A Short-Term Prediction Model at the Early Stage of the COVID-19 Pandemic Based on Multisource Urban Data

Ruxin Wang, Chaojie Ji, Zhiming Jiang, Yongsheng Wu, Ling Yin, and Ye Li, *Senior Member, IEEE*

***Abstract*— The ongoing coronavirus disease 2019 (COVID-19) pandemic spread throughout China and worldwide since it was reported in Wuhan city, China in December 2019. 4 589 526 confirmed cases have been caused by the pandemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), by May 18, 2020. At the early stage of the pandemic, the large-scale mobility of humans accelerated the spread of the pandemic. Rapidly and accurately tracking the population inflow from Wuhan and other cities in Hubei province is especially critical to assess the potential for sustained pandemic transmission in new areas. In this study, we first analyze the impact of related multisource urban data (such as local temperature, relative humidity, air quality, and inflow rate from Hubei province) on daily new confirmed cases at the early stage of the local pandemic transmission. The results show that the early trend of COVID-19 can be explained well by human mobility from Hubei province around the Chinese Lunar New Year. Different from the commonly-used pandemic models based on transmission dynamics, we propose a simple but effective short-term prediction model for COVID-19 cases, considering the human mobility from Hubei province to the target cities. The performance of our proposed model is validated by several major cities in Guangdong province. For cities like Shenzhen and Guangzhou with frequent population flow per day, the values of R^2 of daily prediction achieve 0.988 and 0.985. The proposed model has provided a reference for decision support of pandemic prevention and control in Shenzhen.**

***Index Terms*— Coronavirus disease 2019 (COVID-19), human mobility, multisource urban data, short-term prediction.**

I. Introduction

THE coronavirus disease 2019 (COVID-19), a global pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has caused tremendous social burden and economic losses for many regions and countries in the world. According to the latest statistics of the World Health Organization (up to May 18, 2020), 4 589 526 cases have been infected in over 216 countries, areas, or territories, resulting in approximately 310 391 deaths [1] (Fig. 1). And the number of confirmed cases is still escalating rapidly worldwide, which has led to an unprecedented challenge in front of the international community [2], [3].

The outbreak of the COVID-19 began in early December 2019, which was first reported in Wuhan city, Hubei province of China [4], [5]. The COVID-19 can cause fever, cough, and other symptoms, even affect acute respiratory failure and multiple-organ dysfunction [6], [7]. Of particular note is the fact that human-to-human transmission has been confirmed [8], [9]. The geographic spread of emergent infectious diseases by large-scale human mobility can cause rapid multipoint transmission of the pandemic [10]–[12]. At the initial stage of the COVID-19 outbreak, it is an effective approach that the pandemic situation based on multisource urban data is analyzed. Especially, tracking human mobility from Wuhan and other cities in Hubei province is critical due to the massive transportation in the run-up to the Chinese Spring Festival. Among the infected population, asymptomatic and mild-symptomatic cases are difficult to be detected. Considering the natural transmission efficiency and mortality of the COVID-19 pandemic, the potential influence caused by human mobility cannot be neglected [13]. For the first-tier megacities like Shenzhen city and Guangzhou city, the early prevention and control of importation risk are especially important. For instance, Shenzhen is the largest mobility city with the highest population density in China. There are large inflows of the population from domestic areas, which possibly brings high importation risk. Specifically, there are nearly 1.5 million people who immigrate from Hubei province to Shenzhen according to recent statistics. From the aspect of urban prevention and control of the pandemic, accurate and timely prediction plays an important role [14]–[16].

Recently, many works have focused on forecast and simulation for the pandemic situation from different perspectives [17]–[20]. The most commonly-used prediction approaches are the classical compartmental models based on pandemic transmission dynamics. There have been extensive studies predicting the pandemic trend of COVID-19 [21]–[23]. For instance, Wu *et al.* [8] adopted a susceptible-exposed-infectious-recovered (SEIR) metapopulation model to simulate the pandemic across all major cities in mainland China.

Manuscript received May 30, 2020; revised December 31, 2020; accepted February 10, 2021. Date of publication March 5, 2021; date of current version August 2, 2021. This work was supported in part by the Strategic Priority CAS Project under Grant XDB38040200, in part by the National Natural Science Foundation of China under Grant 41771441, in part by the Bill & Melinda Gates Foundation under Grant INV-005834, in part by the Shenzhen Basic Research Projects under Grant JCYJ20180703145202065, and in part by the Shenzhen Science and Technology Innovation Project under Grant JSGG20170823144843046. *(Ruxin Wang and Chaojie Ji contributed equally to this work.) (Corresponding authors: Ye Li; Ling Yin; Yongsheng Wu.)*

Ruxin Wang, Chaojie Ji, and Ye Li are with the Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: rx.wang@siat.ac.cn; cj.ji@siat.ac.cn; ye.li@siat.ac.cn).

Zhiming Jiang and Ling Yin are with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: zm.jiang@siat.ac.cn; yinling@siat.ac.cn).

Yongsheng Wu is with the Shenzhen Center for Disease Control and Prevention, Shenzhen 518055, China (e-mail: cdc@szcdc.net).

Digital Object Identifier 10.1109/TCSS.2021.3060952

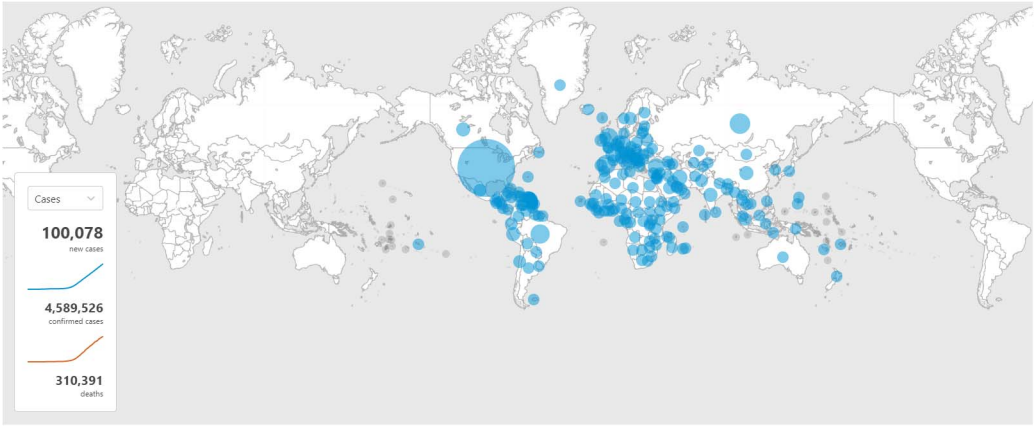


Fig. 1. COVID-19 outbreak situation [1].

Yang *et al.* [21] proposed a modified SEIR model in which population migration data are integrated, and the results show a gradual decline of the pandemic in China by end of April. Kucharski *et al.* [23] employed a stochastic transmission dynamic model to capture the varying dynamics of the pandemic. However, at the early stage of an emerging pandemic, limited knowledge about the transmission characteristics can be obtained, such as the basic regeneration number, transmission capability of asymptomatic cases, and so on. Moreover, the reality could be complicated and change over time and space, which results in that the accurate estimation of these parameters by modeling transmission dynamics becomes challenging. Besides them, some studies have explored the relationship between human mobility and transmission of COVID-19 with respect to statistics and machine learning. To be concrete, Jiang and Luo [19] used the random effect model to evaluate the impact of population mobility on COVID-19 transmission. Kraemer *et al.* [20] utilized the generalized linear model, aiming to estimate the influence of human mobility and control measures for the pandemic in China. The autoregressive time series model was adopted by Maleki *et al.* [24] to analyze the confirmed and recovered COVID-19 cases across the world from April 21, 2020 up to April 30, 2020.

The analysis method based on multisource urban data can be used to analyze and model the transmission of the pandemic due to the limited prior knowledge of the emerging COVID-19 pandemic. Meanwhile, to provide rational suggestions for the importation risk, especially at the early stage of the COVID-19 pandemic, this study proposes a short-term prediction approach for daily new COVID-19 cases based on human mobility data of Hubei province. Our study areas focus on major cities in Guangdong province, especially Shenzhen city. Firstly, considering multisource urban data, we conduct some statistics and correlation analysis between possible factors – human mobility, weather condition, and so on – and the newly confirmed cases. According to the results, it can be observed that the early trend of the COVID-19 pandemic in Shenzhen can be explained well by the population migrated from Hubei province during the Chinese Lunar New Year. Then, a fixed effect multiple regression (MR) model is constructed, serving the short-term prediction of daily new confirmed cases of COVID-19. We finally verify the performance of our method through several major cities of Guangdong province, from February 1st to February 15, 2020. Specifically, the value of R^2 of daily new confirmed cases reaches 0.988 and 0.985 for Shenzhen and Guangzhou, respectively. Therefore, the proposed model can be applied to assist the disease control department to accurately get insight into the pandemic situation, reasonably allocate the emergency resources, precisely arrange the medical staff at the early stage of the pandemic. In summary, the contributions of this work are twofold as follows.

1) In terms of Shenzhen city, the correlation between related multisource urban data and daily new confirmed cases, at the early stage of the local pandemic transmission, is analyzed. The early tendency of COVID-19 in Shenzhen can be explained well by the migrated population from Hubei province during the period of the Chinese Lunar New Year.

2) We put forward a fixed effect MR model so that short-term prediction of daily new confirmed cases for a city becomes available. Especially, the proposed model has been adopted, aiming to provide a reference for the prevention and control of pandemic in Shenzhen at the early stage of the COVID-19 pandemic.

The rest of this article is organized as follows. Section II describes the proposed short-term pandemic prediction model for the early stage of the COVID-19 outbreak based on mobility flow. Section III presents analysis and prediction results for several major cities in Guangdong province. Section IV discusses the contribution and limitations of this proposed model. Finally, this study is concluded in Section V.

II. Methodology

A. Study Design

The occurrence, spread, and spatial distribution of many infectious diseases are closely related to geographical factors and human activities, such as climate change, weather condition, human mobility, and so on [25]–[27]. In our research, we take the epidemic situation of Shenzhen in early February as the main analysis object. In the early phase of the COVID-19 pandemic, owing to the lack of prior knowledge of the emerging disease, we first collect multisource urban data, such as temperature, humidity, air quality and human mobility

in the study area, and the newly confirmed cases. Then the correlation between them is investigated. It is expected that the most relevant factors at the early stage of the local pandemic transmission are screened out, which also makes a contribution to precise modeling. Finally, we merge the established critical factors and the historical confirmed cases to execute short-term predictions. In this way, the need for early and rapid warning is fulfilled.

B. Data Sources

To assess the impact of various factors on early dynamics of transmission in Shenzhen, we collect multiple publicly available urban data in Shenzhen, which consists of confirmed cases, mobility, air quality, and weather data. In particular, the daily data of COVID-19 confirmed cases in Guangdong province are obtained from the Health Commission of Guangdong Province from January 21, 2020 to February 15, 2020. It is composed of the numbers of the accumulated confirmed cases and daily ones of cities in Guangdong province.[1] The data of daily human mobility moving from Hubei to Guangdong province are extracted from Baidu Inc.[2] This information represents the daily mobility trend in form of mobility ratio. The daily air quality data containing air quality index (AQI) and fine particulate matter (PM2.5 and PM10) are derived from the China air quality online monitoring and analysis platform.[3] We acquire the temperature and relative humidity data from the open meteorological information website.[4] Specifically, human mobility, air quality and weather data are collected from January 1st to February 15, 2020.

C. Correlation Analysis

With respect to Shenzhen city, employing correlation analysis, we analyze the correlation between local temperature, relative humidity, air quality (AQI/PM2.5/PM10) and inflow rate from Hubei province, and daily new confirmed cases. Due to the high intraprovince mobilization of Hubei province, the pandemic had spread to every city in Hubei province from Wuhan, during the Spring Festival. Therefore, we consider the daily inflow from the whole Hubei province to the target cities as input rather than separate cities in Hubei province. In Shenzhen city, the first confirmed case was reported on January 19th. Considering the incubation period of the COVID-19, it can spread to others during this period. According to latest literatures [28]–[31], the mean incubation period is defined as around 5–7 days. As a result, we define the incubation period as 6 days before the onset of the disease. At the same time, regarding that some factors may have a lag effect on the transmission, we expand the incubation period to 14 days and analyze them in Section III-B.

We investigate the relevance between the number of newly confirmed patients from January 21th to 30th in Shenzhen and the corresponding human mobility and other weather/climate factors, in terms of different incubation periods. In this article, the Pearson correlation coefficient [32], [33] is employed to quantitate the correlation, which can be defined as follows:

$$\text{pearsonr} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\sum_{i=1}^{T}(X_i-\mu_X)^2}\sqrt{\sum_{i=1}^{T}(Y_i-\mu_Y)^2}} \tag{1}$$

where $X = [X_1, \ldots, X_T]$ and $Y = [Y_1, \ldots, Y_T]$ represent the two vectors to be measured, and the length is T. μ_X and μ_Y denotes the means of X and Y, respectively.

D. Early Short-Term Prediction Model Integrating Multiple Factors

Based on multisource urban data, we construct a fixed effect MR model, in which the historical confirmed cases are integrated to predict the daily new confirmed cases, especially for these cities that take high importation risks. Different from the traditional transmission dynamics models, the model is data-driven. Without loss of generality, the input variables consist of the factors that correlate with the confirmed cases and the amount of historical confirmed cases. For simplicity, we define F as the set of these factors, in which $f \in F$ represents a factor in F (*e.g.* daily population inflow ratio of Hubei toward target city). Fixing the disease incubation period H, $[f(t-2H+1), \ldots, f(t-H)]$ stands for the daily value of the factor f in H time interval, and $[I(t-H+1), \ldots, I(t)]$ denotes the sequence which is composed of corresponding cumulative confirmed case. Based on the above historical observations, the number of cumulative confirmed cases at future time point $t+n$ (n is the time step size of prediction) can be modeled as follows:

$$\hat{I}(t+n) = \gamma_0 + \sum_{f\in F}\sum_{i=0}^{H-1}\left[\beta_{f,i} f(t-H-i)\right] + \sum_{i=0}^{H-1}\alpha_i I(t-i) \tag{2}$$

where γ_0, $\beta_{f,i}$, and α_i are the trainable parameters of the model. $\hat{I}_{t+n}$ refers to the predicted number of cases at $t+n$th day. We apply a supervised machine learning approach to estimate the parameters of the model.

III. Results

A. Evaluation Criteria

In the task of the correlation analysis, the Pearson correlation coefficient and p value are applied. Particularly, the factor with a high value of correlation and p value that is less than 0.05 is considered a statistically significant feature for subsequent prediction mission.

Verifying the performance of the short-term prediction, the mean absolute error (MAE), root mean-squared error (RMSE), and the coefficient of determination-R^2 are adopted. Given a series of forecast values $(\hat{I}_1, \hat{I}_2 \ldots, \hat{I}_N)$ and their corresponding ground truth values $(I_1, I_2, \ldots, I_N)$, above

[1] http://wsjkw.gd.gov.cn
[2] http://qianxi.baidu.com
[3] https://www.aqistudy.cn/historydata
[4] http://lishi.tianqi.com/shenzhen/index.html

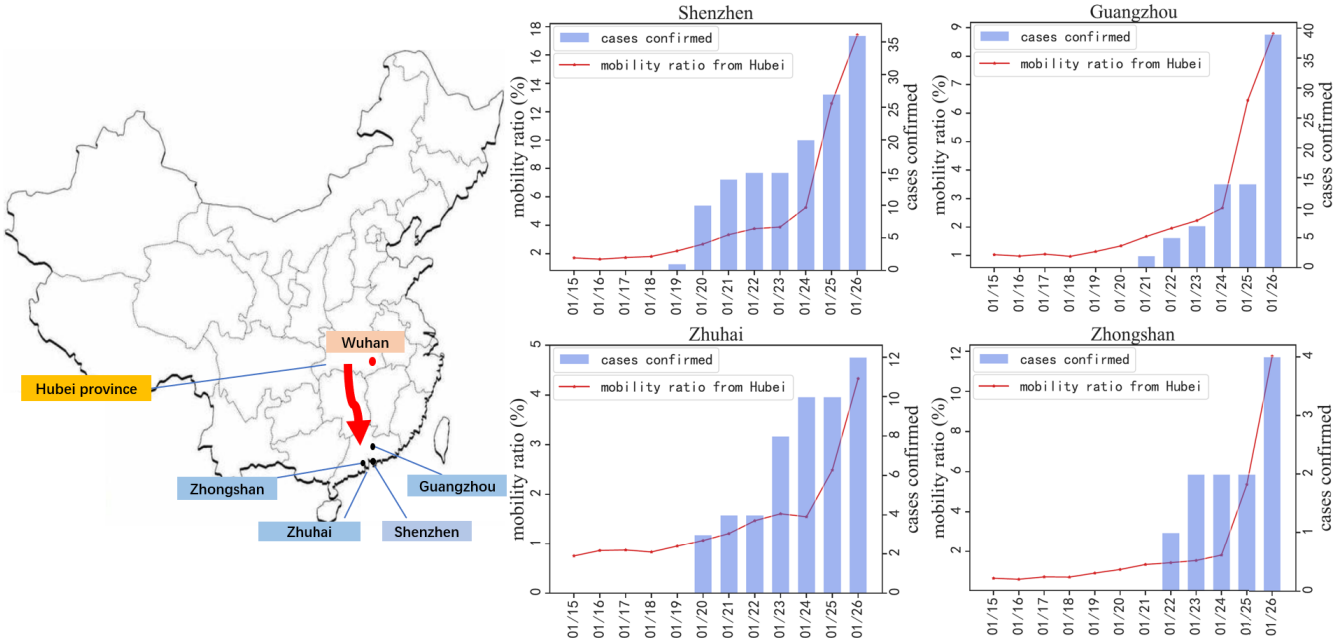


Fig. 2. Confirmed cases from the Health Commission of Guangdong Province and human mobility from Hubei province to Shenzhen, Guangzhou, Zhuhai, and Zhongshan, respectively.

evaluations can be calculated as follows:

$$\text{MAE} = \frac{1}{N}\sum_{j}|I_j - \hat{I}_j| \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{j}||I_j - \hat{I}_j||_2^2} \tag{4}$$

$$R^2 = 1 - \frac{\sum_j ||I_j - \hat{I}_j||_2^2}{\sum_j ||I_j - \bar{I}||_2^2} \tag{5}$$

where N is the total number of forecast days, I_j and $\hat{I}_j$ indicates the amount of ground truth and predicted cases in the jth day, respectively. $\bar{I}$ denotes the mean values of true series. The RMSE and MAE show the degree of deviation, while R^2 represents the extent of interpretation of the independent variable with respect to the dependent variable. A prediction with great performance can be represented as the low values of RMSE and MAE and the high one of R^2.

B. Correlation Analysis for Multiple Factors

The human mobility from Hubei province plays a critical role due to that the cases reported in China are almost derived from Hubei province until January 23, 2020. Fig. 2 depicts the population outflow from Hubei province toward several major cities in Guangdong province (Shenzhen, Guangzhou, Zhuhai, and Zhongshan), during the Chinese Lunar New Year ranging from January 15th to 26th. Especially approaching the Spring Festival, the mobility rate has dramatically increased every day. Fig. 3 shows the Pearson correlation coefficient between the daily new confirmed cases and factors selected with 6 days lag. As Fig. 3(a) illustrates, there is a clear correlation between population inflow from Hubei province and early cases in Shenzhen (pearsonr $= 0.95$), which is consistent with the statistics of the real cases, according to the official data of Shenzhen Health Commission (Fig. 4). According to Fig. 3(b) and (c), the daily average temperature has a weakly positive correlation with the increased number of confirmed cases (pearsonr $= 0.41$), while relative humidity shows negative correlation (pearsonr $= -0.62$). Furthermore, Fig. 3(d)–(f) show the correlation of other air quality factors. The results indicate that the AQI, PM2.5, and PM10 are all negatively correlated with the confirmed cases in the early phase of the outbreak (pearsonr $= -0.31$, -0.25, and -0.43, respectively). Validating whether some factors exhibit a lag effect, the time interval is expanded to 14 days. Table I shows the details of the correlation coefficient between the investigated factors and the number of daily new confirmed cases in Shenzhen with respect to different lag days. From the statistical results, it can be concluded that human mobility shows the highest consistency, while the other factors almost show the negative or weak one. In the meantime, p values of mobility ratio are all less than 0.05. It is noteworthy that the clearest correlation between population mobility and the number of confirmed cases occurs at 6-day lag, which is consistent with the mean incubation period investigated in the literature. Following this observation, we further conduct other experiments to verify the importance of tracking population movements from Hubei province.

C. Short-Term Predictions

According to the above correlation analysis, the impact of population flow on the early diffusion is selected as the main factor for short-term prediction. To further demonstrate the effectiveness of human mobility at the early stage, we only use human mobility as the input feature for our model firstly. And in our work, the other two classical regression models, the least absolute shrinkage and selection operator (LASSO) and ridge regression (RR), are also used for verification and comparison. Based on the above analysis, we estimate the number of infected cases from February 1st to 15th.

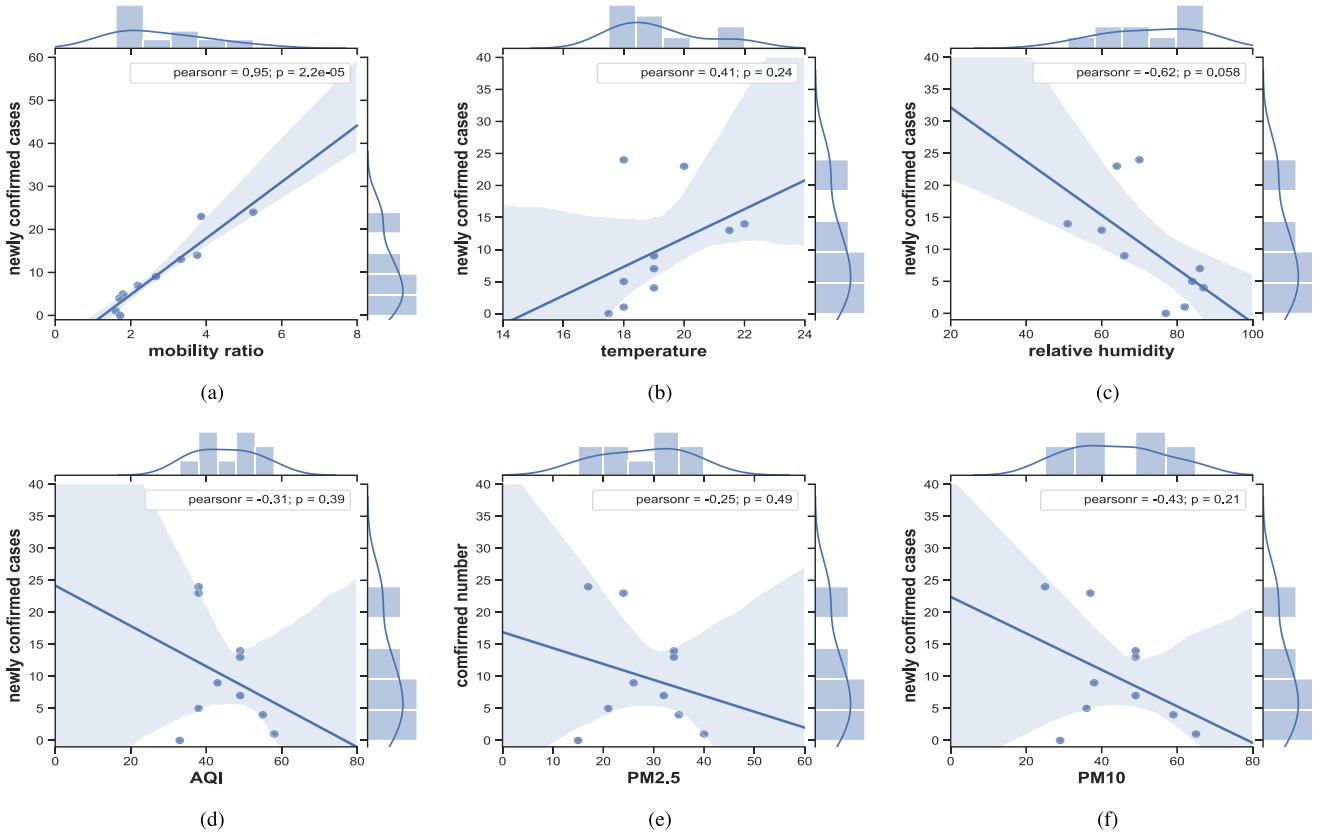


Fig. 3. Correlation analysis between the number of confirmed cases and other factors in the early phase of the COVID-19 pandemic from January 21, 2020 to January 30, 2020, which consists of (a) mobility rate from Hubei province to Shenzhen, (b) temperature, (c) relative humidity, (d) AQI, (e) PM2.5, and (f) PM10.

TABLE I

CORRELATION ANALYSIS BETWEEN THE NUMBER OF CONFIRMED CASES AND OTHER FACTORS IN SHENZHEN

n-day-lag	Mobility ratio		Temperature		Relative humidity		AQI		PM2.5		PM10	
	pearsonr	*p-value*	*pearsonr*	*p-value*	*pearsonr*	*p-value*	*pearsonr*	*p-value*	*pearsonr*	*p-value*	*pearsonr*	*p-value*
$n = 6$	0.95	2.2e-5	0.41	0.240	-0.62	0.058	-0.31	0.386	-0.25	0.489	-0.43	0.212
$n = 7$	0.94	4.5e-5	0.52	0.084	-0.85	0.002	-0.36	0.309	-0.14	0.684	-0.39	0.261
$n = 8$	0.93	7.6e-5	0.27	0.049	-0.88	0.001	-0.20	0.570	0.05	0.891	-0.26	0.456
$n = 9$	0.92	1.1e-4	-0.28	0.433	-0.80	0.005	-0.25	0.482	0.00	1.000	-0.34	0.336
$n = 10$	0.91	1.4e-4	-0.51	0.135	-0.40	0.247	-0.34	0.333	-0.14	0.692	-0.41	0.229
$n = 11$	0.91	2.3e-4	-0.45	0.187	0.39	0.266	-0.39	0.271	-0.19	0.595	-0.43	0.207
$n = 12$	0.92	1.3e-4	-0.47	0.167	0.32	0.370	-0.67	0.034	-0.48	0.157	-0.68	0.029
$n = 13$	0.89	4.1e-4	-0.67	0.031	0.44	0.204	-0.33	0.336	-0.07	0.832	-0.31	0.389
$n = 14$	0.73	0.016	-0.85	0.002	0.29	0.409	0.06	0.865	0.47	0.164	0.24	0.490

And the corresponding daily confirmed data is used to verify the model. Table II gives quantitative results of models with different factors. Only using the human mobility data, the three models achieve high R^2 value, which further illustrates the effectiveness of this factor. However, as shown in the table, human mobility can only reflect the overall trend, and it cannot accurately represent the changes in confirmed cases. Therefore, only using single mobility factor, the prediction performance of the model is limited. With the integration of historical case data, the prediction performance has been further improved. Taking Shenzhen as an example, the R^2 value of daily prediction is increased to 0.988 compared with only considering the population inflow of Shenzhen.

Table III shows the prediction performance from 1- to 5-day-ahead for Shenzhen, Guangzhou, Zhuhai, and Zhongshan. For the daily forecast, our model predicts amounts of daily cases across these cities with relatively high accuracy, in which all the values of R^2 are greater than 0.9. Especially for cities like Shenzhen and Guangzhou with frequent population flow per day, the values of R^2 of daily prediction achieve 0.988 and 0.985, which implies that the magnitude of the early pandemic outside of Hubei province is very well related to the volume of human mobility out of Hubei province. This is also apparent from the case counts by other cities in Guangdong province. With increasing forecast days n, the values of R^2 begin to decline. But the overall trend

TABLE II

PREDICTION PERFORMANCE OF ONE-DAY-AHEAD IN SHENZHEN, GUANGZHOU, ZHUHAI, AND ZHONGSHAN OF GUANGDONG PROVINCE WITH SINGLE OR MULTIPLE FACTORS

Models	Multiple factors	Shenzhen			Guangzhou			Zhuhai			Zhongshan		
		MAE	*RMSE*	R^2	*MAE*	*RMSE*	R^2	*MAE*	*RMSE*	R^2	*MAE*	*RMSE*	R^2
RR	N	27.60	30.23	0.782	25.20	28.66	0.706	7.800	9.088	0.652	7.600	8.694	0.643
LASSO	N	28.73	31.54	0.761	27.13	30.14	0.675	8.333	9.623	0.610	7.733	8.921	0.625
MR	N	26.80	29.38	0.800	24.53	27.47	0.729	8.667	9.338	0.643	6.800	8.618	0.651
RR	Y	9.600	10.63	0.973	7.000	7.496	0.979	4.000	4.367	0.919	2.733	3.454	0.943
LASSO	Y	10.66	11.84	0.966	8.800	9.338	0.968	4.333	4.947	0.896	3.400	4.155	0.919
MR	Y	7.000	7.234	0.988	6.067	6.434	0.985	3.400	3.568	0.946	2.667	2.921	0.960

TABLE III

PREDICTION PERFORMANCE OF n-DAY-AHEAD IN SHENZHEN, GUANGZHOU, ZHUHAI, AND ZHONGSHAN OF GUANGDONG PROVINCE

n-day-ahead	Shenzhen			Guangzhou			Zhuhai			Zhongshan		
	MAE	*RMSE*	R^2	*MAE*	*RMSE*	R^2	*MAE*	*RMSE*	R^2	*MAE*	*RMSE*	R^2
$n=1$	7.000	7.234	0.988	6.067	6.434	0.985	3.400	3.568	0.946	2.667	2.921	0.960
$n=2$	21.60	23.16	0.872	20.80	22.70	0.816	5.000	5.698	0.863	3.400	3.873	0.929
$n=3$	31.10	33.48	0.733	25.90	27.17	0.736	6.670	8.270	0.712	5.667	8.379	0.670
$n=4$	41.20	43.46	0.550	33.60	35.60	0.561	10.33	10.83	0.506	7.867	9.709	0.556
$n=5$	43.53	46.31	0.490	41.67	44.34	0.300	10.66	11.75	0.419	9.867	13.59	0.230

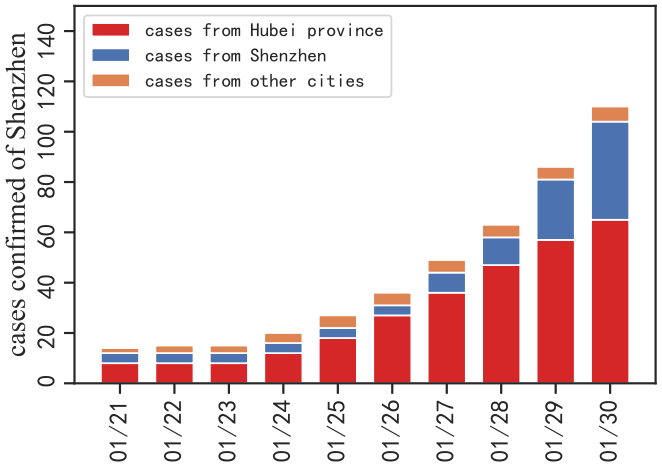


Fig. 4. Distribution of COVID-19 in Shenzhen with respect to the date of onset.

can still be captured. The main reasons for this phenomenon are as follows: first, the number of training samples used for learning decreases with the increase of n; second, because of the strong correlation between early human mobility and disease, the difference in daily mobility is still obvious. Therefore, the lack of real-time data supplement will also have a certain impact on the accuracy of the prediction model.

Also, Fig. 5 shows the pandemic trends for Shenzhen, Guangzhou, Zhuhai, and Zhongshan of Guangdong province from February 1, 2020 to February 15, 2020. From the figure, we can find that there is a fine synchronization between the number of predicted cases and that of ground truth cases for daily prediction. It visually proves the prediction ability of our model in the early phase of the pandemic. The proposed model has served as a reference for decision support of epidemic prevention and control in Shenzhen at the early outbreak of the COVID-19.

IV. DISCUSSION

In the earliest phase of the pandemic, the spatial distribution of COVID-19 cases in China was explained well by human mobility data when the human-to-human transmission is established. Especially, for the first-tier megacities like Shenzhen city, we employ multisource data to verify it. Therefore, cases exported from Hubei province appear to have contributed to the initiation of transmission chains for Guangdong province. Obviously, with booming population outflow from Hubei province, the number of confirmed cases will be significantly increased in the early phase (e.g. Shenzhen and Guangzhou). Therefore, the rapid and massive responses of China are necessary, which substantially slows the spread of COVID-19 through the implementation of unprecedented containment measures, including the implementation of the travel ban, lockdown of the whole province of Hubei, and so on.

At the early stage of the outbreak, the analysis model with the population mobility data can realize the short-term and quick prediction of the pandemic spread. At the same time, it also shows that the early development of strategies, such as the lockdown of cities, is of great significance in blocking the rapid and large-scale spread of the pandemic.

Overall, the proposed model is mainly suitable for the situation of the normal mobility of the population at an early stage. Since the prevention and control measures have been adopted by all cities in the later stage, which leads to limited importation, the prediction of the short-term model may be influenced to some extent. It is worth noting that as of February 2nd, cases of intracommunity transmission have occurred in Shenzhen; thereby the effectiveness of our method that is based on the investigation of importation risk has gradually decreased. With the large-scale return flow of the Spring Festival and the stable state of mobility, the risk of local community transmission overtaking imported cases

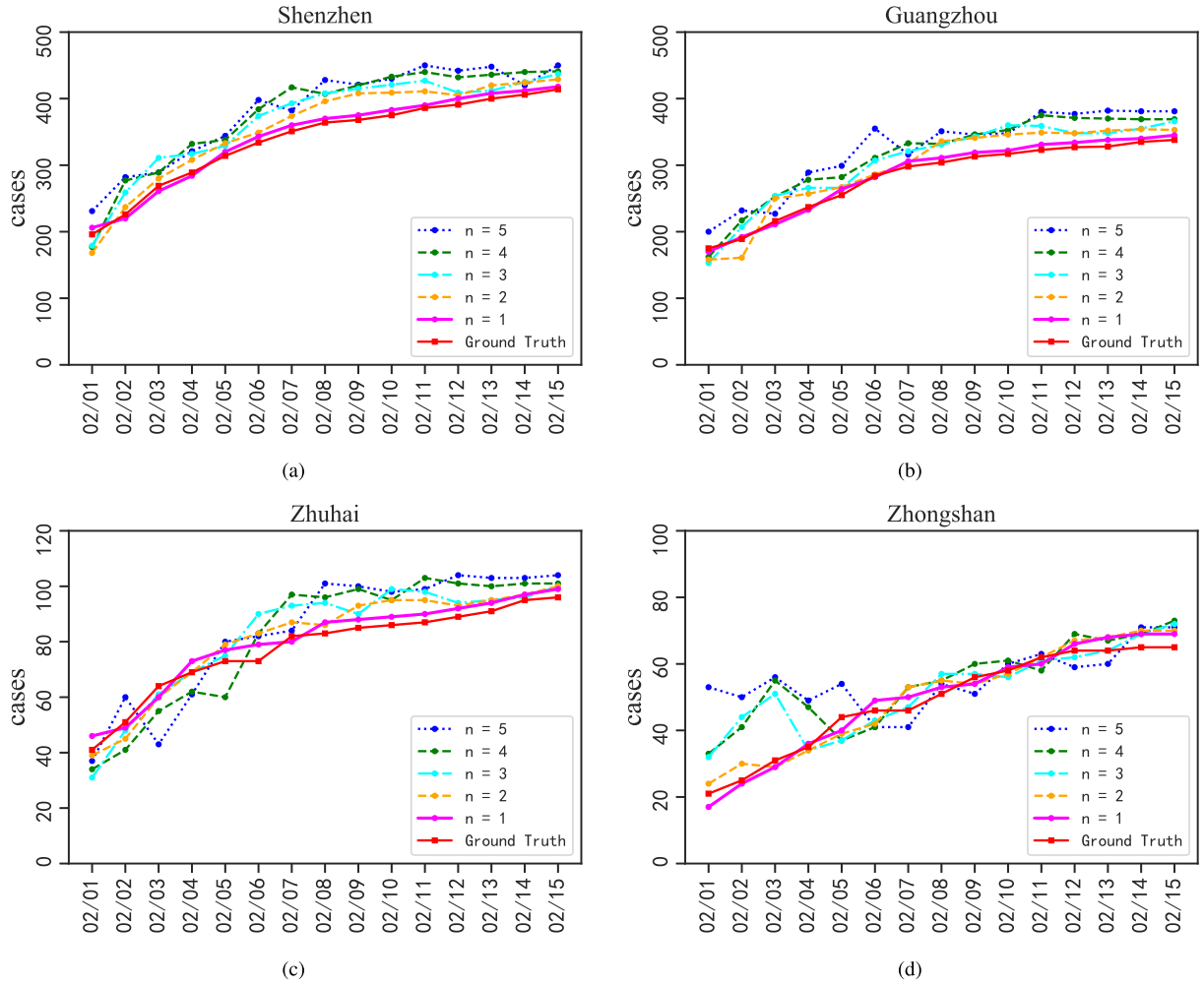


Fig. 5. Estimation of the number of cumulative infected cases for (a) Shenzhen, (b) Guangzhou, (c) Zhuhai, and (d) Zhongshan.

cannot be ignored. At the middle and late stages of the pandemic, to analyze the transmission risk under the local steady-state situation, it is necessary to further combine the actual local situation of disease infection and cure. What is more, thoroughly considering the multidimensional factors such as climate, population behavior, etc., a regional transmission risk model can be built to analyze the evolution of infectious disease dynamics within the city, and more scientific and reasonable suggestions for the daily protection of residents and decision-makers can be also provided.

V. Conclusion

In this article, we first analyze the correlation between some possible factors and the daily new confirmed cases, utilizing multisource urban data where many elements are somewhat or closely related to the transmission of the pandemic. What is more, based on the actual human mobility data and pandemic trend, a fixed effect MR model is proposed for short-term prediction in the early phase of the COVID-19 outbreak. By tracking the population flow in real-time, the model can provide powerful suggestions for decision-makers and epidemiologists. In this way, they can settle down appropriate policies to promptly and effectively prevent and control the pandemic and save lives to the largest extent.

References

[1] WHO. *WHO Coronavirus Disease (COVID-19) Dashboard*. Accessed: 2020. [Online]. Available: https://www.who.int/redirect-pages/page/novel-coronavirus-(covid-19)-situation-dashboard

[2] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, and M. Wang, "Presumed asymptomatic carrier transmission of COVID-19," *Jama*, vol. 323, no. 14, pp. 1406–1407, 2020.

[3] H. Chen *et al.*, "Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: A retrospective review of medical records," *Lancet*, vol. 395, no. 10226, pp. 809–815, Mar. 2020.

[4] V. J. Munster, M. Koopmans, N. van Doremalen, D. van Riel, and E. de Wit, "A novel coronavirus emerging in China—Key questions for impact assessment," *N. Engl. J. Med.*, vol. 382, no. 8, pp. 692–694, 2020.

[5] C. Huang *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet*, vol. 395, pp. 497–506, May 2020.

[6] N. Chen *et al.*, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study," *Lancet*, vol. 395, no. 10223, pp. 507–513, Feb. 2020.

[7] D. Zhao *et al.*, "A comparative study on the clinical features of coronavirus 2019 (COVID-19) pneumonia with other pneumonias," *Clin. Infectious Diseases*, vol. 71, no. 15, pp. 756–761, Jul. 2020.

[8] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study," *Lancet*, vol. 395, no. 10225, pp. 689–697, Feb. 2020.

[9] Q. Li *et al.*, "Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia," *New England J. Med.*, vol. 382, pp. 1199–1207, Dec. 2020.

[10] M. Halloran *et al.*, "Ebola: Mobility data," *Science*, vol. 346, no. 6208, p. 433, Oct. 2014.
[11] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, Jan. 2006.
[12] D. Brockmann and D. Helbing, "The hidden geometry of complex, network-driven contagion phenomena," *Science*, vol. 342, no. 6164, pp. 1337–1342, Dec. 2013.
[13] J. S. Jia *et al.*, "Population flow drives spatio-temporal distribution of COVID-19 in China," *Nature*, vol. 582, pp. 389–394, Jun. 2020.
[14] N. P. Jewell, J. A. Lewnard, and B. L. Jewell, "Predictive mathematical models of the COVID-19 pandemic: Underlying principles and value of projections," *JAMA*, vol. 323, no. 19, pp. 1893–1894, 2020.
[15] V. Colizza *et al.*, "Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions," *PLos Med.*, vol. 4, no. 1, pp. 0095–0110, 2007.
[16] S. Chen, J. Yang, W. Yang, C. Wang, and T. Bärnighausen, "COVID-19 control in China during mass population movements at new year," *Lancet*, vol. 395, no. 10226, pp. 764–766, Mar. 2020.
[17] S. Zhao *et al.*, "Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of Jan. 2020: A data-driven modelling analysis of the early outbreak," *J. Clin. Med.*, vol. 9, no. 2, pp. 1–6, 2020.
[18] M. Chinazzi *et al.*, "The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak," *Science*, vol. 368, no. 6489, pp. 395–400, Apr. 2020.
[19] J. Jiang and L. Luo, "Influence of population mobility on the novel coronavirus disease (COVID-19) epidemic: Based on panel data from Hubei, China," *Global Health Res. Policy*, vol. 5, no. 1, pp. 1–10, Dec. 2020.
[20] M. U. G. Kraemer *et al.*, "The effect of human mobility and control measures on the COVID-19 epidemic in China," *Science*, vol. 368, no. 6490, pp. 493–497, May 2020.
[21] Z. Yang *et al.*, "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," *J. Thoracic Disease*, vol. 12, no. 3, pp. 165–174, Mar. 2020.
[22] M. Catalá, S. Alonso, E. Alvarez-Lacalle, D. López, P.-J. Cardona, and C. Prats, "Empirical model for short-time prediction of COVID-19 spreading," *PLOS Comput. Biol.*, vol. 16, no. 12, Dec. 2020, Art. no. e1008431.
[23] A. J. Kucharski *et al.*, "Early dynamics of transmission and control of COVID-19: A mathematical modelling study," *Lancet Infectious Diseases*, vol. 20, pp. 553–558, May 2020.
[24] M. Maleki, M. R. Mahmoudi, D. Wraith, and K.-H. Pho, "Time series modelling to forecast the confirmed and recovered cases of COVID-19," *Travel Med. Infectious Disease*, vol. 37, Sep. 2020, Art. no. 101742.
[25] V. Charu *et al.*, "Human mobility and the spatial transmission of influenza in the United States," *PLoS Comput. Biol.*, vol. 13, no. 2, pp. 1–23, 2017.
[26] M. Thomson, T. Palmer, A. Morse, M. Cresswell, and S. Connor, "Forecasting disease risk with seasonal climate predictions," *Lancet*, vol. 355, no. 9214, pp. 1559–1560, Apr. 2000.
[27] Y. Zhang, H. Bambrick, K. Mengersen, S. Tong, and W. Hu, "Using Google Trends and ambient temperature to predict seasonal influenza outbreaks," *Environ. Int.*, vol. 117, pp. 284–291, Aug. 2018.
[28] K. Zhang *et al.*, "Clinically applicable AI system for accurate diagnosis, quantitative measurements and prognosis of COVID-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433, 2020.
[29] M. M. Böhmer *et al.*, "Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: A case series," *Lancet Infectious Diseases*, vol. 20, no. 8, pp. 920–928, Aug. 2020.
[30] S. A. Lauer *et al.*, "The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application," *Ann. Internal Med.*, vol. 172, no. 9, pp. 577–582, May 2020.
[31] S. Tian *et al.*, "Characteristics of COVID-19 infection in Beijing," *J. Infection*, vol. 80, no. 4, pp. 401–406, Apr. 2020.
[32] F. Sun, C. Mao, X. Fan, and Y. Li, "Accelerometer-based speed-adaptive gait authentication method for wearable IoT devices," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 820–830, Feb. 2019.
[33] W. Wiedermann and M. Hagmann, "Asymmetric properties of the pearson correlation coefficient: Correlation as the negative association between linear regression residuals," *Commun. Statist.-Theory Methods*, vol. 45, no. 21, pp. 6263–6283, Nov. 2016.

Ruxin Wang received the Ph.D. degree in operational research and cybernetics from the University of Chinese Academy of Sciences, Beijing, China, in 2017.

He is currently an Assistant Research Fellow with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. His current research interests include artificial intelligence in medical big data, graph data mining, machine learning and pattern recognition.

Chaojie Ji received the M.S. degree in software engineering from Nanchang University, Jiangxi, China, in 2018.

He is currently an Engineer with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His current research interests include knowledge graph and graph neural networks.

Zhiming Jiang received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2018. He is currently pursuing the M.S. degree in computer science with the University of Chinese Academy of Sciences, Beijing, China.

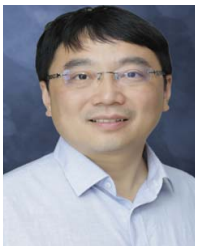
His research interests include data mining and applications and information system.

Yongsheng Wu is currently a Professor and the Director with the Department of Public Health Information, Shenzhen Center for Disease Control and Prevention.

He served as a Member of the Committee of the Public Health Information Committee, China Association of Health Information and Health Medical Big Data, an Executive Committee Member of the Shenzhen Society of Preventive Medicine, a Member of the Public Health Committee of Shenzhen Health Information Association. His current research interests include public health information management medical data statistic analysis.

Ling Yin received the B.S. and M.S. degrees in geographical information system (GIS) from Nanjing University, Nanjing, China, in 2003 and 2006, respectively, and the Ph.D. degree in geographical information science from the University of Tennessee at Knoxville, Knoxville, TN, USA, in 2011.

She has been working with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, since 2011, and has been a Full Professor since 2020. Her research interests include spatiotemporal data mining, spatially epidemic modeling, urban computing and GIS for smart city.



Ye Li (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1999 and 2002, respectively, and the Ph.D. degree in electrical engineering from Arizona State University, AZ, USA, in 2006.

He is currently a Full Professor with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. He has authored or coauthored more than 150 articles in prestigious journals and conferences like *Information Fusion*, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE INTERNET OF THINGS JOURNAL, the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, and the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING. His research interests include medical big data, artificial intelligence, and health informatics.

Dr. Li has been serving as an Editorial Board Member for *Journal Information Fusion* since 2018 and the Program Chair for the IEEE UIC 2021.