Time Series Model for Forecasting the Number of Covid-19 Cases in the United States of America

serhat akay[1] and huriye akay[1]

[1]University of Health Sciences Izmir Bozyaka Education and Research Hospital

September 11, 2020

Abstract

Background: Coronavirus disease-19 (Covid-19) had an unprecendented effect on both nations and health systems. Time series modeling using Auto-Regressive Integrated Moving Averages (ARIMA) models have been used to forecast variables extensively in statistics and econometrics. Objectives: The aim is to predict the total number of Covid-19 cases in the United States of America using ARIMA models of time-series analysis. Methods: We used time series analysis to build an ARIMA model of the total number of cases from January 21, 2020 to August 7, 2020 and used the model to predict cases in the following 7 days, from August 8, 2020 to August 14, 2020. Hyndman and Khandakar algorithm was used to select components of ARIMA models. Percentage error was used to evaluate forecasting accuracy. Results: During the model building period, 4,883,646 cases were diagnosed and during 14 days of validation period additional 313,502 new cases were added. ARIMA model with (p,d,q) components of (5,2,1) was used for forecasting. The mean percentage error of forecast was 0.09% and forecast accuracy was high in the following week. Conclusion: ARIMA models can ve used to forecast the total number of cases of Covid-19 patients in the upcoming first week.

Time Series Model For Forecasting the Number of Covid-19 Cases in the United States of America

Background:

Coronavirus disease-19 (Covid-19) had an unprecendented effect on both nations and health systems. Time series modeling using Auto-Regressive Integrated Moving Averages (ARIMA) models have been used to forecast variables extensively in statistics and econometrics.

Objectives:

The aim is to predict the total number of Covid-19 cases in the United States of America using ARIMA models of time-series analysis.

Methods:

We used time series analysis to build an ARIMA model of the total number of cases from January 21, 2020 to August 7, 2020 and used the model to predict cases in the following 7 days, from August 8, 2020 to August 14, 2020. Hyndman and Khandakar algorithm was used to select components of ARIMA models. Percentage error was used to evaluate forecasting accuracy.

Results:

During the model building period, 4,883,646 cases were diagnosed and during 14 days of validation period additional 313,502 new cases were added. ARIMA model with (p,d,q) components of (5,2,1) was used for forecasting. The mean percentage error of forecast was 0.09% and forecast accuracy was high in the following week.

Conclusion:

ARIMA models can ve used to forecast the total number of cases of Covid-19 patients in the upcoming first week.

Highlights:

- Covid-19 has impacted the health care systems over the world regardless of the advancement and complexity systems.
- Prediction of Covid-19 cases can play a crucial role for the management of health care systems.
- Time series had been used in forecasting in different aspects of science.
- Prediction of number of Covid-19 cases can be made using time-series analysis.

Introduction:

Coronavirus disease 19 (Covid-19) is an infectious disease caused by the Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) which was initially identified in December 2019 before becoming a global pandemic. Human to human spread is the identified form of transmission while exact molecular pathways in this pathway are not fully understood [1, 2].

Respiratory diseases spread by the inhalation of droplets scattered by the infected person. Avoidance of social distancing, failure of using personal protective equipment, late detection of symptoms all contribute to the rapid spread of the disease, and increase the burden in the health system in Covid-19 pandemic [3]. The unexpected increase of infected patients had put tremendous pressure on health systems causing a capacity overload, the premature ending of medical supplies, and exhausted health professionals to name a few. This sudden increase of patients had caused also significant implications for non-Covid-19 patients such as failure of initiating proper workup and treatment of conditions [4].

Such an unexpected increase of patients is not foreseeable for the health system and there are several factors for this. First, modern medicine had faced this scale of the pandemic first time where a systematic approach is used to gather data about how the number of cases evolves across nations. Second, countermeasure for this pandemic, lockdowns, quarantines and curfews, are internationally applied for the first time with no previous experience. Third, as the number of cases increase with different increments between communities, prediction of the total number of cases can be challenging.

Time series analysis with autoregressive integrated moving averages (ARIMA) models was popularized by Box and Jenkins in 1970 with their Box-Jenkins approach [5]. By using only one variable measured in equally spaced points in time, forecasting can be made with the help of the model build using the variable. Time series are used in statistics, weather prediction, and econometrics to name a few. In medicine, time series are used to predict the number of patients admitted in previous studies [6, 7].

In this research, we explored whether ARIMA model is feasible to predict the number of cases for Covid-19 patients. The aim is to forecast the total number of patients in the United States of America (USA) using the time series model and this modeling can provide health systems to provide better health care to patients.

Materials-Methods:

This time-series analysis of Covid-19 data consisted of data starting from identification of the first case from January 21, 2020 to August 14, 2020 in the USA. Data for this cohort study was obtained from Situation Dashboard of the European Center for Disease Prevention and Control website Coronavirus 19 Situation Report from the official report page of the WHO webpage on August 14, 2020 [8]. Data analyzed included daily confirmed cases in the USA between the aforementioned dates.

Modeling consisted two of important steps: (1) building a time series model from January 21, 2019 to August 7, 2020 and (2) validation of the fit model, to forecast the number of confirmed cases from August 8, 2020 to August 14, 2020.

Before building the time-series model, stationarity was evaluated with augmented Dickey-Fuller (ADF) unit root test and the visual diagnosis was used to access trends. If stationarity was not met log transformation and differencing was used to de-trend the series.

Mathematically simple ARIMA model is written as $W[?]=\mu+(\vartheta(B)/\psi(B))\alpha[?]$; where $W[?]$ is the response series $Y[?]$ or difference of the response series, μ is the mean term, $\vartheta(B)$ is the MA operator, $\psi(B)$ is AR operator, B is the backshift operator, that is $BX[?]=X[?]-1$ and $\alpha[?]$ is the independence disturbances also known as the random error [10]. Parameters for the ARIMA method are estimated using the maximum likelihood method.

Auto-correlation and partial auto-correlation functions were used to determine the components of the ARIMA model (p,d,q). Box-Jenkins approach traditionally used to build models for ARIMA models where an iterative process was applied with three steps: Identification, estimation of parameters, and diagnostic checking. Models with the least BIC and AIC tests were used for forecasting. But for this study, the best model was selected based on "auto.arima()" function included in the "forecast" library of the statistical program which uses the Hyndman and Khandakar algorithm [11]. "auto.arima()" function is a step-wise approach to determine the model with the best fit by using models with appropriate and optimized parameters, models with least AIC, and producing point forecasts using the best model. Aim of function is to choose the parsimonious model.

Forecasting accuracy was evaluated by the percentage error (PE) defined as; the difference between forecast and confirmed cases divided by confirmed cases and mean average percentage error (MAPE), mean of PE [5]. $p<0.05$ was considered significant and statistical analysis was conducted using R 4.0.0 (R Core Team, Vienna, Austria).

Results:

Between 207 days of January 21 2019 and August 14 2020, 5,248,242 confirmed cases were identified with 167,110 (3.2%) deaths. For model building variables from the first 200 days were used with total cases of 4,883,646 with 164,104 (3.3%) mortal cases where 364,596 new confirmed cases with 7,006 (1.9%) new deaths were analyzed for the validation part.

ADF unit root test showed there is a unit root and visual diagnosis of cases upward trend in cases so log transformation was used. Because ADF showed unit root for the log-transformed case numbers, differencing was applied to make the series stationary. After differencing with 2 lags, time series became stationary and ADF showed there were no unit roots ($p<0.05$). Since we assumed the time series was stationary, we proceeded to model fitting.

"auto.arima ()" function was used to find the best fitting model with an auto-regressive (AR) component of one order (p=3), moving averages (MA) component of one order (q=1) and differencing of 2 (d=2). The proposed model was ARIMA (5,2,1). The coefficients for AR(3) were -0.3173, -0.0205, -0.1031, -0.1991 and -0.2060 while MA(1) was -0.6376 with the model's AIC of -643.54. The order of differencing was 2 as previously found.

We used the newly formed ARIMA model to forecast the number of cases from August 8 to August 14, 2020 using the "forecast" function (Table 1). By comparing the actual number of cases with predicted ones, the prediction accuracy of forecasting calculated by mean percentage error was 0.09%.

Discussion:

In this study, we used time series modeling to predict the number of cases worldwide in the following 14 days. Although percentage error was minimal in the beginning of prediction, it increased considerably as the prediction interval increased. But as noted at the results section, prediction ability weakens as the predicted days increase. The prediction had percentage error below 1% which indicated the model had a good fit. We conclude that as the predicted time period increases, the 95% confidence interval of the prediction increased.

Covid-19 pandemic had a devastating effect on both nations and their health systems. Although pandemics had been a part of human history and history had faced many pandemics before Covid-19, preparedness for Covid-19 was not ideal. Unpredicted increase of cases had been the main cause of public health overload. It is crucial for scientists to estimate the severity of the total number of cases, deaths, and reproduction numbers to predict the epidemic and its' duration.

Mathematical modeling of infectious diseases had been widely used since described by Kermack in 1927 [12]. In SIR model, a deterministic approach to epidemiologic modeling, the population is divided into compartments, which an individual is assigned to Susceptible (S), Infectious (I) or Recovered (R) compartment and models are made how a disease spreads, the total number of infected or the duration of an epidemic. SEIR modeling with the inclusion of Exposed (E) compartment has been studied in the Covid-19 pandemic where authors predicted epidemic progression in Mainland China to be around 40 thousand to 351 thousand depending on the implementation of control measures [13].

Deterministic models, like SIR models, use precisely determined, known relationships among events and don't have any random variation. Any given input produces the same result resulting in smooth, analytic curves with no noise. On the other hand, stochastic models have randomness where same inputs produce an ensemble of different outputs, eventually a variation of a distribution. Time series modeling is more close to stochastic modeling that randomness and ensemble of different outputs play a role.

Obtaining accurate data in pandemics is challenging. Roda et. al had noted, availability of limited reliable data during the pandemic is the basis of difficulty for accurate prediction [14]. Confirmed cases, by either imaging or DNA studies, are the tip of the iceberg where patients who don't have symptoms, present to the hospital, or misdiagnosed are the part of the iceberg hidden under the water. But for scientifically sound analysis, we included confirmed cases.

Our analysis included the total number of cases in the introduction and acceleration part of the "pandemic phase" of the continuum of pandemic phases, described for the Influenza Pandemic by WHO [15]. We don't know when the peak transmission phase will start or whether it started. Magnitude and time of shifting to transition phase and to inter-pandemic phase are also unknown and probably relies on several factors, different counter-measures for spreading taken by governments and individuals, change in climate conditions, advances in diagnosis and treatment, etc. As the peak transmission of the pandemic period has started, time series modeling with different variables or different models may be suitable for prediction.

Limitations:

First, this study predicted the total number of cases in the USA, where each country has its' own number of cases. Prediction for individual countries needs different ARIMA models for each country. Second, Covid-19 is dynamic, using data from different time periods for model building and validation may lead to models with different AR, MA, and differencing components with different validations. Third, we don't know if we used the variables from the whole or just the tip of the iceberg of cases, but we know that we used all confirmed cases which is more scientifically sound.

Conclusion:

Time series modeling can be used to predict the number of cases of Covid-19 patients worldwide where predictions in one-week interval are accurate.

Ethical approval:

The Study was approved by the institutional ethics committee and Ministry of Health, Republic of Turkey.

Sources of funding:

No funding was received for this study.

References:

Bourouiba L. Turbulent Gas Clouds and Respiratory Pathogen Emissions: Potential Implications for Reducing Transmission of COVID-19. JAMA. 2020 , March 20, online ahead of print. doi: 10.1001/jama.2020.4756

Kannan S, Shaik Syed Ali P, Sheeza A, Hemalatha K . COVID-19 (Novel Coronavirus 2019) - Recent Trends Eur Rev Med Pharmacol Sci. 2020 Feb;24(4):2006-2011. doi: 10.26355/eurrev_202002_20378.

Adhikari SP, Meng S, Wu YJ, Mao YP, Ye RX, Wang QZ, Sun C, Sylvia S, Rozelle S, Raat H, Zhou H. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. Infect Dis Poverty. 2020;9:29. doi: 10.1186/s40249-020-00646-x

Driggin E, Madhavan MV, Bikdeli B, Chuich T, Laracy J, Biondi-Zoccai G, Brown TS, Der Nigoghossian C, Zidar DA, Haythe J, Brodie D, Beckman JA, Kirtane AJ, Stone GW, Krumholz HM, Parikh SA. Cardiovascular Considerations for Patients, Health Care Workers, and Health Systems During the COVID-19 Pandemic. J Am Coll Cardiol. 2020;75:2352-2371. doi: 10.1016/j.jacc.2020.03.031

Box GEP JG, Reinsel GC. Time series analysis: Forecasting and control. Delhi: Pearson Education, 1994. Zhou L, Zhao P, Wu D, Cheng C, Huang H. Time series model for forecasting the number of new admission inpatients. BMC Med Inform Decis Mak. 2018;18:39 doi: 10.1186/s12911-018-0616-8

Juang WC, Huang SJ, Huang FD, Cheng PW, Wann SR. Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. BMJ Open. 2017;7:e018628 doi: 10.1136/bmjopen-2017-018628

Covid-19. Web address: https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html Accessed: August 18,2020

Agha R, Abdall-Razak A, Crossley E, Dowlut N, Iosifidis C and Mathew G, for the STROCSS Group. The STROCSS 2019 Guideline: Strengthening the Reporting of Cohort Studies in Surgery. International Journal of Surgery 2019;72:156-165.

General Notation for ARIMA models, Web address: https://v8doc.sas.com/sashtml/ets/chap7/sect8.htm accessed: June 16 ,2020.

Hyndman, R, Khandakar Y. Automatic Time Series Forecasting: The Forecast Package for R. J Stat Softw 2008;27:1–22 doi: 10.18637/jss.v027.i03

Kermack, WO, McKendrick AG. A Contribution to the Mathematical Theory of Epidemics. P Roy Soc A-Math Phy. 1927;115:700–721.

Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M, Liu P. Modified SEIR and AI prediction of the epidemics trend of Covid-19 in China Under Health Interventions. J Thorac Dis 2020;12:165-174. doi: 10.21037/jtd.2020.02.64

Roda WC, Varughese MB, Han D, Li MY. Why is it difficult to accurately predict the covid-19 epidemic? Infect Dis Model 2020;5:271-281. doi: 10.1016/j.idm.2020.03.001 WHO Pandemic Influenza Risk Management. Web address: https://www.who.int/influenza/preparedness/pandemic/influenza_risk_management_update2017/en/ Accessed June 16,2020.

Hosted file

table 02 09 2020.docx available at https://authorea.com/users/356076/articles/479083-time-series-model-for-forecasting-the-number-of-covid-19-cases-in-the-united-states-of-america