MODELO DE PREVISÃO COMBINADA: UMA APLICAÇÃO À SÉRIE MENSAL DAS NOTIFICAÇÕES DO "DENGUE" NO ESTADO DE PERNAMBUCO

Dirac Moutinho CORDEIRO[1]
Gauss Moutinho CORDEIRO[2]

▪ RESUMO: Desenvolve-se neste artigo um modelo de previsão combinada para explicar o comportamento das notificações mensais do dengue no Estado de Pernambuco. Neste modelo, é aplicada a proposta de Bunn (1985). Duas previsões pontuais, as melhores em termos de uma medida de eficiência – EMQ – erro médio quadrático, foram combinadas linearmente, para a obtenção de melhores previsões por meios estatísticos, análise dos resíduos e no percentual de redução da variância não-explicada da modelagem combinada em relação à previsão individualizada. Constata-se que em todas as abordagens metodológicas imputadas à série de interesse as previsões das notificações decresceram drasticamente quando comparadas com as observadas em 1998, ano de epidemia generalizada na região que, para nossa surpresa, foi o ano de maior racionamento de água nos últimos 30 anos. Com efeito, os resultados das previsões são estáveis e, nesse caso, pode-se extrapolar o seu comportamento para um processo estacionário com poucas oscilações, a série tende para uma endemia controlada.

▪ PALAVRAS-CHAVE: Abordagem bayesiana; Box e Jenkins; Bunn; dengue; distribuição beta; erro de Monte Carlo; modelo de previsão combinada; modelo linear dinâmico; Winters.

1 Introdução

O dengue é transmitido pela fêmea do mosquito Aedes aegypti, que também é o vetor da febre amarela. Qualquer epidemia das duas doenças está diretamente ligada à concentração do mosquito transmissor, ou seja, quanto mais desses insetos, mais doenças far-se-ão presentes.

O Aedes aegypti originou-se provavelmente na Região Nordeste do continente africano e de lá se espalhou para a Ásia e Américas, decorrente do tráfego marítimo

[1]Departamento Matemática, Escola Politécnica, Universidade de Pernambuco - UPE, CEP 50171-900, Recife, PE, Brasil. E-mail: *dirac@emtu.pe.gov.br*

[2]Departamento de Física e Matemática, Universidade Federal Rural de Pernambuco - UFRPE, CEP 50171-900, Recife, PE, Brasil. E-mail: *gausscordeiro@uol.com.br*

entre esses continentes. No Brasil, constatou-se a presença do inseto por volta do século XVIII, provavelmente, com as embarcações que transportavam escravos, já que os ovos do mosquito podem resistir, sem estar em contato com a água, por um período em torno de um ano.

Em 1955, uma grande campanha realizada pela Organização Pan-Americana de Saúde chegou a erradicar o Aedes aegypti não somente do Brasil, mas na grande maioria dos países do continente americano. Entretanto, por falta de recursos, a referida campanha não foi completa e, sendo assim, o mosquito continuou presente nas ilhas do Caribe, mais fortemente, nas Guianas e na Venezuela, de onde voltou a espalhar-se para o Brasil, Bolívia e Colômbia.

O mosquito não foi erradicado do país e, portanto, voltaria a ocorrer no antigo Estado da Guanabara, hoje Rio de Janeiro, uma grande epidemia no final da década de 1920 e, posteriormente, explodindo em todo território nacional. Na era Vargas, a grande luta pela erradicação se tornou um problema nacional de saúde pública epidemiológica e, nos anos 50, por meio de muito esforço no combate ao surto, conseguimos ser certificados por sanitaristas estrangeiros como livres do Aedes aegypti. Ao final da década de 1970, o Brasil figurava com um surto de dengue em suas principais cidades, provavelmente vindo do Caribe em pneus contrabandeados.

No ano de 2002, os dois estados que lideravam o número de ocorrências até o fim de fevereiro eram Rio de Janeiro e Pernambuco; sendo o Rio de Janeiro o que apresentava o maior número de casos e de óbitos. Já no Estado de Pernambuco, o número de doentes em 2001, representa 32% do total registrados em 1998, quando a maior epidemia de dengue já ocorrida em Pernambuco atingiu 53.500 notificações, representando cerca de 1,5% da população dos 14 municípios que definem a Região Metropolitana do Recife. Vale ressaltar, que essa região possui uma alta densidade populacional – a segunda do Brasil – e, além disso, 75% dos domicílios não têm rede de esgoto e um grande percentual ainda não dispõe de água encanada, fazendo uso de reservatórios e pequenos recipientes, no seu dia-a-dia. A Tabela 1, apresenta os percentuais de incremento do número de notificações do dengue em todo Estado de Pernambuco. Está claro, que a grande epidemia de forma generalizada, principalmente na Região Metropolitana do Recife, ocorreu no ano de 1998, que para nossa surpresa, foi o ano de maior racionamento de água nos últimos 30 anos.

A única garantia para que não exista o dengue é a total ausência do vetor transmissor. A Organização Mundial de Saúde (OMS), preconiza que há maior probabilidade de ser deflagrada uma epidemia quando os índices de infestação predial – número de imóveis amostrados com focos positivos de Aedes aegypti sobre o total de imóveis inspecionados – estiverem acima de 0,05, ou seja, 5%. No entanto, não existe um limite inferior no qual se possa afirmar, com absoluta certeza, que não ocorrerão surtos de dengue. É verossímil que em áreas com o Aedes o monitoramento deve ser constante, para que se possa conhecer as áreas infestadas e, com isso, poder deflagrar as medidas de combate ao mosquito. Para mais detalhes, ver Cunha (1997).

Tabela 1 - Número de notificações do dengue: 1995-2001

Ano	N. de casos notificados	%
1995	12.552	–
1996	14.825	18,10
1997	33.527	167,10
1998	53.575	326,82
1999	36.671	192,15
2000	28.729	128,87
2001	16.575	32,05

Fonte: Ministério da Saúde

2 Modelagens individualizadas aplicadas à série mensal das notificações do dengue

A série mensal das notificações do dengue revela, entre outras coisas:

a) assimetria da distribuição da série, ou seja, a amostra é mais populosa em valores baixos do que em valores altos – somente 11% das notificações são acima de 4.000, dos quais 30% são acima de 10.000;

b) abaixo de certo valor mínimo, praticamente não se registra ocorrência de notificações desse agravo. Especificamente, uma notificação mínima seria sempre mantida – "não-erradicação" –, portanto, valores abaixo deste mínimo teriam probabilidade nula de ocorrência e, como a própria OMS preconiza,"não existe limite inferior no qual se possa ter com absoluta certeza de que não ocorrerão surtos de dengue";

c) a construção do "Ploter" da série temporal revela características importantes das componentes não-observáveis, bem como da variabilidade das observações para uso de transformações adequadas.

Além dessa inspeção gráfica, utilizamos testes de hipóteses estatísticos, com intuito de verificar se realmente existem as principais componentes, tendência e sazonalidade. Logo, as hipóteses a serem testadas são:

H_0: não-existe componente
H_1: existe componente.

O teste para avaliação da tendência antes da sua estimação é baseado na série temporal $A[X_t]$ livre da componente sazonal e obtida por meio da aplicação de uma operação linear que transforma $A[X_t] = T_t$–componente tendência para $t = 1, 2, \ldots, N$. A transformação imputada a X_t para a obtenção de T_t é a de um filtro linear representada por uma média móvel centrada e de tamanho igual ao comprimento sazonal $L = 12$. Esse teste é baseado no coeficiente de correlação Spearman (ρ).

$$\rho = 1 - \frac{6 \sum_{t=1}^{N-L} (R_t - t)^2}{(N-L)[(N-L)^2 - 1]}$$

sendo $N = 84$ observações, $L = 12$ o comprimento sazonal e $R_t = r(T_t)$, o posto de T_t dentre as $(N - L)$ observações. O valor desse coeficiente indica a interpretação da correlação; no caso do estudo da série X_t é desprezível, pois $\rho = 0,18$. Sendo assim, admite-se a hipótese nula para T_t.

Quanto ao teste para avaliação da sazonalidade antes da estimação calcula-se a estatística ES dada pela equação:

$$ES = \left(\frac{N-L}{L-1}\right)\left[\frac{\sum_{j=1}^{L} m_j(\overline{X}_j^* - \overline{X}^*)}{\sum_{j=1}^{L}\sum_{i=1}^{m_j}(X_{ij}^* - \overline{X}_j^*)}\right]$$

onde $m_j = 7$, que representa o número de estações sazonais para o mês $j = 1, 2, \ldots, 12$ e X_t^* a série livre da componente tendência. O valor dessa estatística é igual a 12; comparando-se com o valor da estatística $F(11; 72; 0,05) = 1,89$, rejeita-se a hipótese nula para os fatores sazonais.

Mesmo com o fraco efeito da componente tendência para a série das notificações mensais do dengue, visualiza-se claramente na Figura 1, que as amplitudes da série variam sensivelmente com a média – sendo isto um forte indicador para uma modelagem multiplicativa da componente sazonal. Bowerman (1987) sugere que o melhor critério de escolha entre os fatores multiplicativos ou aditivos é calcular os valores da EMA (Erro Médio Absoluto) ou EMQ (Erro Médio Quadrático); e, para ambas as modelagens, a opção recai nos menores erros. Com efeito, os menores erros são observados, quando se imputa um ajuste multiplicativo à componente sazonal.

A Figura 1 também elucida essa tese, que é discernida por meio da série amortecida, usando uma média móvel centrada em $C = 7$ e de tamanho $q = 12$. Já a série sazonalmente ajustada é obtida em função das estimativas dos fatores sazonais multiplicativos. Além disso, está bastante visível, que em $t = 40$ (abril/98), nota-se de forma acentuada, como a magnitude do efeito sazonal influencia no comportamento da série. Já na Figura 2, tem-se uma grande semelhança entre o comportamento da série livre do efeito sazonal com o da série livre da componente estacionária, indicando assim, que os fatores sazonais aproximam-se dos fatores estacionais.

Sabe-se, que a aplicabilidade dos modelos de amortecimento exponencial podem ser estendida para as séries não-sazonais. Constata-se, por meio das autocorrelações seriais r_k – para um número de "lags" $k \leq 21$, que a série X_t livre do efeito sazonal encontra-se em regime estacionário.

Com efeito, são exeqüíveis as implementações dos modelos de amortecimento exponencial de Brown e Bi-paramétrico de Holts. Ambos responderam com razoável adequabilidade em relação ao número de observações não-explicadas e aos valores não-significativos das autocorrelações dos resíduos – aceitando-se a hipótese de $\hat{r}_k(\varepsilon) = 0$ pela estatística de Box-Pierce. Após as aplicações dos modelos de amortecimento exponencial, destacamos para a série X_t o Amortecimento Exponencial Sazonal Multiplicativo de Winters com os seguintes parâmetros: $\alpha = 0,95$ (nível), $\beta = 0,01$ (tendência) e $\gamma = 0,05$ (sazonal).

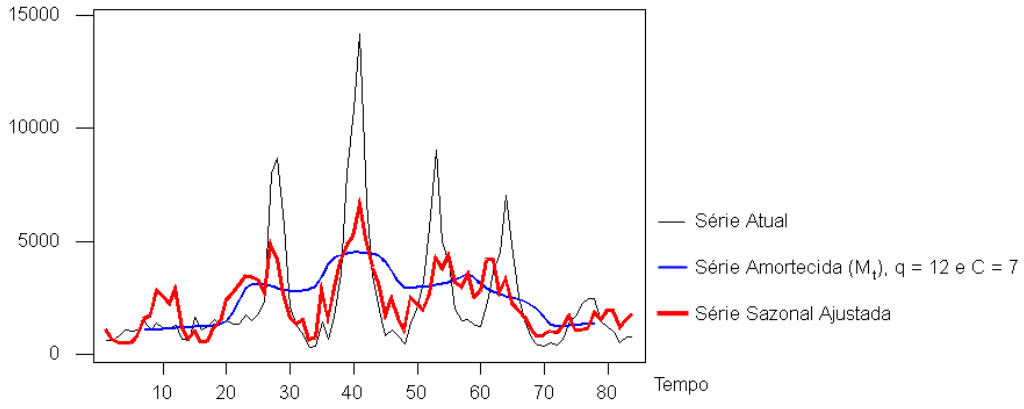

FIGURA 1 - Série mensal das notificações do dengue.

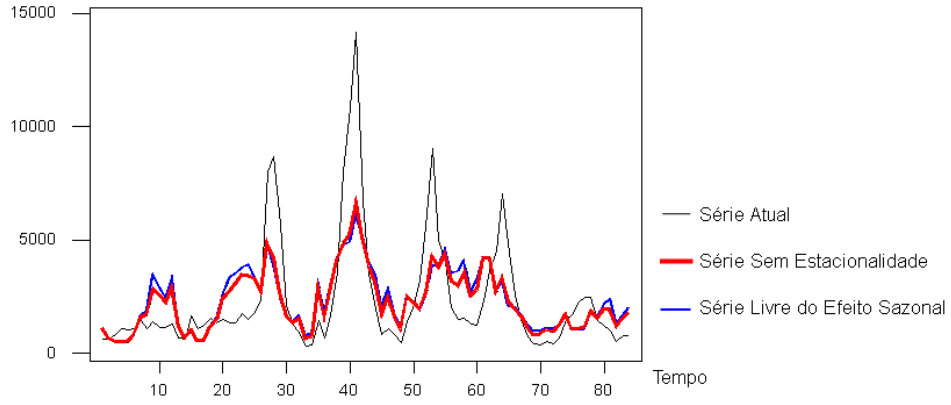

FIGURA 2 - Série mensal das notificações do dengue.

O modelo teve uma boa *performance* com base EMQ; na análise dos resíduos, não constatamos nenhum comportamento sistemático. Finalizando, com base na estatística de Box-Pierce, aceita-se estatisticamente a hipótese nula para as autocorrelações residuais, isto é, $\hat{r}_k(\varepsilon) = 0$ em todo o "lags", $k \leq 21$.

Conforme fora exibida nas figuras anteriores, apresentamos na Tabela 2 os quantitativos mensais das notificações desse agravo no período de janeiro de 1995 a dezembro de 2001. Se tentarmos explicar a ocorrência de valores mais altos das notificações, podemos concluir que estes são decorrentes de "períodos atípicos" representados pela grande crise de racionamento de água ocorrido nesse período, que provocou na população uma grande necessidade de estocar água em reservatórios, nem sempre preparados para esta finalidade. Esse argumento explica o processo heterocedástico, a variabilidade da série no ano de 1998, ano do grande racionamento, é bem maior que nos anos adjacentes.

Tabela 2 – Série mensal das notificações do dengue: janeiro/1995 a dezembro/2001

Mês	1995	1996	1997	1998	1999	2000	2001
1	604	633	1.803	1.617	1.376	2.302	671
2	578	601	2.365	3.590	1.967	4.637	1.473
3	808	1.638	8.006	8.001	3.219	4.495	1.685
4	1.091	1.080	8.687	10.861	5.601	7.017	2.225
5	1.032	1.196	5.691	14.166	9.061	7.841	2.475
6	1.082	1.549	2.117	6.699	4.998	2.576	2.440
7	1.436	1.340	1.198	3.504	3.989	1.498	1.403
8	1.052	1.488	936	1.948	1.983	694	1.206
9	1.388	1.332	300	833	1.452	383	951
10	1.107	1.341	322	1.085	1.546	371	515
11	1.118	1.743	1.436	812	1.270	510	762
12	1.252	1.478	666	459	1.209	405	769

Fonte – Ministério da Saúde.

Logo, a adoção de transformação faz-se necessária principalmente para corrigir a não-normalidade dos dados ao longo do tempo e, também, para corrigir problemas, devido à variabilidade dos dados; isto é, quando a série temporal tem um comportamento instável em torno da média e, nesse caso, melhorando significativamente as estimativas dos parâmetros do modelo. Sendo assim, a imputação de uma transformação não-linear à serie original X_t conseguiu, se não corrigir totalmente essa assimetria, pelo menos, minimizá-la tornando-a mais próxima de uma Gaussiana.

A variável transformada Y_t foi obtida pela aplicação da função potência de Box-Cox, cujo parâmetro da transformação estimado $\hat{\lambda} = -0,225$ foi obtido pela maximização da função suporte $\ell(\lambda)$, conforme é apresentado na Figura 3.

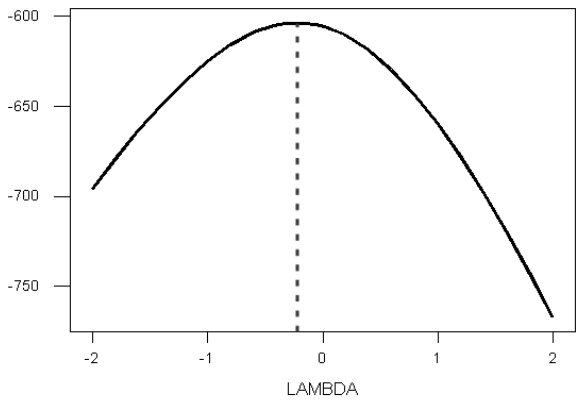

FIGURA 3 - Função suporte × lambda.

Após a transformação de X_t em Y_t, determinamos o correlograma da série Y_t. Esse correlograma é exibido pela Figura 4. Após análise do correlograma da série Y_t, nota-se claramente dois eventos:

a) existência de uma componente periódica de período ou comprimento sazonal $L = 12$;

b) as autocorrelações estimadas apresentam picos decrescentes – exponenciais amortecidas por senóides nos "lags" múltiplos de L.

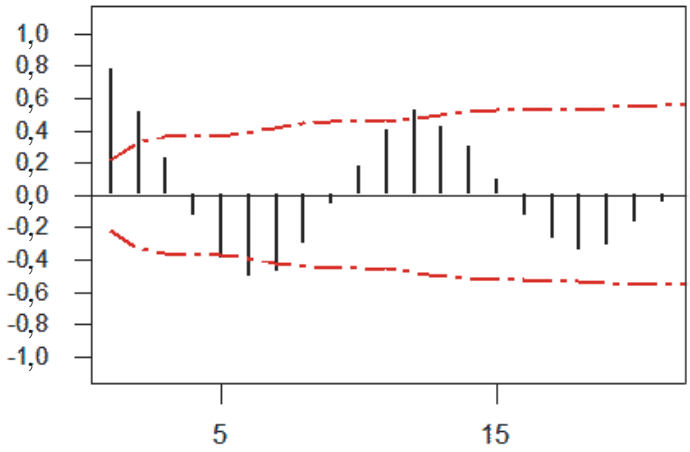

FIGURA 4 - Correlograma da série transformada.

Isso indica, entre outras coisas, que a série Y_t não deve ser submetida a um filtro linear instável $(1 - B^{12}) = \nabla_{12}$. A utilização desse filtro faz-se necessária para eliminar o efeito da não-estacionariedade sazonal; todavia, como os fatores estacionais são muito próximos dos fatores sazonais multiplicativos, preconiza-se a tese da não necessidade de usar o operador de diferenças sucessivas de lag12 e, dessa maneira, evita-se reduzir em 14% o tamanho da série.

Em efeito, está evidente o decréscimo exponencial das autocorrelações, o que é um forte indicativo de que o modelo verdadeiro só contém componentes ARs. Apresenta-se a seguir, a equação do modelo Box-Jenkins – de melhor *performance* – ARIMA $(1,0,0) \times (1,0,0)_{12}$:

$$(1 - 0,6028B^{12})(1 - 0,8238B)Y_t = 0,015805 \qquad (1)$$

Pela estatística t rejeita-se a hipótese nula para os três parâmetros e o modelo atende perfeitamente a condição de estacionariedade, pois o módulo do parâmetro do AR (1) é menor do que 1.

O objetivo deste artigo é propor uma modelagem estatística que substitua adequadamente os modelos paramétricos de Winters e Box-Jenkins. Assim sendo,

busca-se um modelo com melhor *performance*, por meio da análise dos resíduos e minimização EMQ. O modelo proposto é obtido por uma combinação linear entre os modelos SARIMA e de Amortecimento Exponencial Sazonal. A equação de combinação é representada por uma ponderação utilizando inferência Bayesiana para seleção seqüencial dos pesos.

3 Combinação dos modelos de previsão

Essa formulação foi proposta em 1980, por Granger, pela combinação linear de previsões pontuais calculadas por meio de modelos com características distintas. Essa nova metodologia é baseada no seguinte argumento científico:

"Se duas ou mais previsões pontuais, as melhores em termos de uma medida de eficiência (EMA-Erro Médio Absoluto ou EMQ-Erro Médio Quadrático), obtidas em modelos com metodologias distintas, então, se as previsões são combinadas linearmente o resultado da previsão final será sempre melhor que as previsões não-combinadas ou individualizadas".

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

$$\hat{X}_{t+1}^{(c)} = p\hat{X}_{t+1}^{(1)} + (1-p)\hat{X}_{t+1}^{(2)} \tag{2}$$

Sendo p o peso, em que no caso clássico, o seu cálculo é feito de maneira a minimizar a variância do erro de previsão combinada. O argumento citado anteriormente, estende-se facilmente para o caso de mais de dois modelos. Logo, em virtude da equação (2) estar vinculada à minimização da variância residual, é interessante mencionar a abordagem Bayesiana para o referido problema. Em Granger e Newbold (1974) e Granger e Ramanathan (1984), encontra-se uma descrição detalhada do enfoque Bayesiano para a combinação de previsões calculadas pelos modelos especializados. É interessante ressaltar, que a combinação deve ser sempre com os modelos individualizados que apresentam melhores *performances* quanto a uma medida de qualidade, que pode ser o EMA ou o EMQ. Uma outra análise de natureza Bayesiana para a seleção seqüencial do peso p referente à equação (2) é proposta por Bunn (1985), na qual o peso pertencente ao intervalo (0,1) é considerado como uma variável aleatória com distribuição beta, seqüencialmente atualizada no tempo.

Granger e Newbold (1977) apresentam comparações entre previsões pontuais e combinadas com base em 3 modelos clássicos de previsão, a saber: SARIMA na sua forma mais parcimoniosa ($M1$), Auto-regressivo ($M2$) e o Amortecimento Exponencial de Winters ($M3$). As Tabelas 3 e 4 são alusivas às aplicações dos modelos em 106 séries representativas da economia americana. Com efeito tem-se a *performance* como medida quantitativa de superioridade,como, por exemplo, o EMQ. Observe, que a comparação hegemônica dos modelos refere-se sempre ao primeiro; tanto na previsão individualizada com na previsão combinada.

Tabela 3 – Desempenho individual dos modelos

Modelos	*Performance* (%)
$M1 \times M2$	+73
$M1 \times M3$	+68
$M2 \times M3$	+48

Tabela 4 – Desempenho da previsão combinada

Modelos	*Performance* (%)
$(M1 \times M3)^c \times M1$	+56
$(M1 \times M2)^c \times M1$	+55
$(M2 \times M3)^c \times M2$	+69
$(M2 \times M3)^c \times M3$	+86
$(M2 \times M3)^c \times M1$	+46
$(M1 \times M2)^c \times M3$	+63

Em média os modelos $M1$ e $M2$ tiveram uma *performance* 58% superior ao $M3$; este valor cresce para 74% na modelagem combinada.

4 Modelos bayesiano de previsão

Para estabelecer um algoritmo mais flexível em relação às equações recursivas que definem os Modelos de Alisamento Exponencial-MAE, faz-se necessária a não-manutenção da equação de invariância para os estimadores de mínimos quadrados-MQ, ou seja,

$$\underset{\sim t}{b} = \underset{\sim t-1}{b} \tag{3}$$

para qualquer que seja t pertencente ao conjunto $(1, 2, \ldots, N)$. Nesse caso, para a não-manutenção da equação (3), os parâmetros variam com o tempo, ou seja, têm-se dois modelos com a finalidade de obterem $\underset{\sim t}{b}$, bem como, a matriz de variância-covariância C_t. O primeiro, a forma mais simples para a determinação de $\underset{\sim t}{b}$, é o passeio aleatório puro, onde

$$\underset{\sim t}{b} = \underset{\sim t-1}{b} + \underset{\sim t}{w}, \tag{4}$$

sendo $\underset{\sim t}{w}$ um vetor de variáveis aleatórias independentes de dimensão $(k \times 1)$. A sua forma mais complexa se baseia no modelo auto-regressivo Gauss-Markov, onde

$$\underset{\sim t}{b} = A\underset{\sim t-1}{b} + \underset{\sim t}{w}, \tag{5}$$

sendo A uma matriz diagonal de $(k \times k)$ cujos elementos a_{ij} variam entre $(-1, 1)$.

Esses modelos foram disseminados na vasta literatura de séries temporais por Harrison e Stevens em meados de 1976, portanto seis anos após a implementação dos modelos de Box e Jenkins. Os modelos de previsão amparados na estatística

Bayesiana impulsionaram os estudos das séries temporais em virtude de oferecerem uma nova metodologia científica para formulação de verdadeiros sistemas de previsão. A grande virtude desses modelos é consentir que sejam negligenciadas as restrições de estacionariedade ou estacionalidade a série histórica; bem como um tamanho razoável demandado pela quase totalidade dos modelos amparados na estatística clássica. Com efeito, a sua modelagem formulada pela equação (5) pode ser imputada a qualquer tipo de série temporal. Lembramos que a nossa abordagem dar-se-á somente a uma classe restrita, ou seja, os modelos seguem as propriedades gaussianas na qual existe uma relação de linearidade entre as observações X_t e os parâmetros.

Apresentamos a seguir, uma síntese metodológica do Modelo Linear onde os parâmetros e as observações são regidos por uma lei de formação dinâmica linear.

4.1 Modelo linear de Harrison e Stevens

Harrison e Stevens (1976) conceberam para as equações regentes do modelo, uma formulação parametrizada, na qual, as observações da série temporal X_t se relacionam linearmente com os parâmetros b_t que, por sua, vez evoluem por meio de dinâmica de estado, atualizado segundo um passeio aleatório. O princípio do modelo é obter a melhor estimativa atual dos parâmetros (b_t) com base no estado prévio (b_{t-1}). A essa característica dá-se o nome de Modelo Linear Dinâmico. Para mais detalhes ver Souza e Baratojo (1988), Souza (1982) e Souza (1984).

Considere-se o modelo linear descrito a seguir:

$$X_t = b_1 Z_{1t} + b_2 Z_{2t} + b_3 Z_{3t} + \cdots + b_k Z_{kt} + \varepsilon_t$$

Colocando em uma forma matricial, temos

$$\underset{\sim t}{X} = Z_t \underset{\sim t}{b} + \underset{\sim t}{\varepsilon} \tag{6}$$

onde

$\underset{\sim t}{X}$: é o processo estocástico representado por um vetor de N observações no instante atual $t = 1, 2, \ldots$;

$\underset{\sim t}{b}$: vetor de estado dimensão $k \times 1$ representando os parâmetros para todo instante $t = 1, 2, \ldots$;

Z_t : matriz de constantes conhecidas de dimensão $N \times k$ para todo instante $t = 1, 2, \ldots$;

$\underset{\sim t}{\varepsilon}$: vetor aleatório dimensão $N \times 1$ com propriedades Gaussianas de média $\underset{\sim}{0}$ e matriz de co-variância R_t, isto é, $\underset{\sim t}{\varepsilon} \sim N(\underset{\sim}{0}, R_t)$;

Assumindo que os parâmetros variam de uma mesma maneira, que podem ser descritos pela seguinte equação de transição entre o estado atual e o estado prévio, temos, então:

$$\underset{\sim t}{b} = T \underset{\sim t-1}{b} + F \underset{\sim t}{w}, \tag{7}$$

onde

T: matriz de transição de dimensão $k \times k$,

F: matriz de entrada de dimensão $k \times k$,

$\underset{\sim t}{w}$: vetor $k \times 1$ de variáveis aleatórias independentes – ruído branco – média nula e matriz de co-variância Q_t, ou seja, $E(\underset{\sim t}{w}\ \underset{\sim t}{w'}) = I_k w_t^2$; e $\underset{\sim t}{w} \sim N(\underset{\sim}{0}, Q_t)$.

Vale lembrar, que $\underset{\sim t}{\varepsilon}$ e $\underset{\sim t}{w}$ são ruídos gaussianos supostamente independentes. É interessante observar que a melhor estimativa do vetor $\underset{\sim t}{b}$ junto com a matriz de variância-co-variância do erro proveniente da estimativa desse vetor, deve ser obtida recursivamente pela sua melhor estimativa prévia, bem como da sua matriz de co-variância. Logo, a partir das estimativas iniciais subjetivas tanto para os parâmetros como para matriz de co-variância, isto é $\underset{\sim t-1}{b}$ e C_{t-1}, temos condições de predizer – estimar antecipadamente – os novos parâmetros, ou seja, $\underset{\sim t/t-1}{b}$ e $C_{t/t-1}$. Isto é possível devido à lei de regência de $\underset{\sim t}{b}$.

Logo, concluímos que, ao se conhecer a equação de formação dos parâmetros, podemos calcular as estimativas da mesma maneira para todos os parâmetros, mesmo antes de demandar os dados no instante de tempo t. Sendo assim, levando em consideração as propriedades dos vetores $\underset{\sim t}{w}$ e $\underset{\sim t}{\varepsilon}$, tem-se a transição entre a predição e estimação de $\underset{\sim t}{b}$ e C_t. Predição:

$$\begin{aligned} \underset{\sim t/t-1}{\hat{b}} &= T\underset{\sim t-1}{\hat{b}} \\ C_{t/t-1} &= TC_{t-1}T' + FQ_tF' \end{aligned} \tag{8}$$

Correção:

$$\begin{aligned} \underset{\sim t}{\hat{b}} &= \underset{\sim t/t-1}{\hat{b}} + C_{t/t-1}Z_t'(Z_tC_{t/t-1}Z_t' + R_t)^{-1}(\underset{\sim t}{X} - Z_t\underset{\sim t/t-1}{\hat{b}}) \\ C_t &= C_{t/t-1} - C_{t/t-1}Z_t'(Z_tC_{t/t-1}Z_t' + R_t)^{-1}Z_tC_{t/t-1} \end{aligned} \tag{9}$$

Note que T, F, R_t e Q_t são matrizes de entrada e, conseqüentemente, conhecidas *a priori*, além do que R_t está implicitamente relacionada com Q_t e seus valores são subjetivamente estimados pelos analistas que de posse desses valores, a distribuição de $\underset{\sim t}{b}$ e C_t é atualizada recursivamente pelas equações (8) e (9), definidoras do algoritmo desenvolvido por (Kalman e Buck, 1961). Observe-se, que a matriz de variância não depende dos valores observados $\underset{\sim t}{X}$, mas só depende da matriz de co-variância R_t, que, em muitos casos, pode ser pré-computadorizada.

Ao se avaliarem as equações de predição, observamos que há uma restrição vinculada diretamente às respectivas equações. Essa condicionante provém da equação de $\underset{\sim t}{b}$, que pode ser de difícil modelagem, o que a torna complicada para o estabelecimento da sua lei de formação. A nosso ver o arcabouço matemático da lei de formação foi sem dúvida o grande impasse na disseminação do modelo de Harrison e Stevens. Otter (1984) e Cordeiro (2002) apresentam de forma bastante detalhada as equações recursivas do algoritmo de estimação e correção de Kalman e Buck, bem como os seus respectivos diagramas de contexto.

5 Desenvolvimento de um modelo de previsão combinada – série mensal do dengue

A proposta de previsão pontual por meio de combinação linear usando métodos de previsão com características distintas é devida a Granger (1980). Reportando-se a Seção 3, "se duas ou mais previsões pontuais, baseadas em modelagens distintas são combinadas linearmente, a previsão pontual combinada obtida será sempre melhor que as previsões individuais". A equação de combinação é representada por uma ponderação utilizando uma abordagem Bayesiana para a seleção seqüencial dos pesos $p_i,\ i = 1, 2, \ldots, t$. A metodologia implementada é aquela proposta por Bunn (1985), na qual o parâmetro peso p_i é uma variável aleatória seqüencialmente atualizada no tempo segundo uma distribuição beta. O interesse maior na adoção dessa metodologia está focalizada na minimização da variância residual vinculada à redução do número de observações não-explicadas.

5.1 Descrição do processo de modelagem

O principio do método de previsão por meio de combinação linear estabelece que uma combinação linear de modelos lineares também representa um modelo linear. Embora simples e óbvio o método é importante para a construção de modelos mais adequados ao comportamento dos dados. O cerne do método é que um grande modelo linear poderá ser considerado uma combinação linear de modelos lineares mais simples.

Suponhamos que dois processos distintos j geram os dados; como exemplo o efeito sazonal superimposto ao crescimento ou decrescimento em longo prazo da série mensal do dengue. Logo, podemos construir as equações de estado para ambos modelos, ou seja,

$$\begin{aligned} \underset{\sim t}{\hat{X}}^{(j)} &= Z_t^{(j)} \underset{\sim t}{\hat{b}}^{(j)} \\ \underset{\sim t}{b}^{(j)} &= G_t \underset{\sim t-1}{b}^{(j)} + \underset{\sim t}{w}^{(j)} \end{aligned} \tag{10}$$

onde agora:

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

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

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

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

$\underset{\sim t}{w}^{(j)}$: vetor aleatório normal $(k \times 1)$ de média nula e matriz de variâncias e covariâncias calculada por $E\{\underset{\sim t}{w}\ \underset{\sim t}{w'}\} = I_k \sigma_w^2$ e conhecida no instante t.

Com base em dois processos $(j = 1, 2)$, observe-se, que o ruído branco das observações não é especificado, pois $\hat{X}_t^{(1)}$ e $\hat{X}_t^{(2)}$ são valores idealizados e são

observáveis somente em combinação por $X_t^{(c)}$ e, onde $\underset{\sim t}{\varepsilon}$ representa o vetor ruído com distribuição $N(\underset{\sim}{0}, I_N\sigma_\varepsilon^2)$. Tem-se,

$$\underset{\sim t}{X^{(c)}} = \underset{\sim t}{\hat{X}^{(1)}} + \underset{\sim t}{\hat{X}^{(2)}} + \underset{\sim t}{\varepsilon}. \tag{11}$$

O sistema acima pode ser escrito por meio de uma combinação linear conforme apresentamos nas equações (12); vale salientar, que a sua extensão para três ou mais modelos é imediata. Vejamos:

$$\begin{aligned} \underset{\sim t}{X^{(c)}} &= \left(Z_t^{(1)}, Z_t^{(2)}\right) \begin{pmatrix} \underset{\sim t}{\hat{b}^{(1)}} \\ \\ \underset{\sim t}{\hat{b}^{(2)}} \end{pmatrix} + \underset{\sim t}{\varepsilon} \\ \begin{pmatrix} \underset{\sim t}{b^{(1)}} \\ \\ \underset{\sim t}{b^{(2)}} \end{pmatrix} &= \begin{pmatrix} G_{11} & 0 \\ \\ 0 & G_{22} \end{pmatrix} \begin{pmatrix} \underset{\sim t-1}{b^{(1)}} \\ \\ \underset{\sim t-1}{b^{(2)}} \end{pmatrix} + \begin{pmatrix} \underset{\sim t}{w^{(1)}} \\ \\ \underset{\sim t}{w^{(2)}} \end{pmatrix}. \end{aligned} \tag{12}$$

Assim, visando a melhorar as previsões pontuais, imputamos à serie temporal de interesse uma subjetividade para minorar o efeito explosivo dos dados, decorrente, basicamente dos fatores sazonais. Neste sentido, torna-se imprescindível minimizar tal efeito e há, nesse caso, necessidade de reduzir a capacidade de influência das observações nos períodos em que o efeito sazonal sobrepõe-se às observações. Apresenta-se na Seção 5.2 a estrutura canônica do modelo representada pelos pesos p_i, bem como pelas estimativas dos modelos de melhor desempenho, até então analisados com base em estatísticas e medidas de qualidade.

Algumas considerações para elaboração do processo de combinação foram introduzidas objetivando o melhor entendimento da estrutura das equações do modelo e, ao mesmo tempo, facilitar os procedimentos computacionais inerentes à convergência das estimativas dos parâmetros p_i. Com efeito, apresentamos as premissas que nortearam a concepção metodológica Bayesiana para a modelagem da série do dengue.

a) ao número de casos notificados mensalmente ao longo dos 7 anos – r_i, imputou-se uma modelagem consistente com o nível de confiança que temos; ou seja, por meio da subjetividade admitimos que o número médio de notificações em cada mês do ano e ao longo dos 7 anos segue uma uniforme $U[(p_i)(N_i)]$;
b) os parâmetros p_i do modelo são estimados em a); essas estimativas são reavaliadas em cada instante de tempo i, $i = 1, 2, \ldots, 7$;
c) as variáveis aleatórias $p_i s$ são independentes e podem seguir uma distribuição não-informativa do tipo $B(1, 1)$;
d) a distribuição *a posteriori* de p_i é calculada a partir da distribuição *a priori* utilizando o Teorema de Bayes;
e) obtém-se o estimador de Bayes – máxima probabilidade e menor EMQ – a partir da distribuição *a posteriori*, que pode ser qualquer medida de posição dessa distribuição; como por exemplo, moda, média ou mediana. Sendo as três

medidas coincidentes, a função densidade de probabilidade da distribuição *a posteriori* é simétrica;

f) obtida a distribuição *a posteriori* do vetor de parâmetros p_i em cada instante de tempo i, utiliza-se esta informação para inferir as distribuições das futuras previsões.

5.1.1 Discussão e análise dos resultados

A seguir, exibimos na Tabela 5 um resumo das estatísticas referentes à distribuição *a posteriori* dos parâmetros p_i –, em que se define a estimativa média de p ao longo dos 7 anos ou 84 meses. No caso $\hat{p}_1$, temos o valor da estimativa de p para o ano 1995; $\hat{p}_2$ estimativa de p para o ano 1996, e, assim, sucessivamente até o último ano, ou seja, 2001. Para o cálculo de $\hat{p}_i$ faz-se necessário usar um algoritmo de convergência com a finalidade de minimizar o erro-padrão e, conseqüentemente, o erro de Monte Carlo (MC erro); uma forma de se avaliar adequadamente o parâmetro de interesse da distribuição *a posteriori* é calcular o quociente entre o erro de Monte Carlo e o seu respectivo desvio padrão. Caso essa relação seja menor que 5%, então, pode-se afirmar que a estimativa $\hat{p}_i$ representa a verdadeira média da distribuição *a posteriori* –, no caso, tem-se que este valor ficou muito aquém, algo em torno de 2,5%. Ainda em relação à Tabela 5, nota-se a plena caracterização do estimador de Bayes, por meio da igualdade entre a média e a mediana; bem como, o menor desvio padrão.

Tabela 5 – Sumário das estatísticas referentes às distribuições *a posteriori* – beta

par.	média	sd	MC erro	2.5%	mediana	97.5%	iníc.	fim
$\hat{p}_1$	0,1863	0,003435	1,184E-4	0,1798	0,1864	0,1929	1	1000
$\hat{p}_2$	0,1576	0,003089	8,841E-5	0,1517	0,1576	0,1639	1	1000
$\hat{p}_3$	0,06974	0,001376	3,714E-5	0,06691	0,06972	0,07259	1	1000
$\hat{p}_4$	0,04368	8,986E-4	2,477E-5	0,04197	0,04367	0,04546	1	1000
$\hat{p}_5$	0,06376	0,001315	4,479E-5	0,06135	0,06371	0,06643	1	1000
$\hat{p}_6$	0,08145	0,001607	5,412E-5	0,0783	0,08141	0,08452	1	1000
$\hat{p}_7$	0,141	0,0027	6,997E-5	0,1358	0,1409	0,1464	1	1000

A Figura 5 exibe os processos de convergência de p_i em função das iterações, evidenciando, aparentemente, que durante as 1.000 iterações não existiu descontinuidade no processo de convergência. Os gráficos das distribuições dos parâmetros – distribuição *a posteriori*, são exibidos na Figura 6 e está claro, representam distribuições betas com os seguintes *designs*: $p_1 \sim B(2393, 10452)$, $p_2 \sim B(2193, 11720)$, $p_3 \sim B(2389, 31874)$, $p_4 \sim B(2260, 49471)$, $p_5 \sim B(2201, 32319)$, $p_6 \sim B(2360, 26610)$ e $p_7 \sim B(2342, 14271)$.

Observe-se na Figura 7 os valores das correlações obtidos em todas as iterações (1.000); esses valores da matriz de correlação entre os diversos p_i estão configurados no formato de diagrama. Portanto, os valores dos p_i constituem-se

de varáveis aleatórias independentes. Vale frisar que os correlogramas, gráficos das autocorrelações, de cada um p_i em todas as 1.000 iterações são estatisticamente iguais a zero (excluindo o lag 0).

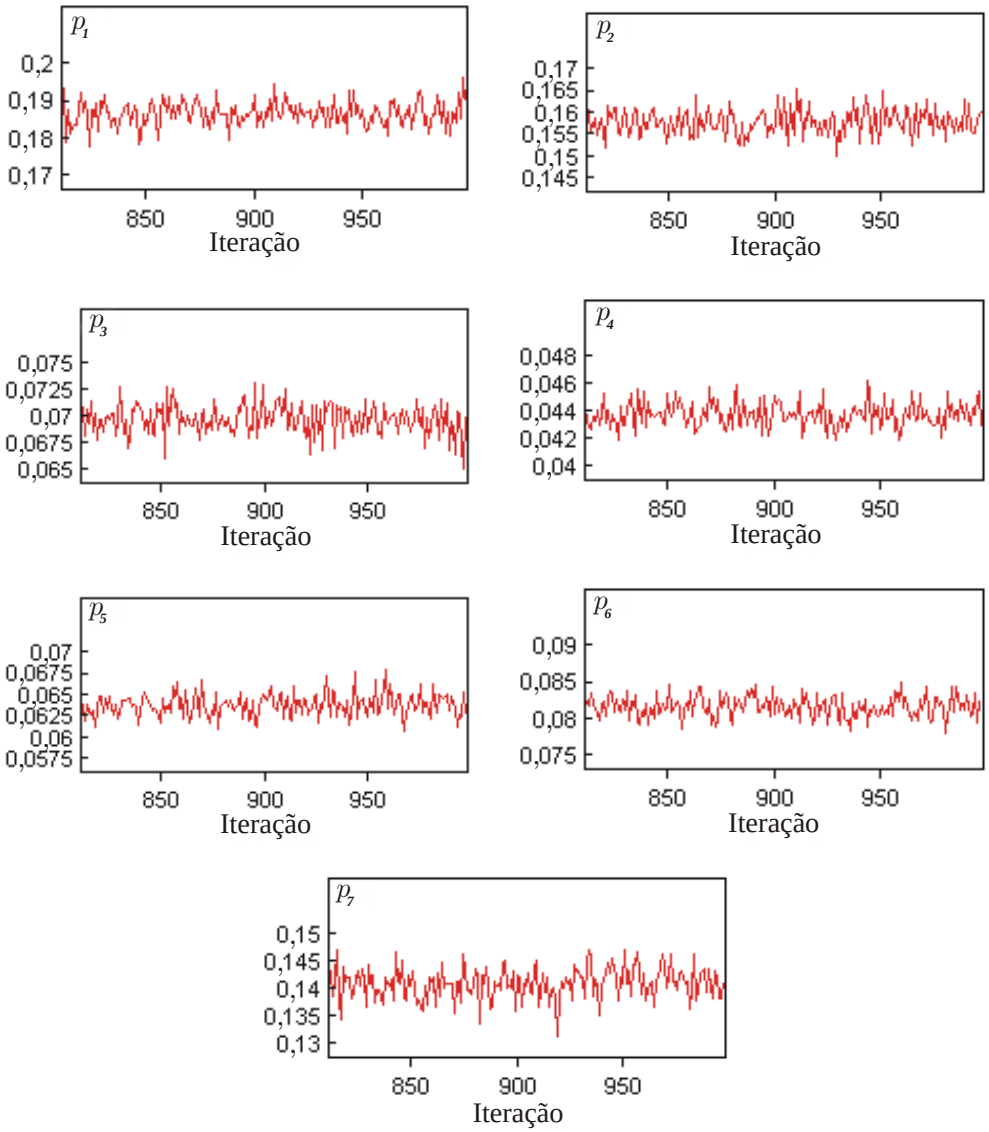

FIGURA 5 - Convergência de $p_i \times n^o$ de iterações.

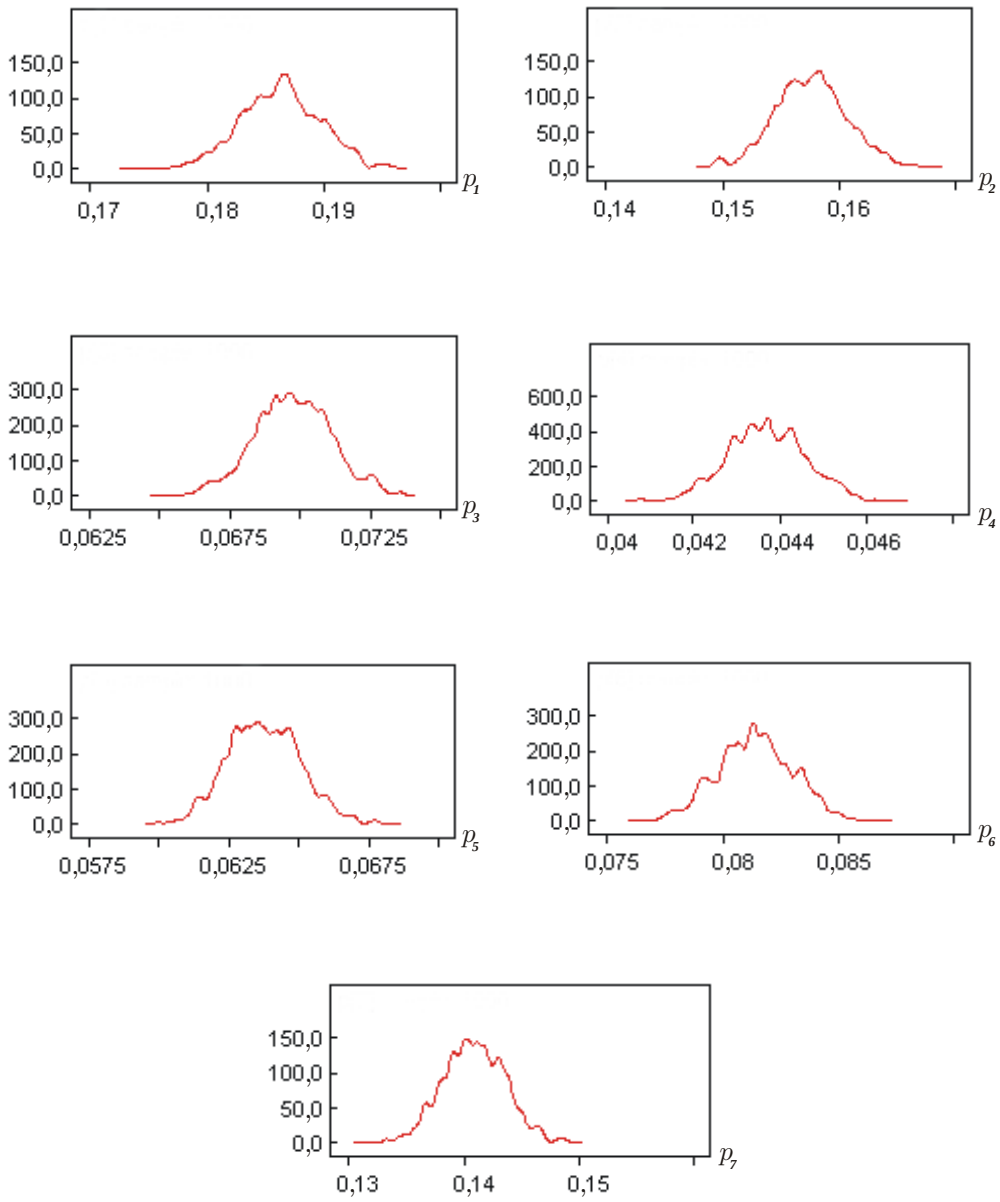

FIGURA 6 - Distribuição *a posteriori* de p_i.

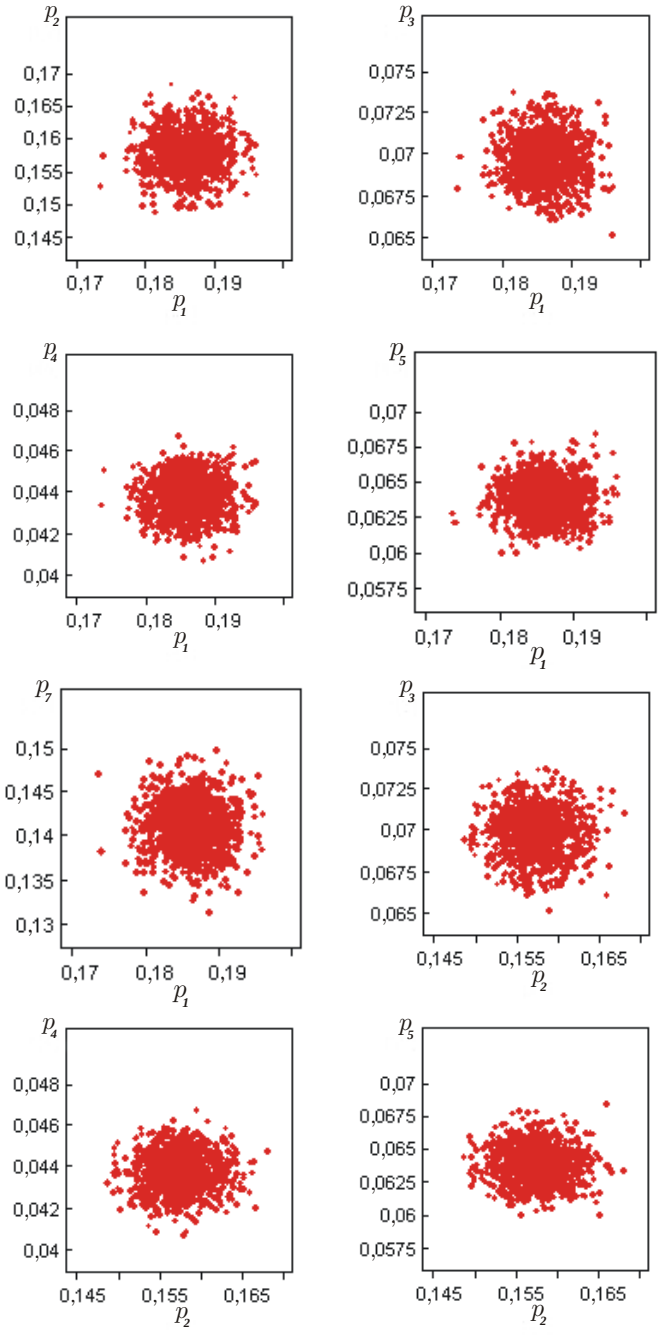

FIGURA 7 - Valores das correlações entre os p_i (continua).

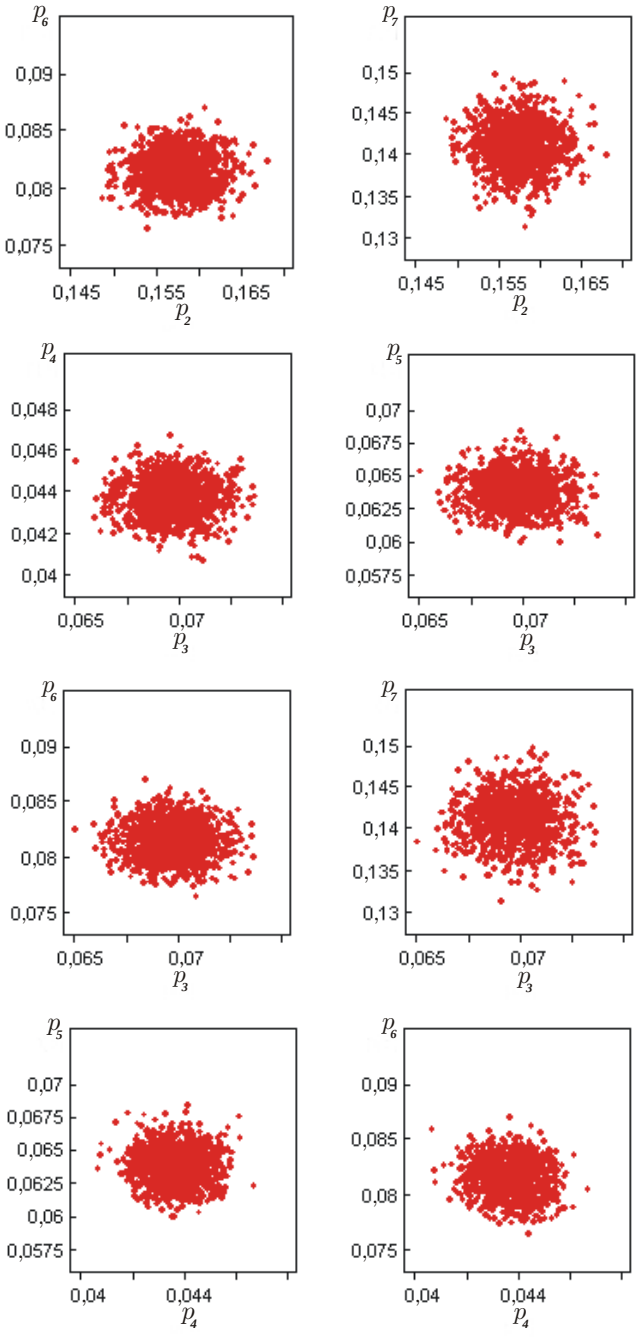

FIGURA 7 - Valores das correlações entre os p_i (continuação).

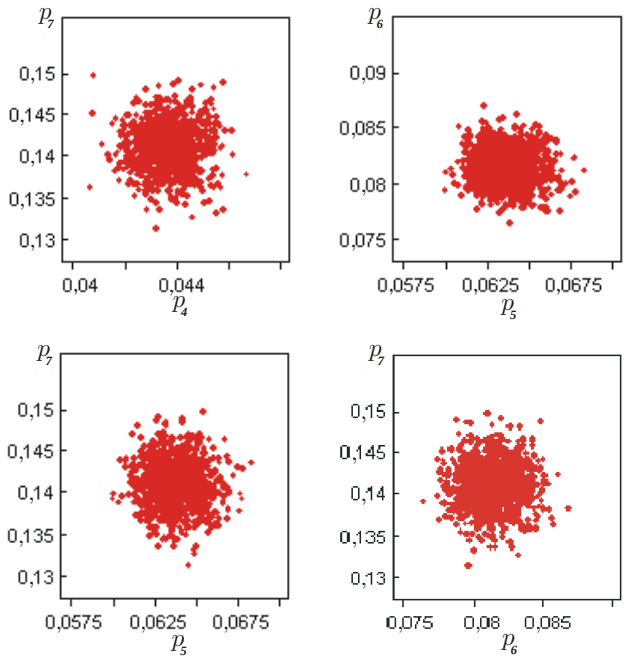

FIGURA 7 - Valores das Correlações entre os p_i.(continuação)

Finalizando, temos na Figura 8 a representação do ajuste da curva, para as diversas estimativas de p_i em função das notificações ocorridas durante o período mensal entre 1995 a 2001. Além do excelente ajuste, 95% da variabilidade dos parâmetros mensais $p_i s$ ao longo dos sete anos, é explicada pelas notificações do dengue. Nota-se também, que os estimadores de p_i decrescem com o crescimento dos dados observados mensalmente, ano a ano; nesse caso, concluímos que os valores estimados, de certa forma, funcionaram como um filtro de amortecimento sazonal de modo a minimizar o efeito da componente sazonal decorrente da imputação de fatores externos não-controláveis, a exemplo do grande racionamento de água advindo da "grande seca" no Estado de Pernambuco no período de 1997 a 1998.

Com efeito e tendo em vista os intervalos de confiança determinados para cada um dos $p_i s$, fica totalmente excluída a possibilidade de nulidade para os parâmetros. Os intervalos obtidos para p_i na seqüência $i = 1, 2, \ldots,$ são:

$$
\begin{array}{llll}
(0{,}184153, & 0{,}188518), & (0{,}155637, & 0{,}159563), \\
(0{,}068886, & 0{,}070614), & (0{,}043110, & 0{,}044250), \\
(0{,}062920, & 0{,}064590), & (0{,}080428, & 0{,}082470), \\
 & (0{,}139280, & 0{,}142710). &
\end{array}
$$

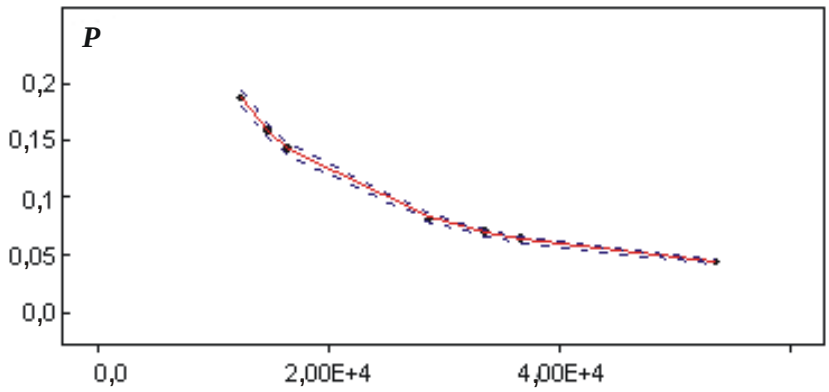

FIGURA 8 - Valores ajustados p_i × valores anuais - mensais da série dengue 1995-2001.

5.2 Combinação das estimativas – previsões

A equação de combinação linear proposta por Bunn (1985), na qual $p_i \in (0,1)$ é considerada sendo uma variável aleatória com distribuição beta – seqüencialmente atualizada no tempo – conforme apresentada na estrutura matricial abaixo

$$\hat{X}_t^{(c)} = (\underset{\sim}{\hat{p}}_t : \underset{\sim}{1} - \underset{\sim}{\hat{p}}_t) \begin{pmatrix} \underset{\sim}{\hat{X}}_t^{(1)} \\ \underset{\sim}{\hat{X}}_t^{(2)} \end{pmatrix}$$

onde:

$\underset{\sim}{\hat{p}}_t$: representa o vetor estimador dos parâmetros p_i $i = 1, 2, \ldots, t$ cuja dimensão é (1×84);

$\underset{\sim}{1}$: vetor unitário dimensão (1×84);

$\underset{\sim}{\hat{X}}_t^{(1)}$: vetor das estimativas da série dengue calculado pelo processo 1;

$\underset{\sim}{\hat{X}}_t^{(2)}$: vetor de estimativas da série dengue calculado pelo processo 2.

A seguir apresenta-se na Tabela 6, os resultados das previsões combinadas usando os modelos sazonais, isto é, SARIMA e Winters.

Tabela 6 – Previsões pontuais com modelagem combinada - 3 passos à frente

Mês	SARIMA	Winters	Combinação
1	1008	1116	1104
2	1549	1779	1754
3	1581	2573	2467

Na Tabela 7, além das análises dos resíduos dos três modelos, têm-se os resultados das *performances* com base no percentual de redução da variância não-explicada do modelo combinado em relação às variâncias individualizadas; logo, em função do redutor da variância adicionado ao número de *outliers* – observações não-explicadas –, obteve-se o percentual de desempenho da modelagem combinada versus os processos individualizados, modelo de amortecimento exponencial sazonal e SARIMA.

Tabela 7 – Análise comparativa modelagem combinada × individualizada

Modelo	Análise dos Resíduos Box-Pierce	Análise da Hipótese $\hat{r}_k(\varepsilon) = 0$	EMQ	N^o. de Obs. Não-Expl. $(+3\sigma)$	Desemp. (%)
$M1$	$Q = 53,56$	Aceita-se	$2,1 \times 10^6$	3	–
$M2$	$Q = 38,68$	Aceita-se	$6,7 \times 10^5$	6	–
$(M1 \times M2)^c \times M1$	$Q = 36,40$	Aceita-se	$6,2 \times 10^5$	3	+70
$(M1 \times M2)^c \times M2$	$Q = 36,40$	Aceita-se	$6,2 \times 10^5$	3	+54

$M1$: SARIMA
$M2$: Winters
$(M1 \times M2)^c$: Combinação

Já as Figuras 9 e 10 exibem as estimativas e o correlograma dos resíduos da modelagem combinada, respectivamente. Nota-se, claramente, que é aceita a hipótese nula para as autocorrelações residuais; ou seja, estatisticamente os resíduos não são correlacionados e, sendo assim, $\hat{r}_k(\varepsilon) = 0$.

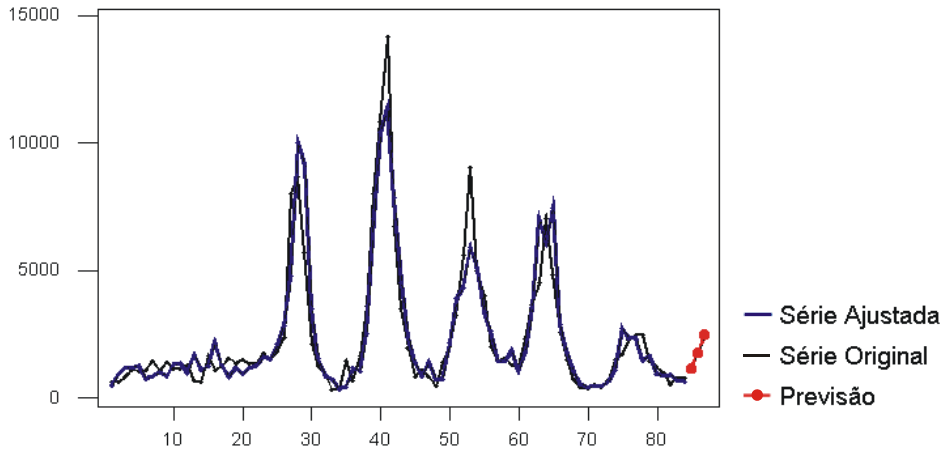

FIGURA 9 - Valores ajustados e previstos da série mensal do dengue usando modelagem combinada × valores observados da série mensal do dengue.

Observe-se também que mesmo usando uma metodologia combinada o correlograma é constituído por exponencial amortecida por meio de senóides e, dessa maneira, permanece ainda, como um processo autorregessivo. Esse fato condiz com a amenização do efeito da componente sazonal, que, de certa forma, foi amortecida pelo processo de combinação linear utilizado.

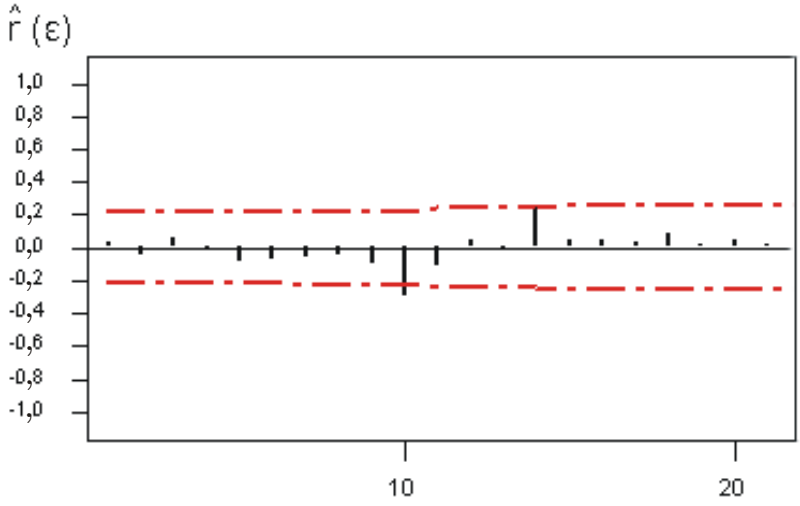

FIGURA 10 - Correlograma dos resíduos - modelagem combinada.

5.3 Comentários finais

A proposta de usar a combinação linear para previsões pontuais é factível e mais robusta. Logo, a afirmação de Granger (1980) é de fato uma exortação verossímil; ou seja, a previsão pontual combinada é sempre melhor que as previsões individuais. Tal conclusão é também reforçada pela abordagem clássica, pois, Pereira e Coqueiro (1983) constatam o argumento de Granger (1980) por meio da combinação linear de previsão usando Filtro Adaptativo para a série histórica de taxas de juros combinadas com opiniões subjetivas de especialistas de operadores do mercado financeiro.

Em síntese, constata-se que em todas as abordagens metodológicas imputadas à série de interesse – Modelos de Suavizamento Exponencial e Modelos de Box e Jenkins – revelam que as estimativas das notificações decresceram drasticamente, quando comparadas com 1997 e 1998. Para nossa surpresa nos quatro modelos utilizados, tantos para previsões pontuais como combinadas, os resultados das previsões são estáveis e, nesse caso, podemos extrapolar o comportamento para um processo estacionário com poucas oscilações, ou seja, a série tende a uma estabilidade em nível, semelhante aos processos epidemiológicos convergindo notadamente para uma endemia controlada. Contudo, há uma exceção que não invalida a nossa

conclusão, no modelo de Winters a previsão no terceiro mês se afastou um pouco da média de previsão da série temporal, ocasionada principalmente pelos fatores sazonais determinísticos como as componentes calculadas nas estações de janeiro/98, fevereiro/98 e março/98. Esse afastamento, provavelmente, foi decorrente do grande racionamento de água verificado no final do verão de 1997 e início de 1998. Logo, esse racionamento de água a ser implementado em qualquer ano, nenhuma dúvida, propiciará um desequilíbrio entre as componentes estacionais e sazonais da série do dengue e, neste caso, uma explosão das notificações nos meses posteriores é muito provável de acontecer. Como estamos em constante alerta para que ações de racionamento sejam implantadas, dessa forma é pouco verossímil afirmar que não estamos sujeitos, em curto espaço de tempo, a uma nova epidemia nos moldes de 1997 e 1998.

Não resta dúvida de que as instituições públicas necessitam informar e esclarecer a população, que a estocagem de água em reservatórios não condizentes representa um grande risco de epidemia desse agravo, uma vez que demandará muito tempo para que essa doença possa ser erradicada totalmente do nosso Estado.

Lembramos que nas séries de forte componente sazonal o efeito sazonal passa a ser de suma importância principalmente como vetor contaminador de longo prazo, que é peculiar às séries epidemiológicas, notadamente, com mesmas características comportamentais, o seu *design* apresenta harmônicos que se propagam ao longo do tempo.

Agradecimentos

Os autores agradecem aos revisores, intelectuais anônimos, pelos comentários e contribuições que foram proveitosos para a melhoria da primeira versão deste artigo em relação à exposição das idéias originalmente discutidas.

CORDEIRO, D. M.; CORDEIRO, G. M. A combined forecast model: an application to the monthly series os dengue reports in Pernambuco state - Brazil. *Rev. Mat. Estat.*, São Paulo, v.22, n.3, p.57-80, 2004.

▪ *ABSTRACT: A combined forecast model is developed to explain the behaviour of monthly dengue reports in Pernambuco state, Brazil. Bunn's proposal (1985) is used. Two punctual forecasts, the best in terms of efficient dimension - RMSE (Mean Square Erros), were linearly associated to achieve the best forecasts by statistical means, residual analysis and perceptual reduction of unknown combined variables. It is observed that all the methodological approaches to the series show a dramatic reduction in report forecasts as compared with those of 1998 - the year when a generalised epidemics occurred. To our surprise, 1998 had been the year of the largest water rationing in the previous thirty years. The results of forecasts are stable and, in this case, it is possible to extrapolate their behaviour to a stationary process with few oscillations. The series tends to a controlled epidemics.*

▪ *KEYWORDS: Bayesian approach; beta distribution; Box and Jenkins; Bunn; combined forecast model; dengue; linear dynamic model; Monte Carlo error; Winters.*

Referências

BUNN, D. W. Statistical efficiency in the linear combination of forecasting. *Int. J. Forecast*, Amsterdan, v.1, p.151-163, 1985.

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

CORDEIRO, D. *Séries temporais, análise quantitativa*: teoria e aplicações. Edupe, 2002. p.91-124 (Série Ciência e Tecnologia).

CUNHA, R. da, *Aspectos clínicos e epidemiológicos da infecção pelo vírus dengue em áreas endêmicas do Brasil*. Rio de Janeiro: Fiocruz, 1997.

GRANGER, C. W. J. *Forecasting in Business and Economics*. New York: Academic Press, 1980.

GRANGER, C. W. J.; NEWBOLD, P. Experience with statistical forecasting and with combining forecasting. *J. R. Stat. Soc.*, Cambridge, v.137 , p.131-165, 1974.

GRANGER, C. W. J.; NEWBOLD, P. *Forecasting economic time series*. New York: Academic Press, 1977.

GRANGER, C.W.J.; RAMANATHAN, R. Improved methods of combining forecasting. *J. Forecast.*, Chichester, v.3, p.197-204, 1984.

HARRISON, P. J.; STEVENS, C. F. Bayesian Forecasting. *J. R. Stat. Soc.*, Series B, Cambridge, v.38, n.3, p.205-267, 1976.

KALMAN, R. E.; BUCK, R. S. New results in linear filtering and prediction theory. *J. Basic Eng.*, New York, v.82D, n.1, p.95-108, 1961.

OTTER, P. W. The discrete Kalman filter applied to linear regression models: Statistic consideration and application. *Stat. Neerlandica*, v.31, n.1, p.41-56, 1984.

PEREIRA, B. B.; COQUEIRO, R. C. O.; PERROTA, A. H. V. Combinação de informações subjetivas e métodos quantitativos para previsões de taxas no open market. *Rev. SOBRAPO*, v.3, p.25-40, 1983.

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

SOUZA, R. C. Modelo bayesiano geral para previsão de demanda esparsa em problemas de controle de estoques. In: CLAIO 2., 1984, Buenos Aires, 1984. p.8-18.

SOUZA, R. C.; BARATOJO, S. C. *Combinação bayesiana de previsões*: aplicação ao IGP-DI, GSC-36/88. Rio de Janeiro: PUC, 1988.

Recebido em 21.05.2004.

Aprovado após revisão em 06.12.2004.