Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Article

A dynamic and classifier-based model for SARS-CoV-2 Omicron variant spillover risk assessment in China

Hongjie Wei[a,1], Jia Rui[a,b,c,1], Yunkang Zhao[a], Huimin Qu[a], Jing Wang[a], Guzainuer Abudurusuli[a], Qiuping Chen[a,b,c], Zeyu Zhao[a,b,c], Wentao Song[a], Yao Wang[a], Roger Frutos[b,*], Tianmu Chen[a,*]

[a] State Key Laboratory of Molecular Vaccinology and Molecular Diagnostics, School of Public Health, Xiamen University, Xiamen 361000, China
[b] CIRAD, Intertryp, Montpellier 34398, France
[c] Université de Montpellier, Montpellier 34090, France

ARTICLE INFO

Article history:
Received 26 September 2022
Received in revised form 27 March 2023
Accepted 29 March 2023
Available online xxx

Keywords:
SARS-CoV-2 Omicron
Spatial spillover risk
Dynamics model
A classifier model
Influencing factors

ABSTRACT

The coronavirus disease 2019 (COVID-19) continues to have a huge impact on health care and economic systems around the world. The first question to ponder is to understand the flow of COVID-19 in the spatial and temporal dimensions. We collected 7 Omicron clusters outbreaks in China since the outbreak of COVID-19 as of August 2022, selected outbreak cases from different provinces and cities, and collected variable indicators that affect spillover outcomes, such as distance, migration index, PHSM index, daily reported cases number and so on. First, variables influencing spillover outcome events were assessed and analyzed retrospectively by constructing an infectious disease dynamics model and a classifier model, and secondly, the association between explanatory variables and spillover outcome events was constructed by fitting a logistics function. This study incorporates 7 influencing factors and classifies the spillover risk level into 3 levels. If different outbreak sites could be classified into different levels of spillover, it may reduce the pressure of epidemic prevention in some districts due to the lack of a uniform standard, which might be more conducive to achieving the goal of "dynamic zero".

1. Introduction

In the two and a half years since the WHO declared the novel coronavirus pneumonia epidemic a global pandemic in Geneva on 11 March 2020 [1], it has had an unusually large impact on the health care and economies of almost every country in the world. And after the WHO declared Omicron a variant of concern on 26 November 2021 [2], it recommended that countries lift or relax travel restrictions, saying that travel restrictions cannot stop the spread of Omicron [3]. The measures taken internationally vary from country to country. However, most developed countries have already announced the removal of vaccination restrictions and free mass COVID-19 testing, such as Denmark, which announced the removal of all vaccination restrictions on 1 January 2022, and the UK, which announced on 1 April 2022 that it would no longer offer free testing for the general public [4]. Estimates published by the Office for National Statistics (ONS) show that for the second consecutive week since the launch of the "Living with COVID-19″ program [5], the number of confirmed cases of COVID-19 in the UK is approaching 5 million in a single week, continuing at the highest level ever recorded, in most parts of the country. In most parts of the country, an average of 1 in 13 people are infected with SARS-CoV-2. In China, according to incomplete statistics, the seven Omicron clusters since March 2022 have generated a total of 676,482 cases and 929 spillover events.

Under the guidance of China's "dynamic zero" and the ninth edition of the prevention and control guidelines, how to prevent the spillover of infected persons from the COVID-19 outbreak in the context of the Omicron variant pandemic in 2022 has become a priority issue in China. In this paper, we first explored the main factors influencing spillover outcome events by building a classifier model to incorporate factors affecting disease spillover, training and adjusting the model accuracy, and setting three different spillover risk levels. This was followed by fitting functions to further demonstrate the extent to which the main factors influencing spillover outcome events contributed to spillover outcomes.

There is a paucity of literature addressing disease spillover, with one study using the COVID-19 Community Index to model the risk rating of COVID-19 spillover by county and region [6], but without a weighted comparison of risk factors for outbreak spillover to analyze its main influences. There appears to be a gap in current research in understanding

[*] Corresponding authors.
E-mail addresses: frutossmt@gmail.com (R. Frutos), 13698665@qq.com (T. Chen).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.fmre.2023.03.014

Please cite this article as: H. Wei, J. Rui, Y. Zhao et al., Dynamic and classifier-based model SARS-CoV-2 Omicron spillover risk assessment in China, Fundamental Research, https://doi.org/10.1016/j.fmre.2023.03.014

how the specifics of infected spillover under COVID-19 pandemic evolve over time and how it spreads to other regions. Therefore, this study hopes to fit a classifier model and infectious disease dynamics model with a classifier model in machine learning that can be more accurate in measuring the severity of spillover within and between regions and predicting the number of infections.

Our findings suggest that spatial distance will be a key factor influencing spillover outcome events in the context of an Omicron variant pandemic, a finding that implies that the likelihood of reduced spillover risk decreases with increasing distance from the site of occurrence, and that spillovers tend to occur in the early stages of an outbreak.

2. Material and methods

2.1. Data collection and variables definition

All case data, mid-risk and high-risk area data sources are from National Health Commission of the People's Republic of China. The migration index is from the official website of Baidu migration (https://qianxi.baidu.com/).

The definition and rationale for the variables included in the classifier model in this study are as follows: 1) distance: calculate the distance between two points on the sphere using the haversine formula, the latitude and longitude of the location are derived from the amap (https://ditu.amap.com/); 2) R_t: real-time reproduction number; 3) cases: number of new cases reported per day; 4) mid-risk and high-risk: risk level definition originated from National Health Commission of the People's Republic of China; 5) migration index: from the number of migrations announced by Baidu Migration; 6) same area: based on seven regions in China based on geographic regions and Hong Kong, Macau and Taiwan, a total of eight areas.

2.2. Dynamic model structure

We considered pre-symptomatic infections based on the basic Susceptible- Exposed- Symptomatic- Asymptomatic- Quarantined- Recovered/Removed (SEIAQR) deterministic model according to the previous research [7–10]. In our model, the whole population were first divided into two groups, completed booster vaccination population and uncompleted booster vaccination population. Furthermore, individuals of each group were divided into six categories: Susceptible (S), Exposed (E), Symptomatic (I_s), Pre-symptomatic (I_p), Asymptomatic (A), Quarantined (Q) and Removed (R) including recovered. The equations of the model were

$$\frac{d}{dt(S_1)} = -\beta_{11} * S_1 * \left(I_{S1} + \kappa * I_{P1} + \kappa * A_1\right) - \beta_{21} * S_1 * \left(I_{S2} + \kappa * I_{P2} + \kappa * A_2\right)$$

$$\frac{d}{dt(E_1)} = \beta_{11} * S_1 * \left(I_{S1} + \kappa * I_{P1} + \kappa * A_1\right) + \beta_{21} * S_1 * \left(I_{S2} + \kappa * I_{P2} + \kappa * A_2\right) - (1-p) * \omega * E_1 - p * \omega_2 * E_1$$

$$\frac{d}{dt(I_{P1})} = (1-p) * \omega * E_1 - \omega_1 * I_{P1} - h * I_{P1}$$

$$\frac{d}{dt(I_{S1})} = \omega_1 * I_{P1} - \gamma * I_{S1} - h * I_{S1}$$

$$\frac{d}{dt(A_1)} = p * \omega_2 * E_1 - \gamma * A_1 - h * A_1$$

$$\frac{d}{dt(R_1)} = \gamma * I_{S1} + \gamma * A_1$$

$$\frac{d}{dt(R_2)} = \gamma * Q_1$$

$$\frac{d}{dt(Q_1)} = h * \left(I_{S1} + I_{P1} + A_1\right) - y * Q_1$$

$$\frac{d}{dt(S_2)} = -\beta_{22} * S_2 * \left(I_{S2} + \kappa * I_{P2} + \kappa * A_2\right) - \beta_{12} * S_2 * \left(I_{S1} + \kappa * I_{P1} + \kappa * A_1\right)$$

$$\frac{d}{dt(E_2)} = \beta_{22} * S_2 * \left(I_{S2} + \kappa * I_{P2} + \kappa * A_2\right) + \beta_{12} * S_2 * \left(I_{S1} + \kappa * I_{P1} + \kappa * A_1\right) - (1-p) * \omega * E_2 - p * \omega_2 * E_2$$

$$\frac{d}{dt(I_{P2})} = (1-p) * \omega * E_2 - \omega_1 * I_{P2} - h * I_{P2}$$

$$\frac{d}{dt(I_{S2})} = \omega_1 * I_{P2} - \gamma * I_{S2} - h * I_{S2}$$

$$\frac{d}{dt(A_2)} = p * \omega_2 * E_2 - \gamma * A_2 - h * A_2$$

$$\frac{d}{dt(R_3)} = \gamma * I_{S2} + \gamma * A_2$$

$$\frac{d}{dt(R_4)} = \gamma * Q_2$$

$$\frac{d}{dt(Q_2)} = h * \left(I_{S2} + I_{P2} + A_2\right) - y * Q_2$$

$$N = S_1 + E_1 + I_{S1} + I_{P1} + A_1 + R_1 + R_2 + Q_1 + S_2 + E_2 + I_{S2} + I_{P2} + A_2 + R_3 + R_4 + Q_2$$

This extended SEIAQR model follows some basic assumptions, including that population is homogeneous and well-mixed interactions without influence by social behavior, age and work. And we add some assumptions to our study:

(1) Susceptible population would be infected with a transmission relative rate of β by contact with pre-symptomatic/ symptomatic/ asymptomatic infections, and their transmission relative rate is the same.

(2) The incubation period of symptomatic infections was $1/\omega + 1/\omega''$, the latent period of an asymptomatic person was $1/\omega'$.

(3) Parameter p $(0 \le p \le 1)$ gave the proportion of individuals who had asymptomatic infections.

(4) Symptomatic infections are communicable in $1/\omega''$ days before developed symptoms.

(5) Individuals in categories I_s and A were transferred into category R after an infectious period of $1/\gamma'$ and$1/\gamma$, respectively.

(6) Case fatality rate (CFR) was 0 and was not simulated in the model because Omicron variant has low CFR.

(7) We assumed that the infectivity and susceptibility would be reduced after vaccination. VEI and VES due to being fully vaccinated were denoted as (1 - x) and (1 - y), respectively.

2.3. Parameter estimation approach

Three parameters were estimated based on real data, which are the total population, asymptomatic infection rate and the coverage rate of COVID-19 booster vaccination. In this study, several parameters were adopted to develop the model, and the description, value, and source are listed in Table S1.

(1) According to the statistical year in 2021, the total population of the seven districts is 21,893,000, 15,618,300, 5464,087, 6255,000, 8782,285, 24,870,000, 10,081,200, respectively. The number of initial infections (I), including symptomatic and asymptomatic, is obtained from the actual reported data, and the initial values I_0 of the seven districts are 1, 12, 3, 2, 10, 1, 1 respectively. The initial values of E and R were set to 0.

(2) The coverage rate of COVID-19 booster vaccination in seven districts is 71.07%, 1.53%, 0, 0 and 0, 42.74%, 30.85% respectively (districts with 0 booster vaccination are assumed to have a 30% booster vaccination rate for all districts with vaccination rates below 30% in the model due to the lack of publicly available data).

(3) The parameter κ refers to the relative transmissibility rate of asymptomatic to symptomatic individuals. Refer to the previous research. κ is set to 1 in this study.

(4) Since reported asymptomatic patients are far more than infected in Shanghai city, the p in Shanghai is set to 0.8 by assumption, while the p in other districts are set to 0.31 according to the previous research.

(5) As of August 23, 2022, no death case was reported in the report data of the seven districts, so this study did not incorporate the case fatality rate (f) in the model.

(6) At present, there were few researches on the incubation period of the symptomatic infections(ω) and latent period of the asymptomatic(ω') of Omicron Variant, so we made assumptions based on the existing literatures of Omicron BA.1, and we assumed that the latent period is the same as the incubation period, which is similar with the previous study in Gauteng and KwaZulu-Natal. According to the outbreak in Norway, the median incubation period was 3 days (interquartile range: 3–4); it was 4.2 days (range, 2–8 days) according to other publicly reported data from Korea; the incubation found by a survey in South Korea median incubation period was 3–4 days; it was 3 days (interquartile range:1–4 days) in the study of a northern region of Spain; We also refer to another study in Japan, mean incubation periods were 3.7 (95% credible interval (CI) 3.4–4.0) and 5.0 (95% CI 4.5–5.6) days for Delta and non-Delta cases, respectively. According to CDC Newsroom report in December 27, 2021, the $1/\omega''$ was 1–2 days.

(7) In this study, the infectious period was set to 4.5 days ($\gamma = \gamma$'=0.22) by our previous research about Delta and CDC Newsroom report in December 27, 2021.

2.4. Statistical analysis

Real-time reproduction number (R_t) was performed by EpiEstim (version 2.2. 4) in R software (version 4.1.2). Other statistical analysis was conducted by using Python (version 3.8.8). Univariate logistic regression analysis was performed to screen related risk factors of Spillover risk factors. We found the point on the ROC curve that is closest (i.e., the shortest distance) to the perfect model (with 100% sensitivity and 100% specificity), which was associated with the upper left corner of the plot. The discrimination ability of the model was evaluated by using receiver operator characteristic (ROC) curve [11] analysis. The AUC > 0.5 indicated better predictive values, the closer the AUC to 1, the better the model performance. The specific process is shown in Fig. 1.

The decision tree model. A tree structure composed of root node, branch node and leaf node, which reflected the mapping relationship between features and tags [12].

Random forest [13]. An ensemble learning method for classification, regression, and other tasks that operates by constructing a large number of decision trees at training time. We used scikit-learn (version 1.1.3) in python for training and prediction of the model. The criterion of function in this model that measures the quality of a split is gini, and we use RandomizedSearch and GridSearchCV to find the optimal parameters of the model, considering and solving the impact of sample imbalance problem on the model.

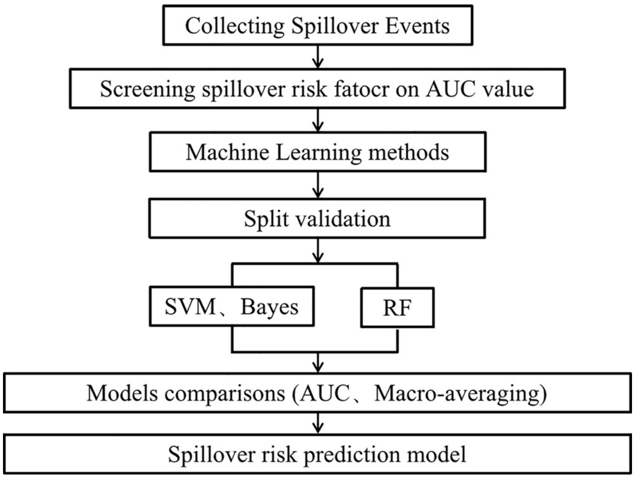
Collecting Spillover Events
Screening spillover risk fatocr on AUC value
Machine Learning methods
Split validation
SVM、Bayes
RF
Models comparisons (AUC、Macro-averaging)
Spillover risk prediction model

Fig. 1. Model selection and comparison flowchart.

Naive Bayes classifier. Bayes' Rule answers [14] the question "based on the predictors that we have observed. In this study, we use gaussian naive bayes, sample imbalance problem solved using imblearn.over sampling in python. The likelihood of the features is assumed to be Gaussian as follow:

$$P(\chi_i \mid \gamma) = \frac{1}{\sqrt{2\pi\sigma_\gamma^2}} \exp\left(-\frac{(\chi_i - \mu_\gamma)^2}{2\sigma_\gamma^2}\right)$$

Support Vector Machines (SVM). The basic idea of SVM [15] learning is to solve the separated hyperplane that correctly partitions the training data set and maximizes the geometric separation. We use svm in scikit-learn (version 1.1.3) in python in this study.

Fitting function

We use SMOTE in imblearn.over_sampling in python (version 3.8.9) to resample the unbalanced data and fit the logit function using statsmodels.

All model code can be posted on github when the article is received.

3. Result and discussion

3.1. Epidemiological description

3.1.1. Descriptive analysis of aggregated outbreaks

In this study, a total of seven Omircon outbreaks were collected, as shown in Fig. 2, and the list from top to bottom are Beijing, Tianjin, Langfang, Nanchang, Quanzhou, Shanghai city and Hannan province. The remaining six cities were all BA.2 among the seven aggregated outbreaks counted, the durations were 91 days, 54 days, 25 days, 44 days, 31 days, and 22 days, respectively. The longest duration of the outbreak was 93 days in Shanghai, and the highest cumulative number of reported cases was 649,354, followed by 19,266 in Hainan (data as of August 23, 2022), and the cumulative number of reported cases during the outbreak in the remaining five cities was 2,283, 834, 3,409, 1,133, and 3,175, respectively. Based on previous research, this study simulates the outbreak curve of 7 areas by establishing the SEIAQR model, as shown in Fig. 2b, from left to right, Beijing, Tianjin, Langfang, Nanchang, Quanzhou, Shanghai city and Hainan province. The bars in the figure are the actual number of cases in the area, while the red curve is the fitted curve, which is the number of new cases per day fitted by the SEIAQR model. Since the reported case data are often lagging or unstable, the risk can be subsequently determined by importing the fitted data into the classifier model to increase the stability and real-world fit of the model. The peak time of the outbreak in the seven areas counted was 53 days, 16 days, 12 days, 8 days, 6 days, 16 days, and 44 days, with Beijing taking the longest time to peak at 53 days, followed by Shanghai

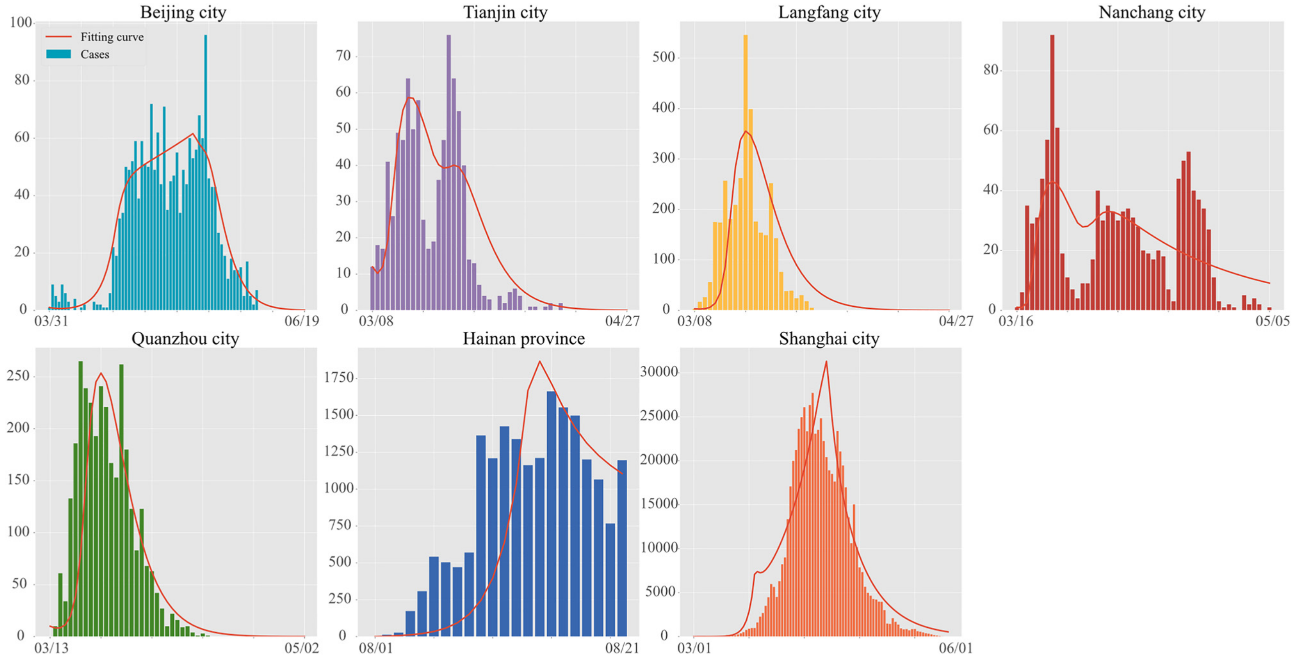

Fig. 2. COVID-19 outbreak curves in seven districts. From left to right, the first district is Beijing and the last district is Shanghai.

at 44 days. The peak number of cases in the seven districts was 96, 76, 546, 92, 265, 1663, and the highest peak number of cases was 27,719 in Shanghai.

Among the seven outbreaks counted, the COVID-19 outbreak in Shanghai was long in duration and large in scale, but we believe that Shanghai took a unique approach [16] to fighting the earliest waves of SARS-CoV-2 outbreaks in China and that this outbreak had the most far-reaching impact and generated the most spillover cases of Omicron variant of SARS-CoV-2. Therefore, it was included in the statistics despite its potential impact on the stability of the data. In a large outbreak such as Shanghai, we can use the number of cases fitted by the dynamics model (SEIAQR) to correct for bias in the number of daily reported cases due to reporting. Therefore, in this study, the dynamics model can be used as a correction for the input classifier model variables.

3.1.2. Spillover risk description

This study counted the specific districts and the cumulative number of spillover cases in each of the 7 areas, as shown in Fig. 3. Fig. 3a shows the number of spillover cases in the 7 areas, the number of which is 12, 3, 5, 4, 154 and 47. Among them, the area with the highest number of spillover cases was Shanghai with 959 cases, followed by Quanzhou, Beijing, Langfang, Nanchang and Tianjin. Fig. 3b shows the specific districts involved in the spillover of each district and the proportion of all spillover cases. According to the geographical division of China, it is divided into 8 different regions. Among them, Shanghai has the largest number of districts involved in the spillover, with 127 districts, followed by 17 in Quanzhou, 14 in Hainan, 10 in Beijing, 3, 3, and 2 in Nanchang, Langfang, and Tianjin.

The reason we divided China into eight regions on the map was to perform a descriptive analysis of the spillover case data, expecting to find the distribution of the number of spillover cases. In the spillover map, it can be seen that the spillover of diseases is mainly concentrated in the surrounding districts, i.e. the same areas, which is perhaps also related to the fact that people's choice of transport [17] for traveling is mostly by rail, car, etc. Perhaps better management of the same area could be more effective in controlling the outbreak and spread of the disease, or depending on the distance from the remaining seven areas, different levels of control in areas with cases may help reduce the pressure to prevent outbreaks.

3.1.3. Descriptive analysis of influencing factors

We conducted a descriptive analysis of the influencing factors based on the number of spillover cases mentioned above, and the cumulative results of the 472 samples included are shown in Table 1, in which the mean value of the distance was 1310.59 km, the interquartile spacing was 1154.07–1649.34 km, the mean value of the R_t was 1.1, and the interquartile spacing was 0.82–1.27 for the samples with successful spillover. The mean values of the number of cases, mid-risk areas, high-risk areas, and migration index were 10,878.85, 18.29, 11.98, 0.73, and 0.89, with interquartile spacing of 557.5–22,248, 13–13, 0–0, and 0.41–0.74.

The results showed that the interquartile range of R_t was (0.82–1.27), which is a low R_t level, and the interquartile range of migration index was (0.41–0.74), both of which indicated that the spillover cases were not in the rapid growth phase of the disease outbreak and tended to be in the early or late phase.

3.2. Analysis of influencing factors

3.2.1. Weight analysis of influencing factors

The random forest model was used in this study, and the explanatory variable weight scores of different districts affecting spillover outcomes were calculated. As shown in Fig. 4, in the seven districts included in the study, distance is almost the most important factor affecting spillover events, and its weight is above 0.3 in many districts. However, in Nanchang and Langfang, it could be detected that the weight of the population of the spillover area is higher than that of distance, but distance is still one of the main influencing factors. In addition to the most important influencing factor of distance, whether it is within the same geographical area is also one of the main influencing factors, the weight ratio of spillover results in it that Tianjin, Quanzhou, and Shanghai has reached more than 0.2. These four influencing factors of spill, mid-risk, high-risk areas and R_t have similar effects on spillover outcome events.

3.2.2. Variable correlation analysis

We conducted variable correlation analysis and variable frequency distribution plotting for the seven included explanatory and response

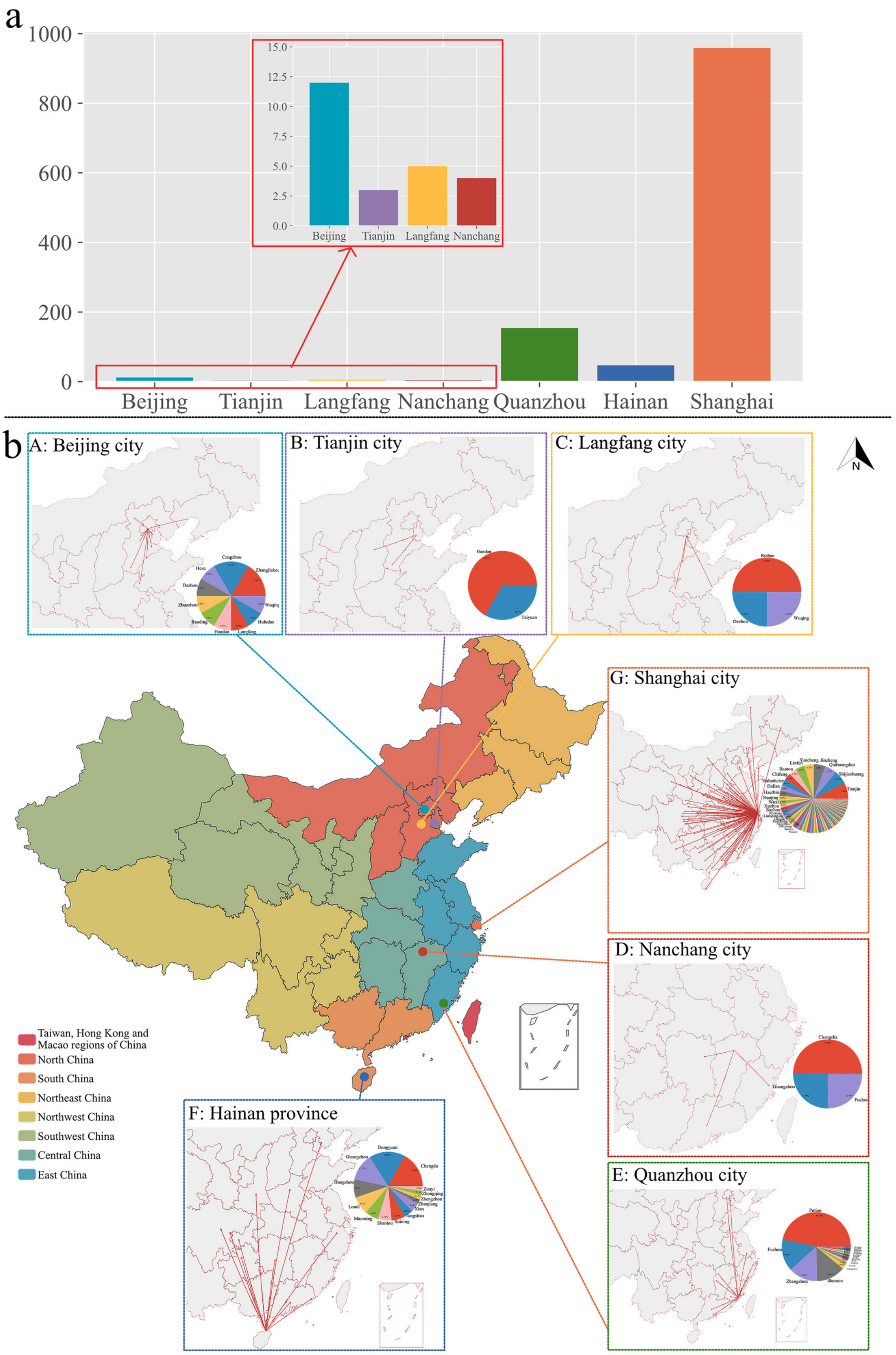

Fig. 3. Spillover case number bar graphs and maps. (a) The specific number of spillover cases in each district, and the red box shows the number of spillover cases in Beijing, Tianjin, Langfang and Nanchang. (b) The number of spillover maps. There are one China map and 7 local maps, where the lines in the local maps point to the spillover direction for the specific district of spillover, and the pie chart shows the number of spillover cases in the districts. The pie chart shows the proportion of districts with the number of spillover cases. (Map approval number: GS (2018) 5572).

Table 1
Statistical description of influencing factors.

	distance	R_t	Cases	Mid-risk	High-risk	Migration index	Same Area	Y
mean	1310.59	1.1	10,878.85	18.29	11.98	0.73	0.89	1.97
std	559.62	0.45	10,370.66	25.96	49.11	0.66	0.32	2.92
min	41.55	0.3	0	0	0	0.05	0	1
25%	1154.07	0.82	557.5	13	0	0.41	1	1
50%	1466.45	0.99	7333	13	0	0.5	1	1
75%	1649.34	1.27	22,248	13	0	0.74	1	2
max	3435.78	3.93	27,719	189	268	4.92	1	36

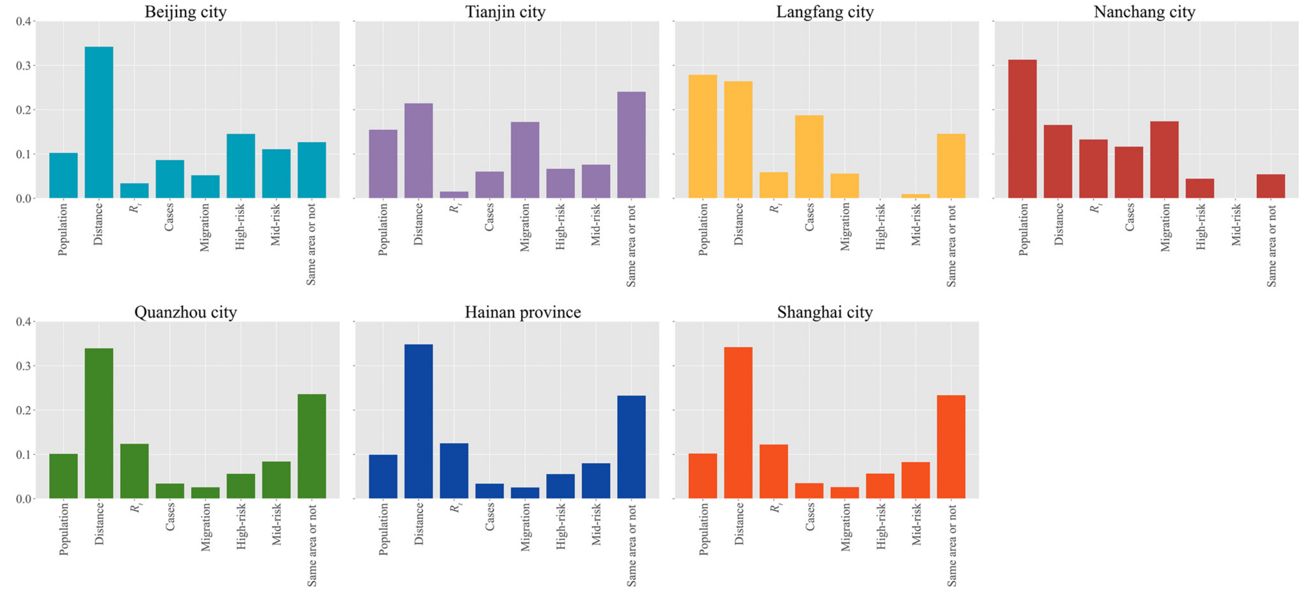

Fig. 4. Influencing factor weights. From left to right and from top to bottom, the weighting ratios of spillover risk influencing factors for different districts are shown, with Beijing as the first district and Shanghai as the last district.

variables, as shown in Fig. 5. The results indicate that the data distributions of R_t, Cases, High-risk, Mid-risk, Migration index, and Y all show positive skewed distributions, while Same Area shows negative skewed distributions. The pairwise correlation analysis performed in it shows that the correlations between distance and other explanatory variables except Same Area are almost negatively correlated. The R_t has little correlation with other explanatory variables. Cases has a strong correlation with distance, but it has a certain negative correlation with other explanatory variables. Among all explanatory variables, Mid-Risk and High-Risk showed a very strong positive correlation, and the correlation with other influencing factors was weak. There is almost no correlation between the other influencing factors.

As shown in Fig. 6, it was found that similar to the results of the correlation analysis above, the mid and high risk areas, R_t and migration index were more inclined to the same category, while the number of cases and distance, and whether the same area were more inclined to the same category, which has similarities with the results in Fig. 4 (Influencing factor weights), where the number of cases and distance were the factors that contributed more to the spillover outcome factors. Therefore, we can pay more attention to the two factors of distance and number of cases when we focus on the spillover outcomes in the real world.

3.2.3. *Analysis of influencing factors of successful spillover events*

In this study, the main influencing factors of the successful spillover outcome events were analyzed, as shown in Fig. 7. The results of the distribution of the influencing factors are similar to Fig. 5, with distance, R_t andmigration index showing a positive skewed distribution. In the analysis of influencing factors for all successful spillover events, the probability of successful spillover events was highest when the distance was around 1600 km, the R_t was close to 1, and the migration index was close to 0.5, the probability of a successful spillover event is the highest.

For the spillover events at the time of the outbreak, this study focused on the analysis of the factors influencing the spillover events. In the analysis results, it is found that the spillover infections are mainly concentrated in the same province or the same geographical area, and the distance is also concentrated within 200–300 km. And according to the analysis of the second most influential factor in the results section, it is also clear that distance is an important factor influencing the spillover results regardless of which district's classifier model is used for simulation. Moreover, there is no correlation between distance and other influencing factors, so it can be considered that spatial distance has the greatest impact on spillover outcome events. The degree of contribution of the remaining explanatory variables to the model varies widely across districts without corresponding stability, but still has a high weight in some models, even higher than the degree of influence of distance on the outcome event. For example, the influence of the number of people moving out of the destination in Langfang and Quanzhou on the outcome is higher than the influence of distance as a factor. Therefore, we should consider various factors, such as the geographical location of the outbreak, the radius of the outbreak, the size of the population, etc., when we subsequently develop different levels of prevention and control measures.

We further integrate the spillover success events and analyze the results of their influencing factors to show that the distance, R_t, and migration index all show a positive skewed distribution, which similarly indicates that the spillover success events are mostly concentrated in the proximity area. In contrast, the positive skewed distribution of R_t and migration index indicates that the spillover of infected persons is more likely to occur in the first and middle phases of the outbreak, which is

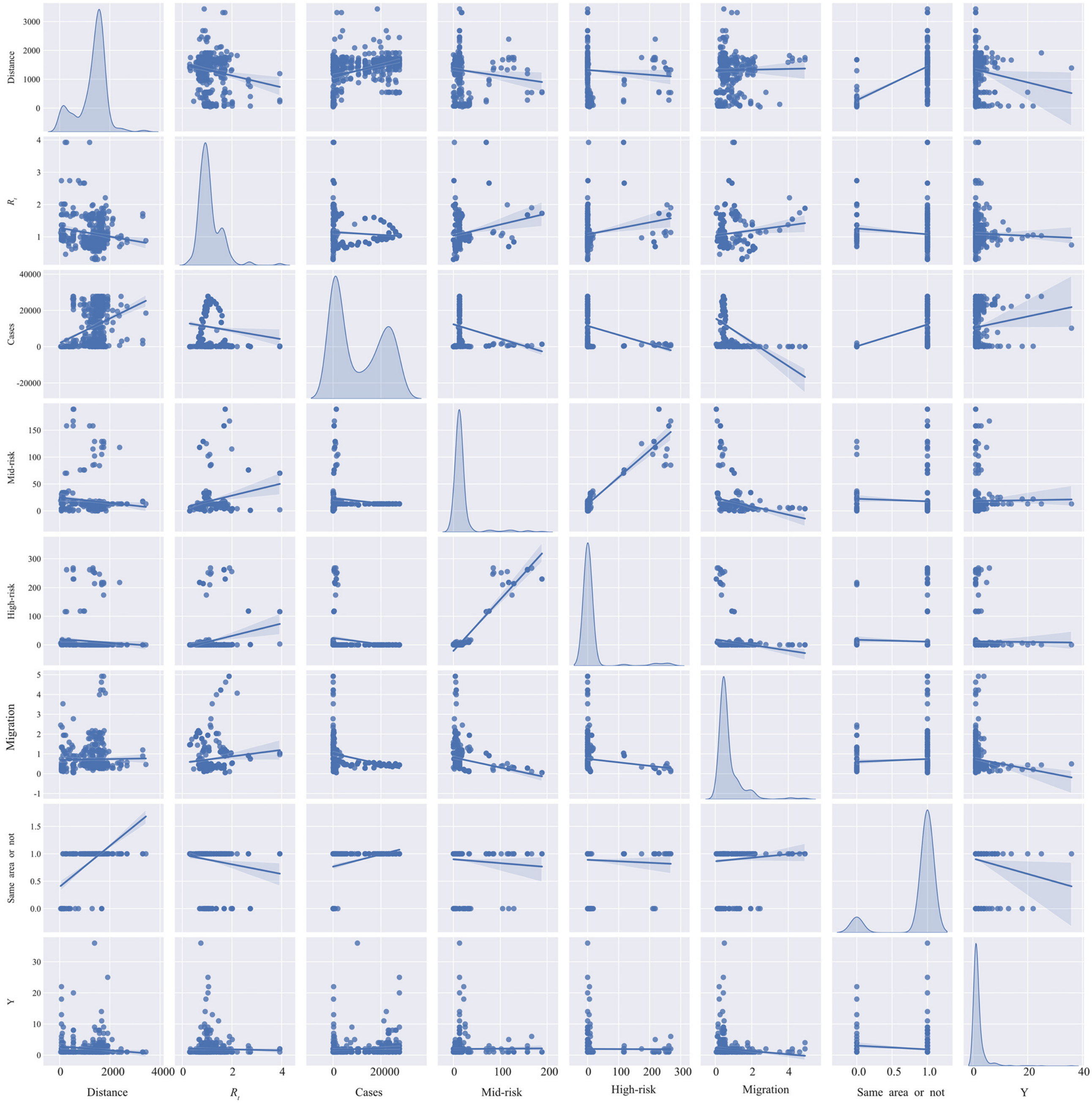

Fig. 5. Correlation analysis of variables. The diagonal line is the distribution of different influencing factors, both horizontal and vertical coordinates indicate the influencing factors, the cross part is the relationship between two influencing factors, and the thick blue line with a range indicates the degree of correlation between them.

also related to the way the outbreak is managed in the first and middle phases. The prevention and control measures are not strong and the frequency of nucleic acid testing is low in the first and middle stages of an outbreak, which makes it easier to generate the number of spillover cases. Therefore, we should strengthen the prevention and control efforts in the early stage of the outbreak, and deal with the outbreak as soon as possible to achieve rapid extinction and prevent spillover in the early stage.

In this study, we made a radiation range map for the spillover risk probability of more than 200 cities generated by the spillover model of seven districts. And the results showed that only Shanghai, Hainan, and Quanzhou have high-risk radiation, while the remaining four districts do not have high risk radiation. But the study does not exclude that there is still high-risk in the same provinces and geographical divisions.

3.3. *Analysis of model results*

3.3.1. *Function fitting results*

In this study, the logit function was fitted using spillover success data, and the specific results of the fit are shown in Table 2. The largest absolute value of coef is same area, followed by Intercept, then R_t, and the smallest is High-risk. p-value is less than 0.005 except for high-risk, which is statistically significant. We considered the multicollinearity of the model, and the results are shown in the Supplementary Table S2. The functions are as follows:

$$logit = \ln \frac{p_1}{p_0} = 1.1333 - 0.0022Dis + 0.9072R_t + 0.001Cases$$
$$+0.0308Mid - 0.0008High - 0.5918Spill - 1.9669Same$$

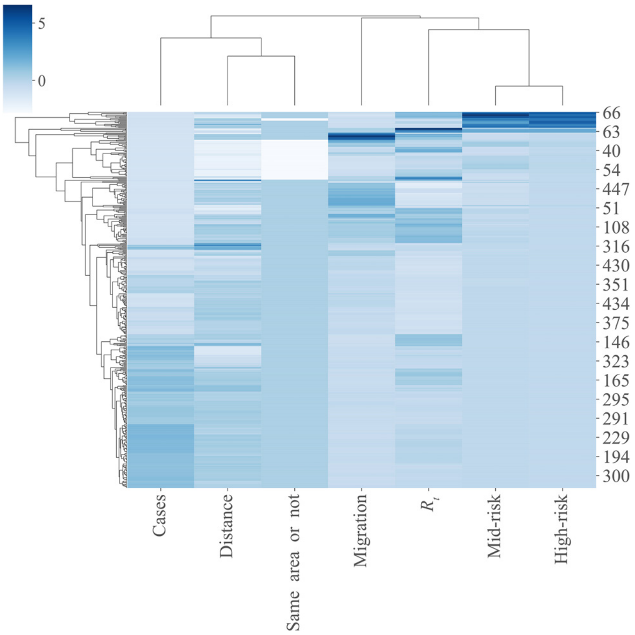

Fig. 6. Clustering analysis. The horizontal coordinates are the influencing factors of spillover risk, while the heat map part indicates the frequency magnitude of different variables, and the top part of the chart indicates the clustering results, where the two influencing factors with higher correlation are grouped into one category. For example, mid-risk and high-risk are classified into the same category in one clustering.

Table 2
Logit regression result.

	coef	std err	z	$P>\|z\|$	[0.025	0.975]
Distance	−0.0022	2.15e-0.5	−102.753	0.000	−0.002	−0.002
R_t	0.9072	0.013	67.529	0.000	0.881	0.933
Cases	0.0010	5.33e-0.5	19.497	0.000	0.001	0.010
Mid-risk	0.0308	0.001	50.954	0.000	0.030	0.032
High-risk	−0.0008	0.000	−1.865	0.062	−0.002	3.84e-0.5
Migration	−0.5918	0.013	−45.583	0.000	−0.617	−0.566
Same_area	−1.9669	0.023	−85.320	0.000	−2.012	−1.922
Intercept	1.1333	0.028	39.796	0.000	1.077	1.189

3.3.2. *Model score results*

We incorporated three classifier models RF, SVM, and Bayes, and compared the accuracy, macro mean, and AUC values of the three classifier models. As shown in Table 3, due to too few positive samples, the three districts of SVM, Bayes, namely Tianjin, Langfang, and Nanchang cannot be used for calculation. The model scoring results of the three models for seven districts, RF, earn accuracy, macro, and accuracy in the three models. The mean value and AUC value were the highest.

This study focuses on the simulation of disease prevalence curves using infectious disease dynamics models. Comparing classifier models in machine learning, it is concluded that the random forest model has the advantages of interpretability as well as high accuracy, so the random forest model is chosen in this study. Although the sample imbalance was considered and solved in the training model, the AUC value of the model could still not be calculated for some regions due to the problem of a small sample size of positive events, i.e., spillover successes. In the end, the results of the three classifier models were synthesized, and the RF model was used to simulate the spillover risk.

3.3.3. *Spillover risk level*

As shown in Figs. 8, and 7 districts are divided into three different risk level ranges. According to the simulation results of the classifier model, if the spillover success probability is greater than 0.5, it is identified as a high-risk area (red), the spillover risk probability is between 0.2 and 0.5 as mid risk (mid risk), and less than 0.2 as low risk (blue). Not every district has three risk areas at the same time, only two districts, Quanzhou and Shanghai, have both high and low-risk areas, but Quanzhou's mid and high-risk areas are concentrated in Fujian province, covering 4 and 2 geographical areas respectively. The low-risk area coverage is 1,274 km, while Shanghai's three risk areas range from high-risk to low-risk coverage is focused on 1,300–166 km, in the order of 1,317, 1,519, 1,586 km, and covers all eight regions of the Tianjin and Langfang, having similar mid and low risk coverage, and there is no high-risk area. But the coverage of Tianjin involves three regions and Langfang covers only one north China. Beijing mid-risk areas are concentrated within Beijing, and low-risk area coverage is mainly concentrated in northern China. Nanchang only exists mid risk area coverage to 6 areas. Hainan only had mid and high-risk areas, mainly concentrated in South China.

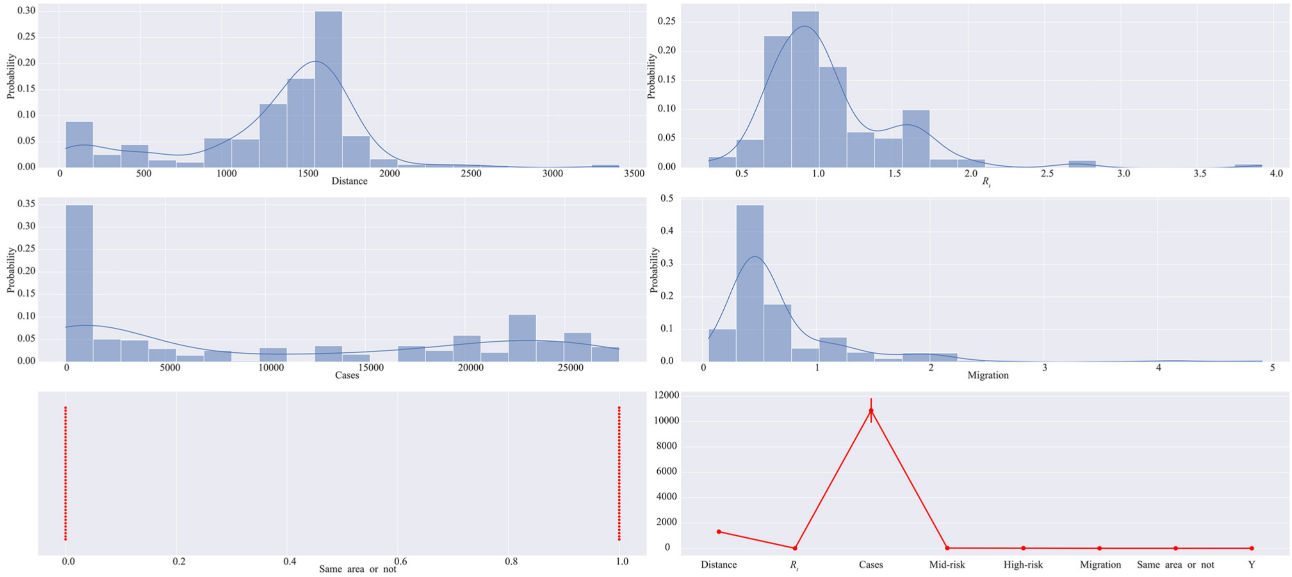

Fig. 7. Distribution probability of spillover success event variables. (a-c, e) The probability density distribution of the five influencing factors of Distance, R_t, Cases, Migration, and Same Area, respectively. (f) represents the distribution of all influencing factors, where the red dot position is the median and the error line is the interquartile spacing.

Table 3
Model comparison results.

Model	Evaluation Indicators	Beijing	Tianjin	Langfang	Nanchang	Quanzhou	Hainan	Shanghai
RF	accuracy_score	0.9989	0.9991	0.998	0.9996	0.9973	0.9946	0.716
	Macro-averaging	0.6247	0.5	0.4993	0.5	0.8179	0.4981	0.5161
	roc_auc_score	0.6247	–	0.4997	–	0.9364	0.4992	0.7081
SVM	accuracy_score	0.9993	–	–	–	0.983	0.8694	0.716
	Macro-averaging	0.4996	–	–	–	0.4978	0.5144	0.5161
	roc_auc_score	0.5	–	–	–	0.4937	0.9344	0.7081
Bayes	accuracy_score	0.9869	–	–	–	0.8667	0.9946	0.9832
	Macro-averaging	0.5071	–	–	–	0.5159	0.4801	0.4919
	roc_auc_score	0.6187	–	–	–	0.9331	0.4992	0.4997

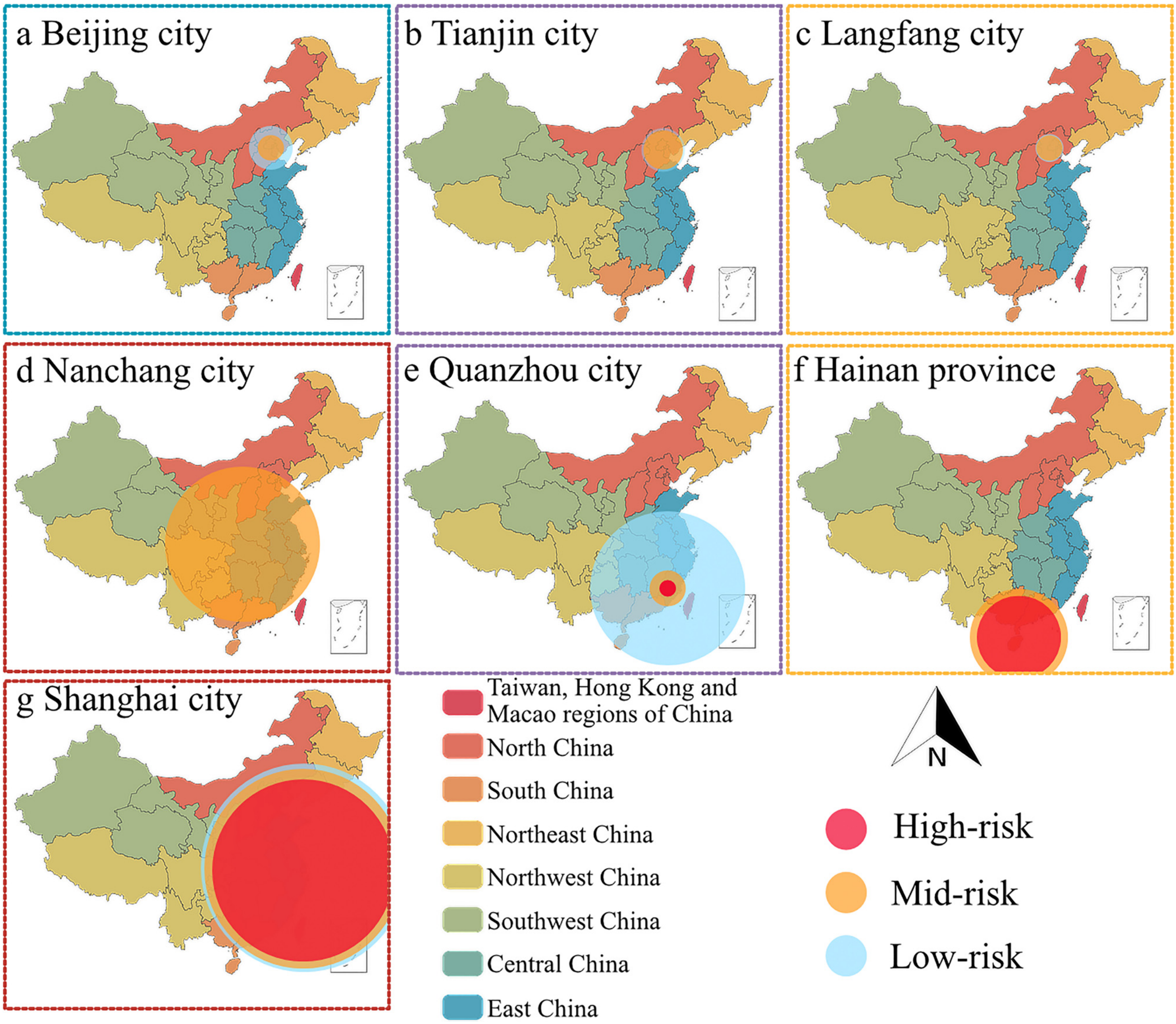

Fig. 8. Spillover risk rating range. From left to right, from top to bottom, the last district is Shanghai, such as Beijing and Tianjin, where the red range indicates that the model simulation results in high risk, the orange part indicates medium risk, and the blue part indicates low risk areas. (Map approval number: GS (2018) 5572).

The results of this part of the study show that the scope of the outbreak affected varies from district to district, but it is the surrounding cities that are most affected. This study divided into three risk levels, and some districts differed greatly in the scope of the three levels, while others differed very little, which may be related to the population mobility preferences of a particular location, an issue not considered in this study.

4. Conclusion

The variables affecting the spillover outcome events are diverse, and the study believes that it is difficult to prevent the spillover of infected people by implementing the same outbreak prevention measures at the provincial, city and county levels. The human factors that lead to the emergence of diseases vary by society, and they can also vary by culture, history and geography, or by the occurrence of black swan events and other factors that ultimately lead to outbreaks. However, the chain of spillover events may vary from society to society, and the only way to prevent future disease spillovers from occurring is to analyze the impact factors on outcome events according to their respective characteristics to stop the chain of events and prevent the occurrence of unexpected events.

The results of all studies show that the most important factor influencing disease spillover is distance. Therefore, different risk radius should be classified according to the spatial distance of outbreak sites, and different risk levels should be classified according to different stages [15] of the outbreak to achieve the purpose of stratified prevention and control, which is more conducive to "Dynamic zero" and in line with the cost-benefit principle. However, many factors influence the development of an epidemic and we need to consider them in the context of various realities; for example, we may need to consider the impact of different local population movements, economic and cultural levels, and other factors on the epidemic. Nonetheless, we can still identify some of the major factors influencing the spread of the epidemic and thus change the level of measures to achieve more effective epidemic control.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

This project was supported by National Key Research and Development Program of China (2021YFC2301604) and the Fundamental Research Funds for the Central Universities (No. 20720230001).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.fmre.2023.03.014.

References

[1] D. Cucinotta, M. Vanelli, WHO declares COVID-19 a pandemic, Acta Biomed. 91 (2020) 157–160.
[2] WHO, WHO News. https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern, 2021 (accessed 25 August 2022).
[3] D. The Lancet Infectious, Emerging SARS-CoV-2 variants: shooting the messenger, Lancet Infect. Dis. 22 (2022) 1.
[4] TheGuardian, TheGuardian Europe, https://www.theguardian.com/world/2022/feb/21/boris-johnson-says-free-covid-tests-in-england-will-end-on-1-april/, 2022 (accessed 25 August 2022)
[5] S. Dhanda, V. Osborne, E. Lynn, et al., Postmarketing studies: can they provide a safety net for COVID-19 vaccines in the UK? BMJ Evid.-Based Med. 27 (2022) 1–6.
[6] J. Ulimwengu, A. Kibonge, Spatial spillover and COVID-19 spread in the U.S, BMC Public Health 21 (2021) 1765.
[7] Q. Zhao, M. Yang, Y. Wang, et al., Effectiveness of interventions to control transmission of reemergent cases of COVID-19 - Jilin Province, China, 2020, China CDC Wkly 2 (2020) 651–654.
[8] T.M. Chen, J. Rui, Q.P. Wang, et al., A mathematical model for simulating the phase-based transmissibility of a novel coronavirus, Infect. Dis. Poverty 9 (2020) 24.
[9] Z.Y. Zhao, Y.Z. Zhu, J.W. Xu, et al., A five-compartment model of age-specific transmissibility of SARS-CoV-2, Infect. Dis. Poverty 9 (2020) 117.
[10] S.N. Lin, J. Rui, Q.P. Chen, et al., Effectiveness of potential antiviral treatments in COVID-19 transmission control: a modelling study, Infect. Dis. Poverty 10 (2021) 53.
[11] L. Gonçalves, A. Subtil, M.R. Oliveira, et al., ROC curve estimation: an overview, REVSTAT-Stat. J. 12 (2014) 1–20-21–20.
[12] B. Charbuty, A. Abdulazeez, Classification based on decision tree algorithm for machine learning, J. Appl. Sci. Technol. Trends 2 (2021) 20–28.
[13] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, et al., Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines, Ore Geol. Rev. 71 (2015) 804–818.
[14] D. Berrar, Bayes' theorem and naive Bayes classifier, Encycl. Bioinform. Comput. Biol.: ABC Bioinform. 403 (2018).
[15] B. Mahesh, Machine learning algorithms-a review, Int. J. Sci. Res. (IJSR) 9 (2020) 381–386.
[16] X. Zhang, W. Zhang, S. Chen, Shanghai's life-saving efforts against the current omicron wave of the COVID-19 pandemic, Lancet 399 (2022) 2011–2012.
[17] X.K. Xu, X.F. Liu, L. Wang, et al., Assessing the spread risk of COVID-19 associated with multi-mode transportation networks in China, Fundam. Res. (2022).

Hongjie Wei, master, is at the State Key Laboratory of Molecular Vaccinology and Molecular Diagnostics, School of Public Health, Xiamen University. His main research direction is research on mathematical modeling of infectious diseases and response strategies for public health emergencies.

Tianmu Chen (BRID: 07167.00.08959) holds the position of associate professor at the School of Public Health and serves as the deputy director of the Department of Preventive Medicine at Xiamen University. His research encompasses the study of over 20 newly emerging and sudden infectious diseases and has involved the construction of more than 60 related models. His work has received support from the Guangzhou Laboratory, the Bill and Melinda Gates Foundation, and the National Key Research and Development Program. He has published more than 100 SCI papers in distinguished academic journals. He is also the director of the Research Center of Surveillance and Early Warning Technology, which is jointly established by the School of Public Health at Xiamen University and Shangrao Center for Disease Control and Prevention.