Analysis of COVID-19 Data Using Arima Time Series Model

Başak Er
Dokuz Eylul University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Murat Emeç
Dokuz Eylul University
8 PUBLICATIONS **1** CITATION
SEE PROFILE

Mehmet Hilal Ozcanhan
Dokuz Eylul University
38 PUBLICATIONS **81** CITATIONS
SEE PROFILE

ANALYSIS OF COVID-19 DATA USING ARIMA TIME SERIES MODEL

Başak Er[1 *], Murat Emeç[2], Mehmet Hilal Özcanhan[3]

[1,2,3] Dokuz Eylul University, Computer Engineering, Turkey

ORCID ID: 0000-0001-7999-5722, basak.er@ceng.deu.edu.tr

ORCID ID: 0000-0002-9407-1728, murat.emec@deu.edu.tr

ORCID ID: 0000-0002-5619-6722, hozcanhan@cs.deu.edu.tr

**** Corresponding Author***

Abstract

The covid-19 caused a pandemic that affected in global scale for the world since December 2019. Globally, research has focused on issues such as detecting, preventing, mitigating and predicting its causes and consequences. Diverse studies limited to specific geographical areas, countries or situations, as well as global modeling have been carried out. The models created in the studies were able to provide information about the progress of the pandemic and the outcomes it will lead to (Benvenuto, 2020)*. The obtained analytical insight allowed governments to make more accurate decisions and make plans for the future* (Fang Y. N., 2020)*. Prior knowledge of upcoming numbers in infections also helped in controlling the anxiety, about the pandemic. This study provides an analysis of the spread of COVID-19 in five countries, over a five-month period of March-Jul 2020. The analysis shows a similar trend for the whole world. Countries which can be associated with geographical region and population similarities were studied using the Auto Regressive Moving Average (ARIMA) time series model. The ARIMA model analysis results showed that the model proved to be accurate for short-term forecasting, while monthly or annual forecasts had large errors. The margin of error in the analysis of COVID-19 cases based on countries was observed to increase for long term; to a point where a satisfactory forecast could not be made. Updating the model frequently with new data proved to produce results, with very small errors. It can be concluded that ARIMA model is capable of giving accurate short-term forecast numbers with very little errors, on COVID-19 pandemic.*

Keywords: *ARIMA, COVID-19, Forecast, Epidemic*

1. Introduction

1.1. COVID-19

The COVID-19 is infected which has come from Wuhan in China on December 2019. Spread across the globe and within six months led to an unprecedented situation. Since the first report of the Corona (COVID-19) has infected millions of people and also provoked as many people, according to the World Health Organization (WHO) (WHO, 2020). SARS and MERS corona viruses which have been occurred examples, they have not seen as infectious and persistent as COVID-19, as is known. Nowadays, there is no clearly cure way or process. Furthermore 180 countries are suffering from the virus. The virus is extremely infectious in its present form and causes death due to respiratory failure.

In Turkey, cases have started to be reported with the first Covid-19 patient on March of 10th in 2020. As the severity of the spread of the virus was revealed, partial lockdown was enforced by the Government of Turkey in order to control the pandemic. The first isolation was announced and was progressively extended on April 10th through to June. Turkey is one of the top ten most affected countries in the world and is now the eighth most affected country in the world (Coronavirus Update, 2020) Statistical analysis of the pandemic has become a research area in our country, as in most countries. In our present study, we analyzed the Covid-19 pandemic using a very popular time series analysis tool.

1.2. ARIMA

The Box-Jenkins technique proposed by Box Jenkins is commonly used for analysis of time series (Box GE, 1976) This technique involves models of ARIMA applied to the series that are non-stationary models, but models are rendered stationary with the series difference operation. The basis of the Box-Jenkins approach is to select an ARIMA model that, depending on the nature of the data considered, contains the most suitable parameter but also limited parameter among the various model options.

The models of ARIMA (p, d, q) are obtained by taking the series difference from the degree of d and adding the stabilization process model of ARMA (p, q). In the ARIMA (p, d, q) models, p is the Autoregressive (AR) model degree, q is the moving average (MA) model degree, and d is the number of differences needed to make the series stationary. If the time series is stationary, the ARIMA model is AR (p), MA(q) or ARMA (p, q). (Asteriou & Hall, 2011)

The rapid spread of the Covid-19 epidemic requires steps to be taken at the right time and some immediately. ARIMA is a very useful tool that can provide the early warnings needed. The results of such an analysis can give an idea to the government and communities about how and where the epidemic is progressing; thus, giving them the opportunity to take necessary measures on time at the exact density locations.

In this study, the number of COVID-19 outbreak cases in the France, Germany, Iran, Italy and Turkey between 12.03.2020 and 09.07.2020 were estimated using the Box-Jenkins (ARIMA) model. Although the start date of the outbreak varies by country, the common time frame for the five countries was taken into account. Forecasts predict that cases will increase as the outbreak continues.

2. Related Works

Several models have been used in recent studies to predict the incidence, prevalence and mortality of COVID-19 in China. Li et al., for example, built a function to forecast the ongoing trend with data-based analysis and estimate the size of China's COVID-19 outbreak (Li, 2020). The temporal dynamics of the COVID-19 pandemic in mainland China, Italy and France were analysed by Fanelli and Piazza (Fanelli, 2020). Wu et al. forecast the spread of COVID-19 to determine the effect of the metropolitan-wide quarantine of Wuhan and its neighbors on a national and global scale (Fang X. L., 2020). The algorithm based on patient awareness was developed by Wang et al. to estimate the mortality rate of COVID-19 in real time, using publicly available data (Zhang, 2017). Accumulated confirmed COVID-19 instances can be viewed as data from time series. ARIMA has recently been used to model the COVID-19 outbreak dataset to estimate its epidemiological pattern (Benvenuto, 2020). Finally, Petropoulos and Makridakis have adopted simple time series estimation approaches in another study closely related to this study, using models from the exponential smoothing family to estimate the number of COVID-19 reported cases on a global scale (Petropoulos & Makridakis, 2020). The methodology adopted by the IHME COVID-19 health service usage forecasting team (Team, 2020) was at the core of statistical modeling, which estimated the strain caused by the pandemic in the United States health system by estimating the number of hospital beds, ICU beds and ventilators required in the next four months as well as the number of deaths. Similar approaches to modeling diseases that occur in cyclical or repeated trends such as seasonal influenza have been followed; a number of studies have been published to forecast future outbreaks using time series modeling. An ARIMA (Adhikari & Agrawal, 2013) model was built in (Song, ve diğerleri, 2016) to estimate the monthly incidence of influenza in China for 2012, whereas a time series prediction model (Tempel) for mutation prediction of influenza A viruses was proposed in (Yin, Luusua, Dabrowski, Zhang, & Kwoh, 2020). In addition, there are studies that the ARIMA time series model is used in the detection of many infectious diseases (U, 1986) (Williamson G, 1999).

3. Materials and Methods

3.1. Materials

Confirmed COVID-2019 cases were obtained daily from the Johns Hopkins University official website (University, 2020) from March 12, 2020 to July 9, 2020, and a time series database was created using Microsoft Excel 2019.

The data set covers the dates of 12 March 2020 and 9 July 2020. It includes information from Turkey, Iran, France, Germany and Italy. It contains a total of 13 columns and 598 rows. Columns are the parameters country name, date, total number of cases, daily number of cases, total number of deaths, daily number of deaths, total number of tests, daily number of tests, smoothed daily number of tests, number of cases per million, number of deaths per million, and number of daily deaths per million. The total number of tests parameter contains two hundred and seventy null values and is not used in the model.

The model was created in Python language on Jupyter Anaconda. The *pandas* and *numpy* libraries were used in the analysis of the dataset, while the *seaborn* library was used in the visualization of the dataset. The *statsmodels* library tools were used to perform tests that had to be performed before the ARIMA time series model. The *pmdarima* library provides the basis of our model with the *auto_arima* function.

3.2. Methods

The ARIMA model was built over a period of 120 days (from March 12, 2020 to July 9, 2020). In time series analysis, ARIMA models are important techniques that can be used in autocorrelated data analysis. These models include the autoregressive (AR) model, the moving average (MA) model and the integrated moving average (SARIMA) seasonal autoregressive model (Fattah J). Time series should be station-dependent based on mean and variance before analysis. In recognizing mean stationarity (Cao S, 2013) and in the Box Cox test, Augmented Dickey-Fuller (ADF) is used to understand whether time series according to variance are stationary. Log conversion and differences are remedial methods for stabilizing time series for variance and mean, respectively (Cheung Y-W, 1995).

Using the *Seasonal-decompose* method (StatsModels, 2020), the seasonality of the series was examined and no seasonality was observed. The autocorrelation function (ACF) function and the partial autocorrelation (PACF) function are calculated in the first issue of the ARIMA model. Comparison was made using the officially given daily case, daily death, total case, total death parameters for the 5 countries examined. In addition to the column chart expressing the total number of cases in comparison of countries, column and line charts expressing the course of the daily number of cases by date were drawn.

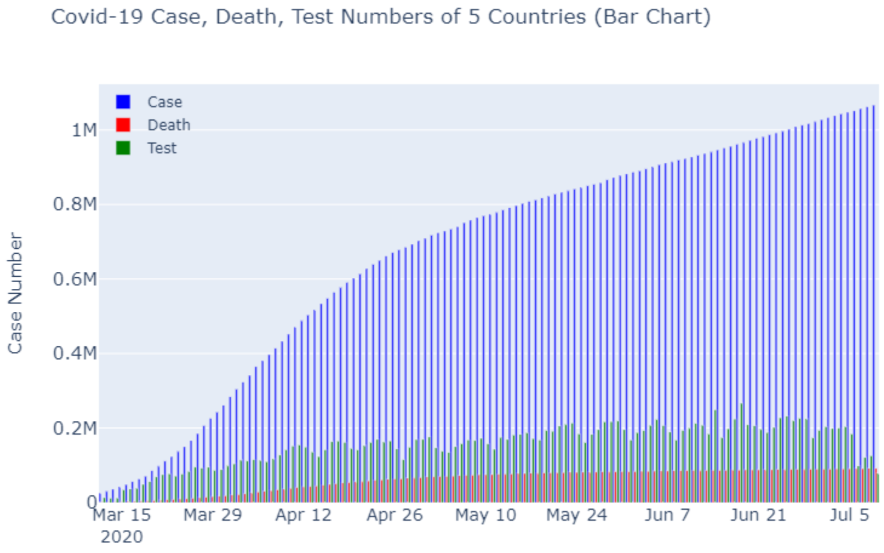

Figure 1: Case, Death, Test Numbers of Five Countries

The distribution and trend of the charts were examined and is shown in figure 1. In Figure 1, the comparison of the total number of cases, the number of deaths and the number of tests for all countries is expressed by a line chart. In the chart, blue lines indicate the number of cases, red lines indicate the number of deaths, and green lines indicate the number of tests. The horizontal axis depends on time, while the vertical axis depends on the total number.

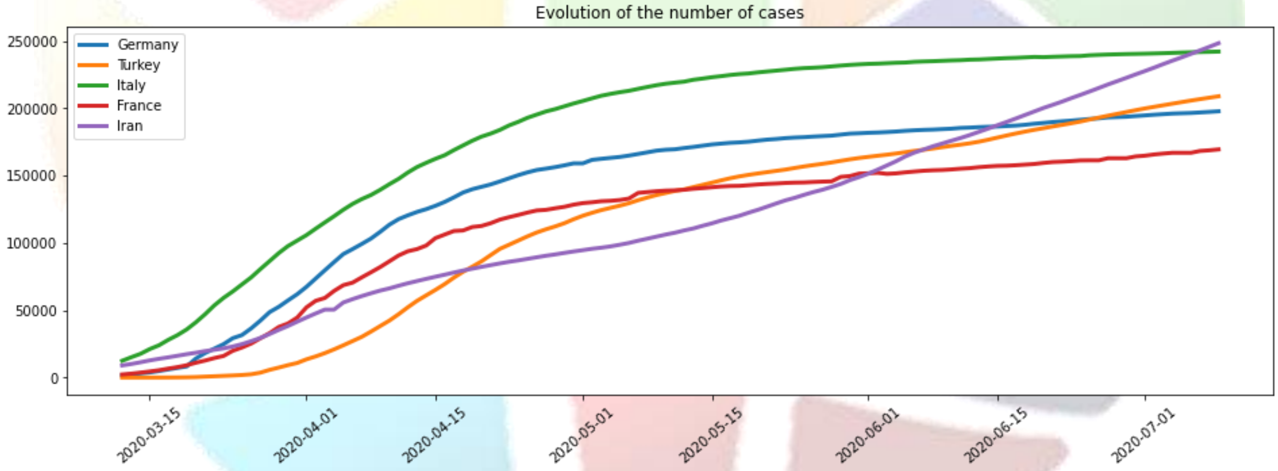

Figure 2: Total Number of Cases on Country Basis

As a result of examining the slope in the line chart, where all countries coexist, it was found that Turkey bears similarities with Germany, especially in the increase in daily cases (Fig 2). As can be seen from the figure, the total number of cases on a country-by-country basis varies according to time. In countries expressed by lines of different colors, especially Turkey, Italy and Germany seem to bear similarities.

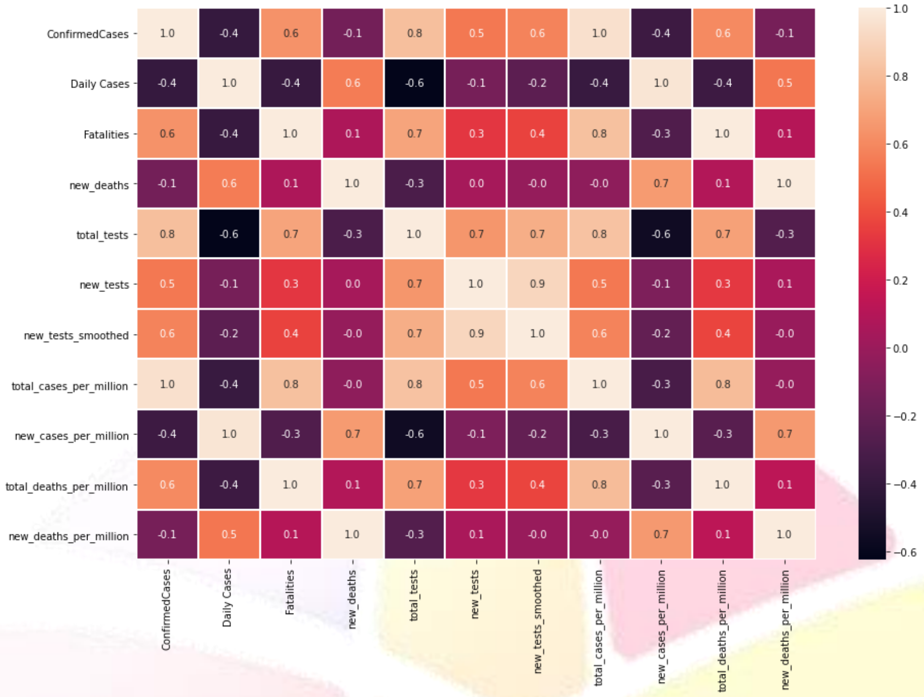

Figure 3: Correlation Matrix of Parameters in the Dataset

In addition to graphs showing the daily number of cases and daily deaths of countries, correlation matrices examining the relationship of parameters in the dataset were also examined (Fig3). The correlation matrix shows the relationship between the parameters of the dataset. In the Matrix, values that change in a range of zero and one indicate the strength of the relationship, while positive or negative indicate the increase and decrease in the same direction or the increase and decrease in the opposite direction.

The correlation of the other four countries with Turkey was also realized. The Dickey Fuller test was used in the total number of cases data of each country to determine the stasis. As a result, it was found that Turkey, Iran and Italy did not show stationary, but France and Germany were stationary

4. Experiment Results

About eighty percent of the one hundred and twenty values in the data set that will be given to the ARIMA time series were used to train the model (97 data) and the remaining part (23 data) was used to predict the model.

As a result of the Dickey Fuller test, Turkey's case number parameter was not stationary when examined, its P-value was 0.055 and it had a unit root. After the change was made by taking the difference, the Dickey Fuller test showed that it became stationary, the P-value was 0.047, and the unit root was not found. Data on the number of cases belonging to Turkey were divided into ninety-seven and twenty-three. These data were charted separately. The model was created with the *auto_arima* (Alkaline, 2020) function of the Pmdarima Library (Python, 2020). It was determined that there was no seasonality in the parameter of this model and that it was a five-month season. Then a line chart was drawn in which the predicted data and the data given to the model complement each other. Although there was an error rate of 0.07 percent for the first day predicted, it rose to 1.84 percent on the last day (Fig4).

Similarly for Germany, the Dickey Fuller test was applied on case numbers. As a result, it was found to be stationary, without a unit root, and the P-value was 0.008. The data set was given to the model after separation. Estimating and case number graphs were drawn. The error rate for the first day was 0.016 percent, while it rose to 1.86 percent on the last day (Fig5).

In the Dickey Fuller test applied to Iran, it was observed that the data was not stationary, that it had a unit root and that the p-value was 0.993. When the difference was taken and retested, it was realized that it was still not stationary, that it had a unit root and that the p-value was 0.253.After that, the second difference was taken and the test was applied. As a result of the second difference, it was possible to conclude that the series was stagnant. After the data set was divided, the model was created and graphs were drawn. Although the margin of error was 0.15 percent on the first day, it was 5.63 percent on the last day (Fig 6).

As a result of the Dickey Fuller test applied for France, it was observed that the series was stationary, there was no unit root, and the p-value was 0.001. The data was divided, the corresponding graphs were drawn and given to the model. On the first day, the margin of error was 0.02 percent, while on the last day it rose to 1.96 percent (Fig 7).

Finally, in the Dickey Fuller test conducted for Italy, the series was not stationary, the unit root and the P-value was 0.11. After the first difference process, it was observed that the series became stagnant, there was no unit root, and the p-value was 0.0001. The data set was divided, given to the model. As a result of the forecast, the margin of error on the first day was 0.03 percent, and on the last day it was 0.8 percent (Fig 8).

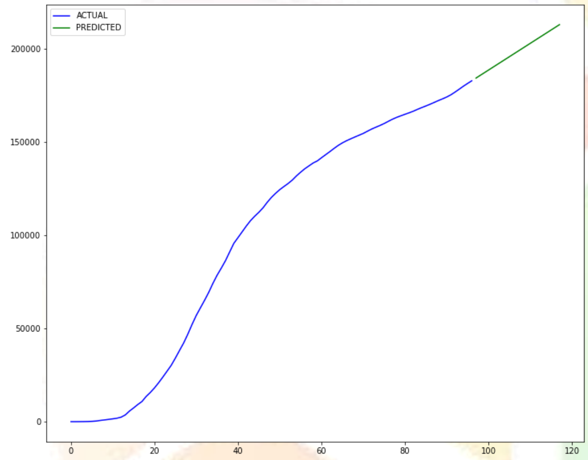

Figure 4: Predicted Number of Cases in Turkey

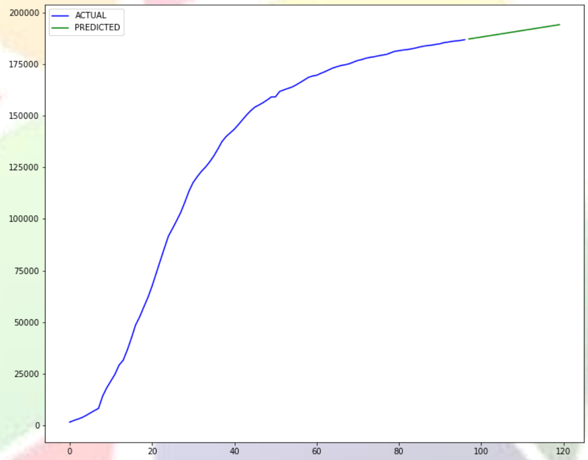

Figure 5: Predicted Number of Cases in Germany

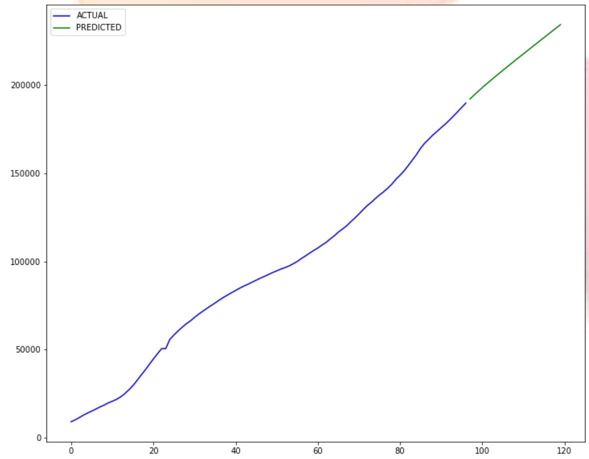

Figure 6: Predicted Number of Cases in Iran

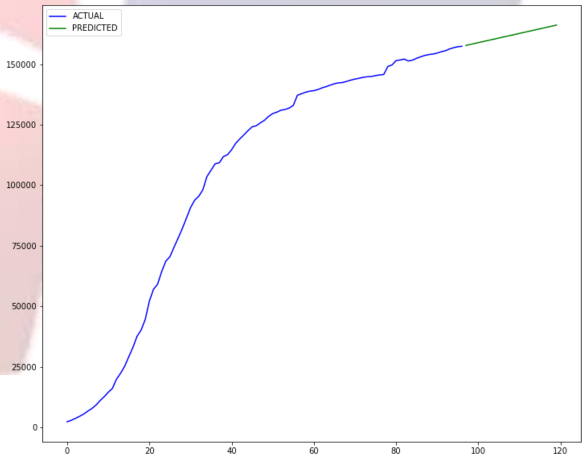

Figure 7: Predicted Number of Cases in France

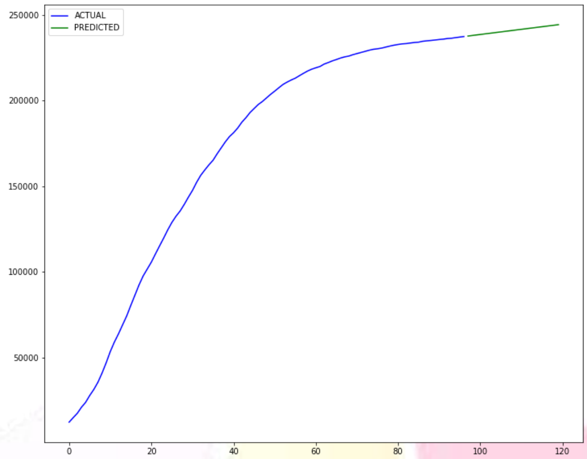

Figure 8: Predicted Number of Cases in Italy

The figures above represent five countries ' estimates of the total number of cases. The blue line refers to the ninety-seven-day actual number of cases, while the green line indicates the result of twenty-three-day forecast using the ARIMA time series model.

5. Conclusion

As a conclusion, the results of ARIMA model analysis showed that the model was correct for short-term forecasts; in monthly or annual forecasts it had significant errors. In the study of COVID-19 cases based on countries, the margin of error has risen to the point that it is not possible to make a satisfactory calculation in the long term. Keeping the model up-to-date, that is, feeding it with up-to-date data, has been found to provide satisfactory results in the short term. It can be concluded that in the COVID-19 pandemic, the ARIMA model was able to give short-term forecast numbers with very few errors. For short to medium-term forecasting, ARIMA can therefore be considered an acceptable model, but the results should be interpreted in a thrifty way. Finally, the continuous updating of these results, the addition of interventions and other real aspects and the extension of the model to other countries and/or regions will provide more useful and more accurate forecasts.

6. References

Adhikari, R., & Agrawal, R. (2013). An introductory study on time series modeling and forecasting. Lambert Academic Publishing.

Alkaline. (2020). alkaline: Retrieved from:
https://alkalineml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html

Asteriou, D., & Hall, S. G. (2011). ARIMA Models and the Box–Jenkins Methodology. . Applied Econometrics (Second ed.) (s. 265-286). içinde Palgrave MacMillan.

Benvenuto, D. G. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Brief.

Box GE, J. G. (1976). Time series analysis: forecasting and control. John Wiley & Sons.

Cao S, W. F. (2013). A hybrid seasonal prediction model for tuberculosis incidence in China. BMC medical informatics and decision making, 56.

Cheung Y-W, L. K. (1995). Lag order and critical values of the augmented Dickey–Fuller test. Journal of Business & Economic Statistics, 80.

Coronavirus Update. (2020). COVID-19 virus pandemic worldmeter: Retrieved from: https://www.worldometers.info/coronavirus/.

Fanelli, D. P. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons and Fractals, 1-12.

Fang, X. L. (2020). Forecasting incidence of infectious diarrhea using randomforest in Jiangsu Province, China. BMC Infect. Dis, 1-8.

Fang, Y. N. (2020). Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions. Journal of medical virology, 645-659.

Fattah J, E. L. (tarih yok). Forecasting of demand using ARIMA model. International Journal of Engineering Business Management.

Li, Q. F. (2020). Trend and forecasting of the COVID-19 outbreak in China. International Journal of Information Security, 469-496.

Petropoulos, F., & Makridakis, S. (2020). Forecasting the novel coronavirus COVID-19. PLoS ONE, 15.

Python. (2020). Pypi. Retrieved from: https://pypi.org/project/pmdarima/.

Song, X., Xiao, J., Deng, J., Kang, Q., Zhang, Y., & Xu, J. (2016). Time series analysis of influenza incidence in Chinese. Medicine Journal.

StatsModels.(2020).Retrieved from https://www.statsmodels.org/stable/generated/statsmodels.tsa.seasonal.seasonal_decompose.html.

Team, I. C.-1. (2020). Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. medRxiv.

U, H. (1986). Box-Jenkins modelling of some viral infectious diseases. Stat Med.

University, J. H. (2020, March). [https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html. John Hopkins University: Retrieved from: https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html

WHO. (2020). COVID19 WHO. World Health Organization (WHO): Retrieved from: https://covid19.who.int

Williamson G, W. H. (1999). A monitoring system for detecting aberrations in public health surveillance reports. Stat Med.

Yin, R., Luusua, E., Dabrowski, J., Zhang, Y., & Kwoh, C. T. (2020). Time-series mutation prediction of influenza A viruses via attention-based recurrent neural networks. Bioinformatics.

Zhang, L. L. (2017). Time prediction models for echinococcosis based on gray system theory and epidemic dynamics. Int. J. Environ. Res. Public Health.