2020 International Conference on Computational Science and Computational Intelligence (CSCI) | 978-1-7281-7624-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/CSCI51800.2020.00150

Time Series Forecasting of COVID-19 Infections in United Arab Emirates using ARIMA

Leila Ismail, Member, IEEE, Shaikhah Alhmoudi, and Sumyah Alkatheri
Distributed Computing and Systems Research Laboratory
College of Information Technology, United Arab Emirates University, Al-Ain, Abu Dhabi, 15551, United Arab Emirates
Correspondence: Leila Ismail (email: leila@uaeu.ac.ae) 0000-0003-0946-1818

***Abstract*— Machine learning time series models have been used to predict COVID-19 pandemic infections. Based on the public dataset from Johns Hopkins, we present a novel framework for forecasting COVID-19 infections. We implement our framework for the United Arab Emirates (UAE) and develop autoregressive integrated moving average (ARIMA) time series forecast model. To the best of our knowledge, this is the only study to forecast the infections in UAE using the time series model.**

Keywords—COVID-19, coronavirus, machine learning, time series, autoregressive integrated moving average (ARIMA)

I. Introduction

The novel coronavirus disease (COVID-19) was first identified in Wuhan, China in December 2019. World Health Organization (WHO) declared the disease as a pandemic three months after it was discovered [1]. To date over 38 million people have been infected by the virus [2]. The virus can be life-threatening, particularly for elderly people or the ones suffering from chronic illnesses such as diabetes, hypertension, respiratory diseases, and cancer [3]. The outspread of the virus not only imposes health risks but affects the economy worldwide. To reduce the infection spread and to ensure the safety of the residents, governments impose several social practices and national strategies, such as social distancing, travel restrictions, a reduced percentage of working staff in the offices, and e-learning. Such measures and strategies imposed during previous pandemics have shown a potential reduction of the disease spread and death cases by 50% [4, 5].

Since the epidemic outbreak, several studies have focused on machine learning time series models to predict the number of COVID-19 infections in different countries [6–14]. This is to design and regulate effective outbreak-mitigating strategies. Apart from the government, the time series forecast models can aid healthcare organizations and allied health professionals such as doctors, nurses, pathologists, and pharmacists to plan resources in health personnel and facilities effectively. For instance, based on forecasted values for the coming days, hospitals can efficiently plan on the number of beds and ventilators required. The pharmacists can increase the stock of medications if the number of COVID-19 infections is likely to increase based on the forecasted values by the time series models. The machine learning time series models are developed based on learning from the spatial distribution of the COVID-19 infections trend over time.

In this paper, we develop a framework for machine learning time series modeling for forecasting COVID-19 infections. We implement the framework using the United Arab Emirates (UAE) number of infections collected from the Johns Hopkins COVID-19 dataset [15]. To obtain reliable predictions leading to an effective pandemic precautionary measure, we develop the autoregressive integrated moving average (ARIMA) model [16] using the proposed framework. The developed model allows government and healthcare organizations to estimate the rate of infection growth. We evaluate the performance of the developed model in terms of Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). MAPE is used as an evaluation metric in addition to RMSE, because it allows the comparison of the model's absolute percentage of variations between countries having heterogeneous scales of infections.

The main aim of this study is two-fold: 1) to provide a framework for machine learning modeling, and 2) to model the spread of COVID-19 infections for the UAE providing an implementation of the framework, based on the spatial distribution of the infections which vary over time based on social and health policies put in place. We develop the model using a cross-sectional study of 265 observations. The model's main objective is to guide policymakers towards effective pandemic precautionary measures. The framework can be used to develop the ARIMA model or other models for other countries.

The rest of the paper is organized as follows. In Section II, we describe the dataset used for the model development, investigate the data distribution trend of infections, along with a description of ARIMA, followed by the description of the development framework. Section III discusses the conducted experiments and provides results analysis and discussion. Section IV concludes the paper.

II. Methodology

A. Dataset

The COVID-19 data used in this paper is collected from the Johns Hopkins COVID-19 dataset [15]. The dataset is updated daily. The dataset includes confirmed infection cases. We extracted the data related to the number of confirmed COVID-19 infections in the UAE. Fig. 1 shows the data trend for COVID-19 infections spread from 22/01/2020 to 12/10/2020. The infection distribution shows an exponential trend over the considered period of time. Based on the study conducted in [17], the time series model ARIMA is the best suited for an

Identify applicable funding agency here. If none, delete this text box.

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00150

exponential distribution of infections to predict future infections.

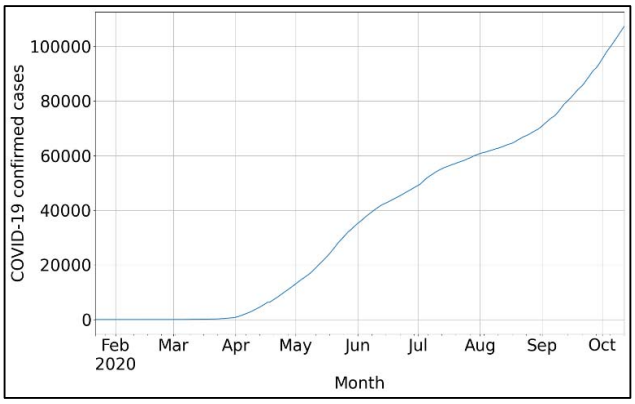

Fig. 1. Number of COVID-19 confirmed cases as of 12 October 2020

B. *Proposed Infections Forecast Framework*

Fig. 2 shows our proposed framework for the machine learning time series models' development. The framework consists of the following stages:

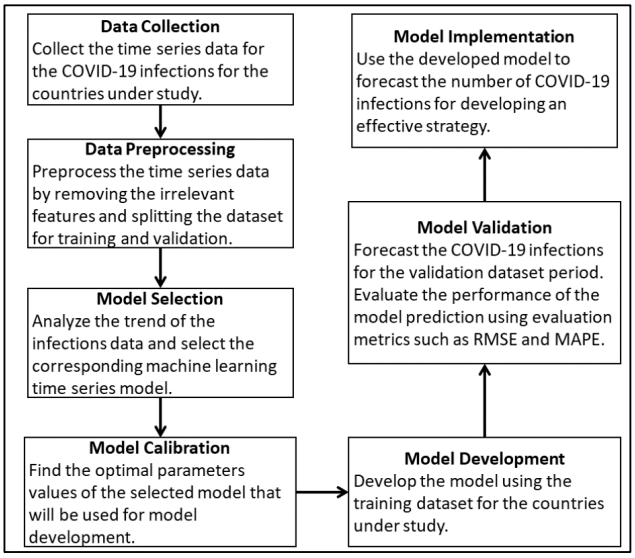

Fig. 2. Proposed infections forecast framework

- Stage 1: Data Collection. One can consider data collected from a hospital database or a survey or publicly available open sources for the countries under study.
- Stage 2: Data preprocessing. The features required for the model are extracted from the database. The data should be cleaned so that a missing value of infections is inserted by including the same value as that of the previous available day. At this stage, the data is split into training and validation datasets.
- Stage 3: Model Selection. The trend of the data distribution is analyzed and the corresponding machine learning time series model is selected [17].
- Stage 4: Model Calibration. The optimal parameters' values of the selected model from stage 3 are computed.
- Stage 5: Model Development. The model is developed using the training dataset which is determined in stage 2, and the optimal parameters computed at stage 4.
- Stage 6: Model Validation. The model is validated using the validation dataset resulted from stage 2, and computing evaluation metrics, such as RMSE and MAPE.
- Stage 7. Model Implementation. The developed model is used to forecast the number of COVID-19 infections; aiming to put in place an effective plan for needed resources.

C. *ARIMA for COVID-19 Forecasting*

Autoregressive integrated moving average (ARIMA) is a combination of the auto-regressive (AR) and the moving average (MA) models [16]. AR formulates a linear model that forecasts the number of COVID-19 infections based on the previous (lagged) infections as stated as shown in (1). On the other hand, MA formulates a linear model that forecasts the number of COVID-19 infections based on the previous (lagged) forecast errors as stated in (2). After adding differencing to AR and MA models, the ARIMA model can be expressed as stated in (3). Differencing is the process of transforming a non-stationary time series to a stationary time series [18]. To implement ARIMA for time series forecasting, the data trend should be stationary.

$$Infections_T = \alpha + \beta_1 Infections_{T-1} + \beta_2 Infections_{T-2} + \cdots + \beta_p Infections_p + \varepsilon_T \quad (1)$$

$$Infections_T = \alpha + \varepsilon_T + \emptyset_1 \varepsilon_{T-1} + \emptyset_2 \varepsilon_{T-2} + \cdots + \emptyset_q \varepsilon_{T-q} \quad (2)$$

$$Infections_T = \alpha + \beta_1 Infections_{T-1} + \beta_2 Infections_{T-2} + \cdots + \beta_p Infections_p + \emptyset_1 \varepsilon_{T-1} + \emptyset_2 \varepsilon_{T-2} + \cdots + \emptyset_q \varepsilon_{T-q} + \varepsilon_T \quad (3)$$

where α, β, and Ø are the regression coefficients, $Infections_T$ represents the number of COVID-19 infections at time T, and ε_T represents the forecast error at time T.

The autoregressive, integrated, and the moving average components, designated by p, d, and q respectively, are explicitly specified in the ARIMA model as parameters. ARIMA model is denoted by ARIMA(p,d,q). The variable p (known as lag order) represents the number of lagged observations on the COVID-19 infections that have to be included in the model. This captures the AR term in the ARIMA model. For instance, if the value of p is set to 10, then the model will use the infections data for 10 previous days for the forecast. The variable d (known as the order of

differencing) represents the number of times the time series data is differenced to obtain stationary time series. This captures the integrated component in the ARIMA model. If the trend of the time series is constant, then no differencing is required, and the value of d is set to 0. For time series data having a linear data trend, first-order differencing is required, and the value of d is set to 1. For an exponential data trend, a second-order differencing is required, and the value of d is set to 2. The variable q (known as the order of moving average) represents the number of lagged error terms that have to be included in the model. This captures the MA component in the ARIMA model.

III. Performance Analysis

In this section, we develop the ARIMA model for forecasting COVID-19 infections in UAE and evaluate the performance of the model in terms of RMSE and MAPE.

A. Experiments

To evaluate the ARIMA model, we create a dataset for the UAE from Johns Hopkins data. We use 70% of the dataset (i.e., 22 January – 24 July 2020) for training to develop the model and 30% of the dataset (25 July – 12 October 2020) for validation. To validate the model, we first forecast the number of COVID-19 infections over time based on the developed model, then we compare the predicted values with the actual ones. In addition, we also present the predicted values for the validation dataset with a 95% confidence interval. To obtain the values of ARIMA model parameters, we conduct two sets of experiments: 1) to determine the order of differencing for the ARIMA model (value of the parameter d). This is by inspecting whether the time series is stationary or not. 2) to determine the lag order and order of moving average for the ARIMA model (values of the parameters p and q). This is after differencing the time series in case the series is non-stationary.

To determine the value of parameter d, we perform a visual test using the ACF autocorrelation function (ACF) and partial autocorrelation function (PACF) plots [19] as well as statistical augmented Dickey-Fuller (ADF) test [20]. The ACF plot demonstrates the autocorrelation between the number of infections at time T ($Infections_T$) and the number of infections at time T-k ($Infections_{T-k}$). If $Infections_T$ and $Infections_{T-1}$ are correlated, then $Infections_{T-1}$ and $Infections_{T-2}$ must also be correlated. However, it might be that $Infections_T$ and $Infections_{T-2}$ might be related as well because they both are connected to $Infections_{T-1}$. The PACF plot measures the partial correlation between $Infections_T$ and $Infections_{T-k}$ after removing the effects of lags 1, 2, 3, …, k-1. Consequently, the partial correlation between $Infections_T$ and $Infections_{T-2}$ can be obtained without the influence of $Infections_{T-1}$. If the ACF plot drops to zero quickly, then the data trend is stationary, and no differencing is required (d=0). If the ACF plot reaches to zero relatively slow, then the data trend is non-stationary, and differencing is required. In such a case, the value of d is equivalent to the number of time differencing is performed before making the time series stationary. The ADF test tests the null hypothesis that the time series is non-stationary. The p-value of less than 5% represents that the null hypothesis can be rejected, i.e., the time series is stationary.

To determine the values of the parameters p and q, we perform the visual test using the ACF and PACF plots. Once, the order of differencing is determined, we plot the ACF and PACF plots using the differenced time series. The ACF plot is used to determine the lag order (p) while the PACF plot is used to determine the order of the moving average (q). The number of lags for which the ACF is outside the significant threshold represents the lag order and the number of lags for which the PACF is outside the significant threshold represents the order of moving average.

After determining the values of the ARIMA model parameters, we develop the model using the training dataset. We then evaluate the performance of the developed model using the validation dataset. We evaluate the performance of the model in terms of RMSE and MAPE. The RMSE and MAPE are calculated using (4) and (5) respectively.

$$RMSE = \sqrt{\frac{\sum_{T=1}^{n}(Infections_T - Infections'_T)^2}{n}} \tag{4}$$

$$MAPE = \left(\frac{1}{n}\sum_{T=1}^{n}\frac{|Infections_T - Infections'_T}{Infections_T}\right) * 100\% \tag{5}$$

where $Infections_T$ is the actual value of infections, $Infections'_T$ is the forecasted value of infections and n is the total number of fitted data points.

B. Experimental Results Analysis

Fig. 3 shows our results on the ACF and PACF plots and the ADF tests. It shows the ACF and PACF plots for no differencing, first-order differencing and second-order differencing. As shown in Fig. 3, most of the autocorrelation values for the ACF plot of the original time series data (without any differencing) are above the significance threshold level and the ACF plot is converging very slowly to zero. Consequently, the time series is non-stationary. This is also confirmed by the ADF test as the p-value is greater than 0.05. The time series remains non-stationary even after performing first-order differencing (Fig. 3). The ACF plot for first-order differenced time series converges very slowly towards zero and the p-value is greater than 0.05. When second-order differencing was applied, the non-stationary time series transformed to stationary. This is evident from Fig 3. as the ACF plot converges quickly to zero and also the p-value is 0, i.e., less than 0.05. Consequently, the value of parameter d for the ARIMA model is set to 2 in our experiments. Fig. 3 shows that the first two lag values are outside the significance level for both ACF and PACF plots of second-order differencing (stationary time series). This indicates the most suitable values for lag order (p) and the order of moving average (d) can be set to 2. Consequently, we develop the ARIMA(2,2,2) model to forecast COVID-19 infections in UAE.

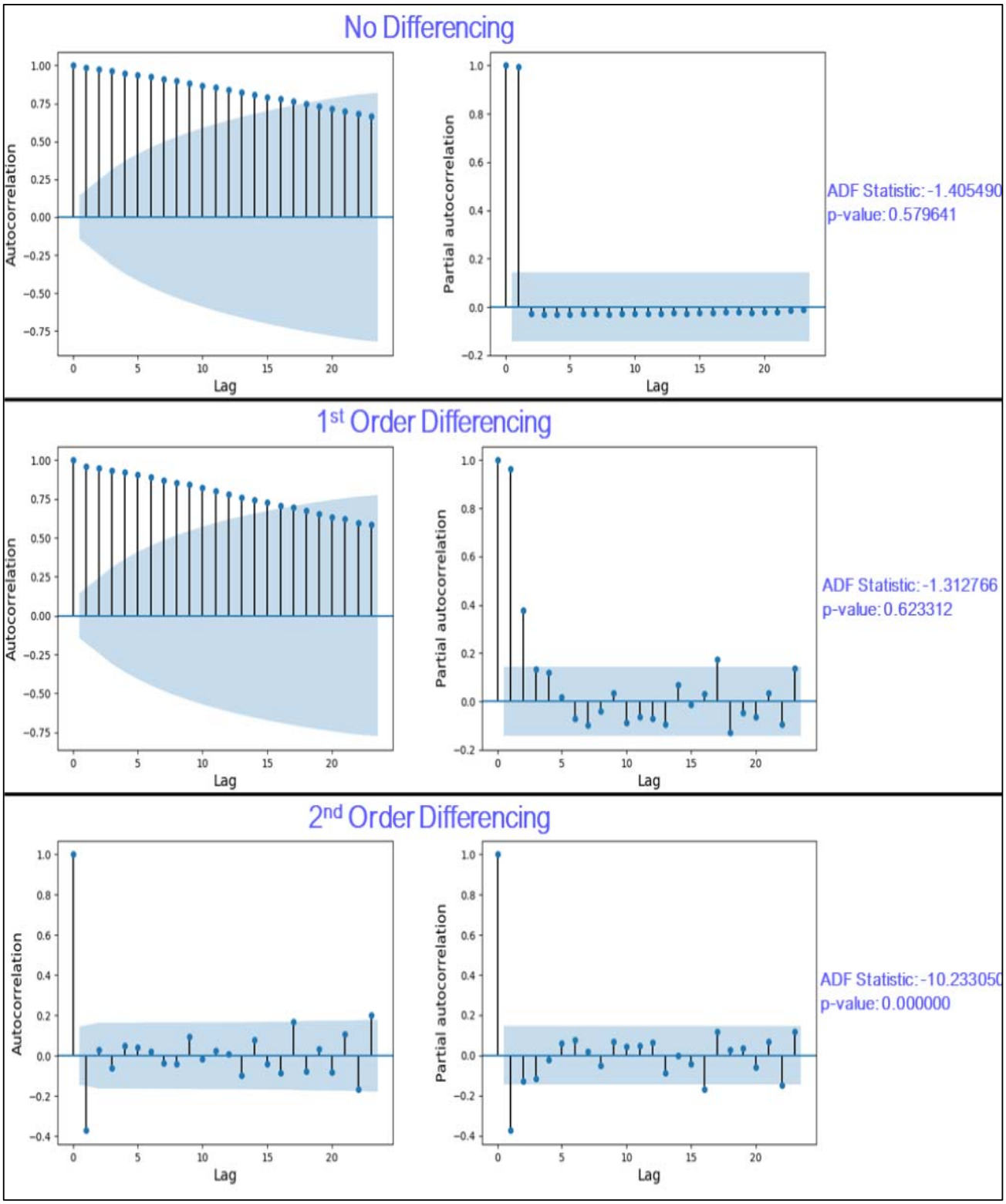

Fig. 3. ACF and PACF plots for no order, first-order and second-order differencing, and the p-value for augmented Dickey-Fuller test

Fig. 4 shows the number of COVID-19 confirmed cases for the training and validation datasets. It also shows the forecast values for the validation dataset using ARIMA(2,2,2). It shows that the number of infections predicted by the model is less than the actual number of infections. Table I shows the RMSE and MAPE value for the developed model.

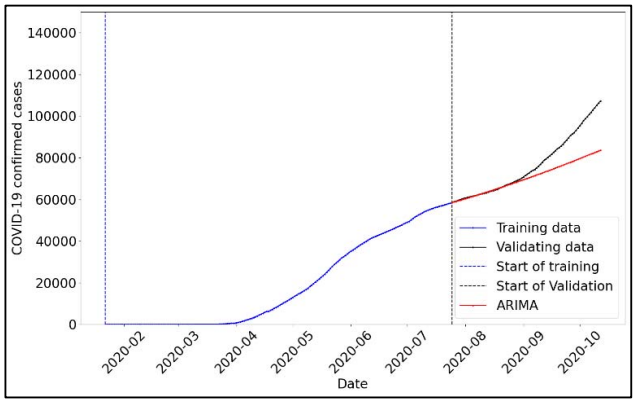

Fig. 4. Prediction of COVID-19 infections in UAE for the validation dataset using ARIMA(2,2,2)

TABLE I. RMSE AND MAPE VALUES FOR COVID-19 FORECASTING USING ARIMA(2,2,2) MODEL

	ARIMA(2,2,2)	
UAE	**RMSE**	**MAPE**
	9284.095	6.385

The forecasted number of COVID-19 infections using ARIMA(2,2,2) along with the 95% confidence interval is shown in Fig. 5. The fig. shows that the actual number of infections for the validation dataset are within the confidence interval.

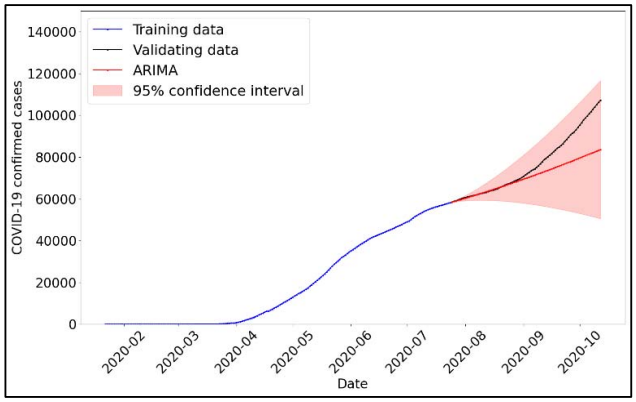

Fig. 5. 95% confidence interval for the predicted COVID-19 infections in UAE for the validation dataset using ARIMA(2,2,2)

IV. CONCLUSIONS

Machine learning time series methods have been used to predict the number of infections of COVID-19 to tackle the pandemic. In this paper, we propose prediction framework for COVID-19 infections. We develop the framework along with the ARIMA machine learning time series model to predict the infections spread in the United Arab Emirates. Our framework is generic and can be used to develop any time series model for the prediction of pandemic infections for any country. The paper describes how the ARIMA model is developed with optimal parameters. The developed model shows an accuracy of 93.615. This accuracy helps policymakers of the country to achieve reliable predictions, to put in place effective precautionary measures, and to plan for needed resources.

REFERENCES

1. World Health Organization COVID-19: A pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019
2. COVID-19 Coronavirus Pandemic. https://www.worldometers.info/coronavirus/. Accessed 15 Oct 2020
3. Jordan RE, Adab P, Cheng K (2020) Covid-19: risk factors for severe disease and death. Br Med J Publ Gr. https://doi.org/https://doi.org/10.1136/bmj.m1198
4. Halder N, Kelso JK, Milne GJ (2010) Analysis of the effectiveness of interventions used during the 2009 A/H1N1 influenza pandemic. BMC Public Health 10:
5. Ahmed F, Zviedrite N, Uzicanin A (2018) Effectiveness of workplace social distancing measures in reducing influenza transmission: a systematic review. BMC Public Health 18:
6. Tandon H, Ranjan P, Chakraborty T, Suhag V (2020) Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. arXiv Prepr arXiv200407859
7. Elmousalami HH, Hassanien AE (2020) Day level forecasting for Coronavirus Disease (COVID-19) spread: analysis, modeling and recommendations. arXiv Prepr arXiv200307778
8. Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus COVID-19. PLoS One 15:
9. Chimmula VKR, Zhang L (2020) Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals
10. Yonar H, Yonar A, Tekindal MA, Tekindal M (2020) Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods. EJMO 4:
11. Jiang X, Zhao BZ, Jinming C (2020) Statistical Analysis on COVID-19. Biomed J Sci Tech Res
12. Panda M (2020) Application of ARIMA and Holt-Winters forecasting model to predict the spreading of COVID-19 for India and its states. medRxiv
13. Patil R, Patel U, Sarkar T (2020) COVID-19 cases prediction using regression and novel SSM model for non-converged countries. J Appl Sci Eng Technol Educ 3:74–81
14. Ahmar AS, del Val EB (2020) SutteARIMA: Short-

term forecasting method, a case: Covid-19 and stock market in Spain. Sci Total Environ

15. Novel coronavirus (covid-19) cases data - humanitarian data exchange

16. Box GE, Pierce DA (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J Am Stat Assoc 65:1509–1526

17. Ismail L, Materwala H, Znati T, et al (2020) Tailoring Time Series Models For Forecasting Coronavirus Spread: Case Studies of 187 Countries. Comput Struct Biotechnol J. https://doi.org/https://doi.org/10.1016/j.csbj.2020.09.015

18. Solo V (1984) The order of differencing in ARIMA models. J Am Stat Assoc 79:916–921

19. ACF and PACF plots

20. Cheung Y-W, Lai KS (1995) Lag order and critical values of the augmented Dickey--Fuller test. J Bus Econ Stat 13:277–280