Published by SET Publisher

Journal of Basic & Applied Sciences

ISSN (online): 1927-5129

Data Fixing by Data Fitting: Estimating the Unreported Cases During the Early COVID-19 Outbreak in Hubei, China

Kamlesh Sarkar and Xiang-Sheng Wang*

Department of Mathematics, University of Louisiana at Lafayette, Lafayette, 70503, LA, USA

Article Info:

Keywords:
COVID-19,
Kermack-McKendrick epidemic model,
basic reproduction number,
data fitting,
data fixing.

Timeline:
Received: June 15, 2024
Accepted: July 14, 2024
Published: August 02, 2024

Citation: Sarkar K, Wang X-S. Data Fixing by Data Fitting: Estimating the Unreported Cases During the Early COVID-19 Outbreak in Hubei, China. J Basic Appl Sci 2023; 19: 92-97.

DOI: https://doi.org/10.29169/1927-5129.2024.20.09

Abstract:

On February 13, 2020, the Health Commission of Hubei Province changed the definition of confirmed cases, resulting in a reported daily case number that is significantly larger than on other dates. Such abnormal data points pose a challenge in data fitting and parameter estimation. To address this, we derive a simple formula from the classical Kermack-McKendrick model and introduce a new quantity to capture the number of unreported cases hidden in the data. We then use this new formula to fit the inconsistent data and estimate key epidemic parameters. Based on the reported cumulative case numbers until February 21, 2020, we estimate that the unreported case number in Hubei is 60856 (95% CI: [33513, 91206]), while the unreported case number in Wuhan is estimated as 29374 (95% CI: [18205, 40665]). The peak times in Hubei and Wuhan are February 6, 2020, and February 8, 2020, respectively. The basic reproduction numbers are 2.334 (95% CI: [2.053, 2.711]) for Hubei and 2.189 (95%CI: [1.992, 2.448]) for Wuhan.

*Corresponding Author
E-mail: xswang@louisiana.edu

INTRODUCTION

Even though we are living in the era of big data, we often need to make inferences with a limited amount of data points. For instance, during the early outbreak of a new epidemic wave, we usually have fewer than 100 reported cumulative case numbers. However, it is crucial to estimate key epidemic parameters such as the basic reproduction number, the final outbreak size, the peak time, and the disease transmission rate. While a complicated mathematical model with many parameters can easily fit the data points, it is safer to reduce the number of fitted parameters and keep the model as simple as possible to avoid over fitting. Ideally, there should be no more than five fitted parameters if the number of data points is less than 100. Another issue that may arise during early data collection is the inconsistency of data points due to sudden changes in the collection method. This poses a challenge in correcting data collected using incorrect methods. The main purpose of this paper is to introduce a novel approach to fixing data through data fitting, using the cumulative case numbers from the early COVID-19 outbreak in Hubei, China, as an illustrative example.

The first confirmed case of the COVID-19 outbreak was reported in Wuhan City, Hubei province, China, on December 31, 2019 [1]. As of February 21, 2020, the reported cumulative case number in China was 75891 [2], most of which were reported in Hubei province. As the capital city of Hubei, Wuhan reported a total of 45346 cases [2, 3]. On February 13, 2020, the Health Commission of Hubei Province changed the definition of confirmed cases and reported a total of 13797 daily cases [3]. This number is significantly larger than the reported daily cases on other dates. Such abnormal data points pose a challenge in data fitting and parameter estimation. To overcome this difficulty, we propose a new method to address the inconsistency problem in the epidemic data. Additionally, we will estimate the total number of unreported cases hidden within the reported data. Our method differs from the calibration method in [4, 5], where an appropriately fitted exponential function is used to replace the reported data points, or the method in [6], where the sudden increments are proportionally distributed into earlier reported data. It also differs from those methods in [7-10].

Data

We use the reported cumulative case number from the websites of the Health Commission of Hubei Province [3] and the National Health Commission of the Peoples Republic of China [2]. Whenever there is a correction on a reported date, we update the corresponding value in the previous reported date. After correction, the reported cumulative case numbers of Hubei province from January 21, 2020 to February 21, 2020 are {270, 375, 444, 549, 729, 1052, 1423, 2714, 3554, 4586, 5806, 7153, 9074, 11177, 13522, 16678, 19665, 22112, 24953, 27013, 29631, 31728, 33366, 47163, 51986, 54406, 56249, 58182, 59989, 61682, 62457, 63088}. It is noted that the number jumps abruptly from 33366 to 47163 on February 13, 2020. Also, the Wuhan cumulative case numbers [3, 11] from January 17, 2020 to February 21, 2020 are {45, 62, 121, 198, 258, 363, 425, 495, 572, 618, 698, 1590, 1905, 2261, 2639, 3215, 4109, 5142, 6384, 8351, 10117, 11618, 13603, 14981, 16902, 18454, 19558, 32081, 35991, 37914, 39462, 41152, 42752, 44412, 45027, 45346}.

METHODS

We adopt the standard epidemic model proposed by Kermack and McKendrick [12]:

$$\begin{cases} S'(t) = -\beta S(t)I(t)/[S(t)+I(t)], \\ I'(t) = \beta S(t)I(t)/[S(t)+I(t)] - \gamma I(t), \\ R'(t) = \gamma I(t), \end{cases} \quad (1)$$

where β is the disease transmission rate, and γ is the removal rate of infected individuals; see also [13]. To simplify our analysis, we make the following assumptions:

(A1) Among all the cumulative infected cases at the end of the outbreak, S(t) accounts for the number of individuals who have not been infected at time t.

(A2) The infected individuals, denoted by I(t), are infectious even during the incubation period with no symptoms.

(A3) The removed individuals R(t) include all infected individuals who are no longer infectious due to recovery, death, or quarantine.

The assumption (A2) is reasonable because asymptomatic infections have been reported in the clinical study [14]. The solution to (1) has a close form (see Appendix):

$$\begin{cases} S(t) = N[1+(R_0-1)e^{(\beta-\gamma)(t-t_p)}]^{-R_0/(R_0-1)}, \\ I(t) = N\,(R_0-1)e^{(\beta-\gamma)(t-t_p)}[1+(R_0-1)e^{(\beta-\gamma)(t-t_p)}]^{-R_0/(R_0-1)}, \\ R(t) = N - N[1+(R_0-1)e^{(\beta-\gamma)(t-t_p)}]^{-1/(R_0-1)}, \end{cases} (2)$$

where N is the final outbreak size, t_p is the peak time when I(t) reaches its maximum, and $R_0 = \beta/\gamma$ is the basic reproduction number which is defined as the average number of secondary infections generated by a primary infection that is introduced into an entirely susceptible population. Due to the limitation of resources, asymptomatic infection, and some other reasons, it is impossible to count the number of infected individuals during an outbreak. Rather, it is reasonable to make the following assumption.

(A4) The daily reported case number D(t) is a proportion of removed infected individuals.

If the method of counting case number is consistent during an outbreak, then we can assume that the proportion in (A4) is a constant p ∈ (0, 1); namely, D(t) = pγI(t). Consequently, the cumulative case number becomes C(t) = pR(t). By data fitting, we can estimate the epidemic parameters such as the basic reproduction number R_0, the peak time t_p, the disease transmission rate β, and the final reported cumulative case number K = pN. Note that the final outbreak size N and the proportion constant p are still undetermined because they are correlated in the expression of C(t).

To address the change of counting method by the Health Commission of Hubei Province on February 13, 2020. We need to make the following assumption.

(A5) There exist a change-point t_c and two proportion constants p_1, p_2 ∈ (0, 1) such that, the daily reported case number D(t) = p_1γI(t) for t < t_c and D(t) = p_2γI(t) for t > t_c.

The proportion constants p_1 and p_2 characterize two different counting methods adopted by the Health Commission of Hubei Province before and after the change-point t_c. It is clear that the cumulative case number C(t) = p_1R(t) for t < t_c. For t > t_c, a new method is adopted and the cumulative case number is C(t) = p_2R(t) − H, where p_2R(t) is the cumulative case number if the second method were used at the very beginning, and H accounts for the hiding case number before the application of new method. A big jump in the reported daily case number on February 13, 2020 indicates that the Health Commission of Hubei Province tried to fill the gap between two counting methods. However, the further t_c is away from the initial outbreak time, the more difficult it is to catch the difference. So, we introduce a new parameter H to capture the hiding information from the data. To sum up, we will fit the reported cumulative case numbers by the following piecewise defined function

$$C(t) = \begin{cases} K_1 - K_1[1+(R_0-1)e^{(\beta-\gamma)(t-t_p)}]^{-1/(R_0-1)}, t < t_c, \\ K_2 - K_2[1+(R_0-1)e^{(\beta-\gamma)(t-t_p)}]^{-1/(R_0-1)} - H, t > t_c \end{cases}, (3)$$

where K_1=p_1N, K_2=p_2N, and H is the hiding case number from the reported data. Note that the final reported cumulative case number is C(∞) =K_2 − H.

To find the confidence intervals of estimated parameters, we make the following assumption on the data of reported daily cases:

(A6) The reported daily case number has an accuracy of 100(1 − ε)%; namely, for each reported data point {(t_k, D_k)}, the daily infected case number D(t_k) follows a uniform distribution in the interval [(1 − ε)D_k, (1 + ε)D_k].

Assume (A6) with a given ε = 0.1 (i.e., each data point is at least 90% accurate), we can sample N = 1,000 datasets from the given reported dataset. For each sampled dataset, we estimate parameter values by data fitting. Now we have N estimated values for each parameter. By ordering these N values, we then define the 100(1 − 2α)% confidence interval of the corresponding parameter as the interval bounded by the 100α and 100(1 − α) percentiles of the ordered values.

RESULTS

The estimated parameter values with different report dates for the Hubei cumulative case numbers [2, 3] and Wuhan cumulative case numbers [3, 11] are listed in Tables **1** and **2**. The fitted curves are also illustrated in Figure **1**.

Based on the reported Hubei cumulative case numbers from January 21, 2020 to February 21, 2020 [2, 3], we estimate that the basic reproduction number is R_0 = 2.334 (95% CI: [2.053, 2.711]), the peak time is 02/06/2020 (95/% CI: [02/06/2020, 02/06/2020]), the transmission rate is β =0.539 (95% CI: [0.497, 0.582]), the reported final cumulative case number is C(∞) = 66329 (95% CI: [65115, 67974]), and the hiding (unreported) case number is H = 60856 (95% CI: [33513, 91206]).

Table 1: The parameter estimation for Hubei cumulative data with different last report dates

Report date	K_1	K_2	H	C(∞)	R_0	β	γ	Peak time t_p
02/15/2020	41236	136679	67791	68888	2.394	0.528	0.22	02/06/2020
02/16/2020	38786	142666	79145	63521	1.997	0.594	0.297	02/06/2020
02/17/2020	38009	155787	93240	62547	1.869	0.629	0.337	02/06/2020
02/18/2020	39017	141742	77734	64008	2.036	0.585	0.287	02/06/2020
02/19/2020	40978	123531	57045	66486	2.356	0.533	0.226	02/06/2020
02/20/2020	41076	122818	56217	66601	2.372	0.531	0.224	02/06/2020
02/21/2020	40811	124939	58630	66309	2.328	0.536	0.23	02/06/2020

Table 2: The parameter estimation for Wuhan cumulative data with different last report dates

Report date	K_1	K_2	H	C(∞)	R_0	β	γ	Peak time t_p
02/15/2020	21652	140829	100268	40561	1.254	1.303	1.039	02/08/2020
02/16/2020	21603	145774	105119	40655	1.253	1.31	1.045	02/08/2020
02/17/2020	22813	107525	64673	42852	1.503	0.813	0.541	02/08/2020
02/18/2020	24464	86501	41087	45414	1.83	0.617	0.337	02/08/2020
02/19/2020	26842	74109	25431	48678	2.288	0.513	0.224	02/08/2020
02/20/2020	27043	73453	24519	48934	2.326	0.507	0.218	02/08/2020
02/21/2020	26432	75872	27684	48188	2.207	0.525	0.238	02/08/2020

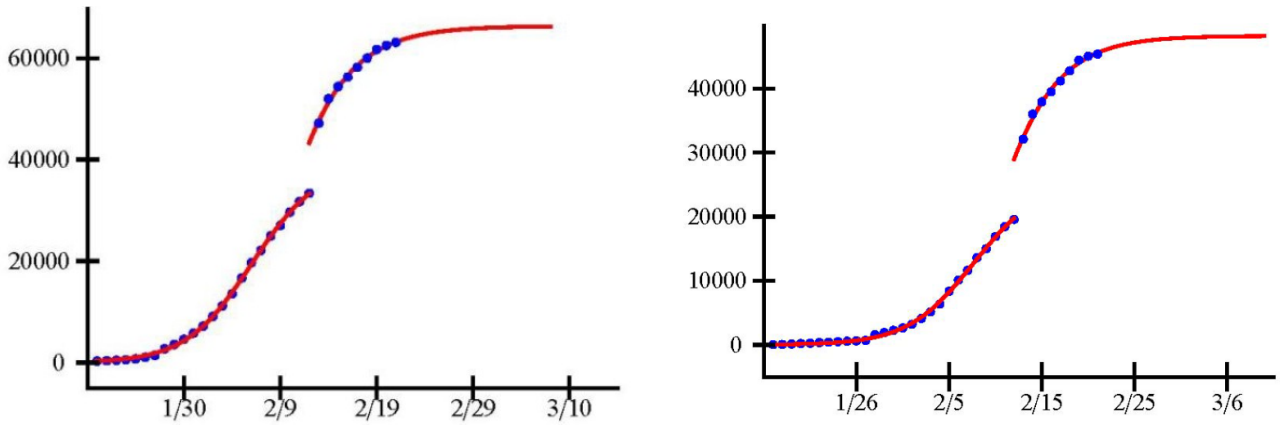

Figure 1: Reported data (dots) and fitted model (curve) for the cumulative cases of COVID-19outbreak in Hubei province (left panel) and Wuhan city (right panel), respectively.

Similarly, we can fit the reported Wuhan cumulative case numbers from January 17, 2020 to February 21, 2020 [3, 11]. The basic reproduction number is R_0 = 2.189 (95% CI: [1.992, 2.448]), the peak time is 02/08/2020 (95/% CI: [02/07/2020, 02/08/2020]), the transmission rate is β =0.530 (95% CI: [0.494, 0.567]), the reported final cumulative case number is C(∞) = 48110 (95% CI: [47345, 49090]), and the hiding (unreported) case number is H = 29374 (95% CI: [18205, 40665]).

DISCUSSION

To summarize, we have proposed a simple and biologically relevant formula (3) to estimate key epidemic parameters using reported cumulative cases with inconsistencies in data collection. This formula not only fits the inconsistent data well but also corrects it by capturing the unreported case numbers hidden within the data through data fitting. Interestingly, our estimation of the basic reproduction number aligns with the results in [15], despite using a different type of data

and method. Based on the reported cumulative case numbers up to February 21, 2020, we estimated that the epidemic peak occurred at the beginning of February 2020, and we predicted that the first epidemic wave would approach its end in March 2020.

ACKNOWLEDGMENT

X.-S. Wang is partially supported by the Louisiana Board of Regents Support Fund under contract No. LEQSF (2022-25)-RD-A-26.

APPENDIX

In this Appendix, we will provide a detailed derivation of the closed form (2) from the differential equations (1). First, we obtain from the first two equations in (1) that

$$\frac{d(S+I)}{dS} = \frac{\gamma(S+I)}{\beta S}.(A1)$$

Solving this equation gives

$$S + I = cS^{\gamma/\beta},(A2)$$

where c is a constant of integration. Substituting the above formula into the first equation in (1) yields

$$S'(t) = -\beta S(t)\{1-[S(t)/N]^{1-\gamma/\beta},(A3)$$

where

$$N = c^{\beta/(\beta-\gamma)}.(A4)$$

The solution to (A3) is given by

$$S(t) = N[1+ae^{(\beta-\gamma)(t-t_p)}]^{-\beta/(\beta-\gamma)},(A5)$$

where a is a constant to be determined, and t_p is the peak time when I(t) reaches its maximum; namely, $I'(t_p)=0$. On account of the second equation in (1) and the formulas (A2), (A4), and (A5), we have

$$\frac{\beta}{\gamma} = \frac{S(t_p)+I(t_p)}{S(t_p)} = cS(t_p)^{\gamma/\beta-1} = [\frac{S(t_p)}{N}]^{\gamma/\beta-1} = 1+a.(A6)$$

Note that $R_0 = \beta/\gamma$ denotes the basic reproduction number. Moreover, the final outbreak size is

$$\int_{-\infty}^{\infty} \frac{\beta S(t)I(t)}{S(t)+I(t)}dt = S(-\infty) - S(\infty) = N.(A7)$$

Thus, (A5) is the same as the first formula in (2). A combination of (A2), (A4), and (A5) gives the second formula in (2). Finally, adding the three equations in (1) implies that the sum S+I+R is independent of the time t. Hence, we have

$$S(t)+I(t)+R(t) = S(-\infty)+I(-\infty)+R(-\infty) = N,(A8)$$

which, together with the first two formulas in (2), yields the third formula in (2). This completes the proof of (2).

REFERENCES

[1] World Health Organization. Pneumonia of unknown cause - China. https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/, 2020. [Online; accessed 1-February-2020].

[2] National Health Commission of the People's Republic of China. News. http://www.nhc.gov.cn/yjb/s7860/new_list.shtml, 2020. [Online; accessed 21-February-2020].

[3] Health Commission of Hubei Province. News. http://wjw.hubei.gov.cn/fbjd/dtyw/, 2020.[Online; accessed 21-February-2020].

[4] Sun GQ, Wang SF, Li MT, Li L, Zhang J, Zhang W, Jin Z, Feng GL. Trans mission dynamics of COVID-19 in Wuhan, China: effects of lockdown and medical resources. Nonlinear Dyn 2020; 101; 1981–1993. https://doi.org/10.1007/s11071-020-05770-9

[5] Wang L, Zhou Y, He J, Zhu B, Wang F, Tang L, Eisenberg M, Song PXK. Anepidemio logical forecast model and software assessing interventions on COVID-19 epidemicin China. J. Data Sci 2020; 18: 409–432. https://doi.org/10.6339/JDS.202007_18(3).0003

[6] Zhou L, Rong X, Fan M, Yang L, Chu H, Xue L, Hu G, Liu S, Zeng Z, Chen M, Sun W, Liu J, Liu Y, Wang S, Zhu H. Modeling and evaluation of the joint prevention and control mechanism for curbing COVID-19 in Wuhan. Bull. Math. Biol 2022; 84: 28. https://doi.org/10.1007/s11538-021-00983-4

[7] Huo X, Chen J, Ruan S. Estimating asymptomatic, undetected and total cases for theCOVID-19 outbreak in Wuhan: a mathematical modeling study. BMC Infect. Dis 2021; 21: 476. https://doi.org/10.1186/s12879-021-06078-8

[8] Li J, Yuan P, Heffernan J, Zheng T, Ogden N, Sander B, Li J, Li Q, B´elair J, Kong JD, Aruffo E, Tan Y, Jin Z, Yu Y, Fan M, Cui J, Teng Z, Zhu H. Fangcang shelter hospitals during the COVID-19 epidemic, Wuhan, China. Bull. World Health Organ 2020; 98: 830–84. https://doi.org/10.2471/BLT.20.258152

[9] Wang L, Wang J, Zhao H, Shi Y, Wang K, Wu P, Shi L. Modelling and assessing the effects of medical resources on transmission of novel coronavirus (COVID-19) in Wuhan, China. Math. Biosci. Eng 2020; 17: 2936–2949. https://doi.org/10.3934/mbe.2020165

[10] Zhang XS, Vynnycky E, Charlett A, Angelis DD, Chen Z, Liu W. Transmission dynamics and control measures of COVID-19 outbreak in China: a modelling study. Sci.Rep 2021; 11: 2652. https://doi.org/10.1038/s41598-021-81985-z

[11] Wuhan Municipal Health Commission. News. http://wjw.wuhan.gov.cn/front/web/list2nd/no/710, 2020. [Online; accessed 21-February-2020].

[12] Kermack WO, McKendrick AG. A contribution to the mathematical theory of epi demics. Proc. R. Soc. Lond. Ser. B 1927; 115: 700–721. https://doi.org/10.1098/rspa.1927.0118

[13] Wang XS, Wu J, Yang Y. Richards model revisited: validation by and application to infection dynamics. J. Theor. Biol 2012; 313: 12–19. https://doi.org/10.1016/j.jtbi.2012.07.024

[14] Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, Xing F, Liu J, Yip CCY, Poon RWS, Tsoi HW, Lo SKF, Chan KH, Poon VKM, Chan WM, Ip JD, Cai JP, Cheng VC, Chen H, Hui CKM, Yuen KY. A familialcluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet 2020; 395: 514–523. https://doi.org/10.1016/S0140-6736(20)30154-9

[15] Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z. Early transmission dynamicsin Wuhan, China, of novel corona virus-infected pneumonia. N. Engl. J. Med 2020; 382: 1199–1207. https://doi.org/10.1056/NEJMoa2001316