RESEARCH ARTICLE

REVISED Tracking and forecasting milepost moments of the epidemic in the early-outbreak: framework and applications to the COVID-19 [version 2; peer review: 2 approved]

Huiwen Wang[1,2], Yanwen Zhang[1], Shan Lu[3], Shanshan Wang[1,4]

[1]School of Economics and Management, Beihang University, Beijing, China
[2]Beijing Advanced Innovation Center for Big Data and Brain Computing,, Beijing, China
[3]School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China
[4]Beijing Key Laboratory of Emergence Support Simulation Technologies for City Operations, Beijing, China

v2 **First published:** 06 May 2020, **9**:333
https://doi.org/10.12688/f1000research.23107.1

Latest published: 18 Sep 2020, **9**:333
https://doi.org/10.12688/f1000research.23107.2

Abstract

Background: The outbreak of the 2019 novel coronavirus (COVID-19) has attracted global attention. In the early stage of the outbreak, the most important question concerns some meaningful milepost moments, including the time when the number of daily confirmed cases decreases, the time when the number of daily confirmed cases becomes smaller than that of the daily removed (recovered and death), and the time when the number of daily confirmed cases and patients treated in hospital, which can be called “active cases”, becomes zero. Unfortunately, it is extremely difficult to make right and precise prediction due to the limited amount of available data at the early stage of the outbreak. To address it, in this paper, we propose a flexible framework incorporating the effectiveness of the government control to forecast the whole process of a new unknown infectious disease in its early-outbreak.
Methods: We first establish the iconic indicators to characterize the extent of epidemic spread. Then we develop the tracking and forecasting procedure with mild and reasonable assumptions. Finally we apply it to analyze and evaluate the COVID-19 outbreak using the public available data for mainland China beyond Hubei Province from the China Centers for Disease Control (CDC) during the period of Jan 29th, 2020, to Feb 29th, 2020, which shows the effectiveness of the proposed procedure.
Results: Forecasting results indicate that the number of newly confirmed cases will become zero in the mid-early March, and the number of patients treated in the hospital will become zero between mid-March and mid-April in mainland China beyond Hubei Province.
Conclusions: The framework proposed in this paper can help people get a general understanding of the epidemic trends in countries where COVID-19 are raging as well as any other outbreaks of new and

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 18 Sep 2020	✓ report	✓ report
version 1 06 May 2020	? report	? report

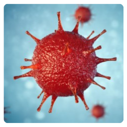
1. **Rosanna Verde**, University of Campania “Luigi Vanvitelli”, Caserta, Italy

2. **Paula Brito**, University of Porto & LIAAD - INESC TEC, Porto, Portugal

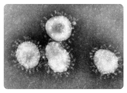
Any reports and responses or comments on the article can be found at the end of the article.

unknown infectious diseases in the future.

Keywords
COVID-19, Prediction Method, Epidemic Development Index System



This article is included in the Disease Outbreaks gateway.



This article is included in the Coronavirus collection.

Corresponding author: Shanshan Wang (sswang@buaa.edu.cn)

Author roles: Wang H: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision; **Zhang Y**: Formal Analysis, Investigation, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Lu S**: Investigation, Validation, Visualization, Writing – Review & Editing; **Wang S**: Formal Analysis, Methodology, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study was supported by the National Science Foundation of China [11701023, 71843006]
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Wang H, Zhang Y, Lu S and Wang S. **Tracking and forecasting milepost moments of the epidemic in the early-outbreak: framework and applications to the COVID-19 [version 2; peer review: 2 approved]** F1000Research 2020, **9**:333 https://doi.org/10.12688/f1000research.23107.2

First published: 06 May 2020, **9**:333 https://doi.org/10.12688/f1000research.23107.1

REVISED **Amendments from Version 1**

In this version, we have carefully incorporated all comments and suggestions raised by the reviewers, with further explanation for some concepts and revision of some typos. A new version of R package is also included.

Any further responses from the reviewers can be found at the end of the article

1 Introduction

The atypical pneumonia caused by the 2019 novel coronavirus (COVID-19), which is a highly infectious human disease, was first reported in Dec 31st, 2019 in Wuhan, the capital of Hubei Province in China (WHO *et al.*, 2020). To mitigate the effect of epidemics spreading across China and other countries, Wuhan was temporarily shut-down from Jan 23th, 2020, which has proved to be efficient in the timely stopping the spread of the coronavirus (Chinazzi *et al.*, 2020). However, due to the "Spring Festival travel rush", there was still a rising number of confirmed cases in China in the following two months, which has caused great strain on medical resources (Li *et al.*, 2020).

The questions that draw the most concerns are how COVID-19 will spread, and when it will end. People were always asking when the number of the daily confirmed cases will become smaller than the previous days, and when the daily confirmed cases will become smaller than that of the removed (recovered and death). These are not only of highly important for the general public, but also for government, who plays an important role in controlling the disease within a short period as much as possible. Since the decline of the number of newly confirmed cases and the number of active cases imply the alleviation of epidemic, the emergence of these turning points convey useful information for decision making on medical resources allocation and isolation policies in the post-stage of the epidemic.

Meanwhile, it is also important to predict when will the number of daily confirmed cases become "zero", as well as when the number of active cases will be "zero". The latter indicates the end of the epidemic. These two "zero points" can also help the government to consider loosening population migration restriction in cities. Additionally, authorities in economic departments can use the forecasting results to assess the impact of the epidemic on the economy in advance, and plan for the restoration of normal production and living order.

There have been various publications on COVID-19 from different perspectives, i.e., the origin of COVID-19, the clinical features as well as epidemic transmission characteristics. Specifically, for the origin of the virus, Fan *et al.* (2019) and Luk *et al.* (2019) pointed out that COVID-19 is an infectious disease caused by a virus closely related to SARS-CoV, while others believed that the COVID-19 virus was originally derived from wild animals (Benvenuto *et al.*, 2020; Huang *et al.*, 2020). For the epidemic transmission characteristics, Holshue *et al.* (2020) and Hui *et al.* (2020) found that the virus can be transmitted from person to person and that it has a high interpersonal transmission rate. Zhao *et al.* (2020) investigated the preliminary estimation of the basic reproduction number R_0, which ranged from 2.24(95%CI : 1.96 − 2.55) to 3.58(95%CI : 2.89 − 4.39) in the early outbreak, while Prasse *et al.* (2020) estimated it around 2.2, Tang *et al.* (2020) applied likelihood-based and model-based methods to the analysis of early reported cases, and the results showed that R_0 could be is as high as 6.47. Zhou *et al.* (2020) used the SEIR model and stated that the range of R_0 of COVID-19 is 2.8–3.3, indicating that the early pathogenic transmission capacity of COVID-19 is close to or slightly higher than SARS. Other studies related to R_0 are Anastassopoulou *et al.* (2020); Zhang *et al.* (2020) and referenced therein. Unfortunately, each of these models may result in different estimations of R_0, which may cause any predictions based on R_0 to be unstable.

Recently, a number of publications have been related to trend prediction of the COVID-19 outbreak in China. Zeng *et al.* (2020) proposed a multi-model ordinary differential equation set neural network and model-free methods to predict the interprovincial transmission in mainland China, especially those from Hubei Province, and predicted that COVID-19 in China is likely to decelerate before Feb 18th and to end before April 2020. Chen *et al.* (2020) made prediction based on epidemiological surveys and analyses, which showed that the total number of diagnoses would be 2–3 times that of SARS, and the peak is predicted to be in early or middle February. Yu *et al.* (2020) revised the SIR model based on the characteristics of the COVID-19 epidemic development, and proposed a time-varying parameter-SIR model to study the trend of the number of infected people. Peng *et al.* (2020) used the SEIR method to predict the end of the epidemic in most cities in mainland China. Wu *et al.* (2020) used the Markov chain Monte Carlo method to estimate R_0, and inferred from the SEIR model that the peak COVID in Wuhan would be reached in April, and other cities in China would be delayed by 1 to 2 weeks.

However, there are some obvious shortcomings of forecasting methods based on epidemic models in terms of outbreak prediction. For example, the SEIR model is a mathematical method relying on an assumption of epidemiological parameters for disease progression, which are absent for a novel pathogen. For instance, the basic infection number R_0, the daily recovery rate, the characteristics of the disease itself (such as the infection rate and the conversion rate of the latent to the infected), the daily exposure rate of the latent and infected, and their initial population infection status (total population, infected, the initial value of the latent, the susceptible, the healer, etc.) and many other key parameters need to be set. For infectious diseases that have already appeared in the past, or those who have a large amount of data, it is not difficult to obtain these parameters. However, for unknown, sudden and early infectious diseases, obtaining these parameters is full of difficulties, which leads to a great uncertainty and limitations in the prediction of the epidemic situation using the SEIR model.

Moreover, there exist many challenges for the prediction of a new epidemic situation similar to COVID-19. First, little prior knowledge that can be refered to or analogized for a brand new epidemic; secondly, the existence of government management will make the development of the epidemic completely different from that under free development, thus how to incorporate the influence of government measures into the fitting process of parameters and build a statistical model from this needs to be considered; thirdly, in the early-outbreak the initial data often fluctuates violently and the data quality is low, thus many commonly used parameter estimation methods are not applicable anymore; furthermore, the amount of data in the early stage is too small, making it difficult to directly rely on the inertia of the data to make forward prediction. In summary, in the early stages of a brand new epidemic, how to use some low-quality and small data sets to make basic and relatively accurate forecast judgements for the entire process of the epidemic, is a long-term pain point.

To cope with these challenges, we propose a simple and effective framework incorporating the effectiveness of the government control to forecast the whole process of a new unknown infectious disease in its early-outbreak, from which we emphasis the prediction of meaningful milepost moments. Specifically, we first propose a series of iconic indicators to characterize the extent of epidemic spread, and describe four periods of the whole process corresponding to the four meaningful milepost moments: two turning points and two “zero” points; then we develop the proposed procedure with mild and reasonable assumptions, specfically without relying on an assumption of epidemiological parameters for disease progression. Finally we apply it to analyze and evaluate COVID-19 using publicly available data from mainland China beyond Hubei Province from the China CDC during the period of Jan 29th, 2020, to Feb 29th, 2020, which shows the effectiveness of the proposed procedure.

From the empirical study, we can suggest that the proposed method may cast a flexible framework and perspective for early prediction of a sudden and unknown new infectious disease with effective government control. Specifically, in the early stage of the epidemic when some regular information is initially displayed, the proposed method can be used to predict the process of epidemic development and to judge which stage of development the situation is at, when the peak will be reached, and when the turning point will appear. Moreover, by continuously accumulating data and updating the model during the development of the epidemic, we can also predict when the epidemic will basically end. Finally, the proposed method enjoys great generalizability, which can be used to understand the epidemiological trend of COVID-19 spread in other counties, which will provide useful guidance for fighting against it.

The reminder of this paper is organized as follows. In Section 2, we proposed the main methodology, where we defined the iconic indicators to characterize the extent of epidemic spread in Section 2.1, yielding four periods of the whole process corresponding to the four meaningful milepost moments: two turning points and two “zero” points in Section 2.2, then Section 2.3 presents the proposed procedure with mild and reasonable assumptions. Then we applied the proposed method to the COVID-19 using the public available data in mainland China beyond Hubei Province from the China CDC during the period of Jan 29th, 2020, to Feb 29th, 2020, and describe the trend of the COVID-19 spread in detail in Section 3. Some conclusions and discussions are finally given in Section 4.

2 Methods

The data we used are provided by China CDC via public data sources, in which the cumulative confirmed cases up to the given day t, the daily confirmed cases at day t, the daily recovered ones and the daily deaths at day t are included. All the data analysis results are done with **R** software, version 3.6.0 and higher is recommended. The main code for the implementation of the proposed procedure as well as the data and its full description are available from **Github** (See data availability for more detail (YuanchenZhu2020, 2020).

In order to assess and predict the epidemic, we first define a set of necessary indicators that can reflect the status of disease contagion. We then divide the cycle of the epidemic into four stages, which are divided by the turning points of the proposed indicators. Finally, we propose a computational framework to predict the turning points.

2.1 The iconic indicators to characterize a epidemic

It is obvious that the contagion process of an unknown virus in different regions would be diverse with respect to the number of patients and the growth pattern of the epidemic, because of population density, population mobility, public health conditions, as well as disease prevention and control measures. Therefore, we first constructed a set of indicators to monitor the essential laws of the development of the disease.

There are several requirements for the monitoring indicators. Firstly, as the number of patients can vary greatly across regions, the scale of the data should be eliminated so that the analysis methods and results are comparable. Secondly, they should well reflect the general laws and characteristics of the epidemic process as well as accurately and coherently describe the entire process of the epidemic from the beginning to the end. Particularly, they should be able to answer the question of when the turning point of the epidemic would appear. Thirdly, they should be as simple and convenient as possible so that it can be applied with publicly available data. Last but not least, the indicators should have clear meaning and be easily interpreted.

Following the above, we first adopt three basic indicators that are published daily by the provincial and municipal governments of China. That is, for time t, the daily confirmed cases E_t, the daily recovered cases O_t, and the daily deaths D_t. Then we define a few monitoring indicators to characterize the epidemic stages, that is the number of active cases N_t, the daily infection rate K_t and the daily removed (the sum of recovered and deaths) rate I_t, which are defined as follows.

- The number of active cases N_t is defined as the cumulative confirmed cases with recovered ones and deaths removed up to t, that is

  $$N_t = \sum_{i=1}^{t}(E_t - O_i - D_i).$$

  Note that N_t is essential for epidemic investigation, since it reflects the size of local patients and the pressure on the medical system.

- The daily infection rate K_t is defined as the ratio of the daily confirmed cases at time t and the number of active cases at time $t - 1$, i.e.

  $$K_t = \frac{E_t}{N_{t-1}}.$$

  Obviously, K_t reflects the rate at which patients enter the treatment system. It is influenced by many factors, including the property of the infectious disease, the average immune capacity of the population, population density, climate condition, public health conditions, public health awareness, the awareness of self-prevention of diseases and the efforts of epidemic prevention and control.

- Similarly, the daily removed rate I_t is defined as the ratio of the daily removed cases at time t and the number of active cases at time $t - 1$, i.e.

  $$I_t = \frac{O_t + D_t}{N_{t-1}},$$

  where I_t reflects the rate at which patients leave the medical system, that is, the rate at which the pressure on medical resource is eased.

Using the above indicators, we further define R_t as the outbreak status on day t as follow:

$$R_t = 1 + K_t - I_t.$$

Obviously, it holds that

$$N_t = N_{t-1}R_t = N_0 \prod_{l=1}^{t}(1 + K_l - I_l), \tag{1}$$

where N_0 denotes the initial number of active cases at the beginning of the outbreak. In particular, when the daily infection rate and removed rate are relatively stable, denoted as K and I respectively, we have the constant epidemic status index $R = 1 + K - I$. Then (1) can be written as:

$$N_t = N_0 \cdot R^t = N_0 \cdot (1 + K - 1)^t, \tag{2}$$

which shows that the epidemic situation is in the form of an exponential curve. And the epidemic status indicator R can well reflect the rate of expansion or convergence of the population with infectious capacity.

2.2 Four stages of an epidemic

In this section, we will describe the whole process of a epidemic under the assumption that the government has implemented effective control measures, which can be divided into four stages, i.e. “outbreak period”, “controlled period”, “mitigation period” and “convergence period” successively. And we will quantify the iconic features for each stage, which corresponds to the two turning points and two “zero” points, respectively.

Stage 1: Outbreak Period

In the initial stage of an epidemic outbreak, there is delay of social response due to the limited knowledge of the epidemic, and the power of contagion prevention and control is inevitably not enough. Thus the daily infection rate K_t would be high. At the same time, the recovery process in the initial stage is relatively long, and the number of severe patients is small, leading the daily removed rate I_t to be close to “zero”. Therefore, the outbreak status indicator R_t during this period is usually much larger than 1, that is:

$$K_t \gg I_t, R_t = 1 + K_t - I_t > 1 \Rightarrow N_t > N_{t-1}.$$

It can be seen that, during the outbreak period, the number of newly diagnosed patients increases sharply, and the number of active cases will increase dramatically correspondingly, which will pose a great burden to medical institutions, especially for hospitals.

As the epidemic exacerbates, if the government begins to intervene through a series of emergency measures, where a disease prevention and control system is quickly established, the daily infection rate K_t will significantly decrease. Usually, the new daily confirmed cases will begin to decline as well. During the epidemic prevention and control process, once the situation improves, we will see the emergence of the first turning point denoted as T_1. Then after the data T_1, the newly diagnosed patients E_t changes from a rapid rise in the outbreak period to a descending channel ($E_t < E_{t-1}$). In summary, the emergence of the first turning point T_1 indicates that the disease control measures have begun to work, which implies the end of the “Outbreak Period”.

Stage 2: Controlled Period

The emergence of the first turning point is a very positive signal, indicating that the public health management measures have obviously taken effect and the epidemic has entered the “controlled period”. However, due to the fact that the completion rate I_t at this stage is still relatively low, the number of active cases will continue to increase. The controlled period will continue until the second turning point T_2 appears, that is, active cases N_t reaches the peak and starts to decline. This is because the completion rate increase so significantly that $K_t = I_t$ is fulfilled after a long period of treatment in the previous stage. When the completion rate I_t surpasses infection rate K_t, the number of patients treated in the hospital begins to decline from peak.

Stage 3: Mitigation Period

The sign of the end of the controlled period is $K_t = I_t$. Thereafter, K_t will continue to fall with the rise of I_t, which gives

$$K_t < I_t, R_t = 1 + K_t - I_t > 1 \Rightarrow N_t < N_{t-1}$$

This indicates that the daily completion rate I_t will start to be greater than the daily infection rate K_t, that is, the value of the outbreak status indicator R_t becomes less than 1. The population size with infectious capacity will be reduced, and the pressure of medical resources will be significantly relieved, marking the beginning of the “mitigation period”. The mitigation period will continue until the appearance of zero reported newly confirmed cases, that is, $E_t = 0$, which we call the first “zero” point Z_1. After the first “zero” point is reached, the intensity of prevention and control in the entire society will be relieved except for hospitals, that is, the “mitigation period” ends and the “convergence period” starts.

Stage 4: Convergence Period

The “convergence period” will end at the second “zero” point Z_2, which means that the number of people treated in the hospital is equal to or close to “zero”. After reaching the second “zero” point, the epidemic is completely over.

For clarity, we summarize the iconic features and the corresponding milepost moments of each stage in the whole process of the epidemic in Table 1.

Table 1. The four stages of an epidemic.

Stage	Outbreak	Controlled	Mitigation	Convergence
Begin with	the number of newly diagnosed increases	the number of newly diagnosed decreases (the first turning point)	the number of patients in hospital decreases (the second turning point)	the number of newly diagnosed equals to 0 (the first "zero" point)
End with	the number of newly diagnosed reaches peak (the first turning point)	the number of active cases reaches peak (the second turning point)	the number of active cases equals to 0 (the first "zero" point)	the number of active cases equals to 0 (the second "zero" point)
	$K \gg I$, $R \gg 1$	$K > I$, $R > 1$	$K < I$, $R < 1$	$K = 0$, $R \ll 1$

2.3 Implementation: the proposed model

According to Section 2.2, the modeling and predicting of the epidemic need to be divided into two parts. The first part corresponds to the outbreak period, where the intervention and disease curing is not effective enough. The infection rate K_t increases rapidly and the completion rate I_t is small. Thus, the number of newly diagnosed patients E_t increases rapidly, and the number of active cases N_t increases. The pressure on medical resources will soon be overwhelmed. According to Equation (2), N_t will be in an exponential growth trend without forming a convex curve, nor will the so-called two turning points or two "zero" points appear.

The second part, which is the focus of this article, is when the K_t starts to decrease and I_t starts to increase due to effective intervention and improved recovery level for individual patients. Only in this situation will the turning points and "zero" points T_1, T_2, Z_1, Z_2 successively appear, and then the epidemic could end. Therefore, we will model the development of the epidemic under the assumption of effective intervention, then we can obtain the early prediction of two turning points and two "zero" points based on the predicting modeling of E_t and N_t.

Suppose that the infection rate K_t and the removed rate I_t change gently with a stable unitary rate of change within a time window m before time t_0, then given m and t_0, denote $V_{K|(t_0,m)}$ and $V_{I|(t_0,m)}$ as the unitary rate of change of K_t and I_t respectively, that is,

$$V_{K|(t_0,m)} = \left\{\frac{K_{t_0}}{K_{t_0-m+1}}\right\}^{1/(m-1)}, \quad V_{I|(t_0,m)} = \left\{\frac{I_{t_0}}{I_{t_0-m+1}}\right\}^{1/(m-1)}. \tag{3}$$

For any $t > t_0$, the infection rate K_t and the removed rate I_t can be predicted as follows:

$$\hat{K}_{t|t_0} := \hat{K}_{t_0}(t-t_0) = \hat{K}_{t_0}(t-t_0-1)\cdot V_{K|(t_0,m)} = \cdots = \hat{K}_{t_0}(1)\cdot V_{K|(t_0,m)}^{t-t_0-1} = K_{t_0}\cdot V_{K|(t_0,m)}^{t-t_0}, \tag{4}$$

$$\hat{I}_{t|t_0} := \hat{I}_{t_0}(t-t_0) = \hat{I}_{t_0}(t-t_0-1)\cdot V_{I|(t_0,m)} = \cdots = \hat{I}_{t_0}(1)\cdot V_{I|(t_0,m)}^{t-t_0-1} = I_{t_0}\cdot V_{I|(t_0,m)}^{t-t_0}. \tag{5}$$

Thus, we can obtain the outbreak status R_t, the number of patients in the hospital N_t, and the number of newly diagnosed E_t as

$$\begin{aligned} \hat{R}_{t|t_0} &= 1+\hat{K}_{t|t_0}-\hat{I}_{t|t_0}, \\ \hat{N}_{t|t_0} &= \hat{N}_{t-1|t_0}\cdot\hat{R}_{t|t_0}, \\ \hat{E}_{t|t_0} &= \hat{N}_{t-1|t_0}\cdot\hat{K}_{t|t_0}. \end{aligned}$$

According to the prediction process, it can be seen that the prediction results mainly depend on $V_{K|(t_0,m)}$ and $V_{I|(t_0,m)}$, whose value is up to the selection of time window m and starting point t_0. However, it is worth noting that the selection of m and t_0 is not arbitrary, which is suggested as in the follow assumption.

Assumption 1. *The time window m and the starting point t_0 should be chosen satisfy $V_{K|(t_0,m)} < 1$ and $V_{I|(t_0,m)} > 1$. Meanwhile, keeping $I_t < 1$ due to interpretability constraints, and the starting point t_0 should be close to the date of the latest published data as much as possible.*

It is worth noticing that the assumption is proposed to make sure that the trend of outbreak development have already emerged and stable, which means that the outbreak have already been controlled. The assumption is an mild requirement, since when some basic condition are satisfied, such as the epidemic prevention policy is effective and steady, the unitary rate of change would be relatively stable. Our method is totally based on the assumption above, thus when any constraint listed above is not satisfied, our algorithm would be inapplicable.

In summary, here we describe details of the proposed procedure in Algorithm 1.

Algorithm 1. **Main Prediction Procedure**

```
1: Initial setting $m$ and $t_0$, which satisfying Assumption 1;
2: Compute $V_K$ and $V_I$ according to (3); Set $t = t_0 + 1$.
3: Prediction: updating the predicted results at time $t$ via the forecasting value ahead of $l = t–t_0$-step as follows:
       $\hat{K}_{t|t_0} = \hat{K}_{t_0}(l) = K_{t_0} \cdot V^l_{K|(t_0,m)}$
       $\hat{I}_{t|t_0} = \hat{I}_{t_0}(l) = I_{t_0} \cdot V^l_{I|(t_0,m)}$
       $\hat{R}_{t|t_0} = 1 + \hat{K}_{t|t_0} - \hat{I}_{t|t_0}$
       $\hat{N}_{t|t_0} = \hat{N}_{t-1|t_0} \cdot \hat{R}_{t|t_0}$
       $\hat{E}_{t|t_0} = \hat{N}_{t-1|t_0} \cdot \hat{K}_{t|t_0}$
4: Prediction of the milepost moments: If $\hat{E}_{t-1|t_0} < \hat{E}_{t|t_c}$, then $T_1 = t - 1$; If $\hat{N}_{t-1|t_0} < \hat{N}_{t|t_0}$, then $T_2 = t - 1$; If $\hat{E}_{t-1|t_0} < E_0 = 1$, then $Z_1 = t - 1$; If $\hat{N}_{t-1|t_0} < N_0 = 1$, then $Z_2 = t - 1$; If none of the above is satisfied, turn to the next step.
5: Set $t = t + 1$, return to Step 2 until $T_1$, $T_2$, $Z_1$, $Z_2$ are obtained.
```

It is also worth noting that in practice, there are many special cases that we need to take into consideration, thus we created a relatively complete computing framework, which has already been implemented and made into R packages and are available from **Github** (See data availability for more detail (YuanchenZhu2020, 2020).

the more data we accumulate, the clearer the underlying law of the epidemic. Therefore, we can also continuously modify the iterative prediction model according to the actual data, so that the prediction of the next stage and the prediction of the long-term situation can be more accurate.

3 Application: Analysis of the COVID-19 in mainland China beyond Hubei Province

We apply our model to analyze and evaluate the COVID-19 using publicly available data from mainland China beyond Hubei Province from the China CDC during the period of Jan 29th, 2020, to Feb 29th, 2020. Here we first show the actual trend of the COVID-19, and then compared with the predicted ones via the proposed method. Finally, we will show the effect of m on the predicted results. All these results are implemented via **R** software.

3.1 The turning points and "zero" points observed

After the shutdown of most parts of Hubei province on Jan 23rd, other parts of China also immediately launched prevention and control strategies, including regional isolation, admission of all confirmed patients, isolating all suspected patients and so on. The effective implementation of these intervention policies quickly controlled the rapid spread of the epidemic in these areas. As can be seen in Figure 1, the parameter infectious rate K_t, which reflects the intensity of the spread of the epidemic, has shown a significant downward trend since Jan 27th after severe fluctuations from Jan 22nd to 26th. As can be seen in Figure 1, we find out that the daily confirmed cases peaked on Jan 30th, 2020, with 761 confirmed cases and then continued to decline for two consecutive days.

However, the migration raised from people returning to work after Chinese New Year on Feb 3rd undermines the continuous decline of E_t. Since Feb 2nd, the number of daily confirmed patients in mainland China beyond Hubei Province has increased for two consecutive days, where the E_t on Feb 3rd has increased by 23% compared to that on Feb 2nd. It can be concluded that these fluctuations are caused by the resuming of social activities, which leads E_t to continue to decline since Feb 4th. In many literature and media reports, Feb 3rd is used as the time point when the number of newly confirmed patients starts to decline. But considering the fact that the epidemic was already under control, here we still view Jan 30th as the first turning point.

After that, the second turning point T_2, which is the time point when the number of active cases N_t starts to decline, is also observed. Figure 2 shows the true curves of the daily infection rate K_t, daily removed rate

I_t, and N_t calculated based on the actual data from mainland China beyond Hubei from Jan 22th, 2020 to Mar 13th, 2020. It can be seen that the second turning point T_2 appeared on Feb 11th, with the emergence of $K_t < I_t$ on that day, and the number of patients in the hospital continued to decrease since then.

As for the first "zero" point Z_1, the definition is the time when the number of daily confirmed cases is equal to "zero", which is too strict for the real situation. Thus, in this article, we take the criteria for cancelling travel warnings developed by the WTO during SARS as a reference, and make some adjustments to the definition of the first "zero" point: the time when the daily confirmed cases E_t continues to be less than 5 for 3 days is revised to be Z_1. Then, if we exclude confirmed cases that originated from abroad, daily confirmed cases has already become less than 5 since Mar 3rd in mainland China beyond Hubei Province, thus according to our revised definition, Mar 5th is Z_1. However, there were still 1,089 active cases on that day. Therefore, it would still take some extra time to reach the second "zero" point Z_2.

3.2 Prediction results

Starting from Jan 29th, we use the proposed forecasting method to make real-time predictions on the two turning points T_1 and T_2 and two "zero" points Z_1 and Z_2 with window size $m = 5$. To clarify, the data before January 26th fluctuates violently, with assumption unsatisfied. Only after January 27th the data becomes stable, thus we waited 2 days to make sure the trend had emerged and began our prediction at January 29th. The specific and predicted results are as follows.

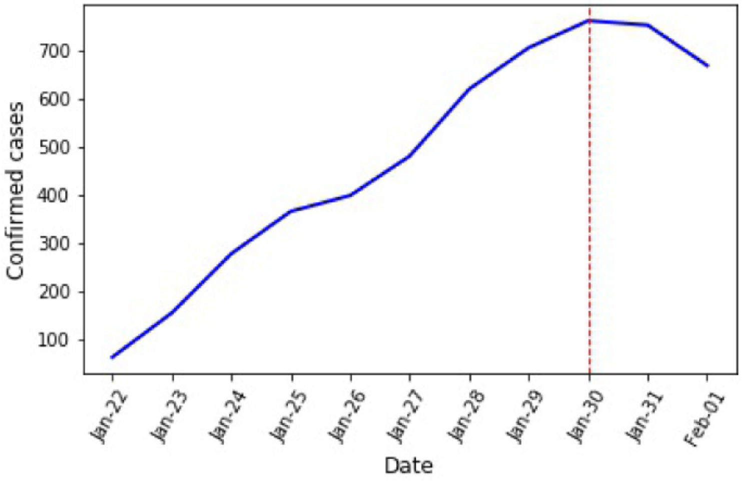


Figure 1. Trend of the daily confirmed cases from 01/22 to 02/01, 2020.

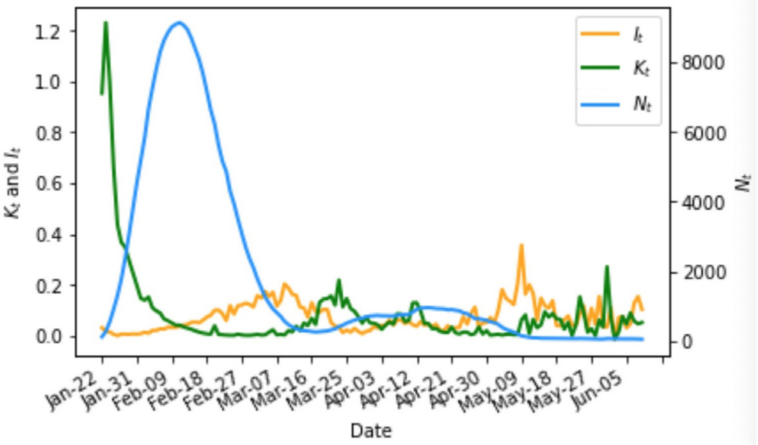


Figure 2. Observed K_t, I_t and N_t of the COVID-19 from Jan 22 to Jun 10, 2020.

We first conducted the proposed prediction model on Jan 29th, which indicated that the first turning point T_1 would arrive on Jan 31st, i.e., $E_t < E_t - 1$. In reality, the first turning point did arrive on Jan 30th, which is only one day away from our predicted result.

As for the second turning point, since the true T_2 occurred on Feb 11th, we summarize the frequency of the prediction results obtained with t_0 varying from Jan 29th to Feb 10th, 2020 and $m = 5$ in Figure 3(a). From it we can see that the prediction of the second turning point mainly concentrated in the range from Feb 9th to Feb 11th, which is consistent with the observed second turning point in reality. It is worth mentioning that we got the general information of T_2 at a very early stage: we predicted on Feb 2nd that the second turning point T_2 would arrive on Feb 11th, which is exactly the same as the second turning point that observed in reality. Since then, we have continuously tracked the rolling predictions, which have not yet changed much.

Similarly, Figure 3(b) and Figure 3(c) show the frequency of the prediction results for two "zero" points obtained with t_0 varying from Jan 29th to Feb 29th, 2020 and $m = 5$, respectively. Specifically, for the predicted first "zero" point Z_1 in Figure 3(b), we divide the prediction results from these days into 5 intervals, which can be seen that the prediction results of the first "zero" point Z_1 are mainly concentrated on Mar 1st to 5th, which is consistent with the actual result. There is also a "pessimistic" prediction as a result of the sudden fluctuation of data on Feb 3rd, which predicted that the first "zero" point would arrive on Mar 17th. For the predicted second "zero" point Z_2 in Figure 3(c), it can be seen that the second "zero" point will be reached from early-March to late-March. However, there is a prediction result that Z_2 will appear on May 11th, which is far away from other results. The reason for this uncommon result is that the starting point of this forecast is Jan 29th, when the epidemic situation in mainland China beyond Hubei was still in the outbreak period with E_t still rising, I_t very small, so the prediction result about the finish of the epidemic may not be accurate.

Furthermore, we also present the forecast results of the four milepost moments together with the trend of the cumulative number of active cases $\hat{N}_t$ and the cumulative number of infectious $\sum_{l=1}^{t} \hat{E}_l$ in Figure 4 when the prediction starting point t_0 fixed at Jan 29th, Jan 31st, Feb 12th and Feb 26th, 2020, respectively. As can be seen from Figure 4(a), on Jan 29th, which is the very early stage of the epidemic, we predicted that the first turning point would appear on Jan 31st, which is only one day behind the actual observation. Additionally, the time of the second turning point result predicted on that day was Feb 14th, which is only 3 days away from the reality. The first ”zero” and second ”zero” forecast results are Mar 7th and May 11th, respectively.

Figure 4(b) shows the prediction results when the first turning point have already appeared, from which we can see that the prediction for T_2 on Jan 31st is accurately with the second turning point possible occurring on Feb 11th. Meanwhile the first "zero" point and the second "zero" point are predicted to appear around Mar 4th and Mar 23rd, respectively.

Similarly, after the arrival of the second "zero" point, Figure 4(c) shows the forecast results of the first and second "zero" points predicted on Feb 12th, which show the forecast results for Z_1 and Z_2 are on Mar 9th and Mar 25th, respectively. From the fitting results, we know that our prediction of the cumulative number of active cases N_t and the total number of confirmed patients is very similar to the actual situation, so our prediction

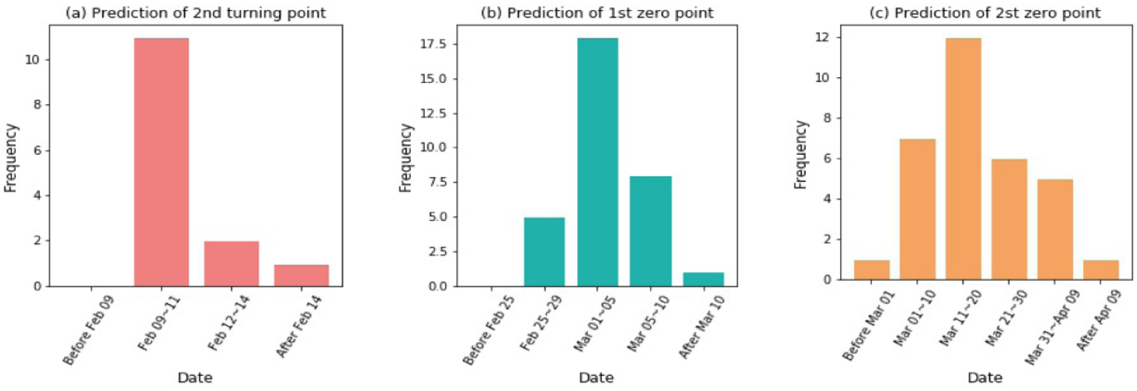


Figure 3. The frequency of prediction results of turning points and ”zero” points.

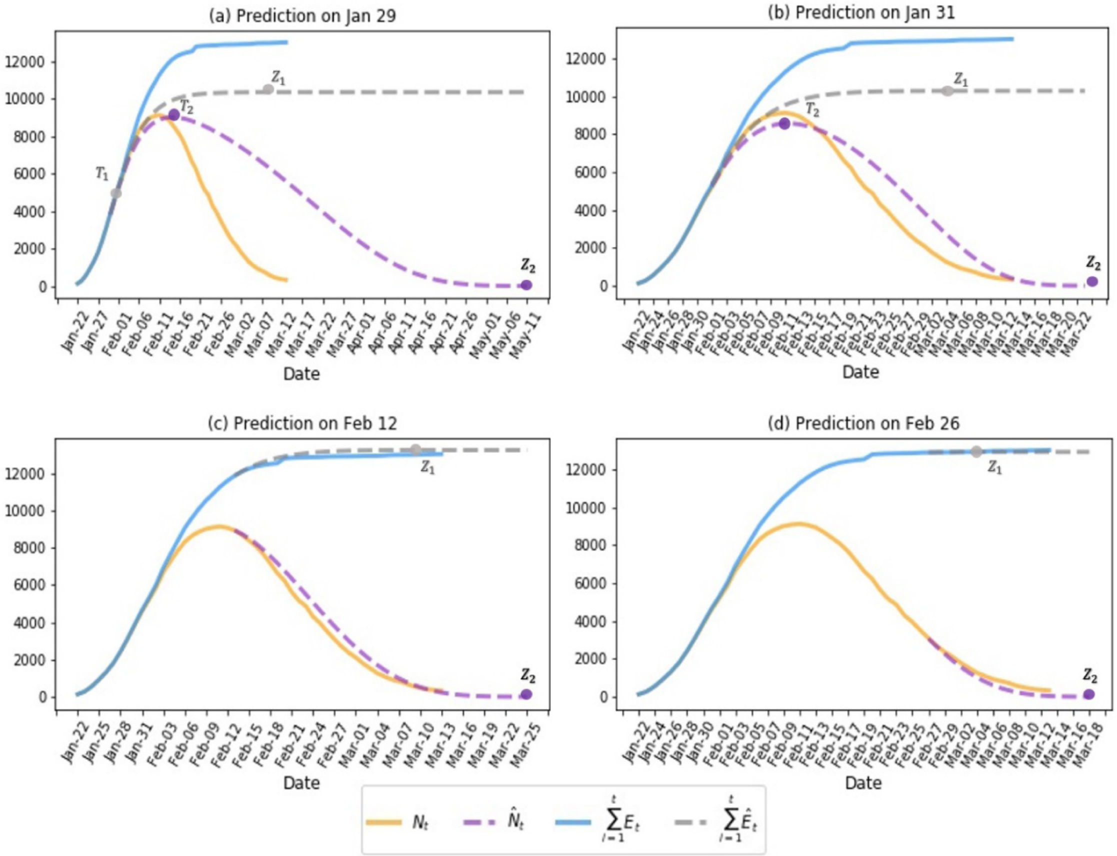


Figure 4. Forecasting results of the four milepost moments together with the trend of the cumulative number of active case $\hat{N}_t$ and the cumulative number of infectious $\sum_{l=1}^{t}\hat{E}_l$ compared with their observed cases N_t and $\sum_{l=1}^{t}E_l$ when the prediction starting point t_0 fixed at Jan 29th (**a**), Jan 31st (**b**), Feb 12th (**c**) and Feb 26 (**d**), 2020, respectively.

results are likely reliable. Finally, we also give a very recent (Feb 26th) forecast in Figure 4(d), which is similar to the results mentioned above.

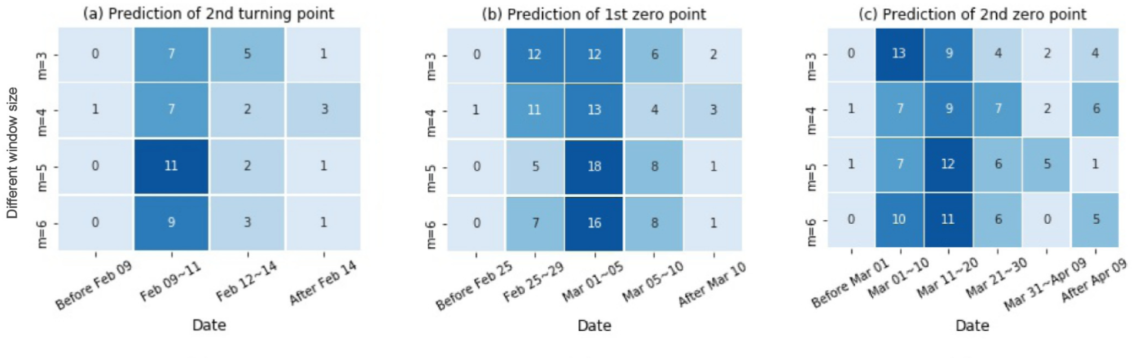
3.3 Results with different window sizes *m*

Note that the number of m plays an important role in the proposed procedure, and all the results we discussed in the section 3.2 are obtained with fixed $m = 5$. In this section, we will illustrate the impact of different choice of m on the results, and give the empirical choice in real data analysis. Parallel to Section 3.2, here we obtain the results for the second turning point and both "zero" points via implementation of the proposed procedure with m =3, 4, and 6, respectively. We summarize all these results for the second turning point and both "zero" points in Figure 5, respectively.

From Figure 5, we can see that the highest frequency of prediction results for the second turning point occur around the period from Feb 9th to 11th for all choice of m, which means that the second turning point is most likely to occur during this period; similar results hold for the forecast of the first "zero" with the most likelihood of appearance around the early March. Both results show the limited influence of m on the results. From Figure 5(c), although the results of forecast frequency distributions for the second "zero" point with different m seem not as concentrated as those for the second turning point and the first "zero", it varies slightly, with its occurrence from mid-March to mid-April. Overall, the choice of m seems not to be a critical value for the forecasting results, and we recommend its empirical choice from 3 to 6.

4 Discussion and conclusion

Focusing on the four meaningful mileposts, we put forward a simple and effective framework incorporating the effectiveness of the government control to forecast the whole process of a new unknown infectious disease in



(a) (b) (c)

Figure 5. Summary of prediction for the second turning point (**a**), the first (**b**) and second (**c**) "zero" points with different *m*.

its early-outbreak. Specifically, we first propose a series of iconic indicators to characterize the extent of epidemic spread, and describe four periods of the whole process corresponding to the four meaningful milepost moments: two turning points and two "zero" points; then we develop the proposed procedure with mild and reasonable assumption, especially without relying on an assumption of epidemiological parameters for disease progression.

We examine our model with COVID-19 data in mainland China beyond Hubei province, which can detect the gross process of the epidemic at its early-outbreak. Specifically, in the first predicting task that conducted on Jan 29, the predicted date when the number of newly confirmed patients E_t would fall for the first time is only one day behind the observation in reality. On Feb 2nd, our model predicted that the date when the number of patients in the hospital N_t reaches its peak is Feb 11th, which is consistent with the real world situation. Later, the forecasting results fluctuated but were overall stable and close to the true observation. Meanwhile, we predict that the first "zero" point Z_1 will arrive between the end of Feb and the beginning of March. And the second "zero" point Z_2 will arrive at mid-March to mid-April. We also checked the robustness of our model under different time windows and found that the selection of the time window has little effect on the prediction of turning points. As a prediction model for the task of early warning of a new epidemic, our prediction model is proved to be quite efficient.

At present, many countries around the world are overwhelmed by the COVID-19 epidemic, which calls for global efforts. While our method is able to depict and predict the trend of an epidemic at a very early stage, it can be used to predict the current COVID-19 epidemic internationally, or any other new, unknown, explosive epidemic in the future. We believe that the prediction results of this method can provide decision support for epidemic control and intervention. It is worth noting that, due to the short-term dependence of our method, our model may show poor performance for wildly fluctuating data. Thus, more data preprocessing methods like data smoothing need to be developed within our framework, in order to allow for wider use of our method.

Data availability

The underlying data and code required to replicate the studies finding are available from GitHub (data: https://github.com/Vicky-Zh/Tracking_and_forecasting_milepost_moments_of_COVID-19/tree/v1.0.0, code: https://github.com/YuanchenZhu2020/DemoPreTurningPointsCOVID19) and archived with Zenodo (data: http://doi.org/10.5281/zenodo.3755197 (Zhang, 2020), code: https://doi.org/10.5281/zenodo.398724 (YuanchenZhu2020, 2020).

Underlying data

Zenodo: Vicky-Zh/Tracking and forecasting milepost moments of COVID-19: First release. http://doi.org/10.5281/zenodo.3755197 (Zhang, 2020).

This project contains the following underlying data:

- Data of China Mainland Beyond Hubei.csv (A csv file with data collected from China CDC and four variables: the cumulative confirmed cases up to the given day *t*, the daily confirmed cases at day *t*, the daily recovered ones and the daily deaths at day *t*, with *t* from Jan 29th to Feb 29th, 2020)

Extended data

Zenodo:
YuanchenZhu2020/DemoPreTurningPointsCOVID19:
Version 1.0.0. https://doi.org/10.5281/zenodo.3987242 (YuanchenZhu2020, 2020).

This project contains the following extended data:

- **DemoPreTurningPointsCOVID19_1.0.0.zip** (R binary package)
- **DemoPreTurningPointsCOVID19_1.0.0.tar.gz** (R source package)
- **DemoPreTurningPointsCOVID19_1.0.0.pdf** (Reference manual for R package)

Data are available under the terms of the Creative Commons Zero ”No rights reserved” data waiver (CC0 1.0 Public domain dedication).

References

Anastassopoulou C, Russo L, Tsakris A, *et al.*: **Data-based analysis, modelling and forecasting of the novel coronavirus (2019-ncov) outbreak.** *medRxiv.* 2020.
Publisher Full Text

Benvenuto D, Giovanetti M, Salemi M, *et al.*: **The global spread of 2019-ncov: a molecular evolutionary analysis.** *Pathog Glob Health.* 2020; **114**(2): 64–67.
PubMed Abstract | Publisher Full Text | Free Full Text

Chen Z, Zhang W, Lu Y, *et al.*: **From sars-cov to wuhan 2019-ncov outbreak: Similarity of early epidemic and prediction of future trends.** *bioRxiv.* 2020.
Publisher Full Text

Chinazzi M, Davis JT, Ajelli M, *et al.*: **The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak.** *Science.* 2020; **368**(6489): 395–400.
PubMed Abstract | Publisher Full Text | Free Full Text

Fan Y, Zhao K, Shi ZL, *et al.*: **Bat coronaviruses in China.** *Viruses.* 2019; **11**(3): pii: E210.
PubMed Abstract | Publisher Full Text | Free Full Text

Holshue ML, DeBolt C, Lindquist S, *et al.*: **First case of 2019 novel coronavirus in the united states.** *N Engl J Med.* 2020; **382**(10): 929–936.
PubMed Abstract | Publisher Full Text | Free Full Text

Huang C, Wang Y, Li X, *et al.*: **Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.** *Lancet.* 2020; **395**(10223): 497–506.
PubMed Abstract | Publisher Full Text | Free Full Text

Hui DS, EI Azhar E, Madani TA, *et al.*: **The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China.** *Int J Infect Dis.* 2020; **91**: 264–266.
PubMed Abstract | Publisher Full Text | Free Full Text

Li Q, Guan X, Wu P, *et al.*: **Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia.** *N Engl J Med.* 2020; **382**(13): 1199–1207.
PubMed Abstract | Publisher Full Text | Free Full Text

Luk HKH, Li X, Fung J, *et al.*: **Molecular epidemiology, evolution and phylogeny of sars coronavirus.** *Infect Genet Evol.* 2019; **71**: 21–30.
PubMed Abstract | Publisher Full Text | Free Full Text

Peng L, Yang W, Zhang D, *et al.*: **Epidemic analysis of covid-19 in China by dynamical modeling.** *medRxiv.* 2020.
Publisher Full Text

Prasse B, Achterberg MA, Ma L, *et al.*: **Network-based prediction of the 2019-ncov epidemic outbreak in the chinese province Hubei.** *arXiv preprint arXiv: 2002.04482.* 2020.
Reference Source

Tang B, Wang X, Li Q, *et al.*: **Estimation of the transmission risk of the 2019-ncov and its implication for public health interventions.** *J Clin Med.* 2020; **9**(2): pii: E462.
PubMed Abstract | Publisher Full Text | Free Full Text

World Health Organization: **Novel coronavirus (2019-ncov) situation reports**. 2020.
Reference Source

Wu JT, Leung K, Leung GM: **Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in Wuhan, China: a modelling study.** *Lancet.* 2020; **395**(10225): 689–697.
PubMed Abstract | Publisher Full Text | Free Full Text

YuanchenZhu2020: **YuanchenZhu2020/DemoPreTurningPointsCOVID19: Version 1.0.0 (Version V_1.0.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3987242

Yu WB, Tang GD, Zhang L, *et al.*: **Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data.** *Zool Res.* 2020; **29**: 1–11.
PubMed Abstract | Publisher Full Text

Zeng T, Zhang Y, Li Z, *et al.*: **Predictions of 2019-ncov transmission ending via comprehensive methods.** *arXiv preprint arXiv: 2002.04945.* 2020.
Reference Source

Zhang Y: **Vicky-zh/tracking and forecasting milepost moments of covid-19: First release**. 2020.
https://www.zenodo.org/record/3755197

Zhang H, Kang Z, Gong H, *et al.*: **The digestive system is a potential route of 2019-ncov infection: a bioinformatics analysis based on single-cell transcriptomes.** *BioRxiv.* 2020.
Publisher Full Text

Zhao S, Lin Q, Ran J, *et al.*: **Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak.** *Int J Infect Dis.* 2020; **92**: 214–217.
PubMed Abstract | Publisher Full Text | Free Full Text

Zhou T, Liu Q, Yang Z, *et al.*: **Preliminary prediction of the basic reproduction number of the wuhan novel coronavirus 2019-ncov.** *J Evid Based Med.* 2020; **13**(1): 3–7.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 14 December 2020

https://doi.org/10.5256/f1000research.28903.r71640

✓

Paula Brito
Faculty of Economics, University of Porto & LIAAD - INESC TEC, Porto, Portugal

In this new version, the authors have clarified the issues raised, mainly about the applicability of the proposed method, and the choice of the starting point.
As mentioned before, the results reported from the application of the proposed methodology to the China Covid-19 data are quite good, showing that the method is well founded and useful.

I look forward to seeing this being applied to data from other regions/countries, and results discussed.

The manuscript is quite well written, and is easy to follow. There are nevertheless some typos/small issues that need correction, as indicated here-below.

I consider that the method is sound, and not complicated, neither relying on extraordinary assumptions.
I consider this is an interesting and nice piece of work, in a pertinent and up-to-date topic, and I am therefore in favour of acceptance.

Minor issues :

Page 3, line 3 : reported in -> reported on
Page 3, line 5 : stopping the spread -> stopping of the spread
Page 3, line 12 : of highly -> highly
Page 3, line 32 : is as high -> as high

Page 4, line 26 : correct “specifically”
Page 4, line 38 : proposed -> propose
Page 4, line 39 : defined -> define
Page 4, line 42 : applied -> apply

Page 5, line 20, formula for N_t : in the sum, is it E_t or E_i ?

Page 5, last line : the right side of the formula should be: $N_0 (1 + K - I)^t$ (there is a '1' instead of 'I')

Page 6, line 22 : data -> date
Page 6, line (-13) : $K_t < I_t$,$R_t = 1 + K_t - I_t < 1 \Rightarrow N_t < N_{t-1}$ - in the center it should be '<' and not '>', $R_t < 1$
Page 6, line (-10) : pressure of -> pressure on
Page 6, lines (-4)-(-3) : number of people treated in the hospital -> number of active cases

Page 7, Table 1, cell in line 1, column 3 : patients in hospital -> active cases
Page 7, line (-11) : number of patients in the hospital -> number of active cases
Page 7, line (-3) : chosen satisfy -> chosen to satisfy

Page 8, lines 1-2 : have already emerged and stable -> has already emerged and is stable
Page 8, line 2 : outbreak have already -> outbreak has already
Page 8, line 3 : are satisfied -> is satisfied
Page 8, Algorithm 1, point 1 : remove "which"
Page 8, line 5 after the Algorithm box : the -> The
Page 8, line (-19) : compared -> compare
Page 8, line (-13) : Figure 1 -> Figure 2

Page 9, line 3 : number of patients in the hospital -> number of active cases
Page 9, lines 13 and 14 : "the data": please indicate what you refer to exactly

Page 10, line 1 : Please correct : $E_t < E_{t-1}$ (index of the second term)
Page 10, line 9 : that -> that was
Page 10, line (-4) : "after the arrival of the second "zero" point, Figure 4(c) shows..." - is it "zero point" or "turning point" here ?...

Page 11, line 4 : number -> value
Page 11, line 5 : section -> Section
Page 11, line (-10) : choice -> choices

Page 12, line 4 : assumption -> assumptions
Page 12, line 6 : that conducted -> that was conducted.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Multivariate Data Analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 09 October 2020

https://doi.org/10.5256/f1000research.28903.r71641

✓

Rosanna Verde
Department of Mathematics and Physics, University of Campania "Luigi Vanvitelli", Caserta, Italy

I completely agree with the new version.
The authors have improved the paper also following the comments and suggestions I gave.
The paper is interesting and the proposed methods could be applied to COVID-19 data of other Countries.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistical Data Analysis; Data Stream Analysis; Functional Data Analysis; Symbolic Data Analysis; Clustering

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 14 July 2020

https://doi.org/10.5256/f1000research.25510.r63123

?

Paula Brito
Faculty of Economics, University of Porto & LIAAD - INESC TEC, Porto, Portugal

In this paper, the authors propose a methodology aimed at forecasting milepost moments of an epidemic in the early outbreak. This is then applied to the COVID-19 outbreak in China.
The method lays on the definition of the relevant involved variables :

- E_t : number of new confirmed cases at date t
- O_t : number of recovered cases at date t
- D_t : number of deaths at date t
- N_t : number of infectious cases in hospital at date t
- K_t : infection rate at date t
- I_t : removal rate at date t
- R_t : outbreak status at date t

The milepost moments to be predicted are

- the first turning point, T1 – when the number of newly diagnosed patients E_t starts decreasing;
- the second turning point, T2 – when number of active cases N_t starts decreasing;
- the first zero point, Z1– when the number of newly diagnosed patients E_t becomes null ;
- the second zero point, Z2– when the number of active cases N_t becomes null.

The procedure depends on the starting point t_0, and on the size of the time window used, m. It is assumed that K_t and I_t change gently within time window m before t_0, so that the average change rate of both K_t and I_t within that period are reliable values.
Under those assumptions, recurrence formulas allow predicting the required time points.

The results reported from the application of the proposed methodology to the China Covid-19 data are quite good, predicting well the observed time points, and hence show that the method is well founded and useful.

The manuscript is quite well written, and is easy to follow.

The pertinence and actuality of the topic go without saying. The method appears sound to me, and is not extremely complicated, neither relying on extraordinary assumptions.

I consider this is an interesting and nice piece of work, and I am therefore in favour of acceptance.

However, a few issues need consideration/correction, as described here below.

The authors should revise the manuscript taking these aspects into account.

Major points

1. The success of the proposed methodology lays on the fact that the values of the average change rate of both K_t and I_t, within the considered time window period, are reliable and stable - so that they may be used for prediction. It is not clear to me how this condition may be assessed, or what are the criteria involved. This should be clarified.
   The fact that the reported predictions for China are indeed good, result, in my opinion, from the fact that the local situation has been stable, measures taken and local conditions not changing before the general end of the outbreak (or so it is perceived from the outside). If this were not the case, I wonder whether the method would still apply – which then questions its applicability in other countries /regions where such stability is not guaranteed/observed.
   This emphasises the need for a clear establishment of applicability conditions.
   This is my major concern.

2. I am curious about the prediction results if the starting point is a bit earlier, and not as close to T1 as January 29th.

3. N_t is defined as “number of infectious cases in hospital at date t”. I believe this is what is generally referred to as “active cases” – also because in many countries not all diseased people are actually in hospital. I suggest using this terminology, so that it becomes clear to a general reader. In that case, the reference to “in hospital” should be updated along the

manuscript.

4. Algorithm 1 : there are some typos here, that require attentive correction: in point 3,
   - formula 1: it should be K and not R , in both left and middle terms
   - formula 3, right term: R should be replaced by K

Minor issues

Page 1, line 15 of the Abstract : assumption -> assumptions
Page 1, line 16 of the Abstract : COVID-19 -> COVID-19 outbreak
Page 1, line 25 of the Abstract : counties -> countries

Page 3, line 35 : results -> result

Page 4, line 30 : generalized -> used
Page 4, line 36 : assumption -> assumptions

Page 8, Algorithm 1, line 2 : satisfying -> satisfy
Page 8, Section 3, line 7 : in Jan 23rd -> on Jan 23rd

Page 9, line 12 : decreases -> decrease

Page 10, line 15 : in Jan 29th -> on Jan 29th
Page 10, line (-13) : zero -> turning

Page 11 – Figure 5 : the labels along the horizontal axis are not centered, therefore the figure is not clear.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Multivariate Data Analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 17 June 2020

https://doi.org/10.5256/f1000research.25510.r63126

?

Rosanna Verde

Department of Mathematics and Physics, University of Campania “Luigi Vanvitelli”, Caserta, Italy

The paper deals with a forecasting procedure to analyse and evaluate the early stage of the COVID-19 outbreak in China. Inspite the classical SEIR model, the authors define a set of iconic indicators to study the status of disease contagion and the extent of epidemic spread. Then, they divide the cycle of the epidemic in four stages corresponding to the four meaningful milepost moments: two turning points and two “zero” points, related to the tuning of the proposed indicators.

The authors characterize the first stage: the *Outbreak period*, by its end, which is represented by the arrival of the first tuning point (T_1), when the newly diagnosed patients (denoted E_t) after a rapid rise, begin to decrease (E_t < E_{t-1}). The second stage is the *Controlled period*, that sees an increasing of hospitalised patients (denoted N_t) until a second turning point (T_2), when N_t, after to have reached a peak, starts to decline.

The *Mitigation stage* corresponds to the third stage and it continues until the daily confirmed cases reach the zero (E_t=0). That represents the first “zero” point (Z_1). This condition is also reformulated and referred to the time when the daily confirmed cases are less then “5” for 3 consecutive days.

The *Convergence Period*, the four stage, is characterized to reach a second “zero” point (Z_2), when the number of hospitalised patients is closed to zero (N_t=0). That signs the end of the epidemic.

The procedure is corroborated on public available data in mainland China beyond Hubei Province from the China CDC during the period of Jan 29th, 2020, to Feb 29th, 2020. The results show the proposed procedure has provided to a prediction of the tuning and “zero” points, very near to the exact dates also in a very large stage of the epidemic.

An analysis of the robustness of the method has also been performed on the time windows and on the four milepost moments. The results have revealed the influence of the parameter “m”, related to the width of the time windows used to estimate the average change rate of the daily diagnosed cases and of the daily removed cases.

The paper is clear and smoothly written. I agree for its indexing after to have kept into consideration the following remarks:

Section 2. Methods
Equation (1)

1) N_l in brackets (1+K_l - N_l) should to be replaced with I_l

Section 2.3

2) I suggest of referring explicitly the exponential model of the infection rate K_t and of the removed rate I_t (in the window m);

3) the term “average” for V_{K|(t0,m)} (and V_{I|(t0,m)}) is confused, it would be better to refer to it as “unitary rate of change” (and “unitary rate of removed”). I think, that is more properly the “rate of change” (and the “rate of removed”) associated with a unitary time step.

Algorithm 1 Main Prediction Procedure.
3 Prediction: updating the predicted results .....

4) The first equation R_t|to = \hat{R_t|to (l)} should to be replaced with K_t|to = \hat{K_t|to (l)}

Section 3: Application

5) All the Figures must be improved, especially the Figure 2. And the Figure 4.

6) The proposed strategy has been verified on data referred to a very early stage of the epidemic and it performs short-term dependency. Given to the spread of the epidemic in world wide, maybe it can be proved on data related to the outbreak of the pandemic in another Countries.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistical Data Analysis; Data Stream Analysis; Functional Data Analysis; Symbolic Data Analysis; Clustering

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.