ISSN 2090-3359 (Print)
ISSN 2090-3367 (Online)

Advances in Decision Sciences

Volume 26
Issue 1
March 2022

Published by Asia University, Taiwan

Modeling COVID-19 Confirmed Cases Using a Hybrid Model

Samya Tajmouati (corresponding author)

Department of Mathematics, Ibn Tofail University, Faculty of Sciences, Kénitra, Morocco

samya.tajmouati@gmail.com

Bouazza El Wahbi

Department of Mathematics, Ibn Tofail University, Faculty of Sciences, Kénitra, Morocco

bouazza.elwahbi@uit.ac.ma

Mohamed Dakkon

Department of Economics and Management, Abdelmalek Essaâdi University, FSJES Tétouan, Morocco

m.dakkoun@gmail.com

Received: November 08, 2021; First Revision: December 28, 2021;

Last Revision: March 10, 2022; Accepted: March 15, 2022;

Published: March 20, 2022

Abstract

Purpose: The COVID-19 virus has caused numerous problems worldwide. Given the negative effects of COVID-19, this study aims to estimate accurate forecasts of the number of confirmed cases to help policymakers determine and make the right decisions.

Design/methodology/approach: This paper uses a hybrid approach for forecasting the daily COVID-19 cases based on combining the Autoregressive Integrated Moving Average (ARIMA) and Autoregressive Neural Network (NNAR) with a single hidden layer. To fit the linear pattern from the data, ARIMA models are used. Then, the NNAR models are used to capture the nonlinear pattern. The final prediction is obtained by adding up the two predictions.

Findings: Using six-time series from January 22, 2020, to June 22, 2021, of new daily confirmed cases of COVID-19 from Pakistan, Tunisia, Indonesia, Malaysia, India and South Korea, this work evaluates the hybrid approach against some benchmark models and generated ten days ahead forecasts. Experiments demonstrate the superiority of the hybrid model over the benchmark models.

Originality/value: Given the complex nature of new confirmed cases, it is assumed that the data contains both linear and nonlinear components. In literature, different studies have tended to forecast future cases of COVID-19. However, most of them have used single models that capture either linear or nonlinear patterns. This paper proposes a hybrid model that captures both linear and nonlinear components from the data.

Keywords: COVID-19, Time series forecasting, Autoregressive Integrated Moving Average, Autoregressive Feedforward Neural Network.

Introduction

In late December 2019, the COVID-19 pandemic was first identified in Wuhan city, China, with similar clinical symptoms to common colds (Huang et al., 2020). It is characterized by a rapid rate of spread and a high level of harm compared to the previous infectious diseases, which make its control hard (Kırbaş et al., 2020). Since its emergence, COVID-19 has negatively affected people's health and different sectors. As a result, on December 20, 2021, the global confirmed cases of the pandemic reached 273900334 with 5351812 deaths. Even though different vaccines have shown up, the coronavirus continues to spread, and different variants have taken place in different parts of the world so that the virus has caused various disruptions in different levels, including the economic, social and financial ones. More precisely, personal lives, businesses and economic activities have been negatively affected due to the restrictions made by the governments. The restrictions include travel banning, social distancing, lockdowns, school closure and public events banning, to name a few. Though these measures are meant to curb the coronavirus spread, they have led to disruptions in household income, production, transportation, and distribution of goods, including food. In particular, the enforcement of these measures is likely to lead to increases in food prices, food crises and food insecurity. The reduction of food and labor supply caused by the restrictions are likely to trigger food security issues. Many studies reported by (Agyei et al., 2021) showed that the world would witness an increase in poverty and food insecurity due to the pandemic, which would avoid achieving sustainable development goals. In this context, Agyei et al. (2021) studied the direct and indirect effects of the COVID-19 on food prices in Sub-Saharan Africa. Moreover, COVID-19 has negatively affected the exportations. For example, Safi et al. (2022), showed how the COVID-19 pandemic has disrupted the flow of goods that China exports to other parts of the world and forecast the China exports during the COVID-19 pandemic using seven different forecasting models.

On the other hand, the COVID-19 pandemic has impacted the financial sector (Moslehpour et al., 2022). Owusu Junior et al. (2021) argued that the global financial markets are likely to respond to the implications of COVID-19 at the economic level, including a reduction in productive activities, an increase in unemployment and a gradual reduction in market participants, to name a few. Implicitly, they investigated the impact of COVID-19 in stock markets by quantifying the flow from COVID-19 to major global equities in different levels of time (short, medium and long terms) by combining the Rényi entropy specification and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) technique.

Given the severe damage caused by COVID-19, forecasting the number of cases infected with this pandemic is a primordial task to help policymakers implement the proper measures regarding different sectors**.** For instance, in the financial sector, forecasting the COVID-19 cases helps quantify the flow from COVID-19 to global equities in the future, allowing policymakers and investors to implement informed decisions (Moslehpour et al. 2022; Owusu Junior et al.,

2021). Similarly, Agyei et al. (2021) concluded a significant relationship between COVID-19 and staple food prices in Sub-Saharan Africa. So, forecasting the number of confirmed cases infected with COVID-19 helps evaluate the future staple food prices, leading governments of sub-Saharan Africa to think about investing in infrastructures that improve efficiencies in the food supply chain during pandemics. In order to forecast the number of confirmed cases, various mathematical and statistical tools were applied. According to Torrealba-Rodriguez et al. (2020), computational and mathematical models like Logistic, Artificial Neural Network (ANN) and Gompertz were employed to forecast the number of cases in Mexico. For Tran et al. (2020), the Autoregressive Integrated Moving Average (ARIMA) model was used to predict the daily total confirmed cases/new cases, the total deaths/ new deaths and the growth rate in confirmed cases/deaths in Iran. Abbasimehr and Paki (2021) proposed three hybrid models that combine Long Short-Term Memory (LSTM), multi-head attention and Convolutional Neural Network (CNN) with the Bayesian optimization algorithm to forecast the number of daily infected cases with COVID-19. Tajmouati et al. (2020) introduced a new approach to forecasting the total number of confirmed cases of COVID-19 based on the k-nearest neighbors' algorithm (k-NN). For more details, the reader should refer to the study conducted by Tajmouati et al. (2021). Moftakhar et al. (2020) compared the ARIMA and ANN models to predict the subsequent thirty daily new infected cases with COVID-19. In (Kırbaş et al. (2020)), confirmed COVID-19 cases of Denmark, Belgium, Germany, France, United Kingdom, Finland, Switzerland and Turkey were modeled with ARIMA, Nonlinear Autoregression Neural Network (NARNN) and LSTM approaches. Chakraborty and Ghosh (2020) presented a hybrid approach based on ARIMA and Wavelet-based forecasting (WBF) models to generate short-term forecasts (ten days ahead) of the number of daily-confirmed cases for Canada, France, India, South Korea, and the UK.

ARIMA models are widely used in linear time series forecasting. These models strongly assume that the future data are linearly dependent on current and past observations. However, many real-world time series data exhibit complex nonlinear patterns which cannot be modeled effectively by ARIMA. Several nonlinear models have been proposed in the literature to overcome this restriction. These include the bilinear model (Anderson et al., 1979), the threshold autoregressive (TAR) (Tong, 2011), and the autoregressive conditional heteroscedastic (ARCH) model (Engle, 1982). However, these methods present some limitations: they can model nonlinear time series that have specific nonlinear patterns. Therefore, their use in general forecasting problems is limited (Zhang, 2003). Artificial Neural Networks (ANNs) have been proposed to forecast time series. Their main power lies in their flexible nonlinear modeling capability. Over the other models can approximate a large class of functions with high accuracy without making assumptions on data or model form. More precisely, with ANN, there is no need to specify a particular model form. Instead, the model is adaptively formed based on data features.

The ANN is suitable for empirical data for which no theoretical guidance is available to suggest data generating process (Zhang, 2003). Some studies report the superiority of linear and nonlinear models in the literature. For example, Aras and Kocakoç (2016), Foster et al. (1992),

and Casey Brace et al. (1991) showed that linear models outperform ANNs. Likewise, Denton (1995), Hann and Steurer (1996), and Callen et al. (1996) reported the success of ANN over linear models when time series present high volatility and collinearity. However, no universal model is suitable under all circumstances (Büyükşahin & Ertekin, 2019). In order to overcome this limitation, various hybrid approaches have been proposed in the literature to take advantage of each employed model. The time-series data is often decomposed into linear and nonlinear patterns; then, one can model them separately using an adequate model.

Khashei and Bijari (2011) proposed a successful ARIMA-ANN that defines a functional relationship between the linear and nonlinear components. Babu and Reddy (2014) devised a new hybrid ARIMA-ANN that proposes a solution to volatile time series using a moving average filter. Zhang (2003) proposed a hybrid ANN-ARIMA model that achieves accurate predictions compared to individual models in many applications such as Adhikari and Agrawal (2013) and Ömer Faruk (2010). In this context, this paper implemented the hybrid ANN-ARIMA proposed by Zhang (2003) to forecast the ten days ahead of the number of new confirmed cases for Pakistan, Tunisia, Indonesia, Malaysia, India, and South Korea. For simplicity, the Feed-forward Autoregression Neural Network with one single hidden layer is chosen among significant variants of neural networks since it is the most widely used model that brings fruitful results (Zhang, 2003; Safi et al. 2022).Experiments showed that the proposed hybrid approach is better than single and hybrid ARIMA-WBF models (Chakraborty & Ghosh, 2020). This result is evident since the COVID-19 confirmed cases contain both linear and nonlinear patterns. Therefore, making decisions based on an individual model would be critical. Moreover, the WBF model is designated to forecast non-stationary patterns (Chakraborty & Ghosh, 2020). As a result, combining it with ARIMA model is adapted to time series that are only linear and non-stationary, which is not the case for the new confirmed cases.

.

The primary motivation of this study is to model the new confirmed cases of COVID-19 accurately. Accurate predictions are essential for the preparedness of outbreak management and anticipation of efficient policies. Due to the complex nature of new confirmed cases, linear models like ARIMA are hardly reasonable for such series. On the other hand, nonlinear models such as ANN are designed for pure nonlinear time series. As real-world time series are rarely pure linear or nonlinear (Zhang, 2003), it is strongly believed that the new confirmed cases contain both linear and nonlinear patterns. Thus, such a series would be better fitted by a model capturing both linear and nonlinear patterns, like the one presented in this study.

The main contributions of this paper are:

- This study confirms the proposed hybrid model ARIMA-NNAR over ARIMA, NNAR, and ARIMA-WBF in capturing the linear and nonlinear patterns of new confirmed cases of COVID19 for Pakistan, Tunisia, Indonesia, Malaysia, India and South Korea. More

concisely, experiments show that the accuracy measures are minimal for the ARIMA-NNAR model compared to the other benchmark models for these countries.

- The results of this study are expected to help policymakers develop appropriate measures and policies regarding the COVID-19 outbreak.
- This study is one of the alternative studies for modeling the number of cases of COVID-19 for other countries. It shares the code used to forecast the COVID-19 confirmed cases with the ARIMA-NNAR model. Hence, the user could employ this code to forecast the updated confirmed cases for different countries.

The rest of the paper is organized as follows: Section 2 presents an overview of ARIMA and Feedforward Autoregression Neural Network with a single hidden layer model and the generated hybrid model. Section 3 describes the experimental results for short-term forecasting of COVID-19 and compares the proposed model's performance to some benchmark models. Finally, section 4 contains a summary.

Methodology

This section briefly describes the models used in hybridization: ARIMA and NNAR models and the produced hybrid model ARIMA-NNAR**.** The motivation for using the hybrid model comes from the following reasons (Zhang, 2003). First, it is difficult for forecasters to determine an adequate model for a given time series. Therefore, several models are tested and selected with the most accurate result. However, the selected model is not necessarily the best for future data because of many factors such as structure change or model uncertainty. Therefore, combining different single models could solve the problem of model selection. Second, real-world time series are rarely pure linear or nonlinear and often contain linear and nonlinear structures. As a result, such series can be modeled neither by ARIMA nor ANN since ARIMA cannot model nonlinear patterns and ANN alone cannot deal with the linear ones. Hence, such data could be accurately modeled by combining both models. Finally, almost universally agreed that no single model is the best in every situation. Moreover, many empirical studies, including several large-scale forecasting competitions, suggest that hybridization could improve the accuracy of single models (Zhang, 2003; Clemen, 1989; Ömer Faruk, 2010; Khashei & Bijari, 2011; Büyükşahin & Ertekin, 2019; Babu & Reddy, 2014).

ARIMA model

ARIMA is a classical time series model based on a description of the linear autocorrelations in the data. Given a time series y_t, an ARIMA(p,d,q) model is given by:

$$y_t^{'} = c + \phi_1 y_{t-1}^{'} + \cdots + \phi_p y_{t-p}^{'} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_p \varepsilon_{t-p} + \varepsilon_t, \quad (1)$$

where $y_t^{'}$ is the differenced series, p and q are the order of the autoregressive model (AR) and the moving average model (MA), respectively, d is the level of differencing and ε_t is the random error at time t. The ϕ_i and θ_j are the coefficients of the ARIMA model. The error terms ε_t are supposed to be independent and identically distributed and follow a distribution with zero mean and constant variance (i.e., white noise). When fitting an ARIMA model to a given time series data, the Box-Jenkins approach is used (Ziegel et al., 1995). It refers to the iterative application of the following three steps:

- *Identification:* involves determining the possible orders of the ARIMA model (*p,d,q*) for a given time series. It consists first of transforming the series in order to make it stationary and stabilize its variance (the number of differentiations determines the order of integration: d), and then identifying the possible ARMA orders (*p,q*) of the transformed series by using the ACF and PACF plots. The first ACF and PACF that are significant are picked.

- *Estimation:* involves estimating the parameters $\varphi_i$and θ_j of the different models obtained in the *Identification* step using the least square or maximum likelihood methods and proceeds to the first selection of models whose information criteria are minimal (*AIC,BIC ou AICc*).

- *Diagnostic checking:* involves determining whether the model(s) specified in the *Estimation* step are adequate. Thus, the residuals obtained from the estimated model(s) are used to check whether they follow a white noise, using a Portmanteau test, usually the Box-Pierce test (See the Appendix).

These steps are applied iteratively until the *Diagnostic checking* step involves the adequacy of the model. The Hyndman-Khandakar algorithm is provided by the *auto.arima* function in R (Hyndman et al., 2008) is used to fit automatically the ARIMA model, which chooses *AICc* as an information criterion. However, this algorithm only takes care of the *Identification* and *Estimation* steps. So, even if someone uses it, he will still need to take the *Diagnostic checking* step himself.

NNAR model

A neural network is a network of "neurons" organized in layers: Input layer, hidden layers and output layer. It is used for complex nonlinear forecasting. Each layer of nodes receives inputs from the previous layers. The outputs of the nodes in one layer are inputs to the next layer. This study considers the autoregressive feedforward network with one hidden layer (Hyndman and Athanasopoulos, 2021). For a time series y_t, the input variables are its lagged values (y_{t-1},y_{t-2},...). In the input layer, the input variables (i.e., lagged values) are combined linearly to give $z_j = b_j + \sum_i w_{i,j} y_{t-i}$. In the hidden layer, the z_j are combined using a nonlinear function, called activation function, such as a sigmoid, $s(z_j) = \frac{1}{1+e^{-z_j}}$, to give the input for the next layer

(ie, output layer). Finally, the $s(z_j)$ are combined linearly to output $\hat{y}_t = w_0 + \sum_j s(z_j).w_j$. The parameters $b_j, w_j \; and \; w_{i,j}$ are learned from the data. The parameters take random values to begin with, and these are updated until the model fits well data. The autoregressive feedfarward network with one hidden layer is usually denoted by NNAR(p,P,k)$_m$ where (p + P) is the number of input nodes (lagged inputs) and k is the number of hidden nodes. More generally, NNAR(p,P,k)$_m$ model has inputs $(y_{t-1}, y_{t-2}, \ldots, y_{t-p}, y_{t-m}, y_{t-2m}, \ldots, y_{t-Pm})$ and k neurons in the hidden layer, where m is the seasonal period. A NNAR(p,P,0)$_m$ is equivalent to an ARIMA(p,0,0)(P,0,0)$_m$ model. To fit an NNAR(p,P,k)$_m$, the *nnetar()* function from the R package *forecast* can be used. If the parameters p, P and k are not specified, *nnetar()* selects them automatically. For non-seasonal time series, *nnetar* takes p as the optimal number of lags (according to AIC) for a linear AR(p). For seasonal time series, the default values are P=1 and p is chosen from the optimal linear model fitted to the seasonally adjusted data. If k is not specified, the default value is $k = (p + P + 1)/2$ (rounded to the nearest integer).

Hybrid ARIMA-NNAR model

This study proposes a hybridization of ARIMA and NNAR models. The new confirmed cases of COVID-19 are complex and can be seen as a combination of linear and nonlinear components. Therefore, the ARIMA model fails to produce accurate forecasts for such data sets. The ARIMA model underestimates the data, and there is still a part of the information that ARIMA cannot explain. This paper tends to remodel the ARIMA residuals using a NNAR model to solve the problem. The final predictions are obtained by summing the predictions from ARIMA and the adjusted residuals from NNAR model. Overall, in the ARIMA-NNAR model, the ARIMA is firstly used to model the linear component. Let e_t denote the residual of the ARIMA model at time t, then:

$$e_t = y_t - \hat{L}_t, \tag{2}$$

where $\hat{L}_t$ is the estimation of y_t by the ARIMA model. After that, the residuals e_t are modeled using the NNAR model. Thus, with n inputs nodes, the NNAR model outputs:

$$\hat{e}_t = w_0 + \sum_j w_j.s\left(b_j + \sum_{i=1}^{n} w_{i,j} e_{t-i}\right), \tag{3}$$

where s is the activation function, $w_j, b_j and \; w_{i,j}$ are the NNAR parameters, and $\hat{e}_t$ is the estimated value of e_t by NNAR. Finally, the prediction of y_t is obtained as follows:

$$\hat{y}_t = \hat{L}_t + \hat{e}_t. \tag{4}$$

To sum up, the general steps of the hybrid ARIMA-NNAR model are reported in figure 1.

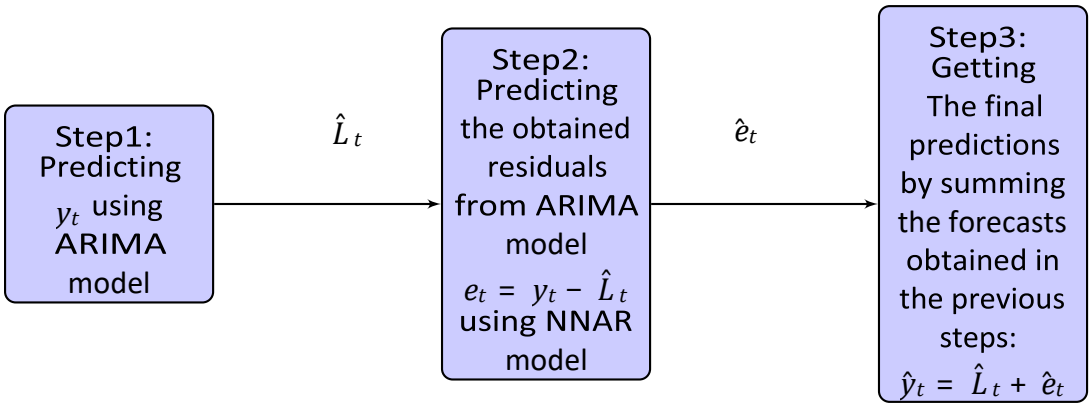


Figure 1. The General Steps of Predicting the Time Series y_t using the Hybrid ARIMA-NNAR Model

Experimentation

In this paper, the dataset, including the number of daily new confirmed cases for six different countries, namely Pakistan, Tunisia, Indonesia, Malaysia, India, and South Korea, compares the proposed hybrid model against some benchmark models. The dataset consists of 528 daily observations ranging from January 22, 2020, to July 02, 2021, out of which 518 data points are used for modeling purposes, and the rest ten are kept for assessing the model. The description of the datasets, the procedure of the hybrid method, and the experimentation's results are presented in the following subsections.

Datasets

The six univariate time series, namely the number of daily new confirmed cases for Pakistan, Tunisia, Indonesia, Malaysia, India and South Korea, are retrieved from the R package *coronavirus* that collects the data from Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) and is taken from January 22, 2020, to July 02, 2021. Overall, each dataset contains 528 observations. Figure 2 depicts the daily new confirmed cases of COVID-19 and shows the trend across the time for the six countries. The plots reveal clearly that the data are not stationary. Descriptive statistics of daily confirmed cases are given in table 1. According to table 1, the mean is greater than the median for all countries, which indicates that the data are positively skewed. Figures A.10, A.11, A.12, A.13, A.14, and A.15 presented in Appendix A.3 display the ACF and PACF plots for the six-time series and their differences at lags one and two. According to these figures, neither the ACF nor PACF cut off strictly at a

certain lag. Instead, they show infinite behavior. Thus, ARIMA models might not be suitable for such series.

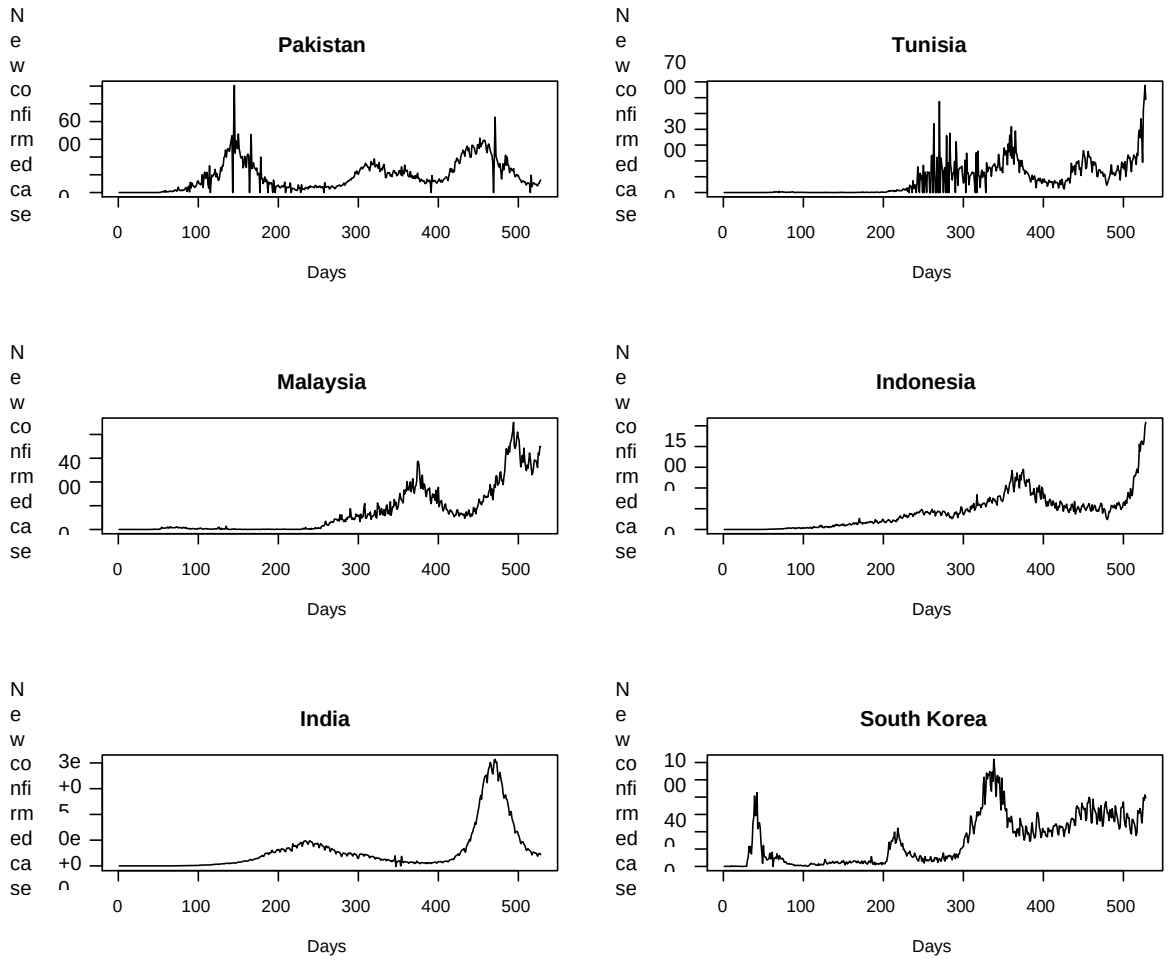


*Figure 2***.** Actual Values of New Confirmed Cases of COVID-19 for the Six Countries

Table 1

Descriptive Statistics of Daily Confirmed Cases

Country	Min	Q_1	Median	Mean	Q_3	max
Pakistan	0	570	1358	1820	2730	12073
Tunisia	0	4	384.5	819.6	1410.2	6776
Indonesia	0	692	3834	4221	5939	25830
Malaysia	0	19.75	576	1450.66	2192.75	9020
India	0	8278	28552	57770	64438	414188
South Korea	0	48.75	201.50	301.78	508	1237

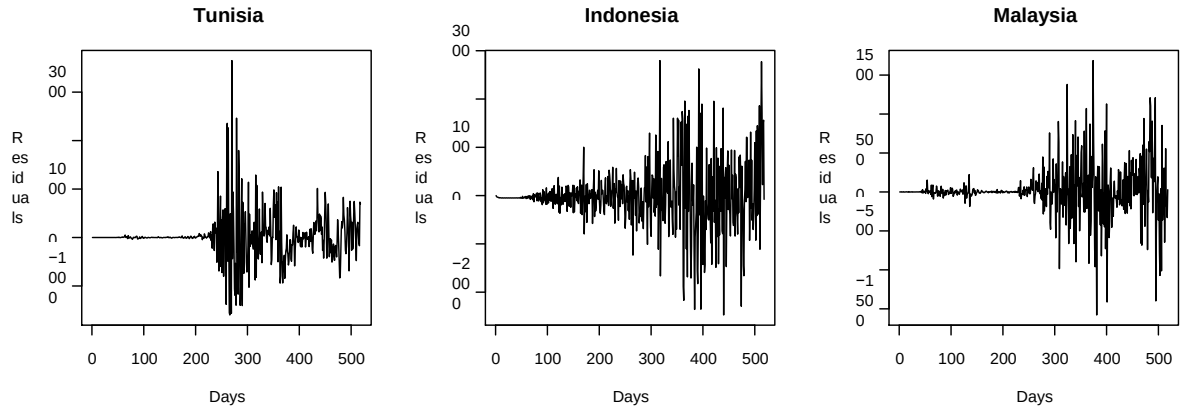
The ARIMA-NNAR model

The ARIMA-NNAR modeling is carried out in three steps:

Step 1: (Linear modeling): In a first step, the best ARIMA models are fitted by using the function *auto.arima* from the R package *forecast*. Overall, the best fitted models are *ARIMA*(4,1,0), *ARIMA*(5,1,3), *ARIMA*(3,1,5) with drift, *ARIMA*(3,1,2), *ARIMA*(0,1,2) and ARIMA(5,1,2) for Pakistan, Tunisia, Indonesia, Malaysia, South Korea, and India respectively. It is interesting to see that the residuals generated by ARIMA are far to be white noise (see figure 3). Consequently, there is a part of the information which ARIMA did not explain. The NNAR model is used to remodel this part, as explained in the next step.

Step 2: (nonlinear modeling): In a second step, the function *nnetar* from the R package *forecast* is used to fit the best model for the obtained residuals from step 1. The neural network is trained 1000 times and use the logistic function (i.e., sigmoid) as the activation function. Overall, the best fitted models are *NNAR*(21,0,11), *NNAR*(21,0,11), *NNAR*(15,0,8), *NNAR*(21,0,11), *NNAR*(27,0,14) and NNAR(27,0,14) for Pakistan, Tunisia, Indonesia, Malaysia, South Korea, and India respectively.

Step 3: (combination): Finally, the obtained results from step 1 and 2 are combined together.



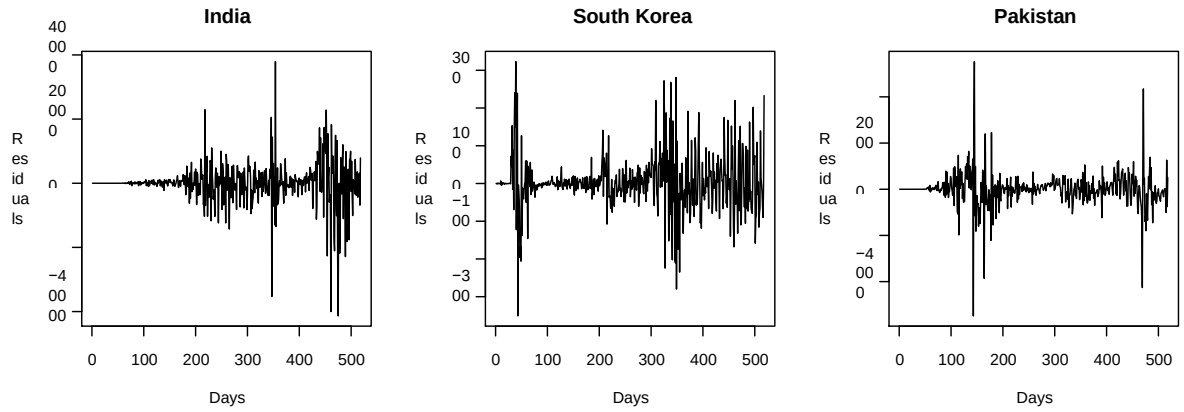


Figure 3. Plots of ARIMA Residuals for the Six Countries

Practically, this paper uses the following R code (see Figure 4):

```
#load packages
library(coronavirus)
library(dplyr)
library(forecast)

#### Code assumes that the time-series data set for a country is loaded as a variable y ####
y=ts(coronavirus %>%
              filter(country=="country's name ",type == "confirmed") %>%
              group_by(date)%>%
              summarise(cases = sum(cases))%>%
              select(cases))

#### Build an ARIMA model
mymodel=auto.arima(y)

#### 10-step ahead forecast from the fitted ARIMA model
fit=forecast(mymodel,h=10)

#### Assign residuals to a variable
res=fit$residuals

#### Fit a NNAR model to the residuals obtained from the previous step
nnetar_res=nnetar(res,P=0,repeats=1000)

#### Add the fitted ARIMA outputs to the fitted NNAR model outputs
hybrid_fit=nnetar_res$fitted+mymodel$fitted

#### Compute accuracy of the hybrid model
accuracy(hybrid_fit,y)

#### Final 10-day-ahead forecast
final_forecast = fit$mean+forecast(nnetar_res,h=10)$mean
```

Figure 4. The code used in Predicting New COVID-19 Confirmed Cases by ARIMA-NNAR model

Note: For a real forecast system, one can update the actual confirmed cases and use the previous code presented in Figure 4 to generate updated predictions.

Results

To determine the efficiency of the ARIMA-NNAR model, it is compared with three benchmark models: ARIMA, NNAR, and ARIMA-WBF. The root mean square error (RMSE), mean absolute error (MAE), and mean percentage errors (MAPE) are used to evaluate the predictive performance of the models. To the best of our knowledge, these are the most common measures used to assess the accuracy of nonlinear and hybrid models (Ravazzolo et al. 2020; Safi et al. 2022). Table 2 shows the accuracy of the prediction of ARIMA-NNAR model over the benchmark models when the 518 observed data are compared with the prediction data. At this stage, we did not use MAPE for comparison since the data contain some zero values, which leads MAPE to take infinity values.

Similarly, table 3 reports the accuracy of the training models' prediction when the last ten observations are compared with the prediction data. According to table 2, the hybrid ARIMA-NNAR provides more accurate results as it outputs the lowest RMSE and MAE. The NNAR model ranks second. Finally, the remaining models have inferior accuracy. After training, the accuracy of trained models is assessed on the last ten observations. According to table 3, the accuracy measures of the hybrid ARIMA-NNAR are minimal for Pakistan, Tunisia, Indonesia, India and South Korea, which indicate its superiority over the other models. However, for Tunisia and Indonesia, though the accuracy measures of the ARIMA-NNAR model are minimal, they are still high, indicating overfitting. This could be explained by the change in structures of both series that ARIMA-NNAR fails to capture. Indonesia has recorded higher values since 24/06/2021 (see table A.7) compared to the period fitted by ARIMA-NNAR.

Similarly, Tunisia recorded higher values since 23/06/2021 (see table A.6) compared to the values fitted by the proposed model. Malaysia is the only country whose the NNAR gives the best results for the ten last observations, which indicates the nonlinearity of the series. On the other hand, the prediction errors by ARIMA are 17.72% and 89.41% for Pakistan and Tunisia, which are reduced by 20.20% and 53.09% by the application of hybrid ARIMA-NNAR. Similarly, the prediction errors by ARIMA-WBF are 45.77% and 15.18% for Indonesia and Malaysia, which are reduced by 30.69 % and 23.25% by ARIMA-NNAR. For India and South Korea, the prediction errors by the NNAR model are decreased by 43.41% and 37.25% by the application of the hybrid ARIMA-NNAR. A comparison of models fitted values and forecasts against actual values has been shown in figures 5, 6, 7, 8, 9 and 10, which reveals that the fitted values of ARIMA-NNAR model are close to the observed data. Furthermore, according to Box-Pierce test applied to the residuals obtained from the ARIMA-NNAR model: $r_t = y_t - \hat{L}_t - \hat{e}_t$, all the p-values are greater than 0.05 (see Appendix A.1). Therefore, there is no autocorrelation in r_t. Thus, the residuals $r_t$are white noise, which indicates that ARIMA-NNAR fits well the six series. On the other hand, the fitted values of ARIMA and ARIMA-WBF are far from the actual data. Moreover, the ten forecasts generated by the models are far from the observed data for Tunisia and Indonesia, which indicate that NNAR and ARIMA-NNAR overfit the two series. In

contrast, the ARIMA-NNAR and NNAR forecasts are close to the actual observations for Malaysia, India, and Pakistan. Finally, for South Korea, only the ARIMA-NNAR forecasts are near to the current data. As a result, the ARIMA-NNAR model is trustable for Pakistan, India, Malaysia, and South Korea, which confirms the results of table 3. For more transparency, the reader could check by himself the reliability of the presented results through the tables A.5, A.6, A.7, A.8, A.9, and A.10 presented in the Appendix A.2, which report the actual values against the predicted values based on the presented models for the period 23/06/2021-02/07/2021.

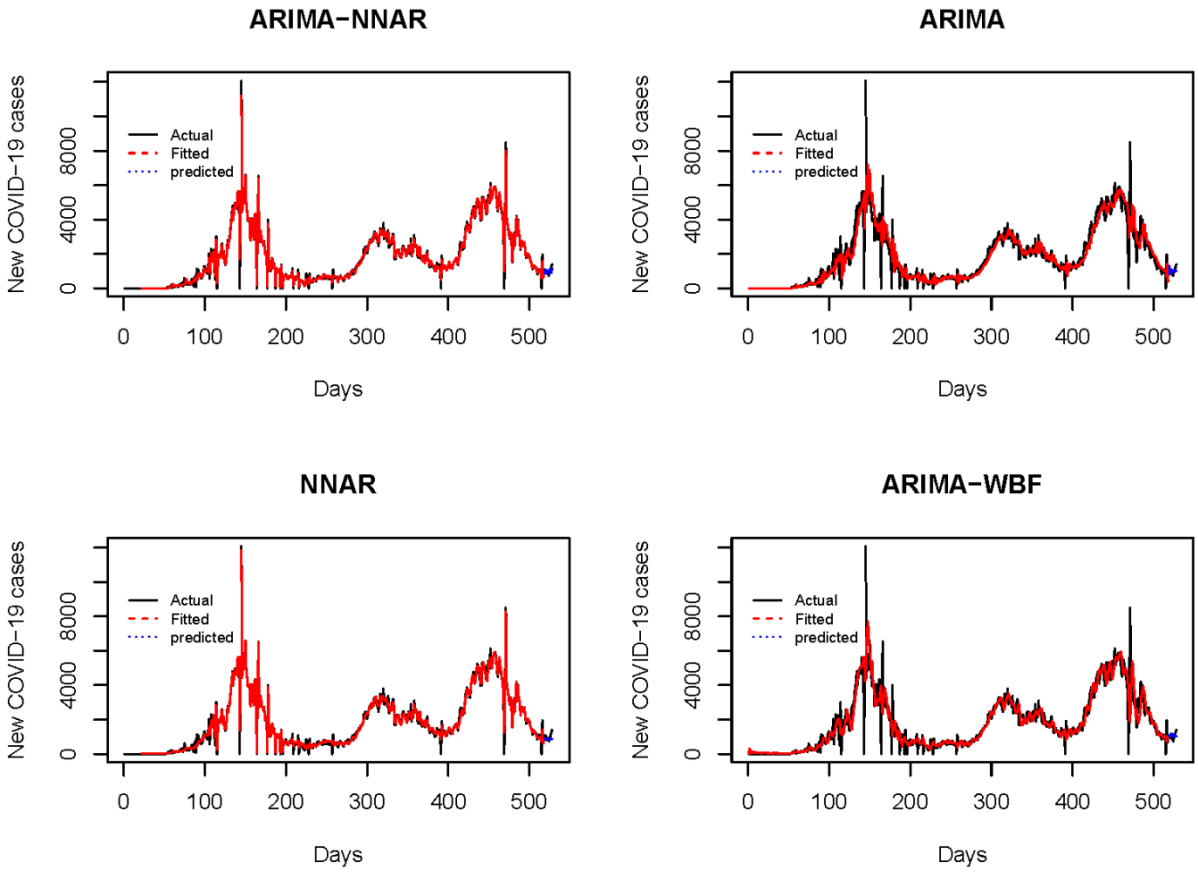


Figure 5. Actual and fitted values along with ten days ahead forecasts based on hybrid ARIMA-NNAR model for Pakistan

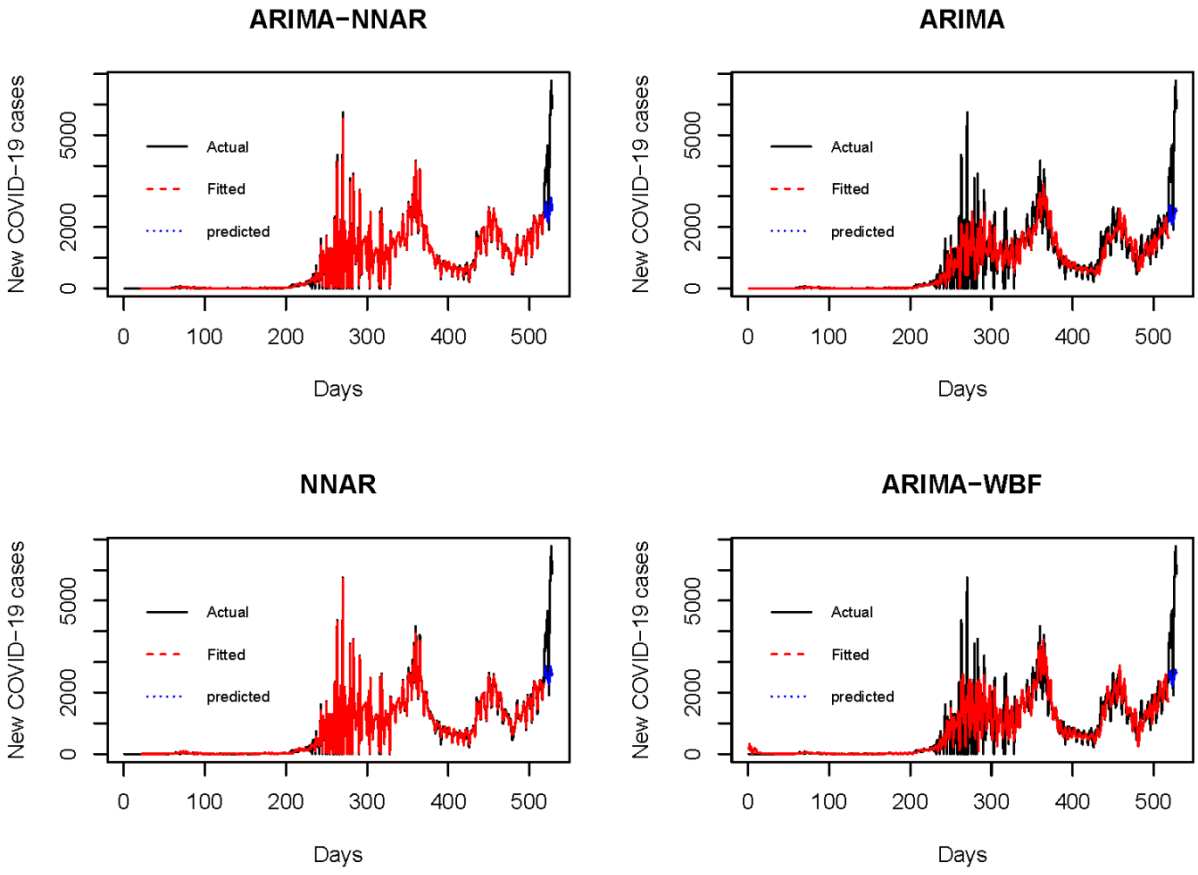


Figure 6. Actual and Fitted Values along with Ten Days Ahead Forecasts Based On Hybrid ARIMA-NNAR Model and Benchmark Models for Tunisia

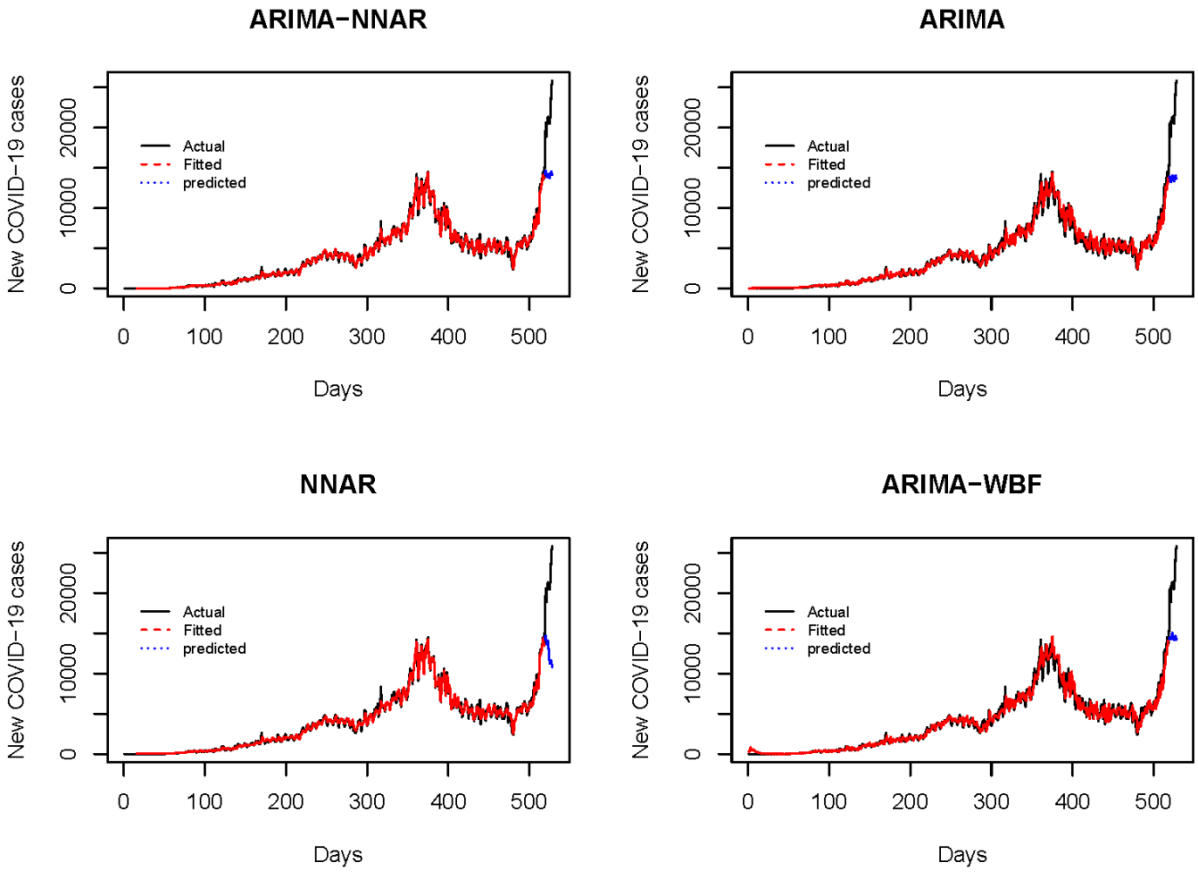


Figure 7. Actual and Fitted Values along with Ten Days Ahead Forecasts Based On Hybrid ARIMA-NNAR Model and Benchmark Models for Indonesia

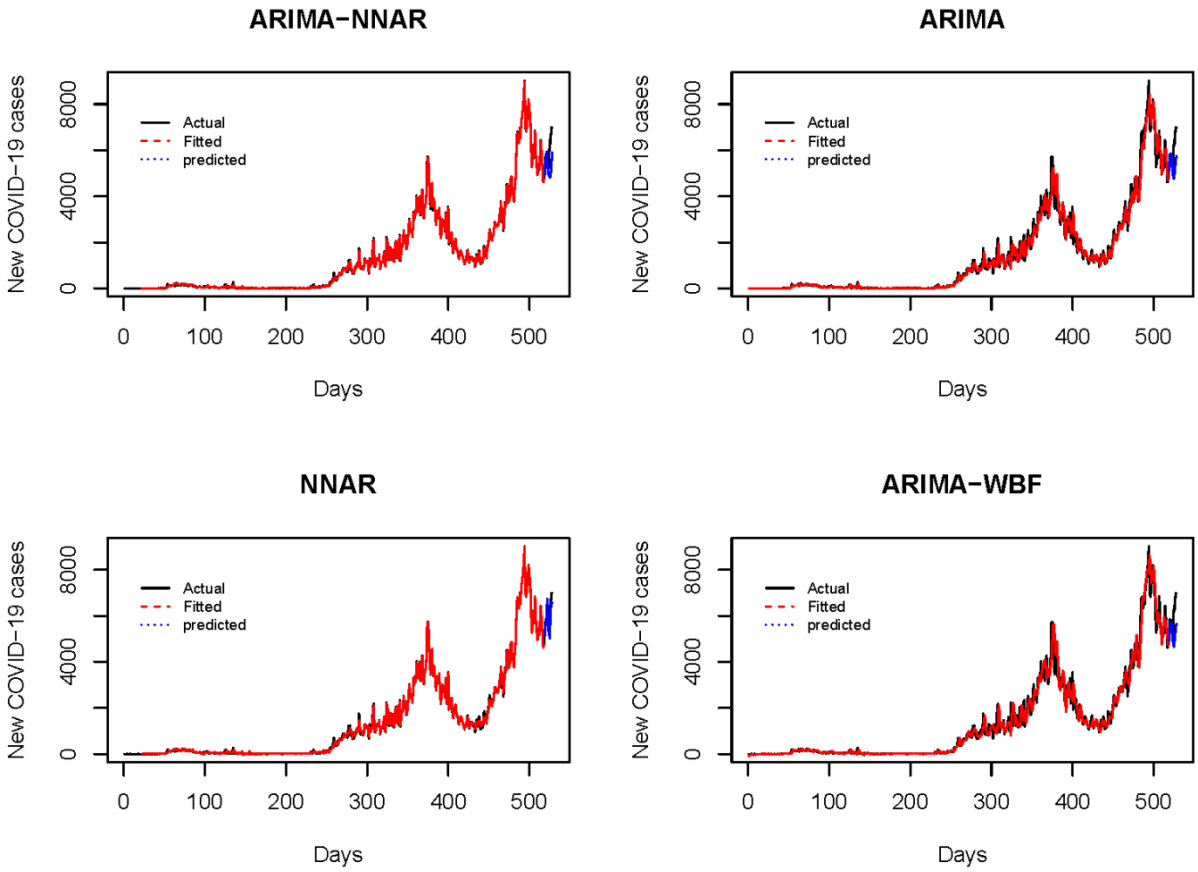


Figure 8. Actual and Fitted Values along with Ten Days Ahead Forecasts Based On Hybrid ARIMA-NNAR Model and Benchmark Models for Malaysia

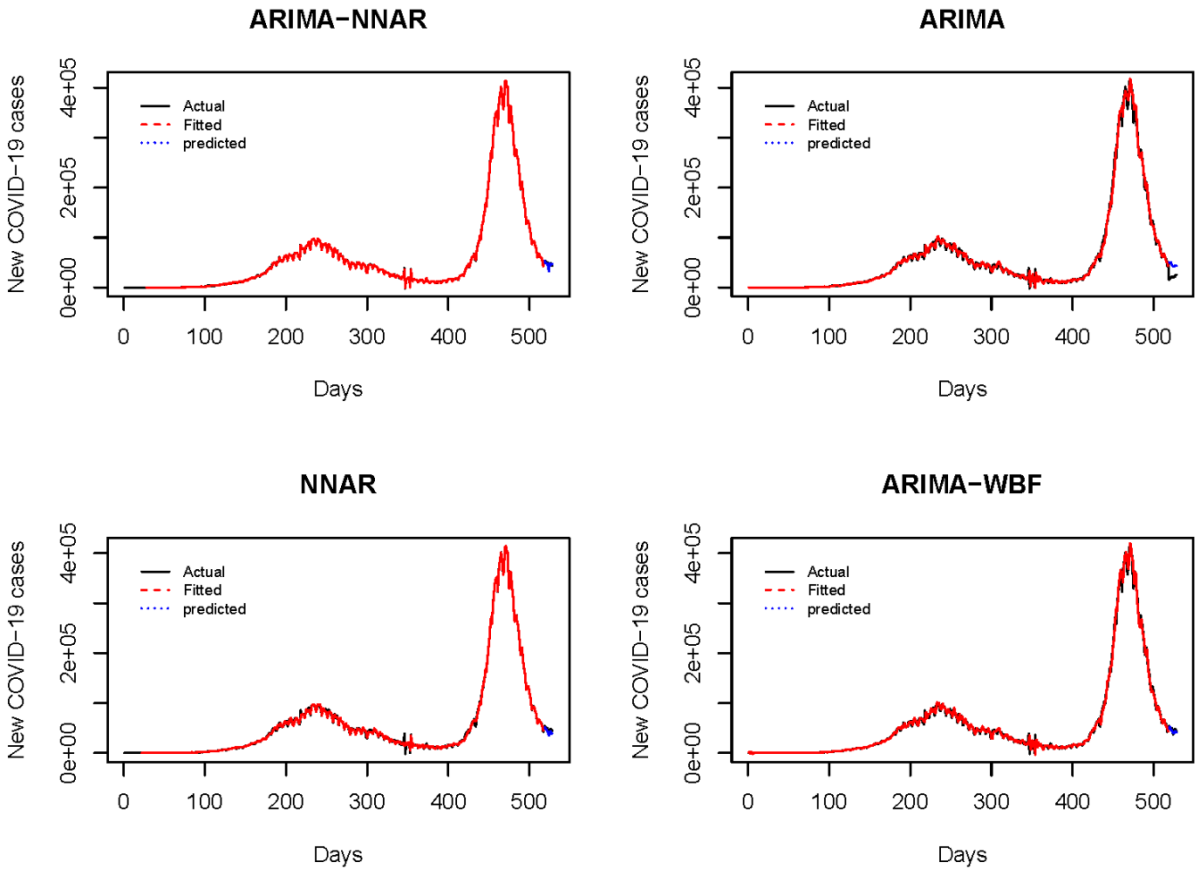


Figure 9. Actual and Fitted Values along with Ten Days Ahead Forecasts Based On Hybrid ARIMA-NNAR Model and Benchmark Models for India

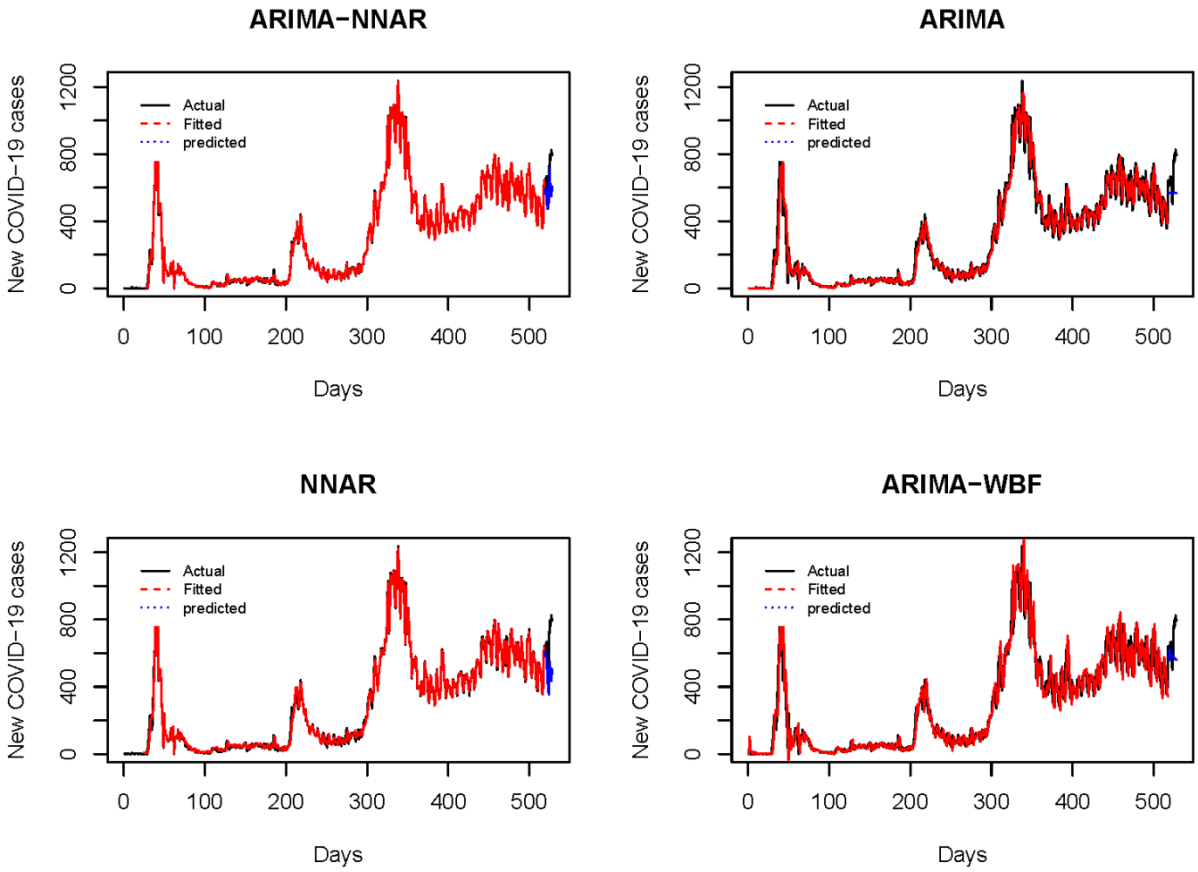


Figure 10. Actual and Fitted Values along with Ten Days Ahead Forecasts Based On Hybrid ARIMA-NNAR Model and Benchmark Models for South Korea

Table 2

RMSE and MAE Values for Different Forecasting Models on Six Time Series Data Sets

Country	Performancemetrics	ARIMA	NNAR	ARIMA-WBF	ARIMA-NNAR
Pakistan	RMSE	712.62	183.20	797.80	152.96
	MAE	394.33	114.52	401.91	87.82
Tunisia	RMSE	463.53	54.36	497.17	42.61
	MAE	243.27	32.06	256.94	24.54
Indonesia	RMSE	664.15	306.05	665.10	249.03
	MAE	420.82	207.97	431.15	167.68
Malaysia	RMSE	319.99	58.16	347.86	37.72
	MAE	171.67	34.53	189.79	24.37
India	RMSE	6350.96	2591.56	6487.26	411.43
	MAE	3448.24	1412.74	3631.24	228.99
South Korea	RMSE	72.08	18.59	80.74	7.00
	MAE	44.40	11.84	50.19	3.70

Table 3

RMSE, MAE, and MAPE Values related to the 10 Days ahead Forecasting of Six Time Series Data Sets

Country	Performance metrics	ARIMA	NNAR	ARIMA-WBF	ARIMA-NNAR
Pakistan	RMSE	213.48	244.94	216.04	182.80
	MAE	180.43	189.98	171.43	149.06
	MAPE	17.72	16.95	15.67	14.14
Tunisia	RMSE	2483.91	2340.54	2371.77	2333.16
	MAE	2156.36	2010.94	2058.16	2048.26
	MAPE	89.41	77.79	80.59	41.94
Indonesia	RMSE	7892.12	9123.16	7161.18	7467.41
	MAE	7402.73	8276.82	6620.70	6949.17
	MAPE	33.91	37.52	45.77	31.72
Malaysia	RMSE	919.51	588.36	989.33	977.88
	MAE	688.04	419.31	777.48	743.18
	MAPE	13.22	7.23	15.18	11.65
India	RMSE	3281.55	4672.74	3671.57	2747.55
	MAE	3010.74	4350.81	3373.23	2370.85
	MAPE	6.49	9.26	7.74	5.24
South Korea	RMSE	152.21	215.54	150.99	139.60
	MAE	125.93	180.92	117.87	114.95
	MAPE	17.23	25.39	20.75	15.93

Conclusion

Time-series forecasting models play an important role in estimating the spread of the COVID-19 pandemic. Accurate predictions are critical for implementing informed decisions aimed at fighting this pandemic and controlling its spread. Given the complex nature of new confirmed cases of COVID-19, it is strongly believed that such data is a combination of linear and nonlinear patterns. In literature, different forecasting models have been used to forecast future cases of COVID-19 (Moslehpour et al., 2022). However, most of them just used single models that cannot capture the linear and nonlinear patterns of the data simultaneously. This paper proposes a hybrid model, called ARIMA-NNAR, that combines Autoregressive Integrated Moving Average and Autoregressive Neural Network with a single hidden layer model. ARIMA models were used to capture the linear patterns from the series, and NNAR fitted their nonlinear patterns corresponding to the ARIMA residuals. The global prediction is obtained by adding up the two predictions**.** For in-depth details, a general code describing the main steps of forecasting time series with ARIMA-NNAR model is given in Figure 4.

In order to evaluate the hybrid ARIMA-NNAR model, this paper has used the new confirmed cases of COVID19 for Pakistan, Tunisia, Indonesia, Malaysia, India and South Korea, in which the last ten observations are kept for testing purposes. According to the experiments, ARIMA-NNAR provides more accurate results than ARIMA, NNAR and ARIMA-WBF models for training and testing phases. More concisely, tables 2 and 3 show how the accuracy measures are minimal compared to the other stipulated models, which confirm the ability of ARIMA-NNAR model to encapsulate the linear and nonlinear components of COVID-19 data sets. However, this model might be subject to overfitting. For example, though ARIMA-NNAR fits precisely against its training data for Tunisia and Indonesia, it cannot perform accurately against testing data. This could be explained by failing ARIMA-NNAR to capture such series' structural changes. Changing the number of hidden layers or changing the function that links linear and nonlinear components could solve this limitation. We will consider these suggestions in future work.

It is worth noting that this study brings many implications. It confirms the efficacy of ARIMA-NNAR model over ARIMA, NNAR, and ARIMA-WBF in fitting linear and nonlinear components of COVID-19 datasets. Forecasting COVID-19 datasets with ARIMA-NNAR model anticipates the real damage caused by this pandemic in different sectors, which allow policymakers to implement the proper policies that reduce the anticipated damage. More importantly, this study shares the code of modeling COVID-19 datasets with ARIMA-NNAR model, allowing the user to forecast the updated data for different countries.

References

Abbasimehr, H., Paki, R., & Bahrini, A. (2021). A novel approach based on combining deep learning models with statistical methods for COVID-19 time series forecasting. *Neural Computing and Applications*, *34*(4), 3135–3149.https://doi.org/10.1007/s00521-021-06548-9

Adhikari, R., & Agrawal, R. K. (2013). A combination of artificial neural network and random walk models for financial time series forecasting. *Neural Computing and Applications*, *24*(6), 1441–1449. https://doi.org/10.1007/s00521-013-1386-y

Agyei, S. K., Isshaq, Z., Frimpong, S., Adam, A. M., Bossman, A., & Asiamah, O. (2021). COVID-19 and food prices in sub-Saharan Africa. *African Development Review*, *33*(S1).https://doi.org/10.1111/1467-8268.12525

Anderson, O. D., Granger, C. W. J., & Andersen, A. P. (1979).An Introduction to Bilinear Time Series Models.*Applied Statistics*, *28*(3), 305.https://doi.org/10.2307/2347208

Aras, S., & Kocakoç, P. D. (2016).A new model selection strategy in time series forecasting with artificial neural networks: IHTS. *Neurocomputing*, *174*, 974–987. https://doi.org/10.1016/j.neucom.2015.10.036

Babu, C. N., & Reddy, B. E. (2014). A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data. *Applied Soft Computing*, *23*, 27–38.https://doi.org/10.1016/j.asoc.2014.05.028

Büyükşahin, M. A., & Ertekin, E. (2019).Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition.*Neurocomputing*, *361*, 151–163. https://doi.org/10.1016/j.neucom.2019.05.099

Callen, J. L., Kwan, C. C., Yip, P. C., & Yuan, Y. (1996).Neural network forecasting of quarterly accounting earnings. *International Journal of Forecasting*, *12*(4), 475–482. https://doi.org/10.1016/s0169-2070(96)00706-6

Casey Brace, M., Schmidt, J., & Hadlin, M. (1991). Comparison of the forecasting accuracy of neural networks with other established techniques. *Proceedings of the First International Forum on Applications of Neural Networks to Power Systems*. https://doi.org/10.1109/ann.1991.213493

Chakraborty, T., & Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons& Fractals*, *135*, 109850. https://doi.org/10.1016/j.chaos.2020.109850

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. https://doi.org/10.1016/0169-2070(89)90012-5

Denton, J. W. (1995), "How good are neural networks for causal forecasting?"*The Journal of Business Forecasting*, 14, 17.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation.*Econometrica*, *50*(4), 987.https://doi.org/10.2307/1912773

Foster, W., Collopy, F., & Ungar, L. (1992). Neural network forecasting of short, noisy time series.*Computers & Chemical Engineering*, *16*(4), 293–297. https://doi.org/10.1016/0098-1354(92)80049-f

Hann, T. H., & Steurer, E. (1996).Much ado about nothing? Exchange rate forecasting: Neural networks vs. linear models using monthly and weekly data. *Neurocomputing*, *10*(4), 323–339. https://doi.org/10.1016/0925-2312(95)00137-9

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., . . . Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497–506. https://doi.org/10.1016/s0140-6736(20)30183-5

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). Otexts.

Hyndman, R. J., &Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, *27*(3).https://doi.org/10.18637/jss.v027.i03

Khashei, M., & Bijari, M. (2011).A novel hybridization of artificial neural networks and ARIMA models for time series forecasting.*Applied Soft Computing*, *11*(2), 2664–2675.https://doi.org/10.1016/j.asoc.2010.10.015

Kırbaş, S., Sözen, A., Tuncer, A. D., & Kazancıoğlu, F. I. (2020).Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. *Chaos, Solitons& Fractals*, *138*, 110015.https://doi.org/10.1016/j.chaos.2020.110015

Moftakhar, L., Seif, M., & Safe, M. S. (2020). Exponentially Increasing Trend of Infected Patients with COVID-19 in Iran: A Comparison of Neural Network and ARIMA Forecasting Models. *Iranian Journal of Public Health*. https://doi.org/10.18502/ijph.v49is1.3675

Moslehpour, M., Al-Fadly, A., Ehsanullah, S. *et al.* (2022). Assessing Financial Risk Spillover and Panic Impact of Covid-19 on European and Vietnam Stock market. *Environ Sci Pollut Res*. https://doi.org/10.1007/s11356-021-18170-2

Ömer Faruk, D. (2010). A hybrid neural network and ARIMA model for water quality time series prediction. *Engineering Applications of Artificial Intelligence*, *23*(4), 586–594 .https://doi.org/10.1016/j.engappai.2009.09.015

Owusu Junior, P., Frimpong, S., Adam, A. M., Agyei, S. K., Gyamfi, E. N., Agyapong, D., & Tweneboah, G. (2021). COVID-19 as Information Transmitter to Global Equity Markets: Evidence from CEEMDAN-Based Transfer Entropy Approach. *Mathematical Problems in Engineering*, *2021*, 1–19. https://doi.org/10.1155/2021/8258778

Ravazzolo, F., Casarin, R., Corradin, F., & Sartore, D. (2020). A scoring rule for factor and autoregressive models under misspecification. Advances in Decision Sciences. Asia University, Taiwan. 24(2), 66-103. https://doi.org/10.47654/v24y2020i2p66-103

Safi, S. K., Sanusi, O. I., & Tabash, M. I. (2022). Forecasting the Impact of COVID-19 Epidemic on China Exports using Different Time Series Models. Advances in Decision Sciences. Asia University, Taiwan. 26(1), 102-127. https://doi.org/10.47654/v26y2022i1p102-127

Tajmouati, S., Wahbi, B. E., Bedoui, A., Abarda, A., & Dakkoun, M. (2021). Applying k-nearest neighbors to time series forecasting: two new approaches. arXiv preprint arXiv:2103.14200. https://arxiv.org/abs/2103.14200

Tajmouati, S., Wahbi, B. W. B., Bedoui, A., ABARDA, A., & Dakkoun, M. (2020).Une nouvelle approche d'application de l'algorithme K-NN pour la prévision du nombre total des cas confirmés de COVID-19. *Journal of Integrated Studies In Economics, Law, Technical Sciences & Communication*, 2(1). https://revues.imist.ma/index.php/JISELSC/article/view/24760

Tong, H. (2011). Threshold models in time series analysis — 30 years on. *Statistics and Its Interface*, *4*(2), 107–118. https://doi.org/10.4310/sii.2011.v4.n2.a1

Torrealba-Rodriguez, O., Conde-Gutiérrez, R., & Hernández-Javier, A. (2020). Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models. *Chaos, Solitons & Fractals*, *138*, 109946.https://doi.org/10.1016/j.chaos.2020.109946

Tran, T., Pham, L., and Ngo, Q. (2020), "Forecasting epidemic spread of SARS-CoV-2 using ARIMA model (Case study: Iran),"*Global Journal of Environmental Science and Management*, 6, 1–10. 10.22034/GJESM.2019.06.SI.01

Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175. https://doi.org/10.1016/s0925-2312(01)00702-0

Ziegel, E. R., Box, G., Jenkins, G., & Reinsel, G. (1995). Time Series Analysis, Forecasting, and Control. *Technometrics*, *37*(2), 238.https://doi.org/10.2307/1269640

Appendix

A. 1

The following table generates the results of the Box-Pierce test applied to the residuals of ARIMA-NNAR model $(i.e., r_t)$ for Pakistan, Tunisia, Indonesia, Malaysia, India and South Korea. Formally, the Box-Pierce test tests the hypothesis: H_0: $\rho_1 = \rho_2 = \cdots = \rho_m = 0$ against H_1: $\exists \rho_i \neq 0$, where ρ_i is the simple autocorrelation function of order i. The statistic used in the test is $Q = n.\sum_{j=1}^{m} \hat{\rho}_j^2$. In this study, we put m = 24.

Table A.4

Results of Box-Pierce Test related to the Residuals r_t for Pakistan, Tunisia, Indonesia, Malaysia, India and South Korea

Country	*Q*	*p-value*
Pakistan	16.92	0.85
Tunisia	31.35	0.14
Indonesia	22.50	0.55
Malaysia	14.33	0.94
India	17.32	0.83
SouthKorea	16.90	0.86

A. 2

Table A.5

Actual Values and Prediction of Daily Confirmed Cases using ARIMA-NNAR and Benchmark Models for Pakistan

Days	Actual values	ARIMA-NNAR	ARIMA	NNAR	ARIMA-WBF
23/06/2021	1097	862	1298	925	1049
24/06/2021	1042	1089	822	856	1099
25/06/2021	935	860	1083	805	1108
26/06/2021	901	1044	1077	864	1194
27/06/2021	914	969	958	836	1003
28/06/2021	735	922	1068	868	1127
29/06/2021	979	1105	1023	916	1056
30/06/2021	1037	1028	1012	904	1044
01/07/2021	1277	905	1044	836	1065
02/07/2021	1400	1158	1020	873	1037

Table A.6

Actual Values and Prediction of Daily Confirmed Cases using ARIMA-NNAR and Benchmark Models for Tunisia

Days	Actual values	ARIMA-NNAR	ARIMA	NNAR	ARIMA-WBF
23/06/2021	3638	2328	2189	2548	2497
24/06/2021	3951	2820	2712	2879	2643
25/06/2021	3467	2639	2675	2577	2709
26/06/2021	4664	2301	2250	2439	2391
27/06/2021	3524	2261	2082	2397	2236
28/06/2021	1914	2352	2184	2283	2325
29/06/2021	5251	2442	2287	2371	2430
30/06/2021	5921	2833	2465	2858	2611
01/07/2021	6776	2956	2603	2698	2737
02/07/2021	5882	2549	2519	2567	2649

Table A.7

Actual Values and Prediction of Daily Confirmed Cases using ARIMA-NNAR and Benchmark Models for Indonesia

Days	Actual values	ARIMA-NNAR	ARIMA	NNAR	ARIMA-WBF
23/06/2021	15308	14206	13914	14263	14270
24/06/2021	20574	14651	13822	14722	14512
25/06/2021	18872	14062	13230	13955	14299
26/06/2021	21095	13790	13350	13573	14366
27/06/2021	21342	14235	14053	14100	15091
28/06/2021	20694	13824	13699	12554	14615
29/06/2021	20467	13695	13322	11219	14152
30/06/2021	21807	14264	13718	11256	14456
01/07/2021	24836	14523	14055	11602	14656
02/07/2021	25830	14082	13634	10811	14201

Table A.8

Actual Values and Prediction of Daily Confirmed Cases using ARIMA-NNAR and Benchmark Models for Malaysia

Days	Actual values	ARIMA-NNAR	ARIMA	NNAR	ARIMA-WBF
23/06/2021	5244	4958	5122	5186	4917
24/06/2021	5841	5662	5626	5688	5497
25/06/2021	5812	5922	5859	5758	5733
26/06/2021	5803	5886	5684	6737	5613
27/06/2021	5586	4968	5235	5908	5188
28/06/2021	5218	4955	4835	5435	4791
29/06/2021	6437	4811	4758	5015	4660
30/06/2021	6276	4887	5044	6072	4963
01/07/2021	6988	5183	5480	6584	5394
02/07/2021	6982	5909	5758	6556	5655

Table A.9

Actual Values and Prediction of Daily Confirmed Cases using ARIMA-NNAR and Benchmark Models for India

Days	Actual values	ARIMA-NNAR	ARIMA	NNAR	ARIMA-WBF
23/06/2021	54069	53316	50818	51947	49786
24/06/2021	51667	48234	49554	48280	52234
25/06/2021	48698	49506	50568	46704	46985
26/06/2021	50040	47439	46409	44443	46574
27/06/2021	46148	43385	43265	41137	42963
28/06/2021	37566	32341	42694	35136	40401
29/06/2021	45951	47926	42616	39697	40179
30/06/2021	48786	46095	43785	43990	43554
01/07/2021	46617	46239	44448	39447	43193
02/07/2021	44111	47192	43384	39363	40856

Table A.10

Actual Values and Prediction of Daily Confirmed Cases using ARIMA-NNAR and Benchmark Models for South Korea

Days	Actual values	ARIMA-NNAR	ARIMA	NNAR	ARIMA-WBF
23/06/2021	610	564	567	609	558
24/06/2021	634	595	567	583	623
25/06/2021	668	580	567	581	577
26/06/2021	614	474	567	484	594
27/06/2021	501	480	567	395	528
28/06/2021	595	522	567	355	579
29/06/2021	794	715	567	551	572
30/06/2021	762	563	567	492	568
01/07/2021	826	548	567	434	563
02/07/2021	793	606	567	505	560

A. 3

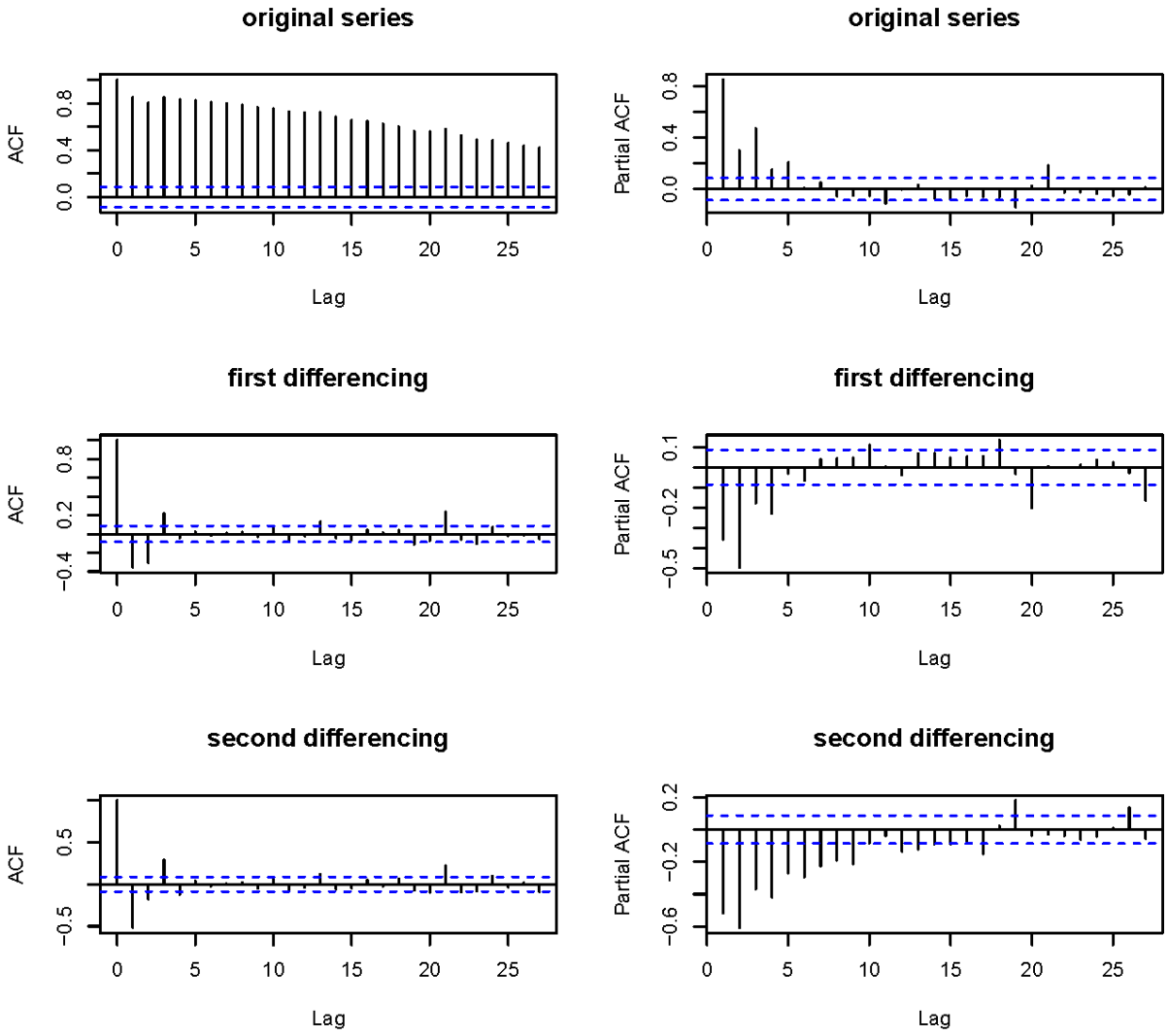


Figure A.10. ACF and PACF for Pakistan

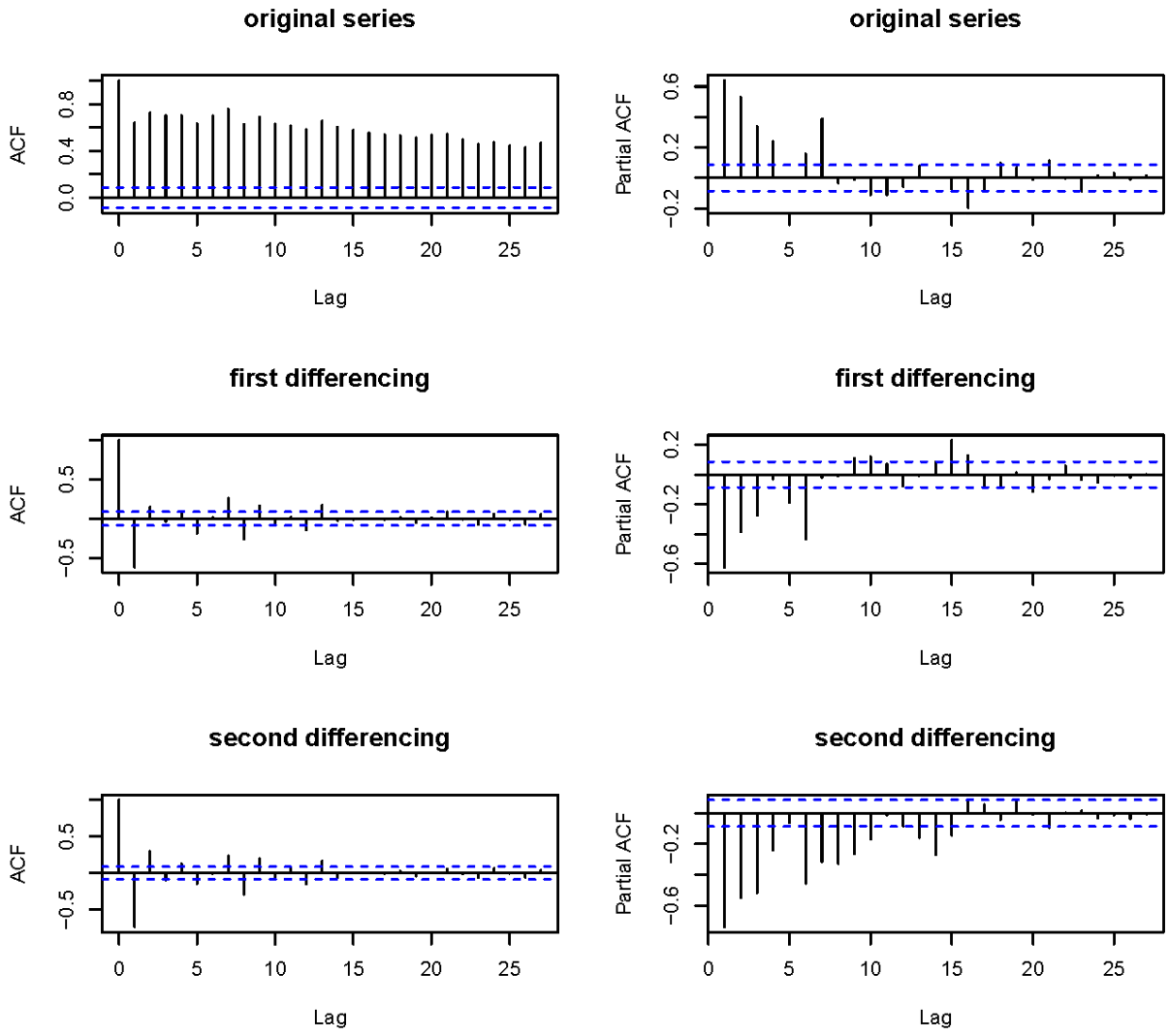


Figure A.11. ACF and PACF for Tunisia

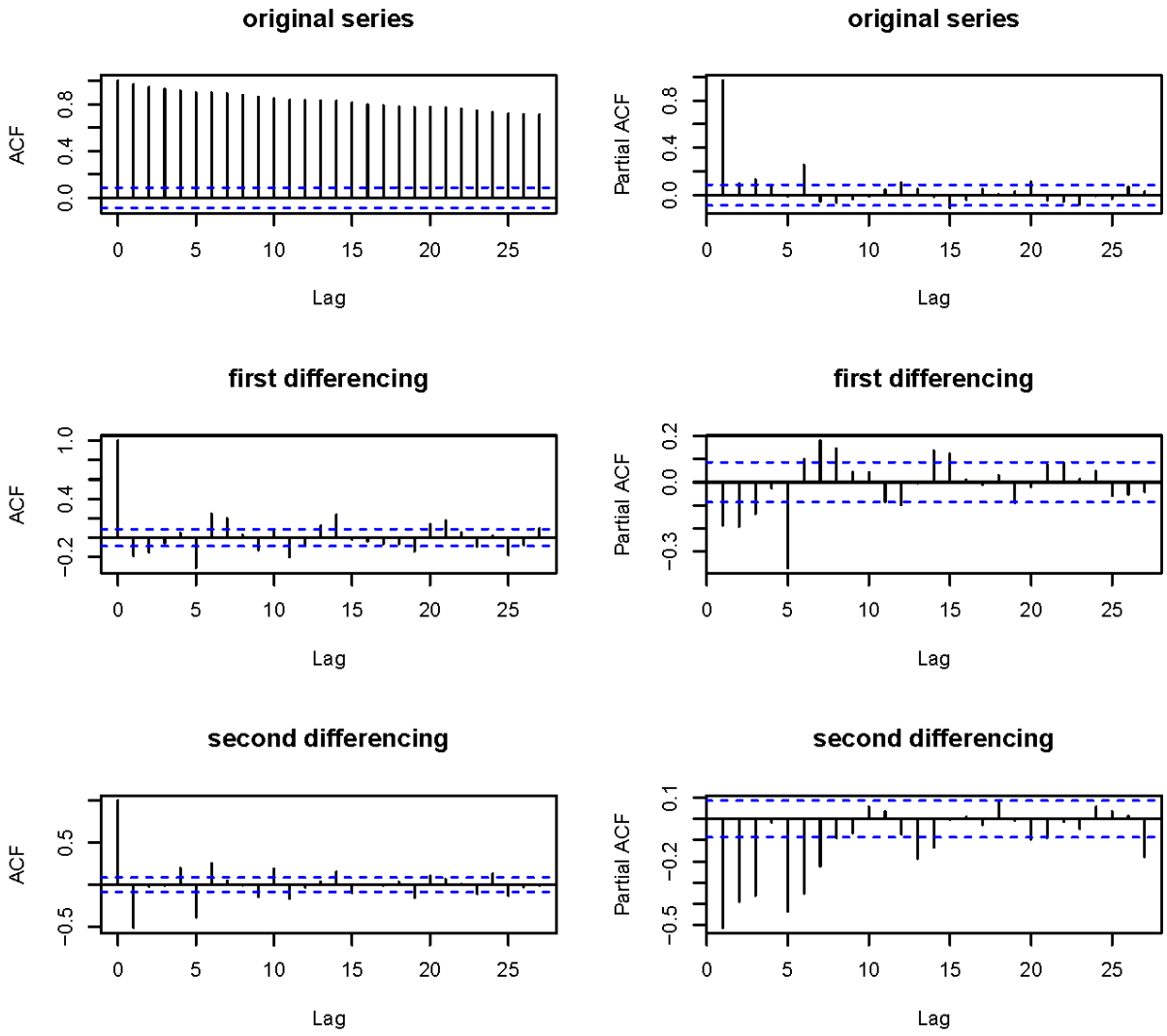


Figure A.12. ACF and PACF for Indonesia

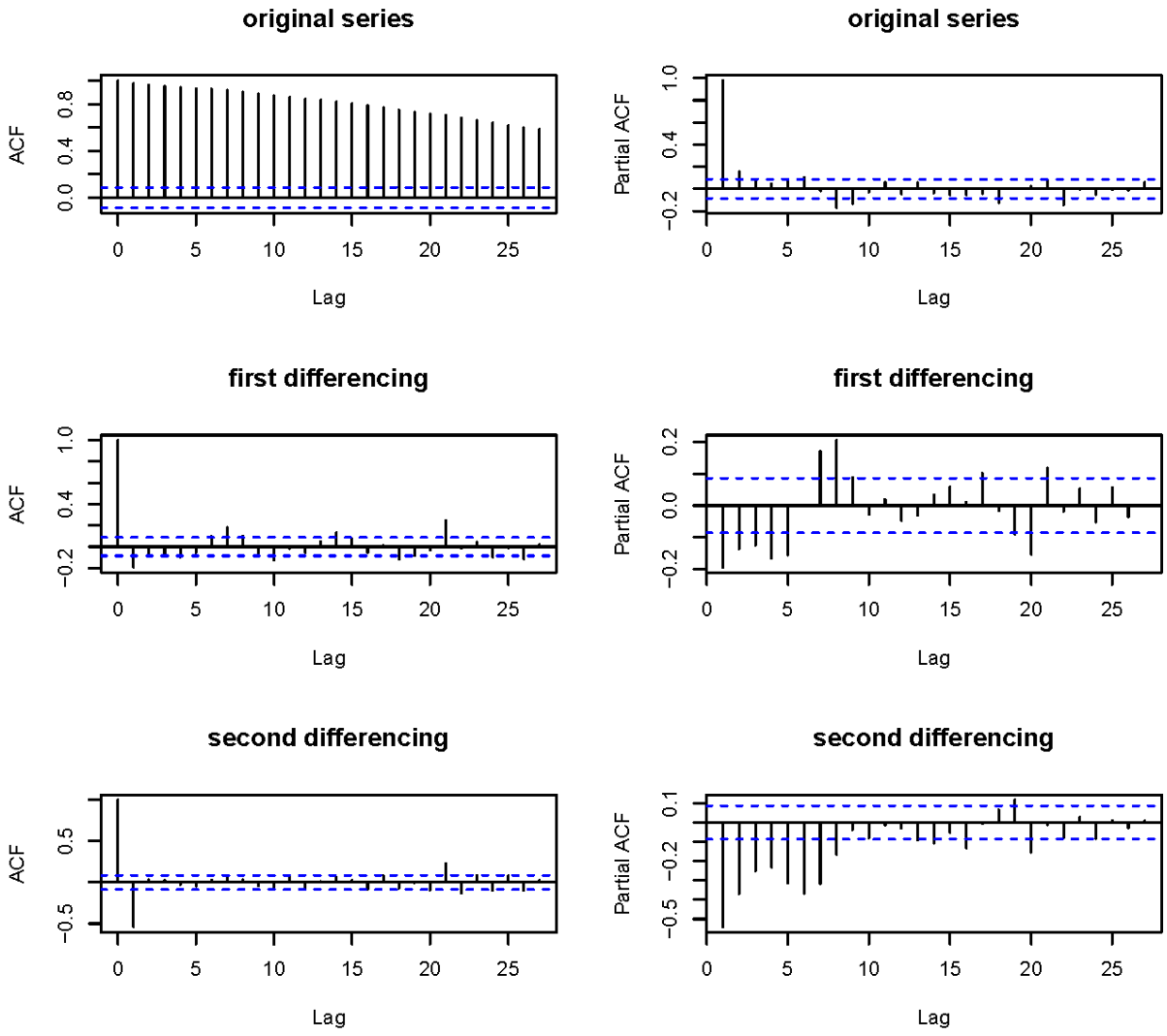


Figure A.13. ACF and PACF for Malaysia

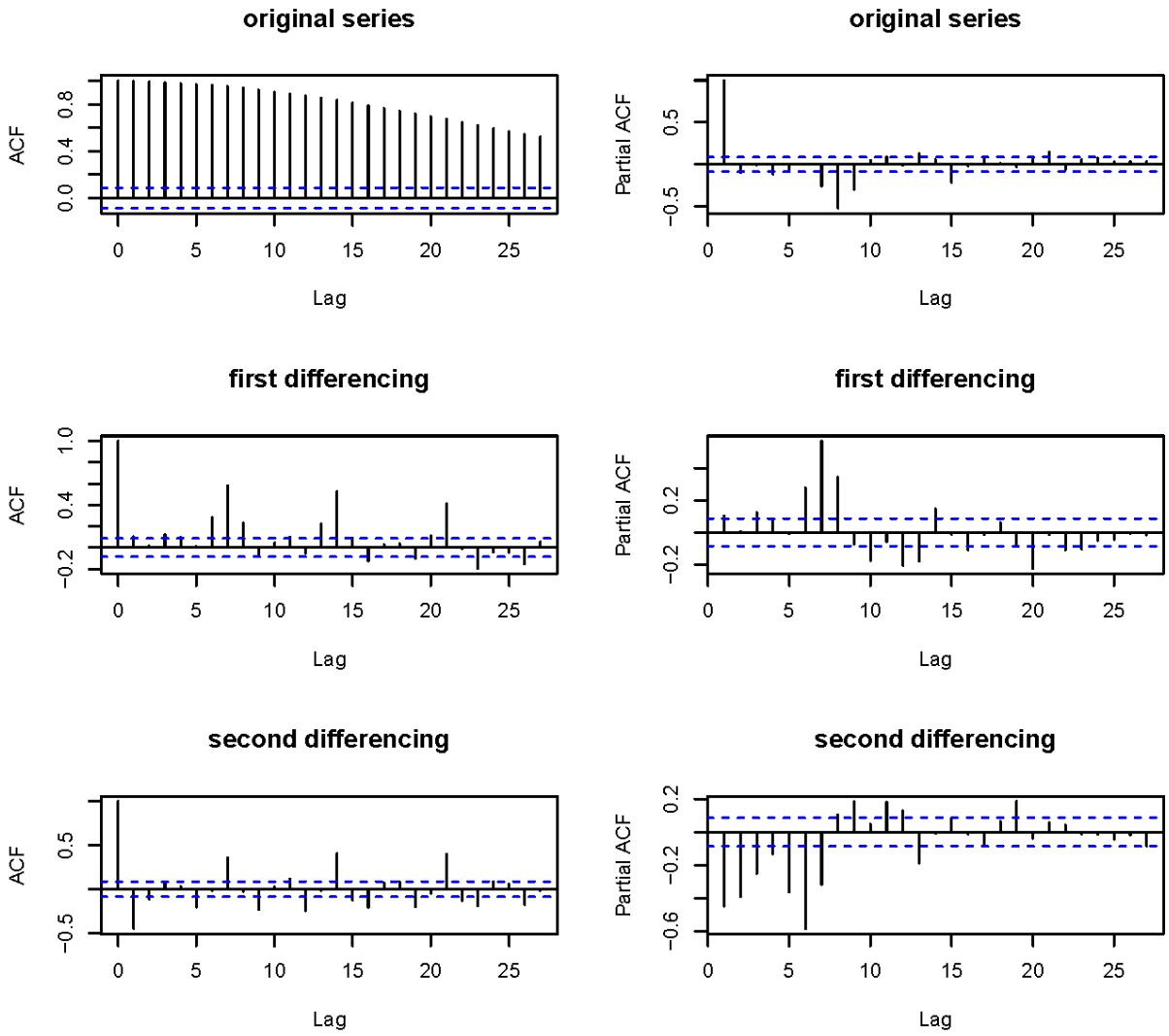


Figure A.14. ACF and PACF for India

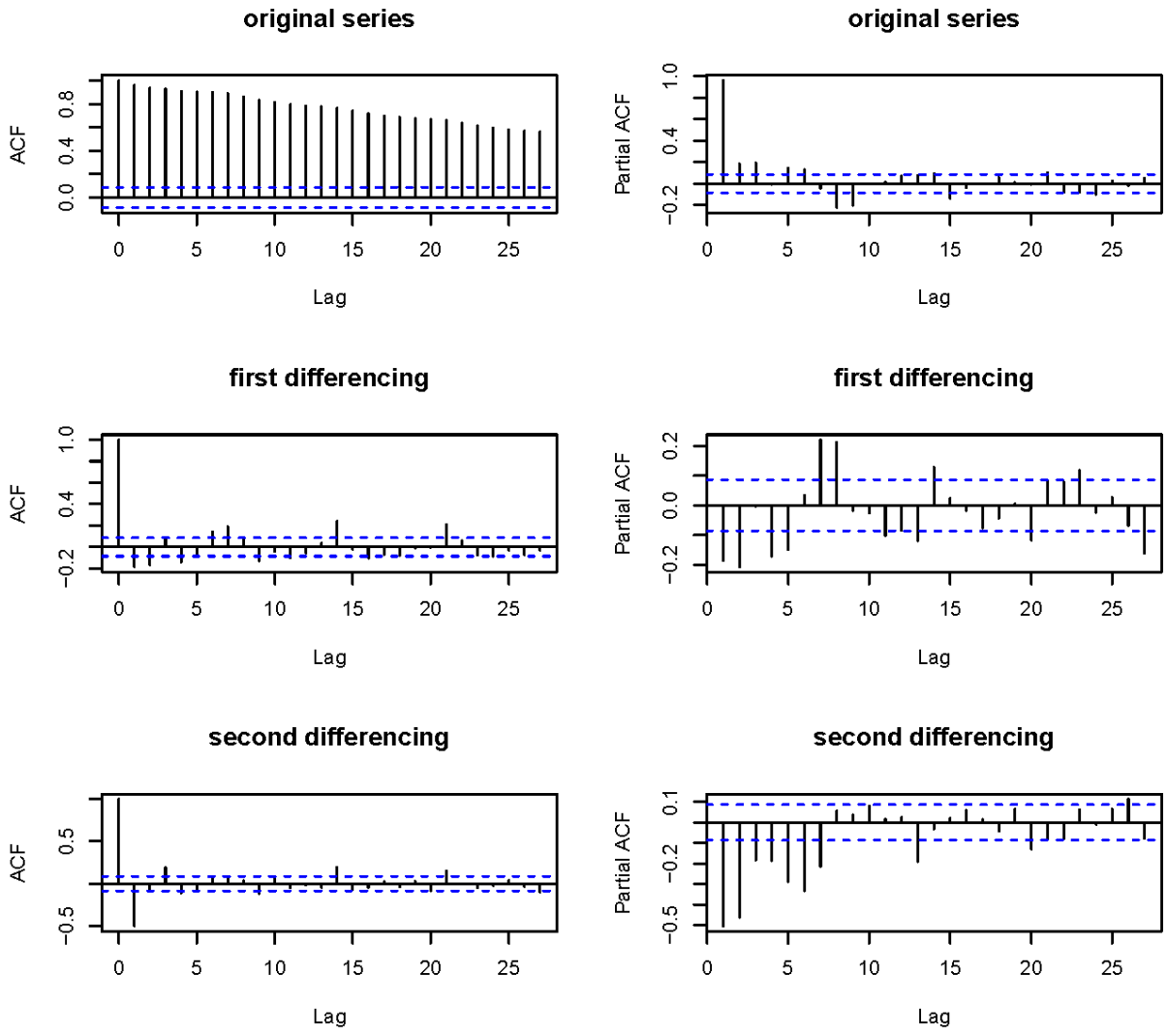


Figure A.15. ACF and PACF for South Korea