

Received June 14, 2020, accepted June 20, 2020, date of publication June 25, 2020, date of current version July 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3004933

Risk Assessment of COVID-19 Based on Multisource Data From a Geographical Viewpoint

YAN ZHANG^{1,2}, YINGBING LI¹, BO YANG³, XIANG ZHENG⁴, AND MIN CHEN¹

¹School of Geodesy and Geomatics, Wuhan University, Wuhan 430072, China

²State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China

³School of Public Administration, Zhongnan University of Economics and Law, Wuhan 430073, China

⁴School of Information Management, Wuhan University, Wuhan 430072, China

Corresponding author: Yingbing Li (ybli@whu.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2018YFC0807000.

ABSTRACT In June 4, 2020, Corona Virus Disease 2019(COVID-19) cases in Wuhan were cleared, and the epidemic situation was basically controlled. Such public safety infectious disease includes influences great pressure on the national economy. At present, some countries and regions in the world are still in epidemic situation, and there is an urgent need to judge the infection situation and travel risk in the region. In a relatively fine scale down to perceive the surrounding situation, and then rational zoning decisions to promote the resumption of production and work. In this study, indicators for the evaluation of COVID-19 epidemic were constructed using multi-sourced data. A computational evaluation of 736 fine-grained grids was performed using the GeoDetector model and the decision tree model. The study found that the risk level in older neighborhoods was much higher than in newer neighborhoods; the population density was the most important determinant of infection; the number of urban people slumped to 37% of that in usual times according to Tencent data after the “city closure”; The model this paper used portrays the major factor in defining low-risk areas and high-risk areas, and offers suggestions and assessment from a geographical perspective to fight COVID-19, thus presenting great practical value.

INDEX TERMS COVID-19, geographical detectors, decision tree, risk map, spatial grid.

I. INTRODUCTION

Infectious diseases are common risks across national boundaries and society; they seriously threaten not only public life and health but also social stability and economic development. Due to the rapid development of urbanization and tourism and the changing ecological environment, as well as the fragile public health system, epidemics are becoming more frequent, more complex and more difficult to prevent and control [1]. Novel infectious diseases, such as Ebola hemorrhagic fever, Middle East respiratory syndrome, and Coronavirus Disease 2019 (COVID-19), are constantly emerging. Some of these diseases are characterized as zoonosis and/or coexist with other epidemics, which greatly increase their infectiousness and pathogenicity. Epidemics that occur in densely populated metropolitans will likely develop into global pandemics. The vulnerability of the population to these situations has increased, and the combination and interaction of these factors and influences have contributed to a more complex epidemic. COVID-19, the

pandemic that is spreading worldwide, has revealed the vulnerability of human society to severe infectious diseases and the difficulty of solving this problem in a globally interconnected complex system. COVID-19 affected more than 100 countries in a span of weeks. As a result, the whole human race should not only collaborate to overcome the epidemic but also reasonably arrange to return to work and production according to the actual situation of each region and carry out geographical risk assessment [2].

In China, there is a key time difference between COVID-19 and other epidemic viruses. The epidemic occurred during the Chinese New Year, which generated an immense population flow. According to the statistics, from January 10 to 27, 2020, a total of 1.202 billion passengers traveled by railways, roads, waterways and civil aviation. In the early morning of January 23, Wuhan announced its “closure”. The city’s public transport systems—subway, airport, railway station, ferry and long-distance passenger transport systems—were temporarily closed. For the first time in human history, a large city with a population of 10 million people adopted “closure” measures to prevent the epidemic.

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott¹.

II. LITERATURE REVIEW

After the outbreak of COVID-19, some researchers employed traditional SEIR (Susceptible-Exposed-Infectious-Removed) model to simulate the spread of the epidemic and applied personnel migration data to modify the model [3]. Zhu *et al.* [4] investigated the city-scale dynamics of the epidemic using mobile phone city data. They also trained and validated models based on the classic SIR (Susceptible-Infectious-Removed) model to predict the future trend in different scenarios. Liu *et al.* [5] designed a flow SEIR model that utilizes Baidu migration data to estimate the risk of epidemic spread when people return to work after holidays. Li *et al.* [6] employed a SEIR model that considers the effect of control measures to predict the spread of virus.

From a statistical point of view, Fu *et al.* [7] applied the Boltzmann function to simulate the cumulative number of confirmed cases in each province/municipality and mainland China and predicted the developing trend of confirmed cases in subsequent weeks. Danon *et al.* [8] constructed a model to estimate early transmission trends and peak times of the disease in England and Wales and analyzed the effects of seasonal variation in transmission rates. Based on the accumulated data sets of reports, deaths, isolation and suspected cases, Tang *et al.* [9] consider that the epidemic trend mainly depends on isolation and suspected cases. Thus, it is very important to continue to strengthen quarantine and isolation strategies and improve the detection rate. Wu *et al.* [10] collected and analyzed medical observation, discharge, infection, nonsevere, critical, cure, and death data and employed this state transfer matrix model to predict the peak inflection time and patient distribution to better allocate medical resources [11]. Anastassopoulou *et al.* [12] evaluated the basic regeneration number (R_0) and other major epidemiological parameters and predicted the infected population three weeks after the outbreak development according to the estimated parameters [13]. Bert-Dufresne *et al.* [14] emphasized the urgent need for tracking close contacts during outbreaks of emerging infectious diseases and the need to use a larger number than the R_0 value when predicting the size of epidemics. Based on data-driven analysis, Zhang *et al.* [15] estimated the basic regeneration number and outbreak scale of COVID-19 during the outbreak stage on the Diamond Princess cruise ship.

The previously mentioned study mainly simulates the time series of the number of infections, immunizations, deaths, cures and time nodes of pandemic influenza based on the classical infectious disease infection model [16], [17]. In addition to the traditional SIR/SEIR infectious disease model, some analytical methods from the Geographic Information System (GIS) and Social Network Analysis (SNA) have been introduced to the derivation of infectious diseases [18]–[22]. Powerful geospatial data collection, management, processing, analysis and display capabilities of Geographic Information System are increasingly employed by scholars for early warning research on infectious disease surveillance based

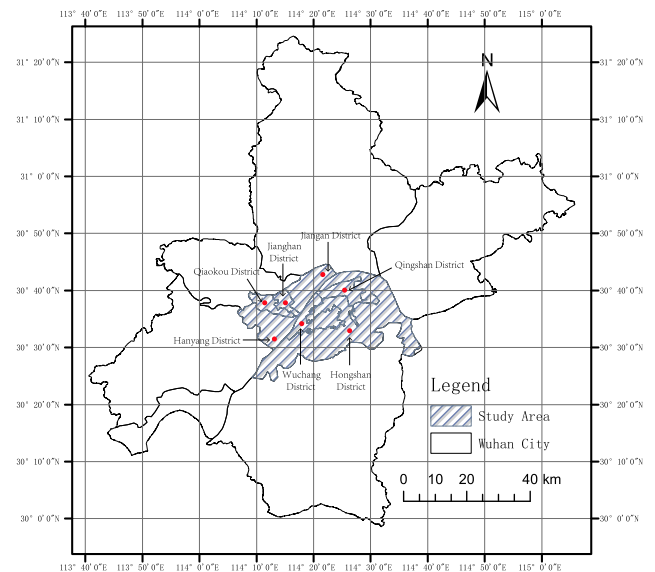


FIGURE 1. The study area.

on a combination of the strength of prevention and control measures, support capacity, support resources and severity of outbreaks in different regions [9], [23].

Previous studies have focused on the macroscopic perspective of the COVID-19 epidemic using Baidu migration and confirmed diagnosis data at the national or provincial scale. This paper applies multisource and open-source data, which is easy to obtain to perceive and classify the spatial situation of the epidemic, and explores the influence and role of various spatial elements on the breeding of the epidemic from a geographical perspective. We constructed a system of indicators for the assessment of major infectious diseases. Based on this evaluation system, a grid risk map was constructed.

III. STUDY AREA AND DATA

A. OVERVIEW OF THE STUDY AREA

The study area is Wuhan city center. As shown in Figure.1, there are 13 districts in Wuhan, which covers approximately 8,570 square kilometers, and a permanent population of 1,089,290. Seven of the districts encompass urban areas. According to the 2018 Wuhan statistical yearbook, there are 652,700, 868,500, 729,600, 962,400, 1,637,500, 1,276,300, and 528,800 people in Qiaokou, Hanyang, Jianghan, Hongshan, Wuchang, and Qingshan, respectively. Wuhan has a population of 6,655,800 permanent residents, excluding college students and migrant workers. According to the 2018 Yearbook, except for a small number of rural residents with household registrations in Hongshan district, other areas have urban residents with household registrations. All coordinates in this paper are transformed into the WGS84 (World Geodetic System 1984) geodetic coordinate system.

B. INTRODUCTION OF DATA

The data set is directly or indirectly related to the epidemic and is utilized to monitor, analyze the trend of the epidemic,

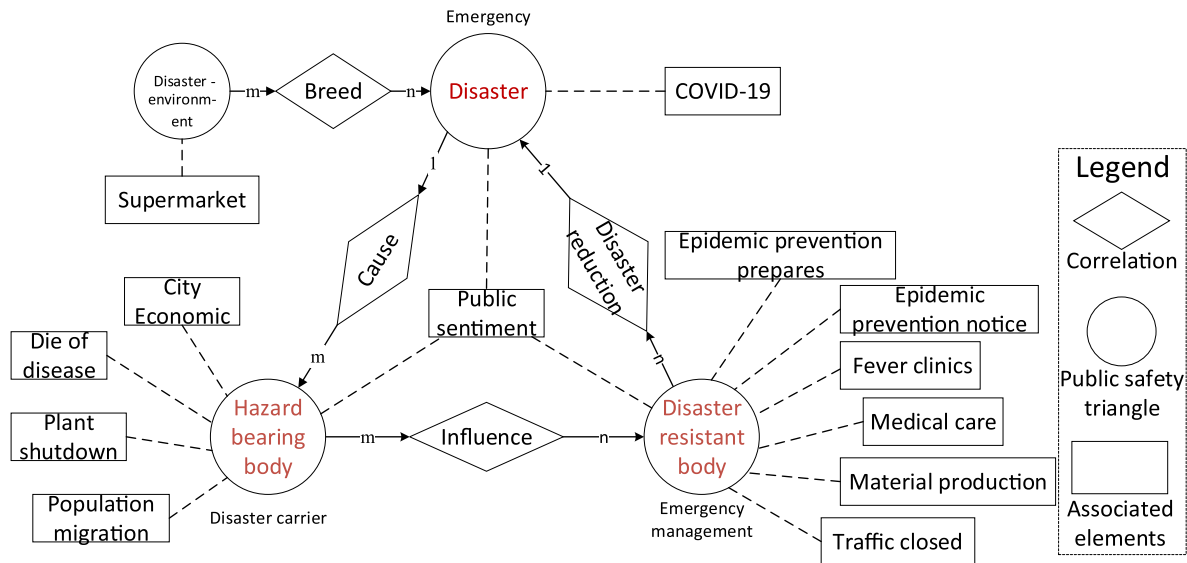


FIGURE 2. Composition of big data of major infectious diseases.

and guide the epidemic prevention strategy. According to the triangle theory framework of public safety, the big data of the epidemic situation is categorized into three types according to the disaster body (hazard, infectious disease and infectious disease breeding environment), hazard bearing body (subject affected by epidemic situation) and disaster resistant body (epidemic prevention and control), as shown in Figure.2. The disaster body is a kind of epidemic situation that damages human life, property and environment, including disease data and disaster environment data. The data include the confirmed infection caused by the epidemic and suspected cases. The environmental data of disasters is employed to reflect the external environment when an epidemic occurs, including densely populated places (such as supermarkets, fairs, public transport and other geographical entities). The hazard-bearing body is the social subject that is affected and damaged by the epidemic, including personnel data, location data and economic data. In this paper, the hazard-bearing body consists of real-time location data obtained by Tencent’s yichuxing heat map [25] and the data of population migration obtained by the transportation departments, such as highway, railway and airplane administrations. The economic data reflect the economic losses of various industries, such as businesses, schools, and public transport areas with dense populations that are closed with suspended operations due to the COVID-19 epidemic. The disaster-resistant body is the emergency response to break the chain of infectious disease spread and gradually control and eliminate the epidemic situation. This response consists of epidemic prevention preparation and response by fever clinics, rescue squads, etc. The epidemic data are organized as $\{time(T), space(x), epidemic(E)\}$ within a unified spatiotemporal framework. This study utilizes some open-source data for model calculation: Baidu map POI (point of interest)

data, Tencent yichuxing thermal data, government open fever clinics and designated hospital location data, as well as information about infected communities. Considering that viruses have an incubation period [26], [27], we disregarded the implications of POI closure. Based on the theory of disease transmission, this paper constructs four prior indicators for risk assessment, including “Infection source”, “Transmission route”, “Susceptible population”, and “Prevention ability”, and nine second-order indicators, including “Fever outpatient distance”, “Population flow”, “Daily population density”, “Population density after closure”, “Market distance”, “Hotel density”, “Quotient hyper density”, “Catering density” and “Traffic density”. The model requires the discretization or classification of data. Different data classification approaches may have different effects on the experimental results. After repeated tests, we divide all the evaluation indexes into 10 levels.

IV. METHOD

The geographic detector [28] consists of a set of statistical methods for detecting spatial heterogeneity and revealing its driving force. The basic assumption is that if an independent variable has an important influence on a dependent variable, the spatial distribution of the independent variable and the dependent variable should be similar. The core of the theory is to detect the consistency of the spatial distribution patterns between the dependent variable and the independent variable by calculating the spatial heterogeneity and then to measure the explanatory degree of the independent variable to the dependent variable. Geographical detectors are extensively applied in space analysis, such as the exploration of the driving factors of green buildings development [29], evaluating indicator of surface water quality [30], influencing factors of industrial sector carbon dioxide

TABLE 1. Interaction condition.

Interaction state	Interaction relationship
$q(X1 \cap X2) < \text{Min}(q(X1), q(X2))$	Nonlinear weakening
$\text{Min}(q(X1), q(X2)) < q(X1 \cap X2) < \text{Max}(q(X1), q(X2))$	Single factor nonlinear weakening
$q(X1 \cap X2) > \text{Max}(q(X1), q(X2))$	Two factor enhancement
$q(X1 \cap X2) = q(X1) + q(X2)$	Independent
$q(X1 \cap X2) > q(X1) + q(X2)$	Nonlinear enhancement

emissions [31], [32], etc. A GeoDetector can be employed in three aspects: measuring the spatial differentiation of given data; finding the largest spatial differentiation of variables; and finding the explanatory variables of dependent variables [33]–[37].

A. FACTOR DETECTION

There is a widely accepted basic hypothesis in geography, i.e., the closer is the distance between two spatial elements, the stronger is the correlation; the farther is the distance between two spatial elements, the weaker is the correlation. Because of the geographical isolation between two elements, the spatial heterogeneity emerges. The purpose of a GeoDetector is to measure the spatial heterogeneity, which is expressed by the statistical parameter q . The calculation indexes of the statistical parameter q are integrated in Eq. 1.

$$q = 1 - \frac{\sum_{h=1}^L N_h \sigma_h^2}{N \sigma^2} = 1 - \frac{SSW}{SST} \tag{1}$$

$$SSW = \sum_{h=1}^L N_h \sigma_h^2, \quad SST = N \sigma^2 \tag{2}$$

The value range of q is [0, 1]. SSW represents the sum of the variance in the feature layers, and SST represents the total variance in the whole region. The higher is the value, the stronger is the spatial heterogeneity. The greater is the effect of the feature layer’s independent variable x on the dependent variable y , the greater is the value of q . When $q = 0$, the independent variable x has no effect on the dependent variable y ; when $q = 1$, the independent variable x is the only determinant of the dependent variable y . Parameter q can be employed to test the significance of the parameters by a simple transformation to obey the noncentral distribution of f in Eq. 3.

$$F = \frac{N - L}{L - 1} \frac{q}{1 - q} \sim F(L - 1, N - L; \lambda) \tag{3}$$

$$\lambda = \frac{1}{\sigma^2} \left[\sum_{h=1}^L \bar{Y}_h^2 - \frac{1}{N} \left(\sum_{h=1}^L \sqrt{N_h} \bar{Y}_h \right)^2 \right] \tag{4}$$

The noncentrality parameter is λ and \bar{Y} is the arithmetic mean of element h .

B. INTERACTION DETECTION

To identify the interaction between different element layers, to interact with different layer elements, and detect the better interpretation and fitting of dependent variables under the

interaction, the interaction relationships among different elements are described in Table 1.

C. ECOLOGICAL EXPLORATION

Ecological detection of a geographical detector is performed to determine whether significant differences exist in the effects of different element layers on the spatial distribution of dependent variables. The degree is expressed by statistic f in Eq. 5.

$$F = \frac{N_{X1} (N_{X2} - 1) SSW_{X1}}{N_{X2} (N_{X1} - 1) SSW_{X2}} \tag{5}$$

$$SSW_{X1} = \sum_{h=1}^{L_1} N_h \sigma_h^2, \quad SSW_{X2} = \sum_{h=1}^{L_2} N_h \sigma_h^2 \tag{6}$$

where N_{X1} and N_{X2} represent the sample size, and SSW represents the variance sum of different feature layers. L_1 and L_2 represent the number of discretization levels for the feature layer. If the hypothesis of zero is rejected at the significant level, there is a significance difference between the two factors.

D. SPATIAL GRID

A regular grid with fine resolution is advantageous over the subjectively defined areal units [38]. The investigated area is divided into the chessboard grid structure according to latitude and longitude with the coding system, as shown in Figure.3. Thus, a spatiotemporal cube model is established with the dimension of time/layer and different grids have different properties. The model is utilized to perform time management and backtracking, discover space-time hotspots, obtain spatial statistics and perform overlay calculation of geographic information data [39]. A space-time cube model is a three-dimensional geographic visualization analysis method that maps spatiotemporal data into cubes and is very useful for discovering spatiotemporal patterns. At the same spatial location, the bars distributed in different time step ranges share the same location ID and form a bar time series. Within the same time range, the bars distributed in different spatial locations share the same time ID and form a time slice. Based on this model, we can systematically describe the data from the three heterogeneous information facets, i.e., semantics, space, and time/layer to take a spatiotemporal snapshot for any attribute. The merits of this model are its strong ability to share, store and acquire, which facilitate dynamic community division and spatial data mining [40], [41].

Each grid of the cube model represents an area in the real space, and the mapping relationship between the research real

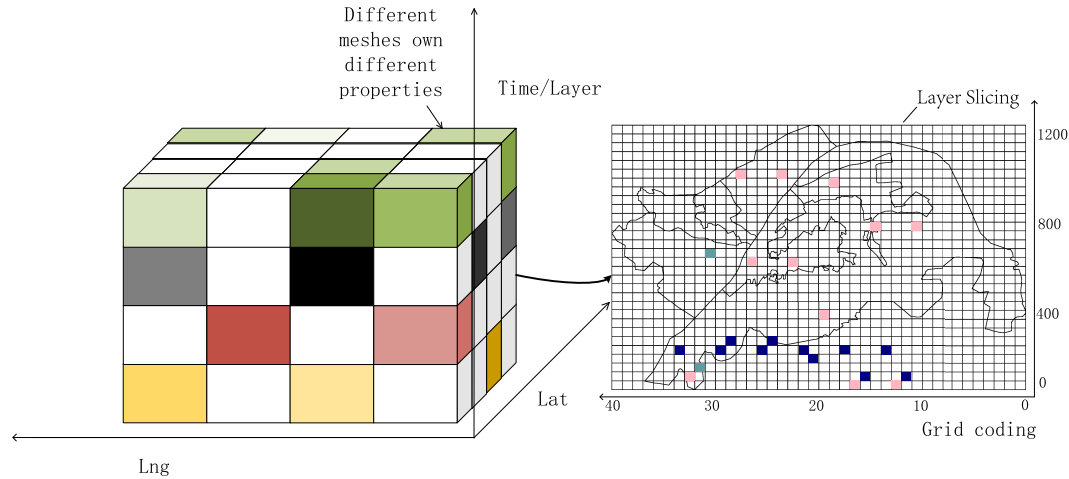


FIGURE 3. Space-time cube model.

space and the attribute space is established according to the longitude and latitude. By sampling the area data (such as the density surfaces of interest points, and population density surface at different times), the data can be assigned to each grid so that they have certain attributes and each attribute has a “slicing”.

V. RESULT

A. DATA PREPROCESSING

The data need to be preprocessed before they are utilized.

1) POI PROCESSING

Considering that all the fever clinics are public hospitals designated by the government, 71 fever clinics designated by the government in the study area are counted and plotted by the Euclidean distance [42]. The results are shown in Figure.4. With the exception of some marginal and suburban areas, most of the urban areas are closely located to fever clinics. We used the Baidu map POI data set, which consists of 9,896 supermarkets, 6,069 restaurants, 5,491 traffic types (bus stations and subway stations), and 1,901 hotel types in the study area.

2) POPULATION DATA PROCESSING

Tencent’s yichuxing location data is employed as the relative change index of the population. Tencent’s yichuxing data is obtained from Tencent’s big data location service window (<https://heat.qq.com/index.php>), which is based on Tencent’s multiple apps (cover nearly a billion people) for user location analysis and calculation [43], with a maximum spatial resolution of 25 meters by 25 meters.

This paper collects the data of Tencent’s yichuxing heatmap for two specific times, namely, 18:00 on October 25, 2019 and 18:00 on January 23, 2020, after closure of the city. The data include latitude, longitude, and counts, which describe the population density information of the region. As shown in Figure.5, data of the permanent population statistical yearbook, Tencent’s yichuxing location data on

October 25, 2019, and the number of infected people reported by each district are illustrated and compared. Pearson’s correlation test was carried out using data of the resident population and Tencent’s yichuxing daily population data. The positive correlation between the two types of data was 0.87, which indicates that the use of Tencent’s yichuxing location to represent the population is reliable [25], [44].

Compared with the statistical yearbook data, Tencent’s yichuxing data shows abnormal values. The population of Hongshan District, which has the largest number of online locations, is three times larger than that of Wuchang District, which is the second largest, while its resident population is only 3,000,000. It is speculated that the phenomena are related to the notion that college students are not included in the resident population in the statistics. The kernel density estimation method is employed to obtain the data of the two periods, which are compared using the same level of classification method. It can be seen from Figure.6 that the number of people has declined significantly after sealing off the city, and the relative data show that the population has decreased to only 37% of the normal period after closure of the city, according to Tencent data. Using the number of permanent residents, we calculate that approximately 4.2 million people have left the study area, which is similar to the official report of approximately 9 million people remaining and 5 million people leaving. By subtracting the grid, we determine that the area with the largest population decline is Hongshan district, and the population even increases in the suburbs of the city. According to the statistics, the attenuation areas shown in Figure.6C, from south to north, include university towns and business districts. Due to the influence of the Spring Festival and the fact that Wuhan has the largest number of college students in Central China, the decreased activity in these areas is very distinctive.

3) DISCUSSION OF POPULATION DISTRIBUTION DIRECTION

The standard deviation ellipse [45] was drawn, which contains 68% of the data, as shown in Figure.7; the two ellipses

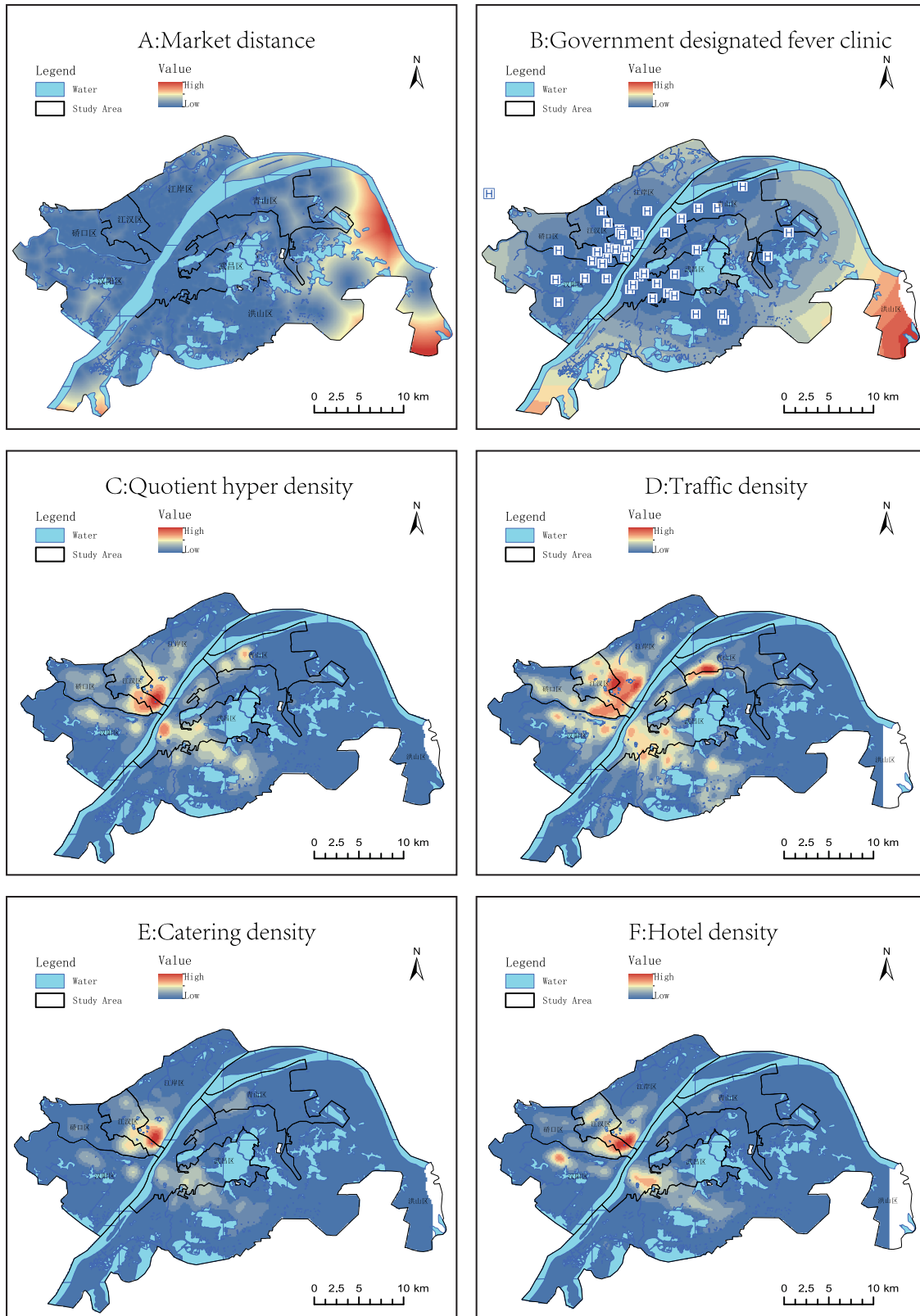


FIGURE 4. Data preprocessing results.

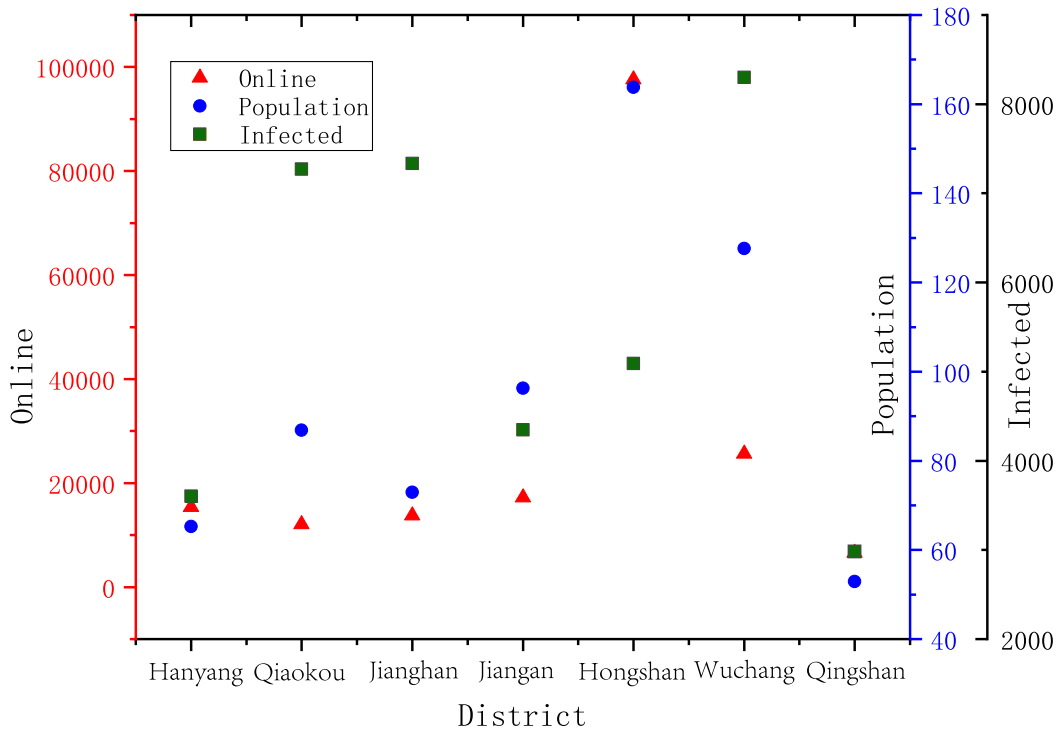


FIGURE 5. Evaluation of population data.

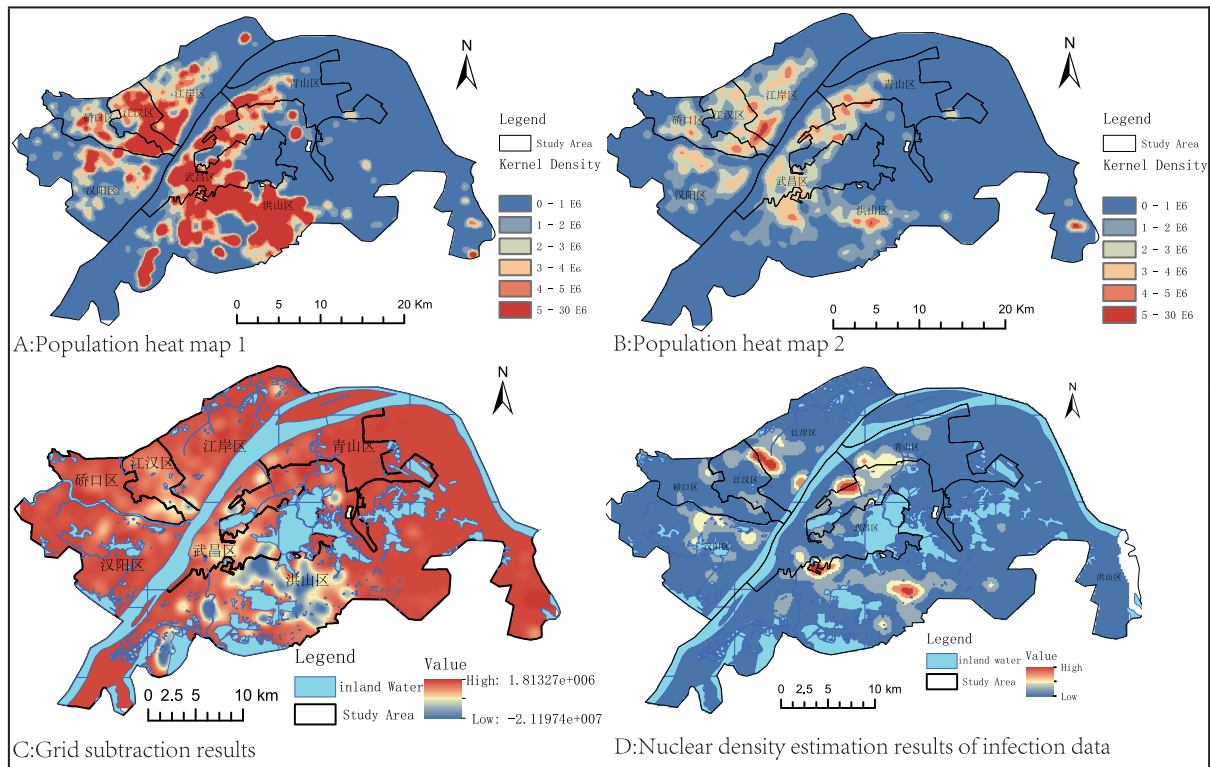


FIGURE 6. Population density data and infect data.

differ significantly between the daily ellipse and the ellipse after lockdown of the city. The population is usually highly concentrated in the ellipse centered between Wuhan Yangtze

River Bridge and Yuemachang, with a small area and a high concentration of population. The population distribution ellipse after closure is parallel to the main flow direction

TABLE 2. Risk factor detector.

	Fever clinic distance	Population reduction	Daily population density	Closed city population density	Market distance
<i>q</i> Statistics	0.18	0.19	0.31	0.48	0.16
<i>p</i> value	0.00	0.00	0.00	0.00	0.00
		Hotel density	Quotient hyper density	Catering density	Traffic density
<i>q</i> Statistics		0.26	0.27	0.29	0.33
<i>p</i> value		0.00	0.00	0.00	0.00

of the Yangtze River and the urban structure (southwest-northeast), and the coverage area increases. This finding shows that, at this time, the proportion of population in the suburbs is increasing, the population distribution is more scattered and the density is smaller.

The previously mentioned parameters are applied to assess the risk of the epidemic. We employ the community infection to fit the model. The community infection data is obtained from the official notification data. After the nuclear density estimation, the data are utilized as the dependent variable field of the model. Because we want to give each grid a certain attribute, we apply the value extraction point tool. If we use the appropriate density of points for spatial sampling, we can describe the spatial distributions of different elements from a certain dimension. Each grid is regarded as an independent cell for data sampling within the grid.

B. GeoDetector RESULT

It can be seen from Table 2 that the single factor detection results that the population density after closure of the city is the most important factor in determining the risk of community infection. The second important factor is the traffic density and daily population density, which show that its reasonable epidemic prevention effect can reduce the areas with a dense floating population and prevent long-term exposure in the population. The factor with the smallest impact is the spatial distance from the market. The second least important factor is the space distance from the fever clinic, which is not as strongly related to the probability of infection as we assume, i.e., even if the patients are located far from hospitals, they can receive better treatment.

We interact two by two with different elements; the detection results are shown in Table 3. After the interaction between the distance index of the market and the population density index after closure of the city, the best explanatory effect for the risk of community infection is achieved, with a *q* value of 0.57. After the interaction between any two indexes, the explanatory effect of infection is greater than that of a single factor.

Ecological detection was carried out using the *F* test with a confidence of 0.05 in Table 4. “Y” means a significant difference, and “N” means no significant difference. After closure of the city, the population density index is significantly different from all the other indicators, which is the most important factor for measuring the risk of community infection. As a comparison, there is no significant difference

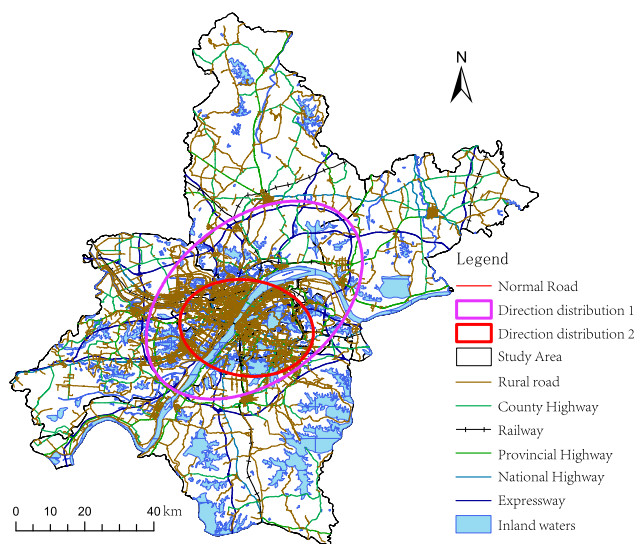


FIGURE 7. The discussion of direction.

between the population density index and the densities of hotels, markets, restaurants and bus stops, which shows that the distribution of these POI is reasonable and consistent with the population distribution and can better serve the vast majority of the population at ordinary times.

C. DECISION TREE

The decision tree training results are shown in the following Figure.8. When the depth of a decision tree is 10, the score in the test set is the highest. As the depth of the tree increases, the model does not improve. When the Gini index of a node is less than or equal to a certain threshold value, the nodes of the decision tree do not need to be further split; otherwise, new partition rules need to be generated [46]. The training concentration accuracy score is 0.96, and the test concentration score is 0.77.

The sub-tree judgment standard with the largest number of sub-trees is that the population density level of the closed city is less than or equal to 3 (a total of 10 levels) and the market distance is greater than or equal to 2. This kind of classified grid has the largest number, that is, the population density of the closed city is not large and there is a certain buffer distance from the market. Thus, the risk of this kind of area is low, and there are more indexes of the population density of the closed city, The level is less than or equal to 3,

TABLE 3. Interaction detector.

	Fever clinic distance	Population reduction	Daily population	Closed city population density	Market distance	Hotel density	Quotient hyper density	Catering density	Traffic density
Fever clinic distance	0.18								
Population reduction	0.36	0.19							
Daily population	0.42	0.44	0.31						
Closed city population density	0.53	0.54	0.51	0.48					
Market distance	0.29	0.4	0.48	0.57	0.16				
Hotel density	0.34	0.35	0.4	0.54	0.4	0.26			
Quotient hyper density	0.36	0.36	0.42	0.54	0.38	0.33	0.27		
Catering density	0.33	0.4	0.42	0.55	0.4	0.37	0.34	0.29	
Traffic density	0.46	0.45	0.43	0.54	0.45	0.49	0.47	0.45	0.33

TABLE 4. Ecological detector.

	Fever clinic distance	Population reduction	Daily population	Closed city population density	Market distance	Hotel density	Quotient hyper density	Catering density	Traffic density
Fever clinic distance	N								
Population reduction	Y	Y							
Daily population	Y	Y	Y						
Closed city population density	Y	Y	Y	Y					
Market distance	N	N	Y	Y					
Hotel density	Y	N	N	Y	Y				
Quotient hyper density	Y	Y	N	Y	Y	N			
Catering density	Y	Y	N	Y	Y	N	N		
Traffic density	Y	Y	N	Y	Y	Y	N	N	

the market distance is less than or equal to 1, and the fever clinic distance index level is less than or equal to 3, which indicates that it is relatively close to the market but it is not far from the fever hospital, which can achieve early detection and early treatment, and that this kind of community is also relatively safe. The population density index level is equal to 4, the population reduction index level is less than or equal to 4, and the minimum market distance is 2, that is, the population density level after closure of the city is medium, which is not very close to the market, and the area with a large attenuation (population reduction statistics are negative), is also relatively safe, such as universities, government agencies, office CBD and other areas.

D. RISK MAP

Using the evaluation model generated by the decision tree, the risks of 726 grids of the urban area are predicted; the

prediction results are shown in the following Figure.9. The Qiaokou district and Jiangnan district, have a higher density and smaller area, and the number of people infected ranks second and third, respectively [47], [48]. According to statistics, 397 grid areas have the lowest risk level, 63 grid areas are divided into the second risk level, 93 grid areas are divided into the third risk level, the highest risk level grid areas have 130, followed by 43 grid areas are divided into the fourth risk level. Using Anjuke and Google street view and other channels to query, there is a common feature among these communities, that is, many of them are old urban areas or old communities that have not yet been transformed by villages in the city. This discovery warrants the attention of the management department.

The risk map applies the grid discrete kernel density analysis results of community infection as the dependent variable of the decision tree model and extracts a part of it for training

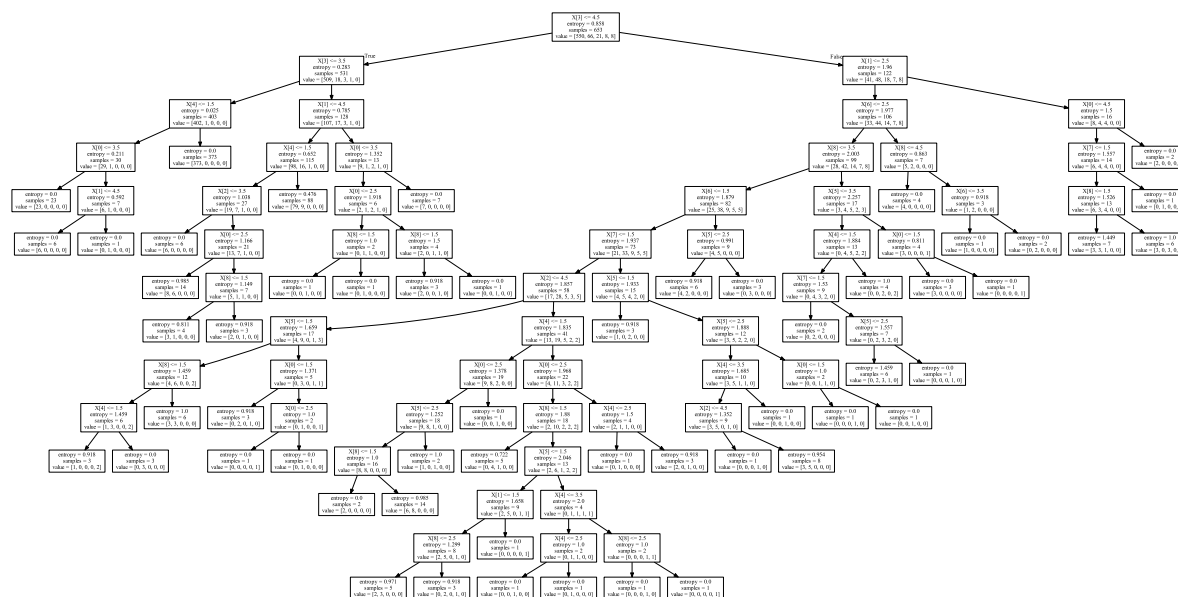


FIGURE 8. The decision tree model result.

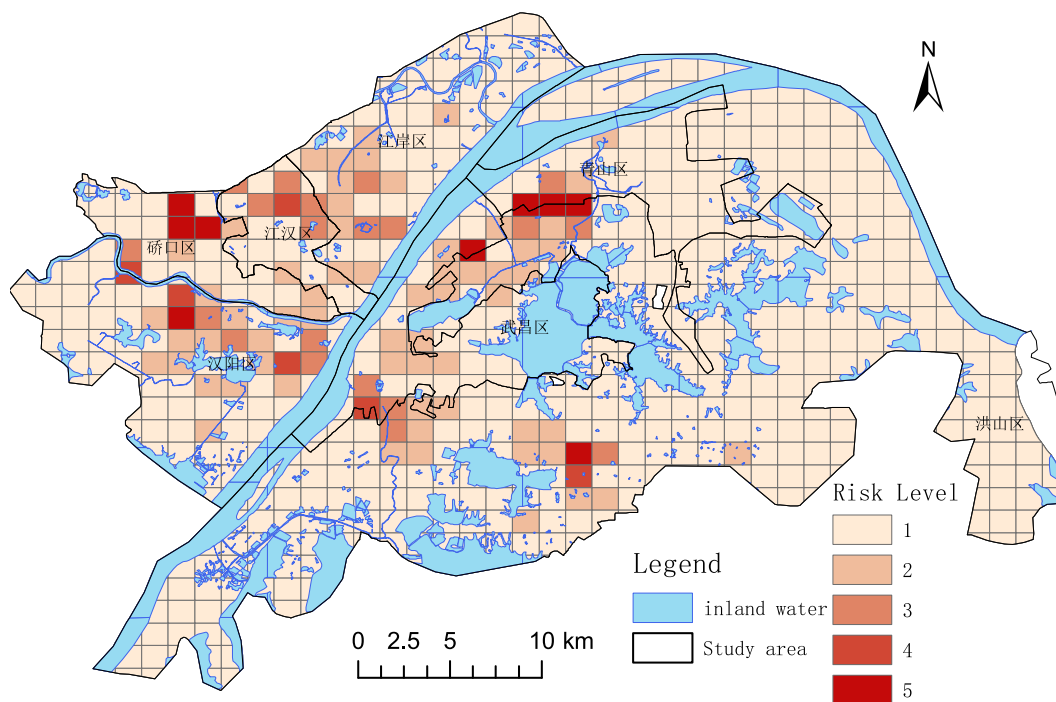


FIGURE 9. The risk map of study area.

in the decision tree model. The decision tree model obtains our judgment criteria for epidemic infection. Using the criteria, we can predict the epidemic situation, and people in the district can conduct a self-assessment of the risk situation in their surroundings. We use the decision-making model to classify all the grids. If the risk level (results of infection data interpolation and dispersion) is reduced to three, the accuracy

of the model will increase correspondingly but the effectiveness of the risk map will be greatly reduced.

VI. DISCUSSION AND CONCLUSION

A. DISCUSSION

In this paper, we divide the study area into 726 grids, use the geographical detector to obtain the major factors that

determine the magnitude of regional infection of COVID-19 by using the attribute difference and spatial difference between different grids. We also employ the machine learning method of the decision tree to calculate the exposure risk of infection and we construct the community risk map for fighting the epidemic.

The most important factor that determined the degree of regional infection is the population density after closure of the city, followed by the traffic density and daily population density. From this point of view, a small area with a small daily population and a small population after the closure that is locate far from the traffic hub is relatively safe. The measures of the lockdown are effective, necessary and timely. To verify the results of the geo-detector and generate the risk criteria from the perspective of the decision tree, three main criteria are generated: The population density after closure of the city is small and there is a certain buffer distance from the market; It is close to the market but not far from the fever hospital; After closure of the city, the population density level is medium, which is not very close to that of the market. The result of the decision tree shows agreement with that of the GeoDetector, which also confirms the correctness of the model.

B. CONCLUSION

According to relevant information, such as open community infection data, patient trajectory data, cell phone signaling, and population density or interest points, and referring to the ideas of this paper, the city can be divided into high-risk areas and low-risk areas with finer spatial granularity according to the actual situation. Because there is no specific medicine or reliable vaccine for the new coronavirus epidemic, the isolation and short-term “closure” of patients remain effective means. After the calculation with reference to the method in this paper, the communities with high risk can appropriately extend the closure time, while the communities with low risk can appropriately shorten or cancel the closure, resume work and production, and then achieve accurate prevention and control.

The limitations of this paper are that the acquired infection data are not comprehensive, the generalization ability of the model needs to be improved, and the resolution of the fine-grained grid needs further adjustment. The evaluation indicators can also be reconstructed to take into account the time factor.

REFERENCES

- [1] R. J. Coker, B. M. Hunter, J. W. Rudge, M. Liverani, and P. Hanvoravongchai, “Emerging infectious diseases in southeast Asia: Regional challenges to control,” *Lancet*, vol. 377, no. 9765, pp. 599–609, Feb. 2011.
- [2] E. Dong, H. Du, and L. Gardner, “An interactive Web-based dashboard to track COVID-19 in real time,” *Lancet Infectious Diseases*, vol. 20, no. 5, pp. 533–534, May 2020.
- [3] N. Al-Rousan and H. Al-Najjar, “Nowcasting and forecasting the spreading of novel coronavirus 2019-nCoV and its association with weather variables in 30 Chinese provinces: A case study,” *Tech. Rep.*, 2020.
- [4] X. Zhu, A. Zhang, S. Xu, P. Jia, X. Tan, J. Tian, T. Wei, Z. Quan, and J. Yu, “Spatially explicit modeling of 2019-nCoV epidemic trend based on mobile phone data in Mainland China,” *medRxiv*, 2020.
- [5] Q. Liu, Z. Liu, D. Li, Z. Gao, J. Zhu, J. Yang, and Q. Wang, “Assessing the tendency of 2019-nCoV (COVID-19) outbreak in China,” *medRxiv*, 2020.
- [6] X. Li, X. Zhao, and Y. Sun, “The lockdown of Hubei province causing different transmission dynamics of the novel coronavirus (2019-nCoV) in Wuhan and Beijing,” *medRxiv*, 2020.
- [7] X. Fu, Q. Ying, T. Zeng, T. Long, and Y. Wang, “Simulating and forecasting the cumulative confirmed cases of SARS-CoV-2 in China by Boltzmann function-based regression analyses,” *J. Infection*, vol. 80, no. 5, pp. 578–606, May 2020.
- [8] L. Danon, E. Brooks-Pollock, M. Bailey, and M. J. Keeling, “A spatial model of CoVID-19 transmission in England and Wales: Early spread and peak timing,” *medRxiv*, 2020.
- [9] B. Tang, F. Xia, S. Tang, N. L. Bragazzi, Q. Li, X. Sun, J. Liang, Y. Xiao, and J. Wu, “The evolution of quarantined and suspected cases determines the final trend of the 2019-nCoV epidemics based on multi-source data analyses,” *Tech. Rep.*, 2020.
- [10] K. Wu, J. Zheng, and J. Chen, “Utilize state transition matrix model to predict the novel corona virus infection peak and patient distribution,” *Tech. Rep.*, 2020.
- [11] S. Lai, I. Bogoch, N. W. Ruktanonchai, A. Watts, Y. Li, J. Yu, X. Lv, W. Yang, H. Yu, and K. Khan, “Assessing spread risk of Wuhan novel coronavirus within and beyond China, January-April 2020: A travel network-based modelling study,” *medRxiv*, 2020.
- [12] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, “Data-based analysis, modelling and forecasting of the novel coronavirus (2019-nCoV) outbreak,” *medRxiv*, 2020.
- [13] X. Li, X. Zhao, and Y. Sun, “The lockdown of Hubei province causing different transmission dynamics of the novel coronavirus (2019-nCoV) in Wuhan and Beijing,” *medRxiv*, 2020.
- [14] L. Hébert-Dufresne, B. M. Althouse, S. V. Scarpino, and A. Allard, “Beyond R_0 : The importance of contact tracing when predicting epidemics,” 2020, *arXiv:2002.04004*. [Online]. Available: <https://arxiv.org/abs/2002.04004>
- [15] S. Zhang, M. Diao, W. Yu, L. Pei, Z. Lin, and D. Chen, “Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the diamond princess cruise ship: A data-driven analysis,” *Int. J. Infectious Diseases*, vol. 93, pp. 201–204, Apr. 2020.
- [16] P.-C. Lai, C. B. Chow, H. T. Wong, K. H. Kwong, Y. W. Kwan, S. H. Liu, W. K. Tong, W. K. Cheung, and W. L. Wong, “An early warning system for detecting H1N1 disease outbreak—A spatio-temporal approach,” *Int. J. Geograph. Inf. Sci.*, vol. 29, no. 7, pp. 1251–1268, Jul. 2015.
- [17] E. Delmelle, C. Dony, I. Casas, M. Jia, and W. Tang, “Visualizing the impact of space-time uncertainties on dengue fever patterns,” *Int. J. Geograph. Inf. Sci.*, vol. 28, no. 5, pp. 1107–1127, May 2014.
- [18] G. Gao, Y. Feng, X. Tong, Z. Lei, S. Chen, and S. Zhai, “Modeling urban growth using spatially heterogeneous cellular automata models: Comparison of spatial lag, spatial error and GWR,” *Comput., Environ. Urban Syst.*, vol. 81, no. 3, pp. 10473–101459, 2020.
- [19] S. Yang, D. He, J. Luo, W. Chen, X. Yang, M. Wei, X. Kong, Y. Li, X. Feng, and Z. Zeng, “Simulation of HIV/AIDS distribution using GIS based cellular automata model,” *Biomed. Res.*, vol. 28, no. 9, pp. 4053–4057, 2017.
- [20] P. Magal, A. Noussair, G. Webb, and Y. Wu, “Modeling epidemic outbreaks in geographical regions: Seasonal influenza in Puerto Rico,” *Discrete Continuous Dyn. Syst.-S.*, vol. 25, no. 6, pp. 2185–2202, 2018.
- [21] R. Sandhu, H. K. Gill, and S. K. Sood, “Smart monitoring and controlling of pandemic influenza a (H1N1) using social network analysis and cloud computing,” *J. Comput. Sci.*, vol. 12, pp. 11–22, Jan. 2016.
- [22] L. Li, Q. Zhang, X. Wang, J. Zhang, T. Wang, T.-L. Gao, W. Duan, K. K.-F. Tsoi, and F.-Y. Wang, “Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo,” *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 2, pp. 556–562, Apr. 2020.
- [23] J. Besag and P. J. Green, “Spatial statistics and Bayesian computation,” *J. Roy. Stat. Soc., B, Methodol.*, vol. 55, no. 1, pp. 25–37, 1993.
- [24] F. Weicheng, L. Yi, and W. Wen-Guo, “Triangular framework and ‘4+1’ methodology for public security science and technology,” *Sci. Technol. Rev.*, vol. 27, no. 6, p. 3, 2009.
- [25] Y. Dong, H. Liu, and T. Zheng, “Does the connectivity of urban public green space promote its use? An empirical study of Wuhan,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 1, pp. 297–316, 2020.

- [26] J. A. Backer, D. Klinkenberg, and J. Wallinga, "Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020," *Eurosurveillance*, vol. 25, no. 5, pp. 10–15, Feb. 2020.
- [27] B. Pirouz, S. S. Haghshenas, S. S. Haghshenas, and P. Piro, "Investigating a serious challenge in the sustainable development process: Analysis of confirmed cases of COVID-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis," *Sustainability*, vol. 12, no. 6, pp. 2427–2448, 2020.
- [28] J. Wang, X. Li, G. Christakos, Y. Liao, T. Zhang, X. Gu, and X. Zheng, "Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China," *Int. J. Geograph. Inf. Sci.*, vol. 24, no. 1, pp. 107–127, Jan. 2010.
- [29] Y. Gao, G. Yang, and Q. Xie, "Spatial-temporal evolution and driving factors of green building development in China," *Sustainability*, vol. 12, no. 7, pp. 2773–2794, 2020.
- [30] J. Wang, M. Hu, F. Zhang, and B. Gao, "Influential factors detection for surface water quality with geographical detectors in China," *Stochastic Environ. Res. Risk Assessment*, vol. 32, no. 9, pp. 2633–2645, Sep. 2018.
- [31] R. Wu, J. Zhang, Y. Bao, and F. Zhang, "Geographical detector model for influencing factors of industrial sector carbon dioxide emissions in Inner Mongolia, China," *Sustainability*, vol. 8, no. 2, pp. 149–161, 2016.
- [32] X. Zhang and Z. Gong, "Spatiotemporal characteristics of urban air quality in China and geographic detection of their determinants," *J. Geograph. Sci.*, vol. 28, no. 5, pp. 563–578, May 2018.
- [33] J.-F. Wang and Y. Hu, "Environmental health risk detection with GeogDetector," *Environ. Model. Softw.*, vol. 33, pp. 114–115, Jul. 2012.
- [34] R. Yang, Q. Xu, and H. Long, "Spatial distribution characteristics and optimized reconstruction analysis of China's rural settlements during the process of rapid urbanization," *J. Rural Stud.*, vol. 47, pp. 413–424, Oct. 2016.
- [35] R. Yang, Y. Liu, H. Long, and L. Qiao, "Spatio-temporal characteristics of rural settlements and land use in the Bohai Rim of China," *J. Geograph. Sci.*, vol. 25, no. 5, pp. 559–572, May 2015.
- [36] H. Bai, Y. Ge, J. Wang, D. Li, Y. Liao, and X. Zheng, "A method for extracting rules from spatial data based on rough fuzzy sets," *Knowl.-Based Syst.*, vol. 57, pp. 28–40, Feb. 2014.
- [37] C. Wu, X. Ye, Q. Du, and P. Luo, "Spatial effects of accessibility to parks on housing prices in Shenzhen, China," *Habitat Int.*, vol. 63, pp. 45–54, May 2017.
- [38] M. Li, X. Shi, X. Li, W. Ma, J. He, and T. Liu, "Sensitivity of disease cluster detection to spatial scales: An analysis with the spatial scan statistic method," *Int. J. Geograph. Inf. Sci.*, vol. 33, no. 11, pp. 2125–2152, Nov. 2019.
- [39] G. Langran, "A review of temporal database research and its use in GIS applications," *Int. J. Geograph. Inf. Syst.*, vol. 3, no. 3, pp. 215–232, Jul. 1989.
- [40] J. Li, S. Chen, W. Chen, G. Andrienko, and N. Andrienko, "Semantics-space-time cube: A conceptual framework for systematic analysis of texts in space and time," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 4, pp. 1789–1806, Apr. 2020.
- [41] S. Shekhar, M. Coyle, B. Goyal, D.-R. Liu, and S. Sarkar, "Data models in geographic information systems," *Commun. ACM*, vol. 40, no. 4, pp. 103–111, 1997.
- [42] C. Huang *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.
- [43] S. Lifan, Z. Chun, L. He, W. Ye, and W. Zijia, "Relationship between urban rail transit commuting and jobs-housing balance: An empirical analysis from Beijing based on big data methods," *Progress Geogr.*, vol. 38, no. 6, pp. 791–806, 2019.
- [44] Y. Chen, X. Liu, X. Li, X. Liu, Y. Yao, G. Hu, X. Xu, and F. Pei, "Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method," *Landscape Urban Planning*, vol. 160, pp. 48–60, Apr. 2017.
- [45] D. W. Lefever, "Measuring geographic concentration by means of the standard deviational ellipse," *Amer. J. Sociol.*, vol. 32, no. 1, pp. 88–94, Jul. 1926.
- [46] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote Sens. Environ.*, vol. 61, no. 3, pp. 399–409, Sep. 1997.
- [47] G. Colletaz, C. Hurlin, and C. Pérignon, "The risk map: A new tool for validating risk models," *J. Banking Finance*, vol. 37, no. 10, pp. 3843–3854, Oct. 2013.
- [48] A. M. Fonseca, J. L. Biscaya, J. Aires-de-Sousa, and A. M. Lobo, "Geographical classification of crude oils by Kohonen self-organizing maps," *Anal. Chim. Acta*, vol. 556, no. 2, pp. 374–382, Jan. 2006.



YAN ZHANG was born in Wocheng, Linying, Henan, China, in 1997. He received the B.S. and M.S. degrees in geographic information system engineering from the University of Wuhan University. He is currently pursuing with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing. His research interests are natural language processing, complex networks, and spatial data analysis.



YINGBING LI was born in Hubei, China, in 1972. He received the B.S. and M.S. degrees in engineering survey from the Wuhan Technical University of Surveying and Mapping, China, in 2000, and the Ph.D. degree in geodesy and geomatics from Wuhan University, China, in 2003. Since 2001, he has been teaching with Wuhan University. In 2008 and 2013, he visited Ohio State University twice as a Visiting Scholar. His main research interests include typical natural disaster evolution process analysis, emergency response scenario deduction, and spatial big data analysis.



BO YANG is currently pursuing the bachelor's degree with the School of Public Administration, Zhongnan University of Economics and Law, Wuhan, China. His research interests include urban and regional development and the applications of geographical information systems.



XIANG ZHENG was born in Zhumadian, Henan, China, in 1996. She received the B.A. degree in information management from Wuhan University, Wuhan, China, in 2018, where she is currently pursuing the master's degree with the School of Information Management. Her research interests are digital information resource management and service.



MIN CHEN is currently pursuing the bachelor's degree in cartography and geographic information science with the School of Geodesy and Geometrics, Wuhan University. Her main research interests include cascading disaster modeling and analysis from the geographic perspective.

...