Predição da evolução da Covid-19 em Porto Alegre-RS: Modelo ARIMA

Alessandra Paranhas
Ana Maria Pinheiro
João Pedro Faria

I. Introdução

Em janeiro de 2020, O Comitê de Emergência da OMS declarou estado de emergência na saúde global com base no crescimento de notificações de casos em locais chineses e internacionais. O surto do novo coronavírus se espalhou em muitos países, a taxa de detecção de novos casos muda diariamente e tem seu rastreio quase em tempo real disponibilizados em sites governamentais. Por volta de fevereiro de 2020, a China carregava o enorme fardo de mortalidade, enquanto outros países asiáticos, na América do Norte e Europa as taxas permaneciam baixas [1].

O Coronavírus é um grande vírus RNA (RNA é um ácido ribonucleico essencial na síntese de proteínas. Formada a partir da molécula de DNA em um processo chamado de transcrição).

Também sendo uma doença infecciosa que se propagou por muitos países, obtendo-se muitos casos de infectados e causando muitas mortes. A doença é causada pelo vírus Covid-19 [2] que é transmitida através de gotículas produzidas nas vias respiratórias das pessoas infectadas que ao espirrar ou tossir, essas gotículas podem ser inaladas ou atingir diretamente a boca, nariz ou olhos de pessoas próximas. Os casos mais graves podem evoluir para pneumonia grave com insuficiência respiratória grave, chegando a falência de vários órgãos e morte. Nos últimos anos o Brasil não conteve a doença e obteve bastante casos de infectados e mortos. Em maio de 2020 o governo do estado do Rio Grande do Sul, passou a adotar estudos com modelos de séries temporais para previsões e acompanhamentos da doença no estado. [3].

Este artigo foi desenvolvido a partir de um modelo econométrico simples: ARIMA (Auto Regressive Integrated Moving Average), que será útil para o acompanhamento das evoluções e prever a propagação do Covid-19 na cidade de Porto Alegre localizada no estado do Rio Grande do Sul.

II. Referencial Teórico

O presente artigo, a modelo ARIMA também conhecido como metodologia de Box-Jenkins [4] será aplicado aos dados diários do Covid-19 da cidade de Porto Alegre-RS, para acompanhar a evolução dos casos e prever futuros novos casos.

Modelos Box-Jenkins

De acordo com Morretin e Toloi (1987) [5], uma série temporal é qualquer conjunto de valores, em que esses valores são observações ordenadas no tempo. Os principais objetivos da análise de série temporal é compreender o mecanismo gerador da série e predizer o comportamento de série futura. Os modelos de Box-Jenkins, também conhecido por ARIMA (Auto Regressive Integrated Moving Averege), são modelos estatísticos lineares para analisar e prever dados de séries temporais. O ARIMA é um algoritmo de previsão que se baseia na ideia de que as informações nos valores anteriores da série temporal podem ser usadas sozinhas para prever os valores futuros, isto é chamado de previsão de série temporal univariada. De acordo com Fava (2000) [6], os modelos ARIMA podem ser delineado da seguinte forma:

Auto-Regressivo (AR): refere-se as defasagens da série transformada (que seria a série estacionária ob-

tida por diferenciação). Sua fórmula matemática é descrito como:

$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + e_t$

Onde c é uma constante e e_t é um erro aleatório (ruído branco).

Integrado (I): representa o processo de diferenciação da série original para torná-la estacionária.

Médias Móveis (MA): refere-se as defasagens dos erros aleatórios. Sua fórmula matemática é descrito como:

$y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q}$

Onde y_t é uma média ponderada dos erros de previsão passados.

Com a combinação entre esses métodos: auto-regressão, diferenciação e média móvel; resultam numa modelo ARIMA. Com os valores de *p*, *d*, e *q* é possível identificar qual processo está sendo modelado. A fórmula matemática do modelo ARIMA é descrito:

$\acute{y_t} = c + \phi_1 \acute{y}_{t-1} + ... + \phi_p \acute{y}_{t-p} + \theta_1 + \theta_1 e_{t-1} + ... + \theta_q e_{t-q} + e_t$

$\acute{y_t}$ é a série diferenciada.
p é a ordem do modelo auto-regressivo.
d é o grau de diferenciação.
q é a ordem do modelo de média móvel.

Uma série temporal pode ser modelada por esses três componentes ou apenas um subconjunto deles, que resultará em vários outros modelos. E a abordagem desse artigo será apenas os modelos ARIMA.

Etapas da metodologia Box-Jenkis

Os modelos Box-Jenkis seguem algumas etapas para fazer previsão de séries temporais, que são as seguintes:

Etapa 1. Identificação:
Nesta etapa o processo é identificar a estrutura do modelo, identificar os parâmetros de p, d e q que caracterizam o processo estocástico. E são utilizadas algumas ferramentas principais na identificação que são: a função de correlação amostral (ACF) e a função de correlação amostral parcial (PACF) [7].

Etapa 2. Estimação:
Nesta etapa consiste em estimar os parâmetros que foram identificados na etapa anterior.

Etapa 3. Verificação:
Nesta etapa consiste em verificar se o modelo ARIMA específico, tendo os parâmetros estimados, é adequado. Caso contrário, é necessário voltar nas etapas anteriores e repetir o processo até o modelo ARIMA correto.

Etapa 4. Previsão:
Depois da verificação do modelo correto, passa-se então para a previsão dos valores futuros.

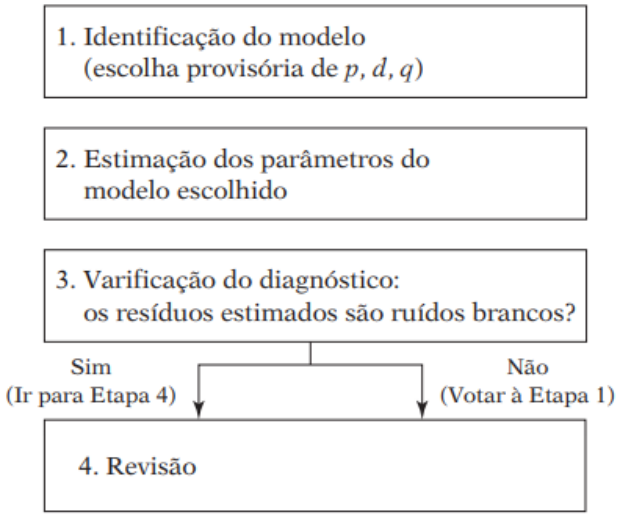

Fig.1: Modelo Box-Jenkins.

Software RStudio

O Rstudio é um software livre de análise de dados,com ambiente de desenvolvimento integrado para linguagem de programação R, para gráficos e cálculos estatísticos.

Pacotes usados na análise e na geração do modelo ARIMA

tidyverse

O tidyverse contém um conjunto de pacotes R utilizados no dia a dia para análise de dados e para ciência de dados.

lubridate:
Pacote para manipulação de datas.

forecast:
O pacote forecast disponibiliza métodos e ferramentas para analisar e exibir previsões de séries temporais univariadas.

ggplot2:
O pacote ggplot2 é um pacote para criar vizualizações gráficas, e sua essência está na construir um gráfico camada por camada.

III. Metodologia

A análise preditiva será a metodologia que utilizaremos para conclusões de previsões de cenários futuros sobre a evolução de casos de Covid-19. E afim de se obter dados foi feita uma pesquisa documental de casos de corona vírus e utilizaremos os registros dos dados do Covid-19 do site oficial da prefeitura de Porto Alegre-RS [8]. Os dados do site são atualizados diariamente pelos serviços de saúde pública e privada, com os principais dados epidemiológicos da Covid-19 em Porto Alegre para monitoramento da doença. Os dados mostram a *ocupação de leitos UTI*, *casos de Covid-19* e *testes em Porto Alegre*.

No presente artigo, aplicamos o modelo ARIMA para prever o número de casos infectados por Covid-19 nos próximos dias em Porto Alegre-RS. As previsões dos dados serão feitas no Software R.

IV. Desenvolvimento

Os dados diários foram coletados no período de 20 de fevereiro de 2020 a 30 de maio de 2021, e o software R foi utilizado para o tratamento dos dados e a aplicação do modelo ARIMA.

Os dados passaram por uma limpeza e transformação para melhor desempenho do modelo, tipo: remoção de outliers, transformação de variáveis, remoção de variáveis...

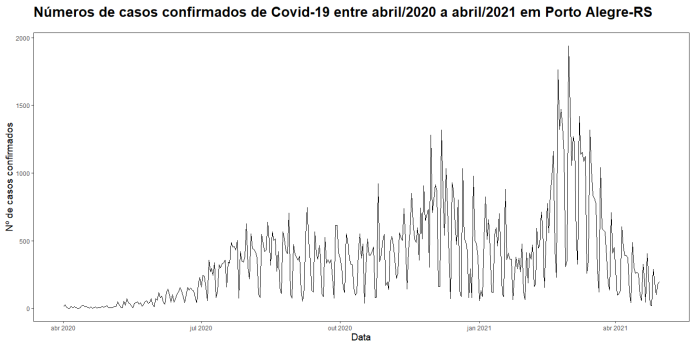

Fig.2: Gráfico de série temporal de casos Covid-19.

O gráfico acima mostra o número de casos de Covid-19 confirmado, contabilizados diariamente na cidade de Porto Alegre-RS no período de abril de 2020 até abril de 2021. As funções de autocorrelação (acf) e autocorrelação parcial (pacf) [7] foram utilizadas para identificar a ordem apropriada dos processos Auto-Regressivo (AR) e Médias Móveis (MA) - p e q.

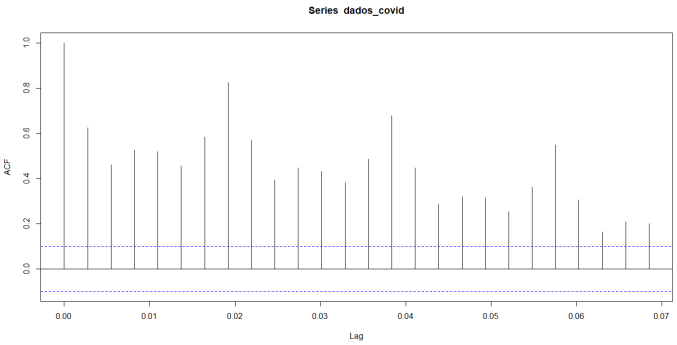

Fig.3: Correlograma de autocorrelação.

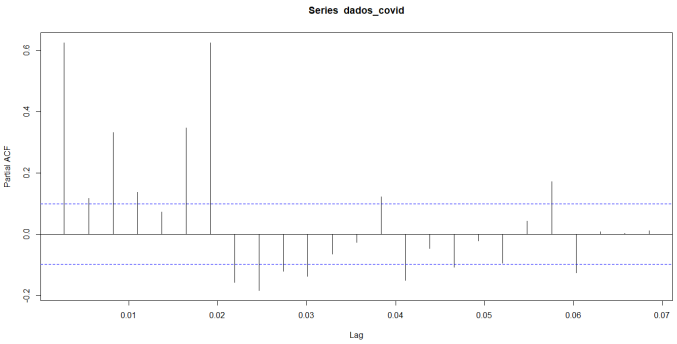

Fig.4: Correlograma de autocorrelação parcial.

Nos gráficos acima as linhas horizontais tracejadas azuis define o nível de um intervalo de significância mínimo, isto é, se as barras passam por essas linhas, significa que têm significado estatístico.

eixo y: pontuação de correlação
eixo x: indicação das defasagens.

Logo em seguida, os dados passaram por uma transformação logarítmica afim de tentar controlar a heterocedasticidade e melhorar as previsões de futuros dados.

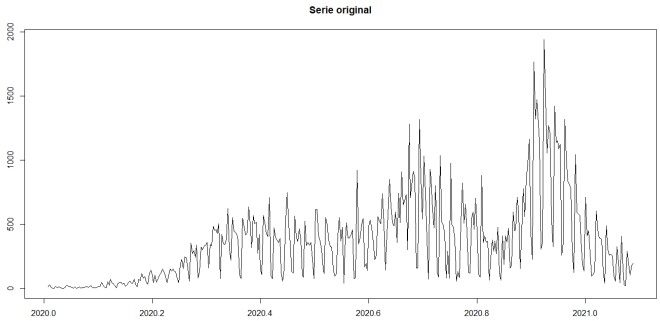

Fig.5: Série original.

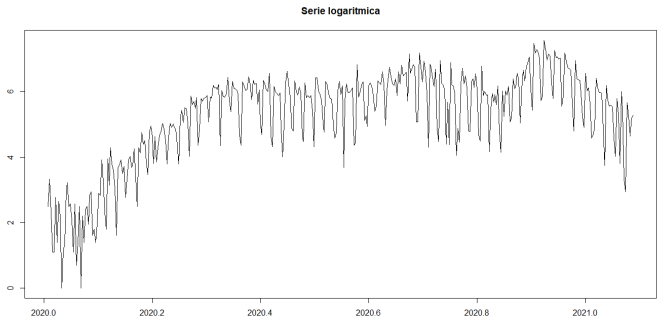

Fig.5: Série logarítmica.

Para o treinamento do modelo ARIMA foram utilizados os dados no período de abril de 2020 a abril de 2021 (395 pontos da série), e para validação do modelo foram utilizados os dados do mês de maio de 2021 (30 pontos).

Para a criação do modelo ARIMA (p, d, q), foi utilizado a função auto.arima() [9] do pacote forecast() [10]. A função auto.arima realiza uma pesquisa sobre os modelos possíveis dentro das restrições de pedido fornecidas e sugere o modelo de melhor ajuste como ARIMA.

Logo após a criação do modelo ARIMA (p, d, q), a função auto.arima sugeriu um modelo (0,1,2) como o melhor ajuste para o ARIMA.

Logo após do treinamento foi aplicado a transformação inversa da função logarítmica para depois fazer as previsões. As previsões de novos casos de Covid-19 foram feitas nos próximos 30 dias, no caso, para o mês de maio de 2021.

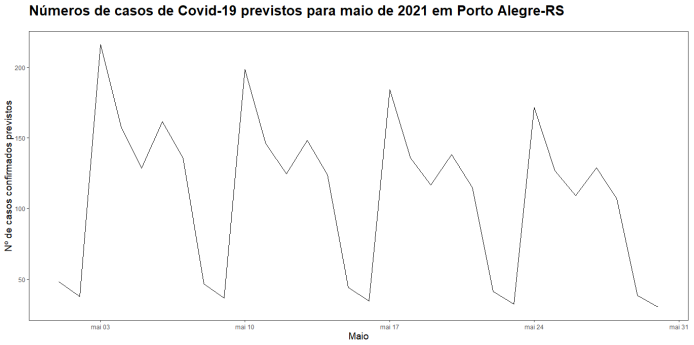

Fig.7: Previsões.

E para analisar se o modelo generalizou e previu bem os dados, comparamos as previsões com a base de validação.

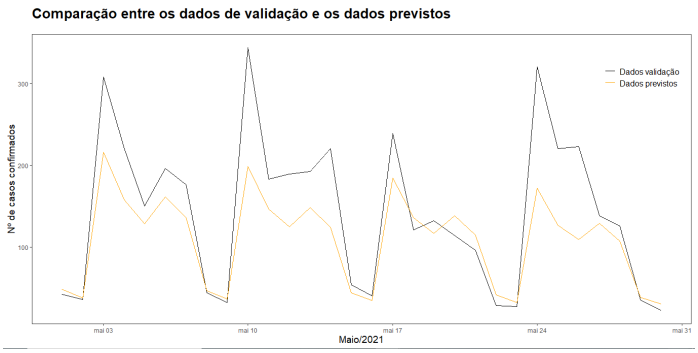

Fig.8: Comparação com as previsões e base de validação.

Pelo gráfico de comparação percebemos que o modelo generalizou bem os dados. E algumas métricas foram aplicadas para analisar o desempenho do modelo, *RMSE*, *MAE*, *MAPE*. E esses são os valores das métricas obtidos do modelo ARIMA:

Métricas	Valores
RMSE	58.46
MAE	40.31
MAPE	31.39

RMSE (Raiz Quadrada Média do Erro): essa métrica calcula a raiz quadrática média dos erros entre valores observados (reais) e predições (hipóteses)

[11].

MAE (Erro Absoluto Médio): essa métrica calcula o erro absoluto médio dos erros entre os valores observados (reais) e predições (hipóteses) [11].

MAPE (Erro Percentual Absoluto Médio): essa métrica definida o erro como o valor real ou observado menos o valor previsto. Os erros de porcentagem são somados independentemente do sinal para calcular o MAPE [12].

V. Conclusão

De a acordo com os dados coletados e apresentado no gráfico *Fig.2, Gráfico de série temporal de casos Covid-19.*, verificamos que inicialmente á partir de maio de 2020 houve um crescimento nos casos de Covid-19 em Porto Alegre, tendo maior pico em fevereiro de 2021 à março de 2021. Após esse período houve uma queda nos casos registrados á partir do mês de abril.

Com base nos dados obtidos pelo modelo ARIMA, as previsões foram próximas dos dados reais, seguindo a tendência dos dados. As previsões indica uma ligeiramente diminuição nos casos infectados pela Covid-19 na cidade de Porto Alegre-RS.

Ao realizar as análises é possível concluir que o método de predição ARIMA é adequado e tende a gerar ótimos resultados.

Referências

[1] Thirumalaisamy P Velavan e Christian G Meyer. "The COVID-19 epidemic". Em: *Tropical medicine & international health* 25.3 (2020), p. 278.

[2] "COVID-19". Em: *Wikipédia: a enciclopédia livre*. Wikimedia, 2021. URL: %5Curl%7Bhttps://pt.wikipedia.org/wiki/COVID-19%7D.

[3] Cristiano Aguiar de Oliveira. "RIO GRANDE DO SUL SOB BANDEIRA PRETA: UMA AVALIAÇÃO DO MODELO DE DISTANCIAMENTO CONTROLADO ATRAVÉS DE UMA ANÁLISE QUASE EXPERIMENTAL BASEADA EM PREVISÕES REALIZADAS COM O AUXÍLIO DE BUSCAS NO GOOGLE". Em: ().

[4] Sunil Bhatnagar, Vivek Lal, Shiv D Gupta, Om P Gupta et al. "Forecasting incidence of dengue in Rajasthan, using time series analyses". Em: *Indian journal of public health* 56.4 (2012), p. 281.

[5] Pedro A Morettin e Clélia Toloi. "Análise de séries temporais". Em: *Análise de séries temporais*. 2006, pp. 538–538.

[6] Vera Lúcia FAVA et al. "Manual de econometria". Em: *Vasconcelos, MAS; Alves, D. São Paulo: Editora Atlas* (2000).

[7] "Significance of ACF and PACF Plots In Time Series Analysis". Em: *https://towardsdatascience.com/significance-of-acf-and-pacf-plots-in-time-series-analysis-2fa11a5d10a8*. Website, 2019.

[8] "Secretaria Extraordinária de ENFRENTAMENTO AO CORONAVÍRUS". Em: *https://prefeitura.poa.br/coronavirus*. Website, 2021.

[9] "auto.arima function - RDocumentation". Em: *https://www.rdocumentation.org/packages/forecast/versions/8.15/topics/auto.arima*. Website.

[10] "CRAN - Package forecast". Em: *https://cran.r-project.org/web/packages/forecast/index.html*. Website.

[11] "RMSE ou MAE? Como avaliar meu modelo de machine learning?" Em: *https://www.linkedin.com/pulse/rmse-ou-mae-como-avaliar-meu-modelo-de-machine-learning-rezende/?originalSubdomain=pt*. Website.

[12] "MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)". Em: *https://doi.org/10.1007/1-4020-0612-8*$_5$80. Website.

[13] Damodar N. Gujarati. *Econometria básica*. Makron Books, 2000.

[14] Farhan Mohammad Khan e Rajiv Gupta. "Arima and nar based prediction model for time series analysis of covid-19 cases in india". Em: *Journal of Safety Science and Resilience* 1.1 (2020), pp. 12–18.

[13] [14]