

UMA COMPARAÇÃO DOS MODELOS SARIMA E ESTRUTURAIIS PARA A PREVISÃO DA INCIDÊNCIA DE DENGUE EM BELO HORIZONTE, MINAS GERAIS

Thiago Rezende dos Santos¹

- RESUMO: Fazer previsões para a incidência de dengue é um dos objetivos de vigilância em saúde pública, trazendo aos gestores em saúde estimativas dos números mensais de novos casos de forma que os mesmos possam, por exemplo, planejar estratégias de atendimentos, deslocamentos de equipes de saúde, etc. Este artigo contém previsões para as observações futuras da série de incidência de dengue em Belo Horizonte baseadas nos modelos *Seasonal Autoregressive Integrated Moving Average* (SARIMA) e Modelos Estruturais (ME), comparando-os quanto aos seus ajustes e às suas capacidades preditivas. As previsões via ME têm um desempenho melhor que as via modelos SARIMA, segundo os critérios erro quadrático médio (EQM) e erro absoluto médio (EAM). Além disso, como os ME decompõem a série em termos de seus componentes não-observáveis, pode-se estudar o comportamento da tendência da incidência dos casos de dengue ao longo do tempo. Este componente têm um decrescimento entre os anos de 2002 a 2005, tornando não significativo em 2005 e tendo um crescimento daí por diante. Também pode-se identificar os meses de maior e menor incidência através dos componentes não-observáveis de sazonalidade. Os primeiros meses do ano apresentam maior incidência, o que já era esperado. Os resultados são muito satisfatórios e servem de base para novos estudos.
- PALAVRAS-CHAVE: Dengue; modelos espaço de estados; inferência clássica; previsão.

1 Introdução

A dengue é arbovirose causada por um Flavivírus, com quatro sorotipos conhecidos. Caracteriza-se como doença febril aguda, com espectro clínico variando desde quadros febris inespecíficos até manifestações graves com hemorragia e

¹Universidade Federal de Minas Gerais, UFMG, Departamento de Estatística, CEP: 31270-901, Belo Horizonte, MG, Brasil, E-mail: thiagords@ufmg.br

choque: a Febre Hemorrágica da Dengue (FHD) e a Síndrome do Choque da Dengue (SCD). Estima-se um total entre 50 a 100 milhões de casos de dengue no mundo (WORLD HEALTH ORGANIZATION , 2014).

No Brasil, entre 1990 e 2000, os sorotipos 1 e 2 se disseminaram, com registro de epidemias principalmente nos grandes centros urbanos do Sudeste e Nordeste. Em 2003, os sorotipos 1, 2 e 3 foram isolados em 23 Estados e observou-se um número crescente de internações por FHD. Em Belo Horizonte, a doença tem ocorrido de forma endêmica, com picos epidêmicos nos primeiros quadrimestres anuais e circulação dos sorotipos 1, 2 e 3. Essa situação, associada à limitada efetividade no controle do vetor e à ausência de uma vacina eficaz, aumentou o risco da ocorrência de FHD/SCD (CORRÊA et al., 2005).

O município de Belo Horizonte (BH) vem apresentando desde 1996 sucessivas epidemias de dengue ao longo do tempo, acompanhadas de uma progressiva infestação e dispersão do vetor *Aedes aegypti* nos vários agregados espaciais da cidade (CORRÊA et al., 2005). Como também pode ser percebido, esses dados têm uma estrutura de correlação temporal que pode ser utilizada para fazer previsão para prever a incidência de dengue é um dos objetivos de vigilância em saúde pública e é essencial para que se tenha agilidade para detectar as epidemias e casos graves e planejar ações (MARTINEZ e SILVA , 2011).

De acordo com Unkel et al. (2012), as metodologias de análise de séries temporais, por exemplo, os modelos ARIMA e de alisamento exponencial têm sido empregados na detecção prospectiva de clusters, bem como os métodos que puramente incorporam a informação espacial (KULLDORF, 1997; KULLDORF et al., 1998). Uma boa revisão sobre esses métodos pode ser encontrada em Unkel et al. (2012). Há vários estudos temporais sobre a dengue nas regiões mais afetadas pela mesma, entre os quais são citados a seguir.

Wu et al. (2007) avaliaram os impactos das variações climáticas sobre a ocorrência de Dengue em uma cidade metropolitana do Sul de Taiwan por meio de análise de séries temporais. Os modelos autoregressivos integrados e de médias móveis (ARIMA) mostraram que a incidência de dengue foi negativamente associada ao desvio de temperatura mensal. Enquanto que, Luz et al. (2008) utilizaram a metodologia de Box-Jenkins para ajustar um modelo auto-regressivo integrado de média móvel (ARIMA) à incidência de dengue no Rio de Janeiro, Brasil, 1997-2004.

Naish et al. (2006) analisaram o impacto da variabilidade de temperatura e das marés sobre a transmissão do *Vírus Barmah Forest* (VBF) e desenvolveram um modelo de previsão baseado na temperatura para estudar o VBF na região de Gladstone na Austrália. O modelo adotado foi o SARIMA com as seguintes variáveis: temperatura mínima e máxima, pluviosidade e umidade relativa, a indicação de maré alta e maré baixa e o tamanho da população estudada.

Dentre os vários modelos na literatura para a modelagem de séries temporais, neste artigo abordam-se os modelos SARIMA (BOX e JENKINS, 1976) e os Modelos Estruturais (ME) (HARVEY, 1989) também conhecidos como modelos lineares dinâmicos (WEST e HARRISON, 1997) sob a perspectiva Bayesiana. Uma das grandes vantagens dos últimos é que a série temporal pode ser decomposta

em termos de seus componentes não-observáveis tais como nível, tendência e sazonalidade, os quais têm uma interpretação direta. Além disso, coeficientes de regressão variando no tempo, dados ausentes e extensões multivariadas são facilmente tratados no contexto dos ME, enquanto que, em um modelo SARIMA, isso pode não ser tão simples (COMMANDEUR e KOOPMAN, 1997). Harvey e Todd (1983) fizeram uma comparação entre os modelos SARIMA e ME para seis séries macroeconômicas da Grã-Bretanha, ilustrando as vantagens dos últimos. Maiores detalhes e diferenças sobre esses modelos podem ser encontrados em HARVEY (1989); COMMANDEUR e KOOPMAN (1997). Nobre et al. (2001) compararam os modelos estruturais e SARIMA com dados epidemiológicos.

Os objetivos deste estudo são:

- Fazer uma comparação de ajuste e preditiva entre dois modelos de séries temporais, a saber, os modelos SARIMA e os modelos estruturais para os dados de dengue em BH;
- Estudar a evolução da incidência ao longo dos anos.

Este trabalho está organizado da seguinte forma. Na Seção 2, os dados e o problema motivante são mostrados. Nas Seções 3 e 4, são apresentados os modelos SARIMA e modelos estruturais, respectivamente, bem como os métodos de inferência e previsão. Na Seção 5, mostram-se os resultados dos ajustes dos modelos, previsão e comparação da capacidade preditiva dos mesmos. Finalmente, na Seção 6, conclui-se o trabalho e algumas considerações finais são discutidas.

2 Materiais

Os dados do número de casos notificados de Dengue em BH no período de 01/2002 a 03/2008, totalizando $n = 75$ observações mensais são analisados. Os dados dos próximos 12 meses (de 04/2008 a 03/2009) são conhecidos, usados como “padrão ouro” e omitidos do ajuste dos modelos para efeito de comparação da capacidade preditiva dos mesmos, isto é, a acurácia das previsões “out-of-sample”. Essas informações foram cedidas pela Secretaria de Estado da Saúde de MG (SES MG, 2010).

A incidência em um instante de tempo é definida como o número de casos notificados dividido pela população sob risco multiplicado por 100 mil. É comum, em estudos de previsão de casos de dengue, devido a grande quantidade de casos, a modelagem da série na escala logarítmica (LUZ et al., 2008, ver). Portanto, a série temporal de interesse deste estudo é dado por

$$y_t = \ln \left(\frac{\text{número de casos no tempo } t}{\text{pop. sob risco no tempo } t} \times 100.000 \right),$$

para $t = 1, \dots, n$. Foram usadas projeções populacionais anuais disponibilizadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE) (<http://www.ibge.gov.br>). A população sob risco mensal na definição da série y_t foi aproximada pela sua respectiva projeção anual.

3 Modelos SARIMA

Uma dos métodos mais utilizadas para a análise de séries temporais é o de Box & Jenkins (1976). De acordo com esse método, modelos autorregressivos (AR) e de médias móveis (MA) são ajustados à série temporal (Box & Jenkins, 1976).

O modelo SARIMA é uma extensão da classe de modelos ARIMA, que inclui a sazonalidade. Suponha uma série temporal sazonal y_t observada em s períodos por ano, $s = 12$ para séries mensais. Um modelo $SARIMA(p, d, q)(P, D, Q)_s$ para o processo y_t é definido da seguinte forma (MORETTIN e TOLOI, 2006):

$$\phi(B)\Phi(B^s)\Delta^d\Delta_s^D y_t = \theta(B)\Theta(B^s)u_t, \quad (1)$$

para $t = 1, \dots, n$ onde:

- $\Phi(B^s) = (1 - \Phi_1 B^s - \dots - \Phi_P B^{Ps})$ é o polinômio autorregressivo sazonal de ordem P ;
- $\Theta(B^s) = (1 - \Theta_1 B^s - \dots - \Theta_Q B^{Qs})$ é o polinômio de médias móveis sazonal de ordem Q ;
- $\phi(B) = (1 - \phi_1 B - \dots - \phi_p B^p)$ é o polinômio autorregressivo de ordem p ;
- $\theta(B) = (1 - \theta_1 B - \dots - \theta_q B^q)$ é o polinômio de médias móveis de ordem q ;
- $\Delta^d = (1 - B)^d$ é o operador de diferença e d indica o número de diferenças necessárias para tornar a série estacionária com respeito à média;
- $\Delta_s^D = (1 - B^s)^D$ é o operador de diferença sazonal e D indica o número de diferenças sazonais necessárias para tratar a sazonalidade da série;
- u_t é um processo ruído branco Gaussiano com média zero e variância σ_u^2 .

Para se obter os estimadores de máxima verossimilhança dos parâmetros, maximiza-se através de procedimentos numéricos a função de log-verossimilhança.

3.1 Previsão em modelos SARIMA

As previsões k passos à frente foram obtidas a partir da equação do modelo ajustado. Seja $\hat{y}_n(k)$ a previsão para y_{n+k} no instante $n+k$. A função de previsão é a esperança condicional dos valores futuros com base nas observações até o instante n (MORETTIN e TOLOI, 2006).

Considerando o modelo na forma da equação de diferença em (1), a função de previsão na origem n e k passos a frente é dada por (MORETTIN e TOLOI, 2006)

$$\hat{y}_n(k) = E[y_{n+k} | y_1, \dots, y_n], \quad (2)$$

assumindo as seguintes condições:

$$\begin{aligned} [y_{n+k}] &= \hat{y}_n(k), \quad k > 0; \\ [y_{n+k}] &= y_{n+k}, \quad k \leq 0; \end{aligned}$$

$$\begin{aligned} [u_{n+k}] &= 0, \quad k > 0; \\ [u_{n+k}] &= e_{k+k}, \quad k \leq 0. \end{aligned}$$

A variância do erro de previsão é

$$V(k) = (1 + \tau_1^2 + \tau_2^2 + \dots + \tau_{k-1}^2)\sigma_u^2,$$

em que os pesos τ_i são determinados através da forma de choques aleatórios ($MA(\infty)$) do modelo SARIMA em (1) (MORETTIN e TOLOI, 2006, p. 225).

Com a previsão pontual e a sua variância, pode-se construir um intervalo de previsão de nível $1 - \kappa$ de confiança para a observação futura y_{n+k} da seguinte maneira:

$$\hat{y}_n(k) \pm |z_{\kappa/2}| \sqrt{V(k)},$$

onde $z_{\kappa/2}$ é o percentil de ordem $\kappa/2$ de uma distribuição Normal-padrão.

4 Modelos estruturais

Os modelos estruturais (ME) (HARVEY, 1989) decompõem uma série temporal y_t em termos de seus componentes não-observáveis ou estados, tais como nível (μ), tendência (β), sazonalidade (γ) e erro (u). O modelo estrutural básico (MEB) é um caso particular dos ME e é comumente usado, possuindo o componente de sazonalidade, dado por (HARVEY, 1989)

$$y_t = \mu_t + \gamma_t + u_t, \quad u_t \sim N(0, \sigma_u^2) \quad i.i.d. \quad (3)$$

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2) \quad i.i.d. \quad (4)$$

$$\beta_t = \beta_{t-1} + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2) \quad i.i.d. \quad (5)$$

$$\gamma_t = -\gamma_{t-1} - \dots - \gamma_{t-s+1} + \omega_t, \quad \omega_t \sim N(0, \sigma_\omega^2) \quad i.i.d. \quad (6)$$

onde s é o número de períodos sazonais e u_t , η_t , ξ_t e ω_t são erros ruído branco mutuamente não-correlacionados. O vetor de estados é dado por $\alpha'_t = (\mu_t, \beta_t, \gamma_t, \gamma_{t-1}, \dots, \gamma_{t-s+1})$. A representação da sazonalidade no modelo acima é feita por fatores sazonais ou *dummies* sazonais (HARVEY, 1989; WEST e HARRISON, 1997; FRANCO et al., 2009). O vetor $\psi = (\sigma_u^2, \sigma_\eta^2, \sigma_\xi^2, \sigma_\omega^2)$, composto pelas as variâncias dos termos de erro do modelo, é conhecido como vetor de parâmetros estáticos ou hiperparâmetros (HARVEY, 1989).

As Equações (3)-(6) são normalmente escritas na forma de espaço de estados, para permitir o uso de algoritmos lineares, tais como o filtro de Kalman (KALMAN, 1960), na estimação dos componentes não-observáveis ou dos estados.

O estimador de máxima verossimilhança (EMV) e suas propriedades assintóticas são usados para fazer inferência para o vetor de hiperparâmetros ψ . É necessário também usar procedimentos de otimização numérica para encontrar $\hat{\psi}_{EMV}$. Para maiores detalhes, consultar HARVEY (1989); FRANCO et al. (2008, 2009).

4.1 Previsão em modelos estruturais

A função de previsão k passos à frente, baseada em toda informação disponível até o instante n , é definida como $\tilde{y}_{n+k|n,\psi} = E(y_{n+k}|\mathbf{Y}_n, \psi)$ (HARVEY, 1989; PETRIS et al., 2009; FRANCO et al., 2009). Para o MEB, a função de previsão é dada por (FRANCO et al., 2009)

$$\tilde{y}_{n+k|n,\psi} = \mathbf{z}'_{n+k} \left(\prod_{i=1}^k \mathbf{T}_{n+i} \right) \mathbf{a}_n, \quad (7)$$

E o seu EQM de $(\tilde{y}_{n+k}|\mathbf{Y}_n, \psi)$ é dado por

$$\begin{aligned} & \mathbf{z}'_{n+k} \left(\prod_{i=1}^k \mathbf{T}_{n+i} \right) \mathbf{P}_n \left(\prod_{i=1}^k \mathbf{T}'_{n+i} \right) \mathbf{z}_{n+k} + \mathbf{z}'_{n+k} \sum_{i=1}^k \left(\prod_{j=1}^{k-i} \mathbf{T}_{n+k-j+1} \right) \mathbf{R}_{n+i} \mathbf{Q}_{n+i} \mathbf{R}'_{n+i} \\ & \left(\prod_{j=1}^{k-i} \mathbf{T}'_{n+k-j+1} \right) \mathbf{z}_{n+k} + \sigma_u^2, \end{aligned} \quad (8)$$

onde:

\mathbf{a}_n e \mathbf{P}_n são a média e a matriz de covariâncias da distribuição de atualização do vetor de estados (obtida pelo filtro de Kalman), $\mathbf{z}'_t = \begin{bmatrix} 1 & 0 & 1 & 0 & \dots & 0 \\ \sim & & & & & \end{bmatrix}_{1 \times (s+1)}$,

$\mathbf{R}_t = \mathbf{I}_{s+1}$ (matriz identidade),

$$\mathbf{T}_t = \begin{bmatrix} 1 & 1 & | & & & & 0 \\ 0 & 1 & | & & & & \sim \\ - & - & | & - & - & - & - \\ & & | & -1 & -1 & \dots & -1 & -1 \\ & & | & 1 & 0 & \dots & 0 & 0 \\ 0 & \sim & | & 0 & 1 & \dots & 0 & 0 \\ & & | & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & | & 0 & 0 & \dots & 1 & 0 \end{bmatrix}_{(s+1) \times (s+1)} \quad \text{e}$$

$$\mathbf{Q}_t = \begin{bmatrix} \sigma_\eta^2 & & & & & & & \\ & \sigma_\xi^2 & & & & & 0 & \\ & & \sigma_\omega^2 & & & & \sim & \\ & & & 0 & & & & \\ & & & & & & & \ddots \\ & 0 & & & & & & \\ & \sim & & & & & & \\ & & & & & & & 0 \end{bmatrix}_{(s+1) \times (s+1)} .$$

Essas matrizes são invariantes no tempo.

A forma da função de previsão abordada acima pode ser denominada como natural ou padrão. O intervalo de previsão, de nível $1 - \kappa$ natural ou padrão é dado por:

$$\left[\tilde{y}_{n+k|n,\psi} \pm |z_{\kappa/2}| \sqrt{\widehat{EQM}(\tilde{y}_{n+k|n,\psi})} \right], \quad (9)$$

onde $z_{\kappa/2}$ é o percentil de ordem $\kappa/2$ de uma distribuição Normal-padrão.

Quando o valor de ψ é desconhecido, deve-se estimá-lo. A prática mais comum é substituir ψ por $\hat{\psi}$ (BROCKWELL e DAVIS, 1996, ver).

5 Resultados

Nesta seção são comparados os modelos SARIMA e os modelos estruturais para a previsão da incidência de dengue em BH. Para tanto, serão considerados os dados do número de casos notificados de Dengue em BH do período de 01/2002 a 03/2008, totalizando 75 observações. Os dados dos próximos 12 meses ($k = 12$) são conhecidos e omitidos do ajuste dos modelos para efeito de comparação da capacidade preditiva dos mesmos.

O gráfico da série, os correlogramas e o gráfico de probabilidade Normal estão na Figura 1. Os dados não parecem seguir uma distribuição Normal de probabilidade e possuem correlações significativas. A série apresenta um comportamento sazonal. Os meses de dezembro a março têm uma incidência maior de dengue, já que se tratam de meses de verão os quais possuem temperaturas elevadas e chuvas (clima propício à proliferação do mosquito transmissor da dengue) (CORRÊA et al., 2005).

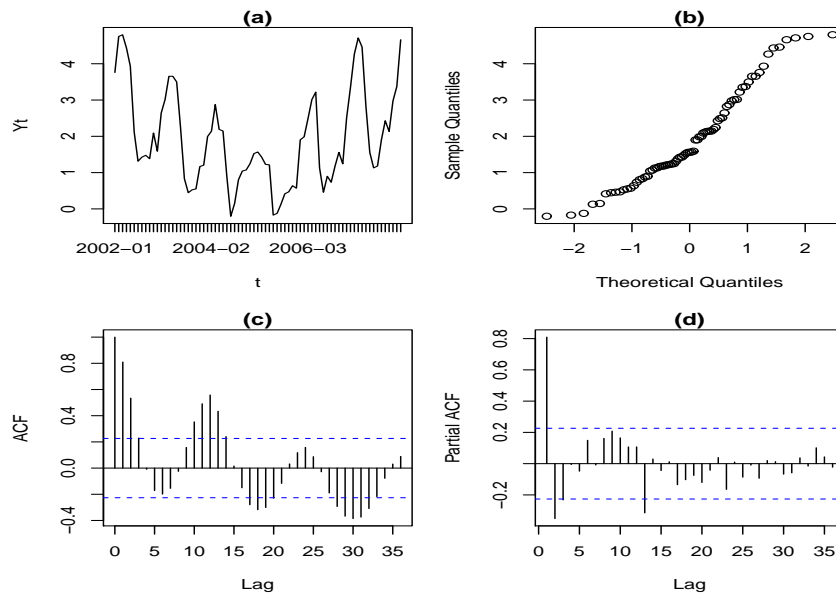


Figura 1 - (a) O gráfico temporal da série do logaritmo natural da incidência de dengue por 100.000 habitantes; (b) O QQ-Plot Normal da série; (c) Função de autocorrelação amostral da série; (d) Função de autocorrelação parcial amostral da série. Nível nominal de confiança assumido é de 99%.

5.1 MEB

Como a série possui os componentes de nível e sazonalidade, de acordo com a Seção 2, o Modelo Estrutural Básico (MEB) com sazonalidade modelada por variáveis *dummy* é ajustado a essa série transformada no software Ox (DOORNIK, 2002). As EMV e o intervalo de confiança *bootstrap* (FRANCO et al., 2008) para os hiperparâmetros do modelo estão na Tabela 1. Com a exceção do hiperparâmetro do componente de nível σ_{η}^2 , os demais hiperparâmetros (variâncias) são bem pequenos, algo comum na prática (FRANCO et al., 2009).

Tabela 1 - Estimativas de máxima verossimilhança e intervalos de confiança *bootstrap* para os hiperparâmetros do MEB ajustado à série do logaritmo da incidência de casos de dengue em BH

Parâmetro	Estimativa	LI 95%	LS 95%
σ_{η}^2	0,136	0,0494	0,1717
σ_{ξ}^2	$1,130 \times 10^{-4}$	$1,519 \times 10^{-13}$	0,00386
σ_{ω}^2	$9,496 \times 10^{-6}$	$5,317 \times 10^{-21}$	0,002593
σ_u^2	$2,651 \times 10^{-5}$	$5,518 \times 10^{-10}$	0,0337

Analisando a Figura 2, percebe-se que o ajuste do MEB é adequado aos dados. Os resíduos não evidenciam nenhum padrão, ou seja, não são autocorrelacionados e se aderem a uma distribuição Normal. Além da análise gráfica, foram feitos os testes de normalidade de Shapiro-Wilks e de Ljung-Box para checar as hipóteses de normalidade e não correlação dos resíduos, respectivamente. O valor da estatística de teste de Ljung-Box e seu p-valor para os lags 6, 12, 24 e 36 foram $Q(6) = 5,86$ (p-valor = 0,06), $Q(12) = 13,46$ (p-valor = 0,10), $Q(24) = 24,61$ (p-valor = 0,22) e $Q(36) = 33,02$ (p-valor = 0,42), respectivamente. Não se rejeita a hipótese de que os resíduos são independentes. Já o p-valor do teste de normalidade foi maior que 0,10, ou seja, não se rejeita a hipótese que os resíduos são oriundos de uma distribuição Normal.

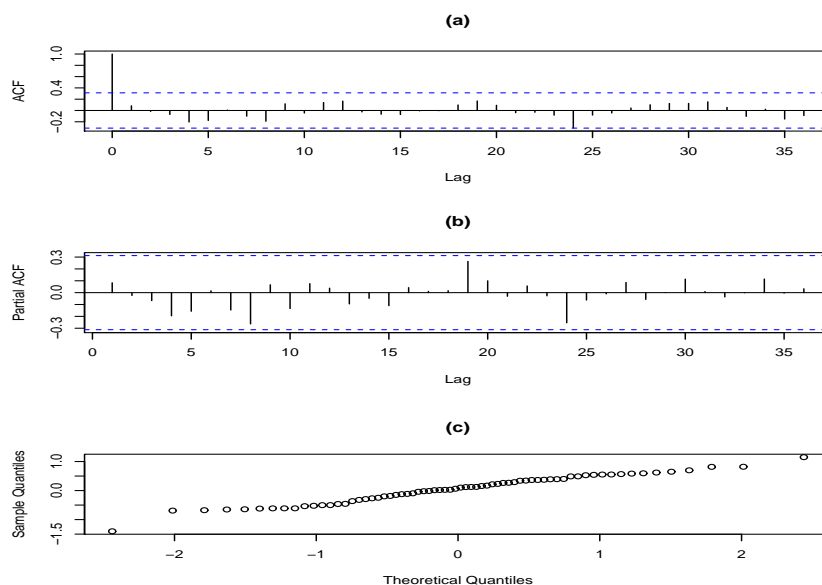


Figura 2 - Análise dos resíduos do MEB. (a) Função de autocorrelação amostral dos resíduos; (b) Função de autocorrelação parcial amostral dos resíduos; (c) O QQ-Plot Normal dos resíduos. Nível nominal de confiança de 99%.

Analisando a Figura 3, obtêm-se informações sobre as componentes não-observações da série, tais como nível, tendência e sazonalidade. Em 1998, Minas Gerais enfrentou sua primeira epidemia que atingiu todo o estado e principalmente a região Metropolitana de Belo Horizonte. Ocorreu uma diminuição no número de casos. No ano seguinte, entretanto apareceu nova tendência de aumento, culminando em um segundo pico em 2002, que corresponde o pico inicial no gráfico de tendência (no topo à direita) em 2002.

Houve um decréscimo na tendência de 2003 a 2005 e, em torno de 2005, se tornou não significativa. A partir 2006, em relação a 2005, apresentou nova tendência de aumento significativa que se manteve até os meses de dados disponíveis deste estudo. Essas informações são muito úteis para formulação de políticas públicas para o combate da dengue.

O gráfico da componente de sazonalidade nos mostra um aumento na incidência de casos no início do ano se comparado com o fim do mesmo, o que já foi também observado pelo gráfico da série (Figura 1).

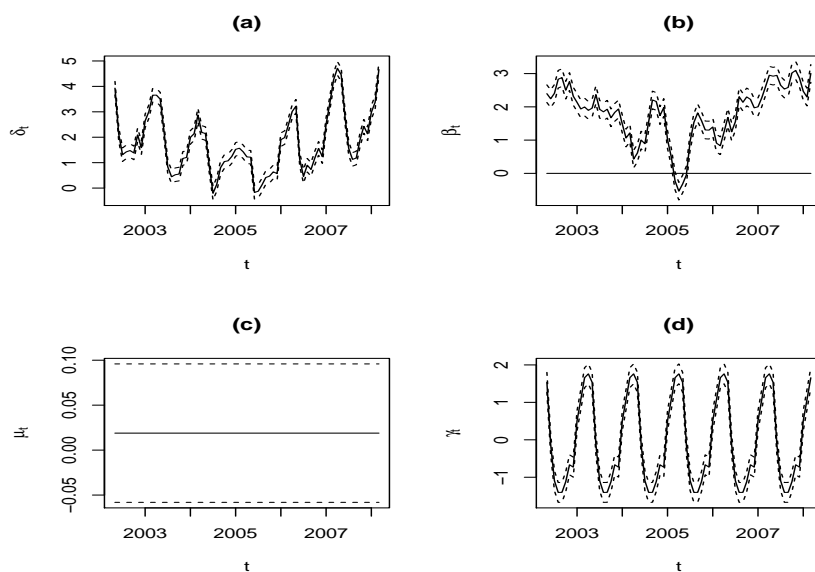


Figura 3 - Estimativas Suavizadas do vetor de estados. (a) A média suavizada dos dados; (b) Estimativa suavizada do componente de tendência (inclinação) da série; (c) Estimativa suavizada do nível da série; (d) Estimativa suavizada da sazonalidade da série. Linha tracejada indica o intervalo de confiança de nível nominal de 95%.

5.2 SARIMA

Os modelos SARIMA (BOX e JENKINS, 1976) são ajustados à série no software R (R CORE TEAM, 2016), para fazer uma comparação sobre a capacidade preditiva desses modelos frente aos ME. Segundo os critérios *AIC* e log-verossimilhança, foi escolhido o modelo $SARIMA(1, 1, 0) \times (1, 0, 0)_{12}$ entre os vários ajustados - que é utilizado para a comparação das previsões com os modelos estruturais.

A Tabela 2 contém as estimativas de máxima verossimilhança dos parâmetros do modelo SARIMA. Note que apenas o parâmetro Φ_1 é significativo ao nível de 5% de significância. O intercepto e o parâmetro autorregressivo de ordem 1 (ϕ_1) não foram significativos. Isso provavelmente ocorreu devido à aplicação de uma diferença não sazonal. Observe que o modelo é estacionário, uma vez que o módulo das estimativas dos parâmetros ϕ_1 e Φ_1 são menores que 1.

Tabela 2 - Estimativas de máxima verossimilhança e erros-padrão para os parâmetros do modelo SARIMA, ajustado à série do logaritmo da incidência de casos de Dengue em BH

Parâmetro	Estimativa	Erro-padrão	LI 95%	LS 95%
ϕ_0	-0,003	0,188	-0,371	0,365
ϕ_1	0,142	0,122	-0,097	0,381
Φ_1	0,781	0,070	0,644	0,918

Analisando a Figura 4, não há evidência de violação das suposições do ajuste do modelo SARIMA. Foram realizados os testes de normalidade de Shapiro-Wilks e de Ljung-Box para checar as hipóteses de normalidade e não autocorrelação dos resíduos, respectivamente. O valor da estatística de teste de Ljung-Box e seu p-valor para os lags 6, 12, 24 e 36 foram $Q(6) = 7,26$ (p-valor = 0,12), $Q(12) = 10,91$ (p-valor = 0,37), $Q(24) = 19,47$ (p-valor = 0,617) e $Q(36) = 21,20$ (p-valor = 0,96), respectivamente. Não se rejeita a hipótese de que os resíduos são independentes. Também não se rejeita a hipótese que os resíduos são provenientes de uma distribuição Normal (p-valor = 0,97).

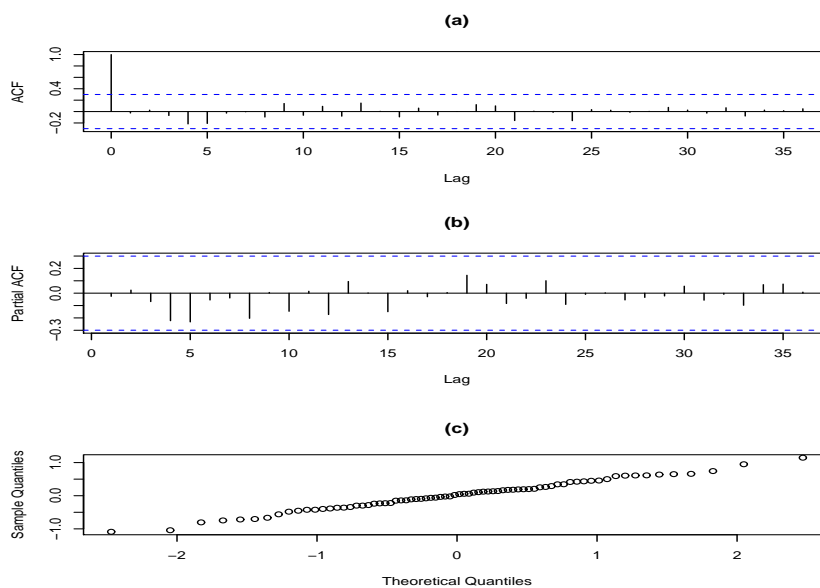


Figura 4 - Análise dos resíduos do modelo SARIMA. (a) Função de autocorrelação amostral dos resíduos; (b) Função de autocorrelação parcial amostral dos resíduos; (c) O QQ-Plot Normal dos resíduos. Nível nominal de confiança de 99%.

5.3 Comparação dos modelos e previsões

Em termos de comparação dos ajustes dos modelos, o MEB teve o $AIC = 93,00$ e a log-verossimilhança = $-43,00$, enquanto que o modelo SARIMA obteve o $AIC = 108,24$ e a log-verossimilhança = $-51,12$. Isso evidencia um melhor ajuste do MEB aos dados. A comparação em termos de ajuste do MEB com o modelo SARIMA não é algo incomum e também já foi realizada por HARVEY (1989, p.94).

As previsões para as observações futuras dos próximos 12 meses (de 04/2008 a 03/2009) da série estão na Tabela 3 e são bem razoáveis para ambos modelos, próximas dos valores observados. As estatísticas de erro de previsão Erro Quadrático Médio (EQM) e Erro Absoluto Médio (EAM) na Tabela 4 mostram um melhor desempenho preditivo dos modelos estruturais frente aos modelos SARIMA para previsões a mais de um passo a frente ($k > 1$). Para $k = 1$, o modelo SARIMA tem um desempenho ligeiramente melhor.

Tabela 3 - Valores observados da série e previsão para a série mensal da log-incidência de dengue em BH usando os modelos SARIMA e MEB

Meses	Valores Observados	Previsão 12 passos a frente	
		MEB	SARIMA(1, 1, 0) × (1, 0, 0) ₁₂
abr/08	5,85	4,65	5,06
mai/08	5,63	3,02	4,44
jun/08	4,24	2,19	3,16
jul/08	2,52	2,28	1,43
ago/08	1,54	2,47	0,33
set/08	1,28	2,74	0,55
out/08	1,77	3,17	1,48
nov/08	2,27	3,04	2,12
dez/08	2,65	4,00	1,85
jan/09	4,22	4,44	2,36
fev/09	4,89	5,04	2,41
mar/09	5,53	5,03	3,99

Tabela 4 - Estatísticas de erro de previsão a 1 e 12 passos a frente usando os modelos SARIMA e MEB

Lags	EQM		EAM	
	SARIMA	ME	SARIMA	ME
1	0,63	1,44	0,79	1,19
12	1,60	0,71	1,10	0,76

Na Figura 5 estão as previsões e intervalos de previsão de 95% para as observações futuras, obtidos através dos modelos ajustados aos dados. A linha contínua após a linha vertical representa os valores reais. Os intervalos de confiança assintóticos cobriram todos os valores reais futuros sob o MEB e o SARIMA, porém

a amplitude dos intervalos do SARIMA é bem maior do que do MEB. Os intervalos de previsão via modelos estruturais seguem bem o comportamento das observações futuras. O intervalo de previsão via o SARIMA parece seguir bem o comportamento das primeiras observações ($k \leq 6$), mas não o das últimas ($k > 6$).

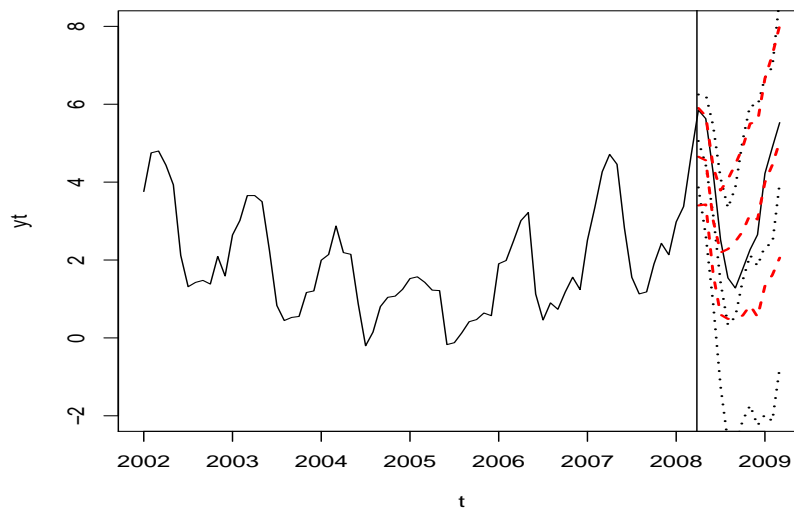


Figura 5 - Previsão para a série mensal da log-incidência de dengue em BH. **Linha contínua em preto:** a série do logaritmo da incidência de casos de Dengue em BH; **Linha pontilhada em preto:** as estimativas pontual e intervalar do modelo SARIMA; **Linha tracejada em vermelho:** a estimativa pontual e o intervalo de previsão assintótico (I^N) do MEB. A linha vertical refere-se a 03/2008 e separa os dados usados na estimação do modelo das previsões. O nível de confiança adotado é de 95%.

6 Conclusão

Neste artigo as previsões para os casos de dengue usando os modelos SARIMA e estruturais são construídas e comparadas. O último tem um melhor desempenho do que o primeiro, de acordo com os critérios EQM e EAM. HARVEY e TODD (1983) e NOBRE et al. (2001) também chegaram em resultados similares indicando o bom desempenho dos modelos estruturais em termos de capacidade preditiva e flexibilidade.

A previsão revela um aumento na incidência de dengue para os próximos anos. Através dos componentes não-observáveis, pode ser avaliada a evolução da tendência

e sazonalidade da incidência de dengue em BH.

Para o ajuste de ME, há muitos softwares e pacotes tais como dlm no software R (R CORE TEAM, 2016), STAMP, SsfPack (KOOPMAN et al. , 1999) no software Ox (DOORNIK, 2002).

Alguns estudos apontam a importância da inclusão de covariáveis ambientais e climáticas (LUZ et al., 2008), que podem ser facilmente incluída no modelo estrutural, tais como temperatura e chuva.

Trabalhos futuros incluem a incorporação dessas covariáveis, funções de transferência (SANTOS et al., , 2010) e informações espaciais no modelo.

Agradecimentos

T. R. Santos foi apoiado pela Universidade Federal de Minas Gerais (PrPq-UFMG), CNPq-Brasil e FAPEMIG. O autor agradece ao Prof. Dr. Ricardo Tavares e Secretaria de Estado da Saúde MG por disponibilizar os dados.

SANTOS, T. R. A comparison of SARIMA and structural models for forecasting of dengue incidence in Belo Horizonte, Minas Gerais. *Rev. Bras. Biom.*, Lavras, v.35, n.1, p.98-114, 2017.

■ **ABSTRACT:** *Forecasts for dengue incidence are essentials for public policies and actions, providing to health managers monthly estimates of the number of new cases in order that they can, for example, plan and organize strategies, health teams, etc. This paper provides forecasts of the future values of dengue incidence in Belo Horizonte, Brazil, based on Seasonal Autoregressive Integrated Moving Average (SARIMA) models and Structural Models (SM), comparing them with respect to their fit and out-of-sample forecasts. In general, forecasts via the SM had a better performance than the SARIMA ones, according to the forecast error statistics. Furthermore, as the SM decompose the time series in terms of its unobservable components, the behavior of the trend component of dengue incidence may be studied over time. This component decreases between 2002 and 2005 and increases after 2005. The months of the highest and lowest incidence can also be identified through the seasonality component. The first months of year have a high incidence, which was already expected. The results are very satisfactory and motivate future studies.*

■ **KEYWORDS:** *Dengue; state space models; classic inference; forecasting.*

Referências

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

- BOX, G. E. P.; JENKINS, G. M. *Time Series Analysis: Forecasting and Control*, 2.ed. San Francisco: Holden Day, 1976.
- BROCKWELL, P. J.; DAVIS, R. A. *Introduction to Time Series and Forecasting*, Springer text in Statistics. New York: Springer, 1996.
- COMMANDEUR, J. J.; KOOPMAN, S. J. *An introduction to state space time series analysis*. Oxford University Press, 2007.
- CORRÊA, P. R. L.; FRANÇA, E.; BOGUTCHI, T. F. Infestação pelo *Aedes aegypti* e ocorrência da dengue em Belo Horizonte, Minas Gerais. *Revista de Saúde Pública*, v.39, n.1, p.33-40, 2005.
- DOORNIK, J. A. *Ox: An Object-Oriented Matrix Language*. 3. ed. London: Timberlake Consultants Press, 2002.
- FRANCO, G. C.; SANTOS, T. R.; RIBEIRO, J. A.; CRUZ, F. R. B. Confidence intervals for hyperparameters in structural models. *Communications in Statistics: Simulation and Computation*, v.37, n.3, p.486-497, 2008.
- FRANCO, G. C.; GAMERMAN, D.; and SANTOS, T. R. *Modelos de espaço de estados: abordagens clássica e bayesiana*. São Paulo: Associação Brasileira de Estatística, 2009.
- HARVEY, A. C. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: University Press, 1989.
- HARVEY, A. C. Comments of Time Series analysis of non-gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society, Ser. B*, v.62, p.34-35, 2000.
- HARVEY, A. C.; TODD, P. H. J. Forecasting economic time series with structural and Box-Jenkins models: A case study. *Journal of Business & Economic Statistics*, v.1, n.4, p.299-307, 1983.
- KALMAN, R. E. A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng.*, v.82, p.35-45, 1960.
- KOOPMAN, S. J., SHEPHARD, N.; DOORNIK, J. A. Statistical algorithms for inference in state space using SsfPack 2.2 *Econometrics Journal*, v.47, p.129-139, 1999.
- KULLDORF, M. A Spatial Scan Statistic. *Communications in Statistics: Theory and Methods*. v.26, n.6, p.1481-1496, 1997.
- KULLDORF, M.; ATHAS, W.F.; FEURER, E.J.; MILLER, B.A.; KEY, C.R. Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American Journal of Public Health*. v.88, n.9, p.1377-1380, 1998.

- LUZ, P. M.; MENDES, V. M.; CODEÇO, C. T.; STRUCHINER, C.J.; GALVANI, A.P. Time series analysis of Dengue Incidence in Rio de Janeiro, Brazil. *American Journal of Tropical Medicine and Hygiene*, v.79, p.933-939, 2008.
- MARTINEZ, E. Z.; SILVA, E. A. S. Predicting the number of cases of dengue infection in Ribeirão Preto, São Paulo State, Brazil, using a SARIMA model. *Cadernos de saude publica*, v.27, p.1809-1818, 2011.
- MORETTIN, P. A.; TOLOI, C. M. C. *Análise de Séries Temporais*. São Paulo: Blucher, 2006.
- NAISH, S.; HU, W.; NICHOLLS, N.; MACKENZIE, J. S.; McMICHAEL, A. J.; DALE, P.; TONG, S. Weather Variability, Tides, and Barmah Forest Virus Disease in the Gladstone Region, Australia. *Environmental Health Perspectives*, v.114, n.5, p.678-683, 2006.
- NOBRE, F. F.; MONTEIRO, A. B. S.; TELLES, P. R.; WILLIAMSON, G. D.. Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology. *Statistics in medicine*, v.20, n.20, p.3051-3069, 2001.
- PETRIS, G.; PETRONE, S.; CAMPAGNOLI, P. *Dynamic linear models with R*. Springer, 2009.
- R CORE TEAM. *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, 2016.
- SANTOS, T.R.; FRANCO, G.C.; GAMERMAN, D. Comparison of classical and Bayesian approaches for intervention analysis in structural models. *International Statistical Review*, v.78, n.2, p.218-239, 2010.
- SECRETARIA DE ESTADO DE SAÚDE DE MINAS GERAIS. Superintendência de Epidemiologia, Gerência de Vigilância Ambiental, 2010.
- UNKEL, S.; FARRINGTON, C. P.; GARTHWAITE, P. H.; ROBERTSON, C.; ANDREWS, N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical society: Series A*. v.175, n.1, p.49-82, 2012.
- WEST, M.; HARRISON, J. *Bayesian Forecasting and Dynamic Models*. New York: Springer, 1997.
- WORLD HEALTH ORGANIZATION. *Dengue hemorrhagic fever. Diagnosis, treatment, prevention*. World Health Organization, Ginebra, Suíça, 2014.
- WU, P.; GUO, H.; LUNG, S.; LIN, C.; SU, H. Weather as an effective predictor for occurrence of dengue fever in Taiwan. *Acta Tropica*. v.103, n.1, p.50-57, 2007.

Recebido em 21.01.2016.

Aprovado após revisão em 29.07.2016.