Towards Understanding the Dynamics of COVID-19: An Approach Based on Polynomial Regression with Adaptive Sliding Windows

Yuxuan Xiu and Wai Kin (Victor) Chan

1 Introduction

Nowadays, COVID-19 is a major threat that all mankind needs to face. Up till now, some countries are approaching the end of the outbreak, thanks to vigorous responses to the epidemic. In other countries, however, the outbreak has not yet been brought under control, or there have been second waves of outbreak. In the current situation, the study of the dynamics of COVID-19 is of great importance. On the one hand, we can compare the impact of different control measures of the epidemic horizontally between different countries. On the other hand, we can refer to the dynamics of the epidemic in different countries to help predict possible future developments in countries where the epidemic has not yet ended.

Intuitively, the dynamics of each wave of COVID-19 can be classified into several stages, such as the early stage, the raising stage and the fading stage. However, the current criteria for partitioning COVID-19 time series are still subjective. Researchers lack quantitative standards to distinguish one stage from another. In other words, a mathematical definition of different stages is still needed. This paper aims at solving this problem by using polynomial regression with adaptive sliding windows as a mathematical standard to partition the COVID-19 time series into different stages.

Existing work on the dynamic analysis of COVID-19 can be broadly classified into the following four categories: epidemic models-based (e.g., SIR, SEIR) methods (Wangping et al., 2020; Calafiore et al., 2020), regression and autoregressive models (Ceylan, 2020; Singh et al., 2020), machine learning-based methods

Y. Xiu · W. K. (Victor) Chan (✉)
Shenzhen Environmental Science and New Energy Technology Engineering Laboratory,
Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School,
Tsinghua University, Shenzhen, Guangdong, People's Republic of China
e-mail: chanw@sz.tsinghua.edu.cn

H. Yang et al. (eds.), *AI and Analytics for Public Health*, Springer Proceedings in
Business and Economics, https://doi.org/10.1007/978-3-030-75166-1_4

(Kapoor et al., 2020), and hybrid models of the above three approaches (Yang et al., 2020). Methods based on epidemic models are well interpretable, that is, their parameters have practical meaning. However, the time-dependent parameters of the epidemic models are very difficult to estimate, due to the complexity of people's responses to COVID-19. Although multiple epidemic models with time-varying parameters have been proposed (Chen et al., 2020; Waqas et al., 2020), the determination of these time-varying parameters is still a difficult problem. Most epidemic models such as SIR and SEIR can not adaptively adjust their parameters. Therefore, when the control measures of COVID-19 change significantly, it is needed to assess and quantify the effect of these measures on the model parameters, which makes prediction rather difficult. Machine learning-based methods, on the other hand, can fit the data very well. In particular, machine learning-based methods can fuse and exploit data of multiple types and sources (e.g., population movement data) (Kapoor et al., 2020), in addition to the time series itself. But they are less interpretable, making it difficult to characterize the dynamics of the COVID-19 development. As a result, the usefulness of these methods is limited to forecasting. They can not quantify the impact of different responses on the development of the epidemic.

Regression and autoregressive models provide a good balance between interpretability and prediction accuracy. Thus, they are widely used in the study of COVID-19 time series. For example, Ceylan (2020) predicted total confirmed cases in Italy, Spain and France based on ARIMA model using data up to April 15, 2020. The optimal order of the ARIMA model is determined for each country. Similarly, Singh et al. (2020) exploited an advanced autoregressive integrated moving average (ARIMA) model to predict future trends in the 15 most affected countries at the time, forecasting confirmed cases, deaths, and recoveries in the following 2 months based on data up to April 24. Pandeya et al. (Gupta et al., 2020) compared the effectiveness of the SEIR model and polynomial regression for predicting the epidemic in India. Their results showed that the SEIR model had better prediction accuracy than polynomial regression.

These existing regression methods have a similar defect, that is, they all use the entire time series to fit the model. However, with the development of the epidemic, diverse responses are adopted, thus leading to changes in the inherent dynamics of the COVID-19 time series. Therefore, a more efficient approach would be to segment the COVID-19 time series according to different dynamic patterns and then fit a separate model for each segment.

In this paper, we adopt an approach based on polynomial regression with adaptive sliding windows to segment the COVID-19 time series. Such method has been proved to be able to efficiently extract the dynamic patterns of time series (Liu et al., 2020). This method uses the sliding windows with adaptive lengths to partition time series into segments. The dynamic pattern of each sliding window is defined as a fitted n-order polynomial. For a sliding window of length L, we consider the data within the sliding window to have the same dynamic pattern if the residual of the n-order polynomial regression is less than a predetermined threshold. We consecutively increase the length of the sliding window until the polynomial is insufficient to fit the newly added data. Subsequently, we set a new sliding window and repeat the above steps. Experimental results show that this method is able to adaptively segment the COVID-19 time series of different countries into segments that are highly intuitively interpretable.

Further, we analyze the segmentation results and define the similarity between segments using the dynamic time wrapping (DTW) distance (Berndt & Clifford, 1994). We visualize the similarity based on complex network analysis method. It can be observed that each wave of the epidemic outbreak can be broadly partitioned into three stages: the early outbreak, the rising stage, and the falling stage. Besides, significant similarities exist between the same stages of different waves and in different countries. In addition, some evidences suggest that the dynamics of the previous segment could provide useful information of the subsequent development of the epidemic.

2 Methodology

2.1 Data Sources

This paper studies the dynamics of COVID-19 through the time series of daily active cases. Our data is collected from https://www.worldometers.info/coronavirus/ at country level, from January 22, 2020 to September 2, 2020. In order to horizontally compare the development of the epidemic across countries, we normalize the number of the daily active cases by the total population of each country, which is collected from https://github.com/datasets/population.

Based on the above data, we first extract the dynamic patterns of different stages of COVID-19 in each country, then compare the dynamic patterns of 15 representative countries. These two steps are described as follows.

2.2 Extracting Dynamic Patterns of COVID-19 Time Series

For extracting the dynamic patterns, we adopt a similar approach to that proposed by Liu et al. (2020), whose original purpose is to study the self-similarity of the fluctuation behaviors of the WTI crude oil price. In this paper, we use a sliding window with adaptive length to partition the entire the time series of COVID-19 daily active cases into segments, fitting each segment with an n-order polynomial as shown in Eq. (1). In this paper, we choose $n = 3$.

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_n x^n + \varepsilon \tag{1}$$

Since the time series within each sliding window is fitted using an n-order polynomial, the initial length of the sliding window is set to $n + 1$. Subsequently, we use Eqs. (2) and (3) as metrics for determining the length of the sliding window, where y_i is the real value of the number of COVID-19 active cases at the i-th day of the sliding window, and $\hat{y}_i$ is the corresponding predicted value by the model. The average number of the daily active cases inside the sliding window is expressed as $\overline{y}$. If both R^2 and $E_{\max}$ are greater than a predetermined threshold T, we increase the length of the window L by 1 day, that is, we move the front edge of the sliding window 1 day forward. We fit the data within the new sliding window and check whether the new fit satisfies $R^2 \geq 0.9$ and $E_{max} \geq 0.9$. We iteratively increase the length of the sliding window forward and perform polynomial regression until either R^2 or E_{max} is less than T. In this paper, we choose $T = 0.9$.

$$R^2 = 1 - \frac{\sum_{i=1}^{L} \left(y_i - \hat{y}_i\right)}{\sum_{i=1}^{L} (y_i - \overline{y})} \tag{2}$$

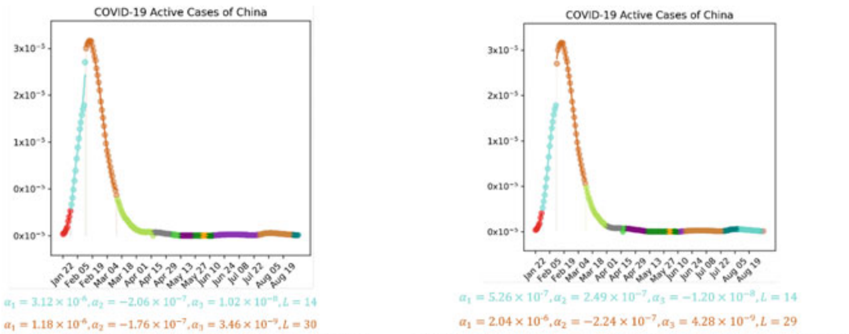
$$E_{\max} = 1 - \max\left(\frac{\left|y_i - \hat{y}_i\right|}{|y_i|}\right) \tag{3}$$

When the R^2 or E_{max} of a sliding window becomes less than T, we fix this sliding window to its current state and create a new sliding window. The new sliding window is created at the front edge of the old one, with the same initial length of $n + 1$. The same operations are performed consecutively until the time series is completely partitioned.

For each partition, its dynamic is described by a vector composed of coefficients of its polynomial regression, which is denoted as $[\alpha_1, \alpha_2, \cdots, \alpha_n]$. Notice that the intercept of the polynomial is not considered, since we are only interested in the variation of the time series in each partition, not in the specific values.

2.3 Dealing with Structural Breakpoints of the Dynamics

In this paper, we use *structural breakpoints* to refer to the data points that divide the time series into two fragments with different dynamics. In the method introduced above, structural breakpoints of the dynamics naturally appear at the ends of segments, and they participate in the polynomial regression. However, this operation is inappropriate, since the occurrence of a structural breakpoint implies that the dynamic pattern has changed. The structural breakpoint should be the beginning point of the next segment, rather than the end point of the previous segment. At the



(a) Original partition and dynamics of the time series

(b) Partition and dynamics after dealing with breakpoints

Fig. 1 An illustrative example for dealing with structural breakpoints. (**a**) Original partition and dynamics of the time series. (**b**) Partition and dynamics after dealing with breakpoints

same time, the structural breakpoint should not be involved in the regression of the previous segment. The reason is explained as follows.

The regression of n-order polynomials can be approximately regarded as an n-order Taylor expansion. Our decision on whether to create a new sliding window depends on whether the n-order Taylor expansion is sufficient to describe the dynamics within the current sliding window. In other words, we use the two criteria (i.e., $R^2 \geq 0.9$ and $E_{max} \geq 0.9$) to determine whether the n-order polynomial regression is a good fit to the data in the current sliding window. If not, a new sliding window will be created. Therefore, if the structural breakpoints are assigned into the previous sliding window, then dynamic pattern within that sliding window cannot be expressed by any n-order Taylor expansion. To ensure that the data within the current window can be well described by the n-order polynomial regression, structural breakpoints need to be assigned to the latter segment. In the rest part of this subsection, we demonstrate this idea through the case study of China.

In the time series analysis of COVID-19, the reason of the occurrence of a breakpoint, in addition to changes in COVID-19 dynamics, may also be an increase in detection capability or a change in statistical methods. Figure 1a illustrates this situation by a case study on China. In the CVOID-19 time series of China, there was an obvious jump up on February 12, dividing the already-occurring downward trend into two segments. This jump up was due to the fact that Hubei changed the statistical method of COVID-19, including clinically diagnosed cases as active cases. The overall downward trend of daily activity cases in China did not change. However, this outlier significantly affects the results of the polynomial regression of the previous segment, which can affect our judgment on the dynamics of the epidemic. More importantly, improvements in the detection capability and changes in the statistical methods can affect all the subsequent data after the structural breakpoint. In fact, these are two reasons that trigger changes in the dynamics of the COVID-19 time series. Therefore, assigning the structural breakpoints to the latter segments is a more proper operation.

In the regression shown in Fig. 1a, there is no special treatment on the structural breakpoints. From the second segment which is marked in blue, it can be clearly observed that the presence of the structural breakpoint in this sliding window results in an unsatisfactory regression result. In fact, the regression result is an approximately linear growth. It does not really reflect the dynamics of the time series within the window, which is actually in a fading stage. Considering that changes in detection capacity or statistical methods can affect both the breakpoints and their subsequent data, we assign breakpoints into the latter segments. Regression results obtained through this method is demonstrated in Fig. 1b, showing that we better extract the dynamics of the time series.

2.4 Measuring the Similarity Among Segments

After dividing the COVID-19 time series of each country into segments according to different dynamics, we further measure the similarity among these segments. Liu et al. (2020) measure the similarity between segments based on the Euclidean distance between the coefficient vectors of each segment's polynomial regression, as is shown in Eq. (4).

$$S_{ij} = \sqrt{\sum_{t=1}^{n} \left[\alpha_t^{(i)} - \alpha_t^{(j)}\right]^2} \tag{4}$$

where $\left[\alpha_1^{(i)}, \alpha_2^{(i)}, \cdots, \alpha_n^{(i)}\right]$ is the coefficient vectors of the polynomial regression of Segment i. Notice that the intercept is ignored.

However, we argue that this method may not be suitable for measuring the similarity between the segments of the COVID-19 time series. This is because there is a huge variation in the percentage of daily active cases per population among different countries. Therefore, we first normalized the time series of each country with its peak value, then use the dynamic time warping (DTW) distance (Berndt & Clifford, 1994) between each pair of segments as the similarity measure. The experimental results show that the DTW distance can effectively identify the similarity between the dynamics of each segment, thus matching the segments of the same stages of the outbreak in different countries.

3 Results

3.1 Partitioning COVID-19 Time Series

The trends of the COVID-19 epidemic can be broadly classified into three categories: rising, falling and equilibrium. Rising and falling mean that the number of active cases is significantly increasing or decreasing, while equilibrium means that the number of active cases does not change significantly. The equilibrium state includes the beginning and the end of the epidemic, as well as the dynamic equilibrium in the middle of the epidemic when the number of new infections and the number of cured cases are approximately equal.

In this paper, the COVID-19 time series (i.e., the percentage of active cases per population of each day) is partitioned into segments with different dynamics. Figure 1 shows the results of the partitioning of the time series for a representative set of 15 countries, where each fragment is labeled with a number. Overall, it can be observed that the COVID-19 time series are partitioned into a number of segments of different lengths. Especially, in the equilibrium state of the epidemic, the time series are partitioned into segments of short lengths, whereas the rising and falling trends are partitioned into long segments.

This result matches very well with our intuition. In both the rising and the falling state, the method adopted in this paper is able to nicely extract the corresponding trend. On the other hand, in the equilibrium state, when the number of active cases fluctuates, it is not possible to discern from the time series itself whether this is a random fluctuation (e.g., controllable imported cases) or a signal of the change in dynamics that indicates an upward or downward trend in the future. The approach adopted in this paper treats the data in the equilibrium state with caution, regarding each of the significant fluctuations as a signal of the change in dynamics. This results in the equilibrium state being split into small segments (Fig. 2).

Moreover, in many countries, a complete wave of an outbreak can be partitioned into four major segments: the beginning stage, the ascending stage, the descending stage, and the end stage. We can observe that the lengths of the beginning stages and end stages exceed the lengths of the short fragments during the equilibrium state, but are significantly smaller than the lengths of the ascending and descending stages. In addition, note that the ascending and descending stages are not segmented from the peak point, rather, these two stages are segmented from a time point before the time series reaches the peak. This is because we focus on the change in the dynamics of COVID-19, which is essentially caused by the epidemic being under control. In other words, the number of new daily infections gradually decreases and the number of cured cases gradually increases after the time point where the dynamics change, which leads to the number of daily active cases reaching the peak after a period of time.

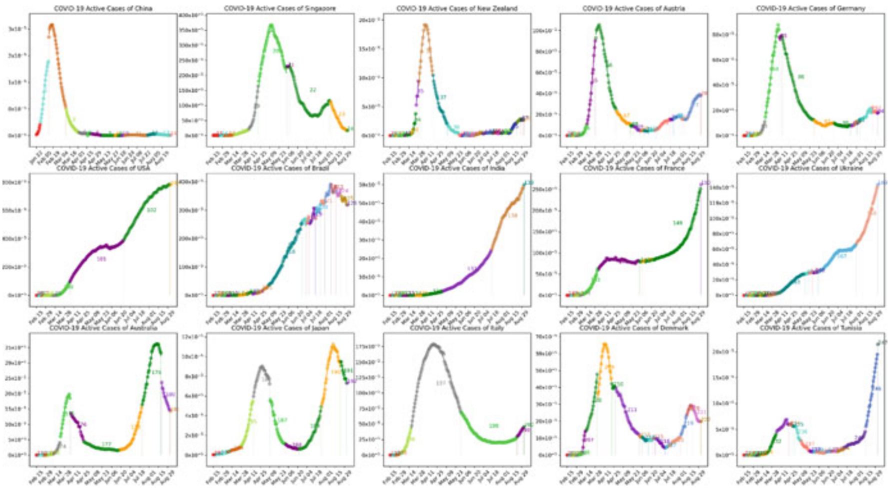


Fig. 2 Partitioning the COVID-19 time series according to different dynamics

3.2 *Analyzing the Dynamics of the First Wave of the Outbreak*

For the countries in Fig. 1, the development of the COVID-19 epidemic can be broadly classified into three categories: (1) the countries with only one peak where the epidemic is almost over, (2) the countries where the first wave of the epidemic is not yet under control, (3) the countries with two peaks where there is a second outbreak. We selected China, the United States, and Australia as representatives of each of the three categories. In this subsection, we compare the dynamics of the first wave of the COVID-19 outbreak in these three countries.

Table 1 selects some representative segments and shows the coefficients of the polynomial regression of these segments. Segment 0, Segment 100, and Segment 174 represent the dynamics of the initial stage of the outbreak in China, the USA and Australia, respectively. We can see that the polynomial coefficients of theses three segments are roughly proportional, suggesting that the same dynamics are present among the three countries at the beginning of the epidemic outbreak. However, the subsequent dynamics of the outbreak become very different. For China, the polynomial regression coefficients in Segment 1 already reflect the decreasing trend in the number of active cases, but the jump up on February 12 disrupts the dynamics we identified. For the U.S., the number of active cases in Segment 101 first decreases and then increases, indicating the occurrence of a second outbreak before the first wave of the epidemic ends. Segment 102 appears to be entering a downward trend, but the future is still unclear. For Australia, Segment 145 is similar to the Segment 1 of China, but the declining speed of Segment 176 is significantly slower than that of China's Segment 3, which may suggest the risk of a second outbreak.

Similar phenomenon can be observed from many other countries such as Japan, Italy and Denmark, where a slower declining speed of the first wave is followed

Table 1 Representative segments of the first wave of the outbreak in China, the U.S. and Australia

Country	Index of Segment	$[\alpha_1, \alpha_2, \alpha_3]$
China	0	$[7.27\times 10^{-8}, 5.87\times 10^{-8}, 4.03\times 10^{-9}]$
	1	$[5.26\times 10^{-7}, 2.49\times 10^{-7}, -1.20\times 10^{-8}]$
	2	$[2.04\times 10^{-6}, -2.24\times 10^{-7}, 4.28\times 10^{-9}]$
	3	$[-1.07\times 10^{-6}, 4.92\times 10^{-8}, -9.36\times 10^{-10}]$
USA	100	$[5.92\times 10^{-6}, 2.20\times 10^{-6}, 1.18\times 10^{-8}]$
	101	$[1.11\times 10^{-4}, -1.61\times 10^{-6}, 8.60\times 10^{-9}]$
	102	$[9.10\times 10^{-5}, -1.56\times 10^{-7}, -4.92\times 10^{-9}]$
Australia	174	$[5.86\times 10^{-7}, 1.25\times 10^{-7}, 1.24\times 10^{-8}]$
	175	$[9.25\times 10^{-6}, 1.03\times 10^{-6}, -6.91\times 10^{-8}]$
	176	$[-4.78\times 10^{-6}, 2.81\times 10^{-7}, -1.39\times 10^{-8}]$

by a second outbreak. By comparing the dynamics of the first wave of the COVID-19 outbreak across different countries, we propose an initial assumption that the dynamics at the end of the first wave may partially reflect the future development of the epidemic. A slower decline at the end of the first wave may suggest the risk of a second outbreak in the future.

3.3 Comparing the Dynamics of the First and Second Outbreak

In this subsection, we analyze a number of other representative segments. The additional segments are shown in Table 2. The main aim is to compare the similarities in epidemic dynamics across countries, as well as the similarities between the first and second outbreak.

Based on the method described in Sect. 2.4, we measure the similarity between the segments listed in Tables 1 and 2. We further normalize each column of the DTW matrix by the following Eq. (5).

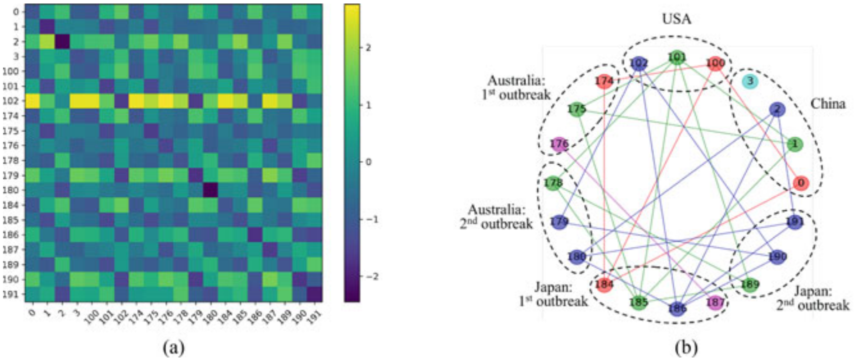
$$S'_{ij} = \frac{S_{ij} - \frac{1}{n}\sum_{i=1}^{n} S_{ij}}{\frac{1}{n}\sum_{i=1}^{n}\left(S_{ij} - \frac{1}{n}\sum_{i=1}^{n} S_{ij}\right)} \tag{5}$$

Figure 3a demonstrates the matrix of DTW distance among those segments. We select the 15% percentile of all elements in the matrix as the threshold t. We drop the DTW distance greater than t to obtain the similarity network shown in Fig. 3b.

We further study the similarity network shown in Fig. 3b based on some complex network analysis methods. In this network, each node represents a corresponding segment and the edge (i,j) represents Segment i and Segment j are similar. The

Table 2 Additional representative segments of the first and second wave of the outbreak

Country	Index of Segment	$[\alpha_1, \alpha_2, \alpha_3]$
Australia: 2nd outbreak	178	$[-3.69\times 10^{-7}, 5.61\times 10^{-8}, 2.45\times 10^{-9}]$
	179	$[8.30\times 10^{-6}, 5.15\times 10^{-7}, -2.06\times 10^{-8}]$
	180	$[-1.52\times 10^{-5}, 1.04\times 10^{-6}, -3.75\times 10^{-8}]$
Japan: 1st outbreak	184	$[4.44\times 10^{-7}, -2.35\times 10^{-8}, 7.22\times 10^{-10}]$
	185	$[-5.04\times 10^{-7}, 2.48\times 10^{-7}, -4.40\times 10^{-9}]$
	186	$[5.10\times 10^{-6}, -2.13\times 10^{-7}, 1.59\times 10^{-9}]$
	187	$[-3.01\times 10^{-6}, -4.23\times 10^{-8}, 3.56\times 10^{-9}]$
Japan: 2nd outbreak	189	$[7.45\times 10^{-8}, 1.64\times 10^{-8}, 4.98\times 10^{-10}]$
	190	$[4.92\times 10^{-6}, 2.57\times 10^{-8}, -5.06\times 10^{-9}]$
	191	$[-2.50\times 10^{-6}, -1.58\times 10^{-7}, 1.66\times 10^{-8}]$



(a)

(b)

Fig. 3 Measuring the similarity among the representative segments. (**a**) The normalized DTW matrix. (**b**) The similarity network of the representative segments

similarity network is composed of six connected components. In Fig. 3b, the connectivity components are marked by different colors, with nodes belonging to the same connected component represented by the same color. After ignoring the outliers 3 and 199, it can be observed that the four major components (i.e., labeled red, blue, brown, and gray) represent roughly four different types of dynamics. The first three components represent the early outbreak, the raising, and the falling, respectively. Meanwhile, the segments that belong to the gray component are followed by second outbreaks.

Based on the above, some preliminary ideas can be suggested. Among the time series of COVID-19 daily active cases, some similarities can be observed in the dynamical structure. First, each wave of the outbreak can be roughly partitioned in to three stages: the early outbreak, the ascending stage, and the descending stage. In particular, the descending stage is characterized by the containment of the epidemic rather than a decrease in the number of daily active cases. Second, if we partition COVID-19 time series in to segments according to different dynamics, significant similarities can also be observed between the same stages in different countries and different waves of outbreak. Finally, the dynamics of the previous segment may implicitly inform the subsequent trends of the epidemic development, e.g., slower decline at the end of the first wave may suggest higher risk of a second outbreak in the future.

4 Conclusion

Understanding the dynamics of development of COVID-19 is crucial in the prediction and containment of the epidemic. This paper extracts the underlying dynamical structures of the time series of COVID-19 daily active cases, based on polynomial regression with adaptive sliding windows. Further, similarities among the partitioned COVID-19 time series of different countries are compared based on DTW distance and complex network analysis. We find that each wave of the outbreak can be roughly divided into three stages: the early outbreak, the ascending stage, and the descending stage. Significant similarities exist between the same stages of different waves in different countries. The dynamics of the previous stage may provide information of the subsequent development of the epidemic.

Acknowledgements This research was funded by the National Natural Science Foundation of China (Grant No. 71971127) and the Hylink Digital Solutions Co., Ltd. (120500002).

References

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, Seattle, WA.

Calafiore, G. C., Novara, C., & Possieri, C. (2020, March 31). A modified SIR model for the COVID-19 contagion in Italy. *arXiv:2003.14391* [physics].

Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France, (in en). *Science of the Total Environment, 729*, 138817.

Chen, Y.-C., Lu, P.-E., Chang, C.-S., & Liu, T.-H. (2020, April 28). A time-dependent SIR model for COVID-19 with undetectable infected persons. *arXiv:2003.00122* [cs, q-bio, stat].

Gupta, R., Pandey, G., Chaudhary, P., & Pal, S. K. (2020). SEIR and Regression Model based COVID-19 outbreak predictions in India. *Public and Global Health*, preprint 3 Apr 2020. Available: http://medrxiv.org/lookup/doi/10.1101/2020.04.01.20049825. Accessed on 22 Sept 2020 03:39:41.

Kapoor, A., et al. (2020, July 6). Examining COVID-19 forecasting using spatio-temporal graph neural networks. *arXiv:2007.03113* [cs].

Liu, S., Fang, W., Gao, X., Wang, Z., An, F., & Wen, S. (2020). Self-similar behaviors in the crude oil market. *Energy, 211*, 118682.

Singh, R. K., et al. (2020). Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced Autoregressive Integrated Moving Average (ARIMA) model, (in en). *JMIR Public Health and Surveillance, 6*(2), e19115.

Wangping, J., et al. (2020). Extended SIR prediction of the epidemics trend of COVID-19 in Italy and compared with Hunan, China. *Frontiers in Medicine, 7*, 169.

Waqas, M., Farooq, M., Ahmad, R., & Ahmad, A. (2020, May 10). Analysis and prediction of COVID-19 pandemic in Pakistan using time-dependent SIR model. *arXiv:2005.02353* [q-bio].

Yang, Z., et al. (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease, 12*(3), 165–174.