Predicting the spread of COVID-19 in China with human mobility data

Shangbin Wu
shangbin@stu.xmu.edu.cn
Xiamen University
Xiamen, China

Xiaoliang Fan*
fanxiaoliang@xmu.edu.cn
Xiamen University
Xiamen, China

Longbiao Chen
longbiaochen@xmu.edu.cn
Xiamen University
Xiamen, China

Ming Cheng
chm99@xmu.edu.cn
Xiamen University
Xiamen, China

Cheng Wang
cwang@xmu.edu.cn
Xiamen University
Xiamen, China

ABSTRACT

The coronavirus disease 2019 (COVID-19) break-out in late December 2019 has spread rapidly worldwide. Existing studies have shown that there is a significant correlation between large-scale human movements and the spread of the epidemic. However, there is a lack of quantification of these correlations, and it is still challenging to predict the spread of the epidemic at early stage. In this paper, we address this issue by conducting a statistical analysis on the spatio-temporal relationship between human mobility and the epidemic spread. Specifically, we proposed an improved SEIR model to adapt to the COVID-19 epidemic, so that we can predict the spread of the epidemic at the early stage using human mobility data and the early confirmed cases. We evaluated our model in various provinces and cities in China, and the results are superior to various baselines, verifying the effectiveness of the method.

CCS CONCEPTS

• **Human-centered computing** → *Collaborative and social computing*.

KEYWORDS

COVID-19, human mobility, SEIR, epidemic

ACM Reference Format:
Shangbin Wu, Xiaoliang Fan, Longbiao Chen, Ming Cheng, and Cheng Wang. 2018. Predicting the spread of COVID-19 in China with human mobility data. In 29th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '21), November 2–5, 2021, Beijing, China. 4 pages. https: /doi.org/10.1145/3474717.3483952

*Corresponding author

SIGSPATIAL '21, November 2–5, 2021, Beijing, China

ACM ISBN 978-1-4503-8664-7/21/11…$15.00
https://doi.org/10.1145/3474717.3483952

1 INTRODUCTION

In December 2019 an outbreak of atypical pneumonia [coronavirus disease 2019 (COVID-19)] has infected more than 70 million people and caused more than 1.5 million deaths by December 2020. Because the outbreak coincided with chunyun, the annual period of mass migration for the Spring Festival holidays that was to begin on January 25, 2020. The disease quickly spread to all provinces of China, and then affected 215 countries and regions in the world. In this situation, we need to use modeling methods and available human mobility data to predict disease outbreaks[9], then provide scientific guidance for human interventions.

Since the classic SEIR(Susceptible-Exposed-Infected-Removed) cannot fit the epidemic situation in China well, many studies have proposed modification to the SEIR model. However, existing studies generally used many excessive parameters to model human interventions, which might inevitably introduce noises and uncertainties to the prediction[4, 17, 18]. Therefore, we introduced only one new parameter to the SEIR model to express the degree of protection of the intervention measures to the susceptible population. We can easily study the correlation between this parameter and human mobility data. We found that the value of this parameter has a linear relationship with human mobility data, which allows us to solve the parameters through human mobility data without parameter fitting.

Our approach differs from prior work linking human mobility and disease spread in terms of: our use of travel intensity data, which represents the attributes of the region itself, not the interaction between regions[10, 19]; our focus on aggregate population flows rather than individual tracking[3, 17]; and our modified SEIR modeling approach.

2 DATA COLLECTION

The data used in this paper consists of two parts:

Human mobility data provided by Baidu[2], which includes:

(1) Inflow/outflow population intensity of each city/province, which reflects the number of inflow/outflow population of this city/Province;

(2) Urban travel intensity, which reflects the ratio of the population traveling in a city to the total population of the city;

(3) Inter regional migration scale index data. It reflects the ratio of the number of people flowing from area A to area B to the total number of people flowing out of area A.

Official confirmed case data(city level and province level) until the date of July 2, 2020, released by the China Health Commission[15]: It contains

- Daily new confirmed cases.
- Daily cured cases.
- Daily deaths.
- Daily new suspected cases.

3 HUMAN MOBILITY DATA ANALYSIS

In this section, we analyze human mobility data. The non-pharmaceutical interventions taken by the Chinese government due to the epidemic have a significant impact on human mobility, and human mobility in turn affects the spread of the epidemic.

3.1 Human Intervention Changed Human Mobility

We conduct visual processing and statistical analysis for human mobility data. Aims to discover what has changed in human behavior due to the human intervention In order to control the spread of the virus, the Chinese government has taken unprecedented intervention strategies. As a result, the intensity of population movement between provinces has dropped by 67% , and the urban travel intensity has dropped by 37%(Figure 1).

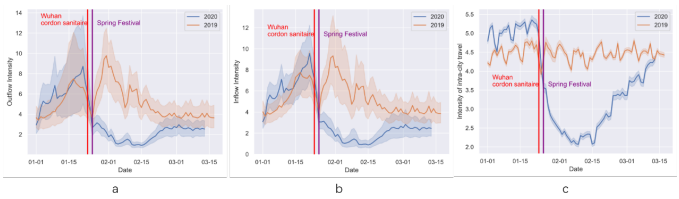

Figure 1: Changes in travel intensity due to the intervention for epidemic. (a. the intensity of provincial outflow; b. the intensity of provincial inflow; c. the intensity of intra city travel. Wuhan cordon sanitaire: January 23; Spring Festival: January 25, 2020. Please note that the date of 2019 has been adjusted relative to the Spring Festival of that year (February 3, 2019))

3.2 Human Mobility Drives the Epidemic

In this section, we analyze the correlation between different types of human mobility data and the epidemic. And we can find out which data can be used to predict the spread of the epidemic.

3.2.1 The Impact of Migrants From Hubei/Wuhan on the Epidemic in Other Regions. Since 31 December 2019, cases have been exported to other Chinese cities and provinces from Wuhan, the capital of Hubei province in China. We found that the Migration scale number from Wuhan is consistent with the number of confirmed cases. This conclusion is the same as that of Jia et al. (using mobile phone data)[10] and Kraemer et al. (using detailed case data including travel history)[11]. While Wuhan City, as the capital of Hubei Province(the data from Wuhan to Hubei were excluded), has a more significant impact on other regions than Hubei Province (Table 1).

Table 1: Correlation between migration scale index and number of confirmed cases

Migration from	Migration to	R-value	P-value	Standard Error	R^2
Wuhan	cities	0.970	3.55E-62	5.770	0.941
Wuhan	provinces	0.801	1.05E-07	35.788	0.642
Hubei	cities	0.683	4.80E-15	6.853	0.467
Hubei	provinces	0.801	1.06E-07	11.147	0.642

3.2.2 The Impact of the Region's Own Human Mobility on the Local Epidemic. We investigate the impact of the inflow intensity, outflow intensity and intensity of intra-city travel on the epidemic. We found that provinces with more active interactions with the outside tend to have more severe epidemics, but active cities do not (Figure 2). We think this is because the active provinces will have more interactions with Wuhan (where the epidemic broke out), but the active cities may only interact more actively with neighboring cities.

Another counter-intuitive phenomenon is the weak correlation between the intensity of intra-city travel and the number of confirmed cases (see Figure 2). The reasons for this are various. On the one hand, the cities' epidemic is mainly affected by the population of Wuhan it receives; On the other hand, in cities with severe epidemics, people tend to adopt stricter restrictions, and intra-city travel also affected by other factors such as education level and geographic factors[4].

3.2.3 Conclusion. By analyzing these human mobility data, we found that the non-pharmaceutical interventions in China have significant effects:

(1)There is a correlation between human mobility data and the spread of the epidemic, the human mobility data of city level has stronger correlation with epidemic than that of province level.

(2)In China, the main factor affecting the severity of the epidemic in a region is the number of people from Wuhan it receives before the cordon sanitaire.

(3)There is no clear correlation between the intensity of intra-city travel and the severity of the epidemic in the city.

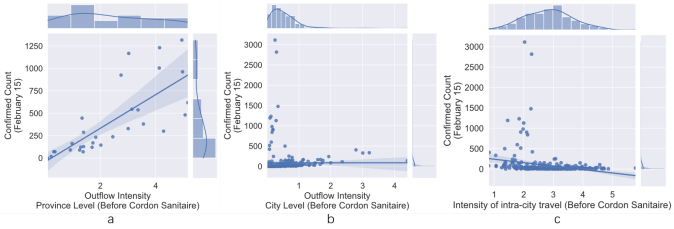

Figure 2: Joint distribution of travel intensity and the number of confirmed cases

4 A MODIFIED SEIR

We predict future trends based on early data on the epidemic. Firstly, we propose a modified SEIR model, and take the initial confirmed case data as the input to calculate the model parameters in different regions. Then we find the linear relationship between the model

parameters and the human mobility data. In the prediction, we use the human mobility data to calculate the value of the model parameters, and then use the model to predict the number of confirmed cases.

4.1 Mathematical Model

Here, we use the classic SEIR ordinary differential equations to model the dynamics of disease outbreaks in China, which will not add parameters to the equations. The transmission dynamics are governed by the following system of equations:

$$\begin{cases} S^{'} = -\beta SI/(S+E+I+R) \\ E^{'} = -\beta SI/(S+E+I+R) - \omega E \\ I^{'} = \omega E - \gamma I \\ R^{'} = \gamma I \end{cases} \tag{1}$$

Where "'" is the derivative with respect to time, and β denoted the coefficient of infection rate; $\omega = 1/T_e$, ω denoted the transition rate of exposed individuals to the infected class, T_e denoted the average latency; and $\gamma = 1/T_i$, γ denoted the removed rate of infected individuals, T_i denoted the mean duration from onset to hospital admission (In China, hospitalization means isolation, therefore, we think that hospitalized patients are no longer infectious[5]). We believe that when using SEIR to model the epidemic in an area, due to the effective human interventions, the initial susceptible population (S_0) is no longer equal to the total population of the area, but equal to effective-size-of-the-populations-at-risk (N_{eff}). That is:

$$S_0 = N_{eff} \tag{2}$$

We assume that N_{eff} is a certain percentage of the total population:

$$N_{eff} = qN \tag{3}$$

Where N denoted the population of a region; and q denoted the proportion of the initial susceptible population to the total population due to the effective non-pharmacological interventions.

4.2 Result

In this study, we set the T_e=5.2 days and T_i=12.5 days[12]. We try to predict the trend of the epidemic through the early information, so we did not adopt time-dependent parameters, which requires long-term data. We used the data from January 19th to January 26th to estimate the β value of 11.37 using the least square method .

We believe that because of the differences in human mobility in different regions, their effective-size-of-the-populations-at-risk proportions (q) are also different. We use the least square method to estimate the value of q in different cities and provinces, and we found that the value of q has a linear relationship with human mobility data(see Figure 3):

$$q \propto \frac{m}{N} \tag{4}$$

Where m denoted the human mobility data, and N, the total population of a region. We found that there are six kinds of human mobility data that satisfy this relationship (Table 2), which means that we can actually estimate a parameter value (q) in the SEIR model through human mobility data. In the Figure 3, we show two examples of the linear relationship, and in Figure 3.b. Pearson correlation coefficient has a downward trend in the long run (similar to previous study from other data[10]), but the correlation is still high. Take advantage of this correlation to estimate q, we do not need long-term data of laboratory-confirmed cases to fit the differential equation.

Table 2: Correlation between q and human mobility data

Human mobility data	R-value	P-value	Standard Error	R^2
Outflow intensity(Province)	0.577	0.001	2.922	0.333
Inflow intensity(Province)	0.405	0.029	2.379	0.164
Wuhan to cities	0.955	7.55E-43	0.919	0.912
Wuhan to provinces	0.734	5.97E-06	5.091	0.538
Hubei to cities	0.222	0.121	9.904	0.049
Hubei to provinces	0.586	0.001	17.073	0.343

Table 3: Prediction results with different human mobility data

method	R-value	P-value	Standard Error	R^2
fit q (province level)	0.760	2.17E-07	0.086	0.589
fit q (city level)	0.508	0.031	0.156	0.318
using outflow intensity (province level)	0.764	1.49E-07	0.104	0.595
using inflow intensity (province level)	**0.771**	**2.78E-08**	**0.079**	**0.604**
using migration from Wuhan to cities	0.515	0.030	0.155	0.325

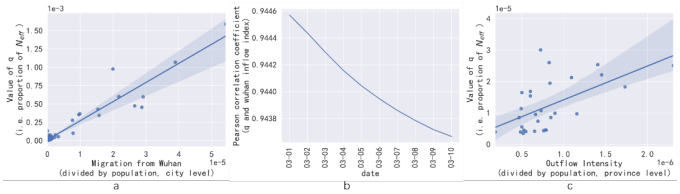

Figure 3: The linear relationship between q and human mobility data

Table 3 shows the statistical analysis on the prediction results. We used different kinds of human mobility data to calculate the q value, then predict the number of confirmed cases, and compare it with the q value fitted by the least square method(as baseline).

We found that the migration from Wuhan to cities/provinces showed the strongest correlation with the number of confirmed cases Table 1) and q (Table 2). However, when prediction, the provincial level inflow and outflow data showed a stronger correlation with the real number of confirmed cases (Table 3). Practically, the correlation distribution between the outflow/inflow intensity of each province and the number of confirmed cases is more uniform, so it shows better effect in forecasting.

5 RELATED WORK

SEIR model divides all people in a region into four categories: susceptible, exposed, infected and removed. While under the effective intervention measures taken by China, the primitive SEIR cannot adapt well to the epidemic in China. Therefore, improvements to the SEIR model[6, 19] can make the model more applicable for actual conditions and can also research the impact of human interventions.

For example, many studies have added asymptomatic infectious or quarantined populations to the model[1, 7, 14, 16]. However, all these extensions lead to the proliferation of free model parameters, and thus require orders of magnitude more data to fit the model than does the basic three parameter SEIR disease process formulation[8]. In the early stages of the epidemic, accurate data on infected persons may be difficult to obtain, and insufficient research on asymptomatic infected persons makes it difficult to estimate model parameters[13].

There are many existing studies using human mobility data to predict epidemic, but most of existing methods only focus on a specific area. For example, Boston metropolitan area[1], ten US metropolitan areas[3], Spain[14], Beijing[16], Wuhan[18], Hubei, Zhejiang and whole China[19]. And some of them require additional fine-grained human mobility data[1][3], international flight data[18], long term (40 days) confirmed cases data[12]. Our method is different from the above methods in twofold: First, we propose a general method and apply it to 367 cities and 31 provinces in China. Second, we only use the data in early stage (2 weeks) to carry out long-term predictions.

6 CONCLUSION AND FUTURE WORK

In this study, we investigated the relationship between different types of human mobility data and disease transmission. After that, we proposed a modified SEIR model, in which human mobility data can be used to solve the parameters of the differential equation, which is a different approach from previous studies. Our forecast results are superior to various baselines.

Contrary to the intuition, we found that there is less correlation between intra-city travel intensity and the number of confirmed cases. We also find that the outflow intensity and inflow intensity of a province are consistent, and they are strongly correlated with the number of confirmed cases, just like the migration from Wuhan/Hubei. This may mean that immigrants from Wuhan and Hubei have gone to various provinces in China with equal probability. The human dynamics behind this will be a promising research direction in the future.

ACKNOWLEDGMENTS

The research was supported by Natural Science Foundation of China (61872306), and Fundamental Research Funds for the Central Universities (20720200031).

REFERENCES

[1] Alberto Aleta, David Martín-Corral, Ana Pastore y Piontti, Marco Ajelli, Maria Litvinova, Matteo Chinazzi, Natalie E. Dean, M. Elizabeth Halloran, Ira M. Longini, Stefano Merler, Alex Pentland, Alessandro Vespignani, Esteban Moro, and Yamir Moreno. 2020. Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nature Human Behaviour* 4, 9 (2020), 964–971. https://doi.org/10.1038/s41562-020-0931-9

[2] Baidu. 2020. *Baidu migration data.* Retrieved March 05, 2020 from http://qianxi.baidu.com/

[3] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. 2020. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* (2020). https://doi.org/10.1038/s41586-020-2923-3

[4] Ben Charoenwong, Alan Kwan, and Vesa Pursiainen. 2020. Social connections with COVID-19–affected areas increase compliance with mobility restrictions. *Science Advances* 6, 47 (nov 2020), eabc3054. https://doi.org/10.1126/sciadv.abc3054

[5] Tian Mu Chen, Jia Rui, Qiu Peng Wang, Ze Yu Zhao, Jing An Cui, and Ling Yin. 2020. A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infectious Diseases of Poverty* 9, 1 (2020), 1–8. https://doi.org/10.1186/s40249-020-00640-3

[6] Yaqing Fang, Yiting Nie, and Marshare Penny. 2020. Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis. *Journal of Medical Virology* 92, 6 (2020), 645–659. https://doi.org/10.1002/jmv.25750

[7] Marino Gatto, Enrico Bertuzzo, Lorenzo Mari, Stefano Miccoli, Luca Carraro, Renato Casagrandi, and Andrea Rinaldo. 2020. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences of the United States of America* 117, 19 (2020), 10484–10491. https://doi.org/10.1073/pnas.2004978117

[8] Wayne M. Getz, Richard Salter, and Whitney Mgbara. 2019. Adequacy of SEIR models when epidemics have spatial structure: Ebola in Sierra Leone. *Philosophical Transactions of the Royal Society B: Biological Sciences* 374, 1775 (2019), 1–7. https://doi.org/10.1098/rstb.2018.0282

[9] K. H. Grantz, H. R. Meredith, Dat Cummings, Cje Metcalf, and A. Wesolowski. 2020. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications* 11, 1 (2020), 4961.

[10] Jayson S Jia, Xin Lu, Yun Yuan, Ge Xu, Jianmin Jia, and Nicholas A Christakis. 2020. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* 582, 7812 (2020), 389–394. https://doi.org/10.1038/s41586-020-2284-y

[11] Moritz U G Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, John S Brownstein, Maylis Layan, Alessandro Vespignani, Huaiyu Tian, Christopher Dye, Oliver G Pybus, and Samuel V Scarpino. 2020. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* 368, 6490 (may 2020), 493 LP – 497. https://doi.org/10.1126/science.abb4218

[12] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy S.M. Leung, Eric H.Y. Lau, Jessica Y. Wong, Xuesen Xing, Nijuan Xiang, Yang Wu, Chao Li, Qi Chen, Dan Li, Tian Liu, Jing Zhao, Man Liu, Wenxiao Tu, Chuding Chen, Lianmei Jin, Rui Yang, Qi Wang, Suhua Zhou, Rui Wang, Hui Liu, Yinbo Luo, Yuan Liu, Ge Shao, Huan Li, Zhongfa Tao, Yang Yang, Zhiqiang Deng, Boxi Liu, Zhitao Ma, Yanping Zhang, Guoqing Shi, Tommy T.Y. Lam, Joseph T. Wu, George F. Gao, Benjamin J. Cowling, Bo Yang, Gabriel M. Leung, and Zijian Feng. 2020. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *New England Journal of Medicine* 382, 13 (2020), 1199–1207. https://doi.org/10.1056/nejmoa2001316

[13] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. 2020. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* 368, 6490 (2020), 489–493. https://doi.org/10.1126/science.abb3221

[14] Leonardo López and Xavier Rodó. 2020. The end of social confinement and COVID-19 re-emergence risk. *Nature Human Behaviour* 4, 7 (2020), 746–755. https://doi.org/10.1038/s41562-020-0908-8

[15] National Health Commission of China. 2020. *Situation report for COVID-19 (in Chinese).* Retrieved July 06, 2020 from http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml

[16] Biao Tang, Xia Wang, Qian Li, Nicola Luigi Bragazzi, Sanyi Tang, Yanni Xiao, and Jianhong Wu. 2020. Estimation of the Transmission Risk of the 2019-nCoV and Its Implication for Public Health Interventions. *Journal of Clinical Medicine* 9, 2 (2020), 462. https://doi.org/10.3390/jcm9020462

[17] Michele Tizzoni, Paolo Bajardi, Adeline Decuyper, Guillaume Kon Kam King, Christian M. Schneider, Vincent Blondel, Zbigniew Smoreda, Marta C. González, and Vittoria Colizza. 2014. On the Use of Human Mobility Proxies for Modeling Epidemics. *PLoS Computational Biology* 10, 7 (2014). https://doi.org/10.1371/journal.pcbi.1003716

[18] Joseph T. Wu, Kathy Leung, and Gabriel M. Leung. 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* 395, 10225 (2020), 689–697. https://doi.org/10.1016/S0140-6736(20)30260-9

[19] Zifeng Yang, Zhiqi Zeng, Ke Wang, Sook San Wong, Wenhua Liang, Mark Zanin, Peng Liu, Xudong Cao, Zhongqiang Gao, Zhitong Mai, Jingyi Liang, Xiaoqing Liu, Shiyue Li, Yimin Li, Feng Ye, Weijie Guan, Yifan Yang, Fei Li, Shengmei Luo, Yuqi Xie, Bin Liu, Zhoulang Wang, Shaobo Zhang, Yaonan Wang, Nanshan Zhong, and Jianxing He. 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease* 12, 3 (2020), 165–174. https://doi.org/10.21037/jtd.2020.02.64