Presidente da República
Luíz Inácio Lula da Silva

Ministro do Planejamento, Orçamento e Gestão
Guido Mantega

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Eduardo Pereira Nunes

Diretor Executivo
José Sant´Anna Bevilaqua

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Maria Martha Malard Mayer

Diretoria de Geociências
Guido Gelli

Diretoria de Informática
Luiz Fernando Pinto Mariano (em exercício)

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 62 número 218 julho/dezembro 2001

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 62, n. 218, p. 1-90, jul./dez. 2001

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

Revista Brasileira de Estatística, ISSN 0034-7175

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

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

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

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

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

Impressão
GráficaDigital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2003

Capa
Renato J. Aguiar – Gerência de Criação – CDDI

Ilustração da Capa
Marcos Balster – Gerência de Criação – CDDI

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

v.

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

Continuação de: Revista de economia e estatística.

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

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

ISSN 0034-7175 = Revista brasileira de estatística.

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

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

Impresso no Brasil/*Printed in Brazil*

Sumário

Nota do Editor 5

Artigos

Uma análise de associação entre poluição atmosférica e saúde, usando um modelo de regressão binomial negativo 7
Jacqueline S. E. David
Silvia L. P. Ferrari

Construção de uma rede ótima para o monitoramento de um sistema bioecológico a partir do variograma 25
Gladys Elena Salcedo Echeverry
Luis Hermano Hurtado Tobón
Maria Dolly Garcia González

Estimação da fração de não-disjunção meiótica em pacientes com Síndrome de Down 45
Glaura da Conceição Franco
Paula Arantes Barros

Estimação de probabilidades em campeonatos de futebol utilizando simulação estocástica, modelos de séries temporais para dados de contagem e o princípio da máxima entropia 59
Eduardo Lima Campos
Cristiano Fernandes
Reinaldo Castro Souza

Política editorial 87

Nota do Editor

É com prazer que assumo a função de Editor Responsável da Revista Brasileira de Estatística - RBEs - em substituição a Pedro Luís do Nascimento Silva. Quero agradecer ao Pedro, em meu nome e em nome de todo o corpo editorial, pela dedicação, interesse e carinho que ele demonstrou à frente da revista nos últimos cinco anos. É uma honra e um privilégio poder continuar sua tarefa, tão importante para nós.

De fato, a RBEs possui uma posição única na comunidade estatística brasileira. Ela é produzida e distribuída pelo IBGE, que arca com todos os seus custos. A partir de um convênio celebrado em 1995 entre a Associação Brasileira de Estatística, na gestão de Pedro Morettin, e pelo IBGE, na gestão de Simon Schartzman, ficou estabelecido que a escolha do corpo editorial seria feita de comum acordo entre o IBGE a ABE.

A Associação Brasileira de Estatística já possui uma revista científica (REBRAPE- Brazilian Journal of Probability and Statistics) editada em inglês e com a intenção de circular internacionalmente. Seu público-alvo são os pesquisadores em estatística e probabilidade das universidades e centros de pesquisa do Brasil e do exterior.

A RBEs tem outro público-alvo: os estatísticos ou profissionais que lidam com coleta e análise de dados nos centros de pesquisa, nas empresas e nas instituições governamentais. Também faz parte desse público os estudantes de graduação e de pós-graduação em estatística. Para atingir este público amplo e diferenciado, a revista é publicada em português. Deste modo, a RBEs atualmente possui um perfil semelhante ao de várias revistas de nossa área, tais como a Statistical Science, American Statistician e Chance. Esta possibilidade de atingir um público tão amplo e diversificado é a característica mais importante da RBEs.

Estabeleci os seguintes três objetivos durante meu período como editor responsável da RBEs: tornar a RBEs mais conhecida do seu publico-alvo; aumentar a freqüência de submissão e publicação de bons artigos à revista; e aumentar o seu número de assinantes. Estes objetivos não são fáceis de serem alcançados. As razões que levam-me a acreditar que será possível atingi-los é o fato de contar com o apoio de um corpo editorial constituído por alguns dos melhores membros de nossa comunidade e com o apoio da ABE e do IBGE.

Este número vem composto de artigos com aplicações de estatística nas mais diversas áreas do conhecimento humano. Tenho certeza que você aprenderá algo interessante e novo, da estimação de probabilidades em jogos de futebol até a possível associação entre poluição atmosférica e saúde, um assunto muito debatido recentemente pela comunidade de saúde pública no mundo todo. Você ainda poderá conhecer como uma rede espacial estações de monitoria ecológica que pode ser especificada e aspectos mais intricados relacionados à estimação de probabilidade em genética moderna. Boa leitura de mais um número da Revista Brasileira de Estatística.

Renato Martins Assunção
Editor Responsável

Uma análise de associação entre poluição atmosférica e saúde, usando um modelo de regressão binomial negativo

Jacqueline S. E. David*
Silvia L.P. Ferrari*

Resumo

Modelos de regressão de Poisson são freqüentemente utilizados para analisar a associação entre dados de contagem e um conjunto de covariadas. A distribuição de Poisson, no entanto, impõe uma importante restrição ao modelo: ela assume que, dados os valores das covariadas, a média e a variância da variável resposta são iguais. Um modelo alternativo que permite que a variância da variável resposta seja maior que a respectiva média é baseado na distribuição binomial negativa. Nesse artigo, usamos o modelo de regressão log-linear binomial negativo para analisar a associação entre o número de atendimentos pediátricos de emergência por causas respiratórias e os níveis de poluição atmosférica na cidade de São Paulo. Os principais aspectos abordados são inferências dos parâmetros de regressão e dispersão, verificação da qualidade de ajuste do modelo e estimação de riscos relativos.

Palavras-chave: diagnóstico, distribuição de Poisson, distribuição binomial negativa, modelo linear generalizado, poluição atmosférica, saúde humana e superdispersão.

1. Introdução

O modelo padrão para a análise da associação entre dados de contagem e um conjunto de covariadas é baseado na distribuição de Poisson. Uma suposição dos modelos de regressão de Poisson é que, dados os valores das covariadas, a média e a variância da variável resposta são iguais. Freqüentemente esta suposição é violada, pois os dados apresentam dispersão maior do que a prevista pelo modelo. Possíveis causas para

* Endereço para correspondência: Departamento de Estatística, Universidade de São Paulo, Caixa Postal 66281, 05311-970, São Paulo - SP - Brasil. E-mail: sferrari@ime.usp.br.

superdispersão incluem heterogeneidade das unidades amostrais, existência de correlação entre as unidades amostrais, não observação de variáveis explicativas de grande importância para a resposta, entre outras. Uma conseqüência importante de se ignorar a superdispersão é que os erros-padrão obtidos através do modelo de regressão de Poisson são incorretos e subestimam a variabilidade dos estimadores de seus parâmetros.

A distribuição binomial negativa é uma alternativa à de Poisson em caso de superdispersão. Nesse trabalho, aplicamos um modelo de regressão binomial negativo a um estudo da associação entre o número de atendimentos pediátricos de emergência por causas respiratórias e os níveis de poluição atmosférica na cidade de São Paulo. Este artigo está relacionado com o de Ferrari e outros (2002) que utilizam modelos de superdispersão para análise de associação entre a concentração de dióxido de enxofre e o número de atendimentos por causas respiratórias em um pronto-socorro infantil em Vitória, Espírito Santo. McNeney e Petkau (1994) abordam o uso de modelos de Poisson superdispersos nesse contexto, embora o foco do trabalho desses autores seja um estudo de simulação. Estudos da associação entre poluição atmosférica e mortalidade ou atendimentos hospitalares têm sido realizados por vários pesquisadores; ver, por exemplo, Schwartz e outros (1996), Morgan e outros (1998), Saldiva e outros (1994) e Wong e outros (2001). Entretanto, o modelo binomial negativo não tem sido considerado em estudos epidemiológicos. Outra alternativa para a análise de contagens superdispersas são os modelos de quase-verossimilhança (Wedderburn, 1974). Um texto abrangente sobre modelos de superdispersão é o de Hinde e Demétrio (1998).

Na seção 2 revisamos o modelo usual de Poisson e o modelo binomial negativo, adequado para contagens superdispersas. A seção 3 é dedicada à verificação da qualidade de ajuste do modelo através de análise de diagnóstico, incluindo gráficos de resíduos e medidas de influência. Notamos que técnicas de diagnóstico são raramente utilizadas em estudos epidemiológicos. Um estudo da associação entre poluição atmosférica e saúde humana é apresentado na seção 4. A análise de diagnóstico indica que o modelo binomial negativo produz um ajuste bem melhor que o modelo de Poisson. Finalmente, na seção 5 apresentamos nossas conclusões.

2. Modelos para dados de contagem

Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes com distribuição de Poisson de médias $\mu_1, \ldots, \mu_n$, respectivamente, e sejam $x_1, x_2, \ldots, x_n$ vetores $p \times 1$ de constantes conhecidas. O modelo de regressão de Poisson assume que

$$g(\mu_i) = \eta_i = x_i^T \beta, \tag{1}$$

para $i=1,\ldots,n$, onde $\beta = (\beta_1, \ldots, \beta_p)^T$ é um vetor de p parâmetros desconhecidos e $g(.)$ é uma função monótona duplamente diferenciável. O modelo de regressão de Poisson log-linear usual assume que $g(\mu_i)=log(\mu_i)$. Aqui, $Var(Y_i)=\mu_i$, isto é, assume-se que, dados os valores das covariadas, a média e a variância da variável resposta são iguais. Este modelo pertence à classe dos modelos lineares generalizados (MLGs); ver McCullagh e Nelder (1989). Estimadores de máxima verossimilhança para os parâmetros da regressão podem ser obtidos pelo

procedimento iterativo de mínimos quadrados reponderados, que está implementado nos aplicativos S-PLUS (Venables e Ripley, 1999), GLIM 4 (Aitkin, Anderson, Francis e Hinde, 1989), STATA (Hardin e Hilbe, 2001) e SAS (Pedan, 2001), entre outros. Reduções no modelo podem ser baseadas na função desvio, que é definida como a diferença entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob investigação (com p parâmetros, sendo $p<n$), isto é, $D_p = \sum_{i=1}^{n} d_i^2$, onde

$$d_i = 2\, sinal(y_i - \hat{\mu}_i)\left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}^{1/2} \qquad (2)$$

sendo $\hat{\mu}_i$ o estimador de máxima verossimilhança de μ_i. Se o modelo inclui intercepto, temos $\sum_{i=1}^{n} (y_i - \hat{\mu}_i) = 0$ e o desvio se reduz a $D_p = 2\sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i)$. Sob a hipótese nula $H_0: \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$, o desvio parcial, isto é, a diferença entre o desvio do modelo (1) e o modelo restrito, tem distribuição qui-quadrado com $p-q$ graus de liberdade assintoticamente.

Existem vários procedimentos para a seleção de modelos de regressão (Paula, 2002). Aqui apresentaremos o método proposto por Akaike (1974), muito utilizado na seleção de MLGs, que se diferencia dos demais por ser um processo de minimização que não envolve testes estatísticos. A idéia básica é selecionar um modelo que seja parcimonioso, e como o máximo do logaritmo da função de verossimilhança $l(\beta;y)$ cresce com o aumento do número de parâmetros do modelo, uma proposta razoável seria encontrar aquele que apresenta menor valor para a função $AIC = -2l(\hat{\beta}\,; y) + 2p$, onde p denota o número de parâmetros e $\hat{\beta}$, a estimativa de máxima verossimilhança de β no modelo em questão. Assim, para o modelo de regressão de Poisson, temos que $AIC = -2\sum_{i=1}^{n} (y_i \log \hat{\mu}_i - \hat{\mu}_i - \log y_i!) + 2p$, com $\hat{\mu}_i = g^{-1}(x_i^T \hat{\beta})$.

A média da variável resposta e os parâmetros de regressão estão relacionados pela função de ligação $g(.)$. A função de ligação logarítmica $g(\mu_i) = \log \mu_i$ tem uma vantagem sobre outras possíveis funções de ligação usuais, pois permite uma interpretação simples para os parâmetros de regressão. Sejam $x_h = (x_{h1}, \ldots, x_{hp})^T$ um vetor de valores das covariadas e $x_{h^*} = (x_{h1}, \ldots, x_{hj} + c, \ldots, x_{hp})^T$. Sob um modelo log-linear, as médias da variável resposta dados x_h e x_{h^*} são $\mu_h = \exp(x_h^T \beta)$ e $\mu_{h^*} = \exp(x_{h^*}^T \beta)$, respectivamente. Note-se que o risco relativo μ_{h^*}/μ_h é dado por $\exp(c\beta_j)$ e, portanto, $c\beta_j$ é o logaritmo desse risco quando o valor da j-ésima covariada é acrescido de c unidades e as outras covariadas permanecem inalteradas.

A distribuição binomial negativa é útil para se definir um modelo de regressão para contagens superdispersas. Se Y tem distribuição binomial negativa com parâmetros $\mu>0$ e $k>0$, e função de probabilidade

$$f_Y(y;\mu,k)=\frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}\left(\frac{k}{k+\mu}\right)^k\left(\frac{\mu}{k+\mu}\right)^y, y=0,1,\ldots, \quad (3)$$

onde $\Gamma(.)$ é a função gama, então

$$E(Y)=\mu \text{ and } Var(Y)=\mu+\frac{\mu^2}{k}.$$

Note-se, portanto, que $Var(Y)>E(Y)$.

Uma possível motivação para assumir que as contagens observadas procedem de uma distribuição binomial negativa é que (3) aparece como a distribuição marginal de Y quando assumimos que $Y|Z\sim Poisson(Z)$ e $Z\sim$ Gama(μ,k), com $\mu>0$, $k>0$ e k independente de μ. A função densidade de probabilidade da distribuição gama nesse caso assume a forma

$$f_Z(z;\mu,k)=\frac{1}{\Gamma(k)}\left(\frac{zk}{\mu}\right)^k \exp\left\{-\left(\frac{kz}{\mu}\right)\right\}, z>0$$

O *modelo de regressão binomial negativo* assume que as variáveis aleatórias independentes $Y_1,\ldots,Y_n$ têm distribuição binomial negativa com médias $\mu_1,\ldots,\mu_n$, respectivamente, e um parâmetro comum k, e que as médias estão relacionadas com as covariadas $x_1,\ldots,x_n$ através da função de ligação (1). Para k conhecido, este modelo pertence à classe dos MLGs.

O logaritmo da função de verossimilhança para β e k é dado por

$$l(\mu,k,y)=\sum_{i=1}^{n}\{y_i\log\mu_i+k\log k-(k+y_i)\log(k+\mu_i)+\log\Gamma(k+y_i) -\log\Gamma(k)-\log y_i!\}.$$

A função escore $U(\beta,k)=(\partial l/\partial\beta_1,\ldots,\partial l/\partial\beta_p,\partial l/\partial k)^T$ tem elementos

$$\frac{\partial l}{\partial\beta_r}=\sum_{i-1}^{n}\frac{(y_i-\mu_i)}{V(\mu_i)g'(\mu_i)}x_{ir}, \quad (4)$$

e

$$\frac{\partial l}{\partial k}=\sum_{i-1}^{n}\left\{\Psi(y_i+k)-\Psi(k)-\log(\mu_i+k)-\frac{k+y_i}{k+\mu_i}+\log k+1\right\}, \quad (5)$$

onde $V(\mu_i)=\mu_i+\mu_i^2/k$, x_{ir} é o (i,r)-ésimo elemento do vetor x_i e $\psi(\cdot)=\Gamma'(\cdot)/\Gamma(\cdot)$ representa a função digama.

É fácil mostrar que a matriz de informação para (β,k) é dada por

$$I(\beta,k)\begin{bmatrix} X^T W X & 0 \\ 0^T & i_{k,k} \end{bmatrix}$$

onde $W = diag\{w_1,\ldots,w_n\}$, $w_i = (d\mu_i/d\eta_i)^2/V(\mu_i)$, X é uma matriz $n\times p$ de posto completo e linhas, e $x_1^T,\ldots,x_n^T$

$$i_{k,k}(\beta,k) = \sum_{i-1}^{n}\left\{-E(\Psi'(Y_i+k))+\Psi'(k)+\frac{1}{\mu_i+k}-\frac{1}{k}\right\}.$$

Observe que $I(\beta,k)$ é uma matriz bloco diagonal e, portanto, β e k são parâmetros ortogonais. A matriz de covariância assintótica do estimador de máxima verossimilhança de β é $(XWX)^{-1}$ se k é conhecido ou mesmo se é estimado a partir dos dados.

A forma bloco diagonal da matriz de informação $I(\beta,k)$ permite que as estimativas de máxima verossimilhança para β e k sejam obtidas pela aproximação de Gauss-Seidel, da seguinte forma:

i •para um valor fixo de k, obtenha a solução de $\partial l/\partial\beta_r=0$, para $r=1,\ldots,p$ (ver (4)), aplicando o algoritmo dos mínimos quadrados reponderados (ver McCullagh and Nelder, 1989);

ii •para um valor fixo de β, obtido no passo anterior, obtenha a solução de $\partial l/\partial k=0$ (ver (5)) aplicando o método iterativo de Newton-Raphson (ver Rustagi, 1994); e

iii •repita os passos i e ii alternando a estimação de β e k até que este processo convirja, obtendo assim as estimativas de máxima verossimilhança $\hat{\beta}$ e $\hat{k}$.

Um valor inicial para k é

$$k^{(0)} = \frac{\sum_{i=1}^{n}\hat{\mu}_i(1-c_i\hat{\mu}_i)}{\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i}-(n-p)},$$

onde aqui $\hat{\mu}_i$ é o estimador de máxima verossimilhança de μ_i obtido pelo modelo de regressão de Poisson e $c_i = x_i^T(X^T\hat{W}X)^{-1}x_i$, com $\hat{W}=\hat{W}(\hat{\mu})$, que é a estimativa da variância assintótica de $\hat{\eta}_i = x_i^T\hat{\beta}$. A idéia vem da comparação da estatística de Pearson do ajuste do modelo de Poisson, isto é, $X^2=\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2/\hat{\mu}_i$ com o valor esperado sob o modelo binomial negativo (ver Breslow, 1984, para detalhes).

Uma alternativa para a estimação de k baseada no método dos momentos é discutida por Breslow (1984). Lawless (1987) mostra que a utilização do método dos momentos para a obtenção de uma estimativa para k é mais robusta que o método de máxima verossimilhança, no entanto é menos eficiente se o modelo de regressão binomial negativo estiver correto. As funções `neg.bin(k)` e `glm.nb` da biblioteca MASS do S-PLUS (Venables e Ripley, 1999) dão as estimativas de máxima verossimilhança dos parâmetros do modelo assumindo que k é conhecido e desconhecido, respectivamente. Ver Hardin e Hilbe (2001, capítulo 13) e Pedan (2001) para detalhes sobre o ajuste desse modelo usando o STATA e o SAS, respectivamente.

Reduções no modelo podem ser baseadas no desvio parcial. Seja $H_0 : b_{\theta+1} = b_{\theta+2} = ... = b_\pi = 0$ a hipótese nula a ser testada contra a hipótese alternativa bilateral. Para k conhecido, os resíduos do desvio são dados por

$$d_i = sinal(y_i - \hat{\mu}_i)\left\{ y_i \left[\log\left(\frac{y_i}{y_i + k}\right) - \log\left(\frac{\hat{\mu}_i}{\hat{\mu}_i + k}\right)\right] + k \log\left(\frac{\hat{\mu}_i + k}{y_i + k}\right)\right\}^2 \qquad (6)$$

e o desvio parcial é igual à estatística da razão de verossimilhanças. Sob a hipótese nula, o desvio parcial tem distribuição assintótica qui-quadrado com p-q graus de liberdade. Se k é desconhecido, uma aproximação para o desvio parcial é obtida estimando k sob o modelo irrestrito.

A função de Akaike (AIC) para o modelo binomial negativo, pela definição apresentada anteriormente, é dada por

$$AIC = -2\sum_{i=1}^{n} \{y_i \log \hat{\mu}_i + \hat{k} \log \hat{k} - (\hat{k} + y_i)\log(\hat{k} + \hat{\mu}_i) + \log\Gamma(y_i + \hat{k}) - \log\Gamma(\hat{k}) - \log y_i!\} + 2p,$$

com $\hat{\mu}_i = g^{-1}(x_i^T \hat{\beta})$.

3. Técnicas de diagnóstico

Técnicas de diagnóstico são de grande relevância para detectar problemas em modelos de regressão, tais como inadequação do modelo utilizado e a presença de *outliers* e observações influentes. Gráficos dos resíduos de Pearson ou resíduos do desvio contra alguma função dos dados, como, por exemplo, as estimativas dos preditores lineares $\hat{\eta}_i = x_i^T \hat{\beta}$, podem ser úteis para detectar falta de ajuste do modelo ou destacar *outliers*. Os resíduos de Pearson são definidos como

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{v}_i}}, \qquad (7)$$

onde $\hat{v}_i$ é a estimativa da variância de Y_i. Para os modelos de regressão de Poisson e binomial negativo, os resíduos de Pearson são dados por (7) com $v_i = \mu_i$ e $v_i = \mu_i + \mu_i^2/k$, respectivamente. Os resíduos do desvio são dados por (2) e (6) para os modelos de regressão de Poisson e binomial negativo, respectivamente. Para mais

detalhes sobre análise de diagnóstico em modelos de regressão para dados de contagem, ver Cameron e Trivedi (1998, capítulo 5). O uso de alisadores para examinar gráficos de resíduos na análise de associação entre poluição atmosférica e saúde humana é discutido em Schwartz (1994).

Uma vez que a distribuição dos resíduos não é conhecida, gráficos de probabilidade meio-normal com envelope simulado são uma ferramenta útil para propósitos de diagnóstico (Atkinson, 1985, Neter e outros, 1996, seção 14.6). A idéia é acrescentar aos gráficos de probabilidade meio-normal um envelope simulado que pode ser utilizado para decidir se as respostas observadas são consistentes com o modelo ajustado.

Gráficos de probabilidade meio-normal com envelope simulado são construídos da seguinte forma:

1. ajuste o modelo e gere uma amostra simulada de n observações independentes utilizando o modelo ajustado como se fosse o modelo verdadeiro;

2. ajuste o modelo para a nova amostra, e calcule os valores absolutos ordenados da medida de diagnóstico de interesse;

3. repita os passos acima 18 vezes;

4. considere os n conjuntos de 19 estatísticas ordenadas; para cada conjunto, calcule as respectivas médias, valores mínimos e máximos; e

5. faça o gráfico desses valores e da medida de diagnóstico ordenada da amostra original contra os escores meio-normais $\Phi^{1}((i+n-1/8)/2(2n+1/2))$ onde $\Phi(.)$ é a função distribuição acumulada normal padrão.

Os valores mínimos e máximos das 19 estatísticas ordenadas constituem o envelope. Um gráfico de probabilidade meio-normal com envelope simulado dos resíduos de Pearson ou resíduos do desvio do ajuste de um modelo de Poisson pode ser utilizado como uma verificação informal de superdispersão. Uma parcela considerável de pontos acima do envelope simulado indica que a variabilidade dos resíduos é maior que a esperada, e portanto um modelo de superdispersão pode ser mais apropriado para o conjunto de dados em estudo.

Uma técnica gráfica para detectar superdispersão é discutida por Lambert e Roeder (1995), e uma verificação formal é descrita por Lawless (1987). Assuma que o modelo de regressão binomial negativo é correto e seja $\delta=1/k$. *Se* $\delta=0$, não há presença de superdispersão e os dados procedem de uma distribuição de Poisson. A estatística da razão de verossimilhanças de $H_0{:}\delta=0$ contra $H_1{:}\delta\neq 0$ é $\omega=-2(l_P-l_{NB})$, onde l_P e l_{NB} são os logaritmos das funções de verossimilhança maximizadas dos modelos de Poisson e binomial negativo, respectivamente. Observe que o valor de δ estabelecido em H_0 encontra-se na borda do espaço paramétrico e, portanto, as propriedades usuais do teste da razão de verossimilhanças não são válidas. Lawless (1987) mostra que, sob H_0, a função distribuição acumulada de ω é $F_\omega(z)=1/2+P(\chi_1^2\leq z)$, para $z\geq 0$, onde χ_1^2 representa uma variável aleatória de distribuição qui-quadrado com um grau de liberdade. Para outros testes de superdispersão, ver Dean (1992).

Para o modelo de regressão binomial negativo, o i-ésimo elemento da diagonal (h_i) da matriz $H=W^{1/2}X(XWX)^{-1}XW^{-1/2}$ pode ser visto como uma medida de alavanca (*leverage*) da observação

correspondente. Um gráfico de $h_1,\ldots,h_n$ contra os valores ajustados $\hat{\mu}_1,\ldots,\hat{\mu}_n$, pode ser útil para detectar observações com altos valores de alavanca. Essas observações são potencialmente influentes no ajuste do modelo. Para o modelo de regressão log-linear binomial negativo, $h_i = k\mu_i/(\kappa+\mu_i)x_i^T(X^TWX)^{-1}x_i$.

A distância de Cook é freqüentemente utilizada como uma medida de influência (Cook, 1986), pois considera o efeito da i-ésima observação sobre todos os valores ajustados pelo modelo. No nosso caso, não é possível escrever essa medida em uma forma fechada, mas esta pode ser aproximada por $r_i^2 h_i/(1-h_i)^2$, com r_i dado em (7). Diagnóstico de influência local e análise dos resíduos do desvio no modelo log-linear binomial negativo são discutidos em Svetliza e Paula (2001).

Sempre que os dados são obtidos em uma seqüência de tempo, uma maneira informal de avaliar se a suposição de que, dados os valores das covariadas, as observações são independentes é através do cálculo das correlações entre os resíduos r_i e r_{i-L} para $L=1,2,3,...$, denominadas autocorrelações de defasagem (*lag*) L (Morettin e Toloi, 1985). A construção de um gráfico com as autocorrelações para diferentes defasagens (função de autocorrelação) é muito útil na análise de diagnóstico, onde se espera que as autocorrelações estejam próximas de zero caso a independência entre as observações esteja satisfeita. Se a correlação entre os resíduos mais próximos no tempo é maior do que entre os resíduos mais distantes, dizemos que existe autocorrelação serial. Isto não acarretará viés nos coeficientes do modelo estimados, mas os erros-padrão poderão ser sub ou superestimados. A adoção do modelo de Zeger (Zeger e Liang, 1986) é uma alternativa para a análise de dados correlacionados por apresentar uma estrutura de dependência incorporada ao modelo, cujos parâmetros são estimados através de equações de estimação generalizadas.

4. Estudo de São Paulo

A cidade de São Paulo, a maior da América do Sul, tem sido objeto de estudos epidemiológicos para avaliar a associação entre poluição atmosférica e saúde humana; ver, por exemplo, Saldiva e outros (1994, 1995), Braga e outros (1999, 2001) e Conceição e outros (2001). São Paulo é um excelente local para avaliar os efeitos da poluição atmosférica na saúde. É o maior centro urbano industrializado da América Latina, com uma população de aproximadamente 10 milhões de habitantes. A frota de veículos automotores da Região Metropolitana de São Paulo é de cerca de 6 milhões de veículos, os quais constituem a principal fonte de poluição do ar, seguidos pelos processos industriais (CETESB, 2001). Devido às suas características geográficas, São Paulo apresenta freqüentes inversões térmicas que dificultam a dispersão de poluentes, resultando em um aumento substancial da poluição atmosférica. Além disso, a agência de controle da poluição atmosférica de São Paulo (CETESB) apresenta o registro diário da concentração de vários poluentes, facilitando a realização de estudos temporais relativos à poluição atmosférica. A grande massa populacional permite sua divisão em menores grupos de especial interesse, já que a literatura aponta que crianças, idosos e pessoas que apresentam doenças crônicas são mais suscetíveis à poluição do ar.

Esse estudo epidemiológico é do tipo ecológico, no qual a unidade de observação é a população e não o indivíduo. Além desse tipo de estudo ser simples e barato por aproveitar dados coletados rotineiramente por autoridades governamentais, ele tem a vantagem de que a população age como seu próprio controle, evitando fatores de confundimento que diferentes populações poderiam gerar.

Nesse artigo nosso foco é investigar a associação entre o número de atendimentos pediátricos de emergência por causas respiratórias e os níveis de poluição atmosférica na cidade de São Paulo. Trata-se de uma nova análise de parte dos dados estudados por Lin e outros (1999), os quais utilizam modelos de regressão de Poisson para avaliar tal associação. Mostramos que o modelo de regressão binomial negativo é mais adequado à modelagem do conjunto de dados considerado do que o modelo de regressão de Poisson.

Registros diários do número de atendimentos pediátricos de emergência por causas respiratórias foram originalmente obtidos no Instituto da Criança da Universidade de São Paulo -ICUSP- no período entre maio de 1991 e abril de 1993. No nosso estudo esses eventos representam o indicador de saúde da população.

O indicador de poluição utilizado foi a concentração média diária de material particulado (MP_{10}, em $\mu g/m^3$) e procuramos controlar o efeito de variáveis de confundimento. Para descrever as características do meio ambiente foram incluídas: a temperatura (^oC), representada pela média diária das temperaturas mínimas observadas em diferentes estações meteorológicas, e a umidade relativa do ar (%), representada pela média diária dos valores registrados ao meio-dia. Para representar fatores sazonais incluímos variáveis indicadoras para cada mês do período em estudo e dia da semana. O número total de atendimentos pediátricos de emergência por causas não respiratórias foi também considerado, objetivando controlar fatores externos como, por exemplo, greve no hospital.

A Tabela 4.1 mostra medidas-resumo anuais de algumas variáveis em estudo. A Figura 4.1 apresenta as séries de dados do número de visitas pediátricas de emergência por causas respiratórias (RESP) e da concentração de MP_{10}. Note-se que há períodos sem registros da variável resposta e alguns dias onde a concentração média de MP_{10} ultrapassou o valor de referência internacional de qualidade do ar ($150\mu g/m^3$).

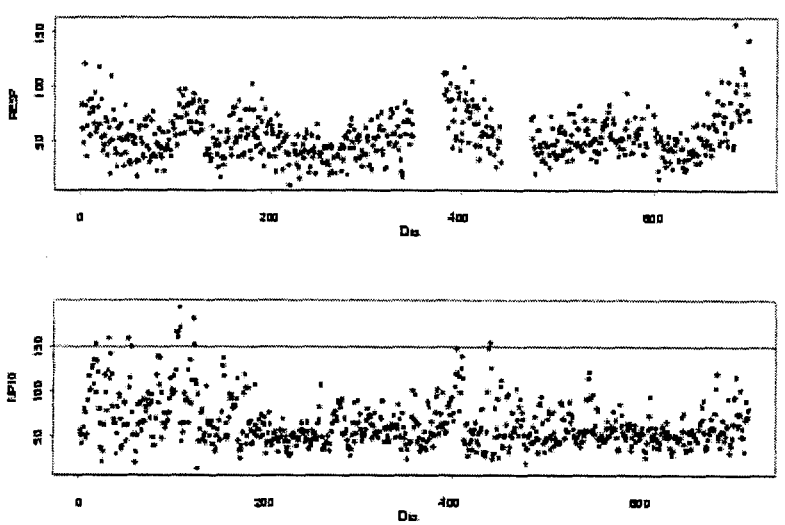
A estratégia de construção do modelo para a análise estatística envolveu a determinação de um modelo básico no qual efeitos devidos a variáveis climáticas, tendências sazonais e outros fatores de confundimento foram removidos. A idéia é explicar ao máximo a variabilidade no número de atendimentos pediátricos de emergência através do controle de variáveis de confundimento, definindo um modelo básico antes de adicionar a concentração do poluente no modelo. Estratégias similares para a construção do modelo de análise são apresentadas por Conceição e outros (2001), Schwartz e outros (1996) e Singer e outros (2002).

Considerando que a exposição ao poluente, à temperatura e à umidade deve preceder o efeito na saúde, incluímos médias móveis para cada uma dessas variáveis. A média móvel para uma variável é a média de seu valor no dia considerado com os dos dias precedentes, isto é, uma média móvel de dois dias é a média da variável no dia presente e no dia anterior. Médias móveis com defasagens de um a sete dias foram avaliadas para

Tabela 4.1 - Número de observações (*n*) e médias (desvios-padrão) anuais do número de visitas pediátricas de emergência por causas respiratórias (RESP) e não-respiratórias (N-RESP), temperatura (T), umidade (U) e concentração de MP_{10}

Ano	RESP	NRESP	T (^oC)	U (%)	MP_{10} ($\mu g/m^3$)
1991	56,95	119,37	14,17	65,23	76,42
(n=226)	(19,59)	(44,90)	(2,86)	(13,45)	(34,10)
1992	54,02	120,97	15,77	69,26	61,83
(n=292)	(18,43)	(54,02)	(2,74)	(13,83)	(22,05)
1993	59,38	153,56	18,24	60,93	53,83
(n=99)	(26,14)	(61,83)	(1,15)	(12,21)	(15,94)
Total	55,95	125,61	15,58	66,45	65,89
(n=617)	(20,33)	(53,60)	(2,94)	(13,74)	(27,69)

Figura 4.1 - Séries de dados do número de visitas pediátricas de emergência por causas respiratórias (RESP) e da concentração de MP_{10} (com valor de referência internacional de qualidade do ar)

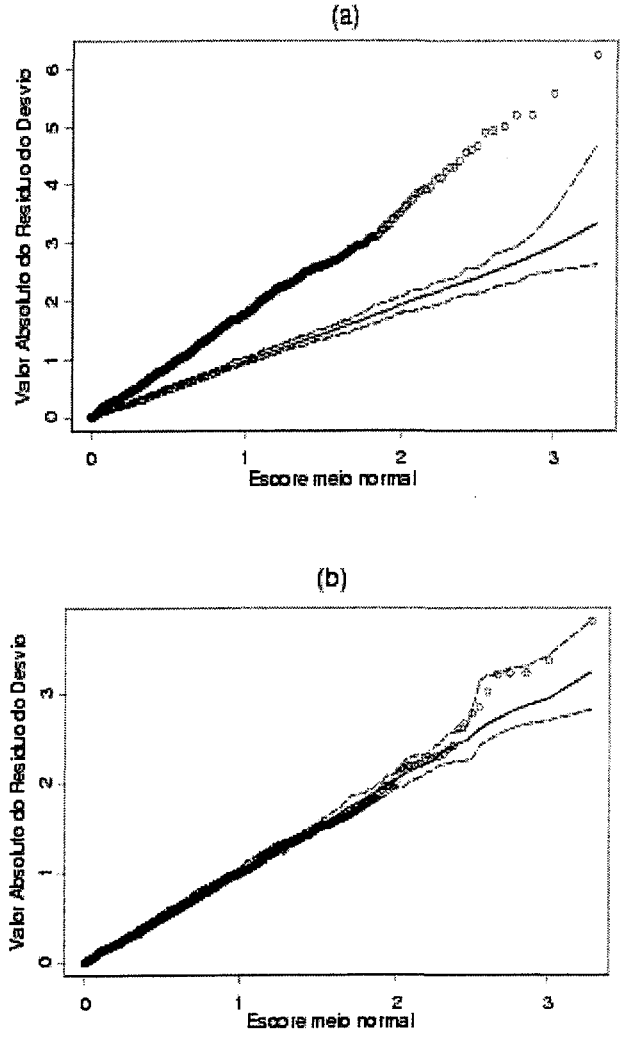


cada uma das variáveis nos dois modelos estudados. Assim, na escolha do modelo básico foram avaliados 49 ajustes, e do modelo final sete ajustes. Os menores valores da função de Akaike, maior significância das variáveis e melhor qualidade de ajuste do modelo, determinaram a escolha das seguintes médias móveis: de dois dias para a temperatura e de sete dias para a umidade e a concentração de MP_{10}, em ambos os modelos. O

conjunto final de dados analisados contou com 611 observações. Toda a análise estatística que se segue foi realizada com o auxílio do software S-Plus 4.5.

Primeiramente, consideramos o modelo de regressão log-linear de Poisson. O modelo básico foi ajustado e associação significativa foi observada para todas as variáveis de confundimento (p<0,001). A inclusão do poluente no modelo básico de Poisson é significante (p<0,001). No entanto, o desvio desse modelo é igual a 1949, 84 com 577 graus de liberdade (g.l.), indicando que o modelo de regressão de Poisson não fornece um bom ajuste. Além disso, todos os pontos do gráfico de probabilidade meio-normal dos resíduos do desvio (Figura 4.2a) estão acima do envelope simulado, levando à conclusão de que há fortes evidências de superdispersão.

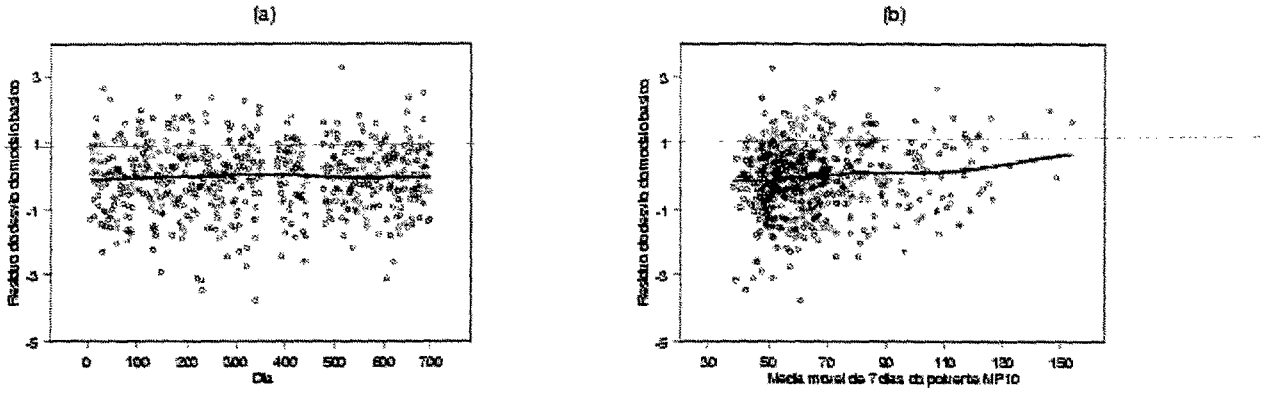
Figura 4.2 - Gráfico de probabilidade meio-normal dos valores absolutos dos resíduos do desvio dos modelos de regressão de Poisson (a) e binomial negativo (b)



Consideremos agora o modelo de regressão binomial negativo com função de ligação logarítmica e parâmetro de dispersão k desconhecido. O modelo básico foi ajustado e associação significativa foi observada

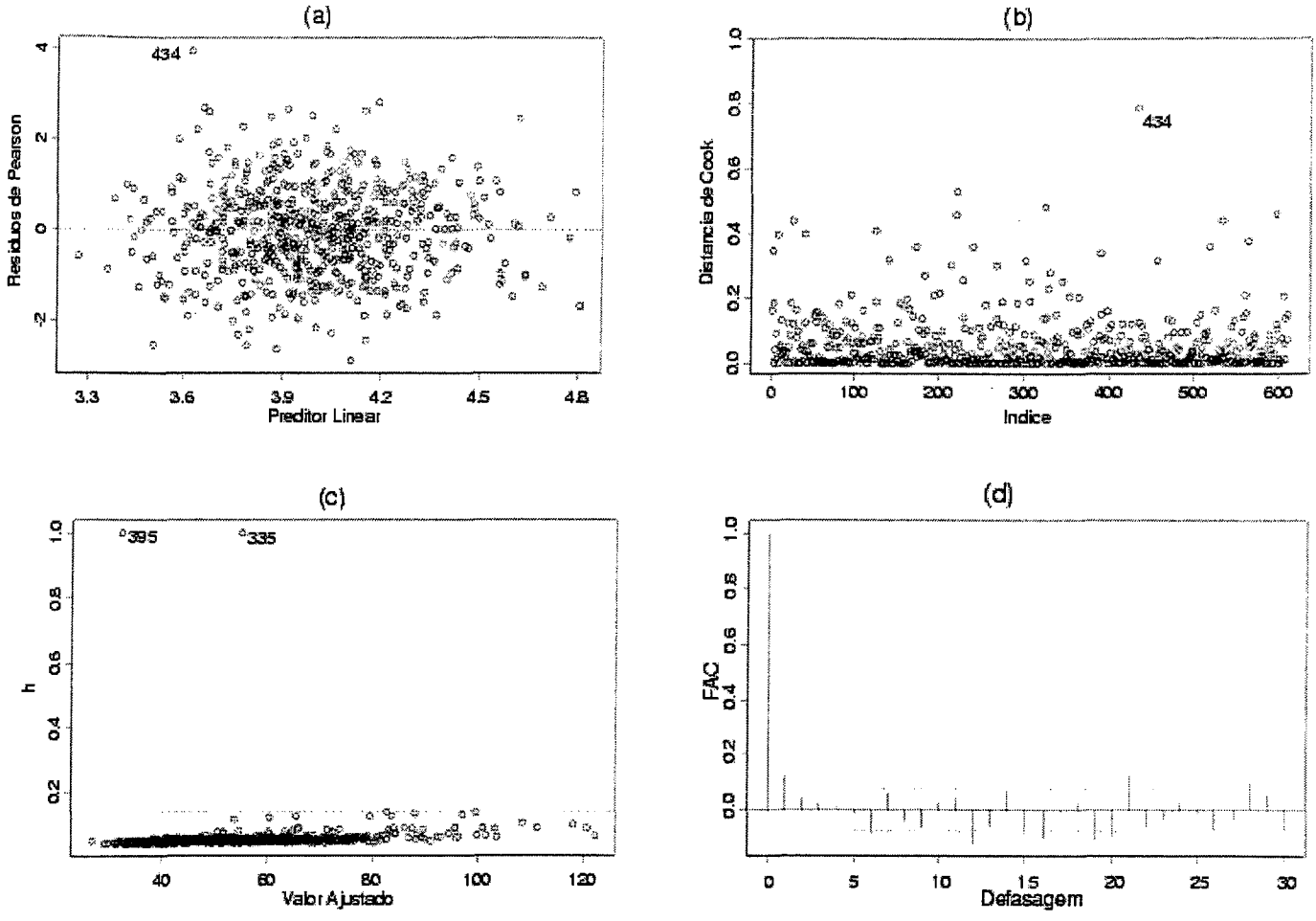
para todas as variáveis de confundimento (p<0,001). O gráfico dos resíduos do desvio desse modelo básico contra o número de dias desde o início do estudo está apresentado na Figura 4.3a, juntamente com a curva de alisamento do tipo *loess* (Fox, 1947). Essa curva está muito próxima de uma reta paralela ao eixo das abscissas, indicando que a tendência temporal dos dados parece controlada no modelo ajustado. Na Figura 4.3b apresentamos o gráfico desses resíduos contra a média móvel de sete dias do poluente MP_{10} com a curva de alisamento do tipo *loess*. Essa curva sugere uma relação linear crescente do número de visitas pediátricas de emergência por causas respiratórias e a concentração do poluente, após o controle das variáveis de confundimento em estudo. A inclusão do poluente no modelo básico binomial negativo é significante (p<0,001), o desvio desse modelo é igual a 625,35 com 577 g.l. e obtivemos k=24,76 com erro-padrão igual a 2,10. O gráfico de probabilidade meio-normal dos resíduos do desvio (Figura 4.2b) indica que o modelo binomial negativo é muito mais apropriado para nossos dados que o modelo de Poisson.

Figura 4.3 - Gráfico dos resíduos do desvio do modelo binomial negativo básico contra os dias observados (a) e contra a média móvel de sete dias do poluente MP_{10} (b), com curvas de alisamento *loess*



O gráfico de resíduos de Pearson contra os preditores lineares (Figura 4.4a) não indica nenhuma evidência de falta de ajuste, e somente a observação 434 se destaca por apresentar maior valor que as demais, indicando que o número observado de visitas pediátricas de emergência foi muito maior que o esperado nesse dia. A Figura 4.4b apresenta um gráfico da distância de Cook para o ajuste binomial negativo. Observamos que a distância de Cook da observação 434 também se destaca das demais. A Figura 4.4c apresenta o gráfico de alavancas h_i contra os valores ajustados. Observamos que os pontos com alta alavanca ($h_i > 2{,}5p/n$) correspondem às únicas observações de maio e agosto de 1992 no nosso conjunto de dados, e essa é a razão pela qual elas são potencialmente influentes. No entanto, como a eliminação de todas essas observações discrepantes não altera as conclusões inferenciais da análise, elas não serão desconsideradas. A Figura 4.4d apresenta o gráfico da função de autocorrelação dos resíduos de Pearson, onde podemos observar que não há evidências de que a suposição de independência do modelo de regressão binomial negativo seja falsa.

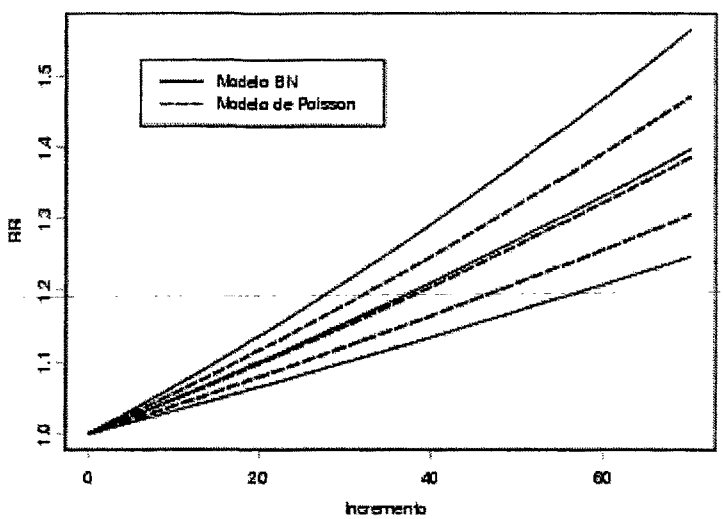
Figura 4.4 - Gráfico dos resíduos de Pearson contra os valores ajustados (a), gráfico da distância de Cook (b), gráfico da medida de alavanca *h* contra os valores ajustados (c) e gráfico da função de autocorrelação (FAC) dos resíduos de Pearson (d) para o modelo binomial negativo



As estimativas do parâmetro relacionado ao poluente MP_{10} em ambos os modelos avaliados foram positivas indicando que quanto maior a concentração desse poluente, maior será o valor esperado do número de atendimentos pediátricos de emergência diários por causas respiratórias. O risco relativo ao poluente MP_{10} que, como descrito na seção 2, é uma medida de interesse em estudos epidemiológicos, e respectivo intervalo de 95% de confiança foram calculados para incrementos de até $70\mu g/m^3$, construindo assim a curva de risco relativo estimada, que está apresentada na Figura 4.5. Para efeito de comparação, foi incluída no gráfico a curva de risco relativo estimada através do modelo de Poisson. Apesar de as estimativas pontuais dos riscos relativos calculados pelos dois modelos estarem muito próximas, os intervalos de confiança diferem consideravelmente de um modelo para outro. Isso ilustra perfeitamente o que foi discutido na seção 1: uma conseqüência importante

de se ignorar a superdispersão é que os erros-padrão obtidos através do modelo de regressão de Poisson são incorretos e subestimam a variabilidade dos estimadores de seus parâmetros. O modelo de Poisson pode indicar falsa associação significante entre a resposta e as covariadas do modelo. Este é um exemplo de que conclusões erradas podem ser tiradas quando as suposições de um modelo teórico são violadas, e daí a importância de uma análise de diagnóstico para verificar a qualidade e a adequação do ajuste de um modelo estatístico. Esse fato não ocorreu nessa análise pela alta significância de todas as variáveis em estudo, não prejudicando assim os resultados apresentados em Lin e outros (1999).

Figura 4.5 - Estimativas pontuais das curvas de risco relativo (RR) associadas ao poluente MP_{10} e respectivos intervalos de 95% de confiança para os modelos binomial negativo (BN) e de Poisson



Para o modelo binomial negativo, as estimativas do risco relativo (intervalo de 95% de confiança) para o incremento de $10\mu g/m^3$ na média móvel de sete dias do poluente e para o incremento da distância interquartílica (75-25 percentil) dessa mesma média móvel que, nesse caso, é igual a $29{,}2\mu g/m^3$, são 1,049 ([1,032;1,066]) e 1,150 ([1,097; 1,205]), respectivamente. Essa última medida foi calculada com o objetivo de representar a razão do número esperado de atendimentos pediátricos de emergência por causas respiratórias em um dia de "muita" poluição comparado com um dia de "pouca" poluição. Assim, podemos dizer, por exemplo, que o aumento da distância interquartílica na média móvel de sete dias do poluente MP_{10} está associado a um aumento médio estimado de aproximadamente 15% no número de atendimentos pediátricos de emergência por causas respiratórias, considerando constantes as demais covariadas do modelo.

5. Conclusões

Neste trabalho utilizamos um modelo de regressão binomial negativo para a análise de associação entre o número de atendimentos pediátricos de emergência diários por motivos respiratórios registrados no Instituto da Criança da USP, em São Paulo, e os níveis de poluição atmosférica, representados pela concentração de material particulado (MP_{10}), no período de maio 1991 a abril de 1993. Mostramos que esse modelo é bem mais adequado que o modelo de regressão usual de Poisson para o conjunto de dados considerado. Nossa análise, baseada no modelo binomial negativo, leva a evidências de associação positiva entre a concentração do poluente e o número de atendimentos. Este resultado reforça a idéia de que a poluição urbana é muito prejudicial à saúde, identificando a necessidade de medidas eficientes de controle de emissão de poluentes visando a proteger a saúde pública. Embora esta mesma conclusão seja alcançada através do modelo de Poisson, a variabilidade dos estimadores dos riscos relativos associados a incrementos no poluente é muito subestimada se este modelo é utilizado, levando a intervalos de confiança enganosos. Nosso estudo ilustra também a importância de uma análise de diagnóstico, em particular para detectar uma possível inadequação do modelo de Poisson quando os dados revelam variabilidade maior do que a suposta por tal modelo.

Referências bibliográficas

AITKIN, M., ANDERSON, D., FRANCIS, B. E HINDE, J. (1989). *Statistical Modelling in GLIM*. Oxford: Oxford University Press.

AKAIKE, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, AU-19, 716-722.

ATKINSON, A. (1985). *Plots, Transformations and Regression*. Oxford: Clarendon Press.

BRAGA, A.L.F, CONCEIÇÃO, G.M.S., PEREIRA L.A.A.; KISHI, H., PEREIRA, J.C.R., ANDRADE, M.F., GONÇALVES, F.L.T., SALDIVA,P.H.N. E LATORRE, M.R.D.O. (1999). Air Pollution and pediatric respiratory admissions in São Paulo, Brazil. *Journal of Environmental Medicine*, 1, 95-102.

BRAGA, A.L.F., SALDIVA, P.H.N., PEREIRA, L.A.A., MENEZES, J.J.C., CONCEIÇÃO, G.M.S., LIN, C.A., ZANOBETTI, A., SCHWARTZ, J. E DOCKERY, D.W. (2001). Health effects of air pollution exposure on children and adolescents in São Paulo, Brazil. *Pediatric Pulmonology*, 31, 106-113.

BRESLOW, N. (1984). Extra-poisson variation in log-linear models. *Applied Statistics*, 33, 38-44.

CAMERON, A.C. E TRIVEDI, P.K. (1998).*Regression Analysis of Count Data*. Econometric Society Monographs 30. Cambridge: Cambridge University Press.

CETESB (2001). *Relatório de qualidade do ar no Estado de São Paulo*. São Paulo, CETESB.

CONCEIÇÃO, M.S.C., MIRAGLIA, S.G.E.K., KISHI, H.S., SALDIVA, P.H.N. E SINGER, J.M. (2001). Air Pollution and child mortality: a time-series study in São Paulo, Brazil. *Environmental Health Perspectives*, 109 (suppl 3), 347-350.

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

DEAN, C. (1992). Testing overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, 87, 451-457.

FERRARI, S.L.P. DAVID, J.S.E, ANDRÉ, P.A. E PEREIRA, L.A.A. (2002). Overdispersed Regression Models for Air Pollution and Human Health. Em: *Statistical Data Analysis Based on the L_1-Norm and Related Methods (Statistics for Industry and Technology*, Y. Dodge (Editor), 429-438. Basel: Birkhäuser.

FOX, J. (1947). *Nonparametric Simple Regression: Smothing Scatterplots.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-130. Thousand Oaks, CA: Sage.

HARDIN, J. E HILBE, J. (2001). *Generalized Linear Models and Extensions.* College Station: Stata Press.

HINDE, J. E DEMÉTRIO, C. (1998). *Overdispersion: models and estimation. Computational Statistics and Data Analysis*, 27, 151-170.

LAMBERT, D. E ROEDER, K. (1995). Overdispersion diagnostics for generalized linear models. *Journal of the American Statistical Association*, 95, 1225-1237.

LAWLESS, J. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, 15, 209-225.

LIN, C.A., MARTINS, M.A., FARHAT, S.C., POPE III, C.A., CONCEIÇÃO, G.M.S., ANASTASIO, V.M., HATANAKA, M., ANDRADE, W.C., HAMAUE, W.R., BÖHM, G.M. E SALDIVA, P.H.N. (1999). Air pollution and respiratory illness of children in São Paulo, Brazil. *Paediatric and Perinatal Epidemiology*, 13, 475-488.

MCCULLAGH, P. E NELDER, J.A. (1989).*Generalized Linear Models* (The Monographs on Statistics and Applied Probability, Vol 37), 2a edição. London: Chapman and Hall.

MCNENEY, B. E PETKAU, J. (1994). Overdispersed Poisson regression models for studies of air pollution and human health. *The Canadian Journal of Statistics*, 22, 421-440.

MORETTIN, P.A. E TOLOI, C.M.C. (1985). *Previsão de Séries Temporais.* São Paulo: Atual.

MORGAN, G., CORBETT, S.,WLODARCZYK, J. E LEWIS, P. (1998). Air pollution and daily mortality in Sydney, Australia, 1989 through 1993. *American Journal of Public Health*, 88, 759-764.

NETER, J., KUTNER, M.H., NACHTSHEIM, C.J. E WASSERMAN, W. (1996). *Applied Linear Statistical Models* (Irwin Series in Statistics), 4a edição. Chicago: Irwin.

PAULA, G.A. (2002). *Modelos de Regressão.* IME-USP. Não publicado. Disponível em `http://www.ime.usp.br/~giapaula/Book.pdf`.

PEDAN, A. (2001). Analysis of count data using the SAS system. *Proceedings of the 26th SAS Users Group International Conference*, P247-26, 1-6. Disponível em `www2.sas.com/proceedings/sugi26/p247-26.pdf`.

RUSTAGI, J.S. (1994). *Optimization Techniques in Statistics.* San Diego: Academic Press.

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

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

SCHWARTZ, J. (1994). Nonparametric smoothing in the analysis of air pollution and respiratory illness. *The Canadian Journal of Statistics*, 22, 471-487.

SCHWARTZ, J., SPIX, C. TOULOUMI, G., BACHÁROVÁ, L., BARUMAMDZADEH, T., LE TERTRE, A., PIEKARKSI, T., PONCE DE LEON, A., PÖNKÄ, A. ROSSI, G., SAEZ, M. E SCHOUTEN, J.P. (1996). Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *Journal of Epidemiology and Community Health*, 50 (Suppl 1), S3-S11.

SINGER, J.M., ANDRÉ, C.D.S., LIMA, L.P. E CONCEIÇÃO, G.M.S. (2002). Atmospheric Pollution and Mortality in São Paulo. Em: *Statistical Data Analysis Based on the L_1-Norm and Related Methods (Statistics for Industry and Technology*, Y. Dodge (Editor), 439-450. Basel: Birkhäuser.

SVETLIZA, C.F. E PAULA, G.A. (2001). On diagnostics in log-linear negative binomial models. *Journal of Statistical Computation and Simulation*, 71, 231-243.

VENABLES, W. N. E RIPLEY, B. D. (1999). *Modern Applied Statistics with S-Plus*. Third edition. New York: Springer.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, 61, 439-447.

WONG,C.-M., MA, S., HEDLEY, A.J. E LAM, T.-H. (2001). Effect of air pollution on daily mortality in Hong Kong. *Environmental Health Perspectives*, 109, 335-340.

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

Agradecimentos

Agradecemos o suporte fincanceiro do CNPq e da FAPESP. Agradecemos também o Laboratório de Poluição Atmosférica Experimental da Faculdade de Medicina da Universidade de São Paulo e, em especial, a Gleice M. S. Conceição, pela fonte de dados.

Abstract

Practitioners frequently use Poisson regression models to analyse the association between count data and a set of covariates. The Poisson distribution imposes, however, an important restriction to the model: it assumes that, given the values of the covariates, the mean and the variance of the response variable are equal. An alternative model that allows the response variance to be greater than the respective mean is based on a negative binomial distribution. In this paper we use a negative binomial log-linear regression model to analyze the association between pediatric emergency hospital visits for respiratory diseases and the levels of air pollution in São Paulo city. The main aspects addressed are inference on the regression and dispersion parameters, model adequacy checking and estimation of relative risks.

Key words: air pollution, diagnostics, generalized linear models, human health, negative binomial distribution, overdispersion, Poisso distribution.

Construção de uma rede ótima para o monitoramento de um sistema bioecológico a partir do variograma

Gladys Elena Salcedo Echeverry*
Luis Hernando Hurtado Tobón*
María Dolly García González*

Resumo

Neste trabalho apresenta-se uma metodologia para construir uma rede de monitoramento em um sistema bioecológico, aproveitando o fato de que na técnica do Krigagem o quadrado médio do erro de predição em cada ponto independe do valor da variável; portanto, a partir de uma amostra inicial da variável em N pontos georreferenciados e do variograma, pode-se eliminar pontos por etapas onde em cada etapa o ponto eliminado é aquele onde a variável apresenta o mínimo erro de predição, e portanto, pode se predizer por Krigagem. O procedimento se repete até identificar um conjunto de n pontos, $n \leq N$, que conformam a rede de monitoramento. O procedimento é ótimo no sentido que identifica um valor mínimo de n pontos a partir dos quais é possível fazer interpolação por Krigagem.

Será feita uma aplicação da construção de uma rede ótima de monitoramento para a variável Salinidade na lagoa "Ciénaga Grande de Santa Marta" localizada no litoral norte da Colômbia.

Palavras-chave: Variograma, Horizonte de autocorrelação espacial, Erro de predição, Krigagem, Rede de monitoramento.

1. Introdução

No estudo dos sistemas bioecológicos, a coleta de informação é um processo lento e custoso que merece a busca de estratégias para sua otimização. Este fato é de particular importância quando um sistema bioecológico é submetido a um seguimento e a informação deve ser tomada repetidamente através do tempo; justifica-se então a identificação de um conjunto de pontos de amostragem que produza a maior informação do sistema com o

* Mestrado em Biomatemáticas Universidade do Quindío AA 460 Armenia - Colômbia

menor esforço possível; este conjunto denomina-se uma **rede de monitoramento**. Esta rede é ótima quando se identifica um número mínimo de pontos que permitam fazer interpolação por Krigagem, tendo pelo menos q vizinhos para interpolar qualquer ponto.

Existem diversas metodologias para construir uma rede de monitoramento, por exemplo, Caselton e Zidek (1984), Warrick e Myers (1987), Samper e Carrera (1990); um procedimento comum consiste em agregar pontos de amostragem a uma rede inicial. Particularmente Samper e Carrera (1990), utilizando a técnica de Krigagem e o fato de que o quadrado médio do erro de predição independe da medida da variável no ponto, porém depende da estrutura de autocorrelação espacial, propõem agregar pontos a uma rede inicial de tal forma que minimizem o erro de predição.

Nossa metodologia é similar à de Samper e Carrera, já que aproveita a mesma propriedade do quadrado médio do erro de predição de Krigagem, porém partindo de uma rede inicial de pontos de amostragem suficientemente grande. Um único variograma é construído a partir desta rede inicial e um procedimento de eliminação de pontos é feito por etapas.

Na segunda seção se apresenta a estrutura de correlação espacial, a terceira fornece alguns conceitos básicos do método de Krigagem, na quarta se descreve o processo de construção da rede e finalmente, na quinta seção se aplica o procedimento para encontrar uma rede ótima para o ecossistema "Ciénaga Grande de Santa Marta" utilizando a variável Salinidade.

2. A estrutura de correlação espacial

A estrutura de correlação de um processo espacial estacionário, $\{Z(s);\ s \in D, D \subset R^n\}$, é dada basicamente pelo **variograma** definido através de

$$2\gamma(h) = E[Z(s+h) - Z(s)]^2,$$

onde h é um vetor com norma e direção. Quando se dispõe de uma amostra do processo, e se fixa uma direção, o variograma nesta direção se estima através de

$$\hat{2\gamma}(h) = \frac{\sum^{n(h)} [Z(s+h) - Z(s)]^2}{n(h)}$$

onde $n(h)$ corresponde ao número de pontos separados por uma distância $\|h\|$.

Os modelos mais comums de variogramas são o *efeito pepita, esférico, exponencial, gaussiano e linear*. Para mais detalhes consultar Cressie(1993), Wackernagel(1995), Stein(1999), Hurtado et al.(1996).

3. A técnica de Krigagem

O Krigagem é uma técnica de interpolação espacial que permite predizer o valor de uma variável aleatória Z em um determinado ponto de uma região a partir dos valores $Z_1, Z_2, \ldots, Z_n$ medidos em n pontos diferentes da região, $t_1, t_2, \ldots, t_n$. O valor predito da variável Z_0 no ponto t_0, chamado $\hat{Z}_0$, obtem-se por meio de uma combinação linear dos valores de Z nos n pontos observados:

$$\hat{Z}_0 = \sum_{i=1}^{n} \lambda_i Z_i .$$

Os valores dos λ_i, $i = 1,2,\ldots,n$, podem ser determinados para produzir o melhor preditor linear que seja não viciado, considerando as seguintes condições:

i) $E(\hat{Z}_0 - Z_0) = 0$ (1)

ii) O quadrado médio do erro de predição de $\hat{Z}_0$ seja mínimo.

Se além disso o processo é fracamente estacionário tem-se que $E(Z) = \mu$ e $V(Z) = c_0$ onde Z é o valor do processo medido numa posição arbitrária.

A condição (i) implica em

$$E(\hat{Z}_0) - E(Z_0) = E(\sum_{i=1}^{n} \lambda_i Z_i) - E(Z_0) = \mu(\lambda_1 + \lambda_2 + \ldots + \lambda_n - 1) = 0 ,$$

conseqüentemente $\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_n - 1 = 0$, ou seja, $\sum_{i=1}^{n} \lambda_i = 1$.

Para analisar as conseqüências da condição (ii), se parte do fato de que o quadrado médio do erro de predição (QME) é dado por

$$QME = E(\hat{Z}_0 - Z_0)^2 = V(\hat{Z}_0 - Z_0) = V(\hat{Z}_0) + V(Z_0) - 2Cov(\hat{Z}_0, Z_0) . \quad (2)$$

Desenvolvendo cada um dos termos do lado direito de (2) obtém-se:

- $$V(\hat{Z}_0) = V(\sum_{i=1}^{n} \lambda_i Z_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j Cov(Z_i, Z_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j c_{ij} \quad (3)$$

onde $c_{ij} = Cov(Z_i, Z_j)$.

- $V(Z_0) = c_0$ (4)

- $$Cov\,(\hat{Z}_0, Z_0) = Cov\,(\sum_{i=1}^{n} \lambda_i Z_i, Z_0) = \sum_{i=1}^{n} \lambda_i Cov\,(Z_i, Z_0) = \sum_{i=1}^{n} \lambda_i c_{i0}\ . \quad (5)$$

Substituindo (3), (4) e (5) em (2), obtém-se:

$$QME = V(\hat{Z}_0 - Z_0) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j c_{ij} + c_0 - 2\sum_{i=1}^{n} \lambda_i c_{i0}. \quad (6)$$

Observe que devido à condição (i), o QME se transforma na variância do erro de predição. Para minimizar esta variância com respeito aos λ_i e com a restrição que $\sum_{i=1}^{n} \lambda_i = 1$, deve-se minimizar a função

$$g(\lambda_1, \lambda_2, \ldots, \lambda_n, \mathrm{M}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j c_{ij} + c_0 - 2\sum_{i=1}^{n} \lambda_i c_{i0} + \mathrm{M}(\sum_{i=1}^{n} \lambda_i - 1) \quad (7)$$

onde M é um multiplicador de Lagrange.

Os λ_i que minimizam a função $g(\lambda_1, \lambda_2, \ldots, \lambda_n, \mathrm{M})$ correspondem à solução do sistema das $n+1$ equações

$$\sum_{i=1}^{n} \lambda_j c_{ij} + \mathrm{M} = c_{io}, \quad i = 1,2,\ldots,n \quad (8)$$

$$\sum_{i=1}^{n} \lambda_i = 1$$

que matricialmente é representado por

$$\begin{bmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2n} & 1 \\ c_{31} & c_{32} & c_{33} & \cdots & c_{3n} & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nn} & 1 \\ 1 & 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_n \\ \mathrm{M} \end{bmatrix} = \begin{bmatrix} c_{10} \\ c_{20} \\ c_{30} \\ \vdots \\ c_{n0} \\ 1 \end{bmatrix}.$$

Substituindo (8) em (6), obtém-se a variância mínima do erro de predição:

$$V(\hat{Z}_0 - Z_0) = c_0 - \sum_{i=1}^{n} \lambda_i c_{i0} + \mathrm{M}.$$

Sob estacionariedade e quando o processo é isotrópico a covariância entre duas variáveis localizadas nos pontos t_i e t_j depende somente da distância $h = |t_i - t_j|$, isto é,

$c_{ij} = Cov(Z_i, Z_j) = C(h)$.

A partir de $C(h)$ constroi-se uma outra função $\gamma(h)$ denominada o variograma onde $\gamma(h) = c_0 - C(h)$ ou $\gamma_{ij} = c_0 - c_{ij}$, onde c_0 é a variância devida aos erros de medição.

Em termos do variograma, o problema de encontrar os λ_i que produzam um preditor não viciado e minimizem a variância do erro de predição, é equivalente a encontrar a solução do sistema

$$\begin{bmatrix} \lambda_{11} & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \cdots & \gamma_{2n} & 1 \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & \cdots & \gamma_{3n} & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \gamma_{n3} & \cdots & \gamma_{nn} & 1 \\ 1 & 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_n \\ \mathrm{M} \end{bmatrix} = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \\ \gamma_{30} \\ \vdots \\ \gamma_{n0} \\ 1 \end{bmatrix}$$

e neste caso a variância do erro de predição é dada por

$$V(\hat{Z}_0 - Z_0) = \sum_{i=1}^{n} \lambda_i \gamma_{i0} + \mathrm{M}.$$

Para mais detalhes consultar Cressie(1993), Krige(1951) e Hurtado et al. (1996).

4. A rede de monitoramento

Para construir a rede de monitoramento começa-se por supor que se tem informação sobre N pontos de amostragem georrefenciados que formam um conjunto **Ro** e que além disso, resume toda a informação sobre um sistema bioecológico. Trata-se de identificar em **Ro**, um subconjunto **R** de n pontos, que contenha aproximadamente a mesma informação que **Ro** embora com n muito menor que N. O subconjunto **R** é a rede de monitoramento.

A incerteza produzida ao predizer com m pontos, $n \leq m \leq N$, um valor de Z em algum ponto de **Ro**, mede-se através do quadrado médio do erro de predição de Krigagem dado por

$$V(\hat{Z}_0 - Z_0) = \sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j c_{ij} + c_0 - 2\sum_{i=1}^{m}\lambda_i c_{i0}$$

ou em termos do variograma por

$$V(\hat{Z}_0 - Z_0) = \sum_{i=1}^{m}\lambda_i\gamma_{i0} + \mathrm{M}$$

onde os λ_i, $i = 1,2,\ldots,m$ são obtidos das equações de Krigagem e os γ_{i0} do variograma.

Observa-se nestas equações que a incerteza diminui quando os c_{ij} são pequenos e os c_{io} são grandes. Também se observa que a incerteza independe do valor da variável em cada ponto, e, portanto, é possível calcular a variância do erro de predição sem conhecer a variável no ponto; o que se necessita é conhecer a estrutura de correlação que se obtém do variograma.

Partindo destas considerações se constrói um procedimento que parte da informação sobre uma variável de interesse em N pontos que são candidatos a entrar na rede e aproveita-se esta informação para estimar o variograma.

A construção da rede de monitoramento baseia-se num procedimento de eliminação de pontos por etapas, que é controlado em cada etapa, pelo incremento da incerteza e pela relação que existe entre o horizonte de autocorrelação espacial e as distâncias entre os pontos que permanecem. O procedimento tem duas regras de parada: a primeira é fixando o número de pontos que se desejam na rede e a segunda, é identificando o número mínimo de pontos que são necessários para fazer interpolação por Krigagem. Isto equivale a deter o processo quando aparece um ponto isolado de ordem q, entendido neste caso, como um ponto com menos de q vizinhos a distâncias menores do horizonte de autocorrelação espacial.

Esta forma de construir a rede garante a utilização de um único variograma que é estimado com o maior número de pontos possíveis, a amostra inicial, e portanto não deve mudar através do processo.

Concretamente o procedimento parte de um conjunto **Ro** de N pontos de amostragem georreferenciados, escolhidos a critério de especialistas e que supostamente reúnem informação suficiente sobre um sistema bioecológico para uma variável Z. Partindo desta informação, constroi-se um variograma $\gamma(h)$ que expressa a estrutura de autocorrelação espacial para a variável de interesse Z, estima-se o horizonte de autocorrelação e logo escolhe-se o conjunto de n pontos que constituem a rede de monitoramento através do procedimento iterativo a seguir:

1. Suprima um dos pontos de **Ro** e calcule a variância do erro de predição neste ponto a partir dos $N-1$ pontos restantes.

2. Coloque novamente o ponto que suprimiu.

3. Repita os passos anteriores para todos os pontos de **Ro**.

4. Identifique o ponto de **Ro** cuja variância do erro de predição seja mínima utilizando o variograma estimado com todos os pontos iniciais.

5. Suprima o ponto identificado no item 4.

Quando se fixa a dimensão da rede por critério de especialistas ou por limitação de recursos para o monitoramento, o procedimento para depois de eliminar τ pontos e, neste caso, $n = N - \tau$.

Adotando o procedimento de otimizar a rede no sentido de minimizar o valor de n, o procedimento deve realizar também os passos a seguir:

6. Calcule as distâncias entre os pontos que permanecem e verifique se nenhum deles é um ponto isolado de ordem q.

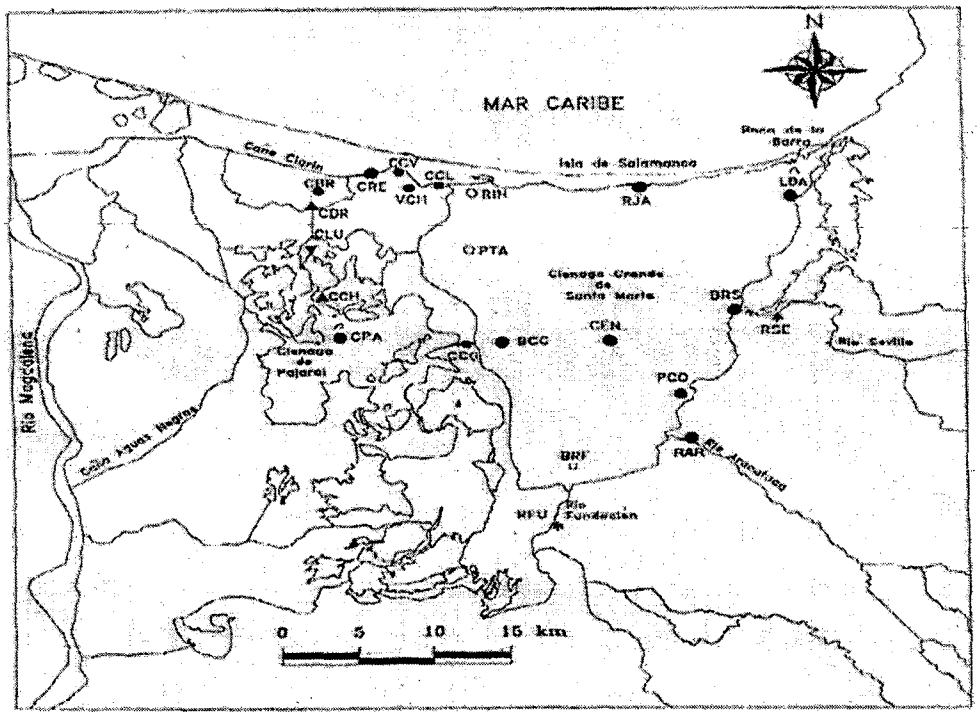
7. Repita o procedimento desde o passo 1 levando em consideração que o conjunto inicial **R** tem, em cada etapa, um ponto a menos.

Este procedimento termina quando aparece o primeiro ponto isolado. Finalizado qualquer dos procedimentos, o conjunto final **R** de n pontos, é a **rede ótima de monitoramento.**

5. Aplicação

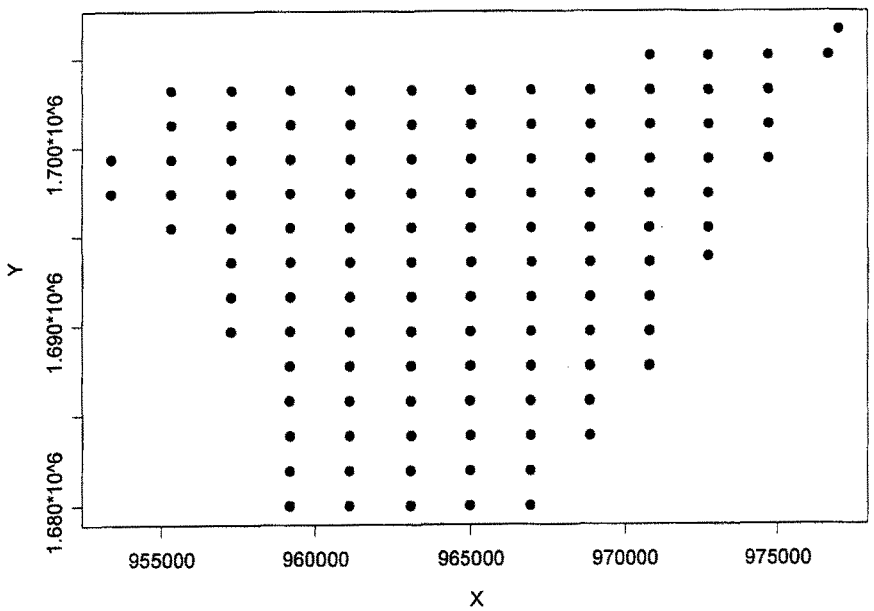
Para aplicar a metodologia apresentada na seção anterior, se encontra uma rede ótima de monitoramento no sistema bioecológico denominado "La Ciénaga Grande de Santa Marta" (CGSM), localizada sobre a costa caribenha na Colômbia entre os departamentos do Atlántico e Magdalena, ver Figura 1. Pelas suas características geomorfológicas, a CGSM é o maior ecossistema lagoa-estuarino do pais. Além do seu significado desde o ponto de vista bioecológico, destaca-se também uma grande importância social, pois da CGSM depende uma população de aproximadamente 300 000 habitantes. Por estes e outros motivos, o "Instituto de Investigaciones Marinas" de Punta Betín INVEMAR, leva pelo menos 25 anos monitorando diferentes aspectos da CGSM e suas zonas vizinhas. Maiores detalhes podem ser consultados em Botero e Mancera (1996).

Figura 1 - Mapa da Ciénaga Grande de Santa Marta – Colômbia



Para analisar o comportamento espacial de um conjunto de variáveis físicoquímicas, e para definir uma rede ótima de monitoramento, o INVEMAR planejou, em dezembro de 1995, uma amostragem composta por 114 pontos regularmente espaçados, cobrindo aproximadamente a área da CGSM. Nosso conjunto inicial **Ro** é composto por estes 114 pontos cuja localização espacial aparece na Figura 2. Em cada uma destas 114 estações registrou-se informação sobre as variáveis Salinidade, Nitrito, Ortofosfato, Nitrato, Silicato, Amônio, Concentração de Oxigênio e Clorofila.

Figura 2 - A rede inicial Ro na CGSM



Para construir a rede ótima se escolhe a variável Salinidade já que esta é uma das variáveis que mudou mais abruptamente através do tempo e teve a maior relevância em diferentes estudos feitos na lagoa, Botero e Mancera (1996). A Salinidade é dada em porcentagem e os dados aparecem na coluna 3 do Apêndice 1.

5.1. Análise Exploratória dos Dados

Antes de aplicar o procedimento descrito na seção 4, realiza-se uma análise exploratória dos dados de Salinidade na CGSM.

A partir de gráficos de Ramo e Folhas se observa a distribuição dos dados e claramente observa-se a presença de dois valores atípicos que alteram a distribuição, Figura 3(a). Esta distribuição melhora depois de eliminar estes valores atípicos, Figura 3(b), no entanto, observa-se uma possível tendência dos dados.

Figura 3 - Ramo e Folhas dos dados de Salinidade

0	0
0	22222223333333333
0	44444444444444444455555555555555555555555555
0	66666666666666666666666666666666666666777777777777
0	9
1	
1	
1	
1	
1	
2	
2	
2	
2	
2	
3	
3	
3	55

(a)

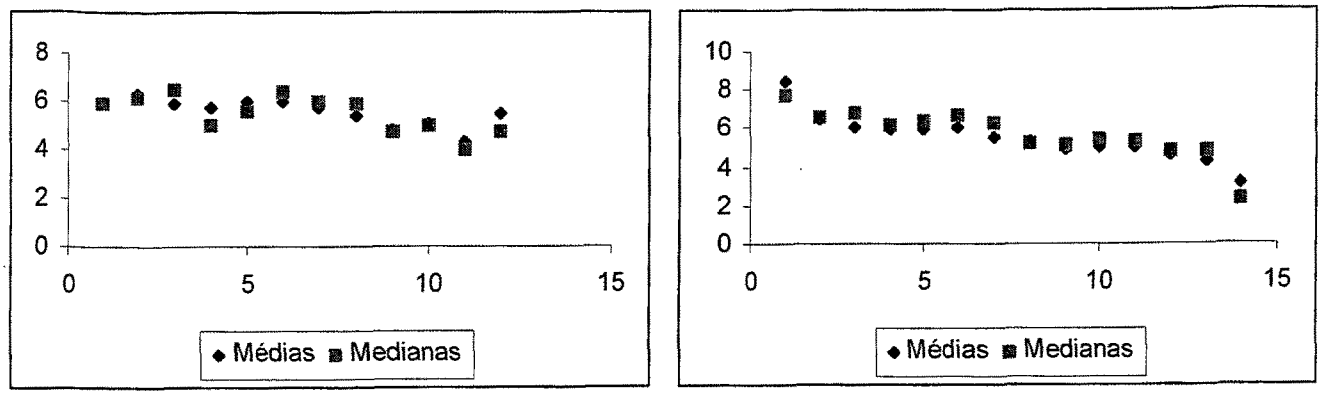
0	9
1	
2	0234568
3	1234667999
4	022455666788899999
5	012223334444555788899999999
6	000011112333333555666677777778899999
7	000011234557
8	
9	9

(b)

Para visualizar a tendência, veja na Figura 4 um gráfico de médias e medianas na direção Leste-Oeste (a) e na direção Norte-Sul (b), Cressie(1993).

Figura 4 -Médias e Medianas nas direções (a) Leste-Oeste, (b) Norte-Sul

(a) (b)



Observa-se claramente uma tendência linear dos dados na direção Norte-Sul devido a que pelo Norte entra no estuário água salgada do Mar Caribe, e em direção Sul desemborcam diferentes rios, fazendo com que a Salinidade seja maior para o Norte e menor para o Sul. Esta tendência foi corrigida mediante o modelo de Regressão Linear Simples na coordenada Y, Esterby (1993), dado por

$$Sal = -386 + 0.000232Y$$

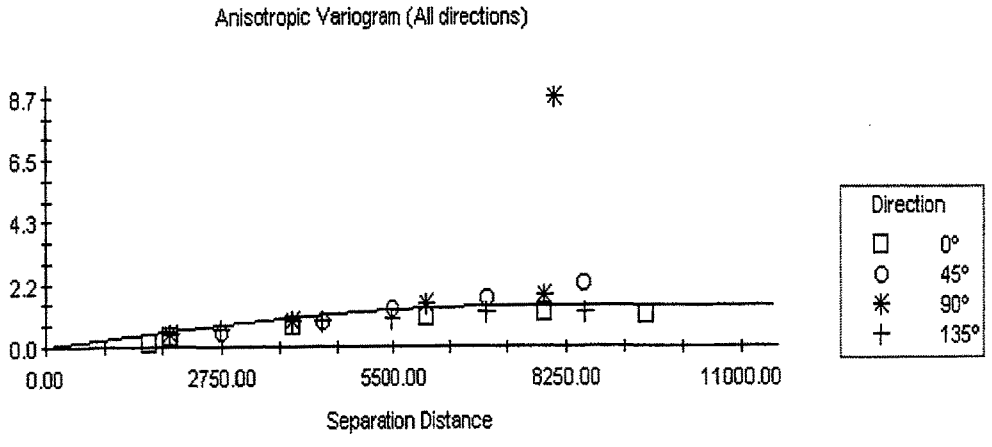
Os resíduos desta regressão aparecem na coluna 4 do Apêndice 1 e através da Figura 5, observa-se a sua distribuição. Para verificar a normalidade destes resíduos se aplica um teste de Kolmogorov-Smirnov, fornecendo um nível de significância maior de 0.15.

Figura 5 - Ramo e Folhas para a Salinidade Corrigida de Atípicos e Tendência

-4	7
-4	30
-3	96
-3	33
-2	
-2	42210000
-1	988886665
-1	433222211110
-0	99999988877777777775555
-0	4331111000
0	00022233334444
0	66778888889999
1	01112334
1	55566
2	11
2	7

A seguir, analisa-se a anisotropia através da comparação dos variogramas estimados nas direções 0, 45, 90 e 135 graus, Figura 6. O comportamento desta figura sugere um processo isotrópico. A quinta estrela na direção 90 graus, não é significativa, pois é um valor estimado com somente $n(h) = 1$.

Figura 6 - Modelos de Variogramas Anisotrópicos



5.2. Construção da rede ótima

Para encontrar a rede ótima inicialmente se ajusta um variograma a partir da rede inicial, aos dados de Salinidade corrigida por tendência (coluna 4 do Apêndice 1), levando em consideração os valores atípicos. O variograma estimado corresponde a um modelo esférico dado pela função

$$\gamma(h) = \begin{cases} s\left(\frac{h}{a}\right)\left[\frac{3}{2} - \frac{1}{2}\left(\frac{h}{a}\right)^2\right] & h \le a \\ s & h > a \end{cases}$$

onde s representa o patamar e a o horizonte de autocorrelação. Neste caso, $s = 7.15$ e $a = 7650$, o gráfico do variograma estimado aparece na Figura 7 e os valores de $h, \hat{\gamma}(h)$ e $n(h)$ na Tabela 1.

Figura 7 - Variograma Isotrópico da Salinidade

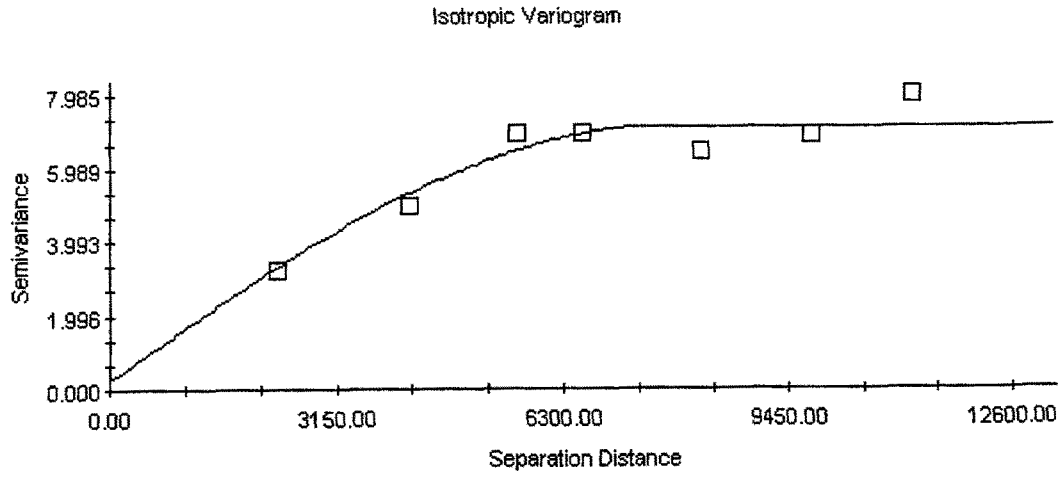
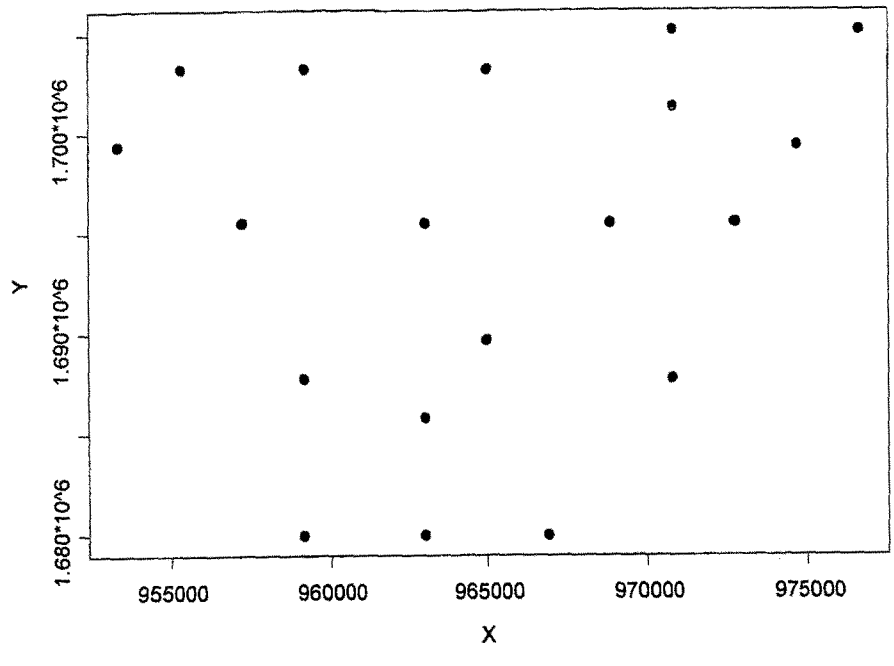


Tabela 1 - Valores do Variograma Ajustado

h	$\hat{\gamma}(h)$	$n(h)$
2319.35	3.243	384
4170.26	4.975	502
5639.27	6.935	301
6537.17	6.933	543
8196.06	6.445	699
9763.16	6.882	497
11172.14	7.985	671

Em seguida se aplica o procedimento que minimiza o valor de n para obter $n = 19$ pontos na rede ótima cuja localização aparece na Figura 8. Nesta rede, a variável Salinidade em qualquer ponto pode ser interpolada por Krigagem pelo menos com dois pontos vizinhos, de modo que neste caso $q = 2$. As coordenadas desta rede ótima aparecem no Apêndice 2.

Figura 8 - A Rede Ótima para a Variável Salinidade



Observe que na rede ótima está incluído um dos pontos atípicos, que é de grande importância já que este fornece informação do comportamento da Salinidade justamente no ponto de comunicação da CGSM com o mar Caribe.

Finalmente, para avaliar a rede se comparam os mapas de Krigagem utilizando os 114 pontos iniciais, Figura 9, e somente os 19 pontos da rede ótima, Figura 10. Em ambos os casos, utiliza-se o mesmo variograma esférico inicial. Observa-se a partir das Figuras 9 e 10 que os mapas de interpolação por Krigagem não apresentam muita diferença.

Um variograma esférico foi refeito com os 19 pontos onde $s = 46.78$ e $a = 7650$; a mudança no patamar se explica porque na rede ótima os dados obviamente apresentam maior variabilidade; no entanto, dado que 19 é o número mínimo de pontos para monitorar a Salinidade, este variograma não é o mais informativo.

Figura 9 - Krigagem para a Salinidade com os 114 pontos

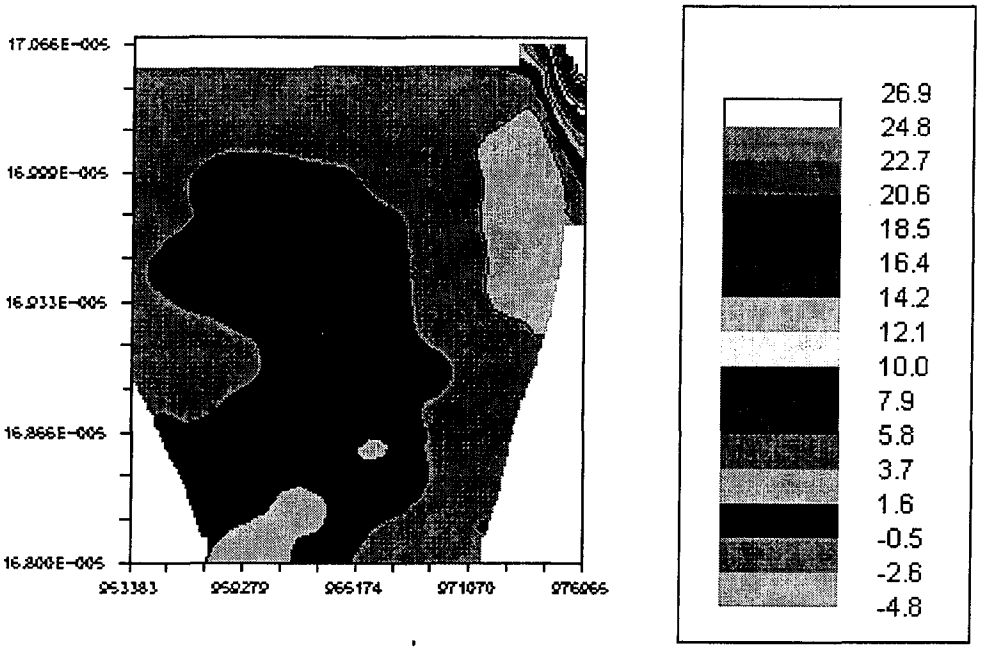
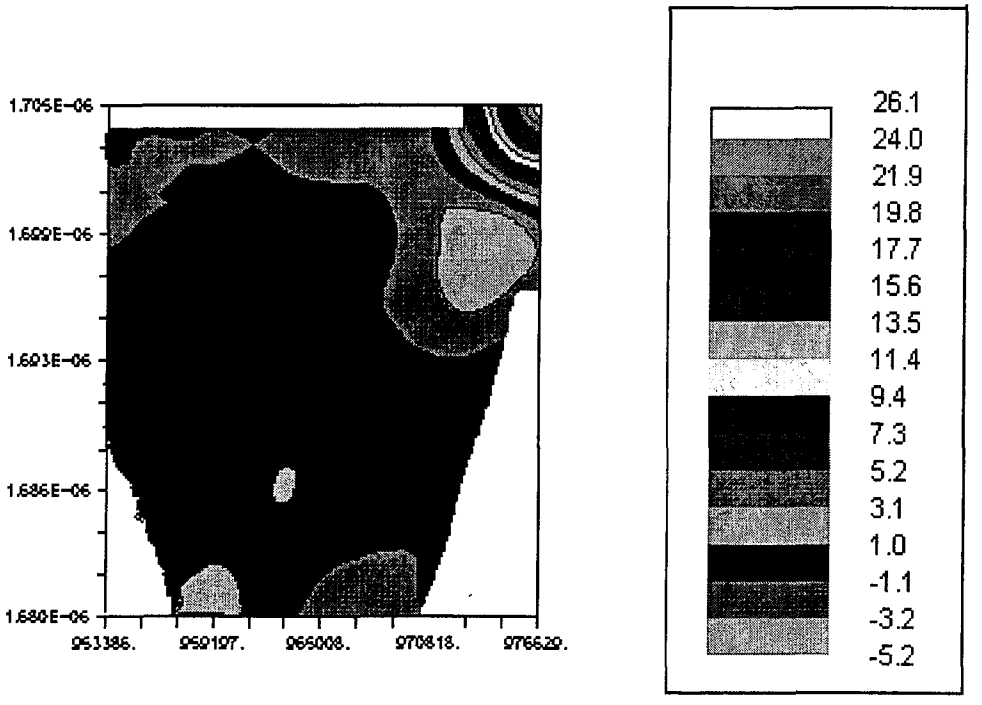


Figura 10 - Krigagem para a Salinidade com os 19 pontos ótimos



6. Conclusões

- A presença de dois valores atípicos de Salinidade altera a distribuição dos dados.

- A Salinidade na CGSM apresenta uma tendência na direção Norte-Sul a qual foi eliminada através de uma regressão linear simples na coordenada Y. Esta tendência se deve à dinâmica das águas doces e salgadas no estuário.

- Em ausência dos atípicos e eliminada a tendência, o fenômeno apresenta um comportamento isotrópico.

- A vantagem de partir de uma rede inicial consiste em aproveitar um único variograma construído com o maior número de pontos possíveis e que portanto não deve mudar em cada etapa da eliminação.

- nosso procedimento reduz a rede inicial de 114 pontos a uma rede ótima de 19 pontos.

- A interpolação por Krigagem a partir da rede ótima quase não varia comparado com aquela, utilizando os 114 pontos iniciais.

- procedimento para achar a rede ótima leva em conta os valores atípicos já que eles fornecem uma informação muito importante do ecossistema. Observa-se então, que na rede ótima aparece um destes valores.

Referências bibliográficas

BOTERO, L. AND E. MANCERA (1996). ***Síntesis de los Cambios de Origen Antrópico Ocurrido en los Últimos 40 años en la Ciénaga Grande de Santa Marta.*** **Revista Académica Colômbiana de Ciencias Exactas, Físicas y Naturales. 20(78), 465-473.**

CASELTON, W.F. AND ZIDEK, J.V. (1984) : ***Optimal Monitoring Network Designs.*** **Statistics & Probability Letters. 2. 223-227. North-Holland.**

CRESSIE N.A.C. (1993) : ***Statistical for Spatial Data,*** **John Wiley and Sons, New York.**

ESTERBY S. R. (1993) : ***Trend Analysis Methods for Environmental Data.*** **Environmetrics. 4(4). 459-481. USA.**

HURTADO L.H., SALCEDO G.E., SERRANO H.J. (1996) : ***Modelos Estadísticos para Información Autocorrelacionada.*** **Cuadernos de Biomatemáticas. 2. Universidad del Quindío. Armenia, Colômbia.**

KRIGE D.G. (1951): ***A Statistical Approach to Some Mine Valuation and Allied Problems at the Witwaterstrand*** **. Unpublished. M.s.c. Thesis, University of Witwaterstrand.**

SAMPER C. F., CARRERA R. J. (1990) : ***Geoestadística, Aplicaciones a la Hidrogeología Subterránea.*** **Centro Internacional de Métodos Numéricos en Ingeniería.Barcelona,España.**

STEIN, MICHAEL L. (1999) : *Interpolation of Spatial Data, Some Theory for Kriging.* Springer Verlag. New York.

WACKERNAGEL HANS. (1995) : Multivariate Geostatistics, An Introduction with Applications. Springer Verlag. Berlin.

WARRICK, A. W. AND MYERS, D. E. (1987). *Optimization of Sampling Locations for Variogram Calculations.* Department of Soil and Water Science, the University of Arizona, Tucson.

Geostatistics for the Environmental Sciences, GS+. (2002). Gamma Design Software. Michigan. USA.

S-Plus 4.5 Professional Edition for Windows. (1998). Mathsoft.

Apêndice 1

Coordenadas da Rede Inicial e Valores da Salinidade

Coordenada X	Coordenada Y	Salinidade (%)	Salinidade Corrigida
976965	1706595	35.47	26.49824202
970821	1705153	7.53	-1.10770648
972756	1705151	7.72	-0.91724316
974693	1705150	9.92	1.2829885
976629	1705149	35.47	26.83322015
955329	1703233	6.91	-1.28292224
957265	1703230	6.6	-1.59222727
959200	1703228	6.79	-1.40176395
961138	1703225	6.36	-1.83106897
963073	1703223	6.54	-1.65060566
965011	1703221	5.93	-2.26014234
966946	1703219	6.17	-2.01967902
968882	1703217	7.41	-0.77921571
970819	1703216	6.24	-1.94898405
972755	1703214	6.54	-1.64852073
974692	1703212	6.05	-2.13805741
955325	1701294	6.54	-1.20373649
957262	1701291	6.85	-0.89304152
959198	1701289	7.03	-0.7125782
961135	1701287	7.03	-0.71211488
963071	1701284	6.97	-0.77141991
965009	1701282	7.03	-0.71095659
966944	1701280	6.79	-0.95049327
968880	1701278	5.93	-1.81002996
970817	1701277	5.33	-2.4097983
972753	1701275	4.07	-3.66933498
974690	1701274	3.37	-4.36910332
953386	1699359	5.93	-1.36547738
955322	1699356	6.17	-1.1247824
957260	1699354	7.28	-0.01431908
959196	1699351	7.53	0.23637589
961133	1699349	7.16	-0.13316079
963069	1699347	6.91	-0.38269747
965006	1699345	6.79	-0.50223416
966942	1699343	6.17	-1.12177084

968878	1699341	5.08	-2.21130752
970816	1699339	5.27	-2.02084421
972751	1699337	3.96	-3.33038089
974689	1699336	2.51	-4.78014923
953383	1697421	5.87	-0.97652328
955320	1697419	6.05	-0.79605997
957257	1697416	6.36	-0.48536499
959193	1697414	6.91	0.06509833
961131	1697411	6.94	0.0957933
963067	1697409	6.86	0.01625662
965004	1697407	6.68	-0.16328007
966940	1697405	6.73	-0.11281675
968876	1697403	5.55	-1.29235343
970814	1697401	4.81	-2.03189011
972750	1697400	2.8	-4.04165846
955317	1695480	6.08	-0.31687422
957255	1695477	6.73	0.33382076
959191	1695475	7.12	0.72428408
961129	1695472	7.31	0.91497905
963065	1695470	7.05	0.65544237
965002	1695468	6.82	0.42590568
966938	1695466	6.62	0.226369
968875	1695464	4.77	-1.62316768
970812	1695463	5.45	-0.94293602
972748	1695461	2.48	-3.91247271
957252	1693539	5.93	-0.01722515
959188	1693537	6.37	0.42323817
961126	1693535	6.62	0.67370148
963063	1693532	6.79	0.84439646
965000	1693530	6.78	0.83485978
966937	1693528	6.3	0.35532309
968873	1693527	4.62	-1.32444525
970811	1693525	3.93	-2.01398193
972739	1693852	2.65	-3.36973425
957250	1691601	4.4	-1.09827106
959186	1691598	4.95	-0.54757608
961124	1691596	5.43	-0.06711277
963060	1691594	6.39	0.89335055
964998	1691592	6.3	0.80381387
966935	1691590	5.92	0.42427718
968871	1691588	4.56	-0.9352595
970809	1691586	4.24	-1.25479618
957247	1689663	3.22	-1.82931697
959184	1689660	3.93	-1.11862199
961122	1689658	4.93	-0.11815867
963058	1689656	6.1	1.05230464
964996	1689654	6	0.95276796
966933	1689652	5.83	0.78323128
968869	1689650	5.32	0.27369459
970807	1689648	4.25	-0.79584209
959181	1687722	4.96	0.3603321
961120	1687719	5.5	0.90102708
963056	1687717	5.98	1.38149039

964994	1687715	5.41	0.81195371
966931	1687713	5.92	1.32241703
968867	1687711	3.63	-0.96711966
970805	1687710	3.71	-0.886888
959179	1685784	4.63	0.47928619
961117	1685782	5.27	1.11974951
963054	1685780	5.31	1.16021283
964992	1685777	5.71	1.5609078
966929	1685776	5.82	1.67113946
968865	1685774	3.42	-0.72839722
959177	1683846	4.56	0.85824028
961115	1683844	5.23	1.5287036
963052	1683842	5.19	1.48916692
964990	1683840	4.89	1.18963023
966927	1683838	4.65	0.95009355
968864	1683836	3.1	-0.59944313
959174	1681908	4.81	1.55719437
961113	1681905	4.9	1.64788935
963050	1681903	5.44	2.18835267
964988	1681901	3.64	0.38881598
966925	1681899	2.37	-0.8807207
959172	1679970	4.97	2.16614847
961111	1679968	5.57	2.76661178
963048	1679966	2.07	-0.7329249
964986	1679963	2.27	-0.53222993
966923	1679961	0.97	-1.83176661

Fonte: Instituto de Investigaciones Marinas-INVEMAR-Santa Marta-Colômbia

Apêndice 2

Coordenadas da Rede Ótima

Coordenada X	Coordenada Y
970821	1705153
976629	1705149
955329	1703233
959200	1703228
965011	1703221
970817	1701277
953386	1699359
974689	1699336
957255	1695477
963065	1695470
968875	1695464
972748	1695461
964996	1689654
959181	1687722
970805	1687710
963054	1685780
959172	1679970
963048	1679966
966923	1679961

Fonte: Grupo de Investigación y Asesoría en Estadística
Universidad del Quindío-Colômbia

Abstract

This paper presents a methodology for creating a network of bio-ecological monitoring utilizing the Kriging method, taking advantage of the independence of the mean square prediction error at each location on the variable value. Thus, starting with an initial sample at N spatial locations, we can eliminate locations, where at each stage the location eliminated is that where the variable presents the smallest prediction error, and therefore may be predicted by kriging. This procedure is repeated until a group of n observations is identified, $n \leq N$, which consist of the monitoring network. This procedure is optimal in the sense that it identifies a minimum value for n, from which one can interpolate using kriging.

An optimal network will be designed for Salinity at the lake "Ciénaga Grande de Santa Marta" in the North of Colômbia.

Key Words: Variogram, Spatial autocorrelation range, Prediction error, Kriging, Monitoring network.

Estimação da fração de não-disjunção meiótica em pacientes com Síndrome de Down

Glaura da Conceição Franco*

Paula Arantes Barros*

Resumo

As trissomias cromossômicas ocorrem principalmente devido ao processo de não-disjunção na meiose. Com o objetivo de entender melhor o mecanismo subjacente da não-disjunção, é necessário estabelecer a proporção de casos que ocorrem na primeira ou segunda divisão meiótica. Pacientes trissômicos apresentam, em estudos de microssatélites empregando a Reação em Cadeia da Polimerase (RCP), três fragmentos de igual intensidade, dois fragmentos a uma taxa de 2:1 ou um único fragmento. Neste trabalho é apresentado um modelo de probabilidade para o número de picos em um *locus* de microssatélite polimórfico, que é uma função da fração de não-disjunção na meiose I, F. Baseado neste modelo, o estimador de máxima verossimilhança para F é obtido, usando a proporção observada de um, dois e três alelos em indivíduos com trissomia do cromossomo 21. A partir das propriedades da teoria de máxima verossimilhança, são calculados a variância assintótica e intervalos de confiança para F. Devido ao fato de que as trissomias cromossômicas são eventos raros, o uso da teoria assintótica por ficar comprometida. Assim, a técnica *bootstrap* é empregada para construir intervalos de confiança para F e comparar os resultados com aqueles obtidos da teoria normal.

Palavras-chave: Trissomia, não-disjunção, máxima verossimilhança, bootstrap, intervalos de confiança.

* Departamento de Estatística – ICEx – UFMG – Caixa Postal 702 - Belo Horizonte – MG - Brazil – 31.270-901

e-mail: glaura@est.ufmg.br

1. Introdução

As anomalias cromossômicas numéricas, chamadas aneuploidias, são causas comuns de morte fetal e de malformações congênitas. Estas ocorrem geralmente como eventos esporádicos de não-disjunção meiótica.

Dentre as aneuploidias, a mais estudada é a trissomia do cromossomo 21, vista em aproximadamente um em 750 nativivos. Ela produz um fenótipo conhecido como síndrome de Down e é a causa mais comum de retardo mental de origem genética em humanos. A síndrome de Down é causada pela presença de três cromossomos 21, sendo um cromossomo extra.

Para melhor entendimento da epidemiologia desta trissomia é necessário estabelecer a função de probabilidade de que ela ocorra na primeira ou segunda divisão meiótica. Até agora a estimativa de F (a probabilidade condicional de que não-disjunções ocorram na meiose I) tem feito uso do estudo comparativo de marcadores genéticos centroméricos na criança afetada e em seus pais. A necessidade do estudo dos pais complica estes estudos e impossibilita o uso de material de arquivo. Uma outra alternativa utilizada pelos pesquisadores é fazer a análise do fracionamento por densitometria *a laser*, utilizando a amplificação através da Reação em Cadeia da Polimerase - PCR (Parra, 1999).

No estudo de *locus* de microssatélites polimórficos, pela reação em cadeia da polimerase (PCR), pacientes trissômicos apresentam: a presença de três picos de mesma intensidade, dois picos com uma dosagem relativa 2:1 ou apenas um pico (casos não-informativos). Para a ocorrência do padrão de 3 picos em um feto trissômico, é necessário que a não-disjunção ocorra na meiose I e que sejam transmitidos 3 alelos diferentes dos pais.

Se conhecermos a proporção F de casos de não-disjunção na primeira divisão meiótica e as freqüências relativas $(p_1, p_2, \ldots, p_m)$ dos m alelos de um *loco* multialélico de microssatélites, podemos aplicar o equilíbrio de Hardy-Weinberg para calcular a freqüência relativa esperada dos padrões de 1, 2 e 3 picos já que as freqüências genotípicas têm distribuição multinomial.

O estimador de máxima verossimilhança (EMV) para F, baseado no modelo de probabilidade para a fração F de não-disjunção na meiose I apresentado em Franco et al. (2003), utiliza a proporção observada de padrões de um, dois ou três picos em uma amostra de indivíduos trissômicos. Através das propriedades assintóticas do EMV, é possível calcular a variância assintótica destes estimadores usando a matriz de informação observada (Welsh, 1996).

O objetivo principal deste trabalho é comparar os Intervalos de Confiança para a fração de não-disjunção F, utilizando a teoria normal e intervalos bootstrap percentílicos, através de uma aplicação a dados reais.

2. GENÉTICA

2.1 - Meiose

A meiose é um processo de reprodução sexual através do qual o número cromossômico das células germinativas diplóides (2n) é reduzido à metade (n) na formação das células reprodutivas maduras ou gametas.

Em um indivíduo normal, essa redução é obtida através de duas divisões celulares sucessivas, durante as quais os cromossomos são duplicados apenas uma vez. Procedendo a primeira das duas divisões, os cromossomos homólogos são pareados e duplicados. Esses cromossomos pareados são separados na primeira divisão meiótica da célula. Na segunda divisão meiótica, as duas cromátides de cada cromossomo separam-se, produzindo um total de quatro células sexuais maduras, cada uma capaz de fertilização. Uma explicação mais detalhada do processo de meiose pode ser vista em Gardener e Swstad (1986).

2.2 - Trissomia do Cromossomo 21

A trissomia do 21 é o resultado da não-disjunção primária, que pode ocorrer em ambas as divisões meióticas e em ambos os pais. Os cromossomos pareados não se separam de forma apropriada para os pólos na anáfase; um dos gametas receberá dois cromossomos 21 e o outro nenhum. As figuras 1a e 1b apresentam processos de meiose com a não-disjunção ocorrendo na 1ª e 2ª fases, respectivamente. Sem perda de generalidade, ilustraremos o processo com a não-disjunção ocorrendo na mãe de um indivíduo trissômico.

Figura 1a - Não-disjunção na 1ª fase da meiose em indivíduos trissômicos

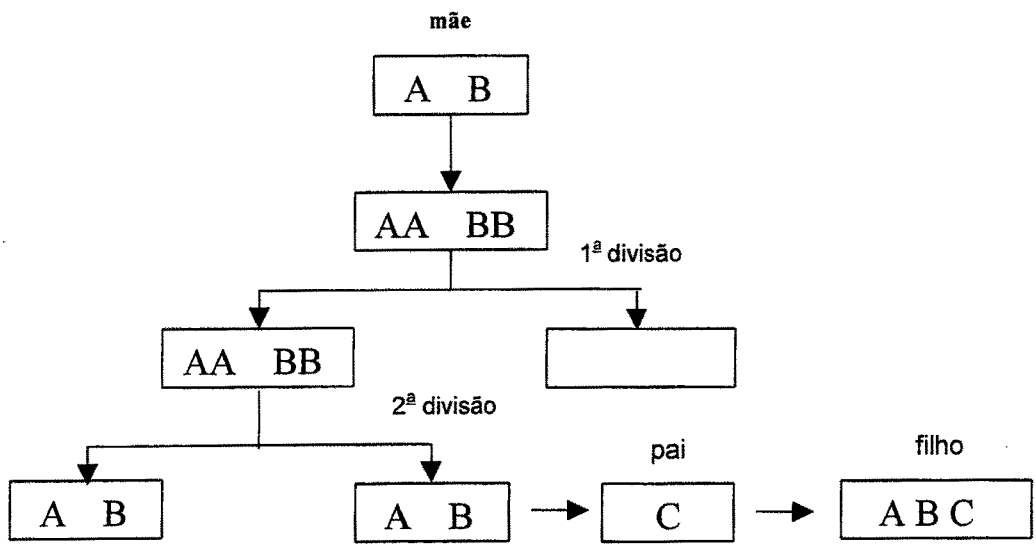
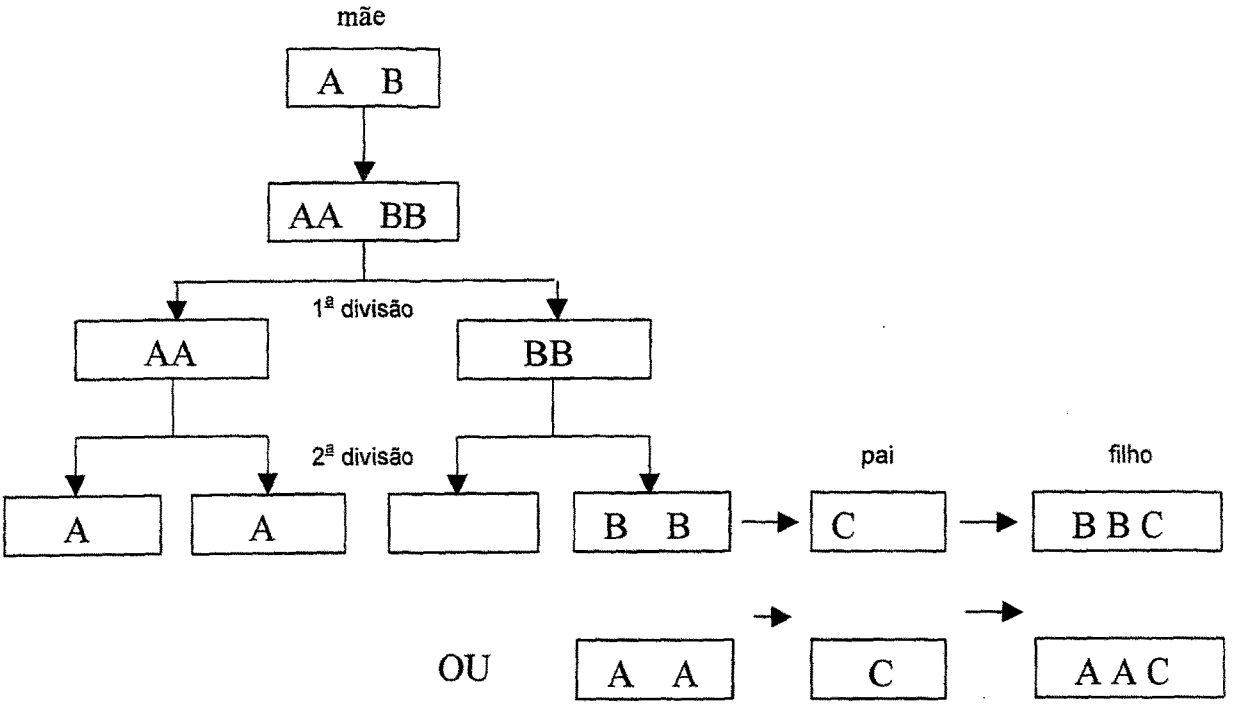


Figura 1b - Não-disjunção na 2ª fase da meiose em indivíduos trissômicos



2.3 - Análise da origem das trissomias através da PCR

A análise da origem das trissomias é feita através da utilização de marcadores moleculares dos cromossomos, explorando regiões polimórficas do DNA. Por serem altamente informativos, os polimorfismos de DNA permitem a determinação da origem das trissomias na maioria dos casos, além de propiciarem uma análise mais objetiva.

A introdução da técnica de amplificação do DNA através da Reação em Cadeia da Polimerase - PCR - permitiu a identificação de um grande número de marcadores denominados microssatélites, mapeados ao longo dos diversos cromossomos (Pena, 1998). Os microssatélites são unidades curtas de DNA de 1 a 5 pares de base que se repetem seguidamente. Estes marcadores puderam ser estudados em conjunto através da PCR, possibilitando um resultado mais rápido e ainda mais informativo.

2.4 - Padrões de picos em indivíduos trissômicos

Em estudos com microssatétites polimórficos, indivíduos portadores de trissomias apresentam padrões de um, dois ou três picos (Figura 3). O padrão de um pico indica a igualdade entre os três alelos herdados pelo indivíduo. Padrões de dois picos, normalmente, estão em uma dosagem relativa de 2:1, indicando a presença de

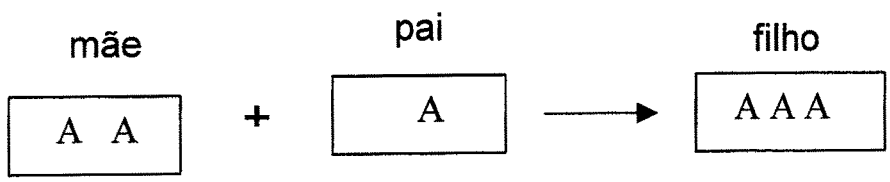
dois alelos iguais e um de tamanho diferente e o padrão de três picos de mesma intensidade equivale à presença de três alelos com tamanhos diferentes.

A proporção relativa de cada um dos padrões de um, dois e três picos vai depender, principalmente, do momento do acidente meiótico, além do índice de heterozigosidade do *loco* na população. Ignorando recombinações, para que ocorra um padrão de três picos, é necessário que o erro na segregação dos cromossomos tenha ocorrido na primeira divisão meiótica e que sejam transmitidos três alelos diferentes dos pais, provenientes de dois cromossomos homólogos de um dos genitores, e um cromossomo diferente, herdado pelo outro genitor. A não ser quando o genitor que transmitiu os dois cromossomos homólogos seja homozigoto, o padrão de dois picos retrata uma situação em que a não-disjunção acontece na segunda divisão, quando as cromátides irmãs permanecem unidas.

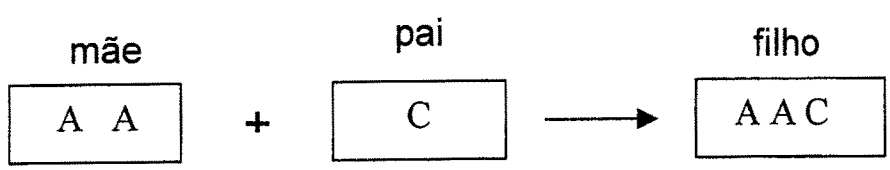
A detecção molecular de trissomias humanas usando PCR e densitometria *a laser* (Pena, 1998) surgiu, portanto, como uma poderosa ferramenta alternativa para o diagnóstico aplicado em tecidos humanos que não estão em divisão, o que não é permitido na citogenética convencional.

Figura 2 - Desenho esquemático no número de picos

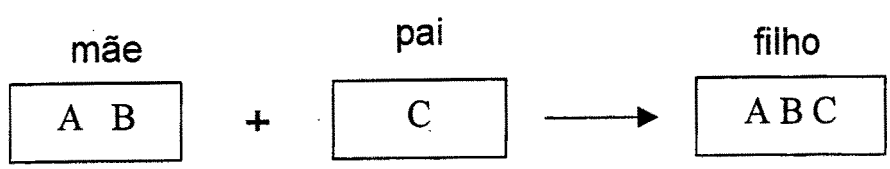
I) 1 PICO (três alelos iguais):



II) 2 PICOS (dois alelos iguais):



III) 3 PICOS (três alelos diferentes):



3. Estimação de F

3.1 - O Modelo de probabilidade

Seja X a variável aleatória "Número de Picos em indivíduos portadores de trissomias cromossômicas". Franco et al. (2003) determinaram a distribuição de probabilidade de X, que é feita da maneira descrita abaixo. Seja

$$p_X(x) = P[(X = x \cap ND_1) \cup (X = x \cap ND_2)]$$

onde ND_1= ocorrência de não-disjunção na 1ª fase da meiose, dado que ocorreu a não-disjunção

ND_2 = ocorrência de não-disjunção na 2ª fase da meiose, dado que ocorreu a não-disjunção

Assim,

$$p_X(x) = P(X = x \cap ND_1) + P(X = x \cap ND_2) - P[(X = x \cap ND_1) \cap (X = x \cap ND_2)]$$

Como a não-disjunção não pode ocorrer em ambas as fases ao mesmo tempo,

$P[(X = x \cap ND_1) \cap (X = x \cap ND_2)] = 0$.

Então, $p_X(x) = P(X = x \cap ND_1) + P(X = x \cap ND_2)$

$\Rightarrow p_X(x) = P(X = x / ND_1) \times P(ND_1) + P(X = x / ND_2) \times P(ND_2)$.

Chamando $P(ND_1) = F$ e $P(ND_2) = 1 - F$, temos

$p_X(x) = P(X = x / ND_1) \times F + P(X = x / ND_2) \times (1 - F)$

Logo, vemos que a função de distribuição de X é função da fração desconhecida, F, de não-disjunção na meiose.

Seja p_i a probabilidade de o alelo i estar presente em um indivíduo, $i=1, \ldots, m$. Para a ocorrência do padrão de *1* pico, a mãe deve ser homozigota e o alelo do pai deve ser igual ao da mãe. Assim,

$$P(X = 1) = F \times P(X = 1 / ND_1) + (1 - F) \times P(X = 1 / ND_2)$$

$$= F \sum_{i=1}^{m} p_i^{\,3} + (1 - F) \sum_{i=1}^{m} p_i^{\,2}$$

Para ocorrer o padrão de *2* picos há duas possibilidades:

se a mãe for homozigota, o alelo do pai, necessariamente, terá que ser diferente dos alelos da mãe;

b) se a mãe for heterozigota, o alelo do pai terá que ser igual a qualquer um dos alelos da mãe.

Logo,

$$P(X = 2) = F \times P(X = 2 / ND_1) + (1 - F) \times P(X = 2 / ND_2)$$

$$= F \times 3 \sum_{i=1}^{m} \sum_{j=1}^{m} p_i^2 p_j + (1-F) \times 2 \sum_{i=1}^{m} \sum_{j=1}^{m} p_i p_j \text{ , com i≠j}$$

Para a ocorrência do padrão de *3* picos, a mãe deverá ser, obrigatoriamente, heterozigota e o alelo do pai deverá ser diferente dos alelos da mãe.

$$P(X=3) = P(X=3/ND_1) \times F + P(X=3/ND_2) \times (1-F)$$

Como o padrão de *3* picos nunca vai ocorrer se a não-disjunção meiótica for na 2ª fase, temos que $P(X=3/ND_2) = 0$. Logo,

$$P(X=3) = F \times \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} p_i p_j p_k \quad , \text{ com } i \neq j \neq m.$$

Dado que as probabilidades p_i's , que correspondem à frequência do alelo $i=1,...,m$ são conhecidas, pode-se concluir que o modelo usado para descrever X será uma *Multinomial*$(n, \theta_1, \theta_2, \theta_3)$, onde os parâmetros θ_i's são dados por:

$$\theta_1 = P(X=1) = F \sum_{i=1}^{m} p_i^3 + (1-F) \sum_{i=1}^{m} p_i^2$$

$$\theta_2 = P(X=2) = F \times 3 \sum_{i=1}^{m} \sum_{j=1}^{m} p_i^2 p_j + (1-F) \times 2 \sum_{i=1}^{m} \sum_{j=1}^{m} p_i p_j \text{ , para } i \neq j \qquad (3.1)$$

$$\theta_3 = P(X=3) = F \times \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} p_i p_j p_k \text{ , para } i \neq j \neq k$$

e $0 \leq F \leq 1$.

3.2 - O Estimador de F

Vimos anteriormente que $f_{\underset{\sim}{X}}(\underset{\sim}{x}, \theta) \sim$ *Multinomial*$(n; \theta_1, \theta_2, \theta_3)$. Devemos notar que os θ_i's , $i = 1, 2, 3$, são funções do parâmetro F (equação 3.1) que desejamos estimar.

Assim, podemos calcular a função de verossimilhança, que é dada por:

$$L(F) = f_{\underset{\sim}{X}}(x,\theta) = \theta_1^{n_1} \times \theta_2^{n_2} \times (1-\theta_1-\theta_2)^{n-n_1-n_2} \qquad (3.2)$$

onde: n = número de indivíduos com trissomia

n_1 = número de indivíduos com o padrão de 1 pico

n_2 = número de indivíduos com o padrão de 2 picos

n_3 = número de indivíduos com o padrão de 3 picos

e os θ_i's são dados na equação (3.1).

O estimador de máxima verossimilhança (EMV) de F é encontrado maximizando-se a função de verossimilhança L em relação a F (Mood et al.(1974)).

Utilizando-se o software Mathematica (Blachman, 1992) para maximizar esta função, encontramos o EMV de F, que é dado por:

$$\hat{F} = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \qquad (3.3)$$

onde:

$$A = n(a-b)(3c-1+b)$$

$$B = (n-n_2)(a-b)(1-b) + b(n-n_1)(3c-1+b)$$

$$C = b(n-n_1-n_2)(1-b)$$

e $\quad a = \sum_{i=1}^{m} p_i^{\,3}$, $\quad b = \sum_{i=1}^{m} p_i^{\,2}$ e $\quad c = \sum_{i=1}^{m}\sum_{j=1}^{m} p_i^{\,2} p_j$, $\quad i \neq j$

3.3 - Variância assintótica

Como foi definido no item anterior $\hat{F} = \hat{F}(X_1,...,X_n)$ é o estimador de máxima verossimilhança de F baseado numa amostra aleatória de tamanho n. Pelas propriedades da teoria assintótica (Mood et al.,1974), temos que, para n suficientemente grande, a distribuição do EMV de F é aproximadamente

$$\hat{F} \sim N(F, I^{-1}(F))$$

onde: $I = -E\left[\frac{\partial^2 \log L(F)}{\partial F^2}\right]$ é conhecido como a matriz de Informação esperada de Fisher.

Como na maioria das vezes a esperança acima é difícil de ser calculada, utilizamos um estimador para I conhecido como a matriz de informação observada (I) (Welsh,1996), dada por:

$$I = \left. \frac{\partial^2 \log L(F)}{\partial F^2} \right|_{F=\hat{F}}$$

No nosso caso, I é dada por[1]:

$$I = -\left. \frac{\partial^2 \log L(F}{\partial F^2} \right|_{F=\hat{F}} = \frac{(a-b)^2 n_1}{\left(b+(a-b)\hat{F}\right)^2} + \frac{n-n_1-n_2}{\hat{F}^2} + \frac{(3c+b-1)^2 n_2}{\left(1-b+(-1+b+3c)\hat{F}\right)^2}$$

onde $\hat{F}$ é calculado como em (3.3).

Assim a variância assintótica de F é dada por:

$$\text{Vass} = 1 \ / \text{I} \qquad (3.4)$$

4. Intervalos de confiança para F

Neste trabalho utilizaremos duas metodologias para se calcular intervalos de confiança para F: a teoria normal e a técnica de bootstrap.

O intervalo com nível de confiança $1\text{-}2\alpha$ utilizando a teoria normal é bem conhecido e dado por

$$\left[\hat{F} - z^{(1-\alpha)} \cdot e\hat{p}; \hat{F} - z^{(\alpha)} \cdot e\hat{p}\right] \qquad (4.1)$$

onde $z^{(\alpha)}$ o $100.\alpha\text{-ésimo}$ ponto percentílico de uma distribuição $N(0,1)$ e $e\hat{p}(\hat{F})$ um estimador razoável do erro-padrão de $\hat{F}$. Este intervalo será denotado aqui por Intervalo Normal.

A técnica bootstrap é basicamente uma técnica de reamostragem, que permite aproximar a distribuição de uma função das observações pela distribuição empírica dos dados, baseado em uma amostra de tamanho finito. No bootstrap não-paramétrico a amostragem é feita, com reposição, da amostra original, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Neste caso, supomos que as observações são obtidas da função de distribuição empírica, $\hat{f}$, que designa uma massa de probabilidade igual a $1/n$ para cada ponto amostral (Efron, 1979).

O procedimento bootstrap básico consiste em gerar repetidas amostras bootstrap, cada uma delas de tamanho n, um número suficientemente grande de vezes. Obtemos assim as amostras bootstrap

$\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$

onde B é o número de replicações bootstrap.

O valor ótimo de B depende da finalidade para a qual a técnica bootstrap está sendo empregada. Se desejamos somente estimar o erro-padrão, Efron e Tibshirani (1993), sugerem valores de B variando de 50 a 200. Já para a construção de Intervalos de Confiança, a sugestão é de B igual a 1000.

[1] A derivada 2ª em relação a F foi calculada utilizando o software Mathematica (Blachman, 1992).

Para cada amostra bootstrap calculamos o valor da estatística de interesse, a qual denotaremos por $\hat{F}^*(b)$, $b = 1, \ldots, B$. Assim, podemos usar a distribuição empírica de $\hat{F}^*(b)$ como uma aproximação para a verdadeira distribuição da estatística $\hat{F}$, e portanto fazer inferência sobre o parâmetro de interesse. O procedimento acima é referido como princípio plug-in (Efron & Tibshirani, 1993).

Seja $G^*(x)$ a função distribuição acumulada de $\hat{F}^*$,

$$G^*(x) = P^*\left[\hat{F}^* \leq x\right] \tag{4.2}$$

O intervalo de confiança bootstrap percentílico de nível $1\text{-}2\alpha$ é calculado como

$$[G^{*-1}(\alpha)\,;\, G^{*-1}(1-\alpha)] \tag{4.3}$$

Na prática, gera-se B amostras bootstrap independentes $x^{*1}, x^{*2}, \ldots, x^{*B}$ e estima-se $\hat{F}^*$ para cada uma delas. Os limites inferior e superior do intervalo percentílico são dados por:

$$\left[\hat{F}_B^{*(\alpha)}\,;\, \hat{F}_B^{*(1-\alpha)}\right] \tag{4.4}$$

onde $\hat{F}_B^{*(\alpha)}$ é o *Bα-ésimo* valor ordenado das replicações $\hat{F}_B^*(b)$.

5. Aplicação

Neste estudo trabalhamos com dois tipos de amostras de sangue de indivíduos trissômicos. O primeiro tipo consiste de uma amostra do sangue de $n = 34$ indivíduos com a síndrome de Down, analisados usando a técnica da PCR, tão logo o sangue seja coletado[2]. O segundo tipo consiste de lâminas de material de arquivo, contendo o sangue de indivíduos trissômicos. A primeira lâmina foi coletada em 1992, com $n = 18$ indivíduos, a segunda em 1995, com $n = 31$ indivíduos e a terceira em 1998, com $n = 33$ indivíduos.

Para se determinar os padrões do número de picos, foi escolhido o microssatélite D21S215, devido ao seu polimorfismo e, principalmente, à sua localização, na região pericentromérica do cromossomo 21, o que indica que não há evidências da ocorrência de recombinação com regiões do seis alelos, que apresentaram um polimorfismo com freqüências de 0.12, 0.45, 0.09, 0.31, 0.01 e 0.02, respectivamente, para os alelos CA-13, CA-14, CA-15, CA-16, CA-17 e CA-19.

A Tabela 1 mostra a proporção de 1, 2 e 3 picos para cada amostra.

Tabela 1 - Proporção de picos para as 4 amostras

	Sangue	Lâmina 92	Lâmina 95	Lâmina 98
1 pico	06 (18%)	09 (50%)	12 (39%)	10 (30%)
2 picos	22 (64%)	06 (33%)	14 (45%)	15 (45%)
3 picos	06 (18%)	03 (17%)	05 (16%)	08 (25%)

Percebemos que a amostra que mais se difere da amostra de sangue é a da Lâmina 92. Neste caso a maior proporção é de *1* pico, enquanto para as outras amostras a quantidade de *2* picos é a mais freqüente. Com o número de picos podemos calcular o estimador de F, $\hat{F}$, e a variância assintótica ($Vass$) para cada uma das amostras, mostrados na Tabela 2.

Tabela 2 - Estimativas de F e Variância assintótica

Amostra	$\hat{F}$	Vass
Sangue	0.6552	0.0481
Lâmina 92	0.3753	0.0368
Lâmina 95	0.4228	0.0274
Lâmina 98	0.6504	0.0336

Podemos observar que as estimativas de F para o sangue e a Lâmina 98 apresentam valores muito próximos, em torno de 0.65. Já as amostras das Lâminas 92 e 95 apresentam valores de F bem inferiores ao encontrado para a amostra de sangue. Estes resultados parecem sugerir que o fato de o material ficar estocado por um longo período de tempo para ser utilizado em análises futuras afeta, de alguma forma, a avaliação dos padrões de picos nas amostras e conseqüentemente as estimativas da fração de não-disjunção, F. A variância assintótica é um pouco maior para a amostra de sangue.

Utilizando a variância assintótica para o caso do intervalo normal, calculamos intervalos de 95% de confiança normal e percentílico para F, obtendo

[2] Esta amostra foi coletada em 1998.

Tabela 3 - Intervalos de 95% de Confiança para F

	Intervalo	
Amostra	Normal	Percentílico
Sangue	[0.2253;1.0000]	[0.2148;0.9804]
	Amp = 0.7747	Amp = 0.7656
Lâmina 92	[-0.0006;0.7512]	[0.0000;0.7855]
	Amp = 0.7518	Amp = 0.7855
Lâmina 95	[0.0986;0.7470]	[0.1370;0.7698]
	Amp = 0.6484	Amp = 0.6328
Lâmina 98	[0.2911;1.0096]	[0.3037;0.9209]
	Amp = 0.7185	Amp = 0.6172

Obs.: Amp = (Limite superior – Limite inferior) do intervalo.

Devemos chamar a atenção para o fato de que o intervalo Normal apresenta limites superiores maiores que 1.00 (caso da Lâmina 98) e limites inferiores menores que 0.00 (caso da Lâmina 92). Assim, devido a este fato, e também por apresentar amplitudes menores (exceto para a Lâmina 92), podemos considerar o Intervalo Percentílico mais preciso neste caso.

Observamos porém que, para amostras grandes, as duas técnicas devem ser equivalentes, o que favorece à técnica tradicional de máxima verossimilhança por ser mais simples. Além disto, a técnica de bootstrap depende do modelo adotado, isto é, a geração de valores aleatórios é feita com base no modelo ajustando pelo método de máxima verossimilhança. Assim sendo, a técnica de simulação por métodos de Monte Carlo deve ser recomendada para valores pequenos e moderados de n.

7. Conclusões

Nesse trabalho é mostrado que o número de picos em um locus de microssatélite polimórfico tem distribuição multinomial, com os parâmetros sendo uma função da fração de não-disjunção, F, na meiose I. Baseado neste modelo, o estimador de máxima verossimilhança para F é calculado, assim como sua variância assintótica. Estimativas de F, para quatro amostras de sangue de indivíduos com síndrome de Down, são calculadas. Os resultados parecem sugerir que a avaliação dos padrões de picos e, por conseguinte, a estimativa de F, ficam comprometidas se material de arquivo, estocado por um longo período de tempo, é utilizado. Como trissomias cromossômicas são eventos raros, é utilizada a técnica bootstrap, apresentando um procedimento alternativo para construir intervalos de confiança para F, que têm uma melhor performance que os intervalos obtidos através da teoria assintótica normal.

Referências bibliográfica

BLACHMAN, N.R., (1992) Mathematica: A Practical Approach. ***New Jersey: Prentice-Hall.***

EFRON, B. (1979) Bootstrap methods: another look at the jackknife. ***The Annals of Statistics,*** **7, 1-26.**

EFRON, B., TIBSHIRANI, R.J., (1993). An Introduction to the Bootstrap. ***New York: Chapman and Hall.***

FRANCO, G.C., LUCIO, P.S., PARRA, F.C., PENA, S.D.J. A probability model for the meiosis I non-disjunction fraction in numerical chromosomal anomalies. *Statistics in Medicine, in press,* 2003.

GARDENER, E E SWSTAD, D. P. (1986) Genética. 4ª edição, Rio de Janeiro: *Editora Guanabara.*

MOOD, A.M., GRAYBILL, F.A. AND BOES, D.C. (1974) Introduction to the Theory of Statisiics. *Third Edition, Singapore: McGraw-Hill.*

PARRA, F.C. (1999) Estudo da Origem de Não-disjunção Meiótica em Indivíduos Brasileiros com Trissomia do Cromossomo 21 Utilizando um Estimador de Máxima Verossimilhança. *Belo Horizonte: Instituto de Ciências Biológicas, UFMG.* 100 p . (Dissertação de Mestrado).

SHEN JJ, SHERMAN SL, HASSOLD TJ. (1998) Centromeric genotyping and direct analysis of nondisjunction in humans: Down syndrome. *Chromosoma* 107: 166-172.

PENA, S.D.J. (1998) Molecular Cytogenetics I: PCR-based diagnosis of human trisomies using computer-assisted laser densitometry. *Genet Mol Biol.* 3: 317-322.

WELSH, A.H., (1996) Aspects of Statistical Inference. *New York : John Wiley.*

Agradecimentos

Este trabalho contou com o apoio do CNPq através da bolsa PIBIC para a aluna Paula Arantes Barros. Gostaríamos também de agradecer ao Prof. Dr. Sérgio Danilo Pena e Flávia C. Parra, do Departamento de Bioquímica e Imunologia da UFMG, pelo material da pesquisa.

ABSTRACT

Chromosome trisomy primarily arises by the process of nondisjunction in meiosis. In order to increase our understanding of the mechanisms underlying nondisjunction, it is necessary to establish the proportion of cases occurring in first or second division of meiosis. Trisomic patients will display, in study of microsatellite by employing Polymerase Chain Reaction (PCR) based approach, three fragment peak of equal intensity, two fragments at an average 2:1 dosage or one individual fragment. In this work we present a probability model for the number of peaks in a polymorphic microsatellite locus, which is a function of the fraction of meiosis *I* nondisjunction, F. Based on this model, we obtain a maximum likelihood estimator for *F*, using the observed proportion of one, two and three allele patterns in individuals with chromosome trisomy. Relying on the properties of maximum likelihood theory, we also calculate the asymptotic variance and confidence intervals for *F*. Due to the fact that chromosome trisomies are rare events, the use of asymptotic theory may be compromised. Thus, we employ the bootstrap technique to build confidence intervals for *F* and compare the results with those obtained from the normal theory.

Keywords: Trisomy, Nondisjunction, Maximum Likelihood, Bootstrap, Confidence Intervals.

Estimação de probabilidades em campeonatos de futebol utilizando simulação estocástica, modelos de séries temporais para dados de contagem e o princípio da máxima entropia

Eduardo Lima Campos*

Cristiano Fernandes**

Reinaldo Castro Souza**

RESUMO

Neste trabalho, apresentamos uma metodologia para estimar probabilidades de eventos relacionados ao desempenho das equipes participantes de um campeonato de futebol. O cálculo exato destas probabilidades é bastante complicado, devido não somente ao elevado número de combinações possíveis de resultados para as partidas, mas também à estrutura de dependência eventualmente existente entre estes resultados. O primeiro passo da solução proposta aqui é o ajuste de um modelo de séries temporais para dados de contagem a séries de gols assinalados em partidas sucessivas e válidas pela competição de interesse. Neste trabalho utilizamos o modelo Poisson-Gama (Harvey e Fernandes, 1989). O efeito de fatores como mando de campo, escalação e alterações na comissão técnica, sobre a performance das equipes, foi incorporado através da introdução de variáveis explicativas no modelo. As distribuições preditivas dos modelos estimados foram então combinadas, através do Princípio da Máxima Entropia, para gerar as distribuições de probabilidade dos resultados das partidas restantes, a partir das quais,

* Escola Nacional de Ciências Estatísticas - ENCE/IBGE, Rio de Janeiro, Brasil, e-mail: limacampos@ibge.gov.br

** Pontifícia Universidade Católica do Rio de Janeiro - PUC-RIO, Rio de Janeiro, Brasil, e-mail: cris@ele.puc-rio.br

*** Reinaldo Castro Souza, Pontifícia Universidade Católica do Rio de Janeiro - PUC-RIO, Rio de Janeiro, Brasil, e-mail: reinaldo@ele.puc-rio.br

utilizando técnicas de simulação estocástica, geramos cenários aleatórios para o campeonato. As probabilidades associadas aos eventos de interesse foram estimadas pelas freqüências relativas de ocorrência destes eventos, em um número suficientemente grande de cenários gerados. A metodologia foi aplicada na primeira fase das três edições mais recentes do Campeonato Brasileiro de Futebol (1999 a 2001), tendo sido obtidas estimativas seqüenciais (ao final de cada rodada) das probabilidades de classificação das equipes para a segunda fase e de rebaixamento para a segunda divisão.

Palavras-chave: Séries Temporais, Dados de Contagem, Simulação, Entropia, Futebol.

1. Introdução

O futebol, assim como outros esportes, tem atraído recentemente o interesse de pesquisadores que trabalham com modelos probabilísticos e métodos estatísticos, como área de aplicação vasta, promissora e de forte e imediato apelo popular. Recentes aplicações de métodos quantitativos a esportes podem ser encontradas, por exemplo, em Bennet (1999). Dentre as aplicações específicas a futebol, podemos citar Harvey e Fernandes (1989), que utilizam modelos de séries temporais para dados de contagem na previsão dos resultados de partidas entre Inglaterra e Escócia, e Arruda (2000), que apresenta uma metodologia para estimar probabilidades associadas aos resultados das partidas e à classificação e rebaixamento de equipes em um campeonato, utilizando modelos de regressão Poisson e simulação estocástica.

Por outro lado, sabemos da importância histórica do futebol para o povo brasileiro, refletida no espaço que notícias relacionadas a este esporte ocupam nos meios de comunicação. Em razão desta popularidade, a imprensa esportiva vem se mostrando cada vez mais interessada em projeções e análises das perspectivas de classificação e rebaixamento das equipes nas diversas competições que fazem parte do calendário do nosso futebol. Em particular, quantificar as possibilidades de uma equipe em uma competição é bastante conveniente para os meios de comunicação, uma vez que condensa a informação relevante, traduzindo a situação da equipe em um único número e permitindo estabelecer conjecturas de significado claro e intuitivo, tais como: "a 4 rodadas do final da primeira fase do campeonato, a equipe A possui 94% de chances de classificar-se para a fase seguinte", ou: "as chances da equipe B permanecer na primeira divisão são de apenas 3%", ou ainda: "caso a equipe C vença suas próximas duas partidas, suas chances de classificação passam a ser de 75%".

Todavia, o cálculo destas probabilidades é um problema bastante complicado. A principal razão é que existe um número muito elevado de combinações possíveis para os resultados das partidas, o que torna inviável a contagem das ocorrências favoráveis aos eventos de interesse. Isto remete ao uso de técnicas de simulação estocástica, que podem ser utilizadas para obter estimativas das probabilidades destes eventos. Um outro problema é que resultados de partidas sucessivas apresentam eventuais estruturas de dependência, que devem ser representadas através de modelos de séries temporais para observações de contagem. Incidentalmente, diversos fatores afetam a performance das equipes e, por conseguinte, as probabilidades de interesse, tais como o mando de campo, as escalações e possíveis alterações na comissão técnica. Estes fatores podem ser incorporados através da introdução de variáveis explicativas nos modelos de séries temporais.

O objetivo deste trabalho é apresentar uma metodologia para estimar probabilidades de eventos relacionados ao desempenho das equipes participantes de um campeonato de futebol. A metodologia proposta combina simulação estocástica, modelos de séries temporais para dados de contagem e o Princípio da Máxima Entropia. O artigo está estruturado em sete seções. A seção 2 apresenta o modelo Poisson-Gama (Harvey e Fernandes, 1989) para séries temporais de dados de contagem. A seção 3 descreve como usar os resultados do ajuste do modelo Poisson-Gama a séries de gols, para obter as distribuições de probabilidade dos resultados de partidas de futebol, através do Princípio da Máxima Entropia, e o procedimento para estimar probabilidades de classificação e rebaixamento das equipes participantes de um campeonato de futebol, através de simulação estocástica e geração de cenários para os resultados das partidas do campeonato. A seção 4 apresenta testes de validação e resultados do ajuste do modelo Poisson-Gama às séries de gols utilizadas neste trabalho. A seção 5 apresenta propostas de solução para alguns problemas de ordem prática do modelo Poisson-Gama, em particular a contrução de intervalos de confiança e de testes de hipóteses para os hiperparâmetros e a obtenção da distribuição da estatística do teste de razão de verossimilhanças para pequenas amostras, utilizando a técnica do *bootstrap*. A seção 6 reporta aplicações no Campeonato Brasileiro de Futebol (1999 a 2001), e a seção 7 dedica-se às conclusões e possíveis extensões do trabalho.

2. O modelo Poisson-Gama para séries temporais de dados de contagem

Séries temporais para dados de contagem são registros da frequência de ocorrência de determinados eventos em sucessivos intervalos de tempo. Exemplos são o número de acidentes de trânsito diários, a incidência semanal de casos de uma doença em uma região e o número de gols marcados ou sofridos por uma equipe em partidas sucessivas de um campeonato de futebol. Estas séries apresentam observações que são números inteiros não-negativos, e portanto modelos com especificação Gaussiana são inadequados.

Diversas propostas para modelar séries de dados de contagem são apresentadas na literatura. Zeger (1988), por exemplo, estende a teoria de modelos de regressão para dados de contagem, para trabalhar com uma estrutura de dependência entre as observações. Outra abordagem parte de modelos de séries temporais para observações Gaussianas. Smith (1979) adapta o modelo de nível local (ver Harvey, 1989) para observaçoes não-gaussianas, utilizando a propriedade de conjugação. West, Harrison e Migon (1985) utilizam esta idéia e diversas aproximações analíticas para tratar a evolução estocástica de outras componentes, além do nível. Ambos os trabalhos adotam um enfoque Bayesiano para a estimação dos hiperparâmetros. Smith e Miller (1986) abordam estes modelos sob um ponto de vista clássico, propondo a estimação dos hiperparâmetros por máxima verossimilhança.

Harvey e Fernandes (1989) utilizam este enfoque para modelar dados de contagem e observações qualitativas. Em particular, os modelos Poisson-Gama e Pascal-Beta foram desenvolvidos para observações de contagem. Nesta seção, apresentamos o modelo Poisson-Gama, utilizado para as séries deste trabalho. Na seção

4.1, justificaremos o uso deste modelo, através de testes de validação e da comparação com o modelo Pascal-Beta via critérios de informação.

2.1 – Especificação, Estimação e Previsão

Seja $\{ y_t ; t = 1,2,...,T\}$ uma série de dados de contagem (gols, neste trabalho). O modelo Poisson-Gama (Harvey e Fernandes, 1989) assume que:

$$y_t \mid \mu_t \sim Poisson(\mu_t), t = 1,2,...,T. \quad (2.1)$$

onde μ_t é a média ou nível do processo gerador no instante t. Observe a diferença em relação ao modelo de nível local para observações Gaussianas, em que a distribuição em (2.1) é normal, e o nível μ_t segue também distribuição normal, cujos parâmetros são estimados seqüencialmente através do Filtro de Kalman (ver Harvey, 1989). No modelo Poisson-Gama, a distribuição de μ_t é, por hipótese, gama, sendo os parâmetros estimados através das equações apresentadas a seguir.

Seja a distribuição *a posteriori* do nível no instante *t-1*, dada por: $\mu_{t-1} \mid Y_{t-1} \sim G(\mathsf{a}_{t-1}, \mathsf{b}_{t-1})$, onde $Y_{t-1} = (y_1, y_2, ..., y_{t-1})'$. Considere a seguinte distribuição *a priori* para μ_t:

$$\mu_t \mid Y_{t-1} \sim G(a_{t|t-1} = \omega a_{t-1}, b_{t|t-1} = \omega b_{t-1}) \quad (2.2)$$

onde $\omega \in (0,1]$ é o hiperparâmetro do modelo. As equações para $a_{t|t-1}$ e $b_{t|t-1}$ em (2.2) são chamadas equações de previsão, e são especificadas desta forma para reproduzir as seguintes características do modelo de nível local: $E(\mu_t \mid Y_{t-1}) = E(\mu_{t-1} \mid Y_{t-1})$ e $VAR(\mu_t \mid Y_{t-1}) > VAR(\mu_{t-1} \mid Y_{t-1})$ (para detalhes, ver Harvey, 1989). Após observarmos y_t, utilizamos o Teorema de Bayes para obter a distribuição *a posteriori* de μ_t:

$$\mu_t \mid Y_t \sim G(a_t = a_{t|t-1} + y_t = \omega a_{t-1} + y_t, b_t = b_{t|t-1} + 1 = \omega b_{t-1} + 1) \quad (2.3)$$

As equações para a_t e b_t em (2.3) são chamadas equações de atualização, e são utilizadas para estimar, seqüencial e recursivamente, os parâmetros do modelo. As equações em (2.2) e (2.3) são conjuntamente denominadas Filtro Poisson-Gama. Para inicializar o filtro, fazemos: $\mathsf{a}_0 = 0$ e $\mathsf{b}_0 = 0$, de tal forma que $\mu_1 \mid Y_0 \sim G(0,0)$. Observe que esta distribuição é imprópria, ou seja, não integra um. Entretanto, é fácil ver que a partir do instante $t = \tau$, em que ocorre a primeira observação diferente de zero, obtemos distribuições próprias para μ_t.

A distribuição preditiva 1-passo-à-frente pode ser obtida de (2.1) e (2.2), da seguinte forma:

$$p(y_t \mid Y_{t-1}) = \int_0^{y_t} p(y_t, \mu_t \mid Y_{t-1})\, d\mu_t = \int_0^{y_t} p(y_t \mid \mu_t)\, p(\mu_t \mid Y_{t-1})\, d\mu_t = \binom{a_{t|t-1} + y_t - 1}{y_t} b_{t|t-1}^{\,a_{t|t-1}} (1 + b_{t|t-1})^{-(a_{t|t-1} + y_t)},$$ que é

a distribuição binomial negativa, com parâmetros dados em (2.4), a seguir:

$$y_t \mid Y_{t-1} \sim \text{NBin}\left(\frac{b_{t|t-1}}{1+b_{t|t-1}}, a_{t|t-1}\right) \tag{2.4}$$

A função de verossimilhança é dada por $L(\omega; y_\tau, y_{\tau+1}, \ldots, y_T) = \prod_{t=\tau}^{T} p(y_t|Y_{t-1})$, e o hiperparâmetro ω é estimado através da maximização da função de log-verossimilhança, dada por:

$$l(\omega) = \ln[L(\omega)] = \sum_{t=\tau+1}^{T} \{\ln[\Gamma(a_{t|t-1} + y_t)] - \ln(y_t!) - \ln[\Gamma(a_{t|t-1})] + a_{t|t-1} \ln(b_{t|t-1}) - (a_{t|t-1} + y_t) \ln[1 + b_{t|t-1}]\} \quad 2.5)$$

sujeita a $\omega \in (0,1]$, que corresponde ao problema de otimização não-linear restrita em uma dimensão.

Pode-se provar (ver Harvey e Fernandes, 1989) que:

$$y_{T+L \mid T} = E(y_{T+L} \mid Y_T) = \frac{a_T}{b_T} = \frac{\sum_{j=0}^{T-1} \omega^j y_{t-j}}{\sum_{j=0}^{T-1} \omega^j} \cong (1-\omega) y_T + \omega y_{T \mid T-1}; \; L = 1,2,\ldots \tag{2.6}$$

onde $Y_T = (y_1, y_2, \ldots, y_T)'$. A aproximação em (2.6) é válida para valores grandes de T, e portanto a função de previsão L-passos-à-frente do modelo pode ser aproximada por um amortecimento exponencial das observações passadas, com constante de amortecimento $\alpha = (1-\omega)$. É fácil ver que, se $\omega = 1$, a função de previsão em (2.6) reduz-se à média aritmética das observações.

Conforme veremos na seção 3, a distribuição preditiva 1-passo-à-frente em (2.4) será utilizada neste trabalho para simular resultados para as partidas da rodada seguinte de um campeonato. No entanto, para gerar os resultados de duas ou mais rodadas à frente, precisamos das distribuições preditivas L-passos-à-frente, para $L > 1$. No caso do modelo Poisson-Gama, não é possível derivar analiticamente estas distribuições. Fernandes (1990) sugere um método para obter numericamente a distribuição preditiva L-passos-à-frente do modelo. O método consiste em calcular $a_{T+k|T+k-1}$ e $b_{T+k|T+k-1}$, através de (2.2), substituindo as observações y_{T+k} pelas previsões $y_{T+k|T}$, $k = 1,2,\ldots,L$. Os valores obtidos são então utilizados para calcular a preditiva, através de (2.4).

2.2 – Introdução de variáveis explicativas

Seja $\underline{X}_t$ um vetor *(kx1)* de variáveis explicativas no instante *t*, o efeito de $\underline{X}_t$ no modelo Poisson-Gama pode ser representado da seguinte forma:

$$\mu_t^+ = \mu_t \exp(\underline{X}_t' \underline{\delta}) \tag{2.7}$$

onde $\underline{\delta}$ é um vetor *(kx1)*, cujas componentes são os coeficientes das variáveis explicativas. Observe que o efeito de $\underline{X}_t$ é incorporado diretamente ao nível, e não na equação das observações, como no modelo de nível local (ver Harvey, 1989). A função de ligação exponencial garante que μ_t^+ assuma somente valores positivos. É fácil provar que, sob (2.7), as equações de previsão em (2.2) tornam-se:

$$\mu_t^+ | Y_{t-1} \sim \mathrm{G}(a_{t|t-1}^+ = \omega a_{t-1}, b_{t|t-1}^+ = \omega b_{t-1} \exp(-\underline{X}_t' \underline{\delta})) \tag{2.8}$$

ou seja: $a_{t|t-1}^+ = a_{t|t-1}$, mas $b_{t|t-1}^+ \neq b_{t|t-1}$. As equações de atualização em (2.3) tornam-se:

$$\mu_t^+ | Y_t \sim G(a_t^+ = \omega a_{t-1} + y_t, b_t^+ = \omega b_{t-1} + \exp(\underline{X}_t' \underline{\delta})) \tag{2.9}$$

ou seja: $a_t^+ = a_t$, mas $b_t^+ \neq b_t$. O vetor $\underline{\delta}$ é estimado conjuntamente com ω por máxima verossimilhança, onde a função de log-verossimilhança $l(\omega, \underline{\delta})$ é dada pela expressão (2.5), substituindo $b_{t|t-1}$ por $b_{t|t-1}^+$. A função de previsão 1-passo-à-frente é dada por:

$$y_{T+1|T} = E(y_{T+1}|Y_T) = \frac{a_T}{b_T} = \frac{\exp(\underline{X}_{T+1}' \underline{\delta}) \sum_{j=0}^{T-1} \omega^j y_{t-j}}{\sum_{j=0}^{T-1} \omega^j \exp(\underline{X}_{T+1}' \underline{\delta})} \tag{2.10}$$

A expressão da função de previsão L-passos-à-frente do modelo é bastante similar a (2.10), e sua derivação é bastante simples (ver Harvey e Fernandes, 1989). A distribuição preditiva L-passos-à-frente pode ser obtida através do método apresentado na seção 2.1.

3. Estimação de probabilidades via máxima entropia e simulação

3.1 – O princípio da máxima entropia e as distribuições dos resultados das pPartidas

O conceito de entropia originou-se na física, nos estudos de Boltzman (1870), mas foi Shannon (1948) quem definiu uma medida de incerteza baseada neste conceito, permitindo associá-lo a uma distribuição de probabilidade. Seja X uma variável aleatória do tipo discreto, cuja função de probabilidade é dada por $P(X = x) = P_x$, $x = 0, 1, ..., k$, a entropia de $P(X = x)$ é definida como:

$$H(P_0, P_1, ..., P_k) = -\sum_{x=0}^{k} P_x \ln(\mathrm{P_x}) \tag{3.1}$$

A expressão em (3.1) quantifica a incerteza associada à $P(X = x)$, ou a ignorância sobre os possíveis resultados do experimento por ela representados (para a justificativa, ver Golan, Judge e Miller, 1996). A idéia é que, quanto menor a quantidade de informação acerca de determinado fenômeno, maior a entropia da distribuição de probabilidade adequada. A extensão para o caso contínuo é imediata.

O Princípio da Máxima Entropia, proposto por Jaynes (1957) e estendido por Kullback (1959), consiste em obter os parâmetros de uma distribuição de tal forma que sua entropia seja maximizada, isto é, que seja introduzido o mínimo possível de informação, permitindo que os dados representem a maior parte do conhecimento sobre o fenômeno correspondente. Observe que a expressão em (3.1) deve ser encarada como função de $P_0, P_1, ..., P_k$, que são os parâmetros de $P(X = x)$. Portanto, neste caso, a aplicação do Princípio da Máxima Entropia consiste na solução do seguinte problema:

$$\text{Maximizar } H(P_0, P_1, ..., P_k) \tag{3.2 a}$$

$$\text{Sujeito a: } P_x \geq 0,\ x = 0,1,2,...,k,\ e \sum_{x=0}^{k} P_x = 1. \tag{3.2b}$$

É fácil ver que a solução de (3.2) é trivial, dada por $P_0 = P_1 = ... = P_k = 1/(k+1)$. Entretanto, na aplicação apresentada a seguir, são impostas restrições adicionais que tornam a solução não-trivial.

Seja uma partida entre duas equipes A e B, considere a representação do resultado desta partida através de duas variáveis aleatórias: X_1 = número de gols de A e X_2 = número de gols de B (por exemplo, se a equipe A vence por *3 x 2*, então $X_1 = 3$ e $X_2 = 2$). Nesta seção, apresentamos um procedimento para estimar as distribuições de X_1 e X_2: $P(X_1 = x)$ e $P(X_2 = x)$, a partir das séries de gols a favor e sofridos pelas equipes em suas partidas anteriores[1]. Neste trabalho, para cada uma destas séries, ajustamos um modelo univariado para dados de contagem, o modelo Poisson-Gama. Para cada modelo ajustado, temos uma distribuição preditiva 1-passo-à-frente, no caso, dada pela expressão (2.4). Portanto, para cada partida, temos um total de quatro distribuições preditivas, que podem ser combinadas para gerar a distribuição associada ao seu resultado.

Sejam $P_{11}(x)$ e $P_{12}(x)$, respectivamente, as preditivas 1-passo-à-frente dos modelos ajustados às séries de gols a favor e sofridos pela equipe A, e sejam $P_{22}(x)$ e $P_{21}(x)$ as preditivas dos modelos para as séries referentes à equipe B, então, $P(X_1 = x)$ deve ser obtida a partir da combinação das informações provenientes de $P_{11}(x)$ e $P_{21}(x)$, enquanto $P(X_2 = x)$ deve ser obtida a partir de $P_{12}(x)$ e $P_{22}(x)$. Uma forma adequada de combinar estas distribuições, de tal forma que as probabilidades resultantes estejam entre *0* e *1*, é através de combinações convexas, isto é:

$$P(X_i = x) = \alpha_i P_{1i}(x) + \beta_i P_{2i}(x);\ x = 0,1,...,7;\ \alpha_i, \beta_i > 0,\ \alpha_i + \beta_i = 1,\ i = 1,2. \tag{3.3}$$

[1] A principal razão para não utilizarmos somente os resultados das partidas anteriores entre A e B, o que levaria ao uso de modelos bivariados, é que, em geral, o espaço de tempo entre tais partidas é amplo o suficiente para que não haja estrutura de dependência significativa, tornando sem sentido a abordagem do problema através de modelos de séries temporais.

onde $\sum_{x=0}^{7} P(X_i = x) = 1;\ i = 1,2$. Observe que fixamos $k = 7$ (valores de $X > 7$ são bastante raros). Os coeficientes α_i, β_i, $i = 1,2$ são obtidos através da aplicação do Princípio da Máxima Entropia, ou seja, da seguinte formulação do problema (3.2):

$$\textit{Maximizar} \sum_{x=0}^{7} P(X_i = x) \ln[P(X_i = x)] \tag{3.4a}$$

$$\textit{Sujeito a:} \sum_{x=0}^{7} P(X_i = x) = 1,\ \alpha_i,\ \beta_i > 0 \text{ e } \alpha_i + \beta_i = 1, \tag{3.4b}$$

onde $P(X_i = x)$, $i = 1,2$ são dadas em (3.3). Para obter as distribuições das partidas de duas ou mais rodadas à frente, basta substituir $P_{ij}(x);\ i,j = 1,2,$ em (3.3)-(3.4), pelas correspondentes distribuições preditivas L-passos-à-frente, obtidas pelo método da seção 2.1.

3.2 – Simulação Estocástica e Estimação das Probabilidades de Classificação e Rebaixamento das Equipes Participantes de um Campeonato

Simulação estocástica é a geração de amostras de variáveis aleatórias em ambiente computacional. Estas amostras podem ser utilizadas para obter uma solução aproximada para determinado problema, cuja solução exata inexiste. No caso do cálculo de probabilidades em campeonatos de futebol, a solução existe, mas é inviável, uma vez que demanda considerar todas as combinações possíveis de resultados para as partidas restantes. A idéia aqui é gerar resultados razoáveis para estas partidas, via simulação, e obter uma projeção da classificação final, à qual chamamos um possível cenário para o campeonato.

A geração de resultados para as partidas restantes pode ser efetuada a partir das distribuições $P(X_j = x)$, $x = 0,1,...,7;\ j = 1,2$ (ver seção 3). A Transformação Integral de Probabilidades (James, 1981) permite simular valores destas distribuições, da seguinte forma: para $= 1,2\ i,...,L$ (número de partidas restantes), gera-se X ~ unif$(0,1)$, e para $x = 0,1,...,7$, fazemos $X_j = x$, se e somente se o valor gerado pertencer ao intervalo $[\sum_{k=0}^{x-1} P(X_j = k), \sum_{k=0}^{x} P(X_j = k)]$, j = 1,2. Sujeito a: $\sum_{x=0}^{7} P(X_i = x) = 1,\ \alpha_i,\ \beta_i > 0 \text{ e } \alpha_i + \beta_i = 1.$

Sejam agora n eventos de interesse E_j, $j = 1,2,...,n$, como por exemplo: E_1 = {"equipe A classifica-se para a segunda fase"} ou E_2 = {"equipe B é rebaixada para a segunda divisão"}. Considere a geração independente de N cenários, e seja X_{ij} uma variável aleatória que assume valor 1 se o evento E_j ocorre no i-ésimo cenário, e valor 0, caso contrário, $i = 1,2,...,N;\ j = 1,2,...,n$. Então: X_{ij} ~ Bernoulli(p), onde p é a probabilidade de E_j. Se $\hat{p} = \sum_{i=1}^{N} X_{ij} / N$, prova-se que $\hat{p} \xrightarrow{p} p$, e portanto $\hat{p}$ é um estimador consistente

para p (ver Mood e Graybill, 1974). Desta forma, para N grande, podemos utilizar a freqüência relativa de ocorrência de cada evento de interesse E_j como uma estimativa confiável para sua probabilidade p_j, $j = 1,2,...,n$. Intervalos de Confiança podem ser obtidos através do Teorema Central do Limite (James, 1981). Maiores detalhes em Campos (2001).

4. Ajuste do modelo Poisson-Gama às séries de gols

Nesta seção, apresentamos os principais resultados do ajuste do modelo Poisson-Gama a 1 836 séries de gols, referentes aos resultados das partidas válidas pela primeira fase das três edições mais recentes do Campeonato Brasileiro de Futebol (1999-2001), e que serão utilizadas nas aplicações deste trabalho.

4.1 – Testes de validação, estimativas e diagnósticos da qualidade do ajuste

4.1.1 – Comparação com o modelo Pascal-Beta e verificação do ajuste da preditiva aos dados

Comparamos o modelo Poisson-Gama com o modelo Pascal-Beta (Harvey e Fernandes, 1989), através do critério de informação de Akaike (AIC) (ver Box e Jenkins, 1976), e verificamos que o modelo Poisson-Gama apresentou menor AIC para 1 734 séries, ou seja, 94,4% do total.

Em seguida, utilizamos testes de validação, para investigar se a distribuição preditiva do modelo Poisson-Gama, no caso a binomial negativa em (2.4), é adequada para representar as séries de gols. A Figura 4.1 ilustra o ajuste desta distribuição a duas séries, e a Tabela 4.1 apresenta resultados de três testes de aderência, aplicados a todas as séries utilizadas neste trabalho.

Figura 4.1 – Ajuste da distribuição binomial negativa às séries de gols a favor do Vasco (gráfico da esquerda) e sofridos pelo Palmeiras (gráfico da direita), nas 20 primeiras partidas da Copa João Havelange/2000

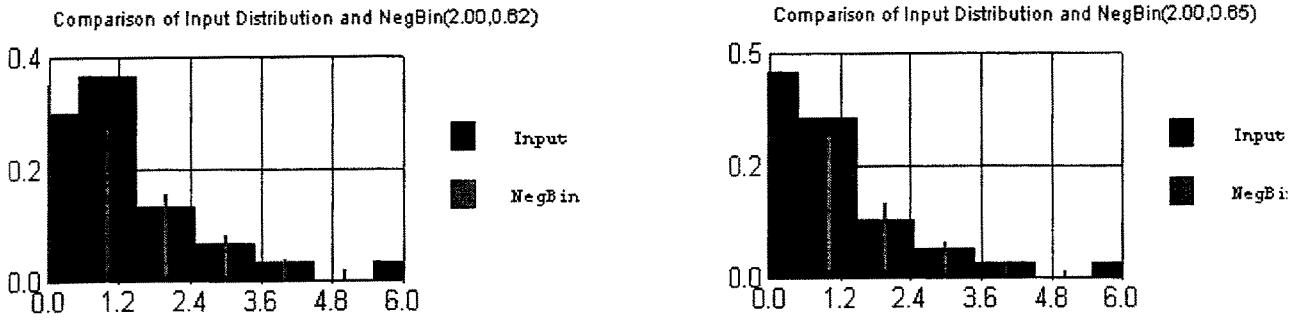


Tabela 4.1 – Testes de aderência da distribuição binomial negativa às séries de gols

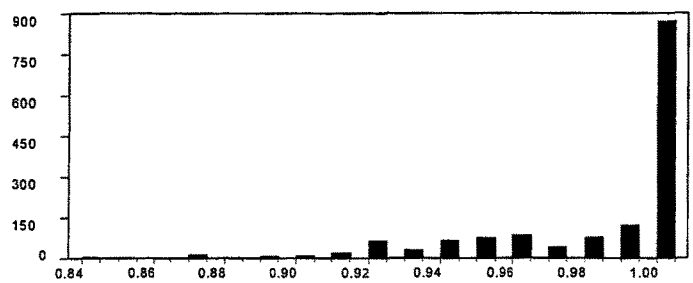
Testes ↓	H_0 não é rejeitada em nível $\alpha = 0.05$	H_0 não é rejeitada em nível $\alpha = 0.01$
Qui-Quadrado	93.1%	97.0%
Kolmogorov-Smirnov	92.4%	93.9%
Anderson-Darling	90.3%	93.4%

Verificamos que, para mais de 90% das séries analisadas, não temos evidência estatística para rejeitar a hipótese de que a distribuição binomial negativa ajuste-se bem aos dados, em nível de 0.01.

4.1.2 – Estimativas de ω, do nível μ_t e previsões 1-passo-à-frente

A Figura 4.2 apresenta as estimativas de máxima verossimilhança de ω, obtidas através do método de Levemberg-Marquardt (ver Bazaara et al., 1993).

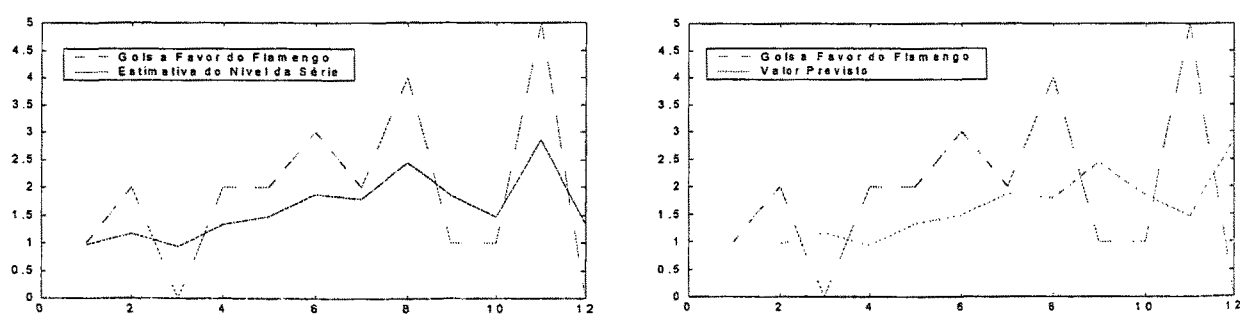
Figura 4.2 – Estimativas de ω, para as 1 836 séries utilizadas neste trabalho



Observe que, para quase 60% das séries, obtivemos $\hat{\omega} = 1$. O menor valor obtido foi $\hat{\omega} = 0.8402$, correspondente à série de gols sofridos pelo São Caetano/SP, nas 12 primeiras partidas do campeonato de 2001. Uma possível explicação é a instabilidade da campanha da equipe do ABC paulista, no início do campeonato, que se reflete em baixos valores para ω (ver Campos, 2001).

Para cada modelo ajustado, foram obtidas estimativas seqüenciais do nível e previsões 1-passo-à-frente das observações. A Figura 4.2 ilustra estas estimativas para a série de gols a favor do Flamengo nas 12 primeiras partidas do Campeonato Brasileiro de 1999.

Figura 4.2 – $\hat{\mu}_t$ (gráfico da esquerda) e $\hat{y}_{t|t-1}$ (gráfico da direita) *versus* tempo, para a série de gols a favor do Flamengo nas 12 primeiras partidas do Campeonato Brasileiro de 1999



Observe que as estimativas do nível acompanham melhor o movimento da série real do que os valores previstos, o que é esperado não somente porque a estimativa do nível incorpora a observação no instante t (o que

não ocorre com as previsões), mas também devido ao pequeno tamanho das séries utilizadas, o que torna as estimativas mais sensíveis à novas observações.

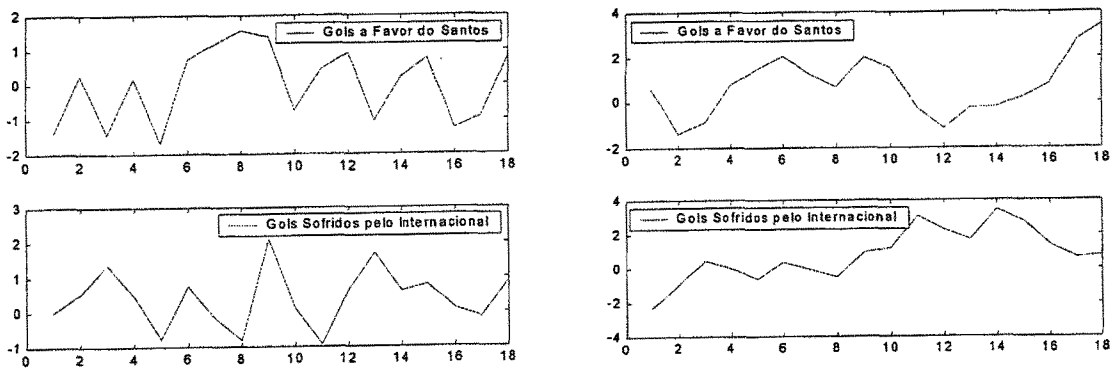
4.1.3 – Testes de diagnóstico

A inovação padronizada do modelo Poisson-Gama, no instante t, é dada por:

$$\upsilon_t = \frac{y_t - E(y_t|Y_{t-1})}{\sqrt{Var(y_t|Y_{t-1})}} \tag{4.1}$$

Os diagnósticos da qualidade do ajuste do modelo são baseados nas séries de inovações padronizadas $\{\upsilon_t; t = \tau+1, \tau+2, ..., T\}$ e acumuladas $\{\text{CUSUM(t)} = \sum_{i=\tau+1}^{t} \upsilon_i; t = \tau+1, \tau+2, ..., T\}$. Sob a hipótese de que o modelo está corretamente especificado, estas séries são estacionárias e não apresentam estrutura de dependência. Para exemplificar, consideramos as séries de gols a favor do Santos e de gols sofridos pelo Internacional, nas 18 primeiras partidas válidas pela Copa João Havelange/2000. A Figura 4.3 apresenta os gráficos pertinentes.

Figura 4.3 – Inovações padronizadas (esquerda) e acumuladas (direita) *versus* tempo, do modelo ajustado às séries de gols a favor do Santos e sofridos pelo Internacional, nas 18 primeiras partidas da Copa João Havelange/2000



Os gráficos acima apresentam comportamento estacionário. Além disso, as estimativas da FAC e FACP (para ambas as séries) estão dentro dos limites de confiança de 95% (ver Campos, 2001).

Aplicamos testes de diagnóstico para todos os modelos ajustados. Para investigar as hipóteses de independência e homocedasticidade, utilizamos, respectivamente, os testes de carreiras e de Goldfeld-Quandt (ver, por exemplo, Gujarati, 2000). Verificamos que a qualidade do ajuste foi satisfatória para 88.5% dos modelos ajustados, em nível de 0.05, e para 94,6%, em nível de 0.01.

4.2 – Variáveis explicativas: estimativas e interpretação dos coeficientes

Uma das principais dificuldades na modelagem probabilística de resultados de partidas de futebol é que o desempenho das equipes é afetado não somente pelos resultados das partidas anteriores, mas também por outros fatores, como a escalação das equipes, alterações na comissão técnica e mando de campo. Nesta seção, veremos como estes fatores podem ser adequadamente representados através de variáveis explicativas binárias no modelo Poisson-Gama, cujos coeficientes estimados permitem quantificar e testar a significância dos efeitos correspondentes. Diversas combinações de variáveis explicativas foram analisadas neste trabalho, com o objetivo de melhorar a qualidade do ajuste dos modelos e investigar o efeito dos fatores supracitados.

Apresentamos a seguir algumas das conclusões mais interessantes oriundas desta análise. São reportadas apenas as estimativas pontuais e sua interpretação. Os procedimentos para testar a significância das variáveis são não-triviais, e serão apresentados na seção 5.

4.2.1 – Escalação das Equipes: Quais são os Jogadores que Desequilibram uma Partida ?

O efeito da escalação das equipes sobre os resultados das partidas pode ser investigado através de variáveis explicativas binárias. Em particular, a influência da participação de um jogador específico, considerado decisivo, sobre o número médio de gols marcados ou sofridos por uma equipe pode ser representada por uma variável X_t, que vale 1 se o jogador disputa a t-ésima partida, e vale 0, caso contrário. Como vimos na seção 2, esta variável pode ser incorporada na média do modelo através de (2.7), neste caso: $\mu_t^+ = \mu_t \exp(X_t\delta)$, onde δ é o coeficiente de X_t. Se $\hat{\delta}$ é a estimativa de δ, é fácil ver que, se $X_t = 1$, $\mu_t^+ = \mu_t \exp(\hat{\delta})$, ou seja, de acordo com esta especificação, o número médio de gols é multiplicado por $\exp(\hat{\delta})$ quando o jogador está em campo.

Reportamos a seguir quatro casos: Guilherme, do Atlético/MG, e Alex Alves, do Cruzeiro (campeonato de 1999), Adhemar, do São Caetano (2000), e Romário, do Vasco (2001). Investigamos o efeito da presença destes jogadores sobre o desempenho dos ataques de suas respectivas equipes, nas referidas competições, utilizando as séries de gols a favor destas equipes ao longo das 15 primeiras partidas disputadas na primeira fase de cada competição.

A Tabela 4.2 apresenta as estimativas de máxima verossimilhança de δ, em cada caso.

Tabela 4.2 – Estimativas de δ, para investigar o efeito da presença de alguns jogadores

Jogador (equipe, ano) ↓	$\hat{\delta}$	$\exp(\hat{\delta})$
Guilherme (Atlético/MG,1999)	0.47	1.60
Alex Alves (Cruzeiro,1999)	0.38	1.46
Adhemar (São Caetano, 2000)	0.16	1.17
Romário (Vasco, 2001)	0.59	1.80

A interpretação é a seguinte: a média de gols a favor do Atlético/MG, em 1999, foi 60% mais elevada nas partidas em que Guilherme estava em campo, enquanto as presenças de Alex Alves e Adhemar aumentaram a

média de gols de suas equipes em, respectivamente, 46% e 17%. Nada comparável ao desempenho de Romário, no campeonato de 2001. Quando o jogador estava em campo, o Vasco marcava, em média, 80% mais gols. Não obstante, Romário não foi artilheiro do campeonato de 2001, ao contrário de Guilherme, em 1999 (Alex Alves foi o vice), e de Adhemar, em 2000.

Testamos também o efeito conjunto da presença de algumas duplas de ataque. Em muitos casos, verificamos que a influência da presença de um jogador é significativa apenas quando outro jogador também está em campo, ou seja, quando ocorre um efeito de interação. Por exemplo, introduzimos uma variável para representar a interação entre Guilherme e Marques (Atlético/MG), em 1999, e verificamos que o efeito da presença de Guilherme ocorre em grande parte através da interação com Marques. Quando consideramos somente as partidas em que Guilherme atuou sem a presença do parceiro, o efeito da participação de Guilherme sobre o número de gols assinalados pela equipe mineira foi de apenas 14% ($\hat{\delta} = 0.13$)[2].

No caso de Romário, analisamos também sua participação nos campeonatos de 1999 (quando atuava pelo Flamengo) e 2000. Em 1999, obtivemos $\hat{\delta} = 0.16$, que corresponde a um efeito de 17% sobre o número médio de gols a favor da equipe rubro-negra. Em 2000, obtivemos $\hat{\delta} = 0.33$, ou seja, um efeito de 39% sobre a média de gols assinalados, maior do que na época do Flamengo, mas substancialmente menor em relação ao seu desempenho no campeonato de 2001, o que evidencia a evolução da performance do jogador[3].

4.2.2 – Mudança de Técnico: Antônio Clemente e a recuperação do Botafogo/RJ em 1999

No futebol brasileiro, é comum o treinador ser responsabilizado por uma seqüência de resultados negativos de uma equipe, sendo freqüentes as mudanças de técnico no decorrer de uma competição. Muitas vezes, observamos que estas mudanças acabam realmente por surtir efeito, seja pela maior identificação do novo treinador com a equipe, ou mesmo por influência psicológica. Considere o caso do Botafogo, no Campeonato de 1999. A equipe fazia uma péssima campanha, sendo inclusive forte candidata ao rebaixamento, até que, após a décima-terceira rodada, Antônio Clemente assumiu o comando, obtendo seis vitórias em oito partidas. Podemos estimar o efeito do trabalho de Clemente através da variável: $X_t = 1$, *se* $t \geq \tau$, e $X_t = 0$, *se* $t < \tau$, onde τ é o índice da partida em que a equipe mudou de treinador (no caso, $\tau = 14$). Ajustamos o modelo Poisson-Gama às séries de gols a favor e sofridos pelo Botafogo ao longo das 21 partidas disputadas e obtivemos, respectivamente, $\hat{\delta} = 0.37$ *e* -0.29. Como $exp(0.37) = 1{,}45$ e $exp(-0.29) = 0{,}75$, podemos concluir que a média de gols a favor aumentou 45% e a média de gols sofridos reduziu 25%, após a mudança.

[2] Vale ressaltar que não devemos investigar simultaneamente o efeito de muitas variáveis e/ou interações entre elas, uma vez que o ganho advindo em termos de ajuste deve ser balanceado com a perda de graus de liberdade, que pode prejudicar ou inviabilizar a interpretação das estimativas, em razão do pequeno número de observações das séries utilizadas.

[3] Na seção 5, descrevemos o procedimento para testar a significância do efeito das variáveis explicativas, via *bootstrap*.

4.2.3 – Fator-Campo: quais equipes tiram proveito do fato de jogar em casa ?

Existem evidências empíricas de que algumas equipes obtêm melhores resultados em partidas disputadas em seu estádio, não somente pela torcida favorável, mas também pelo fato de os jogadores estarem habituados às dimensões do campo e ao estado do gramado. Desta forma, o resultado de uma partida de futebol é claramente afetado pelo chamado "fator-campo", cujo efeito pode ser investigado através do uso de variáveis explicativas binárias no modelo Poisson-Gama. Observe que temos aqui três categorias: mandante, visitante e campo neutro. Assim, introduzimos duas variáveis, digamos x_{1t} e x_{2t}, e a expressão (2.7) torna-se: $\mu_t^+ = \mu_t \exp(\delta_1 x_{1t} + \delta_2 x_{2t})$. Os valores de x_{1t} e x_{2t} são atribuídos de tal forma que as estimativas dos coeficientes sejam interpretadas em relação à categoria "campo neutro". Ou seja, na t-ésima partida, se a equipe detém o mando de campo, $x_{1t} = 1$ e $x_{2t} = 0$, e daí $\mu_t^+ = \mu_t \exp(\delta_1 x_{1t})$; se a equipe joga fora de casa, $x_{1t} = 0$ e $x_{2t} = 0$, e $\mu_t^+ = \mu_t \exp(\delta_2 x_{2t})$; e se a partida é disputada em campo neutro, $x_{1t} = x_{2t} = 0$, e $\mu_t^+ = \mu_t$.

A Tabela 4.3 apresenta as estimativas $\hat{\delta}_1$ e $\hat{\delta}_2$, para algumas situações.

Tabela 4.3 – Estimativas de $\hat{\delta}_1$ e $\hat{\delta}_2$, para investigar o efeito do mando de campo

Série (nº de observações)		$\hat{\delta}_1$	$\exp(\hat{\delta}_1)$	$\hat{\delta}_2$	$\exp(\hat{\delta}_2)$
1	Gols a favor – Flamengo (1999)	0.21	1.23	- 0.13	0.88
2	Gols sofridos – Santos (1999)	0.34	1.40	- 0.28	0.76
3	Gols a favor – Vasco (2000)	0.31	1.36	- 0.19	0.83
4	Gols sofridos – Ponte Preta (2000)	- 0.09	0.91	0.12	1.13
5	Gols a favor – Atlético/PR (2001)	0.38	1.46	- 0.23	0.79
6	Gols sofridos – São Caetano (2001)	0.44	1.55	- 0.31	0.73

A interpretação dos valores da Tabela 4.3 é direta. Por exemplo, no caso do Flamengo, $\exp(\hat{\delta}_1) = 1.23$, o que significa que o número de gols a favor, nas partidas em que a equipe deteve o mando de campo foi, em média, 23% maior, em relação às partidas disputadas em campo neutro (ou seja, no Maracanã, contra Fluminense, Botafogo e Vasco). No caso da série de gols sofridos, exp(-0.13) = 0.88, e assim o Flamengo sofreu, em média, 12% menos gols quando jogou fora de casa.

Através da análise de todas as séries utilizadas neste trabalho, verificamos que o efeito do "fator-campo" é mais proeminente para equipes do interior e/ou de menor investimento, como América/MG, Juventude/RS e Ponte Preta/SP. Uma possível explicação é que os estádios destas equipes ficam, em geral, localizados em cidades do interior, sendo a maior parte da torcida favorável à equipe local. Incidentalmente, as dimensões do campo e o estado do gramado contribuem para que o mando de campo destas equipes seja um fator de desequilíbrio a seu favor. Não obstante, observe as estimativas para as séries de gols da Ponte Preta, no campeonato de 2000, que indicam que a equipe obteve melhor desempenho em partidas disputadas fora de seu estádio.

4.2.4 – Efeito da introdução de variáveis explicativas sobre os diagnósticos de ajuste

A introdução de variáveis explicativas nos modelos ajustados neste trabalho teve dois efeitos substanciais. O primeiro é que as estimativas de ω aumentaram quando obtidas conjuntamente com $\hat{\underline{\delta}}$. Isto porque a introdução de variáveis explicativas contribui (parcialmente) para explicar eventuais instabilidades, e, quanto maior a instabilidade da série, menor a estimativa de ω. Por exemplo, no caso do São Caetano, no campeonato de 2001, a introdução das variáveis para mando de campo modificou a estimativa de ω de 0.8402 para 0.9377. O segundo efeito foi a melhoria significativa dos diagnósticos de ajuste. Considerando os testes da seção 4.1, o percentual de modelos com ajuste satisfatório aumentou de 88.5% para 91,3% (em nível de 0.05) e de 94,6% para 96,7% (em nível de 0.01), respectivamente. Para maiores detalhes, ver Campos (2001).

5 – Bootstrap no modelo Poisson-Gama

5.1 - Intervalos de confiança e testes de hipóteses para os hiperparâmetros

A obtenção de intervalos de confiança e a formulação de testes de hipóteses para os hiperparâmetros de um modelo baseia-se, em geral, na distribuição amostral dos respectivos estimadores. Entretanto, no caso do modelo Poisson-Gama, não conhecemos as distribuições dos estimadores de ω e $\underline{\delta}$. Apresentamos a seguir um procedimento, baseado na técnica do *bootstrap* (Efron, 1979), para obter as distribuições de $\hat{\omega}$ e $\hat{\underline{\delta}}$.

Seguem os passos para a obtenção das distribuições amostrais destes estimadores:

Selecionar, aleatoriamente e com reposição, (T-τ) observações da série de inovações do modelo: $\{\upsilon_t; t = \tau+1,...,T\}$, onde υ_t é dado por (4.1). Desta forma, geramos outra série, de mesmo tamanho e com as mesmas características da série de inovações original. Repetir este processo B vezes, para obter $\{\upsilon_t^b; t = \tau+1,...,T; b = 1,2,...,B\}$ (fizemos $B = 5000$).

A partir das séries do passo 1, geramos B replicações da série original:

$$y_t^b = \hat{y}_{t|t-1} + \upsilon_t^b; t = \tau+1,...,T; b = 1,2,...,B. \tag{5.1}$$

onde $\hat{y}_{t|t-1}$ é obtido de (2.6), fazendo L = 1. As séries em (5.1) são chamadas séries *bootstrap*.

Para cada série *bootstrap*, ajustamos o modelo, obtendo assim B estimativas de ω e das componentes de $\underline{\delta}$. As distribuições dos respectivos estimadores são estimadas a partir das distribuições de freqüências destas estimativas, chamadas distribuições *bootstrap*.

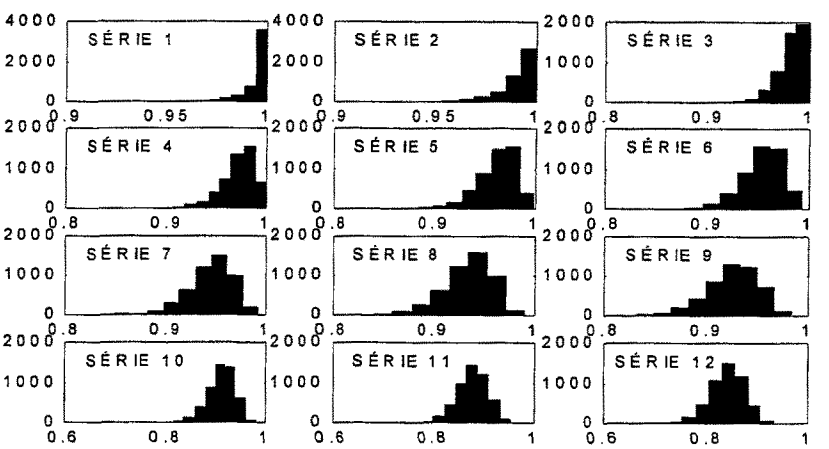
A partir das distribuições *bootstrap*, obtemos Intervalos de confiança para ω e $\underline{\delta}$, selecionando percentis adequados, tais que os tamanhos dos intervalos sejam os menores possíveis. A Tabela 5.1 apresenta as

estimativas pontuais de ω, para algumas séries, e os intervalos de confiança associados (via *bootstrap*). A Figura 5.1 ilustra as distribuições *bootstrap* correspondentes.

Tabela 5.1 – Estimativas pontuais e intervalares (via *bootstrap*) p/ ω

Série (nº de observações)		$\hat{\omega}$	IC p/ $\alpha = 0.05$	IC p/ $\alpha = 0.01$
1	Gols a favor – Vasco (19)	1.0000	[0.9883,1.0000]	[0.9736,1.0000]
2	Gols sofridos – São Paulo (16)	0.9902	[0.9698,0.9979]	[0.9554,0.9996]
3	Gols a favor – Coritiba (13)	0.9805	[0.9492,0.9943]	[0.9341,0.9962]
4	Gols sofridos – Flamengo (18)	0.9711	[0.9304,0.9838]	[0.9210,0.9928]
5	Gols a favor – Gama (21)	0.9622	[0.9182,0.9834]	[0.9077,0.9860]
6	Gols sofridos – Sport (14)	0.9523	[0.9055,0.9812]	[0.8987,0.9830]
7	Gols a favor – Juventude (20)	0.9434	[0.8992,0.9755]	[0.8938,0.9786]
8	Gols a favor – Santos (17)	0.9354	[0.8885,0.9738]	[0.8902,0.9759]
9	Gols sofridos – Ponte Preta (15)	0.9264	[0.8803,0.9685]	[0.8725,0.9707]
10	Gols a favor – Corinthians (12)	0.9091	[0.8552,0.9536]	[0.8399,0.9654]
11	Gols sofridos – Paraná Clube (15)	0.8774	[0.8246,0.9333]	[0.8075,0.9508]
12	Gols sofridos – São Caetano (12)	0.8402	[0.7810,0.9168]	[0.7681,0.9380]

Figura 5.1 – Distribuições *bootstrap* p/ ω

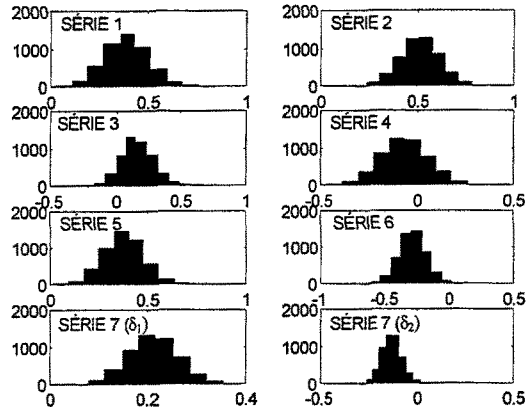


A Tabela 5.2 apresenta estimativas pontuais e intervalares (via *bootstrap*) de $\underline{\delta}$, para algumas séries. A Figura 5.2 ilustra as distribuições *bootstrap* correspondentes.

Tabela 5.2 – Estimativas pontuais e intervalares p/ $\underline{\delta}$

Série (nº de observações)	$\hat{\delta}$	IC p/ $\alpha = 0.05$	IC p/ $\alpha = 0.01$
Gols a favor – Cruzeiro (15) (Alex Alves)	0.38	[0.2473,0.5127]	[0.0815,0.6785]
Gols a favor – Atlético/MG (15) (Guilherme)	0.47	[0.3415,0.5915]	[0.1832,0.7508]
Gols a favor – São Caetano (15) (Adhemar)	0.16	[0.0210,0.2998]	[-0.1507,0.4723]
Gols sofridos – Ponte Preta (15) (campo)	-0.09	[-0.2275,0.0483]	[-0.3611,0.2211]
Gols a favor – Botafogo/RJ (21) (técnico)	0.37	[0.2435,0.4965]	[0.0854,0.6546]
Gols sofridos – Botafogo/RJ (21) (técnico)	-0.29	[-0.4165,-0.1635]	[-0.5746,-0.0054]
Gols a favor – Flamengo (21) (campo)	$\hat{\delta}_1 = 0.21$	[0.1206,0.2994]	[0.0088,0.4112]
	$\hat{\delta}_2 = -0.13$	[-0.2194,-0.0406]	[-0.3312,0.0712]

Figura 5.2 – Distribuições *bootstrap* p/ $\underline{\delta}$



Podemos utilizar os intervalos da Tabela 5.2 para investigar a significância das variáveis explicativas no modelo. Por exemplo, para a série 3, observamos que o intervalo de 99% contém o zero, e assim H_0: $\delta = 0$ não é rejeitada a nível 0.01. Logo, neste nível de significância, não temos evidência estatística de que a presença do jogador Adhemar tenha exercido efeito significativo sobre o número de gols a favor do São Caetano, na Copa João Havelange/2000. No caso da série 4, observamos que os intervalos de 95% e 99% contém o zero, e assim H_0 não é rejeitada em nível de 0.05. Por outro lado, observando os resultados obtidos para as séries 5 e 6, verificamos que o efeito de Antônio Clemente sobre a recuperação do Botafogo/RJ no campeonato de 1999 foi significativo, mesmo em nível de 0.01.

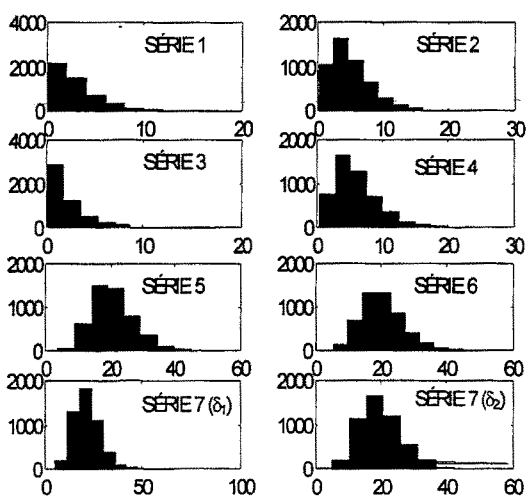
5.2 – Teste de razão de verossimilhanças

Quando temos mais de uma variável explicativa no modelo, um procedimento mais geral, que permite inclusive testar a significância do efeito conjunto das variáveis, é o teste de razão de verossimilhanças. Todavia, existe um problema específico aqui. A distribuição da estatística do teste, sob a hipótese nula, é aproximadamente qui-quadrado (ver, por exemplo, Gujarati, 2000). Entretanto, esta aproximação é válida apenas para grandes amostras, e as séries utilizadas neste trabalho têm poucas observações, variando entre 12 e 23.

Podemos contornar este problema, utilizando a técnica do *bootstrap* para obter a distribuição da estatística de razão de verossimilhanças para pequenas amostras, através de um procedimento análogo ao apresentado na seção 5.1. Neste caso, a região crítica para o teste de razão de verossimilhanças é definida, caso a caso, pelos percentis da distribuição *bootstrap*.

A Figura 5.3 apresenta as distribuições bootstrap da estatística de razão de verossimilhanças, para as séries apresentadas na Tabela 5.2.

Figura 5.3 – Distribuições *bootstrap* p/ a estatística de razão de verossimilhanças

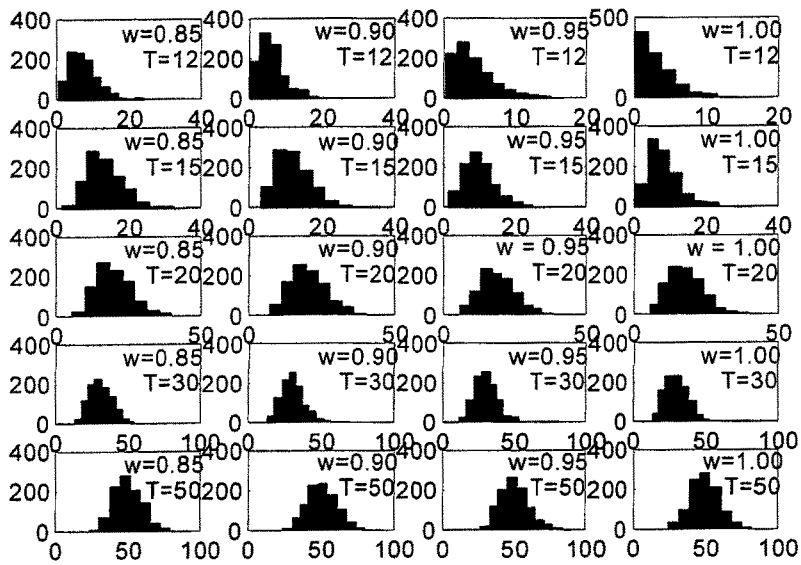


Comparamos as distribuições *bootstrap* com as correspondentes aproximações para grandes amostras, através do teste de aderência qui-quadrado. Verificamos que a qualidade da aproximação foi satisfatória para as séries 5, 6 e 7. Estas séries possuem 21 observações, conduzindo a uma qui-quadrado com mais de 20 graus de liberdade. Cabe observar que o uso da aproximação assintótica conduziu a decisões incorretas no caso das séries 1, 3 e 4. Por exemplo, para a série 3, se aplicarmos o teste sem a correção via *bootstrap*, não rejeitaremos H_0, a nível 0.05. Para detalhes, ver Campos (2001).

5.3 – Distribuição da estatística de razão de verossimilhanças para pequenas amostras

Nesta seção, investigamos, em um contexto mais geral, a qualidade da aproximação para a distribuição da estatística do teste de razão de verossimilhanças. O procedimento é análogo ao utilizado na seção 5.1. Fixamos valores de T e ω e, para cada par (T,ω), geramos B séries *bootstrap* (fizemos B = 1000; T = 12, 15, 20, 30 e 50; ω = 0.85, 0.9, 0.95 e 0.99). Em seguida, ajustamos o modelo para cada série gerada, calculamos os correspondentes valores da estatística do teste, e obtemos sua distribuição *bootstrap*. A Figura 5.4 ilustra as distribuições *bootstrap* da estatística de razão de verossimilhanças para cada par (T,ω).

Figura 5.4 – Distribuição *bootstrap* da estatística de razão de verossimilhanças, variando T e ω



Utilizamos o teste qui-quadrado para comparar as distribuições geradas com as correspondentes aproximações assintóticas, para investigar como a qualidade da aproximação varia com T e ω. A Tabela 5.3 apresenta os níveis descritivos dos testes.

Tabela 5.3 – Níveis descritivos dos testes qui-quadrado para verificar a aderência da distribuição assintótica à distribuição *bootstrap* da estatística de razão de verossimilhanças

ω\T	12	15	20	30	50
0.85	0.03	0.056	0.064	0.083	0.112
0.9	0.019	0.051	0.057	0.078	0.081
0.95	0.025	0.045	0.056	0.069	0.093
0.99	0.012	0.038	0.052	0.062	0.076

Podemos observar que, para T > 20, a aproximação qui-quadrado é satisfatória. A conclusão é que o uso da técnica do bootstrap, para obter a distribuição da estatística do teste de razão de verossimilhanças, somente se justifica para séries com menos de 20 observações.

6. Probabilidades estimadas: Campeonato Brasileiro 1999-2001

A metodologia proposta neste trabalho foi aplicada na primeira fase das três edições mais recentes do Campeonato Brasileiro de Futebol, de 1999 a 2001. Os números de equipes classificadas para a segunda fase e rebaixadas para a segunda divisão são apresentados na Tabela 6.1.

Tabela 6.1 – Número de classificados/rebaixados em cada edição do Campeonato Brasileiro

	1999	2000[4]	2001
Número de equipes participantes	22	25	28
Número de equipes classificadas p/ a segunda fase	8	12	8
Número de equipes rebaixadas p/ a segunda divisão	4	-	4

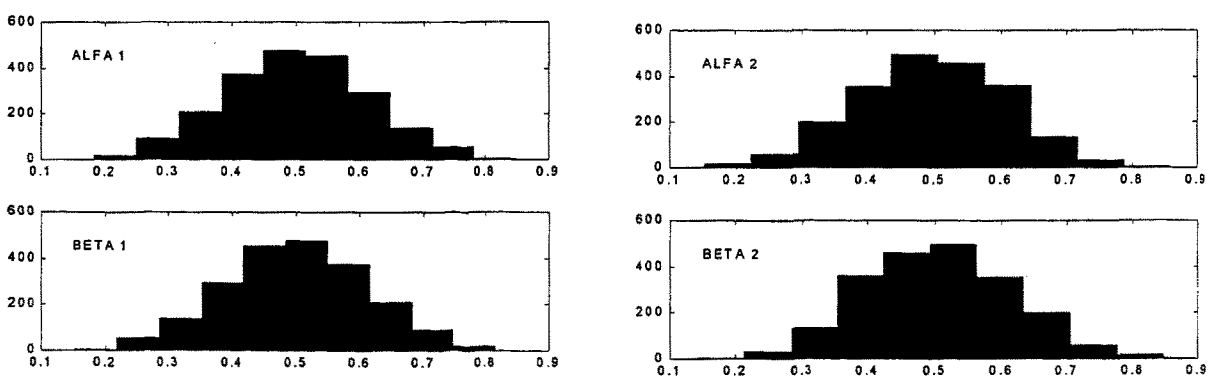
Para cada edição do campeonato, foram obtidas as distribuições dos resultados das partidas restantes, e as estimativas seqüenciais (ao final de cada rodada) das probabilidades de classificação e rebaixamento das equipes participantes (exceto em 2000, quando não houve rebaixamento).

6.1 – Distribuições de probabilidade dos resultados das partidas

6.1.1 - Coeficientes obtidos através do princípio da máxima entropia

Vimos na Seção 3 que as distribuições de probabilidade dos resultados das partidas são obtidas através da combinação convexa das preditivas dos modelos ajustados às séries de gols, e que os coeficientes destas combinações, α_1, α_2, β_1 e β_2, são obtidos pelo Princípio da Máxima Entropia, como em (3.4). A Figura 6.1 apresenta os valores de α_1, α_2, β_1 e β_2, para todas as séries.

Figura 6.1 - Coeficientes α_1, β_1, α_2 e β_2, obtidos pelo princípio da máxima entropia



[4] Considerando apenas o módulo azul da Copa João Havelange/2000.

Observe que todos os coeficientes estão bastante próximos de 0.5. Por exemplo, no caso de β_2, 73,5% dos valores obtidos estão no intervalo [0.45,0.55] e 91.2% estão no intervalo [0.4,0.6]. Para uma comparação deste procedimento com a alternativa: $\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 0.5$, ver Campos (2001).

6.1.2 – Distribuições dos resultados de algumas partidas

Apresentamos a seguir as distribuições de freqüência dos resultados de três partidas, geradas através da simulação de 10.000 valores das distribuições $P(X_i = x)$, i = 1,2, em (3.3):

Tabela 6.2 – Distribuição do resultado de Corinthians/SP (X_1) x Guarani/SP (X_2), 1999

$X_1 \backslash X_2$	0	1	2	3	4 ou mais
0	0.093	0.12	0.089	0.008	0.02
1	0.11	**0.117**	0.054	0.012	0.01
2	0.062	0.08	0.061	0	0
3	0.021	0.029	0.026	0.007	0.012
4 ou mais	0.022	0.02	0.01	0.011	0.01

Tabela 6.3 – Distribuição do resultado de Ponte Preta/SP (X_1) x Fluminense/RJ (X_2), 2000

$X_1 \backslash X_2$	0	1	2	3	**4 ou mais**
0	0.05	0.075	0.108	0.022	0.021
1	0.08	0.106	0.051	0.013	0.017
2	0.065	0.089	0.079	0.017	0.022
3	0.024	0.042	0.026	0.018	**0.011**
4 ou mais	0.025	0.023	0.004	0.02	0.012

Tabela 6.4 – Distribuição do resultado de Botafogo/RJ (X_1) x São Caetano/SP (X_2), 2001

$X_1 \backslash X_2$	0	1	2	3	**4 ou mais**
0	0.074	0.06	0.1	0.03	0.028
1	0.079	0.089	0.058	0.015	0.03
2	0.066	0.067	0.065	0.022	**0.023**
3	0.029	0.031	0.019	0.022	0.014
4 ou mais	0.031	0.025	0.005	0.013	0.005

onde as freqüências em negrito correspondem aos verdadeiros resultados destas partidas, respectivamente: 1x1 (117 ocorrências nas simulações), 3x4 (11) e 2x4 (23). As distribuições marginais fornecem as probabilidades dos três resultados básicos: por exemplo, na Tabela 6.3, temos as seguintes probabilidades: vitória da Ponte Preta: 0.378, empate: 0.265 e vitória do Fluminense: 0.357. Observe que, se tivéssemos considerado a moda desta distribuição, teríamos previsto a vitória da Ponte Preta. Por outro lado, a

probabilidade de vitória do Fluminense (resultado que realmente ocorreu) é bastante elevada. Um método para a avaliação da qualidade destas estimativas será apresentado na seção 6.3. Uma discussão mais completa pode ser encontrada em Arruda (2000).

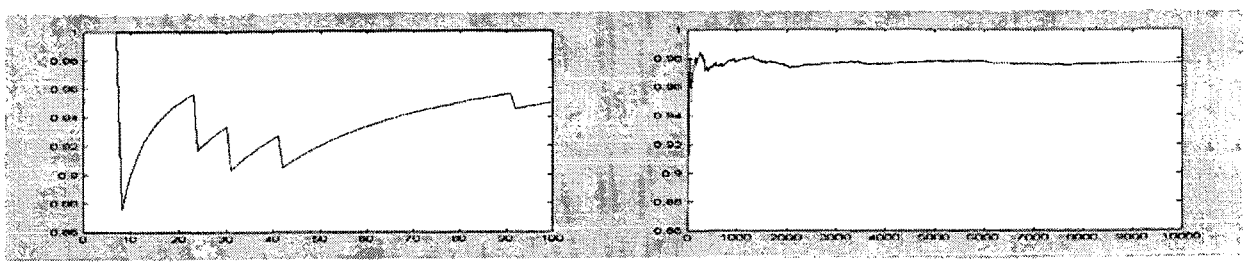
Considerando todas as séries do trabalho, verificamos que em 87.18% dos casos a correlação entre gols a favor e gols sofridos foi positiva. Portanto, o método induz uma dependência positiva entre estas séries, o que é bastante razoável, e consistente com os resultados de Arruda (2000).

6.2 – Estimativas das probabilidades de classificação e rebaixamento

6.2.1 – Determinação do número de cenários gerados x tempo computacional

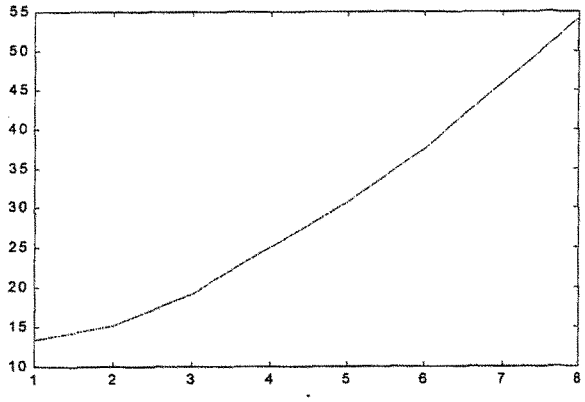
O problema de definir o número N de cenários que devemos gerar para o campeonato está relacionado à convergência das freqüências relativas para as probabilidades de interesse. Por exemplo, considere o problema de definir o número de cenários para estimar a probabilidade de rebaixamento do Botafogo/RJ, no campeonato de 1999. A Figura 6.2 apresenta a evolução das estimativas a cada cenário gerado, considerando 2 casos: N = 100 e N = 10.000.

Figura 6.2 – Evolução das estimativas da probabilidade de rebaixamento do Botafogo/RJ, na 13ª rodada do Campeonato Brasileiro de 1999, p/ N = 100 (esquerda) e N = 10 000 (direita)



Observe que somente para N > 2 000, a trajetória definida pelas estimativas começa a estabilizar. Neste trabalho, fizemos N = 10 000, que satisfez ao critério: $\left|\hat{p}_{N+1} - \hat{p}_N\right| < 0.0001$. A Figura 6.3 ilustra o tempo computacional para gerar 10 000 cenários, em função do número de rodadas restantes.

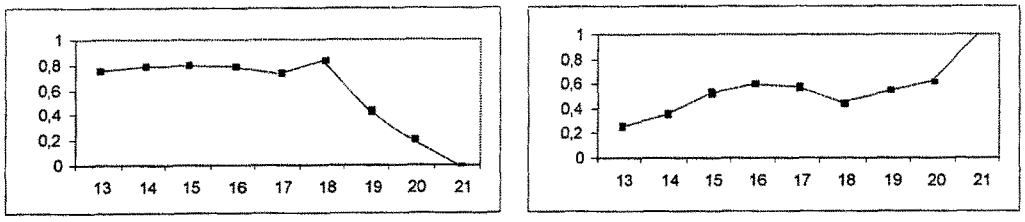
Figura 6.3 – Tempo computacional demandado x número de rodadas restantes



6.2.2 – Probabilidades estimadas para algumas equipes

Nesta seção, apresentamos estimativas seqüenciais das probabilidades de classificação e rebaixamento. Os pontos escuros em torno das estimativas representam os intervalos de confiança[5]. A Figura 6.4 apresenta as probabilidades de classificação de Flamengo/RJ e Atlético/MG, em 1999.

Figura 6.4 – Probabilidades de Classificação: Flamengo/RJ e Atlético/MG, 1999

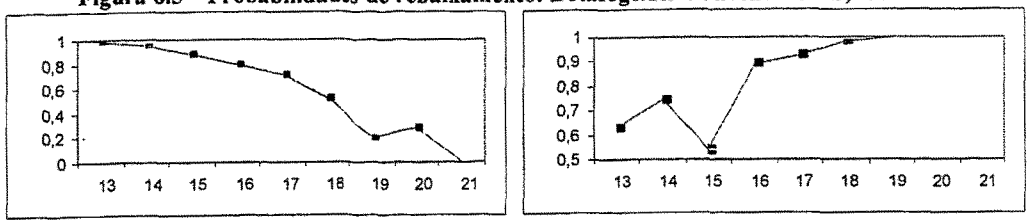


Observe que o Flamengo realizava boa campanha até a 18ª rodada, após a qual a equipe tinha 83% de chances de classificar-se para a segunda fase. Entretanto, nas últimas rodadas, três derrotas sucessivas eliminaram o Flamengo da disputa. A queda brusca na probabilidade de classificação deve-se ao fato de que, nas últimas rodadas, as estimativas são mais sensíveis aos resultados. Situação exatamente oposta ocorreu com o Atlético/MG, cujas estimativas indicam uma campanha irregular até a 18ª rodada, seguida de uma ascensão nas últimas três rodadas. A equipe mineira somente garantiu classificação na 21ª e última rodada, em virtude não somente de haver ganho sua partida, mas também de uma combinação favorável dos outros resultados[6]. A figura 6.5 apresenta as probabilidades de rebaixamento de Botafogo/RJ e Juventude/RS, em 1999.

[5] Estes intervalos, obtidos através do Teorema Central do Limite (ver seção 3), refletem apenas a incerteza inerente à estimação das probabilidades através dos resultados das simulações, não incorporando a parcela referente à previsão dos resultados futuros. Por esta razão, os intervalos são estreitos, mas sua aplicabilidade é restrita. Detalhes em Campos (2001).

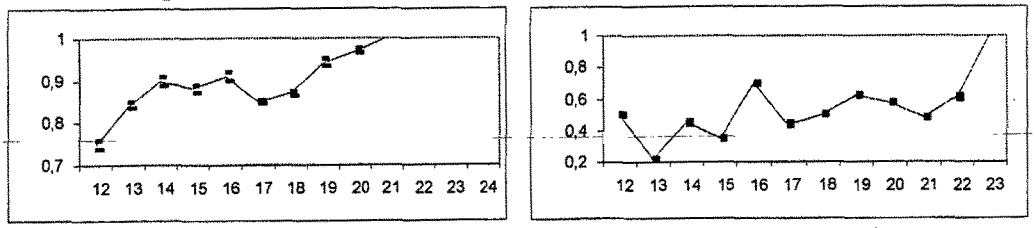
[6] Isto evidencia a natureza multivariada do problema.

Figura 6.5 – Probabilidades de rebaixamento: Botafogo/RJ e Juventude/RS, 1999



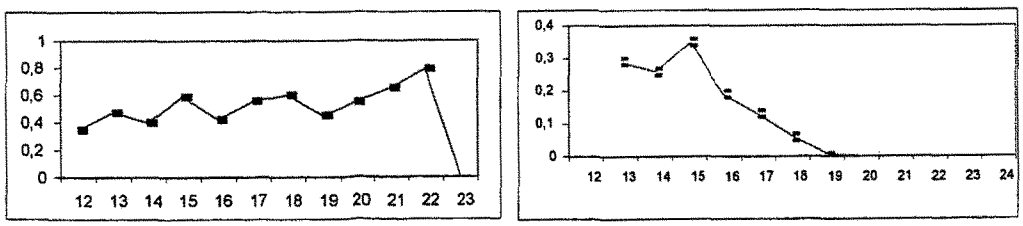
Nas 13 primeiras rodadas, o Botafogo/RJ havia perdido onze partidas e empatado duas, ocupando o último lugar na tabela. Após a 13ª rodada, a probabilidade de rebaixamento estimada foi 0.9761. Sob o comando de Antônio Clemente, a equipe ganhou seis das oito partidas restantes, e permaneceu na primeira divisão (ver seção 4.2.2). Já o Juventude/RS realizou péssima campanha do início ao fim do campeonato e, a duas rodadas do final, já estava rebaixado com probabilidade um. A Figura 6.6 apresenta as probabilidades de classificação de Vasco/RJ e Bahia/BA, na Copa João Havelange/2000.

Figura 6.6 – Probabilidades de Classificação: Vasco/RJ e Bahia/BA, 2000



Embora ambas as equipes tenham obtido a classificação para a segunda fase, observe a diferença entre as campanhas: o Vasco teve um desempenho muito mais regular, e classificou-se com duas rodadas de antecedência. Esta situação refletiu-se não somente nas estimativas das probabilidades, mas também no confronto entre as equipes na segunda fase, do qual o Vasco saiu vitorioso, e seguiu adiante na competição, tendo inclusive conquistado o título. A Figura 6.7 apresenta as probabilidades de rebaixamento de Guarani/SP e Santa Cruz/PE, na mesma competição.

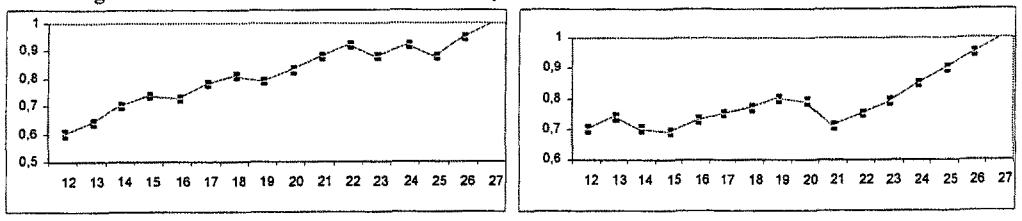
Figura 6.7 – Probabilidades de Classificação: Guarani/SP e Santa Cruz/PE, 2000



O Santa Cruz/PE realizou péssima campanha desde o início do campeonato. Após a 15ª rodada, a equipe só obteve resultados negativos, tendo sido eliminada a cinco rodadas do final da primeira fase. Por outro lado, o Guarani/SP havia obtido bons resultados ao longo da competição e, após a penúltima rodada, suas chances de classificação eram de 78%. Porém, na última rodada, a equipe foi surpreendida por uma derrota para o

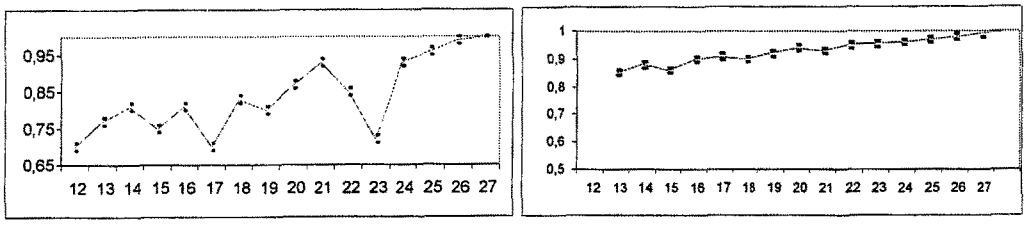
Palmeiras, concorrente direto à vaga, e foi eliminada. A Figura 6.8 apresenta as probabilidades de classificação de Atlético/PR e São Caetano/SP, em 2001.

Figura 6.8 – Probabilidades de Classificação: Atlético/PR e São Caetano/SP, 2001



Observe que as campanhas do campeão e vice brasileiros, naquele ano, foram similares, sendo as trajetórias das estimativas ascendentes e regulares. A Figura 6.9 apresenta as probabilidades de rebaixamento de Sport/PE e Botafogo/SP, na mesma competição.

Figura 6.9 – Probabilidades de Rebaixamento: Sport/PE e Botafogo/SP, 2001



O Sport/PE alternou bons e maus resultados ao longo da competição, enquanto o Botafogo/SP esteve sempre entre os quatro piores colocados, e esta situação reflete-se nas estimativas obtidas.

6.3 – Medidas de ajuste para modelos probabilísticos e comparação com o método ingênuo

Seja a estimação da probabilidade p de um evento E. Observe que não conhecemos o valor verdadeiro desta probabilidade, e portanto não dispomos de uma referência para avaliar a qualidade das estimativas obtidas, ou para comparar a performance de modelos/métodos concorrentes. Todavia, ao final da primeira fase, após observarmos se o evento E realmente ocorreu ou não, podemos utilizar o seguinte critério: se E ocorreu, o melhor modelo é o que tiver fornecido probabilidade mais próxima de 1 para E, e se E não ocorreu, o melhor modelo é aquele cuja probabilidade estimada para E tiver sido mais próxima de 0. Para ilustrar, considere o problema da estimação das probabilidades de classificação das equipes no Campeonato Brasileiro de 1999 (22 equipes, 8 vagas). Seja $\hat{P}_{CL}(i)$ a estimativa da probabilidade de classificação da equipe i. Possíveis medidas de proximidade são o erro quadrático médio (EQM) e o erro médio absoluto (EMA):

$$EQM = \frac{\sum_{i=1}^{8} (\hat{P}_{CL}(i) - 1)^2 + \sum_{i=9}^{22} (\hat{P}_{CL}(i) - 0)^2}{22}; \quad EMA = \frac{\sum_{i=1}^{8} \left|\hat{P}_{CL}(i) - 1\right| + \sum_{i=9}^{22} \left|\hat{P}_{CL}(i) - 0\right|}{22} \qquad (6.1)$$

Cabe observar que, neste caso, o EQM equivale à distância de DeFinetti (ver Arruda, 2000). Utilizamos as medidas em (6.1) para comparar, nas três competições, o desempenho da metodologia proposta com o método ingênuo, que consiste em atribuir probabilidade 1/3 aos três resultados possíveis de uma partida A x B: vitória de A ou B e empate. Em 87,1% das séries analisadas, o EQM e o EMA foram menores, quando utilizamos as distribuições estimadas via modelo Poisson-Gama/Princípio da Máxima Entropia, nas simulações.

7. Conclusões e extensões

Neste trabalho, apresentamos uma solução para o problema de obter probabilidades de classificação, e rebaixamento das equipes em um campeonato de futebol, combinando modelos estatísticos e métodos computacionais. A principal dificuldade deste problema é o elevado número de combinações possíveis para os resultados das partidas, e a solução adotada aqui foi o uso de técnicas de simulação e geração de cenários para estimar as probabilidades. As distribuições dos resultados das partidas, utilizadas nas simulações, foram obtidas a partir do Princípio da Máxima Entropia e das distribuições preditivas de modelos de séries temporais para dados de contagem, ajustados a séries de gols assinalados em partidas sucessivas. O uso destes modelos permite não somente considerar eventuais estruturas de dependência entre os resultados das partidas, mas também incorporar o efeito de fatores como escalação, mudanças de técnico e mando de campo, através de variáveis explicativas.

A metodologia foi aplicada no Campeonato Brasileiro 1999-2001. Com base em testes de validação, critérios para comparação de modelos e diagnósticos de ajuste, escolhemos o modelo Poisson-Gama para representar as séries de gols deste trabalho. Investigamos o efeito de jogadores considerados decisivos, como Romário, cuja influência sobre o desempenho do ataque do Vasco no campeonato de 2001 verificamos ser altamente significativa, e Guilherme, cujo efeito sobre o número de gols a favor do Atlético/MG, em 1999, somente foi significativo nas partidas em que Marques também estava em campo, evidenciando um efeito de interação. Além disto, estimamos o efeito de Antônio Clemente na histórica recuperação do Botafogo/RJ no campeonato de 1999 e verificamos que equipes de "porte médio", como Atlético/PR, Juventude/RS e São Caetano/SP, obtém melhores resultados nas partidas em que detém o mando de campo. Incidentalmente, utilizamos a técnica do *bootstrap* para obter intervalos de confiança para o hiperparâmetro do modelo Poisson-Gama, e para testar a significância dos coeficientes das variáveis explicativas. O teste de razão de verossimilhanças também foi aplicado, porém a distribuição da estatística do teste foi obtida via *bootstrap*.

Aplicações da metodologia proposta incluem não somente atender a demanda da imprensa esportiva, mas também estabelecer prognósticos acerca dos resultados de partidas de futebol, como instrumento de apoio à

decisão de apostadores nas variantes da extinta Loteria Esportiva, ou a criação de uma bolsa de apostas, como um mercado de opções, em que as cotas seriam definidas a partir das projeções da situação futura dos clubes.

Uma possível extensão deste trabalho é a comparação do modelo Poisson-Gama com outros modelos de séries temporais (ver, por exemplo, Zeger, 1988), para representar as séries de gols. Além disto, outros fatores podem ser incorporados ao método, como a tradição das equipes e a opinião de especialistas. Podemos ainda utilizar alternativas ao Princípio da Máxima Entropia, para combinar as preditivas dos modelos ajustados. Por exemplo, os coeficientes das combinações convexas podem ser especificados de modo a atribuir maior peso à série com menor variabilidade. A comparação de métodos alternativos pode ser efetuada através dos critérios definidos na seção 6.3.

Referências bibliográficas

ARRUDA, M. (2000). *"Poisson, Bayes, Futebol e DeFinetti";* Dissertação de Mestrado, IME-USP.

BAZARAA, M..SHERALI, H; SHETTY, C. (1993). *"Nonlinear Programming: Theory and Algorithms"*, 2nd ed.

BENNET, J. (1999). *"Statistics in Sports";* Arnold Publishers.

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

CAMPOS, E.L. (2001). *"Estimação de Probabilidades em Campeonatos de Futebol";* Tese de Doutorado, PUC-RIO.

FERNANDES, C. (1990). *"Non Gaussian Structural Time Series Models"*; Tese de Doutorado, London School of Economics.

GOLAN, A; JUDGE, G.; MILLER, D. (1996). *"Maximum Entropy Econometrics: Robust Estimation with Limited Data"*; John Wiley & Sons, c.3.

GUJARATI, D.N. (2000). *"Basic Econometrics"*; Makron Books, 3rd ed.

HARVEY, A.C. (1989). *"Forecasting, structural Time Series Models and the Kalman Filter"*; Cambridge University Press.

HARVEY, A.C.; FERNANDES, C. (1989). *"Time Series Models for Count Qualitative Observations"*; Journal of Business & Economic Statistics, v.7, nº 4.

JAMES, B. (1996). *"Probabilidade: um Curso em Nível Intermediário";* 2ª edição, Rio de Janeiro: IMPA.

JAYNES, E.T. (1957). *"Information Theory and Statistical Mechanics I and II. Physics Review"*; v.106, p.620-30 and v.108, p.171-90.

KULLBACK, J. (1959). *"Information Theory and Statistics"*; John Wiley, New York.

MOOD, A.M.; GRAYBILL, F.A.; BOES, D.C. (1974). *"Introduction to the theory of statistics"*; 3rd ed.

ZEGER, S.L. (1988). *"A Regression Model for Time Series of Counts"*; Biometrika, v.75, p.621-629.

Abstract

In this paper, we introduce a methodology to estimate probabilities associated with soccer games. The exact calculation is difficult due to the large number of possible results and the lack of independence between them. Iniatilly, we model the time series of goals scored in successive games using a Poisson-Gama model proposed by Harvey and Fernandes (1989). We introduced a number of covariates such as the changes in the coaching group and in the team, among others. Using the Maximum Entropy Principle, we combined the predictive distributions of the estimated models to generate the scores probability distributions of the remaning games. We then used simulation to generate random outcomes of the tournament.Based on their relative frequencies, we estimated probabilities of several events of interest. We applied the methodology to data from the first phase of the three most recent Brazilian Soccer Competition (from 1999 to 2001), obtaining sequential estimates (after each subsequent phase) of the classification probabilities for the next phase and the dropout probabilities from the next year competition.

Key words: Time series, Count data, Simulation, Entropy, Soccer.

Política editorial

A Revista Brasileira de Estatística - RBEs - objetiva promover a Estatística relevante para aplicação em questões sociais, interpretadas, amplamente, para incluir questões educacionais, de saúde, demográficas, econômicas, legais, de políticas públicas e de estatísticas oficiais, entre outras. A revista apresenta artigos num formato que permita fácil assimilação pelos membros da comunidade científica em geral. Os artigos devem incluir aplicações práticas como assunto central. Essas aplicações devem ter conteúdo estatístico substancial. As análises devem ser exaustivas e bem apresentadas, mas o emprego de métodos estatísticos inovadores não é essencial para publicação.

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

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

planejamento de pesquisas;

avaliação e mensuração de erros em pesquisas;

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

novos desenvolvimentos em metodologia de pesquisa;

crítica e imputação de dados;

amostragem e estimação;

disseminação e confiabilidade de dados;

análise de dados;

análise de séries temporais;

modelos e métodos demográficos; e

modelos e métodos econométricos.

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

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBEs

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

Renato Assunção
Editor Responsável
Revista Brasileira de Estatistica - 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