Turk J Public Health 2021;19(2):140-145
Available online at: http://dergipark.gov.tr/tjph/
doi:10.20518/tjph.809201

TURKISH JOURNAL OF PUBLIC HEALTH

ORIGINAL ARTICLE / ORİJİNAL MAKALE

Time series model for forecasting the number of COVID-19 cases in Turkey

Türkiye'de görülen COVID-19 olgu sayılarının tahmininde zaman serisi modelinin kullanılması

Serhat Akay[a], Huriye Akay[b]

[a] Uzm.,Dr.,Emergency Medicine Clinic, Izmir Bozyaka Education and Research Hospital, University of Health Sciences, Izmir, Turkey.
[b] Uzm.,Dr., Emergency Medicine Clinic, Izmir Bozyaka Education and Research Hospital, University of Health Sciences, Izmir, Turkey.

Received: 12.10.2020, Accepted: 12.04.2021

ABSTRACT

Objective: Coronavirus disease 2019 (COVID-19) had an unprecedented 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. We aimed to predict the total number of cases for COVID-19using ARIMA models of time-series analysis in Turkey. **Methods:** We used time series analysis to build an ARIMA model of the total number of cases from March 11, 2020 to August 24, 2020 and used the model to predict cases in the following 14 days, from August 25, 2020 to September 7, 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, 259692 cases were diagnosed and during 14 days of validation period additional 21817 new cases were added. ARIMA model with (p,d,q) components of (4, 2, 0) was used for forecasting. The mean percentage error of forecast was 0.20% and forecast accuracy was highest in the two weeks of forecasting. **Conclusion:** ARIMA models can be used to forecast the total number of cases of COVID-19 patients for the upcoming two weeks in Turkey.

Keywords: COVID-19, time series, forecasting

Correspondence: Serhat AKAY, Emergency Medicine Clinic, Izmir Bozyaka Education and Research Hospital, University of Health Sciences, Izmir, Turkey. **E-mail:** howls_Castle@mynet.com. **Tel:** +905057079350.

Cite This Article: Akay S, Akay H. Time Series Model For Forecasting the Number of COVID-19 Cases in Turkey. Turk J Public Health 2021;19(2):140-145.

ÖZ

Amaç: Koronavirüs hastalığı 2019'un (Covid-19) hem ülkeler hem de sağlık sistemleri üzerinde beklenmedik bir etkisi olmuştur. Otoregresif Entegre Hareketli Ortalama (Auto-Regressive Integrated Moving Averages) (ARIMA) modellerini kullanarak yapılan zaman serisi modellemesi, istatistik ve ekonometride değişkenleri kapsamlı şekilde tahminde kullanılmaktadır. Zaman serisi analizinin ARIMA modellerini kullanarak, Türkiyede Covid-19 için toplam olgu sayısını tahmin etmeyi amaçladık. **Yöntem:** 11 Mart 2020'den 24 Ağustos 2020'ye kadar olan toplam olgu sayısının bir ARIMA modelini oluşturmak için zaman serisi analizini kullandık ve 25 Ağustos 2020'den 7 Eylül 2020'ye kadar takip eden 14 gündeki vakaları tahmin etmek için bu modelden yararlandık. ARIMA modellerinin bileşenlerinin seçiminde Hyndman ve Khandakar algoritması kullanıldık. Öngörme doğruluğunu değerlendirmek için yüzde hata kullanıldı. **Bulgular:** Model oluşturma döneminde 259.692 olgu teşhis edildi ve 14 günlük doğrulama süresi boyunca ek 21.817 olgu vaka eklendi. Öngörü için (4, 2, 0) bileşenli (p, d, q) bileşenli ARIMA modeli kullanıldı. Ortalama tahmin hatası % 0.20 olarak bulundu ve tahmin doğruluğu tahminin iki haftalık döneminde en yüksekti. **Sonuç:** ARIMA modelleri, Türkiye'de önümüzdeki iki hafta boyunca Covid-19 hastalarının toplam olgu sayısını tahmin etmek için kullanılabilir.

Anahtar Kelimeler: COVID-19, zaman serisi, tahmin

Introduction

Coronavirus disease 2019 (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 contributed to the rapid spread of the disease, and increased the burden in the health care 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 was 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 was used to gather data about how the number of cases differed across nations. Second, a countermeasure for this pandemic, lockdowns, quarantines and curfews, were internationally applied for the first time with no previous experience. Third, as the number of cases increased with different increments between communities, prediction of the total number of cases was 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 becoming widely 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 Turkey using time series model and this modeling can provide health systems to provide better health care to patients.

Material and Methods

This time-series analysis of Covid-19 data consisted of data starting from the identification of the first case from March 11, 2020 to September 07, 2020 in Turkey. Data for this methodological study was obtained from the General Covid Report page updated daily at the Republic of Turkey, Ministry of Health.[8] Data analyzed included the total number of daily confirmed cases in Turkey between the aforementioned dates. This study was approved by the institutional ethics committee and Ministry of Health, Republic of Turkey and compiles with STROBE Guidelines. Work has been reported in the line with the STROBE guideline.[9] Modeling consisted two of important steps: (1) building a time series model from March 11, 2020 to August 24, 2020 and (2) validation of the fit model, to forecast the number of confirmed cases in the following 14 days, from August 25, 2020 to September 7, 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 $W_t=\mu+(\theta(B)/\psi(B))\alpha_t$; where W_t is the response series Yt or difference of the response series, μ is the mean term, $\theta(B)$ is the MA operator, $\psi(B)$ is AR operator, B is the backshift operator, that is $BX_t=X_t-1$ and αt 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. The 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) with "forecast" library written by Rob Hyndman et al.[11]

Result

Between 181 days of March 11, 2019 and September 7, 2020, 281509 confirmed cases were identified with 6782 (2.4%) deaths. For model building variables from the first 167 days were used with total cases of 259692 with 6139 (2.4%) mortal cases where 21817 new confirmed cases with 643 (2.9%) 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 differencing was used. 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 four orders (p=4), moving averages (MA) component of one order (q=0) and differencing of 2 (d=2). The proposed model was ARIMA (4,2,0). The coefficients for auto-regressions were 0.0795, -0.1133, -0.1350, and 0.316 with the model's AIC of 2294.4. The order of differencing was 2 as previously found. We used the newly formed ARIMA model to forecast the number of cases from August 25 to September 8, 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 0.20% (Table 1).

Table 1. Forecasts of number of Covid-19 cases from August 25 to September 7, 2020 with associated percentage error.

	Number of Covid-19 patients	**Predicted number of cases (95% Confidence Interval)**	**Percentage Error (%)**
August 25	261194	261135 (260610 – 261660)	0.02
August 26	262507	262578 (261405 - 263751)	0.03
August 27	263998	264021 (262058 - 265984)	0.01
August 28	265515	265464 (262591 - 268337)	0.02
August 29	267064	266907 (263016 - 270797)	0.06
August 30	268546	268350 (263346 - 273354)	0.07
August 31	270133	269793 (263586 - 275600)	0.13
September 1	271705	271236 (263743 - 278728)	0.17
September 2	273301	272679 (263823 - 281535)	0.22
September 3	274943	274122 (263828 - 284415)	0.30
September 4	276555	275565 (263764 - 287365)	0.35
September 5	278228	277008 (263633 - 290382)	0.44
September 6	279806	278451 (263438 - 293464)	0.48
September 7	279806	279894 (263181 - 296607)	0.58

Discussion

In this study, we used time series modeling to predict the number of cases in Turkey for the following 14 days. Although percentage error was minimal at the beginning of the prediction, it increased considerably as the prediction interval increased. But as noted in the results section, prediction ability weakens as the predicted days increase. The prediction had a 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 slightly.

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. An unpredicted increase of cases had been the main cause of public health overload. Scientist must 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. The 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 the 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 inter-pandemic phase are also unknown and probably rely 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.

There are limitations involved in this study. First, this study predicted the total number of cases in Turkey, 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 in countries where predictions in two-week interval are accurate.

Acknowledgment

Ethical Declaration: This study was approved by the institutional ethics committee and Ministry of Health, Republic of Turkey and compiles with STROBE Guidelines. Work has been reported in the line with the STROBE guideline 9.

Financial Support: No funding was received for this study.

Conflict of Interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. 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.
2. 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.
3. Adhikari SP, Meng S, Wu YJ, Mao YP, Ye RX, Wet al. 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.
4. Driggin E, Madhavan MV, Bikdeli B, Chuich T, Laracy J, et al. Cardiovascular Considerations for Patients, Health Care Workers, and Health Systems During the COVID-19 Pandemic. J Am Coll Cardiol. 2020;75:2352-2371.
5. Box GEP JG, Reinsel GC. Time series analysis: Forecasting and control. Delhi: Pearson Education, 1994.
6. 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
7. 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
8. Covid-19. Web address: https://covid19.saglik.gov.tr/TR-66935/genel-koronavirus-tablosu.html# Accessed: September 10, 2020
9. Agha R, Abdall-Razak A, Crossley E, Dowlut N, Iosifidis C et al. for the STROCSS Group. The STROCSS 2019 Guideline: Strengthening the Reporting of Cohort Studies in Surgery. International Journal of Surgery 2019;72:156-165.
10. General Notation for ARIMA models, Web address: https://v8doc.sas.com/sashtml/ets/chap7/sect8.htm accessed: June 16 ,2020.
11. Hyndman, R, Khandakar Y. Automatic Time Series Forecasting: The Forecast Package for R. J Stat Softw 2008;27:1–22.
12. Kermack, WO, McKendrick AG. A Contribution to the Mathematical Theory of Epidemics. P Roy Soc A-Math Phy. 1927;115:700–721.
13. Yang Z, Zeng Z, Wang K, Wong SS, Liang W, et al. Modified SEIR and AI prediction of the epidemics trend of Covid-19 in China Under Health Interventions. J Thorac Dis 2020;12:165-174. d
14. 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.
15. WHO Pandemic Influenza Risk Management. Web address: https://www.who.int/influenza/preparedness/pandemic/influenza_risk_management_update2017/en/ Accessed June 16,2020.