ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

2021, 22(3), pp. 239-259, DOI:10.18038/estubtda.840307

RESEARCH ARTICLE

ESTIMATION OF DAILY CASES OF COVID-19 AND REPRODUCTION NUMBER IN USA, GERMANY, INDIA, RUSSIA, ITALY, SPAIN, FRANCE, UNITED KINGDOM, BRAZIL USING DISCRETE TIME GOMPERTZ MODEL AND ADAPTIVE KALMAN FILTER

Levent ÖZBEK [1,*], **Hakan DEMİRTAŞ** [2]

[1] Department of Statistics, Ankara University, Ankara, Turkey
[2] Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, IL, USA

ABSTRACT

In this study, cumulative and daily cases are estimated online using discrete-time Gompertz model (DTGM) and Adaptive Kalman Filter (AKF) based on the total COVID-19 cases between February 29-July 28, 2020 in USA, Germany, India, Russia, Italy, Spain, France, United Kingdom, Brazil. Employing the data collected between February 29 and July 28, 2020, it is showed that the DTGM in conjunction with AKF provides a good analysis tool for modeling the daily cases made using the in terms of mean square error (MSE), mean absolute percentage error (MAPE), and R^2.

Keywords. COVID-19, Gompertz models, Adaptive Kalman filter, Estimation, Reproduction number

1. INTRODUCTION

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 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. [5] have used these two modes in China, South Korea, Italy, and Singapore. Roosa et al. [6] have used Generalized Logistic Growth Model (GLM) for the data gathered between February 5 and February 24, 2020, for China. Roosa et al. [7] have used the Generalized Logistic Growth Model (GLM) and Richard model for the data gathered between February 13 and February 20, 2020 for China. Munayco et al. [8] have used the Generalized Growth Model for the dates February 29 and March 30, 2020, for Peru. Gompertz, Logistic, and Artificial Neural Network models were applied in [9]. Zuzana et al. [10] used the Gompertz curve to model a trajectory of the number of infections for the USA. Cata et al [11] employed the Gompertz function in several countries to make short-time predictions. Petropoulos et al. [12] adopted simple time series forecasting approaches. In [4], Logistic, Bertalanffy and Gompertz non-linear mathematical growth models are studied and Prediction and analysis is given for Coronavirus Disease. The prediction methods of Logistic model, Gompertz model and Bertalanffy model 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 [5], only Gompertz non-linear mathematical growth model is studied and apllied to China, South Korea and Italy

*Corresponding Author: ozbek@science.ankara.edu.tr
Received: 14.12.2020 Published: 29.09.2021

data. They considered The cumulative number of infected people and stated that this analysis needs to be updated on a daily basis. In [6], The generalized logistic growth model (GLM), exponential growth dynamics model and The Richards models are used and applied to the data from Hubei and other. Mean squared error (MSE) is used as performance criterion. In [7], similar to [6] logistic growth model, the Richards growth model, and a sub-epidemic wave model models are use and the data from Guangdong and Zhejiang provinces in China. In [8], the generalized growth model (GGM) differential equation is used an applied to Lima-Peru data. In [9], 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. In [10], only Gompertz model is used and applied to the USA data. In [11], only Gompertz model is used and applied to data obtained from different provinces in China. In [12], only exponential smoothing model is studied and applied to Global confirmed cases.

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 models and defined using differential equations. Specific algorithms such as mathematical optimization technique need to be employed for parameter estimation. The data used in the models employed need to be updated daily in order to analyze it. The methods used are offline and all data up to a spesific date is needed 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.

The Gompertz model is well known and widely used in many sub-fields of biology. The Gompertz model was originally recommended to explain human mortality curves Gompertz (1825) [13], and it has been further used in the description of growth processes, for example, growing of bacterial colonies Zwietering et al. [14] and tumors Gerlee [15]. Numerous parametrizations and re-parametrizations of the Gompertz model can be found in the literature Kathleen [16].

The model, a stochastic version of the Gompertz model, can be transformed into a linear Gaussian state-space model for convenient fitting to time-series data. The study makes an emphasis on modeling and estimating the cumulative cases and daily cases of COVID-19 in USA, Germany, India, Russia, Italy, Spain, France, United Kingdom, and Brazil using DTGM and AKF in order to make estimations on the COVID-19 progress in these regions. This paper presents the use of AKF in the analysis of the COVID-19 cumulative cases and daily cases. This work presents the modeling and estimation of cumulative cases and daily cases of COVID-19 infection in these regions through mathematical and computational models using only the confirmed cases provided by the daily technical reports of COVID-19 until July 28th. Here, we employ the DTGM to analyze the dynamics of the spreading of COVID-19 to make short-term estimations of the new cases for the subsequent days. We use the DTGM for the growing process, for the modeling of the cumulative cases and daily cases of COVID-19. With the DTGM, we calculated the instantaneous reproduction number with daily case time series at the modeling and estimation stages.

The rest of this article is organized as follows. In Section 2, the mathematical and computational methodologies are specified and mathematical equations which will be used further in this study are given, and the modeling analysis and estimation results are also presented. In Section 3, the computation of the reproduction number with AKF is presented. Finally, the last section presents the conclusions.

2. DISCRETE-TIME GOMPERTZ MODEL

The underlying model we use for COVID-19 cumulative cases is a DTGM. Let n_t denote COVID-19 cumulative cases at time t. The process model is

$$n_t = n_{t-1}\exp(a + b\ln n_{t-1} + e_t) \quad (1)$$

where a and b are constants, and e_t is $e_t \sim N(0,\sigma^2)$. The random variables $e_1, e_2, ..., e_n$ are assumed to be uncorrelated. On the logarithmic scale, the DTGM is a linear, autoregressive time-series model of order 1 [AR (1) process].

$$y_t = y_{t-1} + a + by_{t-1} + y_t = a + cy_{t-1} + e_t \quad (2)$$

where $y_t = \ln n_t$ and $c = b + 1$. The statistical properties of the DTGM are well-known Dennis et al (2006) [17].

2.1. Mathematical and Computational Methodologies

The optimum linear filtering and estimations methods introduced by Kalman (1960) have been considered one of the greatest achievements in estimation theory.

Discrete-time linear state-space models and Kalman filtering (KF) have been employed since the 1960s, mostly in the control and signal processing areas. The KF has been extensively employed in many areas of estimation the extensions and applications of discrete-time linear state-space models can be found in almost all disciplines [18-26].

In this work, Kalman filtering[1] has been used to estimate the time-varying parameter of the discrete-time Gompertz model. KF is a recursive estimator to estimate the time-varying parameters. If $a = 0$ in Eq.(2), n_t being the case counts observed until t and $y_t = \ln n_t$, equation

$$y_t = cy_{t-1} + e_t \quad (3)$$

is acquired. In the case where the c parameter in Eq.(3) is time-varying and presumed as

$$c_t = c_{t-1} + w_t$$

random walk process, state-space model

$$y_t = c_t y_{t-1} + e_t \quad (4)$$

$$c_t = c_{t-1} + w_t \quad (5)$$

is written. Here, the state variable is an unobservable, time-varying c_t parameter, and can be estimated through AKF (explanation regarding AKF is given in the Appendix section). If this time-varying parameter is estimated using on-line AKF, estimation for the total case counts in times $t+1$, $t+2$, ... can be made via this online-estimated parameter.

[1] Kalman filter is in fact an estimator rather than a conventional filter, however it is employed to estimate parameters from a noisy data sequence, hence the name filter.

Actual cumulative case estimations that have been made online using AKF. The number of daily cases can be easily calculated with $i_t = n_t - n_{t-1}$ to show the total number of cases up to n_t, t days. Since we have the estimates of n_t, we can easily find the estimations of i_t with $\hat{i}_t = \hat{n}_t - \hat{n}_{t-1}$. The data used was taken from Johns Hopkins University [31].

Daily cases and estimations are given in odd-numbered figures. As can be seen from these figures, the estimation results obtained from the model used are very close to the real values. According to the estimation results obtained by using the daily number of cases in the Gompertz model, MSE, MAPE, and R^2, were calculated (see Table 1). These calculated values indicate that the compatibility of the model with real data is quite high. This situation tells us that estimating the daily number of cases via the Gompertz model is a reliable method. Since estimation using the AR(1) stochastic process does not require any other model assumption, it is much simpler than the estimation method through the Gompertz model. As for AKF, utilizing only the observation in time t and the preceding estimation is the most advantageous aspect of this method.

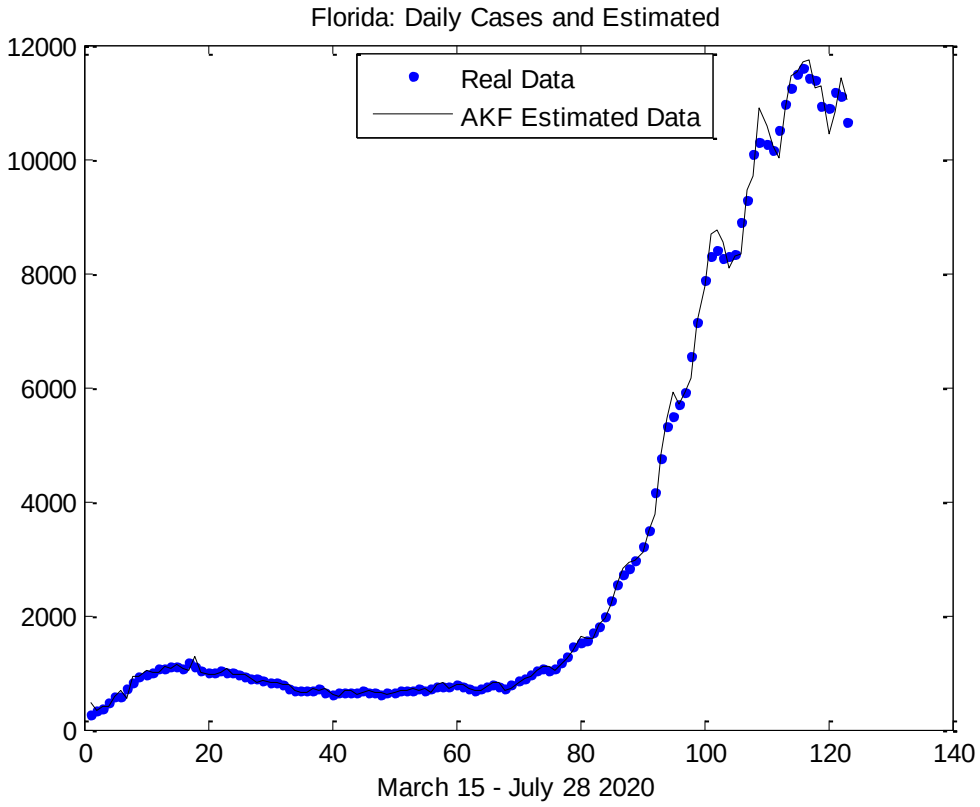

Figure 1. USA-Florida, daily cases and estimated

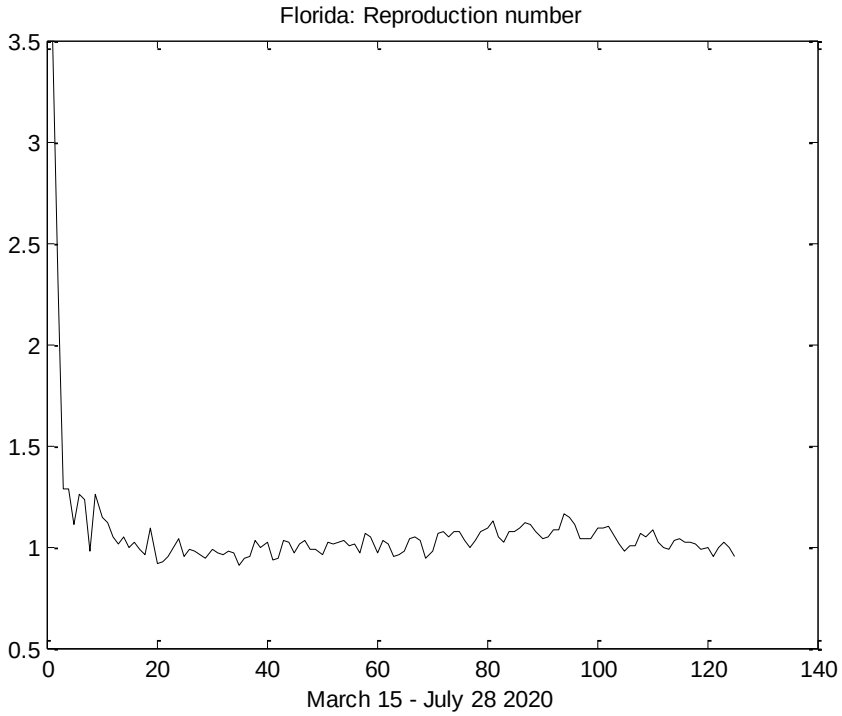

Figure 2. USA-Florida, reproduction number estimated

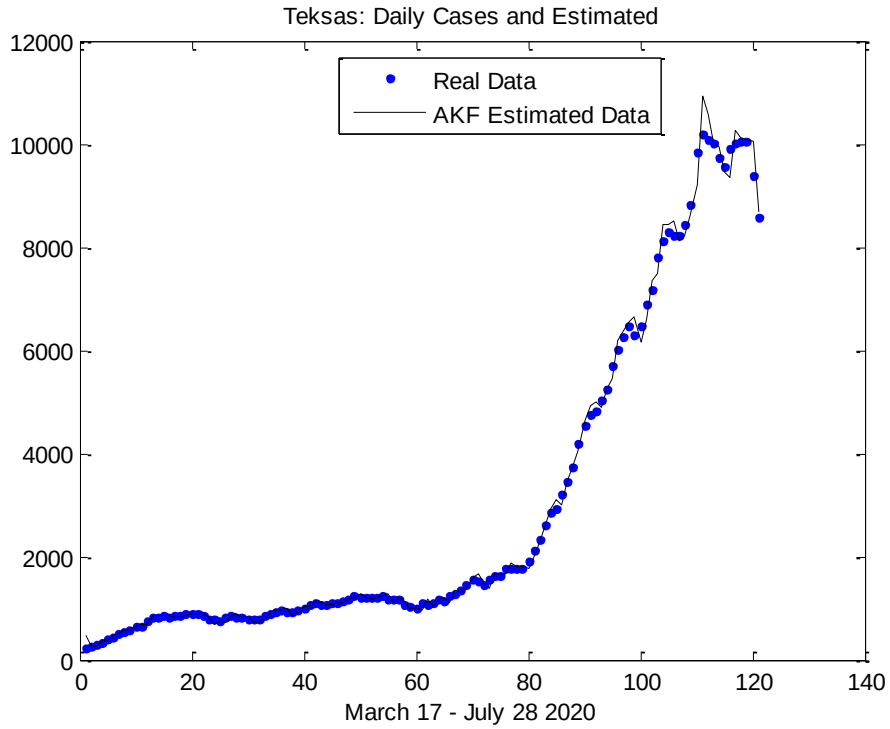

Figure 3. USA-Texas, daily cases and estimated

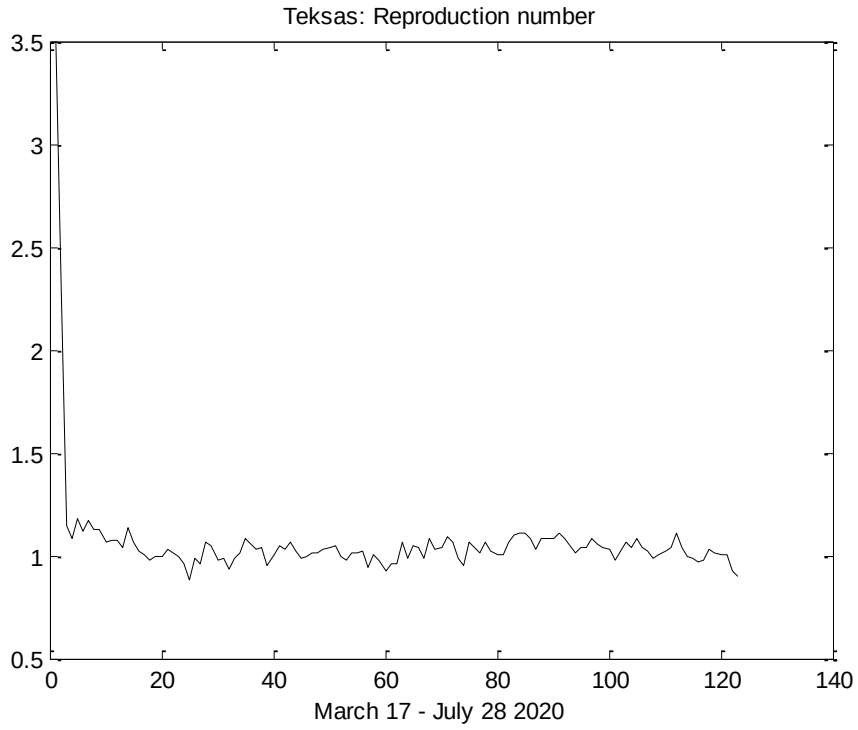

Figure 4. USA-Texas, reproduction number estimated

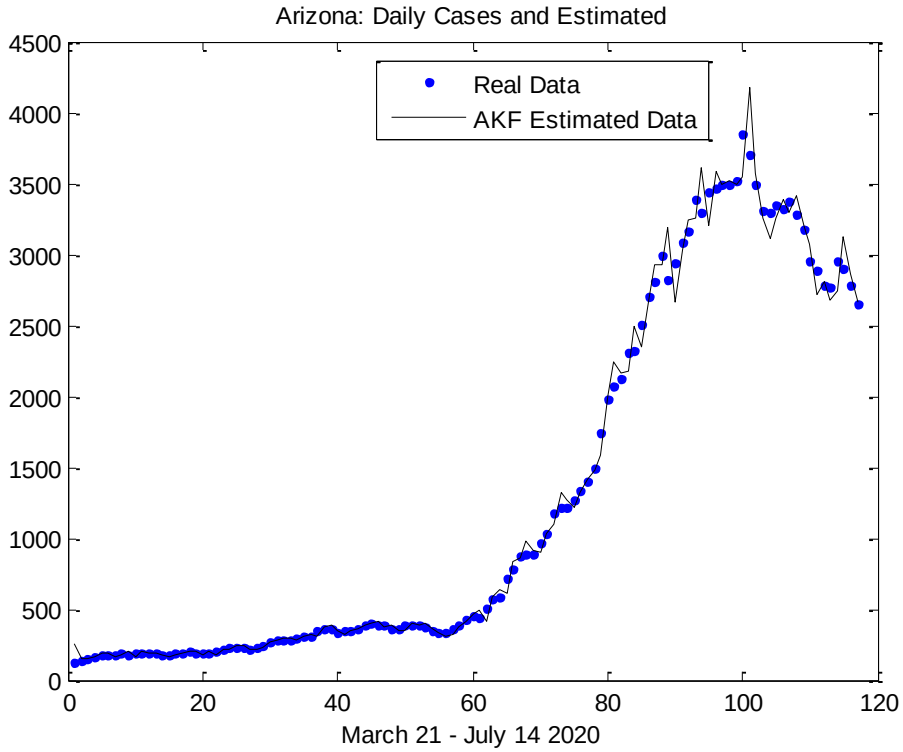

Figure 5. USA-Arizona, daily cases and estimated

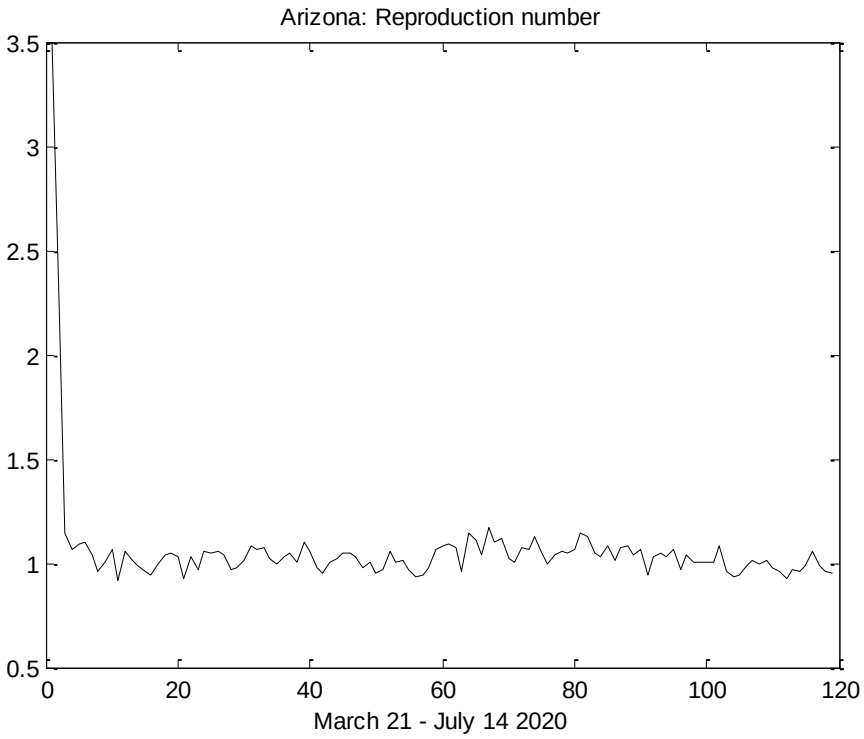

Figure 6. USA-Arizona, reproduction number estimated

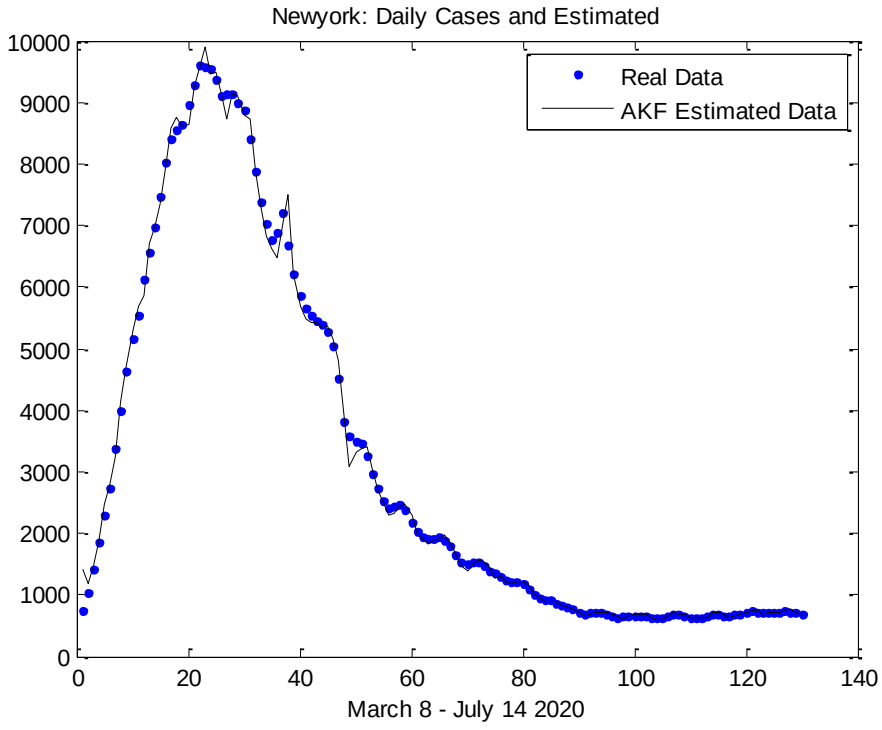

Figure 7. USA-New York, daily cases and estimated

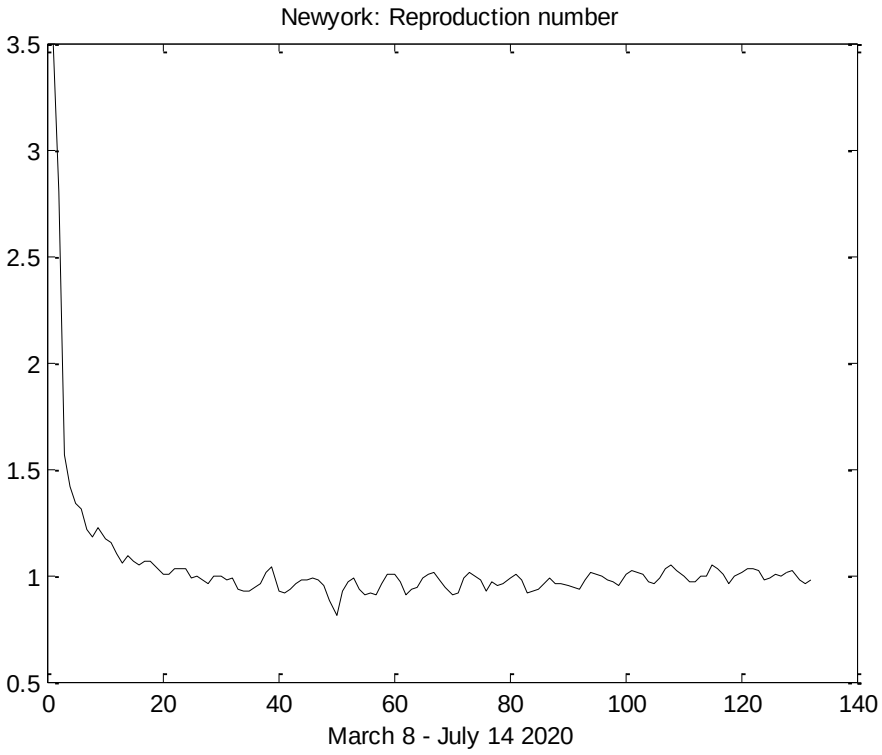

Figure 8. USA-New York, reproduction number estimated

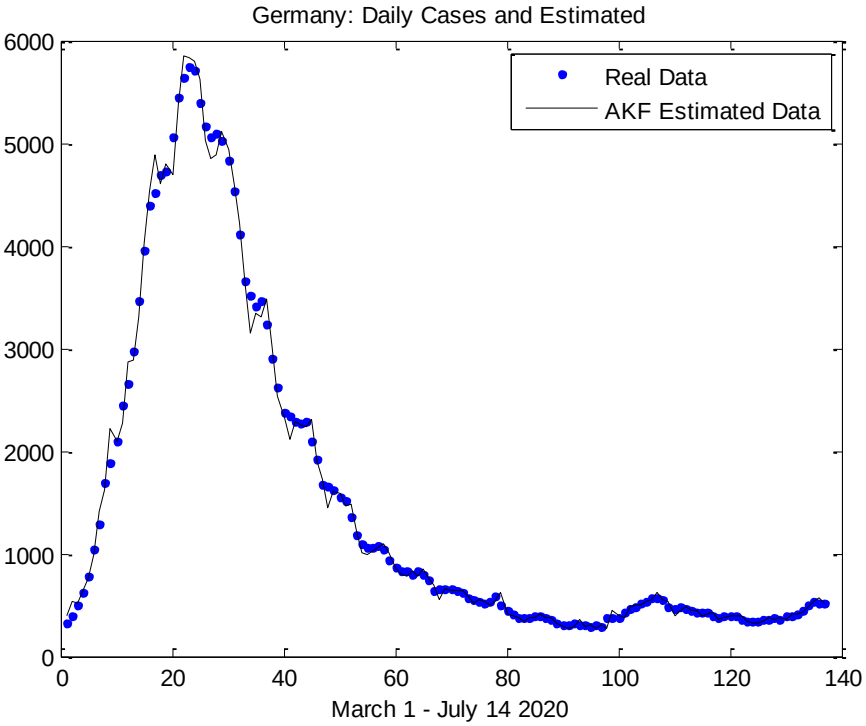

Figure 9. Germany, daily cases and estimated

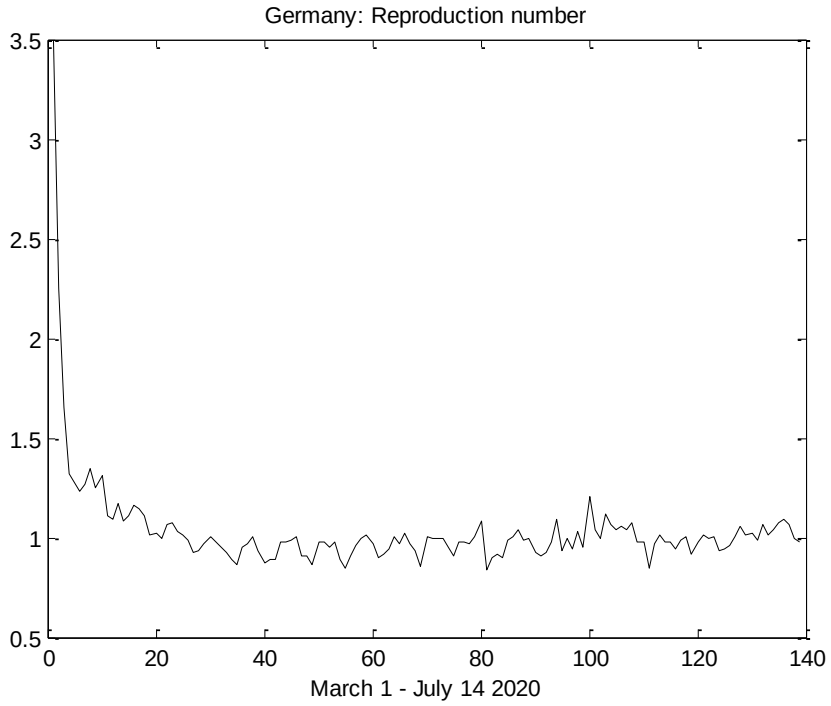

Figure 10. Germany, reproduction number estimated

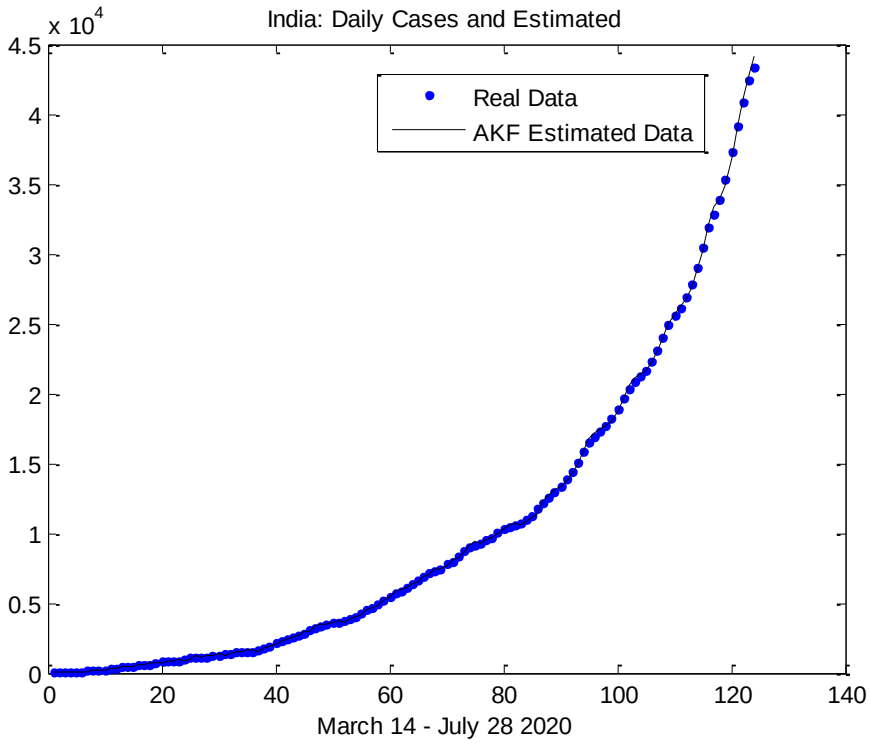

Figure 11. India, daily cases and estimated

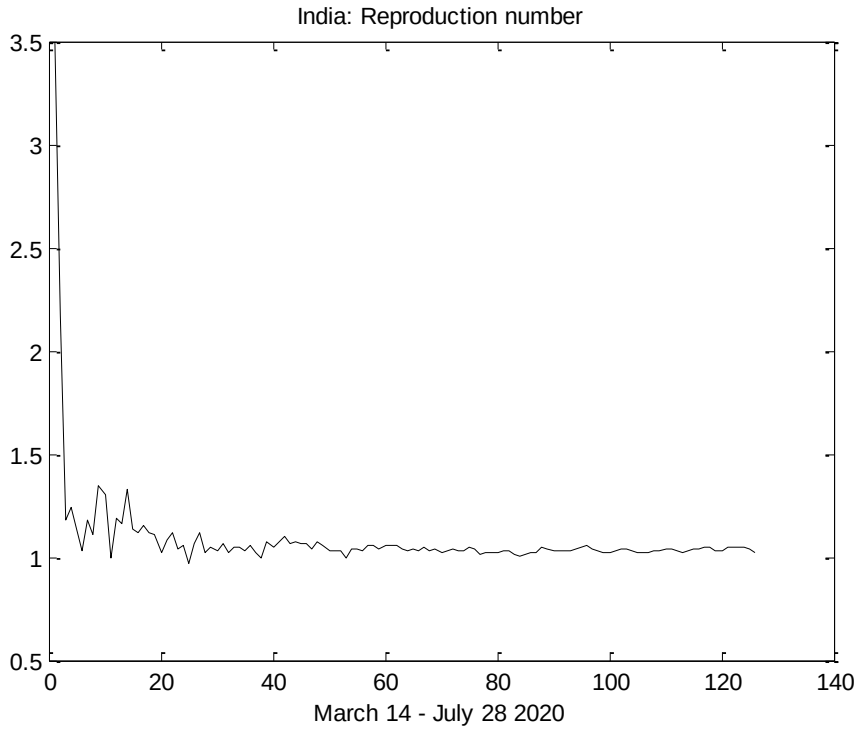

Figure 12. India, reproduction number estimated

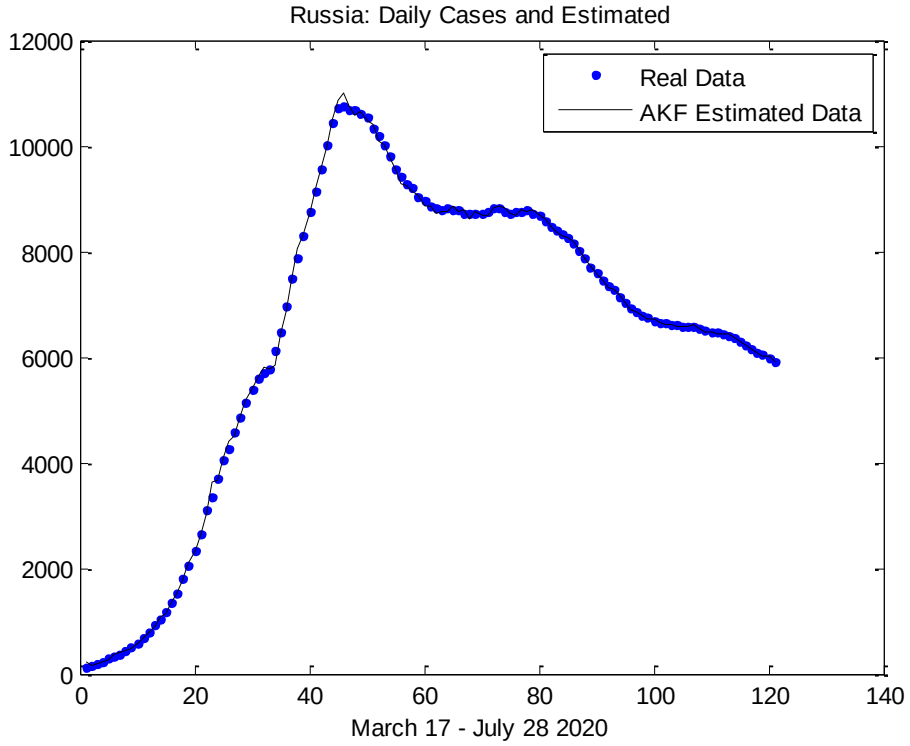

Figure 13. Russia, daily cases and estimated

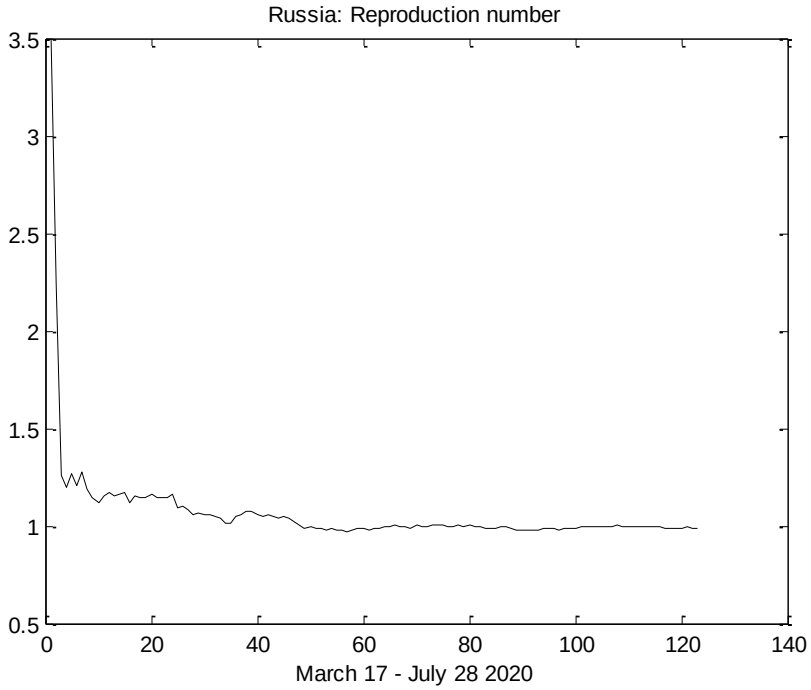

Figure 14. Russia, reproduction number estimated

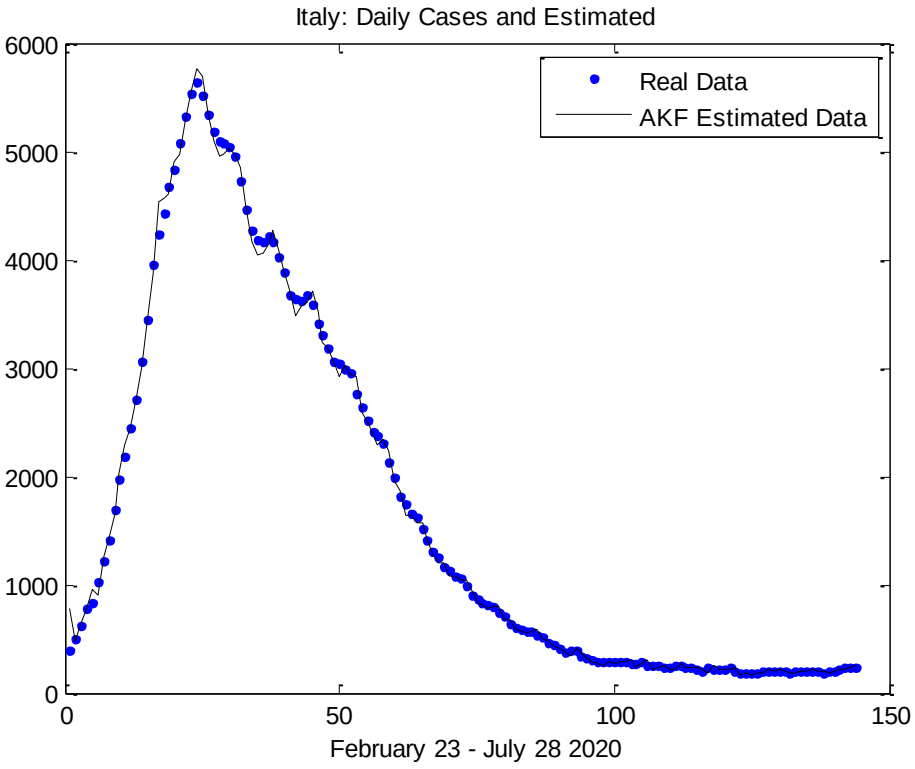

Figure 15. Italy, daily cases and estimated

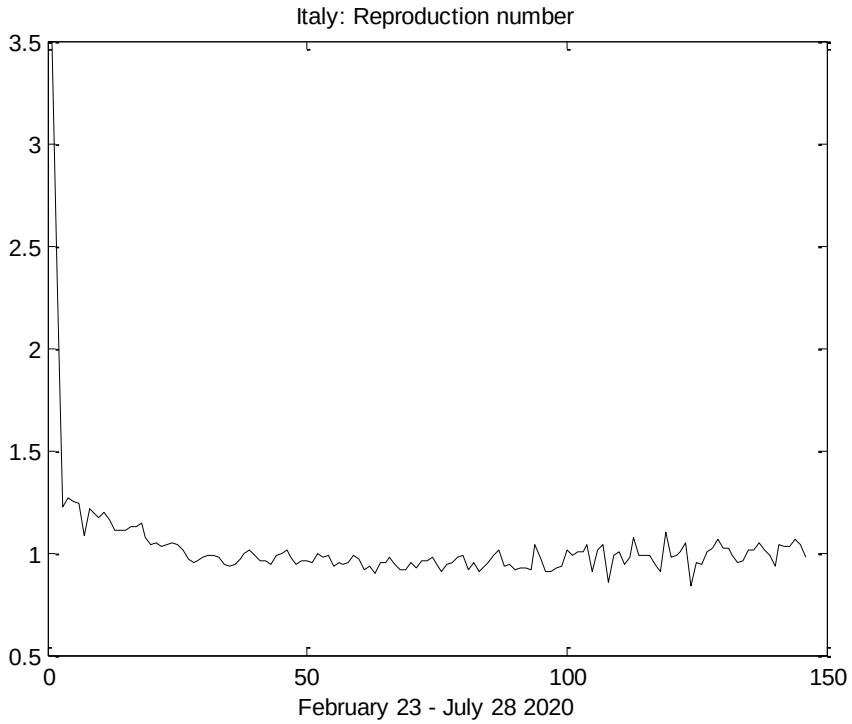

Figure 16. Italy, reproduction number estimated

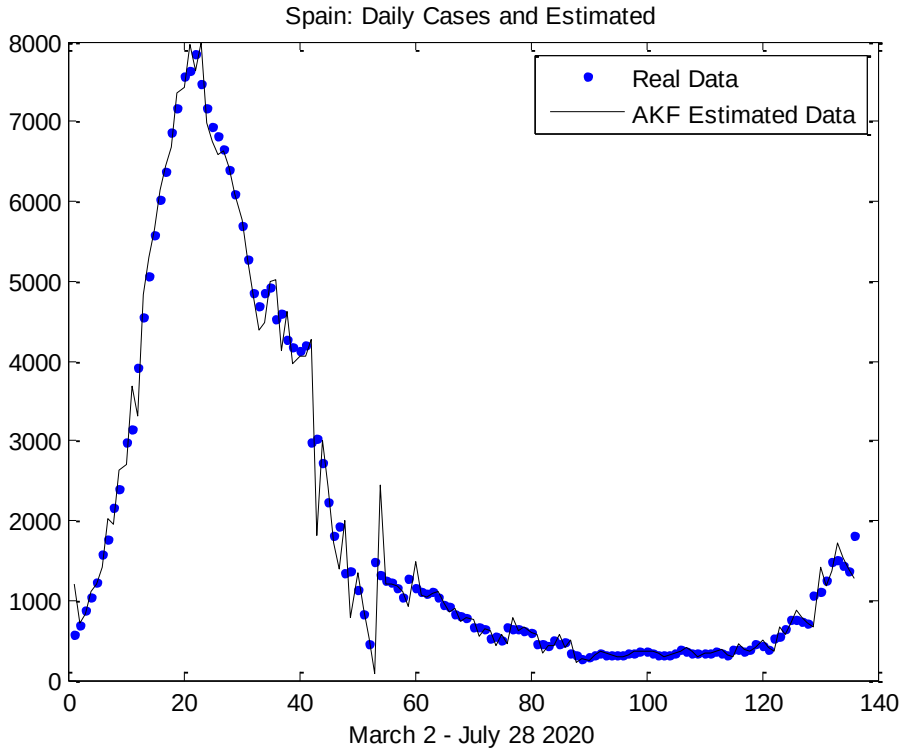

Figure 17. Spain, daily cases and estimated

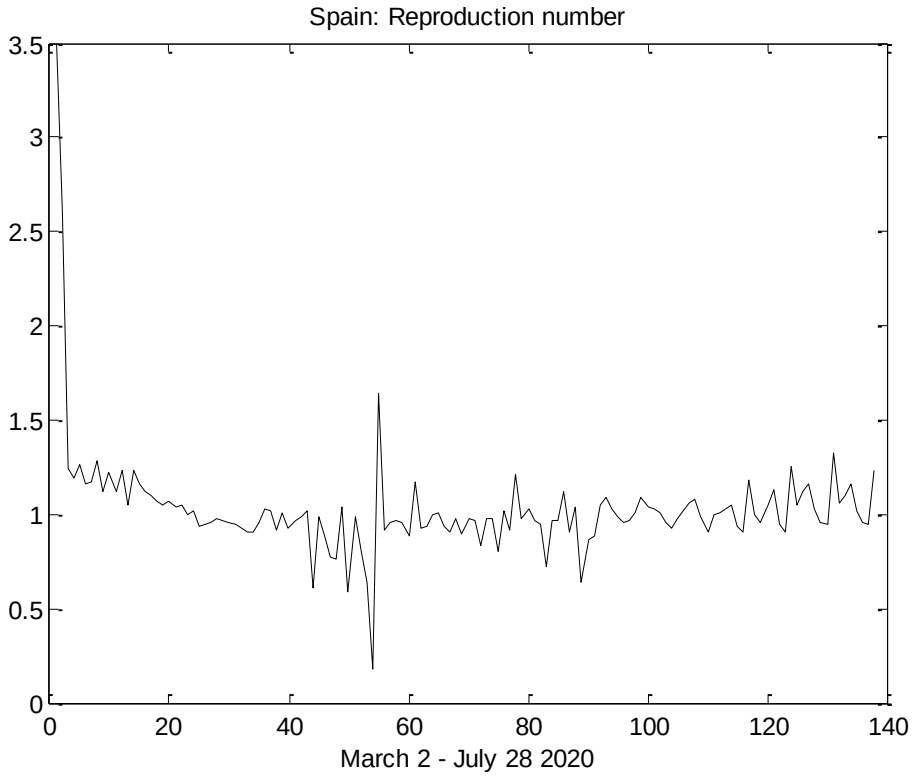

Figure 18. Spain, reproduction number estimated

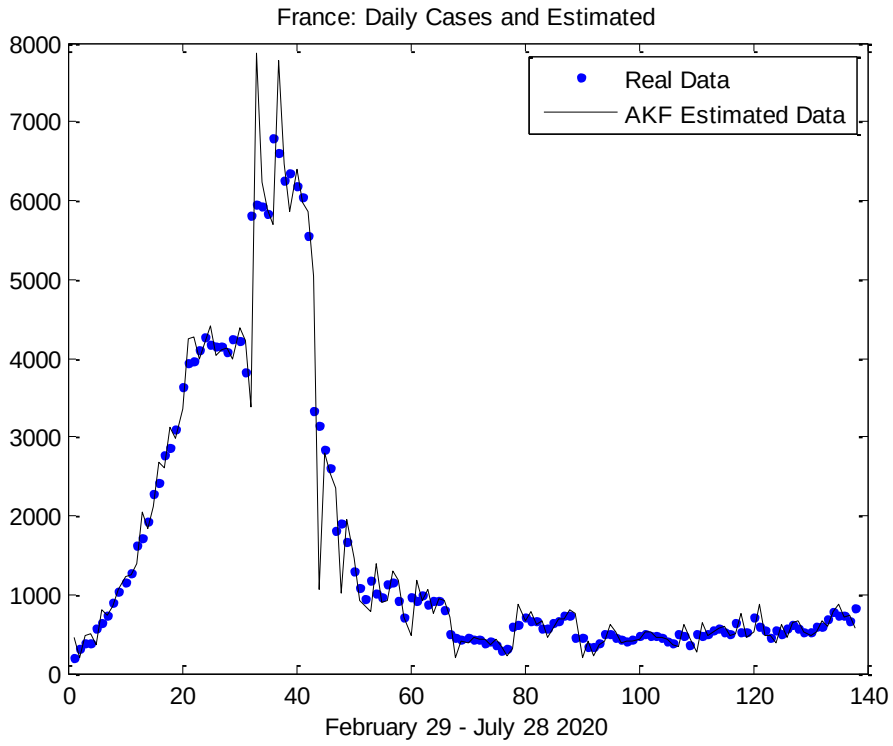

Figure 19. France, daily cases and estimated

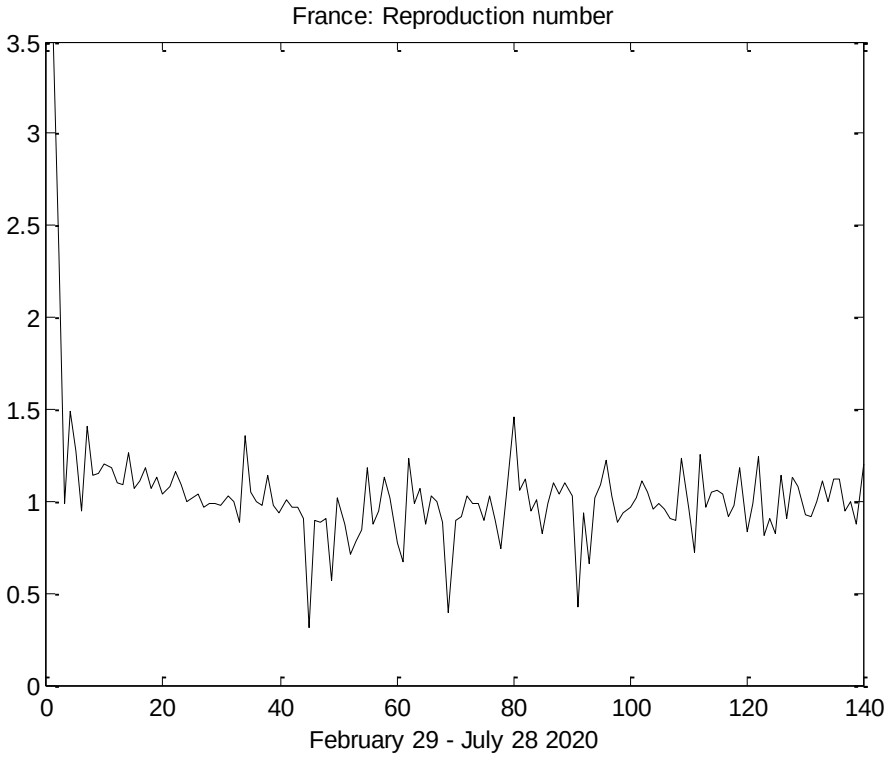

Figure 20. France, reproduction number estimated

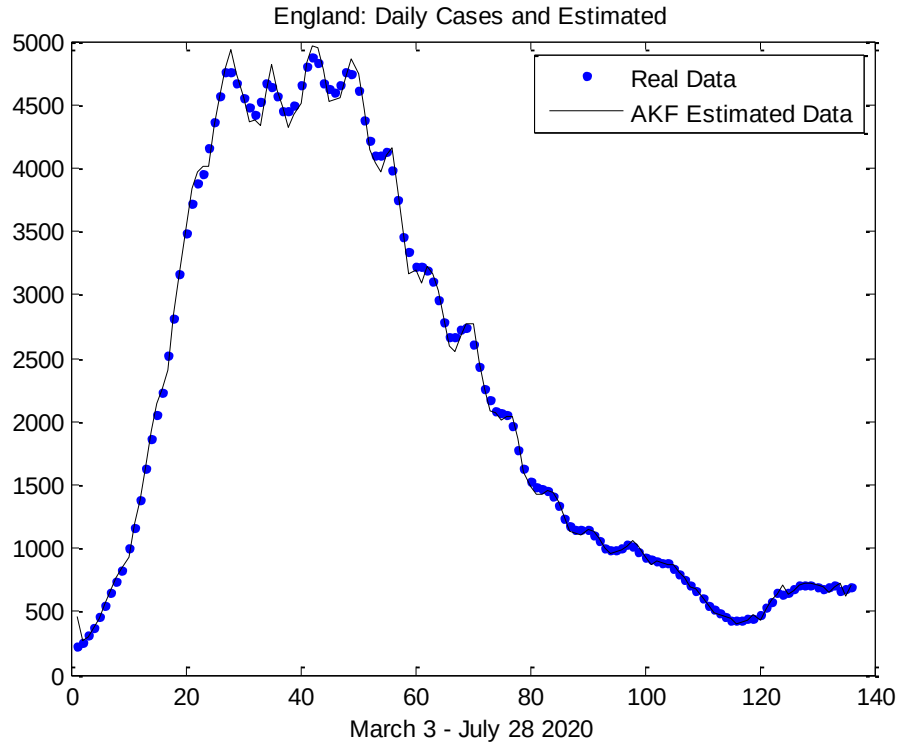

Figure 21. United Kingdom, daily cases and estimated

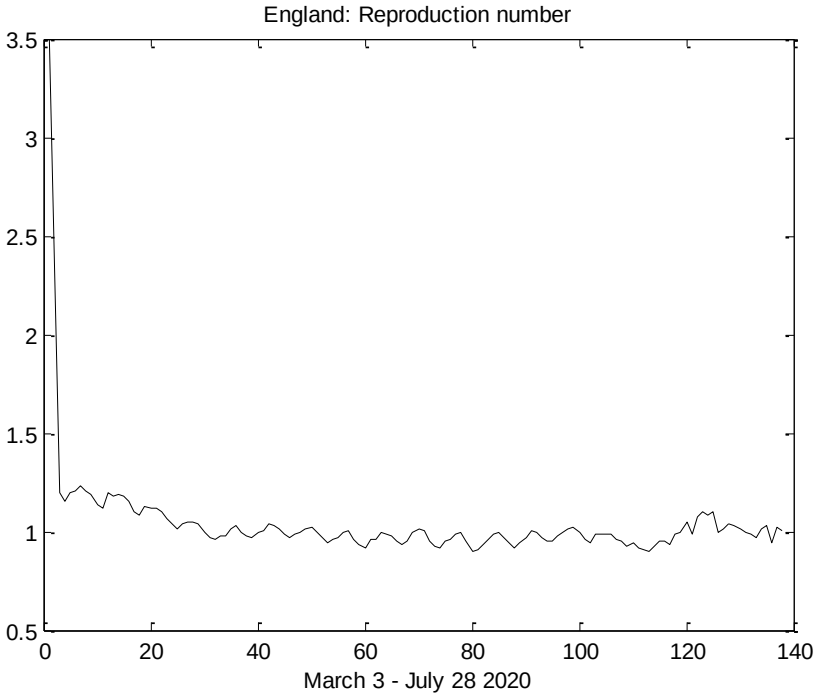

Figure 22. United Kingdom, reproduction number estimated

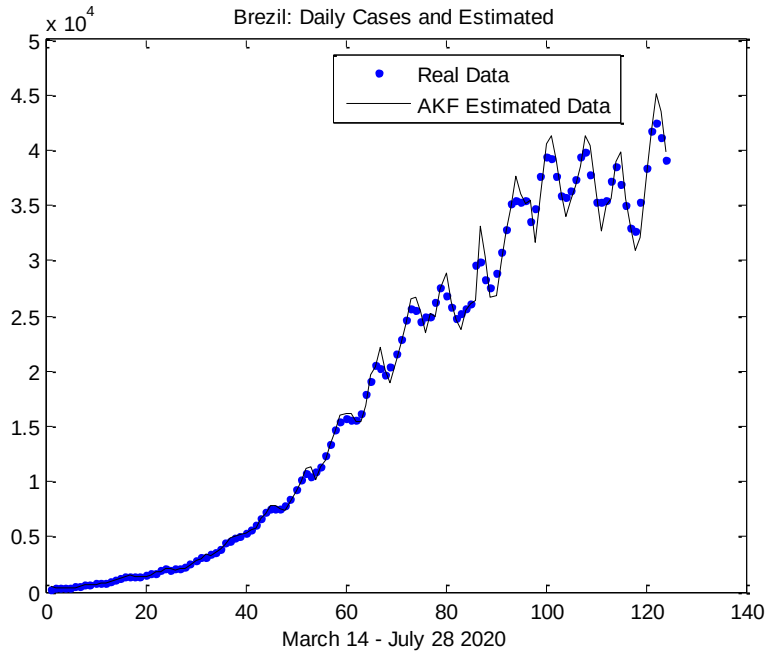

Figure 23. Brazil, daily cases and estimated

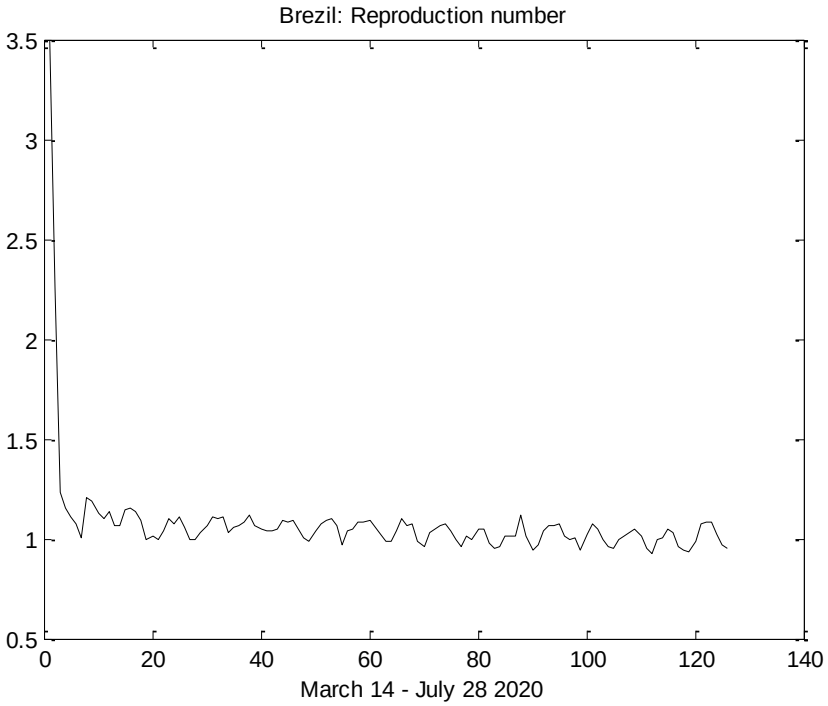

Figure 24. Brazil, reproduction number estimated

Table 1. Calculated R^2, MSE, MAPE

Region	MSE	R^2	MAPE
USA-Florida	27215	0.99806	114,99
USA-Texas	31009	0.99681	113,20
USA-Arizona	10277	0.99359	107,95
USA-New York	24135	0.99723	121,78
Germany	10036	0.99598	127,56
India	23445	0.99982	116,49
Russia	4285	0.99954	113,71
Italy	5360	0.99817	133,45
Spain	88601	0.98109	126,76
France	174127	0.94149	128,88
United Kingdom	5028	0.99802	127,74
Brazil	1209480	0.99418	116,76

3. COMPUTATION OF THE REPRODUCTION NUMBER WITH AKF

The instantaneous reproduction number, R_t at time t can be estimated as in Eq,(8).

$$R_t = \frac{E(i_t)}{\sum_{s=1}^{t} i_{t-s} w_s} \tag{8}$$

where $E(X)$ denotes the expectation of a random variable X [27]. In Eq.(8), w_s is the probability distribution of the infectivity profile which is dependent on time. In practice, w_s is approximated by the distribution of the serial interval. In this study, we have taken the distribution of w_s as a uniform distribution in a $f(w_s)=1,\ s=1$ form. Since $E(i_t)=\hat{i}_t$, Eq.(8) can be written in the form of Eq.(9).

$$R_t^G = \frac{\hat{i}_t}{i_{t-1}},\ t = 2,3,...,n-1 \tag{9}$$

The value of R_t^G (using the Gompertz model) calculated using the Equation (9) is given in even-numbered figures. There is no need for any other model assumption in estimating R_t with this method by using the 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 estimate the number of daily cases and estimate the instantaneous reproduction number without any other operation. It is quite 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.

4. CONCLUSION

In this study, cumulative and daily cases have been estimated online using DTGM and AKF based on the total of COVID-19 cases between February and July 28, 2020 in USA, Germany, India, Russia, Italy, Spain, France, United Kingdom, Brazil. The cumulative case number was modeled with DTGM, and the time-varying parameters of the obtained AR(1) stochastic time series were estimated by on-line AKF. Estimation by acquired data observed between February 29 and July 28, 2020 shows that employing the DTGM model and AKF in terms of MSE, MAPE, and R^2 provides efficient analysis for modeling the total case. It is proposed that the use of DTGM and AKF will be appropriate. After estimating the number of cumulative cases, the estimation of daily cases was made. After estimating the daily case number, the estimation of reproduction number was obtained. The AR(1) model is an appropriate estimation method for the daily cases. As for AKF, utilizing only the observation in time t and preceding the estimation is the most advantageous aspect of this method. Modeling the cumulative case time-series with the time-varying parameter AR(1) stochastic process and estimating the time-varying parameters with AKF both leads to the number of daily cases and the instantaneous reproduction number without any other operation.

It is quite a simple method to model the cumulative case number time series with the time-varying parameter AR(1) stochastic process and estimate the time-varying parameter with online AKF. Among the studies made on COVID-19 pandemic, the progress of modeling the disease is remarked primarily. The progress of modeling the disease is substantial for the precautions which will be taken by countries and interventions, and treatments to be administered. As a result of estimations by acquired data taken

observed between February 29 and July 28, 2020, it is proposed that the efficient analysis for modeling the total case is to be made using the DTGM and AKF in terms of MSE, MAPE, and R^2. It is thought that the method we have proposed will be suitable for the estimation of the forthcoming progress. Our suggestion is that the simplest method for the estimation of the reproduction number can be performed by modeling the daily case number time series using AR(1).

Appendix. State-Space Model and Adaptive Kalman Filter (AKF)

Let us consider a general discrete-time stochastic system represented by the state and measurement models given by

$$x_{t+1} = F_t x_t + G_t w_t \tag{A1}$$

$$y_t = H_t x_t + v_t \tag{A2}$$

where x_t is an $n\times1$ system vector, y_t is an $m\times1$ observation vector, F_t is an $n\times n$ system matrix, H_t is an $m\times n$ matrix, w_t an $n\times1$ vector of zero mean white noise sequence and v_t is an $m\times1$ measurement error vector assumed to be a zero mean white sequence uncorrelated with the w_t sequence. The covariance matrices w_t and w_t are defined by $w_t \sim N(0,Q_t)$, $v_t \sim N(0,R_t)$. The filtering problem is the problem of determining the best estimate of its x_t condition, given its observations $Y_t = (y_0, y_1, ..., y_t)$ Jazwinski (1970) [18-26]. When $Y_t = (y_0, y_1, ..., y_t)$ observations are given, the prediction of state x_t with

$$\hat{x}_t = E(x_t \,|\, y_0, y_1, ..., y_t) = E(x_t \,|\, Y_t)$$

and the covariance matrix of the error with

$$P_{t|t} = E\left[(x_t - \hat{x}_{t|t})(x_t - \hat{x}_{t|t})' \,|\, Y_t \right]$$

when $Y_{t-1} = (y_0, y_1, ..., y_{t-1})$ observations are given, the prediction of state x_t with

$$\hat{x}_{t|t-1} = E(x_t \,|\, y_0, y_1, ..., y_{t-1}) = E(x_t \,|\, Y_{t-1})$$

and the covariance matrix of the error are shown with

$$P_{t|t-1} = E\left[(x_t - \hat{x}_{t|t-1})(x_t - \hat{x}_{t|t-1})' \,|\, Y_{t-1} \right].$$

Let the initial state be assumed to have a normal distribution in the form of $x_0 \sim \mathrm{N}(\bar{x}_0, \mathrm{P}_0)$. The optimum update equations for KF are,

$$\hat{x}_{t|t-1} = F_{t-1}\hat{x}_{t-1} \tag{A3}$$

$$P_{t|t-1} = F_{t-1}P_{t-1|t-1}F_{t-1}^{'} + G_{t-1}Q_{t-1}G_{t-1}^{'} \tag{A4}$$

$$K_t = P_{t|t-1}H_t^{'}(H_tP_{t|t-1}H_t^{'} + R_t)^{-1} \tag{A5}$$

$$P_{t|t} = [I - K_tH_t]P_{t|t-1} \tag{A6}$$

$$\hat{x}_t = \hat{x}_{t|t-1} + K_t(y_t - H_t\hat{x}_{t|t-1}) \tag{A7}$$

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. Also, $P_{t|t-1}$ and $P_{t|t}$ are the covariance of a priori and a posteriori estimations respectively Jazwinski [18], Anderson and Moore [19]. In some cases, divergence problems may occur in the Kalman Filter due to the incorrect installation of the model. In order to eliminate divergence in the Kalman filter, adaptive methods are used Özbek and Aliev [28], Efe and Özbek [29], Özbek and Efe [30]. One of these is the use of the forgetting factor. A forgetting factor is proposed by Ozbek and Aliev [28].

$$P_{t|t-1} = \alpha\left(F_{t-1}P_{t-1|t-1}F_{t-1}^{'} + G_{t-1}Q_{t-1}G_{t-1}^{'}\right) \tag{A8}$$

CONFLICT OF INTEREST

The authors stated that there are no conflicts of interest regarding the publication of this article.

REFERENCES

[1] Gorbalenya AE, Baker SC, Baric RS, et al. The Species Severe Acute Respiratory Syndrome-Related Coronavirus. Classifying 2019-Ncov and Naming It SARS–Cov-2. Nat Microbiol 2020;5.536–44.

[2] Li Q, Guan X, Wu P, Wang X, Zhou L, Tong, Y, Feng Z. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia N Engl J Med 2020; 382:1199-1207.

[3] World Health Organization. Novel coronavirus (2019-nCoV) situation reports.(2020); https.//www.who.int/emergencies/diseases/novel-coronavirus-2019/.

[4] Jia L, Li K, Jiang Y, Guo X, Zhao, T. Prediction and analysis of coronavirus disease 2019. 2020; arXiv preprint arXiv.2003.05447.

[5] Castorina, P, Iorio A, Lanteri D. Data analysis on coronavirus spreading by macroscopic growth laws.2020; arXiv preprint arXiv.20 03.0 0507.

[6] Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, Yan P, Chowell G. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th. Infectious Disease Modelling 2020; 5, 256-263.

[7] Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, Yan P, Chowell G. Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China. February 13–23. 2020; J. Clin. Med. 2020, 9, 596.

[8] Munayco V, Tariq A, Rothenberg R, Gabriela G, Cabezas S, Reyes MF, Valle A, Mezarina LR, Cabezas C, Loayza M, Chowell G. Peru COVID-19 working Group Early transmission dynamics of COVID-19 in a southern hemisphere setting. Lima-Peru. February 29 the March 30th. 2020; Infectious Disease Modelling. 2020; 5, 338-345.

[9] Rodriguez OT, Gutiérrez RAC, Javier ALH. Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models. 2020; Chaos Solitons and Fractals 138, 109946.

[10] Mazurek J, Nenickova Z. Predicting the number of total COVID-19 cases in the USA by a Gompertz curve. 2020; https.//www.researchgate.net/publication/340738553

[11] Catal M, Alonso S, Lacalle EA, L´opez D, Cardona PJ, Prats C. Empiric model for short-time prediction of COVID-19 spreading. 2020; medRxiv https.//doi.org/10.1101/2020.05.13.20101329

[12] Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PLOS ONE. 2020; https.//doi.org/10.1371/journal.pone.0231236 March 31.

[13] Gompertz, B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS &c. Philosophical transactions of the Royal Society of London. 1825;115:13–583.

[14] Zwietering M, Jongenburger I, Rombouts F, Riet VK. Modeling of the bacterial growth curve. Appl Environ Microbiol 1990;56(6):1875–1881.

[15] Gerlee P. The model muddle. in search of tumor growth laws. Cancer research 2013; 73(8):2407–2411.

[16] Kathleen MC, Tjørve E. The use of Gompertz models in growth analyses, and new Gompertz-model approach. An addition to the Unified-Richards family. PLoS ONE 2017;12(6), e0178691.

[17] Dennis B, Ponciano JM, Subhash R, Traper LM, Staples DF. Estimating Density Dependence, Process Noise and Observation Erros. Ecological Monographs 2006;76(3): 323–341.

[18] Jazwinski, AH. Stochastic Processes and Filtering Theory. Academic Press, 1970.

[19] Anderson, BDO, Moore JB. Optimal Filtering. Prentice Hall, 1979.

[20] Chui, C.K, Chen G. Kalman Filtering with Real-time Applications. Springer Verlag, 1991.

[21] Ljung, L, Söderström T. Theory and Practice of Recursive Identification. The MIT Press,1993.

[22] Chen, G. Approximate Kalman Filtering. World Scientific, 1993.

[23] Grewal S, Andrews AP. Kalman Filtering Theory and Practice. Prentice Hall, 1993.

[24] Öztürk F, Özbek L. Mathematical Modelling and Simulation, Pigeon Yay, 2016 (in Turkish).

[25] Özbek L. Kalman Filtresi, Akademisyen Yay. 2017 (in Turkish).

[26] Kalman RE. A new Approach to linear Filtering and Prediction Problems". Journal of Basic Engineering. 1960;82:35-45.

[27] 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):1505-1512.

[28] Özbek L, Aliev FA. Comments on Adaptive Fading Kalman Filter with an Application. Automatica 1998; 34(12): 1663-1664.

[29] Efe M, Özbek L. Fading Kalman Filter for Manoeuvring Target Tracking. Journal of the Turkish Statistical Assocation 1999; 2(3):193-206.

[30] Özbek L, Efe M. An Adaptive Extended Kalman Filter with Application to Compartment Models. Communications In Statistics-Simulation And Computation 2004; 33(1): 145-158.

[31] Johns Hopkins University Center for Systems Science and Engineering, 2020. ttps.//github.com/CSSEGISandData/COVID-19.