Presidente da República
Luíz Inácio Lula da Silva

Ministro do Planejamento, Orçamento e Gestão
Nelson Machado

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Eduardo Pereira Nunes

Diretor Executivo
Sérgio da Costa Côrtes

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Wasmália Socorro Barata Bivar

Diretoria de Geociências
Guido Gelli

Diretoria de Informática
Luiz Fernando Pinto Mariano

Centro de Documentação e Disseminação de Informações
David Wu Tai

Escola Nacional de Ciências Estatísticas
Pedro Luis do Nascimento Silva

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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 64 número 221 janeiro/junho 2003

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 64, n. 221, p. 1-82, jan./jun. 2003

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

Revista Brasileira de Estatística, ISSN 0034-7175

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

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

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

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

Renato Martins Assunção
Editor responsável – RBEs – IBGE.
Av. República do Chile, 500 – Centro
20031-170 – Rio de Janeiro, RJ.

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

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

Editoração
Helem Ortega da Silva - Coordenação de Métodos e Qualidade - DPE/COMEQ/IBGE

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

Capa
Renato J. Aguiar – Coordenação de Marketing/CDDI/IBGE

Ilustração da Capa
Marcos Balster – Coordenação de Marketing/CDDI/IBGE

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

v.

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

Continuação de: Revista de economia e estatística.

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

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

ISSN 0034-7175 = Revista brasileira de estatística.

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

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

Impresso no Brasil/*Printed in Brazil*

Nota do Editor

A Revista Brasileira de Estatística tem mais um número novo com quatro artigos. O primeiro artigo é bastante aplicado, mostrando uso de funções de transferência e equações de estimação para modelar dados de séries temporais. O assunto de funções de transferência é clássico em séries temporais, mas equações de estimação é uma boa aquisição na caixa de ferramentas do estatístico, fornecendo bons métodos de estimação em problema que a verossimilhança usual não é capaz de resolver. O segundo artigo trata de um tema muito recente em estatística, a técnica de *boosting*, para classificação. Uma verdadeira revolução está acontecendo nesta área da estatística nos últimos anos com o aparecimento de técnicas, tais como: *boosting*, *support vector machine*, e outras. É com satisfação que apresentamos um artigo na RBEs de revisão sobre este assunto tão novo. O terceiro artigo também usa uma técnica intensiva computacionalmente e recente, os algoritmos genéticos. Esses algoritmos servem para otimização numérica e fazem uma busca de soluções que imitam o processo evolutivo das espécies. Finalmente, o quarto artigo apresenta uma aplicação de modelos lineares generalizados num interessante problema de atuaria, um campo profissional em franca expansão no Brasil e intimamente associado com estatística e probabilidade.

Boa leitura de mais um número da Revista Brasileira de Estatística.

Renato Martins Assunção
Editor Responsável

Sumário

Aplicação de modelos de função de transferência e equações de estimação para previsão do número de passageiros em ponte aérea

Adriana Bruscato [*]
Rinaldo Artes [*]
Clélia Maria de Castro Toloi [**]

Resumo

Com a grande concorrência no setor aéreo, prever corretamente o controle de gastos torna-se cada vez mais importante. Nesse contexto, prever corretamente o número de refeições embarcadas num vôo é um fator de economia para as companhias. As empresas aéreas precisam informar o número previsto de passageiros duas horas antes de cada vôo. As refeições que não são utilizadas durante o vôo não podem ser reaproveitadas e a falta de refeições obriga a empresa a pagar um valor estipulado para que o passageiro possa fazer uma refeição ao desembarcar. Neste trabalho, são comparados os desempenhos de duas metodologias para previsão de séries temporais em dois conjuntos de dados reais: modelos de função de transferência e uma abordagem de equações de estimação. Em ambos, a série de reservas é utilizada como auxílio para fazer previsões do número de passageiros embarcados em cada vôo. É verificada também a necessidade de inclusão de intervenções nestes dados durante alguns feriados tais como: Ano Novo, Carnaval, e outros.

[*] Endereço para correspondência: Departamento de Estatística, IME – USP – Caixa Postal 66281-970 – CEP 05315 – São Paulo – SP.

[**] Ibmec / SP – Rua Maestro Cardim, 1170 – CEP 01323-001 São Paulo – SP.

1. Introdução

O controle dos gastos com refeições que não são consumidas em um vôo é um problema enfrentado por companhias aéreas. As refeições que sobram não podem ser reaproveitadas, uma vez que já estão em processo de utilização. Se por um lado não há interesse na sobra de refeições, por outro não podem faltar refeições para os passageiros embarcados, pois a empresa é obrigada a pagar para que o passageiro faça a refeição em algum estabelecimento, o que também acarreta gastos para a companhia aérea. Concluindo, a perda de dinheiro da companhia pode ser elevada fazendo com que exista o interesse em previsões mais precisas do número de refeições necessárias em cada vôo.

O embarque das refeições é feito duas horas antes da decolagem. Nesse momento, a empresa que fornece as refeições é informada pela comissária da companhia aérea sobre o número de refeições a serem embarcadas. O avião é abastecido com dez refeições acima do número fornecido pela comissária, e estas refeições destinam-se aos passageiros que chegam no aeroporto na última hora, e que não fazem reservas.

Os dados analisados foram coletados no período de janeiro a abril de 2001. O objetivo do trabalho é ajustar modelos de função de transferência e modelos de equações de estimação para prever o número de passageiros embarcados diariamente em determinados vôos e, baseado nesses modelos, prever o número de passageiros que embarcarão no vôo seguinte. Esse dado pode ser utilizado pela comissária para prever o número de refeições a serem embarcadas.

Os dados analisados referem-se a dois vôos:

- Vôo com destino a Belo Horizonte (SP-BH), ocorrendo de domingo a sexta-feira; e
- Vôo com destino ao Rio de Janeiro (SP-RJ), ocorrendo diariamente.

A variável de interesse é o número de passageiros embarcados em cada vôo. Também está disponível o número de reservas feitas até duas horas antes do vôo; esta variável poderá ser utilizada na explicação do número de passageiros embarcados.

2. Metodologias de análise

Nesta seção, serão apresentadas as metodologias utilizadas para a previsão do número de passageiros no vôo. Inicialmente, descreveremos um modelo de quase verossimilhança para séries temporais de variáveis não normais, desenvolvido por Zeger (1988) e, na seqüência, a metodologia de função de transferência Box, Jenkins e Reinsel, (1994).

2.1. Modelo de quase verossimilhança

O modelo de quase verossimilhança foi estimado pelo método proposto por Zeger e as previsões para os dados de contagem foram baseadas em Vijapurkar e Gotway (2000). A estimação dos parâmetros do modelo de quase verossimilhança foi feita com base em processos iterativos no software SPSS (1997) e a previsão foi programada no software Splus (2000). Koyama (1997) apresenta uma aplicação dessa metodologia em dados de poluição atmosférica.

Considere Y_t o número de passageiros embarcados no vôo no dia t. O modelo estipula que a estrutura de dependência de Y_t seja explicada por um processo fracamente estacionário, ε_t, com

$$E(\varepsilon_t) = 1,\ Var(\varepsilon_t) = \sigma^2 \text{ e } Cov(\varepsilon_t, \varepsilon_{t+\tau}) = \sigma^2 \rho_\varepsilon(\tau),$$

na qual $\rho_\varepsilon(\tau) = Corr(\varepsilon_t, \varepsilon_{t+\tau})$.

Desse modo, assuma que

$$E(Y_t|\varepsilon_t) = \exp(\mathbf{x}_t'\boldsymbol{\beta})\varepsilon_t = Var(Y_t|\varepsilon_t), \tag{1}$$

em que $\mathbf{x}_t$ é o vetor de covariáveis e β é o vetor de parâmetros. Note que a estrutura de momentos do processo condicional é equivalente à de uma distribuição de Poisson.

A partir de (1) temos

$$E(Y_t) = \mu_t = \exp(\mathbf{x}_t'\beta),\quad v_t = Var(Y_t) = \mu_t + \mu_t\sigma^2 \quad \text{e}$$

$$\rho_Y(t,\tau) = corr(Y_t, Y_{t+\tau}) = \frac{\rho_\varepsilon(\tau)}{\left\{\left[1+\left(\sigma^2\mu_t\right)^{-1}\right]\left[1+\left(\sigma^2\mu_{t+\tau}\right)^{-1}\right]\right\}^{\frac{1}{2}}}.$$

Esta modelagem admite a existência de sobredispersão na série Y_t, o que é comum neste tipo de variável (Cox,1983). Admita a existência de um vetor α que especifica completamente $\rho_\varepsilon(\tau)$, ou seja, $\rho_\varepsilon(\tau) = \rho_\varepsilon(\tau;\alpha)$.

Sejam $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$; $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$ e $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)'$ então a função de estimação por quase verossimilhança Zeger (1988) é dada por

$$U(\beta)=\frac{\partial\mu'}{\partial\beta}V^{-1}(\beta,\theta)(Y-\mu)=X'HV^{-1}(\beta,\theta)(Y-\mu)$$

em que $\theta=(\sigma^2;\alpha)'$, $H=diag\{exp(x_t'\beta)\}$, $V=A+\sigma^2AR_\varepsilon A$, $A=diag\{\mu i^{(n\times n)}$ e R_ε é a matriz de correlação do tipo AR(1) com elementos j, k da forma $\rho_\varepsilon(|j-k|)$.

Note que para obter os estimadores da função de quase verossimilhança é necessário usar métodos iterativos, uma vez que os mesmos não possuem forma fechada. Outra dificuldade é a inversão da matriz $V(n\times n)$ devido a sua dimensão. Para contornar esse problema, Zeger (1988) sugere substituir V por $V_z=D^{1/2}R_\varepsilon D^{1/2}$, na qual, $D=diag\{\mu_t+\sigma^2\mu_t^2\}$.

Além disso, ele sugere substituir R_ε por

$$R(\rho)=\begin{pmatrix}1 & \rho & \rho^2 & \cdots & \rho^{n-1}\\ \rho & 1 & \rho & \cdots & \rho^{n-2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1\end{pmatrix},$$

ou seja, pela matriz de correlação de um processo AR(1). A vantagem dessa substituição é que a inversa de $R(\rho)$ tem forma conhecida (Lindsey (1993), por exemplo), sendo dada por $R^{-1}(\rho)=L'(\rho)L(\rho)$, na qual,

$$L(\rho)=\frac{1}{\sqrt{1-\rho^2}}\begin{pmatrix}\sqrt{1-\rho^2} & 0 & \cdots & 0 & 0\\ -\rho & 1 & \cdots & 0 & 0\\ 0 & -\rho & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & -\rho & 1\end{pmatrix}.$$

Admitindo essas aproximações, a estimação é feita utilizando um processo iterativo, denominado mínimos quadrados filtrados e ponderados,

$$\hat\beta_R^{(j+1)}=\left\{\left(LD^{-\frac{1}{2}}\frac{\partial\mu}{\partial\beta}\right)'\left(LD^{-\frac{1}{2}}\frac{\partial\mu}{\partial\beta}\right)\right\}^{-1}\left(LD^{-\frac{1}{2}}\frac{\partial\mu}{\partial\beta}\right)'\left(LD^{-\frac{1}{2}}Z\right),$$

em que $Z=\frac{\partial\mu'}{\partial\beta}\beta+(Y-\mu)$.

A estimação das autocorrelações é feita através do método dos momentos, isto é,

$$\hat{\rho}_{\varepsilon}(\tau)=\hat{\sigma}^{-2}\frac{\sum_{t=\tau+1}^{n}\{(Y_t-\hat{\mu}_t)(Y_{t-\tau}-\hat{\mu}_{t-\tau})\}}{\sum_{t=\tau+1}^{n}\hat{\mu}_t\hat{\mu}_{t-\tau}} \quad \text{e} \quad \hat{\sigma}^2=\frac{\sum_{t=1}^{n}\left\{(Y_t-\hat{\mu}_t)^2-\hat{\mu}_t\right\}}{\sum_{t=1}^{n}\hat{\mu}_t^2}.$$

Sob condições de regularidade, $\sqrt{n}(\hat{\beta}-\beta)$ tem distribuição assintótica normal com média zero e matriz de covariâncias $\mathbf{V}_{\hat{\beta}}=\mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1}$, em que $\mathbf{I}_0=\lim_{n\to\infty}\frac{1}{n}\left(\frac{\partial\mu}{\partial\beta}'\mathbf{V}^{-1}\frac{\partial\mu}{\partial\beta}\right)$ e

$\mathbf{I}_1=\lim_{n\to\infty}\frac{1}{n}\left(\frac{\partial\mu}{\partial\beta}'\mathbf{V}_Z^{-1}\mathbf{V}\mathbf{V}_Z^{-1}\frac{\partial\mu}{\partial\beta}\right)$. $\mathbf{I}_0$ e $\mathbf{I}_1$ podem ser consistentemente estimadas substituindo-se os parâmetros por suas estimativas. Campbel (1994) e Koyama (1997) sugerem outra alternativa para estimar $\mathbf{I}_1$. Note que se $\mathbf{V}_R$ for uma boa aproximação para $\mathbf{V}$, teremos que

$$\mathbf{V}_{\hat{\beta}}\approx\lim_{n\to\infty}n\left(\frac{\partial\mu}{\partial\beta}'\mathbf{V}^{-1}\frac{\partial\mu}{\partial\beta}\right)^{-1}.$$

Vijapurkar e Gotway (2000) desenvolveram preditores por quase verossimilhança para dados correlacionados e que não têm distribuição normal, particionando-se a matriz de covariâncias até ordem $(n+h)$ da seguinte forma

$$\Sigma=Var\begin{pmatrix}\mathbf{Y}\\ \mathbf{Y_0}\end{pmatrix}=\begin{bmatrix}\Sigma_{YY} & \Sigma_{Y0}\\ \Sigma_{0Y} & \Sigma_{00}\end{bmatrix}$$

em que Σ_{yy} é a matriz de covariâncias do processo Y até o instante t, Σ_{00} é a matriz de covariâncias do processo Y do instante $t+1$ até o instante $t+h$.

Obtém-se o preditor linear

$$\hat{\mathbf{Y}}_0=\exp(\mathbf{X_0}\hat{\beta})+\Sigma_{0Y}\Sigma_{YY}^{-1}(\mathbf{Y}-\exp(\mathbf{X}\hat{\beta}))$$

em que $Y_0=(Y_{t+1},\ldots,Y_{t+h})$ é o vetor de variáveis não observadas que desejamos prever e X_0 $(h\times p)$ é a matriz de covariáveis destes valores futuros.

Para calcular uma medida de variabilidade da previsão, foi utilizado o erro quadrático médio dado por

$$\text{EQMP}\left(\hat{\mathbf{Y}}_0;\mathbf{Y}_0\right)=\Sigma_{00}-\Sigma_{0Y}\Sigma_{YY}^{-1}\Sigma_{Y0}+\left(\mathbf{D}_0-\Sigma_{0Y}\Sigma_{YY}^{-1}\mathbf{D}\right)\left(\mathbf{D}'\Sigma_{YY}^{-1}\mathbf{D}\right)^{-1}\left(\mathbf{D}_0-\Sigma_{0Y}\Sigma_{YY}^{-1}\mathbf{D}\right)',$$

em que $D=H'X(n\times p)$ e $D_0=H_0'X_0(h\times p)$.

2.2. Modelo de função de transferência

O modelo de função de transferência é utilizado quando se deseja explicar uma série temporal que tem relação linear com outra série temporal, que será usada como covariável.

Considere um processo estocástico bivariado (X_t, Y_t), $t = 0, \forall 1, \forall 2, ...$, não necessariamente estacionário. A estacionariedade deve ser obtida utilizando-se transformações adequadas, $x_t = T_1(X_t)$ e $y_t = T_2(Y_t)$. É muito comum utilizar para T_1 e T_2 diferenças simples e/ou sazonais das séries originais.

A construção do modelo baseia-se na hipótese de existência de uma relação de causalidade da série x_t para a série y_t. O objetivo dessa metodologia é identificar uma função $v(B)$, denominada função de transferência, tal que

$$y_t = v(B)x_t + N_t = w(B)\delta^{-1}(B)x_t + N_t \qquad (2)$$

em que N_t é um processo estacionário $ARMA\ (p, q)$, ou seja, $\phi(B)N_t = \theta(B)a_t$, a_t é um ruído branco, independente de x_t e $v(B) = w(B)\delta^{-1}(B)B^b$ é a função de transferência, escrita como quociente de dois polinômios finitos com defasagem b.

Existem vários métodos para identificação do modelo de função de transferência, a maioria deles segue um algoritmo semelhante. O método utilizado neste trabalho foi proposto por Haugh e Box (1977).

O procedimento de identificação ocorre em dois estágios. No primeiro estágio, ajustam-se modelos ARMA univariados a cada série do modelo; num segundo estágio, identifica-se um modelo dinâmico para as séries de resíduos. A seguir combinam-se os modelos obtidos nos dois estágios para identificar um modelo dinâmico para as duas séries originais.

Assim, o método proposto consiste das seguintes etapas:

Etapa 1: Ajustar modelos *ARMA* a cada série do modelo, obtendo-se

$$\Phi_y(B)y_t = \Theta_y(B)e_t \text{, ou seja, } \Pi_y(B)y_t = e_t \text{ e}$$

$$\Phi_x(B)x_t = \Theta_x(B)u_t \text{, ou seja, } \Pi_x(B)x_t = u_t \text{.} \qquad (3)$$

Esse procedimento é denominado pré-branqueamento das séries.

Etapa 2: Calcular a função de correlação cruzada entre as séries residuais $\hat{e}_t$ e $\hat{u}_t$, com o objetivo de identificar uma possível relação linear entre e_t e u_t. Pode-se dizer, usando as expressões (1) e (2), que esta relação é dada por

$$e_t = v^*(B)u_t + \psi^*(B)a_t = w^*(B)\delta^{*-1}(B)u_t + \psi^*(B)a_t \text{,}$$

em que $w^*(B)$ e $\delta^*(B)$ são polinômios de graus s e r, respectivamente.

Logo, o padrão da função de correlação cruzada estimada entre $\hat{u}_t$ e $\hat{e}_t$ fornecerá uma identificação preliminar da função de transferência $v^*(B)$, em sua forma racional, ou seja, identificam-se as ordens dos polinômios $w^*(B)$ e $\delta^*(B)$.

Etapa 3: Determinar a forma preliminar do ruído, $\theta^*(B)\phi^{*-1}(B)a_t$, onde $\phi^*(B)$ e $\theta^*(B)$ têm ordens p^* e q^*, respectivamente.

Mostra-se que

$$\phi^*(B) = \delta^*(B) \text{ e } \theta^*(B) \sim MA(q^*) \text{,} \qquad (4)$$

em que $q^* \leq \max(r^*, s^*)$.

Etapa 4: Identificar o modelo de função de transferência, expressão (1), para as séries x_t e y_t.

Combinando os modelos univariados para x_t e y_t, dados pela expressão (2), com o modelo identificado ligando u_t e v_t, expressão (3), obtém-se o modelo de função de transferência preliminar dado por

$$y_t = W^*(B)\delta^{*-1}(B)\Pi_x(B)\,\Pi_y^{-1}(B)x_t + \theta^*(B)\delta^{*-1}(B)\,\Pi_y^{-1}(B)a_t\,.$$

As estimativas preliminares, que servirão como valores iniciais na estimação do modelo de função de transferência dado por (4), serão dadas pelas equações:

$$\hat{w}(B) = \hat{w}^*(B)\hat{\pi}_X(B)$$
$$\hat{\partial}(B) = \hat{\partial}^*(B)\hat{\pi}_Y(B)$$
$$\hat{\theta}(B) = \hat{\partial}^*(B)$$
$$\hat{\phi}(B) = \hat{\theta}^*(B)\hat{\pi}_Y(B)$$

em que todos os operadores do lado direito são obtidos durante o método de identificação.

A verificação de adequação do modelo é feita via função de autocorrelação e autocorrelação parcial dos resíduos do modelo final, e, também através da função de correlação cruzada entre a entrada pré-branqueada e os resíduos do modelo final para ver se não sobrou nenhuma dependência entre X e Y que não foi incluída no modelo final.

Maiores detalhes podem ser encontrados em Morettin e Toloi (1989).

O SCA (1986) foi o software usado para estimar os parâmetros dos modelos de função de transferência.

3. Aplicação

Nesta seção, descrevemos o comportamento das variáveis a serem analisadas, bem como o tratamento preliminar dos dados.

3.1. Análise exploratória

O problema inicial no banco de dados é a ausência de observações (missing values) devido a cancelamentos ou atrasos nos vôos. Nesses casos, a imputação de dados foi feita utilizando-se a média aritmética entre as observações da semana anterior e posterior daquele dia da semana em que seria feita a imputação.

A Figura 1 traz os gráficos das séries de passageiros embarcados e reservas para cada um dos vôos analisados. Com base nesta figura, percebe-se que não há uma clara tendência de crescimento e que as séries parecem apresentar um comportamento sazonal, repetindo-se semanalmente. Além disso, a série de reserva acompanha a série do número de passageiros

embarcados, podendo realmente servir como variável explicativa num modelo de regressão. Nota-se que as séries de reserva têm variabilidade maior e estão acima das séries do número de passageiros embarcados, na maior parte do tempo avaliado.

Figura 1 - Gráfico das séries do número de passageiros embarcados (PAX) e das reservas (01/01/01 a 22/04/01)

Vôo SP-BH **Vôo SP-RJ**

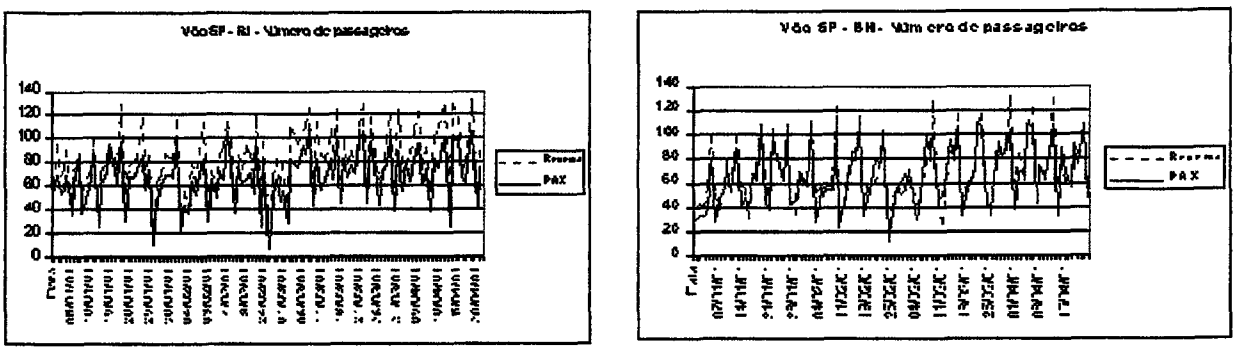
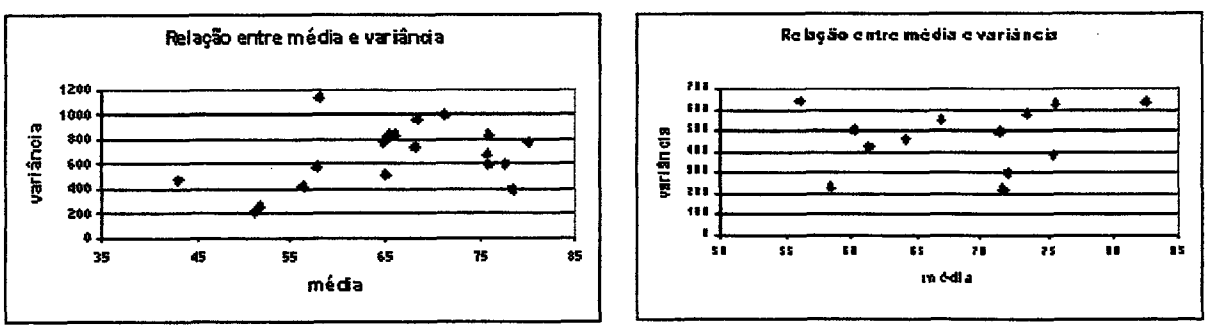


Figura 2 - Gráfico da relação entre a média e a variância das séries do número de passageiros embarcados

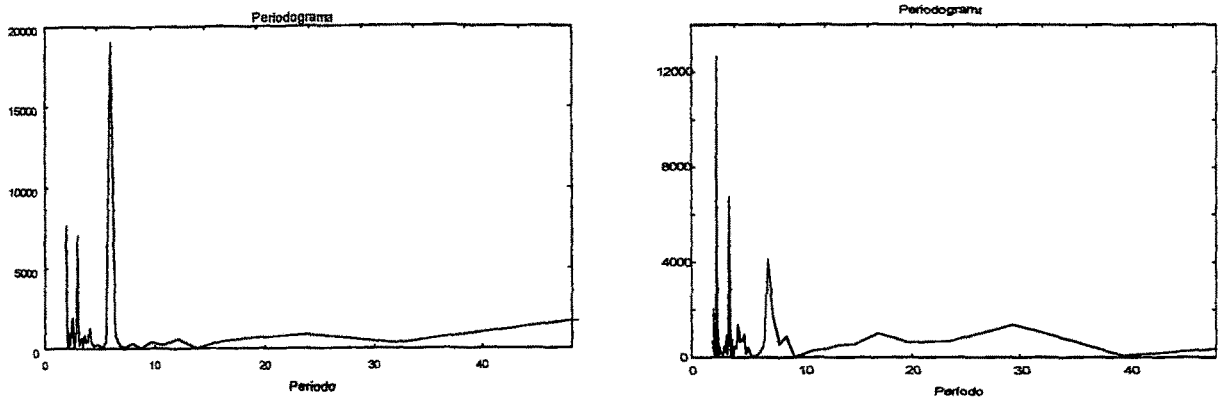
Vôo SP-BH **Vôo SP-RJ**



A Figura 2 apresenta a relação entre média e variância para os vôos avaliados. Não há indícios claros de relação entre a média e a variância das séries para os vôos com destino a Belo Horizonte e Rio de Janeiro; especialmente para as maiores médias parece que a variância não aumenta. Desta forma, não é necessário o uso de transformação para estabilizar a variância destas séries na metodologia de função de transferência.

Figura 3 - Periodograma das séries do número de passageiros embarcados

Vôo SP-BH **Vôo SP-RJ**



A análise dos periodogramas, Figura 3, revela a existência de picos significantes (teste de Fisher) nas freqüências correspondentes às periodicidades seis dias (SP-BH) e sete dias (SP-RJ), confirmando a sazonalidade semanal das séries. Para mais detalhes, ver Priestley (1981).

O efeito sazonal pode ser introduzido no modelo de equações de estimação através da inclusão de senóides com o período de interesse. Por exemplo, no caso do vôo SP-BH verificamos uma periodicidade semanal, então utilizamos funções seno e co-seno com período de seis dias para explicar a sazonalidade semanal do vôo, que ocorre de domingo a sexta-feira.

3.2. Estimação dos modelos

Foi verificada a influência de algumas intervenções em feriados Box et al., (1994), uma vez que nesses dias o comportamento do número de passageiros pode ser diferente dos demais. As intervenções (representadas através de variáveis indicadoras) foram:

1. Ano Novo: SP-BH (01/01 a 03/01) e SP-RJ (01/01 a 05/01);
2. Fundação de São Paulo: SP-BH (25/01);
3. Carnaval: SP-BH e SP-RJ (26/02 a 02/03); e
4. Páscoa: SP-BH (11/04 a 13/04) e SP-RJ (13/04).

A única intervenção estatisticamente significante foi Ano Novo para o vôo com destino a Belo Horizonte, segundo a abordagem de função de transferência. Nos modelos de quase verossimilhança, nenhuma das intervenções apresentou significância estatística.

Para fazer a previsão de cada dia da semana, estimaram-se os parâmetros do modelo usando observações até o dia anterior, ou seja, a estimação foi feita utilizando até a última

observação antes da previsão. A previsão era feita e colocava-se a próxima observação no modelo para refazer as estimativas dos parâmetros para a previsão do vôo seguinte.

A análise de resíduos (Apêndice) foi feita construindo-se gráficos de resíduos padronizados para os modelos de quase verossimilhança e funções de autocorrelação dos resíduos para ambos os modelos.

3.2.1. Abordagem de quase verossimilhança

O modelo de equações de estimação usado para todas as séries foi o seguinte

$$E(Y_t) = \mu_t = \exp[\beta_0 + \beta_1 t + \beta_2 X_t + \beta_3 \cos(2\pi t/k) + \beta_4 \operatorname{sen}(2\pi t/k)],$$

em que k é o número de ocorrências do vôo por semana e X_t é o número de reservas no tempo t.

Na estimação de quase verossimilhança, o dia 20/01/01 foi escolhido como origem dos tempos para retirar um possível efeito do início do ano nas estimativas dos modelos, já que foram feitas muitas imputações de dados no início do ano devido a cancelamentos de vôos.

Para os modelos de quase verossimilhança, os resíduos de Pearson têm um bom comportamento como pode ser visto na Figura 6 (Apêndice). Os valores se distribuem numa faixa horizontal entre −2 e 2 e existem apenas alguns valores fora dessa faixa. Além disso, as funções de autocorrelação destes mesmos resíduos, para cada um dos vôos, está dentro dos limites de confiança, indicando que são ruídos brancos. Desta forma, os modelos de quase verossimilhança parecem estar bem ajustados.

As estimativas para os parâmetros do modelo de quase verossimilhança, bem como os erros padrão destas estimativas estão na Tabela 1.

Utilizando o teste de Wald percebe-se que a tendência não foi estatisticamente significante para todos os modelos. Isto era esperado, pois as séries parecem desenvolver-se ao longo de uma média constante.

A série de reserva foi considerada importante para auxiliar a explicação do número de passageiros embarcados em todos os vôos estudados. O parâmetro β_2 representa a variação média no logaritmo do número médio de passageiros embarcados, quando aumenta-se uma unidade no número de reservas, e mantendo-se as demais variáveis fixas. Para todos os vôos, existe um aumento médio de um passageiro embarcado por unidade de reserva.

Os parâmetros referentes aos senos (β_4) e co-senos (β_3) indicam a contribuição da periodicidade semanal para a explicação do número de passageiros embarcados. Em todos os vôos eles foram significativos, confirmando a sazonalidade semanal nas séries.

Como as estimativas para a variância do processo latente (σ^2) foram próximas de zero, significa que há pouca sobredispersão nos modelos.

3.2.2. Abordagem de função de transferência

O modelo geral utilizado para cada um dos vôos foi proposto utilizando as técnicas de Box, Jenkins e Reinsel (1994) e está a seguir

$$(1-B^k)Y_t = w_0(1-B^k)X_t + \phi^{-1}(B)\,\theta(B)a_t\,.$$

em que k é o número de ocorrências do vôo por semana, X_t é o número de reservas no tempo t e a_t é um processo ruído branco.

Quanto ao ajuste dos modelos de função de transferência, apresentam-se as funções de autocorrelação dos resíduos para cada um dos vôos (Apêndice). Estas funções também estão dentro dos limites de confiança, indicando que são ruídos brancos e que a correlação serial foi bem modelada.

Na Tabela 1 estão as estimativas dos parâmetros para os modelos de função de transferência. Foi aplicada uma diferença sazonal (semana) para todas as séries avaliadas. Com base na análise das estimativas contidas na Tabela 1, nota-se que o número de passageiros embarcados num determinado instante depende do número de passageiros embarcados no mesmo vôo, na semana anterior; da diferença do número de reservas naquele instante e na semana anterior e de um modelo residual.

Por exemplo, o modelo de função de transferência para o vôo com destino a Belo Horizonte leva em conta o número de passageiros embarcados neste vôo na semana anterior, 76.73 % da diferença da reserva desta semana para a semana anterior, um modelo MA(2) e mais o decréscimo médio de sete passageiros no número de passageiros embarcados no Ano Novo (α).

Tabela 1 - Estimativa dos parâmetros dos modelos

Equações de Estimação			Função de Transferência		
Parâmetro	Vôo SP-BH	Vôo SP-RJ	Parâmetro	Vôo SP-BH	Vôo SP-RJ
β_0	3.1982 (0.0546)	3.1329 (0.0426)	w_0	0.7673 (0.0480)	0.7630 (0.0350)
β_1	0.0010 (0.0007)	-0.0006 (0.0003)	ϕ_1	-----	-----
β_2	0.0129 (0.0006)	0.0125 (0.0005)	θ_1	-0.2062 (0.0714)	----
β_3	0.1140 (0.0242)	-0.0208 (0.0149)	θ_2	0.7625 (0.0670)	0.9312 (0.0489)
β_4	-0.0148 (0.0286)	0.0564 (0.0147)	α	-7,0044 (3,0614)	----
ρ_ε	0.2166	-0.1016	R^2	0.881	0.904
σ^2	0.0030	0.0016	Erro padrão residual	8.6453	6.7957

Entre parênteses está o erro padrão

3.2.3. Previsões do número de passageiros embarcados

Foram feitas previsões para valores relativos a um dia à frente, usando as observações até o dia 22/04/01.

As Tabelas 2 e 3 apresentam as previsões dos modelos de quase verossimilhança e função de transferência. Essas previsões serão consideradas não satisfatórias, quando faltarem ou excederem em dez o número de refeições necessárias no vôo, estas estão em negrito e sublinhadas.

Tabela 2 - Previsões do número de passageiros embarcados no vôo SP-BH

		Função de Transferência		Quase Verossimilhança	
Data	Observado	Previsto	Prev – Obs	Previsto	Prev - Obs
23/04/01	48	60 (9)	**12**	51 (9)	3
24/04/01	63	70 (9)	7	65 (11)	2
25/04/01	98	95 (9)	-3	104 (14)	6
26/04/01	65	70 (8)	5	68 (11)	3
27/04/01	106	114 (10)	8	131 (17)	**25**
29/04/01	Cancelado	*****	*****	*****	*****

Entre parênteses está o erro de previsão

Tabela 3 - Previsões do número de passageiros embarcados no vôo SP-RJ

		Função de Transferência		Quase Verossimilhança	
Data	Observado	Previsto	Prev – Obs	Previsto	Prev - Obs
23/04/01	54	53 (7)	-1	50 (7)	-4
24/04/01	86	87 (7)	1	89 (10)	3
25/04/01	81	77 (7)	-4	76 (9)	-5
26/04/01	108	106 (7)	-2	114 (12)	6
27/04/01	48	55 (6)	7	47 (7)	-1
28/04/01	85	87 (6)	2	83 (10)	-2
29/04/01	50	48 (6)	-2	45 (7)	-5

Entre parênteses está o erro de previsão

Observando-se as previsões dos dois modelos nas Tabelas 2 e 3 e também os Gráficos 4 e 5, conclui-se que as previsões foram satisfatórias. A pior previsão do modelo de quase verossimilhança resultou numa sobra de 25 refeições, e a pior no modelo de função de transferência acarretou uma sobra de 12 refeições. Os erros padrões das estimativas foram maiores para o modelo de quase verossimilhança.

Na Figura 5, observa-se o gráfico de dispersão das previsões dos dois modelos estudados. A linha pontilhada na bissetriz do primeiro quadrante do gráfico representa a igualdade entre as duas previsões, ou seja, os pontos mais próximos da reta foram aqueles que tiveram maior concordância entre as previsões dos modelos. O gráfico mostra que existe um alto grau de coerência entre as duas previsões, uma vez que os pontos não estão distantes da reta.

Observando-se o gráfico de dispersão dos erros de previsão (diferença entre o valor previsto e o valor observado) dos modelos de quase verossimilhança e função de transferência na Figura 5, percebe-se que os dois modelos tendem a apresentar erros de previsão no mesmo sentido, isto é, ambos tendem a fazer previsões abaixo do valor observado ou ambos tendem a fazer previsões acima do valor observado pois a maioria dos erros de previsão estão acomodados no primeiro e terceiro quadrantes.

A grande maioria dos valores previstos ficou dentro dos limites esperados pela companhia aérea de dez refeições a mais e dez refeições a menos em ambos os modelos utilizados. A quantidade de previsões fora do esperado (+ ou – 10) nos modelos de função de transferência e quase verossimilhança foi a mesma (uma no vôo SP-BH), indicando também que os dois modelos apresentaram bom desempenho nas previsões.

Figura 4 - Gráfico dos vôos no mês de abril com as previsões dos modelos de função de transferência e de quase verossimilhança

Vôo SP-BH **Vôo SP-RJ**

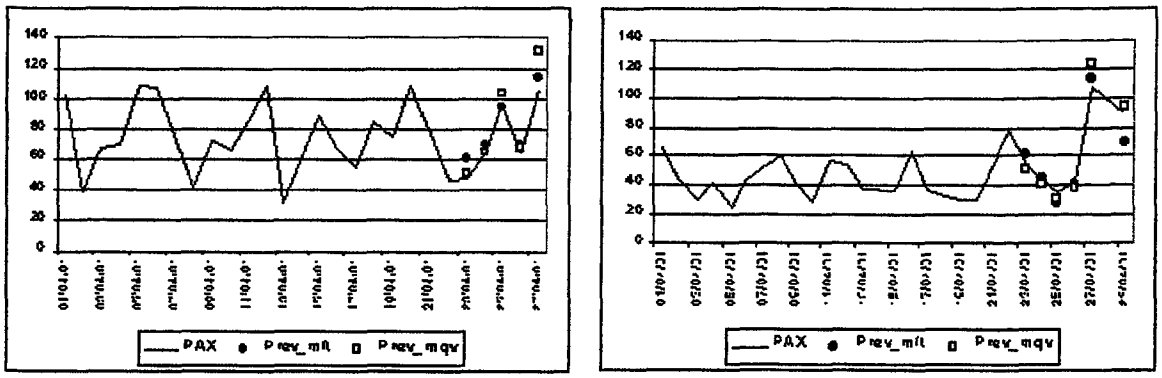
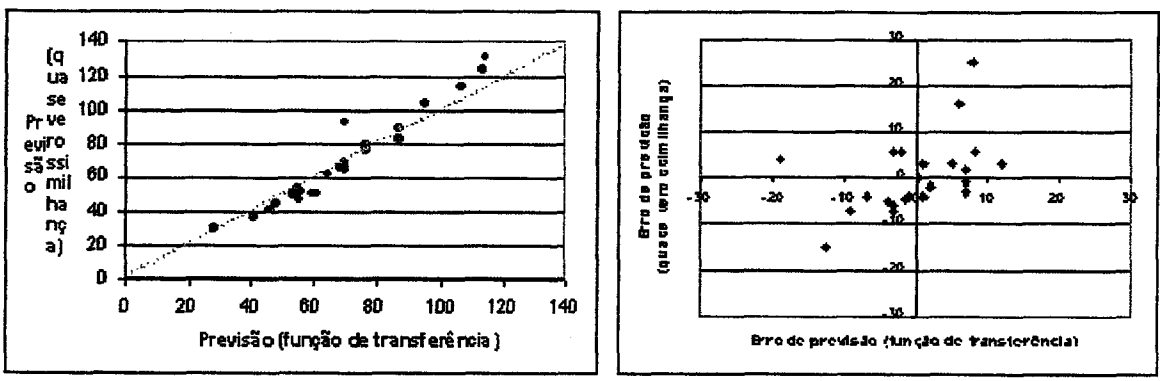


Figura 5 - Gráfico de dispersão das previsões e dos erros de previsão dos modelos de quase verossimilhança e de função de transferência



4. Conclusões gerais

Foram utilizados um modelo de função de transferência e um modelo de quase verossimilhança para estimar os parâmetros relativos a cada série e, com base nestas estimativas, foram calculadas previsões um passo a frente para algumas observações, bem como o erro padrão para estas estimativas.

No caso dos modelos de quase verossimilhança, utilizando o teste de Wald percebe-se que a tendência não foi estatisticamente significante para o modelo de quase verossimilhança, o que era de esperar, pois as séries parecem desenvolver-se ao longo de uma média constante. A reserva pode ser utilizada como variável independente para explicar o número de passageiros embarcados. Isto quer dizer que o número de reservas foi estatisticamente significante para explicar a variabilidade do número de passageiros embarcados, e que as séries dos vôos possuem periodicidade semanal, que foi explicada pelas variáveis seno e co-seno.

Como as estimativas para a variância dos processos latentes (σ^2) foram próximas de zero, significa que há pouca sobredispersão no modelo. A correlação entre os tempos é baixa e próxima de zero, exceto para a série do vôo com destino a Brasília, indicando que existe uma forte dependência em relação aos dias anteriores nesta série.

Nos modelos de função de transferência, pode-se notar a grande influência da semana anterior no número de passageiros embarcados na semana seguinte, bem como a significativa participação da série de reservas na previsão do número de passageiros embarcados.

O ajuste dos modelos foi adequado, podendo ser conferido nos gráficos dos resíduos de Pearson em função dos valores ajustados, para o modelo de quase verossimilhança e, com base nas funções de autocorrelação residuais, para ambos os modelos.

As previsões feitas pelo modelo de quase verossimilhança podem ser consideradas praticamente tão boas quanto as previsões feitas com base no modelo de função de transferência. O erro padrão das estimativas são um pouco maiores para o modelo de quase verossimilhança, mas em compensação ele é mais simples quando comparado ao de função de transferência.

Apêndice (Análise de resíduos)

Figura 6 - Gráfico dos resíduos de Pearson pelos valores ajustados

Vôo SP-BH **Vôo SP-RJ**

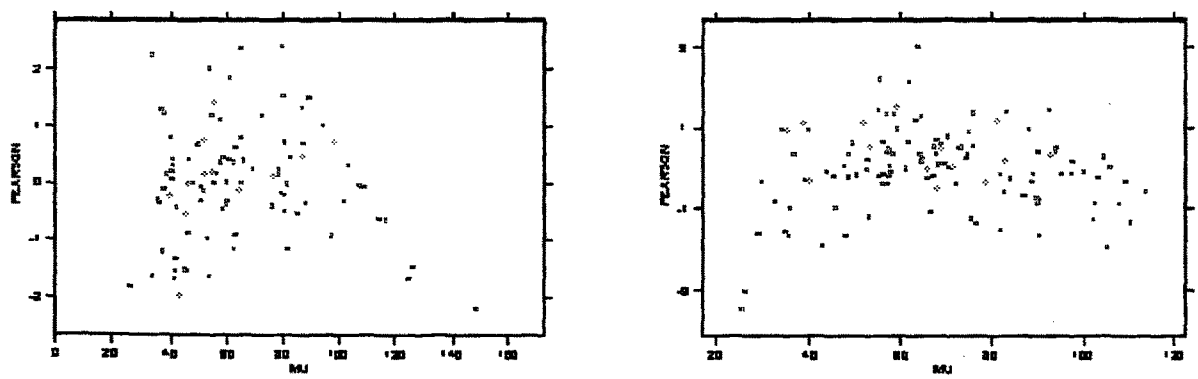
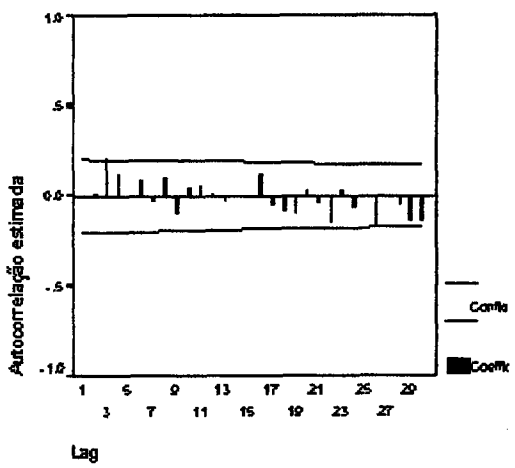
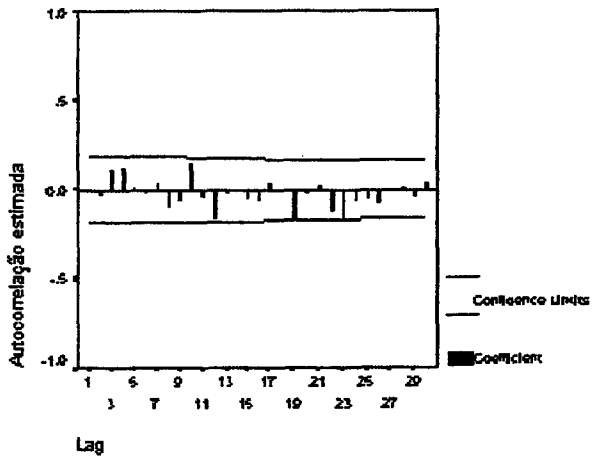


Figura 7 - Fac dos resíduos dos modelos de equações de estimação e função de transferência

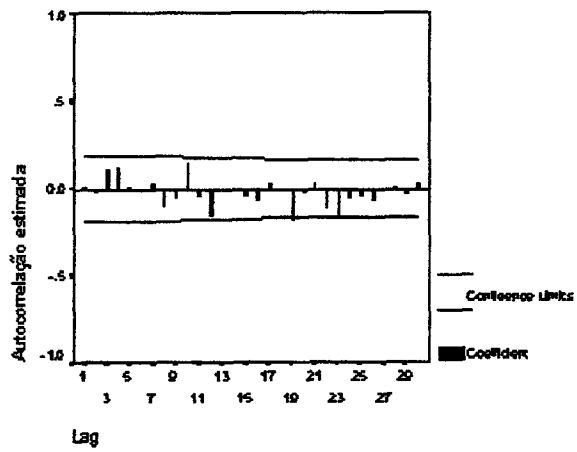
Equações de Estimação - Vôo SP-BH



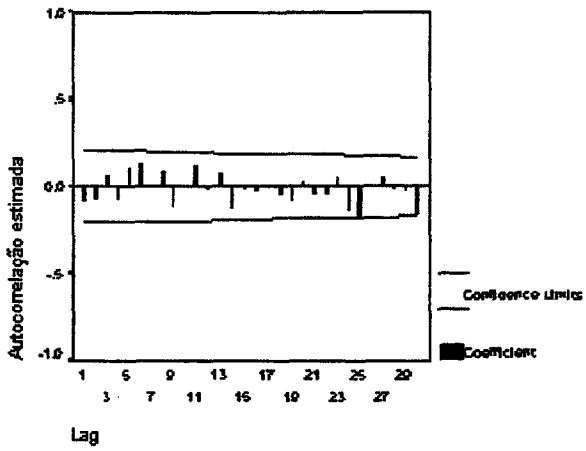
Equações de Estimação - Vôo SP-RJ



Função de Transferência - Vôo SP-BH



Função de Transferência - Vôo SP-RJ



Referências bibliográficas

BOX, G. E. P.; JENKINS, G. M. *Time Series Analysis, Forecasting and Control.* Holden-Day, San Francisco, 1970.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time Series Snalysis: Forecasting and Control.* 3 ed. Prentice Hall, New Jersey, 1994.

BRILLINGER, D. R. *Time Series: Data Analysis and Theory.* Holden-Day, San Francisco, 1981.

CAMPBELL, M.J. Time series regression for counts: an investigation between sudden infant death syndrome and environmental temperature. *Journal of the Royal Statistical Society,* 157, pg. 191-208, 1994

COX, D. R. Some Remarks on Over-dispersion. . *Biometrika,* 70, pg. 267-74, 1983.

HAUGH, L.; BOX, G. E. P. Identification of Dynamic Regression (distributed lag) Models Connecting Two Time Series. *Journal of the American Statistical Association,* 72, pg.121-130, 1977.

KOYAMA, M.A. *O Modelo de Zeger para Análise de Séries de Contagem.* (Dissertação de Mestrado). IME-USP, São Paulo, 1997.

LINDSEY, J.K. *Models for Repeated Measurements.* Oxford University Press, New York, 1993.

MORETTIN, P. A. ; TOLOI, C. M. *Modelos de Função de Transferência.* 3ª Escola de Séries Temporais e Econometria. Associação Brasileira de Estatística, 1989.

PRIESTLEY, M. B. *Spectral Analysis and Time Series,* volume 1, 2. Academic Press, New York, 1981.

SCA. Scientific Computing Associates. Lisle, Illinois, 1986.

Splus 2000. Math Soft Inc. Seatle, Washington, 2000.

SPSS 8.0. SPSS Inc. Chicago, Illinois,1997.

VIJAPURKAR, U. P.; GOTWAY, C. A. Assessment of Forecasts and Forecast Uncertainty Using Generalized Linear Regression Models for Time Series Count Data. *J. Statist. Comput. Simul.,* 68, pg. 321-49, 2000.

ZEGER, S. L. A Regression Model for Time Series of Counts. *Biometrika,* 75, pg. 621-9, 1988.

Abstract

With the large competition of airline companies, to forecast correctly the number of meals that are carried in a flight is an economy factor for the companies. The airline companies need to inform the forecasted number of passengers 2 hours before the flight. The meals that are not consumed during the flight can not be used again and the shortage of meals obliges the company to pay a stipulated value in order that passengers could have a meal on landing.

In this work we intend to compare the performance of two time series forecasting methods: transfer function models and an approach based on estimating equations. In both cases we use the bookings series as an auxiliary tool to forecast the number of passengers on board in each flight. It is also verified the need for inclusion of interventions during some holidays such as New Year and Carnival.

Alocação de clientes em grupos usando classificação via *Boosting*: uma comparação com os métodos tradicionais de classificação

Alexandre Rübesam *
Ronaldo Dias **

Resumo

Os métodos estatísticos tradicionalmente usados em classificação são, basicamente, regressão logística e análise discriminante. Outros métodos que recentemente aparecem nas aplicações são árvores de classificação e redes neurais. Os métodos tradicionais apresentam baixo custo computacional, mas são pouco flexíveis, enquanto os métodos como redes neurais são muito flexíveis, mas caros e com menor interpretabilidade. A metodologia apresentada, *Boosting*, é flexível, fácil de aplicar e tem um custo computacional baixo. *Boosting* funciona combinando seqüencialmente classificadores simples, dando maior peso em cada passo às observações classificadas incorretamente no passo anterior.

Este trabalho apresenta a metodologia *boosting*, e mostra dois exemplos de sua aplicação: uma em dados simulados, e outra em dados reais. A aplicação em dados reais consiste em alocar clientes de uma loja de varejo em grupos que definem os perfis destes clientes. Os resultados obtidos com *boosting* nas duas aplicações são comparados aos resultados dos métodos tradicionais.

* Endereço para correspondência: Deptº de Estatística - IMECC - Universidade Estadual de Campinas – e-mail: rubesam@ime.unicamp.br, financiado pela FAPESP.

** dias@ime.unicamp.br

1. Introdução

Os métodos de classificação (ou discriminação) mais conhecidos e mais utilizados em estática aplicada são, provavelmente, regressão logística Agresti (1996), Hosmer e Lemeshow (1989) e análise discriminante Johnson e Wichern (2002). Esses métodos, cujos fundamentos estão estabelecidos na teoria estática há bastante tempo, são lineares, têm baixo custo computacional, e em geral produzem resultados razoáveis, apesar de não raro serem utilizados em situações que contradizem suas suposições paramétricas distribucionais. As propriedades estatísticas destes métodos são conhecidas há muito tempo, e suas interpretações são claras. Mais recentemente, surgiram métodos do tipo árvores de classificação e regressão (CART) Breiman (1984), um método altamente interpretável, e cujo uso está bastante disseminado. Este método gera regras de decisão que podem ser interpretadas diretamente em termos das variáveis explanatórias. O desempenho deste método, entretanto, é baixo quando a fronteira de decisão ótima não pode ser aproximada por um conjunto de hiper-retângulos.

O aumento da capacidade computacional também tornou populares os chamados métodos de "inteligência artifical", como as redes neurais. Esses métodos têm, em geral, desempenho superior em relação aos métodos citados acima, especialmente quando a fronteira de decisão ótima é não linear mas com um custo computacional bem mais alto. Além disso, os algoritmos de otimização popularmente utilizados são sensíveis aos ajustes específicos que devem ser feitos (características dos algoritmos de otimização numérica, por exemplo, métodos de descida de gradiente, métodos de Newton), e é preciso atenção no processo de estimação para evitar o problema de *overfitting*. Outra questão problemática é a escolha da topologia da rede (número de camadas intermediárias e número de unidades em cada camada). Métodos recentes procuram tratar a questão da generalização ou do desempenho da rede através do controle automático dos parâmetros da rede. A interpretação do modelo pode ser realizada através do uso de algoritmos de extração de regras ou da simplificação do mesmo.

Um método de classificação moderno que tem apresentado bons resultados é o método conhecido como *boosting*. Este método, que surgiu na área de computação Schapire (1990), atinge um desempenho muito bom nas aplicações testadas, com um custo computacional relativamente baixo, mas sofre do problema de interpretação comum aos métodos modernos. O algoritmo de *boosting* funciona aplicando-se, seqüencialmente, um algoritmo de classificação qualquer (chamado de base) a versões iterativamente reponderadas do conjunto de dados de treinamento. Em cada iteração, as observações classificadas incorretamente na iteração anterior recebem um peso maior. A saída final combina os classificadores construídos em cada iteração,

produzindo um comitê de classificação. O desempenho deste comitê é perto de ótimo na maioria dos casos, e a questão de *overfitting* pode ser facilmente controlada Friedman, Hastie e Tibshirani (2000) mostraram que o algoritmo AdaBoost, um algoritmo de *boosting* apresentado neste trabalho, é um algoritmo que ajusta, aproximadamente, um modelo logístico aditivo, no qual o número de iterações é o número de funções usadas na representação aditiva, ou seja, o número de funções somadas para aproximar a função estimada, o que torna o assunto atraente de um ponto de vista estatístico. Mais recentemente, foi mostrado que os vários algoritmos de *boosting* existentes pertencem a uma classe de algoritmos que minimizam um funcional de custo suave através de algum método numérico Mason (1999), e a atenção da comunidade estática voltou-se para características como consistência do método, analisada em uma variedade de artigos (Breiman 2000, Jiang 2000a, Jiang 2000b, Lugosi e Vayatis 2004).

O objetivo deste trabalho é apresentar a metodologia *boosting*, pouco conhecida na área de estatística, especialmente no Brasil, e mostrar que esta ferramenta possui desempenho melhor, quando comparada às metodologias tradicionais, com um custo computacional relativamente baixo. As aplicações apresentadas visam a mostrar que o método é de fácil implementação, mesmo em tarefas muito complexas, onde a quantidade de informação é extremamente grande, assim como o número de classes.

O trabalho está dividido da seguinte maneira. A seção 2 apresenta o problema de classificação. A seção 3 apresenta os algoritmos *AdaBoost* e *LogitBoost* para o caso de classificação em duas classes e em J classes. A seção 4 apresenta os resultados obtidos com simulação de dados, e a seção 5, os resultados com dados reais. A seção 6 contém considerações finais e conclusão.

2. Classificação

Um procedimento de classificação, regra de classificação ou classificador é algum método que (possivelmente de maneira automática) classifique objetos em classes. Em geral, um procedimento de classificação é construído com base na experiência passada, e o interesse é utilizá-lo para classificar novos objetos.

Para definir uma tarefa de classificação, considere que temos uma amostra de treinamento $L = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, onde $\mathbf{x}_i = (x_i, 1, \ldots, x_{i,p})$ é um vetor p-dimensional que contém medições feitas no indivíduo i, y_i é o rótulo de classe do indivíduo i e N é o tamanho da amostra. Considere que o número possível de classes é J, e que estas estão contidas em um conjunto $C = \{1, 2, \ldots, J\}$.

Além disso, o espaço de todos os possíveis valores de $\mathbf{x}$ é designado por X.

A tarefa de classificação consiste em obter uma regra $F: X \to C$ a partir da amostra L, tal que para cada $\mathbf{x} \in X, F(\mathbf{x})$ designa uma classe em C. É desejável que o classificador F tenha erro pequeno não somente sobre a amostra L, mas em todo o espaço X. Se definirmos uma medida de probabilidade P em $X \times C$, podemos definir a medida de erro por $pce^{nc} = P(F(\mathbf{X}) \neq Y)$, a probabilidade não condicional de classificação incorreta (ou taxa de erro), onde Y é uma variável aleatória que assume valores em C. Obviamente, $pce^{nc} \in [0,1]$. Este erro pode ser estimado de algumas maneiras: usando o próprio conjunto de treinamento, usando um conjunto de dados independente, chamado de conjunto de teste e ainda usando validação cruzada.

No primeiro caso, a estimativa da taxa de erro para o classificador F, denotada $\hat{pce}^{nc}(F)$, é dada abaixo:

$$\hat{pce}^{nc}(F) = \frac{1}{N} \sum_{i=1}^{N} I(F(\mathbf{x}_i) \neq y_i), \tag{1}$$

onde $I(\cdot)$ denota a função indicadora do evento dentro dos parênteses, isto, vale 1 se o argumento é verdadeiro, e 0 se for falso. Esta estimativa é chamada de estimativa por ressubstituição Breiman (1984). Como ela usa os mesmos dados que foram usados para estimar o classificador, em geral a estimativa é otimista. Intuitivamente, é razoável que nos dados de treinamento o desempenho seja ligeiramente melhor do que em dados que não foram usados na estimação uma vez que o classificador é construído de maneira a minimizar algum critério relacionado à taxa de erro usando os valores específicos da amostra de treinamento.

Uma maneira de se obter uma estimativa mais honesta da taxa de erro é a estimação por amostra de teste. Esta consiste em dividir o conjunto de dados de treinamento $L = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$ em duas amostras independentes, amostras L_1, de tamanho N_1, e L_2, de tamanho N_2. Usamos L_1 para construir o classificador F (ou seja, L_1 é agora a amostra de treinamento) e o testamos na amostra L_2, que funciona agora como amostra de teste (*test sample*). A estimativa de pce^{nc} é então

$$\hat{pce}^{nc}_{teste}(F) = \frac{1}{N_2} \sum_{i=1}^{N_2} I(F(\mathbf{x}_i) \neq y_i) \tag{2}$$

Uma divisão sugerida heuristicamente é usar 2/3 dos dados para estimação, e 1/3 para teste. O problema com esta estimativa é que, se o conjunto de dados não for muito grande, perde-se informação que poderia ser utilizada na estimação do classificador.

A estimação da taxa de erro por validação cruzada consiste em dividir o conjunto original de tamanho N em V partes, ou seja, temos $L_1, L_2, \ldots, L_V$, cada um com tamanho N_v. Para cada parte v, construa o classificador $F^{(v)}$ usando as $v-1$ partes e teste na restante. Então a estimativa de pce^{nc} usando a parte v pode ser calculada por (2), substituindo N_2 por N_v, $v = 1,2,\ldots,V$ e fazendo a soma sobre os casos em N_v. Então a estimativa de pce^{nc} por validação cruzada é dada por

$$\hat{pce}^{nc}_{VC}(F) = \frac{1}{V}\sum_{v=1}^{V} pce^{nc}_{teste}(F^{(v)}). \qquad (3)$$

A comparação do método apresentado como os métodos tradicionais será feita tendo como medida as estimativas das taxas de erro usando amostra-teste.

2.1. A regra de *Bayes*

A regra de *Bayes* é o melhor classificador construível, no sentido de minimizar a perda (5) definida abaixo. Esta regra só pode ser construída quando se sabe a distribuição dos dados (o que não ocorre na prática), mas além de dar contribuição teórica no desenvolvimento de classificadores, ela é útil na comparação de classificadores em dados simulados, por fornecer uma base de referência, com uma taxa de erro que é a princípio a menor atingível Ripley (1996).

Denotemos por $\mathbf{X}$ o vetor aleatório das variáveis medidas em cada objeto e por Y a variável aleatória que assume valores em C, ou seja, a classe a que o objeto pertence. Vamos assumir que o vetor $\mathbf{X} \in \Re^p$, ou seja, assumimos que as variáveis são todas contínuas.

Sejam:

$p(\mathbf{x}|k)$ a densidade de $\mathbf{X}$ dado $Y = k$

$\pi_k = P(Y = k)$

$p(k|\mathbf{x}) = P(Y = k|\mathbf{X} = x)$ *a posteriori* da classe k dado $\mathbf{X} = x$.

O objetivo é obter um classificador $\hat{F}: X \to C$.

Se $p(\mathbf{x}|k)$ e π_k são conhecidos, então $p(k|\mathbf{x})$ pode ser obtido através da fórmula de *Bayes*:

$$p(k|\mathbf{x}) = \frac{\pi_k p(k|\mathbf{x})}{\sum_{j=1}^{J} \pi_j p(\mathbf{x}|j)}. \tag{4}$$

Seja $L(k,l)$ a perda que se tem por tomar a decisão l, quando a classe verdadeira é k. Se os erros de classificação são homogêneos nas classes, isto é, se uma decisão errada é igualmente ruim, independente de em qual classe o objeto foi erroneamente alocado, a seguinte função perda pode ser utilizada:

$$L(k,l) = \begin{cases} 0 & se\ l = k\ (decisão\ correta) \\ 1 & se\ l \neq k\ (decisão\ errada) \end{cases} \tag{5}$$

Quando $p(\mathbf{x}|k)$ e π_k são conhecidos, o melhor classificador, considerando a perda (5), é dado na proposição abaixo:

Proposição 2.1 - *A regra de classificação que minimiza o risco de Bayes sob a perda (5) é*

$$F_B(\mathbf{X}) = k \ \ se \ \ p(k|\mathbf{x}) = \max_{j=1,2,\ldots,J} p(j|\mathbf{x}) \tag{6}$$

A regra de *Bayes* diz que se deve alocar um objeto na classe com maior probabilidade *a posteriori*. Se duas classes atingem o mesmo valor de $p(k|\mathbf{x})$, o objeto pode ser alocado em qualquer uma delas arbitrariamente.

Para a aplicação com dados simulados (seção 4), comparamos o desempenho dos algoritmos de *boosting* com os métodos tradicionais, e ainda ao desempenho do classificador de *Bayes*, que pode ser construído neste caso específico.

3. *Boosting*

O método conhecido como *boosting* nasceu na área de computação (de maneira geral, uma comunidade conhecida pelo nome *machine learning*, aprendizado de máquina, numa tradução literal). Dentro dessa comunidade, foi proposto um problema teórico chamado de problema de *boosting*, que pode ser informalmente exposto da seguinte maneira: "Suponha que existe um método de classificação que é ligeiramente melhor do que uma escolha aleatória, para qualquer

distribuição em X. Esse método é chamado de *weak learner*, ou classificador fraco. A existência de um classificador fraco implica na existência de um classificador forte (*strong learner*), com erro pequeno sobre todo o espaço X?"

Em Estatística, isso eqüivale a perguntar se, dado um método razoável de estimação, é possível se obter um método próximo de ótimo.

Este problema foi resolvido por Schapire (1990), que apresentou um algoritmo que transformava um classificador fraco em um forte. A partir de então, foram desenvolvidos vários algoritmos dentro do contexto de *boosting*. Um dos mais recentes e bem-sucedidos deles é o algoritmo conhecido como *AdaBoost*, apresentado em Freund e Schapire (1997). Este nome vem de *Adaptative Boosting*, e é oriundo do fato de que *AdaBoost* gera em cada passo (de forma determinística, mas adaptativa) uma distribuição sobre as observações da amostra, dando maior peso (maior probabilidade de estar na amostra perturbada) às observações classificadas incorretamente no passo anterior.

Breiman (1998) chamou os algoritmos do tipo *boosting*, em particular do tipo apresentado por Freund e Schapire (1997), de *arcing*, um acrônimo para Adaptatively Resampling and Combining. Breiman também criou um algoritmo do tipo *arcing*, que mostrou ter desempenho tão bom quanto *AdaBoost*. Isso mostrou que o funcionamento do algoritmo *AdaBoost* não estava relacionado diretamente à sua forma específica, e que existia uma classe de algoritmos que operava daquela maneira. Friedman, Hastie e Tibshirani (2000) mudaram totalmente o modo como *boosting* é visto, pelo menos na comunidade estatística. Eles colocaram *boosting* como um ajuste de um modelo aditivo na escala logística, usando máxima verossimilhança da Bernoulli como critério. Ademais, sugeriram uma aproximação mais direta, o que levou ao algoritmo *LogitBoost*, um algoritmo para ajustar uma regressão logística aditiva que dá resultados praticamente idênticos ao *AdaBoost* de Freund e Schapire.

Mais recentemente, notou-se que, se um algoritmo de *boosting* for executado por um tempo (número de iterações) muito grande, da ordem de dezenas de milhares, isso ocasionará *overfitting*. Friedman, Hastie e Tibshirani (2000) dão um exemplo em que isso ocorre. Algumas abordagens para este problema foram tentadas Jiang (2000a) mostrou que, sob certas condições de regularidade, o algoritmo *AdaBoost* é consistente em processo (process *consistent*), no sentido de que, durante o treinamento, ele gera uma seqüência de classificadores com erro que converge para o erro do classificador (regra) de *Bayes*. Lugosi e Vayatis (2004) mostraram um resultado importante de consistência para algoritmos de *boosting* com regularização.

Apresentamos, primeiramente duas versões do algoritmo *AdaBoost* e o algoritmo *LogitBoost* no caso em que existem $J=2$ classes. Após isso, mostramos o caso em que há $J>2$ classes.

3.1. Caso de duas classes

O algoritmo, discreto pode ser descrito da seguinte maneira. Considere o conjunto $L=(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_N,y_N)$, onde as classes estão rotuladas $\{-1,1\}$.

Defina $F(\mathbf{x})=\sum_1^M c_m f_m(\mathbf{x})$, onde f_m é um classificador-base que retorna valores $\{-1,1\}$, os valores c_m são constantes e a predição correspondente é o sinal de $F(\mathbf{x})$, ou seja, $sign(F(\mathbf{x}))$. O algoritmo *AdaBoost* ajusta classificadores-base f_m em amostras reponderadas do conjunto de treinamento, dando maior peso, ou ponderação aos casos que são classificados erroneamente. Os pesos são ajustados, adptativamente, em cada iteração, e o classificador final é uma combinação linear dos classificadores f_m.

Algoritmo 3.1 (*AdaBoost* discreto)

1. *Inicialize os pesos* $w_i=1/N, i=1,2,\ldots,N$
2. *Repita para* $m=1,2,\ldots,M$:

   (a) Ajuste o classificador $f_m(\mathbf{x})\in\{-1,1\}$ *usando os pesos* w_i *nos dados de treinamento*

   (b) Calcule $\varepsilon_m=E_w[I_{(y\neq f_m(\mathbf{x}))}]=\frac{1}{N}\sum_{i=1}^N w_i I(y_i=f_m(\mathbf{x}_i))$,

   $c_m=\log((1-\varepsilon_m)/\varepsilon_m)$

   (c) Faça $w_i\leftarrow w_i\exp[c_m I_{(y_i\neq f_m(\mathbf{x}_i))}], i=1,2,\ldots,N$ *e renormalize para que* $\sum_i w_i=1$

3. *Saia com o classificador final sign* $(F(\mathbf{x}))=sign\left(\sum_{m=1}^M c_m f_m(\mathbf{x})\right)$

No algoritmo acima, E_w representa a média ponderada (esperança no conjunto de treinamento) com pesos $w=(w_1,\ldots,w_N)$. Em cada iteração m, os ε_m são calculados com base no vetor de pesos w e no acerto do classificador-base f_m. Nos algoritmos de *boosting*, os classificadores base são ajustados da maneira usual com que seriam ajustados se fossem usados normalmente, mas os dados com que eles são ajustados (os dados de treinamento) são gerados, a princípio, de uma distribuição de probabilidade baseado nos pesos w (na maioria dos

algoritmos não há a amostragem aleatória em si; os pesos são incorporados diretamente no algoritmo).

Outra versão do algoritmo *AdaBoost* é o *AdaBoost* real, onde o classificador-base f_m é um classificador que retorna a estimativa $p_m(\mathbf{x}) = \hat{P}_w(y = 1|\mathbf{x}) \in [0,1]$.

O algoritmo usa essas estimativas para construir as contribuições reais $f_m(\mathbf{x})$.

Algoritmo 3.2 (*AdaBoost* real)

1. *Inicialize os pesos* $w_i = 1/N, i = 1,2,\ldots,N$
2. *Repita para* $m = 1,2,\ldots,M$ *:*

   (a) Ajuste o classificador para se obter uma estimativa de probalidade de classe $p_m(\mathbf{x}) = \hat{P}_w(y = 1|\mathbf{x}) \in \{0,1\}$ *usando os pesos* w_i *nos dados de treinamento*

   (b) Faça $f_m = \frac{1}{2}\log p_m(\mathbf{x})/(1 - p_m(\mathbf{x})) \in \Re$,

   (c) Faça $w_i \leftarrow w_i \exp[-y_i f_m(\mathbf{x}_i)], i = 1,2,\ldots,N$ *e renormalize para que* $\sum_i w_i = 1$
3. *Saia com o classificador final sign* $\left(\sum_{m=1}^{M} f_m(\mathbf{x})\right)$

A seguir apresentamos o algoritmo *LogitBoost*, um algoritmo de *boosting* desenvolvido por Friedman, Hastie e Tibshirani (2000), o qual é baseado num modelo logístico aditivo. Neste algoritmo, a resposta $y^* = \frac{y+1}{2}$ é um valor em *{0, 1}*.

Algoritmo 3.3 (*LogitBoost*)

1. *Inicialize os pesos* $w_i = 1/N, i = 1,2,\ldots,N, F(\mathbf{x}) = 0$ *e estimativas de probabilidades* $p(\mathbf{x}_i) = \frac{1}{2}$.
2. *Repita para* $m = 1,2,\ldots,M$ *:*

   (a) Calcule

$$z_i = \frac{y_i^* - p(\mathbf{x}_i)}{p(\mathbf{x}_i)(1 - p(\mathbf{x}_i))}$$

$$w_i = p(\mathbf{x}_i)(1 - p(\mathbf{x}_i))$$

   (b) Ajuste a função classificador $f_m(\mathbf{x})$ *fazendo a regressão de* z_i *em* $\mathbf{x}_i$ *por mínimos quadrados ponderados usando os pesos* w_i

(c) Atualize $F(\mathbf{x}) \leftarrow F(\mathbf{x}) + \frac{1}{2} f_m(\mathbf{x})$ *e* $p(\mathbf{x}) \leftarrow (e^{F(\mathbf{x})}) / (e^{F(\mathbf{x})} + e^{-F(\mathbf{x})})$

3. *Saia com o classificador final* $sign(F(\mathbf{x})) = sign\left(\sum_{m=1}^{M} c_m f_m(\mathbf{x})\right)$

3.2. Caso de *J* classes

No caso de J classes, considere J respostas y_j para o problema de J classes, cada uma assumindo valores em $\{-1,1\}$, ou seja, para cada objeto i, há J respostas y_{ij} que identificam se o objeto pertence à classe j $(y_{ij} = 1)$ ou não $(y_{ij} = -1)$. As classes são consideradas auto-exclusivas.

As versões para J classes do *AdaBoost* (*AdaBoost*.MH) e *LogitBoost* (*Logit-Boost*.MH) são apresentadas em Friedman, Hastie e Tibshirani (2000).

Algoritmo 3.4 (*AdaBoost*.MH)

1. *Expanda as* N *observações originais em* $N \times J$ *pares dados por :*
   $((\mathbf{x}_i, 1), y_{i1}), ((\mathbf{x}_i, 2), y_{i2}), \ldots, ((\mathbf{x}_i, J), y_{iJ}), i = 1, \ldots, N$. *A resposta* y_{ij} *é a resposta para a classe j, observação i.*
2. *Execute o algoritmo AdaBoost Real no conjunto aumentado, produzindo uma função* $F : X \times (1, \ldots, J) \rightarrow \Re$; $F(x, j) = \sum_m f_m(x, j)$.
3. *Saia com o classificador* $\arg\max_j F(x, j)$.

Na prática, a implementação deste algoritmo envolve executar o algoritmo *Ada-Boost* Real em J conjuntos de dados. No j-ésimo conjunto de dados, as respostas usadas devem ser $y_{ij}, i = 1, \ldots, N$. Isso é equivalente a executar J processos de classificação, onde em cada um deles ajusta-se um modelo de uma classe contra as outras. Os classificadores resultantes são $F(x, j)$, para $j = 1, 2, \ldots, J$. Para classificar um objeto representado pelo vetor x, toma-se a classe cujo classificador assume o maior valor, ou seja, o argumento que maximiza as $F(x, j)$ em J.

Para apresentar o algoritmo *LogitBoost* para J classes, definiremos a transformação logística simétrica múltipla.

Definição 3.1- *Para um problema de classificação em J classes, seja $p_j(x) = P(y_j = 1|x)$. Definimos a transformação logística simétrica múltipla por*

$$F_j(\mathbf{x}) = \log p_j(\mathbf{x}) - \frac{1}{J}\sum_{k=1}^{J} \log p_k(\mathbf{x}). \tag{7}$$

Equivalentemente,

$$p_j(\mathbf{x}) = \frac{e^{F_j(\mathbf{x})}}{\sum_{k=1}^{J} e^{F_k(\mathbf{x})}}, \quad \sum_{k=1}^{J} F_k(\mathbf{x}) = 0. \tag{8}$$

A seguir apresentamos o algoritmo *LogitBoost* para J classes, que é uma generalização natural do *LogitBoost* para duas classes.

Algoritmo 3.5 (*LogitBoost* (*J* classes))

1. *Inicialize os pesos* $w_{ij} = 1/N$, $i = 1,2,\ldots,N$, $j = 1,\ldots,J$, $F_j(\mathbf{x}) = 0$ *e estimativas de probabilidades* $p_j(\mathbf{x}_i) = \frac{1}{J} \forall j$
2. *Repita para* $m = 1,2,\ldots,M$ *:*

   (a) Repita para $j = 1,\ldots,J$ *:*

   i. Calcule

$$z_{ij} = \frac{y_{ij}^* - p_j(\mathbf{x}_i)}{p_j(\mathbf{x}_i)(1 - p_j(\mathbf{x}_i))}$$

$$w_{ij} = p_j(\mathbf{x}_i)(1 - p_j(\mathbf{x}_i))$$

   ii. Ajuste a função $f_{mj}(\mathbf{x})$ *fazendo a regressão de* z_{ij} *em* $\mathbf{x}_i$ *por mínimos quadrados ponderados usando os pesos* w_{ij}

   (b) Faça $f_{mj}(\mathbf{x}) \leftarrow \frac{J-1}{J}\left(f_{mj}(\mathbf{x}) - \frac{1}{J}\sum_{k=1}^{J} f_{mk}(\mathbf{x})\right)$, e atualize $F_j(\mathbf{x}) \leftarrow F_j(\mathbf{x}) + f_{mj}(\mathbf{x})$

   c) Atualize $p_j(\mathbf{x})$ *via (8)*

3. *Saia com o classificador final* $\arg\max_j F_j(\mathbf{x})$

3.3. O que *Boosting* faz

Os algoritmos de *boosting* ajustam modelos aditivos através da otimização de um critério, ou funcional de custo. Cada base da expansão é um classificador simples que pode ser escolhido pelo usuário. Comparando com um modelo de regressão logística, por exemplo, o critério utilizado é a verossimilhança da binomial, e o modelo aditivo é linear. No algoritmo *AdaBoost* discreto, pode-se mostrar que o critério que está sendo otimizado é $e^{-yF(\mathbf{x})}$. A quantidade $yF(\mathbf{x})$ é negativa se, e somente se, o classificador F errou a predição de y. O critério $e^{-yF(\mathbf{x})}$ é, portanto, um funcional de custo suave baseado na medida $yF(\mathbf{x})$. A razão para utilizar um funcional deste tipo é a tratabilidade matemática. Ainda é possível mostrar que esse critério é uma aproximação da log-verossimilhança da binomial.

Existem muitos algoritmos de *boosting* disponíveis, que utilizam funcionais de custo diferentes e métodos diferentes de otimização. O algoritmo *LogitBoost*, por exemplo, otimiza a log-verossimilhança da binomial através de um algoritmo do tipo *Newton-Rhapson*. A interpretação desses algoritmos é clara: eles são combinações lineares de classificadores simples, os quais são construídos de maneira a dar mais peso aos erros cometidos. Métodos lineares, contudo, não são beneficiados por *boosting* Breiman (1998). Isso ocorre porque os métodos lineares são estáveis, no sentido de que pequenas pertubações nos dados não produzem grandes pertubações nos classificadores gerados, e, portanto, o método de *boosting*, quando aplicado a esses métodos, apenas combina vários classificadores muito parecidos, produzindo um classificador final muito parecido ao original.

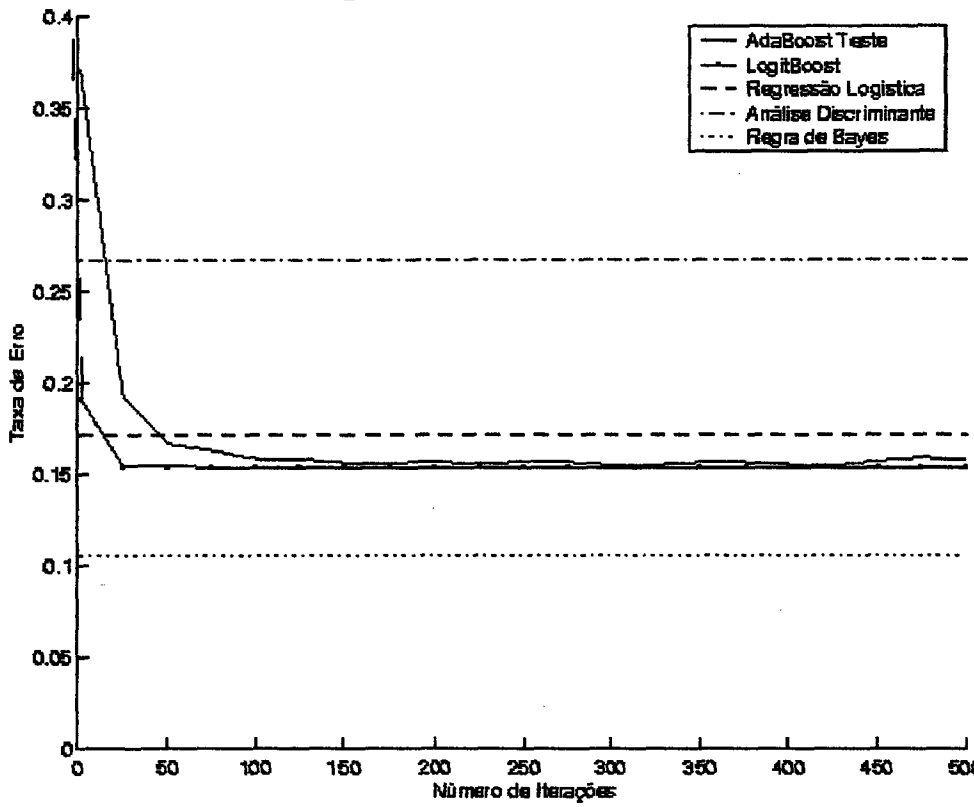
4. Simulação

Nesta seção damos um exemplo de aplicação do algoritmo *AdaBoost* Real em um conjunto de dados simulados, descrito em Breiman (1998) (conjunto denominado *three-norm*).

Os dados possuem duas classes e são gerados da seguinte maneira: os dados da classe 1 são gerados com probabilidade $\frac{1}{2}$ de uma distribuição normal multivariada com média $(a, a, \ldots a)$, e com probabilidade $\frac{1}{2}$ de uma distribuição normal com média $(-a,-a, \ldots,-a)$, com matrizes de variância identidades nos dois casos, com $a = 2/(20)^{1/2}$. Os dados da classe 2 são gerados de uma distribuição normal mul-tivariada com média $(a,-a, a,-a,\ldots, a)$ e matriz de variância identidade.

O vetor de variáveis explanatórias tem 20 variáveis. Foram simulados 1000 valores para amostra de treinamento e 1000 valores para amostra de teste. Executamos os algoritmos *AdaBoost* Real e *LogitBoost*, com o número de iterações variando de 1 a 500, de 25 em 25. Todo o procedimento (simulação dos dados e execução dos algoritmos) foi executado 100 vezes, e foram calculadas as estimativas das taxas de erro para cada número de iterações como uma média das taxas de erro no conjunto de teste nas 100 repetições. Os modelos baseados no algoritmo *AdaBoost Real* foram ajustados usando-se o *software R* versão 1.7.1, com o pacote *gbm*. Os classificadores-base utilizados nesta implementação são árvores de classificação. Os modelos baseados no algoritmo *LogitBoost* foram implementados no mesmo software. Para efeito de comparação, foram ajustados modelos tradicionais de classificação, a saber, regressão logística e análise discriminante linear. Os modelos foram ajustados sem seleção de variáveis. Para o ajuste desses modelos foi utilizado o software SAS versão V8. Ambos os modelos foram ajustados em cada uma das amostras, em cada uma das 100 iterações. As taxas de erro em amostras-teste foram calculadas em cada iteração, e foi utilizada a taxa de erro média para comparação.

Gráfico 1 – Estimativas das taxas de erro



O Gráfico 1 sumariza os resultados obtidos. A linha pontilhada mais abaixo, paralela ao eixo das abscissas, representa o erro do classificador de *Bayes*, que neste exemplo é de 0, 105. Esta é, a princípio, a menor taxa de erro atingível. Todas as outras linhas representam taxas de erro médias estimadas usando amostras-teste. A linha tracejada representa a taxa de erro média dos modelos de regressão logística. A linha de traços e pontos representa a taxa de erro média dos modelos de análise discriminante linear. A linha sólida representa a taxa de erro média dos modelos construídos pelo algoritmo *AdaBoost*, de acordo com o número de funções. Note que esta taxa de erro atinge à estabilidade após, aproximadamente, 75 iterações (no de funções) do algoritmo. A linha sólida com pontos sobrepostos representa a taxa de erro média do algoritmo *LogitBoost*.

Nota-se, através do gráfico, que o desempenho do algoritmo *AdaBoost* é superior ao dos métodos regressão logística e análise discriminante linear. No caso da análise discriminante linear, o ganho é bem maior, e no caso da regressão logística, a diferença não é tão grande, mas o *AdaBoost* ainda é melhor. O algoritmo *LogitBoost* é ainda ligeiramente superior ao algoritmo *AdaBoost*.

Uma questão de interesse prático é o critério de parada para o algoritmo. Na maioria das vezes, executar o algoritmo por centenas de iterações não irá causar *overfitting*. Esse problema só surge, em geral, com números de iterações muito maiores, da ordem de milhares. Assim, se for utilizada uma amostra-teste, pode-se parar o algoritmo quando o erro na amostra-teste atinge um patamar de estabilidade.

Se não houver amostra-teste disponível, pode-se usar validação cruzada e parar quando a estimativa do erro por validação cruzada deixar de cair. Para evitar este tipo de critério, pode-se usar algoritmos de *boosting* regularizados. Estes algoritmos baseiam-se em otimizar funcionais de custo adicionados de um termo de penalização. Referências para algoritmos deste tipo são Mason (1999), Jiang (2000b), Evgenious (2002).

5. Aplicação em dados reais

A aplicação em dados reais deste trabalho foi feita em dados de uma rede de lojas de varejo. O problema consiste em classificar clientes em grupos que definem seus perfis. Cada cliente pode pertencer a um de dez grupos. O conjunto de dados possui, para cada cliente, observações em variáveis de interesse, assim como a classe (grupo) correspondente. O conjunto original continha 184 variáveis de vários tipos. Foram excluídas 14 variáveis que ou não se encaixavam no tipo de variável que pudesse ser utilizada pelo algoritmo de *boosting*, ou que

tivessem muitos dados faltantes (na implementação de *boosting* usada, só é possível utilizar variáveis contínuas, de maneira que excluímos as variáveis de qualquer outro tipo). O tamanho da base original era de aproximadamente meio milhão de clientes. Desta base foi selecionada uma amostra aleatória simples de tamanho aproximadamente 50 000. Esta amostra foi dividida em duas: uma de tamanho aproximadamente 34 000, para treinamento, e outra, de tamanho aproximadamente 16 000, para teste.

Utilizando a amostra de treinamento, ajustamos inicialmente modelos de análise discriminante linear e quadrática. A análise discriminante linear foi superior à quadrática, de maneira que só reportamos os resultados para a linear. Com este mesmo conjunto de dados de treinamento, aplicamos o algoritmo 3.4, *AdaBoost*.MH, variando o número de funções de 25 a 400 e calculando, para cada passo, a estimativa da taxa de erro na amostra-teste. A partir de 300 funções não houve melhora nesta taxa. Os modelos foram ajustados seguindo as mesmas considerações da seção 4.

Os resultados obtidos estão dispostos na Tabela 1.

Tabela 1 - Taxas de erro estimadas com amostra-teste

Método	Taxa de Erro (Teste)
Análise Discriminante Linear	0,1549
AdaBoost com 25 função	0,1210
AdaBoost com 50 funções	0,0991
AdaBoost com 75 funções	0,0880
AdaBoost com 100 funções	0,0802
AdaBoost com 200 funções	0,0712
AdaBoost com 300 funções	0,0662
AdaBoost com 400 funções	0,0662

Como a tabela evidencia, o algoritmo *AdaBoost* foi bastante superior ao método tradicional de análise discriminante linear. Com uma combinação de apenas 25 iterações, essa superioridade já é evidente, e com 300 iterações, a melhora é notável.

Aumentar o número de iterações acima de 300 não trouxe melhora, de maneira que um modelo final poderia conter este número de funções. O método de análise discriminante supõe distribuições *a priori* normais, e produz uma fronteira de decisão linear. A superioridade dos métodos de *boosting* se deve à sua capacidade de aproximar funções mais complexas, e, portanto, produzir fronteiras de decisão que proporcionam classificadores mais poderosos.

O custo computacional de executar o algoritmo *AdaBoost*.MH num conjunto de dados razoavelmente grande e complexo como este não foi muito superior ao custo computacional de executar o modelo de análise discriminante. O ganho obtido em precisão, não obstante, foi substancial. De fato, com poucas iterações, o custo computacional do *AdaBoost* é muito similar ao de Análise Discriminante. Conforme o número de funções cresce, entretanto, o custo computacional aumenta.

6. Considerações finais

Apresentamos uma metodologia pouco conhecida na comunidade estática em geral e, em especial, no Brasil. Nas aplicações apresentadas, esta metodologia, chamada *boosting* e apresentada em dois algoritmos diferentes, *AdaBoost* e *Logit*-Boost, apresentou resultados muito bons em comparação aos métodos estatísticos tradicionais, como regressão logística e análise discriminante. *Boosting* pode ser implementado, facilmente, a partir de um classificador razoável, transformando-o em um comitê de classificação com desempenho muito superior ao classificador original. Seu custo computacional não é elevado, apesar de ser superior ao dos métodos tradicionais, e o método pode lidar com grandes quantidades de informação, o que é o caso em muitas aplicações atuais.

A utilização de metodologias modernas e eficientes, como *boosting*, pode trazer muitos benefícios em diversas áreas onde o problema de classificação aparece.

Referências bibliográficas

AGRESTI, A. *An Introduction to Categorical Data Analysis*. [S.l.]: John Wiley and Sons, Inc, 1996.

BREIMAN, L. et al. *Classification and Regression Trees*. [S.l.]: Chapman-Hall/CRC, 1984.

BREIMAN, L. Arcing classifiers. *The Annals of Statistics*, v. 26, n. 3, p. 801–849, 1998.

BREIMAN, L. Some infinity theory for predictor ensembles. *Technical Report 577, Statistics Department, University of California*, 2000.

EVGENIOUS, T. et al. Regularization and statistical learning theory for data analysis. *Computational Statistics and Data Analysis*, v. 38, p. 421–432, 2002.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, v. 55, p. 119–139, 1997.

FRIEDMAN, J. H.; HASTIE, T.; TIBSHIRANI, R. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, v. 28, p. 337–407, 2000.

HOSMER, D.; LEMESHOW, S. *Applied Logistic Regression*. [S.l.]: John Wiley and Sons, Inc, 1989.

JIANG, W. Process consistency for adaboost. *Technical Report 05*, Department of Statistics, Northwestern University, 2000.

JIANG, W. Is regularization unnecessary for boosting ? *Technical Report 04*, Department of Statistics, Northwestern University, 2000a.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. [S.l.]: Prentice Hall, New York, NY, 2002.

LUGOSI, G.; VAYATIS, N. On the bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, v. 32-1, 2004. To Appear.

MASON, L. et al. Functional gradient techniques for combining hypotheses. *Advances in Large Margin Classifiers*, 1999. MIT Press.

RIPLEY, B. D. *Pattern Recognition and Neural Networks*. [S.l.]: Cambridge University Press, 1996.

SCHAPIRE, R. E. The strength of weak learnability. *Machine Learning*, v. 5, p. 197–227, 1990.

Abstract

The traditional statistical methods used in classification are, basically, logistic regression and discriminant analysis. Recently, other methods have been used in applications, e.g., classification trees and neural networks. The traditional methods are computationally cheap, but lack flexibility, while methods such as neural networks are flexible, but computationally expensive and complicated. The method presented in this work, boosting, combines simple base classifiers, increasing at each step the weights of the observations that were previously misclassified.

This work presents the method of boosting, as well as two applications. One is on simulated data, and the other on real data. The real-data application consists of classifying clients of a retail store into previously defined groups, which define the clients' profiles. The results obtained with boosting in both applications are compared to the results of the traditional methods.

Estimação pontual de parâmetros de distribuições de probabilidade utilizando algoritmos genéticos

M.E. Martínez [*]
A.B. Cheung [**]
P.B. Cheung [***]

Resumo

Um problema em quase todas as áreas da pesquisa científica é a estimação pontual de parâmetros de uma distribuição de probabilidade. Um método muito utilizado para tal estimação é o Método de Máxima Verossimilhança. Este método consiste na maximização da função de verossimilhança. Isto geralmente é feito pela diferenciação do logaritmo desta função com relação a um parâmetro θ. No entanto, em muitas situações reais é necessário resolver esta função através de métodos iterativos, principalmente quando a dimensão do espaço de parâmetros é grande. Neste caso, algumas dificuldades em relação à utilização destes métodos podem ser citadas: necessidade de derivar o logaritmo da função de verossimilhança, grande número de parâmetros, etc. Portanto, o objetivo deste trabalho é utilizar Algoritmos Genéticos simples - AGs para a estimação pontual de parâmetros de distribuições de probabilidade. Uma das vantagens dos AGs é que as derivadas do logaritmo da função de verossimilhança não necessitam ser calculadas.

Palavras-chave: Estimadores de máxima verossimilhança, Algoritmos Genéticos simples, distribuições de probabilidade.

---

[*] Endereços para correspondências: Departamento de Estatística - ICET – UFMT, Av. Fernando Corrêa da Costa. S/N., Coxipó Cuiabá - MT – Brasil CEP: 78060 – 900, E-mail: dest@cpd.ufmt.br.

[**] LaMEM-EESC-USP, Av. Trabalhador São Carlense, 400, São Carlos - SP – Brasil CEP: 13566-590, Telefone: (16) 3361-4536, E-mail: acheung@sc.usp.br

[***] E-mail: peter.cheung@bordeaux.cemagref.fr.

1. Introdução e objetivo

Um problema em quase todas as áreas da pesquisa científica é a estimação pontual de parâmetros de uma distribuição de probabilidade (modelo estatístico) relacionada a um conjunto de dados experimentais. Um método muito utilizado para tal estimação é o Método de Máxima Verossimilhança, que consiste na maximização da função de verossimilhança KALBFLEISCH (1985). O princípio da verossimilhança afirma que devemos escolher aquele valor do parâmetro desconhecido que maximiza a probabilidade de obter a amostra particular observada, ou seja, o valor que torna aquela amostra a "mais provável". O uso desse princípio conduz a um método de estimação pelo qual se obtém os chamados estimadores de máxima verossimilhança MORETIN, BUSSAB (2003). Para obter estes estimadores, em geral, requerem-se cálculos de derivadas, que em muitas situações reais são de difícil resolução analítica, recorre-se então a aplicação de métodos de otimização não convencionais para a estimação destes parâmetros. Um destes métodos podem ser os Algoritmos Genéticos simples.

Algoritmos Genéticos simples - AGs são métodos de busca estocástica baseados no processo de seleção natural das espécies. A idéia é a evolução de uma população de indivíduos, de modo que melhore a adequação média dos indivíduos que formam essa população em relação ao ambiente a que ela está submetida, GEN e CHENG (1996).

Os AGs foram inicialmente desenvolvidos por John Holland, em 1975, porém quem os popularizou foi um dos seus alunos, GOLDBERG (1989). Os objetivos destes pesquisadores foram investigar e projetar sistemas artificiais, análogos aos mecanismos naturais das espécies.

Basicamente, os Algoritmos Genéticos transformam uma população de indivíduos, cada um com um valor associado de adaptabilidade, chamado de aptidão, numa nova geração de indivíduos usando os princípios Darwianos de reprodução e sobrevivência dos mais aptos, pela aplicação de operações genéticas tais como recombinação e mutação, DA SILVA (2002).

Portanto, o objetivo deste trabalho é utilizar os AGs para a estimação pontual de parâmetros de distribuições de probabilidade e investigar a adequação de tal técnica no problema em questão. Três estudos de caso foram considerados, o primeiro apresenta uma distribuição de probabilidade exponencial, do qual sua solução analítica é conhecida. Já no segundo e terceiro caso a distribuição de probabilidade Weibull e de Birnbaum-Saunders foram consideradas, nas quais somente suas soluções numéricas são conhecidas. Estas soluções serviram de referência para serem comparadas com as soluções produzidas pelos AGs como forma de validar e verificar a técnica.

A utilização dos AGs, na estimação pontual de parâmetros de distribuições de probabilidades é justificada por várias vantagens, que são descritas sucintamente a seguir:

- otimização de problemas com muitas variáveis e um espaço de solução de dimensão elevada;
- não há necessidade de se calcular derivadas do logaritmo da função de verossimilhança;
- são versáteis, no sentido que o mecanismo de evolução é separado da representação particular do problema considerado;
- são resistentes aos ótimos locais;
- varrem todo espaço de busca; e
- adaptam-se bem a computadores paralelos.

Desta forma, os AGs vêm se tornando uma das técnicas computacionais mais robustas e poderosas em todas as áreas do conhecimento. Descrições detalhadas da sistemática desta técnica a das vantagens apresentadas podem ser encontradas em literaturas especializadas como os livros texto de GOLDBERG (1989), MICHALEWICZ (1992) e GEN e CHENG (1996). Além disso, atualmente existem diversos trabalhos publicados em que os Algoritmos Genéticos são aplicados a problemas estatísticos clássicos e bayesianos, destacando a seguinte literatura: KNAN, et al. (2002), PITTMAN, MURTHY (2000), PELIKAN, et al. (1999), PETROVSKI, et al., (1998), CHATTERJE, et al. (1996).

2. Fundamentação teórica

2.1. Estimadores de máxima verossimilhança

Um dos melhores métodos para obter estimadores pontuais de um parâmetro é o método da máxima verossimilhança. O estimador de máxima verossimilhança (EMV) de um parâmetro θ é o valor de θ que maximiza a função de verossimilhança $L(\theta)$ dada pela seguinte equação:

$$L(\theta) = \prod_{i=1}^{n} f(y_i, \theta) \qquad (1)$$

onde $f(y_i, \theta)$ é a função densidade de probabilidade discreta ou contínua. Observe que $f(y_i, \theta)$ pode ter mais de um parâmetro, CORDEIRO (1992).

O EMV de θ é usualmente denotado por $\hat{\theta}$, sendo baseado em uma amostra aleatória $y_1, y_2, \cdots, y_n$, geralmente representada por E. Assim, o estimador do parâmetro θ ($\hat{\theta}$) que melhor explica os dados E, é o valor de θ que maximiza a probabilidade dos E sob o modelo estatístico ENGINEERING STATISTICS HANDBOOK (2002).

Observe que $L(\theta)$ são produtos de termos, o que facilita para trabalhar com logaritmos, pois o logaritmo do produto é a soma do logaritmo dos fatores. Assim o logaritmo da função de verossimilhança é naturalmente o logaritmo de $L(\theta)$, isto é:

$$l(\theta) = \ln[L(\theta)]. \qquad (2)$$

Assim, o valor de θ que maximiza $L(\theta)$ do mesmo modo maximiza a $l(\theta)$. Cabe destacar que, na prática, em geral, é mais fácil trabalhar com o logaritmo da função de verossimilhança. Assim o EMV $\hat{\theta}$ é o valor de θ que maximiza o logaritmo da função de verossimilhança.

Portanto, para o cálculo de $\hat{\theta}$, é necessário maximizar $l(\theta)$ para todos os possíveis valores de θ. Isto geralmente é feito pela diferenciação do $l(\theta)$ com relação a θ. Fazendo a derivada igual a zero, encontra-se $\hat{\theta}$.

A equação (2) em alguns casos especiais pode ser resolvida algebricamente e o valor de $\hat{\theta}$ pode ser obtido diretamente da formulação. No entanto, em muitas situações é necessário resolver esta função através de métodos iterativos ou métodos numéricos clássicos. Os métodos iterativos de cálculo dos EMV, são bastante utilizados na prática e, em geral, mostram-se imprescindíveis quando a dimensão do espaço de parâmetros é grande, CORDEIRO (1992). Porém, algumas dificuldades em relação à utilização destes métodos podem ser citadas: necessidade de derivar $l(\theta)$, grande número de parâmetros e necessidade da verificação de convergência. Além disso, os métodos numéricos clássicos podem não garantir a obtenção do ótimo global. Portanto, para minimizar e/ou evitar estas dificuldades, podemos utilizar os *métodos evolucionários* conhecidos como AGs.

2.2. Fundamentos dos Algoritmos Genéticos simples - AGs

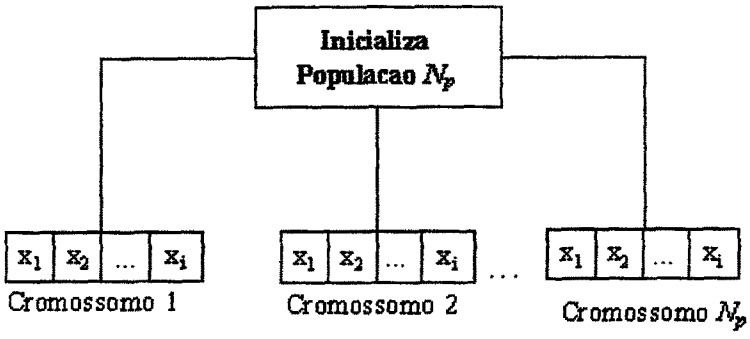
Os Ags possuem vários elementos que os caracterizam. Essas características podem ser utilizadas baseando-se em diferentes especificações, dependendo do problema a ser tratado. Nesta seção são descritos sucintamente alguns conceitos relacionados com os AGs, a fim de facilitar o entendimento deste trabalho.

2.2.1. População inicial

Um Algoritmo Genético começa com uma população inicial de indivíduos com as características representadas pelos cromossomos. O cromossomo é um vetor unidimensional (*string*) cujos valores (variáveis de decisão) representam características (genes) da possível solução do problema e podem ser codificadas através da representação binária ou real, entre outras. A codificação depende do problema que está sendo tratado, por exemplo, em problemas combinatoriais a codificação mais apropriada é a binária. Já em problemas que contemplam funções contínuas, a codificação real é a mais adequada. A partir dessa escolha inicial, os vetores soluções (indivíduos) são gerados aleatoriamente através de rotinas computacionais existentes nos compiladores. Cabe observar que se uma pequena população for gerada, alguns locais no espaço de busca podem não estar sendo representados.

Os AGs diferem das técnicas de busca convencionais, pois iniciam o procedimento computacional gerando um conjunto inicial de soluções possíveis aleatoriamente. Este conjunto é denominado "população" (Figura 1). Vários conjuntos de soluções (populações) são produzidos ao longo do processo iterativo. Cada vetor-solução (indivíduo) pertencente a estes conjuntos (população), intuitivamente denominado "cromossomo", constitui uma representação completa da solução para o problema. Estes cromossomos tendem a evoluir-se gradativamente através de sucessivas iterações, produzindo novas "gerações" de indivíduos melhores, em termos de uma função de aptidão, esta função será definida na Seção 2.2.2.

Figura 1- Representação da população com seus indivíduos (cromossomos)



2.2.2. Função de aptidão

A avaliação da população é realizada pela função de aptidão, que representa uma medida que avalia o vetor-solução (a qualidade de cada indivíduo na população) durante o processo de

evolução. Seu valor servirá como referência (qualidade da resposta) na classificação dessas soluções, indicando a chance de alguns vetores solução estarem presentes na população da próxima geração. A classificação pode ser feita através de uma ordenação das soluções de forma decrescente ou crescente em relação às suas aptidões, representando, respectivamente, o problema de maximização ou minimização. Assim, para problemas de otimização a função de aptidão está intimamente ligada à função objetivo DA SILVA (2002). Uma função-objetivo é um modelo matemático que precisa ser maximizado ou minimizado, considerando um número finito de restrições ou condições SIERKSMA (1996). Porém, os valores da função-objetivo nem sempre são adequados para serem utilizados como valores de aptidão. Pois, a função-objetivo pode fornecer valores negativos ou fornecer valores muito próximos entre os indivíduos presentes na população. No caso de valores negativos, alguns métodos de seleção não funcionam, por exemplo, o Roda Roleta. Em se tratando de valores muito próximos entre os indivíduos pertencentes à população, aumenta-se a possibilidade de uma solução ruim se encontrar na próxima geração. Com isso, deve-se buscar outro procedimento de seleção, por exemplo, seleção por torneio ou aplicar algum procedimento de parametrização dado pela equação (3):

$$f_k = 2\left(\frac{N_p - k}{N_p - 1}\right) \tag{3}$$

onde N_p é o tamanho da população e k é o índice do cromossomo na população. O tamanho da população indica o número de indivíduos em cada população, normalmente constante durante a evolução.

No método de Ordenamento Linear BAKER (1987) a aptidão é dada pela equação (4):

$$f_k = Min + (Max - Min)\left(\frac{N_p - k}{N_p - 1}\right) \tag{4}$$

onde k é o índice do cromossomo na população em ordem decrescente do valor da função objetivo. Vale notar que deste modo a aptidão representa o número de filhos esperados do cromossomo e $Max - Min$ representa uma grande possibilidade de seleção (razão entre a maior aptidão e aptidão média, $\frac{f_{\max}}{\bar{f}}$).

2.2.3. Seleção

O processo de seleção baseia-se no princípio de sobrevivência dos melhores indivíduos. Nesse processo, os indivíduos com melhor aptidão têm maior probabilidade de serem escolhidos para reprodução ou cruzamento. Em contrapartida, os indivíduos com baixa aptidão podem não

ser considerados para reprodução, conforme a pressão de seleção do esquema utilizado DA SILVA (2002). Assim, o objetivo da seleção é escolher os indivíduos que servirão de base (pais) para o processo de reprodução. Existem várias formas de efetuar a seleção, podendo-se citar a seleção por posição (*rank*), seleção proporcional à aptidão (*roulette wheel*), seleção por torneio dentre outras, LEMONGE (1999).

O método clássico proposto por Holland (1975), denominado seleção proporcional ou roda roleta, tem como idéia básica determinar a probabilidade de seleção para cada cromossomo proporcionalmente ao seu valor de aptidão. Para cada cromossomo k com avaliação f_k, onde a probabilidade de seleção P_k pode ser calculada conforme a equação (5).

$$p_k = \frac{f_k}{\sum_{k=1}^{N_p} f_k} \tag{5}$$

2.2.4. Cruzamento (crossover) e mutação

O processo de seleção não introduz novos indivíduos na população, apenas os chamados genitores, que servirão como "pais" para a nova geração, composta pelos "filhos".

É na etapa de reprodução na qual o algoritmo tenta criar novas e melhores soluções (indivíduos mais aptos). Para isso, operadores genéticos são utilizados. Os principais são os operadores de Crossover e Mutação, descritos resumidamente a seguir, DA SILVA (2002).

O operador cruzamento ou recombinação é aplicado em pares de cromossomos retirados da população, para gerar seus descendentes. Cada um dos cromossomos pais tem sua cadeia de dados secionada em uma posição aleatória, produzindo dois novos indivíduos (pares). Segundo DEJONG (1975) deve-se escolher uma probabilidade de cruzamento igual para cada par, em geral entre 0,6 a 0,9.

Várias técnicas de recombinação são propostas na literatura, cruzamento uniforme, cruzamento médio, cruzamento baseado na média geométrica, cruzamento *BLX* – α (*Blended crossover*), cruzamento simples, cruzamento aritmético e cruzamento linear. Maiores detalhes podem ser obtidos em DEB (2001).

O operador de mutação é necessário para a introdução e manutenção da diversidade genética entre os novos indivíduos na população. Ele fornece, assim, meios para a introdução de novos indivíduos ou elementos na população, assegurando que a probabilidade de se chegar a qualquer ponto do espaço de busca seja zero, com o intuito de tentar contornar o problema de ótimos locais, DA SILVA (2002). Este operador é utilizado após a aplicação dos operadores de

recombinação. No caso de codificação binária, o operador de mutação efetua uma troca aleatoriamente no *bit*, de 1 para 0 ou de 0 para 1, na codificação real a troca aleatória respeita os limites de cada variável de decisão. Mutação uniforme, mutação não-uniforme e mutação de deslocamento são algumas das técnicas de mutação propostas na literatura DEB (2001).

O operador de mutação é aplicado aos indivíduos com uma probabilidade dada pela taxa de mutação. Geralmente se utiliza uma taxa de mutação pequena (como na genética natural), pois a mutação é um operador genético secundário, DA SILVA (2002).

Os operadores genéticos acima descritos são os fatores responsáveis pela evolução dos AGs e têm como finalidade principal a obtenção de conjuntos de soluções melhores (populações) através de sucessivas iterações (gerações). GOLDBERG (1989) estudando tais operadores, afirma que através do processo iterativo podem-se perder informações (características) relativas à solução ótima, desta forma o operador de mutação faz com que estas informações sejam novamente incorporadas no processo de busca.

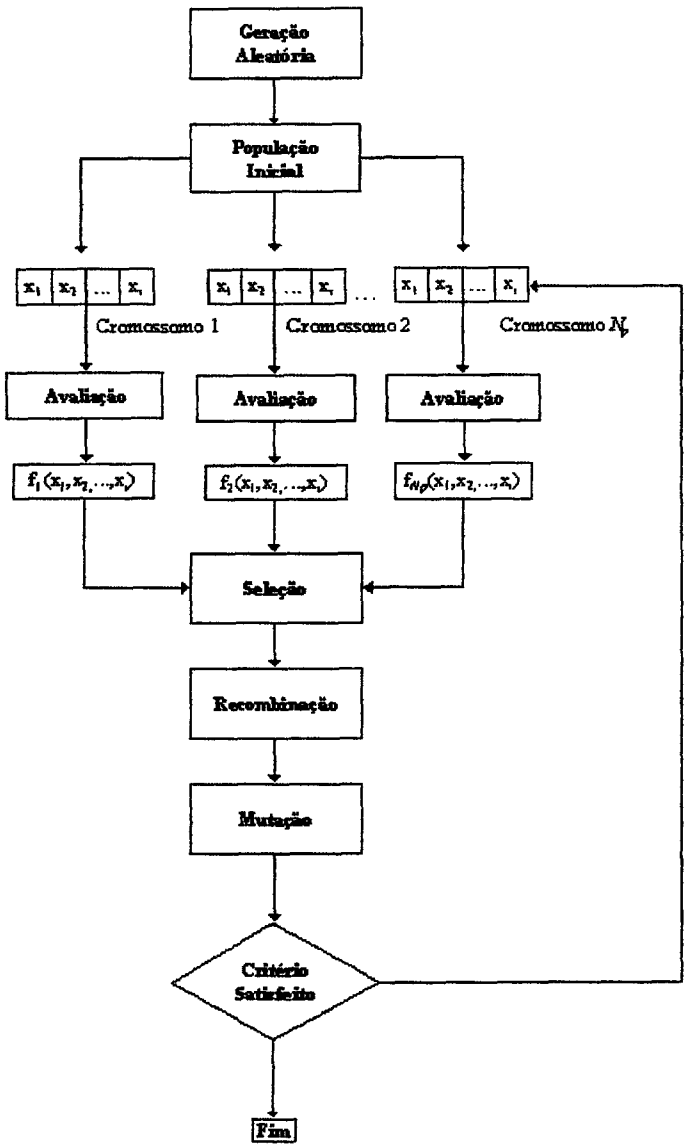
2.2.5. Gerações seguintes

Determinada a primeira geração, o procedimento se repete até que o critério de parada seja satisfeito. Normalmente, o critério de parada considera a convergência e número de geração. Quando uma dessas condições for satisfeita, tem-se a população final, que é o conjunto de possíveis soluções do sistema. A convergência consiste em verificar a proximidade da aptidão com um valor esperado. Fica a cargo do usuário estabelecer um limite entre o valor esperado e o valor de parada. Como alternativa ao teste de convergência, é estabelecido também um número máximo de gerações que possam ser produzidas. Essa estratégia é utilizada como meio de garantir que o algoritmo não fique executando por um tempo muito longo, já que não se sabe quando ele irá (e se irá) convergir, DA SILVA (2002).

2.2.6. Implementação e aplicação dos AGs

Diversas são as possibilidades de implementação e aplicação dos AGs. Geralmente, os AGs seguem o procedimento apresentado na Figura 2.

Figura 2 - Processo esquemático geral dos Algoritmos Genéticos convencionais - AGs



2.2.7. Tipos de algoritmos genéticos simples - AGs

Os AGs podem ser classificados em: geracional, geracional elitista e estacionário (*steady state*). Nos AGs geracionais, todas as soluções são avaliadas (funções de aptidão) e substituídas a cada iteração (geração). Nos AGs geracionais elitista, a estratégia de evolução do método garante que as melhores soluções de cada geração sejam preservadas para compor a próxima geração. Nos AGs estacionários (*steady state*), algumas soluções (por exemplo, duas soluções)

são geradas inicialmente a mais que a dimensão da população total, previamente estipulada (N_p). Por exemplo, geram-se $N_p + 2$ soluções iniciais que são avaliadas e classificadas em ordem decrescente ao valor de suas respectivas funções de aptidão. Uma análise estatística é realizada nessa etapa, onde as N_p melhores soluções são escolhidas e as duas piores soluções são descartadas. Depois disso, duas novas soluções são selecionadas e avaliadas para substituir os vetores descartados e constituir novamente uma população com dimensão $N_p + 2$. A partir da segunda geração, a avaliação da função de aptidão, nos AGs estacionários, é realizada apenas para dois novos indivíduos e não em todos como nos AGs geracionais.

Com a finalidade de exemplificar a utilização dos AGs em problemas reais de estimação pontual de parâmetros de distribuições de probabilidade, é apresentada a seguir a metodologia proposta.

3. Materiais e métodos

3.1. Distribuições de probabilidade utilizadas

Nesta seção são apresentadas e descritas as três distribuições de probabilidade utilizadas (Exponencial, Weibull e Birnbaum-Saunders) com suas respectivas funções de verossimilhança. A motivação desta escolha foi testar os AGs na estimação de parâmetros de uma distribuição de probabilidade com solução analítica conhecida e as outras com unicamente soluções numéricas.

3.1.1. Distribuição Exponencial

A função densidade de probabilidade de uma variável aleatória Y com distribuição exponencial e parâmetro θ é dada pela seguinte equação:

$$f(y) = \frac{1}{\theta} e^{-y/\theta}, \tag{6}$$

para $y > 0$ e $\theta > 0$. Considerando uma amostra aleatória de tamanho n a função de verossimilhança é:

$$L(\theta)=\prod_{i=1}^{n}\frac{1}{\theta}e^{-y_i/\theta}=\theta^{-n}e^{-\sum_{i=1}^{n}y_i/\theta}, \tag{7}$$

e o logaritmo da $L(\theta)$ é:

$$l(\theta)=-n\ln\theta-\sum_{i=1}^{n}\frac{y_i}{\theta}. \tag{8}$$

A derivada primeira de $l(\theta)$ denotada por $s(\theta)$ é:

$$S(\theta)=\frac{dl(\theta)}{d\theta}=-\frac{n}{\theta}+\sum_{i=1}^{n}\frac{y_i}{\theta^2}. \tag{9}$$

Para encontrar $\hat{\theta}$, determinamos as raízes da equação de máxima verossimilhança $S(\theta)=0$, dado por:

$$\hat{\theta}=\frac{\sum_{i=1}^{n}y_i}{n}. \tag{10}$$

3.1.2. Distribuição Weibull

A função densidade de probabilidade de uma variável aleatória com distribuição de Weibull de uma variável aleatória Y é dada pela seguinte equação:

$$f(y)=\frac{\xi}{\psi}\left(\frac{y}{\psi}\right)^{\xi-1}\exp\left\{-\frac{y}{\psi}\right\}^{\xi}, \tag{11}$$

para $y>0$, $\xi>0$ e $\psi>0$, onde ξ e ψ são os parâmetros de forma e escala, respectivamente. Considerando uma amostra aleatória de tamanho n a função de verossimilhança é:

$$L(\psi,\xi)=\prod_{i=1}^{n}\frac{\xi}{\psi}\left(\frac{y_i}{\psi}\right)^{\xi-1}\exp\left\{-\frac{y_i}{\psi}\right\}^{\xi}=\left(\frac{\xi}{\psi}\right)^{n}\prod_{i=1}^{n}\left(\frac{y_i}{\psi}\right)^{\xi-1}\exp\left\{-\sum_{i=1}^{n}\left[\frac{y_i}{\psi}\right]^{\xi}\right\}. \tag{12}$$

Uma outra forma é obtida pela re-parametrização de $\gamma=\xi$ e $\theta=\psi^{\xi}$, a função de verossimilhança em função de γ e θ é dada por:

$$L(\theta,\gamma)=\left(\frac{\gamma}{\theta}\right)^n\prod_{i=1}^{n}y_i^{\gamma-1}\exp\left\{-\sum_{i=1}^{n}\frac{y_i^{\gamma}}{\theta}\right\}, \tag{13}$$

e o logaritmo da função de verossimilhança (da parte informativa) para é γ e θ é,

$$l(\theta,\gamma)=n\ln(\gamma)-n\ln(\theta)+(\gamma-1)\sum_{i=1}^{n}\ln(y_i)-\frac{1}{\theta}\sum_{i=1}^{n}y_i^{\gamma}\ . \tag{14}$$

Derivando a equação (14) em relação a γ e θ, e igualando a zero obtemos as seguintes equações de verossimilhança:

$$s(\gamma)=\frac{\partial l(\theta,\gamma)}{\partial\gamma}=\frac{n}{\hat{\gamma}}+\sum_{i=1}^{n}\ln(y_i)-\frac{1}{\hat{\theta}}\sum_{i=1}^{n}y_i^{\hat{\gamma}}\ln(y_i)=0\,, \tag{15}$$

e

$$s(\theta)=\frac{\partial l(\theta,\gamma)}{\partial\theta}=-\frac{n}{\hat{\theta}}+\frac{1}{\hat{\theta}^2}\sum_{i=1}^{n}y_i^{\hat{\gamma}}=0\,. \tag{16}$$

Resolvendo a equação (16) encontramos uma expressão para o EMV de θ em termos de $\hat{\gamma}$, a qual é dada por:

$$\hat{\theta}=\frac{\sum_{i=1}^{n}y_i^{\hat{\gamma}}}{n}, \tag{17}$$

onde $\hat{\gamma}$ é um EMV do parâmetro γ, observe que não é possível determinar estes estimadores algebricamente. Portanto, deve-se recorrer a técnicas numéricas para obter estes estimadores, resolvendo a seguinte equação:

$$g'(\hat{\gamma}) = \frac{\sum_{i=1}^{n} y_i^{\hat{\gamma}} \ln(y_i)}{\sum_{i=1}^{n} y_i^{\hat{\gamma}}} - \frac{1}{\hat{\gamma}} - \frac{1}{n}\sum_{i=1}^{n} \ln(y_i) = 0. \qquad (18)$$

Observe que, para o caso da distribuição de Weibull, além de derivar é preciso utilizar algum método numérico para determinar os parâmetros da função. Em geral, este problema aumenta quando o número de parâmetros é maior que dois, o qual não é trivial na prática, principalmente na presença de variáveis explicativas. Esta é a principal justificativa para a utilização de AGs na determinação de parâmetros de uma distribuição multiparamétrica, pois os algoritmos genéticos se ajustam eficientemente quando se tem um grande número de parâmetros como é o caso do modelo polinomial ortogonal múltiplo da distribuição de Birnbaum-Saunders dado a seguir.

3.1.3. Modelo polinomial ortogonal múltiplo da distribuição de Birnbaum-Saunders

O modelo polinomial ortogonal múltiplo da distribuição de Birnbaun-Saunders é dado na forma matricial por:

$$Y_i = \ln(N) = \tilde{X}\tilde{\theta} + \phi_i, \qquad (19)$$

onde $\phi_i = \ln(\delta_i)$ é o erro aleatório do modelo o qual tem distribuição Senh-Normal com parâmetros α, $\gamma = 0$ e $\upsilon = 2$, MARTÍNEZ (2001).

Um caso especial do modelo dado pela equação (19), para $k = 2$ variáveis independentes é o modelo quadrático dado por:

$$Y_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{1i}^2 + \theta_4 x_{2i}^2 + \theta_5 x_{1i} x_{2i} + \phi_i, \qquad (20)$$

para $i = 1,2,\cdots,n$ e $\phi_i \sim SN(\alpha,0,2)$, onde os parâmetros do modelo são obtidos pelo método de máxima verossimilhança. O modelo dado pela equação (19) é utilizado para o estudo da fadiga em madeira e derivados e outros materiais submetidos à fadiga. Para o caso específico de dois fatores com três níveis cada um, utilizar o modelo dado pela equação (20), MARTÍNEZ, et al. (2003).

Sejam $y_1,\cdots,y_n$, n observações independentes do modelo dado pela equação (20), com $\phi_i = y_i - (\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{1i}^2 + \theta_4 x_{2i}^2 + \theta_5 x_{1i} x_{2i})$ onde $\phi_i \sim SN(\alpha,0,2)$. A função densidade de probabilidade de ϕ_i é dada na seguinte forma:

$$f(\phi_i) = (\frac{2}{2\sqrt{2\pi}}) \times W_i \times \exp\{-\frac{1}{2} Z_i^2\}, \tag{21}$$

onde,

$$W_i = \frac{2}{\alpha} \cosh\left[\frac{y_i - (\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{1i}^2 + \theta_4 x_{2i}^2 + \theta_5 x_{1i} x_{2i})}{2}\right],$$

$$Z_i = \frac{2}{\alpha} \operatorname{senh}\left[\frac{y_i - (\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{1i}^2 + \theta_4 x_{2i}^2 + \theta_5 x_{1i} x_{2i})}{2}\right],$$

para $-\infty < \theta_0,\theta_1,\cdots,\theta_5 < \infty$; $\alpha > 0$ e $-\infty < y_i < \infty$ (MARTÍNEZ, 2001 e MARTÍNEZ, et al., 2003).

Portanto a função de verossimilhança para $\theta_0,\theta_1,\cdots,\theta_5$ e α é dada por:

$$L(\theta_0,\theta_1,\cdots,\theta_5,\alpha) = [\frac{2}{2\sqrt{2\pi}}]^n \prod_{i=1}^{n} W_i \times \exp\{-\frac{1}{2}\sum_{i=1}^{n} Z_i^2\}, \tag{22}$$

e o logaritmo da função de verossimilhança (da parte informativa) para $\theta_0,\theta_1,\cdots,\theta_5$ e α é,

$$l(\theta_0,\theta_1,\cdots,\theta_5,\alpha) = \sum_{i=1}^{n} \ln(W_i) - \frac{1}{2}\sum_{i=1}^{n} Z_i^2 + k, \tag{23}$$

onde $k = n \ln(\alpha)$.

Assim derivando a equação (23) em relação a $\theta_0,\theta_1,\cdots,\theta_5$ e α, e igualando a zero, obtemos as seguintes equações de verossimilhança:

$$\frac{\partial l(\theta_0,\theta_1,\cdots,\theta_5,\alpha)}{\partial \theta_0} = \frac{1}{2}\sum_{i=1}^{n} \{Z_i W_i - \frac{Z_i}{W_i}\} = 0, \tag{24}$$

$$\frac{\partial l(\theta_0,\theta_1,\cdots,\theta_5,\alpha)}{\partial\theta_1}=\frac{1}{2}\sum_{i=1}^{n}x_{1i}\{Z_iW_i-\frac{Z_i}{W_i}\}=0, \tag{25}$$

$$\frac{\partial l(\theta_0,\theta_1,\cdots,\theta_5,\alpha)}{\partial\theta_2}=\frac{1}{2}\sum_{i=1}^{n}x_{2i}\{Z_iW_i-\frac{Z_i}{W_i}\}=0, \tag{26}$$

$$\frac{\partial l(\theta_0,\theta_1,\cdots,\theta_5,\alpha)}{\partial\theta_3}=\frac{1}{2}\sum_{i=1}^{n}x_{1i}^2\{Z_iW_i-\frac{Z_i}{W_i}\}=0, \tag{27}$$

$$\frac{\partial l(\theta_0,\theta_1,\cdots,\theta_5,\alpha)}{\partial\theta_4}=\frac{1}{2}\sum_{i=1}^{n}x_{2i}^2\{Z_iW_i-\frac{Z_i}{W_i}\}=0, \tag{28}$$

$$\frac{\partial l(\theta_0,\theta_1,\cdots,\theta_5,\alpha)}{\partial\theta_5}=\frac{1}{2}\sum_{i=1}^{n}x_{1i}x_{2i}\{Z_iW_i-\frac{Z_i}{W_i}\}=0, \tag{29}$$

$$\frac{\partial l(\theta_0,\theta_1,\cdots,\theta_5,\alpha)}{\partial\alpha}=-\frac{n}{\alpha}+\frac{1}{\alpha}\sum_{i=1}^{n}Z_i^2=0, \tag{30}$$

Desta maneira resolvendo a equação (30) encontramos uma expressão para o estimador de máxima verossimilhança de α^2 em termos de $\hat{\theta}_0,\hat{\theta}_1,\cdots,\hat{\theta}_5$, o qual é dado por:

$$\hat{\alpha}^2=\frac{4}{n}\sum_{i=1}^{n}\operatorname{senh}^2\left[\frac{y_i-(\hat{\theta}_0+\hat{\theta}_1x_{1i}+\hat{\theta}_2x_{2i}+\hat{\theta}_3x_{1i}^2+\hat{\theta}_4x_{2i}^2+\hat{\theta}_5x_{1i}x_{2i})}{2}\right], \tag{31}$$

onde $\hat{\theta}_0,\hat{\theta}_1,\cdots,\hat{\theta}_5$ são os estimadores de máxima verossimilhança de $\theta_0,\theta_1,\cdots,\theta_5$, os quais são obtidos numericamente (MARTÍNEZ, et al., 2003).

Foram utilizados três exemplos para comparar e validar a utilização dos algoritmos genéticos na estimativa dos parâmetros das distribuições de probabilidade consideradas. No caso da distribuição exponencial, os parâmetros foram determinados por resolução analítica, já na distribuição de Weibull e Birnbaum-Saunders os parâmetros foram obtidos através do método de Newton KALBFLEISCH (1985).

3.2. Exemplos e resultados experimentais

Nesta seção são apresentados os exemplos e resultados experimentais. Estes dados serão utilizados para testar o método proposto.

Exemplo 1- Os tempos de vida (em dias) de 10 corpos-de-prova tratados com um verniz foram testados no Laboratório de Madeiras e de Estruturas de Madeira (LaMEM) da EESC-USP. Os dados obtidos são: 70, 11, 66, 5, 20, 35, 40, 29, 8. Supondo que os dados seguem uma distribuição exponencial, estime o valor do parâmetro θ.

Exemplo 2- Para estudar o tempo de vida de um compensado, uma amostra aleatória de tamanho 20 foi utilizada. Os tempos de vida em horas até a falha dos ensaios realizados no Laboratório de Madeiras e de Estruturas de Madeira (LaMEM) da EESC-USP são: 100, 150, 200, 200, 350, 400, 450, 500, 550, 700, 750, 800, 800, 1000, 1200, 1500, 1700, 2000, 2500, 3000. Supondo que os dados seguem uma distribuição Weibull estime os valores dos parâmetros ψ e ξ.

Exemplo 3- Na Tabela 1, são apresentados os resultados experimentais de números de ciclos correspondentes à vida à fadiga (N) nos corpos-de-prova de madeira sem emendas, com os fatores tensão (S) e freqüência (f), para a espécie de Eucalipto Grandis. Os ensaios foram realizados no Laboratório de Madeiras e de Estruturas de Madeira (LaMEM) da EESC-USP MARTÍNEZ (2001).

Tabela 1 - Dados da vida à fadiga nos corpos-de-prova de madeira sem emendas, para a espécie de Eucalipto Grandis, obtidos no LaMEM, no 2001 MARTÍNEZ (2001)

Ensaio (i)	RESPOSTA N						Fatores	
	N_{i1}	N_{i2}	N_{i3}	N_{i4}	N_{i5}	N_{i6}	S	f
1	300060	447727	256633	483367	428444	180865	60	1
2	153472	124896	81872	94734	77639	114332	75	1
3	13	45	204	104	34	134	90	1
4	1601572	902906	1310678	867411	1413229	1887691	60	5
5	339606	255226	356979	249317	182182	290385	75	5
6	181	120	236	639	594	897	90	5
7	963383	1367100	1410236	1145759	1145760	1535709	60	9
8	175454	241199	310731	473653	332578	329510	75	9
9	355	157	430	284	745	1077	90	9

Nota: N_{ip^*} para $p^* = 1,\cdots,6$ são réplicas.

Considerando que dados da vida à fadiga nos corpos-de-prova de madeira sem emendas se aderem a um modelo polinomial ortogonal múltiplo da distribuição de Birnbaum-Saunders MARTÍNEZ, et al. (2003), estime os valores dos parâmetros $\theta_0,\theta_1,\cdots,\theta_5$ e α.

3.3. Breve descrição do programa utilizado

Em se tratando da técnica de otimização, Algoritmos Genéticos simples - AGs, pode-se afirmar que a literatura apresenta diversas e diferentes implementações. No presente estudo, optou-se por investigar a adequação de duas das principais proposições clássicas no problema proposto: geracional elitista e estacionário.

Para demonstrar à aplicação da metodologia sugerida, um programa computacional foi desenvolvido, em linguagem de programação C++, em dois módulos principais. O primeiro módulo tem como objetivo avaliar as funções-objetivo e enviar os resultados ao módulo de otimização, que processa a busca de novas soluções (candidatas à solução ótima) utilizando tais resultados. Embora, neste trabalho as funções-objetivo são definidas computacionalmente pelas funções de máxima verossimilhança, pois o objetivo é maximizar tais funções. Conceitualmente tais funções são diferentes conforme definidas nas seções 2.1 e 2.2.2. Depois que o módulo de otimização encontra novas soluções (população), recorre-se novamente ao módulo de avaliação para que as funções-objetivo sejam novamente avaliadas. O processo terminado quando se atinge o número máximo de iterações definidas inicialmente, pois o processo de otimização utilizando os AGs tem esta característica peculiar descrito no item 2.2.5. O módulo de otimização desenvolvido neste trabalho conta com suporte da biblioteca computacional de AGs (GAlib), escrita em C++ e de domínio público, desenvolvida por WALL (1996) no *Massachusetts Institute of Technology* (MIT).

4. Análises e discusões

Nesta seção são apresentadas as análises e discussões dos exemplos e resultados experimentais da aplicação de uma técnica de otimização não convencional (AGs), como ferramenta alternativa à estimação dos parâmetros de funções de máxima verossimilhança, que em muitos casos, não apresentam solução analítica. Para tal finalidade e com objetivo de introduzir a técnica dos AGs nessa área do conhecimento, três funções de probabilidade (Exponencial, Weibull e Birnbaun-Saunders) foram consideradas para estudo de caso.

Foram utilizadas neste trabalho diferentes tamanhos de população para os exemplos um e dois e para os dois AGs considerados. Os resultados de tais simulações são apresentados nas Figuras 3, 4, 7 e 8. Observa-se que os tamanhos da população variaram no intervalo 5-100. O número de gerações, a probabilidade de recombinação e a probabilidade de mutação foram considerados como fixos, respectivamente, 100, 0,9 e 0,1. Assim, quatro combinações de parâmetros foram determinadas, às quais são apresentadas na Tabela 2. Para o exemplo três foram realizadas outras combinações de parâmetros para as simulações conforme a Tabela 3 e os resultados são apresentados nas Figuras 11 e 12.

Tabela 2 - Combinação dos parâmetros dos AGs para os Exemplos 1 e 2

Combinação	Parâmetros			
	População	Gerações	Recombinação	Mutação
1	5	100	0,9	0,1
2	20-25	100	0,9	0,1
3	50	100	0,9	0,1
4	100	100	0,9	0,1

Tabela 3 - Combinação dos parâmetros dos AGs para o Exemplo 3

Combinação	Parâmetros			
	População	Gerações	Recombinação	Mutação
1	50	1000	0,9	0,1
2	100	1000	0,9	0,1

Observa-se, nas Figuras 3, 4, 7 e 8, que um tamanho de população igual a 5, para os exemplos um e dois (Exponencial e Weibull) e para os dois AGs considerados, os resultados não se aproximaram dos obtidos através dos valores referência obtidos pelo método analítico para a distribuição exponencial e pelo método de Newton para a distribuição Weibull . Com isso pode-se concluir que um tamanho de população muito pequeno (5) não oferece bons resultados em termos de convergência. Por motivo de melhor visualização e análise das outras dimensões da

população, os resultados referentes ao tamanho de população 5 foram retirados das Figuras 3, 4, 7 e 8, os resultados remanescentes são apresentados nas Figuras 5, 6, 9 e 10. Considerando o exemplo 2 (Weibull) e observando as Figuras 9 e 10, nota-se que os AGs geracionais elitistas necessitam de um menor número de iterações para atingir a convergência do que os AGs estacionários para um tamanho de população igual a 25.

Nota-se que, para o Modelo polinomial ortogonal múltiplo da distribuição de Birnbaum-Saunders os AGs estacionários tiveram um melhor desempenho comparando-se com os resultados obtidos nos AGs geracionais elistitas e que é mostrado nas Figuras 11 e 12. Pode-se observar que para populações de tamanhos diferentes ambos os AGs tiveram pouca influência as diferentes populações adotadas.

Figura 3 - AG Geracional Elistista Exponencial

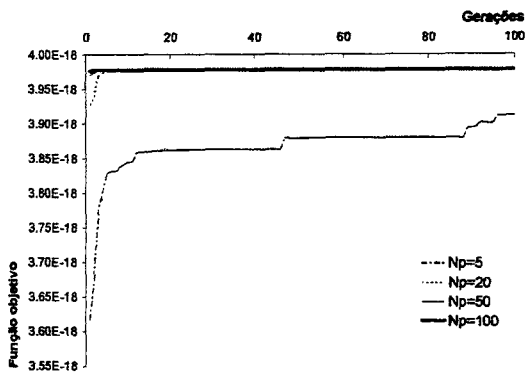


Figura 4 - AGs Estacionários Exponencial

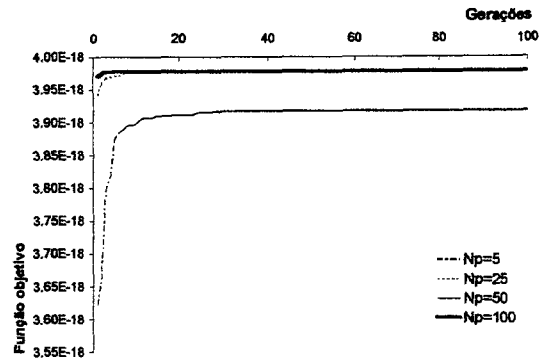


Figura 5 - AG Geracional Elistista - Exponencial (s/ Np 5)

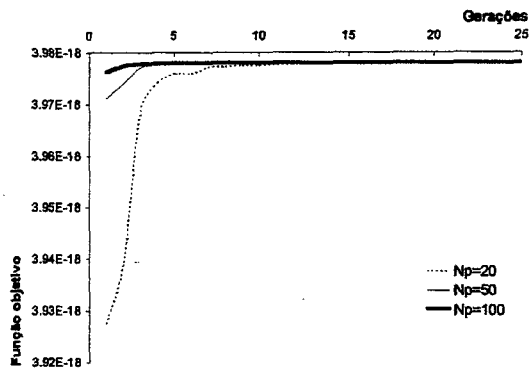


Figura 6 - AG Estacionário - Exponencial (s/ Np 5)

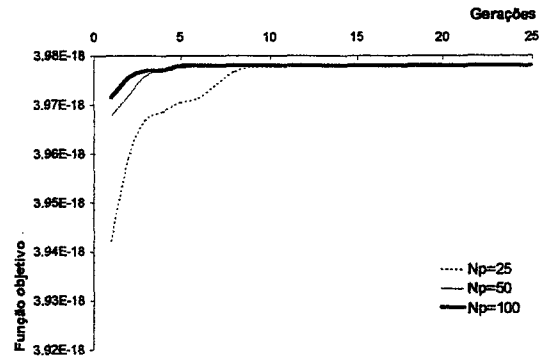


Figura 7 - AG Geracional Elistista – Weibull

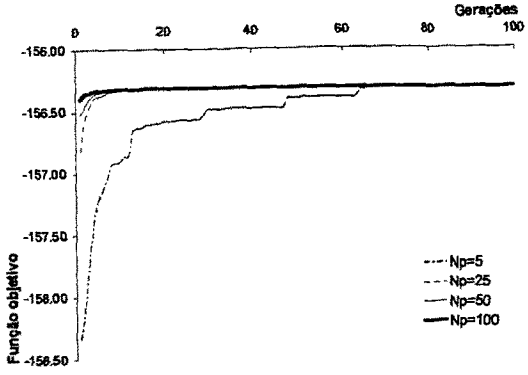


Figura 8 - AG Estacionário – Weibull5

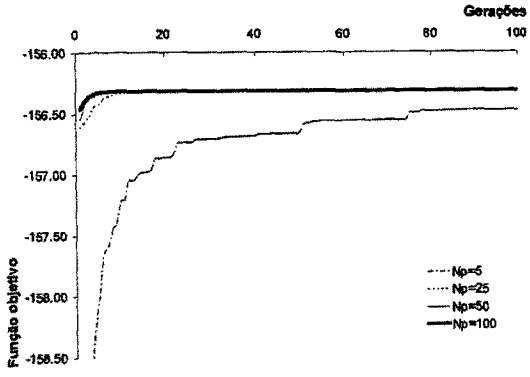


Figura 9 – AG Geracional Elistista – Weibull (s/ Np 5)

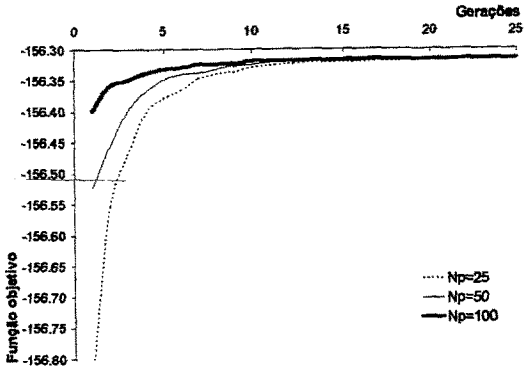


Figura 10 - AGs Estacionário – Weibull (s/ Np 5)

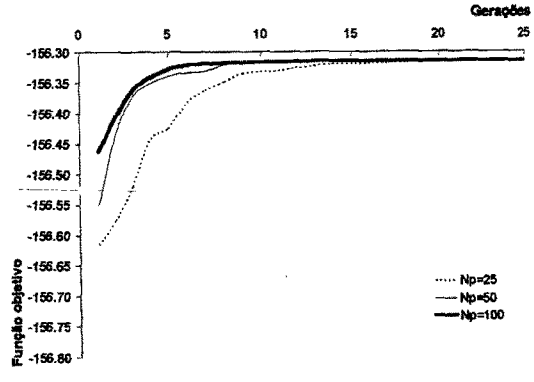


Figura 11– AG Geracional Elistista – Bimbaum-Saunders.

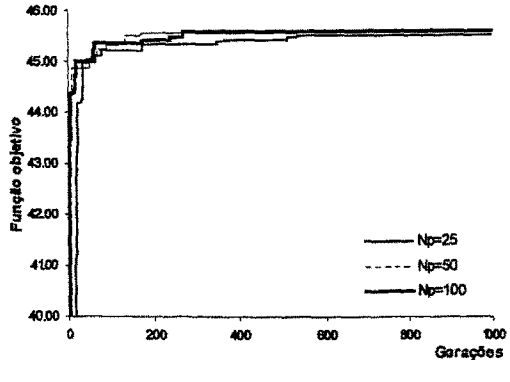
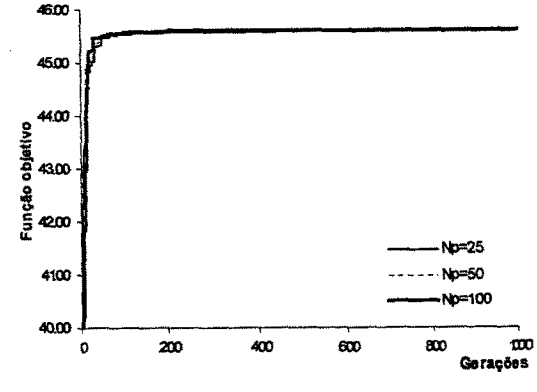


Figura 12 - AGs Estacionário – Birnbaum-Saunders.



Em termos de custo computacional, não se pode afirmar que os AGs geracionais elitistas são melhores que os AGs estacionários, pois os AGs geracionais avaliam todas soluções a cada iteração, já os AGs estacionários avaliam apenas duas soluções a cada iteração (2.2.7). Observou-se que para os exemplos um e dois os AGs estacionários necessitaram de um número menor de iterações para a estimação dos parâmetros, entretanto esta diferença é melhor observada no Exemplo 3 e mostrada nas Figuras 11 e 12.

Após, realizadas todas as simulações, foram obtidos os resultados referentes à estimação dos parâmetros das distribuições consideradas (Exponencial, Weibull e Birnbaum-Saunders). Estes parâmetros são apresentados nas Tabelas 3, 4 e 5 e comparados com suas respectivas soluções referências (analítica e numérica).

Tabela 3 - Parâmetros estimados para o Exemplo1 pelos métodos analítico exato e AGs

Método de cálculo	Parâmetro estimado $(\hat{\theta})$
Analítico exato	31,55
AGs geracionais elitistas	31,55
AGs estacionários (Steady State)	31,55

Tabela 4 - Parâmetros estimados para o Exemplo 2 pelos métodos de Newton, AGs geracionais elitistas e AGs estacionários (Steady State)

Método	Parâmetros estimados	
	$\hat{\xi}$	$\hat{\psi}$
Newton	1,23	1012,10
AGs geracionais elitistas	1,23	1012,10
AGs estacionários (Steady State)	1,23	1012,10

Tabela 5 - Parâmetros estimados para o Exemplo 3 pelos métodos de Newton, AGs geracionais elitistas e AGs estacionários (Steady State)

Parâmetros estimados	Método		
	Newton	AGs geracionais elitistas	AGs estacionários (Steady State)
$\hat{\theta}_0$	10,3790	10,3786	10,3791
$\hat{\theta}_1$	-4,1416	-4,1416	-4,1416
$\hat{\theta}_2$	0,7310	0,7307	0,7309
$\hat{\theta}_3$	-0,9227	-0,9228	-0,9228
$\hat{\theta}_4$	-0,2176	-0,2175	-0,2176
$\hat{\theta}_5$	0,1708	0,1712	0,1709
$\hat{\alpha}$	0,5381	0,5380	0,5381

5. Conclusões

Este trabalho permitiu demonstrar a aplicação da sistemática de Algoritmos Genéticos simples - AGs ao problema de estimação de parâmetros de distribuições de probabilidades através da maximização da função de verossimilhança. O procedimento foi ilustrado em três problemas-exemplos, com soluções conhecidas, como forma de verificar a adequação e desempenho da técnica no problema proposto. Cabe destacar que o presente trabalho não tem como objetivo substituir os métodos convencionais e sim mostrar a facilidade e eficiência dos algoritmos genéticos como um método alternativo para a estimação pontual de parâmetros de distribuições de probabilidade.

Os AGs se apresentaram robustos em relação ao problema e demonstraram que são ferramentas computacionais eficientes para estimação de parâmetros de distribuições de probabilidades. No entanto, como os AGs são métodos de busca estocásticas, proporcionaram um aumento do tempo computacional requerido que não pode ser considerado como uma desvantagem, pois em problemas onde a função verossimilhança apresenta várias inflexões locais, os métodos numéricos convencionais não têm bom desempenho, ao contrário dos AGs, que além de conseguirem investigar todo espaço de busca, geralmente apresentam soluções pertencentes à região do ótimo global. Portanto, pode-se afirmar que os AGs são técnicas computacionais adequadas para estimar parâmetros de distribuição de probabilidades.

Referências bibliográficas

BAKER, J. Reducing Bias and Inefficiency in the Selection Algorithm. *In: Grefenstette, J. (ed.). Proceedings of the Second International Conference on Genetic Algorithms and their Applications.* New Jersey: Hillsdale: Lawrence Erlbaum Association. p. 14-21, 1987.

BUSSAB, W.; MORETTIN, P. A. *Estatística Básica.* Editora Saraiva. São Paulo, 2003.

CHATTERJEE, S.; LAUDATO, M.; LYNCH, L. Genetic algorithms and their statistical applications: an introduction. *Computational Statistics and Data Analysis*, 22, 6, pp. 633-651, 1996.

CORDEIRO, G. *Introdução à teoria de verossimilhança.* 10° Simpósio Nacional de Probabilidade e Estatística (SINAPE). Universidade Federal do Rio de Janeiro, 1992.

DA SILVA, A. R. Projeto e Implementação de uma Ferramenta para o Pós-Processamento de Regras de conhecimento utilizando Algoritmos Genéticos. ICMC-USP, São Carlos SP, Novembro de 2002.

DEB, K. Multi-Objective Using Evolutionary Algorithms. John Wiley & Sons, Ltd, 2001.

DEJONG, K. The Analysis and Behavior of a Class of Genetic Adaptive Systems. University of Michigan. (PhD thesis), 1975.

Engineering Statistics Handbook. NIST/SEMATECH and Handbook of Statistical Methods, Sematech, Inc. http://www.itl.nist.gov/div898/handbook/,date. Copyright 2002.

KALBFLEISCH, J.G. *Probability and Statistical Inference.* 2 ed., New York, Spring-Verlag, v. 2: Statistical inference, 1985.

GEN, M.; CHENG, R. *Genetic Algorithms and Engineering Design.* John Wiley & Sons, INC, 1996.

GOLDBERG, D.E. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading, Massachusetts, 1989.

KNAN, N.; GOLDBERG, D.E.; PELIKAN M. Multi-objective Bayesian optimization algorithm. Illinois Genetic Algorithms Laboratory. University of Illinois at Urbana-Champaing. IlliGAL report No. 200209. March 2002.

LEMONGE, A.C. de C. Aplicação de Algoritmos Genéticos em Otimização Estrutural. Tese (Doutorado). Universidade Federal do Rio de Janeiro – RJ. 202 p, 1999.

MARTÍNEZ, E. M.; LOUZADA, F. N.; CALIL, C. J. A Multiple Orthogonal Polynomial Birnbaum-Saunders Model for Fatigue Data. Artigo submetido à publicação na Revista de matemática e Estatística, 2003.

MARTÍNEZ, M. E. Desenvolvimento de um Modelo Estatístico para Aplicação no Estudo de Fadiga em Emendas Dentadas de Madeira. São Carlos, SP, Tese de Doutorado - Escola de Engenharia de São Carlos/Instituto de Física de São Carlos/Instituto de Química de São Carlos - Universidade de São Paulo, 2001.

MICHALEWICZ, Z. Genetic Algorithms + Data Structures = Evolution Programs. Springer-Verlag, New York, N.Y, 1992.

PELIKAN M.; GOLDBERG, D.E.; CANTÚ-PAZ, E. BOA: The bayesian optimization algorithm. Illinois Genetic Algorithms Laboratory. University of Illinois at Urbana-Champaing. IlliGAL report No. 99003. January 1999.

PETROVSKI, A.; WILSON, A.; McCALL, J. Statistical analysis of genetic algorithms and inference about optimal factors. School of Computer and Mathematical Sciences. Faculty of Science and Technology. The Robert Gordon University. SCMS Technical Report 1998/2. December 1998.

PITTMAN, J.; MURTHY, C.A. Fitting optimal piecewise linear functions using Genetic Algorithms. IEEP Pattern Analysis and Machine Intelligence, 22, 7, 701-718, 2002.

SIERKSMA, G. Linear and Integer Programming: Theory and Practice. Marcel Dekker, Inc. New York, 1996.

SILVA, E. E. Otimização de estruturas de concreto armado utilizando algoritmos genéticos. Dissertação (Mestrado em Engenharia de Estruturas) Escola de Politécnica da Universidade de São Paulo, São Paulo - SP - Brasil, 2001.

WALL, M. Galib: A C++ Library of Genetic Algorithm Components (version 2.4). Mechanical Engineering Department, Massachusetts Institute of Technology, 1996.

WHITLEY, D. A genetic Algorithm Tutorial. Statistic and Computing, Vol. 4, pp. 65-85, 1994.

Abstract

A problem in almost all areas of scientific research is the point estimation of parameters of the probability distribution. A method most used for such estimation is the maximum likelihood method. This method consists of the maximization of the likelihood function. This can usually be done by the differentiation of the logarithm of this function with relationship to a parameter θ. However, in many real situations it is necessary to resolve this function by the iterative methods, mainly when the dimension of the parameters space is big. In this case, same disadvantages with respect to using these methods may be are: need of derivative the logarithm of the likelihood function, great number of parameters, etc. Therefore, the objective of this work is to use Simple Genetic Algorithms (SGA) for the point estimation of parameters of probability distributions. One of the advantages of SGA is that it does not require that derivative of the logarithm of the likelihood function is calculated.

Keywords: maximum likelihood estimators, Simple Genetic Algorithms, probability distributions.

Uso da distribuição de Poisson para avaliar a evolução da taxa de ocorrência de sinistros em uma carteira

Ary Elias Sabbag Junior*

Resumo

No presente trabalho é apresentada a solução para um problema prático, considerando-se para tanto a modelagem da taxa de ocorrência de sinistros através da abordagem de Modelos Lineares Generalizados. A questão investigada é se existe um aumento significativo da taxa de ocorrência de sinistro ao longo do tempo, e em caso afirmativo, se o mesmo é decorrente de condições externas ao controle da companhia, é decorrente de políticas de subscrição ou é devido a estes dois fatores. Para tanto, considerou-se o número de sinistros como sendo a variável resposta de interesse e os períodos de ocorrência e de subscrição como sendo as variáveis explicativas. A quantidade de sinistros foi modelada através do modelo de Poisson, relacionando-se a taxa de ocorrência de sinistros com as variáveis explicativas através da função de ligação canônica para distribuição de Poisson. As conclusões de interesse foram obtidas através de testes sobre os parâmetros estimados do modelo.

* Endereço para correspondência: HSBC SEGUROS, Fundação de Estudos Sociais do Paraná, UFPR.

1. Introdução

No processo de administração de uma carteira de seguros é de extrema importância a avaliação de índices, que determinam a qualidade do negócio. A referência à qualidade envolve, entre outros aspectos, o retorno financeiro obtido pela seguradora, o qual está diretamente relacionado com a evolução do prêmio de risco. Por definição, este prêmio é resultante do produto entre a taxa de ocorrência de sinistros e a indenização média e um aumento deste prêmio ao longo do tempo determina prejuízos para carteira, se não forem feitas as devidas correções do prêmio comercial praticado. Considerando que a seguradora esteja inserida em um mercado competitivo, pode-se, facilmente, perceber serem limitadas as possibilidades de reajustes periódicos no preço do seguro sem que os mesmos não venham a determinar uma diminuição de participação no mercado. Dadas estas considerações, fica evidenciada a necessidade de um controle sobre a evolução do prêmio de risco e, por conseguinte dos seus componentes, de forma que medidas corretivas possam ser tomadas no decorrer do tempo. Neste trabalho será dada atenção à taxa de ocorrência de sinistros, uma vez que uma evolução da mesma, ao longo do tempo, comprometeria o retorno com o negócio, justificando, quando desta constatação, a necessidade de serem tomadas medidas corretivas que possibilitem uma retomada de estabilidade. Uma das características associadas, diretamente, à taxa de ocorrência de sinistros, onde pode haver uma intervenção do gerenciador, é o processo de subscrição, no qual são estabelecidas características desejáveis para aceitação do segurado. Assim, por exemplo, na carteira do automóvel pode-se especificar a não aceitação de veículos esportivos ou na carteira de incêndio não se aceitar construções de madeira ou exigir-se a existência de extintores de incêndio no imóvel. Outra dimensão, onde pode haver uma intervenção do gerenciador é na alteração da tarifa, ajustando-a quando for evidenciada esta necessidade. O processo de subscrição é constituído por regras e procedimentos operacionais vigentes na rotina da seguradora e uma mudança só se justifica se for identificado objetivamente ser o mesmo o causador do aumento da taxa de ocorrência de sinistros. Assim, a questão que surge quando da identificação do referido aumento, é se este aumento é devido a condições associadas ao período de ocorrência, sendo, portanto, um aumento devido a fatores externos ao domínio da seguradora, ou se o aumento é devido a uma deterioração das condições estabelecidas no processo de subscrição, ou ainda, se o aumento é devido a ambos os fatores. A identificação de qual destes fatores é o responsável pelo aumento do referido índice, possibilitará a tomada de decisão pela manutenção ou não dos critérios estabelecidos no processo de subscrição de

segurados e pelo reajuste ou não da tarifa. O presente trabalho estabelece uma metodologia de análise que permite a referida tomada de decisão, considerando a quantidade de sinistros como sendo a variável resposta no problema proposto e os períodos de ocorrência do sinistro e período de subscrição da apólice geradora do sinistro, como as variáveis explicativas. Para a modelagem estatística considerou-se o número de sinistros como seguindo a distribuição de Poisson, adotando-se como fator ponderador para a estimação do modelo a quantidade de itens expostos ao risco em cada período.

Na seção 2, será apresentada a notação considerada, bem como a especificação do modelo adotado. Na seção 3, é apresentado um exemplo de aplicação do modelo juntamente com a interpretação dos resultados. Finalmente na seção 4, são apresentadas as conclusões obtidas.

2. Modelo adotado

Sejam n_{ij} e q_{ij} respectivamente o número de sinistros e quantidade de expostos ao risco referentes a apólices subscritas no período i com ocorrência no período j. Sendo n_{ij} uma variável quantitativa discreta, optou-se por uma aproximação para sua distribuição através do modelo probabilístico de Poisson, considerando-se como parâmetro de interesse λ_{ij}, onde este parâmetro corresponde à taxa anual de sinistros por exposto, com subscrição no período i e ocorrência no período j. Dada esta especificação, n_{ij} seguirá uma distribuição de Poisson com parâmetro $\theta_{ij} = \lambda_{ij} q_{ij}$ onde este parâmetro corresponde à taxa anual de sinistros no período de subscrição i e período de ocorrência j.

Sejam $X_i(X_i = i)$ e $X_j(X_j = j)$ as variáveis explicativas associadas ao i-ésimo período de subscrição da apólice geradora do sinistro $(i = 1,2,3,...,n)$ e j-ésimo período de ocorrência do sinistro $(j=1,2,3,...,p)$, respectivamente. A relação destas variáveis com o parâmetro λ_{ij} foi considerada através da função de ligação canônica para a distribuição de Poisson, ou seja:

$$ln\lambda_{ij} = \eta_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_j \qquad i=1,...,n\,;\, j=1,...,p \qquad (1)$$

onde os $\beta_l's$ correspondem aos parâmetros, da equação preditora linear, a serem estimados com base em um histórico de quantidades de sinistros e de exposição da seguradora.

Para estimação dos $\beta_l's$ considerou-se o estimador de máxima verossimilhança obtido numericamente através do método iterativo de mínimos quadrados ponderados; Dobson (1983).

Adotando-se a notação matricial, seja $\mathbf{b}^{(m)}$ o vetor de dimensão (3x1), associado à estimativa dos $\beta_l's$ na m-ésima iteração. A equação iterativa para o problema de estimação fica sendo:

$$\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}'\mathbf{W}_z \tag{2}$$

onde $\mathbf{X}$ corresponde à matriz de variáveis explicativas, com dimensão $(t \times 3)$, tendo a primeira coluna constituída por $1's$. A dimensão t corresponde ao número de pares (X_i, X_j) considerados na análise, ou seja, os períodos de ocorrência e de subscrição combinados segundo uma determinada ordenação. Em (2) $\mathbf{W}$ é uma matriz diagonal de dimensão $(t \times t)$, cujo elemento da posição *(k,k)* é dado por:

$$W_{kk} = q_k . e^{\eta_k} \; ; \qquad k=1, 2, ..., t \tag{3}$$

Para entendimento da notação a ser considerada, como exemplo, na equação (3), q_k corresponde a quantidade exposta ao risco para o *k*-ésimo período, onde este período é equivalente a um particular par *(i,j)* associado ao *i*-ésimo período de subscrição e *j*-ésimo período de ocorrência.

Finalmente, em (2), $\mathbf{z}$ é um vetor de dimensão *(t x 1)* com *k*-ésimo elemento dado por:

$$z_k = \eta_k + \frac{n_k - q_k . e^{\eta k}}{q_k . e^{\eta k}}; \quad k = 1,2,...,t . \tag{4}$$

Observando-se (2), percebe-se ter esta equação a mesma forma da equação de um modelo linear, obtido por mínimos quadrados ponderados, com a exceção de que aqui a solução tem de ser obtida iterativamente, em função de que na *m*-ésima iteração, tanto $\mathbf{z}$ como $\mathbf{W}$ dependem de $\mathbf{b}^{(m-1)}$, ou seja, da estimativa $\mathbf{b}$ obtida na iteração $m-1$ Nelder (1983).

Para estimação de $\mathbf{b}$ na equação (2), começa-se o processo iterativo considerando-se uma estimativa inicial $\mathbf{b}^{(0)}$ calculando-se com base na mesma $\mathbf{W}$ e $\mathbf{z}$, dadas pelas equações (3) e (4) respectivamente. Com base nestes resultados, por (2), estima-se $\mathbf{b}^{(1)}$, calculando-se novamente $\mathbf{W}$ e $\mathbf{z}$. O processo iterativo continua até que a diferença entre as sucessivas aproximações $\mathbf{b}^{(m-1)}$ e $\mathbf{b}^{(m)}$ seja suficientemente pequena. Neste caso houve convergência, considerando-se $\mathbf{b}^{(m)}$ como sendo o estimador de máxima verossimilhança $\hat{\boldsymbol{\beta}}$.

Um resultado importante para testes de hipóteses, é que a matriz de variância e covariância de $\hat{\boldsymbol{\beta}}$ pode ser aproximada por

$$(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \tag{5}$$

onde $\mathbf{W}$ é calculado com o estimador de máxima verossimilhança $\hat{\boldsymbol{\beta}}$.

Dadas as considerações acima, o estimador para λ_{ij} fica sendo

$$\hat{\lambda}_{ij} = e^{(\hat{\beta}_0 + \hat{\beta}_1 . X_i + \hat{\beta}_2 . X_j)}$$

Já o estimador para a taxa anual de sinistros no *j*-ésimo período de ocorrência, referente a apólices subscritas no *i*-ésimo período, fica sendo:

$$\hat{\theta}_{ij} = \hat{\lambda}_{ij} \, q_{ij}$$

Um resultado de interesse, para o estabelecimento de intervalos de confiança, vem a ser a variância do $\ln \hat{\lambda}_{ij}$, a qual é dada por:

$$\text{var}(\ln \hat{\lambda}_{ij}) = \text{var}(\hat{\beta}_0) + X_i^2 \, \text{var}(\hat{\beta}_1) + X_j^2 \, \text{var}(\hat{\beta}_2) + 2[X_i \, \text{cov}(\hat{\beta}_0, \hat{\beta}_1) + X_j \, \text{cov}(\hat{\beta}_0, \hat{\beta}_2) + X_i X_j \, \text{cov}(\hat{\beta}_0, \hat{\beta}_2)]$$

onde as variâncias e covariâncias das estimativas dos parâmetros são obtidas da matriz (5).

Fica sendo de interesse prático, a investigação se está havendo influência do processo de subscrição na taxa de ocorrência de sinistros $(\beta_1 \neq 0)$ ou se está havendo influência de fatores externos ao controle da seguradora na taxa de ocorrência de sinistros $(\beta_2 \neq 0)$. Assim, no caso de $\beta_1 > 0$, pode-se concluir que o processo de subscrição necessita ser revisto, em função do aumento da taxa de ocorrência de sinistros ser devido a este fator. Já $\beta_2 > 0$ implica que o aumento na taxa de ocorrência de sinistros é devido a fatores externos ao controle da seguradora (por exemplo, condições ambientais ou socioeconômicas), evidenciando uma necessidade de correção da tarifa.

Para a tomada de decisão, no contexto apresentado, podem ser consideradas as hipóteses:

$$H_0 : \beta_l = 0$$

$$H_1 : \beta_l \neq 0 \qquad l = 1,2$$

Para os testes acima, considerou-se a estatística

$$\frac{\hat{\beta}_l}{\sqrt{Var(\hat{\beta}_l)}} \qquad l = 1,2$$

onde $Var(\hat{\beta}_l)$ é obtido de (5).

Para efeito de tomada de decisão, sobre não-rejeição ou rejeição de H_0, deve-se comparar o resultado acima com os valores de uma distribuição Normal Padronizada, considerando-se para tanto um nível de significância α, pré-selecionado.

Observe-se que quando da não-rejeição da hipótese H_0 tanto para β_1 como para β_2, não se rejeita a hipótese de uma estabilidade na taxa de ocorrência de sinistros.

No caso de rejeição da hipótese nula

para os dois parâmetros, pode-se testar as hipóteses

$$H_0 : \beta_1 = \beta_2$$
$$H_1 : \beta_1 \neq \beta_2$$

Para este teste pode-se considerar a estatística.

$$\frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1; \hat{\beta}_2)}} \qquad (6)$$

onde $Var(\hat{\beta}_1), Var(\hat{\beta}_2)$ e $Cov(\hat{\beta}_1, \hat{\beta}_2)$ são obtidos da matriz (5).

Aproximando-se a distribuição da estatística acima por uma Normal Padronizada e escolhendo-se um nível de significância α, tem-se estabelecida uma regra de decisão para a escolha de uma hipótese.

O intervalo, com grau de confiança $(1-\alpha)$, para λ_{ij}, considerando-se a aproximação Normal, fica sendo dado por:

$$e^{\left(\ln \hat{\lambda}_{ij} \pm z_{1-\alpha/2} \sqrt{\text{var}(ln\hat{\lambda}_{ij})}\right)}$$

onde $z_{1-\alpha/2}$ corresponde ao valor de uma variável Normal Padronizada com área acima deste valor igual a $\alpha/2$.

3. Exemplo de aplicação

Os dados a serem analisados são referentes a uma particular garantia, em uma carteira constituída por contratos com vigência anual, considerando-se períodos trimestrais associados aos períodos de subscrição e de ocorrência de sinistros. Na tabela abaixo são apresentadas as quantidades expostas ao risco anualizadas, para cada um dos trimestres considerados no estudo.

Tabela 1 - Quantidades expostas ao risco (q_{ij})

TRIMESTRE DE SUBSCRIÇÃO	TRIMESTRE DE OCORRÊNCIA								
	04/99 a 06/99	07/99 a 09/99	10/99 a 12/99	01/00 a 03/00	04/00 a 06/00	07/00 a 09/00	10/00 a 12/00	01/01 a 03/01	TOTAL
04/98 a 06/98	1,698.60								1,698.60
07/98 a 10/98	3,183.38	1,610.37							4,793.75
10/98 a 12/98	2,904.06	2,848.82	1,243.99						6,996.87
01/99 a 03/99	2,705.72	2,497.68	2,448.55	1,349.68					9,001.63
04/99 a 06/99	1,967.34	3,675.91	3,457.13	3,351.05	1,707.88				14,159.31
07/99 a 09/99		1,988.37	4,212.88	3,913.56	3,823.02	2,177.80			16,115.63
10/99 a 12/99			1,849.49	3,175.14	3,058.76	3,045.65	1,399.28		12,528.32
01/00 a 03/00				1,535.99	3,000.27	2,937.12	2,900.91	1,508.19	11,882.48
04/00 a 06/00					2,039.81	3,969.00	3,933.81	3,808.75	13,751.37
07/00 a 09/00						2,420.76	5,315.29	5,096.02	12,832.07
10/00 a 12/00							3,156.87	5,508.76	8,665.63
01/01 a 03/01								2,545.02	2,545.02
TOTAL	12,459.10	12,621.15	13,212.04	13,325.42	13,629.74	14,550.33	16,706.16	18,466.74	114,970.68

As quantidades expostas anuais, apresentadas na tabela acima, correspondem aos q_{ij} definidos na seção 2, sendo dadas por:

$$q_{ij} = \frac{\text{nº total de dias de exposição dos itens no trim. de ocor. } j \text{ referentes à apol. com subsc. no trim. } i}{365}$$

Nesta tabela, pode-se observar que para cada trimestre de subscrição, no acompanhamento das exposições ao longo dos trimestres de ocorrência, há uma quantidade exposta menor no primeiro trimestre, havendo uma estabilidade nos trimestres seguintes, com uma diminuição no último trimestre, isto deve-se ao fato de que os itens vão tendo início de vigência durante o trimestre de subscrição, levando a uma menor exposição durante este trimestre. Considerando vigência anual haveria uma igualdade nas quantidades expostas nos trimestres intermediários caso não ocorressem cancelamentos das apólices. No exemplo sendo

considerado, pode-se observar que não há igualdade nas quantidades, em função de terem sido considerados, na contabilização das exposições, os cancelamentos.

Na tabela abaixo, são apresentadas as quantidades de sinistros, classificados de acordo com o trimestre de subscrição da apólice e o trimestre de ocorrência. Estes resultados correspondem aos n_{ij} definidos na seção 2.

Tabela 2 - Quantidades de sinistros (n_{ij})

TRIMESTRE DE SUBSCRIÇÃO	TRIMESTRE DE OCORRÊNCIA								
	04/99 a 06/99	07/99 a 09/99	10/99 a 12/99	01/00 a 03/00	04/00 a 06/00	07/00 a 09/00	10/00 a 12/00	01/01 a 03/01	TOTAL
04/98 a 06/98	21								21
07/98 a 10/98	38	20							58
10/98 a 12/98	42	44	16						102
01/99 a 03/99	34	41	33	18					126
04/99 a 06/99	33	59	68	61	25				246
07/99 a 09/99		27	65	54	47	27			220
10/99 a 12/99			38	74	54	37	23		226
01/00 a 03/00				32	51	50	42	23	198
04/00 a 06/00					45	85	75	78	283
07/00 a 09/00						70	119	107	296
10/00 a 12/00							60	132	192
01/01 a 03/01								72	72
TOTAL	168	191	220	239	222	269	319	412	2040

Na tabela abaixo é apresentada a taxa anual de sinistros observada no período analisado, a qual é dada por

$$f_{ij} = \frac{n_{ij}}{q_{ij}}.$$

Tabela 3 - Taxa anual de sinistros por exposto

TRIMESTRE DE SUBSCRIÇÃO	TRIMESTRE DE OCORRÊNCIA								
	04/99 a 06/99	07/99 a 09/99	10/99 a 12/99	01/00 a 03/00	04/00 a 06/00	07/00 a 09/00	10/00 a 12/00	01/01 a 03/01	TOTAL
04/98 a 06/98	0.0124								0.0124
07/98 a 10/98	0.0119	0.0124							0.0121
10/98 a 12/98	0.0145	0.0154	0.0129						0.0146
01/99 a 03/99	0.0126	0.0164	0.0135	0.0133					0.0140
04/99 a 06/99	0.0168	0.0160	0.0197	0.0182	0.0146				0.0173
07/99 a 09/99		0.0136	0.0154	0.0138	0.0123	0.0124			0.0137
10/99 a 12/99			0.0205	0.0233	0.0177	0.0121	0.0164		0.0180
01/00 a 03/00				0.0208	0.0170	0.0170	0.0145	0.0153	0.0167
04/00 a 06/00					0.0221	0.0214	0.0191	0.0205	0.0206
07/00 a 09/00						0.0289	0.0224	0.0210	0.0231
10/00 a 12/00							0.0190	0.0240	0.0222
01/01 a 03/01								0.0283	0.0283
TOTAL	0.0135	0.0151	0.0167	0.0179	0.0163	0.0185	0.0191	0.0223	0.0177

Observando-se as taxas de ocorrência de sinistros, nas marginais da tabela acima, constata-se um aumento tanto na dimensão associada ao trimestre de ocorrência como na dimensão associada ao trimestre de subscrição. A questão a ser analisada fica sendo a determinação se este aumento é devido a uma deterioração dos critérios associados ao processo de subscrição das apólices ou se está havendo uma alteração de fatores externos ao controle da seguradora, os quais implicam o referido aumento.

O modelo ajustado, conforme descrito na seção 2, foi

$$\ln \hat{\lambda}_{ij} = -4.3336 + 0.1144 X_i - 0.0580 X_j .$$

A taxa anual de sinistros, estimada para cada combinação de período de ocorrência e de subscrição, fica sendo dada por:

$$\hat{\theta}_{ij} = q_{ij} \hat{\lambda}_{ij}$$

Para os testes de hipóteses quanto à influência, na taxa de ocorrência, do período de subscrição e do período de ocorrência, considerou-se um nível de significância igual a 0.05. A interpretação para este valor é a de que há 5% de chance de rejeição da hipótese H_0 quando a mesma é verdadeira, ou seja, 5% de chance de se concluir de que há influência de período de subscrição (ou de ocorrência) sobre a taxa de ocorrência de sinistros, quando na verdade não há este efeito.

Neste estudo, as variâncias estimadas para $\hat{\beta}_1$ e $\hat{\beta}_2$ foram respectivamente iguais a 0.0003327 e 0.0004432.

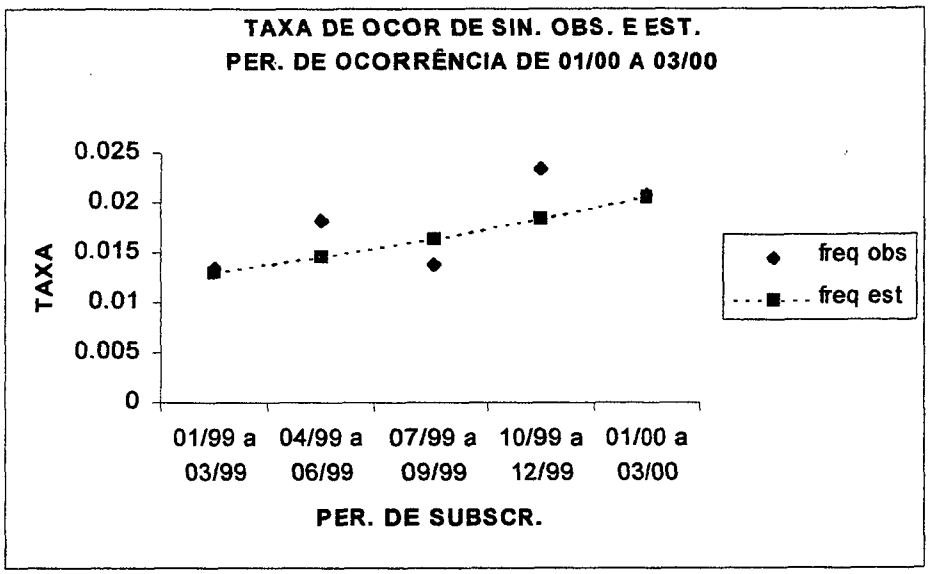
Assim, para o período de subscrição, rejeita-se a hipótese de que não haja influência do referido período na taxa de ocorrência de sinistros, considerando-se o nível de significância de 0.05 ($p<0.0001$). Desta forma, há fortes subsídios de que o período de subscrição tem um efeito altamente significativo na tendência de aumento da taxa de ocorrência de sinistros. Já para o período de ocorrência, também se rejeita a hipótese de que não haja influência do referido período, na taxa de ocorrência de sinistros, considerando-se o nível de significância de 0.05 ($p=0.0059$). Assim, pode-se concluir que os dados fornecem subsídios para se afirmar que, dado um período de subscrição, o período de ocorrência tem influência significativa na diminuição da taxa de ocorrência de sinistros.

A combinação dos dois testes aponta para necessidade de uma investigação sobre o processo de subscrição, analisando-se eventuais modificações ocorridas ou estabelecendo-se novos critérios para aceitação de segurados.

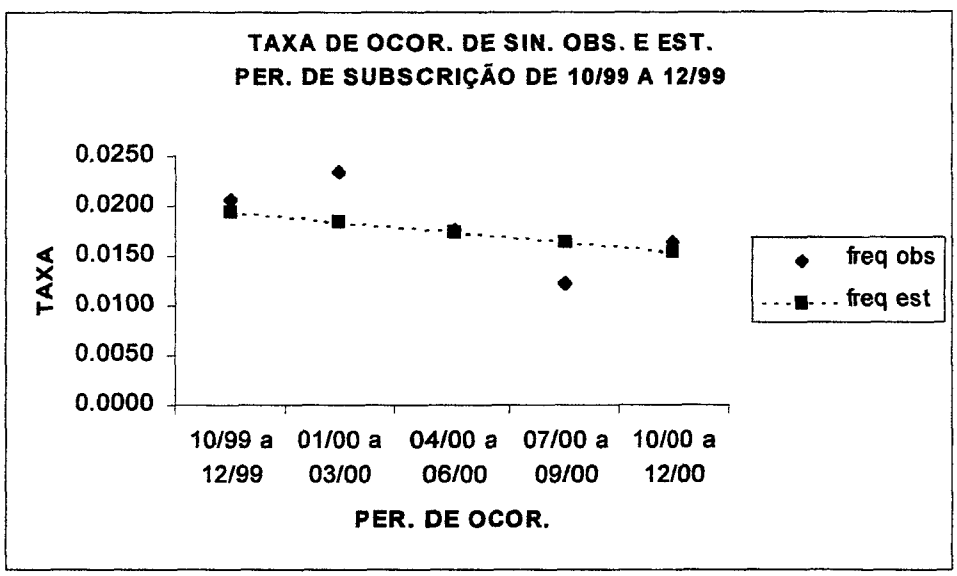
Quanto à comparação dos efeitos dos períodos de subscrição e de regulação sobre a taxa de ocorrência de sinistros, pode ser adotado o teste cuja estatística é apresentada em (6). A covariância entre $\hat{\beta}_1$ e $\hat{\beta}_2$, estimada em (5), foi igual a -0.000341 (correlação igual a −0.8883) e o teste estatístico indicou a rejeição da hipótese de igualdade entre os parâmetros considerando-se o nível de significância de 0.05 ($p<0.0001$).

A forma de interpretação dos resultados deve ser condicionada, ou seja, o comportamento da taxa de ocorrência, em relação ao período de subscrição, dado um período de ocorrência e o comportamento da taxa de ocorrência, em relação ao período de ocorrência, dado um período de subscrição. No primeiro caso há uma tendência significativa de aumento e no segundo caso há uma tendência significativa de diminuição.

Para efeito de visualização gráfica, considerando-se o trimestre de ocorrência de 01/00 a 03/00, têm-se as taxas de ocorrência observadas e estimadas, para os diferentes períodos de subscrição, são apresentadas no gráfico abaixo.

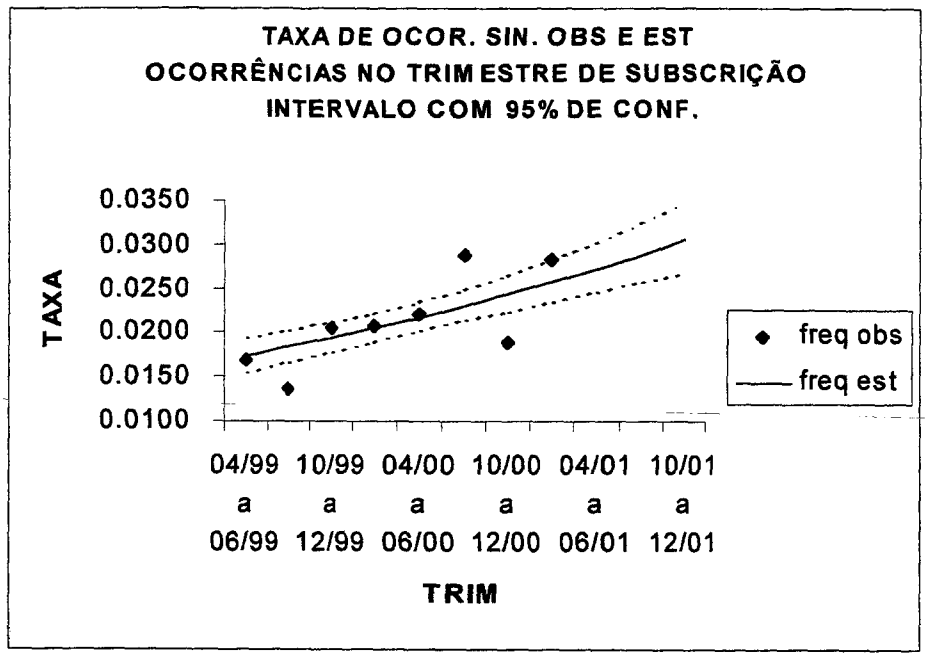


Na outra situação, considerando-se o período de subscrição de 10/99 a 12/99, têm-se as taxas de ocorrência observadas e estimadas, para os diferentes períodos de ocorrência, são apresentadas no gráfico abaixo



Estes resultados permitem identificar que há uma tendência à diminuição da taxa de ocorrência de sinistros, a qual não ocorre na prática em função de que o processo de subscrição não está bem ajustado. Assim sendo, o aumento da referida taxa, o que pode ser observado nas taxas marginais da Tabela 3, tem como motivo fatores sob controle do gerenciador, ou seja, uma redefinição dos critérios de aceitação dos segurados, não sendo portanto necessário um reajuste de tarifa.

Outra possibilidade de análise é a estimação da taxa de ocorrência em períodos futuros, caso sejam mantidas as condições do período de histórico. Como exemplo, considerando-se a taxa de ocorrência de sinistros com ocorrência no trimestre de subscrição, tem-se que a estimativa para o trimestre outubro/01 a dezembro/01 é de 0.0301, com intervalo de 95% de confiança igual a 0.0264; 0.0344. No gráfico abaixo, são apresentadas as taxas de ocorrência observadas e estimadas, juntamente com o intervalo de 95% de confiança



4. Conclusão

Cada vez mais se faz necessária a tomada de decisões com bases objetivas. Esta necessidade se acentua, quando a decisão deve ser tomada em um ambiente de incerteza, onde os custos associados a erros na escolha da ação levam a prejuízos financeiros ou mudanças operacionais sem resultados práticos. O modelo estabelecido neste trabalho possibilita ao gerenciador identificar a existência, ou não, de tendência no aumento da taxa de ocorrência de sinistros, e em esta ocorrendo, decidir sobre a necessidade de revisão dos critérios adotados nos processos de subscrição de apólices ou estabelecer uma revisão de tarifa.

Para uma boa tomada de decisão, se faz necessário ter conhecimento da probabilidade de cometerem-se determinados tipos de erro. Particularmente neste artigo, a informação

probabilística fornecida diz respeito à chance de optar-se pela revisão dos processos quando na verdade não havia esta necessidade. Esta probabilidade é derivada do teste de hipótese associado à estimativa do parâmetro relacionado com os períodos sob análise.

No exemplo considerado na seção 3, identificou-se uma probabilidade menor do que 0.0001 de que, não havendo influência do período de subscrição no aumento da taxa de ocorrência, um valor tão extremo para β_1 seja observado. Isto é, se a hipótese de não-influência fosse verdadeira, um evento muito raro teria ocorrido.

No exemplo considerado neste trabalho, os resultados permitiram identificar um aumento na taxa de ocorrência de sinistros, verificando-se estar este aumento associado ao período de subscrição da apólice. Esta constatação fornece subsídios ao gerenciador da carteira para proceder a uma revisão dos critérios adotados na subscrição das apólices, restringindo aceitação ou tornado-se mais rígido em determinados critérios. Outra evidência obtida é a de não haver um aumento na taxa de ocorrência, devido a fatores externos ao controle da seguradora.

Outro resultado derivado do uso do modelo proposto, é que no caso da tomada de medidas corretivas no processo de subscrição, a eficácia das mesmas pode ser avaliada por comparar-se a taxa de ocorrência de sinistros posterior a estas medidas com a taxa de ocorrência de sinistros estimada, para este período, com o modelo ajustado, uma vez que esta taxa de ocorrência estimada leva em conta a tendência estabelecida com base na série histórica anterior as mudanças efetuadas.

Referência bibliográfica

MCCULLAGH, P. and NELDER, J.A. (1983) ***Generalized Linear Models,*** **Chapman and Hall, London.**

DOBSON, ANNETTE J. (1983) ***Introduction to Statistical Modelling,*** **Chapman and Hall, London.**

Abstract

In this paper, the solution for a practical problem is presented, considering the modeling of the claims rate through Generalized Linear Models. The investigated question is whether a significant increase occurs in the claims rate over time, and in affirmative case, if this increase results from external factors, from underwriting policy or due to both factors. In this analysis, the number of claims was considered as being the response variable of interest and the periods of occurrence and underwriting as being the explanatory variables. The number of claims was modeled through the Poisson distribution, relating the claims rate with the explanatory variables through the canonical link function for distribution of Poisson. The conclusions were achieved through tests on the estimated parameters of the 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 permite fácil assimilação pelos membros da comunidade científica em geral. Os artigos devem incluir aplicações práticas como assunto central. Essas aplicações devem ter conteúdo estatístico substancial. As análises devem ser exaustivas e bem apresentadas, mas o emprego de métodos estatísticos inovadores não é essencial para publicação.

Artigos contendo exposição de métodos são aceitáveis, desde que estes sejam relevantes para as áreas cobertas pela revista, auxiliem na compreensão do problema e contenham interpretação clara das expressões matemáticas apresentadas. A apresentação de aplicações ilustrativas envolvendo dados adequados é requerida. Tratamentos algébricos extensos devem ser evitados.

A RBEs tem periodicidade semestral e publicará, também, artigos escritos a convite e resenhas de livros, bem como artigos abordando os diversos aspectos de metodologias relevantes para órgãos produtores de estatísticas, incluindo:

planejamento de pesquisas;

avaliação e mensuração de erros em pesquisas;

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

novos desenvolvimentos em metodologia de pesquisa;

crítica e imputação de dados;

amostragem e estimação;

disseminação e confiabilidade de dados;

análise de dados;

análise de séries temporais;

modelos e métodos demográficos; e

modelos e métodos econométricos.

Todos os artigos submetidos serão avaliados quanto à qualidade e à relevância por dois especialistas indicados pelo Comitê Editorial da Revista Brasileira de Estatística. Os artigos submetidos deverão ser inéditos e não deverão ter sido simultaneamente submetidos a qualquer outro periódico nacional. O processo de avaliação é do tipo duplo cego, isto é, os artigos são avaliados sem identificação da autoria, e os comentários dos avaliadores também são repassados aos autores sem identificação.

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBEs

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

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

Para cada artigo publicado, serão fornecidas gratuitamente 20 separatas.

Instruções para preparo de originais

Os originais entregues para publicação devem obedecer às seguintes normas:

1. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se, para cada um, a filiação e endereço para correspondência. Agradecimentos a colaboradores e instituições, e auxílios recebidos devem figurar, também, nesta página;

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

3. O artigo deve ser dividido em seções numeradas progressivamente, com títulos concisos e apropriados. Todas as seções e subseções devem ser numeradas e receber título apropriado;

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

5. As tabelas e gráficos devem ser precedidos de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas seqüencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração, tais como as demais seções do trabalho;

6. Gráficos e diagramas para publicação devem ser incluídos nos arquivos com os originais do artigo, sempre que possível. Quando isto não ocorrer, devem ser traçados em papel branco, com nitidez e boa qualidade, para permitir que a redução seja feita mantendo qualidade. Fotocópias não serão aceitas. É fundamental que não existam erros, quer no desenho, quer nas legendas ou títulos; e

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

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

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

atendimento
0800 218181