Análise da série temporal dos casos COVID-19

Kamalpreet Singh Bhangu, Jasminder Kaur Sandhu e Luxmi Sapra
Instituto de Engenharia e Tecnologia da Universidade de Chitkara, Universidade de Chitkara, Punjab, Índia

Abstrair

Propósito – Este estudo analisa a epidemia prevalente da doença coronavírus (COVID-19) utilizando algoritmos de aprendizagem de máquina. O conjunto de dados usado é um dado de API fornecido pelo centro de recursos da Universidade John Hopkins e usou o rastreador da Web para coletar todos os recursos de dados, como casos confirmados, recuperados e de morte. Devido à indisponibilidade de qualquer droga COVID-19 no momento, a verdade sem verniz é que este surto não deve terminar em um futuro próximo, de modo que o número de casos deste estudo seria muito específico. A análise demonstrada neste artigo tem como foco a análise mensal dos casos confirmados, recuperados e de óbito, o que auxilia a identificar a tendência e a sazonalidade nos dados. O objetivo deste estudo é explorar os conceitos essenciais dos algoritmos de séries temporâneais e usar esses conceitos para realizar análises de séries temporâneas sobre os casos infectados em todo o mundo e prever a disseminação do virus nas próximas duas semanas e, assim, auxiliar nos serviços de saúde. Resultados médios de erro percentual absoluto mais baixos obtidos do intervalo de tempo de previsão validam a credibilidade do modelo.

Design/metodologia/abordagem – Neste estudo, a análise em série temporal dessa previsão de surto foi feita utilizando-se o modelo auto-regressivo integrado da média móvel integrada (ARIMA) e também as médias móveis integradas auto-regressivas sazonais com regressor exógeno (SARIMAX) e otimizadas para obter melhores resultados.

Resultados – As inferências dos modelos de previsão de séries temporáticas ARIMA e SARIMAX foram eficientes para produzir resultados aproximados exatos. Os resultados de previsão indicam que uma tendência crescente é observada e há um aumento elevado nos casos de COVID-19 em muitas regiões e países podem enfrentar um de seus piores dias a menos e até que medidas sejam tomadas para conter a propagação dessa doença rapidamente. O padrão do aumento da disseminação do vírus nesses países está exatamente imitando alguns dos países da primeira adoção do COVID-19 como Itália e EUA. Além disso, os números obtidos dos modelos são específicos para que a execução mais recente do modelo retorne resultados mais recentes. O escopo futuro do estudo envolve análise com outros modelos, como memory de longo prazo e, em seguida, comparação com modelos de séries temporidas.

Originalidade/valor – Uma série de tempo é um conjunto de dados carimbado em que cada ponto de dados corresponde a um conjunto de observações feitas em uma determinada instância de tempo. Este trabalho é novo e aborda o COVID-19 with a ajuda da análise de séries temporais. As inferências dos modelos de previsão de séries temporências ARIMA e SARIMAX foram eficientes para produzir resultados exatos aproximados.

Palavras-chave COVID-19, Análise da série Time, ARIMA, SARIMAX, Previsão

Artigo de pesquisa de papel

Revista Mundial de Engenharia
19/1 (2022) 40-48

[DOI 10.1108/WJE-09-2020-0431]

1. Introdução

A doença coronavírus (COVID-19) surgiu em Wuhan (China) em dezembro de 2019 (Yonar, 2020) e agora se tornou uma pandemia. Como muitos países, a Índia relatou seu primeiro caseonto COVID-19 no 30º janeiro.Quando algo como oCOVID-19ison se espalha, passa por quatro estágios; o primeiro é o estágio de importação da doença e seguido pelo estágio de contato da doença (Malki et al. , 2020). O terceiro estágio torna-se a comunhão difundida e, finally, a quarta etapa é declarada como uma pandemia global. O surto do COVID-19 está se transformando em uma grande crise internacional e está emprestando para influence os aspectos mais importantes de nossas vidas diárias. Por exemplo, lugares na maioria das cidades do mundo, onde a comunidade se fecha, como shoppings, mercearias, salas de cinema e pubs.

Os algoritmos de aprendizado de máquina podem ser usados para treinar a partir dos dados COVID-19 disponíveis publicamente e, em seguida, prever e inferir informações úteis a partir dele (Lalmuanawma et al. , 2020). A extração de diferentes tipos de casos ajudaria a alertar o governo na tomada de decisões de extensão de bloqueios para salvaguardar a disseminação dessa pandemia dentro de suas respectivas regiões. Isso nos permite o uso ineficaz dos algoritmos de timeseriestopredictandforecastCOVID-19casesaccurately.

Uma série de tempo é um conjunto de dados com carimbo de tempo no qual cada ponto de dados corresponde a um conjunto de observações feitas em uma instância de tempo específica (Nabi, 2020). Uma lista ilustrativa de dados de séries temporais em diversos domínios inclui figures diários do mercado de ações, vendas ou receitas ou figures de profitabilidade de empresas por ano e informações demográficas como população, taxa de natalidade, figures de mortalidade infantil, alfabetização, renda per capita, figures de matrícula escolar por ano e também pressão arterial e outros sinais vitais do corpo por unidades de tempo (Singh e Dhiman, 2018). A série temporal tem uma forte dependência temporal (baseada no tempo) – cada um desses conjuntos de dados

Revista Mundial de Engenharia
Volume 19 · Número 1 · 2022 · 40–48

consiste essencialmente de uma série de observações carimbadas pelo tempo, ou seja, cada observação está ligada a uma instância de tempo especific (Benvenuto et al. , 2020). Assim, desapedendo a regressão,

Recebido em 10 de setembro de 2020
Revisado em 5 de novembro de 2020
Aceito em 28 de novembro de 2020

a ordem dos dados é importante em uma série temporal. Além disso, em série temporal, a relação entre a resposta e a variável explicativa não é levada em consideração porque há apenas uma variável ,ou seja, tempo (horas, minutos, segundos, mês ou ano) no modelo que é uma variável independente (Abdulmajeed et al. , 2020). A causa por trás das mudanças na variável de resposta é muito uma caixa preta.

Neste estudo, a análise da série temporal desta previsão de surto foi feita utilizando-se o modelo auto-regressivo integrado da média móvel (ARIMA) e também as médias móveis integradas auto-regressivas sazonais com regressor exógeno (SARIMAX) e otimizadas para melhores resultados. Este trabalho é organizado da seguinte forma: a Seção 2 sublinha a metodologia das séries temporais que consistem em conjunto de dados e desenvolvimento de modelos. A seção 3 discute a avaliação, o resultado e o erro percentual absoluto (MAPE). Finalmente, a Seção 4 apresenta a conclusão.

2. Metodologia

Os dados da série temporal são os dados medidos em período de tempo sucessivo em intervalo de tempo uniforme. Alguns dos exemplos são os preços de fechamento de S&P, NSE, BSE e Dow Jones e receita trimestral e profit de uma empresa. A série temporal é geralmente univariada (uma variável) e ao longo do tempo variável de mudanças de juros em relação à variável univariada. O primeiro passo para a exploração de dados em séries temporalhais é a análise de tendências e é utilizado para a previsão da variável de interest, como profit e perda na maioria dos cenários. Essas etapas são cobertas por resultados e seção de avaliação e há três variáveis de interesse neste estudo que são confirmadas, recuperadas e casos de óbito de COVID-19.

2.1 Conjunto de dados

Os casos confirmados, recuperados e de morte por infecção por COVID-19 são coletados no site oficial da Universidade Johns Hopkins (https://gisanddata.maps.arcgis.com/ apps/opsdashboard/index.html) para o período de 22 de janeiro de 2020 a 12 de agosto de 2020. A Tabela 1 mostra as características do conjunto de dados.

2.2 Desenvolvimento de modelos

Vale mencionar sobre os componentes da série temporal que compreendem componentes previsíveis e imprevisíveis. Os subcomponidores de componentes previsíveis consistem em componente local (Xt) e global (tendência [Tt] e sazonalidade [S_t]) enquanto componente consist imprevisível de ruído (Z_t). O componente global sob previsibilidade tem comportamentos de tendência (Tt) e sazonalidade (St) (Fattah et al. , 2018). O componente tendência (T_t) refere-se a qualquer padrão que fale sobre o aumento global ou diminuição dos valores, enquanto a estatempoonalidade (S_t) refere-se a um padrão repetitivo de valores vistos nos dados (Dhiman e Kaur, 2018). O componente local (X_t) é a súbita inflação errática ou efeito imediato e fluctuations na variável dependente em comparação com fatos de decisão anteriores. O ruído (Z_t) é o componente de erro e não pode ser previsto. Assim, em uma série temporal, todos os componentes são combinados, ou seja, séries temporizais (TS) = local (X_t) 1 tendência (Tt) 1 sazonalidade (St) 1 ruído (Zt) e tendência de remoção (Tt) e sazonalidade (St), do componente original dá componentes deixados de fora que é mistura de local (Xt) e ruído (Z_t) (Li e Li, 2017). Existem diferentes maneiras pelas quais esses componentes da série temporal podem estar relacionados entre si (Dhiman e Kaur, 2019a). Sua combinação pode ser aditiva ou multiplicativa. O modelo aditivo é dado como TS = T_t 1 S_t 1 Xt 1 Nt e o modelo multiplicativo é dado como TS = T_t St Xt Nt. Quando a magnitude do padrão sazonal nos dados aumenta com uma increase nos valores dos dados e diminui com uma diminuição nos valores dos dados, o modelo multiplicativo pode ser uma escolha melhor. Quando a magnitude do padrão sazonal nos dados não se correlaciona diretamente com o valor das teses, o modelo aditivo pode ser uma escolha better.

Table 1 Descrição dos recursos do conjunto de dados

Nome do recurso	Descrição do recurso
Data de observação	A data em que o caso COVID-19 foi observado
Província/estado	Nome da província o caso ocorreu

2.2.1 Stationarity de uma série temporal e sua significance

Diz-se que uma série é "estritamente estacionária" se a distribuição marginal de Y no momento t, p(Yt), é a mesma que em qualquer outro momento. Portanto, p(Y_t) = p(Y_{t1k}) e p(Yt, Yt1k) não depende de t. (Aqui, t > = 1 e k é qualquer inteiro). Isso implica que a média, variância e covariância da série Yt são invariantes do tempo.

Para uma série temporal estacionária, propriedades como média e variância serão as mesmas para quaisquer janelas bi-tempo. Uma série temporal estacionária não terá padrões previsíveis de longo prazo, como tendências ou sazonalidade. As tramas de tempo mostrarão a série a ter uma tendência horizontal com a variação constante aproximadamente. Após a remoção do padrão global dos dados originais, a série obtida consiste em componentes deixados de fora ou séries residuais em termos de componentes locais (Xt) e ruído (Z_t). A seguir, a findar todas as correlações de quaisquer janelas bi-tempo da série residual. Se alguma dependência de valores em valores passados com essas janelas ocorre, então isso confirms a existência de padrões locais e tais previsões devem ser modeladas (Reddy et al. , 2017). Em palavras de ordem, estacionário ajuda a fiar o previsível componente local (Xt) e também chamado de estacionária fraca. Some dos testes para verificar o estacionário são o teste de histograma e o teste de enredo Q-Q. A função de correção automática (ACF) e a função de correção automática parcial (PACF) são os testes tradicionais utilizados para determinar estacionários. Se os dados

Revista Mundial de Engenharia

Kamalpreet Singh Bhangu, Jasminder Kaur Sandhu e Luxmi Sapra Volume 19 · Número 1 · 2022 · 40–48

País/região	País em que o caso ocorreu
Última atualização	A data em que os dados foram registrados no caso COVID-19
Inveterado	Número de casos confirmados na data da observação
Mortes	Número de casos de óbito na data de observação
Recuperado	Número de casos recuperados na data de observação

séries temporudes; em outras palavras, não deve haver correlação entre lags.

Deixe {et} denotar tal série que tenha zero média E(et) = 0, tem uma variância constante V(et)= s2, é um E(etes) não corrigido = 0 e variável aleatória ou auto-covariance é zero (Garg e Dhiman, 2020). Cada observação não é corrigida com outras observações na sequência e o enredo de dispersão de tal série ao longo do tempo não indicará nenhum padrão e, portanto, prever os valores futuros de tal série não é possível (Dhiman, 2019). O componente imprevisível na série temporal é um erro de dados chamado ruído branco (Zt) e também chamado de stationarity forte. Se a fonte de dados da série de tempo mostrar recursos de ruído branco, então não há razão para realizar séries temporáticas.

2.2.3 Classificação da série temporal estacionária através da função de correção automática e função de autocorrelação parcial
Modelos de séries temporpulais como ARIMA e SARIMAX são expressos na forma (p, d, q), onde p, q e d são os parâmetros representados como regressão automática, média móvel (MA) e ordem de diferença, respectivamente. Esses parâmetros combinados para determinar a ordem do modelo e são usados para otimizar tais modelos. Os gráficos ACF e GRÁFICO PACF são usados para determinar os valores iniciais de p e q. Na ACF, os valores passados estão correlacionados com os valores atuais, tornando a série uma sequência de valores independentes, não corrigidos, e no PACF, a influência de qualquer um dos valores intermediários são nulidades e, assim, isola a correlação direta entre os valores. O ADF e o KPSS são alguns outros testes para verificação de estacionalidade (Chi, 2018). No teste ADF, a hipótese nula pressupõe que a série não está estacionária; não existem ruídos brancos e componentes previsíveis. Se o valor p dado por este teste for inferior a 5%, a hipótese nula é rejeitada. No teste KPSS, a hipótese nula pressupõe que a série está parada. Os testes de normalidade (trama histograma/Q-Q) como o teste de não está parado, então pode ser feito estacionário através da técnica de diferenciamento (Dhiman, 2020).

2.2.2 Ruído branco
Uma série é chamada de ruído branco se é puramente aleatória na natureza. É um conjunto de valores independentes e não corrigidos e possui características de distribuição gaussiana com um desvio padrão zero e constante. Séries temporais são classificadas como ruído branco se tem média igual a zero, desvio padrão ou volatilidade de séries temporais ao longo do tempo é constante e a correlação entre lags é zero, o que significa que não deve haver correlação entre a série temporal e uma versão defasada do
Shapiro também são úteis para determinar forte estacionário (ruído branco).

2.2.4 Modelagem de componentes previsíveis
Todos os testes mencionados no parágrafo anterior identificam-se parados em séries temporâneas. O componente de ruído branco ou componentes estacionários ou imprevisíveis fortes podem ser negligenciados e nenhuma modelagem pode ser realizada. No entanto, componentes stationary ou locais fracos devem ser modelados e os modelos prevalentes para este fim são o modelo auto-regressivo (AR). A sequália do modelo AR está representada como:

$$x_t \; 1/4 \; a1Xt11Zt$$

onde xt é um componente previsível ou local e Zt é um componente de ruído. Os valores de lag anteriores estão efetuando os valores de lag atuais neste modelo, por exemplo, na equação AR(2) = a1 2 x_{t1} 1 3Xt2, os dois valores de componentes de lag reais anteriores estão contribuindo para o Xt atual valor do componente. Agora, os componentes previsíveis ou estacionários fracos ainda podem ter componentes de ruído e o modelo para prever o componente de ruído é o modelo MA. A equação do modelo MA é representada como: Xt = a 1 bZt11 Zt. Como observado no parágrafo anterior; qualquer modelo de série temporal é expresso em (p, d, q) forma e p neste modelo denota ar(p) e q denota ma(q) modelos o que significa p dá o número de valores componentes reais anteriores contribuindo para o valor original presente e q dá o número de valores de componentes sonoros anteriores que afetam o valor original atual (Nelson, 1998). Além disso, se ambos os componentes reais anteriores e componentes de ruído anteriores estão contribuindo para o valor original atual, então o modelo utilizado é um combinado de AR e ma modelo chamado ARMA (p, q) e a equação formada para este modelo é representada como Xt = a 1 bXt1 1 cZt1 1 Zt , onde os parâmetros p e q denotam os valores

Revista Mundial de Engenharia
Volume 19 · Número 1 · 2022 · 40–48

de lag anteriores e os valores de ruído de lag anteriores, respectivamente, e b e c são coefficients do modelo estimado por métodos de probabilidade semelhantes no caso aos modelos de regressão. Também mencionado no parágrafo anterior, utilizando PACF, o valor de p é determinado e da mesma forma utilizando ACF, o valor de q é determinado.

2.2.5 Modelando séries temporesistente stationary
É um processo em que as previsões globais (tendência e sazonalidade) são removidas e as séries originais e o resíduo intermediário obtido são testados se uma determinada série temporal é semanalmente estacionária ou não e se essa série temporal é identificada como estacionária semanal, então sua equação é figurada.

2.3A extensão do modelo médio móvel integrado-regressivo O modelo médio móvel integrado ARMA (p, q) para incluir diferenciamento (d) no modelo, ARIMA (p, d, q) é formado, onde d é o parâmetro diferente. É um dos algoritmos mais efficientes na previsão de séries temporais (Contreras et al. , 2003). Este modelo é representado como: ARIMA (p, d, q): x_t = a1 x_{t1} 1a2* x_{t2} 1 b1*z_{t1} 1b2* Z $_{t2}$ 1 z_t. Para ser especific em termos de dados COVID19, Xt é o número previsto de casos confirmados, recuperados e de morte no dia específico (tth). No presente estudo, a tendência de incidências futuras para o próximo dia 14 é estimada a partir das observações anteriores após a formação do modelo no período de 21 de janeiro de 2020 a 30 de junho de 2020 e os dados de testes variaram de 1º de julho de 2020 a 12 de agosto de 2020. As tendências de previsão dos casos futuros são reconhecidas para as datas de 13 de agosto de 2020 a 27 de agosto de 2020.

2.4Asseasonal auto-regressivas médias móveis com modelo regressor exógeno
É frequentemente chamado de modelo ARIMA sazonal e componente sazonal também é coberto juntamente com componentes p, d e q como no modelo ARIMA tradicional. O modelo SARIMAX tem quatro parâmetros como (p, d, q, s) é o comportamento de sazonalidade que pode ser mensal ou anualmente, conforme o conjunto de dados. Além disso, outros fatores também influem os resultados em séries tempoesas, como a determinação do preço das ações, também depende de outras variáveis independentes que não o fator tempo, como o PIB do país. O X denota os regressores externos e integrados com modelos como ARIMA para formar modelos como SARIMAX e ARIMAX e, portanto, representam combinação de ARIMA e regressão linear (Soebiyanto et al. , 2010). Os variables independentes são passados com a ajuda do parâmetro exog nos modelosARIMAX e SARIMAX (Arunraj et al. , 2016).

2.5Steps para modelar modelo de séries temporâneas As etapas de alto nível do processo incluem: visualizar as séries temporâneais; reconhecer tendência e componente sazonal; aplicar regressão ao modelo de tendência e sazonalidade; remover a tendência e componente sazonal da série; o que resta é a parte estacionária: uma combinação do AR, e ruído branco; modelar esta série temporal estacionária; combinar a previsão deste model com a tendência e componente sazonal; findar a série residual subtraindo o valor previsto do valor real observado; e verificar se a série residual é puro ruído branco.

Os parâmetros do modelo são então estimados para otimizar o modelo ARIMA. O modelo de previsão de futuros casos COVID-19 é representado como:

$$\text{ARIMA } p\eth;\ d;\ qTH : Xt$$

$$1/4\ a1Xt1\ 1a2Xt2\ 1b1Zt1\ 1b2Zt2\ 1Zt \quad (1)$$

onde:

$$z_t\ 1/4\ Xt\ x_{t1} \quad (2)$$

Aqui, x_t é o número previsto de casos covid-19 confirmados no dia tº anda1,a2, b1 e b2 são parâmetros, enquanto Zt é um termo anual para o décimo dia (Xue e Lai, 2018).

3. Resultado e avaliação

Os dados originais do COVID-19 são, desfinariamente, um dado de tendência ascendente e o enredo de diferentes tipos de casos – confirmados, recuperados e casos de morte na Figura 1 – mostra claramente a tendência de um número crescente desses casos. Em meados de março, os casos confirmados são vistos a subir tremendamente seguidos por casos de recuperação e morte que apresentam tendência e sem orsonalidade.

3.1 Estratégias específicas da Timeseries
Esta seção apresenta os enredos dos casos como tendências nos dados, decomposição de dados, gráficos ACF e PACF. Na Figura 1, os casos de confirmados, recuperados e de óbito têm tendência de alta e, à medida que o tempo aumenta, os casos aumentam. Eles também exibem estacionalidade nos dados. Esse plot justifies geração dos dados da série temporal como o recurso – os dados de observação é o fator tempo e os casos diários DO COVID-19 dependem desse recurso de fator de tempo.

Os dados são descompotados para separar tendência, sazonalidade, componentes residuais e reais, e as parcelas resultantes são mostradas na Figura 2 e o modelo aditivo é usado para esta decomposição. A linha de tendência na figura é bastante linear na natureza, juntamente com essa estacionária está presente nos dados em alta Figura1 Tendência ascendente de casos confirmados, recuperados e de óbito

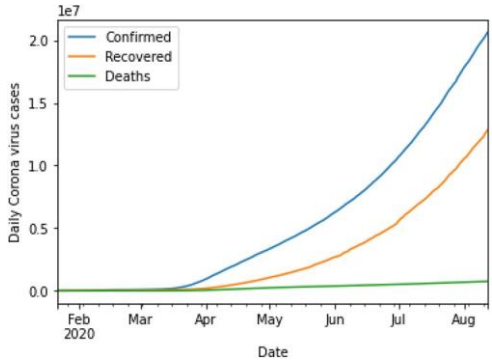
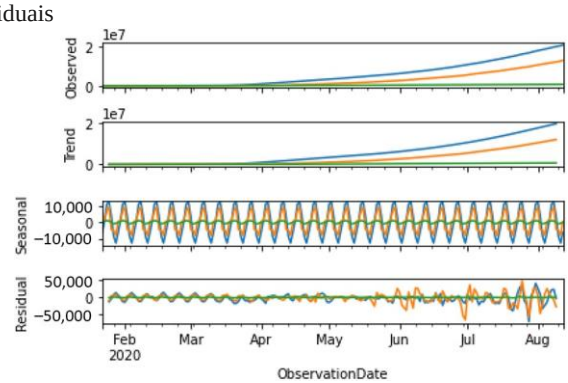


Figura2 Decomposto em dados observados, tendência, sazonais e residuais



escala. O componente residual pode ser um componente local ou ter apenas ruídos ou irregularidades nos dados e precisa ser verificado para qualquer estacionário fraco e estacionário forte usando parcelas ACF e PACF. Para calcular o valor de p (atraso AR), é utilizado o lote PACF, e para o valor de q, o MA é calculado utilizando-se o plot ACF.

3.2 Parcela de correlação automática parcial

A identificação de um modelo AR é muitas vezes melhor feita com o PACF. Para um modelo AR, o PACF teórico "desliga" além da ordem do modelo. A frase "shuts off" significa que, em teoria, as autocorrelações parciais são iguais a 0 além desse ponto. Colocando de outra forma, o número de autocorrelações parciais não zero dá a ordem do modelo AR. Pela "ordem do modelo", queremos dizer o lag mais extremo de x que é usado como preditor (Kaur et al. , 2020).

A correlação parcial é de natureza condicional e útil na detecção da ordem do processo AR. A correlação entre observações em dois pontos de tempo, dado que consideramos que ambas as observações estão correlacionadas a observações em outros pontos de tempo. O PACF é usado para determinar os termos utilizados no modelo AR. Apenas termos significantes serão escolhidos. O número de termos determina a ordem do modelo.

O enredo PACF de casos confirmados é mostrado na Figura 6, e no componente AR, p é de ordem 3 porque o 4º lag está dentro do intervalo de confidência. Este valor de p obtido é utilizado no modelo ARIMA para casos confirmados.

O enredo PACF de casos recuperados é mostrado na Figura 7, e no componente AR, p é de ordem 2 porque o 3º lag está dentro do intervalo de confidência. Esse valor de p obtido é utilizado no modelo ARIMA para casos recuperados.

O enredo pacf de casos de morte é mostrado na Figura 8, e no componente AR, p é de ordem 3 porque o 4º lag está dentro do intervalo de confidência. Esse valor de p obtido é utilizado no modelo ARIMA para casos de morte.

3.3º plano de correlação doAuto

A identificação de um modelo de MA é muitas vezes melhor feita com o ACF em vez do PACF. Para um modelo MA, o PACF teórico não shut off, mas, em vez disso, tapers para 0 de alguma forma. Um padrão mais claro para um modelo MA está no ACF. A ACF terá autocorrelações não zero apenas em lags envolvidos no modelo (Dhiman e Kaur, 2019b).

A correlação automática é a correlação entre as observações no local de tempo atual e as observações em pontos de tempo anteriores. É a semelhança entre as observações em função do intervalo de tempo entre elas. ACF é usado para determinar os termos no modelo DE.

O enredo ACF de casos confirmados é mostrado na Figura 3 e claramente os valores de correlação são previstos, assim os dados exibem componente estacionário e MA fraco, e q é da ordem 1. Este valor obtido de q é utilizado no modelo ARIMA para casos confirmados.

Além disso, a ACF dos casos recuperados é mostrada na Figura 4 e claramente os valores de correlação são previstos, assim os dados apresentam estacionários fracos no caso de casos recuperados e componente MA, e q é da ordem 4. Este valor obtido de q é utilizado no modelo ARIMA para casos recuperados.

Em seguida, o enredo acf de casos de morte é mostrado na Figura 5 e claramente valores de correlação são previstos novamente, assim os dados exibem fracos estacionários em casos de morte também e o componente MA, q é da ordem 1. Este valor obtido de q é usado no modelo ARIMA para cases de morte.

3.4 Teste e dados de trem

Os dados são divididos em treinamento para datas que começam em 21 de janeiro de 2020 até 30 de junho de 2020 e os dados do teste são reservados para data de início de 1º de julho de 2020 a 12 de agosto de 2020. Dados de treinamento

Figura3 ACF parcela de casos confirmados

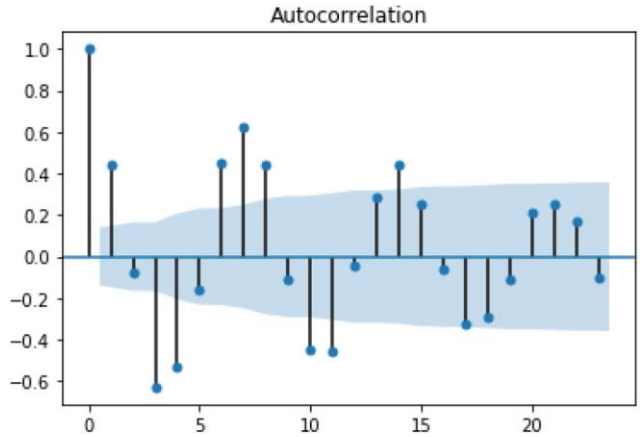


Figure4 ACF parcela de casos recuperados

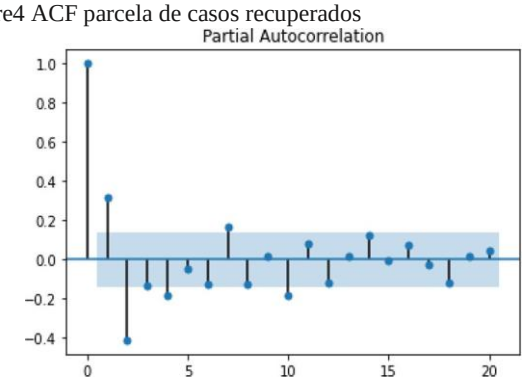


Figura5 ACF parcela de casos de morte

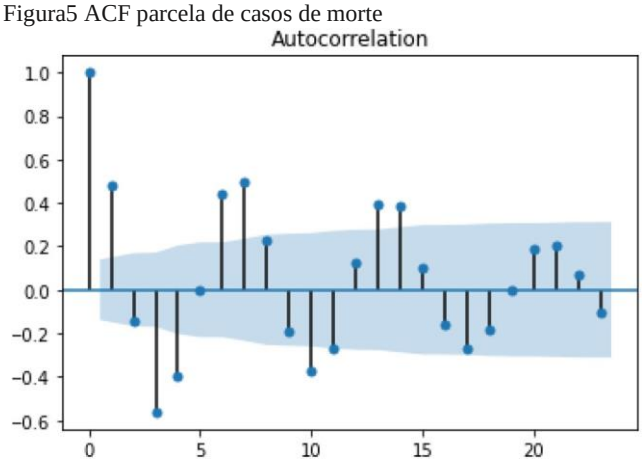


Figura6 PACF parcela de casos confirmados

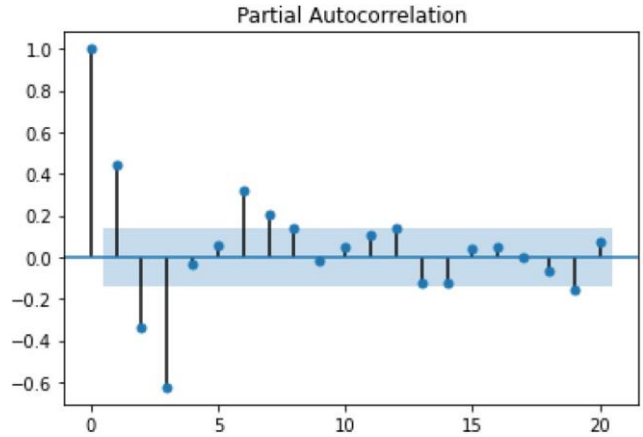
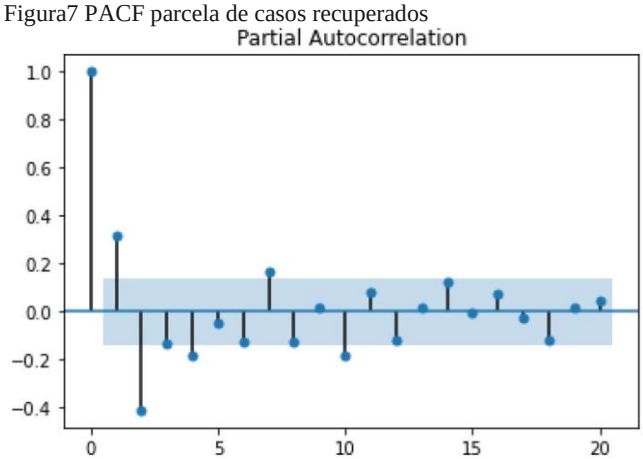


Figura7 PACF parcela de casos recuperados



ajuda na identificação e fitação de modelos certos e os dados de teste são usados para validar o mesmo.

No caso de dados de séries temporidas, os dados de teste são a parte mais recente da série para que o pedido nos dados seja preservado.

Figura8 PACF parcela de casos de morte

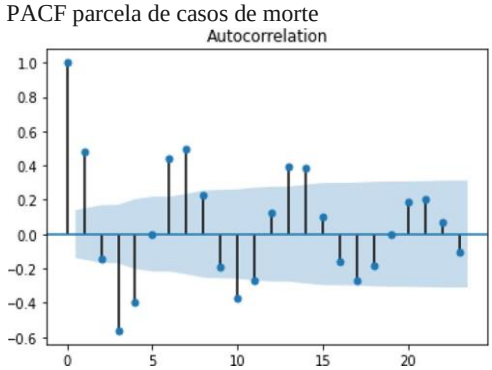


Tabela 2 Resultados do modelo ARIMA de casos confirmados, recuperados e de morte

3.5 Avaliação de modelo médio móvel integrado-regressivo

Os valores alcançados de p e q de parcelas ACF e PACF para os três casos foram utilizados para aplicar separadamente o modelo

ARIMA com diferenciamento (d) de 1. A valorização da probabilidade e a AIC determina o desempenho dos modelos. Em geral, o maior valor de probabilidade de log e o valor mínimo de AIC são considerados melhores valores de parâmetro para a avaliação do modelo. Uma vez que as previsões são feitas, uma medida que é amplamente utilizada é oMAPE.

Os modelos foram otimizados selecionando valores apropriados de p, d e q que resultaram em valores mínimos de AIC e máxima probabilidade de log. Como mostrado na Tabela 2, é

Dep. variável variável	Parâmetros do modelo	Probabilidade de registro	AIC	p-valor	Coeficientes
D.Confirmado	ARIMA (2,1,0)	1733.128	3474.256	0.00 0.00	em. L1. D.Confirmado: 0.53 em diante. L2. D.Confirmado: 0.45
D.Recuperado	ARIMA (2,1,0)	1807.762	3623.523	0.00 0.00	em. L1. D.Recuperado: 0,52 em diante. L2. D.Recuperado: 0.45
Mortes em D.	ARIMA (2,1,2)	1334.813	2681.626	0.00 0.08 0.00 0.05	faria. L1. D.Mortes: 1.23 ar. L2. D.Mortes: 0,2 ma. L1. D.Mortes: 0,54 ma. L2. D.Mortes: 0.17

Figura9 Padrão residual, histograma, Q-Q normal e parcelas de correlogram

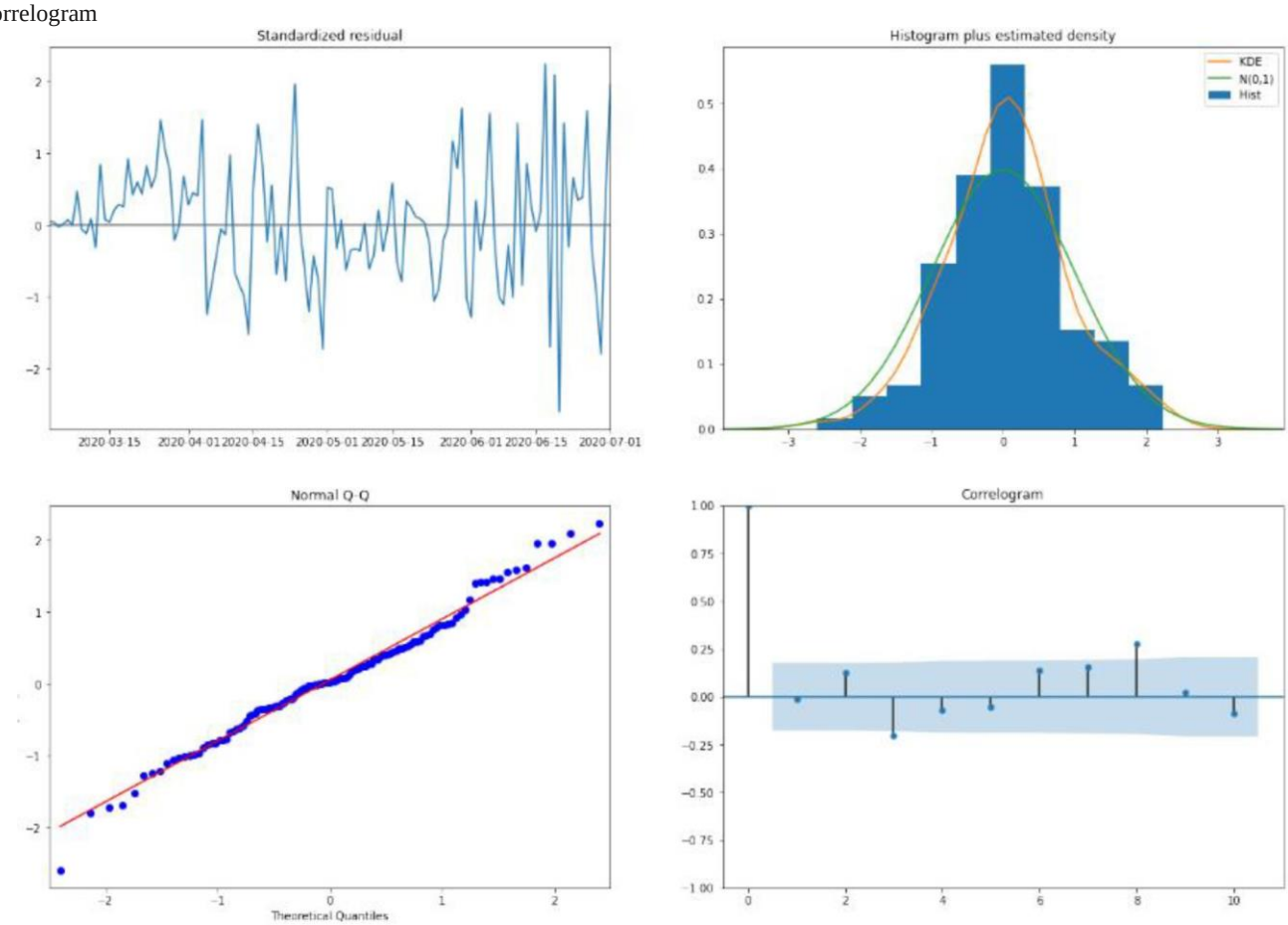
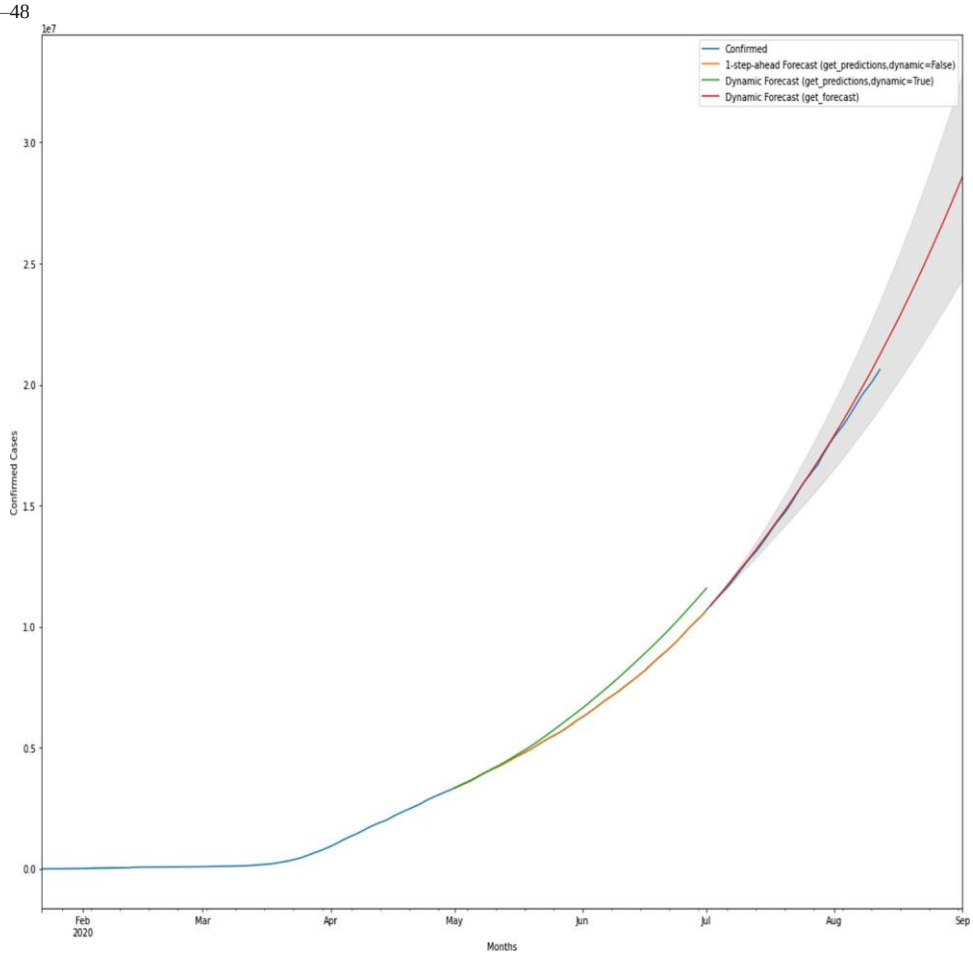


Figura10 Resíduo padrão, histograma, q-q normal e parcelas de correlogram



observou-se que os valores p dos parâmetros AR e MA de todos os tipos de casos implicam que os modelos possuem parâmetros significantes. O menor valor AIC para casos confirmados passou a ser 3474,25 e ARIMA (2, 1, 0) é um modelo otimizado para esses casos. Da mesma forma, o menor value AIC para casos recuperados passou a ser 3623.5225, e a ARIMA (2, 1, 0) é um modelo otimizado para os casos recuperados, e o menor valor AIC para casos de morte passou a ser 2681,62 e ARIMA (2, 1, 2) é um modelo otimizado para esses casos.

Assim, os parâmetros definidos para o modelo The ARIMA para cada um dos casos confirmados, recuperados e de óbito são otimizados levando em conta os valores da AIC. Além disso, os valores dos coefficients ar e MA são retratados nos figures sumários de resultado acima. Os valores p de variáveis inferiores a 5% decidem a importância da variável para o modelo. Assim, ARIMA (2, 1, 0) para confirmed, ARIMA (2, 1, 0) para recuperado e ARIMA (2, 1, 2) para casos de morte são os modelos mais otimizados que são selecionados com base nos valores AIC.

3.6 Médias de moving auto-regressivas sazonais com avaliação de modelo de regressor exógeno

Semelhante ao modelo ARIMA, o modelo SARIMAX é aplicado no conjunto de dados, o parâmetro de sazonalidade é mantido 12 e os demais parâmetros p, d e q são otimizados para cada um dos casos.

O menor AIC é 2648,74 para casos confirmados e modelo SARIMAX (1, 1, 2) (2, 1, 2, 12) ou seja, p, d, q valores como 2, 1, 2 é o melhor modelo. Uma vez que o modelo é fitoado, as parcelas diagnósticas de resíduos como mostrado na figura são usadas para findar qualquer correlação nos dados. As tramas histogramas, gráficos residuais padronizados e Q-Q normais retratam se componentes deixados de fora têm ruído puro ou quaisquer componentes previsíveis. Não existe ruído e correlação nas parcelas e, portanto, nenhum componente local ou previsível existe e o componente de ruído é o único componente deixado de fora.

As análises da série temporal mostram medidas estatísticas significantas para dados COVID-19 e verifica a Figura 9 se a correlação significativa de casos confirmados, recuperados e de morte existe e exibe os lotes residuais para todos os tipos de casos. Observa-se um desvio de resíduos na parcela e indica que os erros estão próximos do normal e, portanto, a suposição de normalidade é seguida com o histograma residual. A variância constante é satisfiada pelo modelo como shown no gráfico de resíduos e correlogram. Os lotes de resíduos exibem a não correção dos dados

O enredo abaixo na Figura 10 delineia a previsão de 1º de julho de 2020 a 12 de agosto de 2020, e prevê dados até 1º de setembro de 2020 para casos confirmados. As linhas verdadeiras dinâmicas e falsas e dinâmicas são representadas separadamente com linhas laranja e verde na trama e a previsão está dentro do intervalo de confidência representado com uma linha vermelha. A tonalidade cinza é o nível de confidência que determina que os casos não

excedem o nível de confidência para esse intervalo de tempo previsto em particular. O valor dos valores previstos está próximo dos valores reais para o período até 12 de agosto de 2020 e também o modelo foi capaz de prever valores para o período de duas semanas. Essa semelhança dos dados previstos com dados reais é clara a partir da trama e também o enredo retrata uma tendência crescente sugerindo um aumento acentuado nos casos consícidos do COVID-19 e casos de morte com tempo; no entanto, a taxa de recuperação é maior do que a taxa de mortalidade. Assim, espera-se uma baixa taxa de mortalidade nos próximos meses.

3,7 Erro percentual absoluto médio

A previsão para os dados de teste para casos confirmados é de 2,37%, o que é considerado excelente no caso dos modelos de séries temporáticas. Da mesma forma, a MAPE para a previsão de casos recuperados nos dados do teste é de 4,21% e a previsão de casos de óbito nos dados do TEST é de 1,51%.

4. Conclusão

As inferências dos modelos de previsão de séries temporáticas ARIMA e SARIMAX foram efficientes para produzir resultados aproximados exatos. Os resultados de previsão indicam que uma tendência crescente é observada e há um aumento elevado nos casos de COVID-19 em muitas regiões e países que podem enfrentar um de seus piores dias a menos e até que medidas sejam tomadas para conter a propagação dessa doença rapidamente. O padrão do aumento da disseminação do vírus nesses países está exatamente imitando alguns dos países da primeira adoção do COVID-19, como Itália e EUA. Além disso, os números obtidos dos modelos são data especific para que a tendência mais recente possa ser analisada com a ajuda de modelos de séries temporais em tempo real.

O escopo futuro do estudo envolve a análise da previsão de casos COVID-19 com modelos como memória de curto prazo e, em seguida, comparação de resultados com modelos de séries temporais abordados neste artigo, como ARIMA e SARIMAX, que ajudariam a prever mais ondas potenciais de pandemia. Outra phase da análise COVID-19 em curso é a construção do classe COVID-19 usando PyTorch e determinando diferenças-chave com abordagens de séries temporais.

Referências

Abdulmajeed, K., Adeleke, M. e Popoola, L. (2020), "Previsão on-line de casos COVID-19 na Nigéria usando dados limitados", Dados em Breve, Vol. 30, p. 105683, doi: 10.1016/j.dib.2020.105683.

Arunraj, N.S., Ahrens, D. e Fernandes, M. (2016), "Aplicação do modelo SARIMAX para prever vendas diárias na indústria do varejo de alimentos", International Journal of Operations Research and Information Systems, Vol. 7 No. 2, pp. 1-21, doi: 10. 4018/ijoris.2016040101.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S. e Ciccozzi, M. (2020), "Aplicação do modelo ARIMA no conjunto de dados epidêmico COVID-2019", Dados em Breve, Vol. 29, p. 105340, doi: 10.1016/j.dib.2020.105340.

Chi, W.L. (2018), "Previsão de preços das ações com base na análise de séries temporais", AIP Conference Proceedings, 1967 (maio). doi: 10.1063/1.5039106.

Contreras, J., Espinola, R., Nogales, F.J. e Conejo, A.J. (2003), "Modelos ARIMA para prever os preços da eletricidade no dia seguinte", Transações IEEE em Sistemas de Energia, Vol. 18 No. 3, pp. 1014-1020, doi: 10.1109/TPWRS.2002.804943.

Dhiman, G. (2019), "ESA: uma abordagem híbrida de otimização metaheurística bioinscrito para problemas de engenharia" , Engenharia com Computadores, doi: 10.1007/s00366-01900826-w.

Dhiman, G. (2020), "MOSHEPO: uma abordagem multi-objetiva híbrida para resolver problemas econômicos de despacho de carga e micro-grade" , Inteligência Aplicada, Vol. 50 No. 1, pp. 119-137, doi: 10.1007/s10489-019-01522-4.

Dhiman, G. e Kaur, A. (2018), "Spotted hyena optimizer for resolvendo problemas de projeto de engenharia", Proceedings - 2017 International Conference on Machine Learning and Data Science, MLDS 2017, 2018-Janeiro, pp. 114-119. doi: 10.1109/MLDS.2017.5.

Dhiman, G. e Kaur, A. (2019a), "Um algoritmo híbrido baseado em enxame de partículas e otimizador de hienas para otimização global", Avanços em Sistemas Inteligentes e Computação, Springer Singapura. doi: 10.1007/978-981-13-1592-3_47.

Dhiman, G. e Kaur, A. (2019b), "STOA: um algoritmo de otimização baseado em bioinscrito para problemas de engenharia industrial", Aplicações de Engenharia de Inteligência Artificial, Vol. 82 Não. Junho de 2018, pp. 148-174, doi: 10.1016/j. engappai.2019.03.021.

Fattah, J., Ezzine, L., Aman, Z., El Moussami, H. e Lachhab, A. (2018), "Previsão de demanda using modelo ARIMA", InternationalJournalofEngineeringBusinessManagement , Vol.10,pp.1-9,doi:10.1177/1847979018808673.

Garg, M. e Dhiman, G. (2020), "Uma nova abordagem de recuperação de imagem baseada em conteúdo para a aplicação de classificação usando recursos GLCM e variantes LBP fundidas de textura" , Neural Computing and Applications, Springer, Londres, 0123456789. 10.1007/ s00521-020-05017-z.

Kaur, S., Awasthi, L.K., Sangal, A.L. e Dhiman, G. (2020), "Algoritmo de enxame tunicate: um novo paradigma metaheurístico baseado em bioinscrito para otimização global", Aplicações de Engenharia da Inteligência Artificial, Vol. 90 No. Novembro de 2019, pp. 103541,doi: 10.1016/j.engappai.2020.103541.

Lalmuanawma, S., Hussain, J. e Chhakchhuak, L. (2020), "Aplicações de aprendizado de máquina e inteligência artificial para COVID-19 (SARS-CoV-2) pandemia: uma revisão", Chaos, Solitons e Fractais,p.139,doi:10.1016/j.chaos.2020.110059.

Li, S. e Li, R. (2017), "Comparação da previsão do consumo de energia em Shandong, China usando o modelo ARIMA, modelo GM e modelo ARIMA-GM", Sustentabilidade (Suíça) , Nº 7, pp. 9, doi: 10.3390/su9071181.

Malki,Z.,Atlam, E.S., Hassanien,A.E., Dagnew,G.,Elhosseini, M.A. e Gad, I. (2020), "Associação entre dados meteorológicos e covid-19 previsão de taxa de mortalidade: machine learning aproxima", Caos, Solitons e Fractals, Vol. 138, p.110137,doi:10.1016/j.chaos.2020.110137.

Nabi, K.N. (2020), "Previsão de pandemia COVID-19: uma análise baseada em dados", Caos, Solitons e Fractais, Vol. 139, p. 110046, doi: 10.1016/j.chaos.2020.110046.

Nelson, B.K. (1998), "Metodologia estatística: v. Análise da série temporal utilizando modelos autoregressivos de média móvel integrada (ARIMA)", Medicina de Emergência Acadêmica, Vol. 5 No. 7, pp. 739-744, doi: 10.1111/j.1553-2712.1998. tb02493.x.

Reddy, JR, Ganesh, T., Venkateswaran, M. e Reddy, P.R.S. (2017), "Previsão de chuvas médias mensais em Rayalaseema",, International Journal of Current Research e Resenha,Vol.9No.24,pp.20-27,doi :10.7324/ijcrr.2017.9244.

Singh, P. e Dhiman, G. (2018), "Um modelo híbrido de previsão de séries temporais difusas baseado em abordagens de computação granular e otimização bioinspired" , Journal of Computational Ciência,Vol.27,pp.370-385,doi : 10.1016/j. jogos.2018.05.008.

Soebiyanto, R.P., Adimi, F. e Kiang, R.K. (2010), "Modelando e prevendo transmissão sazonal de influenza em regiões quentes usando parâmetros climatológicos", PLoS One, Vol. 5 No. 3, pp. 1-10., doi: 10.1371/journal. pone.0009450.

Xue, M. e Lai, C.H. (2018), "Da análise de séries temporificadas a uma equação diferencial ordinária modificada", Journal of Algorithms & Computational Technology, Vol. 12 No. 2, pp. 85-90, doi: 10.1177/1748301817751480.

Yonar, H. (2020), "Modelagem e previsão para o número de casos da pandemia COVID-19 com os modelos de estimativa de curvas, os métodos de alisamento de Box-Jenkins e de suavização exponencial",, Eurasian Journal of Medicine and Oncology, Vol. 4 No. 2, pp. 160-165, doi: 10.14744/ ejmo.2020.28273.

Autor correspondente

Kamalpreet Singh Bhangu pode ser contatado em: pesquisador.kamalpreet.bhangu@gmail.com