Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 187 (2021) 512–517

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Identification, Information and Knowledge in the internet of Things, 2020

Estimate the Trend of COVID-19 Outbreak in China: a Statistical and Inferential Analysis on Provincial-level Data

Kun Li[a], Yangyang Zhang[b], Chao Wang[c,*]

[a]*Business School, Beijing Normal University, Houzhulou Building, 19 Xinjiekouwai Street, Beijing 100875, China*
[b] *Department of Pharmacy, Armed Police Beijing Corps Hospital, Beijing 100600, China*
[c] *Research Department, ZCE Futures & Derivatives Institute Co., LTD., 31 Longhuwaihuan Eest Road, Zhengzhou, 450000, China*

Abstract

The ongoing COVID-19 epidemic spreads with strong transmission power in every part of China. Analyses of the trend is highly need when the Chinese government makes plans and policies on epidemic control. This paper provides an estimation process on the trend of COVID-19 outbreak using the provincial-level data of the confirmed cases. On the basis of the previous studies, we introduce an effective and practical method to compute accurate basic reproduction numbers (R_0s) in each province-level division of China. The statistical results show a non-stop downward trend of the R_0s in China, and confirm that China has made significant progress on the epidemic control by lowering the provincial R_0s from 10 or above to 3.21 or less. In the inferential analysis, we introduce an effective *AR(n)* model for the trend forecasting. The inferential results imply that the nationwide epidemic risk will fall to a safe level by the end of April in China, which matches the actual situation. The results provide more accurate method and information about COVID-19.

Keywords: COVID-19; Basic Reproduction Numbers; Trend; China; Provincial-level Data; Statistical and Inferential Analysis;

Kun Li, *E-mail address: kunli@bnu.edu.cn; Yangyang Zhang, E-mail address: landy17@163.com.*
* Chao Wang, Corresponding author. *E-mail address:* wangchao@zcefdi.com.cn / cwang118@hawk.iit.edu

10.1016/j.procs.2021.04.092

1. Introduction

The outbreak of COVID-19 from Hubei, China at the end of December 2019 has instantly spread to become a nationwide and even global pandemic. A number of recent studies had dedicated into the analyses of the trend of COVID-19. For example, Jung et al., 2020 take the basic reproduction number as a key measure to estimate the risk of death from the COVID-19 infection [1]. Roosa et al. 2020 generate short-term forecasts of cumulative reported cases in Guangdong and Zhejiang provinces of China, and detect the different trends across locations [2]. Inspired by these recent publications, we figure out that the diversity of COVID-19 trends in different areas is critical to concern, although the majority of cases (56,249) were reported in Hubei as of February 16th, 2020.

In this paper, we explore the diversified trends of COVID-19 in province-level divisions of China. We first introduce an effective and accurate method to estimate the basic reproduction numbers. Second, we use the daily data of the cumulative confirmed COVID-19 cases and a news gather technique to acquire the accurate inputs of the key component factors for the computation of basic reproduction numbers. As a contribution, we provide a time-series array of detailed daily basic reproduction numbers in each province-level division of China. Third, using the computed provincial-level data, we conduct a statistical and inferential analysis on the data. Our statistical results indicate that the outbreak of COVID-19 is still in the process to be contained and does not move into a steady state with all individuals healthy in any region of China. But the regional basic reproduction numbers had been close to each other. We construct an Auto-regression (AR) model and quantify how the historical basic reproduction numbers account for the future values. Our model state that the nationwide trend of COVID-19 would fall to a safe level in late April 2020.

2. Method of Basic Reproduction Numbers

According to Anderson and May (1992), Wang and Zhao (2012), Pastor-Satorras et al (2015), the basic reproduction number R_0 is one of the most popular and important indicators in epidemic studies [3, 4, 5]. It quantifies the infection risk of the epidemic by estimating the expected number of secondary cases caused by the initial infective case. The basic reproduction number has been widely used for different epidemics. In this study, we use the method by Zhou et al (2020) to calculate the basic reproduction numbers of the COVID-19 [6].

The basic reproduction number R_0 can be computed by

$$R_0 = 1 + \lambda_t \, GT + \rho(1-\rho)(\lambda_t \, GT)^2 \tag{1}$$

To compute R_0, we need information about the three component factors: λ_t, GT and ρ. λ_t is the growth rate of the confirmed cases at the day *t* since the onset of the infection, written as

$$\lambda_t = \frac{ln\,Y(t)}{t} \tag{2}$$

Where *Y(t)* is the cumulative number of the COVID-19 confirmed cases by day *t*. GT refers to the generation time, which is defined as the time interval between symptom onset in an index case and symptom onset in a secondary case. ρ is the ratio of the incubation period and GT.

3. Data and Inputs of the Three Component Factors

In this section, we compute inputs of the three component factors, λ_t, GT and ρ, to calculate R_0 in each date and each province-level division of China. We collect the numbers of the confirmed cases in each of the 34 province-level divisions of China. We start from January 21st, 2020, the first day when the NHC China provided the real-time platform immediately after the State Council of China had launched the nationwide alert to COVID-19 epidemic control one day before, till February 16th, 2020. The main data source is the National Health Commission of China (NHC China). It provides a real-time platform to update the numbers of COVID-19 confirmed cases sorted by the province-level divisions (including Hong Kong, Macau and Taiwan). We also acquire the data from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

First, we compute λ_t. According to (2), we collect *Y(t)*, the cumulative number of the COVID-19 confirmed cases,

in each date and each province-level division. We then track the onsets of the infection to determine *t*.

Previous studies that use the only date as the onset for every province-level division. As a progress, we use the precise onset dates of COVID-19 in each province-level division. In each province-level division, we find the date when the first case was confirmed in official announcements of the NHC China and its provincial health commission branch, and define the day before it as the onset (*t=0*). To precisely track the information and thus lock the onset date in each province-level division of China, we use the iFinD database for news subscription, a news gather technique which is widely used in financial studies [7]. According to our knowledge, it is the first study that provides the accurate onset dates in each province-level division of China. The accurate onsets improve the reliability and accuracy of our results.

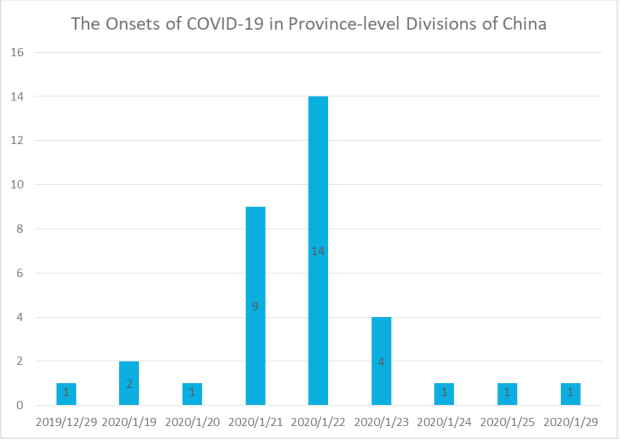

Fig. 1. The Onsets of COVID-19 in Province-level Divisions of China

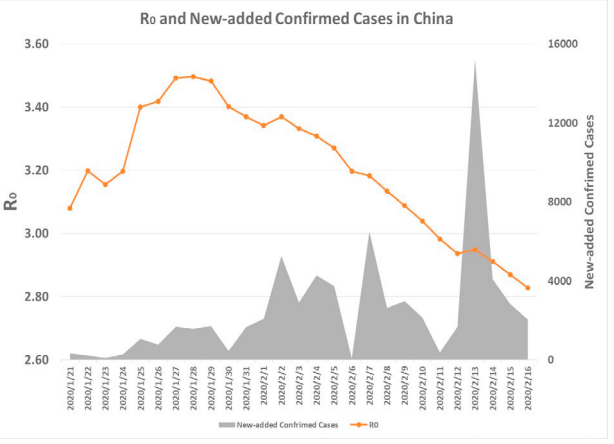

Fig. 2. R_0 and New-added Confirmed Cases in China

Table 1 and Figure 1 summarize the variety of COVID-19 onsets across province-level divisions. As the origin of the COVID-19 outbreak, Hubei had the first confirmed case early on December 29th, 2020, according to the official report of Chinese Center for Disease Control and Prevention (CCDC) published on January 28th, 2020. Following Hubei, Beijing and Guangdong detected their first confirmed cases on January 19th, 2020, and Shanghai one more day later. Thus, the top three developed provinces and municipalities started the epidemic alarm ahead of the rest of China. The overall outbreak of COVID-19 started on January 21st, when nine more provinces and municipalities announced the first confirmed cases in their domains. On January 22nd, 14 province-level divisions had the first confirmed cases, including Hong Kong and Macau. The COVID-19 spread the majority of China. Tibet was the last province that had the first confirmed case till January 29th, 2020.

Table 1. The Onsets of COVID-19 in Province-level Divisions of China

Date	Number	Province-level divisions
2019/12/29	1	Hubei
2020/1/19	2	Beijing; Guangdong
2020/1/20	1	Shanghai
2020/1/21	9	Chongqing; Henan; Hunan; Jiangxi; Shandong; Sichuan; Taiwan; Tianjin; Zhejiang
2020/1/22	14	Anhui; Fujian; Guangxi; Guizhou; Hainan; Hebei; Hong Kong; Jiangsu; Jilin; Liaoning; Macau; Ningxia; Shanxi; Yunnan
2020/1/23	4	Gansu; Heilongjiang; Shaanxi; Xinjiang
2020/1/24	1	Inner Mongolia
2020/1/25	1	Qinghai
2020/1/29	1	Tibet

Given *Y(t)* and *t*, we compute λ_t, the daily-updated growth rate of the confirmed cases between January 21st, and

February 16th, in each province-level division of China. We also compute the $\lambda_t s$ of the overall China, by choosing the onset in Hubei as the onset of China.

Next, we estimate ρ and GT. Previous studies state that the ρ of SARS is between 0.5 and 0.8 [8] [9]. Some recent studies treat SARS as a comparable target of COVID-19 and thus set up the ρ of COVID-19 equal to 0.65 [5] [24] under an optimistic estimation. However, according to the trend of COVID-19 in China during our observation period, this new epidemic has larger impact than SARS in 2003. Therefore, we set up the ρ of COVID-19 equal to 0.5 in this study, such as to maximize the coefficient $\rho(1-\rho)$ in (1) and thus depict the basic reproduction number in a more reasonable level.

As discussed previously, ρ is the ratio of the incubation period and GT, $\rho = \frac{\text{incubation period}}{GT}$. So the generation time GT can be computed as $GT = \frac{\text{incubation period}}{\rho}$. Previous studies usually use the incubation period of SARS as the substitute when the incubation period of COVID-19 is unavailable. According to the latest research, the median incubation period was 3.0 days [9]. So we choose 3.0 days as the incubation period of COVID-19. Thus, the generation time GT is 6.0 days.

In summary, the function of R_0 is

$$R_0 = 1 + 6\lambda_t + 0.5(1-0.5)(6\lambda_t)^2 \tag{3}$$

4. Results and Inferences

4.1. Results of the Basic Reproduction Number of COVID-19

We input the data of λ_t, GT and ρ, to calculate daily updated R_0 in province-level division of China. It is a common case to allow one (the unity) as the epidemic threshold. If $R_0 < 1$, the relative size of the epidemic is negligibly small and the epidemic will die out soon; If $R_0 > 1$, an infective individual can cause an outbreak of the epidemic in a certain size of population and will keep generating new infected individuals. In summary, R_0 can be considered as the expected number of cases directly generated by one case of an infection in a population.

We first look at the trend nationwide. Figure 2 depicts the nationwide daily R_0 and new-added confirmed cases. The peak of R_0 appeared on January 28th, 2020, when it reached 3.50. Since then, R_0 had decreased constantly. The increase of the new-added confirmed cases did not cause R_0 to bounce up to a higher level, even when some boom of confirmed cases occurred on February 2nd (5257 new cases confirmed), February 7th (6483 new cases confirmed), and February 13th (15216 new cases confirmed). Till February 16th, 2020, R_0 was controlled at 2.83, implying that one case of COVID-19 may lead to 2.83 new cases on average. Although the latest R_0 is still higher than one, the downward trend has explicitly shown the positive outcomes from current epidemic prevention and control.

Next, we study R_0s in the province-level divisions. Till February 16th, 2020, the lowest R_0 was equal to one from Tibet, where only one confirmed case appeared, while the rest of China had R_0s ranged between 1.60 and 3.21. The results are in line with our findings from the nationwide data. They indicate that in any region of China, the outbreak of COVID-19 is still in the process to be contained and does not move into a steady state with all individuals healthy.

Although the provincial R_0s are still higher than one, however, changes of the provincial R_0s clearly show the outcomes of epidemic prevention and control. Figure 3 presents the highest R_0s (peak R_0) in each province-level divisions of China during the observation period. The average peak R_0s of the 34 province-level divisions is 12.01. 15 of them had their peak R_0s between one and 10. 19 provinces or municipalities had their peak R_0s above 10, and three of them even reached up to 30 (Chongqing, Hainan and Zhejiang). By contrast, the latest average R_0 on February 16th has been lower to 2.56, which is 20% of the average peak R_0s occurred about twenty days ago. Figure 3 shows that the latest R_0s are very close to each other, indicating that nationwide epidemic prevention and control have been synchronized on the same page in different parts of China.

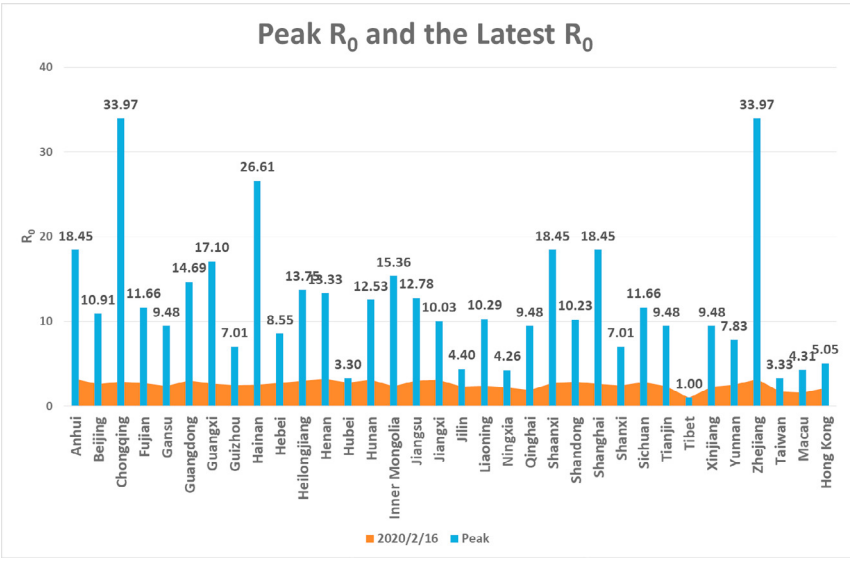

Fig. 3. Peak R_0 and the Latest R_0 (2020/02/16) in Province-level Divisions

4.2. Inferences of the trend of COVID-19

In this section we make a forecasting model of the R_0s. The descriptive statistical results above depict the downward trend of R_0 in both nationwide and provincial levels, implying that the R_0s may follow certain time-varying processes. According to Box and Jenkins (1976) and Ruud (2000) [10, 11], we construct the AR model and quantify how the current outcomes of R_0s will account for the future values. We first use the partial autocorrelation to determine the appropriate lag length. After determining the lag length n, we construct the AR(n) model for one array of R_0s, as described in (4).

$$R_0(t) = a + \sum_{i=1}^{n} bR_0(t-i) + e \tag{4}$$

We use the AR(n) model to forecast the nationwide trend of COVID-19. Depicted in Figure 4, the R_0 will keep moving down by a certain extent every day and eventually fall below one approximately on April 19th, in other words, in two months. Compared with the two-month time frame of the 2003 SARS outbreak [12], this result is reasonable.

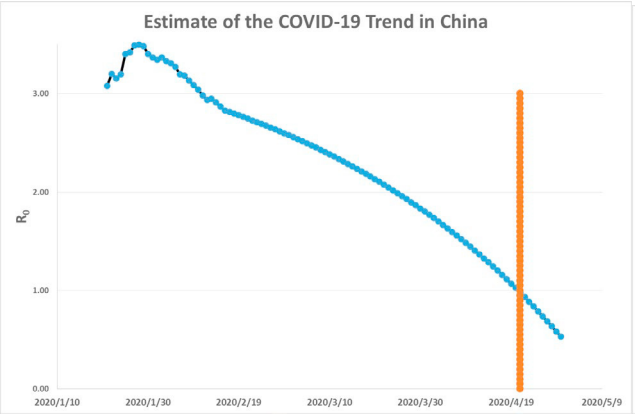

Fig. 4. Estimate of the COVID-19 Trend in China

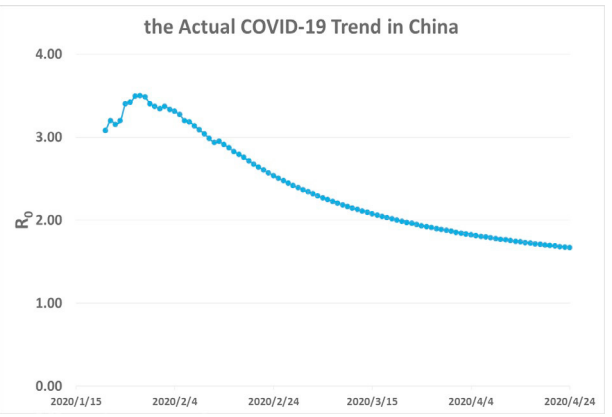

Fig. 5. The Actual COVID-19 Trend in China

We use the actual data to reexamine the validity of this estimated trend. Figure 5 depicts the actual trend of R_0 from January to the end of April, 2020. The actual R_0 shows a decreasing trend and gradually approaches one. The actual result is consistent with our estimated result using the first month data. It confirms the validity and reasonability of our estimating model.

5. Summary

This paper provides an estimation process on the trend of COVID-19 outbreak using the provincial-level data of the confirmed cases. We present daily updated basic reproduction numbers R_0s in each province-level division of China. We employ an effective and practical method to compute R_0s. The three component factors of this method are achievable through the daily updated data of the cumulative confirmed COVID-19 cases and a news gather technique. The outcomes of R_0s show a monotonic downward trend in the nationwide level since January 28th regardless of sudden booms of newly confirmed cases, which confirms the effects of current epidemic prevention and control in China. But the latest R_0s were still above the unity threshold, indicating that the outbreak of COVID-19 had not been contained into a steady state with all individuals healthy in any region of China.

Through the statistical analysis, we also find that China has made significant progress on the epidemic prevention and control. In some provinces the peak R_0s ran up to 33 and 2/3 of the province-level divisions had their peak R_0s once above 10. The Chinese government successfully lower the provincial R_0s below 3.21 and maintain the efforts of epidemic prevention and control on the same page across province.

In the inferential analysis section, we introduce an effective AR(n) model for the trend forecasting. We use it to forecast both the nationwide trend. Our results imply that the epidemic risk will fall to a safe level by the end of April in China, which matches the actual situation. The results provide more accurate method and information about COVID-19.

Acknowledgements

The authors, including Kun Li (https://orcid.org/0000-0001-5862-1959), Yangyang Zhang (https://orcid.org/0000-0002-7492-3022), and Chao Wang (https://orcid.org/0000-0002-6096-7550), gratefully acknowledge the support from the National Natural Science Foundation of China (Grant No. 71803012).

References

[1] Jung, Sung-mok, A.R. Akhmetzhanov, K. Hayashi, et al. (2020) Real-Time Estimation of the Risk of Death from Novel Coronavirus (COVID-19) Infection: Inference Using Exported Cases. *Journal of Clinical Medicine* **9** (**2**): 523.

[2] Roosa, Kimberlyn, Y. Lee, R. Luo, et al. (2020) "Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China: February 13–23, 2020." *Journal of Clinical Medicine* **9** (**2**): 596.

[3] Anderson, Roy M., and R. M. May. (1992) *Infectious diseases of humans: Dynamics and control*, Oxford, Oxford University Press.

[4] Wang, Wendi, and X. Q. Zhao. (2012). "Basic reproduction numbers for reaction-diffusion epidemic models." *SIAM Journal on Applied Dynamical Systems* **11** (**4**): 1652-1673.

[5] Pastor-Satorras, Romualdo, C. Castellano, P. V. Mieghem, et al. (2015) "Epidemic processes in complex networks." *Review of Modern Physis* **87** (**3**): 925-979.

[6] Zhou, Tao, Q. Liu, Z. Yang, et al. (2020) "Preliminary prediction of the basic reproduction number of the Wuhan novel coronavirus 2019-nCoV." *Chinese Journal of Evidence-Based Medicine* **13** (**1**): 3-7.

[7] Li, Kun. (2018). "Reaction to News in the Chinese Stock Market: A Study on Xiong'an New Area Strategy." *Journal of Behavioral and Experimental Finance* **19**: 36–38. https://doi.org/10.1016/j.jbef.2018.03.004

[8] Donnelly, Christl. A., A. C. Ghani, G.M. Leung, et al. (2003) "Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong." *Lancet* **361** (**9371**): 1761-1766.

[9] Leo, Y. S., M. Chen, B. H. Heng, et al. (2003) "Severe acute respiratory syndrome-Singapore 2003." *Morbility & Mortality Weekly Report* **52** (**18**): 405-432.

[10] Box, George E. P., and G. M. Jenkins. (1970) *Time Series Analysis, Forecasting and Control*, San Francisco, HoldenDay.

[11] Ruud, Paul A. (2000). "An introduction to classical econometric theory." *Southern Economic Journal* **67** (**4**): 1025-1028.

[12] Deng, Sheng-Qun, and H.-J. Peng. (2020) "Characteristics of and public health responses to the coronavirus disease 2019 outbreak in China." *Journal of Clinical Medicine* **9** (**2**): 575.