ISSN: 2317-0840

Análise Temporal da Incidência de Leptospirose e sua Relação com o Índice Pluviométrico na Cidade de Recife – PE, 2007 - 2016

Jesy K. S. dos Santos[1†], Carlos R. A. Daniel[2], André L. P. dos Santos[3], Guilherme R. Moreira[4]

[1]Universidade Federal de Sergipe – UFS. E-mail: jesy.sales.comunic@gmail.com
[2]Universidade Federal de Sergipe – UFS. E-mail: raphael_crad@yahoo.com.br
[3]Universidade Federal Rural de Pernambuco – UFRPE. E-mail: andredefensor@hotmail.com
[4]Universidade Federal Rural de Pernambuco – UFRPE. E-mail: guirocham@gmail.com

Resumo: *A leptospirose é uma doença infecciosa causada pelo contato com a urina de ratos e outros animais contaminados pela bactéria leptospira. A disseminação e persistência da doença são facilitadas pelas enchentes e inundações que ocorrem nos períodos chuvosos. Observou-se em estudos anteriores que a doença apresenta um comportamento sazonal, ocorrendo com maior frequência nos meses de março, abril, maio e início de junho para o litoral da região Nordeste, coincidindo com os maiores níveis pluviométricos. O estado de Pernambuco registrou 47% do total de casos da região Nordeste no ano de 2017, enquanto sua capital representou neste mesmo ano 66% dos casos do estado. Como existe uma associação entre os períodos de chuva e maior incidência de leptospirose, os dados de pluviosidade podem auxiliar na descrição da variável de interesse. O objetivo deste trabalho é comparar o desempenho de modelos temporais, dinâmicos, e regressão beta por meio da análise de previsões, para a cidade de Recife no período de 2007 a 2016. Foi possível observar que as previsões da incidência de leptospirose ajustada apenas com os valores passados, e o modelo de regressão beta com os índices pluviométricos como variável explicativa resultaram em melhor desempenho que o modelo dinâmico incluindo a quantidade de chuva. Conclui-se que tanto a utilização de técnicas de séries temporais quanto à inclusão da pluviosidade como variável explicativa permitem prever com antecipação a ocorrência da leptospirose.*

Palavras-chave: *Sazonalidade; Precipitação; Modelos dinâmicos; Regressão Beta; Desempenho de previsões.*

Abstract: *Leptospirosis is an infectious disease caused by the contact with the urine of rats and other animals contaminated by the leptospira bacteria. The spread and persistence of the disease is facilitated by the floods that occur in the rainy periods. It was observed in previous studies that the disease presents a seasonal behavior, occurring more frequently in the months of March, April, May and early June for the coast of the Northeast region, coinciding with the highest rainfall levels. The state of Pernambuco recorded 47% of the total cases in the Northeast region in 2017, while its capital represented 66% of the state's cases in the same year. As there is an association between periods of rain and a higher incidence of leptospirosis, rainfall data can assist in describing the variable of interest. The objective of this work is to compare the performance of seasonal time series, dynamic time series, and beta regression models through the analysis of predictions, for the city of Recife in the period from 2007 to 2016. It was possible to observe that the predictions of the incidence of leptospirosis adjusted only with the past values, and the beta regression model with pluviometric indices as an independent variable resulted in better performance than the dynamic model including the amount of rain. It is concluded that both the use of time series techniques and the inclusion of rainfall as an explanatory variable allow to anticipate the occurrence of leptospirosis.*

Keywords: *Seasonality; rainfall; dynamic model; Beta regression; forecasting performance.*

† Autora correspondente: jesy.sales.comunic@gmail.com.

Introdução

A leptospirose é um grande problema de saúde pública em países em desenvolvimento, inclusive no Brasil, devido ao seu crescimento desordenado, que leva ao surgimento de aglomerações de pessoas em locais de más condições sanitárias (BERNARDI, 2012; MARTINS, 2012). O ambiente insalubre, devido à escassez de saneamento básico nos grandes centros urbanos, principalmente nas favelas, e a cotidiana exposição à contaminação ambiental durante intenso período chuvoso são considerados os fatores fundamentais para a ocorrência das epidemias de leptospirose em área urbana. (TASSINARI et al., 2004).

Trata-se de uma doença infecciosa reemergente que tem registrado surtos em diversos locais do mundo. É causada pelo gênero l*eptospira*, uma bactéria que infecta os animais, principais transmissores para o homem que, em contato com a área contaminada, torna-se receptor da bactéria desenvolvendo sintomas que podem até levar à morte. No Brasil, as epidemias da doença ocorrem anualmente, principalmente em comunidades carentes, após enchentes, inundações e desastres naturais de grande magnitude (PELISSARI et al., 2011).

Estudos avaliam a leptospirose diante do contexto socioambiental, estabelecendo ou evidenciando uma relação entre a doença, a urbanização e o clima (GONÇALVES et al., 2016; GUIMARÃES et al., 2014; TASSINARI, 2004; COSTA, 2001). A relação entre clima e saúde é destacada no estudo de Aleixo e Neto (2017, p. 83): "Quando na vinda dos Holandeses ao Brasil, Willen Piso, Médico de Maurício de Nassau, estudou em 1641 as inundações do Rio Capibaribe no Recife, que ocasionou muitas perdas de vidas humanas" apresentando em seguida uma discussão temporal entre doenças e clima dentro de um contexto histórico.

Sob a mesma perspectiva de análise relacional da leptospirose e o clima, ainda estabelecendo a distribuição temporal, Costa et al. (2001) analisaram descritivamente o padrão da doença no caso específico da cidade de Salvador-BA, mostrando que o nível de precipitação pluviométrica na cidade é um fator que pode orientar quando recursos específicos devem ser alocados para evitar epidemias de leptospirose, particularmente as formas graves da mesma.

Para o município de Fortaleza, Magalhães, Zanella e Sales (2010) também em uma comparação entre casos da enfermidade e chuva de 2004 a 2007, verificaram que na primeira metade do ano, período de maiores precipitações, foi registrado a maior parte dos casos.

O estudo de Guimarães et al. (2014) considerou a série histórica de leptospirose de 2007 a 2012, e levando em consideração a ocorrência de desastres por inundações, mostrou que, para a cidade do Rio de Janeiro, a precipitação mensal apresentou-se como um fator fortemente correlacionado.

Nesse contexto, verifica-se que a região Sudeste, juntamente com a região Sul e Norte, registraram os maiores índices pluviométricos e também os maiores números de casos de leptospirose no período de 10 anos (2007 a 2016). O Sudeste brasileiro registrou mais de 13 mil casos da doença, seguido da região Sul com 12,7 mil casos.

A região Nordeste foi a quarta em números de registros de leptospirose, com um total de 6.040 nesse período. O estado de Pernambuco é notório dentre todos os estados nordestinos, tendo acumulado 34% dos casos registrados na região. Só em 2017 foram 221 pessoas diagnosticadas com a *leptospira* no estado e, destes indivíduos, 66% residiam na cidade de Recife (DATASUS, 2019).

Tendo em vista o número limitado de estudos utilizando técnicas estatísticas de modelagem e previsão para a região Nordeste, diante da situação do estado de Pernambuco e

mais especificamente sua capital, este estudo se propõe a comparar o desempenho de três abordagens na descrição da leptospirose na cidade de Recife no período de 2007 a 2017: um modelo de séries temporais para dados sazonais; um modelo de regressão com a pluviosidade como variável explicativa; e uma terceira estratégia baseada em modelos dinâmicos que utiliza tanto a correlação temporal quanto a contribuição de uma variável explicativa.

Material e Métodos

Para analisar a ideia proposta neste texto, foram obtidos do SINAM (2019) - Sistema de Informação de Agravos de Notificação - os casos confirmados e notificados da leptospirose. O cálculo da incidência da doença foi obtido usando, a cada ano, a população estimada da cidade de Recife pelo Instituto Brasileiro de Geografia e Estatística - IBGE (2019). Os dados pluviométricos acumulados em milímetros foram obtidos do sítio eletrônico da APAC (2019) - Agência Pernambucana de Águas e Climas, utilizando a média de registros dos seis postos de contagem localizados na cidade de Recife (Alto da Brasileira, Codecipe/Santo Amaro, Lamepe/Itep, Santo Amaro, Várzea e PCD).

Nas análises, foi utilizado o software livre R, versão 3.5.2 (R CORE TEAM, 2018), e as respectivas rotinas `auto.arima` do pacote `forecast`, `dynlm` do pacote `dynlm` e `betareg` do pacote de mesmo nome.

Medida de Correlação

Uma maneira de medir o grau de associação entre duas variáveis, ou seja, a forma como o comportamento de uma delas acompanha as variações na outra, é utilizando a Correlação Linear de Pearson (*Pearson Product-Moment Correlation).* Essa medida foi calculada no intuito de identificar uma possível associação entre as variáveis analisadas e verificar o grau de concordância dos valores assumidos pelas variáveis ao longo do tempo.

Os teoremas e fundamentos da Correlação apresentadas por Pearson (1896, 1920), são resumidos matematicamente por seu coeficiente, dado pela expressão:

$$\rho_{(X,Y)} = \frac{cov(X,Y)}{\sqrt{cov(X,X)}\sqrt{cov(Y,Y)}}$$

sendo

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y})$$

onde N é o número de elementos do conjunto de variáveis X e Y, x_i é o i-ésimo elemento de X e x a respectiva média amostral, analogamente para os elementos de Y.

Os valores do coeficiente de correlação de Pearson situam-se no intervalo [-1,1], sendo que os valores próximos dos extremos indicam mais alto grau de associação, ou seja, correlação forte, positiva ou negativa entre as variáveis, enquanto valores próximos de zero sugerem ausência de correlação linear.

Modelos de Séries Temporais Sazonais

Foi proposto o método de análise temporal autorregressivo integrado de médias móveis sazonais (*Seasonal AutoRegressive Integrated Moving Average* - SARIMA) para modelar a incidência de leptospirose, tendo em vista o comportamento cíclico da incidência da doença.

Segundo Silva (2005) os modelos autorregressivos foram desenvolvidos com propósito de explicar a observação presente Z_t em função dos seus respectivos p termos passados, Z_{t-1}, Z_{t-2}, ..., Z_{t-p}, em que p determina o número de passos entre as observações passadas e a observação atual. Segundo o mesmo autor, os modelos Auto-Regressivos e Integrados de Média Móvel (ARIMA) adequam-se às previsões de séries temporais cujo processo estocástico não é estacionário, para tanto são necessárias diferenciações da série a fim de alcançar tal estabilidade. A estacionariedade de um processo é determinada pela imutabilidade das propriedades estatísticas da série ao longo do tempo, isto significa, por exemplo, que as observações tendem a variar sobre o mesmo nível, ou seja, em torno de uma mesma média (IHAKA, 2005). Caso uma série temporal seja não-estacionária, é possível alcançar a estacionariedade aplicando sucessivas diferenças nos dados.

Além da não estacionariedade, outra característica do modelo autorregressivo integrado de média móvel sazonal, é a presença do componente de sazonalidade. Uma série é dita sazonal quando ocorrem similaridades na mesma depois de s intervalos básicos de tempo. Tendo como exemplo, quando s é igual a doze e o intervalo básico de tempo for um mês, a cada doze meses um padrão é repetido, sugerindo uma relação entre os mesmos meses de diferentes anos. Dessa forma, o modelo SARIMA pode ser definido como um modelo ARIMA em que são incorporados componentes sazonais (BOX; JENKINS; RENSEL, 1994, p. 327).

O modelo SARIMA possui parâmetros $(p,d,q)(P,D,Q)$ e cada um dos seus parâmetros foi definido em Morettin e Toloi (2006) e Silva (2005) da seguinte forma:

1. Parâmetro do processo autorregressivo de ordem p: $AR(p)$ ou $\varphi(B)$;
2. Parâmetro do processo de médias móveis de ordem q: $MA(q)$ ou $\theta(B)$;
3. d: Número de diferenças (ordinárias) para tornar a série estacionária: $\Delta^d = (1-B)^d$
4. P: Parâmetro autorregressivo sazonal de ordem P: $\Phi(B^s)$;
5. Q: Parâmetro de médias móveis sazonal de ordem Q: $\Theta(B^s)$;
6. D: Número de diferenças sazonais: $\Delta_s^D = (1-B^s)^D$;
7. a_t é o ruído branco.

O modelo SARIMA foi expresso matematicamente por Silva (2005) usando a seguinte notação:

$$\varphi_p(B)\Phi_p(B^s)\Delta^d\Delta_s{}^D Z_t = \theta_q(B)\Theta_Q(B^s)a_t \quad (1)$$

O modelo (1) implica que são necessárias *d* diferenças simples e *D* diferenças sazonais da série Z_t para que o processo $W_t = \Delta^d\Delta_s{}^D Z_t$ seja estacionário.

Séries Temporais Dinâmicas ou Modelo de Regressão dinâmico

Para verificar a influência da precipitação de chuva na variação do comportamento da taxa de leptospirose, foi utilizado o modelo temporal dinâmico. Este modelo é denominado dinâmico, pois permite a inclusão de uma série temporal na estrutura de uma regressão que, além de considerar a variável de interesse e seus valores defasados, também incorpora o efeito da variável explicativa. A estimação dos parâmetros do modelo é determinada por mínimos quadrados ordinários (ZANINI, 2000).

Os modelos de regressão dinâmica devem ser usados quando há dependência entre a variável de interesse e as variáveis explicativas (possivelmente causais), nesse caso estruturadas em séries temporais, ao passo que a estrutura de correlação da variável dependente não permite supor independência dos erros. (LOPES; PIMENTA, 2011).

A estruturação do modelo dinâmico é representada em forma vetorial na equação (2) como sendo uma regressão em que seus termos variam no tempo *t*, e a variável dependente é escrita em função de suas defasagens e uma ou mais variáveis independentes com respectivas defasagens:

$$\varphi(B)Z_t = \beta_t x_t + \varepsilon_t \quad (2)$$

tem-se que, no tempo *t*, Z_t: É a série temporal das observações da variável dependente; x_t: $(1,x_{t2},...,x_{tk})^T$ é um vetor $k\times1$ de regressores, com o primeiro componente usualmente igual a 1 (ZEILEIS et al., 2005); β_t: é um vetor $k\times1$ de coeficientes de regressão; ε_t: é o ruído aleatório associado ao modelo; $\varphi(B)$: $(1 - \varphi_1 B - \varphi_2 B^2 - ... - \varphi_m B^m)$ é um polinômio autorregressivo de ordem *p*, onde *B* é o operador de defasagem.

A relação anteriormente mencionada, entre a variável dependente, as explicativas e respectivas defasagens eq. (2), pode ser também apresentada na seguinte forma da eq. (3):

$$Y_t = \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_k x_{t-k} + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_m Y_{t-m} + \varepsilon_t \quad (3)$$

A inclusão do vetor de variáveis exógenas (ou explicativas) $\beta_t x_t$ é o que diferencia a regressão dinâmica dos modelos univariados ARIMA de Box e Jenkins, pois nestes apenas os valores passados da série Y_t e dos erros defasados são usados na modelagem e previsão de Y_t (DIAS, 2008).

Regressão Beta

Os modelos de regressão beta têm por objetivo modelar a variável resposta, ou dependente, definida no intervalo (0,1), por meio de uma estrutura de regressão que se baseia em uma função de ligação, covariáveis e coeficientes desconhecidos. Ferrari e Cribari-Neto (2004) propuseram a classe de modelos de regressão beta que é baseada na suposição de que a variável resposta segue uma distribuição beta e incorpora uma possível variação dos erros, associada a cada variável dependente.

No modelo de regressão beta, proposto por Ferrari e Cribari-Neto (2004) como uma forma de suprir algumas das limitações associadas aos modelos de Regressão Linear, principalmente, no que se refere à estrutura da variável reposta, foi considerado uma reparametrização da densidade da distribuição beta que permite modelar a média da variável resposta envolvendo parâmetros da regressão e um parâmetro de precisão.

A função de densidade da distribuição beta foi dada por:

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1$$

sendo que $p, q > 0$ e $\Gamma(p)$ é a função gama avaliada no ponto p, ou seja,

$$\Gamma(p) = \int_0^{\infty} y^{p-1} e^{-y} dy$$

A média e variância de $Y \sim B(p,q)$ são respectivamente

$$E(Y) = \frac{p}{p+q}$$

$$Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}$$

Para obter uma estrutura de regressão para a média da variável resposta com um parâmetro de dispersão, Ferrari e Cribari-Neto (2004) utilizaram uma reparametrização da densidade beta, fazendo $\mu = p/(p+q)$, enquanto $\phi = p+q$, ou seja, $p = \mu\phi$, e ainda $q = (1-\mu)\phi$. Com essa reparametrização, a densidade da distribuição beta pode ser reescrita como

$$f(y;\ \mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \qquad (4)$$

em que $0 < \mu < 1$ e $\phi > 0$, dessa forma $E(Y) = \mu$ e $Var(Y) = V(\mu)/(1+\phi)$. Sendo que o termo $V(\mu) = \mu(1-\mu)$ é a função de variância.

O modelo de regressão beta pode ser escrito em termos de uma função de ligação e um preditor linear. Para tal, sejam $y_0,...,y_n$ variáveis aleatórias independentes e identicamente distribuídas seguindo a distribuição beta reparametrizada, e seja μ_t a média de cada uma das t variáveis, tem-se

$$g(\mu_t) = \eta_t = \sum_{i=1}^{k} x_{ti}\beta_i = X_t^T\beta \quad (5)$$

em que x_i é uma variável explicativa e $\beta = (\beta_1, ..., \beta_k)^T$ são os parâmetros a serem estimados. A estimação dos η_t é feita por máxima verossimilhança.

Avaliação de desempenho das Previsões

Medidas de Erro

As medidas que avaliam a distribuição do erro, ou também chamadas medidas de acurácia, como o Erro Médio Absoluto (do inglês *Mean Absolute Error - MAE*) e o Erro Quadrático Médio (EQM, do inglês *Mean Square Error - MSE*), são amplamente utilizadas na literatura para verificação da performance de modelos de previsão (CHAI; DRAXLER, 2014). Tais medidas dependem da escala dos dados e são úteis quando se compara métodos de previsão aplicados a dados de mesma ordem de grandeza (KIM; KIM, 2016).

Para avaliação do desempenho das previsões utilizou-se o *MAE*, medida de desempenho que permite mensurar a magnitude média dos erros absolutos em um conjunto de valores estimados, logo quanto maior a precisão do modelo menor o erro médio absoluto, dado por

$$MAE = \frac{1}{n} \times \sum_{t=1}^{n} |e_t|$$

Foi também aplicado para análise de erro, o Erro Quadrático Médio (EQM), que quando na comparação de modelos permite identificar os mais eficazes de acordo com a média dos desvios ao quadrado, e pode ser calculado de acordo com a seguinte expressão

$$EQM = \frac{1}{n} \times \sum_{t=1}^{n} e_t^2$$

onde $e^t = Y_t - \hat{T}_t$ é o erro (diferença entre os valores da série Y_t, e os valores previstos por um modelo de tendência $\hat{T}_t$ em um período genérico t).

Tanto o MAE quanto o EQM recorrem a algum artifício matemático para tornar os desvios positivos, já que a utilização do erro médio poderia levar a valores próximos de zero simplesmente pela ocorrência de erros positivos e negativos que acabam se anulando, porém, o erro quadrático médio é mais sensível à presença de *outliers*, já que os erros estão ao quadrado.

Critérios de Informação de Akaike e Bayesiano

Os critérios de seleção de modelos de informação têm por objetivo apresentar a expressão que melhor descreve as observações sob análise, isto é, comparar a performance dos modelos quanto ao viés produzido por eles (EMILIANO et al, 2010).

Akaike (1974) define o critério de informação estabelecendo que o viés é obtido assintoticamente pelo número de parâmetros ajustados, sendo definido como:

$$AIC(\theta) = -2\ log\ (máximo\ da\ função\ de\ verossimilhança) + 2k_j$$

em que k_j: é a dimensão do modelo, ou número de parâmetros estimados; θ: é um vetor de parâmetros.

O critério de informação Bayesiano, proposto por Schwarz (1978), considera o comportamento assintótico dos estimadores de Bayes sobre uma classe especial de distribuições *a priori*.

Para cada modelo *j* tem-se

$$BIC = log(máximo\ da\ função\ de\ verossimilhança) - \frac{1}{2}k_j log(n)$$

em que k_j: é a dimensão do modelo, ou número de parâmetros estimados; *n*: é o número de observações da amostra.

Qualitativamente ambos os modelos de Akaike e Bayesiano dão uma formulação matemática do princípio da parcimônia no modelo construído, ou seja, evitar a inclusão de muitos parâmetros, dando preferência a modelos mais simples como critério de desempate. Quantitativamente, o critério Bayesiano difere do Critério de Akaike apenas na dimensão que é multiplicada *(½)log(n)*, portanto inclina-se mais ainda aos modelos de menor dimensão. Para grande número de observações os modelos diferem nitidamente um do outro, tendo o Bayesiano melhor desempenho (SCHWARZ, 1978).

Resultados e Discussão

A incidência média de leptospirose em Recife no período estudado (2007-2016) foi de 8,01 casos por 100 mil habitantes, sendo os meses de maio, junho e julho aqueles com maior incidência média mensal. O índice pluviométrico médio para o período foi de 164,64 mm e os meses de maiores quantidades de chuva coincidem com os meses de maiores incidências da doença. Numa análise anual, é possível observar que o ano de 2011 apresentou os maiores valores de ambas variáveis.

A taxa de leptospirose e os níveis pluviais apresentaram correlação de Pearson positiva significativa a nível de 5% (0,67, p-valor < 2.2e-16), portanto é possível supor que há uma associação entre o aumento do número de casos de leptospirose e o aumento da quantidade de chuva.

A variação mensal da incidência de leptospirose e do índice pluviométrico pode ser visualizada na Figura 1, que mostra a ausência de tendência em ambas as séries no período de análise, e a presença de padrões sazonais, ou seja, comportamentos que se repetem ao longo do tempo sempre no mesmo período do ano, o que pode ser observado pelo pico da série que se repete nos meses de maio, junho e julho. Esses resultados iniciais permitiram identificar e utilizar os padrões propostos para as análises que se sucedem.

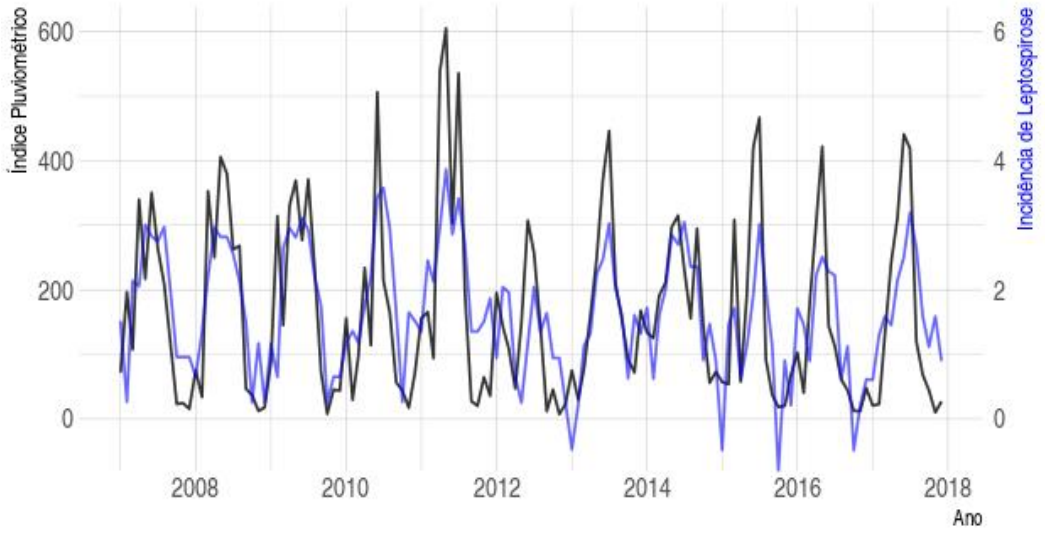

Figura 1 - Comportamento da incidência de leptospirose e do índice pluviométrico em Recife no período de 2007 a 2016.
Fonte: Autores.

Num primeiro momento, a distribuição sazonal da incidência de leptospirose foi estimada pelo modelo SARIMA(1,0,0)(1,1,1), nomeado de *M1*, em que o número de parâmetros foi escolhido com base nos critérios de informação de Akaike e Bayesiano, que apontaram melhor performance em comparação aos demais modelos SARIMA construídos.

A análise dos resíduos para o modelo SARIMA não apresenta significativa inobservância às suposições de independência serial e homocedasticidade, ou seja, os resíduos se comportam como ruído branco, com média zero e variabilidade constante, como mostra a Figura 2. De acordo com o teste de Ljung-Box (maiores detalhes sobre o teste em MCELROY e MONSELL, 2014), o p-valor foi igual a 0,42, não havendo evidências para rejeitar a hipótese nula de que os resíduos são independentes e identicamente distribuídos. Esse modelo permitiu captar o comportamento sazonal com base nas observações passadas da série de leptospirose.

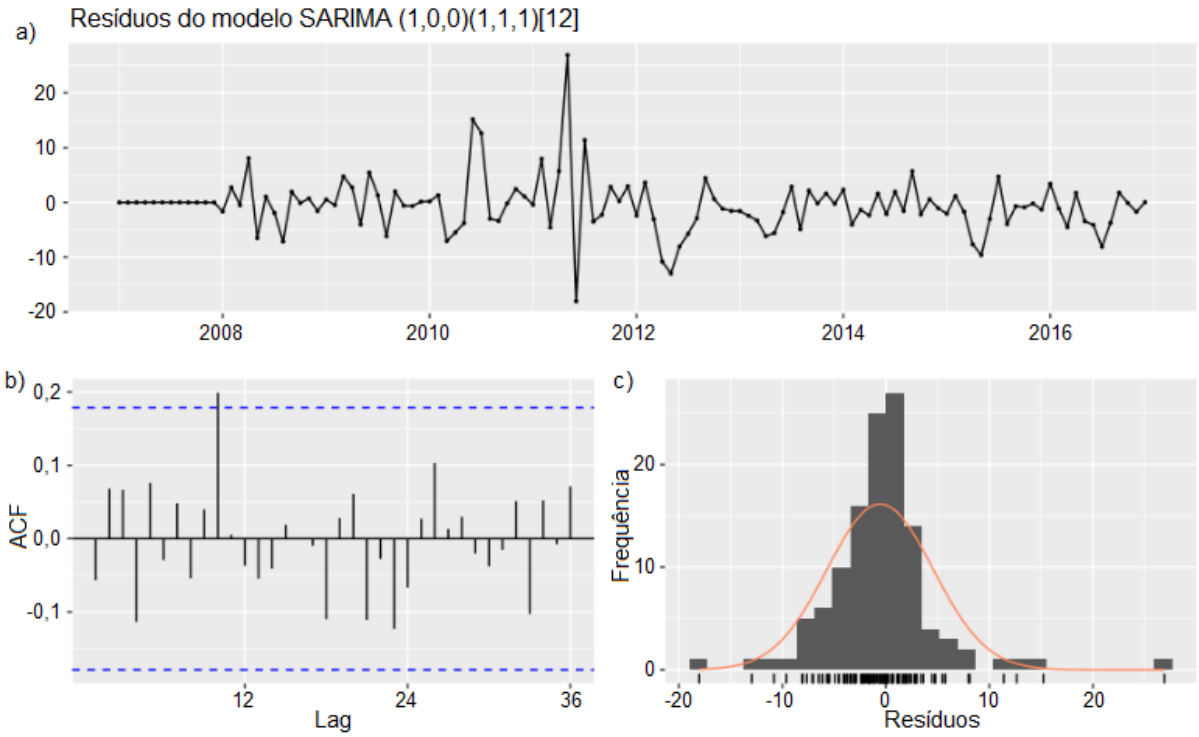

Figura 2: Análise de diagnóstico do Modelo Temporal Sazonal - M_1. (a) Resíduos vs ordem da observação, (b) gráfico de auto correlação, (c) histograma dos resíduos.
Fonte: Autores.

Com o modelo linear dinâmico foi possível, através da regressão, obter ajustes com observações variando no tempo. Assim, foi obtido o modelo *M2* para a variável índice de leptospirose utilizando além da série defasada, também o índice pluviométrico como variável explicativa. Este modelo apresentou alguns pontos aberrantes correspondentes aos meses de junho e julho de 2010 e o mês de maio de 2011, para os quais as incidências se afastaram do padrão estabelecido, como é possível observar pela análise dos resíduos do Figura 3a e 3b.

A presença de pontos discrepantes levou a um desvio das suposições de normalidade e aleatoriedade dos erros do modelo proposto, o que poderia prejudicar o desempenho do modelo para a previsão da incidência da leptospirose. No entanto, foi possível verificar que, com a inclusão do índice pluviométrico para explicar a incidência da doença, o modelo dinâmico captou o comportamento sazonal da série de modo aproximado e, ao utilizar o ano de 2017 para fazer previsões, a estimativa do modelo antecipou em dois meses a maior incidência da doença em comparação aos valores reais, como mostra a Figura 4.

Numa terceira proposta, a modelagem da incidência de leptospirose ocorreu por meio da regressão beta. Assim como no trabalho de Filho (2017), assumindo que a série não apresenta uma variação muito grande decorrente de tendência, sugerindo que a componente sazonal explica a maior parte da variabilidade e esta pode ser bem descrita pela pluviosidade, foi feito um ajuste ignorando a relação temporal a fim de verificar se um modelo usando apenas pluviosidade fornece melhores resultados na descrição da incidência. Para tanto, aplicou-se o modelo de regressão beta à taxa de incidência visto que a variável de interesse está no domínio da distribuição. Ao analisar os resíduos do modelo da regressão beta verifica-

se que foram atendidas as suposições de normalidade e aleatoriedade dos resíduos do modelo, como mostra a Figura 3c e 3d.

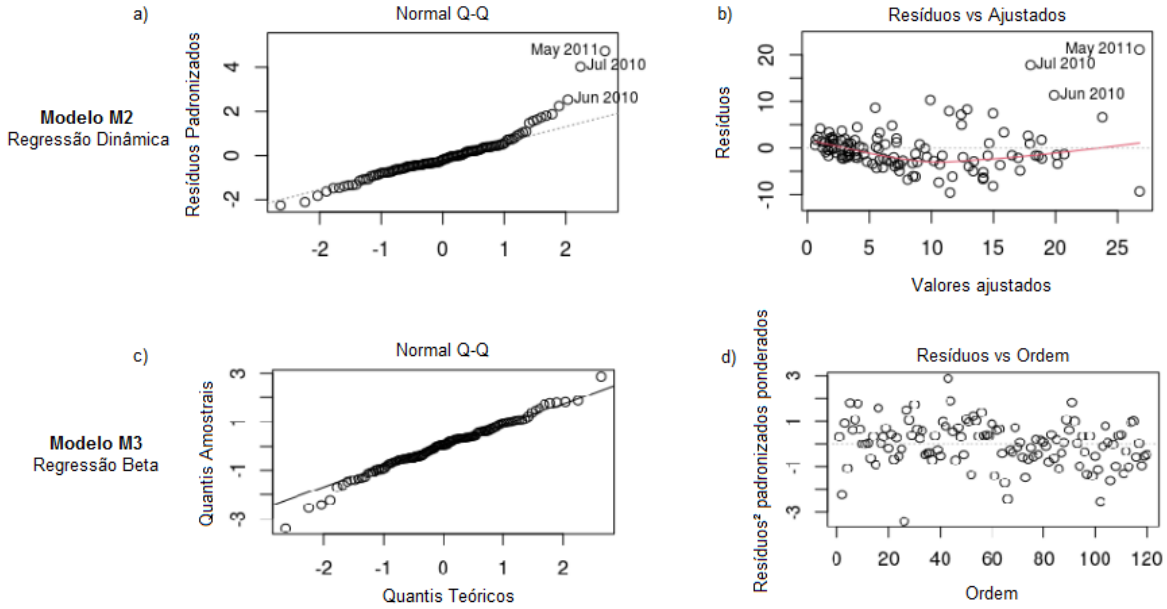

Figura 3 - Análise de diagnóstico dos modelos de Regressão Dinâmica (*M2*) e Regressão Beta (*M3*). (a) Q-Q plot do modelo *M2*, (b) resíduos vs valores ajustados para o modelo *M2*, (c) Q-Q plot do modelo *M3*, e (d) resíduos padronizados vs ordem para o modelo *M3*.
Fonte: Própria autoria

O resultado observado do modelo *M3* demonstra que a regressão beta conseguiu uma descrição semelhante aos ajustes anteriores, possivelmente porque, com exceção da componente sazonal, que é explicada indiretamente pelos dados de pluviosidade, a correlação temporal não foi essencial para o ajuste, de forma que a associação com o índice pluviométrico foi suficiente para explicar a maior parte da dispersão da série.

A Tabela 1 mostra algumas medidas de qualidade do ajuste para comparação dos três modelos. Dentre os modelos apresentados, percebe-se pelos critérios para análise de erro que o modelo *M1* apresenta menor erro quadrático médio de previsão, visto que o padrão sazonal foi bem ajustado, contudo este modelo não pôde aproximar-se das taxas mais altas da série original, dessa forma subestima os meses de maio, junho e julho.

Tabela 1 - Medidas de erro dos ajustes da série da incidência de leptospirose

Modelo	MAE	EQM
M1	3,27	10,03
M2	3,21	63,69
M3	3,45	18,12

Fonte: Autores.

O modelo *M2*, por sua vez, apresenta menor MAE em relação aos valores observados, isto é, devido a maior aproximação dos valores previstos com os valores originais. Apesar de

este modelo descrever o padrão sazonal da série original, ele não o descreve no mesmo período, com isso os picos de incidência que na série original se dão no segundo trimestre do ano, na previsão são antecipados em dois meses, o que explica o alto valor do erro quadrático médio, tal comportamento pode ser visualizado no Figura 4.

Com o modelo dado pela regressão beta, identifica-se que o mesmo apresenta os padrões similares aos modelos temporais e consegue prever a incidência com maior precisão que o modelo dinâmico.

Contudo, é importante mencionar que o modelo linear dinâmico foi o mais fortemente afetado pela presença de observações discrepantes e não faria sentido negligenciar esses pontos extremos, pois parte importante do objetivo do modelo era antecipar exatamente esses momentos mais críticos da série, porém é possível que essa metodologia tivesse um desempenho melhor na ausência dos pontos aberrantes, suposição que pode ser investigada em trabalhos futuros.

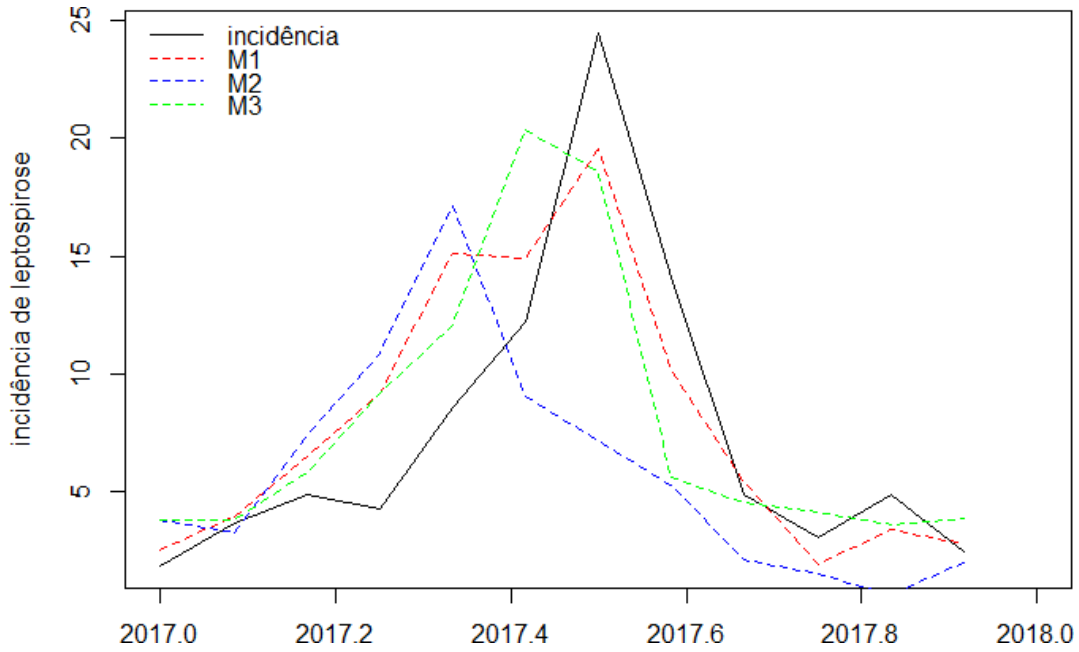

Figura 4 - Comparação dos valores reais e dos previstos pelos modelos *M1*, *M2* e *M3*
Fonte: própria autoria

Conclusões

Da comparação dos modelos propostos, foi possível observar que as previsões da incidência de leptospirose ajustada apenas com os valores passados, assim como o modelo de regressão beta com os índices pluviométricos como variável explicativa resultaram em melhores desempenhos que o modelo dinâmico incluindo a quantidade de chuva.

Levando em consideração que as previsões foram calculadas para um intervalo de doze meses no futuro e que, em um contexto mais realista, o modelo seria constantemente reajustado com os dados mais recentes, os resultados obtidos, principalmente pelos modelos *M1* e *M3* supracitados, foram satisfatórios tanto por fornecerem aproximações razoáveis para os valores reais quanto pela possibilidade de identificar quando a incidência deve aumentar, permitindo que medidas preventivas sejam tomadas a tempo.

Portanto, verifica-se que tanto a utilização de técnicas de séries temporais quanto a utilização da pluviosidade como variável explicativa permitem prever com antecipação a ocorrência da leptospirose.

Por fim, salienta-se que este trabalho, utilizando uma abordagem diferente ao considerar modelos dinâmicos e regressão beta no tratamento da incidência de leptospirose, corrobora o que a literatura apresenta em outras regiões.

Referências bibliográficas

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v. 19, n. 6, p. 716-723, 1974.

ALEIXO, N; NETO, J, Clima e saúde: diálogos geográficos, *Revista Geonorte*, Amazonas, V.8, n.30, p.78-103, 12 nov. 2017.

APAC. Meteorologia. *Base de dados do monitoramento pluviométrico*. Disponível em: http://www.apac.pe.gov.br/meteorologia. Acesso em: 26 fev. 2019.

BERNARDI, I. *Leptospirose e saneamento básico*. Florianópolis, SC: UFSC, 2012. Monografia (Especialização em Saúde Pública).

BOX. G.; JENKINS, G.; REINSEL, G. *Time Series analysis: forecasting and control*. 3rd ed. New Jersey: Prentice Hall, 1994.

BRASIL. SINAN - *Dados Epidemiológicos Sinan*. Acesso em: 26 fev. 2019. Disponível em: http://portalsinan.saude.gov.br/dados-epidemiologicos-sinan.

CHAI, T.; DRAXLER, R.R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, v. 7, n.3, p. 1247–1250, 2014.

COSTA, E. et al. Formas graves de leptospirose: aspectos clínicos, demográficos e ambientais. *Revista da Sociedade Brasileira de Medicina Tropical*, v. 34, n. 3, p. 261-267, 2001.

DATASUS. Ministério da Saúde. *Base de dados*. Acesso em: 26 fev. 2019. Disponível em: http://www2.datasus.gov.br/DATASUS/.

DIAS, E. D. M. *Previsão de Médio Prazo do Consumo de Energia Elétrica no Brasil*: Estimação Via Metodologia Box e Jenkins e Regressão Dinâmica. 2008. 109 f. Dissertação (Mestrado) - Curso de Pós-Graduação em Economia Aplicada, Universidade Federal de Juiz de Fora, Juiz de Fora, 2008.

EMILIANO, P.C.; VEIGA, E.P.; VIVANCO, M.J.F.; MENEZES, F.S. Critério de Informação de Akaike versus Bayesiano: Análise Comparativa, 2010. *In*: Anais do 19° Simpósio de Nacional de Probabilidade e Estatística – SINAPE. 2010.

FERRARI, S. L. P; CRIBARI-NETO, F. Beta regression for modelling rates and proportions, *Journal of Applied Statistics*, n.31, p. 799–815. 2004.

FILHO, J. *Tendência da incidência por leptospirose e a sua relação com os níveis pluviométricos na população do estado de Santa Catarina no período de 2005 a 2015*, Palhoça: UNISUL, 2017. Dissertação (Mestrado Ciências da saúde).

GONÇALVES, N. et al. Distribuição espaço-temporal da leptospirose e fatores de risco em Belém, Pará, Brasil. *Ciência & Saúde Coletiva*, v. 21, n. 12, p. 3947-3955, 2016.

GUIMARÃES, R. et al. Análise temporal da relação entre leptospirose e ocorrência de inundações por chuvas no município do Rio de Janeiro, Brasil, 2007-2012. *Ciência & Saúde Coletiva*, v. 19, n. 9, p. 3683-3692, 2014.

IBGE. *Estimativas da População* | IBGE. Acesso em: 26 fev. 2019. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?=&t=o-que-e.

KIM, S.; KIM, H. A new metric of absolute percentage error for intermittent demand forecasts *International Journal of Forecasting*, v. 32, n.3, p. 669-679, July–September 2016.

LOPES, J.; PIMENTA, L. *Previsão da Série de Preços Globais de Metanol Através dos Modelos Box & Jenkins e Regressão Dinâmica*. Dissertação (Mestrado). Universidade Federal de Juiz De Fora, 2011.

MAGALHÃES, G.; ZANELLA, M.; SALES, M. A ocorrência de chuvas e a incidência de leptospirose em Fortaleza-CE, *Hygeia: Revista Brasileira de Geografia Médica e da Saúde*, Uberlândia, V. 5, n. 9, p. 87-97, 02 fev. 2010.

MARTINS, K. *G. Expansão Urbana Desordenada e Aumento dos Riscos Ambientais à Saúde Humana*: Caso Brasileiro, Planaltina: Faculdade UnB Planaltina, Universidade de Brasília, 2012. Monografia (Bacharelado em Gestão Ambiental).

MCELROY, Tucker; MONSELL, Brian. The multiple testing problem for Box-Pierce statistics. *Electronic Journal Of Statistics*, Washington, v. 8, n. 1, p. 497-522, 2014. Institute of Mathematical Statistics. http://dx.doi.org/10.1214/14-ejs892.

MORETTIN, P.; TOLOI, C. *Análise de séries temporais*, 2. ed. São Paulo: Egard Blucher, 2006.

PEARSON, K. Notes on the history of correlation. *Biometrika*, v. 13, n. 1, p. 25-45, 1920.

PEARSON, K. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London*. Series A, Containing Papers of a Mathematical or Physical Character, v. 187, p. 253-318, 1896.

PELISSARI, D. et al. Revisão sistemática dos fatores associados à leptospirose no Brasil, 2000-2009. *Epidemiologia e Serviços de Saúde*, v. 20, n. 4, p. 565-574, 2011.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 2008. Disponível em: http://www.R-project.org/.

SCHWARZ, G. Estimating the Dimension of a Model. *The Annals of Statistics*, v.6 n.2, p. 461-464, 1978.

SILVA, M. O. Metodologia de Previsão de Séries Temporais- Box & Jenkins. *In*: Uma aplicação de árvores de decisão, redes neurais e knn para a identificação de modelos ARMA não-sazonais e sazonais, 2005, p.15- 4.

TASSINARI, W. et al. Distribuição espacial da leptospirose no Município do Rio de Janeiro, Brasil, ao longo dos anos de 1996-1999. *Cadernos de Saúde Pública*, v. 20, n. 6, p. 1721-1729, 2004.

ZANINI, A. *Redes Neurais e regressão dinâmica: um modelo híbrido para previsão de curto prazo da demanda de gasolina automotiva no Brasil*. Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro. Dissertação (Mestrado em Engenharia Elétrica: Teoria de Controle e Estatística). 2000.

ZEILEIS, A. et al. Monitoring structural change in dynamic econometric models. *Journal of Applied Econometrics*, v. 20, n. 1, p. 99-121, 2005.