Time Series Analytics Reading Report

By Ricardo S. Chao 趙上涵 R08546036

Background

Since the beginning of the year 2020, infected cases of the notorious Coronavirus disease 2019 (COVID-2019) have kept surging and surging. According to the data provided by the Center for Systems Science and Engineering (CSSE) at John Hopkins University (JHU), the number of infected cases has reached a new peak of 77 million cases, along with 1.7 million deaths are recorded, by the date of December 19th, 2020. While astonished by the quick spreading and disastrous impact caused by the disease, we're still being curious about the future trend of the evolution of this Coronavirus disease. With the same motivation, the paper we've chosen has proposed an Auto Regressive Integrated Moving Average (ARIMA) model to further model and predict the confirmed cases of the next 10 days in four European countries. The study has also concluded the Auto ARIMA turns out to be an ideal tool when it comes to forecasting the confirmed case of the next 10 days. Moreover, the paper also concluded that for the next 10 days, Spain is expected to have the highest number of new confirmed cases, followed by Germany and France, while Italy is expected to have the lowest number of new confirmed cases among the four countries.

Dataset

The dataset utilized by this paper is directly collected from the GitHub repository (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports), a repository maintained by the Center for Systems Science and Engineering (CSSE) at John Hopkins University (JHU). The paper used data from January 22nd, 2020 to March 28th, 2020 for modeling, which includes 66 samples in total. The four countries chosen among all European countries are based on the selection methodology of selecting the highest daily growth of confirmed COVID-19 cases by calculating the first difference.

Model & Result

To select the best parameters to construct the ARIMA models, the paper has fit several parameter combinations for each four countries. The best parameters combination is selected based on the Akaike Information Criterion (AIC), where parameter combination with the lowest AIC level is often viewed as the best one to apply when constructing the ARIMA model. The details of the model parameters with the corresponding AIC levels are in Table 1. Based on AIC, the paper has

thus chosen the best parameter combination respectively for Spain, Germany, France, and Italy, which are highlighted in gray color inside the table.

Table 1. Country-wise Best Model Selection using auto.ARIMA.[1]

Spain		Germany		France		Italy	
Model	AIC	Model	AIC	Model	AIC	Model	AIC
ARIMA(2,2,2)	Inf	ARIMA(2,2,2)	988.5937	ARIMA(2,2,2)	931.1487	ARIMA(2,1,2) with drift	Inf
ARIMA(0,2,0)	1141.338	ARIMA(0,2,0)	1044.264	ARIMA(0,2,0)	1048.41	ARIMA(0,1,0)with drift	1007.747
ARIMA(1,2,0)	1062.727	ARIMA(1,2,0)	1023.181	ARIMA(1,2,0)	1015.389	ARIMA(1,1,0) with drift	1009.532
ARIMA(0,2,1)	Inf	ARIMA(0,2,1)	993.1037	ARIMA(0,2,1)	985.305	ARIMA(0,1,1) with drift	1009.242
ARIMA(2,2,0)	1051.972	ARIMA(1,2,2)	Inf	ARIMA(1,2,2)	936.8493	ARIMA(0,1,0)	1007.44
ARIMA(3,2,0)	1044.932	ARIMA(2,2,1)	986.5106	ARIMA(2,2,1)	945.2363	ARIMA(1,1,1) with drift	1009.462
ARIMA(4,2,0)	1046.47	ARIMA(1,2,1)	990.2094	ARIMA(3,2,2)	933.5196	Best model	ARIMA(0,1,0)
ARIMA(3,2,1)	1037.447	ARIMA(2,2,0)	997.9065	ARIMA(2,2,3)	Inf		
ARIMA(2,2,1)	1037.541	ARIMA(3,2,1)	988.5688	ARIMA(1,2,1)	966.6439		
ARIMA(4,2,1)	1039.886	ARIMA(3,2,0)	994.1668	ARIMA(1,2,3)	932.3201		
ARIMA(3,2,2)	Inf	ARIMA(3,2,2)	990.9985	ARIMA(3,2,1)	943.4252		
ARIMA(4,2,2)	Inf	Best model	ARIMA(2,2,1)	ARIMA(3,2,3)	Inf		
Best model	ARIMA(3,2,1)			Best model	ARIMA(2,2,2)		

Figure 1. 10-days daily forecast of confirmed cases for Italy (top left), Spain (top right), Germany (bottom left), and France (bottom right).[2]
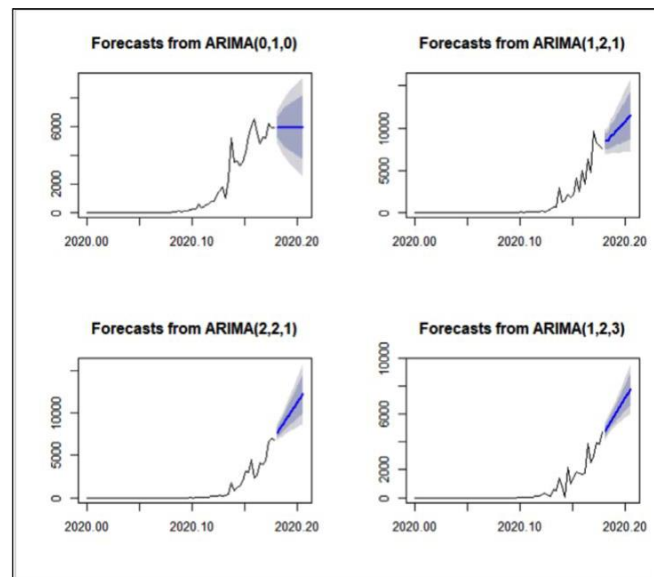
[1] Tahir Mumtaz Awan, Faheem Aslam. (2020). Prediction of daily COVID-19 cases in European countries using automatic ARIMA model. Journal of Public Health Research 2020, 26(8), 228. https://jphres.org/index.php/jphres/article/view/1765/539
[2] Tahir Mumtaz Awan, Faheem Aslam. (2020). Prediction of daily COVID-19 cases in European countries using automatic ARIMA model. Journal of Public Health Research 2020, 26(8), 230. https://jphres.org/index.php/jphres/article/view/1765/539

After model selection, the best fit ARIMA models are then used to forecast the next 10-day confirmed cases of each country, under consideration based on 80% and 95% of confidence level (CI). According to the forecasting result given by the ARIMA models with respective parameters, Spain, Germany, France, and Italy are expected to have an average addition of 11410, 9966, 6937, and 6190 cases per day, respectively. Inside Figure 1, the blue line indicates the forecast value predicted by the models, while the dark and light gray region indicates 95% and 80% confidence interval of the prediction. The paper argues that by predicting the confirmed cases of the next 10 days, the government of each country has thus gained an opportunity to combine the forecast result inside their policy decision process, and probably could then be able to come up with steadier and more complete schemes when it comes to minimizing the damage and impact caused by the diseases, as well as come up with policies to fight against the pandemic.

Critiques and Concluding Remark

To further analyze the results, I would like to first compare the forecast result by the paper's model with the actual data I've collected from the "Our World in Data" website, which is maintained by Oxford University (https://ourworldindata.org/covid-cases). According to the data I've collected, Spain, Germany, France, and Italy have reported an average of 8913, 6714, 5243, 5555 additional confirmed cases per day in the 6-day period after March 28th, 2020. Still, we could somehow observe a certain extent of difference compared with the forecast result by the paper's model, which says that there would be an increase of 11410, 9966, 6937, and 6190 cases per day. Nevertheless, I think it's safe to say that the ARIMA model proposed by this paper has still paved a path when it comes to applying time series models within the decision-making process of policies toward surging pandemic.

Table 2. Numbers of Recorded Cases in Four European Countries[3]

	France	Italy	Spain	Germany
3/28/2020	4645	5974	7516	6824
3/29/2020	2603	5217	6875	4400
3/30/2020	4354	4050	7846	4790
3/31/2020	7629	4053	7967	4923
4/1/2020	4844	4782	8195	6064
4/2/2020	2106	4668	7947	6922
4/3/2020	5224	4585	7134	6365
Average Number of Additional Cases Per Day	5234.17	5554.83	8913.33	6714.67

---

[3] Our World in Data. (2020, December 23th). Coronavirus (COVID-19) Cases. Our World in Data. https://ourworldindata.org/covid-cases

Besides, some other causes may also need to be put into consideration when we are analyzing the results. I've noticed that at the beginning of the pandemic, it seems that the number tallied by the government is not that accurate and some post-revisions have thus been implemented to adjust the number of confirmed cases inside the data. As Table 2 has shown, the French government has corrected their number of confirmed cases record at April 4th and 7th, which shows that the record held by the government of each country might be not that accurate, especially in the beginning of the spreading of the pandemic, and that may undermine the forecasting capability of the ARIMA model constructed by the paper. Furthermore, to stop the pandemic from spreading, plenty of measures have been carried out by the government of each country, and no doubt that these measures would also have an impact on the growth of confirmed cases, and thus may also decrease the predicting capability of the models.

Table 3. Correction on Confirmed Case Record by the French Government[4]

Location	Date	Total Cases	New Cases
France	4/2/2020	59228	2106
France	4/3/2020	64452	5224
France	4/4/2020	47378	**-17074**
France	4/5/2020	48233	855
France	4/6/2020	50887	2654
France	4/7/2020	47396	**-3491**
France	4/8/2020	51251	3855

At last, since we're living in, compare with the series formulated by the model, a far more complicated world, the data are thus not likely to be modeled that easily thanks to the bountiful elements and conditions that may also need to be put into consideration. Thus, maybe a more complicated stochastic model, such as SARIMA or VSARIMA, would be more suitable when it comes to modeling this pandemic series. Also, some other models proposed in the past 10 years, such as the RNN (Recurrent Neural Network) model and LSTM (Long Short-Term Memory) model may perform better when predicting the future trend and the growth of the confirmed cases of COVID-19 pandemic.

[4] Our World in Data. (2020, December 23th). Coronavirus (COVID-19) Cases. Our World in Data. https://ourworldindata.org/covid-cases