Analysis on the Time Series Forecasting on COVID-19 New Infections Declining in Shanghai, China

Peicheng Ji

Department of Biology, Case Western Reserve University, Ohio, United States

pxj161@case.edu

Abstract. This research focus on the forecast of the effect of policy influence on COVID-19 in Shanghai, China, during March to May in 2022. By using new infection numbers in 50 days as the train model, four different models were used for forecasting the new infection numbers in future 14 days. Data was published on China Centers for Disease Control and Prevention. Simple exponential smoothing, Holt's linear method, ARIMA model and ETS model were utilized and compared the accuracy with each other to find the best fit forecast model. As the result, the ARIMA (2,2,2,) model best fits the decline of the new infection. This result needs a condition of similar infectiousness to COVID-19 and the policies and efforts similar to those used to control the outbreak in Shanghai from March to May 2022. This research helps the forecast for policy effects in future epidemiologic disease control.

Keywords: Time series, COVID-19, Forecast, Epidemiology.

1. Introduction

1.1. Research Background and Significance

COVID-19, official named as SARS-CoV-2, was firstly discovered and outbreak in Wuhan province in China in 2019 [1]. During 2002-2003, the coronavirus had association with serve acute respiratory syndrome, which is known as SARS [2]. Research in February 2020 in China summarized that the case-fatality rate at that time was 2.3% [3]. After an outbreak of COVID-19 in China, the coronavirus epidemic has also begun to sweep the world, which has already caused more than 676 million cases and near 7 million deaths [4]. Research on COVID-19 is extremely important and necessary. Research on COVID-19 will help to better build the cognition on the characteristics of the epidemic disease, which can help governments develop preventive measures to protect the public. Policies need to be formulated using scientific research as data to weigh the pros and cons of policy impacts on epidemics and society. Moreover, research on COVID-19 can also provide lessons learned about similar infectious diseases in the future, understanding the rate of transmission as well as the rate of control, which can better assist governments in formulating preventive policies.

1.2. Literature Review

With the increasing severity of new crown epidemics worldwide, research on new crown epidemics is emerging. In terms of forecasting, there have been studies on the forecasting of COVID-19 in many countries around the world. Kumer and Susan have used daily data from the north America, Europe, and middle East to forecast the spread of the COVID-19 and to use the ARIMA model and the Prophet temporal forecasting model to make predictions [5]. Researchers have found that the ARIMA model performed better and had higher accuracy in forecasting the transition of COVID-19. The forecasts have the potential to help the government develop policies to limit the new infections of the COVID-19 virus.

Also, for predictive data, Chimmula used long short-term memory networks for COVID-19 transmission forecast in Canada [6]. The LSTM is a network that utilizes both artificial intelligence (AI) and deep learning to make forecasts. And this research is one of the earliest studies worldwide to utilize this technique to study and forecast the COVID-19 virus spreading. While the results of the study suggesting that June 2020 could be the end point of a COVID-19 outbreak may be far from

reality, applying the techniques of artificial intelligence to the prediction of epidemics was a novel and worthwhile idea at the time.

Even if forecasting is set aside, the time series itself is indispensable for the study of COVID-19. In 2020, a study was conducted to analyze the time series recording for COVID-19 spreading rate in the mainland of the People's Republic of China [7]. The analysis proved that the infection of COVID-19 was affected by both temperature and humidity. When the temperature and humidity are higher, the COVID-19 infections will be less. Therefore, countries or regions with low temperatures or humidity need to be more protected against the spread of the virus.

1.3. Research Contents and Framework

Since there are many studies that have used time series predictive models, this study will also focus on the application of predictive models to outbreaks [8]. The difference is that this study focuses on predicting the decline in the number of new infections in a COVID-19 outbreak within the peak of an outbreak. In this study, multiple prediction models, simple exponential smoothing, Holt's linear method, ARIMA and ETS model, will be used to forecast the spread of the epidemic from March to mid-May in 2022 under the policy intervention in Shanghai, China.

2. Method

This study uses the daily data that was published on China Centers for Disease Control and Prevention (CDC) weekly. Daily new asymptomatic and symptomatic cases are calculated in the time series. The China CDC gives all daily data in 2022. This study uses number of new infections (both symptomatic and asymptomatic) in Shanghai Municipality 50 days from March 1st to forecast new infections in the next 14 days. ARIMA, simple exponential smoothing, Holt's exponential smoothing and ETS models were used to compare the accuracy.

2.1. Simple Exponential Smoothing

Simple exponential smoothing, known as SES, is the simplest model of exponential smoothing method. This method can be used to forecast without a trend or seasonal feature. It uses naïve method; therefore, all the forecast results will be equal to the last value in the training set, which is also the last value of training data from the time series.

$$\check{y}_{T+h|T} = y_T \tag{1}$$

Since the simple exponential smoothing model uses the naïve method, the observed data before does not affect the forecast. The last observed data determines the forecast result.

2.2. Holt's Trend Method

Holt's linear trend method is a forecast model for data with trend but without seasoning feature. It is a special case in ETS models. The forecast equation is:

$$\check{y}_{T+h|T} = l_T + hb_T \tag{2}$$

The model considers both initial level and trend components and forecasts by means of exponential smoothing. It is applicable to linearly trending time series data.

2.3. ETS model

ETS model combines three features of data: error, trend and seasonal pattern. The ETS model can be flexibly adjusted based on the features of the time series, including simple exponential smoothing, which was the model in 2.1, double exponential smoothing, and triple-exponential smoothing. The ETS model can be applied in the absence of significant seasonality, and is suitable for different types of time series data. The model code in ETS function has inputs of "N", "A", "M" and "Z". Input "N" stands for none inputs, while input "A" stands for additive and "M" stands for multiplicative error,

trend or seasonal pattern. Input "Z" stands for automatic selection. The ETS function in R code requires a three-letter input for the error, trend and seasonal pattern separately.

2.4. ARIMA Model

The ARIMA model is a classic time series forecasting method that takes both autoregressive (AR) and moving average (MA) models into consideration with differencing (I) operations. The ARIMA model supposes that future values of the time series data can be affected from past values, lagged terms, and white noise errors. The parameters of the ARIMA model are usually chosen by observing both the ACF and PACF of the time series. With the plot and observation of ACF and PACF graph from the training data, the parameters of the ARIMA model can be chosen and input. The ARIMA model is a common forecast tool for data with some degree of trend and seasonality.

The autoregressive in ARIMA model indicates the correlation between observations at the current moment and observations at past moments. The AR part takes into account the autocorrelation of the time series and denotes the order by *p*, which means that *p* observations at past moments are used to predict the observations at the current moment.

Difference is used to deal with non-stationary time series. By differencing a time series, a non-stationary series can be reformed and considered as a stationary time series, thus making the modeling more reliable. d denotes the number of differencing.

Moving average indicates that there is a correlation between the observations at the current moment and the random errors of the past moments.MA partially takes into account the lag error of the time series and denotes the order by q, which means that q random errors of the past moments are used to predict the observations at the current moment.

The accuracy measurement will be depending on RMSE, MAE and MAPE by accuracy function in R code. RMSE stands for root mean square error. This means the square root of the average of the squares of the prediction error of forecast. MAE stands for mean absolute error, which indicates the average absolute value of the prediction error. And the MAPE value means the mean absolute percentage error, which is the average absolute percentage of the prediction error. As an assessment of the accuracy of the forecast models, smaller RMSE, MAE and MAPE value represent better forecast models.

3. Result

The daily data of new COVID-19 infections from March 1, 2022 till May 31, 2022 in Shanghai is shown in the figure 1.

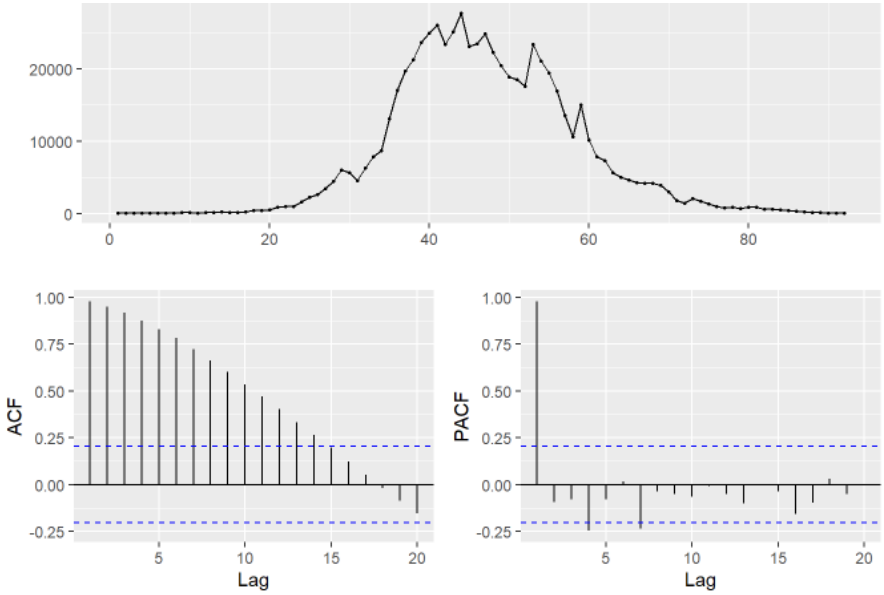


Fig. 1 Daily reported infection in Shanghai

The Figure 1 also contains the autocorrelation function and partial autocorrelation function.

Simple exponential smoothing model, Holt's method, ETS model and ARIMA models were used for the forecast. Table 1 shows the accuracy results for each model.

Table 1. Accuracy result of SES, Holt, ETS and ARIMA models

Fitted models	RMSE	MAE	MAPE
SES	1546.131	975.6518	24.73259
Holt	1426.167	826.1401	22.22041
ETS	1622.597	934.8589	21.46116
ARIMA (1,1,1)	1378.882	818.9508	24.14777
ARIMA (2,2,2)	1226.123	705.2289	22.55816

Among all the models. The ARIMA (2,2,2) model has the lowest RMSE and MAE. Because of the larger actual value of the data itself, MAPE will be considered with less priority than RMSE and MAE. Although the ARIMA (2,2,2) does not have the lowest MAPE value, the data from accuracy function proves that the ARIMA model best fits the forecast among these models (in Figure 2).

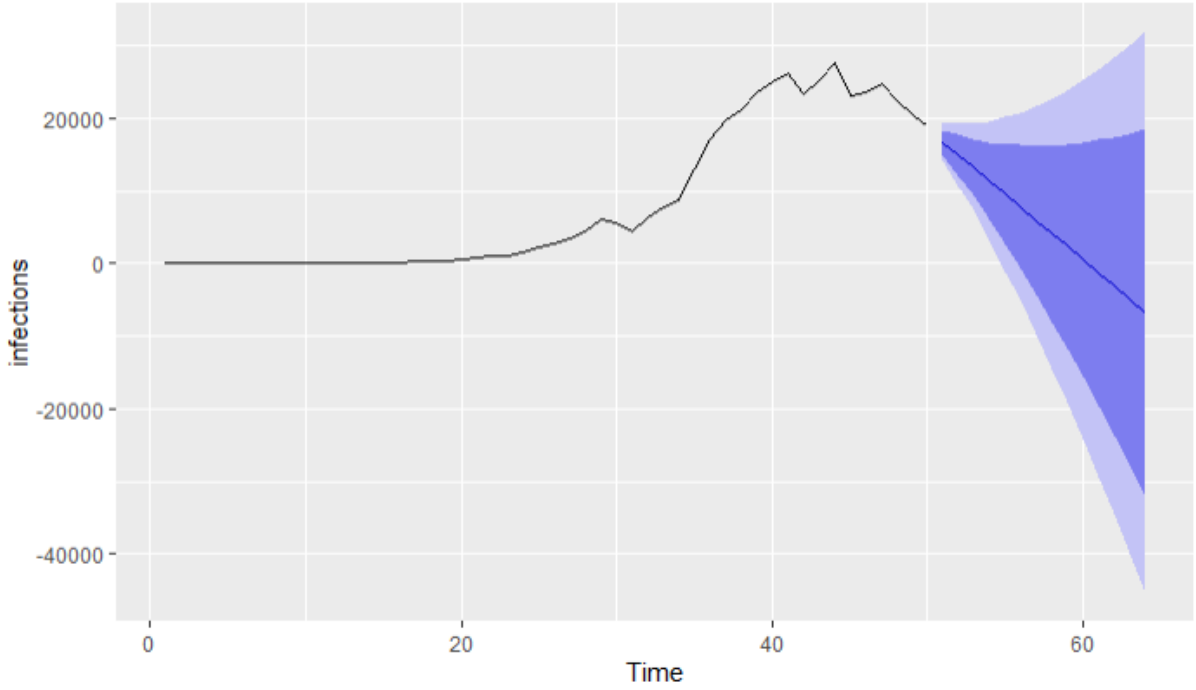


Fig. 2 Forecasts from ARIMA (2,2,2) model with 80% and 95% confidence interval

4. Discussion

ARIMA Models can be used in outbreak forecasting to analyze historical outbreak data and predict future trends and magnitude of outbreak spread. ARIMA model fits the time series data with certain trends and seasonality. In outbreak forecasting, ARIMA models can help public health departments, government policy makers and researchers to better understand the trend of outbreaks so that more targeted prevention and control measures can be taken.

The data selected this time corresponds to a peak infection in Shanghai in 2022. The corresponding forecast part is exactly the stage where the number of infections decreases after the policy control.

The ARIMA model has been used in the prediction of COVID-19 outbreaks in many other countries and regions, such as the Makkah region of Saudi Arabia and India [9,10]. And this literature is mainly used to explore what model should be utilized for forecasting the trend of decreasing number of COVID-19 virus new infections under policy control.

Of the multiple models chosen for this exercise, although ARIMA (2,2,2) had the lowest RMSE and MAE results, it was still not accurate enough. And the downward trend depends on the strength

of government control. If the strength is weak and does not force citizens to restrain their traveling, then the number of new infections may not decrease in a short period of time. If a very strong control policy is published, the spreading of the COVID-19 virus will be less transmissible, which can be expressed as the number of new infections may decline with a fast speed. Like how the ETS model performed in this forecast. It gave a more modestly declining result (in Figure 3).

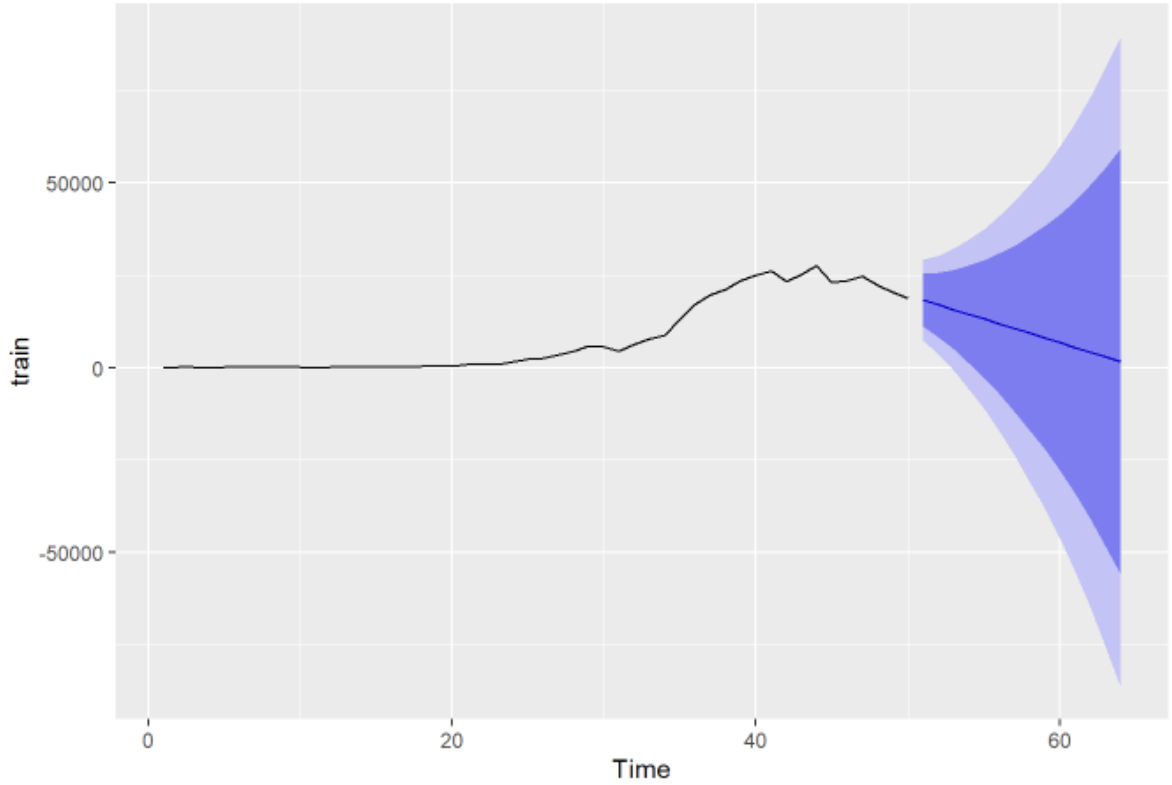


Fig. 3 Forecasts from ETS (M,A,N) with 80% and 95% confidence interval

In considering the modeling of ARIMA, this study starts with the auto. arima function. it gives the result of ARIMA (1,1,1). After getting the result of ARIMA (1,1,1), the authors refer to another article for an example using ARIMA (2,2,2) [11]. It was eventually found that in terms of accuracy, ARIMA (2,2,2) did outperform ARIMA (1,1,1) in this prediction model (in Table 2).

Table 2. Result of auto.arima function

Seires	Train	
ARIMA(1,1,1)		
Coefficients	ar1	ma1
	-0.3877	0.8471
s.e.	0.1756	0.0968
Sigma^2	2022677	
Log likelihood	-424.56	
AIC	855.12	
AICc	855.66	
BIC	860.8	

The advantage of the ARIMA model is that the short-term forecasts with ARIMA model are more accurate. It is therefore more accurate in forecasts that are in the frequency of "days". However, the ARIMA model may fall short if the time period covered by the training model is extended, or if the prediction is made for a longer period of time.

The time in this study covers only one peak in a two-month period and only in the Shanghai area. Therefore, the effect of season was excluded. If the observation is extended to two years, the effects

of season and virus strain need to be considered. Since ARIMA is suitable for predicting outcomes in a stationarity time series, the accuracy of the predictions may be improved by using more observations. Longer training days would give more accurate results, such as using six months of data to train and forecast the results for two weeks.

The most important of these is still the combination of specific and practical local policies. The tightness of policies largely determines the trend in the number of new infections. For example, if an area's policy changes from mandatory to non-requirement of nucleic acid testing, the number of infections may appear to be lower, but in fact fewer people are getting tested.

Therefore, the research in this article is directed towards the prediction of regulatory efforts in the Shanghai municipality-wide from March 1, 2022 to May 31, 2022, and the control of diseases with similar infectious intensity to the strain of COVID-19 at that time. Future studies can compare the control of the same outbreak in different countries or regions within the same country, or the control of the same outbreak in the same region over time. Seasonal, geographic, and past experience may be taken into account in this process, and the predictions of the ARIMA model may give the government some useful predictions and expectations.

5. Conclusion

COVID-19 outbreaks, as large epidemics on a global scale, have been the subject of numerous studies on COVID-19 within the last few years. Records and data on COVID-19 epidemics are consistent with time-series characteristics, and therefore analyses and forecasts of COVID-19 are emerging. Based on the studies and predictions of COVID-19 in other articles, this study used various time series prediction models to analyze and forecast the time series of the number of new COVID-19 infections in Shanghai in the spring of 2022. Since the number of new infections decreased significantly during this time period, the prediction models in this study can be considered as a study of the spread of a similarly infectious epidemic of COVID-19 under policy control. In this paper, simple exponential smoothing, Holt's linear method, ARIMA and ETS models were used to forecast the number of new infections in Shanghai over a 14-day period and compared with the accuracy, The ARIMA model was found to be the best performer among them with the highest accuracy. When the government needs to promulgate new policies to control the spread of epidemics, it can refer to the cities in this paper as well as their time periods to determine the degree of control needed. Therefore, this study is helpful in the control of epidemics. Moreover, this study did not consider the factors of time or virus strains, so future research could focus on the study and analysis of such variables. More importantly, it is important to consider the local epidemic situation and the degree of government management. Control policies may vary in the same area during an epidemic, so it is important to identify the time period in which the policies are implemented. Future studies can use similar methods to compare the control of epidemics in different regions to come up with more appropriate control policies under similar circumstances.

References

[1] Koichi Yuki, Miho Fujiogi, Sophia Koutsogiannaki, COVID-19 pathophysiology: A review, Clinical Immunology, 2020,215, 108427.

[2] Ksiazek Thomas G., Erdman Dean, et al. A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome, New England Journal of Medicine, 2003, 1953-1966.

[3] Wu Zunyou, McGoogan Jennifer M., 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, 2023, 1239-1242.

[4] Covid-19 Map. Johns Hopkins Coronavirus Resource Center, coronavirus. URL: jhu.edu/map.html. Last accessed 2023/8/23.

[5] N. Kumar, S. Susan, COVID-19 Pandemic Prediction using Time Series Forecasting Models, 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020: 1-7.

[6] Vinay Kumar Reddy Chimmula, Lei Zhang. Time series forecasting of COVID-19 transmission in Canada using LSTM networks, Chaos, Solitons & Fractals, 2020,135, 109864.

[7] Hongchao Qi, Shuang Xiao, et al, COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis, Science of The Total Environment, 2020, 728, 138778.

[8] R. Kumar, A. Jain, et al, COVID-19 Outbreak: An Epidemic Analysis using Time Series Prediction Model, 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021: 1090-1094.

[9] N. Darapaneni, D. Reddy, et al, Forecasting of COVID-19 in India Using ARIMA Model, 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2020: 0894-0899.

[10] Awwad Fuad A., Mohamoud Moataz A. Estimating COVID-19 cases in Makkah region of Saudi Arabia: Space-time ARIMA modeling, PLOS ONE, 2021.

[11] Somyanonthanakul, R., Warin, K., Amasiri, W. et al. Forecasting COVID-19 cases using time series modeling and association rule mining. BMC Med Res Methodol 2022, (22), 281.