NAME: Justin Clark
DATE: 08/01/2020
Scientific Article
STAT/CS 387

1 Introduction

The study of infectious diseases has been of great interest within the field of epidemiology for many years due to unique characteristics of emergence and reemergence innately related to these viruses [17]. Infectious diseases demonstrate numerous characteristics distinct from other human disease including a potential for unpredictable and explosive global impact, human to human transmission, prevention potential, and dependence on the nature and complexity of human behavior [12]. The transmission rate and potential impact of infectious disease have been directly related to human behavior and lifestyle due to modes of transmission such as social gatherings, public transportation, occupation, and the current healthcare environment. The distinct characteristics of unique infectious diseases further drives the necessity for proper intervention and containment methods based upon the possibility of vaccination. For many newly emerging infectious diseases, the effectiveness and efficiency of vaccination development is unknown. This discrepancy in vaccination has occurred for many well-known, contemporary, infectious diseases.For instance, vaccine manufacturers were able to quickly develop and produce an intervention strategy to curve the spread of pandemic influenza A (H1N1) while no effective vaccination has been developed to curve the spread of HIV/AIDS, which was discovered over 30 years ago [15].

An outbreak of serve acute respiratory syndrome coronavirus 2 (SARS-CoV-2) or novel coronavirus has resulted in over three million confirmed cases worldwide, first identified in Wuhan, China in December, 2019. [21]. The clinical spectrum associated with this disease, coronavirus disease 2019 (2019-nCoV), has been associated with acute respiratory distress in severe cases while the majority of cases display only mild or sub-clinical symptoms [32]. Currently, 2019-nCoV severity has been correlated with health characteristics including age and underlying conditions such as diabetes, hypertension, and cardiovascular disease [31]. Initial, retrospective, single-centre studies in Wuhan, China have outlined a comprehensive exploration of the epidemiology and clinical features of confirmed 2019-nCoV patients. This study demonstrated evidence of human-to-human transmission while also suggesting that the novel coronavirus is most likely to affect older males with underlying comorbidities due to the weaker immune functionality of patients. The discussed study observed a larger number of infected males when compared to females, similar to SARS-CoV and MERS-CoV, and attributed this reduced susceptibility to viral infection to the innate and adaptive immunity from the X chromosome and specific sex hormones [5] [18].An outbreak of 2019-nCoV aboard a cruise ship, *Diamond Princess*, offered a particular opportunity to understand non-identifiable features compared to a larger population due to the closed population, allowing researchers to investigate confirmed, asymptomatic cases [25]. This research and other accumulating evidence indicates the extreme possibility for transmission in confined environments such as hospitals, cruise ships, and prisons [28].

The absence of a vaccine or effective intervention method tied to 2019-nCoV has highlighted the importance of disease containment in terms of transmission interruption. Global efforts have been implemented to aid in the fast yet thorough identification and isolation of all individuals infected with 2019-nCoV. The uncertainty related to disease transmission and incidence has outlined the possibility of modeling as a way of disease assessment and prediction. Moreover, this variability in disease activity and transmission highlights the necessity of rapidly available assessments of disease intensity and severity to guide the public health domain in overall containment such as prevention techniques and treatment discussion to a public audience [3]. Specifically, increasing the number of recovered cases while minimizing the number of confirmed cases (i.e 'flattening the curve') is necessary to stop the spread of infectious diseases and leads to more stable rates of infection that our approachable within the public health domain. This occurs through the use of intervention strategies targeted at reducing social interaction in different environments to generally lower transmission of infectious disease throughout the population [29].

Mathematical models in combination with the field of epidemiology have been identified as useful experimental tools for hypothesis evaluation, answering disease specific questions, and estimating key dynamic quantities from the data related to infectious disease. This type of modeling effort can identify important data, highlight hidden trends, forecast incidence and quantify uncertainty within model forecasts [16]. Inherently, the reliability of public health surveillance and model forecasts depends upon the ability of models to make the extrapolations necessary to predict the most likely course of a public health emergency given the incoming, recorded information [2]. The modeling possibilities of infectious disease, combined with a constantly updating stream of information highlights the importance and necessity of numerous, accurate models for characterizing and discovering both short and long term trends of infectious disease data. Specifically, we must acknowledge that any single modeling technique will not be the optimal choice for analyzing different disease characteristics during distinct stages of an infectious disease life-cycle [10].

In this piece, we propose multiple methods as techniques for predicting incidence of 2019-nCoV in the United States and the transmission potential of the virus. The discussed mathematical and statistical methods are those able to estimate or predict key epidemiological parameters from available, physical data [14].

First, we propose and define a compartmental SEIRD model out of the classical deterministic epidemiological toolkit as a baseline analysis of the exponential growth phase of the current epidemic. Moreover, we look to employ this type of compartmental model to analyze short term incidence and estimate the basic reproductive number R_0. This epidemiological parameter is of interest due to its ability to quantify infectious disease transmission over a population [27]. Simply put, the value of this parameter will describe epidemic severity based on novel transmission.

Second, we look to forecast the cumulative number of confirmed cases within the United States. This type of data is reported daily by numerous organizations around the globe with the caveat that the number of confirmed or recovered cases on a specific day have strong correlations with the values of previous days. Therefore, we look to implement a statistical based method to aid in forecasting confirmed disease incidence within the United States based on the use of these previous values. The different types of auto regressive time series models have proven to be strong, flexible tools for analyzing overall patterns within time series data and have been used to estimate and forecast many practical problems. Specifically, the ARIMA model is often used for the prediction of infectious disease [4].

2 Methods and Materials

2.1 Data Sources

The 2019-nCoV data used for analysis was drawn from the Johns Hopkins CSSE Github Repository. This data source contains daily updated files relaying location specific information of the confirmed, recovered, and death incidence of 2019-nCoV cases. This data set categorizes information by relevant characteristics such as country, city and coordinate location and contains novel coronavirus information for 249 locations,for both individual cities and entire countries. The time series data contains case numbers for a majority of the outbreak beginning in January 2020 (01/22/2020). The discussed data is available for each day with the cutoff for analysis within this report being June 2020(06/26/2020). This data will be used to summarize location relationships related to 2019-nCoV and to evaluate conclusions drawn from appropriate model derivation.

2.2 Mathematical/Mechanistic State-Space Models

Mechanistic state-space models such as the simple SIR model and more complicated SEIR model are useful mathematical tools for addressing hypothesis related to parameter estimation, transmission characteristics, and population dynamics related to infectious disease. The currently understood transmission dynamics related to 2019-nCoV highlight a weakness within the simple susceptible-infected-removed model. Initial diagnostics of the COVID outbreak in Wuhan, China indicate a difficulty identifying and isolating infected cases at earlier stages of the disease life cycle [23]. This criteria of disease detection promotes the implementation of a more complex susceptible-exposed-infected-recovered-dead (SEIRD) compartmental model. Given a model based of ODE's, certain assumptions related to disease and population dynamics must be defined. We assume homogeneous mixing of individuals within the population such that the entire population is defined to be susceptible at the beginning of simulation. Furthermore, we assume a lack of reoccurring infection such that recovered individuals are protected against reinfection. Similarly, due to the fast timescale of the epidemic we do not include the effect of natural birth or natural death within the model [6] [1]. These assumptions related to disease transmission are illustrated in Diagram 1.

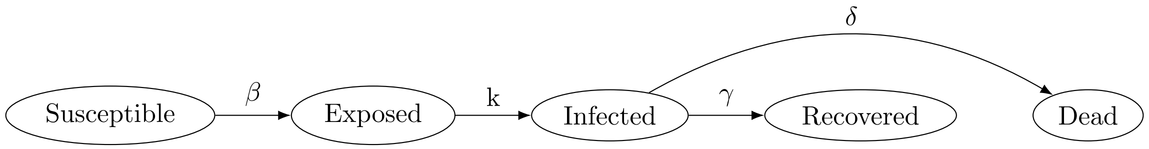


Diagram 1: Transfer Diagram SEIRD Model

Variable	Notation
Susceptible Population	S(t)
Exposed Population	E(t)
Infected Population	I(t)
Recovered Population	R(t)
Dead Population	D(t)
Transmission Rate	β
Latent/Incubation Period	1/k
Recovery Rate	γ
Mortality Rate	δ
Social Distancing Rate	λ

Table (1) SEIRD Compartment and Parameter Notation

Given the model notation and assumptions defined above, we are able begin with differential equations:

$$\begin{aligned} dS/dt &= -\beta I(t)S(t)/N(t) \\ dE/dt &= \beta I(t)S(t)/N(t) - kE(t) \\ dI/dt &= kE(t) - (\gamma + \delta)I(t) \\ dR/dt &= \gamma I(t) \\ dD/dt &= \delta I(t) \end{aligned}$$

where S(t), E(t), I(t), R(t), and D(t) are the respective incidence of each class at time t such that S(t) + E(t) + I(t) + R(t) + D(t) = N(t), the total population, for all values of t. Therefore, S(0) = S_0 quantifies the susceptible population at t=0. The mortality rate, δ, is computed given the equation $\delta = \gamma(\frac{CFP}{1-CFP})$ such that the CFP is the mean case fatality proportion or the percentage of confirmed infectious cases that end in death.

We look to use the nonlinear differential equations defined above for two reasons. First, to simulate the effect of social distancing measures on the growth of infectious disease based on model parameters drawn from literature. This simulation will occur through the inclusion of an additional model parameter, λ. will be equal to interaction effect within the population such that a value of $\lambda = 1$ is simply the basic SEIRD model described above. Alternatively, a value of $\lambda = 0$ implies no interaction between the compartments such that all susceptible individuals are quarantined. Thus, we slightly redefine the ODEs above such that $dS/dt = -\lambda\beta I(t)S(t)/N(t)$ and $dE/dt = \lambda\beta I(t)S(t)/N(t) - kE(t)$. We look to examine the effect of this parameter on 'flattening the curve' of infected/exposed populations.

Second, the described ODEs will be used to model the cumulative number of cases notifications in the United States and estimate key dynamic quantities related to infectious disease such as the basic reproductive number R_0. The basic reproductive number quantifies the transmissibility of an infection and is defined as the mean number of secondary infections caused by a typical infectious agent into an entirely susceptible population [7]. More specifically, we look to employ a least-squares fitting procedure to identify the model that best matches the exponential growth phase of the 2019-nCoV time series data. This optimization of model parameters will occur through the comparison of the observed cumulative confirmed case notifications within the United States to a simulated value C(t) such that $dC/dt = k{*}E$. Finally, through the minimization of model residuals, we look to compare the simulated value of R_0 such that $R_0 = \beta/(\gamma + \delta)$

2.3 Statistical-based Methods for Epidemic Surveillance

Auto-Regressive Integrated Moving Average (ARIMA) models have been previously established as useful for epidemic time series forecasting given their leveraging of the flexibility- interpretability trade off. [34]. The ARIMA model is labeled as ARIMA(p,d,q) given model parameters[11]:

- p: number of auto regressive terms or periods of lag
- d: number of differences required for stationary series
- q: number of moving averages

This type of model requires the initialization of of auto regressive(p),moving average (q) and integration (d) terms. This type of flexibility of implementation allows for the control of seasonal variation and temporal correlation before attempting forecasting techniques. A stationary time series is required as a condition for the use of the ARIMA model. Statistical characteristics of stationary behavior include both the mean and auto-correlation to be consistent over time. Therefore, time series that demonstrate constantly increasing, constant decreasing, and/or distinct trends related to the dependent variable (time) are not stationary. The required condition of the ARIMA model can be satisfied using different methods of data transformation. These modifications to standard regression processes include using the logarithm of lagged disease incidence to control for auto-correlation within the data [19]. Additionally, differencing techniques were applied to the time series data by computing the differences between consecutive observations within the data. The use of logarithmic transformations can help to stabilise the variance of the time series while the use of differencing can help to stabilise the mean of the time series due to removing changes in the level of the series. We define the backward shift operator B such that $By_t = y_{t-1}$. Therefore, a first-order difference $y_t^{'} = y_t - y_{t-1}$ can be written as $(1 - B)y_t$. Using this notation we are able to write a second-order difference as $(1 - B)^2 y_t$. After verification of stationary behavior, we look to determine the optimal model for forecasting confirmed cases within the United States.

The time series data was divided into a training and testing subset where the size of the test subset was equal to ten percent of the total number of days within the initial data set. Therefore, the size of the training set was determined to be 100 days and the size of the testing set was 12 days. The identification of the optimal ARIMA model order, which is based upon discussed parameters, was assessed using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These two objective metrics of model suitability seek to leverage the trade off between the fit of the model and its complexity. These two criterion differ based on penalization that occurs due to increasing the complexity of the model. Therefore, given a training set of observations defined Y = (Y(0), Y(1),...,Y(n), where n is the size of the training set, we define M_j(Y) to be the maximum value of likelihood for the jth model such that k_j is defined to be the number of free parameters within the jth model [20]. Thus, we look to choose the model that minimizes both the AIC [Equation (1)] and BIC [Equation (2)].

$$\text{AIC} = -2lnM_j(Y(0)..., Y(n)) + 2k_j \tag{1}$$

$$\text{BIC} = -2lnM_j(Y(0)..., Y(n)) + k_j lnn \tag{2}$$

It is important to note the fact these these information criterion for optimizing model performance are not strong indications of the appropriate order of time series difference, parameter d, within the ARIMA model. This is due to the definition of differencing such that the value of parameter d changes the data which is used for computation of the maximum value of likelihood of the model, intrinsic to the definition of both discussed criterion.

Due to the inherent importance of time series forecasting and its application to real-world scenarios, numerous performance metrics are necessary to estimate forecast accuracy and aid in the comparison of different, unique models. The selected metrics were chosen based off their ability to measure both bias and accuracy of forecast performance. These performance metrics are measured as a function of the expected and observed values within the time series. Forecast performance metric notation is defined such that y_t is the observed value at time t, $\hat{y_t}$ is the expected or forecast value at time t, and $e_t = y_t - \hat{y_t}$ is the forecast error where n is the size of the test set. The two forecast evaluation metrics,RMSE and MAPE, are described below in Equations (3) and (4) and were chosen through appropriate literature [33]. Strong forecasts will seek to minimize these metrics and observation of inherent properties related to Equations (3) and (4) will be evaluated to properly analyze forecast performance [30].

$$\text{Root Mean Squared Error(RMSE)} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \tag{3}$$

$$\text{Mean Absolute Percentage Error(MAPE)} = \frac{1}{n}\sum_{t=1}^{n} |\frac{e_t}{y_t}| \times 100 \tag{4}$$

3 Results

3.1 Results: SEIRD

Initially, we look to examine the SEIRD curve using certain parameters drawn from model literature to assess transmission characteristics related to 2019-nCoV and the affect of social distancing using the varying parameter λ. Given a defined population of 10,000 individuals for this simulation, varying values of λ were employed to simulate no social distancing ($\lambda = 1$), one third of the population practicing social distancing ($\lambda = .66$) and finally two thirds of the population in quarantine($\lambda = .33$). Figure 1 seeks to visualize the affect social distancing has on reducing transmission of an infectious disease, specifically the novel coronavirus. It is clear from the curves that flattening of the exposed and infected case numbers occurs for smaller values of λ. This quarantine affect is able to push the peak infectious curve further into the future allowing for proper healthcare and surveillance procedures to be put in place.

After examining the effect of social distancing efforts, we look to estimate the basic reproductive number from the data. A least squares fitting procedure was implemented by fitting the cumulative number of cases, defined C(t), in the model to the cumulative number of confirmed cases in the United States. The latent or incubation period was fixed to 1/k = 4.8 [24] while the rate of recovery (γ) was initially set to 0.2 with a minimum possible value of 0 and maximum possible value of 1. Similarly, δ was initially set to 0.2 and optimized in range [0,1]. The total population size N was fixed to the population of the United States [8] with other compartments initialized at the initial values of the physical data. The least squares fitting procedure estimated the value of $\beta = 2.99$ and when inserted into the equation above, the value of the basic reproductive number was calculated to be $R_0 = 2.13$. The results of the fitted and physical results are observed in Figure (2). This figure illustrates certain limitations of the simple SEIRD model.

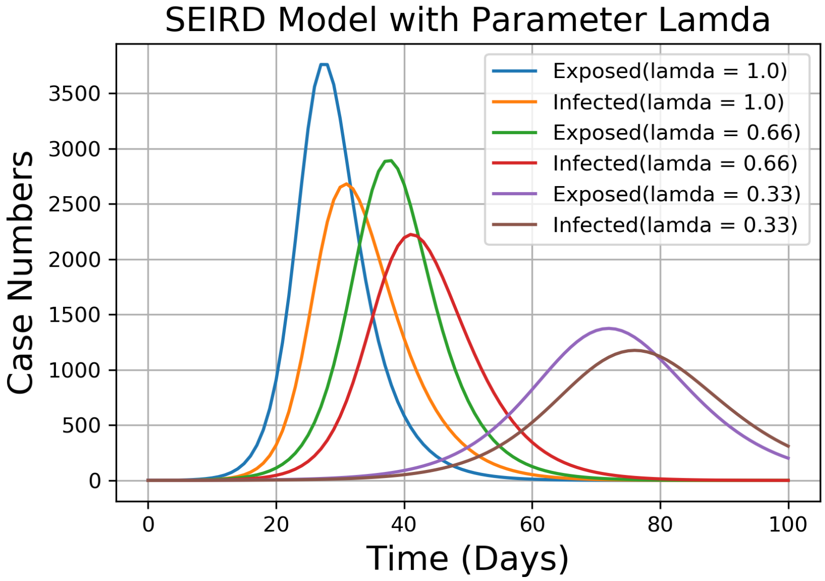


Figure (1) Evaluation of exposed and infectious populations using the SEIRD model with the inclusion of varying values of the social distancing parameter λ

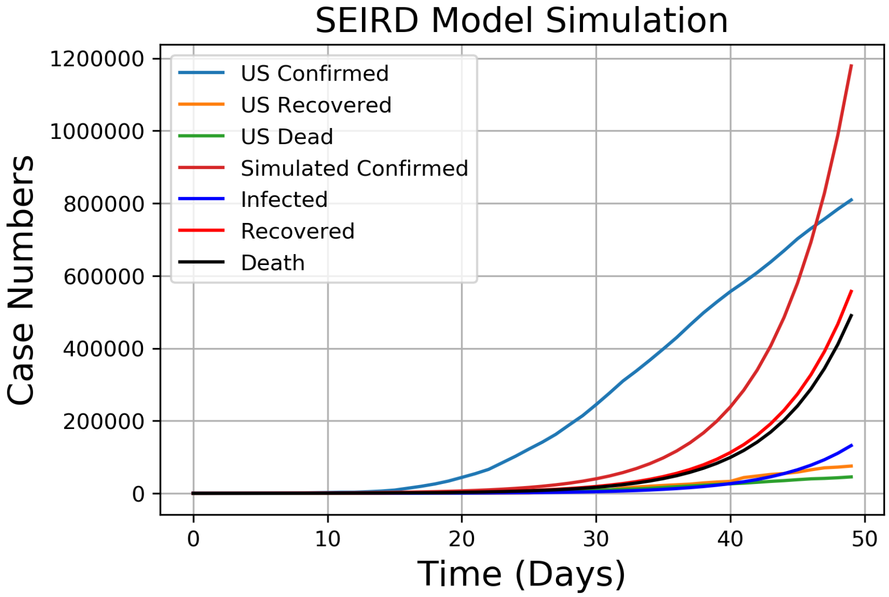
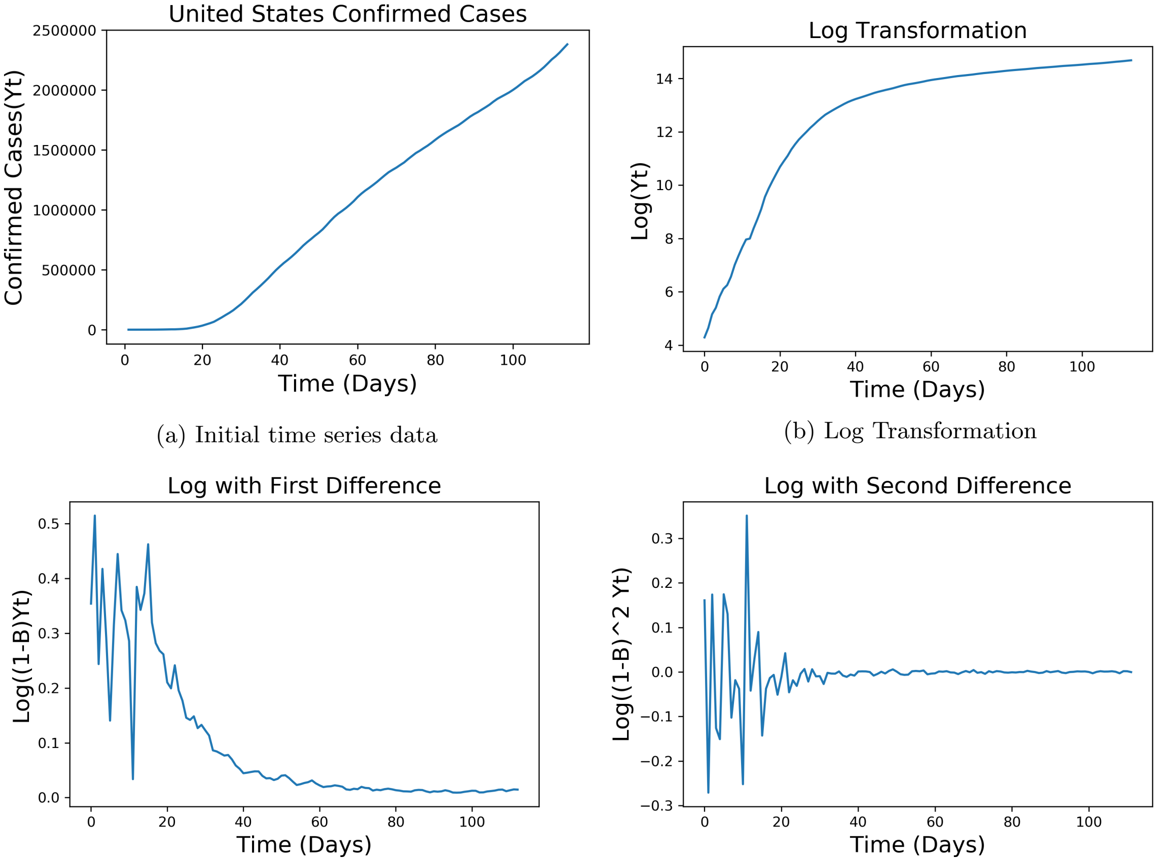


Figure (2) Comparison of the fitted SEIRD model with optimized parameters to the cumulative confirmed case notifications of 2019-nCoV in the United States.

3.2 Results: ARIMA

Initial diagnostics of the time series data demonstrated a clear lack of stationary behavior due to a constantly increasing trend (Figure (3a)). This behavior was expected due to the intrinsic growth rate related to infectious disease. Numerous transformations and differencing strategies were applied to the initial data to obtain and verify stationary behavior within the data. These strategies included taking the logarithm of the initial time series, and both the first and second-order differences of the transformed, logarithmic data. The initial logarithmic transformation (Figure 3b) and logarithmic transformation with first-order differencing (Figure 3c) both failed to obtain stationary behavior of the time series data. The logarithmic transformation combined with second-order differencing (Figure 3d) appears to show stationary behavior.



(a) Initial time series data

(b) Log Transformation

(c) Log Transformation with first-order differencing

(d) Log Transformation with Second-Order Differencing

Figure (3) Different strategies of time series manipulation to identify stationary behavior. Figure (a) visualizes the growth of confirmed cases within the United States. Figure (b),(c),(d) represent different possible transformations to Figure (a) to remove trends/seasonality from the data.

Statistical hypothesis tests of stationarity within the data were applied to determine whether the transformation of the time series was successful in obtaining stationary behavior. Two unit root tests, the Augmented Dickey-Fuller(ADF) and Kwiatkowski-Phillips-Schmidt-Shin(KPSS) test, were implemented with opposite null and alternative hypothesis. The null and alternative hypothesis for the ADF test are defined such that:

$$H_0 : \text{The series has a unit root.}$$

$$H_a : \text{The series does not have a unit root. The series is stationary.}$$

Alternatively, the null and alternative hypothesis for the KPSS test are defined opposite when compared to the ADF test: [22].

$$H_0 \text{ :The series does not have a unit root.}$$
$$H_a \text{ :The series has a unit root. The series is not stationary.}$$

The contents of Table (2) summarize the test results of both the ADF and KPSS tests for stationarity on the twice differenced logarithmic series (Figure 1(d)). With a significance level $\alpha = 0.05$ and associated p-value $< \alpha$ for the ADF test we reject the null hypothesis and assume of the modified time series data to be stationary.Similarly, p-value $> \alpha$ for the KPSS test, we fail to reject the null hypothesis and assume to transformed time series data to be stationary. Therefore, given the results displayed in Table (2), we are able to begin the implementation of decomposition methods such as the ARIMA model.

Variable	ADF Result	KPSS Result
p-value	0.024517	0.1
# Lags Used	13	13
# Observations Used	98	98
Test Statistic	-3.128466	0.0722138
Critical Value	-2.891516	0.146

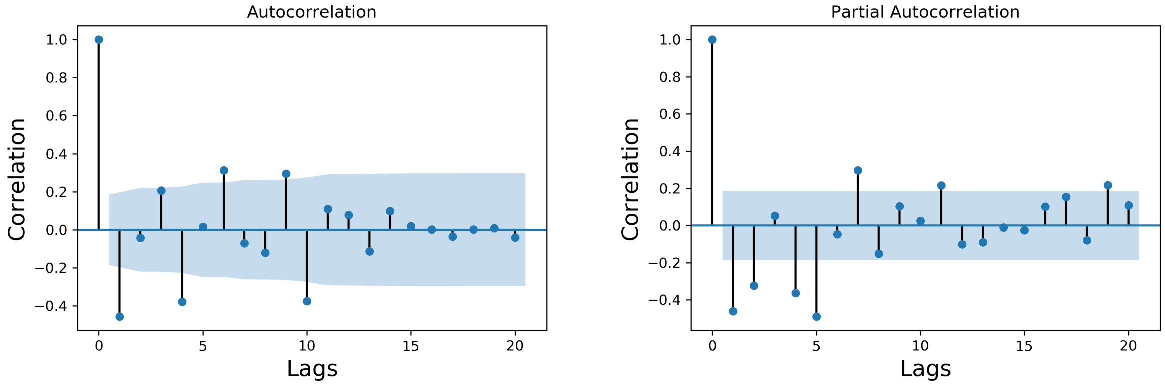
Table (2) Results of Augmented Dickey-Fuller(ADF) and KPSS statistical hypothesis tests for stationary behavior of time series data.

After the application of an appropriate differencing strategy to make the time series stationary, identification of seasonal and non-seasonal orders within the data occurs through the use of auto-correlation functions (ACF) and the similar partial-auto-correlation function (PACF) on the transformed data. The auto-correlation function serves as a tool to measure correlation between values of different temporal occurrence. Similarly, the partial- auto-correlation function quantifies the relationship between the case incidence and its lag that is not explained by correlation at low-order lags. The results shown in Figure (4) reinforce the conclusion drawn from both the ADF and KPSS tests for stationary time series behavior. This is demonstrate by the correlation value dropping to zero very quickly in both Figure (4a) and Figure (4b). Additionally, the ACF and PACF correlograms, found in Figure (4), further illustrate that the necessary ARIMA model parameters are not purely AR or purely MA. We look to examine varying model parameters to identify the best choices of p and q.

The inherent noise within the 2019-nCoV confirmed data outlines the possibility for several strong ARIMA models with varying values of parameters (p,q) to be computed. Therefore, the selection of the optimum model occurred through the evaluation of varying models based upon previously discussed performance metrics, the Akaike Information Criterion (AIC) and Scwartz Bayesian Criterion (SBC).

The optimal model was found to be an ARIMA model with parameter $p = 5$ and parameter $q = 2$. The best choice for parameter d was found to be 2 based upon the discussed stationarity diagnostics above. This ARIMA(5,2,2) model was fit to the training subset of the original data containing 100 observations with associated informative diagnostics AIC = -328.594 and BIC = -305.147. This model was further selected due to the model coefficients being significantly different from zero based on t-test with $\alpha = 0.05$.

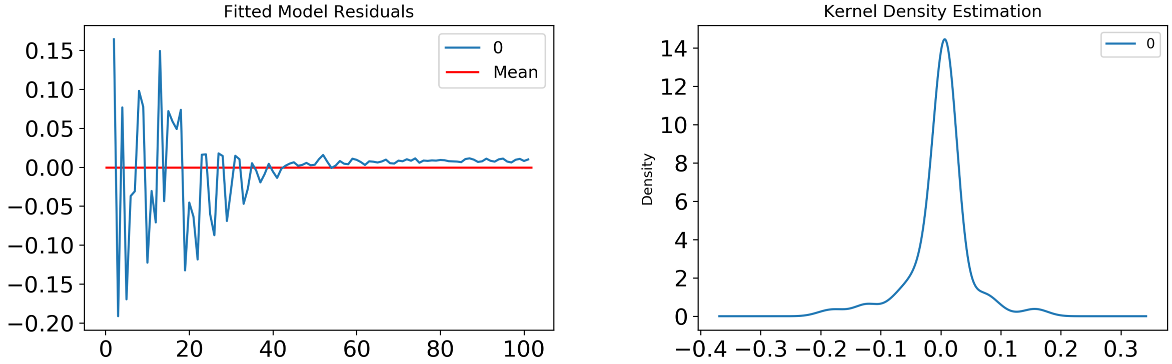
Once we have identified the optimal model, we look to analyze the residuals of the fitted model to check whether our model has captured or included all available information from the data. More specifically, we look to ensure that the model residuals remain uncorrelated with a mean of zero. Additionally, we look to further verify that the fitted model residuals have a constant variance and are normally distributed with mean equal to zero. These additional two properties will make the calculation of prediction intervals easier. The results of Figure (5) verify the required assumptions of residual diagnostics demonstrating a lack of correlation between residuals such that the model residuals are normally distributed with mean of zero.



(a) Auto-correlation Function

(b) Partial Auto-correlation Function

Figure (4) Auto-correlation function and Partial-Auto-correlation function of confirmed cases within the United States using the transformed time series. These correlograms are used to verify stationary behavior and aid in the estimation of ARIMA parameters



(a) Model Residuals

(b) Kernel Density Estimation

Figure (5) Model residual plot and resulting kernel density estimation obtained after fitting the optimized ARIMA model to the training subset of the time series data. The mean value of the residuals (-0.0008) is plotted in red on Figure 5(a)

Post model definition, we then examine the accuracy of the model to forecast the number of confirmed cases in the United States on a held-out test set of 12 days. The results of Figure (6) visualizes the forecast values against the observed values with an associated 95% confidence interval shown in grey. The forecasted values appear to slowly diverge from the observed values over the duration of the test set indicating the strength of the model in smaller, daily predictions. Performance metrics of the forecast were calculated with RMSE = 0.186 and MAPE = 1.026.

Finally, Figure (7) demonstrates the overall results of the optimized ARIMA model including the values of the test set. This figure seeks to illustrate the general accuracy of forecast performance due to the tight confidence interval containing both the observed and forecasted results. This figure is an extension of the visualization shown in Figure (6) to demonstrate the trend of the transformed time series data.

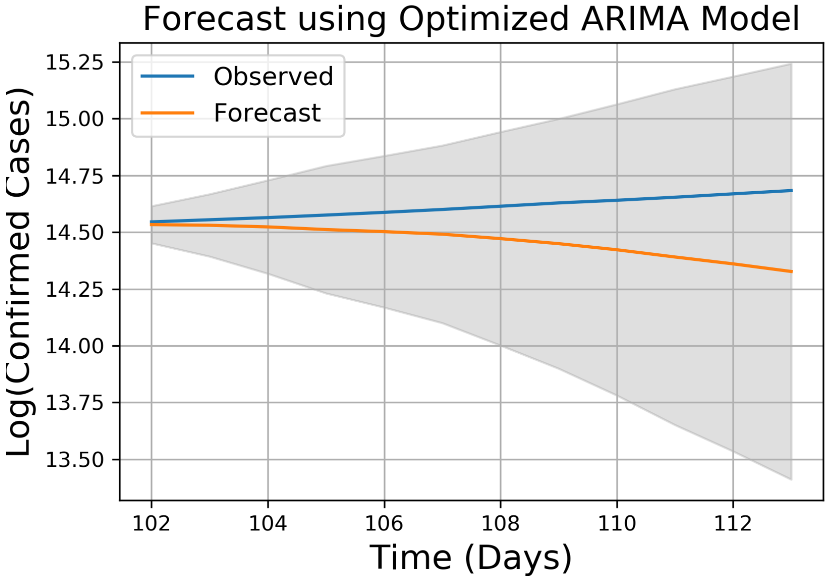


Figure (6) The result of the optimized ARIMA model forecast compared to the held-out test set

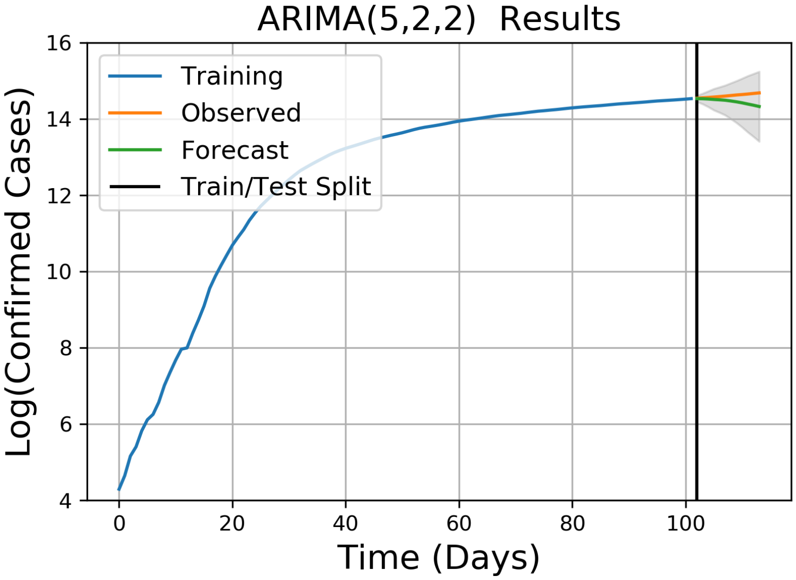


Figure (7) The overall result of model fit and forecast compared to the physical values observed during the novel coronavirus outbreak

4 Discussion

We used two distinct, unique approaches to model the progression and dynamics of the 2019 novel coronavirus in the United States from January 2020 - June 2020. This analysis occurred through the use of daily case notification data relaying information of the cumulative confirmed, recovered, and dead cases recorded over the entirety of the United States.

The first method, a stochastic compartmental SEIRD model, was used to quantify the effect of social distancing measures and estimate the basic reproductive number during the initial exponential growth phase of the epidemic. This type of model served as an exploratory tool into the dynamics of 2019-nCoV and may act as a gateway to more complicated models to be used for infectious disease diagnostics. The result of simulation, using the social distancing parameter λ, demonstrated clear advantages of isolation measures and their potential to delay the peak of the epidemic curve while also reducing the general severity of an epidemic. More specifically, these outbreak control measures provide health-care systems with the opportunity to properly respond to cases of infectious disease. The basic reproductive number quantifies how quickly the novel coronavirus can spread during the exponential growth phase of the outbreak. In our analysis, we used least squares fitting procedure to optimize model parameters to available data. The model estimated $R_0 = 2.13$ which compares nicely to estimates within more complex literature ($R_0 = 2.2$, Kucharski et al.) Analysis limitations occur due to the inherent definition and assumptions related to compartmental models such as the SEIRD. The strong assumptions related to compartmental models may lead to poor inference during application/comparison to physical data [9]. One such limitation is the realistic representation of infectious disease transmission dynamics. One of the most important features of transmission for the novel coronavirus, an infectious case being asymptomatic, is not accounted for in the simple SEIRD model above. A more complex compartmental model would be required to properly diagnose and include such features in a stochastic model. Furthermore, the interpretation of the basic reproductive number derived through the SEIRD model heavily relies on the assumptions that define compartmental models. Particularly, this model assumes homogeneous mixing of the population which in reality does not occur for many reasons such as behavioral changes amongst the susceptible population to avoid contact with infectious agents. Behavior changes have been shown to modify model parameters, contact structure, and the current disease state of individuals [26].

The second method, a statistical-based method of epidemic surveillance, was used to forecast the number of confirmed cases within the United States based on previous values within the time series. The implemented ARIMA model is able to construct valuable information from the historical data of the novel coronavirus epidemic. This occurs through the autoregressive portion considering past values in the times series while the moving average is able to consider current and previous values of the data. The forecast drawn from the fitted ARIMA(5,2,2) model demonstrates a high level of accuracy based on the RMSE and MAPE. The small RMSE value of the fitted model indicates small forecast range relative to the test set due to the emphasis of RMSE in the penalization of extreme errors.

One main advantage of decomposition models such as the ARIMA model is the general interpretability that comes with model definition. Due to the fact that this type of model does not require complex mathematical or statistical calculations, the ARIMA model is much simpler to explain to individuals within, for example, the healthcare domain. These individuals may be more receptive to model forecasts and hold a higher level of confidence in the decisions that this type of model may dictate. This draws contrast to the use of machine learning models of infectious disease that are commonly perceived as 'black boxes', where recent advances have clarified predictive capability[13].

Conversely, certain disadvantages or limitations arise with the use of these type of decomposition methods. For instance, the model hypothesis and/or assumptions may differ from that of the physical infectious disease behavior such that model performance may suffer in certain instances. For example, the ARIMA model extracts linear relationships found within the time series data which may not account for nonlinear relationships of disease occurrence due to unknown factors such as population dynamics.

When analyzing the effectiveness of discussed methods, it is very important to highlight the limitations that occur with the use of daily notification data. First, we acknowledge that many scientific papers investigating infectious disease use aggregated weekly or monthly data to aid in analysis. Daily case reports are defined by smaller numbers and are therefore more susceptible to changes in reporting rates and general population dynamics. Data aggregation is able to highlight more definite trends within the data due to this lack of inherent noise that occurs with daily time series data. This limitation due to the type of data has a direct effect on the ability of the model to accurately forecast the number of confirmed cases in the United States. Another limitation that is apparent with the utilized data occurs due to the accessibility and accuracy of infectious case identification during the early time frame of 2019-nCoV. This inherent change due to population surveillance and detection may have affected the relative accuracy of the data due to under-reporting of the number of confirmed cases in the United States.

5 Conclusion

The occurrence of emergent infectious diseases poses an extreme threat to human health and livelihood. Infectious disease modeling techniques are able to identify important dynamic quantities related to specific viruses, highlight trends, forecast incidence, and even quantify levels of uncertainty for different model forecasts. These tools in combination with the healthcare domain drive decisions that impact the entire United States. In this context, we developed two unique types of model for simulating transmission characteristics, estimating dynamic disease quantities, and forecasting incidence of the novel coronavirus within the United States. Further work regarding accurate forecasting of epidemic events based on daily case notification data should be conducted to enhance the quality and usability of the available data. Similarly, more complex or sophisticated models related to infectious disease forecasting may aid in the estimation of transmission characteristics of infectious disease and help curve the fight against global pandemics. Further possible work within this field include the implementation of a more detailed compartmental model that accounts for more dynamic characteristics of viruses such as 2019-nCoV. This may include the inclusion of probability distributions to define parameter initialization and the addition of more detailed compartments within the model to accurately categorize all individuals within a population. Likewise, interesting work within the field of machine learning models, such as SVMs or neural networks, indicates the importance of their inclusion within the field of epidemiological models. The comparison of different models within this field is important to its advancement in predicting and quantifying emergent infectious diseases.

References

[1] Adam, D. (2020). Special report: The simulations driving the world's response to COVID-19. Nature, 580(7803), 316–318. https://doi.org/10.1038/d41586-020-01003-6

[2] Bettencourt, L. M. A., Ribeiro, R. M., Chowell, G., Lant, T., Castillo-Chavez, C. (2007). Towards Real Time Epidemiology: Data Assimilation, Modeling and Anomaly Detection of Health Surveillance Data Streams. In D. Zeng, I. Gotham, K. Komatsu, C. Lynch, M. Thurmond, D. Madigan, B. Lober, J. Kvach, H. Chen (Eds.), Intelligence and Security Informatics: Biosurveillance (Vol. 4506, pp. 79–90). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72608-1_8

[3] Biggerstaff, M., Kniss, K., Jernigan, D. B., Brammer, L., Bresee, J., Garg, S., Burns, E., Reed, C. (2018). Systematic Assessment of Multiple Routine and Near Real-Time Indicators to Classify the Severity of Influenza Seasons and Pandemics in the United States, 2003–2004 Through 2015–2016. American Journal of Epidemiology, 187(5), 1040–1050. https://doi.org/10.1093/aje/kwx334

[4] Chae, S., Kwon, S., Lee, D. (2018). Predicting Infectious Disease Using Deep Learning and Big Data. International Journal of Environmental Research and Public Health, 15(8), 1596. https://doi.org/10.3390/ijerph15081596

[5] Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases of 2019

novel coronavirus pneumonia in Wuhan, China: A descriptive study. The Lancet, 395(10223), 507–513. https://doi.org/10.1016/S0140-6736(20)30211-7

[6] Chowell, G., Nishiura, H., Bettencourt, L. M. A. (2007). Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. Journal of The Royal Society Interface, 4(12), 155–166. https://doi.org/10.1098/rsif.2006.0161

[7] Cintron-Arias, A., Castillo-Chavez, C., Bettencourt, L. M. A., Lloyd, A., Banks, H. T. (2009). The estimation of the effective reproductive number from disease outbreak data. Mathematical Biosciences and Engineering, 6(2), 261–282. https://doi.org/10.3934/mbe.2009.6.261

[8] Dukic, V., Lopes, H. F., Polson, N. G. (2012). Tracking Epidemics With Google Flu Trends Data and a State-Space SEIR Model. Journal of the American Statistical Association, 107(500), 1410–1426. https://doi.org/10.1080/01621459.2012.713876

[9] Dureau, J., Kalogeropoulos, K., Baguelin, M. (2013). Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. Biostatistics, 14(3), 541–555. https://doi.org/10.1093/biostatistics/kxs052

[10] Erraguntla, M., Zapletal, J., Lawley, M. (2019). Framework for Infectious Disease Analysis: A comprehensive and integrative multi-modeling approach to disease prediction and management. Health Informatics Journal, 25(4), 1170–1187. https://doi.org/10.1177/1460458217747112

[11] Fattah, J., Ezzine, L., Aman, Z., El Moussami, H., Lachhab, A. (2018). Forecasting of demand using ARIMA model. International Journal of Engineering Business Management, 10, 184797901880867. https://doi.org/10.1177/1847979018808673

[12] Fauci, A. S., Morens, D. M. (2012). The Perpetual Challenge of Infectious Diseases. New England Journal of Medicine, 366(5), 454–461. https://doi.org/10.1056/NEJMra1108296

[13] Fountain-Jones, N. M., Machado, G., Carver, S., Packer, C., Recamonde-Mendoza, M., Craft, M. E. (2019). How to make more from exposure data? An integrated machine learning pipeline to predict pathogen exposure. Journal of Animal Ecology, 88(10), 1447–1461. https://doi.org/10.1111/1365-2656.13076

[14] Gambhir, M., Bozio, C., O'Hagan, J. J., Uzicanin, A., Johnson, L. E., Biggerstaff, M., Swerdlow, D. L. (2015). Infectious Disease Modeling Methods as Tools for Informing Response to Novel Influenza Viruses of Unknown Pandemic Potential. Clinical Infectious Diseases, 60(suppl_1), S11–S19. https://doi.org/10.1093/cid/civ083

[15] Halloran, M. E., Longini, I. M. (2014). Emerging, evolving, and established infectious diseases and interventions. Science, 345(6202), 1292–1294. https://doi.org/10.1126/science.1254166

[16] Hethcote, H. W. (1989). Three Basic Epidemiological Models. In S. A. Levin, T. G. Hallam, L. J. Gross (Eds.), Applied Mathematical Ecology (Vol. 18, pp. 119–144). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-61317-3_5

[17] Hethcote, H. W. (2000). The Mathematics of Infectious Diseases. SIAM Review, 42(4), 599–653. https://doi.org/10.1137/S0036144500371907

[18] 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

[19] Imai, C., Armstrong, B., Chalabi, Z., Mangtani, P., Hashizume, M. (2015). Time series regression model for infectious disease and weather. Environmental Research, 142, 319–327. https://doi.org/10.1016/j.envres.2015.06.040

[20] Koehler, A. B., Murphree, E. S. (1988). A Comparison of the Akaike and Schwarz Criteria for Selecting Model Order. Applied Statistics, 37(2), 187. https://doi.org/10.2307/2347338

[21] Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R. M., Sun, F., Jit, M., Munday, J. D., Davies, N., Gimma, A., van Zandvoort, K., Gibbs, H., Hellewell, J., Jarvis, C. I., Clifford, S., Quilty, B. J., Bosse, N. I., . . . Flasche, S. (2020). Early dynamics of transmission and control of COVID-19: A mathematical modelling study. The Lancet Infectious Diseases, S1473309920301444. https://doi.org/10.1016/S1473-3099(20)30144-4

[22] Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics, 54(1–3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

[23] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S. M., Lau, E. H. Y., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., . . . Feng, Z. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. New England Journal of Medicine, 382(13), 1199–1207. https://doi.org/10.1056/NEJMoa2001316

[24] Liu, T., Hu, J., Xiao, J., He, G., Kang, M., Rong, Z., Lin, L., Zhong, H., Huang, Q., Deng, A., Zeng, W., Tan, X., Zeng, S., Zhu, Z., Li, J., Gong, D., Wan, D., Chen, S., Guo, L., . . . Ma, W. (2020). Time-varying transmission dynamics of Novel Coronavirus Pneumonia in China [Preprint]. Systems Biology. https://doi.org/10.1101/2020.01.25.919787

[25] Mallapaty, S. (2020). What the cruise-ship outbreaks reveal about COVID-19. Nature, 580(7801), 18–18. https://doi.org/10.1038/d41586-020-00885-w

[26] Meloni, S., Perra, N., Arenas, A., Gómez, S., Moreno, Y., Vespignani, A. (2011). Modeling human mobility responses to the large-scale spreading of infectious diseases. Scientific Reports, 1(1), 62. https://doi.org/10.1038/srep00062

[27] Metcalf, C. J. E., Lessler, J. (2017). Opportunities and challenges in modeling emerging infectious diseases. Science, 357(6347), 149–152. https://doi.org/10.1126/science.aam8335

[28] Mizumoto, K., Chowell, G. (2020). Transmission potential of the novel coronavirus (COVID-19) onboard the diamond Princess Cruises Ship, 2020. Infectious Disease Modelling, 5, 264–270. https://doi.org/10.1016/j.idm.2020.02.003

[29] Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., Jit, M., Klepac, P., Flasche, S., Clifford, S., Pearson, C. A. B., Munday, J. D., Abbott, S., Gibbs, H., Rosello, A., Quilty, B. J., Jombart, T., Sun, F., Diamond, C., . . . Hellewell, J. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study. The Lancet Public Health, S2468266720300736. https://doi.org/10.1016/S2468-2667(20)30073-6

[30] Ratnadip Adhikari, R. K. Agrawal. (2013). An Introductory Study on Time series Modeling and Forecasting. LAP Lambert Academic Publishing. https://doi.org/10.13140/2.1.2771.8084

[31] Sun, Y., Koh, V., Marimuthu, K., Ng, O. T., Young, B., Vasoo, S., Chan, M., Lee, V. J. M., De, P. P., Barkham, T., Lin, R. T. P., Cook, A. R., Leo, Y. S. (2020). Epidemiological and Clinical Predictors of COVID-19. Clinical Infectious Diseases, ciaa322. https://doi.org/10.1093/cid/ciaa322

[32] Wu, Z., McGoogan, J. M. (2020). Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA, 323(13), 1239. https://doi.org/10.1001/jama.2020.2648

[33] Yang, E., Park, H., Choi, Y., Kim, J., Munkhdalai, L., Musa, I., Ryu, K. (2018). A Simulation-Based Study on the Comparison of Statistical and Time Series Forecasting Methods for Early Detection of Infectious Disease Outbreaks. International Journal of Environmental Research and Public Health, 15(5), 966. https://doi.org/10.3390/ijerph15050966

[34] Zhang, X., Zhang, T., Young, A. A., Li, X. (2014). Applications and Comparisons of Four Time Series Models in Epidemiological Surveillance Data. PLoS ONE, 9(2), e88075. https://doi.org/10.1371/journal.pone.0088075