entropy

MDPI

Article

Network Autoregressive Model for the Prediction of COVID-19 Considering the Disease Interaction in Neighboring Countries

Arash Sioofy Khoojine [1], Mahdi Shadabfar [2,*], Vahid Reza Hosseini [3] and Hadi Kordestani [4]

[1] Faculty of Economics and Business Administration, Yibin University, Yibin 644000, China; arashsioofy@yibinu.edu.cn
[2] Center for Infrastructure Sustainability and Resilience Research, Department of Civil Engineering, Sharif University of Technology, Tehran 145888-9694, Iran
[3] Institute for Advanced Study, Nanchang University, Nanchang 330031, China; v.r.hosseini@ncu.edu.cn
[4] Department of Civil Engineering, Qingdao University of Technology, Qingdao 266033, China; hadi@qut.edu.cn
* Correspondence: mahdi.shadabfar@sharif.edu

Abstract: Predicting the way diseases spread in different societies has been thus far documented as one of the most important tools for control strategies and policy-making during a pandemic. This study is to propose a network autoregressive (NAR) model to forecast the number of total currently infected cases with coronavirus disease 2019 (COVID-19) in Iran until the end of December 2021 in view of the disease interactions within the neighboring countries in the region. For this purpose, the COVID-19 data were initially collected for seven regional nations, including Iran, Turkey, Iraq, Azerbaijan, Armenia, Afghanistan, and Pakistan. Thenceforth, a network was established over these countries, and the correlation of the disease data was calculated. Upon introducing the main structure of the NAR model, a mathematical platform was subsequently provided to further incorporate the correlation matrix into the prediction process. In addition, the maximum likelihood estimation (MLE) was utilized to determine the model parameters and optimize the forecasting accuracy. Thereafter, the number of infected cases up to December 2021 in Iran was predicted by importing the correlation matrix into the NAR model formed to observe the impact of the disease interactions in the neighboring countries. In addition, the autoregressive integrated moving average (ARIMA) was used as a benchmark to compare and validate the NAR model outcomes. The results reveal that COVID-19 data in Iran have passed the fifth peak and continue on a downward trend to bring the number of total currently infected cases below 480,000 by the end of 2021. Additionally, 20%, 50%, 80% and 95% quantiles are provided along with the point estimation to model the uncertainty in the forecast.

Keywords: COVID-19; Iran timeseries prediction; infected cases; ARIMA model; correlation matrix; network autoregressive (NAR) model

Citation: Sioofy Khoojine, A.; Shadabfar, M.; Hosseini, V.R.; Kordestani, H. Network Autoregressive Model for the Prediction of COVID-19 Considering the Disease Interaction in Neighboring Countries. *Entropy* **2021**, *23*, 1267. https://doi.org/10.3390/e23101267

Academic Editors: Rafał Rak and José A. Tenreiro Machado

Received: 8 August 2021
Accepted: 25 September 2021
Published: 28 September 2021

1. Introduction

SARS-Cov-2 (COVID-19) is on the rise and it is quickly infecting new people every day. Currently, two years after the onset of this pandemic, this ascending trend has not yet stopped and it is even multiplying in some countries [1]. When a person is determined to be infected with the disease in a country, there may be two possibilities where he is infected:

1. The first case concerns the situation where both carriers and recipients of the disease are in the same country. This type of disease transmission is considered "local";
2. The second is for cases infected in another country and transferred to a second country by travel. This type of disease transmission is called "imported".

Communication among nations is one of the main causes of disease transmission, and is called disease interaction between countries in this paper. In addition to disease progress in target communities when examining its spreading profile, it is also of the

utmost importance to reflect on its prevalence rate in other countries, including those with a high volume of travel [2]. The number of cases infected with this health condition can be thus deemed as a timeseries, taking account of the related statistics in the form of data over time [3].

In this regard, numerous researchers have thus far attempted to utilize a wide range of statistical tools to predict the number of cases of COVID-19 in the future to guide health care officials to make informed decisions [4]. For example, Shadabfar et al. used a susceptible–exposed–infected–vaccinated–recovered (SEIVR) model combined with the Monte Carlo (MC) sampling method to probabilistically investigate the COVID-19 spreading profile in the United States (USA) [5,6].

In general, different stochastic computations [7,8] and numerical methods [9–14] are exploited to assess the various aspects of the COVID-19 outbreak. In this sense, Katoch et al. used the autoregressive integrated moving average (ARIMA) model to forecast the COVID-19 dynamics in India [15]. Kumar Sahai et al. also modeled and predicted this pandemic via the ARIMA model [16]. Using the same ARIMA model, Malki et al. further predicted the second rebound of this disease; they also projected the end of the pandemic based on the ARIMA model [17]. Chaurasia et al. additionally used ARIMA and a regression model to forecast mortality rates in this respect [18]. Furthermore, Kumar et al. employed timeseries methods to analyze the COVID-19 spreading profile in ten affected countries [19]. Using the α-Sutte indicator and ARIMA, Attanayake et al. modeled COVID-19 [20]. Hernandez et al. correspondingly forecasted COVID-19 per region using the ARIMA model and polynomial functions [21]. Moreover, Yang et al. defined the data as timeseries and predicted the COVID-19 spreading profile in Wuhan, China [22].

Even though these studies have been to take advantage of different regression and optimization techniques to obtain the best fit of the data and consequently provide reliable timeseries forecasting, they typically suffer from one limitation, that is, their prediction remains independent of the disease spreading profile in other nations in the region. In fact, concerning the development trends of the disease interactions in neighboring countries, it seems ideal to measure the relationship between the disease spreading profiles in relevant nations to consider its impact on predicting the disease timeseries in target countries and regions.

To fill this gap, this paper utilizes a Network Autoregressive (NAR) Model. For this purpose, the COVID-19 data are initially retrieved from the World Health Organization (WHO) and the Johns Hopkins University online official websites and databases for seven different countries, namely, Iran, Turkey, Iraq, Azerbaijan, Armenia, Afghanistan, and Pakistan [23]. Thereafter, by constructing a network in the region, in which each vertex corresponds to a country and each edge represents the correlation of the total number of currently infected cases, the correlation matrix of the area is established. After that, the timeseries forecasting for Iran is performed using the NAR model, providing the number of infected cases up to December 2021. Comparing the root mean square error (RMSE) and mean absolute percentage error (MAPE) between autoregressive integrated moving average (ARIMA) and NAR models demonstrate that a better fit is obtained over the data once interactions among neighboring countries are taken into account. The method proposed in this paper can thus be implemented systematically to provide a reference for the investigation of the disease spreading profile in other countries and regions.

The rest of this study is organized as follows. Section 2 introduces the study area and then reviews the disease progression across the countries in the region concerned, from the onset of the COVID-19 pandemic in February 2020. Section 3 sheds light on the details of both methods implemented in this study, namely, the ARIMA and the NAR models, and subsequently describes how to consider the disease interactions in the neighboring nations in the proposed formulation. Next, in Section 4, the ARIMA and the NAR models are fitted to the existing data. In addition, upon comparing both methods, it is settled that the consideration of the disease interactions in the neighboring countries can enhance the prediction accuracy. Thus, the NAR model is employed to forecast the number of cases

infected in Iran until the end of December 2021, and the results are reported. Then, in Section 5, the criteria for choosing the threshold are clarified in more detail. Finally, the contents are summarized and concluded in Section 6.

2. Target Region and Data Description

To implement this work, the records of the COVID-19 data from the WHO and Johns Hopkins University official websites are used [24]. It should be noted that the data reported by the WHO contain some uncertainty and do not reflect the complete and accurate status of the disease in society [25,26]. However, the approach presented in the current research is implemented based on the disease statistics provided by the WHO as the reference dataset. The authors do not claim that the prediction made in the paper is the real state of the disease in society but acknowledge that it will be the disease's future according to WHO data. The data show confirmed cases, daily recovery, and death rates. The total of currently infected patients is accordingly calculated as follows:

$$\text{Total Currently Infected} = \text{Total Confirmed} - \text{Total Recovered} - \text{Total Death}. \quad (1)$$

As mentioned earlier, the primary purpose of this study is to address the impact of COVID-19 interactions in the neighboring countries on the timeseries forecasting model of the number of cases infected in Iran. As a result, some neighboring nations, including Turkey, Iraq, Azerbaijan, Armenia, Afghanistan, and Pakistan, are considered the target region here. The COVID-19 data from Turkmenistan are not publicly available, so they are not reflected in this study. A comparison of the geographic locations of these countries with Iran is further depicted in Figure 1. The timeseries of the rate of infected cases and infected cases in these nations as of 10 September 2021 are shown in Figure 2. A closer look at Figure 2 also reveals that different countries have so far experienced similar trends of this condition at the same time, which reinforces the hypothesis that the nations located in this region interact with the spread of the disease. For example, Iran and Turkey simultaneously experienced three peaks in March 2020, December 2020, and April 2021.

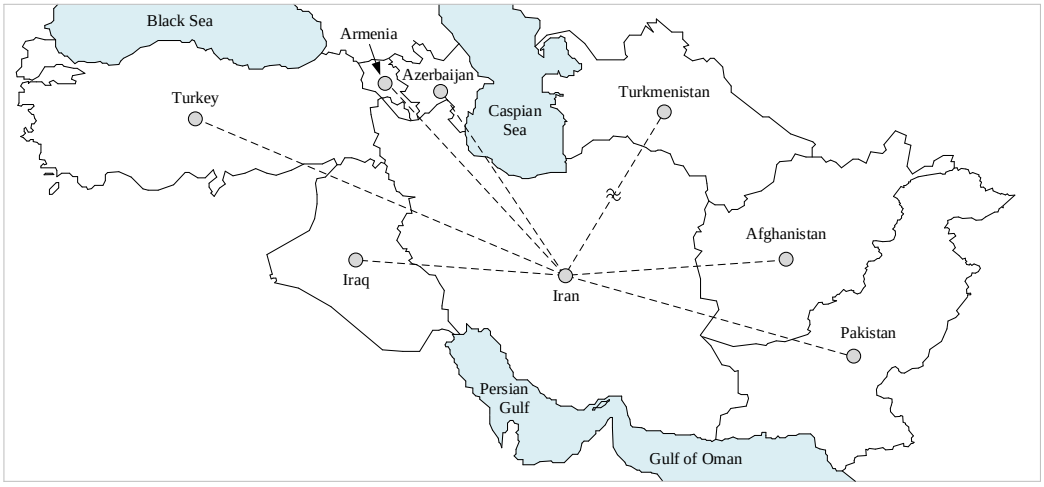

Figure 1. The region of interest investigated in the present study.

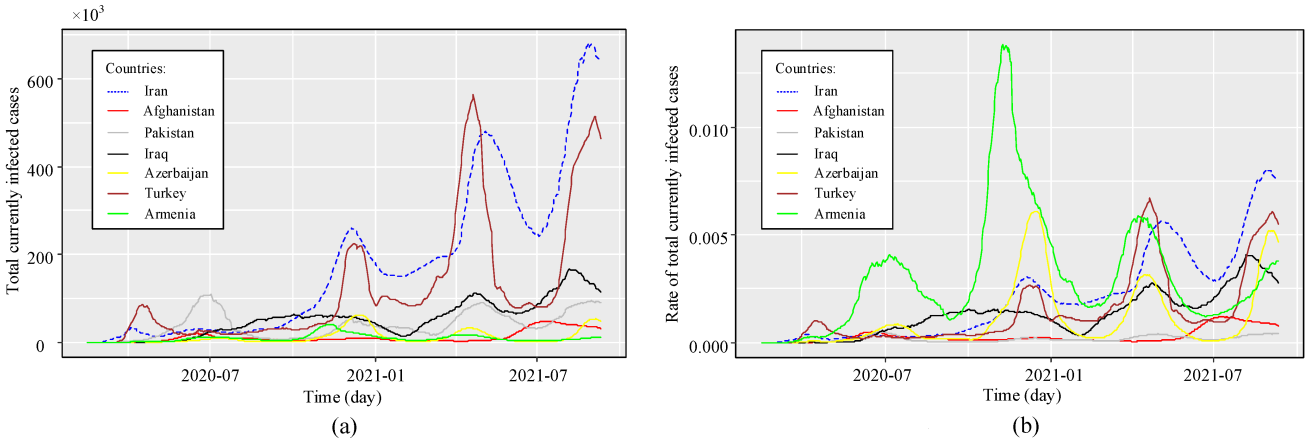

Figure 2. WHO COVID-19 data for seven different countries; (**a**) total currently infected cases, (**b**) rate of total currently infected cases.

3. Model Formulation

3.1. NARI Model

In this section, the NAR model for total infected people, hereafter referred to as NARI, is provided, and its characteristics are explained. First, the data on the infected are transformed to a new timeseries, after that the correlations between countries are calculated and a network is created over the countries in which each edge denotes the correlation between a pair of countries. This network is then introduced into the NARI model to predict the time history of the target country given the correlation values. Assuming N is the number of countries, the difference of logs of the total infected people, s_{it}, is defined as follows:

$$s_{it} = \delta \log\left(\frac{\text{infected } (i,t)}{\text{infected } (i,t-1)}\right), \tag{2}$$

where δ is a constant, which is a hyper parameter in the model. Using trial and error in the countries concerned, the optimal value of δ is computed as 0.5. Next, expanding this equation gives:

$$s_{it} = \frac{1}{2}\log\left(\frac{\text{infected}(i,t)}{\text{infected}(i,t-1)}\right) = \log(\sqrt{\text{infected}(i,t)}) - \log(\sqrt{\text{infected}(i,t-1)}), \tag{3}$$

wherein $\text{infected}(i,t)$ refers to the number of total infected cases from country i at time t. The NARI model is formulated as follows:

$$s_{it} = \alpha_0 + \alpha_1 n_i^{-1} \sum_{j=1}^{N} a_{ij} s_{j(t-1)} + \alpha_2 s_{i(t-1)} + \epsilon_{it}, \tag{4}$$

where

(i) N is the number of countries;
(ii) s_{it} represents the difference of logs of infected cases from country i at time t;
(iii) $\mathbf{A} = (a_{ij})_{N\times N}$ shows the adjacency matrix of the correlation between the log-returns of N countries;
(iv) n_i is the sum of the ith row at the adjacency matrix $\mathbf{A}$;
(v) ϵ_{it} follows the normal distribution.

The assumption of a normal distribution for the error term in Equation (4) has also been adopted in other studies such as [27]. The main idea behind assuming a normal distribution for the noise term is that with this assumption we have the smallest variance between all of the estimators. This assumption helps the algorithm to approximate the

MLE in a straightforward process and to facilitate the time prediction process. For more details, refer to [28].

The cross-correlations between the infected cases in different countries are similarly considered in terms of matrix **C**, whose elements are given through the following equation [29]:

$$c_{ij} \equiv \frac{\langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle}{\sqrt{\left(\langle s_i^2 \rangle - \langle s_i \rangle^2\right)\left(\langle s_j^2 \rangle - \langle s_j \rangle^2\right)}}, \tag{5}$$

where the brackets mean the temporal average over the infected cases during the time considered. Then, c_{ij} can vary between $[-1, 1]$. The case of $c_{ij} = 1(-1)$ also denotes that two countries i and j are completely correlated (anti-correlated), while $c_{ij} = 0$ implies that they are uncorrelated.

Suppose that $\mathbf{A} = (a_{ij})_{(N\times N)}$ is the adjacency matrix of the correlations among N countries; by adjusting a threshold as θ, $-1 \leq \theta \leq 1$, the matrix **A** is defined as follows [30]:

$$a_{ij} = \begin{array}{ll} 1 & if\ i \neq j \text{ and } c_{ij} \geq \theta \\ 0 & otherwise, \end{array} \tag{6}$$

and $n_i = \sum_{j=1}^{N} a_{ij}$ refers to the ith row sum of the adjacency matrix, and $E(\epsilon_{it}) = 0$ and $var(\epsilon_{it}) = \sigma^2$. This threshold value and its calculation details are described in Section 5. For convenience, Equation (4) can be rewritten in a matrix form as:

$$\begin{aligned}(s_{1t}, \ldots, s_{Nt})^\intercal = (1, \ldots, 1)^\intercal \alpha_0 + \alpha_1 \begin{pmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{n_N} \end{pmatrix} \mathbf{A} + \alpha_2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} \\ + (s_{1(t-1)}, \ldots, s_{N(t-1)}) + (\epsilon_{1t}, \ldots, \epsilon_{Nt})^\intercal,\end{aligned} \tag{7}$$

or in a concise form as:

$$\mathbb{S}_t = \mathbb{A}_0 + \mathbf{G}\mathbb{S}_{t-1} + \varepsilon_{St}, \tag{8}$$

in which $\mathbb{S}_t = (s_{1t}, ..., s_{Nt}) \in \mathbb{R}^N$, $\mathbb{A}_0 = \alpha\mathbf{1}$ wherein $\mathbf{1} = (1, ..., 1)^\intercal$, $\mathbf{G} = \alpha_1\mathbf{W} + \alpha_2\mathbf{I}$ in which $\mathbf{W} = diag\{n_1^{-1}, ..., n_N^{-1}\}\mathbf{A}$, and **I** is an identity matrix.

Under the NARI model framework, the model might be based on three factors; first, s_{it} might be affected by itself but from the previous time point, $s_{i(t-1)}$, called the autoregressive effect; second, s_{it} might be influenced by its neighbors, which are collected by $\{j : a_{ij} = 1\}$, labeled the "neighborhood effect". The unexplained variation should also be attributed to an independent random noise, ϵ_{it}. For example, for a country $i = 1$ at $t = 3$, s_{13} is as follows:

$$\begin{aligned} s_{13} &= \alpha_0 + \underbrace{\alpha_1 n_1^{-1} \sum_{j=1}^{N} a_{1j} s_{j2}}_{\text{neighborhoods effect}} + \overbrace{\alpha_2 s_{12}}^{\text{authoregresive effect}} + \underbrace{\epsilon_{13}}_{\text{independent noise}} \\ &= \alpha_0 + \alpha_1 \frac{a_{12}s_{22} + \cdots + a_{1N}s_{N2}}{a_{12} + \cdots + a_{1N}} + \alpha_2 s_{12} + \epsilon_{13}. \end{aligned} \tag{9}$$

Therefore, $\alpha_2 s_{i(t-1)}$ is not incorporated into the first term. It can be proved that $\mathbb{S}_t$ has a stationary property (for more details see [31]). To estimate $\alpha = (\alpha_0, \alpha_1, \alpha_2)$, maximum likelihood estimation (MLE) is also used as follows:

$$min_\alpha \|\mathbb{S}_t - \mathbb{A}_0 - \mathbf{G}\mathbb{S}_{t-1}\|. \tag{10}$$

For estimating the unknown parameter α, the NARI model (Equation (4)) is rewritten as:

$$s_{it} = \alpha_0 + \alpha_1 w_i^\top \mathbb{S}_{t-1} + \alpha_2 s_{i(t-1)} + \epsilon_{it}. \tag{11}$$

Then, it is written as:

$$s_{it} = Y_{i(t-1)}^\top \alpha + \epsilon_{it}, \tag{12}$$

where $Y_{i(t-1)} = (1, w_i^\top \mathbb{S}_{t-1}, s_{i(t-1)})^\top$ and $w_i = (a_{ij}/n_i : 1 \le j \le N)^\top$ indicate the ith row vector of W. Suppose $\mathbb{Y}_t = (Y_{1t}, ..., Y_{Nt})^\top$, then the above model can be written as:

$$\mathbb{S}_t = \mathbb{Y}_{t-1}\alpha + \varepsilon_t. \tag{13}$$

Then, a maximum likelihood (ML) estimator in the logarithmic form can be obtained as follows:

$$\mathcal{L}(\alpha, \sigma^2) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(\sum_{t=1}^{T}\mathbb{S}_t - \sum_{t=1}^{T}\mathbb{Y}_{t-1}\alpha)^\top(\sum_{t=1}^{T}\mathbb{S}_t - \sum_{t=1}^{T}\mathbb{Y}_{t-1}\alpha). \tag{14}$$

Differentiating this expression with respect to α, the ML estimates will be as follows:

$$\frac{\partial \mathcal{L}}{\partial \alpha^\top} = -\frac{1}{2\sigma^2}\Big(-2\sum_{t=1}^{T}\mathbb{Y}_{t-1}^\top \mathbb{S}_t + 2\sum_{t=1}^{T}\mathbb{Y}_{t-1}^\top \mathbb{Y}_{t-1}\alpha\Big) = 0. \tag{15}$$

Therefore,

$$\hat{\alpha} = (\sum_{t=1}^{T}\mathbb{Y}_{t-1}^\top \mathbb{Y}_{t-1})^{-1}\sum_{t=1}^{T}\mathbb{Y}_{t-1}^\top \mathbb{S}_t. \tag{16}$$

Substituting Equation (13) into the estimator $\hat{\alpha}$ in Equation (16), there is:

$$\hat{\alpha} = \alpha + (\sum_{t=1}^{T}Y_{t-1}^\top Y_{t-1})^{-1}\sum_{t=1}^{T}Y_{t-1}^\top \epsilon_{St}. \tag{17}$$

3.2. ARIMA Model

Box and Jenkins [32] published a technique to merge both autoregressive (AR) and moving average (MA) models, called the ARMA (p, q) model, as a union of AR (p) and MA (q) models, generally deployed for univariate timeseries modeling. The ARMA (p, q) model is thus presented as follows:

$$Y_t = c + \varepsilon_t + \sum_{i=1}^{p}\varphi_i X_{t-i} + \sum_{i=1}^{q}\theta_i \varepsilon_{t-i}, \tag{18}$$

where the $\theta_1, \ldots, \theta_q$ and $\varphi_1, \ldots, \varphi_p$, are the parameters of the model and ε is the white noise. If the series is not stationary at the first level, there is a need to subtract it by d ($d = 1, 2, 3, \ldots$) times to make it stationary. Such a timeseries model is called an ARIMA (p, d, q) model.

There are three steps in ARIMA model creation, namely identification, parameter estimation, and diagnostic checking [33,34]. In this regard, for the identification process of the model, after checking the stationarity of the timeseries, the AR and MA terms are derived from the auto-correlation function (ACF) plot. ACF is a statistical metric of the correlation that is used to check if previous values in the timeseries analysis have a certain relationship with the latest values or not [35]. After that, ARIMA parameters, namely (p, d, q), are estimated by the least square method. The three main methods commonly used to select appropriate models are Akaike's Information Criterion (AIC), the Bayesian

Information Criterion (BIC) and the Second-order Akaike's Information Criterion (AICc), which are presented in Equations (19)–(21) for AIC, BIC and AICc, respectively [32,36].

$$\mathrm{AIC} = -2\log(L) + 2k = -2\log(L) + 2(p+q+P+Q) \tag{19}$$

$$\mathrm{BIC} = -2\log(L) + k\ln(n) = -2\log(L) + (p+q+P+Q)\ln(n) \tag{20}$$

$$\mathrm{AICc} = -2\log(L) + 2k(n/(n-k-1)), \tag{21}$$

where n refers to the size of the series and k denotes the number of the parameters of the ARIMA method. It is experimentally proved that the given model becomes efficient when the AIC value is smaller. According to [34], an optimal forecast model is selected based on the best fitting that has the minimum AIC value of the group.

To evaluate the prediction models, the following statistical measures are used for $i(1, \cdots, 7)$ as follows:

$$\mathrm{RMSE}(i) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(s_{it} - \hat{s}_{it})^2} \tag{22}$$

$$\mathrm{MAPE}(i) = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{s_{it} - \hat{s}_{it}}{s_{it}}\right|, \tag{23}$$

where s_{it} denotes the actual value and $\hat{s}_{it}$ and T are the modeled values and the total number of days.

4. Implement the Model for Each Country in the Region

4.1. ARIMA Model Results

The ARIMA model is used in this study as a benchmark to compare the results with the proposed NAR method. To implement the ARIMA model on it, the data are split from 15 February 2020 to 10 September 2021 into two parts; the first part, the training dataset, from 15 February 2020 to 20 May 2021, and the second part, the testing dataset, from 21 May 2021 to 10 September 2021. This data division process is applied to the COVID-19 data of all countries. Therefore, it is shown once in Figure 3. Considering the training part, the ARIMA model is conducted; then, the results are validated with the testing dataset once the parameters are estimated.

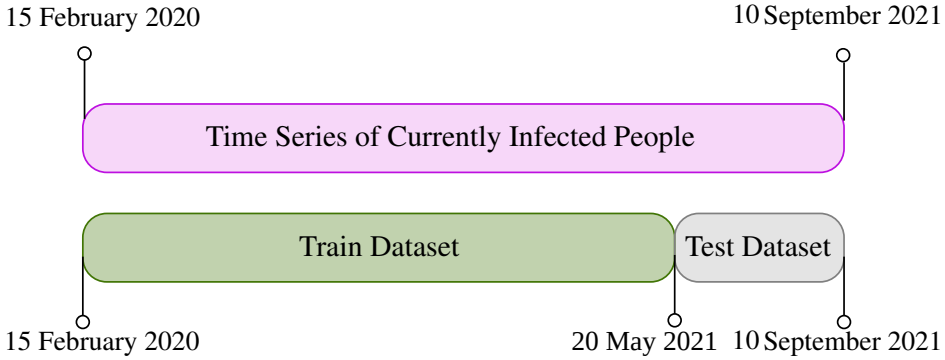

Figure 3. Timetable of train and test datasets.

After splitting the data into the training and test sets, they are transformed to s_{it} for smoothing purposes as expressed in Equation (2). If the data are found to be non-stationary for each country; they are made so by subtracting them from the previous day. The number of times the timeseries data become stationary through difference disposal becomes the value of parameter d.

Upon stabilizing the data, the parameters are estimated. First, the ACF of the d-order difference timeseries is calculated. The order of the Auto-correlation Function (ACF)

exceeding the confidence boundary lag also becomes the value of q. Second, the value of p is computed, which is the order of the Partial Auto-correlation Function (PACF) exceeding the confidence boundary lag. PACF gives the partial correlation of a stationary timeseries with its own lagged values. By observing the ACF and PACF of the residuals, it is determined whether they are white noise or not. Consequently, the fit of the model is assessed by checking the R^2 value. Ultimately, the model is validated and evaluated by applying the ARIMA method to predict the remaining 10% of the data. After that, the RMSE is used, as explained in Section 2, to evaluate the model. The whole process is depicted in Algorithm 1.

Algorithm 1: The procedure of modeling using ARIMA.

```
Data: Obtaining infected data
Result: Choosing the optimized ARIMA model
Initialization;
Splitting data to train and test datasets;
Transforming data by Equation (2);
while minimum of ACF and PACF do
    if (p,q) is acceptable then
        Evaluate the model with RMSE and MAPE;
    else
        Obtaining the value of d;
    end
end
```

The ARIMA models are fit for the datasets of each country to compare them with the proposed model. Both Augmented Dickey–Fuller (ADF) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests also authenticate that the training timeseries are stationary at a 5% significance level. Accordingly, there is a need to apply the differencing method two times for Iran's dataset. Later, diverse models are designed by adjusting various parameters for the MA and AR components of the ARIMA model, as summarized in Table 1. The ARIMA (2,2,2) model additionally assumes AICc criteria. Therefore, the chosen model is checked for some assumptions. The residual analysis of the model is presented in Figure 4. The Ljung–Box test on the residuals, as well as the squared residuals, is also statistically significant at a 5% level with p-value $= 0.9153$. Therefore, the selected model with the minimum summary measures is appropriate with the lowest AIC, BIC, and AICc values.

Table 1. Summary measures for AICc in ARIMA model candidates—Iran series.

Model	AICc
ARIMA (2, 2, 2)	7510.893
ARIMA (0, 2, 0)	7568.916
ARIMA (1, 2, 0)	7568.436
ARIMA (0, 2, 1)	7565.978
ARIMA (1, 2, 2)	7561.546
ARIMA (2, 2, 1)	7561.854
ARIMA (3, 2, 2)	7566.861
ARIMA (2, 2, 3)	7541.272
ARIMA (1, 2, 1)	7563.57
ARIMA (1, 2, 3)	7563.364
ARIMA (3, 2, 1)	7564.825
ARIMA (3, 2, 3)	7512.12

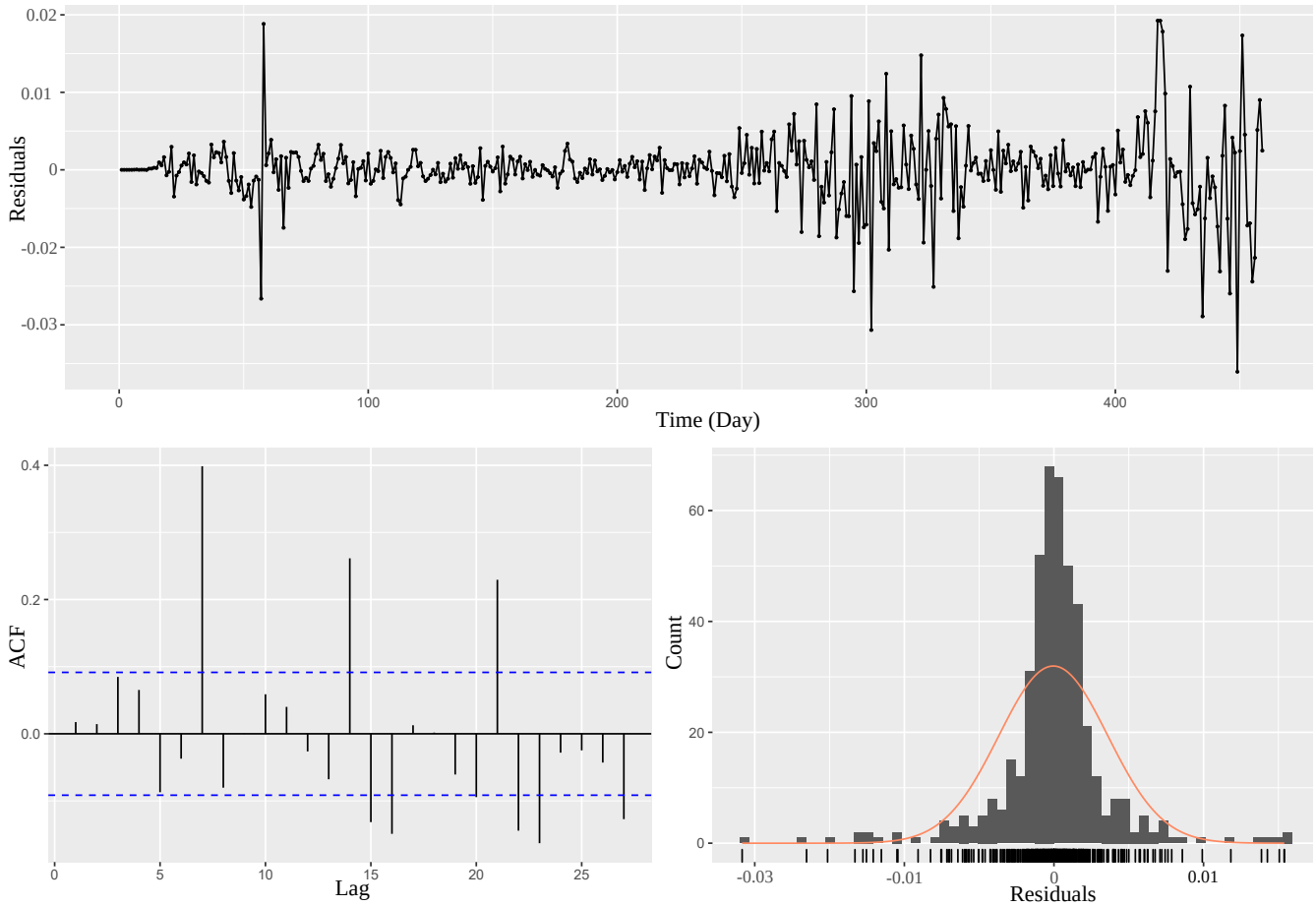

Figure 4. Residuals from ARIMA (2, 2, 2), Iran.

The same approach is implemented in the training data series of other countries. The best model for each country is also selected, which has minimum AIC, BIC, and AICc measures, whose results are summarized in Table 2. The residual analysis of the best model in each country is depicted in Figures A1–A6 in the Appendix A.

Table 2. The results of ARIMA orders and residual analysis.

Country	(p, d, q)	AIC	AICc	BIC
Azerbaijan	$(2, 1, 2)$	3959.84	3959.71	3939.21
Afghanistan	$(2, 1, 3)$	3797.35	3797.17	3772.59
Pakistan	$(2, 1, 3)$	3797.35	3797.17	3772.59
Turkey	$(4, 1, 1)$	3358.17	3357.98	3333.41
Armenia	$(5, 1, 4)$	2886.98	2886.48	2845.71
Iraq	$(1, 1, 3)$	3959.84	3959.71	3939.21

4.2. NARI Model Results

In this section, the constructed NARI model is deployed on the COVID-19 dataset. As mentioned in the previous section, the data are split into two subsets of training and test datasets, as shown in Figure 3. The NARI model is then applied to the training dataset of each country. For this purpose, the training dataset is first transformed into a smooth timeseries, calculated by Equation (2). Then, the correlation matrix of countries is created by Equation (5), as reported in Figure 5. This allows the NARI model to be established for the whole network using Equation (4), given the parameter α as $\alpha = (0.014, 0.005, 0.79)$. The overall process of estimating α is presented in Algorithm 2.

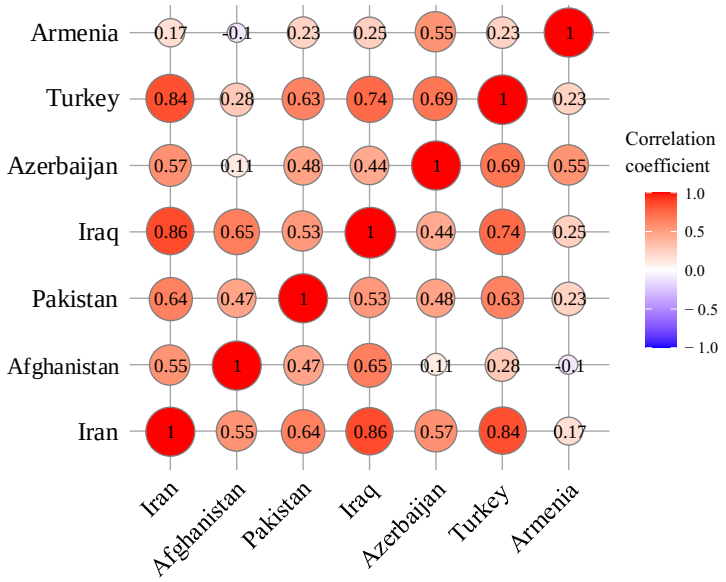

Figure 5. Correlation among countries.

Algorithm 2: The procedure of modeling using NARI.

Data: obtaining infected data
Result: estimating α.
Initialization;
Transforming data by Equation (2);
Creating correlation matrix;
while *setting threshold*, θ **do**
  if $\mathbb{Y}_t$ *is non-singular* **then**
    Evaluate the model with RMSE and MAPE;
  else
    Change the threshold and go to the previous step;
  end
end

Upon estimating the parameter θ, the constructed NARI model needs to be validated. For this purpose, the 50 days are predicted using the constructed ARIMA models in the previous section and are then compared with the 50 days of the test dataset through the RMSE and MAPE metrics. The same procedure is also repeated for the NARI model, and the results are subsequently compared. The outcomes for both models are summarized in Table 3, wherein the NARI model outperforms the ARIMA one.

Table 3. RMSE and MAPE metrics of ARIMA and NARI models for 7 countries.

		Iran	Afghanistan	Pakistan	Iraq	Armenia	Azerbaijan	Turkey
ARIMA model	RMSE	5.42	1.82	4.20	3.04	3.85	3.01	10.1
	MAPE	8.21	0.68	3.33	2.43	4.1	3.53	3.93
NARI model	RMSE	3.06	1.02	1.75	1.41	1.09	1.96	8.33
	MAPE	2.15	0.42	2.87	1.01	1.81	1.55	1.85

The graphs of the transformed data of the total infected cases in Iran and the modeled data using the NARI and ARIMA models are shown in Figure 6. Upon finding the best COVID-19 modeling, the next 110 days of the disease are predicted according to the infected cases. The forecast results are plotted in Figure 7.

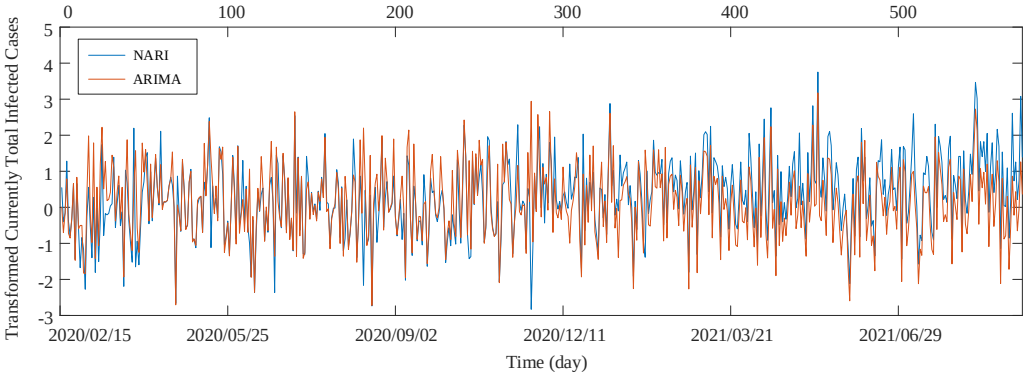

Figure 6. NARI model and ARIMA model for transformed currently total infected cases in Iran.

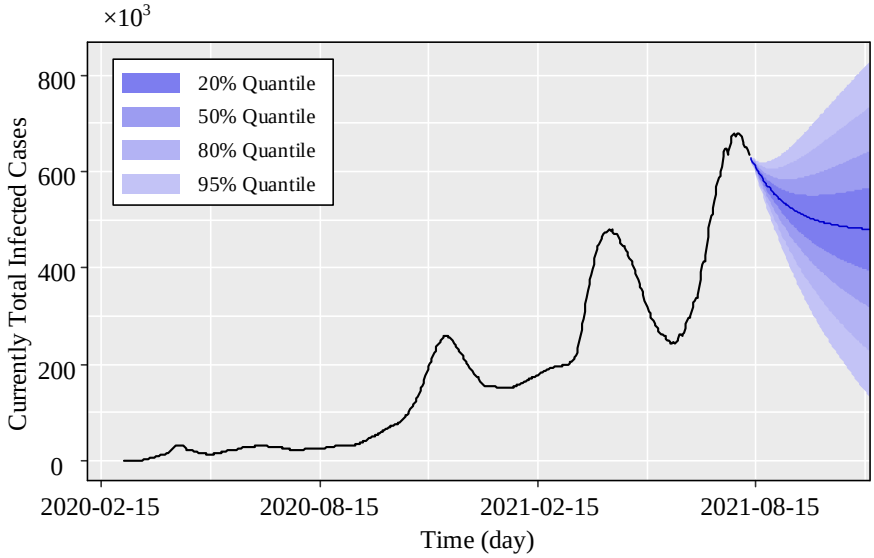

Figure 7. Forecast of currently total infected cases in Iran for 110 days.

Timeseries forecasting for the spreading disease profile can be implemented considering disease interactions in different countries or without any correlation among them. Using the NARI model, considering these interactions, helps the constructed model to examine one of the most critical characteristics of disease transmission between societies and significantly enhances the accuracy of the timeseries prediction. This issue can be assessed by comparing the two models of ARIMA and NARI. It is shown in Table 3 that the RMSE for Iran, considering the disease interactions among neighboring countries, equals 3.06 for the NARI model and without considering it equals 5.42 for the ARIMA model. This means that the deliberation of the disease communications in neighboring countries promotes the prediction certainty substantially. Therefore, more reliable determinations can be made by policymakers to control the disease. It is of note that this prediction is based on the current pandemic situation. The results will also be altered in the case of events that change the present circumstances, such as enforcing stricter social distancing or wider public vaccination.

The forecast shows that the disease trend in Iran has passed the fifth peak, and the downward trend of the disease will continue after September 2021, so that the total number of infected cases per day will fall to less than 480,000 by the end of 2021. This is the point estimation prediction obtained from the NARI model, which is the most statistically possible case. However, Figure 7 also provides the quantiles of 20%, 50%, 80% and 95%, indicating the prediction uncertainty. The lower bounds of the 20%, 50%, 80% and 95% quantiles, respectively, indicate that the downward trend could be relatively more severe, bringing the total number of infected cases below 390,000, 320,000, 220,000 and 130,000

by the end of 2021. In addition, the upper limits of these quantiles indicate a possible new peak in Iran's new COVID-19 data. In this case, the total number of infected cases corresponding to 20%, 50%, 80% and 95% quantiles may reach 560,000, 640,000, 730,000 and 820,000, respectively, by the end of 2021.

5. Discussion

In Section 3, it was discussed that the adjacency matrix, representing the disease interaction among nations, is formed by adopting a threshold and comparing it to the correlation matrix. In order to determine this threshold and explain how to implement the process, additional analysis is required, which is discussed in this section.

To explain the approach adopted to compute the correlation threshold, it should be explained that there are two constraints to meet. First, none of the n_i in Equation (7) should be zero; otherwise, an infinity term would appear in this equation. Besides, a value of n_i, which can minimize the RMSE, is preferred as it helps the algorithm gain better accuracy. To implement an algorithm that can satisfy these two conditions, the threshold value is defined as a decision variable. An external loop is then added to the main algorithm to change the value of θ and calculate the corresponding n_i and RMSE. The results for different θ values from 0 to 1 with an increment of $\Delta\theta = 0.1$ are reported in Table 4.

Table 4. Different correlation threshold and corresponding n_i and RMSE.

θ	n_1	n_2	n_3	n_4	n_5	n_6	n_7	RMSE
0.1	7	6	7	7	7	7	6	3.98
0.2	5	4	6	6	5	6	4	3.85
0.3	5	3	5	5	5	4	1	3.64
0.4	5	3	5	5	5	4	1	3.64
0.5	5	2	3	4	3	4	1	3.06
0.6	3	1	2	3	1	4	0	NaN
0.7	2	0	0	2	0	2	0	NaN
0.8	2	0	0	1	0	1	0	NaN
0.9	0	0	0	0	0	0	0	NaN

As seen, θ values greater than 0.6 give infinity values for n_i, thus cannot be selected as a correlation threshold in the algorithm. Moreover, for the rest of the cases, $\theta = 0.5$ gives the minimum amount of RMSE. Hence, it is selected as the optimal case and is utilized in the model implementation.

6. Summary and Conclusions

In this paper, the COVID-19 spreading profile in Iran is predicted in view of the influence of the severity and correlation of the disease in neighboring countries. To this end, the timeseries of COVID-19 infection among seven countries in the region, including Iran, Turkey, Iraq, Azerbaijan, Armenia, Afghanistan, and Pakistan, are downloaded from the online databases provided by the WHO and Johns Hopkins University. Then, a network is formed in the region to establish the correlation matrix among the countries concerned. Furthermore, by incorporating the correlation matrix into the proposed formula and calculating the model coefficients, the NARI model is used to predict the number of infected cases in Iran up until the end of September 2021, taking into account the impact of the disease in neighboring countries. The main results obtained in this study are as follows:

1. The correlation matrix obtained from the network of the countries in the region shows that the greatest impact of COVID-19 on Iran comes from Iraq, Turkey, Pakistan, Azerbaijan, Afghanistan and Armenia, with correlation coefficients of 0.86, 0.83, 0.64, 0.56, 0.55, 0.16, respectively. This result can also be seen in the trend of infected cases. The increasing/decreasing trend and the number of disease peaks in Iran, Iraq, and Turkey are very similar and have occurred within a short period of time. This indicates that the proposed correlation criterion is able to capture the similarity between infected data and disease peaks;

2. Timeseries predictions can be made with or without considering disease interactions in different countries. Incorporating the disease interaction not only helps the algorithm assess one of the most important components of disease transmission between societies but also significantly increases the accuracy of the timeseries prediction. This issue can be examined by comparing the two models of ARIMA and NARI. The RMSE with and without considering the disease interactions among neighboring countries is equal to 5.42 and 3.06 for ARIMA and NARI, respectively. This means that the consideration of the disease interactions in neighboring countries improves the prediction accuracy considerably. As the model's accuracy in predicting disease increases, more reliable tools are provided for policymakers to take informed controlling decisions;
3. The point estimation obtained from the NARI model indicates that the number of infected cases in Iran declines after September 2021, so the total currently infected cases will fall below 480,000 by the end of 2021. According to the prediction corresponding to the lower limit of 20%, 50%, 80%, and 95% quantiles, the total number of infected persons will fall below 390,000, 320,000, 220,000 and 130,000, respectively, by the end of 2021.

Iran's close neighbors, sharing common borders, and their impacts on the COVID-19 spreading profile in Iran are examined in this paper. However, ideally, more distant countries in the region that have direct or indirect demographic relationships with Iran can be also considered. Such a high volume of interactions between the countries requires the construction of a larger network to cover more countries and to subsequently provide a more reliable prediction. Such a model imposes more complexities on the problem, making the prediction results more accurate and reliable. Moreover, various factors, such as hospitalization, social distancing, quarantine, and so forth, can affect the number of people infected with COVID-19 in a society. However, the spreading profile of disease under the effects of the involved factors is not in the scope of the current research. Simulating the disease spread, taking into account the factors involved, requires establishing a system of differential equations in a so-called compartmental model and solving it incrementally to simulate the disease profile in the future. This topic is under investigation by the authors.

Author Contributions: Conceptualization, A.S.K. and M.S.; Data curation, A.S.K., M.S., V.R.H. and H.K.; Formal analysis, A.S.K. and M.S.; Investigation, A.S.K., M.S. and V.R.H.; Methodology, A.S.K., M.S. and V.R.H.; Project administration, A.S.K., M.S. and V.R.H.; Resources, A.S.K. and V.R.H.; Software, A.S.K., M.S., V.R.H. and H.K.; Supervision, A.S.K.; Validation, A.S.K., M.S., V.R.H. and H.K.; Visualization, A.S.K., M.S. and H.K.; Writing—original draft, A.S.K., M.S., V.R.H. and H.K.; Writing—review and editing, A.S.K., M.S., V.R.H. and H.K. All authors have read and agreed to the published version of the manuscript.

Funding: The work described in this paper is supported by the Faculty of Economics and Business Administration of Yibin University.

Data Availability Statement: Data, models, or code that support the findings of this study are available from the authors upon reasonable request.

Acknowledgments: The first author gratefully acknowledges the Faculty of Economics and Business Administration of Yibin University for the support during the preparation of this paper. Additionally, the second author appreciatively thanks the Iran's National Elites Foundation for a postdoctoral fellowship at Sharif University of Technology for supporting this research.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

This section presents the residuals obtained from the ARIMA model of six countries, including Afghanistan, Pakistan, Iraq, Armenia, Azerbaijan, and Turkey. As discussed in Section 4.1, the results of these graphs are used to find the optimized ARIMA model for the number of infected people in the tracked countries. Further details of the calculations and modeling process of these graphs can be found in Section 4.

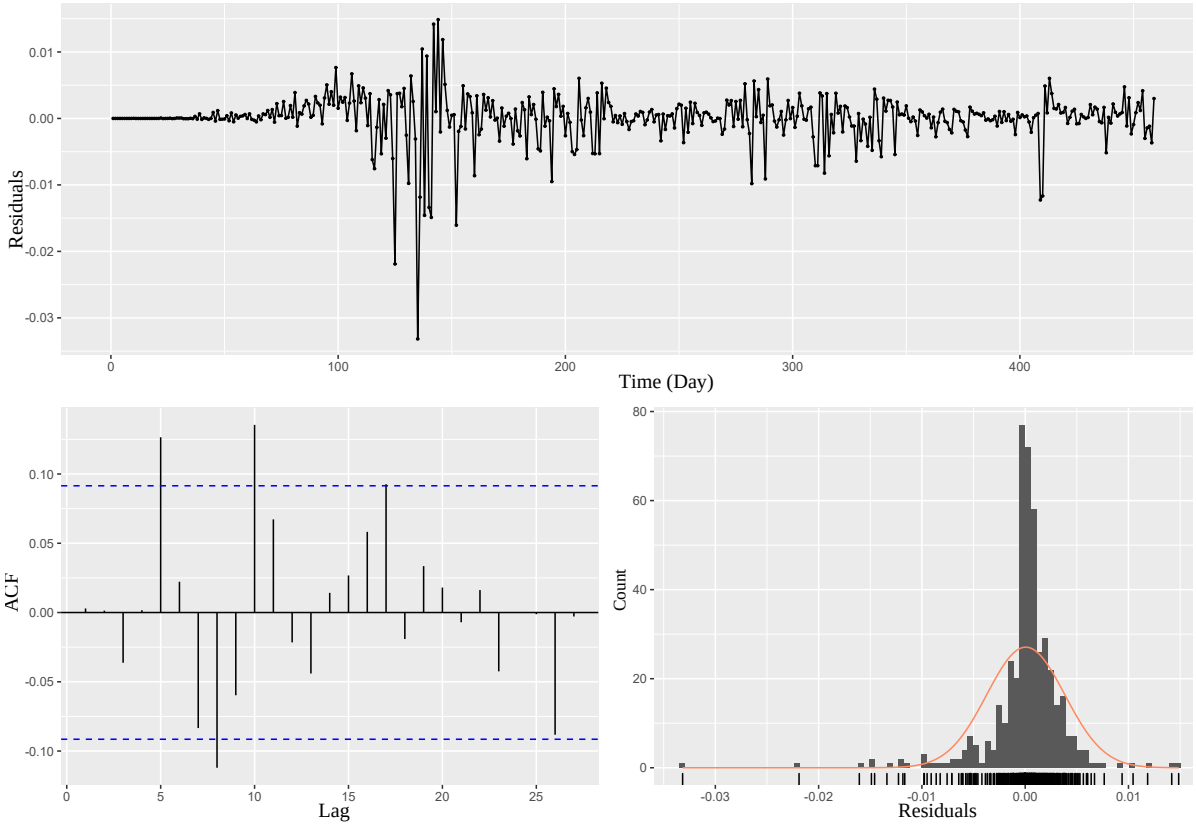

Figure A1. Residuals from ARIMA (2, 1, 3), Afghanistan.

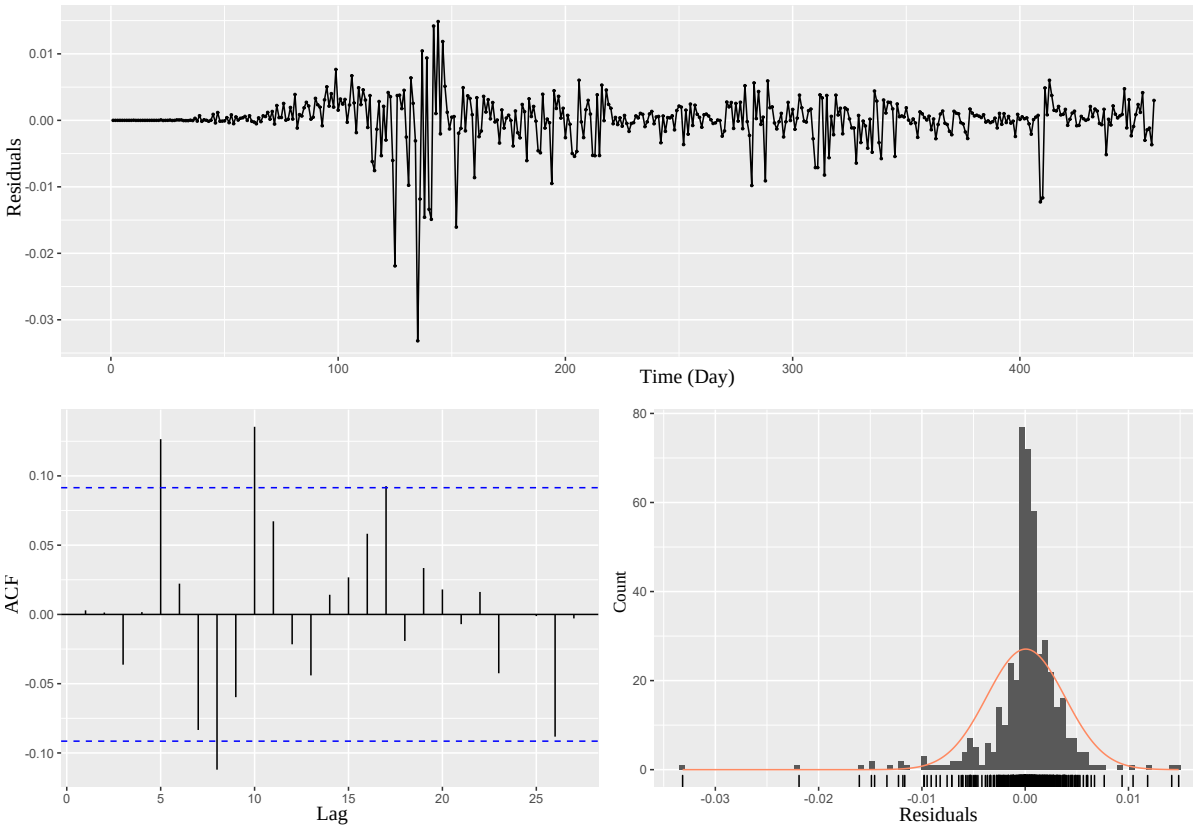

Figure A2. Residuals from ARIMA (2, 1, 3), Pakistan .

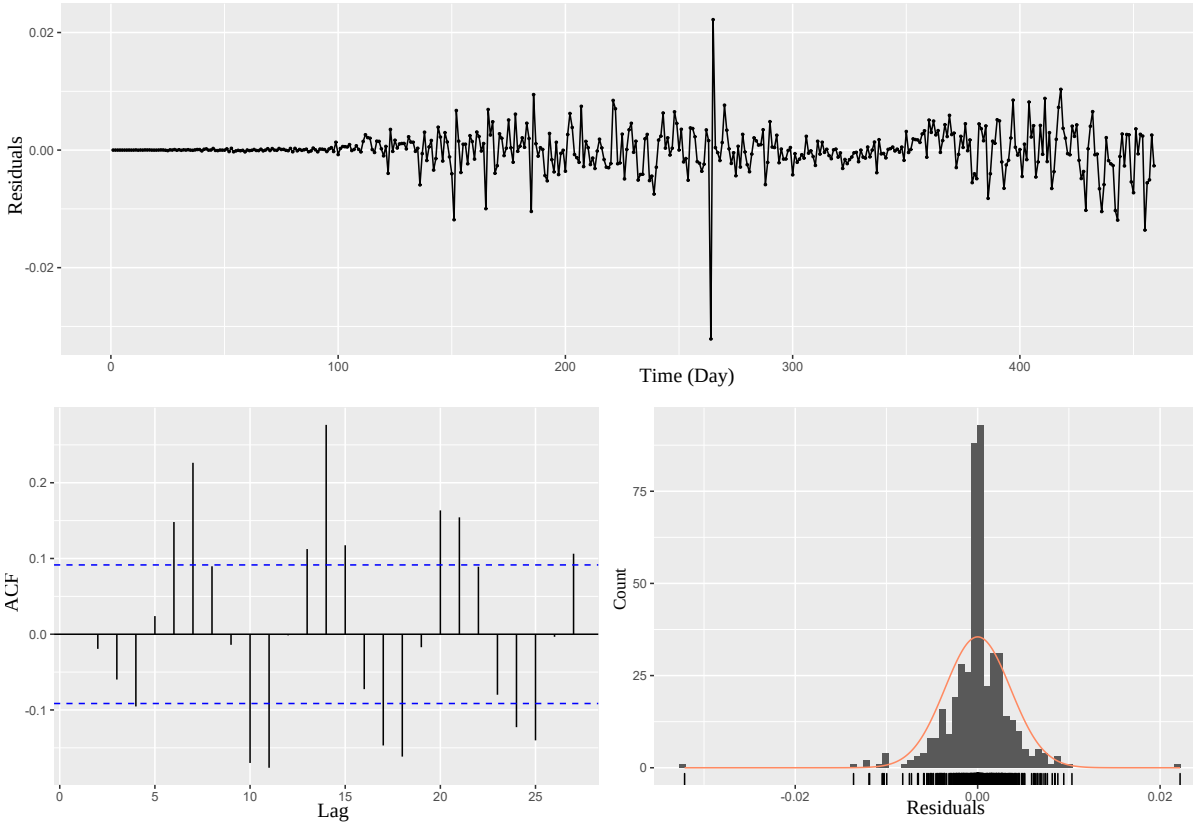

Figure A3. Residuals from ARIMA (1, 1, 3), Iraq.

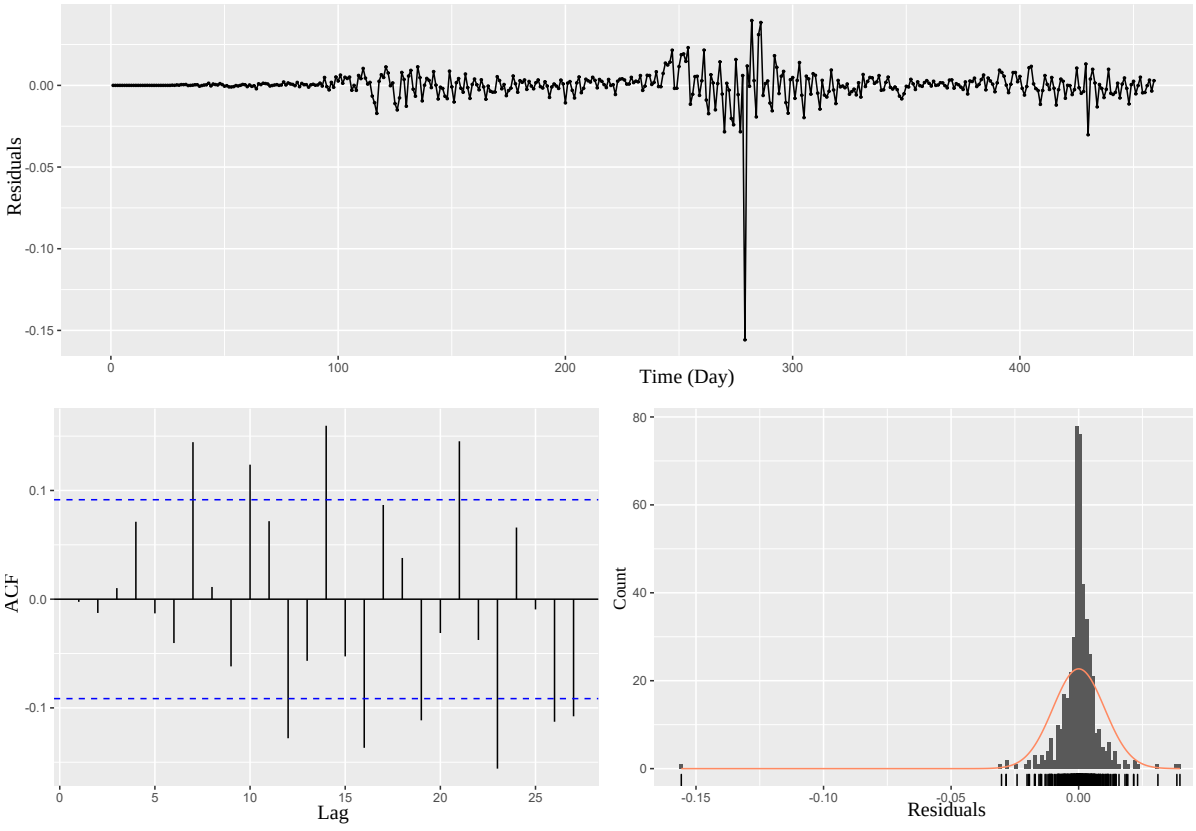

Figure A4. Residuals from ARIMA (5, 1, 4), Armenia.

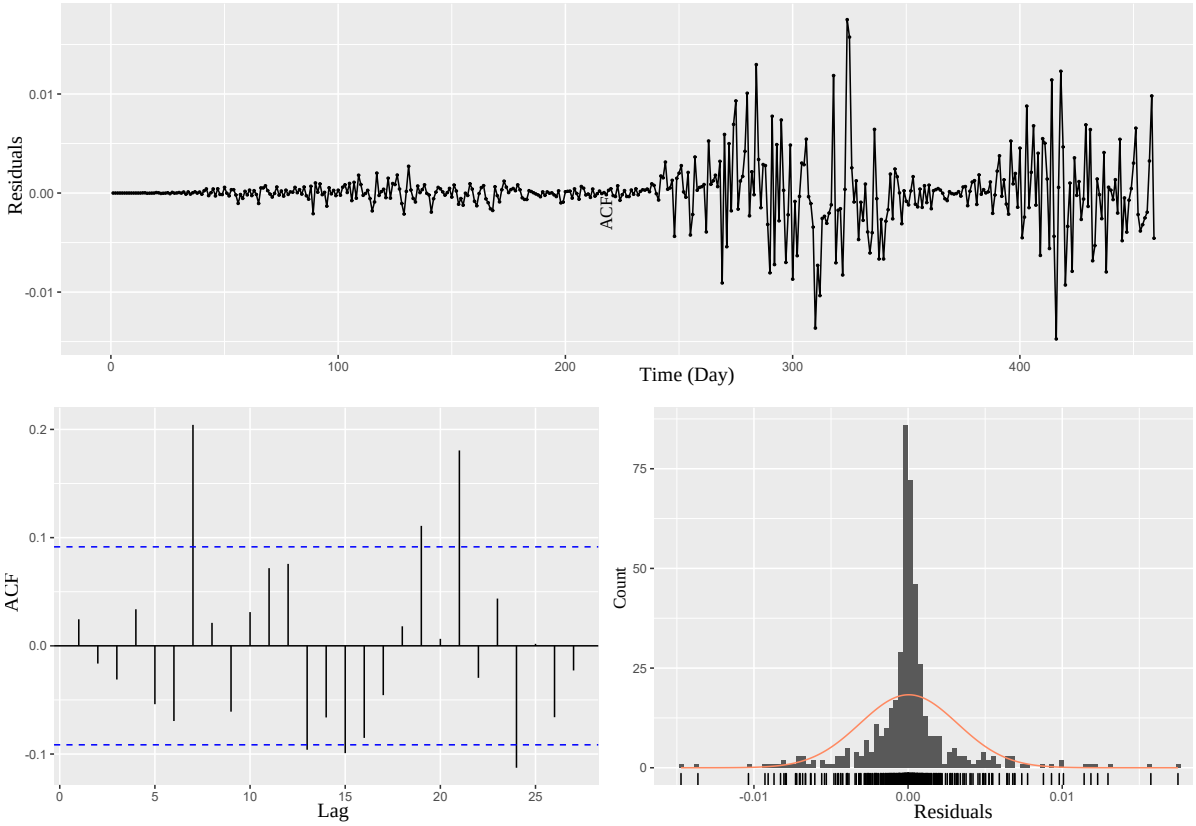

Figure A5. Residuals from ARIMA (2, 1, 2), Azerbaijan.

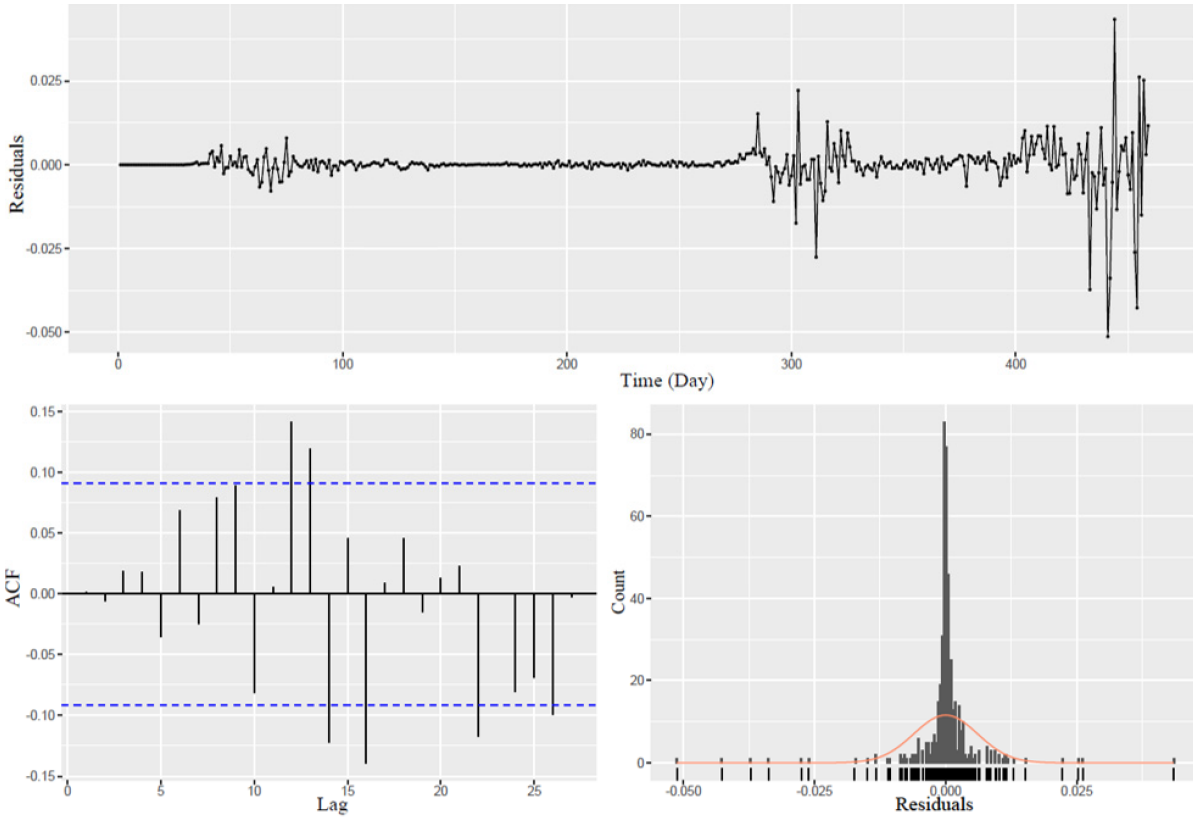

Figure A6. Residuals from ARIMA (4, 1, 1), Turkey.

References

1. Lai, C.C.; Shih, T.P.; Ko, W.C.; Tang, H.J.; Hsueh, P.R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *Int. J. Antimicrob. Agents* **2020**, *55*, 105924. [CrossRef] [PubMed]
2. Davis, R.A. Of Borders and Bodies: The Second Wave Begins. In *The Spanish Flu*; Palgrave Macmillan US: New York, NY, USA, 2013; pp. 47–68. [CrossRef]
3. Anne R. ARIMA modelling of predicting COVID-19 infections. *medRxiv* **2020**. [CrossRef]
4. Adiga, A.; Dubhashi, D.; Lewis, B.; Marathe, M.; Venkatramanan, S.; Vullikanti, A. Mathematical Models for COVID-19 Pandemic: A Comparative Analysis. *J. Indian Inst. Sci.* **2020**, *100*, 793–807. [CrossRef] [PubMed]
5. Shadabfar, M.; Cheng, L. Probabilistic approach for optimal portfolio selection using a hybrid Monte Carlo simulation and Markowitz model. *Alex. Eng. J.* **2020**, *59*, 3381–3393. [CrossRef]
6. Shadabfar, M.; Mahsuli, M.; Khoojine, A.S.; Hosseini, V.R. Time-variant reliability-based prediction of COVID-19 spread using extended SEIVR model and Monte Carlo sampling. *Results Phys.* **2021**, *26*. [CrossRef]
7. Babaei, A.; Jafari, H.; Banihashemi, S.; Ahmadi, M. A stochastic mathematical model for COVID-19 according to different age groups. *Appl. Comput. Math.* **2021**, *20*, 140–159.
8. Babaei, A.; Jafari, H.; Banihashemi, S.; Ahmadi, M. Mathematical analysis of a stochastic model for spread of Coronavirus. *Chaos Solitons Fractals* **2021**, *145*, 110788. [CrossRef]
9. Boudaoui, A.; El hadj Moussa, Y.; Hammouch, Z.; Ullah, S. A fractional-order model describing the dynamics of the novel coronavirus (COVID-19) with nonsingular kernel. *Chaos Solitons Fractals* **2021**, *146*, 110859. [CrossRef]
10. Singh, H.; Srivastava, H.M.; Hammouch, Z.; Sooppy Nisar, K. Numerical simulation and stability analysis for the fractional-order dynamics of COVID-19. *Results Phys.* **2021**, *20*, 103722. [CrossRef]
11. Sahoo, P.; Mondal, H.S.; Hammouch, Z.; Abdeljawad, T.; Mishra, D.; Reza, M. On the necessity of proper quarantine without lock down for 2019-nCoV in the absence of vaccine. *Results Phys.* **2021**, *25*, 104063. [CrossRef]
12. Danane, J.; Allali, K.; Hammouch, Z.; Nisar, K.S. Mathematical analysis and simulation of a stochastic COVID-19 Lévy jump model with isolation strategy. *Results Phys.* **2021**, *23*, 103994. [CrossRef]
13. Zamir, M.; Nadeem, F.; Abdeljawad, T.; Hammouch, Z. Threshold condition and non pharmaceutical interventions's control strategies for elimination of COVID-19. *Results Phys.* **2021**, *20*, 103698. [CrossRef]
14. Babaei, A.; Ahmadi, M.; Jafari, H.; Liya, A. A mathematical model to examine the effect of quarantine on the spread of coronavirus. *Chaos Solitons Fractals* **2021**, *142*, 110418. [CrossRef]
15. Katoch, R.; Sidhu, A. An Application of ARIMA Model to Forecast the Dynamics of COVID-19 Epidemic in India. *Glob. Bus. Rev.* **2021**. [CrossRef]
16. Sahai, A.K.; Rath, N.; Sood, V.; Singh, M.P. ARIMA modelling & forecasting of COVID-19 in top five affected countries. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2020**, *14*, 1419–1427. [CrossRef]
17. Malki, Z.; Atlam, E.S.; Ewis, A.; Dagnew, G.; Alzighaibi, A.R.; ELmarhomy, G.; Elhosseini, M.A.; Hassanien, A.E.; Gad, I. ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound. *Neural Comput. Appl.* **2021**, *33*, 2929–2948. [CrossRef]
18. Chaurasia, V.; Pal, S. COVID-19 Pandemic: ARIMA and Regression Model based Worldwide Death Cases Predictions. *SSRN Electron. J.* **2020**, 1–23. [CrossRef]
19. Kumar, N.; Susan, S. COVID-19 Pandemic Prediction using Time Series Forecasting Models. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020, Kharagpur, India, 1–3 July 2020. [CrossRef]
20. Attanayake, A.M.C.H.; Perera, S.S.N. Forecasting COVID-19 Cases Using Alpha-Sutte Indicator: A Comparison with Autoregressive Integrated Moving Average (ARIMA) Method. *Biomed Res. Int.* **2020**, *2020*. [CrossRef] [PubMed]
21. Hernandez-Matamoros, A.; Fujita, H.; Hayashi, T.; Perez-Meana, H. Forecasting of COVID19 per regions using ARIMA models and polynomial functions. *Appl. Soft Comput.* **2020**, *96*, 106610. [CrossRef] [PubMed]
22. Yang, Q.; Wang, J.; Ma, H.; Wang, X. Research on COVID-19 based on ARIMA modelΔ—Taking Hubei, China as an example to see the epidemic in Italy. *J. Infect. Public Health* **2020**, *13*, 1415–1418. [CrossRef] [PubMed]
23. Dong, E.; Du, H.; Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **2020**, *20*, 533–534. [CrossRef]
24. World Health Organization. COVID-19 Coronavirus Pandemic. 2020. https://covid19.who.int/(accessed on 24 September 2021).
25. Noh, J.; Danuser, G. Estimation of the fraction of COVID-19 infected people in U.S. states and countries worldwide. *PLoS ONE* **2021**, *16*, e0246772. [CrossRef]
26. Hui, D.S.; Azhar, E.I.; Madani, T.A.; Ntoumi, F.; Kock, R.; Dar, O.; Ippolito, G.; Mchugh, T.D.; Memish, Z.A.; Drosten, C.; et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. *Int. J. Infect. Dis.* **2020**, *91*, 264–266. [CrossRef] [PubMed]
27. Hayashi, F. *Econometrics*; Princeton University Press: Princton, NJ, USA, 2000; p. 49.
28. Giada, L.; Marsili, M. Algorithms of maximum likelihood data clustering with applications. *Phys. Stat. Mech. Its Appl.* **2002**, *315*, 650–664. [CrossRef]
29. Casella, G.; Berger, R.L. *Statistical Inference*, 2nd ed.; Cengage Learning: Boston, MA, USA, 2007.

30. Khoojine, A.S.; Han, D. Network analysis of the Chinese stock market during the turbulence of 2015–2016 using log-returns, volumes and mutual information. *Phys. Stat. Mech. Its Appl.* **2019**, *523*, 1091–1109. [CrossRef]
31. Khoojine, A.S.; Han, D. Stock price network autoregressive model with application to stock market turbulence. *Eur. Phys. J. B* **2020**, *93*. [CrossRef]
32. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C. *Time Series Analysis, Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
33. Brockwell, P.J.; Davis, R.A. *Introduction to Time Series and Forecasting*; Springer: Berlin/Heidelberg, Germany, 2016.
34. Paolella, M.S. ARMA Model Identification. In *Linear Models and Time-Series Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2018; [CrossRef]
35. Chakrabarti, A.; Ghosh, J.K. AIC, BIC and Recent Advances in Model Selection. In *Philosophy of Statistics*; Elsevier: Amsterdam, The Netherlands, 2011; pp. 583–605. [CrossRef]
36. Burnham, K.P.; Anderson, D.R. (Eds.) *Model Selection and Multimodel Inference*; Springer: New York, NY, USA, 2004. [CrossRef]