Research on Covid-19 Infections based on SEIR and Arima models

Jackie Cui, Andy Lu

June 9, 2023

1 Introduction

The COVID-19 pandemic has struck the world many times, causing huge losses and vast hysteria. This study relies on the topic of the COVID pandemic and will be divided into two separate sections.

For Section 1: Effective Reproductive Number and the SEIR model R0 (Basic Reproductive Number) is often used as a caliber measuring the infectivity of epidemics. As explained in the article Epidemic theory (effective and basic reproduction numbers, epidemic thresholds) and techniques for analysis of infectious disease data (construction and use of epidemic curves, generation numbers, exceptional reporting and identification of significant clusters), the Basic Reproductive Number is "the average number of cases per infectious case in a population made up of both susceptible and non-susceptible hosts". The calculated value of R0 is often compared with the number 1: if R0¿1, an infector could transmit the virus to more than one person on average, likely causing a dramatic outbreak; if R0=1, the disease is endemic; if R0¡1, an infector will transmit disease to less than one person on average, which means that the disease will decline or even be eliminated. R0 has a prerequisite. It requires that there is no external force interfering with the transmission of disease, and nobody in a population is immune to it. Since this situation is barely present in the real world, R0 becomes an ideal measurement of the transmissions, but it is also statistically significant to compare the value of R0 for different epidemics.

The SEIR model is a mathematical model that could analyze pandemic of epidemics. The four letters of SEIR represent the initials of four factors in a pandemic——S represents the number of susceptible individuals, E represents the number of people exposed to a certain disease, I represents the number of people infected by the disease, and R represents the recovered individual (the removed).

For Section 2: The prediction for COVID infection. Time series analysis is the most rapidly developed method for predicting infectious diseases in recent years. A time series is a set of data arranged in chronological order. For example, observations of a variable x (t) or a set of measurements will be made at a series of moments t1, t2... tn, (t is the independent variable and t1¡t2¡t3... ¡tn) is

the sequence set of discrete digit combinations, X(t1), X(t2) ..., X(tn) is the time series. Time series analysis is to predict the future value based on its past value and present value according to the internal relationship between the series. Due to the occurrence, development, and outcome of infectious diseases, the migration of time, the interference of external factors, the mutation of the virus itself, and the constant change of internal factors in the human body have formed a set of unique rules. Time series analysis contains the comprehensive effects of various factors, including unknown factors, affecting the incidence of infectious diseases in the time variable. Due to the outstanding advantages of this method, it has been widely used in the prediction and early warning analysis of infectious diseases. Therefore, this study uses the time series analysis method to model and analyze the epidemic situation in 2023, and predicts the incidence trend of the epidemic to provide a scientific basis for epidemic prevention and control.

2 Background Research

For Section 1: Effective Reproductive Number

To analyze the difference in infectivity between COVID and influenza, many researchers use R(Effective Reproductive Number) as a caliber. As explained in the article Epidemic theory (effective and basic reproduction numbers, epidemic thresholds) and techniques for analysis of infectious disease data (construction and use of epidemic curves, generation numbers, exceptional reporting and identification of significant clusters), the Effective Reproductive Number is "the average number of secondary cases per infectious case in a population made up of both susceptible and non-susceptible hosts". The calculated value of R is often compared with the number 1: if R¿1, an infector could transmit the virus to more than one person on average, likely causing a dramatic outbreak; if R=1, the disease is endemic; if R¡1, an infector will transmit disease to less than one person on average, which means that the disease will decline or even be eliminated.

For Section 2: Time series prediction model

Statistical analysis of time series data has a longer history than forecasting with time series data. Time series analysis originated in 1927 when mathematician Yuel introduced the concept of the AR model for the first time. The model was then used to regulate the economic data obtained by observation and prediction. Subsequently, the MA model was also proposed on the basis of the AR model. Until the prediction work laid down the milestone of time series analysis. Since then, time series analysis has been widely used in the field of engineering, and its theory has gradually developed to a new height. In recent years, with the development of computer technology and information processing technology, the theory and method of time series analysis have been further developed. The steps involved in the modeling process, such as parameter estimation, model recognition, and order determination, can be realized simply and conveniently through computer operation, which is very practical. Early time series analysis

methods usually reveal the law of phenomena changing over time through intuitive historical data comparison or drawing observation, namely the so-called descriptive time series analysis. The application of traditional time series analysis in practice is mainly deterministic time series analysis methods, including the exponential smoothing method, sliding average method, time series decomposition, and so on. However, in real life, the influence of many uncertain factors is getting bigger and bigger. In 1970, Box and Jenkins proposed the time series analysis method based on stochastic theory, which made the time series analysis theory rise to a new height and the accuracy of prediction is also greatly improved. The theory and method of control show great superiority in the processing and analysis of dynamic data, the processing and extraction of complex information, and the prediction of the future. Time series analysis is the use of this set of data, the application of mathematical statistics to process, in order to predict the future development of things. It includes the exponential smoothing method, moving average method (MA), autoregressive model (AR), autoregressive moving average model (ARMA), and the summation autoregressive moving average model, namely ARIMA model, which is widely used at present.

For the study of non-stationary time series, the ARIMA model is a very important one. It uses the difference to transform non-stationary time series into stationary time series, thus transforming the ARIMA process into the ARMA process. ARIMA is widely used in many fields at home and abroad. For example, the ARIMA model was used in the United States to monitor life expectancy in 2010, and it was also used in Australia to analyze the relationship between the incidence of the Ross River virus and climate change. ARIMA model is also widely used in various fields in China. In the medical field, many scholars use this model for epidemic prediction and monitoring. The ARIMA model was also used in this study to analyze and predict the spread of Covid-19.

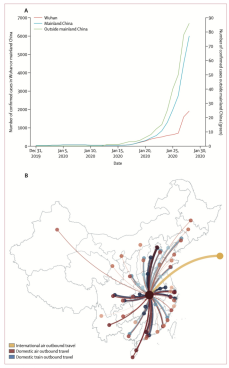
3 Research Methods

3.1 Section1

3.1.1 Data Collection

The study obtained data within the interval starting from Dec. 25, 2019, to Jan. 26, 2020, denoting the starting day as day Ds and the ending day as De. For any epidemic, the source of infection and routes of transmission are critical factors that determine its infectivity. The first case of COVID patients occurred after being exposed to zoonotic sources in Huanan seafood wholesale market, and the Chinese Center for Disease Control and Prevention (CDC) reported that 43 out of 198 confirmed cases were exposed to the source at first. We used the function z(t) to represent the number of people exposed to the zoonotic source. Due to the difficulties of tracing early cases, we assumed that during the last month before the market was closed and disinfected, 86 (twice of 43) people had been exposed to the source of zoonotic infection. After shutting down the market, Z(t) immediately became 0.

Since the starting spot, Wuhan could transmit COVID to other cities in China through air, rail, and road transportation, we used data from the Official Aviation Guide (OAG), Tencent, and Wuhan Municipal Transportation Management Bureau to calculate L, the number of people traveling in and out of Wuhan on a daily basis. Further, LI,W denotes the average international inbound passengers in Greater Wuhan. In reverse, LW,I denote the average international outbound passengers in this area. LI,W=3546 and LW,I =3633. People traveling to and from Hong Kong are excluded because the social unrest in August 2019 caused a huge shrink in Chinese people traveling there. Since other forms of traveling have highly fluctuated, we designate functions to show their values. LW,C(t) is the function measuring the number of all domestic outbound travelers; LC,W(t) measures the number of all domestic inbound travelers.



For other prerequisites or determinants, this research simulates the case of COVID with that of previous epidemics. We assumed that the serial interval of COVID is the same as SARS-CoV, which has a mean of 8.4 days. Also, the incubation period of COVID is similar to that of SARS-CoV and MERS-CoV, with a mean of 6 days.

$$\frac{dS(t)}{dt} = -\frac{S(t)}{N}\left(\frac{R_0}{D_I} I(t) + z(t)\right) + L_{I,W} + L_{C,W}(t) - \left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N}\right) S(t)$$

$$\frac{dE}{dt} = \frac{S(t)}{N}\left(\frac{R_0}{D_I} I(t) + z(t)\right) - \frac{E(t)}{D_E} - \left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N}\right) E(t)$$

$$\frac{dI(t)}{dt} = \frac{E(t)}{D_E} - \frac{I(t)}{D_I} - \left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N}\right) I(t)$$

We will then apply our SEIR model to the data we had. The equations are shown on above.

3.1.2 SEIR Model process

To calculate the value of R0 with a 95 percent confidence interval, we are supposed to solve a likelihood function as shown in Formula(1)(2)(3). A likelihood function shows the joint probability of data considered as a function of the parameters of a statistical model. To calculate the joint probability, all of the probabilities should be multiplied together, resulting in the last equation in (1)(2)(3) below. Then, by taking the derivative of the function, we could find

the critical value(s) where our observed data are the most probable.

However, the actual calculation will be too complex to get the results. Therefore, we used the Markov Chain Monte Carlo method to simulate the case. Monte Carlo is a method for randomly sampling a probability distribution and approximating a desired quantity. It assumes that people could efficiently draw samples from a distribution, and by analyzing the samples, we could get the big picture of the features of the distribution. Markov Chain is a systematic method that produces a sequence of random variables. In the sequence, each current value is probabilistically dependent on the value of the prior variable. Noteworthy, selecting the next variable is only dependent upon the last variable in the chain. Combining the two methods, Markov Chain Monte Carlo randomly samples multi-dimensional probability distribution in which there is a dependence between samples.

$$\lambda(t) = \frac{L_W, I}{N}(E(t) + I(t)) \tag{1}$$

$$\lambda(t) = \int_{d-1}^{d} e^x \, dx \tag{2}$$

$$L(R_0) = \prod_{d=D_S}^{D} \frac{e^- \lambda_d \lambda_d^x d}{x_d} \tag{3}$$

3.1.3 Results

The estimated value of R0 of COVID is 2.68, and the 95

The reported value of R0 of influenza is 1.3 in Wikipedia. Comparing it with our confidence interval, it is far below its lower bound. Therefore, the infectivity of COVID is higher than that of influenza in China. This result shatters the claim that COVID is just a version of influenza from the perspective of infectivity, and it shows the importance of reinforcing protection mechanisms in cities.

3.2 Section2

3.2.1 Theoretical Basis

The Autoregressive Integrated Moving Average model (ARIMA) is widely used in all aspects of influenza prediction. The model with the following structure is an autoregressive summation moving average model

$$\phi(B)(1-B)^d xi = \theta(B)\varepsilon \tag{4}$$

$$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) \tag{5}$$

$$E(x_s \varepsilon_t) = 0 \tag{6}$$

$$\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p \tag{7}$$

$$\theta(B) = 1 - \theta B - ... - \theta_p B^p \tag{8}$$

Formula (7) is the autoregressive coefficient polynomial of the smoothly reversible ARIMA (p, d, q) model. (8) is the moving smooth coefficient polynomial of the smoothly reversible ARIMA (p, d, q) model. "AR" is Autoregression; "p" is the autoregressive coefficient can be estimated by autocorrelation graph; "MA" is the moving average; q is the number of moving average terms, which can be estimated with partial autocorrelation graph; "d" is the order of difference made when the time series becomes stationary.

3.2.2 Establish an Arima model

1. Model Recognition: The basic form of the ARIMA model is ARIMA (p, d, q), "p" is the autoregressive coefficient, "d" is the order of difference made when the time series becomes stationary, "q" is the number of moving average terms. First, perform a stationarity test on the time series, if the sequence is not stationary, use methods such as difference and logarithm to make the sequence stationary. Then use the autocorrelation function (ACF) graph and partial autocorrelation function (PACF) graph to recognize and rank the model. For the order of P and Q, usually try from low-level to high-level, check the fit of each model and compare, generally more than 2 levels are rare.

2. Estimation of Model Parameters and Test of Goodness of Fit: Perform statistical tests on the parameters in the model, and determine whether it is statistically significant. Use Ljung-Box statistics to test whether the residual sequence is a white noise residual. Determine whether the model fully extracts all trend information of the original sequence. At the same time, use the identified Smooth R2, Bayesian Information Criterion (BIC), and other different models to compare the goodness of fit and select the best model.

3.Predict Application: Use the established model to make predictions. This study uses Python for data processing and modeling analysis.

3.2.3 Model Recognition

The time series of the number of new cases in the United States from January 1, 2020, to April 8, 2023, is shown in Figure 1.

The sequence time span is 3 years. At the same time, the sequence itself does not show periodic changes and there is no need to consider seasonal factors. Therefore, a general difference is performed on the sequence to eliminate the trend influence. The original sequence is stationary series by testing the ADF test. The null assumption of ADF test is the existence of unit root. Note that the ADF value is generally negative or positive, but it can only be considered as a significant rejection of the null hypothesis if it is less than 1 percent. If the test statistic is less than the critical value, the null hypothesis is rejected and the series is considered to be stationary. As shown in figure 2, the P-value for ADF test is 0.0049, which is smaller than the significant level, so it is stationary.

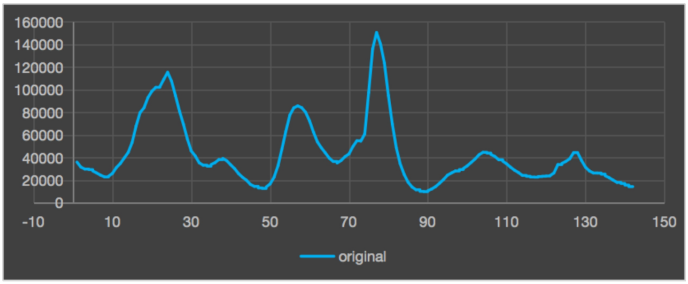


Figure 1: Time Series of the Increasing Number of New Coronary Pneumonia Cases in the United States from 2020 January to 2023 April

```
(-3.6450223117107345, 0.004956390047260187, 6, 140, {'1%': -3.4779446621720114, '5%': -2.8824156122448983, '10%': -2.577901887755102}, 2564.7663499729583)
Test Statistic                  -3.645022
p-value                          0.004956
#Lags Used                       6.000000
Number of Observations Used    140.000000
Critical Value (1%)             -3.477945
Critical Value (5%)             -2.882416
Critical Value (10%)            -2.577902
dtype: float64
是否平稳(1/0): 1
```

Figure 2: ADF test result

Therefore, the series can be regarded as a stationary series. ACF shows obvious tailing which is 5. PACF shows censoring which is 0 (Figure 2). Use maximum likelihood estimation method to estimate the parameters of the model. Combining Akaike information criterion (AIC) and Bayesian Information Criterion (BIC), autocorrelation function graph, partial autocorrelation function graph, etc., the optimal model is MA (0, 0, 5).

3.2.4 Model Parameter Estimation and Model Checking

To ensure continuing modeling, we have to determine whether the coefficient of

	coef	std err	z	p > |z|
ma.L1	2.9136	0.137	21.203	0.000
ma.L2	4.1885	0.391	10.709	0.000
ma.L3	4.0742	0.517	7.876	0.000
ma.L4	2.6697	0.388	6.884	0.000
ma.L5	0.8683	0.134	6.465	0.000

Table1

1. To find the best parameters of the model, we first have to make sure that the series is not a white noise series. Thus, we introduce the Ljung-box test. The null hypothesis is that the values in the series are independent, meaning that the series is an independent identically distributed white noise series. H1 is that the values in the series are correlated, while the series is not a white noise

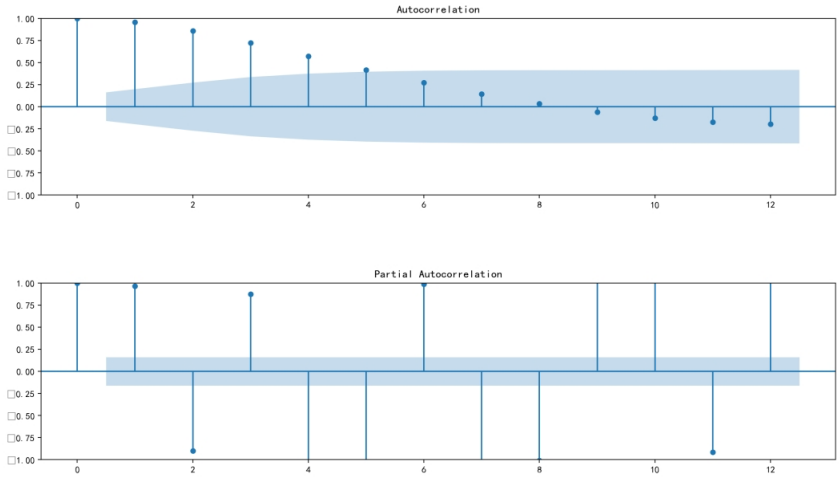


Figure 3: ACF and PACF Diagrams

series.

$$Q_{LB} = n(n+2)\sum_{k=1}^{m}\frac{p_k}{n-k} \tag{9}$$

By computing Q_{LB}

	Q statistics	P-value
Ljung-box	95.16	0.00
Jarque-Bera	3.71	0.16

Table 2

According to table 2, We could see that the p-value is zero, meaning that we reject H0, which indicates that the series is not a white noise series

2. The second test is the Jarque-Bera test, which is the residuals test, Residuals are useful in checking whether the prediction model fully captures the information in the data. The residuals of a good forecasting method have the following characteristics: (1) There is no correlation between Residuals: If there is a correlation between residuals, there is still information in the residual that can be used in the forecast. (2) The mean of Residual is 0: If the Residual mean is not 0, then there is bias in the forecast. The null hypothesis test is that the residual is normally distributed; H1: The residual is not the normal distribution. In our model, the residual is normally distributed and the mean is 0, as shown in Table 2, and the p-value is 0.16, which indicates that we fail to reject H0, the residual is normally distributed, and the mean is 0 as shown in figure 4.

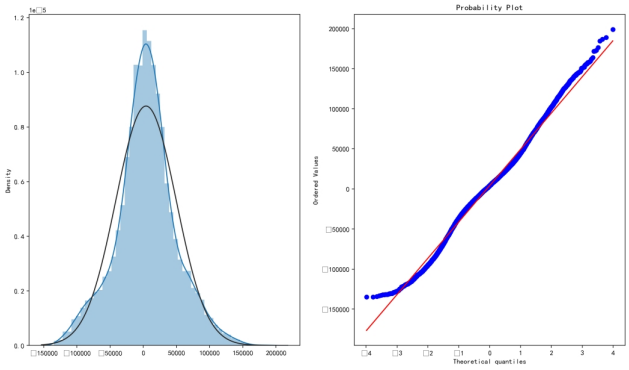


Figure 4: Residual distribution

3.2.5 Evaluation results

	MAE	MSE	R	R^2	RMSE
values	5941	11026	0.94	0.89	10500

Table 3

In table 3, we can see that the R^2, which is the fitness of modeling, is 0.89, this indicates that our model is well-fit for the prediction. It predicts a decent and believable result.

3.2.6 Forecast and Evaluation

Select the optimal model of ARIMA (0,0,5) as shown in Table 1. Use the calculation formula of the predicted value in the ARIMA (p,d,q) prediction model, according to the parameters given by Python, the calculation formula of predicted data is obtained 5. The predicted data and actual data of the number of newly diagnosed people in the United States from April 15 to April 28 are calculated from Formula (10) as shown in Table 4.

$$y(t) = 43260.11+\varepsilon_t+2.9136\varepsilon_{t-1}+4.1885\varepsilon_{t-2}+4.0742\varepsilon_{t-3}+2.6697\varepsilon_{t-4}+0.8683\varepsilon_{t-5} \tag{10}$$

According to Table 4, the overall dynamics predicted by the model are basically consistent with the actual situation, the model makes a good prediction of the changing trend of the number of cases in the short term in the future.

week	True data	predicted	UCL	LCL	absolute error
22-Apr	11779	10758.50465	19198.93934	2318.073824	1020.495346
29-Apr	10150	7853.223587	16293.65442	-587.2072426	2296.776413

Table 4

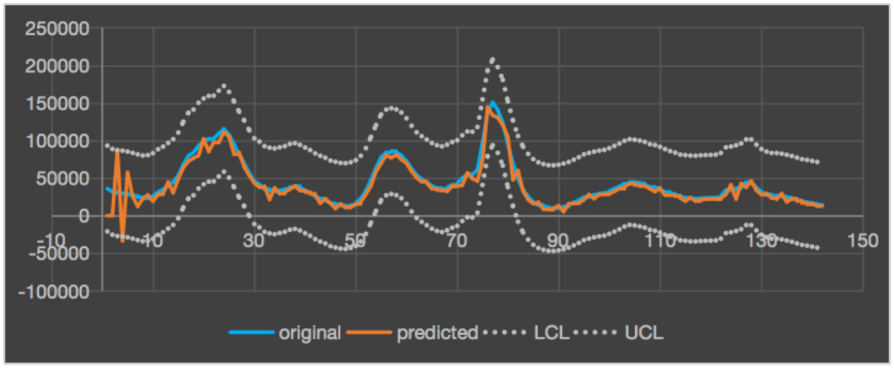


Figure 5: Fitting and Prediction Results of ARIMA (0, 0, 5) Model on the Number of Newly Confirmed COVID-19 Cases in the United States

4 Conclusion and Future expectation

Since all the case numbers and details are locked by the Chinese government at the beginning of 2023, the data we could use to apply our model is limited, and data after Jan. 2020 might affect the value of R0 as well due to herd immunity and policy adjustments. We suggest that further research could focus on diseases that are more traceable than COVID.

The ARIMA model is one of the most commonly used time series models. Not only the dependence on the time series is considered but also the interference of random fluctuations. Therefore, the prediction accuracy of short-term trends is relatively high, which is often used in the fields of epidemic trend prediction of infectious diseases and non-communicable diseases. In this article, the number of new confirmed cases in the United States from April 15 to April 28, 2023 were used to establish prediction models of new cases, It is shown in the results that the ARIMA (0, 0, 5) optimal model was closer to the actual observed value for the number of new cases. According to the figure 5, we can see that the fitness of predicted and actual value is close and the value is all in the UCL and LCL. Also, the respected error was only about 8 percent. the Short-term forecasts can be made very well by the ARIMA model. Compared with the cumbersome degree of the traditional SEIR model, the ARIMA is more accurate for forecasting results and the forecasting method is simpler. And the ARIMA model requires too high for residual white noise testing to get the appropriate data, resulting in fewer predictable indicators. For the future expectation, We would make the data processing, and according to the residual plot for our model, there is still some parameters that we need to figure out and fix it. Moreover, the fitness of the model still needs to improve, the R^2 currently is still not so ideal which we need to improve in the future. On the other hand, for the covid-19 cases analysis, we will combine SEIR and Arima models together and make the predication, we will make the predicted data more believable.

References

[1] Wang LS, Wang YR, Ye DW, Liu QQ. A review of the 2019 novel coronavirus (COVID-19) based on current evidence.Int J Antimicrob Agents.2020;55:105948. doi:10.1016/j.ijantimicag.2020.105948.[PMC free article][PubMed] [CrossRef][Google Scholar]

[2] Mohamed A R, Najem Moussa, Abdellah Madani, Abdessadak Aaroud, Khalid Zine-dine.Forecasting Covid-19 Transmission with ARIMA and LSTM Techniques in Morocco. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8758931/

[3] Epidemic Theory (Effective and Basic Reproduction Numbers, Epidemic Thresholds) and Techniques for Analysis of Infectious Disease Data (Construction and Use of Epidemic Curves, Generation Numbers, Exceptional Reporting and Identification of Significant Clusters) and Health Knowledge. www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/epidemic-theory.

[4] Chinese Center for Disease Control and Prevention. www.chinacdc.cn

[5] Wu, Joseph T., et al. "Nowcasting and Forecasting the Potential Domestic and International Spread of the 2019-nCoV Outbreak Originating in Wuhan, China: A Modeling Study." Obstetrical and Gynecological Survey, vol. 75, no. 7, Lippincott Williams and Wilkins, July 2020, pp. 399–400. https://doi.org/10.1097/01.ogx.0000688032.41075.a8.

[6] Wikipedia contributors. "Basic Reproduction Number." Wikipedia, May 2023.

[7] CDC. "COVID Data Tracker." Centers for Disease Control and Prevention, 28 Mar. 2020, covid.cdc.gov/covid-data-tracker/trendsweeklyhospitaladmissionsselect00.