Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2021;13(1):91-102

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2020-80186

A Study on the Estimation of COVID-19 Daily Cases and Reproduction Number Using Adaptive Kalman Filter for USA, Brazil, Germany, India, Russia, Italy, Spain, United Kingdom, France, Turkey

ABD, Brezilya, Almanya, Hindistan, Rusya, İtalya, İspanya, Birleşik Krallık, Fransa, Türkiye İçin Uyarlanabilir Kalman Filtresi Kullanarak COVID-19 Günlük Vakaları ve Üreme Sayısı Tahmini Üzerine Bir Çalışma

Levent ÖZBEK[a], Hakan DEMİRTAŞ[b]

[a]Department of Statistics, Ankara University Faculty of Science, Ankara, TURKEY
[b]Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, USA

ABSTRACT Objective: In the literature, non-linear mathematical growth models are often used to estimate the number of coronavirus disease-2019 (COVID-19) cases. Specific algorithms such as mathematical optimization technique need to be employed for parameter estimation. In this work, a novel method to estimate COVID-19 daily cases and reproduction number is proposed for COVID-19. **Material and Methods:** In this study, the daily number of COVID-19 cases between January 01 and November 16, 2020 has been estimated online via AR(1) (autoregressive time-series model of order 1) and the adaptive Kalman filter (AKF). After calculating the estimate for daily cases, the reproduction number estimate was obtained. **Results:** It is quite a simple method to model the daily case number by time series with the time-varying parameter AR(1) stochastic process and estimated the time-varying parameter with online AKF. The method is online. Only the data points on the last day are sufficient. **Conclusion:** The COVID-19 data have been modeled in state space, and the AKF has been employed to estimate the number of daily cases. The estimation results were obtained for the number of daily cases using the AR(1) model. Since the estimation using the AR(1) stochastic process does not require any other modeling assumption, it is a simple approach to model the daily case number time series with the time-varying parameter AR(1) stochastic process and estimated the time-varying parameter with online AKF. We suggest that the simplest method for the reproduction number estimation will be obtained by modeling the daily case via an AR(1) model.

Keywords: COVID-19; state-space modelling; AR(1); adaptive Kalman filter; the reproduction number estimation

ÖZET Amaç: Literatürde, doğrusal olmayan matematiksel büyüme modelleri, koronavirüs hastalığı-2019 [coronavirus disease-2019 (COVID-19)] vakalarının sayısını tahmin etmek için sıklıkla kullanılmaktadır. Parametre tahmini için matematiksel optimizasyon tekniği gibi özel algoritmaların kullanılması gerekir. Bu çalışmada, COVID-19 için günlük COVID-19 vakalarını ve çoğalma sayısını tahmin etmek için yeni bir yöntem önerilmiştir. **Gereç ve Yöntemler:** Bu çalışmada, 01 Ocak ve 16 Kasım 2020 tarihleri arasında günlük COVID-19 vakalarına dayalı olarak AR(1) (1 gecikmeli oto regresif zaman serisi modeli) ve uyarlanabilir Kalman filtresi (UKF) aracılığıyla günlük vaka tahmini çevrim içi olarak yapılmıştır. Günlük vakalar için tahmin, çoğalma sayısı tahmini elde edilmiştir. **Bulgular:** Günlük vaka sayısı zaman serilerini zamanla değişen AR(1) stokastik süreç ile modellemek ve çevrimiçi UKF ile zamanla değişen parametreyi tahmin etmek oldukça basit bir yöntemdir. Yöntem çevrim içidir. Yalnızca son gündeki veri noktaları yeterlidir. **Sonuç:** COVID-19 verileri durum uzayında modellenmiştir ve günlük vaka sayısını tahmin etmek için UKF kullanılmıştır. AR(1) modeli kullanılarak günlük vaka sayısı için elde edilen tahmin sonuçları. AR(1) stokastik sürecini kullanan tahmin, başka herhangi bir modelleme varsayımı gerektirmediğinden, basit bir yaklaşımdır. Günlük vaka sayısı zaman serilerini zamanla değişen AR(1) stokastik süreci ile modellemek oldukça basit bir yöntemdir ve zamanla değişen parametreyi çevrim içi UKF ile tahmin edilmiştir. Çoğalma sayısı tahmini için en basit yöntemin, günlük vakayı bir AR(1) modeli aracılığıyla modelleyerek elde edileceğini öneriyoruz.

Anahtar kelimeler: COVID-19; durum uzayı modellemesi; AR(1); uyarlamalı Kalman filtresi; çoğalma sayısı tahmini

In December 2019, a new coronavirus disease emerged characterized as a viral infection with a high level of transmission in Wuhan, China. Coronavirus 19 (COVID-19) is caused by the virus known as severe

Correspondence: Levent ÖZBEK
Department of Statistics, Ankara University Faculty of Science, Ankara, TURKEY/TÜRKİYE
E-mail: ozbek@science.ankara.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 24 Nov 2020 ***Received in revised form:*** 11 Jan 2021 ***Accepted:*** 12 Jan 2021 ***Available online:*** 29 Apr 2021

acute respiratory syndrome coronavirus 2 (SARS-CoV-2) established by the ICTV.[1-3] Gompertz and logistic models have been used to estimate the number of COVID-19 cases in China by Jia et al.[4] Cas torina et al. have used these two modes in China, South Korea, Italy, and Singapore.[5] Roosa et al. have used Generalized Logistic Growth Model (GLM) for the data gathered between February 5 and February 24, 2020, for China.[6] Roosa et al. have used the GLM and the Richards model for the data gathered between February 13 and February 20, 2020 for China.[7] Munayco et al. have used the Generalized Growth Model (GGM) for the dates February 29 and March 30, 2020, for Peru.[8] Gompertz, Logistic, and Artificial Neural Network models were applied in.[9] Zuzana et al. used the Gompertz curve to model a trajectory of the number of infections for the USA.[10] Català et al. employed the Gompertz function in several countries to make short-time predictions.[11] Petropoulos et al. adopted simple time series forecasting approaches.[12] In logistic, Bertalanffy and Gompertz models, non-linear mathematical growth is studied, and prediction and analysis are given for the coronavirus disease.[4] The prediction methods of logistic, Gompertz, and Bertalanffy models are similar, but the mathematical models are different. Specific algorithms such as mathematical optimization technique need to be employed for parameter estimation. The authors use the regression coefficient (R^2) for model evaluation. The paper applies these models to the Wuhan and non-Hubei data in China and stated that "The prediction results of three different mathematical models are different for different parameters and in different regions". Moreover, the authors state that "We have collected some COVID-19 epidemic predictions of other researchers, as shown in Table 3. It can be seen from Table 3 that the total prediction results of different models are quite different". In, only Gompertz non-linear mathematical growth model is studied and applied to China, South Korea, and Italy data.[5] They considered the cumulative number of infected people and stated that this analysis needs to be updated on a daily basis. In, the GLM, exponential growth dynamics model and The Richards models are used and applied to the data from Hubei and other.[6] Mean squared error (MSE) is used as performance criterion. In, similar to logistic growth model, the Richards growth model, and a sub-epidemic wave model models are used and the data from Guangdong and Zhejiang provinces in China.[6,7] In, the GGM differential equation is used an applied to Lima-Peru data.[8] In, non-linear the logistic growth model, Gompertz ve Artificial Neural Networks models are used and non-linear least-squares method is used for parameter estimation.[9] In, only the Gompertz model is used and applied to the USA data.[10] In, only the Gompertz model is used and applied to data obtained from different provinces in China.[11] In, only exponential smoothing model is studied and applied to global confirmed cases.[12]

The papers cited in our manuscript all utilize "the cumulative number of infected people" as the data. Also, the models employed in those papers are non-linear mathematical growth models and there are more than one parameter to be estimated in those models. The models are non-linear mathematical ones and defined using differential equations. Specific algorithms such as mathematical optimization technique are to be employed for parameter estimation. The data used in the models employed need updating daily in order to analyze them. The methods used are offline and all data up to a specific date are necessary for parameter estimation in those models where the estimation needs to be updated on a daily basis with the inclusion of the new set of data. There are other growth models is addition to logistic, Bertalanffy, and Gompertz non-linear matematical models and they are given in Table 1.

TABLE 1: Non-linear models and their mathematical notations.

Model name	Statistical model
Brody	$y(t;\alpha,\beta,k)=\alpha(1-\beta\exp(-kt))+\varepsilon$
Bertalanffy	$y(t;\alpha,\beta,k,m)=(\alpha^{1-m}-\beta\exp(-kt))^{1/(1-m)}+\varepsilon$
Logistic	$y(t;\alpha,\beta,k)=\alpha/(1+\beta\exp(-kt))+\varepsilon$
Generalized logistic	$y(t;\beta,k,\gamma)=\alpha/((1+\beta\exp(-kmt)^{1/m})+\varepsilon$
Richards	$y(t;\alpha,k,m)=\alpha(1-\exp(-kt))^{1/m}+\varepsilon$
Negative exponential	$y(t;\alpha,k)=\alpha(1-\exp(-kt))+\varepsilon$
Stevens	$y(t;\alpha,\beta,p)=\alpha-\beta(k^t)+\varepsilon$
Tanaka	$y(t;\alpha,\beta,k,m)=(1/\sqrt{\beta})\ln\lvert 2\beta.(t-m)+2\sqrt{k^2(t-m)^2+\alpha\beta}\rvert+\varepsilon$
Gompertz	$Y(t)=\alpha\exp(-\beta\exp(-kt))+\varepsilon$

State-space models have been employed since the 1960's, mostly in the control and signal processing areas. The Kalman filtering (KF) has emerged as the most common tool. The KF has been extensively employed in many areas of estimation. The extensions and applications of state-space models can be found in almost all disciplines. This paper presents the use of adaptive KF (AKF) in the analysis of the COVID-19 daily cases.

The rest of this article is organized as follows: In material and methods, section the mathematical and computational methodologies are described, mathematical equations of the models used in this study are given, and analysis and estimation results are presented. In section estimating the reproduction number with AKF, the computation of the reproduction number with AKF is presented. Finally, the last section presents the conclusions.

MATERIAL AND METHODS

We can model the time series data on the number of daily cases in a simple fashion. Let us assume that the number of daily cases i_t is in the form of the AR(1) stochastic process (autoregressive time-series model of order 1) is given with Eq. 1.

$$i_t = \theta i_{t-1} + v_t \tag{1}$$

where θ is constants. v_t is $v_t \sim N(0,\sigma_1^2)$. The random variables $v_1, v_2, ..., v_n$ are assumed to be uncorrelated. Let us assume that the θ parameter of Eq. (1) is time varying and it is a stochastic process in the form of a random walk process. In this case, θ random walk process can be written as in Eq. (2)

$$\theta_t = \theta_{t-1} + w_t \tag{2}$$

w_t is $w_t \sim N(0,\sigma_2^2)$.The random variables $w_1, w_2, ..., w_n$ are assumed to be uncorrelated. Considering Eq. (1) and Eq. (2) together, the following state-space model can be written:

$$\theta_t = \theta_{t-1} + w_t \tag{2}$$

$$i_t = \theta_t i_{t-1} + v_t \tag{3}$$

Here, the state variable is unobservable, time-varying θ_t parameter and can be estimated using AKF (See details in Appendix). If this time-varying parameter is estimated using online AKF, estimation for the daily case counts in times $t+1$, $t+2$, ... can be made through this online-estimated parameter. The data used were taken from Johns Hopkins University.[13] We utilized Matlab 2013a in the statistical analysis.

RESULTS

Actual daily case and estimations that have been made online using AKF are given in Figure 1, Figure 2, Figure 3, Figure 4, Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10. Time varying parameter estimation is made online via AKF. According to the estimation results obtained by using daily number of cases in AR(1) model, MSE, mean absolute percentage error (MAPE), and R^2 were calculated. These calculated values are given in Table 2.

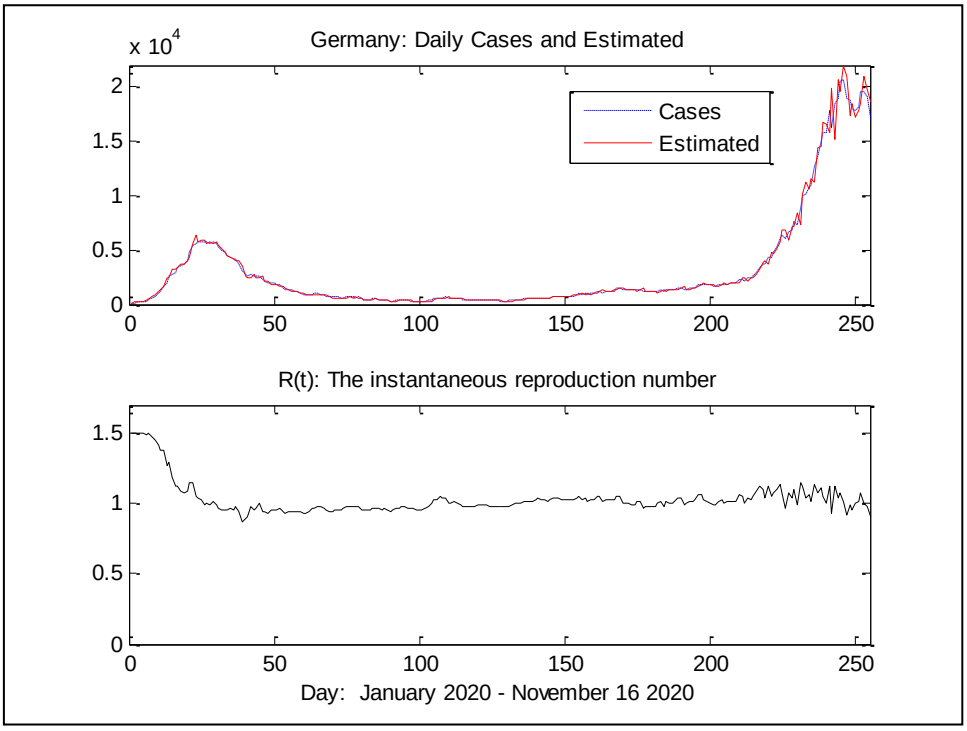

FIGURE 1: Germany: Daily cases and estimated-reproduction number estimated.

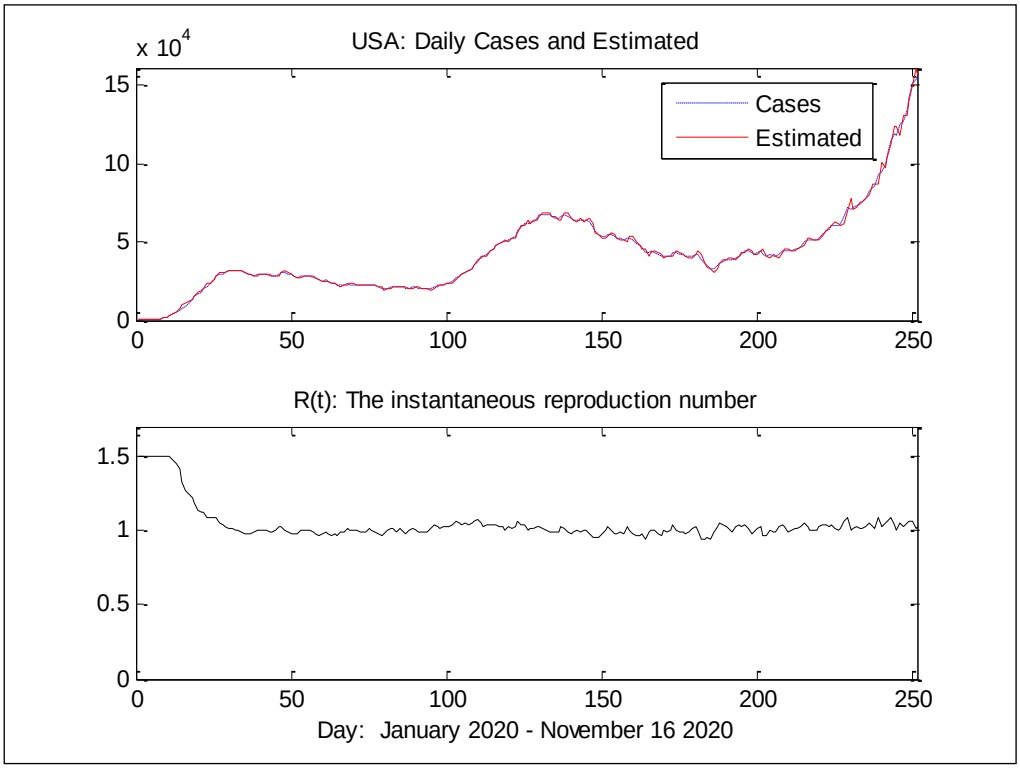

FIGURE 2: USA: Daily cases and estimated-reproduction number estimated.

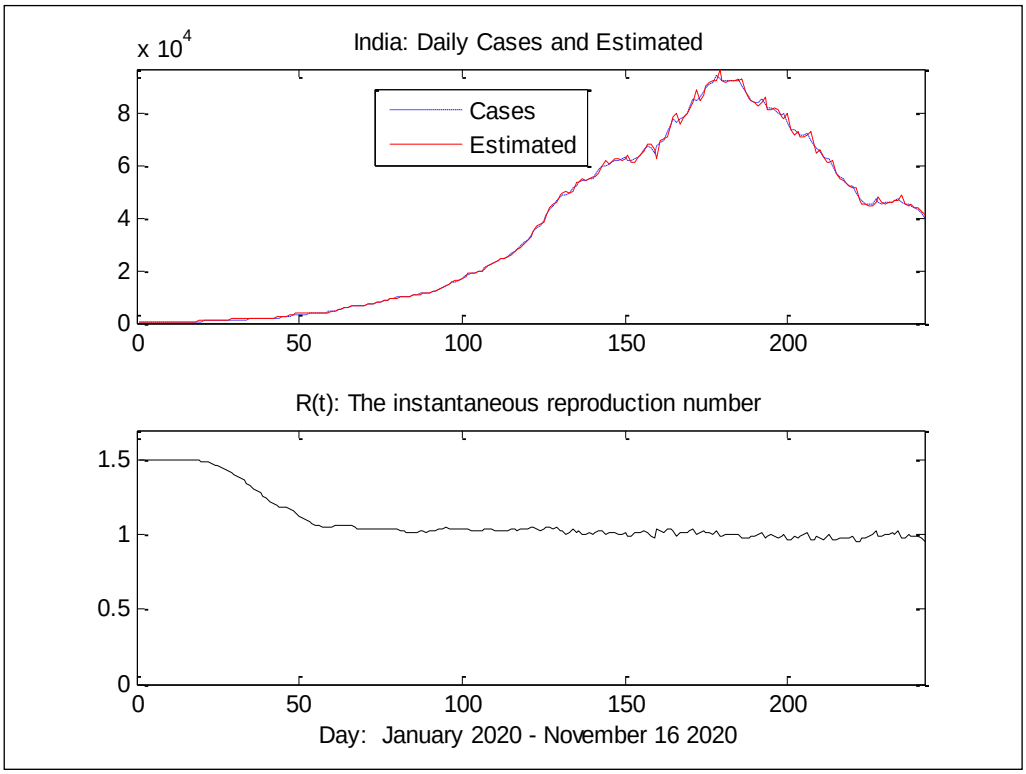

FIGURE 3: India: Daily cases and estimated-reproduction number estimated.

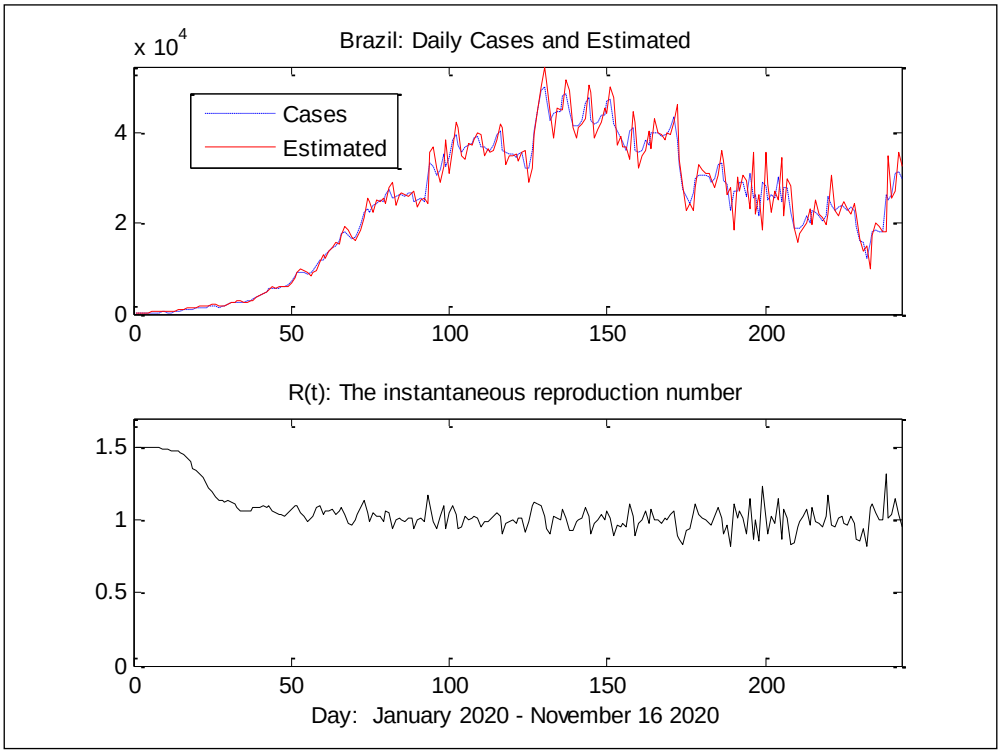

FIGURE 4: Brazil: Daily cases and estimated-reproduction number estimated.

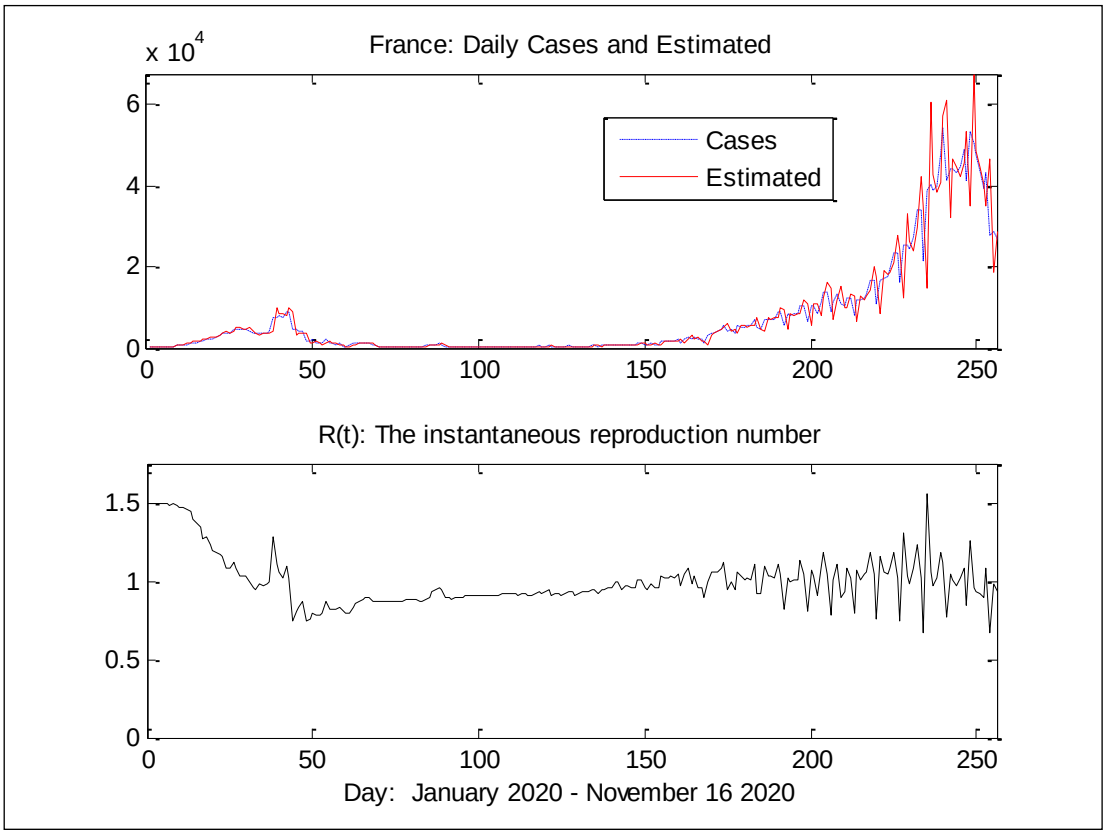

FIGURE 5: France: Daily cases and estimated-reproduction number estimated.

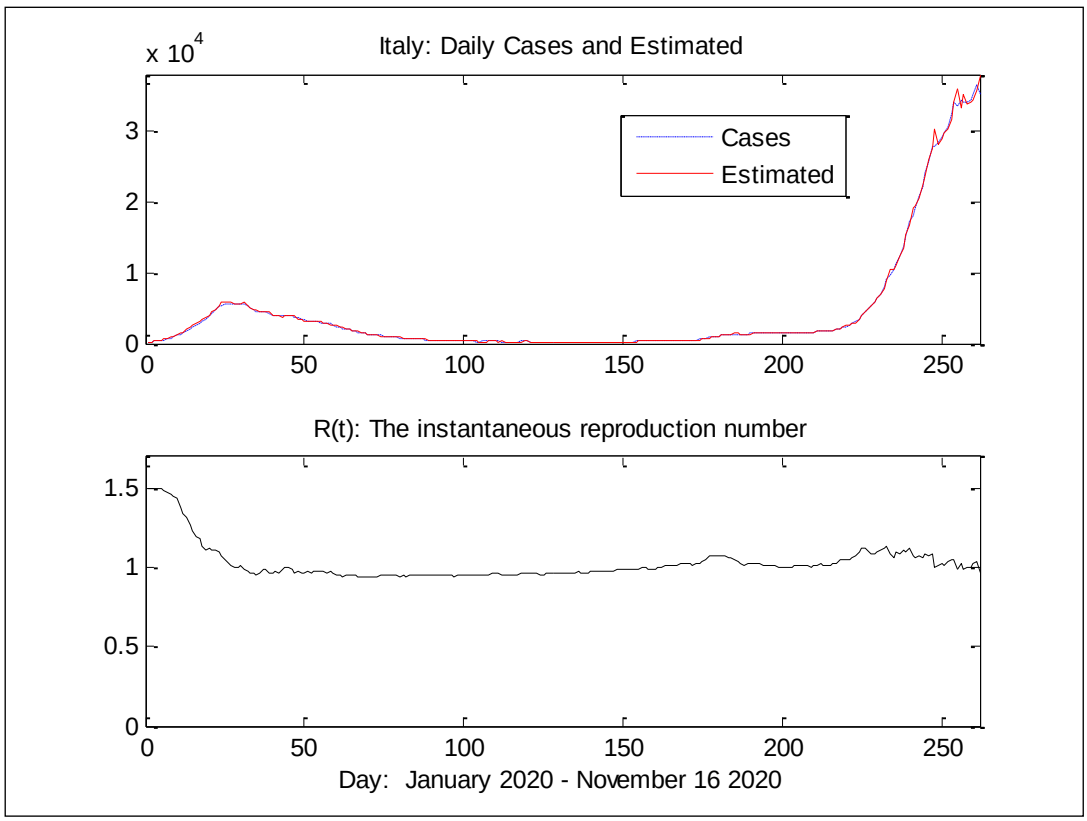

FIGURE 6: Italy: Daily cases and estimated-reproduction number estimated.

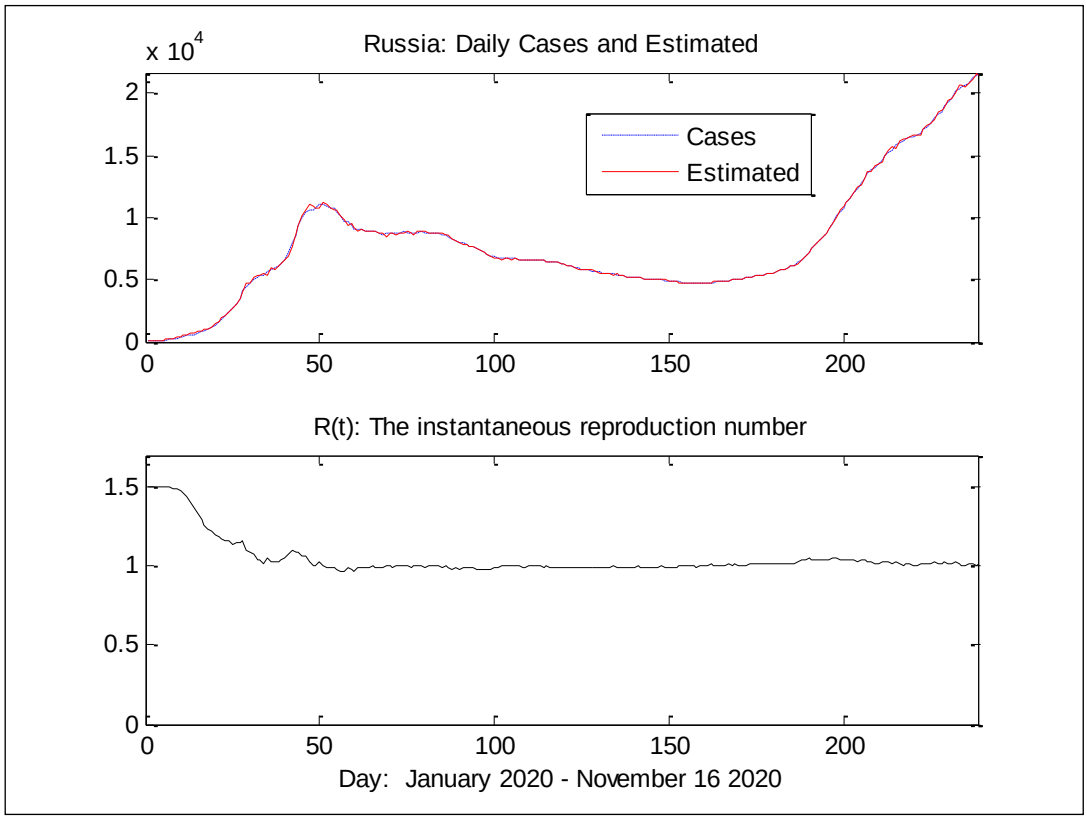

FIGURE 7: Russia: Daily cases and estimated-reproduction number estimated.

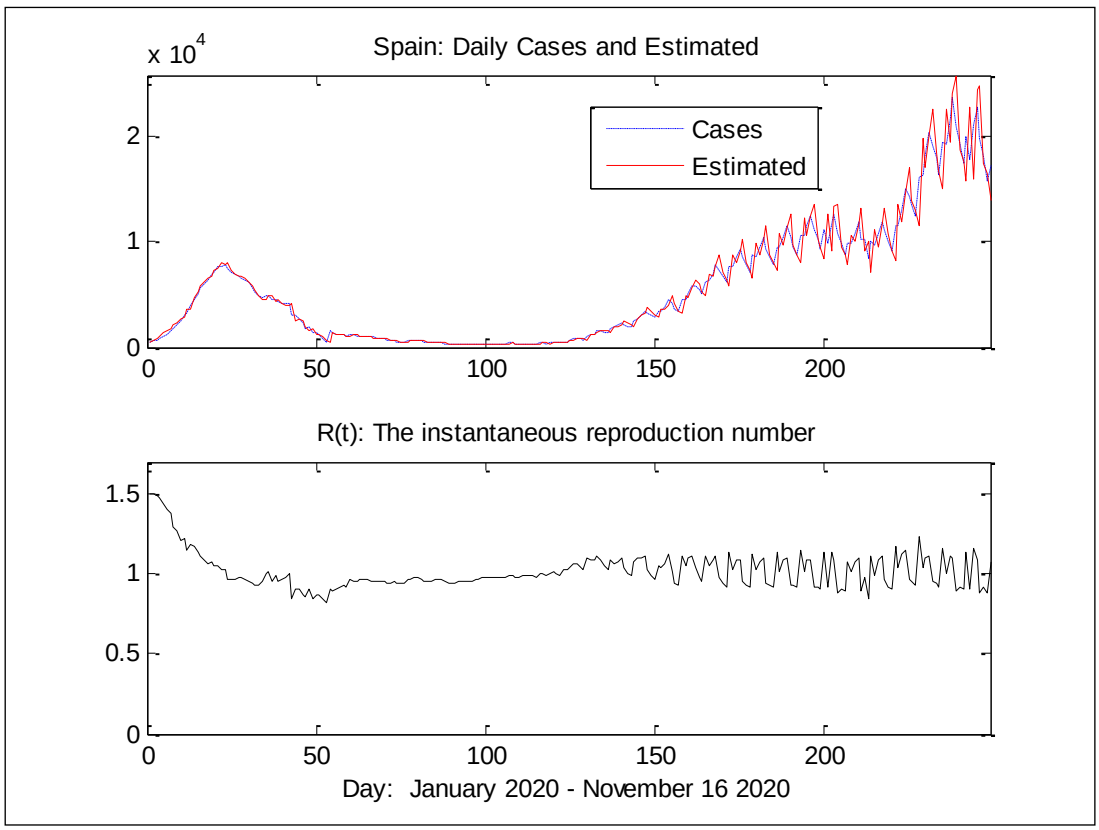

FIGURE 8: Spain: Daily cases and estimated-reproduction number estimated.

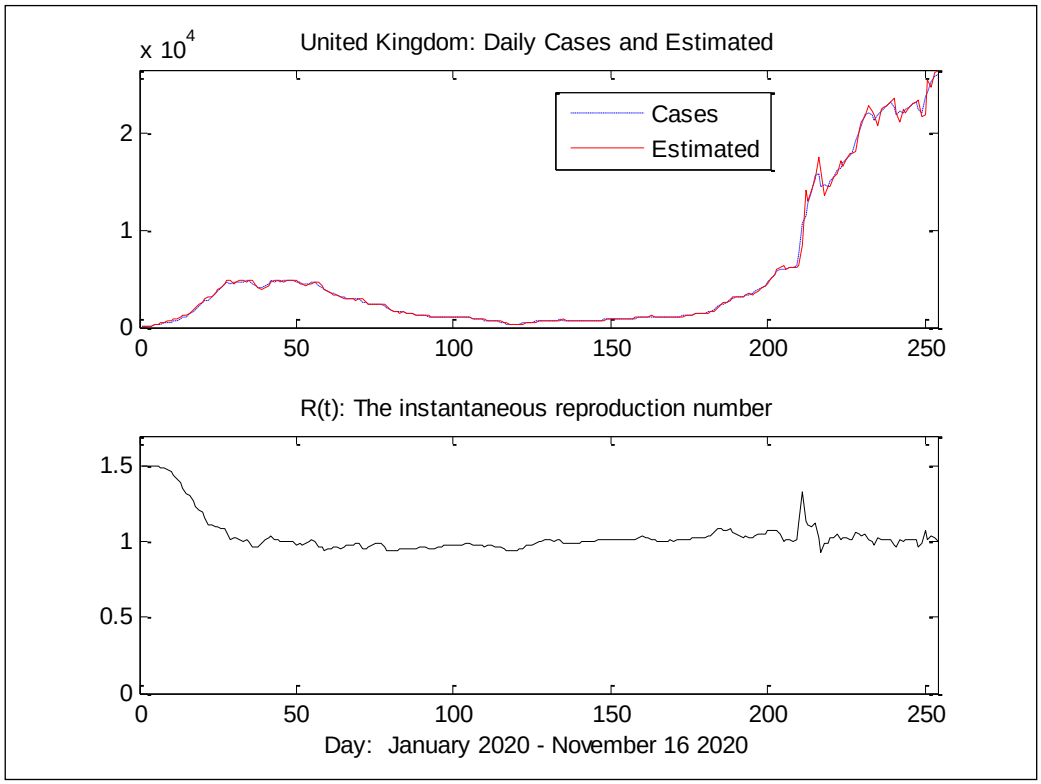

FIGURE 9: United Kingdom: Daily cases and estimated-reproduction number estimated.

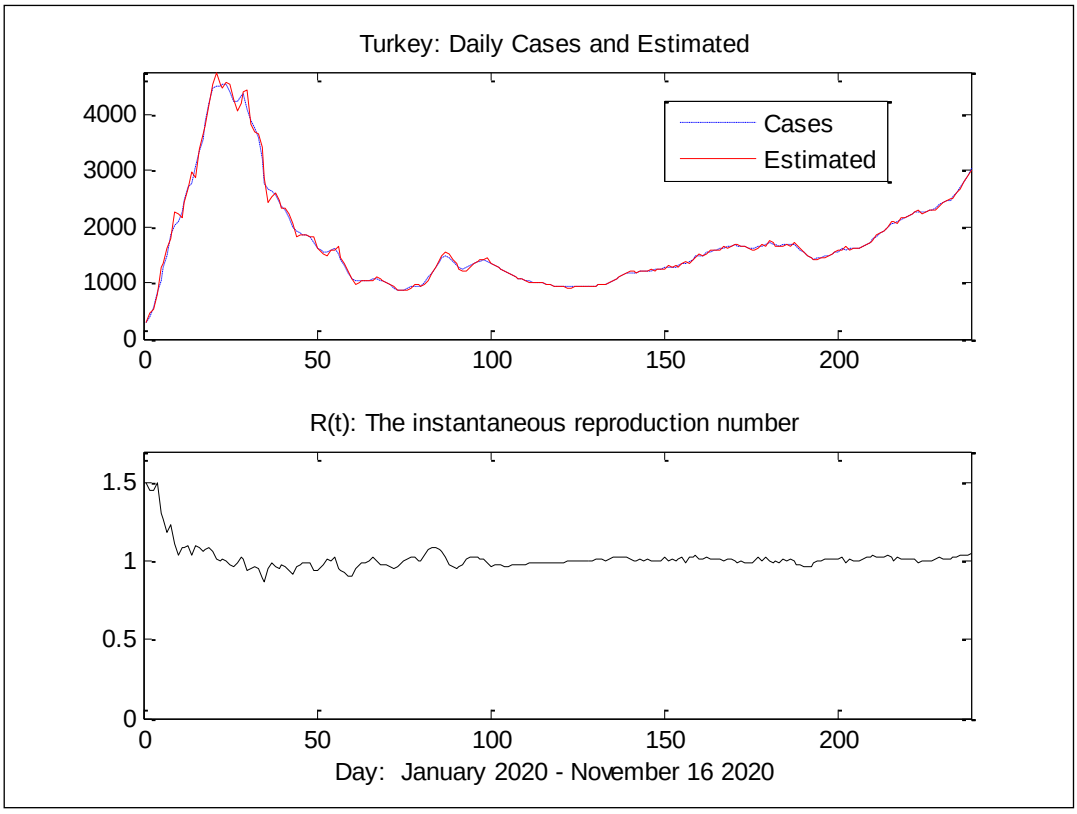

FIGURE 10: Turkey: Daily cases and estimated-reproduction number estimated.

TABLE 2: Calculated R^2, MSE, MAPE.

Region	R^2	MSE	MAPE
USA	0,9969	2281384	6,83
Turkey	0,9953	3508	4,36
Germany	0,9891	238967	18,74
Brazil	0,9597	8091266	23,78
India	0,9991	916202	16,51
Russia	0,9996	9748	5,41
Spain	0,9543	1489086	25,15
France	0,8962	15774196	71,79
Italy	0,9984	102483	12,60
United Kingdom	0,9970	147300	11,64

MSE: Mean squared error; MAPE: Mean absolute percentage error.

It is quite straightforward to model the daily case number time series with the time-varying parameter AR(1) stochastic process and estimated the time-varying parameter with online AKF. This is a rather simple but very effective approach, and to the authors' best knowledge it has not been employed for this purpose before. It is easy to estimate the daily cases COVID-19 by this method.

ESTIMATING THE REPRODUCTION NUMBER WITH ADAPTIVE KALMAN FILTERING

The instantaneous reproduction number, R_t at time t can be estimated as in Eq. (4).

$$R_t = \frac{E(i_t)}{\sum_{s=1}^{t} i_{t-s} w_s} \tag{4}$$

.[14] In equation (4), i_t stands for the number of new infections generated at time step t. w_s is the probability distribution of the infectivity profile which is dependent on time since the infection of the case. In practice, w_s is approximated by the distribution of the serial interval. Let us express the value of R_t calculated using the AR(1) model with R_t^{AR}. If s = 1 and $w_1 = 1$ are taken in Eq. (4), then Eq. (4) can be written as

$$R_t^{AR} = \frac{E(i_t)}{i_{t-1}} = \frac{\hat{i}_t}{i_{t-1}} = \frac{\hat{\theta}_t i_{t-1}}{i_{t-1}} = \hat{\theta}_t \ , \quad t = 2,3,...,n \tag{5}$$

The estimated R_t value using the AR(1) model is equal to the estimate of the time-varying parameter of the AR(1) model. The value of R_t^{AR} calculated using the Eq. (5) is given in Figure 1, Figure 2, Figure 3, Figure 4, Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10. There is no need for any other modeling assumptions in estimating R_t with this method by using AR(1) model. Modeling the daily case time-series with the time-varying parameter AR(1) stochastic process and estimating the time-varying parameter with AKF both estimates the number of daily cases and the estimation of the instantaneous reproduction number without any other component. It is a simple method to model the daily case number time series with the time-varying parameter AR(1) stochastic process and estimating the time-varying parameter with online AKF.

DISCUSSION

In the literature, non-linear mathematical growth models are often used to estimate the number of COVID-19 cases. Specific algorithms such as mathematical optimization technique need to be employed for parameter estimation. In this paper, a novel method for estimating some key parameters of real COVID-19 data is proposed for COVID-19 daily cases analysis. The COVID-19 data have been modeled in state space, and AKF has been employed to estimate the number of daily cases. The estimation results obtained for the number of daily cases using the AR(1) model. The estimation of the daily number of cases using the AR(1) model is a simple method. Since the estimation using the AR(1) stochastic process does not require any other modeling assumption, it is a simple approach. As for AKF, utilizing only observations in time t is the most advantageous aspect of this method.

CONCLUSION

Modeling the daily case time-series with the time-varying parameter AR(1) stochastic process and estimating the time-varying parameter with AKF both estimates the number of daily cases and the estimation of the instantaneous reproduction number without any other operation. It is a simple method to model the daily case number time series with the time-varying parameter AR(1) stochastic process and estimated the time-varying parameter with online AKF. Among the studies made on COVID-19 pandemic, modeling the disease progress is emphasized primarily. Modeling the disease progress is substantial for the precautions which will be taken by countries, interventions, and treatments to be administered. It is thought that the method we have proposed will be suitable for the estimation of the forthcoming progress. We suggest that the simplest method for the reproduction number estimation will be obtained by modeling the daily case time series via an AR(1) model.

APPENDIX: Discrete time state-space model and AKF

Let's consider a discrete-time state-space model stated as

$$x_{t+1} = F_t x_t + G_t w_t$$

$$y_t = H_t x_t + v_t$$

where, x_t is a system, y_t is an observation vector. w_t and v_t are white noise sequences. The covariance matrices w_t and v_t are Q_t and R_t. The matrices F_t, H_t, Q_t, R_t are assumed that they are known at time *t*. The filtering problem is the problem of determining the best estimate of its x_t condition, given its observations $Y_t = (y_0, y_1, ..., y_t)$.[13-15] The covariance matrices w_t and w_t are defined by $w_t \sim N(0, Q_t)$, $v_t \sim N(0, R_t)$. Let the initial state be assumed to have a Gaussian distribution in the form of $x_0 \sim N(\bar{x}_0, P_0)$. The optimum update equations for KF are

$$\hat{x}_{t|t-1} = F_{t-1}\hat{x}_{t-1}$$

$$P_{t|t-1} = F_{t-1}P_{t-1|t-1}F_{t-1}^{'} + G_{t-1}Q_{t-1}G_{t-1}^{'}$$

$$K_t = P_{t|t-1}H_t^{'}(H_t P_{t|t-1}H_t^{'} + R_t)^{-1}$$

$$P_{t|t} = [I - K_t H_t]P_{t|t-1}$$

$$\hat{x}_t = \hat{x}_{t|t-1} + K_t(y_t - H_t\hat{x}_{t|t-1})$$

.[15-17] In the above equations $\hat{x}_{t/t-1}$ is the a priori estimation and $\hat{x}_t$ is the a-posteriori estimation of x_t. Furthermore, $P_{t|t-1}$ and $P_{t|t}$ are the covariance of a priori and a-posteriori estimations, respectively. In order to eliminate divergence in the KF, adaptive methods are used forgetting factor is proposed by Özbek, Özbek and Aliev.[18,19]

$$P_{t|t-1} = \alpha\left(F_{t-1}P_{t-1|t-1}F_{t-1}^{'} + G_{t-1}Q_{t-1}G_{t-1}^{'}\right)$$

α is the forgetting factor proposed by Özbek and Aliev.[18]

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

All authors contributed equally while this study preparing.

REFERENCES

1. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020;5(4):536-44. [Crossref] [PubMed] [PMC]
2. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med. 2020;382(13):1199-207. [PubMed] [PMC]
3. World Health Organization, Weekly operational update on COVID-19 - 6 November 2020, date of access: November 7 2020. [Link]
4. Jia L, Li K, Jiang Y, Guo X, Zhao T. Prediction and analysis of coronavirus disease 2019. 2020; arXiv. [Link]
5. Castorina P, Iorio A, Lanteri D. Data analysis on coronavirus spreading by macroscopic growth laws. International Journal of Modern Physics C. 2020;31(7):1-12. [Crossref]
6. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infect Dis Model. 2020;5:256-63. [Crossref] [PubMed] [PMC]
7. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China: February 13-23, 2020. J Clin Med. 2020;9(2):596. [Crossref] [PubMed] [PMC]
8. Munayco CV, Tariq A, Rothenberg R, Soto-Cabezas GG, Reyes MF, Valle A, et al; Peru COVID-19 working group. Early transmission dynamics of COVID-19 in a southern hemisphere setting: Lima-Peru: February 29th-March 30th, 2020. Infect Dis Model. 2020;5:338-45. [Crossref]
9. Torrealba-Rodriguez O, Conde-Gutiérrez RA, Hernández-Javier AL. Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models. Chaos Solitons Fractals. 2020;138:109946. [Crossref] [PubMed] [PMC]
10. Mazurek J, Neničková Z. Predicting the number of total COVID-19 cases in the USA by a Gompertz curve. 2020. [Link]
11. Català M, Alonso S, Alvarez-Lacalle E, López D, Cardona PJ, Prats C. Empirical model for short-time prediction of COVID-19 spreading. PLoS Comput Biol. 2020;16(12):e1008431. [Crossref] [PubMed] [PMC]
12. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PLoS One. 2020;15(3):e0231236. [Crossref] [PubMed] [PMC]
13. Johns Hopkins University Center for Systems Science and Engineering, 2020 date of access, November 18, 2020. [Link]
14. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013;178(9):1505-12. [Crossref] [PubMed] [PMC]

15. Kalman RE. A new Approach to linear filtering and prediction problems. J Basic Eng. 1960;82(1):35-35. [Crossref]
16. Anderson BDO, Moore JB. Optimal Filtering. 1st ed. New Jersey/ABD: Prentice Hall; 1979.
17. Grewal MS, Andrews AP. Kalman Filtering Theory and Practice. 1st ed. New Jersey/ABD: Prentice Hall; 1993.
18. Özbek L, Aliev FA. Comments on adaptive Fading Kalman Filter with an application. Automatica. 1998;34(12):1663-4. [Link]
19. Ozbek L, Efe M. An adaptive extended kalman filter with application to compartment models. Communications in Statistics-Simulation And Computation. 2004;33(1):145-58. [Crossref]