Urban spatial risk prediction and optimization analysis of POI based on deep learning from the perspective of an epidemic

Yecheng Zhang [a], Qimin Zhang [b], Yuxuan Zhao [a], Yunjie Deng [a], Hao Zheng [c,*]

[a] College of Architecture & Art, Hefei University of Technology, Hefei, China
[b] School of Mechanical Engineering, Hefei University of Technology, Hefei, China
[c] Stuart Weitzman School of Design, University of Pennsylvania, Philadelphia, United States

ARTICLE INFO

ABSTRACT

From an epidemiological perspective, previous research on COVID-19 has generally been based on classical statistical analyses. As a result, spatial information is often not used effectively. This paper uses image-based neural networks to explore the relationship between urban spatial risk and the distribution of infected populations, and the design of urban facilities. To achieve this objective, we use spatio-temporal data of people infected with new coronary pneumonia prior to 28 February 2020 in Wuhan. We then use kriging, which is a method of spatial interpolation, as well as core density estimation technology to establish the epidemic heat distribution on fine grid units. We further evaluate the influence of nine major spatial risk factors, including the distribution of agencies, hospitals, park squares, sports fields, banks and hotels, by testing them for significant positive correlation with the distribution of the epidemic. The weights of these spatial risk factors are used for training Generative Adversarial Network (GAN) models, which predict the distribution of cases in a given area. The input image for the machine learning model is a city plan converted by public infrastructures, and the output image is a map of urban spatial risk factors in the given area. The results of the trained model demonstrate that optimising the relevant point of interests (POI) in urban areas to effectively control potential risk factors can aid in managing the epidemic and preventing it from dispersing further.

1. Introduction

1.1. Research background

The COVID-19 epidemic first broke out in China towards the end of the year 2019. From the beginning of 2020 to the present, the overall reported cases exceed 400 million, with a mortality rate of 1.3 %. However, since COVID-19 was included as a category B infectious disease by the Chinese National Center for Disease Control and Prevention (Velavan and Meyer, 2020), there has been an overall downward trend in pneumonia outbreaks in the country (Bauch et al., 2005; Walker et al., 2020). Based on the past 2,500 years of experience in combating infectious diseases around the world (Badawi and Ryoo, 2016), we know that disease outbreaks are associated with a number of risk factors (Li et al., 2020b; Lin et al., 2020; Prem et al., 2020). Currently, the majority of methods used to study coronary pneumonia from an epidemiological perspective are based on classical statistical analyses, and spatial information is often not used effectively. However, with the growing prevalence of artificial intelligence (AI) and the corresponding increase in computational capabilities, the currently available technology allows for prediction of outbreaks, enabling more effective epidemic prevention and control (Kissler et al., 2020; Perkins and Espana, 2020).

1.2. Literature review

Since the start of the 21st century, machine learning theory has been applied to the construction of transmission models for schistosomiasis, dengue fever, and drug abuse (Alpren et al., 2020; Carvajal et al., 2018; Solano-Villarreal et al., 2019), with beneficial technical results. For instance, Salami et al. (2020) presented a machine learning-based analytical model for the spread of the dengue fever epidemic in Africa and compared four common classifier methods such as partial least squares (PLS) and glmnet. The final maximum sensitivity score of the training model is 0.88, which provides a new design solution for dengue fever early warning surveillance system. Gong et al. (2021) developed a neural network-based prediction model for the transmission of

schistosomiasis. In addition, the implementation of machine learning algorithms in predicting HIV transmission as a function of opioid use, which can be approximated as a model for the spread of infectious diseases, also offers a new solution (Campo et al., 2020). For example, Yedinak et al. (2021) proposed a machine learning-based predictive model for opioid abuse in the United States and applied it to estimate relative vulnerability scores in Rhode Island at a local scale. This approach was influenced by complex factors such as housing density, drug sales, and employment levels. Furthermore, Schneider et al. (2021) proposed a machine learning method for predicting dengue fever based on climatic factors in addition to geographical data, which enhanced the infectious disease transmission model and made it more realistic.

With the growing popularity of machine learning in urban data analysis (Cao and Zheng, 2021; He and Zheng, 2021; Shou et al., 2021; Sun et al., 2020), predictive models for coronary diseases can be broadly divided into two categories: the prediction of trends in the spread of epidemics based on human activity and related information (Li et al., 2020a), and the construction of regression models to predict trends in outbreaks based on geographic information data (Sebastianelli et al., 2021). An example of the former is the study conducted by Niu et al. (2021), which examined the influence of spatio-temporal factors on the Italian epidemic. Spatial autocorrelation, spatio-temporal cluster analysis, and kernel density methods were employed to analyse the spatial clustering of new coronary pneumonia cases, and it was observed that factors such as masks, thermometers, and medical masks contributed highly to the model; the results were then compared with publicly available data to validate the model. Wu et al. (2020) reflected on potential social and non-pharmacological preventive interventions. They proposed a predictive model that used a Markov chain Monte Carlo (MCMC) based approach, combining their data with data from major cities in other countries, and found that the epidemic would affect most cities in the first half of 2020. In addition, Mackey et al. (2020) proposed a retrospective big data information surveillance study based on machine learning, which effectively linked search data on Twitter with the prevention and control of the epidemic, thus successfully exploiting social media information to provide a novel means of epidemic management.

Currently, the development of regression models to predict geographic trends in disease outbreaks based on spatial data is one of the major avenues of research (Hamer et al., 2020). Wang et al. (2020) used COVID-19 outbreak data from China up to June 2020 to develop and train a logistic model — a time series prediction model developed based on machine learning — which predicted the outbreak trend of new crowns. Validation with confirmed cases revealed that the model's predictions were relatively accurate until mutant strains such as Omicron emerged. Amar et al. (2020) used the case of the COVID-19 outbreak in Egypt as an example and constructed various regression models based on machine learning methods. However, due to the influence of multiple complex variables such as policy, nature, and social factors, the proposed model was not as accurate as the results predicted by He and Zheng (2021). Sahai et al. (2022) proposed a proximity regression-based short-term forecasting model using the random forest (RF) algorithm in combination with machine learning theory to analyse the US outbreak and the characteristics of lagging epidemic information. This outperformed the complex Bayesian model and successfully predicted disease incidence based on 3–4 days of epidemic data, which was helpful in analysing the Ohio outbreak and even local-scale outbreaks with appropriate guidance. In the long term, Bloise and Tancioni (2021) chose to use a machine learning approach to predict concentrated outbreaks in Italy based on 77 risk factors, and estimated that the optimal time for manual intervention was in early June. Their elastic net model outperformed traditional predictive analysis tools by testing outbreaks in 19 locations, thus providing a novel solution for the prevention and control of epidemics. However, their model did not take into account the complex co-linear effects of different geographical facilities on the spread of the epidemic. Watson et al. (2021) proposed an

epidemiological compartmental model within a Bayesian time series model to predict new coronary pneumonia in the United States. This model performed well over a period of 20 days; however, accuracy was observed to fluctuate after 20 days due to factors such as epidemic mortality.
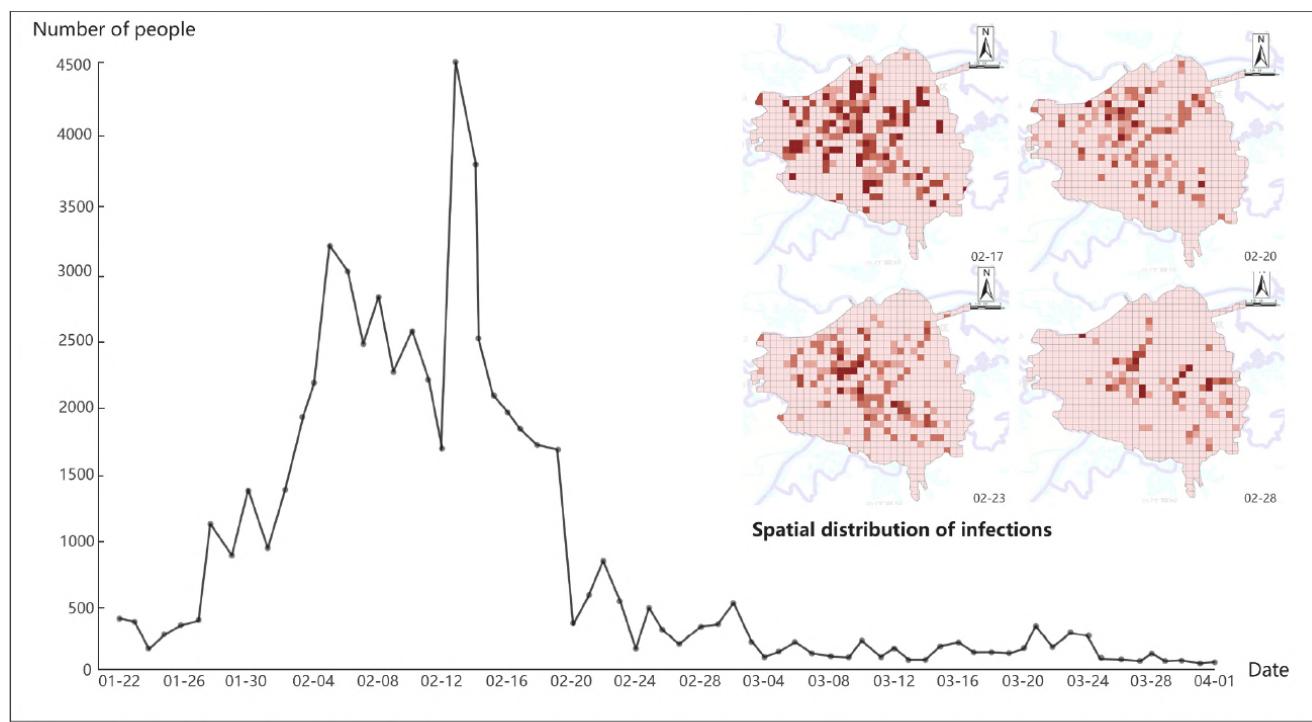
Metric geography and spatial statistics have proven suitable for predicting the epidemic spread of pathogens. However, in this context, most machine learning models based on urban infrastructure information lack an appropriate time-based explanation. The granularity of the study is large. A combination of geostatistical regionalisation, machine learning methods, and long-term pathology data series have led to verifiable predictions. For example, Wu et al. (2021) analysed the spatial and temporal patterns of NCC, quantifying the risk factors for NCC using point-grid maps with calendar-based visualisation to improve resource allocation and emergency response decisions. Various other geographical factors have also been incorporated into studies that aim to predict NCC transmission. Yao et al. (2021) developed an RF algorithm-based community- and place-scale 'spatial variable-infection risk' model to assess the risk of NCC. This model, coupled with community-level epidemic transmission data, point-of-interest data, LBS population density data, and road network data, provided a more accurate assessment of the epidemic. However, the granularity of the study was large and could not fit generalisations to unknown areas.

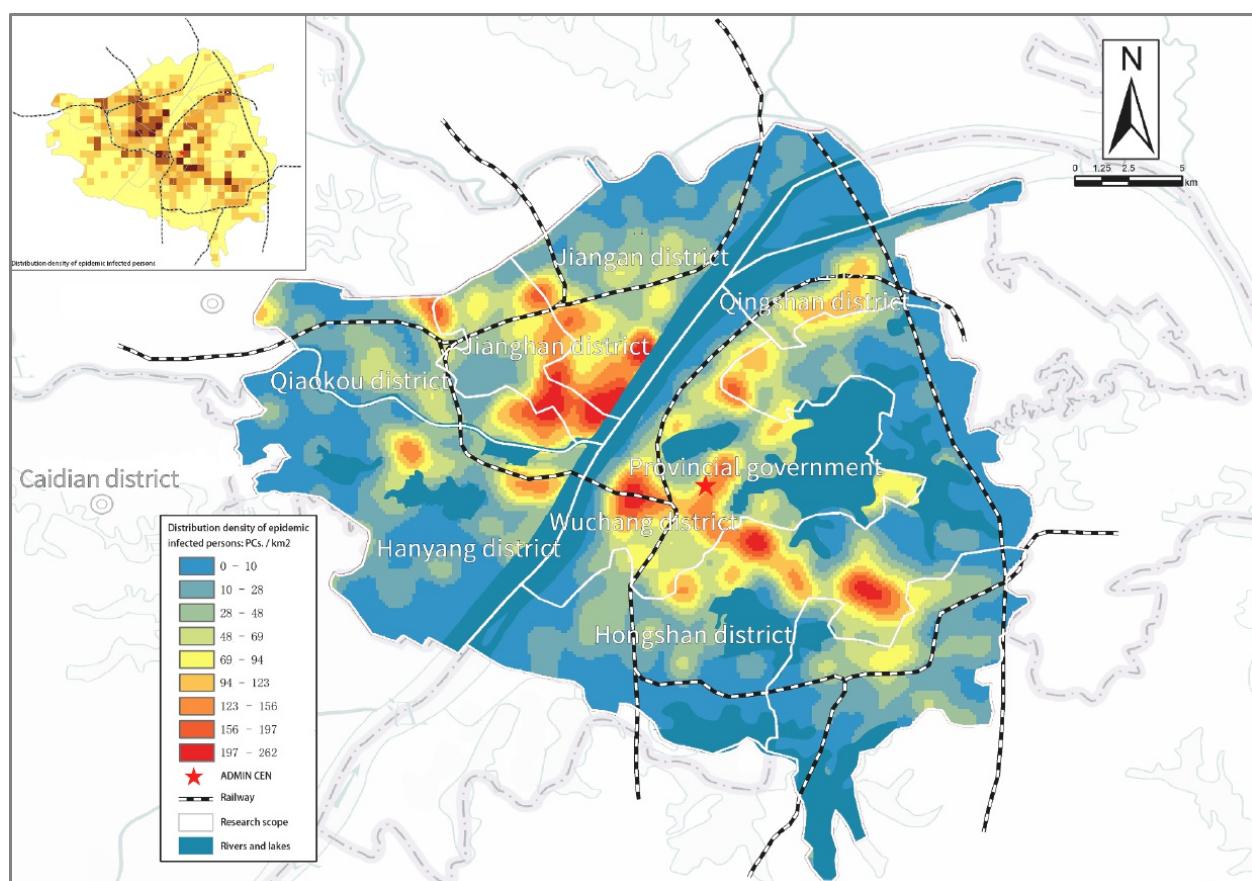## 1.3. Problem statement and objectives

Most domestic and international studies aiming to predict the distribution of COVID-19 new coronary pneumonia based on statistical analyses have focused on the influence of human activity on outbreak trends (Remond and Remond, 2020), and the methods used are unable to quantify complex urban environmental factors (Wu et al., 2021). The existing AI-based epidemic risk prediction models are generally large in granularity, low in timeliness, unable to fit unknown areas, and are not suitable for exploring urban environmental risk characteristics (Wu et al., 2020).

The corresponding author of this paper had used GAN neural network to build a city plan prediction model and corresponding crime distribution map to fit the influence of various environmental features on the crime rate of the city, and brought into the facility planning map of Seattle, New York and other cities to predict the crime rate (He and Zheng, 2021). In addition, the authors had used machine learning algorithms to study the connection between human behavior patterns and facility planning, compared two error analysis models to improve the prediction accuracy (Cao and Zheng, 2021). In this paper, we improve the data type, and propose to use two GAN models to enhance the interpretability of incidence prediction. Compared with the previous research (He and Zheng, 2021) that took the urban map as input, our method uses the POI distribution to reflect the intrinsic influencing features of the urban design, and better predict the spatial risk factors. Therefore, based on past research, we uses the GAN network model to analyze and predict the relationship between spatial risk factors and the outbreak of the epidemic in Wuhan as the research object, and optimizes the model by using multiple error analysis results.
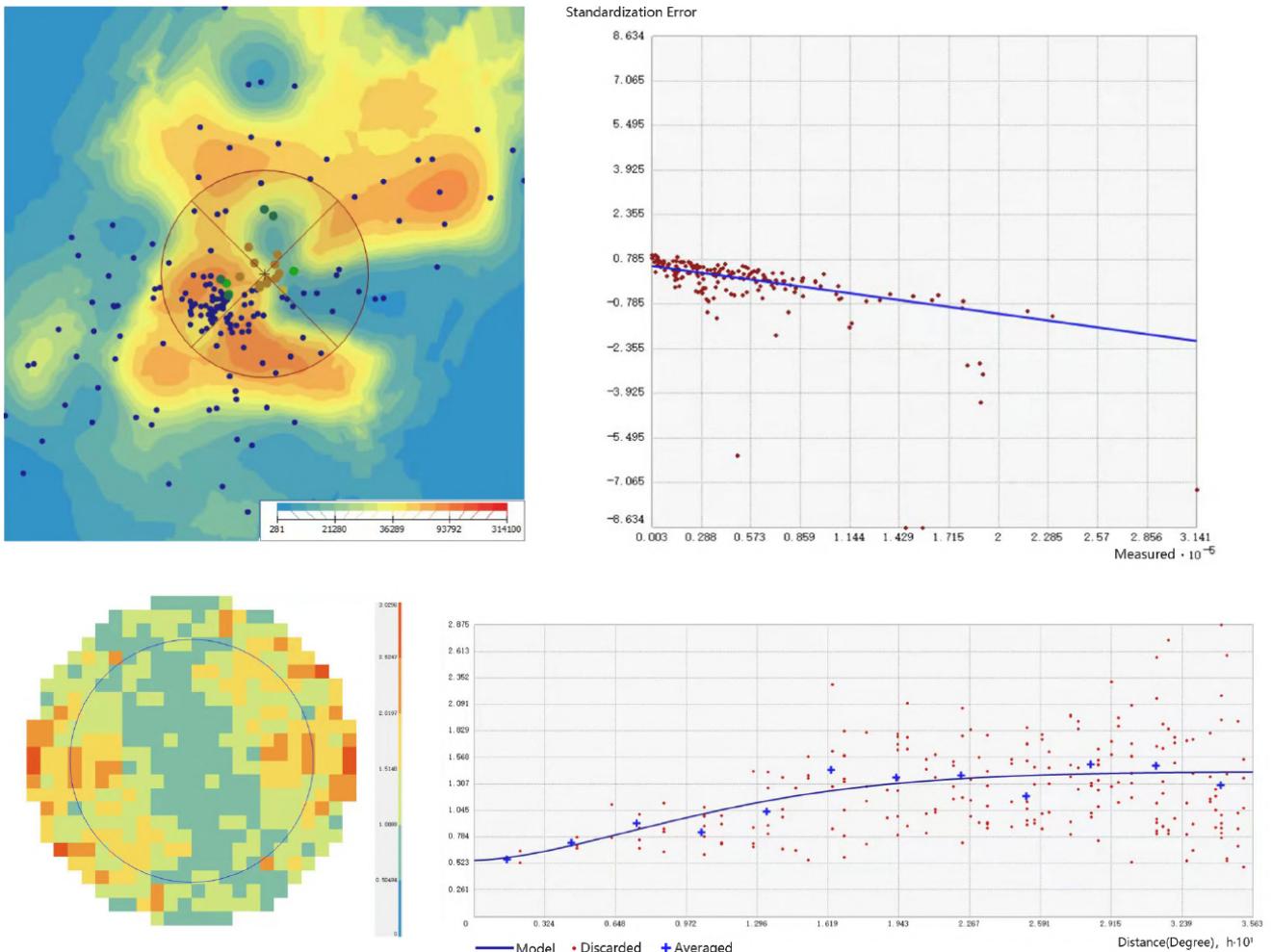
Therefore, this paper proposes an image-based neural network to investigate the relationship between urban spatial incidence and urban facility design, and to predict and validate the covariance between various urban epidemic risk factors and their influence on urban outbreaks of COVID-19. We have constructed a relatively complete set of algorithmic processes that may be applicable to other types of space research. The process is as follows: (1) Establish accurate urban outbreak incidence samples through kernel density analysis and geostatistical interpolation. (2) Explore the correlation mechanism between urban epidemic spread risk and spatial environmental factors through Pearson analysis and grey relational analysis. (3) Simulate and validate the outbreak in Wuhan as a case study. (4) By experimenting with various inputs and model types using open source data and Generative

Fig. 1. Statistics for daily new infections in Wuhan.

Fig. 2. Thermal and spatial risk factors of the epidemic in Wuhan.

Fig. 3. Top: Semi-variogram of kriging interpolation model. Bottom: Search neighborhood graph.

Adversarial Networks (GANs), we can easily compare the effects of different environmental factors, revealing the interactions between urban environmental risk and urban design. The generative adversarial neural networks approach is innovative in terms of data types used and learning methods adopted.

2. Data and spatialization

2.1. Study area and data sources

The central urban area within the third ring road of Wuhan, covering an area of about 591 km^2, was selected for the study. This region covers the seven administrative districts in Wuhan where the epidemic had the greatest impact as of 28 February. The study used a multi-source data fusion approach, in which map data were modified from the standard map provided by the website of the Hubei Provincial Department of Natural Resources (ES(2020)003), and epidemic distribution data were obtained from various public websites and from information released by the official website of the Wuhan Health and Wellness Commission using Scalable Web Crawler technology. According to Wu et al. (2020), the spread of the epidemic in Wuhan appeared to level off after 28 February (Fig. 1). The scale of the Wuhan outbreak estimated by the SEIR model reached its peak in mid-March, hence the data obtained for the purpose of this study can be considered to be well-interpreted and representative. As can be seen in Fig. 1, the number of new infections fluctuated considerably on and after 17 February; the spatial distribution of patients over the four days in that interval shows that there was a strong spatial clustering of newly crowned infected persons, predominantly in densely populated areas such as hospitals and shopping malls. The spatial migration of infected persons was generally centered on the central part of the city, with a decreasing trend towards the peripheral areas. After 23 February, only the central city and sporadic peripheral areas continued to report new infections.

2.2. Epidemic distribution

The spatial analysis method requires prior geo-referencing of the patients' information based on the available records, i.e., converting the patients' information from text format into point data with a specified spatial location. Method selection is based on the main pathways of transmission of the infectious disease and the objective of the study. New coronary pneumonia can be transmitted from human-to-human in case of close contact via droplets, and hospital and home transmission are the two major known transmission routes.

After excluding data points located outside the Third Ring Road, a total of 30,617 data points were obtained for this study. The community-level population data released by the Census were combined with these data points to obtain a complete heat map of the distribution of the new crown epidemic (indicative of incidence rate), which would better reflect the overall status of outbreaks in the region.

This study uses kernel functions (one of the most widely used methods for spatial point pattern analysis) to calculate the volume per unit area based on point elements to fit each outbreak distribution point to a smooth cone-shaped surface(Equation (1)).
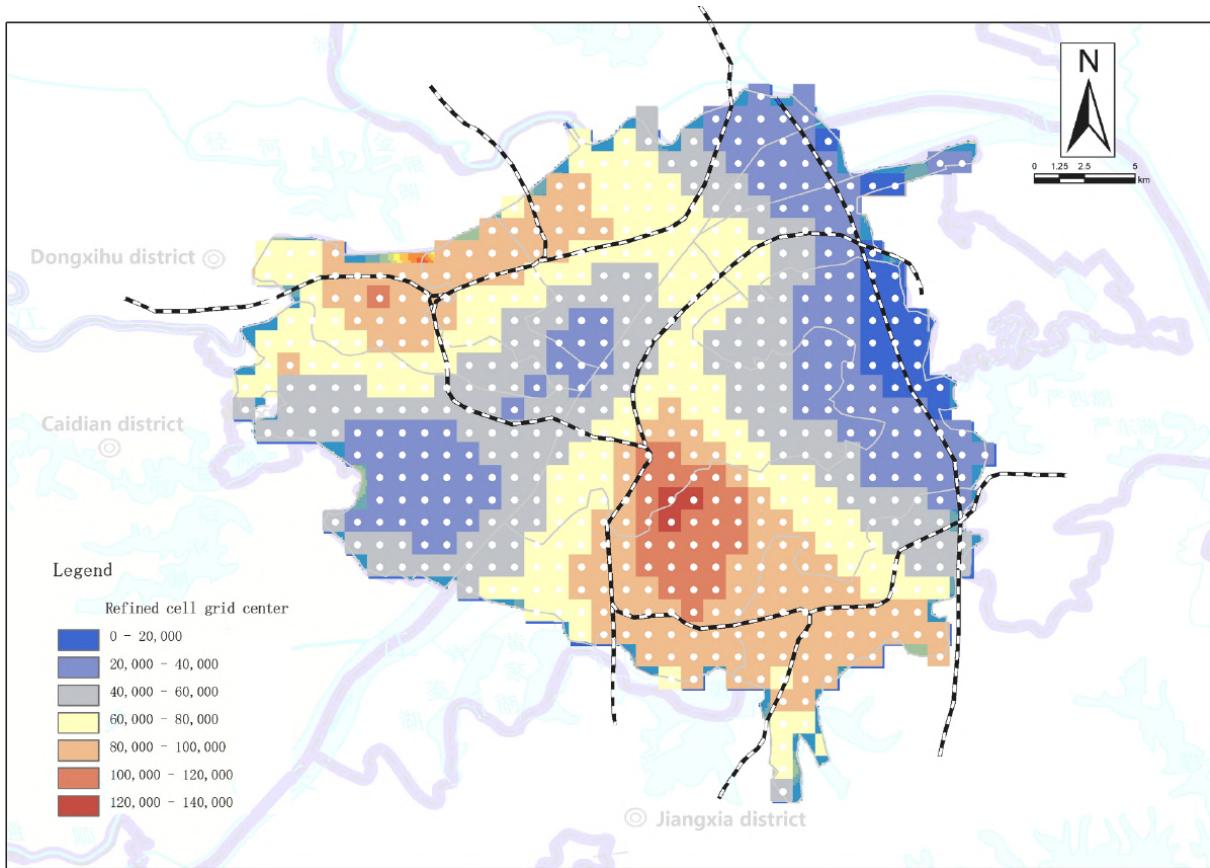
Fig. 4. Population density.

$$\lambda_\tau^\Lambda(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{s - s_i}{\tau}\right) \tag{1}$$

Fig. 2 shows a heat map of the distribution of the outbreak after preliminary kernel density processing. It can be seen that the outbreak in Wuhan before 13 February was distributed in a north–south direction along the river, concentrated in Wuchang, Hankou, and Hanyang. The outbreak in Jiangbei was concentrated in the Hankou area, and the first outbreak site, the South China Market, is located in the Jiangan district of the Hankou area, with an average confirmed density of over 150 PCs/km^2. The main outbreak areas in Jiangbei were concentrated in commercial areas, such as the Jianghan Road business district and the Linjiao Lake business district. These regions have an average confirmed density of 200 PCs/km^2 or more around the Wuhan Passenger Port and Wuhan Pass Terminal along the river, thus allowing for a more continuous cluster of hotspots. The outbreaks south of the Yangtze River in Wuhan were concentrated in Wuchang and Hongshan regions; the three hotspots with the highest outbreak intensity from north to south were: around Dacheng Fresh Market in Wuchang District, around Huazhong Normal University in Hongshan District, and around Guanggu Square, with an average confirmed density of more than 200 PCs/km^2. In addition, outbreaks were also concentrated around Steel Flower New Village in Qingshan District, Fruit Lake in Wuchang District, and Baishazhou Community in Wuchang District. Due to natural features such as water bodies and mountains, the distribution of the epidemic in Wuhan was patchy. Further, the data on the distribution of the epidemic in the community were limited by fluctuations in the number of people in the community and the frequency of activities, and cannot accurately reflect the extent of the epidemic in the region. Therefore, this paper introduces the definition of the incidence rate as described in the severe acute respiratory syndrome (SARS) transmission study for Guangzhou by Cao et al. (2008). The corresponding census data was obtained for a visual representation of the specialization of the population.
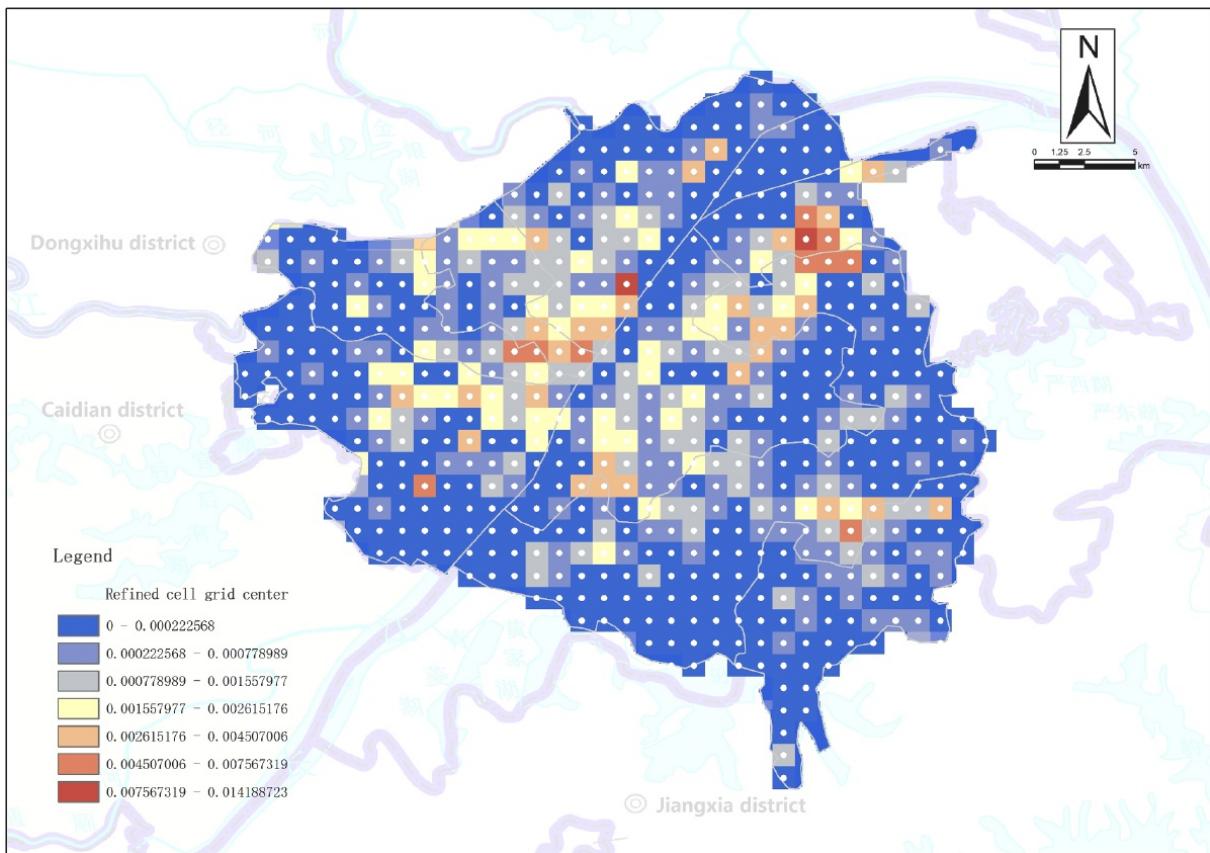
2.3. Incidence of outbreaks and palatalization

Incidence rate is defined as the frequency of new cases reported over a certain period and can represent the regional distribution of cases. This can be visualized using the same criteria as the associated geographical factors (Fig. 3). The formula for this quantity can be given as, Incidence Rate = the number of new cases occurring during the observation period / average population during the same period. Since this paper studies the influence of various urban spatial risk factors at the microscopic scale, the incidence map based on grid cells, which expresses the relevant information, is used to spatialize the population in advance. The demographic data based on administrative divisions are assigned to these grid cells based on a certain model.

Based on the 7th National Census data, the kriging surface interpolation model is used to assign the population statistics to a fine grid cell with a spatial resolution of 1 km (591 cells). Kriging interpolation takes into account the spatial autocorrelation of population density in Wuhan during the data gridding process. This ensures that the estimated population distribution is more consistent with the actual scenario.

The calculation process for the kriging interpolation model for population density is as follows:

(a) The histogram and normal QQ plot of the population distribution data are analysed using the Geostatistical Analyst tool, the log transformations obtained from the tests forming the normal distribution are observed, and the observation in Trend Analysis that the set of points can be fitted in space to the form of a quadratic surface.

Fig. 5. Incidence rate.

(b) The ordinary Kriging interpolation tool in the Geostatistics Wizard is selected, and log-normal transformation and second-order detrending are performed to ensure that the population data meets the smoothness requirements for this method.

(c) Theoretical semi-variance functions for population density are obtained by fitting the empirical semi-variance values using the STABLE model.

(d) Based on the theoretical semi-variance functions obtained, the population density at the centroid of each grid cell can be estimated by using known population density values.

This study uses the core density estimation method to estimate the spatial distribution density of new infections in Wuhan over a 1 km × 1 km grid cell using a standard Gaussian curve function (Equation (2)). There is no uniform standard for the optimal selection of the nuclear radius, which needs to be chosen according to requirements Cao et al. (2008). In this study, the size of the grid cell set by the GIS fishery network is 1 km, while the radius of the entire study area is only about 20 km.

$$k(u) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}u^2\right) \tag{2}$$

This can retain sufficiently detailed information and reflect the overall spatial distribution trend. The spatial incidence of the disease outbreak was obtained by dividing the spatial density of infected persons within the grid cells by the corresponding population density. Since the grid cells do not contain enough information to support the features required for network training, the morbidity midpoint grid points are again kernel densitized according to the above parameters, and a finer morbidity feature is fitted by reducing the output image element. The highest GIS classification of 32 classes under geometric discontinuity is used to achieve the smoothest state to obtain a sample of morbidity

label.

As can be seen in Fig. 2, the Wuhan New Crown epidemic was mainly concentrated in the dense road transport network, the complex, mobile population of Wuhan Railway Station, and the urban center with residential areas forming the core, which is the most economically active commercial area in Wuhan. The South China Seafood Market, where this outbreak first occurred, is located in the core of the high incidence area. As a whole, the spatial distribution of infections in Wuhan is clearly clustered, primarily in the core areas of Wuhan city — Jiangan, Jianghan, Qiaokou, Hanyang, Wuchang, Qingshan, and Hongshan. It is easy to see from the incidence rate (Fig. 5) and the incidence map (Fig. 6) that two locations had incidence rates exceeding 7.567 %, namely the area surrounding the Huainan Seafood Market in Jiangan District, and the Steel Flower Xincun community in Qingshan District. The highest outbreak density was reported at Steel Flower Xincun, where the incidence rate reached or exceeded 4.507 % in several locations around the area, and around the Guanggu Square in Hongshan District, along the river. The outbreak hotspots around the Wuhan Passenger Port and Wuhan Pass Wharf are secondary hotspots of the current outbreak, with incidence rates exceeding 4.507 % at their core locations. In contrast, the outbreak hotspots along the river are linked to the outbreak hotspots at the South China Seafood Market, which, when combined with the local transport network, are found to be highly influenced by population movement. When viewed in conjunction with the population density map of Wuhan (Fig. 4), it can be observed that there is no simple linear relationship between the disease outbreak and population density. The outbreak rate does not exceed 1.164 %. After research by the author and his team, it was found that spatial risk factors such as schools, shopping malls, underground stations, hospitals, and hotels were also important factors influencing the high incidence of the new coronary pneumonia epidemic, for which we conducted a modelling analysis of spatial risk factors and incidence rates.
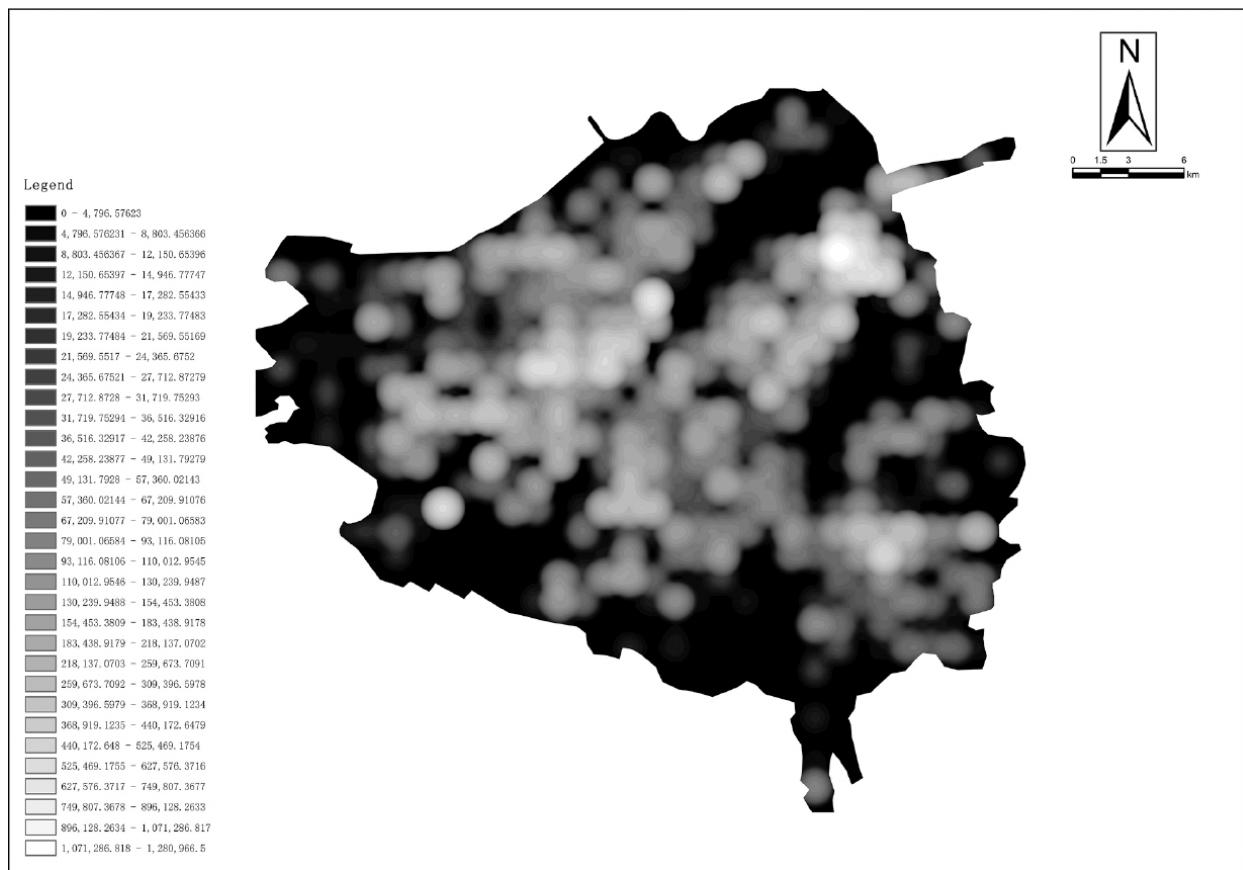
Fig. 6. Sample labels of incidence rate.

2.4. Spatial risk factors and model construction

According to Tobler's first law of geography, geographic space affects regional correlation, which is the basis for the spatial spread of epidemics. Scholars such as Cao et al. (2008), who have performed spatio-temporal modelling of infectious diseases, have pointed out that the spatial spread of infectious diseases such as SARS is closely related to factors such as population, the environment in which people live, and the spatial distribution of various other influencing factors. This paper draws on Li Xin et al.'s definition of spatial risk factors for the spread of infections, including population density, road traffic, the spatial distribution of various public facilities, and so on. The final degree of epidemic spread results from the combined effect of the driving forces for the infectious disease and these spatial factors in the surrounding environment.

A strong positive correlation was found between disease incidence and the densities of various spatial risk factors over the 591 grid cells that were used as a sample. The results showed that the two-tailed test with a significance level of p = 0.000 (<0.001) showed a significant positive correlation between each risk factor and the incidence of new coronary heart disease. The incidence rates in schools, supermarkets, metro stations, hospitals, parks and city squares, and around hospitals all exceeded 0.2 infections per 1,000, with the incidence rate in hospitals, in particular, reaching 0.383 per 1,000. These areas have the following two characteristics: high population density and high population circulation. Therefore, since hospitals are the primary locations for early outbreaks, where members of the staff are at higher risk of infection while treating patients, these are the areas with the highest incidence rates. While places such as metro stations, government and park squares have a high degree of overlap in the spatial distribution of outbreaks, especially since the locations of the highest outbreak points

in these areas are almost overlapping, it can be clearly observed that apart from the special nature of hospitals, the mobility of the population has a major influence on the spread of outbreaks in the city.

In order to obtain a more accurate correlation coefficient, this paper measures the correlation coefficient between the spatial risk factor and the incidence rate based on the grey relational analysis degree to obtain the correlation weight (Fig. 8). Therefore, this paper uses the above correlation weights with POI data distribution (Fig. 7) to build an urban spatial risk factor map with a resolution of 11871*12630 as a feature sample for GAN training (Fig. 8). By image processing of the incidence label sample (Fig. 6), the entire risk factor base map was cut into small images with a resolution of 512×512 using the PIL library in Python. This operation resulted in a uniform size for each image, which is a suitable format for machine learning methods, and obtained the input for the labelled samples. The feature samples for spatial risk factors were also obtained by intercepting the window of the area within the effective range of the Wuhan Third Ring Road in ArcGIS and overlapping it with the incidence label samples. The same cut image method was then used to obtain the input of the feature samples, with a total of 552 sets of 23*24 slices. The validation scope for the study was the Wuhan Third Ring Road urban area, where the regional boundary data was not highly reliable due to kernel density and interpolation analysis. After removing this data, a total of 275 sets of feature and label slices were obtained. Additionally, the uneven distribution of urban spatial elements resulted in the presence of a few features and labels in areas such as rivers and greenery. Therefore, unnecessary identification elements were manually eliminated and a total of 225 sets of slices were obtained.

In order to more conveniently describe the incidence distribution after the training model is run, the value domain of the incidence on the fine spatial cell was used to establish its weights and implement data vectorization in Rhino with the help of tools such as Grasshopper. This
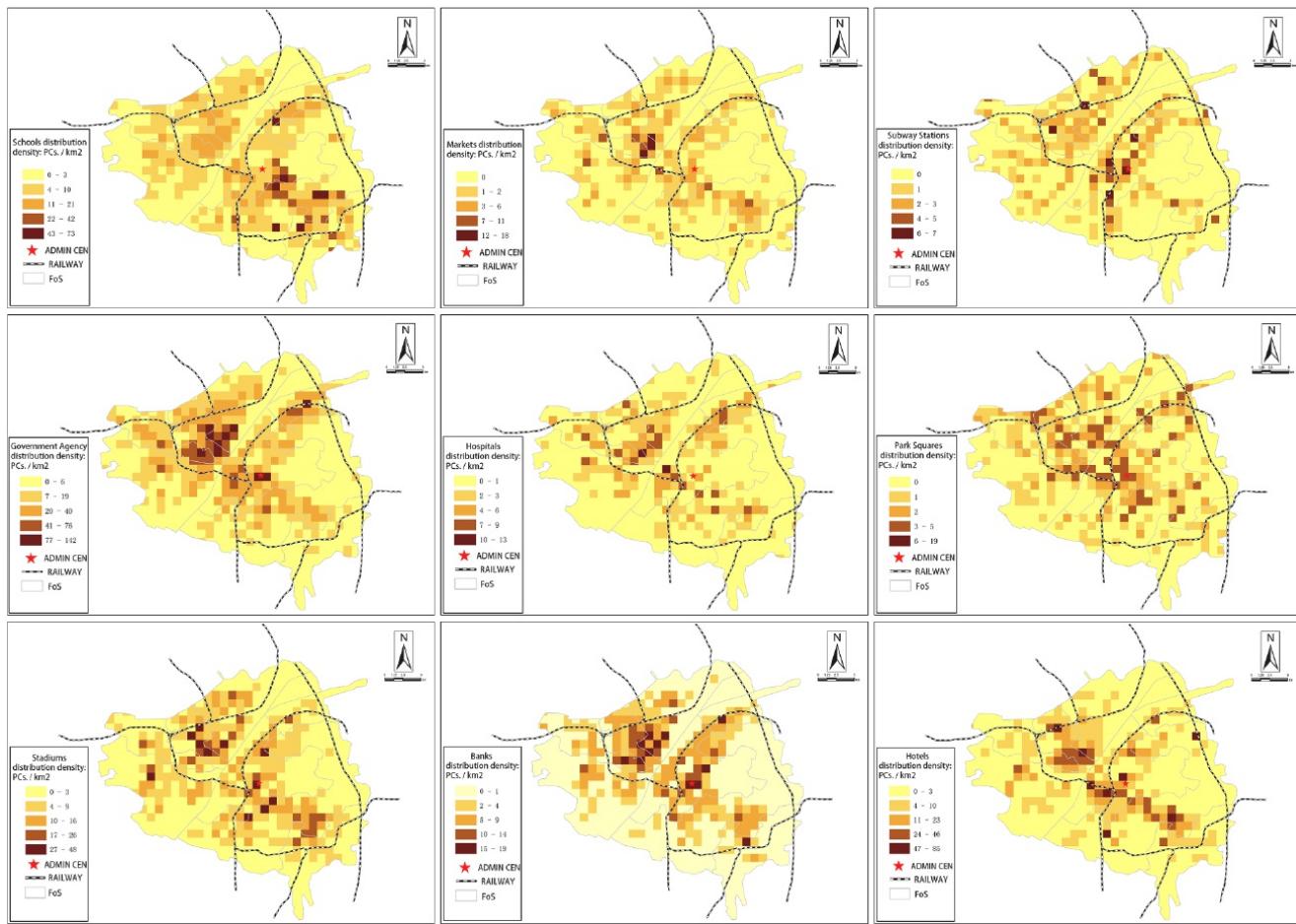
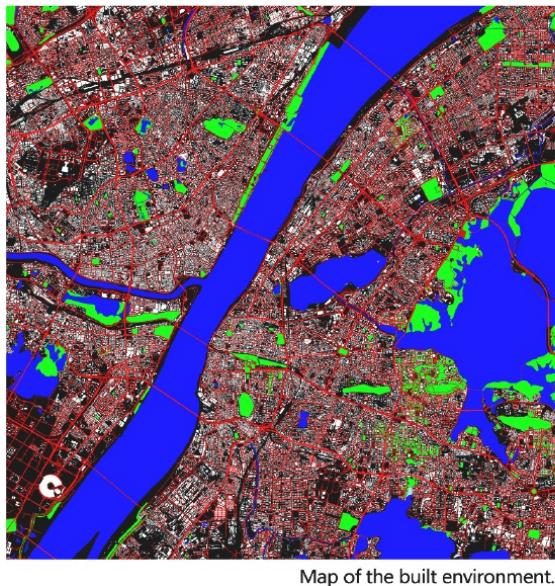**Fig. 7.** Density distribution of COVID-19 spatial risk factors in Wuhan.

Map of the built environment.

Correlation analysis between incidence rate of Covid-19 and spatial risk factors (pearson's ρ)

Risk Factor	School	Market	Bank	Stadium	Park Square	
Incidence Rate	0.209*	0.224*	0.120*	0.170*	0.257*	
Risk Factor	Hospital	Hotel	Government Agency		Subway Station	
Incidence Rate	0.383*	0.295*	0.188*		0.285*	

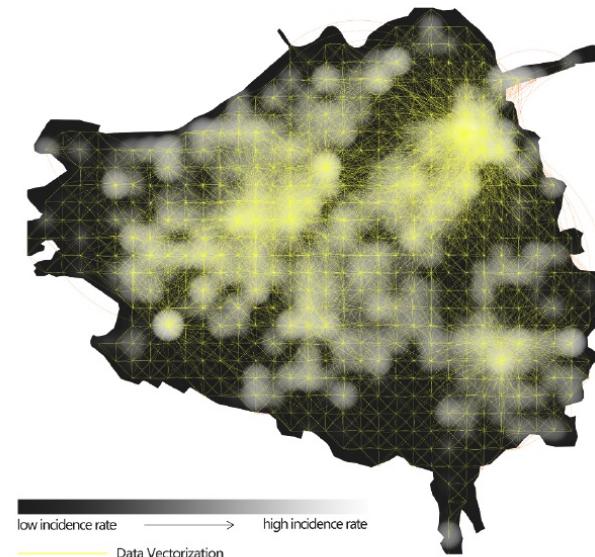Note: * means significant at the significance level of 0.01(two-tailed test)

Determination of Correlation Coefficient of Spatial Risk Factors Based on GRA

Risk Factor	Hotel	Stadium	Market	Government Agency	Bank
Coefficient	0.9149	0.9143	0.9142	0.9137	0.9134
Risk Factor	Hospital	School	Subway Station		Park Square
Coefficient	0.9094	0.9072	0.9017		0.8983

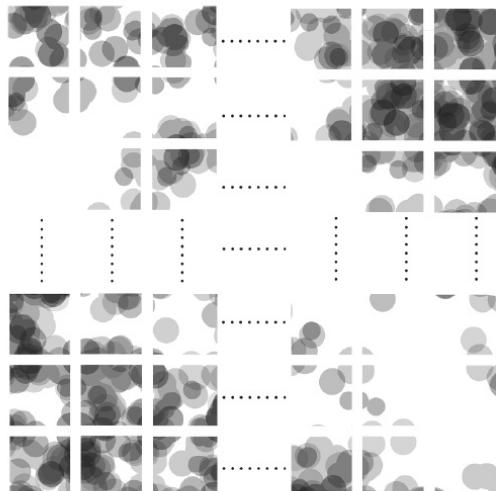**Fig. 8.** Measurement of correlation coefficient.

process (Fig. 9), by which the predicted urban incidence grey-scale map can be extracted from the grid values, enables a better visual representation in Grasshopper.
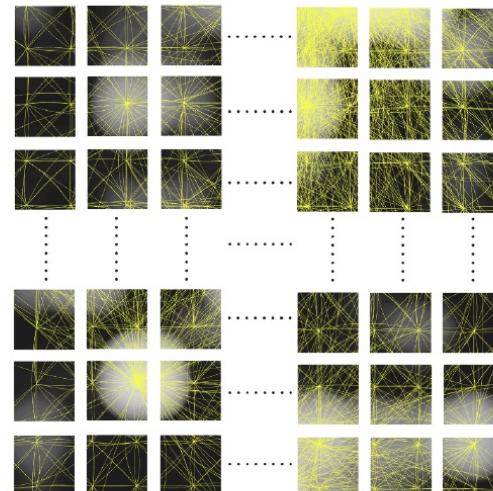
The risk factor map.

The incidence heat map.

Cut images of the risk factor map.

Cut images of the heat map.
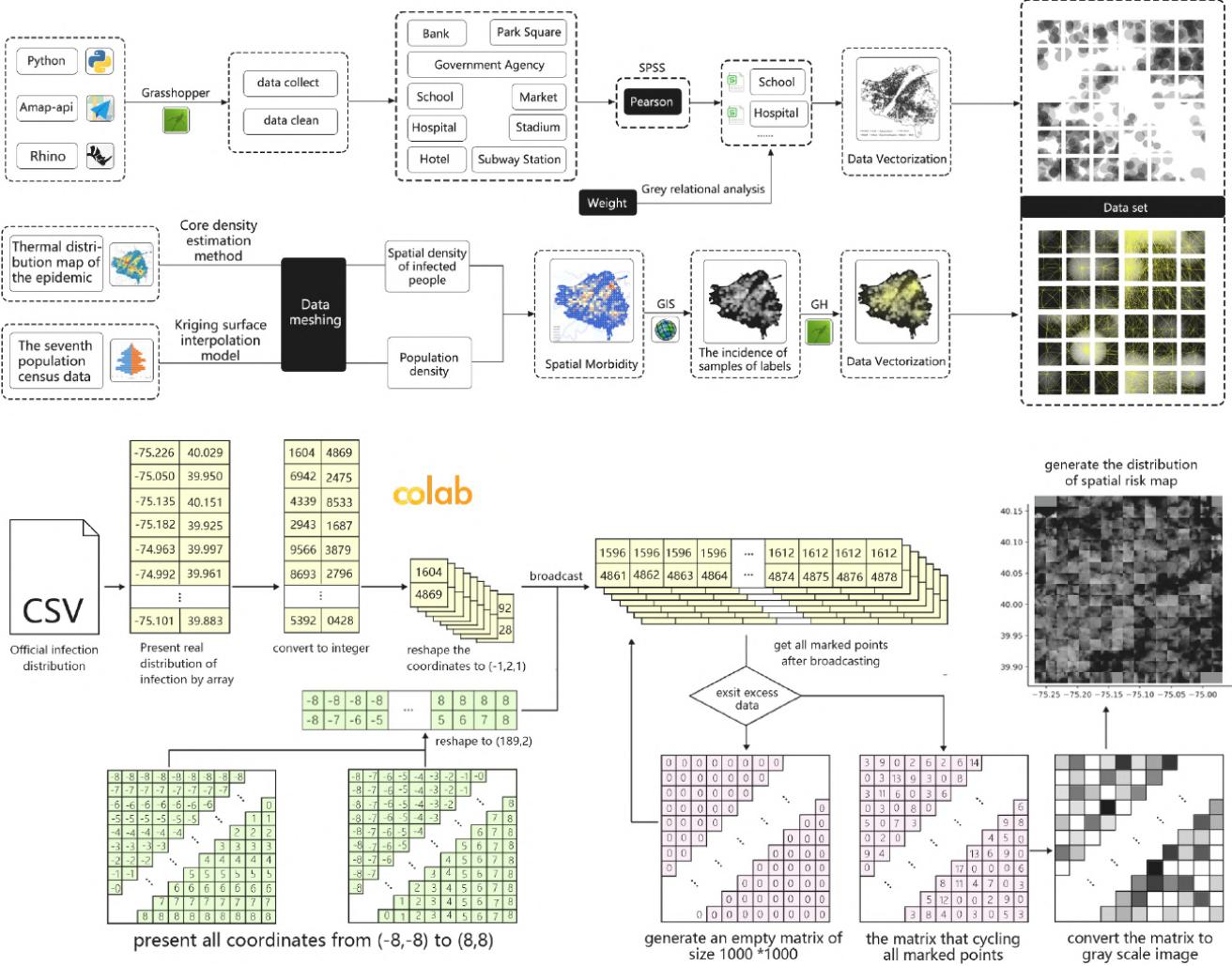
Fig. 9. Processing of features and labeled samples.

3. Machine learning

3.1. Model selection and data augmentation

The task of the machine learning model is to understand the relationship between the input urban facility layout and spatial risk information to determine the output epidemic incidence distribution, based on an image-based GAN framework with convolution and deconvolution kernels for training. Fig. 10 shows the data processing of the input and the output (He and Zheng, 2021). Conditional GAN as implemented by Goodfellow et al. (2014) and pix2pixHD (an open-source project) developed by Isola et al. (2017) were used to build the models for this study. Pix2pixHD implies a pixel-to-pixel transformation where the size of the input and output images remain constant. The adversarial neural network achieves the best generative results through a dual network of generator and discriminator adversaries. The generation model uses a training input transformation encoder, which converts the input into a parameter (code) and then trains an output transformation decoder, which converts the parameter into an image and then computes the

result. The decoder is then trained to convert the parameters into images, and the mean square error between the resulting images and the inputs is calculated. The generator feeds these generated results into the discriminator, which then feeds the loss and gradient to the generator. As a result, the generator is trained to generate false images that are closer to the true labels, while the discriminator is trained to better distinguish between false images, resulting in a cycle of good prediction results.

To apply pix2pixHD, it is necessary to define some important hyperparameters in addition to supplying the program with the image pairs described earlier. Firstly, there is no instance mapping; hence, the functions that read and use instance mapping should be turned off. Secondly, we want the program to use red-blue-green (RGB) colours directly as input, so we set label_nc to 0. Thirdly, during our experiments, we found that the first 70 training periods with a constant learning rate did not improve the network; therefore, we set the learning rate for the first 70 periods of the training process to a constant value but set it to a decaying value for the next 30 periods (Fig. 11 top). All other settings are the same as the default settings.

Fig. 10. Top: Point of interest (POI) distribution (input). Bottom: Incidence distribution (Output).

For the model training process, the input feature samples are spatially weighted distribution maps of various facilities in the city with spatial risk information, and the output label samples are disease incidence distribution maps generated by the kriging interpolation method. Since the data used are open source data for a single city in Wuhan, 80 % (179 sets) of the 275 sets of slices obtained are included in the training sample set, and 20 % (46 sets) are included in the test set to verify the accuracy of the model (Fig. 11 bottom). Therefore, the training set includes all areas of the Third ring urban area of Wuhan, and the images of the test set are identified to represent different areas of the city, in which two vertical columns and one horizontal column from the middle of the study area are selected to reduce the influence of uneven spatial elements, resulting in a total of 46 groups of slices.
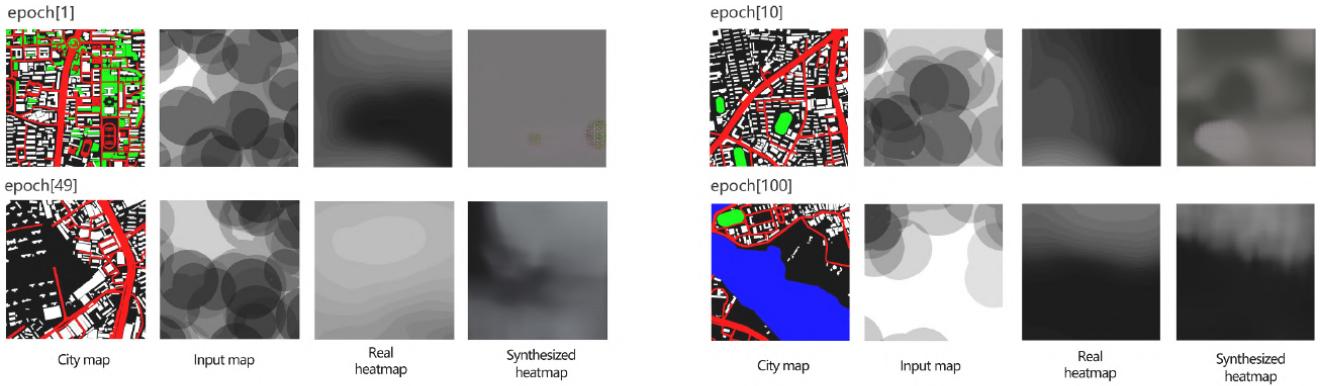
In addition, to solve the problem of the small size of data, the data enhancement strategy proposed by Zoph et al. (2020) is implemented. This form of automatic data augmentation can improve the generalization performance of the model on small training datasets. Considering that, the influence of the urban environment, such as wind in different directions of the base map, may be hidden in the high-dimensional features of the training set, different data augmentation methods such as Gaussian transformation, mirror operation, and adding noises are selected conditionally to enlarge the training set from 179 image pairs to 2148 image pairs. The machine learning models trained by the small and the augmented datasets are named training A and training B, respectively.
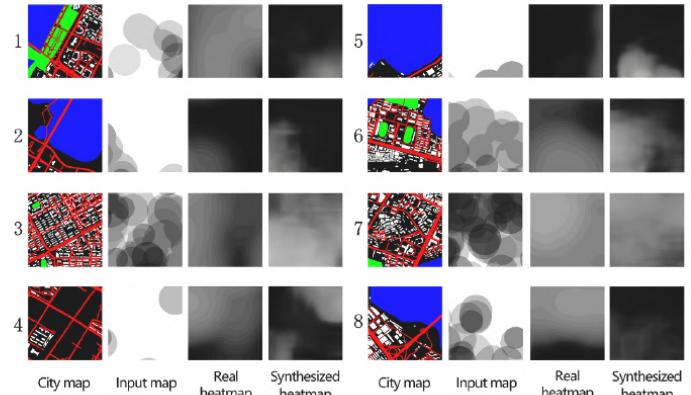
3.2. Accuracy analysis

GAN confrontation is marked by constant fluctuations and competition between the loss values of the generators and discriminators. There is no such thing as one side completely overpowering the other; constant confrontation between the two sides is a characteristic of a successfully trained and accurate GAN. The training process was also successful when the loss values of the generator were low and the loss values of the discriminator were relatively high (Zheng et al., 2020). It can be seen that successive optimisation models can successfully determine the relationship between the incidence heat map and the spatial risk map.

However, the loss values of the GAN models (Fig. 12) cannot directly reflect the training status, because the loss values of the generator and discriminator can only show that the training is being proceeded correctly. Therefore, we also record the generated images during each training epoch (Fig. 13) to visually evaluate the training process. By observing and comparing the generated images with the real images, it is determined whether the model can better meet the prediction of incidence and whether to stop the training (Zheng et al. 2020). Considering the overfitting problem, we train four models for final comparison, they are (1) the models using training A dataset with 100 epochs; (2) the models using training B dataset with 60 epochs; (3) the models using training B dataset with 100 epochs; (4) the models using training B dataset with 200 epochs.
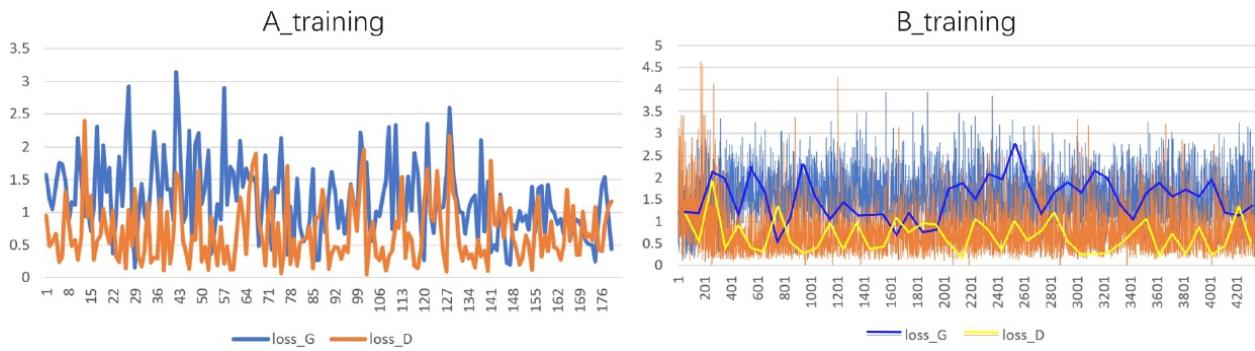
Next, to evaluate the accuracy of the images on the evaluation index,

Training images and synthesized images in four training periods.

Left: training area and test area; Right: example of prediction results.
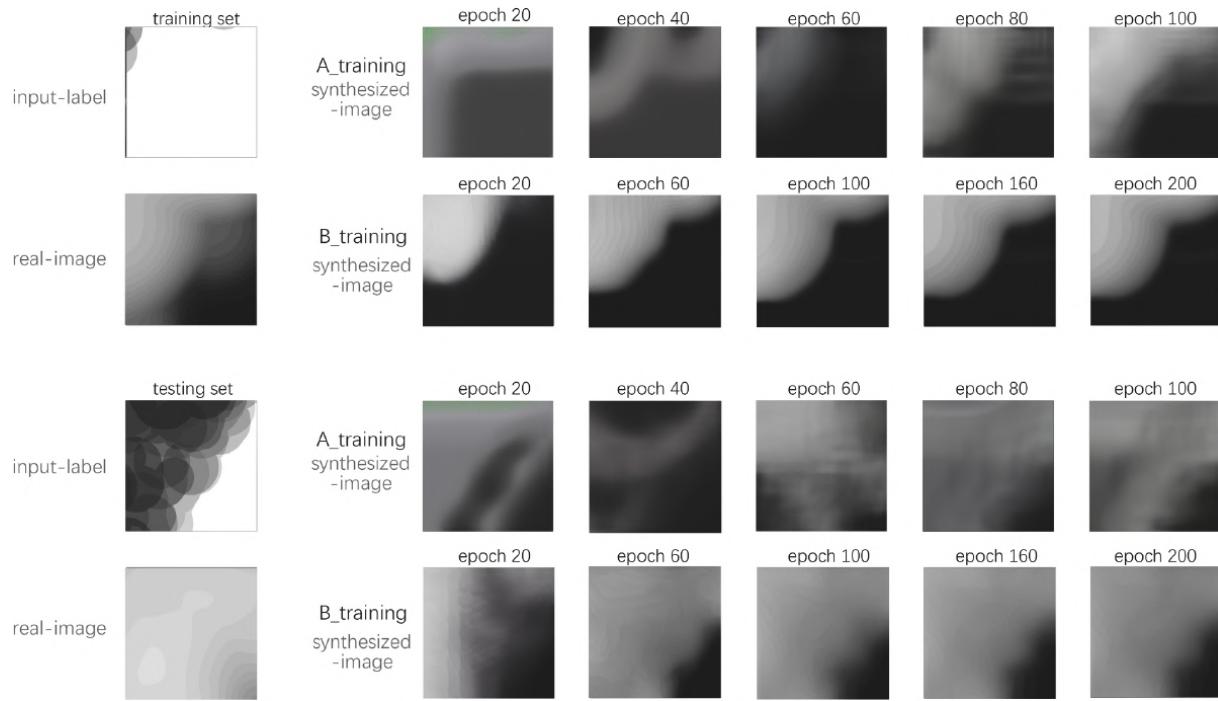
Fig. 11. Top: Result comparison of the A_training set in different periods. Bottom: the distribution of the training and the testing dataset.

Fig. 12. Generator loss (LOSS_G) and Discriminator loss (LOSS_D) during training A and training B.

some scholars have pointed out several representative sample-based GAN evaluation indexes (Xu et al., 2018). In order to better evaluate the performance of heat map fitting, we adopt Inception (IS) (Salimans et al., 2016), Frech et perception distance score (FID) (Heusel et al., 2017), structural similarity index metric (SSIM) (Wang et al., 2004), and perceptual metric (LPIPS) (Zhang et al., 2018) (Fig. 14).
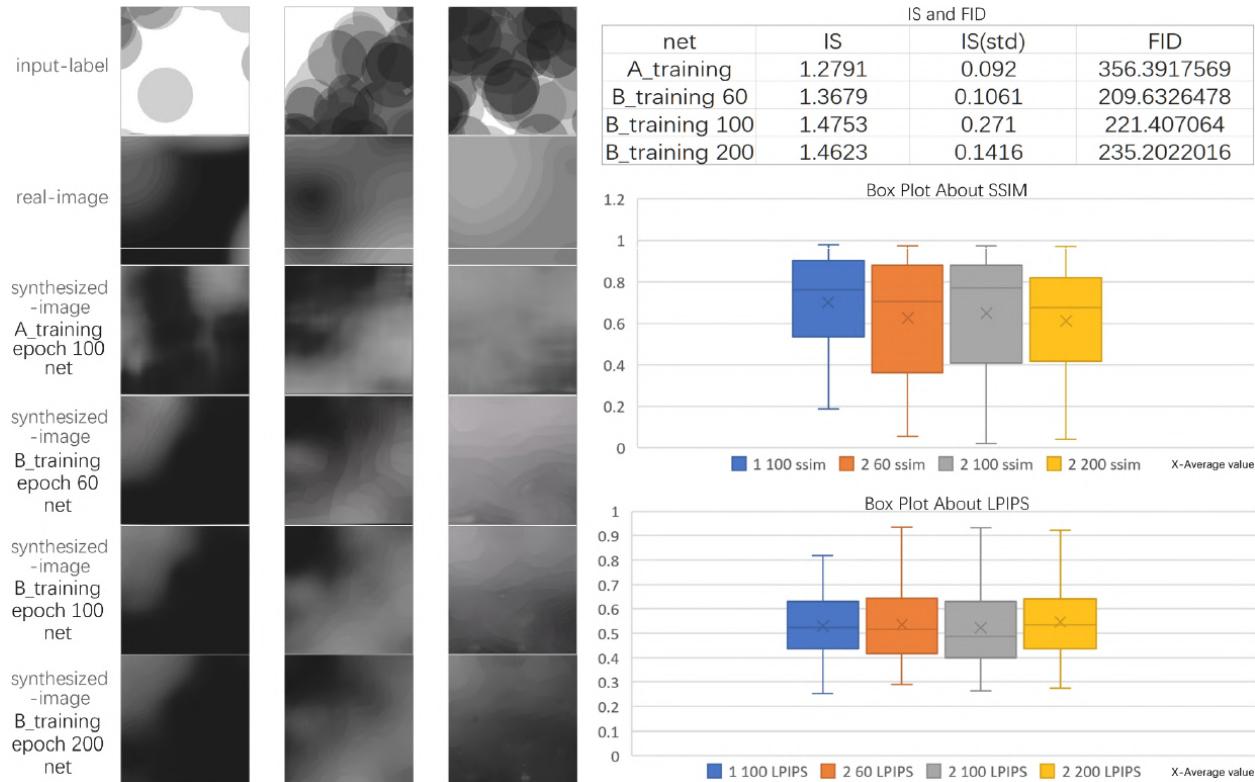
IS is the most widely used measurement method to evaluate the quality and diversity of generated images. It first calculates the edge distribution of generated images in each category, and then calculates

the KL divergence between the high-dimensional vector obtained by Inception model and the edge distribution calculated between them. Therefore, the higher the value of IS, the better the effect of the generated model. The IS index of training B at epoch 100 is 1.4753, which is larger than the other three test results. Also, the standard deviation of training A is the smallest, which shows the model of training A is more stable than other models. Therefore, from the IS point of view, training A and training B at epoch 100 are better.

FID is improved on the basis of IS, and has better robustness than IS.

Fig. 13. Training and testing image pairs of training A and training B.

IS and FID			
net	IS	IS(std)	FID
A_training	1.2791	0.092	356.3917569
B_training 60	1.3679	0.1061	209.6326478
B_training 100	1.4753	0.271	221.407064
B_training 200	1.4623	0.1416	235.2022016

Fig. 14. Four types of model tests (left) and four types of evaluation metrics (right).

FID uses the mean and covariance matrix to calculate the distance between two distributions. The smaller the distribution is, the closer the generated distribution is to the real image. The results show that training B at epoch 60 and epoch 100 are better.

SSIM is used to evaluate the similarity level between two images, ranging from 0 to 1. The closer the value is to 1, the more similar the images are. Observing the three-stage test box diagrams of training A and training B, it can be found that the test result of training A is the best, with the average SSIM reaching 0.7018.

Compared with SSIM, LPIPS is more in line with human perception, and the lower the value, the more similar the two images are. From the box diagram, the highest test value of training A is lower than that of the
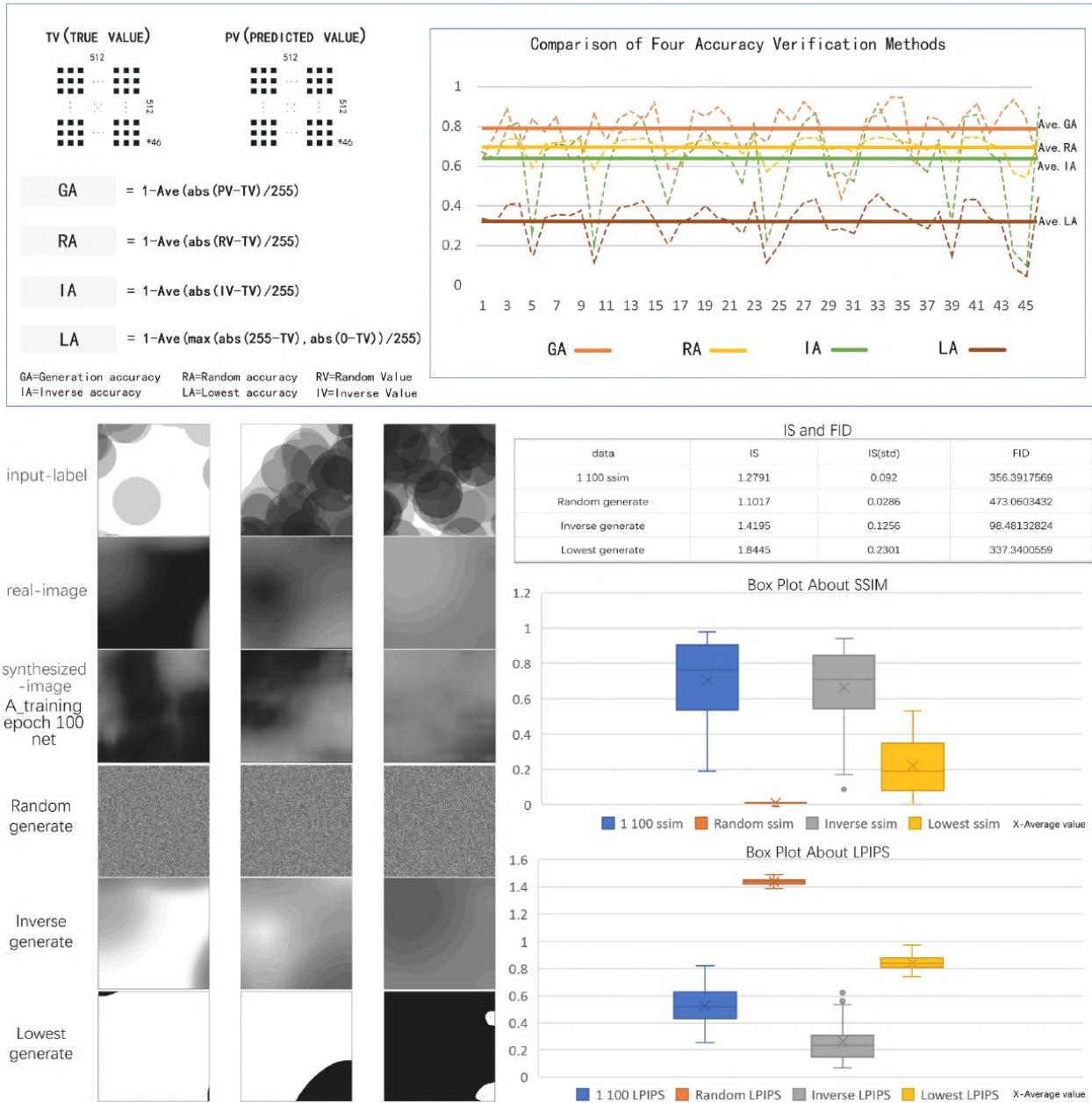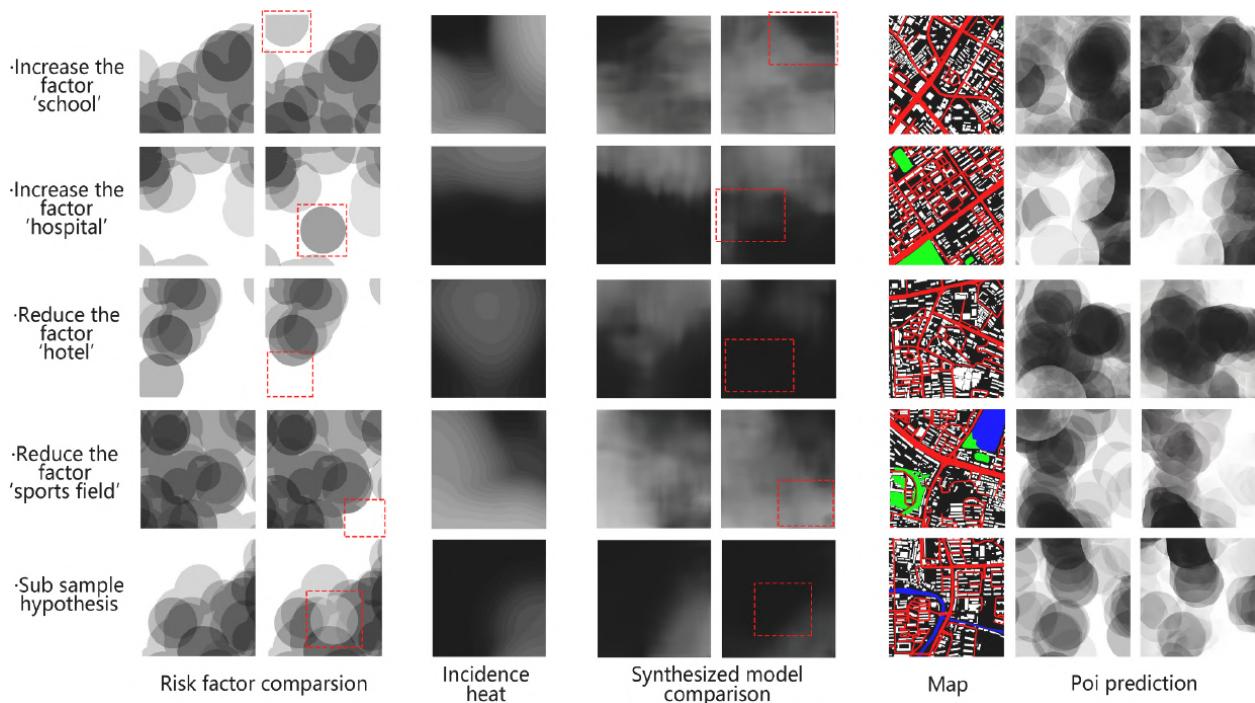
Fig. 15. Accuracy verification methods with extreme values.

other three models, indicating that most of the models are in line with human perception. From the average point of view, training B at epoch 100 performs best on average. Therefore, based on the above indicators, training A and training B at epoch 100 can be used as the best models, the former is relatively stable, while the latter has higher generalization ability through data augmentation.

Considering the above comparison of the accuracy, training A model at epoch 100 is selected as the final model of layout optimization for further research. Therefore, through the model of training A, we could find that the synthetic images performed accurately, with the disease heat map showing a clear pattern corresponding to the overlaid distribution of POIs with different densities shown in shades of grey at epoch 100. The predictive model at this point was therefore stored as the final model.

Next, four additional methods are used to verify the accuracy of the trained model (Fig. 15). First, we traverse the pixel value m (m = r*0.299 + g*0.587 + b*0.114) for each pixel of 46 groups of samples (of

512*512 resolution) in the entire test set using the getpixel method in the PIL library. We then obtain the average difference between the predicted value and the true value of all pixel points in each group of slices. The average generation accuracy (GA) of this training model test set was 0.7895. This method demonstrates that the absolute difference between the predicted value and the true value is small. Second, to test whether the model is a good fit to the spatial distribution, we set randomly generated pixel values to compare with the true value, and obtained an average random accuracy (RA) of 0.6959, which shows that the predicted value of the model is significantly better than the random scenario. Third, after inverting the true pixel value, we calculate the difference between the inverted value and the true value. We obtained an average inversion accuracy (IA) of 0.6401. This indicates that the predicted values are entirely false as compared to the real value. Finally, we take the maximum difference between 0 and 255 and the true value to maximise the relative error rate, and get an average lowest accuracy (LA) of 0.3201. The comparison between GA, RA, IA, and LA is shown in

Fig. 16. The interplay between urban spatial risk and urban design.

Fig. 15 top, in which the accuracy of our generated values is much higher than that of the comparative groups. When the mathematical expectation of RA is regarded as 0.5, and the IA is regarded as 0, the predicted value of our model increases by 167.74 % as compared to the random guess.

Besides, for the three generated images (random-generated, inverse-generated, and lowest-generated), we implement the IS, FID, SSIM, and LPIPS to double-compare the accuracy (Fig. 15 bottom). We re-measure the similarity between the dataset generated by extreme values and by training A model, comparing with the real values. As a result, from the perspective of IS and FID, compared with the random-generated images, the machine-learning-generated images increase the IS by 16.1 % and decrease the FID by 24.7 %. Also, according to SSIM, the highest and average values of the machine-learning-generated results are much higher than those of the three extreme cases. The distance between the machine-learning-generated results and 1 is increased by 332 % compared with the random-generated results. From the perspective of LPIPS, the machine-learning-generated results also decrease the values by 62.9 % compared with the random-generated results, which shows that the structure of our machine-learning-generated images conforms to human perception.

3.3. Urban layout optimisation

With the help of the trained model that can predict spatial risk using neural networks, we can continuously adjust design elements to optimise urban design plans for specific areas and to achieve the lowest public health risk (Fig. 16). Changing the spatial attributes of the POI source data and then testing the model reveals that by increasing the number of schools and hospitals and decreasing the number of hotels and sports grounds, the predicted disease incidence rate can be changed. In addition, assigning virtual negative correlation POIs also corresponds to decreasing incidence rates. This indicates that this model learns better a particular mapping relationship between POI distribution and incidence rates and determine the various risk factors weighting this relationship.
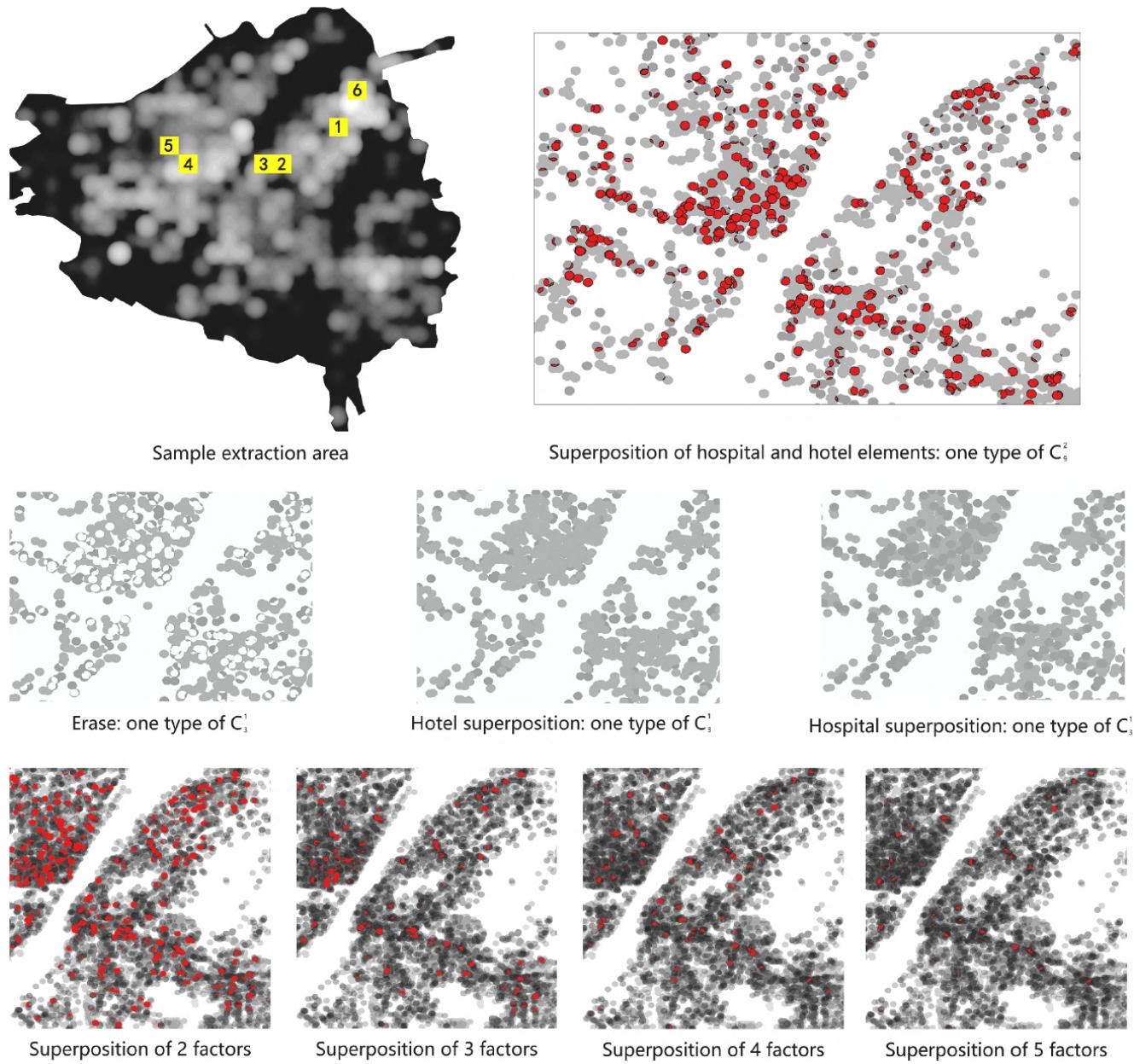
Based on the results of the Pearson test above, the correlation coefficients can be ranked as hospitals, hotels, subways, parks, shopping

malls, schools, government, sports fields, and banks. The proximity of hospitals and hotels does not necessarily indicate the impact of these facilities. The machine learning model can reveal the overall complex impact more accurately, while lacking the visualisation of individual elements in the output.

In 2018, scholars studied the association between the degree of population concentration and behavioral activities in Wuhan, China, through POI. But the method studied a single variable and was not suitable for constructing links between large-scale outbreaks and multiple environmental characteristic factors (Peipei and Yinghui, 2020); and one year after the COVID-19 outbreak in Wuhan, scholars mapped the accessibility of medical facilities in Wuhan (Zhou et al., 2021). WU analysed the impact of multiple factors on urban layout by means of kernel density analysis in a study on the spatial structure of polycentric cities in Guangdong, China (WU et al.,2020); in contrast to the former, the approach mentioned in this paper analyses the risk of outbreaks through the analysis of risk factors such as hospitals, parks and underground stations on the Compared to the former, the method mentioned in this paper is more accurate and credible by analysing the weights of risk factors such as hospitals, parks and metro stations on the risk of an outbreak, using a neural network model for prediction validation and optimizing the model with the results of error analysis.

This paper, therefore, examines the relationship between each type of element by means of a categorical element intersection transformation (intersected to remove overlapping influences). The relative independent influence of each type of element is examined so as to provide an optimisation strategy for the spatial layout of the corresponding single or multiple facilities. Some scholars have already quantified the impact of various urban facilities on urban life, and the degree of impact of various types of POI facilities varies under different circumstances. This direction of research can be taken further by combining the existing quantitative analysis and research with more precise measurement of the epidemic in the post-epidemic era.

In order to maintain consistency in the range of urban spatial risk factors in the training set, the 200 m range of each type of element displayed in the previous section continues to be used. It can be seen from Equation (3) that the distribution state N has 502 superimpositions of two to nine categories (with the presence of co-linear influences), and

Sample extraction area

Superposition of hospital and hotel elements: one type of C_9^2

Erase: one type of C_3^1

Hotel superposition: one type of C_3^1

Hospital superposition: one type of C_3^1

Superposition of 2 factors

Superposition of 3 factors

Superposition of 4 factors

Superposition of 5 factors

Fig. 17. Correlation superposition classification analysis of elements.

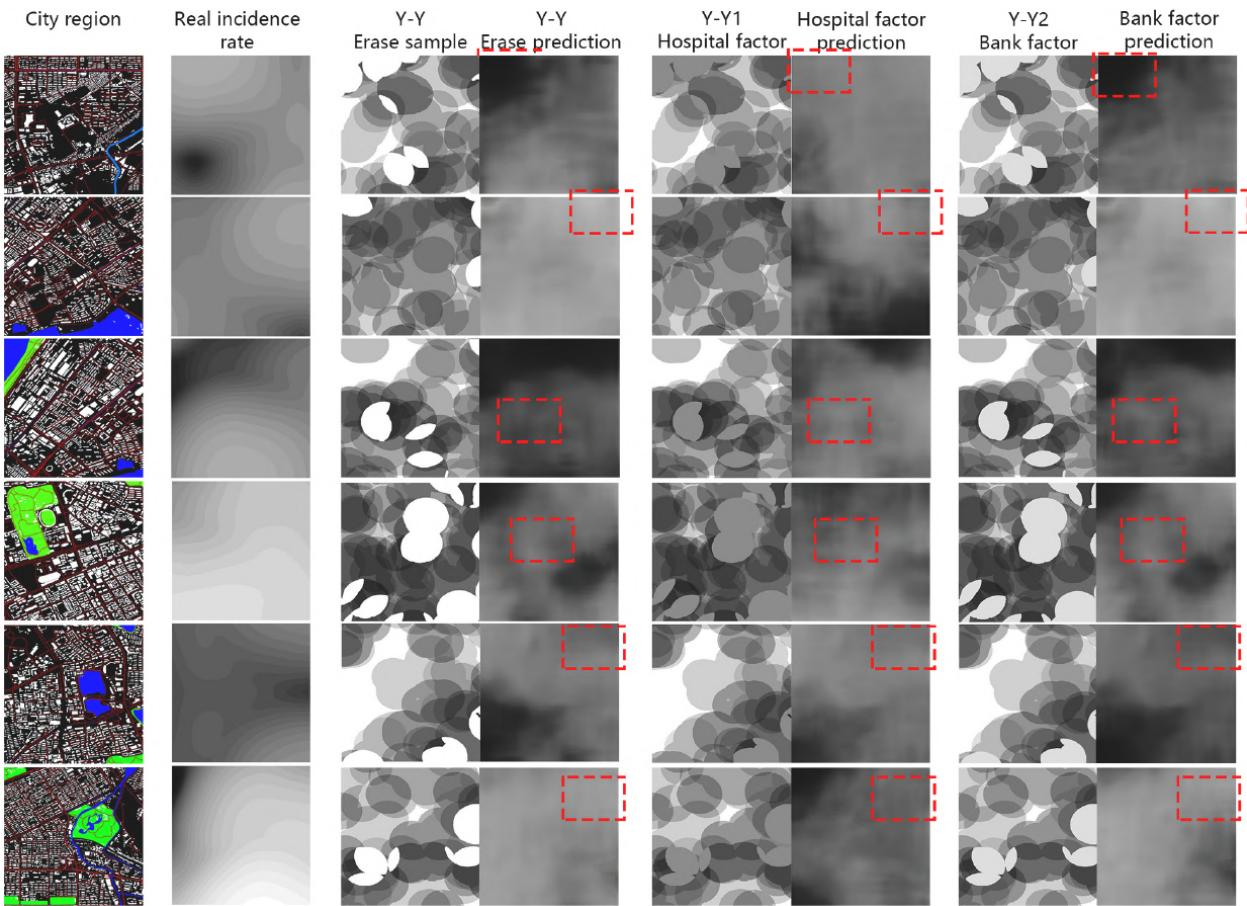each superimposition state has a kind of existence: erasure, factor 1, factor 2…factor n (Fig. 17).

$$N = \sum_{1}^{9} C_9^n * C_{n+1}^1 \qquad (3)$$

In order to make the maximum possible relative change to enhance the level of analysis, hospitals and banks, which have the largest correlation gap, are selected as the impact test samples in case of two categories of spatial risk factor elements. Hospitals, shopping malls and banks are used as the three categories of spatial risk elements, because shopping malls can act as the intermediate elements to enable the co-linear areas to reach the maximum relative difference. Similarly, hospitals, subways, governments, and banks are selected in the case of four categories of spatial risk elements, and hospitals, subways, shopping malls, governments, and banks are chosen as the five categories of risk elements. For more than six categories, there are fewer common lines, and the complexity of the area is increased, so they are discarded. For ease of calculation, the two common elements of hospital and metro are

taken here as the overlay pattern except for erasure, and the erasure pattern is used as the test control group.

The sample selected for analysis is located on the far left-hand side of Fig. 18, from top to bottom, in Qingshan District, Wuchang District, Wuchang District, Jianghan District, Jianghan District and Qingshan District of Wuhan City. Of these, the urban area located in Wuchang District was the primary outbreak site for the current epidemic, and the true outbreak rates for the six locations were extracted in conjunction with the results of the analysis above (Fig. 18).

We tested overlaps between the two risk factors, hospitals and banks, after completing the erasure process, and found that the predicted results of one location using banks as the overlapping risk (i.e., the overlapping portion uses the spatial risk correlation represented by banks) are consistent with the results of the erasure test. This indicates that banks are less spatially risky here, while the use of hospitals is close to the true outbreak rate, indicating that hospitals are riskier. The effect of the spatial risk set, which can be seen from the urban base map, is mainly distributed in residential areas here, suggesting that hospitals are at a

Fig. 18. Two-class element superposition mode.

higher risk of possible epidemic transmission for residential areas.

In contrast, analysis sample 2 for Wuchang district demonstrates the instability of the machine learning model. The prediction model that uses hospitals as the stacked risk differs significantly from the true incidence distribution as compared to the erasure model, while the model that uses banks as the stacked risk is closer to the real scenario. It is evident from the bottom panel of sample 2 that the current model is insufficient for making predictions in regions such as water systems and greenery due to the settings of the training set and region selection. In order to further improve the accuracy of machine learning of epidemic incidence rates, further multi-dimensional natural environmental elements such as ventilation environment and the characteristics of street layouts such as street openness can be input as a supplement to spatial risk, and more urban samples and regional data can be collected and used according to the method described in this paper for training new models.
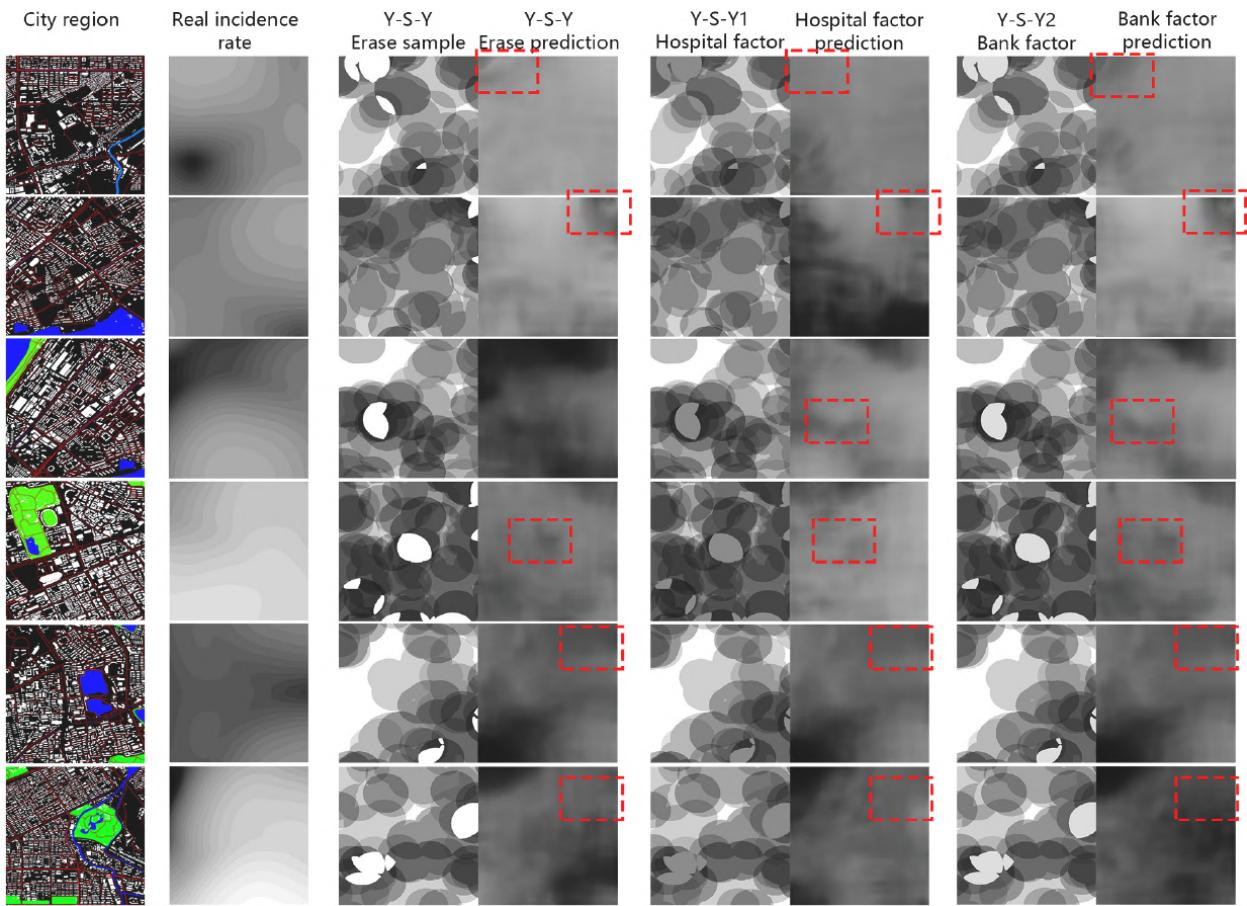
Similar to the results of the two types of factor overlay analysis, hospitals and banks still exhibit the largest relative differences in the predicted results for the overlaid areas when shopping malls are used as intermediate factors. This phenomenon is more pronounced in Sample 1 (Qingshan District) and Sample 2 (Wuchang District), which may be because these are business districts, a factor that significantly influences the study population. In particular, the Wuchang District sample, as the first outbreak area, was more significantly influenced by factors such as hospitals, banks, and shopping malls along Xinhua Road, corresponding to the predicted results as marked in Fig. 19. When the results of the different categories of analysed elements are compared, sample 2 shows a different distribution of outbreaks using the test results of the hospital overlay risk. The difference between the two is that the areas overlaid with hospital and bank data in the sample 2 test in the two-category

overlay analysis corresponds to the overlay of all risk sets in the three-category overlay, indicating that the spatial risk elements excluding hospitals and banks in this location have a greater influence on the distribution of outbreaks. This could be due to the influence of the distribution of metro stations and shopping malls.

Additionally, by observing samples 2 and 3, it was found that the water system and green areas were not directly affected by the POIs directly related to the distribution of the area. However, changes in the superimposed model had a significant impact on the disease incidence in such areas. Sample 3 erasure tests performed particularly well because the training set removed a large number of characteristic samples of water systems and green areas to modify the urban ventilation environment. One of the unspecified routes of transmission of NCCV is aerosol transmission, which is largely dependent on air movement.

On the other hand, as the scope of impact of urban facilities varies, the parks and rivers in samples 3 and 4 aim to serve as recreational support for the local business district, and removing the impact of such facilities would influence this support. Therefore, based on the analysis of the superposition of the two types of elements in the first three samples, it is necessary to strengthen the input of multi-dimensional and multi-scale spatial risk elements in order to improve the accuracy of machine learning. To reduce disease incidence and strengthen the management of urban infectious diseases, urban planning needs to reconsider the proportion of different facilities, appropriately decrease the distribution of dense facilities and control the service support of various types of such urban service facilities.

Through the superposition analysis of five factors, namely, hospitals, subways, shopping malls, government and banks, it can be observed that there is still a collinear relationship among these factors, which is more evident in the research samples of Qingshan District and Wuchang

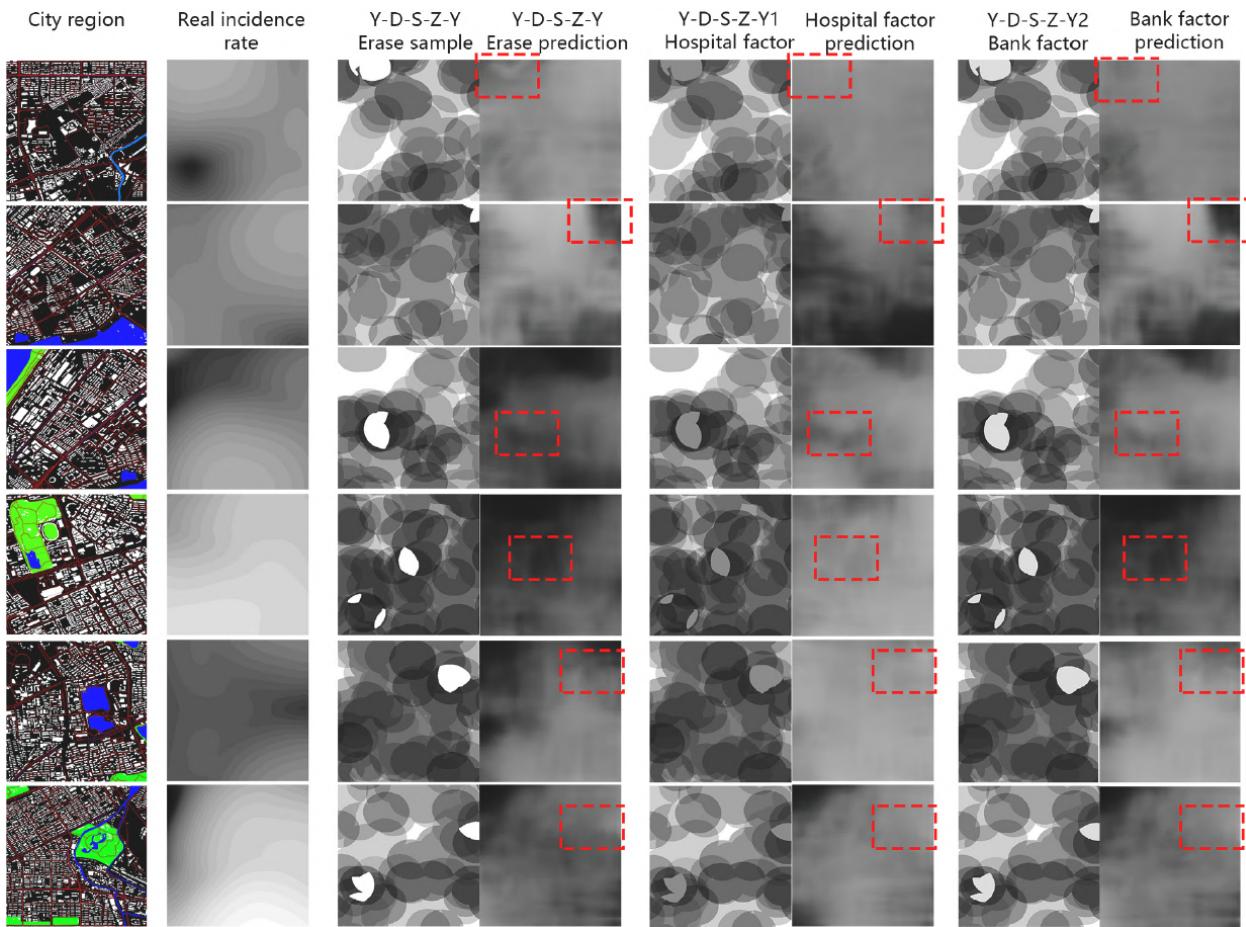**Fig. 19.** Three-class element superposition mode.

District (Fig. 20). Corresponding to the map, it is the area where the epidemic is concentrated — near Dacheng Seafood Market and Ganghua New Village. However, the collinearity characteristics of other locations are no longer as remarkable as that observed on superposition of two types of elements and three types of elements, which may be caused by changes in local epidemic prevention and control policy (in this case, 3 January 2020, when Wuhan closed the city) during the outbreak period. This would mean that the primary influencing factors of the epidemic are now human activity and interactions. At the same time, by comparing the three types of superposition patterns of these elements, it is observed that the results of the erasure test are getting closer and closer to the real prediction. For example, in the upper left corner of sample 1, the erasure of different types of superposition patterns of elements in the same position produces different results, which indicates that the spatial risk factors at different positions have different interactive influences. This may be due to the different central positioning of the influence of urban facilities. For example, the two erased factors in sample 1 contain commercial facilities of different levels, and there may be a commercial complementary relationship between them. Different hospitals, subway stations, banks, etc. also have similar complementary and intersecting influences. Therefore, the machine learning model is more suitable for making accurate predictions, which is difficult to achieve by using linear regression when facing the problem of developing a disease prediction model in complex urban systems.

3.4. Model refinement

The geographical relationship between urban POI facilities (spatial risk factors) and the incidence of new coronaries was mainly constructed through precise numerical values as shown above, which is intuitively not good enough to derive morphological correspondence from the urban base map. The geographical relationship between the Wuhan base map and POIs was also tested. The model was trained again using the above training parameters, and it was found that the prediction of POIs from the base map of the city resulted in better fitting and passed the accuracy validation method above. This suggests that we can use the urban base map for morbidity prediction and to establish a mapping function between urban texture and new crown morbidity that can be commonly used by urban planners for creating more efficient designs. However, although the direct prediction of incidence from the urban base map improves the study's intuitiveness and ease of correspondence in the field of design practice, such maps may lack a precise definition of the characteristic inputs that is necessary derive a quantitative identification of risk elements. Fig. 21 reveals the optimisation possibilities of combining multiple learning models in a logically closed loop, which can verify the interactions between urban environmental risk and urban design. Modifying the urban base map makes it possible to predict epidemics and thus iterate urban design quantitatively.

In order to investigate the potential applications and practical value of the model in the field of urban design, iterating urban design through a predictive model of the relationship between urban spatial risk and epidemic transmission would be possible in planning and design by combining two models from the two constructs above (prediction of POI distribution with spatial weights from the city base map and prediction of COVID-19 incidence from POI spatial risk factors, hereafter referred to as the iterative model). To verify the generalisability of the iterative model given the strict policy control and traffic regulation in China, incidence rates can be predicted for several representative cities in China, such as Xi'an, Hong Kong, and Shanghai. Of these, Shanghai best fits the application scenario of this model. Most Chinese cities have

Fig. 20. Five-class element superposition mode.

performed relatively well in terms of suppressing the epidemic, but as of 5 March 2022, there have been small to medium-sized aggregated outbreaks in cities of varying sizes. The main causes of occurrence are more complex in the major cities, where the implementation of prevention and control policies and socio-demographic movements are considered to be broadly unchanged. Xi'an experienced a cluster outbreak in late 2021 with thousands of cases, occurring in a setting and context broadly similar to the Wuhan outbreak in 2019. The well-developed financial and foreign trade and good internal and external mobility in Shanghai may have allowed for fluctuations in disease incidence as well as both intermittent and sustained outbreaks. Hong Kong's social environment is closer to that of foreign countries, with a higher density of urban building and greater socio-demographic mobility, allowing for a more generalised case to test the fit of the iterative model and suggest room for improvement.

3.5. Predictive application

To practical effectiveness of the main model constructed in this study (Table 1, row 1) was tested with a case study. The model was used to predict the distribution of the disease outbreak in Shanghai (Fig. 22). First, POI distribution data was collected for the central city of Shanghai, the pixel point greyscale were collected by using the PIL library, the distribution values were counted and the predicted incidence value range was established using Grasshopper. The previous incidence label sample data vectorisation was then used to perform disease incidence mapping for Shanghai. It was found that the two flood plains across the river are the likely areas of concentration of disease incidence in the event of an outbreak, which is close to the concentration of the

distribution of an epidemic in this area of the city in the early 20th century (Xujiahui Garden, Houtan Park, Shanghai Station, and other areas) and illustrated the predictive fitting ability of the model.

4. Conclusion

In the context of today's global mega-scale, ultra-high density, ultra-high frequency, and large-scale urban mobility, new conflicts have arisen between urban development and disease prevention. The emergence of COVID-19 poses an urgent challenge to the existing planning systems of cities around the world.

This paper aims to analyse the distribution of the new crown epidemic at the city and municipality level in China from the perspective of geographically diverse facilities, integrating an analytical framework that includes generating a heat map of the epidemic by using nuclear density analysis, a general kriging interpolation model for population density analysis, and a spatial risk factor distribution through Pearson correlation indicators. The analytical framework can be further improved using multiple sources of big data to study the spatial and temporal characteristics of the epidemic (Yang et al., 2019). The proposed framework combines visualisation methods, correlation analysis, spatial interpolation models and machine learning techniques. It can be extended to other areas of spatial epidemiological research.

In this paper, we chose the GAN model instead of the ANN or CNN model to adapt to the characteristics of this type of spatial image generation needs. For the ANN or CNN model, to predict the distribution of the heat map, we need to convert image features into a numeric vector formed by pixels and RGB values, which makes it difficult to converge. Besides, in the case of reforming the output as one numeric value to
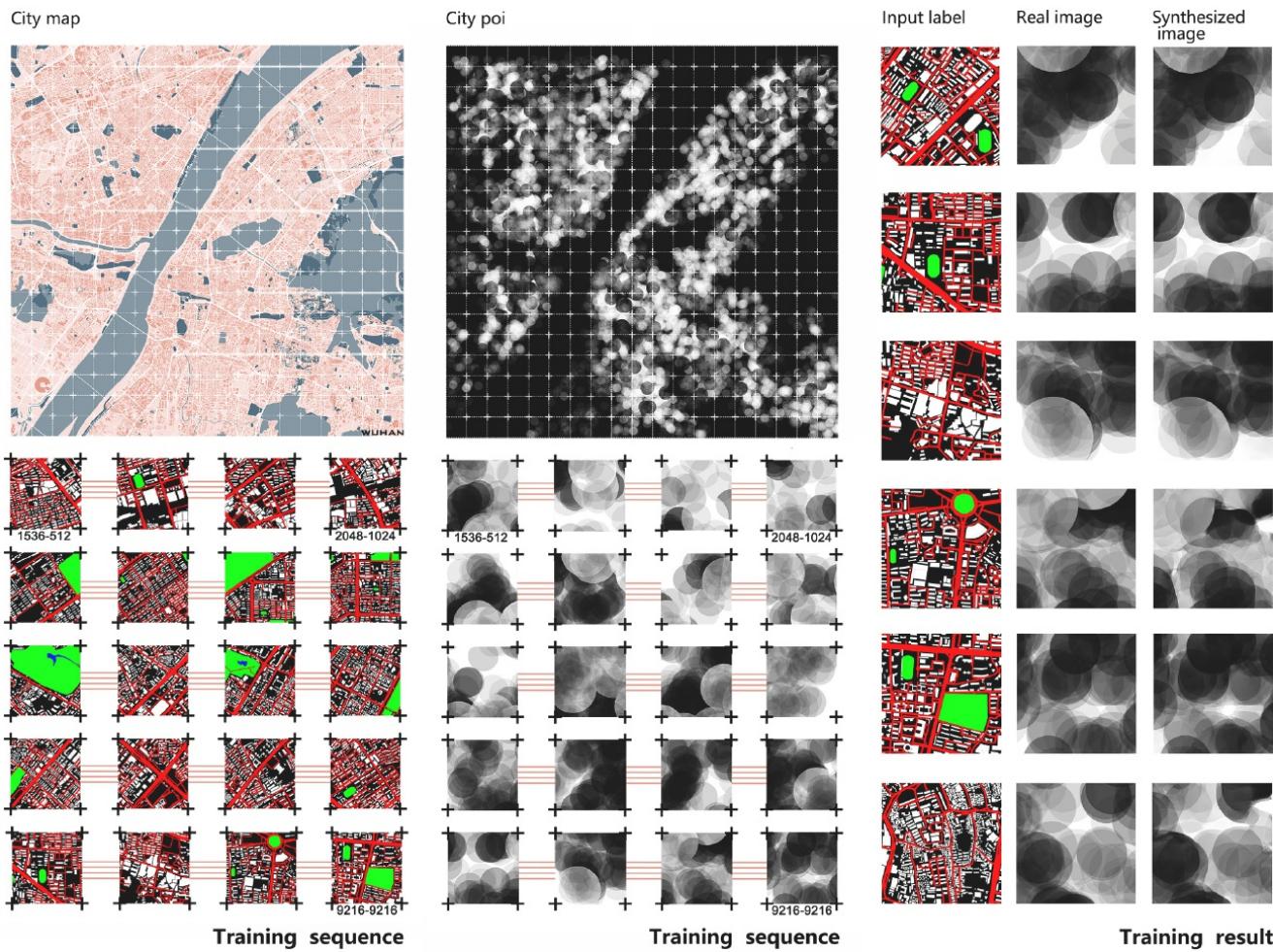
Fig. 21. Predicting POI facilities based on urban maps.

Table 1

Comparison of different forecasting processes.

Prediction mode	Advantage	Disadvantage
City poi (spatial risk) - Incidence rate	Accurate data	Lack of intuition and need data preprocessing
City map - Incidence rate	Intuitive display	Lack of explanation and the degree of risk cannot be measured
City map - City POI - Incidence rate	Accurate data and intuitive display	–

indicate the sum of the grayscale values, the result can reflect the spatial distribution of this type of geographic problem. Thus the image-to-image GAN model can best meet the requirements with affordable training costs and limited size of data.

Suitable neural network models were developed from Wuhan epidemic distribution data and census plot-level data to explore and validate the model as follows: (a) The spread of an epidemic is due to the interactions among the driving forces of the infectious disease and the multiple spatial factors and movements of the human population in the environment. (b) The top three urban facility types in terms of increasing infection are hospitals, hotels, and subway stations, which address the shortcomings of subjective evaluation and the difficulties of data acquisition as compared to previous methods. (c) The generated machine learning models could be extensively used for big data, as well as provide new capabilities for studying environmental behaviour. The geographical relationship between urban facilities and the onset of

epidemics can be learned by neural networks and used to predict the spread of epidemics in other cities. (d) By building multiple machine learning models, not only can the principles of interactions and determinants of behaviour be derived, but real-time feedback on the assessment results can also support urban designers in improving their designs to achieve optimal safety in the urban environment with respect to disease outbreaks.
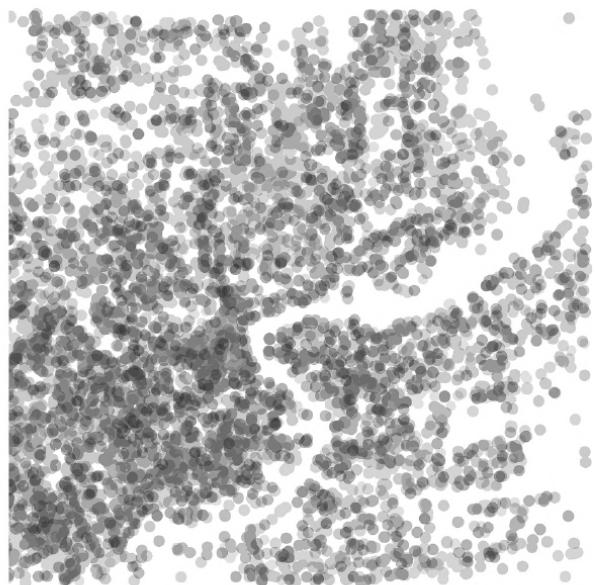
This study proposes a new approach to risk assessment based on fine scales, which can provide new ideas for future disease risk assessment, decision support for epidemic prevention and control, and security for the people. Future research would be based on a multi-scale epidemic prevention system, and existing systems can be improved by incorporating environmental elements such as ventilation and sunlight, and social elements such as population movement in predictive models. This would help to generate accurate digital maps of epidemics and transmission models for scientific and effective prediction and prevention and control of epidemics. At the same time, the symmetric cross entropy method proposed by Ye et al. (2021) can be used to further improve the performance of the model, and SL can be integrated into the existing training to solve the problem of under-learning and over-fitting in the presence of noise labels.
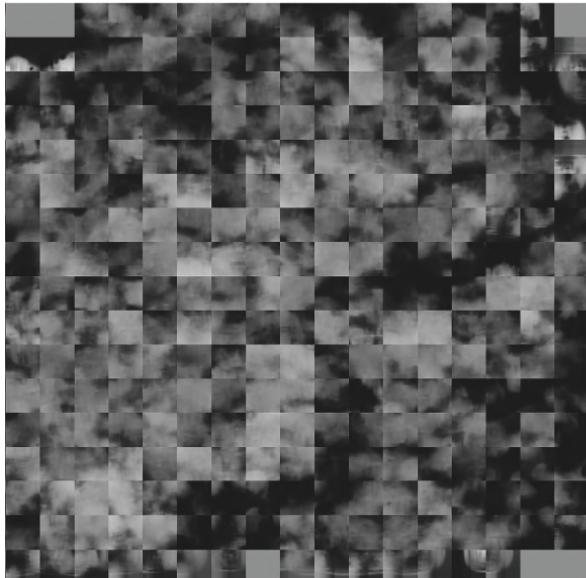
CRediT authorship contribution statement

Yecheng Zhang: Writing – original draft, Software, Visualization, Data curation. **Qimin Zhang:** Software, Writing – original draft. **Yuxuan Zhao:** Formal analysis, Conceptualization. **Yunjie Deng:** Investigation, Writing – review & editing. **Hao Zheng:** Methodology, Project
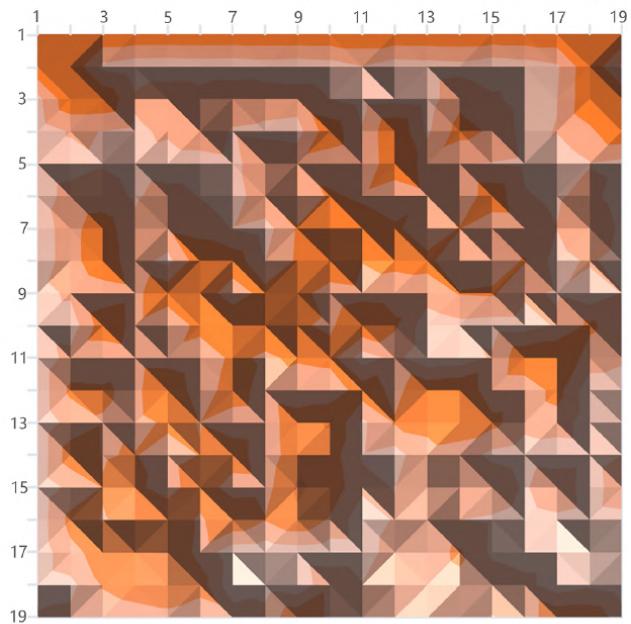
The built environment map of Shanghai.

The risk factor map of Shanghai.

The incidence heat map of Shanghai.

The vectorized incidence heat map of Shanghai.

Fig. 22. Using generative models to predict the incidence of Shanghai City.

administration, Supervision, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

The code and datasets of the research is available at:
https://github.com/24kchengYe/Urban-spatial-risk-prediction-with-GAN

References

Alpren, C., Dawson, E.L., John, B., Cranston, K., Panneer, N., Fukuda, H.D., Roosevelt, K., Klevens, R.M., Bryant, J., Peters, P.J., Lyss, S.B., Switzer, W.M., Burrage, A., Murray, A., Agnew-Brune, C., Stiles, T., McClung, P., Campbell, E.M., Breen, C., Randall, L.M., Dasgupta, S., Onofrey, S., Bixler, D., Hampton, K., Jaeger, J.L., Hsu, K. K., Adih, W., Callis, B., Goldman, L.R., Danner, S.P., Jia, H., Tumpney, M., Board, A., Brown, C., DeMaria, A., Buchacz, K., 2020. Opioid Use Fueling HIV Transmission in an Urban Setting: An Outbreak of HIV Infection Among People Who Inject Drugs—Massachusetts, 2015–2018. Am. J. Public Health 110 (1), 37–44.

Amar, L.A., Taha, A.A., Mohamed, M.Y., 2020. Prediction of the final size for COVID-19 epidemic using machine learning: A case study of Egypt. Infect. Disease Model. 5 (2020), 622–634.

Badawi, A., Ryoo, S.G., 2016. Prevalence of comorbidities in the Middle East respiratory syndrome coronavirus (MERS-CoV): a systematic review and meta-analysis. Int. J. Infect. Diseases 49, 129–133.

Bauch, C.T., Lloyd-Smith, J.O., Coffee, M.P., Galvani, A.P., 2005. Dynamically Modeling SARS and Other Newly Emerging Respiratory Illnesses: Past, Present, and Future. Epidemiology 16 (6), 791–801.

Bloise, F., Tancioni, M., 2021. Predicting the spread of COVID-19 in Italy using machine learning: Do socio-economic factors matter? Structural Change and Economic Dynamics 56 (2021), 310–329.

Campo, D.S., Gussler, J.W., Sue, A., Skums, P., Khudyakov, Y., Blackard, J., 2020. Accurate spatiotemporal mapping of drug overdose deaths by machine learning of drugrelated web-searches. PLoS ONE 15 (12), e0243622.

Cao, S., Zheng, H., 2021. A POI-Based Machine Learning Method for Predicting Residents' Health Status. Proceedings of the 3rd International Conference on Computational Design and Robotic Fabrication (CDRF), Shanghai, China. 2021: 139-147.

Cao, Z., Wang, J., Gao, Y., Han, W., Feng, X., Zeng, G., 2008. Risk Factors and Autocorrelation Characteristics on Severe Acute Respiratory Syndrome in Guangzhou. Acta Ecologica Sinica 2008 (09), 981–993.

Carvajal, T.M., Viacrusis, K.M., Hernandez, L.F.T., Ho, H.T., Amalin, D.M., Watanabe, K., 2018. Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. BMC Infect Dis 18 (1).

Gong, Y.-F., Zhu, L.-Q., Li, Y.-L., Zhang, L.-J., Xue, J.-B., Xia, S., Lv, S., Xu, J., Li, S.-Z., 2021. Identification of the high-risk area for schistosomiasis transmission in China based on information value and machine learning: a newly data-driven modeling attempt. Infect Dis Poverty 10 (1).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al., 2014. Generative adversarial nets. Advances in neural information processing systems 27.

Hamer W.B., Birr, T., Verreet J.A., et al., 2020. Spatio-temporal prediction of the epidemic spread of dangerous pathogens using machine learning methods. ISPRS International Journal of Geo-Information, 2020, 9(1): 44.

He, J., Zheng, H., 2021. Prediction of crime rate in urban neighborhoods based on machine learning. Eng. Appl. Artif. Intell. 106 (2021), 104460.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30.

Isola, P., Zhu, J.Y., Zhou, T., et al., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.

Kissler, S.M., Tedijanto, C., Goldstein, E., Grad, Y.H., Lipsitch, M., 2020. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science.

Li, Q., Guan, X., Wu, P., 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 2020; published online Jan 29.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., 2020b. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). Science.

Lin, Q., Zhao, S., Gao, D., Lou, Y., Yang, S., Musa, S.S., 2020. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. Int J Infect Dis. 2020 (93), 211–216.

Mackey, T, et al., 2020. Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: retrospective big data infoveillance study. JMIR public health and surveillance 6.2 (2020): e19509.

Niu, B., et al., 2021. Epidemic analysis of COVID-19 in Italy based on spatiotemporal geographic information and Google Trends. Transboundary and Emerging Diseases 68.4 (2021): 2384-2400.

Peipei, S., Yinghui, X., 2020,8,3.. Qingming Z.A Study on Spatial and Temporal Aggregation Patterns of Urban Population in Wuhan City based on Baidu Heat Map and POI Data. International Review for Spatial Planning and Sustainable Development.

Perkins, A., Espana, G., 2020. Optimal control of the COVID-19 pandemic with non-pharmaceutical interventions. medRxiv.

Prem, K., Liu, Y., Russell, T.W., Kucharski, A.J., Eggo, R.M., Davies, N., 2020. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. Lancet. Public Health.

Remond, J., Remond, Y., 2020. A new virus-centric epidemic modeling approach, 1: General theory and machine learning simulation of 2020 SARS Cov 2 (COVID-19) for Belgium, France, Italy, and Spain. Mathematics and Mechanics of Complex Systems 8.3 (2020): 233-247.

Sahai, S.Y., et al., 2022. A machine learning model for nowcasting epidemic incidence. Math. Biosci. 343 (2022), 108677.

Salami, Donald, et al., 2020. Predicting dengue importation into Europe, using machine learning and model-agnostic methods. Scientific Reports 10.1 (2020): 1-13.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. In: Advances in Neural Information Processing Systems, p. 29.

Schneider, R., et al., 2021. Climate-based ensemble machine learning model to forecast dengue epidemics.

Sebastianelli, A. et al., 2021. AIRSENSE-TO-ACT: A Concept Paper for COVID-19 Countermeasures Based on Artificial Intelligence Algorithms and Multi-Source Data Processing. IJGI 10, 34 (2021).

Shou, X., Chen, P., Zheng, H., 2021. Predicting the Heat Map of Street Vendors from Pedestrian Flow through Machine Learning. Proceedings of the 26th International Conference on Computer-Aided Architectural Design Research in Asia (CAADRIA), Hong Kong, China. 2021: 2.569-578.

Solano-Villarreal, E., Valdivia, W., Pearcy, M., Linard, C., Pasapera-Gonzales, J., Moreno-Gutierrez, D., 2019. Malaria risk assessment and mapping using satellite imagery and boosted regression trees in the Peruvian Amazon. Sci Rep. 2019;9(1):15173.

Sun, Y., Jiang, L., Zheng, H., 2020. A Machine Learning Method of Predicting Behavior Vitality Using Open Source Data. Proceedings of the 40th Annual Conference of the Association for Computer Aided Design in Architecture (ACADIA), Philadelphia, USA. 2020: 160-168.

Velavan, T.P., Meyer, C.G., 2020. The COVID-19 epidemic. Tropical medicine & international health 25.3 (2020): 278.

Walker, P.G., Whittaker, C., Watson, O., Baguelin, M., Ainslie, K., Bhatia, S., 2020. The global impact of COVID-19 and strategies for mitigation and suppression. Imperial College London.

Wang, P., et al., 2020. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos, Solitons Fractals 139 (2020), 110058.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Watson, G.L., et al., 2021. Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model. PLoS computational biology 17.3 (2021): e1008837.

Wu, J.T., Leung, K., Leung, G.M., 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet 395.10225 (2020): 689-697.

Wu, C., Zhou, M., Liu, P., Yang, M., 2021. Analyzing COVID-19 using multisource data: An integrated approach of visualization, spatial regression, and machine learning. GeoHealth, 5, e2021GH000439.

Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., Wu, F., Weinberger, K., 2018. An empirical study on evaluation metrics of generative adversarial networks. arXiv preprint arXiv: 1806.07755.

Yang, J., Wu, H., Zheng, Y., 2019. Research on the spatial characteristics and optimization strategies of urban street walkability based on multi-source big data - Taking the central urban area of Nanjing as an example[J]. Urban Planning International 2019 (5), 33–42.

Yao, Y., Yin, H., Li, X., Guo, Z., Ren, S., Wang, R., Guan, Q., 2021. Fine-scale risk assessment of COVID-19 in Wuhan based on multisource geographical data. Acta Ecologica Sinica, 2021,41 (19):7493-7508.

Ye, S., Chen, D., Han, S., Liao, J., 2021. Learning with noisy labels for robust point cloud segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, 6443–6452.

Yedinak, J.L., et al., 2021. Machine learning takes a village: assessing neighbourhood-level vulnerability for an overdose and infectious disease outbreak. International Journal of Drug Policy 96 (2021), 103395.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, 586–595.

Zheng, H., An, K., Wei, J., Ren, J., 2020. Apartment floor plans generation via generative adversarial networks. Proceedings of the 25th International Conference on Computer-Aided Architectural Design Research in Asia (CAADRIA), 2:601-610.

Zhou, Z., Xu, Z., Liu, A., Zhou, S., Mu, L., Zhang, X., 2021. Mapping the Accessibility of Medical Facilities of Wuhan during the COVID-19 Pandemic. ISPRS Int. J. Geo-Inf. 10, 318.

Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T.Y., Shlens, J., Le, Q.V., 2020. Learning data augmentation strategies for object detection. European conference on computer vision. Springer, Cham 2020, 566–583.