Journal of Statistical and Econometric Methods, Vol.10, No.3, 2021, 15-27
ISSN: 2241-0384 (print), 2241-0376 (online)
https://doi.org/10.47260/jsem/1032
Scientific Press International Limited

A Modeling Study on the Estimation of COVID-19 Daily and Weekly Cases and Reproduction Number Using the Adaptive Kalman Filter: The Example of Ziraat Bank, Turkey

Dr. İlker Met[1], Dr. Levent Özbek[2], Himmet Aksoy[3] and Ayfer Erkoç[4]

Abstract

Since the beginning of 2020, the world has been struggling with a viral epidemic (COVID-19), which poses a serious threat to the collective health of the human race. Mathematical modeling of epidemics is critical for developing such policies, especially during these uncertain times. In this study, the reproduction number and model parameters were predicted using AR(1) (autoregressive time-series model of order 1) and the adaptive Kalman filter (AKF). The data sample used in the study consists of the weekly and daily number of cases amongst the Ziraat Bank personnel between March 11, 2020, and April 19, 2021. This sample was modeled in the state space, and the AKF was used to estimate the number of cases per day. It is quite simple to model the daily and weekly case number time series with the time-varying parameter AR(1) stochastic process and to estimate the time-varying parameter with online AKF. Overall, we found that the weekly case number prediction was more accurate than the daily case number ($R^2 = 0.97$), especially in regions with a low number of cases. We suggest that the simplest method for reproduction number estimation can be obtained by modeling the daily cases using an AR(1) model.

JEL classification numbers: C02, C22, C32.
Keywords: COVID-19, Modeling, Reproduction number estimation, AR(1), Kalman filter.

[1] Technology Management, KHO, Head of Human Resources, Ziraat Bank, Ankara, Turkey.
[2] Department of Statistics, Ankara University Faculty of Science, Ankara, Turkey.
[3] Business Administration, Selçuk University, Human Resources, Ziraat Bank, Ankara, Turkey.
[4] Statistics, Ankara University, Career Management Officer, Ziraat Bank, Ankara, Turkey.

Article Info: Received: July 15, 2021. Revised: August 9, 2021.
Published online: August 12, 2021.

1. Introduction

The ability to make future predictions helps us to navigate uncertain times. Mathematical, deterministic, and statistical modeling methods are invaluable tools at the disposal of policymakers to do just that, which is especially salient during these trying times. Just as it was with previous viral outbreaks, scientists have conducted numerous studies modeling the rate of disease transmission and the dates when we will reach maximum capacity. These studies are not specific to the COVID-19 outbreak, and these epidemic models are scientifically important for policy development [1].

Trend and peak estimations are tough obstacles due to the changes and restrictions in the data disclosure policies of different countries. The deviation and margin of error are seen to be high in studies with peak estimation [2]. Owing to these constraints, predictive models generally rely on the reproduction numbers to determine model trends. As other researchers have noted, the uncertainty of available official data, especially regarding the actual number of infected cases, may lead to uncertain results and false estimates [3].

A total of 244 oft-cited articles since January 2020 showed that 46% of studies used compartmental models, 32% used statistical models, and 1% used individual-based models. A majority of the studies were conducted in Asian (78.93%) and European (59.09%) countries. A majority of them used compartmental models (SIR and SEIR) (46.1%) and statistical models (growth models and time series) (31.8%). The remaining studies employed artificial intelligence (6.7%), the Bayesian approach (4.7%), network models (2.3%), and agent-based models (1.3%) [4].

The biggest contribution of modeling studies is their ability to show the number of reproductions and the progress of the epidemic trend. Highlighting the true impact of the epidemic at an early stage is an important step towards effective planning on restrictions, hospital SOPs, vaccination policy, etc. As the COVID-19 pandemic continues, mathematical epidemiologists have continued to share their models on how the disease has spread, the current state of play, and what still needs to be done. Since the beginning of COVID-19, nonparametric methods have been used in modeling studies more frequently, especially logarithmic and exponential models [5].

Along with the growth models, non-linear methods were also used for parameter estimation. The spread of the epidemic in China was estimated using a logistic growth model, while non-linear least squares (NLS) were used to estimate model parameters. In this study, we found that the growth rate of Covid-19 is significantly different between China, South Korea, and Iran [6].

In Nigeria, estimations were made using logistic and exponential models, while model parameters were estimated using ordinary least squares [7].

The logistic growth model, generalized logistic growth model, generalized Richards model, and generalized growth model to the reported number of infected cases for the whole of China, 29 provinces in China, and 33 countries and regions that have been or are undergoing major outbreaks are used to estimate peak times [8].

Nine non-linear models (Brody, Bertalanffy, Logistic, Generalized Logistic, Richard, Negative Exponential, Stevens, Tanaka, Gompertz) for the US, Brazil, Germany, India, Russia, Italy, Spain, France, the United Kingdom, and Turkey were studied to model the estimation of reproduction number and daily number of cases by estimating parameters using a Kalman filter [9].

A support vector machine (SVM) with fuzzy granulation was used to predict the growth range of confirmed new cases, new deaths, and new cured cases in China. The experimental results showed that the Elman neural network and SVM can predict the development trend of cumulative confirmed cases, deaths, and cured cases, whereas LSTM (long short-term memory) is more suitable for predicting cumulative confirmed cases [10].

Using mathematical and statistical methods, we attempted to estimate the reproduction number and the time range between waves in South Korea. The findings of the model study support the effectiveness of control measures against COVID-19 in Korea [11]. The purpose of managing and controlling COVID-19 in Iran mortality trends was modeled using regression, spatial modeling, risk mapping, and change detection using the random forest machine learning technique [12]. The maximum likelihood (ML) value of reproductive number (R0) can be estimated by the Poisson distribution determined by daily infectiousness [13].

It is seen in the literature that the types of Kalman filters are also used in the prediction of the transmission of epidemic diseases. An adaptive unscented Kalman filter (AUKF) -based optimal controller has been designed to control unknown tuberculosis dynamics in individuals treated with active tuberculosis, at home or in hospital. In this way, even in the presence of a small group of infectious people, the long-term persistence of the disease is thought to be prevented [14]. Our study proposes a method different from the models of epidemic diseases so far.

2. Materials and Methods

This study consists of four steps:

1. Data collection
2. Trend Analyses with Moving Average Graphs Modeling
3. Parameter Estimation with time series (AR) models
4. Parameter estimation with adaptive Kalman filters (AKF).

2.1 Data Collection Phase

In this study, we used the data of Ziraat Bank as a sample because of their wide presence throughout Turkey. Another reason for using bank data is the policy change affecting data disclosure within the country. The data were collected through an application that was developed by the Bank's technology team, starting from the day the first case was observed on March 11, 2020, to April 19, 2021. Data collected in the timespan of these 325 days were analyzed, (timespan is >365 and 325 days modeled, because no cases days are excluded). The total number of cases used in this study represents 21% of bank employees.

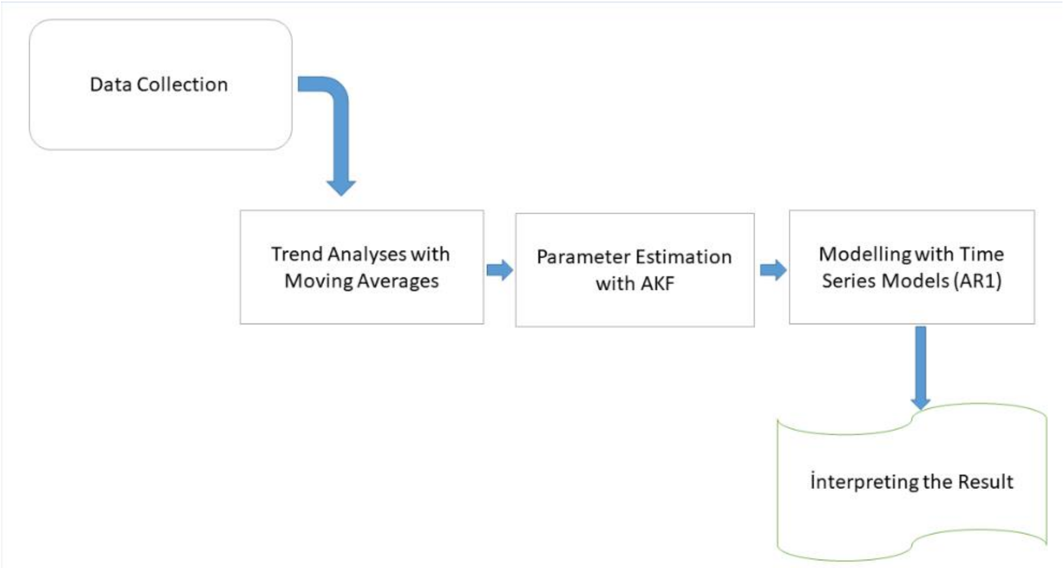
Data Collection
Trend Analyses with Moving Averages
Parameter Estimation with AKF
Modelling with Time Series Models (AR1)
İnterpreting the Result

Figure 1: The flowchart of methodologies modeling the spread of coronavirus in Ziraat Bank

2.2 Trend Analyses

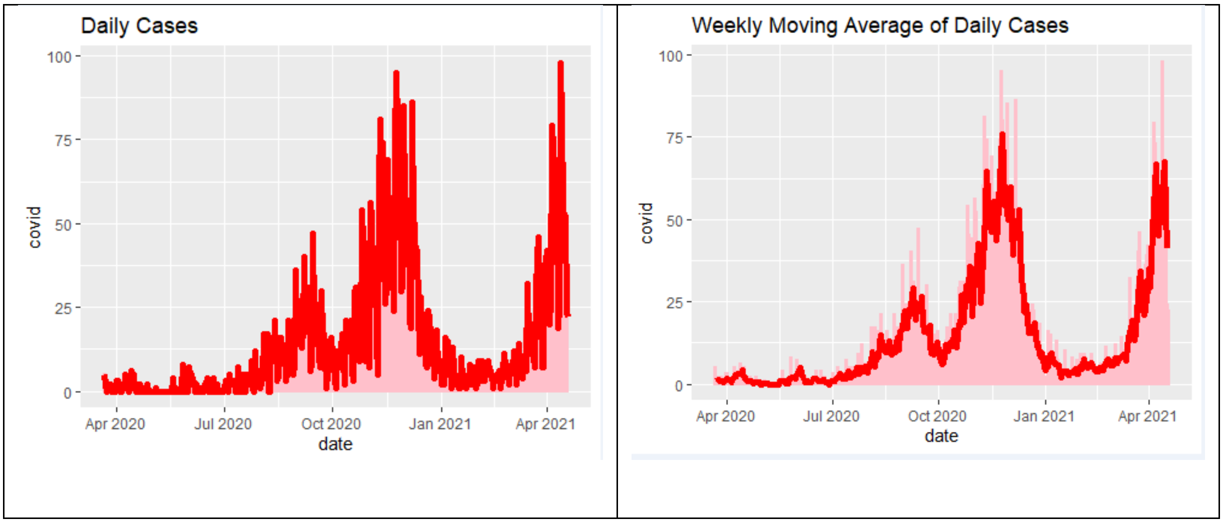

Figure 2: COVID-19 cases trend graphs between 11 March 2020-19 April 2020 in Ziraat Bank Case

As shown in Figure 2, we can see the 3rd peak occurring in the moving average (MA) charts. In this study, MA was used as a checkpoint. Daily cases are more smoothing than weekly cases.

2.3 Parameter Estimation with Time Series AR1 Models

In this study, time-series data (number of daily cases) was modeled. It is assumed that the number of daily cases i_t is in the form of AR(1) is given by Equation (1).

$$i_t = \theta \mathrm{i}_{t-1} + v_t \tag{1}$$

where θ is a constant and v_t is v_t ~ N(0,σ_1^2). The random variables $v_1, v_2, \dots, v_n$ are assumed to be uncorrelated. It is also assumed that the θ parameter in Equation (1) is time-varying and is a stochastic process in the form of a random walk process. In this case, the θ random walk process can be expressed as in Equation (2)

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

w_t is normally distributed with N(0,σ_1^2), and the random variables $w, w_2, \dots, w_n$ are assumed to be uncorrelated. Considering Equations (1) and (2), the following state-space model can be written:

$$\begin{aligned} \theta_t &= \theta_{t-1} + w_t \\ i_t &= \theta \mathrm{i}_{t-1} + v_t \end{aligned} \tag{3}$$

The state variable is unobservable, and the time-varying θ_t parameter can be estimated using the AKF.

Model codes are written in Matlab 2013a program.

Figure 3 shows the estimation of the number of cases per day and the estimation of the reproduction number with the Kalman filter. In Figure 4, the estimation of the weekly case numbers and the estimation of the reproduction number with the Kalman filter are shown. The graphs are created using Matlab. Because of daily cases variability, weekly cases are averaged and smoothed. Calculated MSE, MAPE and R^2 are shown in Table 1. If there is no variation in the daily cases, calculated values can be expected close to the weekly cases calculated values. As can be seen from Table 1 calculated values differ between daily and weekly cases. Estimations made by averaging weekly cases are suitable for using R(t).

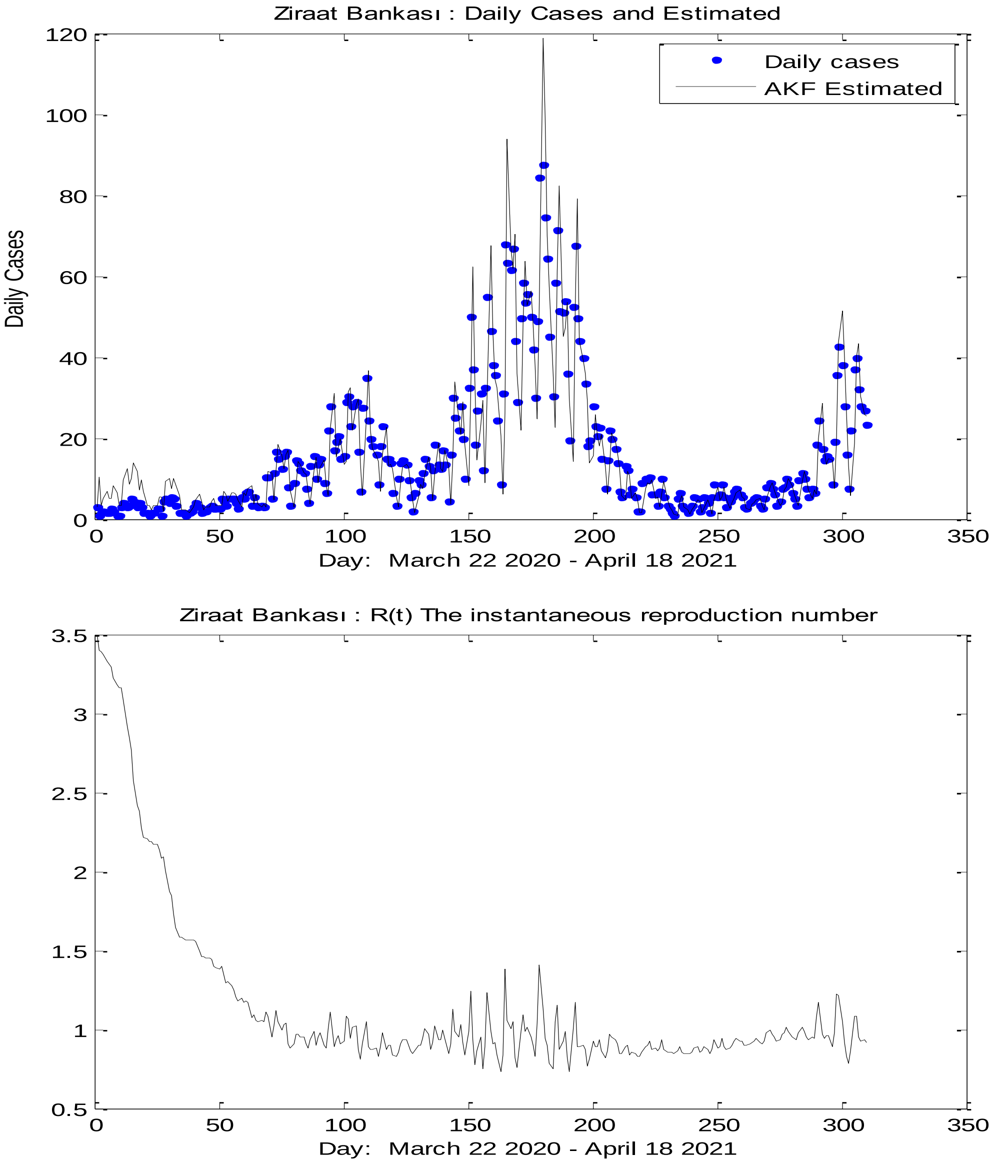

Figure 3: Daily estimated cases and estimated reproduction number

It is quite simple to model the daily and weekly case number time series with the time-varying parameter AR(1) stochastic process and estimate the time-varying parameter with online AKF. It can be seen weekly case number prediction was more precise than the daily case number ($R^2 = 0.97$).

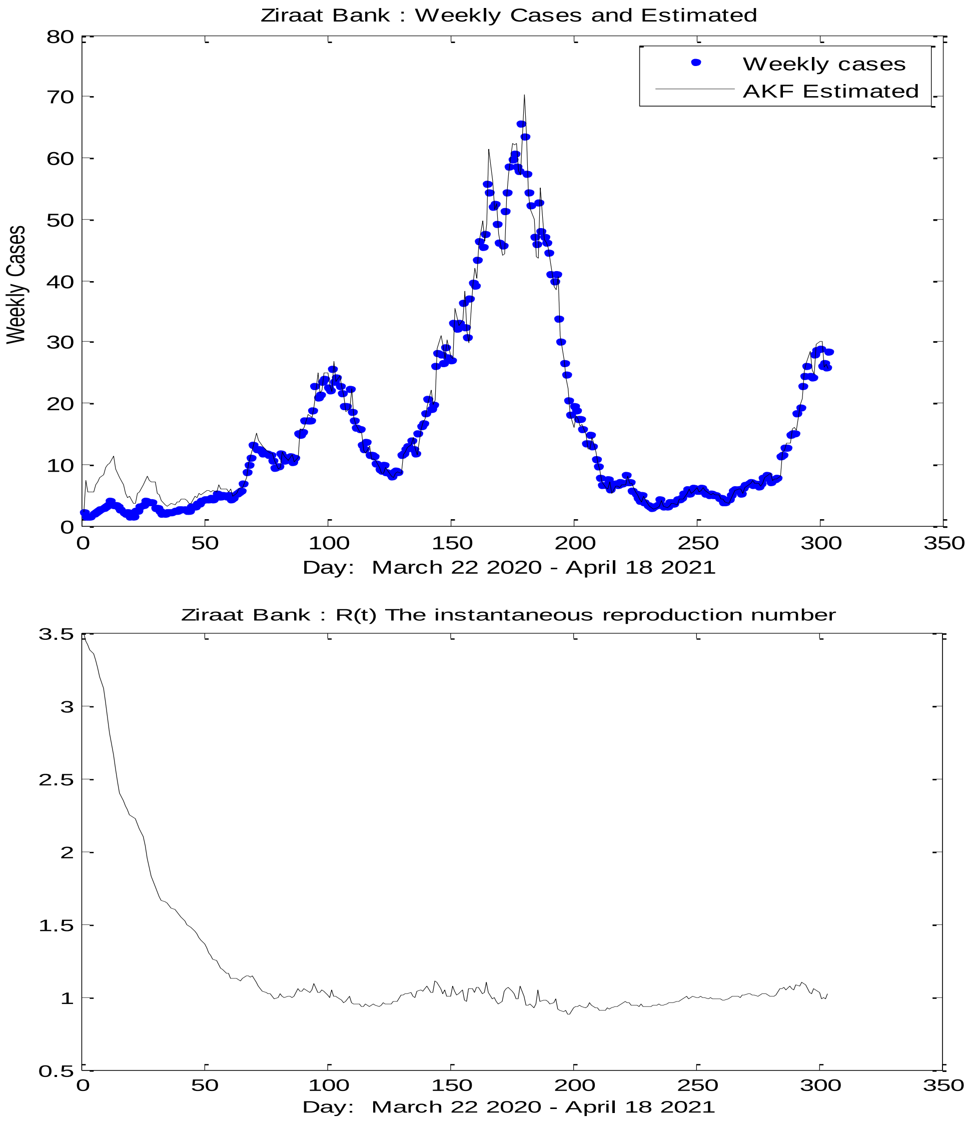

Figure 4: Weekly cases and estimated reproduction number

Table 1: Calculated MSE, MAPE, R2

Time	**MSE**	**MAPE**	**R^2**
Daily	84	154.19	0.69
Weekly	6	62.23	0.97

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

2.4 Estimating the Reproduction Number with Adaptive Kalman Filtering

The instantaneous reproduction number, Rt, at time t can be estimated using Equation (4). [15]

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

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 elapsed since the infection of the case. In practice, w is approximated by the distribution of serial intervals. Let us express the value of R_t calculated using the AR(1) model with R_t^{AR}. İf s=1 and w_1=1 are given in (4), then Equation (4), can be written as

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

The Kalman filter is a popular estimation method used to solve the state estimation problem in dynamic systems. As long as the system characteristics are known correctly, Kalman filter works with the best prediction performance. However, in cases where the system characteristics are partially known or uncertain, it is inevitable that there will be serious losses in the prediction performance of the filter. In order to overcome the performance loss problem in the Kalman filter, the adaptive Kalman filter method has been adapted. In adaptation forgetting factor proposed by Özbek and Aliev is used [15].

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 values of R_t^{AR}calculated using Equation 5 are shown in Figure 3 and Figure 4 [16].

3. Conclusion and Results

Theta in the AR(1) model called it the 'multiplication' factor. In time series models, this is called a parameter. In classical time series methods, this parameter is assumed to be constant and estimated accordingly. Here we assumed the parameter as time varying and random walk stochastic process. When this assumption is made, classical time series methods are not used. We took this assumption and AR(1) model together and turned it into a state-space model and estimated the parameter based on time using AKF. This method does not require all the data, only the arrival of the last observation is sufficient to obtain an estimate of the parameter and is an online estimation method. This is one of the advantages of the method. When the classical AR(1) model is used, the stationarity condition must be met in order to make the predictions. In the time-varying parameter assumption we use, this condition does not need to be met. This is the second advantage of the method we use. Since the Adaptive Kalman Filter is a self-adaptive estimation method and the variance of the noise processes in the model is not known exactly, it makes better predictions than the normal Kalman filter. This is stated in the references. In addition to generally known methods, it is quite simple to model the daily case number time series with the time-varying parameter AR(1) stochastic process and estimate the time-varying parameter with online AKF. It does not require any assumptions and it produces good results in cumulative weekly data when the model results are examined. The model is so simple that it can be reproduced and applied without detailed knowledge of epidemiology or programming languages. We must stress again that the purpose of this study is merely to show that it is possible to model epidemic trends using simple methods.

Using AR(1), the stochastic process for estimation is a common approach as it does not require any other modeling assumptions. Due to the simplicity of AR models, we highly recommend this approach for reproduction number estimation. As seen in our study, it is simple enough to model the daily and weekly case number time series with the time-varying parameter AR(1) stochastic process and estimate the time-varying parameter with online AKF.

We also noted that the weekly case number prediction was more precise than the daily case number prediction (R^2 = 0.97).

References

[1] Gog, J. R. (2020). How you can help with COVID-19 modelling. Nat Rev Phys 2(6):274–275 https://doi.org/10.1038/s42254-020-0175-7.

[2] Tiwari, A. (2020). Modelling and analysis of COVID-19 epidemic in India. J Saf Sci Resil. 1(2):135–140. ISSN 2666-4496. https://doi.org/10.1016/j.jnlssr.2020.11.005.

[3] Anastassopoulou, C., Russo, L., Tsakris, A., Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLOS ONE 15(3):e0230405. https://doi.org/10.1371/journal.pone.0230405.

[4] Gnanvi, J.E., Salako, K. V., Kotanmi, G. B., Glèlè Kakaï, R. G. (2021). On the reliability of predictions on COVID-19 dynamics: A systematic and critical review of modelling techniques. Infect Dis Modell. 6:258–272. ISSN 2468-0427. https://doi.org/10.1016/j.idm.2020.12.008.

[5] Sornette, D., Mearns, E. and Schatz, M. (2020). Interpreting, analysing and modelling COVID-19 mortality data. Nonlinear Dyn. 101:1751–1776. https://doi.org/10.1007/s11071-020-05966.

[6] Shen, C. Y. (2020). Logistic growth modelling of COVID-19 proliferation in China and its international implications. Int J Infect Dis. 96:582–589. ISSN 1201-9712. https://doi.org/10.1016/j.ijid.2020.04.085.

[7] Ayinde, K., Lukman, A. F., Rauf, R. I., Alabi, O. O., Okon, C. E. and Ayinde, O. E. (2020). Modeling Nigerian COVID-19 cases: A comparative analysis of models and estimators. Chaos Solitons Fractals. 138: article 2020. DOI: 10.1016/j.chaos.2020.109911.

[8] Wu. K., Darcet, D., Wang, Q. and Sornette, D. (2020). Generalized logistic growth modeling of the COVID-19 outbreak: Comparing the dynamics in the 29 provinces in China and in the rest of the world. Nonlinear Dyn. 101:1–21. https://doi.org/10.1007/s11071-020-05862-6.

[9] Özbek, L., and Demirtaş, H. (2021). Turkiye Klinikleri Journal of Biostatistics. 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. 13(1):3. doi: 10.5336/biostatic.2020-80186.

[10] Hao, Y., Xu, T., Hu, H., Wang, P. and Bai, Y. (2020). Prediction and analysis of Corona Virus Disease 2019. PLOS ONE. 15(10):e0239960. doi:10.1371/journal.pone.0239960.

[11] Shim, E., Tariq, A. and Chowell, G. (2020). Spatial variability in reproduction number and doubling time across two waves of the COVID-19 pandemic in South Korea, February to July 2020. Int J Infect Dis (2021) 102:1–9, ISSN 1201-9712. https://doi.org/10.1016/j.ijid.2020.10.007.

[12] Pourghasemi, H. R., Pouyan, S., Heidari, B., Farajzadeh, Z., Fallah Shamsi, S. R., Babaei, S., Khosravi, R., Etemadi, M., Ghanbarian, G., Farhadi, A., Safaeian, R., Heidari, Z., Tarazkar, M. H., Tiefenbacher, J. P., Azmi, A., and Sadeghian, F. (2020). Spatial modeling, risk mapping, change detection, and

outbreak trend analysis of coronavirus (COVID-19) in Iran (days between February 19 and June 14, 2020). Int J Infect Dis (2020) 98:90–108, ISSN 1201-9712. https://doi.org/10.1016/j.ijid.2020.06.058.

[13] Zhang, S., Diao, M., Yu, W., Pei, L., Lin, Z. and Chen, D. (2020). Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. Int J Infect Dis. 93:201–204, ISSN 1201-9712. https://doi.org/10.1016/j.ijid.2020.02.033.

[14] Cetin, M. and Beyhan, S. (2020). Adaptive Kalman Filtering Based Optimal Control of Tuberculosis Dynamics with Exogenous Reinfections. Journal of Engineering Sciences and Design, 8 (4), 1260-1268. DOI: 10.21923/jesd.717130

[15] Özbek, L. and Aliev, F. A. (1998). Comments on adaptive Fading Kalman Filter with an application.Available at:https://www.researchgate.net/publication/263595161_Adaptive_Fading_Kalman_Filter_with_an_Application. Automatica (1998) 34(12):1663–1664. DOI: 10.1016/S0005-1098(98)80025-3.

[16] Cori, A., Ferguson, N. M., Fraser C. and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol, 178(9):1505–1512. DOI: 10.1093/aje/kwt133, PMID: 24043437, PMCID: PMC3816335.

[17] Jazwinski, A. H. (1970). Stochastic Processes and Filtering Theory. Academic Press.

[18] Anderson, B. D. O. and Moore, J. B. (1979). Optimal Filtering. Prentice Hall.

[19] Chui, C. K. and Chen, G. (1991). Kalman Filtering with Real-time Applications. Springer Verlag.

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

[21] Chen, G. (1993). Approximate Kalman Filtering. World Scientific.

[22] Grewal, S. and Andrews, A. P. (1993). Kalman Filtering Theory and Practice. Prentice Hall.

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

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

[25] Kalman. R. E. (1960). A new Approach to linear Filtering and Prediction Problems". Journal of Basic Engineering. 82:35-45.

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

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

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

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

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 [17-25].

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×1 system vector, y_t is an m×1 observation vector, F_t is an n×n system matrix, H_t is an m×n matrix, w_t an n×1 vector of zero mean white noise sequence and v_t is an m×1 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 $Y_t = (y_0, y_1, \ldots, y_t)$ [17-25]. When $Y_t = (y_0, y_1, \ldots, y_t)$ observations are given, the prediction of state x_t with

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

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

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

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,

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

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

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

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

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

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 estimation respectively [17],[18]. 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 [26], [27], [28]. One of these is the use of the forgetting factor. A forgetting factor is proposed by Ozbek and Aliev [26].

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