Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/13/20230729

A novel multivariate approach for COVID-19 pandemic prediction in the United States

Han Yu
University of Cambridge, Cambridge, U.K.

yuhancam@163.com

Abstract. Since December 2019, the COVID-19 pandemic has caused enormous economic and social disorder. The analysis and forecasts of pandemic has received considerable attention worldwide. Previous research has primarily focused on predictions based merely on historical data and thus has been unable to identify the effect of external factors such as government policy responses. This study aims to develop a multivariate VECM model to predict the change in confirmed cases with Government Stringency Index taken into consideration. This study carried out exponential smoothing and VECM cointegration test using COVID-19 related data of the United States. By exploring the results of the two forecasters, it becomes evident that VECM has shown a much higher accuracy than Exponential Smoothing in the long run. In particular, when huge changes of tendencies in confirmed cases occur, the improvement appears to be more obvious.

Keywords: COVID-19, time-series forecasts, exponential smoothing, VECM.

1. Introduction

The SARS-CoV-2, a type of coronavirus originally discovered in 2019, has caused a novel global outbreak in last three years. On January 30th, 2020, the World Health Organization (WHO) reconvened the Emergency Committee and announced a high global level of risk. Since then, the COVID-19 pandemic has led to enormous damage to the society and the economy. Consequently, collaborative efforts have been made to fight against COVID-19 and bring everything back on track.

Apart from the efforts made to study the medication and prevention from infection, modelling and forecasting the development of the pandemic has also played an important role. Accurate models and predictions are the key to allocate medical resources in an optimal way, which is extremely crucial especially in the early stage of the pandemic where medical resources were in shortage. Therefore, a group of mathematicians and biologists suggested using the SEIR(Susceptible-Exposed-Infectious-Recovered) model, which is a commonly used tool for infectious disease, to evaluate how the pandemic would evolve [1]. Afterwards, when adequate data has been collected, there has been several studies using time-series forecasting to predict epidemic tendencies. In particular, Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing are the ones being in good fit with the actual data [2].

Both SEIR model and univariate time-series forecasting, however, have certain limitations. This is due to the fact that COVID-19 evolution is a very complicated problem with various factors affecting the pandemic. For instance, government policies, vaccination rate and even the number of cases in

DOI: 10.54254/2755-2721/13/20230729

adjacent countries. In the long run, when a dramatic change in one of the factors occur, such as stricter government policies being imposed, the models mentioned above could fail without frequently adjusting hyperparameters. It is hence reasonable to take into consideration the impact of more factors instead of staring at the confirmed cases itself while doing time-series forecasting. In this paper, Vector Error Correction Mechanism (VECM) model will be used to investigate the COVID-19 confirmed cases forecast with some determining factors introduced.

Amongst all factors having an impact on the COVID-19 pandemic, government policies and vaccination rate seem to be of more significance. China, the first country suffering from COVID-19 was instead one of the first few countries successfully controlled their domestic epidemic. In this process, stricter policies such as mandatory face mask and no public gatherings contributed a lot. Nevertheless, it is difficult to quantify the stringency of government policies. The Oxford Coronavirus Government Response Tracker (OxCGRT), has calculated and published their government stringency index, which is a composite measure of nine government response metrics. In this paper, the government stringency index will be imposed for VECM multivariate forecasting and comparisons with exponential smoothing will be made to observe if there are improvements by using a multivariate time-series forecasting model.

2. Data and methodology

2.1. Data and relevant sources

In this paper, the COVID-19 daily confirmed cases in the United States and the Government Stringency Index since April, 2020 will be imposed as training data and the actual data to evaluate the performance of the forecasting models. The source of the data is Our World in Data where the confirmed cases data is originally from John Hopkins University and the Government Stringency Index is originally calculated and collected by the Oxford Coronavirus Government Response Tracker (OxCGRT) [3].

The Government Stringency Index is calculated as the mean score of nine metrics [4]. The metrics used in the calculation include the following: school closures; workplace closures; cancellation of public events; restrictions on public gatherings; closures of public transport; stay-at-home requirements; public information campaigns; restrictions on internal movements; and international travel controls. The calculated value is ranged from 0 to 100, where a higher score indicates stricter policies imposed.

2.2. Model specifications

In this paper, exponential smoothing and VECM will be used to demonstrate univariate and multivariate time-series forecasting models respectively. The general introduction of these two methods will be included in this section.

2.2.1. Exponential smoothing: Holt-Winters' method

The Exponential Smoothing technique is one model which predicts the next value based on a weighted average of all previous values [5]. The model assigns a greater significance of recent values than older values of the time series. The component form of Simple Exponential Smoothing is given by:

Forecasting equation

$$\hat{y}_{t+h|t} = \ell_t \tag{1}$$

and Smoothing equation

$$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1} \tag{2}$$

where $\hat{y}_t$ denotes the predicted value at time t, y_t denotes the actual value at time t in the training data and ℓ_t denotes the level of the time series at time t. $0 \leq \alpha \leq 1$ is just the smoothing parameter, a hyperparameter included in the model.

The Simple Exponential Smoothing does not take into account trend and seasonality of the data. Hence it is suitable for data which exhibits no clear trend or seasonality. Holt later extended the SES model to Holt's linear trend method to reflect the trend in the training data. Holt and Winters then

DOI: 10.54254/2755-2721/13/20230729

extended Holt's method in order to process data with seasonality. The component form of the Holt-Winters additive method equation used in this paper is given by

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)} \tag{3}$$

$$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{4}$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} \tag{5}$$

$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \tag{6}$$

where m is the frequency of the seasonality, k is the integer part of $\frac{h-1}{m}$. b_t denotes the trend component at time t,s_t denotes the season component at time t. β and γ are hyperparameters.

2.2.2. Vector error correction models

Vector Error Correction Models (VECM) models are a special application of VAR or Vector Autoregressive Models, both models can perform multivariate time-series analysis [6]. VECM methodology is used when the variables in the system have a long-run relationship or when they are cointegrated. In other words, VECM is a representation of cointegrated VAR.

VAR model can be specified in the form of VECM by differencing the variables and introducing error correction terms. However, VECM is used only in the presence of cointegrating or long-run relationships. For time series with no cointegration or when the variables are stationary, VAR model should be applied. VAR model can be specified in the matrix form as follows:

$$Y_t = v + \gamma_1 Y_{t-1} + \gamma_2 Y_{t-2} + \cdots + \gamma_\rho Y_{t-\rho} + \mu_t \tag{7}$$

Where Y_t is a vector of K parameters $(K \times 1)$, v is a vector of constants $(K \times 1)$, γ_1 to γ_ρ are matrices of parameters $(K \times K)$ at different lags (lag 1 to ρ) and μ_t is a vector of impulses $(K \times 1)$. When the time-series are non-stationary and cointegrated, VECM should be applied. VECM models allow a wide range of short-run dynamics while using error correction term and cointegrating equations to maintain long-run cointegrating relationships. Long-run equilibrium is reached by gradually making adjustments in multiple partial short-run time periods. The general form of VECM can be specified as follows:

$$\Delta y_t = \mathbf{v} + \Pi Y_{t-1} + \sum_{i=1}^{\rho-1} \boldsymbol{\theta_i} \Delta \mathbf{y_{t-i}} + \boldsymbol{\mu_t} \tag{8}$$

$$\Delta y_t = v + \Pi Y_{t-1} + \theta_1 \Delta y_{t-1} + \theta_2 \Delta y_{t-2} + \cdots + \theta_{\rho-1} \Delta y_{t-(\rho-1)} + \mu_t \tag{9}$$

The rank of the coefficient matrix associated with Yt-1 shows the number of cointegrating vectors ("r") [7]. The number of cointegration terms can be determined using Johansen's test of Cointegration.

2.3. Performance metrics

In this paper, MAPE (mean absolute percentage error) will be used as the performance metric to evaluate the performance of the forecasting models. It represents the average of the absolute percentage errors of each entry in a dataset, showing, on average, how accurate the forecasted quantities were in comparison with the actual quantities. The formula for calculating the MAPE is as follows:

$$MAPE = \sum_i \frac{|y_i - x_i|}{n|y_i|} \tag{10}$$

where y_i is the actual value at point i, x_i is the forecasted value at point i and n is the sample size.

DOI: 10.54254/2755-2721/13/20230729

3. Results and Discussion

3.1. Result obtained from exponential smoothing

There has already been a few studies analyzing COVID-19 using Exponential Smoothing method, most of which obtained some fairly accurate predictions [8]. In order to obtain more clear comparisons between predictions, daily confirmed cases will be analyzed and compared. By observing the data, a weekly periodic pattern is found in daily confirmed cases. Consequently, seasonality has been imposed in exponential smoothing parameters.

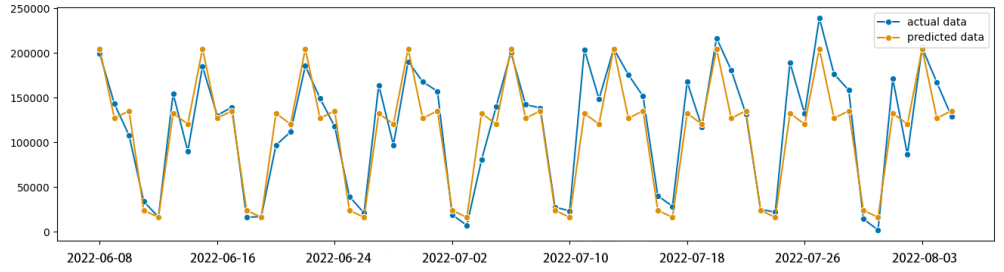

Figure 1. Two-month exponential smoothing starting from June 6th.

In this paper, all the data starting from the very beginning of the COVID-19 pandemic 2020-03-10 till the starting date of forecasts has been imposed as training data for the model. It can be observed from Figure 1 that Exponential Smoothing has achieved prediction within reasonable error from 2022-06-08 to 2022-08-08, the MAPE is calculated to be only 6%.

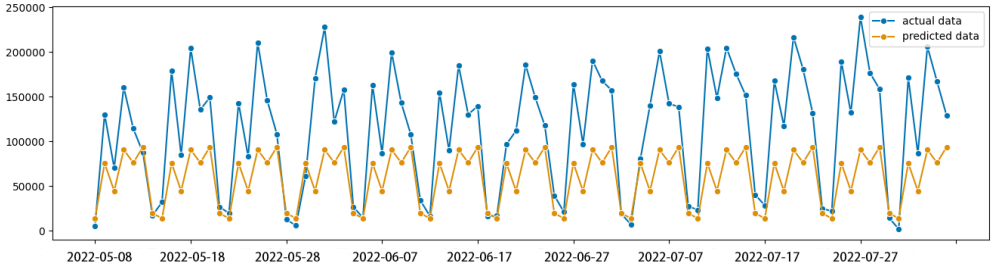

Figure 2. Three-month exponential smoothing starting from May 6th.

A three-month prediction based on exponential smoothing has also been shown on figure 2, from 2022-05-08 to 2022-08-08. This prediction has however, failed with a high MAPE of over 30%. The possible cause for this failure is a lower-down of government stringency index in March from 38.35 to 29.84, which caused a dramatic increase in confirmed cases later in late April. As a univariate forecaster, exponential smoothing failed to predict sudden change in data due to an impulse of factors affecting the current data.

1.1. 3.2 Result obtained from VECM

Figure 3 illustrates the prediction using VECM approach with government stringency index taken into consideration.

DOI: 10.54254/2755-2721/13/20230729

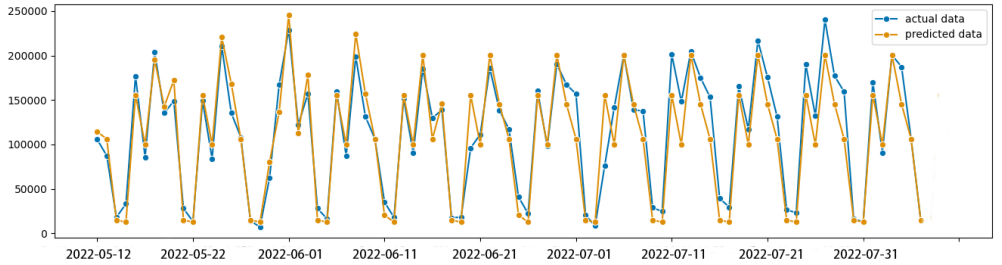

Figure 3. Three-month exponential smoothing starting from May 6th.

It can be observed from figure 3 that within the same three-month period, VECM successfully predicted an increase in the number of confirmed cases. The MAPE is 10.3%. For a prediction dealing with daily confirmed cases with long duration, this MAPE suggests a fairly satisfactory result.

1.2. 3.3 Comparisons between ES and VECM

In order to reach a more convincible conclusion, more time periods have been analyzed with multiple durations and MAPE is calculated and listed in table 1.

Table 1. MAPE of different time periods for ES and VECM.

starting date	end date	ES model	VECM	duration
2022.05.03	2022.05.09	5.1	5.1	weekly
2022.05.10	2022.05.16	5.3	4.6	weekly
2022.05.17	2022.05.23	4.8	5.1	weekly
2022.05.24	2022.05.30	6.2	5.6	weekly
2022.05.31	2022.06.06	5.3	4,8	weekly
2022.06.07	2022.06.13	4.7	4.5	weekly
2022.06.14	2022.06.20	5.2	5.4	weekly
2022.06.21	2022.06.27	4.5	4.6	weekly
2022.06.28	2022.07.04	5.2	4.7	weekly
2022.07.05	2022.07.11	5,3	5,1	weekly
2022.07.12	2022.07.18	4.6	5.1	weekly
2022.07.19	2022.07.25	5.7	5.3	weekly
2022.07.26	2022.08.01	5.2	5.3	weekly
2022.08.02	2022.08.08	6.2	6.1	two months
2021.08.08	2021.10.08	9.1	7.1	two months
2021.10.08	2021.12.08	8.2	7.4	two months
2021.12.08	2022.02.08	17.8	10.1	two months
2022.02.08	2022.04.08	10.2	6.9	two months
2022.04.08	2022.06.08	7.3	6.8	two months
2022.06.08	2022.08.08	5.8	5.6	two months
2021.08.08	2021.11.08	20.7	7.7	three months
2021.11.08	2022.02.08	33.4	9.1	three months
2022.02.08	2022.05.08	15.4	8.1	three months
2022.05.08	2022.08.08	30.2	9.7	three months

By comparing the MAPE of different time periods using Exponential Smoothing and VECM, it is clear that in the short run, Exponential Smoothing and VECM are giving similarly accurate predictions. However, in the long run, despite the rise in error for both model, VECM is giving far more accurate predictions than Exponential Smoothing, with Government Stringency Index taken into consideration.

DOI: 10.54254/2755-2721/13/20230729

4. Conclusion

It can be observed that there is some reasonable error in our predictions, this is due to the effect of a pandemic is an extremely complicated issue. There are multiple factors contributing to the final confirmed cases, such as vaccination, the evolve of the virus and even the climate of a certain place. This paper only considered the effect of government policies using the government stringency index.

In fact, there are multiple factors not taken into account and some factors are hard to quantify. Even the government stringency index used in this paper could merely reflect the stringency of government policies to some extent but does not necessarily give a precise evaluation on the effectiveness of government policies. Finally, like all the other models, adequate training data is required to build a precise forecaster. All the limitations listed above has caused reasonable error for our model. It can be observed from the table that VECM and exponential smoothing have similar predictions in the short run or when the confirmed cases are stable.

However, despite the reasonable error in forecasts, VECM is still giving far more accurate predictions when government stringency index changes and in the long run. Based on the predictions, a precise multivariate forecaster could suggest how an impulse of a factor affect the issue, hence provide guidance for decision making and resource allocation.

References

[1] S. Mwalili, M. Kimathi, V. Ojiambo, D. Gathungu, and R. Mbogo, "SEIR model for COVID-19 dynamics incorporating the environment and social distancing," *BMC Research Notes*, vol. 13, no. 1, p. 352, Jul. 2020, doi: 10.1186/s13104-020-05192-1.

[2] Y. Wang *et al.*, "Prediction and analysis of COVID-19 daily new cases and cumulative cases: Times series forecasting and machine learning models," *BMC Infectious Diseases*, vol. 22, no. 1, p. 495, May 2022, doi: 10.1186/s12879-022-07472-6.

[3] E. Mathieu *et al.*, "Coronavirus pandemic (COVID-19)," *Our World in Data*, 2020.

[4] T. Hale *et al.*, "A global panel database of pandemic policies (oxford COVID-19 government response tracker)," *Nature Human Behaviour*, vol. 5, no. 4, pp. 529–538, Apr. 2021, doi: 10.1038/s41562-021-01079-8.

[5] Hyndman Rob J. and G. Athanasopoulos, *Forecasting: Principles and practice*, 2nd Edition. OTexts.com, 2014, pp. 291 pages;

[6] B. Pfa, "VAR, SVAR and SVEC models: Implementation within r package vars," *Journal of Statistical Software*, vol. 27, pp. 1–32, Jul. 2008.

[7] "Testing for the order of integration," in *Analysis of integrated and cointegrated time series with r*, New York, NY: Springer New York, 2008, pp. 91–105. doi: 10.1007/978-0-387-75967-8_5.

[8] T. Oladunni, S. Tossou, M. Denis, E. Ososanya, and J. Adesina, "A time series analysis and predictive modeling of COVID-19 impacts in the african american community," *medRxiv*, 2021, doi: 10.1101/2021.05.13.21257189.