Modelagem de série temporal de óbitos de residentes em Porto Alegre - RS para previsão de óbitos em 2020 utilizando a metodologia Box-Jenkins

Fabian Corrêa Cardoso[1]
Viviane Leite Dias de Mattos[2]
Andrea Cristina konrath[3]
Luiz Ricardo Nakamura[4]
Antonio Cezar Bornia[5]

Resumo

Neste estudo, apresenta-se a modelagem da quantidade de óbitos registrados de residentes em Porto Alegre (RS), a partir de dados obtidos na plataforma TabNet do DataSUS de 1996 até 2019, utilizando a metodologia de Box-Jenkins, com o objetivo de fazer uma estimativa da previsão para 2020. Para ajuste do modelo, foram utilizados dados do período de 1996 a 2016, sendo os restantes utilizados em sua validação. O modelo encontrado fornece medidas de acurácia de previsão muito boas, entre as quais um erro absoluto percentual médio (MAPE) inferior a 0,5%, além de apresentar todos os dados observados no período de validação dentro dos respectivos intervalos de confiança. Espera-se, entretanto, que o valor observado para 2020 extrapole o intervalo de confiança construído, superando o seu limite superior, que é de 12.739,87 óbitos, pois a série temporal que representa o número de óbitos deverá apresentar uma quebra estrutural neste ano em decorrência à pandemia da COVID-19.

Palavras-chave: Séries temporais; Metodologia *ARIMA*; COVID-19.

Abstract

In this paper we model the number of registered deaths of residents in Porto Alegre (RS), based on data from TabNet platform, which belongs to DataSUS, from 1996 up to 2019, through the Box-Jenkins methodology, in order to forecast these numbers for 2020. For the fitting process we used data from 1996 up to 2016, and the rest was used in the validation phase. The fitted model returned reasonable predictive accuracy measures, including a mean absolute percentage error (MAPE) smaller than 0.5%. Furthermore, all observed data on the validation phase were within their respectively confidence intervals. Nonetheless, it is noteworthy that the observed value for 2020 may exceed the fitted confidence

[1] Universidade do Rio verde – E-mail: fabian@unirv.edu.br
[2] Universidade Federal do Rio Grande
[3] Universidade Federal de Santa Catarina
[4] Universidade Federal de Santa Catarina
[5] Universidade Federal de Santa Catarina

interval, overcoming its upper limit given by 12,739.87 deaths, since it is expected that this time series present a structural break this year due to the COVID-19 pandemic

Keywords: Time series; *ARIMA* methodology; COVID-19.

1. Introdução

Uma série temporal pode ser definida como um conjunto de observações ordenadas sequencialmente ao longo do tempo, em que a ordem dos dados é fundamental para o seu estudo, uma vez que observações vizinhas são dependentes. Sua análise tem como finalidade identificar padrões, como tendência, sazonalidade, ciclos e mudanças de nível, na maioria das vezes com o objetivo de fazer previsões (GUJARATI; PORTER, 2011; MORETTIN; TOLOI, 2006), o que é de interesse para a gestão de organizações públicas e privadas.

Conforme Zhang et al. (2014), modelos de séries temporais aplicados na área da saúde costumam ter o objetivo de prever comportamentos epidemiológicos. Isto pode ser bastante útil, pois as previsões feitas a partir de uma série histórica podem ser fundamentais para um bom planejamento na área da saúde. Nesta área, várias previsões fundamentadas em dados históricos já foram realizadas com sucesso, utilizando métodos de modelagem para séries temporais, estatísticas, matemáticas, computacionais ou híbridas. Nos estudos que empregaram métodos estatísticos, nos quais têm sido utilizados diferentes métodos e técnicas, assumem importante papel os modelos Autorregressivos Integrados de Médias Móveis (*ARIMA*).

Earnest et al. (2005) os utilizaram como ferramenta no monitoramento e previsão do número de leitos ocupados, durante uma epidemia de Síndrome Respiratória Severa Aguda (SARS), obtendo previsões com erro absoluto percentual médio (MAPE) de 8,7%, enquanto Akhtar e Rozi (2009) obtiveram um MAPE de 6,5% em um horizonte de seis meses para previsões da prevalência de soro positivo para o vírus da Hepatite C (HCV) em doadores de sangue do sexo masculino. Já Nunes (2018) realizou um estudo para comparar o desempenho dos modelos Autorregressivos de Médias Móveis (ARMA) e

Autorregressivos de Médias Móveis Generalizados (GARMA) na modelagem de óbitos por neoplasias malignas de pele. Entretanto, neste caso, o modelo *GARMA*-Poisson (MAPE=3,53%) apresentou melhores estimativas e previsões quando comparado ao modelo *ARIMA* (MAPE=7,13%). Barros (2019) analisou dados de casos notificados de dengue obtidos junto ao Centro de Controle de Zoonoses (CCZ), utilizando os métodos de suavização exponencial simples, de Holt e de Holt Winters, bem como o modelo Autorregressivo Integrado de Médias Móveis com entradas Exógenas (ARIMAX) com variáveis exógenas climáticas. Embora os modelos ARIMA tenham apresentado menores erros de previsão, a autora considerou que nenhum modelo foi suficientemente bom para fazer previsões confiáveis, o que pode ter ocorrido pelo fato de a variável ter apresentado múltiplos padrões.

Atualmente, a pandemia de COVID-19 está alarmando o mundo. De acordo com Cascella et al. (2020 apud ALZAHRANI; ALJAMAAN; AL-FAKIH, 2020), este é o sétimo Coronavírus a infestar humanos. Os quatro primeiros não tiveram muita importância, o quinto e o sexto desenvolveram sintomas graves, enquanto o sétimo – COVID-19 – causou uma pandemia. Países que pensavam estar com a doença sob controle, atualmente enfrentam a segunda onda. Diante deste cenário, diversos pesquisadores têm feito uso de métodos de modelagem de séries temporais para, a partir de dados históricos, tentar encontrar algum padrão de comportamento nas diversas variáveis que estão sendo utilizadas, para avaliar não só o comportamento do vírus, como também suas consequências humanas, sociais e econômicas. Dentre as diversas metodologias utilizadas, novamente sobressai-se a modelagem *ARIMA*.

Alzahrani, Aljamaan e Al-Fakih (2020) fizeram uso desta metodologia para prever o número diário de novos casos de COVID-19 na Arábia Saudita para um horizonte de previsão de quatro semanas, obtendo previsões com MAPE de 2,16%. Yanga et al. (2020) empregaram a metodologia ARIMA para modelar a ocorrência diária de novos casos e novos óbitos por COVID-19 em Hubei-China, avaliando os modelos encontrados pelo erro absoluto médio (MAE), que foram considerados satisfatórios. Estes modelos foram usados para prever o comportamento destas duas variáveis na Itália, após um lockdown, produzindo

intervalos de confiança que continham os valores observados. Kirbas et al. (2020) modelaram a quantidade diária de casos de COVID-19 em oito países utilizando modelagem *ARIMA*, NARNN (Redes Neurais Autorregressivas Não Lineares) e LSTM (Memória de Curto e Longo Prazo), fazendo previsões para um horizonte de 14 dias. Os MAPEs para LSTM, *ARIMA* e NARNN ficaram nos intervalos de 0,16-2,55%, 0,34-5,46% e 0,27-7,95%, respectivamente. Neste estudo, os modelos *ARIMA* foram superados pelos modelos LSTM. Benvenuto et al. (2020) ajustaram um modelo *ARIMA* aos dados diários de prevalência e incidência de COVID-19 a partir de dados coletados no site oficial da instituição John Hopkins. Os autores ressaltam que, para comparação posterior ou previsões futuras, a definição de caso e a coleta de dados devem ser mantidas em tempo real. Já Tran, Pham e Ngo (2020), desenvolveram modelos para previsão do total de casos diários confirmados, total de casos confirmados, novos casos, total de óbitos, total de novos óbitos, taxa de crescimento em casos confirmados e taxa de crescimento de óbitos em um país a partir de dados diários SARS-CoV-2, os quais foram coletados do site oficial do Centro Europeu de Prevenção de Doenças e Controle. Os autores concluíram que o modelo *ARIMA* é uma modelagem fácil de ser utilizada e adequada para realizar previsões na propagação de SARS-CoV-2. Bayyurt e Bayyurt (2020) realizaram um estudo para prever a disseminação da COVID-19 em três países. Foi utilizado o modelo *ARIMA* em dados obtidos junto ao Centro Europeu de Prevenção e Controle de Doenças para prever o número de casos confirmados e número de óbitos de COVID-19. Os autores ressaltam que os modelos encontrados podem ser utilizados para fazer previsões futuras, uma vez que o valor MAPE do modelo encontrado foi inferior a 10%.

No Brasil, a pandemia causada pela COVID-19 provocou um grande número de casos, superior a quatro milhões, assim como um número de mortes significativo, superior a cem mil óbitos, conforme dados do Ministério da Saúde do Brasil (BRASIL, 2020). Entretanto, existem algumas dúvidas em relação à exatidão das taxas de mortalidade (quociente entre o número de indivíduos que morrem por uma causa específica e o número total de indivíduos na população) e de letalidade (quociente entre o número de indivíduos que morrem por causa

de uma doença e o número de indivíduos infectados por esta mesma doença) oficialmente divulgados, não apenas em função da falta de testagem da população de uma maneira geral, como também pelos critérios adotados para identificar a COVID-19 como causa do óbito (exames de laboratório e/ou análise clínica).

Observando este cenário, questiona-se: será possível utilizar o número total de óbitos de uma região para avaliar a quantidade de óbitos decorrentes da pandemia por COVID-19? Parte-se do princípio de que um bom modelo permitiria a construção de um intervalo de confiança para a previsão de óbitos no ano de 2020 que contivesse o número de óbitos para este ano. Neste caso, os valores excedentes poderiam ser atribuídos ao “efeito COVID-19”.

Na literatura, sabe-se que a previsão fundamentada no Método Estatístico está bastante consolidada e que, com base no estudo de séries temporais, identificando um padrão de comportamento na evolução do número anual de óbitos é possível fazer a previsão por meio do modelo (*ARMA*), com uma taxa de acerto bem elevada. Assim, o presente estudo tem como objetivo realizar a modelagem da série temporal de óbitos de indivíduos residentes em Porto Alegre, Rio Grande do Sul (RS), com a sua posterior previsão para o ano de 2020. A escolha de Porto Alegre se deu em função deste município ser aquele com maior número de óbitos no Rio Grande do Sul, estado localizado no extremo sul do Brasil, com temperaturas muito baixas nos meses de inverno, o que pode ser um fator de risco para incidência da referida doença.

2. Referencial teórico

2.1 Metodologia *ARMA*

A metodologia de Modelos Autoregressivos de Médias Móveis (*ARMA*), segundo Morettin e Toloi (2006), descreve três classes de processos:

a) Processos lineares estacionários;
b) Processos lineares não-estacionários homogêneos; e,
c) Processos de memória longa.

Ainda segundo os mesmos autores, dentro dos processos lineares estacionários existem três modelos: processo regressivo de ordem *p* (*AR*(*p*)); processo de médias móveis de ordem *q* (*MA*(*q*)); e a união dos dois, de ordem *p* e *q* (*ARMA*(*p,q*)), dados respectivamente por

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad (1)$$

$$y_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (2)$$

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_0 + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (3)$$

em que y_t representa a variável analisada no *t-ésimo* tempo (*t* = 1, 2, 3,..., *T*); *T* é a amplitude da série analisada; *p* é a ordem do modelo AR; ϕ_j é o coeficiente do *j-ésimo* termo do modelo *AR* (*j*=0,1,2,...,*p*), *q* é a ordem do modelo *MA;* θ_l é o coeficiente do *j-ésimo* termo do modelo *MA* (*l*=0,1,2,...,*q*); e ε_t é o erro aleatório do *t-ésimo* termo ($\varepsilon_t RB(0, \sigma^2)$).

Segundo Morettin e Toloi (2006), para se usar um dos modelos supracitados na modelagem de uma série temporal são necessários três estágios: identificação, estimação e diagnóstico. Ainda segundo o mesmo autor, o procedimento de identificação deve ser executado em três etapas.

Inicialmente, verifica-se a necessidade de uma transformação de dados na série original, com o objetivo de estabilizar sua variância. Nesta etapa, é bastante usual utilizar a transformação logarítmica, ou alguma outra função da família de transformações Box-Cox, com o objetivo de satisfazer a condição de homogeneidade de variâncias.

Após, verifica-se se a série pode ser considerada estacionária. Para utilização destes modelos, deve-se garantir que a série seja estacionária, apresentando propriedades que não dependam do tempo (HYNDMAN; ATHANASOPOULOS, 2018). Segundo Morettin e Toloi (2006), neste caso, a média e a variância devem ser constantes, e a covariância deve depender apenas da defasagem. Vários testes podem ser utilizados com esta finalidade, entre os quais o teste de Dickey-Fuller Aumentado (ADF) é um dos mais

utilizados na prática. Este teste considera a presença de raiz unitária como hipótese nula. Suas extensões, o teste Phillipe-Perron (PP) e o teste Dickey-Fuller-GLS, entre outras, também são bastante utilizadas.

Não havendo estacionariedade, diferenciações podem ser utilizadas para tornar a série estacionária. Neste caso faz-se uso do modelo Autoregressivo Integrado de Médias Móveis (*ARIMA*).

Finalizando esta primeira etapa, deve-se identificar o processo *ARMA*(*p,q*) ou *ARIMA*(*p,d,q*). Esta identificação baseia-se na Função de Autocorrelação (FAC) e na Função de Autocorrelação Parcial (FACP). Segundo Hyndman e Athanasopoulos (2018), a Função de Autocorrelação mede a relação linear entre os valores defasados de uma série temporal. Já as Autocorrelações Parciais medem a relação entre e depois de remover os efeitos das defasagens. Considerando a subjetividade de uma interpretação gráfica, é bastante usual serem identificados alguns modelos candidatos, definidos a partir da análise gráfica da FAC e FACP.

Após esta identificação, pode-se iniciar a estimação dos parâmetros (segunda etapa) o que normalmente é realizado com o Método de Máxima Verossimilhança. Alguns algoritmos podem ser utilizados para otimizar a sua realização, entre os quais o de Nelder- Mead.

Esta etapa é complementada com a hierarquização dos modelos candidatos, o que é feito pelos critérios de informação, entre os quais estão o de Akaike (AIC, AKAIKE, 1976), Akaike corrigido (AICc, BOZDOGAN,1987) e o Bayesiano (BIC, SCHWARZ, 1978), sendo estes normalmente os mais utilizados. Busca-se o modelo mais parcimonioso, minimizando os valores de *p* e *q*, ou seja, avalia-se a precisão dos resultados penalizando-a com a quantidade de coeficientes que devem ser estimados.

Com os modelos hierarquizados, inicia-se a terceira etapa, denominada de diagnóstico, com o objetivo de verificar se é possível considerar que o modelo encontrado representa os dados analisados, o que pode ser feito por meio de uma análise dos resíduos. Parte-se do princípio de que os resíduos devem se comportar como um ruído branco para que o modelo seja válido, ou seja, devem ser independentes. Para tal, pode ser utilizado o teste proposto em Ljung e Box

(1978), que verifica a hipótese nula de que os resíduos são independentes, utilizando a estatística de teste *Q* que apresenta distribuição *Qui*-quadrado. Além disso, é desejável que os resíduos apresentem distribuição normal, o que pode ser verificado pelo teste proposto em Jarque e Bera (1987), que a avalia a partir das propriedades de assimetria e curtose. Também é desejável que os resíduos sejam homocedásticos, o que ser verificado pelo teste ARCH (A*utoregressive Conditional Heteroskedasticity*), proposto em Engle (1982).

Quando o objetivo do estudo é fazer previsões, normalmente são selecionados os modelos considerados melhores no ajuste aos dados para realizar sua validação para previsões que, de acordo com Hyndman e Athanasopoulos (2018), consiste no ato de prever eventos futuros com o uso de alguns recursos e informações passadas que possam influenciar os resultados. A identificação do melhor modelo para previsão é feita pela mensuração de sua acurácia, que usa como critério de avaliação: a raiz quadrada do erro quadrático médio (RMSE - *Root Mean Squared Error*), o erro absoluto médio (MAE - *Mean Absolute Error*) e o erro absoluto percentual médio (MAPE - *Mean Absolute Percentage Error*), entre outros

Mais detalhes sobre os métodos e técnicas citados na realização desta modelagem podem ser encontrados em Bueno (2011); Gujarati e Porter (2011); Hyndman e Athanasopoulos (2018); Morettin e Toloi (2006), entre outros.

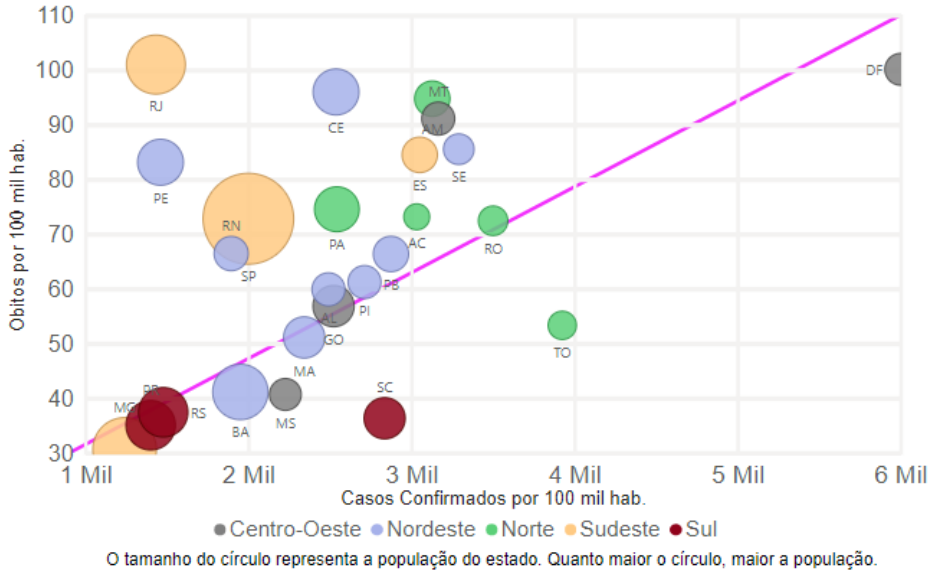
2.2 A COVID- 2019

De acordo com Organização Mundial de Saúde (WHO, 2020), a síndrome respiratória oficialmente denominada COVID-19 foi identificada na cidade de Whuan, na China, em dezembro de 2019, sendo o novo tipo de Coronavírus isolado nos primeiros dias de janeiro de 2020. Sua sequência genética foi logo compartilhada pelos chineses, embora não tivessem encontrado evidências de que ele passasse de pessoa para pessoa. Além disso, oficialmente, entre as 41 pessoas identificadas inicialmente com a doença havia ocorrido apenas um óbito.

Segundo Centro Estadual de Saúde do Rio Grande do Sul (CEVS-RS, 2020a), no final de janeiro de 2020, entretanto, a Organização Mundial da Saúde (OMS) alertou a população mundial para o perigo de uma nova síndrome respiratória grave, modificando a categoria de risco de transmissão desta síndrome para Emergência de Saúde Pública de Importância Internacional. A partir de então, a doença se espalhou rapidamente, gerando uma pandemia. Ainda de acordo com CEVS-RS (2020), o primeiro caso de COVID-19 no Brasil foi confirmado em 26 de fevereiro de 2020 na cidade de São Paulo, enquanto o primeiro caso no Rio Grande do Sul foi registrado em 29 de fevereiro 2020.

Atualmente, de acordo com o Ministério da Saúde (BRASIL, 2020), no Brasil, até o dia 21 de setembro 2020 havia 4.558.068 casos confirmados, o que representa uma incidência de 2.169 casos a cada 100.000 habitantes. Destes, 3.887.199 foram recuperados, 533.597 ainda se encontram em acompanhamento e 137.272 tiveram como desfecho o óbito, levando a uma taxa de mortalidade de 65,3 a cada 100.000 habitantes e a uma taxa de letalidade aparente de 3,0%. Já no Rio Grande do Sul, de acordo com Secretaria de Saúde do estado (SES-RS, 2020), até 22 de setembro de 2020, havia 177.485 casos confirmados, o que representa uma incidência de 1.560 casos a cada 100.000 habitantes. Destes, 162.695 foram recuperados, 10.318 ainda se encontram em acompanhamento e 4.472 tiveram como desfecho o óbito, levando a uma taxa de mortalidade de 39,3 a cada 100.000 habitantes e a uma taxa de letalidade aparente de 2,5%. Observa-se então que o estado do Rio Grande do Sul apresenta indicadores melhores que as médias do Brasil, quando são considerados valores relativos: casos confirmados por 100.000 habitantes e óbitos por 100.000 habitantes, assim como também menor taxa de letalidade aparente, conforme evidenciado no Figura 1. Observe que os estados da região Sul, principalmente Rio Grande do Sul e Paraná, apresentam menores taxas de óbitos e casos confirmados, enquanto os da região Sudeste, com exceção de Minas Gerais, têm as maiores taxas.

Figura 1. Óbitos e casos confirmados a cada 100.000 habitantes no Brasil em setembro de 2020



Nota: População estimada em julho de 2019
Fonte: Ministério da Saúde (BRASIL, 2020) e Secretaria Estadual de Saúde do Rio Grande do Sul (SES-RS, 2020b).

De acordo com SES-RS (2020b), a partir do mês de junho, ocorreu um aumento significativo na densidade da incidência de hospitalizações e óbitos por COVID-19 no Rio Grande do Sul, com estabilização a partir da semana epidemiológica 30, sendo que a região COVID-19 de Porto Alegre foi uma das mais afetadas. Também foi detectado que surtos localizados tiveram impacto na propagação da doença nesta região. Ainda, neste estado, de acordo com SES-RS (2020b), idosos têm maior risco de hospitalização (taxa 5,6 vezes maior) e maior risco de óbito (taxa 18,5 vezes maior), e que a maior parte dos hospitalizados e dos que vieram a óbito apresentavam comorbidade pregressa.

Em Porto Alegre, de acordo SES-RS (2020a), até 22 de setembro de 2020, havia 24.616 casos confirmados, o que representa uma incidência de 1.659 casos a cada 100.000 habitantes. Destes, 946 tiveram o óbito como desfecho, levando a uma taxa de mortalidade de 63,8 a cada 100.000 habitantes.

3. Material e métodos

A confiabilidade dos dados analisados é muito importante para qualquer tipo de análise. No presente estudo, os dados relativos à quantidade de óbitos no período 1996-2018 foram obtidos nas estatísticas vitais disponíveis no Portal da Saúde do Ministério da Saúde (DATASUS, 2020a). A informação relativa ao ano de 2019 foi obtida na seção de dados preliminares do mesmo portal (DATASUS, 2020b). Os dados foram selecionados em ambas as bases de dados da seguinte forma:

i. Opção - Município;
ii. Óbitos - Por residência;
iii. Período - 1996-2019;
iv. Cidade desejada - Porto Alegre.

Para selecionar os dados nestes endereços, é possível escolher entre óbitos por residência ou por local de morte. Optou-se por óbitos por residência, porque foi a metodologia adotada pelo Governo Federal para registros de mortes por COVID-19, isto é, a morte foi contabilizada para o município no qual a pessoa de fato residia e não no qual faleceu pela doença.

A modelagem foi feita com dados relativos ao período 1996-2016, de forma que a validação foi feita para dados entre 2017 e 2019. Os procedimentos descritos foram feitos utilizando-se os pacotes *forecast* (HYNDMAN et al., 2020); *urca* (PFAFF, 2008); *tseries (*TRAPLETTI; HORNIK, 2019*);* rugarch *(*GHALANOS ,2020*), dentre outros do software* R (R Core Team, 2020). Para as análises de inferência, foram considerados os níveis de significância 0,01 e 0,05, e as estimações por intervalo com nível de confiança de 0,95.

A aplicação da metodologia *ARMA* seguiu o protocolo usualmente utilizado. Considerando os dados do período de ajuste, inicialmente foi realizada uma análise exploratória com a construção de alguns gráficos (gráfico em linhas, histograma e *box plot*) e cálculo de algumas medidas descritivas que representassem tendência central, dispersão, assimetria e curtose.

Após, a estacionariedade foi avaliada pelo teste ADF. Este teste foi executado de forma iterativa a partir de um número máximo de defasagens,

considerando o critério proposto em Hyndman e Athanasopoulus (2018). Foram consideradas três situações: modelo com constante e tendência, modelo apenas com constante e modelo sem constante e sem tendência. Na constatação de não-estacionariedade seriam realizadas diferenças sucessivas com repetição dos testes, até consegui-la. A análise gráfica das funções de autocorrelação e autocorrelação parcial permitiram atribuir valores para *p* e *q*, a partir dos quais foram identificados alguns modelos candidatos, somando e subtraindo a estes uma unidade $(p \pm 1;\ q \pm 1)$, sendo consideradas todas as combinações possíveis.

Após, foi feita a modelagem com a determinação dos seus diversos coeficientes e respectiva significância, utilizando o método de máxima verossimilhança. Estes modelos foram então hierarquizados por sua parcimonialidade de acordo com os critérios de informação de Akaike, de Akaike corrigido e de Schwarz.

Seguiu-se o diagnóstico feito pela análise das suposições do modelo teórico, que foi executada diretamente nos resíduos: teste de Ljung-Box para avaliar autocorrelação, teste de Jarque-Bera para avaliar normalidade e teste ARCH, para avaliar homocedasticidade.

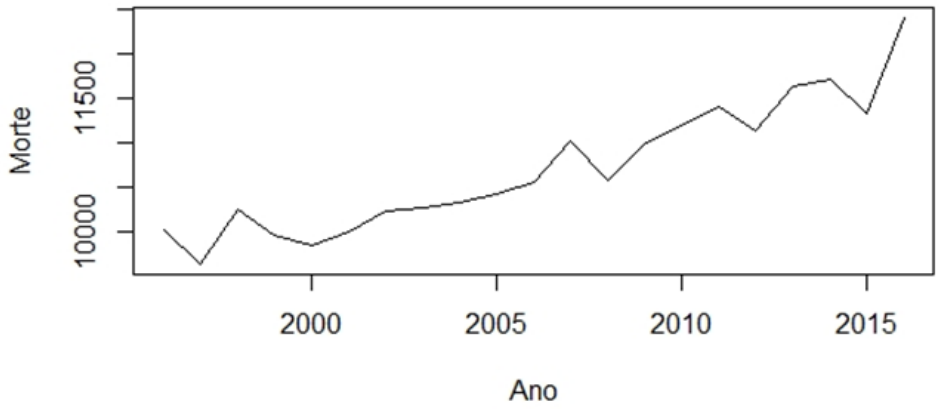
Finalizando, foi realizada a validação para o período 2017-2019 dos modelos para previsão, com a determinação dos indicadores: erro absoluto médio (MAE); raiz quadrada do erro quadrático médio (RMSE); e erro absoluto percentual médio (MAPE), com identificação dos melhores modelos para previsão. Complementaram a análise a construção de intervalos de confiança para os valores previstos.

4. Resultados e discussões

Começou-se pela análise exploratória dos dados para caracterizar o conjunto analisado: a quantidade de óbitos em Porto Alegre (RS) no período entre 1996-2016, apresentados na Figura 2. Eles apresentam uma tendência crescente, embora tenha havido queda em alguns anos, e variaram de 9.647 a 12.402, concentrando-se em torno da média 10.714,52, com desvio-padrão de

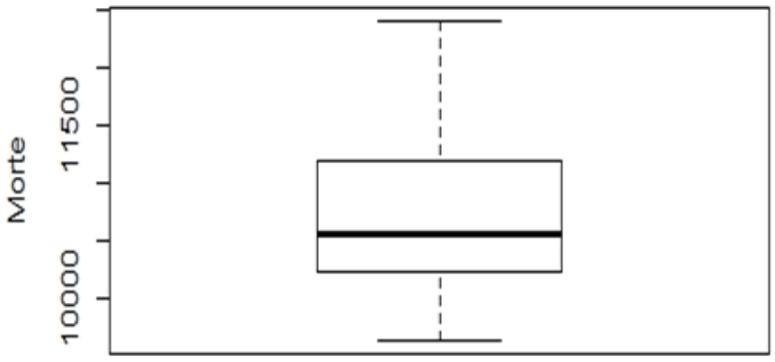
722,23. Sua mediana foi de 10.561. Sua distribuição caracteriza-se como assimétrica positiva, pois o coeficiente assume o valor 0,5469 > 0,50, indicando uma maior concentração dos valores mais baixos, conforme evidenciado pelo *box plot* mostrado na Figura 3. Já o coeficiente de curtose assume o valor -0,4813, classificando-a como platicúrtica. O histograma pode ser visto na Figura 4 e evidencia a existência de 2 picos na distribuição: um entre 10.000 e 10.500 e outro entre 11.000 e 11.500.

Figura 2. Óbitos registrados em Porto Alegre, estado do Rio Grande do Sul, no período entre 1996-2016



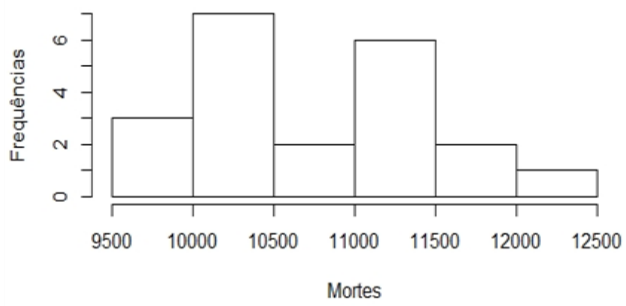
Fonte: Elaborado pelos autores a partir dos dados da pesquisa

Figura 3. Representação de gráfico *box plot* dos óbitos registrados em Porto Alegre, estado do Rio Grande do Sul, no período entre 1996-2016



Fonte: Elaborado pelos autores a partir dos dados da pesquisa

Figura 4. Histograma dos óbitos registrados em Porto Alegre, estado do Rio Grande do Sul, no período entre 1996-2016

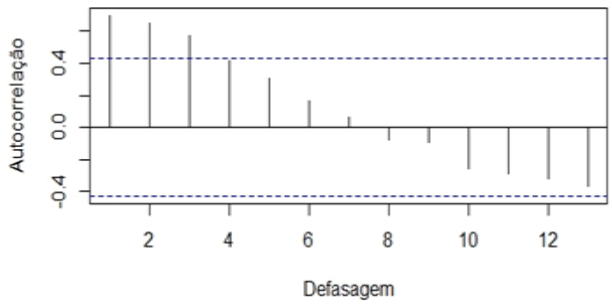


Fonte: Elaborado pelos autores a partir dos dados da pesquisa.

Para aplicação do teste ADF foi considerado o critério de Hyndman, que forneceu o valor 3 para a quantidade máxima de defasagens a ser considerada no início do processo iterativo. Para uma defasagem, a tendência ($\varphi_{calculado} = 17{,}04 < \varphi_{crítico} = 10{,}61$) e a constante ($\varphi_{calculado} = 17{,}49 < \varphi_{crítico} = 8{,}21$) foram significativas (valor p<0,01), ou seja, foram encontradas evidências de que a série possui tendência e é constante. Além disso, também foram encontradas evidências de que a série com tendência e constante é estacionária ($\tau_{calculado} = -5{,}48 < \tau_{crítico} = -4{,}38$).

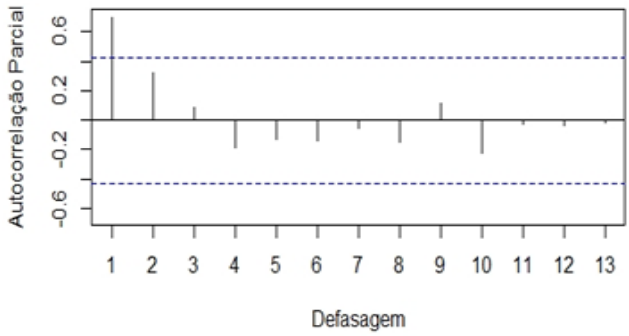
A função de autocorrelação (FAC) apresentada na Figura 5 mostra decaimento lento e de forma senoidal, enquanto a função de autocorrelação parcial (FACP), apresentada na Figura 6, mostra uma queda maior da primeira para a segunda defasagem. Todas demais não são significativas, caracterizando-se como processo *AR*(1) ou *ARMA*(1,0).

Figura 5. Função de autocorrelação



Fonte: Elaborado pelos autores a partir dos dados da pesquisa.

Figura 6. Função de autocorrelação parcial



Fonte: Elaborado pelos autores a partir dos dados da pesquisa.

Pode-se interpretar, então, que os modelos candidatos seriam *p*, variando entre 0 e 2, e *q*, variando entre -1 e 1. Desta forma, foram utilizados 6 modelos candidatos: *ARMA*(0,0); *ARMA*(1,0); *ARMA*(0,1); *ARMA*(1,1); *ARMA*(0,2); e *ARMA*(1,2).

Pela hierarquização dos modelos a partir dos critérios de informação AIC, AICc e BIC, mostrados na Tabela 1, considerou-se que o mais parcimonioso foi o modelo *ARMA*(0,0) com constante e tendência, seguido por *ARMA*(0,1) com constante e tendência, e pelo modelo *ARMA*(1,0), também com constante e tendência.

Tabela 1. Resultados dos critérios de informação nos modelos candidatos *ARMA*(*p,q*)

Modelo	Tendência	Constante	Critérios de informação		
			AIC	AICc	BIC
(0,0)	Com	Com	296,76	298,17	299,89
(0,1)	Com	Com	295,84	298,34	300,01
(1,0)	Com	Com	296,85	299,35	301,02
(1,1)	Com	Com	297.77	301.77	301.77
(2,0)	Com	Com	297.68	301.68	301.68
(2,1)	Com	Com	299.27	305.27	305.53

Fonte: Elaborado pelos autores a partir dos dados da pesquisa.

A análise de resíduos dos modelos selecionados na etapa anterior está apresentada na Tabela 2. Os resultados dos testes Ljung-Box, Jarque-Bera e ARCH, demonstram que os modelos selecionados cumprem todas as suposições do modelo teórico, a um nível de significância de 0,01, pois não há evidências de autocorrelação, não normalidade e nem heterocedasticidade.

Tabela 2. Resultados dos testes de Ljung-Box, ARCH e Jarque-Bera para os resíduos dos modelos candidatos

Modelo	Ljung-Box		Jarque-Bera		ARCH	
	Estatística	Valor *p*	Estatística	Valor *p*	Estatística	Valor *p*
(0,0)	5,9874	0,6486	1,8398	0,3986	7,6981	0,4635
(0,1)	5.803	0.5629	1.8507	0.3964	4,9936	0,7583
(1,0)	5.4217	0.6086	1.8552	0.3955	6,9065	0,5468

Fonte: Elaborado pelos autores a partir dos dados da pesquisa.

Na validação dos modelos candidatos, foi determinada a sua capacidade preditiva, encontrando-se os resultados apresentados na Tabela 3. A melhor capacidade preditiva é a do modelo *ARMA*(0,1) por ter apresentado os menores indicadores RMSE, MAE e MAPE, seguido pelo modelo *ARMA*(0,0).

Tabela 3. Capacidade preditiva dos modelos candidatos *ARMA* para os óbitos em Porto Alegre, estado do Rio Grande do Sul, no período entre 2017-2019

ARMA(*p,q*)	Tendência	Constante	RMSE	MAE	MAPE (%)
(0,0)	Com	Com	64,5428	53,9131	0,4523
(1,0)	Com	Com	135,541	115,685	0,9702
(0,1)	Com	Com	60,2632	51,4890	0,4310

Fonte: Elaborado pelos autores a partir dos dados da pesquisa.

A Tabela 4 apresenta valores observados e previstos para o período 2017-2020, evidenciando a boa capacidade preditiva dos dois melhores modelos, pois todos os valores observados pertencem aos respectivos intervalos de confiança.

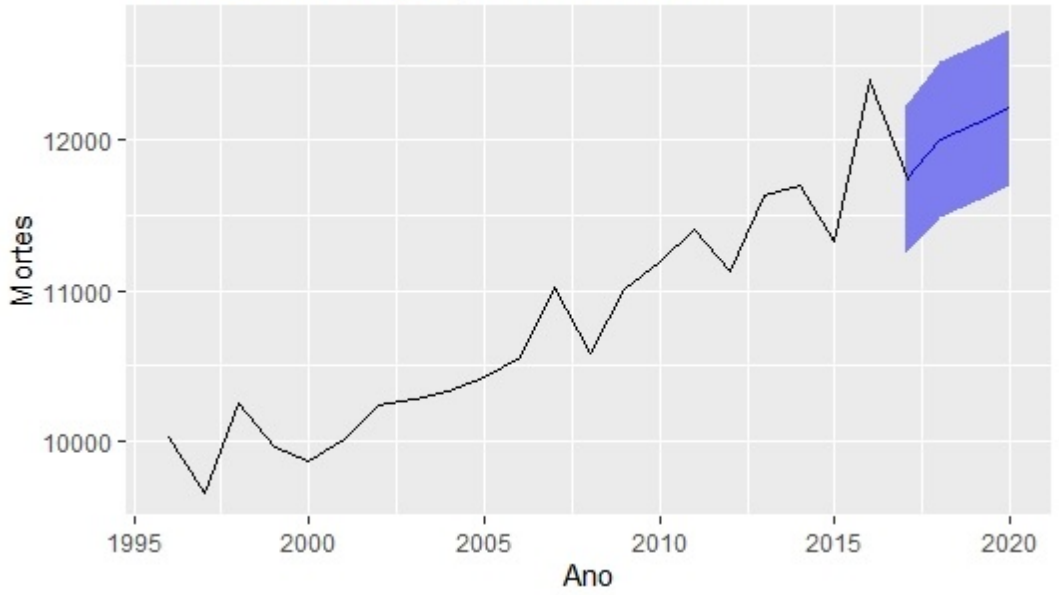
Tabela 4. Valores observados e previstos para os óbitos em Porto Alegre, estado do Rio Grande do Sul, no período entre 2017-2020

	Ano	*Observado*	*Previsão*	*IC (0,95)*
ARMA(0,0)	2017	11.816	11.914,64	11.408,5–12.420.78
	2018	12.075	12.023,74	11.517,6–12.529.88
	2019	12.121*	12.132,84	11.626,7–12.638.98
	2020	-	12.241,94	11.735,8–12.748.08
ARMA(0,1)	2017	11.816	11.738,97	11.243,20–12.234,74
	2018	12.075	12.004,95	11.487,36–12.522,54
	2019	12.121*	12.113,61	11.596,02–12.631,20
	2020	-	12.222,27	11.704,68–12.739,87

Fonte: Elaborado pelos autores a partir dos dados da pesquisa. Símbolo (*) indica dado preliminar, e símbolo (-) indica ausência do dado.

Para o ano de 2020, se o contexto permanecesse inalterado, o modelo *ARMA*(0,1), que apresenta um MAPE de apenas 0,43%, estima um valor esperado de 12.222 óbitos, sendo que o intervalo de confiança [11.705-12.740] possui a probabilidade de 0,95 de conter o verdadeiro valor que ocorrerá neste ano. A representação gráfica dos valores observados no período 1996-2019 da série analisada e das previsões para o período 2017-2020, com respectivos intervalos de confiança, estão apresentadas no gráfico da Figura 7, no qual é possível constatar a qualidade do modelo e de suas previsões.

Figura 7. Representação do modelo *ARIMA*(0,1) para os óbitos em Porto Alegre, estado do Rio Grande do Sul, no período entre 2017-2020



Fonte: Elaborado pelos autores a partir dos dados da pesquisa.

No ano de 2020, entretanto, espera-se a ocorrência de uma quebra estrutural nesta série em função da pandemia de COVID-19, o que provavelmente causará um aumento neste valor esperado, que poderá ser maior que o limite superior do intervalo estimado, ou seja, superior a 12.740 óbitos.

5. Conclusões

A metodologia de Box-Jenkins é bastante empregada na modelagem de séries temporais, em especial, os modelos autorregressivos de médias móveis (*ARMA*), resultante da combinação de modelos autorregressivos (AR) com modelos de médias móveis (MA). O presente estudo buscou modelar a série temporal de óbitos de indivíduos residentes em Porto Alegre (RS), com a sua posterior previsão para o ano de 2020. Os resultados observados indicam que a metodologia Box-Jenkins forneceu um modelo apropriado para previsão de óbitos por COVID-19, pois, além de erros pequenos no período de validação, todos os valores observados ficaram dentro dos intervalos de confiança. Ressalta-se, entretanto, que se espera haver discrepância entre os valores previsto e observado para o ano de 2020, sendo este último superior ao primeiro,

devido ao “efeito da COVID-19”.

Considerando a fragilidade das informações em relação à incidência da doença (taxa de mortalidade e taxa de letalidade), acredita-se que, a partir do valor observado para o ano de 2020, seja possível estimar a quantidade de óbitos decorrentes da referida enfermidade. Como trabalhos futuros, sugere-se replicar esta modelagem a dados de outros municípios e estados, assim como compará-los com modelos encontrados por outros métodos, como por exemplo o modelo G*ARMA*.

Referências

AKAIKE, H. Maximum likelihood identification of Gaussian autoregressive moving average models. **Biometrika,** v. 60, 255–265,1973. DOI: https://doi.org/10.1093/biomet/60.2.255.

AKHTAR, S.; ROZI, S. An autoregressive integrated moving average model for short-term prediction of hepatitis C virus seropositivity among male volunteer blood donors in Karachi, Pakistan. **World Journal of Gastroenterology**¸ n. 15: 1607-1612, 2009.

ALZAHRANI, S. I.; ALJAMAAN, I. A.; AL-FAKIH, E. A. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using *ARIMA* prediction model under current public health interventions. **Journal of Infection and Public Health***,* v. 13: 914–919, 2020.

BARROS, T. V. S. de. **Dengue em Natal/RN: uma análise do período 2000-2016 via séries temporais**. 2019. 56 f. Dissertação - Programa de Pós-Graduação em Matemática Aplicada e Estatística, Universidade Federal do Rio Grande do Norte, Natal, Rio Grande do Norte, 2019.

BAYYURT,L.; BAYYURT, B.. Forecasting of COVID-19 Cases and Deaths Using *ARIMA* Models. **MedRxiv,** 2020. No prelo. DOI: https://doi.org/10.1101/2020.04.17.20069237

BENVENUTO, D.; GIOVANETTI, M.; VASSALLO, L.; ANGELETTI, S.; CICCOZZI, M.. Application of the *ARIMA* model on the COVID-2019 epidemic dataset. **Data in Brief**, v. 29: 1-5, 2020.

BOZDOGAN, H. Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions. **Psychometrica**. n.52, 345-370, 1987. DOI: https://doi.org/10.1007/BF02294361.

BRASIL. Ministério da Saúde. **Painel Coronavírus – COVID 19.** Brasília, DF: Ministério da Saúde do Brasil. 23 set. 2020. Painel de dados como veículo oficial

de comunicação sobre a situação epidemiológica da COVID-19 no Brasil. Disponível em: https://covid.saude.gov.br/. Acesso em: 24 set. 2020.

BUENO, R. de L. da S. **Econometria de séries temporais.** 2 edição. São Paulo: Editora Cengage Learning, 2011.

CEVS-RS. Centro Estadual de Saúde do Rio Grande do Sul. **Boletim epidemiológico – COVID-19.** Disponível em: https://coronavirus.rs.gov.br/upload/arquivos/202005/07181723-boletim-epidemiologico-covid-19-coers-08-04-20.pdf. Acesso em: 24 set. 2020.

DATASUS. Departamento de Informática do SUS. **Informações de Saúde (TABNET) - Estatísticas Vitais, mortalidade de 1996 a 2018, pela CID-10**. Brasília, DF: 2020a. Disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area=0205&id=6937. Acesso em: 24 set. 2020.

DATASUS. Departamento de Informática do SUS. **Informações de Saúde (TABNET) - Estatísticas Vitais, dados preliminares de 2019**. Brasília, DF: 2020b. Disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area=0205&id=6937. Acesso em: 24 set. 2020.

EARNEST, A.; CHEN, M. I; NG, D.; SIN, L. Y. Using autoregressive integrated moving average (*ARIMA*) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. **BMC Health Services Research**, v.5, n.1: 36, 2005.

ENGLE, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflations. *Econometrica,* n. 50: 987-1007, 1982.

GHALANOS, A. **rugarch: Univariate GARCH models. R package version - 1.4-4**. 2020. Disponível em: https://cran.r- project.org/web/packages/rugarch/rugarch.pdf. Acesso em: ago. 2020.

GUJARATI, D. N.; PORTER, D.C. Econometria Básica. 5 edição. AMGH Editora, 2011.

HYNDMAN, R. et al. **forecast: Forecasting functions for time series and linear models. R package version - 13**. 2020. Disponível em https://pkg.robjhyndman.com/forecast/. Acesso em: ago. 2020.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. Melbourne, Australia: OTexts, 2018. Disponível em: http://otexts.com/fpp2/. Acesso em: ago. 2020.

JARQUE, C. M.; BERA, A. K. A Test for Normality of Observations and

Regression Residuals. **International Statistical Review***,* v. 55, n. 2: 163-172, 1987.

KIRBAŞ, İ.; SÖZEN, A.; TUNCER, A. D.; KAZANCIOĞLU, F. Ş. Comparative analysis and forecasting of COVID-19 cases in various European countries with *ARIMA*, NARNN and LSTM approaches. **Elsevier Public Health Emergency Collectio***n*, v. 138: 110015, 2020.

LJUNG, G. M.; BOX, G. E. On a measure of lack of t in time series models. *Biometrika*, v. 65, n. 2: 297–303, 1978.

MORETTIN, P. A.; TOLOI, C. M. C. Análise de Séries Temporais. São Paulo: Blucher, 2006.

NUNES, V. P. **Estudo comparativo entre modelos os *ARIMA* e G*ARMA*: uma aplicação da série temporal de óbitos por neoplasias malignas de pele**. 2018. 132 f. Dissertação - Programa de Pós-Graduação em Modelagem Computacional, Universidade Federal de Rio Grande, Rio Grande, Rio Grande do Sul, 2018.

PFAFF, B. Urca: **Analysis of integrated and cointegrated time series with R.** Nova York: Springer, 2008.

R CORE TEAM. R: **A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2020. Disponível em: https://www.R-project.org. Acesso em: jun. 2020.

SES-RS. Secretaria Estadual de Saúde do Rio Grande do Sul. **Boletim epidemiológico COVID-2019.** Análise das hospitalizações por síndrome respiratória aguda grave e óbitos, Semana Epidemiológica 37**.** Porto Alegre, RS, 2020b. Disponível em: https://coronavirus.rs.gov.br/upload/arquivos/202009/16111423-boletim-covid19-se-37-res.pdf. Acesso em: 24 set. 2020.

SES-RS. Secretaria Estadual de Saúde do Rio Grande do Sul. **Painel Coronavírus – RS.** Porto Alegre, RS, 24 set. 2020a. Disponível em: https://ti.saude.rs.gov.br/covid19/. Acesso em: 24 set. 2020.

SCHWARZ, G. Estimating the dimension of a model [versão eletrônica], **Annals of Statistics,** 6(2), 461–464, 1978. Disponível em : https://projecteuclid.org/download/pdf_1/euclid.aos/1176344136. Acesso em: 20 set. 2020.

TRAN, T. T.; PHAM, L. T.; NGO, Q. X. Forecasting epidemic spread of SARS-CoV-2 using *ARIMA* model (Case study: Iran). **Global Journal of Environmental Science and Management**, v. 6: 1-10, 2020.

TRAPLETTI, A.; HORNIK, K.; LEBARON, B. **tseries: Time Series Analysis and**

Computational Finance. R package version 0.10-47. Vienna, Áustria: R Foundation for Statistical Computing, 2019. Disponível em: https://cran.r-project.org/web/packages/tseries/index.html. Acesso em: 24 set. 2020.

WHO. World Health Organization. **Novel Corona vírus – China**. Genebra, Suíça, 12 jan. 2020. Disponível em: https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/. Acesso em: 24 set. 2020.

YANG, Q.; WANG, J.; MA, H.; WANG, X. Research on COVID-19 based on *ARIMA* model—Taking Hubei, China as an example to see the epidemic in Italy. **Journal of Infection and Public Health**, 2020. No prelo. DOI: https://doi.org/10.1016/j.jiph.2020.06.019.

ZHANG, X; ZHANG, T.; YOUNG, A. A.; LI, X. Applications and comparisons of four time series models in epidemiological surveillance data. **PLoS ONE***,* v. 9, n. 2: e88075, 2014.