Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2022.00036

Research Article

A study on the estimation of COVID-19 case with the adaptive kalman filter (AKF) in USA, Germany, United Kingdom, Italy, France, Russia, Brazil, India, Türkiye, Spain, Peru, Colombia, South Africa, Argentina, Iran, Pakistan

Levent ÖZBEK[1]

[1] *Ankara University, Faculty of Science, Department of Statistics, System Modeling and Simulation Laboratory, Ankara, Türkiye*

ARTICLE INFO

Article history
Received: 17 October 2020
Accepted: 28 June 2021

Key words:
COVID-19; Dynamic Linear Model; State-Space Modeling; Adaptive Kalman Filter; Reproduction Number Estimation

ABSTRACT

In this paper, COVID-19 cumulative cases are estimated with AKF based on total COVID-19 cases between January-September 9, 2020 in USA, Germany, United Kingdom, Italy, France, Russia, Brazil, India, Turkey, Spain, Peru, Colombia, South Africa, Argentina, Iran, Pakistan. The cumulative covid-19 cases time-series was modeled with a stochastic dynamic linear model (DLM). The estimation performance of the models is measured by the calculation of mean square error (MSE) and coefficient of determination (R^2). Ca lculated MSE an d R^2 values showed that the model and AKF could be used to estimate the number of cases in these countries. In this study, firstly, the cumulative number of cases was estimated. Secondly, using these estimates number of daily cases was calculated. Thirdly, the reproduction number was obtained by using these number of daily cases. The model and estimation method used is suitable. The AKF algorithm uses only the number of cases in the last day. We propose that the model and estimation method under consideration is a convenient tool for calculating the reproduction number depending on time.

Cite this article as: Özbek L. A study on the estimation of COVID-19 case with the adaptive kalman filter (AKF) in USA, Germany, United Kingdom, Italy, France, Russia, Brazil, India, Turkey, Spain, Peru, Colombia, South Africa, Argentina, Iran, Pakistan. Sigma J Eng Nat Sci 2022;40(2):323–343.

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

***Corresponding author.**
*E-mail address: ozbek@science.ankara.edu.tr

This paper was recommended for publication in revised form by Regional Editor Alper Yılmaz

Published by Yıldız Technical University Press, İstanbul, Türkiye

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.

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 de-fined 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 es-timation 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.

State-space models have been employed since the 1960's, mostly in the control and signal processing areas. 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. The KF has also been utilized in electrophysiological signal analysis and it compares favorably with other approaches [13].

In this work, COVID-19 cumulative cases are estimated with AKF based on total COVID-19 cases between January-September 9, 2020 in USA, Germany, United Kingdom, Italy, France, Russia, Brazil, India, Turkey, Spain, Peru, Colombia, South Africa, Argentina, Iran, Pakistan. The cumulative covid-19 cases time-series was modeled with a stochastic dynamic linear model (DLM). The estimation performance of the models is measured by the calculation of mean square error (MSE) and coefficient of determination (R^2).

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 three 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

Materials

In this paper, AKF has been used to estimate the actual COVID-19 cases. If we introduce state-space models and AKF representation at this point, it will be easier to see the suitability of the AKF approach to this specific problem. Let's consider a discrete-time state-space model stated as

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

$$y_t = H_t x_t + v_t \tag{2}$$

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

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	$y(t;\beta,k,m) = \alpha/((1 + \beta\exp(-kmt)^{1/m}) + \varepsilon$
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	$y(t;\alpha, \beta,k,m) = (1/\sqrt{\beta})\ln\vert 2\beta.(t\text{-}m)+2\sqrt{k^2(t-m)^2+\alpha\beta}\,\vert+\varepsilon$
Gompertz	$Y(t) = \alpha\ \exp(-\beta\exp(-kt)) + \varepsilon$

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 $Y_t = (y_0, y_1, ..., y_t)$ [14–16]. 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 AKF equations are

$$\hat{x}_{t|t-1} = F_{t-1}\hat{x}_{t-1} \tag{3}$$

$$P_{t|t-1} = \alpha(F_{t-1}P_{t-1|t-1}F'_{t-1} + Q_{t-1}) \tag{4}$$

$$K_t = P_{t|t-1}H'_t(H_tP_{t|t-1}H'_t + R_t)^{-1} \tag{5}$$

$$P_{t|t} = (I - K_tH_t)P_{t|t-1} \tag{6}$$

$$\hat{x}_t = \hat{x}_{t|t-1} + K_t(y_t - H_t\hat{x}_{t|t-1}) \tag{7}$$

where $\hat{X}_{t|t-1}$ is the a priori and $\hat{X}_t$ is the a posteriori estimation of x_t. $P_{t|t-1}$ and $P_{t|t}$ are the covariance of a priori and a posteriori estimation respectively [14–16]. α is the forgetting factor proposed by Özbek and Aliev [17–18].

MODEL AND ESTIMATION RESULTS

In [19] a simple linear model was proposed to describe a stochastic time-series. This s o-called " dynamic l inear model" (DLM) is defined in terms of state-space representation through

$$y_t = \mu_t + \varepsilon_t \tag{8}$$

$$\mu_{t+1} = \mu_t + \beta_t + \xi_t \tag{9}$$

$$\beta_{t+1} = \beta_t + \eta_t \tag{10}$$

where, y_t is the logarithmic actual COVID-19 cumulative cases. In Eq. 9, μ_t represents the trend. We specified the trend component as a random walk with drift. Eq. 10 describes the evolution of the drift which depends on its previous value. It is assumed that ε_t, ξ_t, and η_t are i.i.d. with zero means and constant variances Gaussian white noise. It would be useful to put these equations in vector-matrix form to obtain the state-space model for COVID-19 cases. The simple model introduced above can easily be represented in a state space form, where the state equation and the observation equation are displayed as:

$$\begin{bmatrix}\mu_{t+1}\\ \beta_{t+1}\end{bmatrix} = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\begin{pmatrix}\mu_t\\ \beta_t\end{pmatrix} + \begin{pmatrix}\xi_t\\ \eta_l\end{pmatrix} \tag{11}$$

and

$$y_t = (1 \quad 0)\begin{pmatrix}\mu_t\\ \beta_t\end{pmatrix} + \varepsilon_t \tag{12}$$

Table 2. Calculated MSE and R^2

Country	Cumulative MSE	Cumulative R^2	Daily MSE	Daily R^2
USA	3471865	0.99999	860410	0.99726
Germany	73248	0.99998	19033	0.99067
UK	42561	0.99999	8748	0.9960
Italy	52014	0.99999	8308	0.99667
France	475426	0.99993	748904	0.84456
Russia	117834	0.99999	8982	0.99885
Brazil	4538890	0.99999	4428191	0.98385
India	696734	0.99999	498780	0.99929
Turkey	47872	0.99999	14435	0.98513
Spain	370666	0.99997	598625	0.92229
Peru	222356	0.99999	431187	0.92447
Colombia	94034	0.99999	87392	0.99447
South Africa	146233	0.99999	52605	0.99669
Argentina	28211	0.99999	21020	0.99786
Iran	19691	0.99999	4292	0.98935
Pakistan	73443	0.99999	91626	0.96639

where Eq. 11 is the state equation that defines the evolution of system states and Eq. 12 is the observation equation that relates system states to the observations. In above equations, process noises and observation noise sequences are assumed to be Gaussian and independent of each other.

Given the observation values, AKF estimates unobservable state variables. Since the COVID-19 time-series is written as the state-space model above, the AKF equations can be employed. Initial value of the AKF was used as $x_0 = (51)'$. The selection of the initial values is not critical as the properly constructed model will yield these initial values to converge to the measurements. Forgetting factor α was 1.5. The data used was taken from Johns Hopkins University [20]. Actual cumulative case estimations have been obtained online using AKF. The number of daily cases can be easily calculated with $i_t = y_t - y_{t-1}$ to show the total number of cases up to y_t, t days. Since we have the estimates of y_t, we can easily find the estimations of i_t with $\hat{i}_t = \hat{y}_t - \hat{y}_{t-1}$. Daily cases and estimations are given in Figure 1-Figure 16. According to the estimation results obtained by using the cumulative and daily number of cases in the DLM, MSE and R^2, were calculated (see Table 2). These calculated values indicate that the compatibility of the model with actual data is quite high. This situation tells us that estimating the daily number of cases via the DLM is a reliable method. As for AKF, utilizing only the observation in time t and the preceding estimation is the most advantageous aspect of this method. These results have revealed that with the given system model and the assumptions, AKF could successfully be used to estimate actual COVID-19 cases. The method estimates online.

REPRODUCTION NUMBER ESTIMATION WITH AKF

The instantaneous reproduction number, R_t at time t can be estimated as following equation

$$R_t = \frac{E(i_t)}{\sum_{s=1}^{t} i_{t-s} w_s} \tag{13}$$

where $E(X)$ denotes the expectation of a random variable [21]. In Eq. 13, i_t stands for the number of new infections generated at time step t. In practice, w_s is approximated by the distribution of the serial interval. In this article, we have taken the distribution of w_s as a uniform distribution in $f(w_s) = 1/7$, s = 1,2,...,7 form. Since $E(i_t) = \hat{i}_t$, Eq. 13 can be written as following equation

$$R_t = \frac{\widehat{i}_t}{\frac{1}{7}\sum_{s=1}^{t} i_{t-s}},\ t = 88,0,\ldots,n-1 \tag{14}$$

The value of R_t calculated using Eq. 14 is given in Figure 1-Figure 16. There is no need for any other model assumption in estimating R_t with this method by using the DLM. By modeling the cumulative case time-series COVID-19 with DLM stochastic process and estimating them with AKF the number of daily cases and the instantaneous reproduction number is calculated without any other operation. It is quite a convenient method to model the cumulative case number time series with the DLM stochastic process and estimate them with online AKF.

RESULTS AND DISCUSSION

In this study, cumulative and daily cases of COVID-19 have been estimated online using DLM and AKF based on the total COVID-19 cases between January and September 9 2020, in USA, Germany, United Kingdom, Italy, France, Russia, Brazil, India, Turkey, Spain, Peru, Colombia, South Africa, Argentina, Iran, Pakistan. The cumulative case number was modeled with DLM, and the time-series were estimated by online AKF. Estimation by acquired data observed between January and September 9, 2020, shows that employing the discrete-time DLM and AKF in terms of MSE and R^2 provides efficient analysis for modeling the total case. It is proposed that the use of discrete-time DLM and AKF is appropriate. After estimating the number of cumulative cases, the computation of daily cases was made. After calculation of the estimation of number of daily cases, reproduction number was obtained. The DLM is an appropriate estimation method for the cumulative and daily cases. As for AKF, utilizing only the observation in time t and preceding estimation is the most advantageous aspect of this method. Modeling the cumulative case time-series with the DLM and estimating them 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 time series with the DLM stochastic process and estimate them with online AKF. Among the studies made on the COVID-19 pandemic, the progress of modeling the disease is remarked primarily. The progress of modeling the disease is substantial for the precautions to be taken and, interventions and treatments to be administered by the countries. As a result of estimations computed by data observed between January and September 9, 2020, it is proposed that the efficient analysis for modeling the total case is to be made using the DLM and AKF in terms of MSE and R^2. It is thought that the method we have proposed is suitable for the estimation of the forthcoming progress. Our suggestion is that the most convenient method for the estimation of the reproduction number can be performed by modeling the cumulative case number time series using DLM.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

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;5:536–544. [CrossRef]

[2] Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-ınfected pneumonia. N Engl J Med 2020;382:1199–1207. [CrossRef]

[3] World Health Organization. Novel coronavirus (2019-nCoV) situation reports. 2020. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/ Accessed on May. 18, 2022.

[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. Int J Modern Physics 2020;42:2050103. [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;5:256–263. [CrossRef]

[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;9:596. [CrossRef]

[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;5:338–345. [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.

[10] Mazurek J, Neničková Z. Predicting the number of total COVID-19 cases in the USA by a Gompertz curve. Mol Biol 2020. Preprint. doi: 10.13140/RG.2.2.19841.81761[CrossRef]

[11] Cata M, Alonso S, Alvarez-Lacalle E, Lopez D, Cardona P-J, Prats C, et al. Empiric model for short-time prediction of COVID-19 spreading. medRxiv 2020. Preprint. doi:10.1101/2020.05.13.20101329. [CrossRef]

[12] Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PLoS One 2020;15:e0231236.

[13] Özbek L. Kalman Filtresi. Ankara: Akademisyen Yayınevi; 2017. [Turkish] [CrossRef]

[14] Kalman RE. A new approach to linear filtering and prediction problems. J Basic Eng 1960;82:35–45. [CrossRef]

[15] Jazwinski AH. Stochastic Processes and Filtering Theory. Cambridge, Massachusetts: Academic Press; 1970.

[16] Anderson BDO, Moore JB. Optimal Filtering. New Jersey: Prentice Hall; 1979. [CrossRef]

[17] Özbek L, Aliev FA. Comments on adaptive fading kalman filter with an application. Automatica 1998;34:1663–1664. [CrossRef]

[18] Ozbek L, Efe M. An adaptive extended kalman filter with application to compartment models. Commun Stat Simul Comput 2004;33:145–158. [CrossRef]

[19] Harrison PJ, Stevens CF. Bayesian forecasting (with discussion). J Roy Stat Soc Ser B 1976;38:205–247. [CrossRef]

[20] Johns Hopkins University Center for Systems Science and Engineering, 2019. Available at: https://github.com/CSSEGISandData/COVID-19 Accessed on May 17, 2022.

[21] 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;178:1505–1512. [CrossRef]

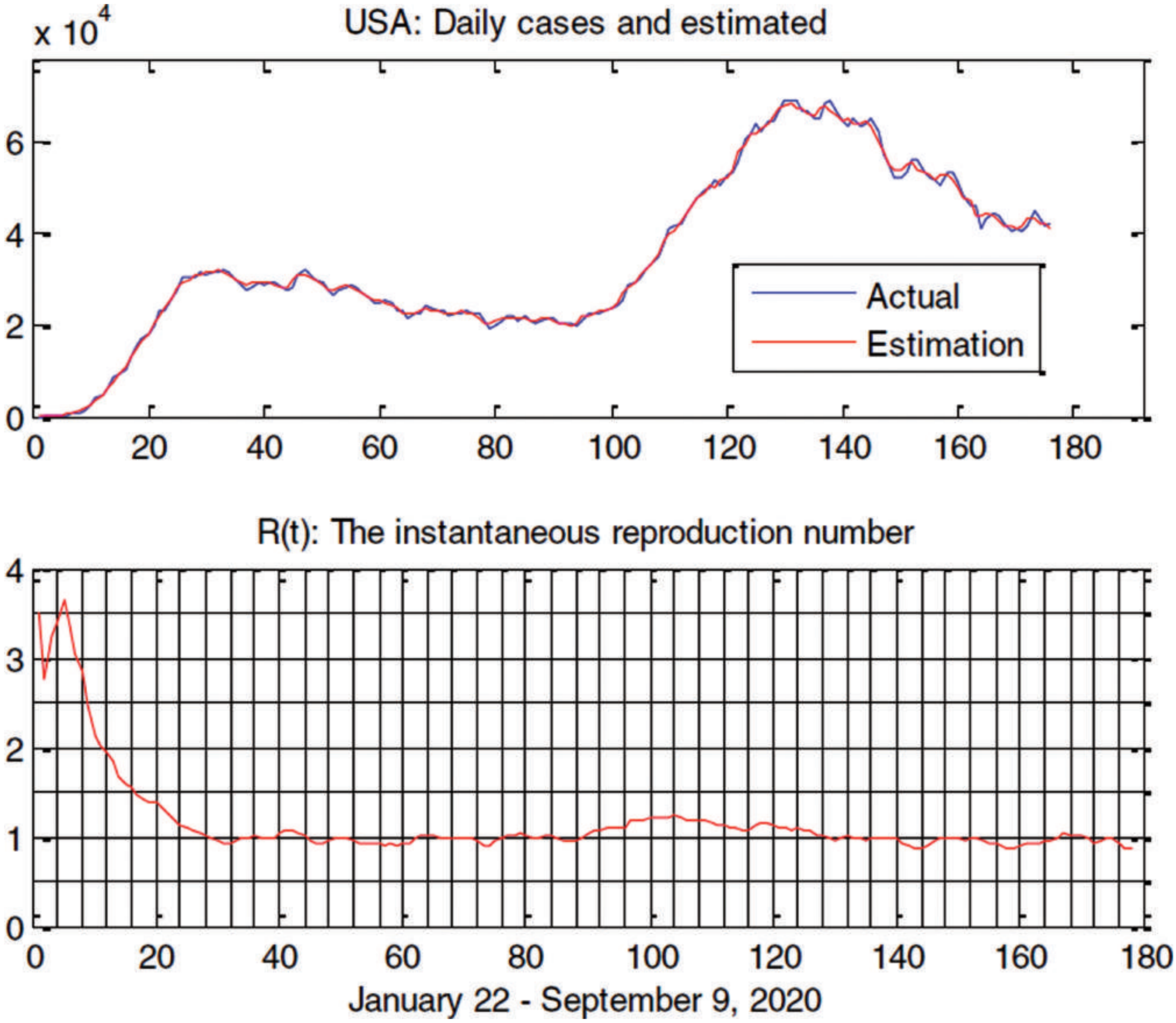

Figure 1. USA: Daily cases and estimated, reproduction number estimation.

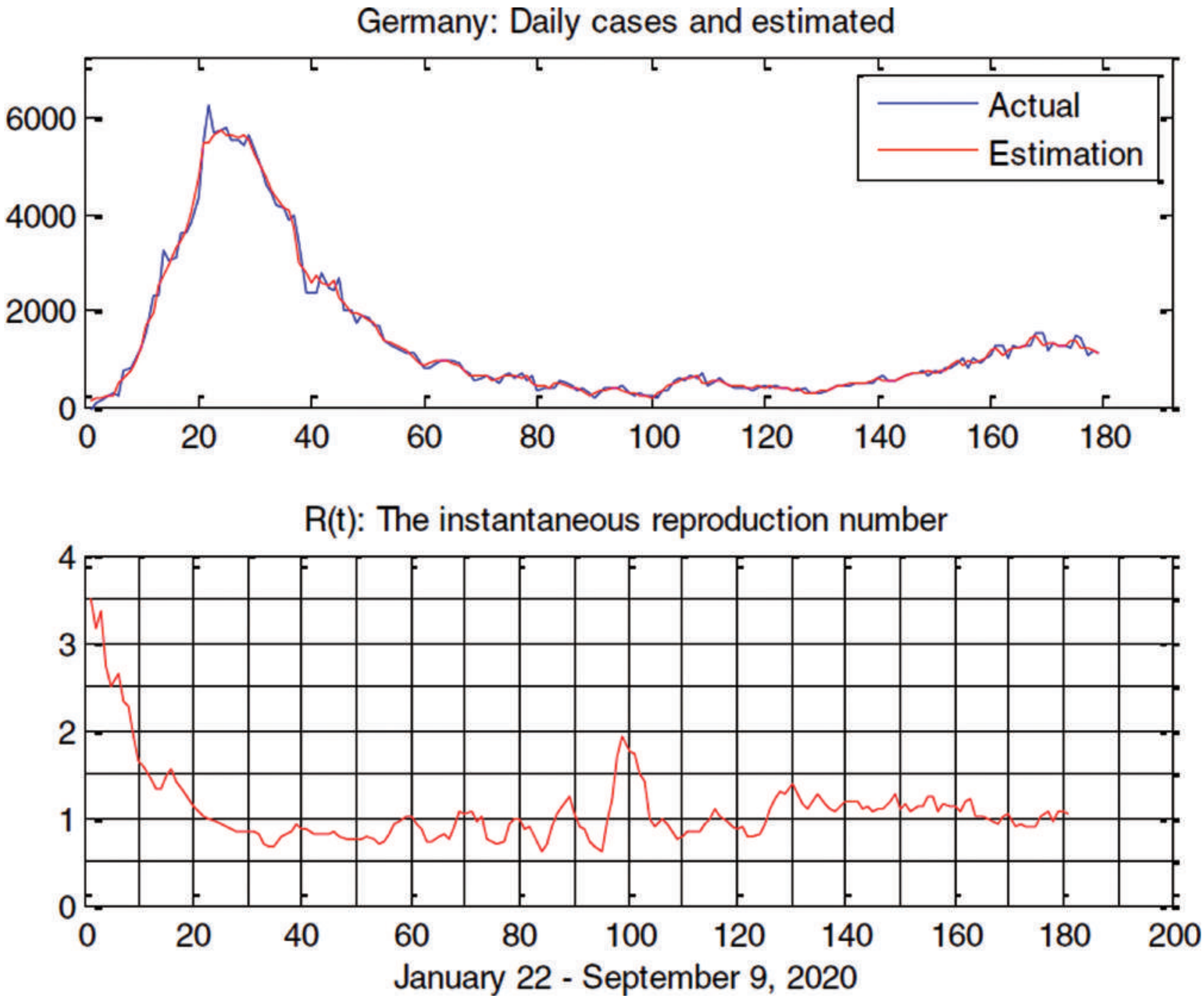

Figure 2. Germany: Daily cases and estimated, reproduction number estimation.

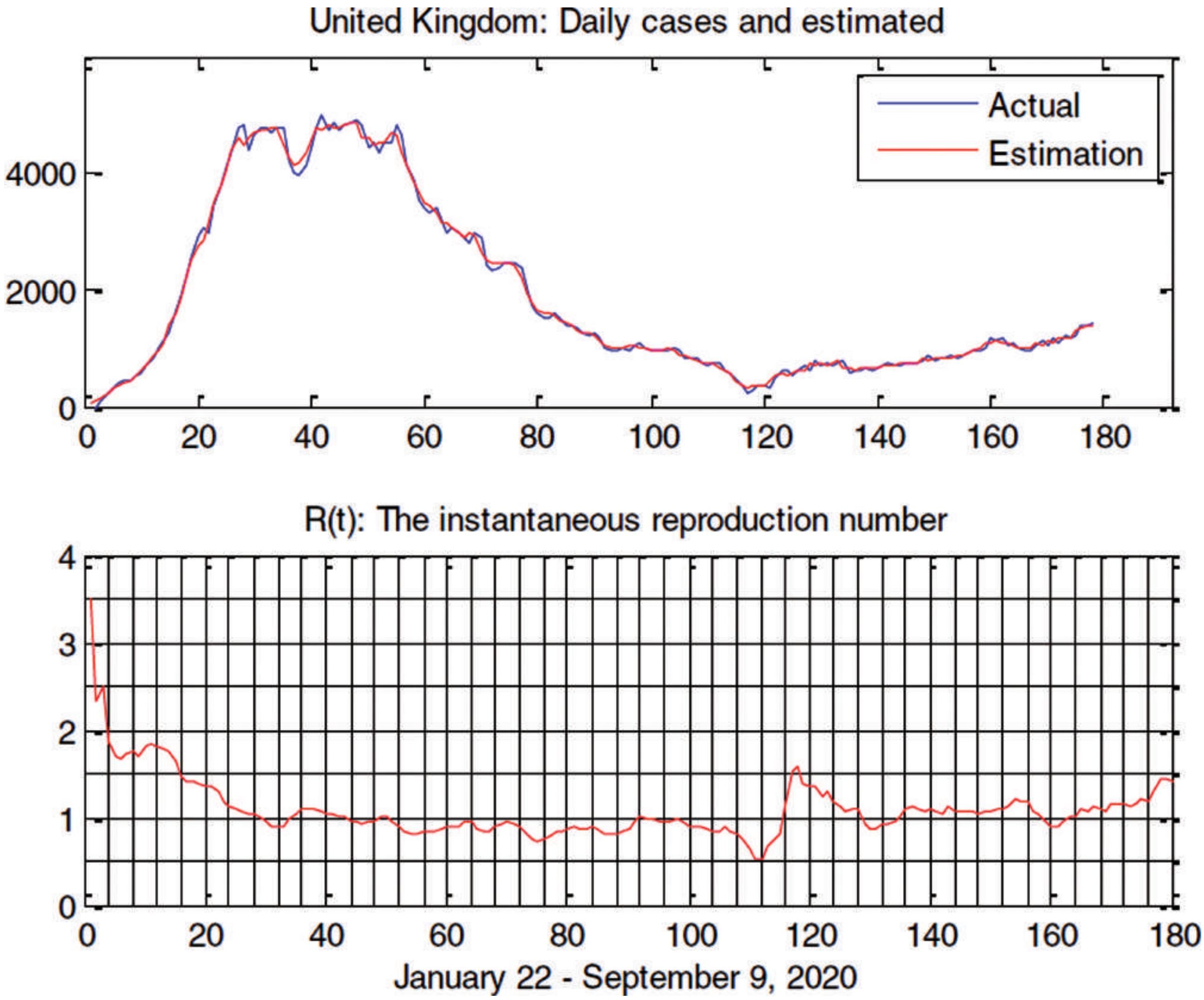

Figure 3. UK: Daily cases and estimated, reproduction number estimation.

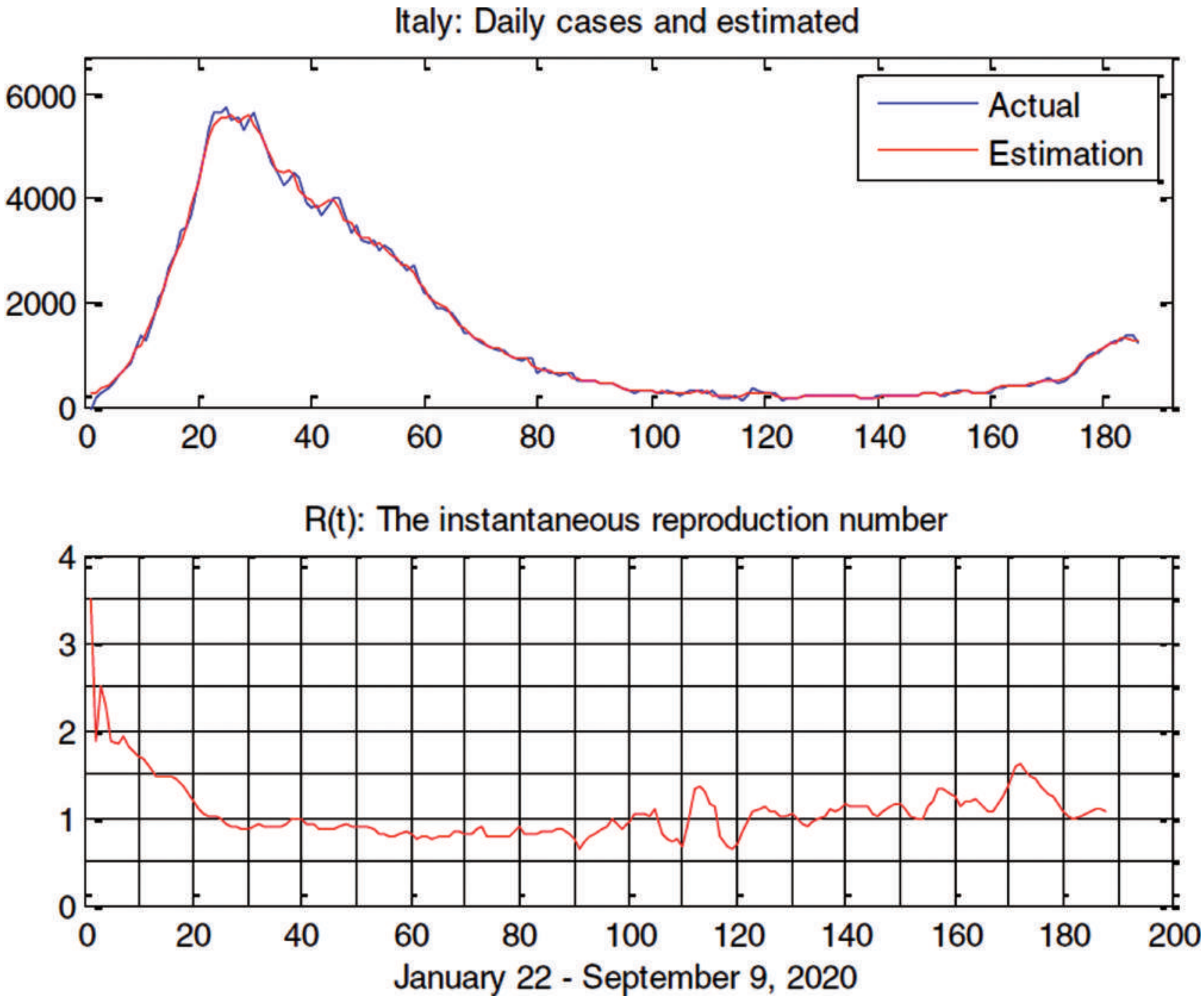

Figure 4. Italy: Daily cases and estimated, reproduction number estimation.

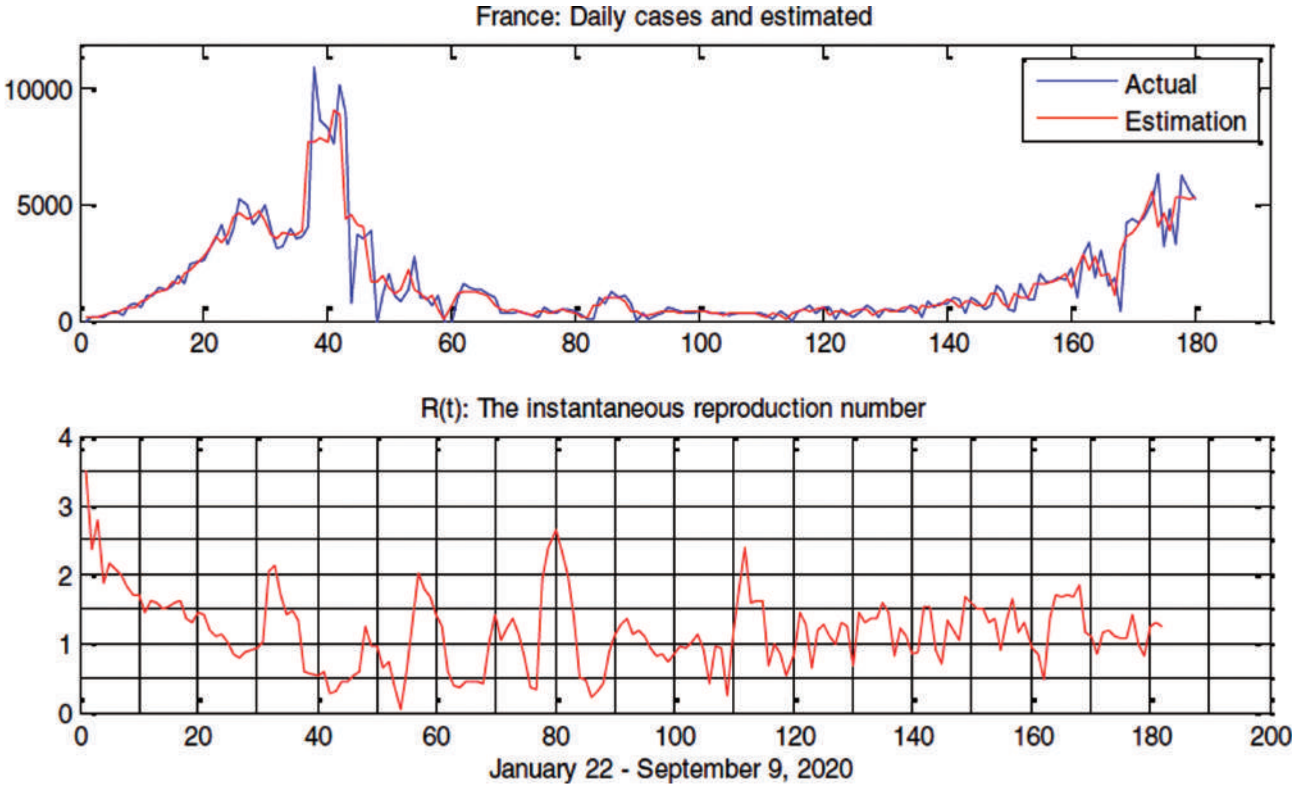

Figure 5. France: Daily cases and estimated, reproduction number estimation.

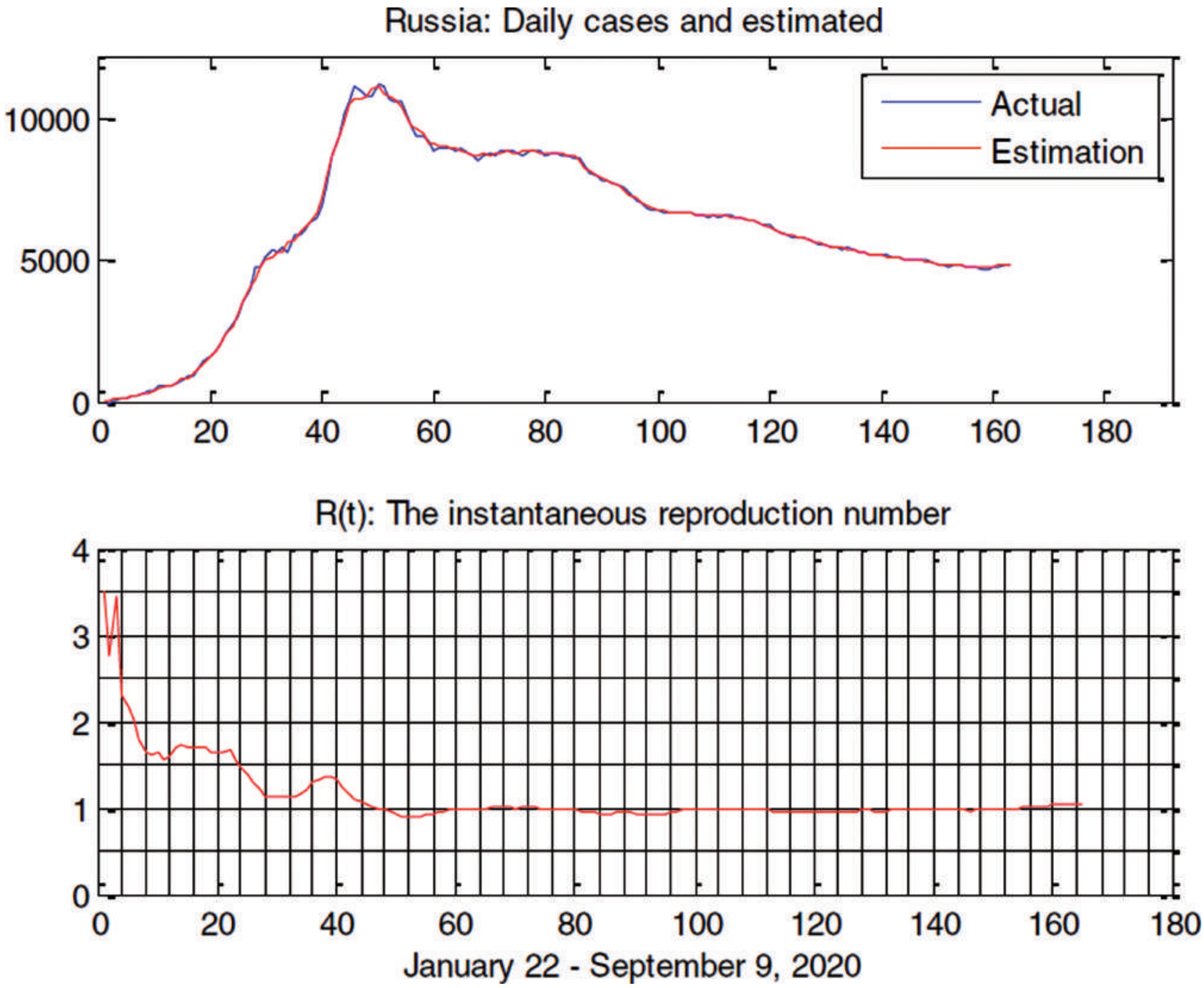

Figure 6. Russia: Daily cases and estimated, reproduction number estimation.

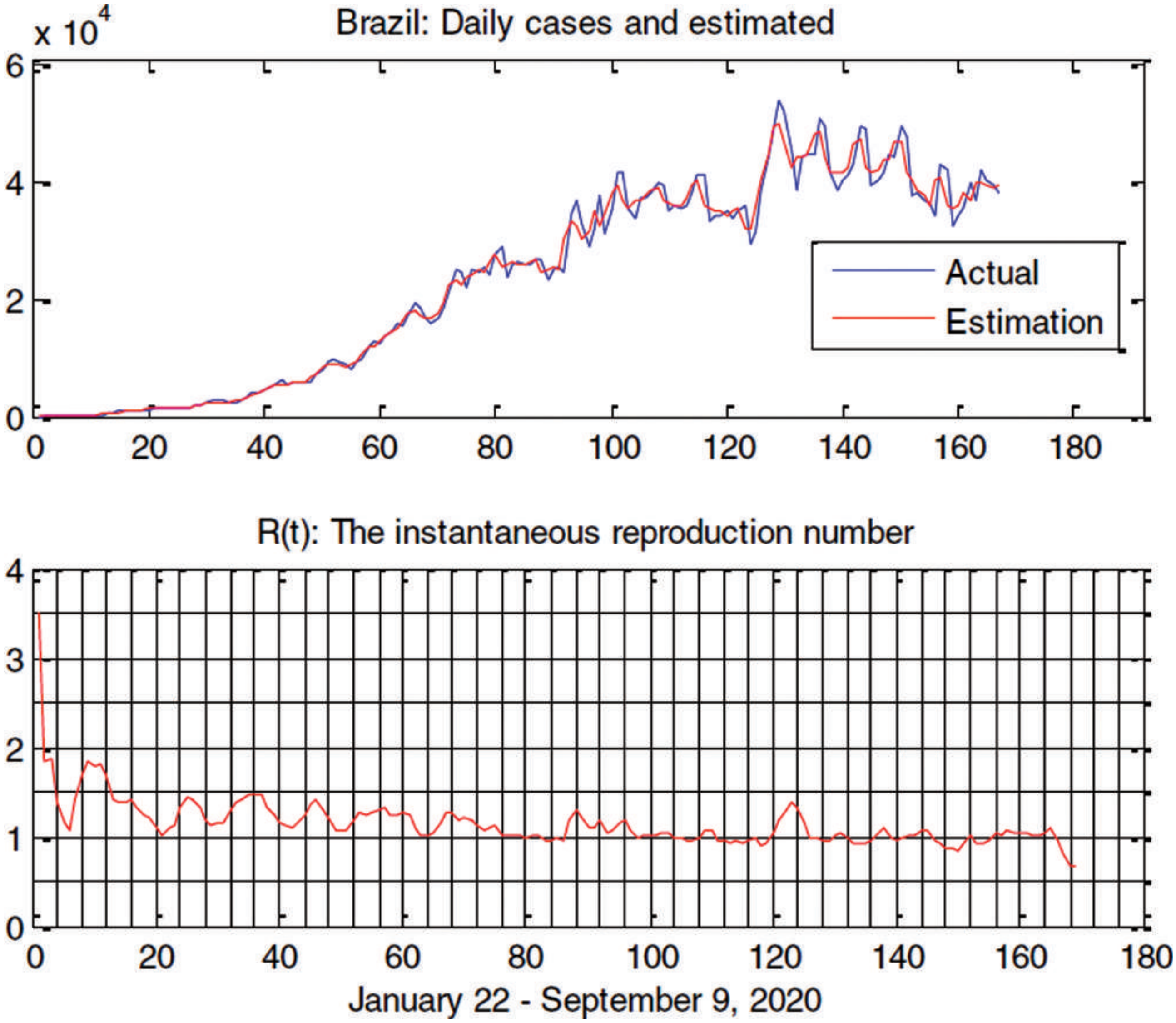

Figure 7. Brazil: Daily cases and estimated, reproduction number estimation.

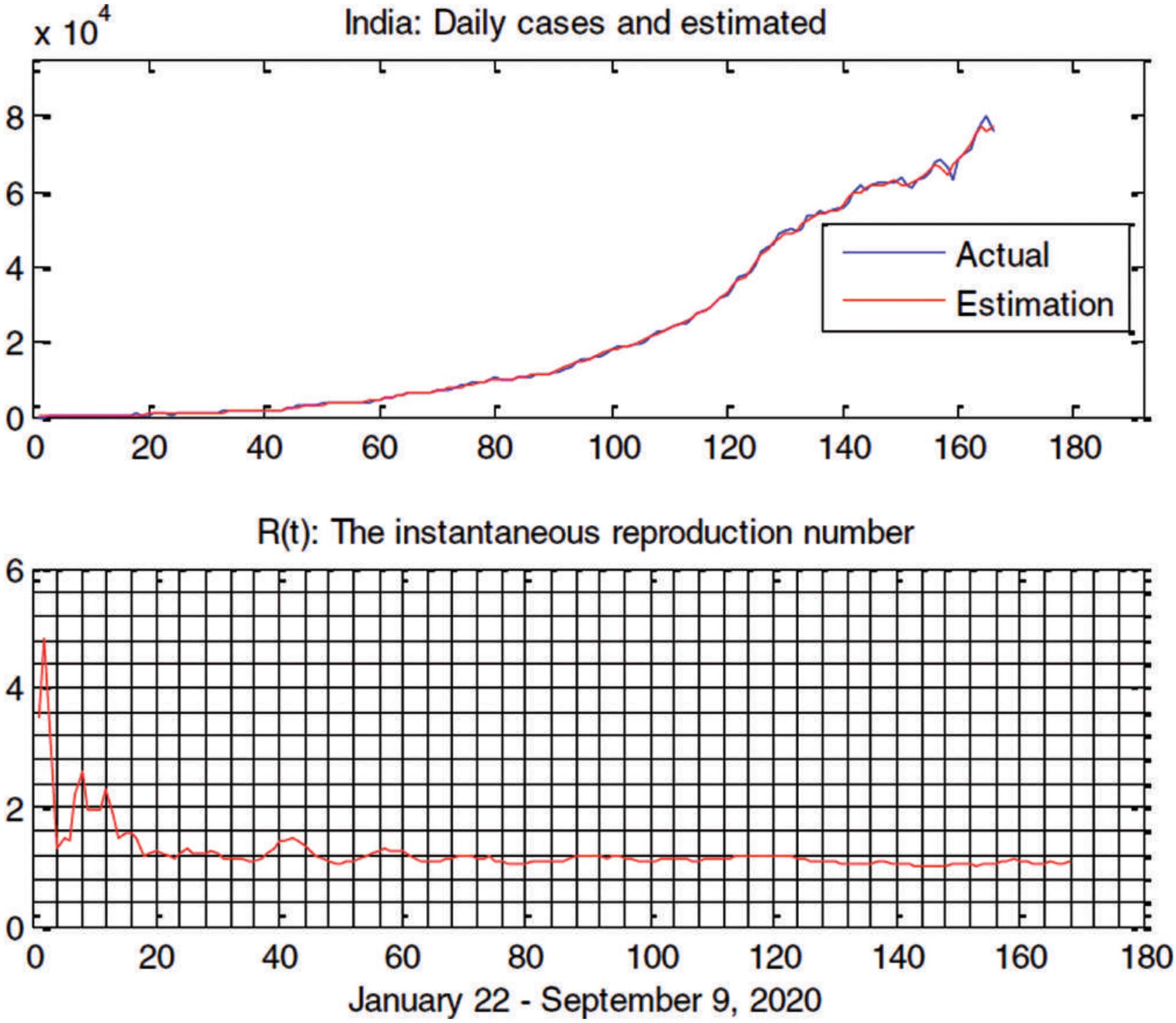

Figure 8. India: Daily cases and estimated, reproduction number estimation.

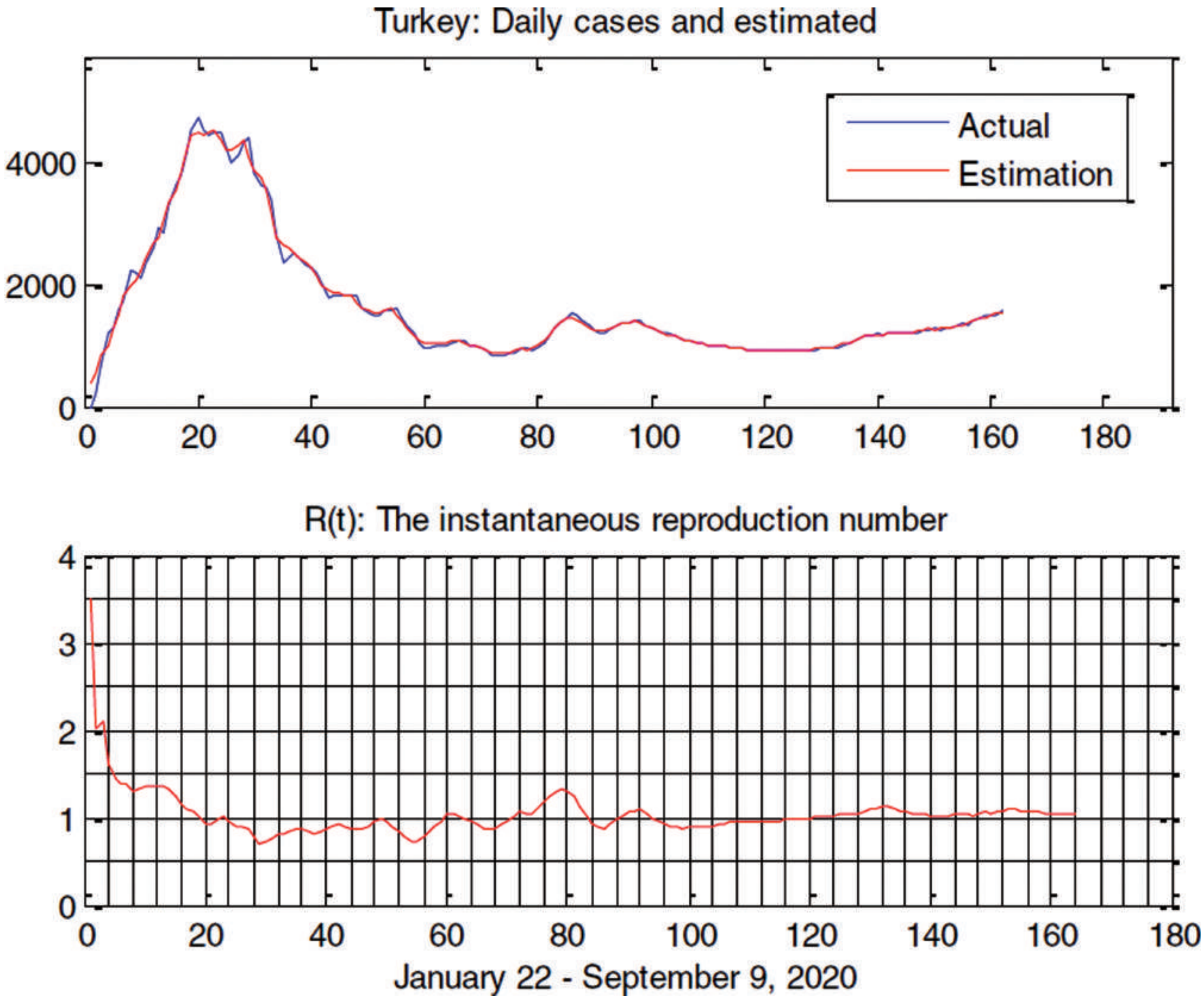

Figure 9. Turkey: Daily cases and estimated, reproduction number estimation.

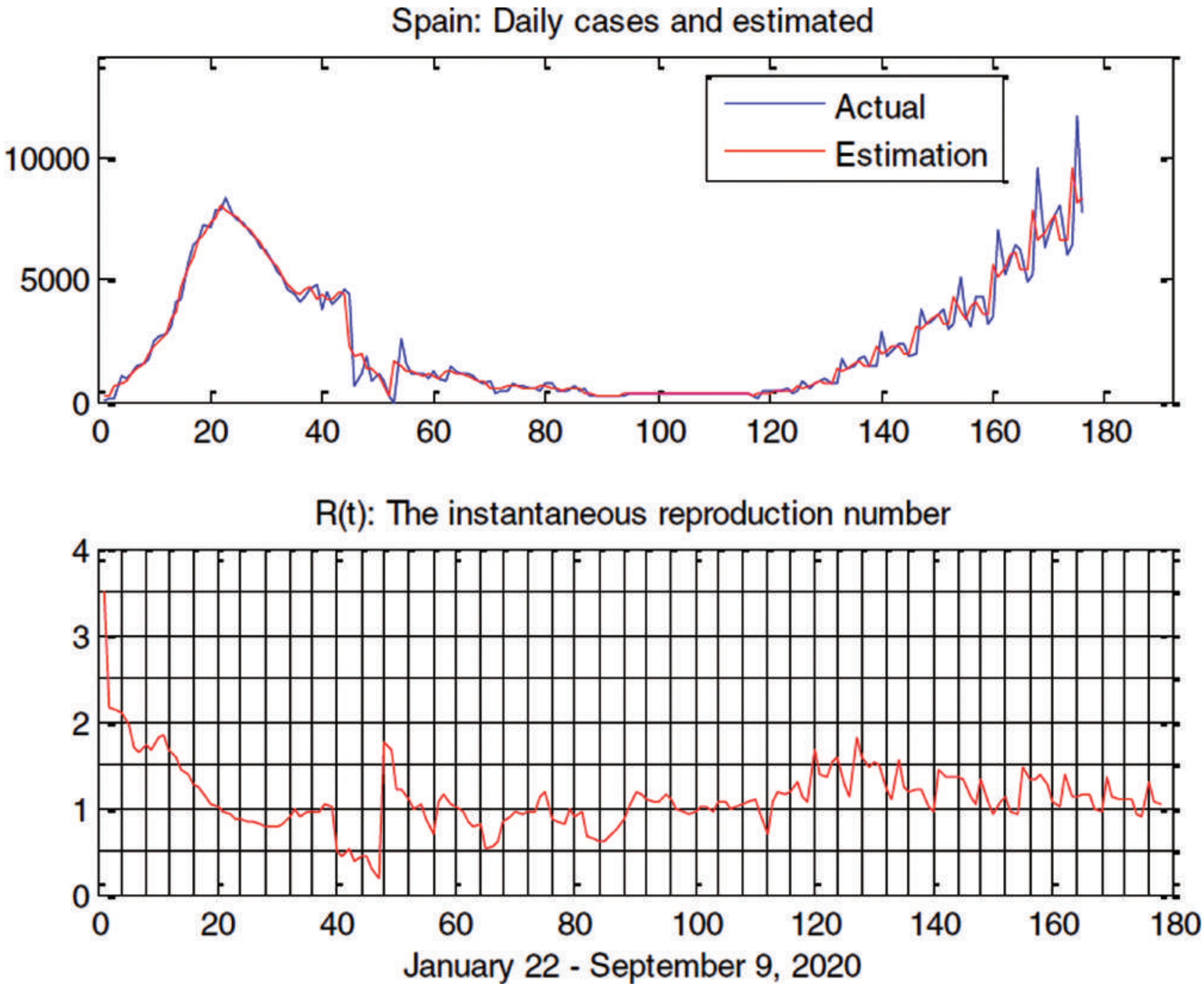

Figure 10. Spain: Daily cases and estimated, reproduction number estimation.

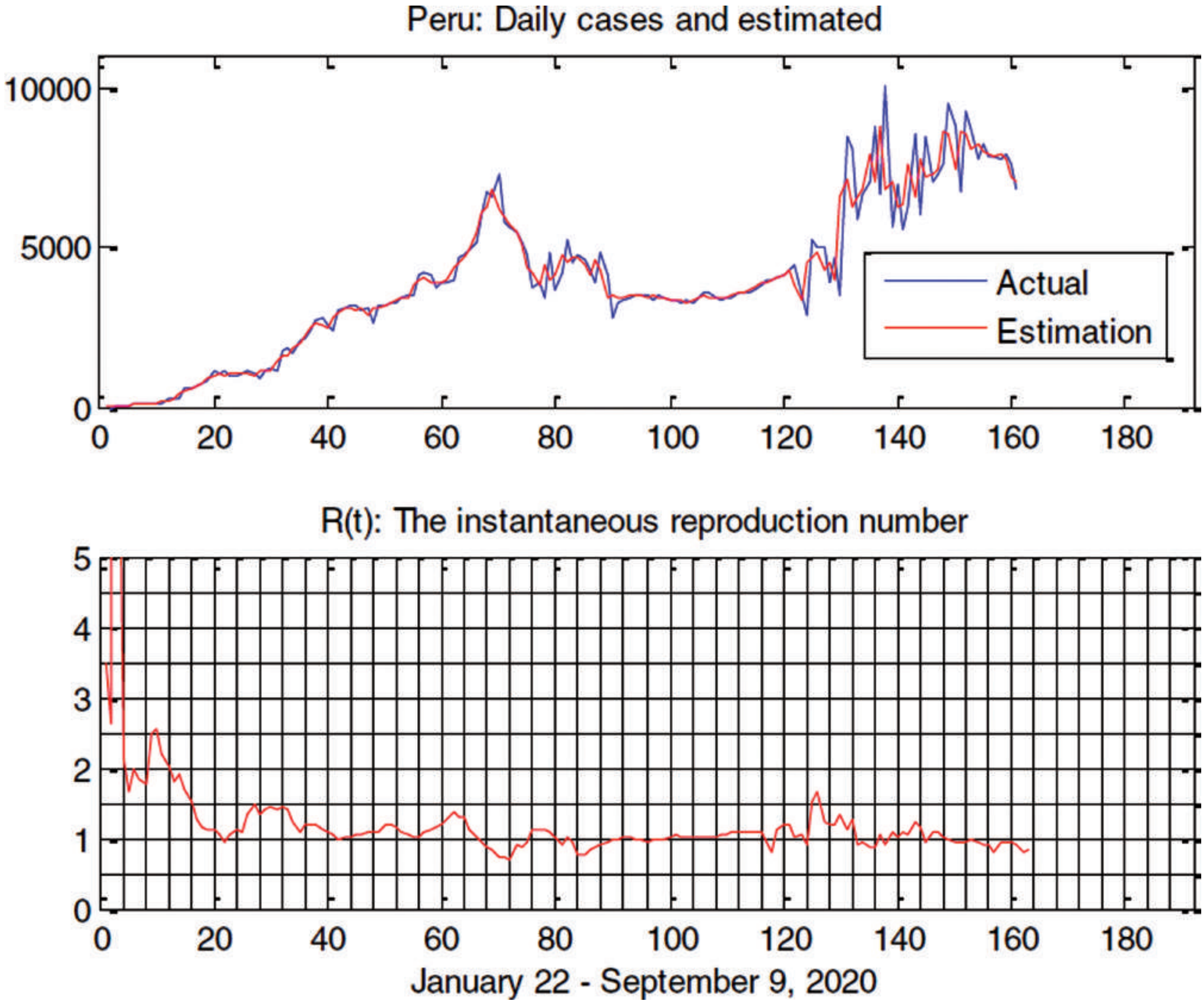

Figure 11. Peru: Daily cases and estimated, reproduction number estimation.

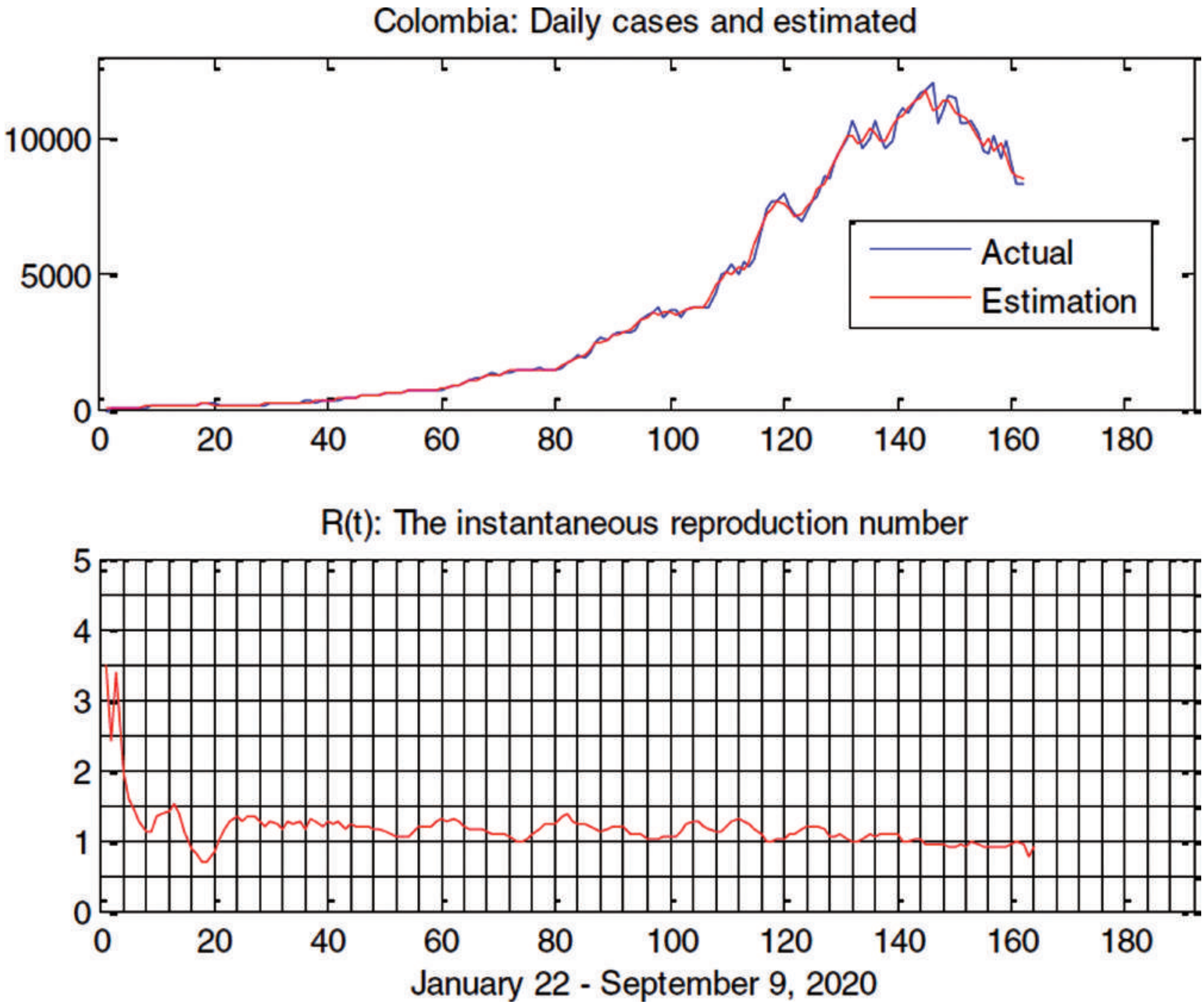

Figure 12. Colombia: Daily cases and estimated, reproduction number estimation.

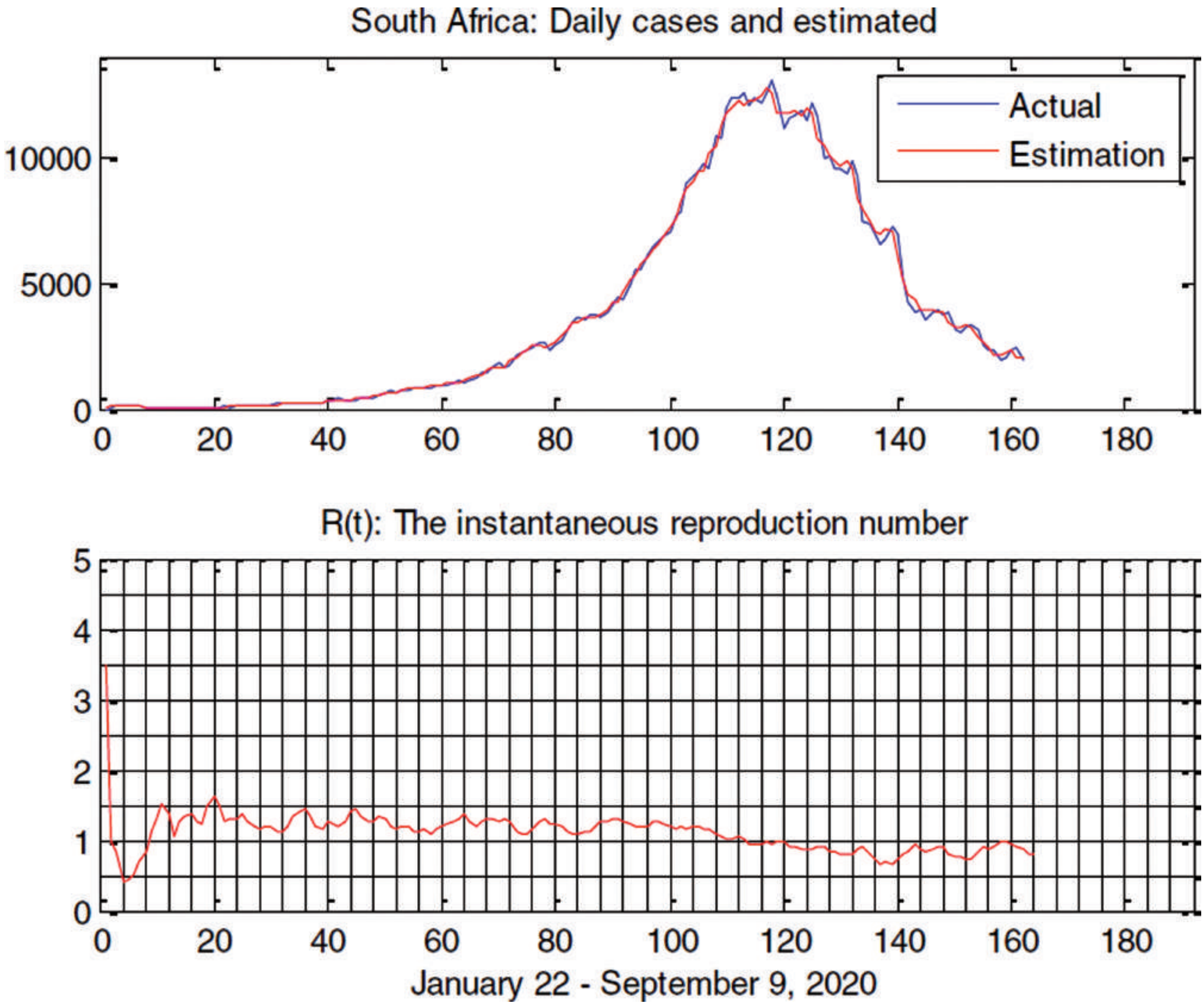

Figure 13. South Africa: Daily cases and estimated, reproduction number estimation.

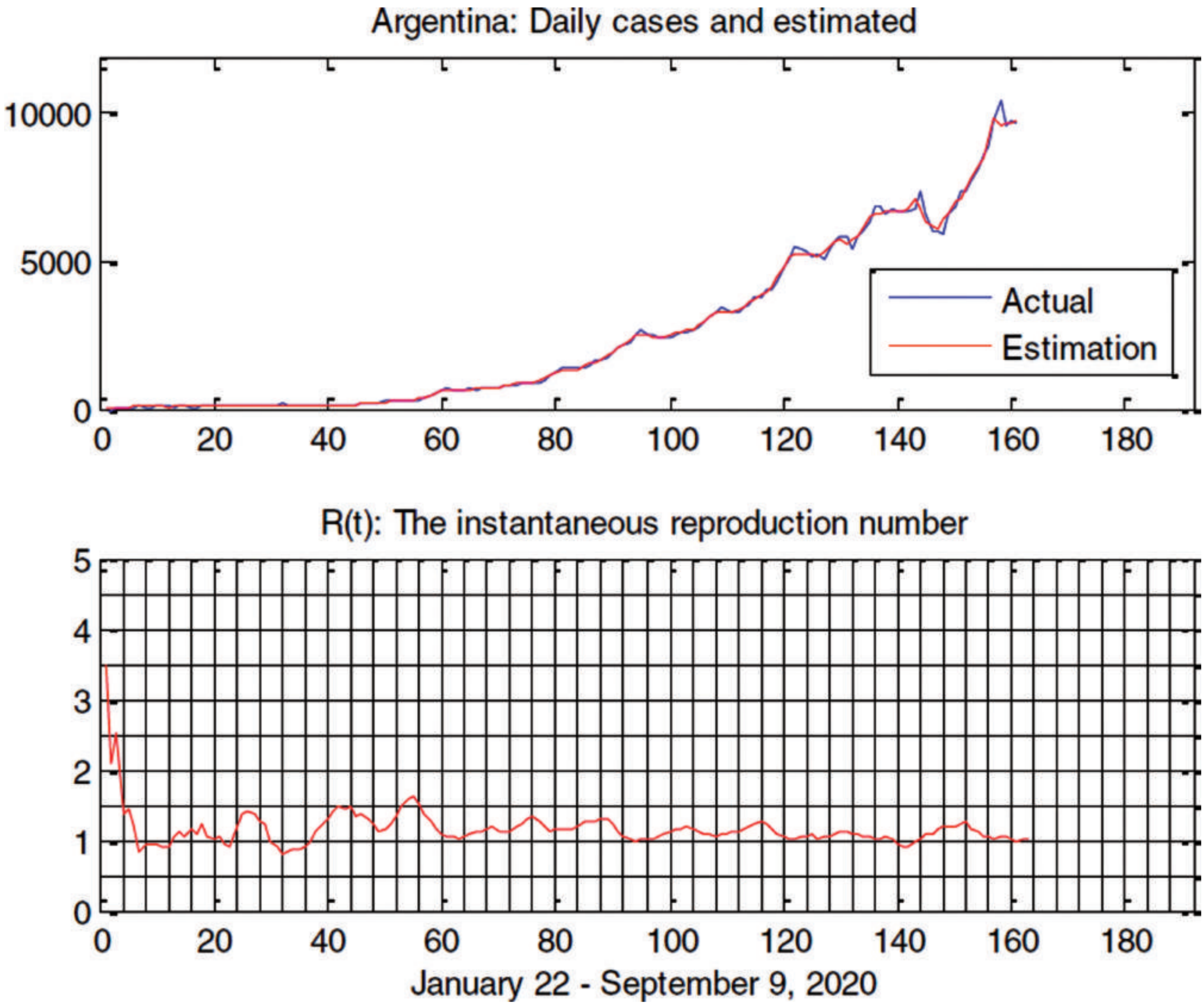

Figure 14. Argentina: Daily cases and estimated, reproduction number estimation.

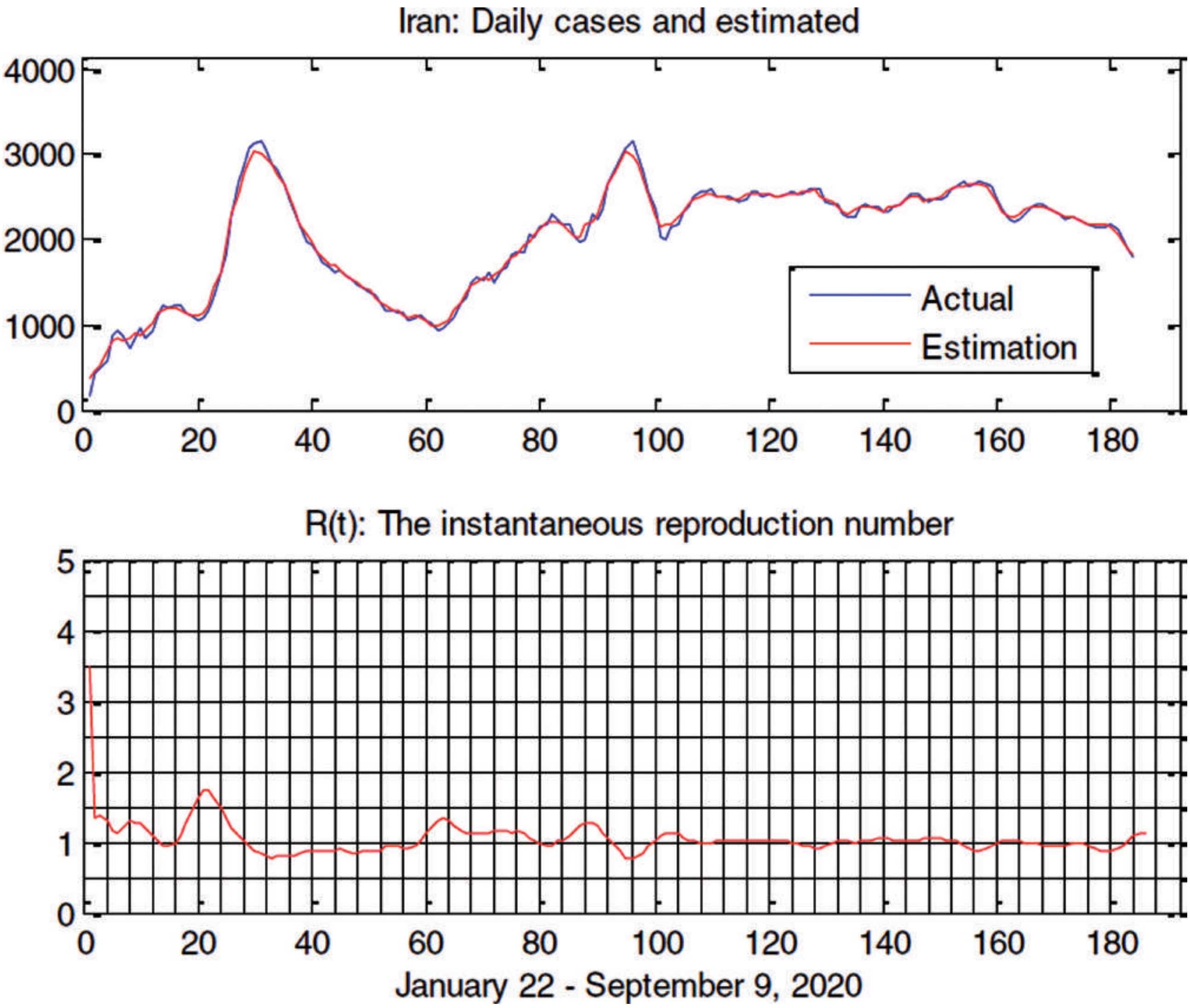

Figure 15. Iran: Daily cases and estimated, reproduction number estimation.

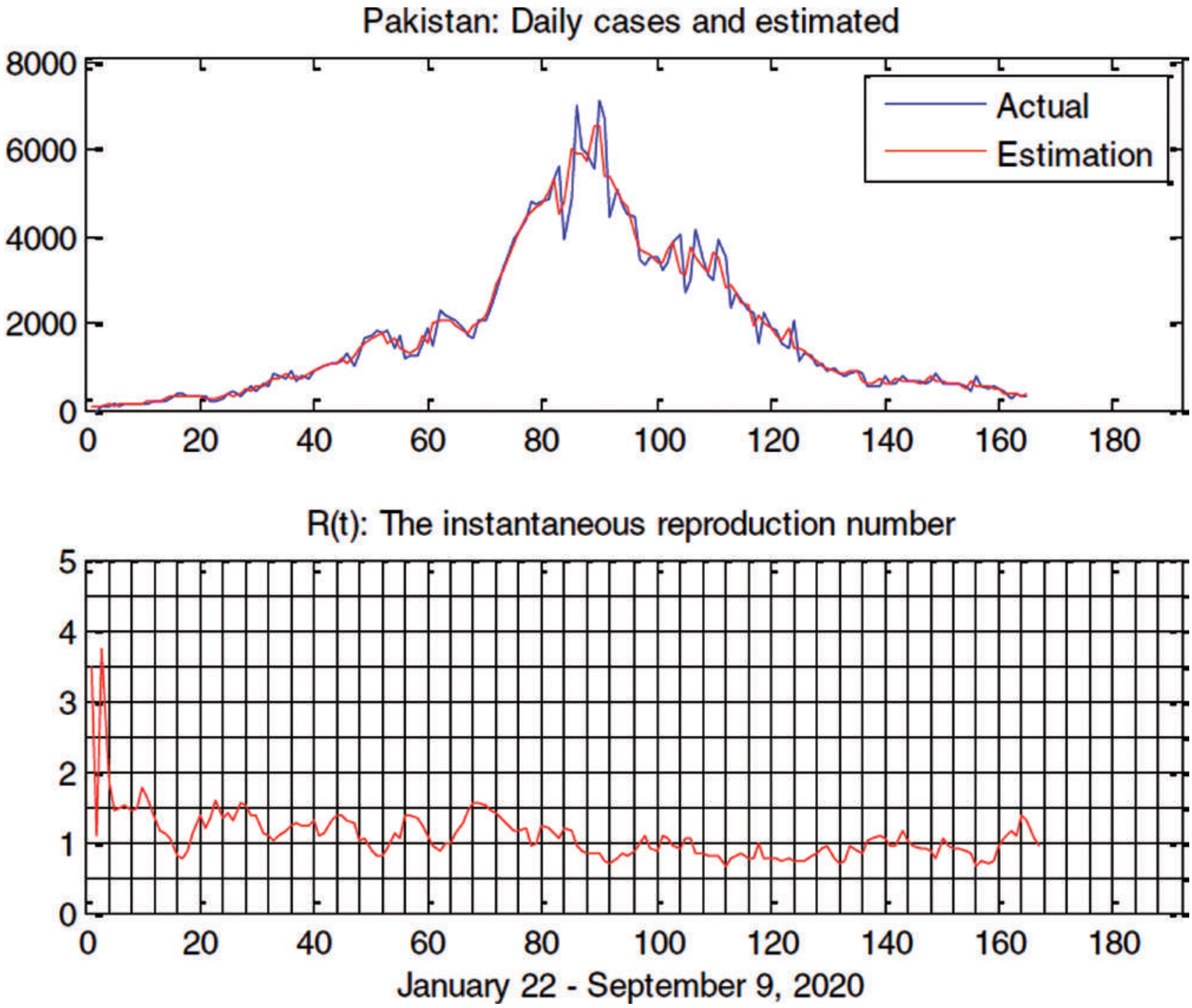

Figure 16. Pakistan: Daily cases and estimated, reproduction number estimation.