_0^{-1}\phi$, e considerando uma versão assintótica $R^{(0)}$ de $R$, já que esta matriz não é um operador linear pois depende de $y_i$ por meio dos pesos $w_i$, $i=1,\ldots,n$. Obtém-se então,

$Var(\hat{\eta}) \approx R^{(0)} W_0^{-1} R^{(0)\prime}\phi$.

Para amostras finitas, considera-se a aproximação

$Var(\hat{\eta}) \approx RW^{-1}R'\phi$.

Similarmente,

$Var(\hat{\mathbf{f}}_j) \approx R_j W^{-1} R_j'\phi$,

onde $R_j$ é a matriz que gera $\hat{\mathbf{f}}_j$ a partir de $z$.

As mesmas condições de regularidade requeridas para o desenvolvimento de resultados assintóticos para os MLGs (ver, por exemplo, Sen e Singer, 1993) são consideradas aqui também. Os resultados obtidos permitem mostrar que $\hat{\eta}$ tem distribuição assintótica $N(\eta_0, R^{(0)} W_0^{-1} R^{(0)\prime}\phi)$ e bandas de confiança pontuais aproximadas para $f_j(x_{ij})$ são dadas por

$$\hat{f}_j(x_{ij}) \pm 2 \left[diagonal_i \left(\mathbf{R}_j \mathbf{W}^{-1} \mathbf{R}_j' \phi\right)\right]^{1/2},$$

$i=1,\ldots,n$, onde $diagonal_i\left(\mathbf{R}_j \mathbf{W}^{-1} \mathbf{R}_j' \phi\right)$ corresponde ao i-ésimo elemento da diagonal da matriz $\mathbf{R}_j \mathbf{W}^{-1} \mathbf{R}_j' \phi$.

Nos modelos semiparamétricos o estimador de $\hat{\beta}$ explicitado em (21) não é consistente. Quando taxas de convergência "típicas" são utilizadas para os termos não-paramétricos, a consistência de $\hat{\beta}$ não pode ser demonstrada (Opsomer e Ruppert, 1999 e Speckman, 1988). Uma expressão aproximada para a matriz de covariância de $\hat{\beta}$ (Thurston et al., 2000) é dada por

$Var(\hat{\beta}) = (X'W(I - \mathbf{R}_{NP})X)^{-1}$.

### 3.5. Seleção do parâmetro de suavização

Um possível critério para selecionar parâmetros de suavização $\lambda_1,\ldots,\lambda_p$ em um MAG, no qual existem $p$ termos não-paramétricos $f_1,\ldots,f_p$ é baseado na estatística

$$AIC = \frac{1}{n}\sum_{i=1}^{n} D(y_i;\hat{\mu}_i) + \frac{2}{n} traço(\mathbf{R})\,\phi, \qquad (26)$$

inspirada no critério de informação de Akaike (ver Hastie e Tibshirani, 1990, por exemplo), que leva em consideração tanto o desvio quanto o *número de parâmetros* do modelo. Valores pequenos desta estatística indicam um bom ajuste do modelo.

Embora esta estatística seja muito empregada na prática, não existem resultados sobre sua utilização como um critério para a seleção do parâmetro de suavização e gráficos com a curva suavizada estimada sobreposta aos dados podem ser usados com esse objetivo.

Além disso, a escolha de $\lambda$ pode depender da natureza do problema estudado. No caso do método *loess*, por exemplo, $\lambda$ está relacionado ao tamanho da vizinhança, ou seja, ao número de pontos em cada ajuste local e, por este motivo, pode ser selecionado subjetivamente. Na análise de séries cronológicas para estudar os efeitos da poluição atmosférica sobre a morbimortalidade, recomenda-se que a escolha do parâmetro de suavização para o termo que controla a sazonalidade de longa duração seja tal que inclua as observações referentes a períodos de aproximadamente 180 dias. Por outro lado, Schwartz (1999) comenta que valores de $\lambda$ muito pequenos podem induzir autocorrelação dos resíduos e este fato também deve ser considerado na escolha de $\lambda$.

## 4. Aplicação

Para efeito de ilustração, consideremos agora os dados descritos em Pereira et al. (1998). No ajuste de (1), a modelagem do efeito do tempo de observação, das variáveis climáticas e do poluente foi evitada por intermédio de sua categorização. A razão para isso é a falta de conhecimento de uma função paramétrica adequada que relacione cada uma das variáveis preditoras com a resposta. Por exemplo, espera-se um comportamento sazonal tanto para a natimortalidade quanto para a mortalidade por causas respiratórias em idosos (Miraglia et al., 1997) e em idosos e crianças (Conceição, 2001). Portanto, espera-se uma relação não-linear entre a variável resposta e a variável de controle no período de observação (dia ou mês de observação). A forma paramétrica dessa relação funcional é difícil de ser explicitada e, portanto, é razoável incluir no modelo um termo não-paramétrico para descrevê-lo. O mesmo procedimento pode ser empregado para as variáveis climáticas e para as concentrações dos poluentes caracterizando um MAG.

A estratégia de análise foi inspirada em outros estudos na área de epidemiologia ambiental e busca controlar adequadamente as variáveis confundidoras e eliminar a autocorrelação dos resíduos. Neste caso, um MAG foi adotado para descrever a relação entre a morbimortalidade e a concentração dos poluentes. Outros autores (por exemplo, Schwartz, 1999 e Conceição et al., 2001), também adotaram uma estratégia semelhante em outro contexto. O método de suavização utilizado na estimação dos termos não-paramétricos foi o *loess*. Modelos ajustados via MLG e MAG foram comparados quanto aos seus desvios e comportamento dos resíduos.

O primeiro passo da análise foi modelar a sazonalidade de longa duração, isto é, a variação da natimortalidade em função dos 726 dias de observação. Para isto, foi considerado o modelo

$$log[E(NATMOR)] = \alpha + f_1(DIAS). \quad (27)$$

Na estimação de $f_1$, o parâmetro de suavização foi fixado inicialmente em $\lambda_1 = 0,25$. Este valor foi determinado visando a obter vizinhanças com cerca de 180 observações em cada ajuste local. Vale lembrar que observações feitas em dias consecutivos podem apresentar resíduos autocorrelacionados. Segundo Schwartz (1999), cada internação hospitalar ou morte é um evento independente e a existência de autocorrelação dos resíduos indica que alguma covariável dependente do tempo pode ter sido omitida da análise; conseqüentemente a variabilidade associada a essa variável pode ter confundido o efeito do poluente. Se a autocorrelação for removida, a variabilidade remanescente associada à covariável omitida não depende do tempo e o confundimento é menos provável. Uma medida resumo apropriada para avaliar a estrutura de autocorrelação dos resíduos é sua Soma dos Quadrados - SQA. O objetivo desta fase da análise é obter um ajuste que forneça o menor valor da SQA e com o número de observações dentro da vizinhança limitada por um ano. Para o ajuste do modelo (27) obteve-se $SQA = 0,06$; e considerando vizinhanças contendo 363 dias de observação (isto é, $\lambda_1 = 0,5$) obteve-se $SQA = 0,05$. Então, o valor de $\lambda_1$ considerado foi $0,5$.

Em seguida, seis variáveis indicadoras dos dias da semana para o controle da sazonalidade de curta duração e termos não-paramétricos para as variáveis climáticas M2TEMP e M2UMID foram adicionados, gerando o modelo

$$log[E(NATMOR)] = \alpha + f_1(DIAS) + \beta_1 SEG + ... + \beta_6 SAB + f_2(M2TEMP) + f_3(M2UMID).$$

A escolha dos valores dos parâmetros de suavização para os termos $f_2(M2TEMP)$ e $f_3(M2UMID)$, respectivamente $\lambda_2$ e $\lambda_3$, foi restrita a valores pré-fixados de 0,65, 0,8 e 0,95, pois segundo Schwartz (1994), embora a relação entre mortalidade e variáveis climáticas possa ser não-linear, parâmetros de suavização muito pequenos para estas variáveis não têm plausibilidade biológica. Foram ajustados modelos que consideram todas as possíveis combinações dos valores de $\lambda_2$ e $\lambda_3$ descritos acima e também modelos com termos lineares para *M2TEMP e M2UMID*. Entre eles, o modelo

$$log(NATMOR) = \alpha + f_1(DIAS) + \beta_1 SEG + ... + \beta_6 SAB + \beta_7 M2TEMP + \beta_8 M2UMID \quad (28)$$

foi selecionado por ter sido aquele que apresentou o menor valor da estatística *AIC* dada em (26).

Finalmente, cada poluente foi adicionado isoladamente ao modelo (28) por intermédio de um termo não- paramétrico. O parâmetro de suavização de cada um deles também foi selecionado de forma que o modelo final apresentasse o menor valor para a estatística *AIC*. Descrevemos, a seguir, o ajuste do modelo que incluiu o $NO_2$ representado por sua média móvel de cinco dias. Os ajustes dos modelos com $PM_{10}$, $SO_2$, $O_3$ e $CO$ não são apresentados por não terem sido detectados efeitos significantes desses poluentes. O modelo final ajustado foi

$$log(NATMOR) = = \alpha + f_1(DIAS) + \beta_1 SEG + ... + \beta_6 SAB + \quad (29)$$

$$\beta_7 M2TEMP + \beta_8 M2UMID + f_4(M5NO_2),$$

com $\lambda_4 = 0,8$ representando o valor selecionado do parâmetro de suavização para $f_4$. A escolha deste modelo foi baseada em ajustes com valores de $\lambda_4$ pré-fixados em *0,6*, *0,7* e *0,8*, e também no ajuste de um modelo com um termo linear para $M5NO_2$.

Para cada função não-paramétrica ajustada, digamos $\hat{f}(X)$, é possível testar a existência de efeito linear e não-linear de $X$ utilizando a estatística do desvio parcial. Por exemplo, a existência de efeito linear de $M5NO_2$ foi avaliada considerando o ajuste do modelo

$$log(NATMOR) = \alpha + f_1(DIAS) + \beta_1 SEG + ... + \beta_6 SAB + \beta_7 M2TEMP + \beta_8 M2UMID + \beta_9 M5NO_2 \quad (30)$$

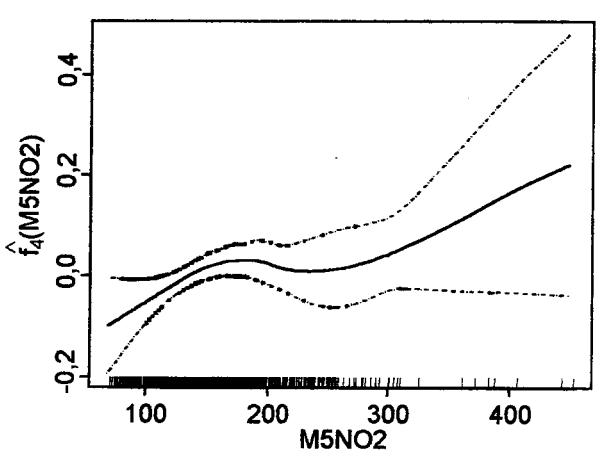
e testando $H_0$: $\beta_9 = 0$, contra $H_1$: $\beta_9 \neq 0$. Já a existência de efeito não-linear de $M5NO_2$ foi testada considerando que o modelo (30) é um submodelo de (29).

A Tabela a.2 do Apêndice apresenta os resultados obtidos sob o modelo (29). De acordo com essa tabela, conclui-se que existem efeitos linear ($p = 0,033$) e não-linear ($p = 0,043$) de $M5NO_2$ relativamente à natimortalidade. Em relação às variáveis de controle da sazonalidade, foram significativos os efeitos linear ($p < 0,001$) e não-linear ($p = 0,004$) de *DIAS*, e o efeito das variáveis indicadoras dos dias da semana ($p < 0,001$), sugerindo uma tendência. Não foram encontrados efeitos significativos para *M2TEMP* e *M2UMID*.

O desvio residual para o modelo (29) foi de 747,7 com 710,9 graus de liberdade, sugerindo um bom ajuste. Comparado com o modelo (1) ajustado via MLG, o ajuste via MAG apresentou uma economia de quase 27 graus de liberdade, embora, neste tipo de estudo, em que o número de observações é grande, tal economia pode não ser muito vantajosa.

É possível exprimir graficamente o efeito estimado de cada variável preditora do modelo na variável resposta, mantidas constantes as outras variáveis. A curva representada na Figura 3 é a contribuição estimada da $M5NO_2$ para o preditor aditivo, após o controle das variáveis confundidoras. Esta figura indica uma relação não-linear entre a resposta e o poluente. As linhas pontilhadas representam os valores ajustados $\pm 2$ vezes os respectivos erros padrões estimados.

**Curva estimativa de $f_4(M5NO_2)$ no modelo (29) e bandas de confiança pontuais.**



O maior interesse neste tipo de análise recai na estimação de medidas que avaliem o impacto do poluente na *NATMOR*, por exemplo, o risco relativo.

Na análise via MAG, o risco de mortalidade fetal tardia para uma dada concentração $M5NO_2$ do poluente, denotada por $P_j$, em relação à menor concentração observada desse poluente, denotada por $P_0$, mantidos constantes os valores de todas as outras variáveis preditoras no modelo, é dado por

$$RR_j = \frac{E(NATMOR \mid P_j)}{E(NATMOR \mid P_0)} = exp[\, f_4(P_j) - f_4(P_0)\,]. \qquad (31)$$

Denotando por $\hat{f}_4(P_j)$ o valor ajustado de $f_4$ em $P_j$ e por $\hat{f}_4(P_0)$ o valor ajustado de $f_4$ para uma concentração de referência, um estimador para (31) é

$$\hat{RR}_j = exp(\hat{d}_j),$$

com $\hat{d}_j = \hat{f}_4(P_j) - \hat{f}_4(P_0)$. Pode-se calcular o valor de $\hat{RR}_j$ para todas as concentrações observadas, e então construir o gráfico de $\hat{RR}_j$ x $P_j$, obtendo-se a curva estimada do risco relativo.

Bandas de confiança aproximadas para essas curvas podem ser construídas representando no gráfico, os pontos

$$\hat{RR}_j \pm 2\,\hat{ep}(\hat{RR}_j),$$

sendo $\hat{ep}(\hat{RR}_j)$ o erro padrão estimado de $\hat{RR}_j$. Uma sugestão para obter $ep(\hat{RR}_j)$ seria expandir $exp(\hat{d}_j)$ em série de Taylor em torno de $d_j = f_4(P_j) - f_4(P_0)$ até a primeira ordem, obtendo

$$exp(\hat{d}_j) \approx exp(d_j) + exp(d_j)\,(\hat{d}_j - d_j).$$

Então,

$$\mathrm{Var}(\hat{RR}_j) = \mathrm{Var}(\exp(\hat{RR}_j)) \approx \exp^2(\mathrm{d_j})\,\mathrm{Var}(\hat{RR}_j)$$

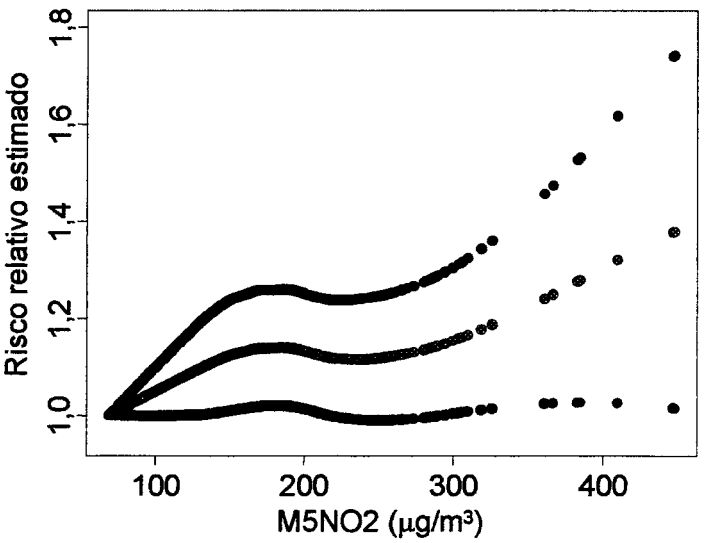
$$= \exp^2(\mathrm{d_j})\left(Var[\hat{f}_4(P_j)] + Var[\hat{f}_4(P_0)] - 2Cov[\hat{f}_4(P_j), \hat{f}_4(P_0)]\right). \qquad (32)$$

Um estimador da $\mathrm{Var}(\hat{RR}_j)$ pode ser obtido substituindo-se $d_j$ na expressão (32) por $\hat{d}_j = \hat{f}_4(\mathrm{P_j}) - \hat{f}_4(\mathrm{P_0})$, assumindo consistência de $\hat{f}$. A estimação das variâncias e covariâncias nessa expressão requer, entretanto, o conhecimento de $\boldsymbol{R}_4$, que é a matriz que produz $\hat{\mathbf{f}}_4$ a partir de $z$. Conforme dito anteriormente, o cálculo de $\boldsymbol{R}_4$ é computacionalmente dispendioso. Um artifício para obtenção dessa matriz é dado em Hastie e Tibshirani (1987).

A Figura 4 apresenta estimativas pontuais do risco relativo de mortalidade fetal tardia para diferentes concentrações da $M5NO_2$. Por esta figura, observa-se um aumento acentuado do risco relativo estimado a partir da concentração mínima até concentrações em torno de 190 $\mu g/m^3$, seguindo-se um platô e um novo aumento constante a partir de uma concentração de aproximadamente 270 $\mu g/m^3$. Não se observa a presença de uma

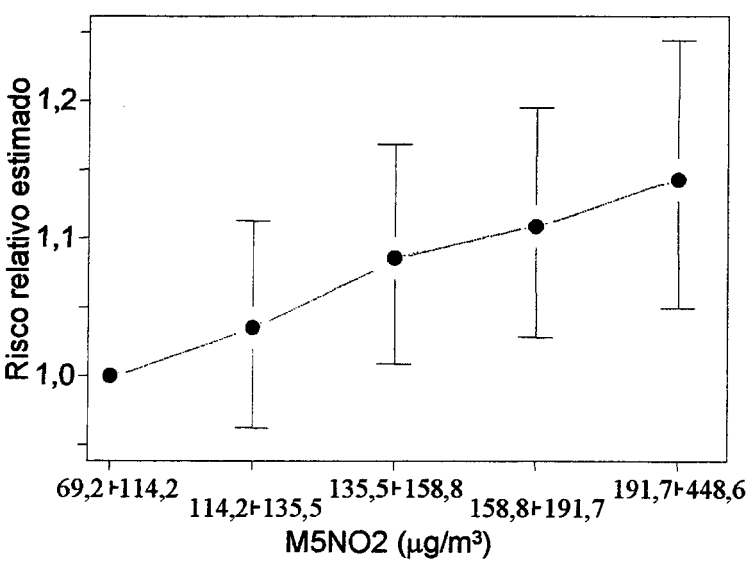
concentração limiar abaixo da qual não exista relação entre *NATMOR* e $M5NO_2$, o que sugeriria concentrações seguras.

**Figura 4 - Estimativa por ponto para o risco relativo obtido do modelo (29) e bandas de confiança pontuais aproximadas**




A ocorrência do "platô" na curva de risco estimada foi observada em outros estudos (Conceição et al., 2001, por exemplo) e uma explicação para esse comportamento está ainda sendo discutida. A idéia mais aceita é que esse fenômeno ocorre devido à existência de grupos de pessoas com diferentes suscetibilidades ao poluente dentro da população.

**Figura 5 - Estimativa por ponto e intervalo de confiança (95%) para o risco relativo obtido da análise via MLG, modelo (1)**



A Figura 5 mostra o risco relativo estimado segundo o modelo (1) para os quintis do poluente. Vários autores utilizam gráficos similares a este para representarem os resultados obtidos dos riscos relativos estimados (ver, por exemplo, Conceição et. al, 2001 ou Pereira et. al, 1998). Com esta simplificação, o risco relativo estimado para concentrações mais elevadas do poluente é limitado ao quintil de ordem cinco. Se substituíssemos a categorização da distribuição do poluente por um termo linear, a relação não-linear sugerida pela Figura 4 não seria detectada.

Um outro parâmetro de avaliação do impacto do poluente na variável resposta nesse tipo de estudo (Schwartz, 1999) é o acréscimo relativo no valor esperado da resposta, devido à variação de um intervalo interquartil na concentração do poluente, isto é,

$$\Delta = \left( \frac{\exp(f_4(P_{Q3}))}{\exp(f_4(P_{Q1}))} - 1 \right) 100\% = (exp[f_4(P_{Q3}) - f_4(P_{Q1})] - 1)\ 100\%, \qquad (33)$$

onde $P_{Q1}$ e $P_{Q3}$ são, respectivamente, os valores observados do primeiro e terceiro quartis da concentração do poluente.

No caso do modelo (29), uma estimativa para $\Delta$ pode ser obtida substituindo os valores $f_4(P_{Q3})$ e $f_4(P_{Q1})$, na expressão (33), pelos correspondentes valores estimados obtidos no ajuste desse modelo, ou seja,

$$\hat{\Delta} = (\exp[\hat{f}_4(\text{M5NO2}_{Q3}) - \hat{f}_4(\text{M5NO2}_{Q1})] - 1)\ 100\% = \qquad (34)$$

$$= (\exp[\hat{f}_4(180{,}7) - \hat{f}_4(119{,}9)] - 1)\ 100\% = 5{,}41\%.$$

Expressões para o erro padrão podem ser obtidas com base nos mesmos argumentos utilizados para o risco relativo. Então,

$$Var(\hat{\Delta}) = Var(exp[\hat{f}_4(\text{M5NO2}_{Q3}) - \hat{f}_4(\text{M5NO2}_{Q1})])$$

pode ser obtida a partir de (32), substituindo-se $P_j$ por $M5NO_{2Q3}$ e $P_0$ por $M5NO_{2Q1}$. No exemplo, tem-se $\hat{Var}(\hat{\Delta}) = 0{,}07\%$ e $\hat{ep}(\hat{\Delta}) = 2{,}57\%$. Vale notar que $\hat{\Delta}$ pode assumir valor negativo. Por esta razão, a rigor, $\hat{\Delta}$ é melhor definido como sendo o máximo entre zero e o valor da expressão (34).

Análises de resíduos para todos os modelos ajustados sugerem sua adequação aos dados.

De uma forma geral, as análises via MLG e MAG geraram conclusões coerentes sobre o efeito dos poluentes na natimortalidade, indicando que o $NO_2$ possui associação estatisticamente significante com a mortalidade fetal tardia. Este resultado está de acordo com estudos prévios que mostram existir associação entre mortalidade por causas respiratórias e $NO_X$ em crianças na cidade de São Paulo (Saldiva et al., 1994). O fato de as duas técnicas produzirem as mesmas conclusões salienta que os efeitos lineares e não-lineares das mesmas variáveis preditoras estão sendo controlados pelos modelos.

# 5. Conclusões

Neste trabalho, apresentamos de forma resumida os modelos aditivos generalizados. Embora os resultados apresentados sejam gerais, utilizamos o suavizador *loess* em seu processo de estimação devido a sua simplicidade e disponibilidade em diferentes aplicativos estatísticos. A utilização de outros suavizadores como o *B-spline* e o *P-spline* (Marx e Eilers, 1998) pode ser considerada. O uso desses suavizadores reduzem um MAG a um *MLG*, e o retroajuste é eliminado do processo de estimação.

A técnica MAG permite controle adequado das variáveis confundidoras, sem haver a necessidade de assumir uma forma funcional rígida entre essas variáveis e a resposta.

Embora a análise via MAG não forneça uma expressão analítica para descrever a relação funcional entre a resposta e as preditoras, esta técnica possui um forte apelo visual onde a forma de cada função estimada sugere a relação das variáveis preditoras com a resposta que é determinada pelos próprios dados. Além disso, este apelo visual é muito útil para sugerir um modelo paramétrico. Por exemplo, a curva $\hat{f}_4$ $(M5NO_2)$, representada na Figura 3, sugere uma relação cúbica entre a resposta e o poluente. Então, um modelo paramétrico com o termo $(M5NO_2)^3$ poderia ser utilizado em uma nova análise dos dados.

Na ausência de conhecimento sobre a função que relaciona o poluente à variável resposta, a análise via MAG permite a construção de uma curva de risco relativo que fornece melhor informação sobre o comportamento do risco relativo do que os resultados obtidos categorizando-se as concentrações do poluente.

# Apêndice

## Ilustração numérica do método de suavização *loess*

Para ilustrar a obtenção de uma curva suavizada pelo método *loess*, considere o conjunto de dados apresentados na Tabela 1. Os valores $x_i$, $i=1,\ldots,10$, foram gerados a partir de uma distribuição $N(25,16)$. De forma independente, foram gerados erros $e_i$, $i=1,\ldots,10$, segundo uma distribuição Normal padrão. As respostas $y_i$ foram obtidas pela relação $y_i = 10 + 0{,}1x_i + e_i$. O diagrama de dispersão de $(x_i,y_i)$ é mostrado na Figura 6a.

**Tabela 1- Dados (ordenados pelos valores de $x_i$) gerados segundo o *modelo* $y_i = 10 + 0{,}1x_i + e_i$, $i=1,\ldots,10$**

| i | $x_i$ | $y_i$ |
|---|---|---|
| 1 | -2,58 | 10,18 |
| 2 | 3,69 | 11,55 |
| 3 | 7,29 | 11,25 |
| 4 | 11,31 | 12,57 |
| 5 | 12,32 | 11,33 |
| 6 | 14,49 | 10,50 |
| 7 | 17,22 | 12,20 |
| 8 | 24,76 | 12,81 |
| 9 | 39,07 | 15,82 |
| 10 | 62,04 | 16,30 |

Para este exemplo, foram fixados:

• $d = 1$ (isto é, um ajuste linear local) pois o diagrama de dispersão (Figura 6a) sugere uma tendência linear

• $\lambda = 0,5$ (portanto, $r = 0,5 \times 10 = 5$ é o número de pontos efetivamente usados em cada regressão local).

A seguir são descritos os passos para obtenção da curva suavizada.

**Passo 1** – Cálculo das distâncias de um ponto alvo.

Por exemplo, para o ponto alvo $(x_3, y_3)$, calcularam-se as distâncias entre $x_3$ e $x_k$, $k=1,\ldots,10$. Na Tabela 2 observam-se os $r = 5$ valores mais próximos de $x_3$ colocados em destaque, sendo que a quinta menor distância é $h_3 = 7,21$.

**Tabela 2 - Distâncias relativas a $x_3$**

| k | $x_k$ | $\lvert x_3 - x_k \rvert$ |
|---|---|---|
| 1 | –2,58 | 9,86 |
| **2** | **3,69** | **3,59** |
| **3** | **7,29** | **0,00** |
| **4** | **11,31** | **4,03** |
| **5** | **12,32** | **5,04** |
| **6** | **14,49** | **7,21** |
| 7 | 17,22 | 9,94 |
| 8 | 24,76 | 17,47 |
| 9 | 39,07 | 31,79 |
| 10 | 62,04 | 54,76 |

**Passo 2** - Cálculo dos pesos $u_{x_3,j}$, $j=1,\ldots,10$, a serem usados na regressão local.

Os pesos $u_{x_3,j}$, $j=1,\ldots,10$, são obtidos a partir de (7) e estão dispostos na Tabela 3. A função $u_{x_3,j}$ tem seu máximo em $x_j = x_3$ e decresce à medida que os valores $x_j$ se distanciam deste valor, tornando-se zero para os pontos que satisfazem $\left|h_3^{-1}(x_j - x_3)\right| \geq 1$. Assim, apenas os pontos entre as linhas verticais pontilhadas na Figura 6b são efetivamente considerados na obtenção de $\hat{f}(x_3)$.

**Tabela 3 - Conjunto de pesos relativos a $x_3$**

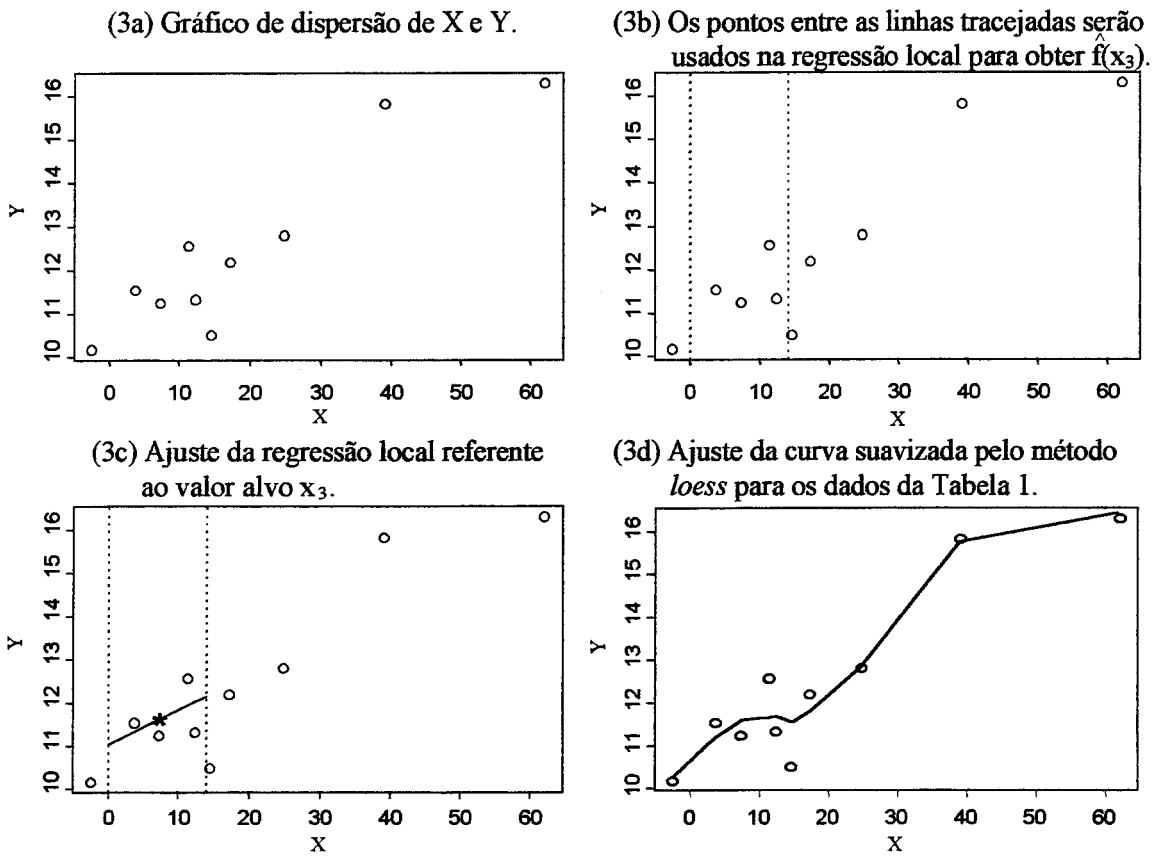
| j | $u_{x_3,j}$ |
|---|---|
| 1 | 0 |
| 2 | 0,673 |
| 3 | 1,000 |
| 4 | 0,563 |
| 5 | 0,28[illegible] |
| 6 | 0,00[illegible] |
| 7 | 0 |
| 8 | 0 |
| 9 | 0 |
| 10 | 0 |

**Passo 3** - Cálculo das estimativas dos parâmetros da reta de regressão por *MQP* e obtenção do valor ajustado.

Com pesos dados por $u_{x_3,j}$, $j=1,\ldots,10$, a reta de regressão ajustada é $\hat{y}_i = 11{,}03 + 0{,}08x_i$, que fornece o valor previsto $\hat{f}(x_3) = 11{,}59$, representado na Figura 6c como *.

Repetindo-se os passos 1 a 3, são calculados os valores ajustados correspondentes às demais observações $(x_i, y_i)$ quando cada uma delas é considerada como ponto alvo. As estimativas de $y_i$, $i=1,\ldots,10$, correspondentes estão representadas na Figura 6d.

**Figura 6 - Etapas do ajuste *loess* para os dados da Tabela 1**



## Cálculo da matriz S

Considere o ajuste da regressão ponderada local tendo $(x_3, y_3)$ como ponto alvo. Fixando $\lambda = 0,5$ e d = 1, a matriz $S_{x_3}$ é dada por

$$S_{x_3} = X\,(X'U_{x_3}X)^{-1}X'U_{x_3} = \begin{bmatrix} 0 & 1,44 & 0,61 & -0,62 & -0,44 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0,73 & 0,48 & -0,11 & -0,10 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0,32 & 0,41 & 0,18 & 0,09 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0,13 & 0,32 & 0,51 & 0,30 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0,25 & 0,30 & 0,59 & 0,35 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0,49 & 0,26 & 0,77 & 0,47 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0,80 & 0,20 & 0,99 & 0,61 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1,65 & 0,04 & 1,60 & 1,01 & 0 & 0 & 0 & 0 & 0 \\ 0 & -3,27 & -0,26 & 2,76 & 1,77 & 0 & 0 & 0 & 0 & 0 \\ 0 & -5,87 & -0,74 & 4,62 & 2,99 & 0 & 0 & 0 & 0 & 0 \end{bmatrix};$$

sua terceira linha, [0 0,32 0,41 0,18 0,09 0 0 0 0 0], de acordo com (8), corresponde à terceira linha de $S$. Repetindo o procedimento para outros pontos $(x_i, y_i)$, $i=1,\ldots,10$, tomados como alvo, obtêm-se as demais linhas e então,

$$S = \begin{bmatrix} 0,94 & 0,15 & 0,09 & -0,00 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0,18 & 0,5 & 0,30 & 0,01 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0,3 & 0,41 & 0,18 & 0,09 & 0 & 0 & 0 & 0 & 0 \\ 0 & & 0,18 & 0,37 & 0,32 & 0,13 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0,42 & 0,38 & 0,20 & 0,00 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0,15 & 0,21 & 0,30 & 0,34 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0,06 & -0,07 & 0,25 & 0,88 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0,00 & -0,04 & 0,06 & 0,98 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0,01 & 0,04 & 0,96 & 0,01 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0,00 & -0,05 & 0,09 & 0,96 \end{bmatrix}.$$

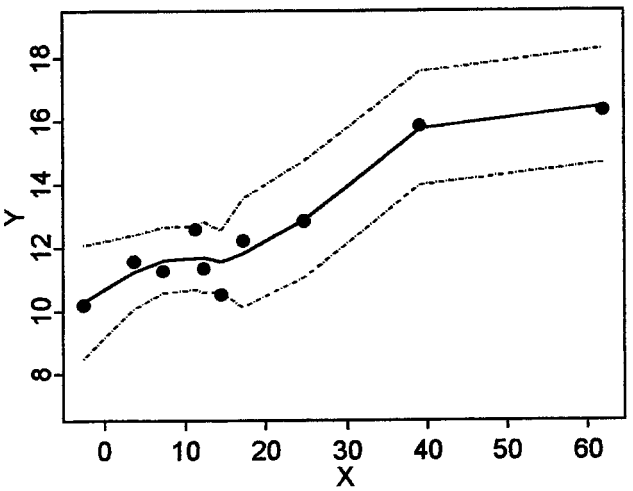
para ilustrar a influência de $\lambda$ nos elementos da matriz $S$, considere agora o mesmo conjunto de dados e assuma $\lambda = 0,4$; note que, neste caso,

$$S = \begin{bmatrix} 0,95 & 0,14 & -0,09 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0,17 & 0,52 & 0,31 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0,29 & 0,45 & 0,26 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0,72 & 0,41 & -0,13 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0,42 & 0,38 & 0,20 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0,36 & 0,36 & 0,28 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0,11 & 0,20 & 0,91 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0,04 & 0,05 & 0,99 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0,01 & 0,02 & 0,99 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0,06 & 0,09 & 0,97 \end{bmatrix}$$

apresenta maior número de elementos nulos e tem maior traço que a matriz $S$ calculada com $\lambda = 0,5$. Quanto menor o valor de $\lambda$, maior será o número de elementos iguais a zero na diagonal de $U_{x_i}$ e maior o número de elementos nulos nas linhas da matriz $S$.

A Figura 7 apresenta a curva suavizada obtida pelo método *loess* com $\lambda = 0,5$. As curvas pontilhadas representam as bandas de confiança pontuais para os elementos de $f$ com coeficiente de confiança aproximadamente igual a 95%.

**Figura 7 - Curva suavizada obtida pelo método *loess* (com $\lambda = 0,5$) para os dados da Tabela 1 e bandas de confiança pontuais**




Para os dados da Tabela 1, fixando $\lambda = 0,5$ tem-se traço$(S)$, traço$(SS')$ e traço$(2S - SS')$, respectivamente, iguais a 6,70, 6,21 e 7,19, e fixando $\lambda = 0,4$ esses valores são, respectivamente, 7,22, 6,87 e 7,59.

## Tabelas dos resultados dos ajustes dos modelos (1) e (31)

**Tabela a.1 - Estimativas dos parâmetros do modelo (1) e outras estatísticas**

| Termo | Estimativa | Erro padrão | $\chi^2$ | Graus de liberdade | Valor de p |
|---|---|---|---|---|---|
| Intercepto | 2,033 | 0,238 | | | |
| $M5NO_2a$ | 0,034 | 0,044 | 0,625 | 1 | 0,429 |
| $M5NO_2b$ | 0,082 | 0,045 | 3,499 | 1 | 0,061 |
| $M5NO_2c$ | 0,103 | 0,046 | 5,303 | 1 | 0,021 |
| $M5NO_2d$ | 0,133 | 0,052 | 6,892 | 1 | 0,009 |
| FEV91 | -0,069 | 0,088 | | | |
| MAR91 | -0,055 | 0,086 | | | |
| ABR91 | -0,182 | 0,090 | | | |
| MAI91 | -0,147 | 0,094 | | | |
| JUN91 | -0,205 | 0,098 | | | |
| JUL91 | -0,232 | 0,107 | | | |
| AGO91 | -0,421 | 0,107 | | | |
| SET91 | -0,263 | 0,102 | | | |
| OUT91 | -0,208 | 0,093 | | | |
| NOV91 | -0,267 | 0,093 | | | |
| DEZ91 | -0,316 | 0,095 | | | |
| JAN92 | -0,092 | 0,086 | | | |
| FEV92 | -0,280 | 0,095 | | | |
| MAR92 | -0,203 | 0,090 | | | |
| ABR92 | -0,203 | 0,090 | | | |
| MAI92 | -0,277 | 0,093 | | | |
| JUN92 | -0,387 | 0,099 | | | |
| JUL92 | -0,464 | 0,107 | | | |
| AGO92 | -0,383 | 0,104 | | | |
| SET92 | -0,288 | 0,097 | | | |
| OUT92 | -0,257 | 0,097 | | | |
| NOV92 | -0,216 | 0,092 | | | |
| DEZ92 | -0,223 | 0,089 | 49,825 | 23 | < 0,001 |
| SEG | 0,169 | 0,052 | | | |
| TER | 0,258 | 0,051 | | | |
| QUA | 0,263 | 0,051 | | | |
| QUI | 0,238 | 0,051 | | | |
| SEX | 0,222 | 0,051 | | | |
| SÁB | 0,152 | 0,053 | 40,507 | 6 | < 0,001 |
| M2UMID | 0,003 | 0,002 | 2,488 | 1 | 0,115 |
| M2TEMP | -0,009 | 0,012 | 0,584 | 1 | 0,445 |
| UMID1 | 0,055 | 0,049 | | | |
| UMID2 | -0,030 | 0,046 | | | |
| UMID3 | -0,036 | 0,061 | 4,229 | 3 | 0,238 |
| TEMP1 | -0,017 | 0,051 | | | |
| TEMP2 | -0,016 | 0,068 | | | |
| TEMP3 | 0,029 | 0,083 | 1,517 | 3 | 0,678 |

**Desvio Residual: 726,339 com 684 graus de liberdade.**
**AIC aproximado: 813,630.**

**Tabela a.2 - Estimativas dos parâmetros da parte linear do modelo (31) e outras estatísticas**

| Parte não paramétrica | | | | | | |
|---|---|---|---|---|---|---|
| | Efeito linear | | | Efeito não-linear | | |
| Termo | $\chi^2$ | Graus de liberdade | Valor de p | $\chi^2$ | Graus de liberdade | Valor de p |
| $f_1$(DIAS), $\lambda = 0,5$ | 19,280 | 1,0 | < 0,001 | 11,603 | 2,3 | 0,004 |
| $f_4$(M5NO$_2$), $\lambda = 0,8$ | 4,547 | 1,0 | 0,033 | 5,938 | 1,8 | 0,043 |

| Parte paramétrica | | | | | |
|---|---|---|---|---|---|
| Termo | Estimativa | Erro padrão | $\chi^2$ | Graus de liberdade | Valor de p |
| intercepto | 1,729 | 0,120 | | | |
| SEG | 0,162 | 0,051 | | | |
| TER | 0,252 | 0,050 | | | |
| QUA | 0,266 | 0,050 | | | |
| QUI | 0,240 | 0,051 | | | |
| SEX | 0,227 | 0,051 | | | |
| SÁB | 0,145 | 0,052 | 42,691 | 6,0 | < 0,001 |
| M2TEMP | 0,005 | 0,004 | 0,000 | 1,0 | > 0,999 |
| M2UMID | 0,002 | 0,002 | 2,197 | 1,0 | 0,138 |

**Desvio residual: 747,727 com 710,896 graus de liberdade.**
**AIC aproximado: 778,797.**

## Referências bibliográficas


**ANDRÉ, P.A.; BRAGA, A.L.F.; LIN, C.A.; CONCEIÇÃO, G.M.S.; PEREIRA, L.A.A.; MIRAGLIA, S.G.E.K.; BÖHN, G.M. Environmental epidemiology applied to urban atmospheric pollution: a contribution from the Experimental Air Pollution Laboratory (LPAE).** ***Rep. Public Health*****, v. 16, n. 3, p. 619-628, 2000.**

**BRAGA, A.L.F.; CONCEIÇÃO, G.M.S.; PEREIRA, L.A.A.; KISHI, H.S.; PEREIRA, J.C.R.; ANDRADE, M.F.; GONÇALVES, F.L.T.; SALDIVA, P.H.N.; LATORRE, M.R.D.O. Air pollution and pediatric respiratory admissions in São Paulo, Brazil.** J. Environmental Med., **v. 1, p. 95-102, 1999.**

**BUJA, A.; DONNELL, D.; STUETZLE, W. Additive principal components.** ***Technical Report, Department of Statistics, University of Washington, Seattle*****, 1986.**

**BUJA, A.; HASTIE, T.J.; TIBSHIRANI, R.J. Linear smoothers and additive models.** *Ann. Statist.*, **v. 17, n. 2, p. 453-555, 1989.**

**CLEVELAND, W.S. Robust locally weighted regression and smoothing scarterplots,** *J. Amer. Statist. Assoc.*, **v. 74, n. 368, p. 829-836, 1979.**

**CONCEIÇÃO, G.M.S.; MIRAGLIA, S.G.E.K.; KISHI, H.S.; SALDIVA, P.H.N.; SINGER, J.M. Air pollution and children mortality: a time series study in São Paulo, Brazil.** *Environmental Health Perspect.* **109, 347-350, 2001.**

**DOCKERY, D.W.; POPE III, C.A. Acute respiratory effects of particulate air pollution.** *Annual Rev. Public Health*, **v. 15, p. 107-132, 1994 .**

**EILERS E MARX, 1996**

**GOUVEIA, N.; FLETCHER, T. Time series analysis of air pollution and mortality: effects by cause, age and socioeconomic status.** *J. Epidemiol. Community Health*, **v. 54, p. 750-755, 2000.**

HASTIE, T.J.; TIBSHIRANI, R.J. Generalized additive models: some applications. *J. Amer. Statist. Assoc.*, v. 82 , n. 398, p. 371-386, 1987.

HASTIE, T.J.; TIBSHIRANI, R.J. Generalized additive models. *London, Chapman & Hall,* 335 p., 1990.

HOAGLIN, D.C.; WELSCH, R.E. The hat matrix in regression and ANOVA. *Amer. Statist.*, v. 32, 17-22, 1978.

LIMA, L.P. Modelos aditivos generalizados: aplicação a um estudo epidemiológico ambiental. *São Paulo.* 215p, 2001. ***Dissertação (Mestrado) – Instituto de Matemática e Estatística da Universidade de São Paulo.***

MARX, B.D.; EILERS, P.C.H. Direct generalized additive modeling with penalized likelihood. *Computational Statistics Data Analysis.* v. 28, 193-209, 1998.

MCCULLAGH, P.; NELDER, J.A. Generalized linear models. *2.ed. London, Chapman and Hall,* 511 p, 1989.

MIRAGLIA, S.G.E.K.; CONCEIÇÃO, G.M.S.; SALDIVA, P.H.N.; STRAMBI, O. Analysis of the impact of fuel consumption on mortality rates in São Paulo. In: *Third international conference on urban transport and the environment for the 21st century, Boston: Computational Mechanics Publications*, 434-444, 1997.

MORGENSTERN, H. (1995). Ecologic studies in epidemiology: concepts, principles and methods. *Annual Rev. Public Health,* v. 16, p. 61-81.

NETER, J.; KUTNER, M.H.; NACHTSHEIM, C.J.; WASSERMAN, W. (1996). Applied linear statistical models. *4.ed. Boston, McGraw-Hill,* 720 p.

OPSOMER, J.D. (2000). Asymptotic properties of backfitting estimators. *J. Multivariate Anal.*, v. 73, n. 2, p. 166-179.

OPSOMER, J.D.; RUPPERT, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. J. Comput. Graph. Statist., v. 4, n. 8, p. 715-732.

PEREIRA, L.A.A.; LOOMIS, D.; CONCEIÇÃO, G.M.S.; BRAGA, A.L.F.; ARCAS, R.M.; KISHI, H.S.; SINGER, J.M.; BÖHM, G. M.; SALDIVA, P.H.N. (1998). Association between air pollution and intrauterine mortality in São Paulo, Brazil. Environmental Health Perspect., v. 106, n. 6, p. 325-329.

ROTHMAN, K.J.; GREENLAND, S. (1998). Modern epidemiology. 2.ed. Philadelphia, Lippincott-Raven, 738 p.

SALDIVA, P.H.N.; LICHTENFELS, A.J.F.C.; PAIVA, P.S.O.; BARONE, I.A.; MARTINS, M.A.; MASSAD, E.; PEREIRA, J.C.R.; XAVIER, V.P.; SINGER, J.M.; BÖHM, G.M. (1994). Association between air pollution and mortality due to respiratory diseases in children in São Paulo, Brazil: a preliminary report. Environmental Res., v. 65, p. 218-225.

SALDIVA, P.H.N.; POPE III, C.A.; SCHWARTZ, J.; DOCKERY, D.W.; LICHTENFELS, A.J.F.C.; SALGE, J.M.; BARONE, I.A.; BÖHM, G.M. (1995). Air pollution and mortality in elderly people: a time-series study in São Paulo, Brazil. Arch. Environmental Health, v. 50, p. 159-163.

SCHWARTZ, J. (1994). Nonparametric smoothing in the analysis of air pollution and respiratory illness. Canad. J. Statist., v. 22, n. 4, p. 471-487.

SCHWARTZ, J. (1999). Air pollution and hospital admissions for heart disease in eight U.S. countries. Epidemiology, v. 10, p. 17-22.

SEN, P.K.; SINGER, J.M. (1993). Large sample methods in statistics: an introduction with applications. London, Chapman e Hall, 382 p.

SINGER, J.M.; ANDRÉ, C.D.S.; LIMA, L.P.; CONCEIÇÃO, G.M.S. (2002). Atmospheric pollution and mortality in São Paulo. Statistical Data Analysis Based on the L1-Norm and Related Methods. Switzerland, Yadolah Dodge, p. 439-450.

SPECKMAN, P.E. (1988). Regression analysis for partially linear models. J. Roy. Statist. Assoc. Ser. B, v. 50, p. 413-436.

THURSTON, S.W.; WAND, M.P.; WIENCKE, J.K. (2000). **Negative binomial additive models.** Biometrics, v. 59, p. 139-144.

ZANOBETTI, A.; WAND, M.P.; SCHWARTZ, J.; RYAN, L.M. (2000). **Generalized Additive Distributed lag Models: quantifying mortality displacement.** Biostatistics, v. 1, n. 3, p. 279-292.

## Abstract

**Generalized additive models (GAM) are an extension of generalized linear models in which the effect of each covariate on the response variable is adjusted in a very flexible form using unspecified functions. In this paper we introduce this class of models and discuss related estimation and testing procedures. For illustration purposes, this methodology is applied to the data from an environmental study focusing on the association between late fetal mortality and atmospheric pollution in the city of São Paulo, Brazil. The results are compared with those obtained from a GLM based analysis of the same data.**

**Key words: environmental study, generalized linear models, nonparametric regression, smoothing methods.**

# Modelos semiparamétricos: uma revisão

Cláudio Pereira Bidurin*

Luis Aparecido Milan **

**Resumo**

**Modelos semiparamétricos vem sendo estudados por diversos autores em função principalmente da sua flexibilidade. Estes modelos consideram um componente paramétrico, associado a uma forma preestabelecida e com suposições rígidas, e um componente não-paramétrico que provê a flexibilidade do modelo, permitindo ao mesmo se adequar ao comportamento dos dados. Estaremos abordando estes modelos, destacando as formas tradicionais de estimação do componente não-paramétrico e apresentando o modelo de regressão por *splines*. Em complementação aos modelos, uma discussão da seleção apropriada dos parâmetros de suavização é feita, destacando-se a medida de validação cruzada generalizada.**

**Palavras-chave : Modelos semiparamétricos; suavização via *kernel*; suavização por *splines*; regressão por *splines*; e validação cruzada generalizada.**

## 1. Introdução

Uma das abordagens mais utilizadas em análise de dados experimentais é o ajuste de modelos, visando a estabelecer a relação entre duas ou mais quantidades. Dentre estes modelos, destacam-se os modelos de regressão. Neste trabalho abordamos uma classe de modelos de regressão que associa um componente paramétrico e um não-paramétrico, os chamados modelos semiparamétricos ou parcialmente lineares, nomenclatura usada por Speckman (1988).

* Endereço para correspondência: Departamento de Ciências Exatas Centro Universitário Moura Lacerda.

** Departamento de Estatística da Universidade Federal de São Carlos - Caixa Postal: 676 13565-905 – São Carlos - SP - Brasil e-mail: dlam@power.ufscar.br

Tem-se ampliado bastante o uso destes modelos, principalmente devido às possibilidades de aplicação na modelagem de formas funcionais complicadas. O surgimento de resultados assintóticos sobre a eficiência e consistência dos estimadores, como em Schick (1986, 1993 e 1996) e Bhattacharya e Zhao (1997), tem ajudado a consolidar a utilização destes métodos.

A utilização com sucesso destes modelos depende fortemente do conhecimento do componente não-paramétrico via processos de suavização e, por conseqüência, existe uma dependência de processos de estimação dos parâmetros de suavização. Neste sentido, apresentamos na seção 2.1 uma definição mais geral para estes modelos, partindo dos modelos aditivos, segundo resultados obtidos por Stone (1985) e Härdle (1990).

Na seção 2.2, estendemos a discussão dos modelos aditivos para os modelos semiparamétricos, a partir de resultados mostrados por Hastie (1990) e Buja *et al.* (1989). Na seção 2.3 apresentamos uma forma de estimação do componente não-paramétrico, baseada na suavização *kernel*, segundo resultados apresentados por Silverman (1986). Na seção 2.4 a suavização por *splines* é apresentada e uma comparação entre esta e a suavização por *kernel* é feita. Na seção 2.5, apresentamos um modelo em particular, definido como de regressão por *splines*, conforme apresentado por Hastie *et al.* (1990), que apresenta uma dependência muito forte da estimação de uma malha de junções para o ajuste do modelo. Na seção 3 é feita uma discussão sobre os métodos automáticos de seleção do parâmetro de suavização associados aos modelos discutidos. Em particular, apresentamos a medida de validação cruzada generalizada, discutida por Craven *et al.* (1979). Na seção 4 fazemos uma discussão geral sobre a importância da seleção do parâmetro de suavização.

# 2. Modelos aditivos

Consideremos n pares de observações $(x_i, y_i)$, $i = 1,2,...,n$, e tomemos uma função $f$ que estabelece a relação entre as variáveis $x$ e $y$ da forma

$$y_i = f(x_i) + \varepsilon_i, \tag{2.1}$$

sendo $f$ uma função a ser estimada e supondo que, para os erros aleatórios $\varepsilon_i$, tenhamos $E(\varepsilon_i) = 0$ e $Var(\varepsilon_i) = \sigma^2$. Alternativamente, se considerarmos um conjunto de $p$ variáveis explicativas representadas em uma matriz $X$, de dimensão $n \times p$, com i-ésima linha dada por $X_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$, teremos uma função $f$ de modo que $y_i = f(x_{i1}, x_{i2}, \ldots, x_{ip}) + \varepsilon_i$.

Sobre este modelo, Stone (1985) destaca importantes aspectos: a flexibilidade, ou seja, a habilidade do modelo de prover bons ajustes a um número variado de situações; a dimensionalidade, que representa o número de observações necessárias para evitar um aumento exagerado na variância das estimativas na medida em que se aumenta a dimensão do modelo (número de parâmetros) e a interpretabilidade do modelo.

Considerando a abordagem paramétrica, como, por exemplo, a regressão linear múltipla, devemos supor uma forma *a priori* para a função $f$, totalmente especificada por um número finito de parâmetros, e ainda algumas suposições adicionais sobre a distribuição de probabilidade para os erros aleatórios.

Admitindo uma abordagem não-paramétrica, podemos adotar uma maior flexibilidade para a função $f$ em relação a sua forma. Härdle (1990) destaca que esta abordagem apresenta pelo menos quatro pontos fortes em relação à abordagem paramétrica: possibilita um método versátil para explorar a relação entre as variáveis do modelo; fornece um modelo preditivo sem a necessidade de referências a um modelo paramétrico fixo; é uma boa ferramenta para a detecção de pontos espúrios, pelo estudo da influência que cada ponto exerce no ajuste; e por ser um método com uma certa flexibilidade na substituição de observações desconhecidas *missing* pela interpolação de observações adjacentes.

Visando a tornar o modelo linear no efeito das variáveis regressoras, Buja *et al.* (1989) e Hastie e Tibshirani (1990) consideraram uma função $f$ como uma soma de funções $f_j$, $j = 1, ..., p$, para cada uma das colunas de $X$, sendo que para a i-ésima linha de $X$ temos

$$f(X_i) = f_1(x_{i1}) + f_2(x_{i2}) + ... + f_p(x_{ip})$$

e desta forma, o modelo passa a ser

$$y_i = f_1(x_{i1}) + f_2(x_{i2}) + ... + f_p(x_{ip}) + \varepsilon_i, \qquad (2.2)$$

onde $y$ é um vetor $n \times 1$ de respostas e $x_j$ é um vetor $n \times 1$ de variável explicativa, para $j = 1,...,p$, com $E(\varepsilon_i) = 0$ e $Var(\varepsilon_i) = \sigma^2$.

Cada função $f_j$, para $j = 1,...,p$, é uma função univariada arbitrária e suave a ser estimada por meio de algum mecanismo de suavização. O modelo (2.2) é chamado por Buja *et al.* (1989), e também por Friedman e Silverman (1989), de regressão aditiva ou modelo aditivo.

Sendo o modelo aditivo um modelo tipicamente não-paramétrico, as estimativas das funções $f_j$ e, por conseqüência, as estimativas para a variável resposta $\hat{y}$, são obtidas através dos chamados suavizadores lineares, ou mais simplificadamente, como adotado por Buja *et al.* (1989), tomando uma matriz de suavização $S$, e assim $\hat{y} = Sy$.

A estimação de um modelo aditivo na conceituação atual, implica conhecimento da matriz $S$, que dependerá da definição do método de suavização a ser empregado.

Os modelos aditivos podem também considerar uma mistura das abordagens anteriores, adotando um componente paramétrico e um componente não-paramétrico, modelo este chamado de modelo semiparamétrico ou parcialmente linear por Stone (1985), Hastie e Tibshirani (1990) e por Lee (1990).

## 2.1. Modelos aditivos semiparamétricos

Modificamos o modelo definido em (2.2) de forma que

$$y_i = \sum_{l=1}^{k} \beta_l x_{il} + \sum_{j=1}^{p} f_j(t_{ij}) + \varepsilon_i ,$$

admitindo um componente linear, paramétrico, juntamente com o componente não-paramétrico já existente no modelo.

Este modelo tem sido bastante explorado na literatura, devido a esta peculiaridade de compor a flexibilidade do modelo aditivo não-paramétrico com um componente paramétrico. Seguindo esta idéia, temos alguns modelos mais usados e que podem de certa forma ser considerados como modelos semiparamétricos, como discutido por Stone (1985), Hastie e Tibshirani (1990) e por Lee (1990), podendo-se destacar os modelos de regressão com erro nas variáveis, os modelos estimados por mínimos quadrados generalizados, e o mais tradicional que é o modelo parcialmente linear. Este último modelo pode ser representado de uma forma matricial, dividindo o conjunto de covariáveis ou variáveis regressoras em dois conjuntos $X$ e $T$, admitindo que $f(T) = f1(t1) + f2(t2) + ... + fp(tp)$ e escrevendo

$$Y = X\beta + f(T) + \varepsilon, \tag{2.3}$$

onde $Y$ é o vetor de respostas $n \times 1$; $X$ é uma matriz $n \times k$ de variáveis explicativas do componente paramétrico, $T$ é uma matriz de variáveis explicativas do componente não-paramétrico, com $T$ real; $f$ é uma função desconhecida, suavizável; $\beta$ é um vetor de parâmetros desconhecido e $\varepsilon$ é o vetor de erros aleatórios, independentes com $E(\varepsilon_i)=0$ e $Var(\varepsilon_i)=\sigma^2$.

Segundo Buja *et al.* (1989) pode-se entender o modelo (2.3) como uma generalização dos modelos aditivos, e tomando observações $(y_i, x_i, t_i)$ independentes, *com* $i = 1,...,n$, temos que o conjunto de parâmetros $\beta$ e $f$ serão estimados considerando-se um procedimento qualquer de suavização.

Green *et al.* (1985), utiliza este tipo de modelo introduzindo o conceito de mínimos quadrados penalizados, abordando um problema de experimentos fatoriais, com estimações simultâneas dos componentes paramétricos e não-paramétricos utilizando uma matriz de suavização $S$ arbitrária. Engle *et al.* (1986), adota o modelo dado por (2.3), utilizando o mesmo conceito de mínimos quadrados penalizados com uma matriz de suavização obtida através de suavização por *splines*, como originalmente proposto por Whittaker (1923).

Utilizando suavização por *splines*, Heckman (1986) apresenta resultados assintóticos e razão de convergência $\sqrt{n}$ para a estimativa do componente paramétrico do modelo.

Rice (1986) obtem o vício assintótico dos estimadores para o modelo (2.3) e, similarmente, Chen (1988) obtem estimativas para o componente paramétrico de (2.3), com uma razão de convergência de ordem $\sqrt{n}$ com variância mínima.

Speckman (1988) apresentou uma extensão do trabalho apresentado por Green *et al.* (1985), adotando agora uma matriz de suavização $S$ obtida por *kernel*, obtendo os vícios assintóticos para os componentes paramétrico e não-paramétrico do modelo.

Uma contribuição de interesse é feita por Buja *et al.* (1989), onde foram apresentados diversas técnicas de estimação aplicáveis a este tipo de modelo, sendo apresentado um critério baseado nos autovalores e decomposição em valores singulares das matrizes de suavização para avaliar o desempenho das mesmas.

Cuzick (1992a) aborda o modelo semiparamétrico (2.3) segundo duas possíveis matrizes de suavização, uma baseada em médias locais e outra, em mínimos quadrados locais, obtendo estimativas eficientes para o

componente paramétrico supondo a distribuição dos erros aleatórios conhecida. Este trabalho foi estendido em Cuzick (1992b), onde os resultados são obtidos admitindo-se a distribuição dos erros desconhecida.

Além destes trabalhos, podemos destacar outros, como Schick (1986), Schick (1993), Schick (1996), Bhattacharya e Zhao (1997) que enriqueceram a teoria assintótica acerca do modelo semiparamétrico (2.3), representando assim uma ponte entre os resultados assintóticos já conhecidos para os modelos puramente não-paramétricos e os modelos semiparamétricos. Por (2.3) ser um tipo de modelo aditivo, e conter um componente não-paramétrico, da mesma forma como os autores citados anteriormente, devemos admitir algum procedimento de estimação não-paramétrica (algum mecanismo de suavização), para a estimação do modelo. De acordo com Silverman (1986), Buja *et al.* (1989) e Härdle (1990), podemos destacar dois dos principais suavizadores, via *kernel* e *splines*.

## 2.2. Suavização por *Kernel*

Seja o modelo (2.1) e as observações $(y_i, x_i)$ para $i=1,...,n$, e, por simplicidade e sem perda de generalidade, tomemos o domínio $D$ de interesse para $f$ como sendo $D=[0,1]$. Um procedimento bastante simplificado para estimar $f$ consiste em estabelecer uma média localmente ponderada,

$$\hat{f}(x_0) = n^{-1} \sum_{i=1}^{n} W_n(x_0) y_i ,$$

onde a matriz $W_n$ representa os pesos a serem atribuídos às observações que estarão presentes na composição da média ponderada. A definição de $W_n$ estabelece o funcionamento do método de suavização. Na suavização por *kernel*, esta matriz é obtida através da função *kernel* $K(u)$. Segundo Silverman (1986), $K(u)$ deve ser uma função real, contínua, simétrica e limitada, tal que $\int K(u)du = 1$. Dessa maneira, a suavização em um ponto $x_0$ é dada por,

$$\hat{f}(x_0) = \frac{\sum_{i=1}^{n} K_h(x_0 - x_i) y_i}{\sum_{i=1}^{n} K_h(x_0 - x_i)},$$

com a função $K_h$ dependendo da função *kernel* $K(u)$ definida na forma

$$K_h = \frac{K\left(\frac{x_0 - x_i}{h}\right)}{h},$$

onde $h$ é o parâmetro de suavização e determina como a vizinhança de $x_0$ será utilizada para a composição da média ponderada. A função *kernel* $K(u)$ determina os pesos, podendo se apresentar de várias formas como, por exemplo, uma distribuição normal padrão,

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right).$$

A aplicação desta técnica depende da escolha da função *kernel* e da definição do valor de $h$. O primeiro aspecto parece pouco relevante, visto que a mudança de $K(u)$ não causa diferenças significativas no ajuste. No entanto, o segundo aspecto é relevante, visto que, se tomarmos o valor de h muito baixo em relação a variação de $x$, podemos estar deixando de suavizar a função $f$ e, se tomarmos $h$ muito alto, podemos estar suavizando em excesso. Para ilustrar, tomemos uma função simulada da forma

$$y_i = 4\, sen(0.83\, x_i) + e_i, \tag{2.4}$$

onde $x_i$ é uma uniforme (0,10) e $e_i$ um erro aleatório com distribuição normal (0,1), para $i = 1,...,20$. Vejamos três possibilidades de suavização, utilizando um $h$ ótimo e dois valores de $h$ extremos. Para tanto, estaremos fixando o *kernel* triangular, definido por,

$$K(u) = \begin{cases} 1 - |u| & |u| < 1 \\ 0 & cc \end{cases}$$

O resultado da simulação pode ser observado nas Figuras de 1 a 3 que seguem. Nas Figuras, a função gerada por (2.4) é mostrada pelos pontos, a curva teórica é mostrada pela linha pontilhada e a curva gerada mostrada pela linha cheia.

A Figura 1 mostra um $h$ adequado que produz um bom ajuste, a Figura 2 mostra o ajuste resultante de um $h$ pequeno, com falta de ajuste e a Figura 3 mostra o excesso de ajuste resultante de um $h$ muito alto.

Esta dependência sobre o parâmetro de suavização vem reforçar a importância do uso adequado de medidas automáticas de seleção.

**Figura 1: Suavização por *kernel* triangular com h = 2,2.**

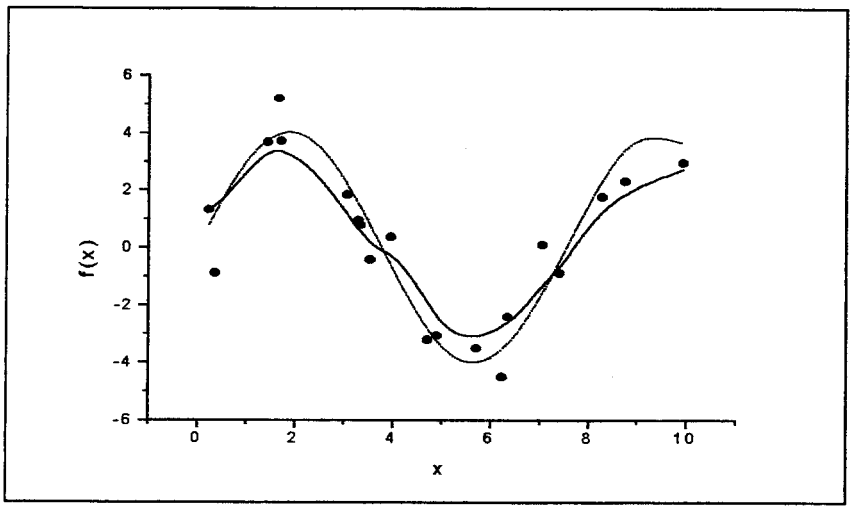



**Figura 2: Suavização por *kernel* triangular com h = 0,5.**

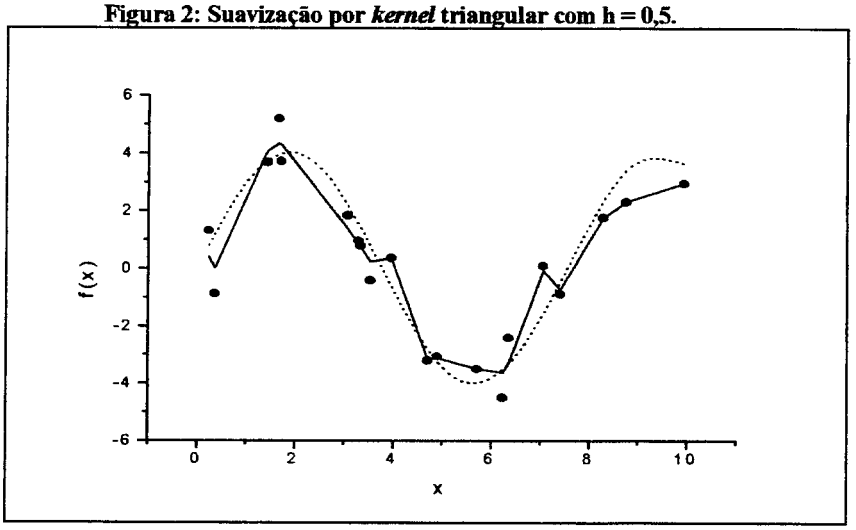
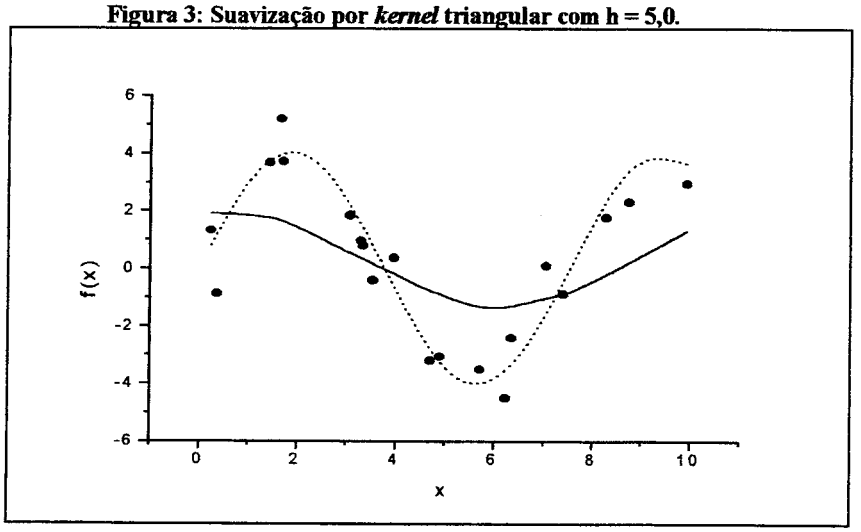



**Figura 3: Suavização por *kernel* triangular com h = 5,0.**



## 2.3. SUAVIZAÇÃO POR *SPLINES*

O termo *splines* tem sua origem na área de desenho, onde se refere a um instrumento utilizado para se traçar curvas suaves, passando por pontos pré-fixados, de forma a fazer concordâncias entre duas ou mais curvas. A transposição desta idéia ao ajuste de modelos é feita de forma a termos uma regressão polinomial por pedaços, chamada de *smoothing splines*.

A função *splines* a ser ajustada obedece a alguns requisitos quanto a aspectos relativos à segmentariedade, continuidade e suavidade. Segmentariedade significa que entre dois pontos pré-fixados consecutivos devemos ter uma função completamente definida. Continuidade implica a continuidade matemática da função em todo o seu domínio D. Suavidade implica a continuidade matemática das derivadas da função.

A aplicação de *smoothing splines* exige o estabelecimento de pontos pré-fixados ao longo do domínio $D$ da função. Estes pontos possuem diversas denominações na literatura, tais como: *knot*, nó, junção ou junta. Estaremos adotando a nomenclatura junção. Temos então um conjunto de junções associadas à função *splines*, sendo que o conjunto destas junções mais os pontos extremos do domínio $D$ é denominado malha do *splines* e é denotado por $\Lambda$.

Um *Splines* polinomial $S_\Lambda(X)$ com junções em $\Lambda = \{\lambda_0, \lambda_1, ..., \lambda_{k-1}, \lambda_k\}$ de grau m é uma função real tal que

i) em cada intervalo $\{\lambda_{i-1}, \lambda_i\}$ para $i=1,...,k$, a função *splines* $S_\Lambda(X)$ é dada por um polinômio de grau no máximo $m$ com pelo menos um intervalo onde o grau do polinômio é $m$;

ii) a função $S_\Lambda(X)$ é contínua em $D$; e

iii) as $(m-1)$ primeiras derivadas de $S_\Lambda(X)$ são contínuas em $D$.

Partindo desta definição de *splines*, temos diversas famílias de funções, destacamos aqui os *splines* cúbicos naturais, considerados inicialmente por Handscomb (1966).

O ajuste de *splines* considera inicialmente uma medida para se verificar o ajustamento dos dados a uma determinada curva $f(x)$ que é a soma de quadrados de resíduos

$$\sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2 .$$

Se $f$ é assumida com uma forma irrestrita (conceito não-paramétrico), então o valor da soma acima pode ser reduzida a zero para alguma curva em particular, e para algum comportamento particular dos dados. A curva estimada não é única. O alisamento por *splines* busca um equilíbrio entre o interesse de se produzir um bom ajuste para os dados com o desejo de se produzir suavizações sem variações locais bruscas.

Como mostrado por Handscomb (1966), as derivadas da função $f$ medem a falta de suavidade do ajuste. Com isso podemos definir uma penalização para a falta de suavidade do alisamento da forma

$$\int_a^b f^{(p)}(z)^2 \, dz$$

para a e b tais que $a \leq x_1 \leq \ldots \leq x_n \leq b$ e $p$ um parâmetro relacionado com o grau dos polinômios a serem ajustados. No caso dos *splines* cúbicos discutidos anteriormente temos $p = 2$.

Compondo então estes dois critérios, temos a forma tradicional de *smoothing splines*, que na verdade é a procura por uma função $f(x)$ com duas derivadas contínuas que minimiza a soma de quadrados de resíduos penalizada,

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int_a^b f^{(2)}(z)^2 dz , \tag{2.5}$$

onde a primeira parcela penaliza a falta de ajuste da função de regressão aos dados e a segunda parcela penaliza a falta de suavidade da função $f$ e $\lambda$ é o parâmetro de suavização.

A solução de (2.5) é única e corresponde ao *splines* cúbico natural com junções coincidentes com os valores observados de $x_i$. A solução dependerá fortemente da constante $\lambda$, sendo que para $\lambda \rightarrow \infty$ o termo de penalização domina, forçando que $f^{(2)}(x) = 0$. Nesta situação, a solução se reduz ao ajuste de uma regressão por mínimos quadrados tradicional. Para $\lambda \rightarrow 0$, o termo de penalização torna-se irrelevante e assim a solução será uma função interpoladora duplamente diferenciável.

Uma comparação é inevitável, entre o parâmetro de suavização $\lambda$ e o seu correspondente em *kernel*, o parâmetro h. Para ilustrar o comportamento da solução em função de $\lambda$, considerando a mesma simulação da seção 2.3, variamos o valor de $\lambda$ para a estimação da função $f(x)$. O resultado desta simulação pode ser observado nas Figuras de 4 a 6, onde a função gerada por (2.4) é mostrada pelos pontos, a curva teórica é mostrada pela linha pontilhada e a curva ajustada é mostrada pela linha cheia.

A Figura 4 mostra um $\lambda$ adequado que produz um bom ajuste, a Figura 5 mostra o excesso de suavização resultante de um valor alto de $\lambda$ e a Figura 6 mostra o ajuste resultante de um $\lambda$ muito pequeno, com falta de ajuste. Existe uma relação entre o parâmetro $\lambda$ adotado e o número de junções, e de fato, considerando-se que o número de graus de liberdade do modelo é igual ao traço da matriz de suavização associada $S_\lambda$ e o número de junções igual ao número de graus de liberdade menos o número de parâmetros do modelo.

**Figura 4 - Ajuste por *smoothing spline* com 7 junções.**

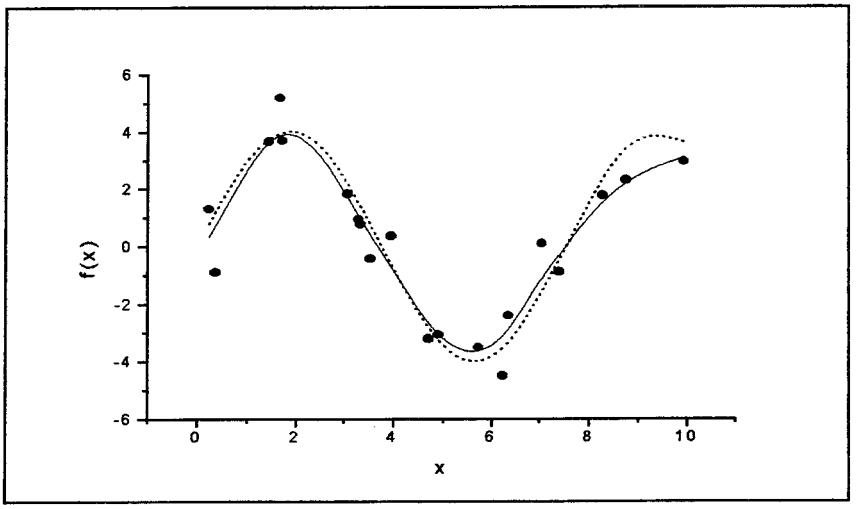



**Figura 5 - Ajuste *smoothing spline* com 4 junções**

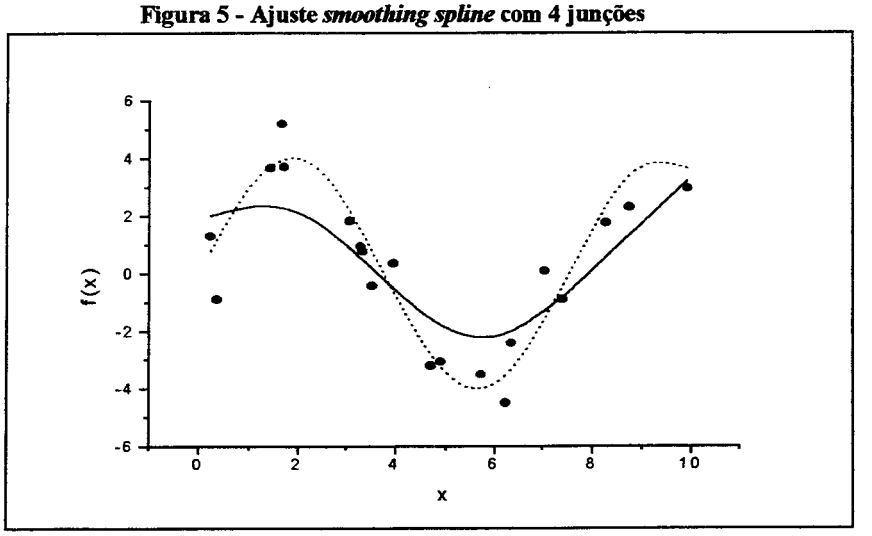
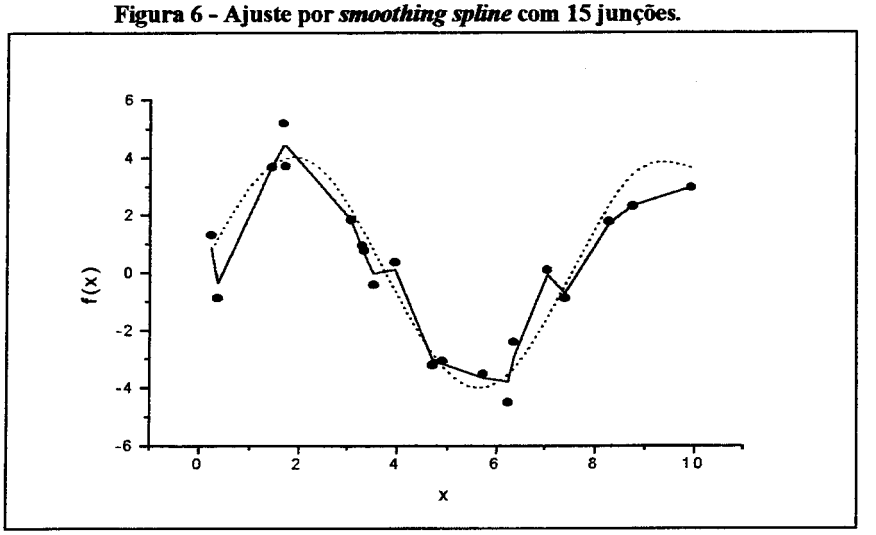



**Figura 6 - Ajuste por *smoothing spline* com 15 junções.**



*Splines* cúbicos têm sido utilizados por vários autores. Silverman (1985) aborda a flexibilidade dos *splines* na estimação de densidades e o uso dos resíduos do modelo como ferramenta para diagnóstico. Speckman (1985) apresenta um estimador minimax baseado em *splines* naturais, em relação ao erro quadrático médio. Além destes, podemos citar Eubank (1994) que descreve de forma bastante didática a formulação e aplicação dos *splines* cúbicos.

Restringimos a apresentação de *smoothing splines* para o modelo estritamente não-paramétrico. Nosso interesse agora é colocar esta abordagem na forma dos modelos aditivos semiparamétricos, ou seja, adotando-se o modelo (2.3). Engle *et al.* (1986) apresenta a forma de penalização

$$\sum_{i=1}^{n}[y_i - x_i'\beta - f(x_i)]^2 + \lambda\int_a^b f^{(2)}(z)^2\,dz\,, \tag{2.6}$$

com as mesmas propriedades já discutidas anteriormente para a forma (2.5). A mesma forma foi adotada por Heckman (1986) e Cuzick (1992a).

Engle *et al.* (1986), Buja *et al.* (1989) e também Härdle (1990) adotam um operador diferença de segunda ordem $\Omega$. Admitindo que o intervalo (a,b) possa ser subdividido em um número $m$ de intervalos ($m$ grande) e definindo $\delta_1,\ldots,\delta_m$ como os respectivos pontos médios dos intervalos, obtiveram estimativas para o modelo.

Inicialmente, reescrevemos o modelo (2.3) como proposto por Engle *et al.* (1986), da forma matricial,

$$Y = X\beta + L\delta + \varepsilon, \tag{2.7}$$

onde $Y$ é um vetor de respostas $n \times 1$, $X$ é uma matriz $n \times p$ de variáveis explicativas, $\beta$ é um vetor $p \times 1$ de parâmetros desconhecidos, $L$ é uma matriz de suavização $n \times n$ e $\delta$ é um vetor $n \times 1$ de parâmetros desconhecidos e $\varepsilon$ é o vetor de erros aleatórios, independentes com $E(\varepsilon_i)=0$ e $Var(\varepsilon_i)=\sigma^2$.

Desta forma, podemos representar a penalização dada em (2.6) através de normas quadráticas

$$||\, Y - X\beta + L\delta||^2 + \lambda||V\delta||^2,$$

onde $||V\delta||^2$ aproxima o termo que penaliza a falta de suavidade do ajuste, e $V$, como mostrado por Buja *et al.* (1989), é uma matriz tridiagonal de ordem $n \times n$ definida da forma

$$V = \Omega' C^{-1} \Omega$$

onde $C$ é uma matriz tridiagonal simétrica de ordem $(n-2)$ com elementos $c_{i-1,i} = c_{i,i-1} = d_i/6$ e $c_{i,i} = (d_i + d_{i+1})/3$, com $d_i = x_{i+1} - x_i$ e $\Omega$ é uma matriz tridiagonal de ordem $(n-2)$ x $n$ com elementos $\Omega_{ii} = 1/d_i$, $\Omega_{i,i+1} = -(1/d_i + 1/d_{i+1})$ e $\Omega_{i,i+2} = 1/d_{i+1}$, para $i = 1, \ldots, n-1$.

Reescrevendo (2.7) de forma a unificar as matrizes $X$ e $L$ e os vetores $\beta$ e $\delta$ obtemos

$$Y = Z\theta + \varepsilon, \tag{2.8}$$

onde $Z$ é uma matriz $n$ x $(p+n)$ composta pelas submatrizes $X$ e $L$, $Z = [X:L]$, e $\theta$ é um vetor de ordem $(p+n)$ x $1$ composto pelos subvetores $\beta$ e $\delta$, $\theta = [\beta:\delta]^t$.

Com esta mudança, a função a ser minimizada passa a ser

$$G(\theta) = || Y - Z\theta ||^2 + \lambda ||U\theta||^2,$$

onde a matriz $U$ de ordem $n$ x $(p+n)$ é formada pela matriz $V$ e por uma matriz de zeros de ordem $n$ x $p$ de forma que $V\delta = U\theta$. O objetivo então passa a ser estimar o vetor de parâmetros $\theta$ que minimiza $G(\theta)$.

Derivando a função $G(\theta)$ em relação a $\theta$ e igualando a zero temos

$$\frac{\partial G(\theta)}{\partial \theta} = 0 \Rightarrow \frac{\partial \|Y - Z\theta\|^2}{\partial \theta} + \frac{\partial \lambda \|U\theta\|^2}{\partial \theta} = 0$$

$$-2 Z^t Y + 2 Z^t Z \hat{\theta} \beta + 2 \lambda U^t U \hat{\theta} = 0 \quad \Rightarrow \quad (Z^t Z + \lambda U^t U) \hat{\theta} = Z^t Y$$

e, portanto, temos o estimador

$$\hat{\theta} = (Z^t Z + \lambda U^t U)^{-1} Z^t Y$$

Note que este estimador é similar ao estimador obtido em regressão *ridge*, sendo que para obtê-los basta fazer $U^t U = I$.

Podemos mostrar facilmente que para um dado valor de $\lambda$,

$$E[\hat{\theta}] = (Z^t Z + \lambda U^t U)^{-1} Z^t Z \theta$$

e que

$$\mathrm{Var}[\hat{\theta}] = \sigma^2 (Z^t Z + \lambda U^t U)^{-1} Z^t Z (Z^t Z + \lambda U^t U)^{-1}.$$

Portanto, $\hat{\theta}$ é um estimador viciado do parâmetro $\theta$ do modelo (2.8) e sua variância cresce conforme o valor de $\lambda$ decresce.

Podemos definir uma matriz de suavização associada a este modelo da forma

$$S_\lambda = Z (Z^t Z + \lambda U^t U)^{-1} Z^t.$$

Com esta matriz podemos obter os valores preditos $\hat{y} = S_\lambda y$ e o número de graus de liberdade associados aos resíduos da regressão dado pelo traço da matriz $I - S_\lambda$ ou $Tr(I-S_\lambda)$. Note que a forma colocada para os valores preditos é similar a forma geral apresentada para os valores suavizados na seção 2.

Nos estimadores anteriores, a dependência em relação à matriz de suavização $S$ e do parâmetro de suavização $\lambda$ é evidente. Neste sentido, coloca-se a questão da obtenção desta matriz. Isto foi tratado nos trabalhos apresentados por Silverman (1984) e Messer (1991) que estabeleceram as relações entre as estimativas por *smoothing spline* e por *kernel*.

Cada uma das técnicas possui uma forma de aplicar suavização que está diretamente ligada à forma da matriz de suavização e, portanto, interferindo na obtenção das estimativas de interesse. Por este fato é que uma discussão de técnicas de seleções destes parâmetros será considerada mais adiante. Isto evidência a importância da escolha do método suavização e dos parâmetros envolvidos.

## 2.3. Regressão por *SPLINEs* (*Regression Splines*)

Segundo Hastie e Tibshirani (1990) uma desvantagem de se adotar o modelo semiparamétrico com estimação baseada em métodos não-paramétricos é que a matriz de suavização $S$, de ordem $n$ e de posto completo, não pode ser assumida como uma matriz de projeção. Com isso, uma alternativa natural é procurar uma abordagem paramétrica flexível para o ajuste de cada um dos termos do modelo aditivo. Neste sentido, teríamos a matriz de suavização que seria, na verdade ,uma matriz de projeção com dimensão menor que $n$.

Uma possibilidade é adotar a regressão por *splines*, onde temos um preditor aditivo representado da forma

$$Y = \sum_{j=0}^{k} \beta_j X^j + \sum_{i=1}^{h} \theta_i (x-\lambda_i)_+^k + \varepsilon ,$$

com $\varepsilon$ atendendo às mesmas suposições já estabelecidas,

$$(x-\lambda_i)_+ = \begin{cases} (x-\lambda_i) & (x-\lambda_i) > 0 \\ 0 & (x-\lambda_i) \le 0 \end{cases},$$

e $\lambda_i$, $i = 1, 2, ..., h$, são as junções fixas sujeitas à restrição $\lambda_1 < \lambda_2 < ... < \lambda_h$.

Particularizando para o caso de *splines* cúbicos naturais, chega-se a um modelo da forma

$$E\{Y\} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{i=1}^{h} \theta_i (x-\lambda_i)_+^3 . \qquad (2.9)$$

Assim como discutido na seção 2.3, assumimos que o modelo (2.9) seja um polinômio cúbico para qualquer subintervalo $[\lambda_i , \lambda_{i+1})$ e que tenha duas derivadas contínuas. Este modelo pode ser representado

matricialmente como $Y = X\beta + \varepsilon$, onde $Y$ é um vetor de respostas $n \times 1$, $X$ é uma matriz de variáveis explicativas de ordem $n \times p$ (com $p = h+4$), $\beta$ é um vetor de parâmetros $p$ dimensional e $\varepsilon$ é um vetor de erros aleatórios normais, independentes com $E(\varepsilon_i) = 0$ e $\text{Var}(\varepsilon_i) = \sigma^2$.

Neste caso, a aplicação do método de mínimos quadrados para a estimação de $\beta$ é imediata e resulta no estimador

$$\hat{\beta} = \left(X^t X\right)^{-1} X^t y .$$

A obtenção deste estimador depende de dois aspectos: O primeiro relacionado ao número de junções a serem consideradas, que definirá o número de parâmetros no modelo e, por conseguinte, o número de graus de liberdade. O segundo está relacionado com a matriz $X$, que pode apresentar problemas de quase colinearidade entre suas colunas. Dependendo do número e da distribuição das junções, podemos ter o primeiro vetor do polinômio truncado praticamente igual ao vetor obtido, tomando-se $x^3$. Por esta razão, adota-se uma mudança na base do espaço coluna de $X$. Esta mudança de base pode ser feita adotando-se a base B-*splines*, devida a De Boor (1978), que transforma o modelo (2.9) na forma

$$E\{Y\} = \sum_{j=-1}^{h+2} \psi_j B_j(x) ,$$

onde os B-*splines* são definidos considerando médias de diferenças divididas e $\psi_j$ são os novos parâmetros do modelo.

Para *splines* cúbicos via B-*splines* temos a forma

$$B_j(x) = \frac{\sum_{k=j-2}^{j+2} (x - \lambda_k)_+^3}{\prod_{\substack{s=t-2 \\ s \neq j}}^{t+2} (\lambda_j - \lambda_s)} ,$$

e as junções adicionais $\lambda_j$ são definidas por

$$\lambda_j = \begin{cases} \lambda_1 - (1-j)(\lambda_1 - x_{\min}), & \text{para } j \leq 0 , \text{ e} \\ \lambda_h + (j-h)(x_{\max} - \lambda_h), & \text{para } j \geq h+1 , \end{cases}$$

sendo que $x_{min}$ e $x_{max}$ representam, respectivamente, o menor valor e o maior valor observado para a variável explicativa $x$. A base B-*splines* satisfaz a propriedade $B_j(x) = 0$, para $x > \lambda_{j+2}$ e para $x < \lambda_{j-2}$.

Nota-se que o número de parâmetros desconhecidos $\psi_j$ é o mesmo que o número de parâmetros livres adotados no modelo (2.9) e espera-se que a utilização da base B-*splines* propicie um processo de estimação mais estável do ponto de vista numérico.

# 3. Seleção do parâmetro de suavização

É freqüente a utilização de métodos subjetivos, tais como análises gráficas, para a escolha do parâmetro de suavização ou a aplicação de critérios que estabelecem uma relação entre o parâmetro de suavização e o tamanho da amostra. Em contrapartida, surgiram diversos métodos de seleção automática do parâmetro de suavização para modelos puramente não-paramétricos. Alguns destes métodos vêm sendo adaptados para o uso com os modelos semiparamétricos.

Uma técnica de seleção automática do parâmetro de suavização, que denotaremos de forma geral por $\lambda$, é a minimização da esperança do erro quadrático médio *(MSE)*, que adotando o modelo (2.1) é definido por

$$MSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E\left\{\hat{f}(x_i) - f(x_i)\right\}^2 .$$

Uma outra medida relacionada com esta última é o erro quadrático predito *(PSE)*, definido por

$$PSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E\left\{\hat{f}(x_i) - y_i^*\right\}^2 ,$$

onde $y_i^*$ é uma nova observação para $x_i$, da forma $y_i^* = f(x_i) + \xi_i$, com $\xi_i$ independente de $\varepsilon_i$.

Desta medida surgiu a validação cruzada ordinária, que leva em conta os valores preditos considerando *(n-1)* observações no modelo. Esta medida que denotaremos por $CV(\lambda)$, é da forma

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{f}^{-i}(x_i)\right\}^2 ,$$

onde $\hat{f}^{-i}(x_i)$ indica o valor predito por um modelo ajustado sem a utilização da i-ésima observação. Adotando esta medida, temos que estabelecer um método de busca do valor do parâmetro $\lambda$ que minimiza o valor de $CV(\lambda)$.

O uso de $CV(\lambda)$ se justifica pelo fato das medidas anteriores terem um relacionamento indicando que a minimização da validação cruzada ordinária ocorre para um $\lambda$ que também minimiza as demais medidas.

Este relacionamento pode ser visto considerando que

$$E\left\{y_i - \hat{f}^{-i}(x_i)\right\}^2 = E\left\{y_i - f(x_i) + f(x_i) - \hat{f}^{-i}(x_i)\right\}^2 = \sigma^2 + E\left\{f(x_i) - \hat{f}^{-i}(x_i)\right\}^2 , \qquad (3.1)$$

visto que o produto cruzado envolvido na expressão acima tem esperança igual a zero, já que $\hat{f}^{-i}(x_i)$ não envolve $y_i$. Analogamente temos que

$$E\left\{y_i^* - \hat{f}(x_i)\right\}^2 = \sigma^2 + E\left\{f(x_i) - \hat{f}(x_i)\right\}^2 ,$$

ou seja, temos a relação

$$PSE(\lambda) = MSE(\lambda) + \sigma^2. \tag{3.2}$$

Assumindo que $\hat{f}^{-i}(x_i) \approx \hat{f}(x_i)$ e comparando (3.1) com (3.2), vemos que $E[CV(\lambda)] \approx PSE(\lambda)$ e, portanto, $E[CV(\lambda)] \approx MSE(\lambda) + \sigma^2$.

Apesar de ser uma medida bastante útil, como mostrado por Rice (1984) considerando modelos não-paramétricos e por Lee (1990) na conceituação de modelos semiparamétricos, esta é uma medida que exige grande esforço computacional. Uma modificação apresentada por Craven e Wahba (1979) e discutida por Golub *et al.* (1979) chamada de validação cruzada generalizada *(GCV)*, estabelece uma relação entre $\hat{f}^{-i}(x_i)$ e a diagonal da matriz de suavização S associada ao modelo, da forma

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E\left\{\frac{y_i - \hat{f}(x_i)}{1 - Tr(S)/n}\right\}^2, \tag{3.3}$$

ou matricialmente da forma

$$GCV(\lambda) = \frac{\frac{1}{n}\|(I - A)y\|^2}{\left[\frac{1}{n}Tr(I - A)\right]^2},$$

onde $A$ é uma matriz de projeção e $Tr(I\text{-}A)$ é o traço da matriz $(I\text{-}A)$.

Um aspecto de interesse desta medida, além da maior simplicidade de cálculo em comparação com a validação cruzada ordinária, é a relação que esta possui com a estatística $C_p$ de Mallow, como mostrado por Hastie e Tibshirani (1990).

A estatística $C_p$ de Mallow é definida da forma,

$$C_p(\lambda) = ASR(\lambda) + \frac{2Tr(S)\hat{\sigma}^2}{n},$$

onde $ASR$ é o quadrado médio de resíduos.

De (3.3), utilizando a aproximação $(1\text{-}x)^{-2} \approx 1+2x$, obtemos

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{f}(x_i)\right\}^2 + 2Tr(S)\frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{f}(x)\right\}^2.$$

Então, tomando

$$\sum_{i=1}^{n}\left\{y_i - \hat{f}(x_i)\right\}^2 = \hat{\sigma}^2,$$

e considerando $ASR(\lambda) \approx \sum_{i=1}^{n} \{y_i - \hat{f}(x_i)\}^2$, temos que $GCV(\lambda) \approx C_p(\lambda)$.

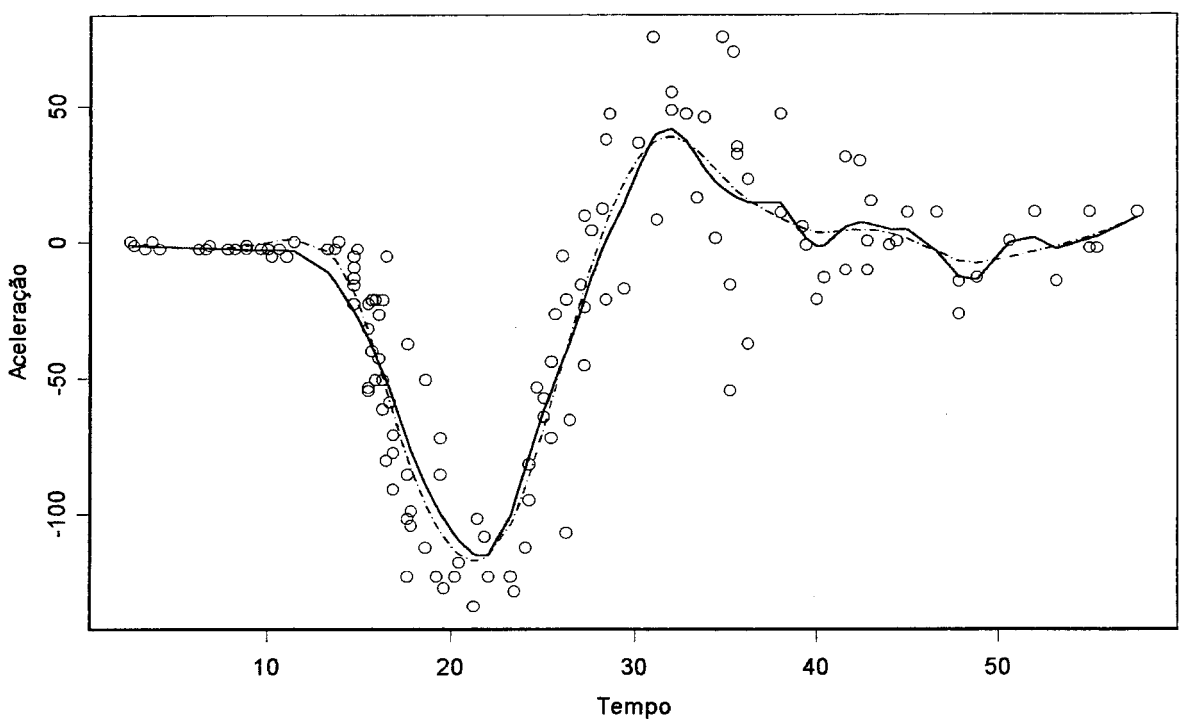
Justifica-se assim a utilização da medida GCV para o propósito de seleção do melhor parâmetro de suavização, ou no caso de *regression splines*, para a seleção do número de junções. É claro que além desta medida, podemos adotar outras, como, por exemplo, o próprio quadrado médio de resíduos, o coeficiente de explicação ajustado e também uma estatística $F$ para o modelo.

# 4. Aplicação

Ilustramos as técnicas descritas através de um conjunto de dados que se refere à simulação de colisões com motocicletas para verificar a eficácia da proteção de capacetes. As observações consistem em leituras da aceleração *versus* tempo. Este conjunto de dados foi originalmente descrito em Schmidt *et al.* (1981) e desde então tem sido amplamente analisado na literatura relacionada aos modelos semiparamétricos.

Estes dados não são igualmente espaçados no tempo e em alguns casos existem múltiplas observações para alguns tempos. Outra característica é que as observações são sujeitas a erros. Nosso interesse aqui é estimar a curva de aceleração.

**Figura 7 - Curva de aceleração *versus* tempo**




O gráfico da Figura 7 mostra os pontos que correspondem às observações, a linha contínua que corresponde ao modelo obtido via *kernel* e a linha pontilhada correspondente ao modelo obtido via *splines*.

No modelo ajustado via *kernel*, foi utilizado o *kernel* normal com parâmetro de suavização h=3.0.

No ajuste do modelo via *splines* a *GVC* obtida foi 568,3. O parâmetro de suavização obtido foi $\lambda=0.000068$, o traço da matriz $S_\lambda$, graus de liberdade associados ao ajuste do modelo, foi $Tr(S_\lambda) = 13,6$.

# 4. Discussão

Neste trabalho, apresentamos uma breve revisão da teoria de modelos semiparamétricos enfocando a importância da utilização de métodos automáticos para a seleção do parâmetro de suavização na estimação destes modelos.

Quando utilizando *kernel*, o parâmetro de suavização tem a função de estabelecer uma vizinhança para a composição de uma média ponderada. Já em *smoothing splines*, o parâmetro de suavização, representado por $\lambda$, tem a finalidade de ajustar os pesos entre a soma de quadrados dos resíduos e a penalidade por falta de suavidade do ajuste e também fixa a quantidade de junções a serem adotadas. No caso de regressão por *splines*, o controle da suavização aplicada é feito através da malha de junções $\Lambda$, onde o número de junções a ser utilizado e as posições das mesmas determinam a suavização.

Apesar de resultarem em processos diferentes, alguns autores já mostraram as relações existentes entre *smoothing splines* e *kernel*, por exemplo Messer (1991) e Silverman (1984). Neste sentido, a escolha do método de suavização não deve produzir resultados muito diferentes, desde que o parâmetro de suavização seja selecionado corretamente.

## Referências bibliográficas


**BHATTACHARYA, P.K. E ZHAO, P.L. (1997). Semiparametric Inference in a Partial Linear Model.** ***The Annals of Statistics*****, 25, 244-262.**

**BUJA, A., HASTIE, T. E TIBSHIRANI, R. (1989). Linear Smoothers and Additive Models (with discussion).** ***The Annals of Statistics*****, 17, 453-555.**

**CHEN, H. (1988). Convergence Rates for Parametric Components in a Partly Linear Model.** ***The Annals of Statistics*****, 16, 136-146.**

**CRAVEN, P. E WAHBA, G. (1979). Smoothing Noisy Data with Spline Functions. Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation.** ***Numerical Mathematics*****, 31, 377-403.**

**CUZICK, J. (1992a). Semiparametric Additive Regression.** ***Journal of the Royal Statistical Society*** **B, 54, 831-843.**

**CUZICK, J. (1992b). Efficient Estimates in Semiparametric Additive Regression Models with Unknown Error Distribution.** ***The Annals of Statistics*****, 20, 1129-1136.**

**DE BOOR, C. (1978).** ***A Practical Guide to Splines*****. Springer-Verlag. New York.**

**ENGLE, R.F., GRANGER, C.W.J., RICE J. E WEISS, A. (1986). Semiparametric Estimates of the Relation Between Weather and Electricity Sales.** ***Journal of the American Statistical Association*****, 81, 310-320.**

**EUBANK, R.L. (1994). A Simple Smoothing Spline.** ***The American Statistician*****, 2, 103-106.**

**FRIEDMAN, J.H. E SILVERMAN, B.W. (1989). Flexible Parsimonius Smoothing and Additive Modeling (with discussion).** ***Technometrics*****, 31, 3-21.**

GOLUB, G.H., HEATH, M. E WAHBA, G. (1979). Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21, 215-223.

GREEN, P.J., JENNISON, C. E SEHEULT, A. (1985). Analysis of Field Experiments by Least Squares Smoothing. *Journal of the Royal Statistical Society* B, 47, 299-315.

HANDSCOMB, D.C. (1966). Spline Functions. In *Methods of Numerical Approximation*. Pergamon Press. Oxford.

HÄRDLE, W. (1990). *Applied Nonparametric Regression* (Econometric Society Monographs). Cambridge University Press. New York.

HASTIE, T.J. E TIBSHIRANI, R.J. (1990). *Generalized Additive Models* (Monographs on Statistics and Applied Probability). Chapman and Hall. London.

HECKMAN, N.E. (1986). Spline Smoothing in a Partly Linear Model. *Journal of the Royal Statistical Society* B, 48, 244-248.

LEE, D.K.C. (1990). Cross-Validation in Semiparametric Models: Some Monte Carlo Results. *Journal of Statistical Computation and Simulation*, 37, 171-187.

MESSER, K. (1991). A Comparison of a Spline Estimate to its Equivalent Kernel Estimate. *The Annals of Statistics*, 19, 817-829.

RICE, J. (1984). Bandwidth Choice for Nonparametric Regression. *The Annals of Statistics*, 12, 1215-1230.

RICE, J. (1986). Convergence Rates for Partially Splined Models. *Statistics & Probability Letters*, 4, 203-208.

SCHICK, A. (1986). On Asymptotically Efficient Estimation in Semiparametric Models. *The Annals of Statistics*, 14, 1139-1151.

SCHICK, A. (1993). On Efficient Estimation in Regression Models. *The Annals of Statistics*, 21, 1486-1521.

SCHICK, A. (1996). Root-n-consistent and Efficient Estimation in Semiparametric Additive Regression Models. *Statistics & Probability Letters*, 30, 45-51.

SCHMIDT, G., MATTERN, R. AND SCHUCLER, F. (1981) Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without protective helmet under the effects of impact. EEC Research Program on Biomechanics of Impacts, Final Report, Phase III, Project G5, Institut fur Rechtsmedizin, University of Heidelberg, West Germany.

SCHUMAKER, L.L. (1981). *Spline Functions: Basic Theory*. John Wiley & Sons. New York.

SILVERMAN, B.W. (1984). Spline Smoothing: The Equivalent Variable Kernel Method. *The Annals of Statistics*, 12, 898-916.

SILVERMAN, B.W. (1985). Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting. *Journal of the Royal Statistical Society* B, 47, 1-52.

SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis* (Monographs on Statistics and Applied Probability). Chapman and Hall. London.

SPECKMAN, P. (1985). Spline Smoothing and Optimal Rates of Convergence in Nonparametric Regression Models. *The Annals of Statistics*, 13, 970-983.

SPECKMAN, P. (1988). Kernel Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society* B, 50, 413-436.

STONE, C.J. (1985). Additive Regression and Other Nonparametric Models. *The Annals of Statistics*, 13, 689-705.

WHITTAKER, E. (1923). On a New Method of Graduation. *Proceedings of the Edinburgh Mathematical Society*, 41, 63-75.

## Abstract

**Semiparametric models are useful tools for modelling phenomena where parametric and nonparametric effects are present. It can be used in a wide variety of situations, since it combines the flexibility of nonparametric modelling with the easy interpretation of parametric modelling. It consists of two components: One is a parametric component, which assumes a fixed and prestablished form of relationship between response and explanatory variables. The other is a non parametric component which provides the flexibility for the modelling. Here we do a review of Semiparametric Models, showing the different ways of estimating its components. We also discuss the methods used for choosing the smoothing parameters, with emphasis to the generalized cross validation method.**

# Análise bayesiana do modelo "open loop threshold autoregressive"

Thelma Sáfadi *

Pedro Alberto Morettin**

**Resumo**

**A análise estatística de séries temporais é facilitada se a média e as covariâncias não mudam com o tempo. Entretanto, esta não é uma hipótese realista em várias áreas de aplicações. Variações sazonais e relações não-lineares entre variáveis levam à utilização de modelos não-lineares. Neste trabalho, estamos interessados na análise Bayesiana do modelo "open loop threshold autoregressive" (abreviadamente, TARSO) e, para isto, consideramos uma *a priori* própria e *a priori* de Jeffreys. A análise é feita por meio do amostrador de Gibbs e, considerando o método de cadeias múltiplas, verificamos a convergência através do fator $\hat{R}$ proposto por Gelman e Rubin (1992). Para exemplificar a metodologia foram consideradas as séries de números de óbitos causados por problemas cardíacos e a série de temperatura mínima diária na cidade de São Paulo. Observou-se que, para temperaturas entre 12,90 e 15,23 graus, há um acréscimo de 300 óbitos causados por problemas cardíacos.**

**Palavras-chaves: Amostrador de Gibbs, Análise Bayesiana, Modelo "Threshold".**

# 1. Introdução

Nos últimos anos muito se tem estudado sobre os modelos não-lineares, principalmente, com o progresso da área computacional, facilitando a análise desses modelos. Os modelos lineares usuais falham em detectar certas propriedades observadas em séries temporais tais como: ciclos limites e saltos *jump phenomena*.

* Endereço para correspondência: Departamento de Ciências Exatas da Universidade Federal de Lavras, Caixa Postal 37, Lavras - Minas Gerais - CEP 37.200-000 - e-mail: safadi@ufla.br

** Instituto de Matemática e Estatística - USP, Caixa Postal 66281 - São Paulo - SP - CEP 05315-970 - e-mail: pam@ime.usp.br

Vários modelos não-lineares têm sido propostos para séries com tais propriedades. Entre os modelos de maior sucesso podemos citar os modelos "patamar", os modelos bilineares e os modelos auto-regressivos com coeficientes aleatórios.

Várias são as áreas de aplicações, entre elas podemos citar ecologia, população, física solar, entomologia, dinâmica de população, economia, finanças, saúde, poluição, geofísica e hidrologia conforme exemplos apresentados por Tong (1990, p.358).

A análise estatística de séries temporais é facilitada se a média e as covariancias não mudam com o tempo. Entretanto, esta não é uma hipótese realista em hidrologia, por exemplo. Variações sazonais e relações não-lineares entre variáveis metereológicas e o rio implicam que a vazão não é nem Gaussiana e nem estacionária. Para alguns rios, a precipitação associada pode alternar entre chuva e neve. Além disso, pode-se ter geleiras em áreas de drenagem. Em tais casos, a variável temperatura tem um papel importante, o "patamar" próximo ao ponto de congelamento tem um significado hidrológico-metereológico prontamente identificado aqui e parece razoável esperar que explique alguma não-linearidade.

A idéia de modelar uma série temporal discreta usando modelos lineares por partes foi enunciada, primeiramente, por Tong (1970). Uma apresentação compreensível, juntamente com inúmeras aplicações e discussões, foi considerada por Tong e Lim (1980). O uso de variáveis e valores "patamar" nos permite representar uma série temporal por meio de modelos lineares por partes.

A análise Bayesiana de processos "threshold autoregressive" (TAR) tem várias características que são análogas aos problemas de pontos de mudança em estatística. Pole e Smith(1985) observaram esta similaridade e desenvolveram uma análise Bayesiana não-assintótica para modelos "patamar" com regimes determinados por variáveis externas ou exógenas, utilizando métodos numéricos.

Neste artigo, estamos interessados na análise Bayesiana do modelo "open loop threshold autoregressive" (abreviadamente, TARSO).

Para a análise vamos considerar uma *a priori* própria e *a priori* de Jeffreys e a inferência será feita através do amostrador de Gibbs. Para a verificação da convergência vamos utilizar o método de cadeias múltiplas proposto por Gelman e Rubin (1992).

A extensão para o modelo "open loop threshold ARMA" (TARMASO) é imediata e será considerada, brevemente, na seção 5. Observemos, ainda, que se $Z_{t-d} = Y_{t-d}$ então teremos o modelo TARMA já analisado por Sáfadi e Morettin (2000).

Este trabalho está organizado da seguinte forma. O modelo TARSO é introduzido na seção 2 e a análise *a posteriori* com *a priori* própria e *a priori* de Jeffreys é dada na seção 3. Na seção 4 é feita uma aplicação da análise considerando as séries de número de óbitos diários por problemas cardíacos e temperatura mínima diária na cidade de São Paulo. Conclusões e considerações finais são apresentadas na seção 5.

# 2. Modelo TARSO

Suponhamos que $\{Y_t, Z_t, t = 0,\pm1,\pm2,\ldots\}$ seja um processo estocástico satisfazendo

$$Y_t = \phi_{l^\circ} + \sum_{i=1}^{pl} \phi_{li} Y_{t-i} + a_t^{(l)} + \sum_{v=0}^{m_l} \beta_{lv} Z_{t-v}, \quad se \quad Z_{t-d} \in R_l,$$

para $l = 1, \ldots, k$, e as suposições:

i) $\{a_t^{(l)}\}$ é uma seqüência de variáveis aleatórias i.i.d. $\sim N(0, \tau_l^{-1})$, onde $\tau_l > 0, \tau_l^{-1} = \sigma_l^2 = Var(a_t^{(l)}$ é desconhecida;

ii) $R_l$ é um intervalo da reta $IR$ tal que $R_i \cap R_j = 0$, se $i \neq j$ e $\cup_{l=1}^{k} R_l = IR$, isto é, $\{R_1, \ldots, R_l\}$ é uma partição da reta com k-1 variáveis "patamar";

iii) $Z_t$ é uma variável exógena, podendo ser, por exemplo, uma função indicadora para modelagem de séries econômicas, $Z_{t-d}$ é uma variável "patamar", independente de $a_t^{(l)}$, para todo $l$ e todo $t$, $d$ é o parâmetro de retardo. $Z_{t-d}$ determina os regimes do modelo; e

iv) As $k$ seqüências são supostas independentes entre si.

Tal sistema $(Y_t, Z_t)$ é chamado TARSO, com $Y_t$ sendo dados observáveis de saída e $Z_t$ dados de entrada.

O sistema $\{Y_t, Z_t\}$ é denominado "closed loop threshold autoregressive" (abreviadamente, TARSC), se $(Y_t, Z_t)$ e $(Z_t, Y_t)$ são ambos TARSO.

Para simplificar a análise vamos considerar o modelo com dois regimes (k=2). A extensão para um maior número de regimes não é difícil. Portanto, seja o modelo $TARSO(2; p_1, p_2; m_1, m_2)$ onde as séries $\{Y_t, Z_t, t = 1, \ldots, n\}$ são conhecidas. O modelo é dado por

$$Y_t = \begin{cases} \phi_{10} + \sum_{i=1}^{p_1} \phi_{1i} Y_{t-i} + a_t^{(1)} + \sum_{v=0}^{m_1} \beta_{1v} Z_{t-v}, & se\ Z_{t-d} \leq r; \\ \phi_{20} + \sum_{i=1}^{p_2} \phi_{2i} Y_{t-i} + a_t^{(2)} + \sum_{k=0}^{m_2} \beta_{2v} Z_{t-k}, & se\ Z_{t-d} > r. \end{cases} \tag{2.1}$$

Vamos supor que as ordens $p_1, p_2, m_1, m_2$ são conhecidas. Os parâmetros para o modelo são $\gamma_i = (\phi_{i0}, \phi_{i1}, \ldots, \phi_{ip_i}, \beta_{i0}, \beta_{i1}, \ldots, \beta_{im_i})', \tau_i, i = 1,2, r$ e $d$.

Fazendo $X_{it} = (Y_{t-1}, \ldots, Y_{t-p_i}, Z_t, Z_{t-1}, \ldots, Z_{t-m_i})', i = 1,2$ podemos reescrever o modelo (2.1) como

$$Y_t = \begin{cases} \gamma_1' X_{1t} + a_t^{(1)}, & se\ Z_{t-d} \leq r; \\ \gamma_2' X_{2t} + a_t^{(2)}, & se\ Z_{t-d} \leq r. \end{cases} \tag{2.2}$$

A função de verossimilhança aproximada condicionada às p-primeiras observações é dada por

$$L(\gamma_1, \gamma_2, \tau, \tau_2, r, d | D) \ \alpha \ \tau_1^{\frac{n_2}{1}} \tau_2^{\frac{n_2}{2}} \exp\{-\frac{\tau_1}{2} \sum_1 (Y_t - \gamma_1' X_{1t})^2 - \frac{\tau_2}{2} \sum_2 (Y_t - \gamma_2' X_{2t})^2\} \tag{2.3}$$

onde

$D=\{Y_t, Z_t, t=p+1,\ldots,n\}$ é o conjunto de todas as observações, $n_1, n_2$ são os números de observações em cada regime, $\Sigma_1$ é a somatória em $t$ para $\{t=p+1,\ldots,n;\ Z_{t-d}\le r\}$, $\Sigma_2$ é a soma em $t$ para $\{t=p+1,\ldots,n; Z_{t-d}>r\}$ sendo que $p=\max\{p_1, p_2, m_1 m_2\}$.

Observemos que

$$\sum_1 (Y_t-\gamma_i' X_{it})^2 = \sum_i (Y_t-\gamma_i' X_{it})(Y_t-\gamma_i' X_{it})'$$

$$= \sum_i Y_t^2 - \sum_i Y_t X_{it}'\gamma_i - \gamma_i' \sum_i X_{it} Y_t + \gamma_i' \sum_i X_{it} X_{it}' \gamma_i, \quad i=1,2.$$

Denotando por

$$\sum_i X_{it} Y_t = \mathbf{B}_i = \begin{pmatrix} \mathbf{B}_{1i} \\ \mathbf{B}_{2i} \end{pmatrix}, \quad \sum_i X_{it} X_{it}' = \mathbf{A}_i = \begin{pmatrix} \mathbf{A}_{11}^i & \mathbf{A}_{12}^i \\ \mathbf{A}_{21}^i & \mathbf{A}_{22}^i \end{pmatrix}, i=1,2.$$

$$\mathbf{B}_{1i} = \begin{pmatrix} \Sigma_i Y_t \\ \Sigma_i Y_t Y_{t-1} \\ \cdot \\ \cdot \\ \cdot \\ \Sigma_i Y_t Y_{t-p_i} \end{pmatrix}, \quad \mathbf{B}_{1i} = \begin{pmatrix} \Sigma_i Y_t Z_t \\ \Sigma_i Y_t Z_{t-1} \\ \Sigma_i Y_t Z_{t-2} \\ \cdot \\ \Sigma_i Y_t Y_{t-m_i} \end{pmatrix},$$

$$\mathbf{A}_{11}^i = \begin{pmatrix} \Sigma_i 1 & \Sigma_i Y_{t-1} & \cdots & \Sigma_i Y_{t-p_i} \\ \Sigma_i Y_{t-1} & \Sigma_i Y_{t-1} Y_{t-1} & \cdots & \Sigma_i Y_{t-1} Y_{t-p_i} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \Sigma_i Y_{t-p_i} & \Sigma_i Y_{t-1} Y_{t-p_i} & \cdots & \Sigma_i Y_{t-p_i} Y_{t-p_i} \end{pmatrix},$$

$$\mathbf{A}_{12}^i = \begin{pmatrix} \Sigma_i Z_t & \Sigma_i Z_{t-1} & \Sigma_i Z_{t-2} & \cdots & \Sigma_i Z_{t-m_i} \\ \Sigma_i Y_{t-1} Z_t & \Sigma_i Y_{t-1} Z_{t-1} & \Sigma_i Y_{t-1} Z_{t-2} & \cdots & \Sigma_i Y_{t-1} Z_{t-m_i} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \Sigma_i Y_{t-p_i} Z_t & \Sigma_i Y_{t-p_i} Z_{t-p_i} & \cdots & \Sigma_i Y_{t-p_i} Y_{t-p_i} & \end{pmatrix},$$

e $\mathbf{A}_{22}^{i}$ é a matriz cujo jk-ésimo elemento é $\sum_i Z_{t-j}Z_{t-k}$, a função de verossimilhança (2.3) pode ser reescrita como

$$L(\gamma_1,\gamma_2,\tau_1,\tau_2,r,d|D) \;\; \alpha \;\; \tau_1^{\frac{n_2}{2}}\tau_2^{\frac{n_2}{2}} \exp\{\sum_{i=1}^{2} -\frac{\tau_i}{2}[\gamma_i'Y_i - \mathbf{B}_i'\mathbf{A}_i^{-1}\mathbf{B}_i$$

$$+(\gamma_i - \mathbf{A}_i^{-1}\mathbf{B}_i)'\mathbf{A}_i(\gamma_i - \mathbf{A}_i^{-1}\mathbf{B}_i)]\} \qquad (2.4)$$

onde $Y_i$ é o vetor das observações do i-ésimo regime.

# 3. Análise *a posteriori*

Neste trabalho, vamos supor que $r$ seja conhecido, caso contrário uma *priori* uniforme pode ser considerada.

Assim, a função de verossimilhança (2.4) é escrita como $L(\gamma_1,\gamma_2,\tau_1,\tau_2,d|r,D)$.

Vamos considerar uma *a priori* própria independente para os parâmetros $\gamma_1,\gamma_2,\tau_1,\tau_2,d$ com:

1) $\gamma_1$ e $\gamma_2$ independentes $\sim$ Normal $(\gamma_{0i},\mathbf{Q}_i^{-1})$ , $i=1,2$;

2) $\tau_1$ e $\tau_2$ independentes $\sim$ Gama $(\alpha_i/2,\beta_i/2)$ , $i=1,2$; e

3) $d \sim$ Uniforme $\{1,2,\ldots,d_0\}$.

Os hiperparâmetros $\gamma_{0i},\mathbf{Q}_i,d_0,\alpha_i$ e $\beta_i$, $i=1,2$, são supostos conhecidos, com $\alpha_i>0,\beta_i>0,\lambda_{0i}\in R^{p_i+m_i+1}$ e $\mathbf{Q}_i$ é uma matriz definida positiva $(p_i+m_i+1)\times(p_i+m_i+1)$.

*A priori* pode ser escrita como

$$P(\gamma_1,\gamma_2,\tau_1,\tau_2,d) \;\alpha\; \tau_1^{\alpha_1/2-1}\tau_2^{\alpha_2/2-1}\exp\{-\frac{1}{2}\sum_{i=1}^{2}[\tau_i\beta_i+(\gamma_i\beta_i+(\gamma_i-\gamma_{0i})'\mathbf{Q}_i(\gamma_i-\gamma_{0i})]\} \qquad (3.1)$$

**Teorema 1:** A distribuição *a posteriori* obtida da verossimilhança (2.4) e da distribuição *a priori* (3.1) será uma distribuição normal-gama dada por

$$P(\gamma_1,\gamma_2,\tau_1,\tau_2,d|r,D)\alpha\;\tau_1^{\frac{n_1+\alpha_1}{2}-1}\tau_2^{\frac{n_2+\alpha_2}{2}-1}\exp\{-\frac{1}{2}\sum_{i=1}^{2}R_i - S_i(\gamma_{0i})$$

$$+(\gamma_i-\mu_i)'(\tau_i\mathbf{A}_i-\mathbf{Q}_i)(\gamma_i-\mu_i)\}, \qquad (3.2)$$

onde

$$\mathbf{R}_i = \tau_i(\beta_i+Y_i'Y_i)+\gamma_{0i}'\mathbf{Q}_i\gamma_{0i},$$

$$S_i(\gamma_{0i}) = (\tau_i \mathbf{B}_i + \mathbf{Q}_i\gamma_{0i})'(\tau_i \mathbf{A}_i + \mathbf{Q}_i)^{-1}(\tau_i \mathbf{B}_i + \mathbf{Q}_i\gamma_{0i})$$

e

$$\mu_i = (\tau_i \mathbf{A}_i + \mathbf{Q}_i)^{-1}(\tau_i \mathbf{B}_i + \mathbf{Q}_i\gamma_{0i}).$$

A inferência para os parâmentros será feita por meio do Amostrador de Gibbs - A.G.- , que é uma técnica para gerar amostras aleatórias de uma distribuição (marginal) sem que se conheça a sua densidade. Embora o A.G. tenha sido muito aplicado em modelos bayesianos, ele também é extremamente útil para se amostrar a função de verossimilhança nos métodos freqüentistas.

Em modelos complicados, raramente se consegue obter amostras diretamente das distribuições *a posteriori*. A idéia do método de Monte Carlo via Cadeias de Markov (MCMC) é simular um passeio aleatório no espaço do parâmetro $\theta$, o qual converge para uma distribuição estacionária, que é a distribuição *a posteriori* $P(\theta|Y)$, onde Y é o vetor de observações.

O A.G. é um particular algoritmo MCMC e tem sido extremamente útil na resolução de problemas multidimensionais e é definido em termos de subvetores de $\theta$. Pode-se pensar no A.G. como uma implementação prática do fato que o conhecimento das distribuições condicionais é suficiente para determinar a distribuição conjunta (se ela existir).

Suponhamos que o vetor de parâmetros $\theta$ seja dividido em $k$ subvetores, $(\theta_1, \theta_2, \ldots, \theta_k)'$ e que as distribuições condicionais de cada parâmetro $\theta$ dado todos os outros, sejam conhecidas. Essas distribuições são denotadas por

$f_1(\theta_1|\theta_2, \ldots, \theta_k, Y)$, $f_2(\theta_2|\theta_1, \theta_3, \ldots, \theta_k, Y), \ldots, f_k(\theta_k|\theta_1, \ldots, \theta_{k-1}, Y)$ e denominadas distribuições a posteriori condicionais completas.

Dados os valores iniciais $\theta^0 = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)})'$ retiramos

$$\theta_1^{(1)} \text{ de } f_1(\theta_1|\theta_2^{(0)}\theta_3^{(0)}, \ldots, \theta_k^{(0)}, Y),$$

$$\theta_2^{(1)} \text{ de } f_2(\theta_2|\theta_1^{(1)}\theta_3^{(0)}, \ldots, \theta_k^{(0)}, Y),$$

$$\vdots$$

$$\theta_k^{(1)} \text{ de } f_k(\theta_k|\theta_1^{(1)}\theta_2^{(1)}, \ldots, \theta_{k-1}^{(1)}, Y),$$

obtendo-se na primeira iteração $\theta^1 = (\theta_1^{(1)},\ldots,\theta_k^{(1)})'$. Após um número grande de iterações, digamos M, obtém-se $\theta^M = (\theta_1^{(M)},\ldots,\theta_k^{(M)})'$. As distribuições marginais a posteriori desejadas podem ser aproximadas pelas distribuições marginais dos $N$ valores $(\theta_1^{(i)},\ldots,\theta_k^{(i)})_{i=M+1}^{M+N}$.

Um problema encontrado é a escolha apropriada dos valores de $M$ e $N$. Uma solução geral para a escolha de $N$ é monitorarmos a convergência da seqüência de Gibbs sob algum aspecto. Cowles e Carlin(1996) apresentam um estudo comparativo entre vários métodos para monitoração da convergência, alguns com uma única cadeia, outros com cadeias múltiplas e a monitoração é feita com métodos gráficos, quantitativos ou qualitativos.

Maiores detalhes sobre o Amostrador de Gibbs poderão ser encontrados em Casella e George(1992). Para a verificação da convergência vamos considerar o método de cadeias múltiplas apresentado por Gelman e Rubin(1992).

**Corolário 1**. As distribuições condicionais completas para $\gamma_i$ e $\tau_i$, $i = 1,2$ obtidas da distribuição *a posteriori* são dadas por

$$P(\gamma_1|\tau_1,\tau_2,d,r,D) \quad \alpha \quad \exp\{-\frac{1}{2}(\gamma_i-\mu_i)'(\tau_i\mathbf{A}_i+\mathbf{Q}_i)(\gamma_i-\mu_i)\},$$

isto é,

$$(\gamma_1|\tau_1,\tau_2,d,r,D) \quad \sim \quad Normal(\mu_i,[(\tau_i\mathbf{A}_i+\mathbf{Q}_i)]^{-1}), i=1,2,$$

$$P(\tau_1|\gamma_1,\gamma_2,d,r,D) \quad \alpha \quad \tau_i^{(n_i+\alpha_i)/2}\exp\{-\frac{\tau_i}{2}[\beta_i+Y_i'Y_i-\mathbf{B}_i'\mathbf{A}_i^{-1}\mathbf{B}_i$$

$$+(\gamma_i-\mathbf{A}_i^{-1}\mathbf{B}_i)'\mathbf{A}_i(\gamma_i-\mathbf{A}_i^{-1}\mathbf{B}_i)]\},$$

isto é,

$$\tau_1|\gamma_1,\gamma_2,d,r,D) \quad \sim \quad Gama\,(\frac{n_i+\alpha_i}{2},\frac{1}{2}(\beta_i+Y_i'Y_i-\mathbf{B}_i'\mathbf{A}_i^{-1}\mathbf{B}_i+(\gamma_i-\mathbf{A}_i^{-1}\mathbf{B}_i)'\mathbf{A}_i(\gamma_i-\mathbf{A}_i^{-1}\mathbf{B}_i))$$

$i = 1,2$.

**Corolário 2.** A distribuição condicional completa de $d$ obtida de (3.2) é uma multinomial com probabilidade

$$P(d|\gamma_1,\gamma_2,\tau_1,\tau_2,r,D)=\frac{L(\gamma_1,\gamma_2,\tau_1,\tau_2,d|r,D)}{\sum_{d=1}^{d_0}L(\gamma_1,\gamma_2,\tau_1,\tau_2,d|r,D)},$$

onde $L(\gamma_1,\gamma_2,\tau_1,\tau_2,d|r,D)$ é a função de verossimilhança dada em (2.4).

No caso de termos pouca ou nenhuma informação *a priori*, poderíamos usar *a priori* de Jeffreys,

$$P(\gamma_1,\gamma_2,\tau_1,\tau_2)\alpha\,\tau_1^{-1}\tau_2^{-1}.$$

Se *a priori* de Jeffreys é combinada com a função de verossimilhança (2.4), nós ainda teremos uma posteriori normal-gama. Os resultados obtidos para *as posterioris* condicionais completas são modificados fazendo-se $\beta_i \to 0, \mathbf{Q}_i \to 0, \alpha_i \to -(p_i+m_i+1), i=1,2$ na distribuição *a posteriori* (3.2).

**Corolário 3.** Se *a priori* de Jeffreys é combinada com a função de verossimilhança (2.4), temos as distribuições condicionais completas dadas por

$$\gamma_1|\tau_1,\tau_2,d,r,D) \sim Normal(\mathbf{A}_i^{-1}\mathbf{B}_i(\tau_i\mathbf{A}_i)^{-1}), i=1,2$$

$$\tau_1|\gamma_1,\gamma_2,d,r,D) \sim Gama\,(\frac{n_i-(p_i+m_i+1)}{2},\frac{1}{2}(Y_i'Y_i-\mathbf{B}_i'\mathbf{A}_i^{-1}\mathbf{B}_i$$
$$+(\gamma_i-\mathbf{A}_i^{-1}\mathbf{B}_i)'\mathbf{A}_i(\gamma_i-\mathbf{A}_i^{-1}\mathbf{B}_i)),$$

e a distribuição de $d$ é como no **Corolário 2**.

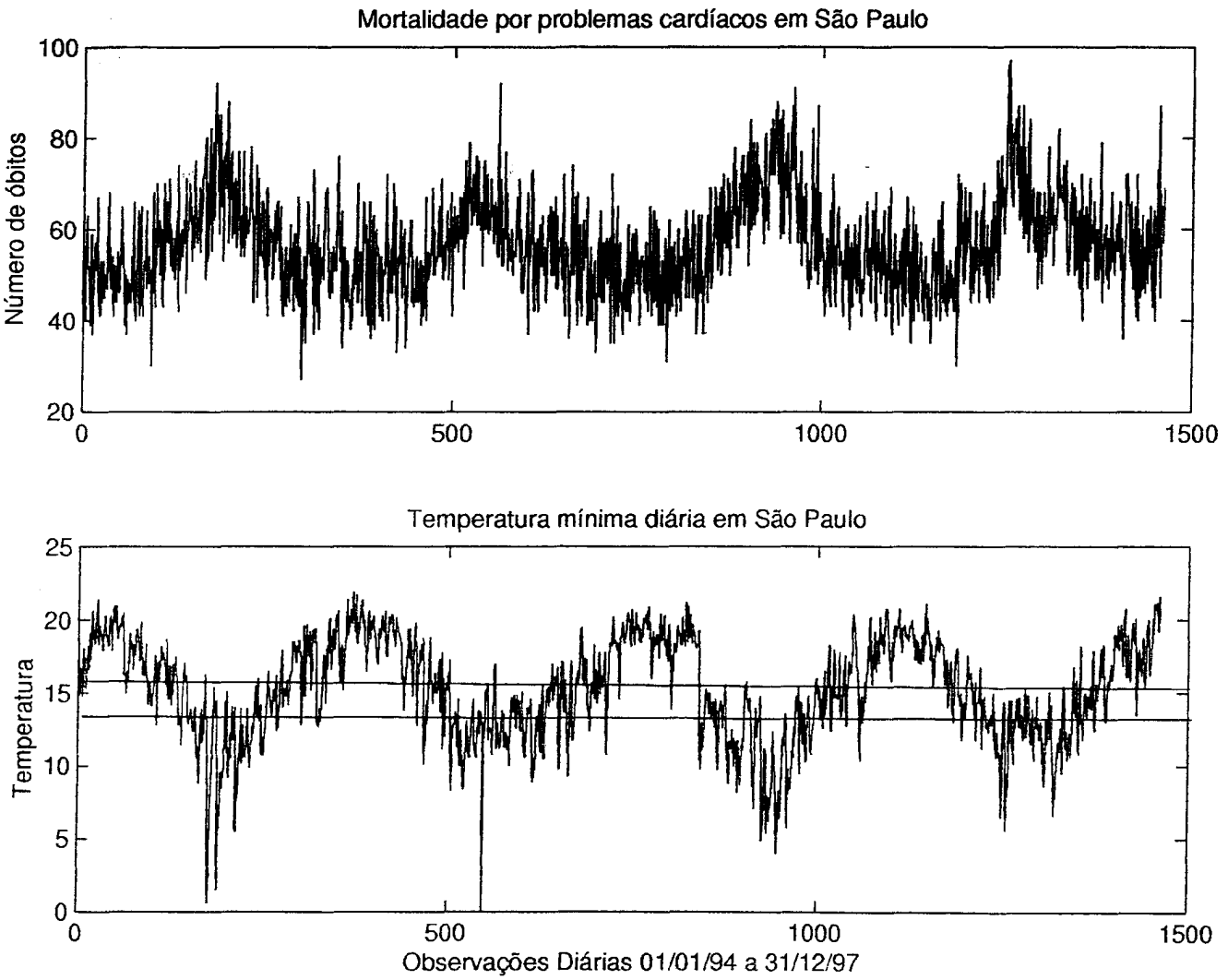
# 4. Aplicação

Para esta aplicação vamos considerar a série do número de óbitos por problemas cardíacos e a série de temperaturas mínimas diárias na cidade de São Paulo. As duas séries são de dados diários no período de 01/01/1994 a 31/12/1997.

Nesta linha, Shumway, Azari e Pawitan (1998) estudam a mortalidade, em Los Angeles, como função da poluição e temperatura.

Na Figura 1 apresentamos o gráfico das duas séries.

**Figura 1 - Mortalidade e temperatura mínima na cidade de São Paulo**



Neste gráfico, podemos observar que as séries são periódicas com período de 365 dias e que para baixos valores de temperatura temos um maior número de mortes causadas por problemas cardíacos. Para a análise, vamos considerar as séries sazonalmente ajustadas.

Nosso objetivo, neste exemplo, é apenas exemplificar o método sem, contudo, tentarmos explicar o fenômeno do ponto de vista clínico. A série de entrada $Z_t$ é a série de temperatura mínima diária e a série de saída $Y_t$ é a série de números de mortes por problemas cardíacos.

Vamos considerar os valores "thresholds" r = temperatura média mínima (15,23) e r=$1^{Q}$ quartil (12,90), representados na Figura 1. A priori de Jeffreys foi utilizada e as distribuições condicionais completas para o uso do amostrador de Gibbs são dadas pelo **Corolário 2.**

Para a verificação da convergência consideramos o método de cadeias múltiplas apresentado por Gelman e Rubin(1992). Aqui, devemos ter $\sqrt{\hat{R}} \approx 1$.

Para cada parâmetro, tomamos três cadeias em paralelo com valores iniciais distintos. Fizemos 6 000 iterações para cada cadeia, desprezamos as 1500 primeiras, e das 4500 restantes, tomamos uma observação a cada 25, de modo a obtermos uma amostra aproximadamente i.i.d.

Durante a análise, para cada valor de r, observamos que para os retardos $d$=1 e $d$=2 não obtivemos diferença entre os números de observações em cada regime. Sendo assim, vamos apresentar o resultado obtido em ambos os casos, ou seja, não estamos considerando, nesta análise, *a posteriori* de d e sim valores pré-fixados iguais a um e dois.

A Tabela 1 apresenta o modelo TARSO(2;2,3;2,1) para r = 15,23 e d =1 (Tabela 1a) e d=2 (Tabela 1b), o primeiro regime com 576 óbitos, e o segundo, com 517. Na Tabela 2, temos o modelo TARSO(2;2,1;1,1) com r=12,90, d=1 (Tabela 2a) e d=2 (Tabela 2b) tendo o primeiro regime 276 óbitos, e o segundo, 817. Podemos observar que, para temperatura entre 12,90 e 15,23 graus, há um acréscimo de 300 óbitos causados por problemas cardíacos.

**Tabela 1 - Modelo TARSO(2;2,3;2,1); r=15,23(temperatura média mínima), regimes com 576 e 517 óbitos, respectivamente**
**Tabela 1a: d=1**

| Parâmetro | Estimativa | d.p. | $\sqrt{\hat{R}}$ |
|---|---|---|---|
| $\phi_{11}$ | 0,1075 | 0,0419 | 1,0043 |
| $\phi_{12}$ | 0,1206 | 0,0408 | 0,9993 |
| $\beta_{12}$ | -0,9866 | 0,1757 | 0,9988 |
| $\phi_{21}$ | 0,1550 | 0,0404 | 0,9976 |
| $\phi_{23}$ | 0,0989 | 0,0416 | 1,0035 |
| $\beta_{21}$ | -1,2613 | 0,1994 | 1,0004 |
| $\tau_1$ | 0,0072 | 0,0003 | 1,0041 |
| $\tau_2$ | 0,0071 | 0,0003 | 1,0006 |

**Tabela 1b: d=2**

| Parâmetro | Estimativa | d.p. | $\sqrt{\hat{\mathbf{R}}}$ |
|---|---|---|---|
| $\phi_{11}$ | 0,1437 | 0,0424 | 1,0002 |
| $\phi_{12}$ | 0,1409 | 0,0439 | 1,0044 |
| $\beta_{12}$ | -1,0129 | 0,1800 | 1,0013 |
| $\phi_{21}$ | 0,1212 | 0,0419 | 0,9988 |
| $\phi_{23}$ | 0,0824 | 0,0424 | 0,9979 |
| $\beta_{21}$ | -1,2435 | 0,1874 | 1,0084 |
| $\tau_1$ | 0,0071 | 0,0003 | 1,0055 |
| $\tau_2$ | 0,0072 | 0,0003 | 0,9973 |

**Tabela 2 - Modelo TARSO(2;2,1;1,1); r=12,90 ($1^{\underline{o}}$ quartil), regimes com 276 e 817 óbitos, respectivamente**
**Tabela 2a: d=1**

| Parâmetro | Estimativa | d.p. | $\sqrt{\hat{\mathbf{R}}}$ |
|---|---|---|---|
| $\phi_{11}$ | 0,1339 | 0,0589 | 1,0057 |
| $\phi_{12}$ | 0,1702 | 0,0551 | 1,0044 |
| $\beta_{11}$ | -0,9542 | 0,1910 | 1,0022 |
| $\phi_{21}$ | 0,1651 | 0,0340 | 0,9998 |
| $\beta_{21}$ | -0,5143 | 0,1443 | 1,0071 |
| $\tau_1$ | 0,0066 | 0,0004 | 0,9978 |
| $\tau_2$ | 0,0070 | 0,0002 | 1,0009 |

**Tabela 2b: d=2**

| Parâmetro | Estimativa | d.p. | $\sqrt{\hat{\mathbf{R}}}$ |
|---|---|---|---|
| $\phi_{11}$ | 0,0987 | 0,0613 | 1,0078 |
| $\phi_{12}$ | 0,1671 | 0,0589 | 0,9980 |
| $\beta_{11}$ | -1,0408 | 0,2282 | 1,0007 |
| $\phi_{21}$ | 0,1745 | 0,0337 | 0,9996 |
| $\beta_{21}$ | -0,6772 | 0,1603 | 0,9978 |
| $\tau_1$ | 0,0067 | 0,0004 | 1,0061 |
| $\tau_2$ | 0,0070 | 0,0002 | 1,0024 |

# 5. Conclusões e considerações finais

Apresentamos uma abordagem Bayesiana para estimar os parâmentros de um modelo $TARSO(k; p_1, p_2; m_1, m_2)$. Foram desenvolvidas distribuições *a posteriori* condicionais completas para os coeficientes $\phi, \beta$, para as precisões dos erros de cada regime $\tau$ e para o parâmetro de retardo $d$. Para a aplicação consideramos as séries do número de óbitos causados por problemas cardíacos e temperatura mínima diária na cidade de São Paulo no período de 01/01/1994 a 31/12/1997 sendo os regimes determinados pela série de temperatura. Observamos que para retardos de 1 ou 2 dias não houve diferenças entre o número de óbitos em cada regime, enquanto para temperaturas entre 12,90 e 15,25 graus houve um acréscimo de 300 óbitos causados por problemas cardíacos.

Poderíamos, também, considerar para a análise as séries do número de óbitos por problemas respiratórios e índice de poluição: série de vazões e índices pluviométricos, entre outras.

Uma extensão do modelo TARSO é o modelo TARMASO ( "Open loop threshold ARMA") dado por (5.1)

$$Y_t = \phi_{lo} + \sum_{i=1}^{pl} \phi_{li} Y_{t-i} + a_t^{(l)} + \sum_{j=1}^{ql} \theta_{lj} a_{t-j}^{(l)} + \sum_{v=1}^{m_l} \beta_{lv} Z_{t-v}, \quad se \quad \in R_l, \qquad (5.1)$$
$$l = 1,2,\ldots,k.$$

Assim como antes, um sistema $\{Y_t, Z_t\}$ é denominado "closed loop threshold autoregressive moving average", TARMASC, se $(Y_t, Z_t)$ e $(Z_t, Y_t)$ são ambos TARMASO.

Considerando o modelo $TARM\ ASO(2; p_1, p_2, q_1, q_2; m_1, m_2)$ com dois regimes, temos

$$Y = \begin{cases} \phi_{10} + \sum_{i=1}^{p_1} \phi_{1i} Y_{t-i} + a_t^{(1)} + \sum_{j=1}^{q_1} \theta_{1j} a_{t-j}^{(1)} + \sum_{v=1}^{m_1} \beta_{1v} Z_{t-v}, & se \quad Z_{t-d} \le r; \\ \phi_{20} + \sum_{i=1}^{p_2} \phi_{2i} Y_{t-i} + a_t^{(2)} + \sum_{j=1}^{q_2} \theta_{2j} a_{t-j}^{(2)} + \sum_{k=1}^{m_2} \beta_{1k} Z_{t-k}, & se \quad Z_{t-d} > r. \end{cases} \qquad (5.2)$$

onde estamos supondo conhecidas as ordens $p_1, p_2, q_1, q_2; m_1, m_2$ e as séries $\{Y_t, Z_i, t = 1,\ldots,n\}$.

Os parâmetros para o modelo são $\gamma_i = (\phi_{i0}, \phi_{i1},\ldots,\phi_{ip_i}, \phi_{i1},\ldots,\phi_{iq_i}, \beta_{i1},\ldots,\beta_{im_i})'$, $\tau_i, i = 1,2, r$ e $d$.

Fazendo $X_{it} = (1, Y_{t-1},\ldots,Y_{t-p_i}, a_{t-1}^{(i)},\ldots,a_{t-q_i}^{(i)}, Z_{t-1},\ldots,Z_{t-m_i})', i = 1,2$ podemos escrever o modelo (5.2) como

$$Y_t = \begin{cases} Y_1' X_{1t} + a_t^{(1)}, & se \quad Z_{t-d} \le r; \\ Y_2' X_{2t} + a_t^{(2)}, & se \quad Z_{t-d} > r. \end{cases} \qquad (5.3)$$

A função de verossimilhança é dada pela equação (2.4) substituindo as matrizes $\mathbf{B_i}$ e $\mathbf{A_i}$ por

$$\mathbf{B}_i = \begin{pmatrix} \mathbf{B}_{1i} \\ \mathbf{B}_{2i} \\ \mathbf{B}_{3i} \end{pmatrix}, \quad \mathbf{A}_i = \begin{pmatrix} \mathbf{A}_{11}^i & \mathbf{A}_{12}^i & \mathbf{A}_{13}^i \\ \mathbf{A}_{21}^i & \mathbf{A}_{22}^i & \mathbf{A}_{23}^i \\ \mathbf{A}_{31}^i & \mathbf{A}_{32}^i & \mathbf{A}_{33}^i \end{pmatrix}, i = 1,2,$$

sendo

$$\mathbf{B}_{1i} = \begin{pmatrix} \Sigma_i Y_t \\ \Sigma_i Y_t Y_{t-1} \\ \cdot \\ \cdot \\ \cdot \\ \Sigma_i Y_t Y_{t-p_i} \end{pmatrix}, \quad \mathbf{B}_{2i} = \begin{pmatrix} \Sigma_i Y_t \hat{a}_{t-1}^{(i)} \\ \Sigma_i Y_t \hat{a}_{t-2}^{(i)} \\ \cdot \\ \Sigma_i Y_t \hat{a}_{t-q_i}^{(i)} \end{pmatrix}, \quad \mathbf{B}_{3i} = \begin{pmatrix} \Sigma_i Y_t Z_{t-1} \\ \Sigma_i Y_t Z_{t-2} \\ \cdot \\ \Sigma_i Y_t Z_{t-m_i} \end{pmatrix},$$

$$\mathbf{A}_{11}^i = \begin{pmatrix} \Sigma_i 1 & \Sigma_i Y_{t-1} & \cdots & \Sigma_i Y_{t-p_i} \\ \Sigma_i Y_{t-1} & \Sigma_i Y_{t-1} Y_{t-1} & \cdots & \Sigma_i Y_{t-1} Y_{t-p_i} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \Sigma_i Y_{t-p_i} & \Sigma_i Y_{t-1} Y_{t-p_i} & \cdots & \Sigma_i Y_{t-p_i} Y_{t-p_i} \end{pmatrix},$$

$$\mathbf{A}_{12}^i = \begin{pmatrix} \Sigma_i \hat{a}_{t-1}^{(i)} & \Sigma_i \hat{a}_{t-2}^{(i)} & \cdots & \Sigma_i \hat{a}_{t-q_i}^{(i)} \\ \Sigma_i Y_{t-1} \hat{a}_{t-1}^{(i)} & \Sigma_i Y_{t-1} \hat{a}_{t-2}^{(i)} & \cdots & \Sigma_i Y_{t-1} \hat{a}_{t-q_i}^{(i)} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \Sigma_i Y_{t-p_i} \hat{a}_{t-1}^{(i)} & \Sigma_i Y_{t-p_i} \hat{a}_{t-2}^{(i)} & \cdots & \Sigma_i Y_{t-p_i} \hat{a}_{t-q_i}^{(i)} \end{pmatrix},$$

$$\mathbf{A}_{13}^i = \begin{pmatrix} \Sigma_i Z_{t-1} & \Sigma_i Z_{t-2} & \cdots & \Sigma_i Z_{t-m_i} \\ \Sigma_i Y_{t-1} Z_{t-1} & \Sigma_i Y_{t-1} Z_{t-2} & \cdots & \Sigma_i Y_{t-1} Z_{t-m_i} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \Sigma_i Y_{t-p_i} Z_{t-1} & \Sigma_i Y_{t-p_i} Z_{t-2} & \cdots & \Sigma_i Y_{t-p_i} Z_{t-m_i} \end{pmatrix},$$

$$\mathbf{A}_{23}^i = \begin{pmatrix} \Sigma_i \hat{a}_{t-1}^{(i)} Z_{t-1} & \Sigma_i \hat{a}_{t-1}^{(i)} Z_{t-2} & \cdots & \Sigma_i \hat{a}_{t-1}^{(i)} Z_{t-m_i} \\ \Sigma_i \hat{a}_{t-1}^{(i)} Z_{t-1} & \Sigma_i \hat{a}_{t-2}^{(i)} Z_{t-2} & \cdots & \Sigma_i \hat{a}_{t-2}^{(i)} Z_{t-m_i} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \Sigma_i \hat{a}_{t-q_i}^{(i)} Z_{t-1} & \Sigma_i \hat{a}_{t-q_i}^{(i)} Z_{t-2} & \cdots & \Sigma_i \hat{a}_{t-q_i}^{(i)} Z_{t-m_i} \end{pmatrix},$$

$A_{22}^{i}$ é a matriz cujo jk-ésimo elemento é $\sum_i \hat{a}_{t-j}^{(i)}\,\hat{a}_{t-k}^{(i)}$,

$A_{33}^{i}$ é a matriz cujo jk-ésimo elemento é $\sum_i Z_{t-j}Z_{t-k}$ e

$\hat{a}_t^{(i)}$ o estimador de mínimos quadrados de $a_t^{(i)}$ obtidos minimizando a soma de quadrados

$$S(\phi_i,\theta_i,\beta_i)=\sum_{t=p+1}^{n}(a_t^{(i)})^2$$

em relação a $\phi_i,\theta_i$ e $\beta_i$ utilizando um algoritimo de regressão não-linear. Veja Harvey (1981) para detalhes.

A análise *a posteriori* é feita de maneira análoga, obtendo as distribuições condicionais completas dadas nos **Corolários 1 a** 3 de acordo com *a priori* escolhida.

Um caso particular do modelo TARMASO é obtido se consideramos $k=2, m_1=m_2=0$ e $Z_t$ uma variável aleatória desconhecida. O modelo $TARMASO\,2;p_1,p_2,q_1,q_2;0,0)$ neste caso é dado por

$$Y_t=\begin{cases}\phi_{10}+\sum_{i=1}^{p_1}\phi_{1i}Y_{t-i}+a_t^{(1)}+\sum_{j=1}^{q_1}\theta_{1j}a_{t-j}^{(1)}, & se \quad Z_t\le 0;\\ \phi_{20}+\sum_{i=1}^{p_2}\phi_{2i}Y_{t-i}+a_t^{(2)}+\sum_{j=1}^{q_2}\theta_{2j}a_{t-j}^{(2)}, & se \quad Z_t>0.\end{cases} \tag{5.4}$$

Os parâmetros para o modelo (5.4) são $\gamma_i=(\phi_{i0},\phi_{i1},\ldots\phi_{ip_i},\theta_{i1},\ldots,\theta_{iq_i})'$, $\tau_i, i=1,2, r, d$ e $Z_t, t=1,\ldots,n$.

As distribuições *a priori* para os parâmetros são escolhidas como antes e para $Z_t$ vamos supor $Z_t$ *i.i.d.* $\sim N(z_0,\tau_0^{-1})$, $t=1,\ldots,n$.

Neste caso, para utilizarmos o Amostrador de Gibbs, serão necessárias as distribuições condicionais

$$\begin{aligned}
&\gamma_1|\tau_1,\tau_2,r,d,Z_{p+1},\ldots,Z_n,D;\\
&\gamma_2|\tau_1,\tau_2,r,d,Z_{p+1},\ldots,Z_n,D;\\
&\tau_1|\gamma_1,\gamma_2,r,d,Z_{p+1},\ldots,Z_n,D;\\
&\tau_2|\gamma_1,\gamma_2,Z_{p+1},\ldots,Z_n,r,d,D;\\
&r|\gamma_1,\gamma_2,\tau_1,\tau_2,d,Z_{p+1},\ldots,Z_n,D;\\
&d|\gamma_1,\gamma_2,\tau_1,\tau_2,r,Z_{p+1},\ldots,Z_n,D;\\
&Z_{p+1}|\tau_1,\tau_2,\gamma_1,\gamma_2,Z_{(p+1)},r,d,D;\\
&Z_{p+2}|\tau_1,\tau_2,\gamma_1,\gamma_2,Z_{(p+2)},r,d,D;\\
&\qquad\vdots\\
&Z_n|\tau_1,\tau_2,\gamma_1,\gamma_2,Z_{(n)},r,d,D.
\end{aligned}$$

Para a distribuição de $Z_t|\tau_1,\tau_2,\gamma_1,\gamma_2,Z_{(t)},r,d,D$, $t=p+1,\ldots,n$, temos

$$Z_t|\gamma_1,\gamma_2,\tau_1,\tau_2,Z_{(t)},D \sim N(z_0,\tau_0^{-1}).\ Gama(\frac{3}{2},\frac{Y_t^2-2\gamma_i'\mathbf{B}_t^{(i)}+\gamma_i'\mathbf{A}_t^{(i)}\gamma_i}{2}).$$

## Referências bibliográficas


CASELLA, G. AND GEORGE, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46, 167-174.

COWLES,M.K. AND CARLIN, B.P.(1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Associations*, 91 (434).

GELMAN, A. AND RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequence. *Statistical Science*, 7, 457-511.

HARVEY, A.C. (1981). *Time Series Models*. Wiley, New York.

POLE, A.M. AND SMITH, A.F.M. (1985). A Bayesian analysis of some threshold switching models. *Journal of Econometrics*, 29, 97-119.

SÁFADI, T. AND MORETTIN, P.A. (2000). Bayesian analysis of threshold autoregressive moving average models. *Sankhyā, Series B*, 62 , Pt. 3 , 353-371.

SHUMWAY, R.H., AZARI, A.S. AND PAWITAN, Y. (1998) Modeling mortality fluctuations in Los Angeles as functions of pollution and weather effects. *Environmental Research*, 45, 224-241.

TONG, H.(1970). Discussion of a paper by A.J.Lawrence and N.T.Kottegoda. *Journal of the Royal Statistical Society A*, 140, 34-35.

TONG, H. AND LIM, K.S.(1980). Threshold autoregression, limit cycles and ciclical data (with discussion). *Journal of the Royal Statistical Society B*, 42, 245-292.


## Abstract


**The statistical analysis of time series is greatly facilitated if the mean and covariances do not change with time. However, this is not a realistic assumption in some areas of applications. Seasonal variations and non linear relationships between variables imply that non linear models may be used. In this paper we consider a Bayesian analysis of open loop threshold autoregressive models (TARSO) using a proper prior and Jeffreys' prior. Inference on the parameters was obtained through the Gibbs sampler and in order to show the convergence, multiple sequences methods are used via the $\hatR$ factor proposed by Gelman e Rubin (1992). The methodology is exemplified considering the number of deaths caused by heart problems and the minimal daily temperature in São Paulo city. We have noted that for temperatures between 12.90 and 15.23 degrees there was an increase of 300 deaths.**



**Key words: Bayesian analysis; Gibbs sampler; Threshold model.**

# Política editorial

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

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

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

planejamento de pesquisas;
avaliação e mensuração de erros em pesquisas;
uso e combinação de fontes alternativas de informação; integração de dados;
novos desenvolvimentos em metodologia de pesquisa;
crítica e imputação de dados;
amostragem e estimação;
disseminação e confiabilidade de dados;
análise de dados;
análise de séries temporais;
modelos e métodos demográficos; e
modelos e métodos econométricos.

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

## INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBEs

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

Pedro Luis do Nascimento Silva
Editor Responsável
Revista Brasileira de Estatística - RBEs
Av. República do Chile 500, 10º andar
Rio de Janeiro – RJ – 20031-170
Tel.: +55 - 21 - 2514 4548
Fax: +55 - 21 - 2514 0039
E-mail: pedrosilva@ibge.gov.br

Para cada artigo publicado, serão fornecidas gratuitamente 20 separatas.

### Instruções para preparo de originais

Os originais entregues para publicação devem obedecer às seguintes normas:

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

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

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

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

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

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

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

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

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

atendimento
0800 218181