SimAUD 2021
SYMPOSIUM ON SIMULATION FOR ARCHITECTURE + URBAN DESIGN

Applications of Machine Learning Methods for Predicting the Risk of COVID-19 in Beijing Based on Multifaceted Data

Xuan Liu[1], Hao Zheng [2]

[1]Harvard University
Beijing, China
xuanliu@gsd.harvard.edu

[2]University of Pennsylvania
Philadelphia, USA
zhhao@design.upenn.edu

ABSTRACT

COVID-19 has been widely considered as the greatest threat to global public health of the century. Urban areas with highly densified population have been worst hit in this pandemic. It can be argued that various factors embedded in urban spatial structures play important roles in influencing people's travel behavior, which further contributes to the transmission of COVID-19. This paper deploys machine learning models for exploring how various factors of urban spatial structure affect the distribution of the risk level of COVID-19. Linear regression models were applied for correlation analysis, the results of which shows that factors of land use and POI distribution are more correlated with COVID-19 risk distribution than others. Based on the correlation analysis, the eight variables, including land use, POI entropy, POI richness, population density, and distance to the breakout location, were selected for COVID-19 risk prediction models. Three risk prediction models were conducted using Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Network (ANN) respectively. As a result of the evaluation of these models, the ANN model achieved the best performance. Other latent factors related to the disease spread are also suggested to be considered for modelling the risk level prediction for future research.

Keywords

COVID-19; risk prediction; land use classification; POI diversity; urban structure; machine learning

1 INTRODUCTION

COVID-19 is an emerged threat to human, especially for densified urban population in major cities. Factors such as land use planning [1], POI distribution [2] and transportation networks [3] have been recognized to have deep influence on travel behaviors and interactions of urban residents, which is believed to be accounted for the spreading of the virus [4]. An understanding of the transmission of COVID-19 in urban areas is critical since the risk level prediction results could aid the virus control process in decision-making. This paper builds upon the idea of deploying machine learning and deep learning models to explore the relationship between urban spatial structures and the risk level distribution of COVID-19.

1.1 Urban Studies on Epidemics

The occurrence, transmission and affection of epidemics like COVID-19 is a complicated process. Researches from different aspects attempted to explain this mechanism. From the climatic perspective, temperature, humidity and air pollution significantly correlate with the transmission of the virus [5]. From the social perspective, economic similarities and geographic proximities both play important roles [6]. More discussions focus on the relationship between urban spaces and epidemics. From the historical point of view, John Snow (1854) depicted a "Cholera Map", which provided essential clues to the discovery of the cause of Cholera. The locations of the wells with contamination introduced the initial research methods for POI distribution. Back to COVID-19, Yaping Huang reflected on the spatial characteristics of employment center in downtown area of Wuhan (the breakout city of COVID-19). The single center of employment is a determinant factor for the concentration of the cases of COVID-19, which also accelerates cross-region mobility, and speeds up the spread of the virus.

1.2 Machine Learning Methods for Urban Studies

Urban agglomerations are complex systems, which requires advanced methods to decode the patterns embedded in the systems. Machine learning and deep learning models have gained popularity due to their great advantages in solving complex non-linear questions. Over recent years, attempts of deploying machine learning and deep learning models for urban studies have been made in a variety of researches, such as urban land use mapping using Random Forest [7], eco-environment system analysis [8] and vulnerability assessment using SVM [9], urban growth simulation using ANN [10], and land cover classification using CNN [11]. These data-driven models, compared to conventional knowledge-driven models, achieve better estimations based on quantitative measures of spatial associations between evidential features and prospective targets.

SimAUD 2021 April 15-17 Virtually Hosted?

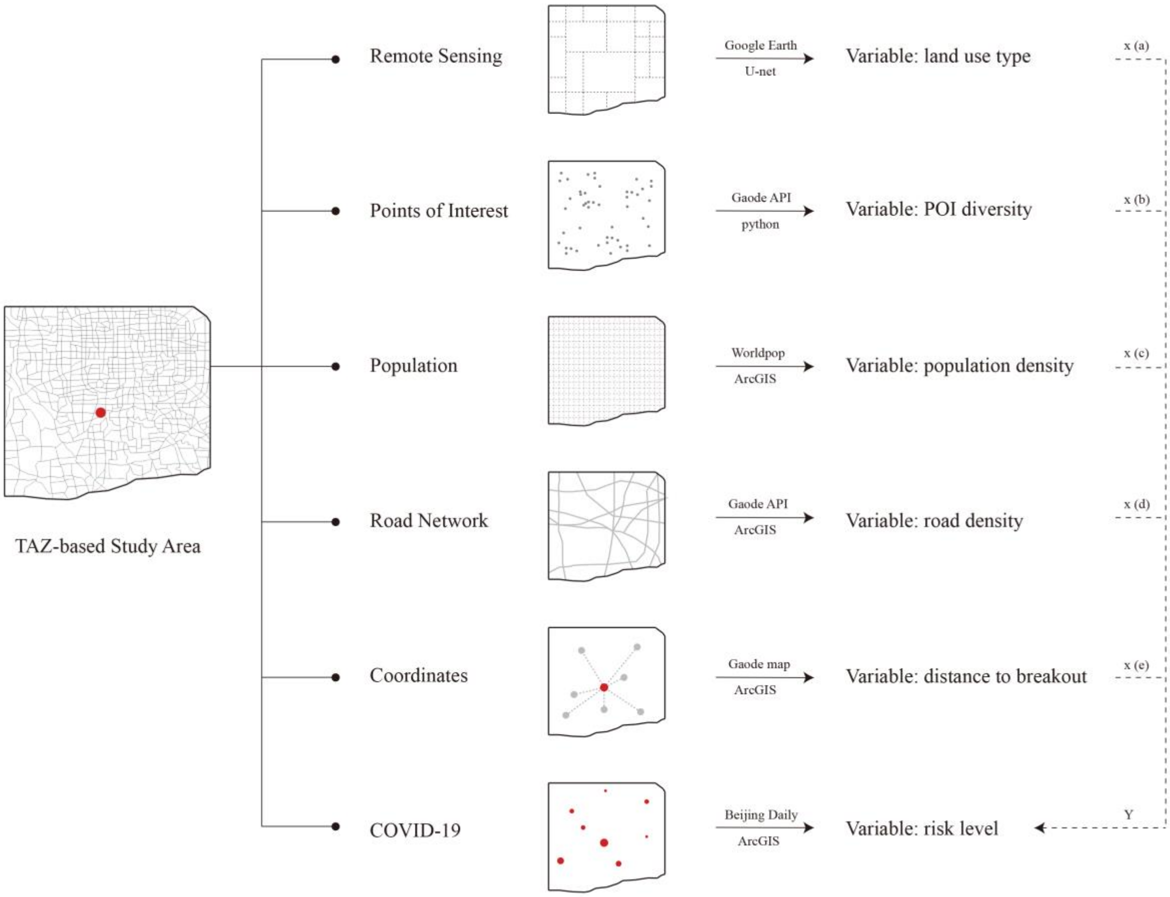


Figure 1. Research framework of each step for prediction.

1.3 Integration of Machine Learning Methods with COVID-19 Analysis

After the outbreak of COVID-19, some researchers employed machine learning methods to simulate the spread of the epidemic in urban spaces. For example, indicators for the evaluation of COVID-19 were constructed using multi-sourced data [12]. By using the GeoDetector and decision tree model, the study found that the older neighborhoods suffered higher risk level in contrast to newer neighborhoods, and the population density contributed most to infection. Also, a timely and novel methodology for COVID-19 forecasting was proposed [13]. The agent-based mechanistic model enabled the exploitation of geo-spatial activities and could be easily extended to most Chinese provinces. Both of the researches mentioned above aimed to make more effective response to COVID-19 situations ahead of time.

This paper, employing multifaceted data, provides a novel application to integrate machine learning methods with COVID-19 prediction in urban areas. The principal objective is to use machine learning models of Random Forest, SVM and ANN to explore the influence and role of various spatial elements on the risk level of COVID-19 from a geographical perspective. Based on this prediction system, future research of scenario planning is prospective.

2 METHODOLOGY

The framework of this study is illustrated in Figure 1. We applied the following steps to identify the risk level in each TAZ. First, we collected the data from multiple sources, and constructed the database with multifaceted variables, including land use, POI diversity and population density. Second, we applied correlation analysis of the input indicators obtained from the previous step. Third, three machine learning models were constructed, including Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN). Then we evaluated and compared the performance of these three models.

2.1 Data Preprocessing

2.1.1 Unet for Land Use Classification

Fully Convolutional Neural Networks allow for semantic segmentation to train pixel-based dataset for prediction. For image segmentation, Unet model is used. Unet contains: 1) with the encoder half, detecting input image features by the process of down-sampling; and 2) with the decoder half, establishing output image features by the process of up-sampling.

2.1.2 Indices for Measuring POI Diversity

Hill numbers was a form to unify diversity indices by ecologists. Yang considered Hill numbers (Richness, Entropy, and Simpson) as a better measurement of POI diversity for reflecting multifaceted, multidimensional urban land mixed use [2]. Hill numbers achieve diversity measurements by order *q*. When *q* = 0, 1, 2, it is the Richness, Entropy (orderliness), and Simpson (concentration) index respectively:

$$ {}^{q}D \equiv \left(\sum_{i=1}^{s} \mathrm{p}_i^q \right)^{1/(1-q)} . \tag{1} $$

There is also another metric of diversity indices except for Hill numbers, namely Gini Coefficient (measuring distribution inequality):

$$ G = \frac{1}{n}\left(n + 1 - 2\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i}\right) \tag{2} $$

2.2 Correlation Analysis

Linear regression is a basic and commonly used type of predictive analysis. It is comprised of two parts: 1) to evaluate the overall performance of the independent variables for outcome prediction; and 2) to evaluate the significance of each variable contributing to the prediction. A multiple linear regression model can be written as:

$$ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \epsilon. \tag{3} $$

2.3 COVID-19 Risk Level Prediction Models

2.3.1 Random Forest (RF)

Random Forest is a supervised learning algorithm with bagging technique for regression and classification. It is operated by constructing multiple parallel decision trees in the training process. We select this algorithm because it can efficiently leverage a large number of features with high overall accuracy.

2.3.2 Support Vector Machine (SVM)

Support Vector Machine is a supervised method with dichotomy classification of multidimensional features. The basic principle is to transform the input features into a higher-dimension space with linear separation. It is adopted because of its excellent properties of boosting generalization ability and global optimal solution.

2.3.3 Artificial Neural Networks (ANN)

Artificial Neural Networks is the most common method to develop nonparametric and nonlinear classification/ regression. The basic elements of an artificial neural network contain units (layers that connect information flow unidirectionally from the input layer, to the hidden layers and to the output layer) and nodes (interconnected with the corresponding links). To train an ANN model and to optimize the performance, it is needed to initialize the structure (number of hidden layers and nodes per layer), the weights, learning rate, and the regularization.

3 DATASET AND EXPERIMENT SETUP

3.1 Study Area

The study area is located in the central area of Beijing, the capital city of China, where 280 COVID-19 cases were confirmed within 15 days, from June 11 to June 25, 2020. Combined with the distribution of middle-and-high risk street blocks, a study area of 25kmx35km was selected near the Sixth Ring Road in Beijing. In the study area, 619 TAZs were further identified to increase the amount of training samples.

3.2 Sources of Data

3.2.1 Remote Sensing and Geospatial Data

The data deployed in this research is mainly collected from three sources, including Google Earth imagery, Gaode Map data, and Worldpop data, for risk level prediction. The Google Earth imagery consists of three bands (RGB), with a spatial resolution of approximately 2 meters.

Geospatial data, including Gaode Map road networks and Gaode POIs, were used to complement HSR-image extracted features and enrich additional information for land use identification in the study area. POIs in our study were collected from Gaode Map Services, which is one of the most popular and largest web map service providers in China. We obtained POIs from 11 main categories in the study area via Gaode Map APIs, including accommodation service, living service, shopping service, sports and entertainment, medical, company, residence, education, transportation, tourist attraction, and dining.

The population density data was obtained from the website of Worldpop. The online data has a raster format with a spatial resolution of 100 meters. Based on this, we obtained the average population density of each TAZ for the experiment in ArcGIS.

3.2.2 Risk Level Data

The locations of the 280 COVID-19 cases in the 15 days were identified from the official report by the Beijing government. To draw a high-grained map of its distribution, we collected 144 point locations according to the occurrence of these cases, and recorded in GIS with the weights of their risk level.

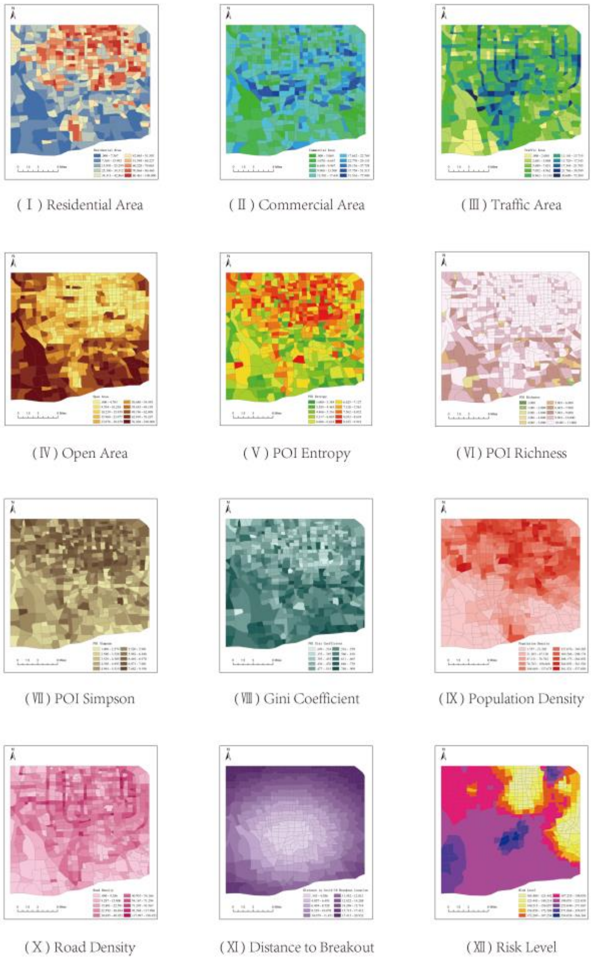
3.3 Construction of Database

3.3.1 Land Use Variables

To make the semantic segmentation label of HSR-image, we divided land use type into four categories, including residential area, commercial area, open area, and traffic area. After finishing the label, we trained the model with the U-net architecture, and obtained the map of land use classification in the study area. The label was divided into 80% training dataset and 20% validation dataset. The model was trained by Pytorch. Finally, we computed statistical proportion of different land use in each TAZ in GIS.

3.3.2 POI Diversity Variables

We used Python to calculate diversity indices of POI based on its density in each TAZ. With Richness, Entropy, and Simpson under the Hill numbers framework, as well as Gini Coefficient, we portrayed a more complete picture of mixed land use among neighborhoods by measuring POI diversity.


(Ⅰ) Residential Area

(Ⅱ) Commercial Area

(Ⅲ) Traffic Area

(Ⅳ) Open Area

(Ⅴ) POI Entropy

(Ⅵ) POI Richness

(Ⅶ) POI Simpson

(Ⅷ) Gini Coefficient

(Ⅸ) Population Density

(Ⅹ) Road Density

(Ⅺ) Distance to Breakout

(Ⅻ) Risk Level

Figure 2. Geographic visualization of initial variables in the study area (TAZ-based).

3.3.3 Transportation Network Variables

We applied the following three steps to identify the spatial structures in each TAZ. First, we extracted the road network information from Gaode Map, and calculated the average road density as a measurement of regional connectivity. Second, the average population density in each TAZ was obtained from the Worldpop raster data. Third, we collected the average distance to the COVID-19 breakout location (Xinfadi market) of each TAZ. This was based on the Euclidean distance calculation in GIS.

3.3.4 COVID-19 Risk Variables

IDW (Inverse Distance Weighted) is a commonly used and simple spatial interpolation method. It uses the distance between the interpolation point and the sample point as the weight to perform a weighted average. The closer the sample point is to the interpolation point, the greater will the weight be. Using this method, we transferred the distribution of COVID-19 cases into a raster format analysis. This transformation helped us to figure out the average risk level of each TAZ. Initial variables mapped in GIS are shown in Figure 2.

3.4 Correlation Analysis of Input Variables

Five models were used to examine the effects of 11 variables on risk level. Model 1 contained demographic variables (population density, and distance to the breakout location). For comparison purposes, Model 2 added road density as an additional independent variable. Model 3 added Hill numbers (Entropy, Richness, and Simpson) based on Model 1, Model 4 extended Model 3 by adding Gini-coefficient, Model 5 extended Model 3 by adding land use type. Referring to the contrast experiments, we selected the most relevant indicators as the input of machine learning models. Table 1&2 illustrate the details of the variables and the contrast experiment.

Type	No.	Variable name	Purpose discription	Range of variation	Source
Inputs	1	Residential area	Land use type	0-100, mean: 42.87	Google Earth
	2	Commercial area	Land use type	0-77.90, mean: 16.29	Google Earth
	3	Traffic area	Land use type	0-73.58, mean: 10.49	Google Earth
	4	Open area	Land use type	0-100, mean: 30.34	Google Earth
	5	POI richness	Diversity indices	1-11, mean: 10.12	Gaode API
	6	POI entropy	Diversity indices	1-9.9, mean: 6.71	Gaode API
	7	POI simpson	Diversity indices	1-9.20, mean: 5.46	Gaode API
	8	Gini-coefficient	Diversity indices	0.25-0.91, mean: 0.52	Gaode API
	9	Population density	Numbers of people	1.56-537, mean: 136	Worldpop
	10	Road density	Road connectivity	0-198, mean: 33.44	Gaode API
	11	Distance to breakout	Proximity to infection	0.34-20.93, mean: 11.08	Gaode map
Outputs	12	COVID-19 risk level	Target of prediction	1.04-5.66, mean: 1.88	Beijing Daily

Table 1. Indictors used for correlation analysis.

Variables	Model 1 (base)	Model 2 (base+road density)	Model 3 (base+Hill numbers)	Model 4 (base+Hill+Gini)	Model 5 (base+Hill+land use)
Population	√	√	√	√	√
Distance	√	√	√	√	√
Road density	-	√	-	-	-
Richness	-	-	√	√	√
Entropy	-	-	√	√	√
Simpson	-	-	√	√	√
Gini	-	-	-	√	-
Residential	-	-	-	-	√
Commercial	-	-	-	-	√
Traffic	-	-	-	-	√
Open	-	-	-	-	√

Table 2. Summary of independent variables used in the 5 models.

3.5 Construction of Machine Learning Models

3.5.1 Random Forest

Standardscaler was initialized for the dataset. For the regression model, the parameters of 6, 7, 8 were set for the depth of the model; and 100, 200, 300 were set for the estimators of the model. Validation dataset counted for 20%. Gridsearch was deployed for detecting the best parameters. R2 Score performed as the benchmark.

For the classification model, the dataset of 619 TAZs was divided into three categories including low risk level (184), middle risk level (349), and high risk level (86). SMOTE was used for the training sample. The parameters of 5, 6, 7, 8, 9, 10 were set for the depth of the model; and 100, 200, 300 were set for the estimators of the model. Validation dataset counted for 20%. Gridsearch was deployed for detecting the best parameters. Accuracy Score performed as the benchmark.

3.5.2 Support Vector Machine

Standardscaler was initialized for the dataset. For the regression model, the parameters of 1, 10, 100, 1000 were set for C in the model; and RBF was selected for kernel in the model. Validation dataset counted for 20%. Gridsearch was deployed for detecting the best parameters. R2 Score performed as the benchmark.

For the classification model, SMOTE was used for the training sample. The parameters of le0, le1, le2, le3, le4, le5 were set for C in the model. Validation dataset counted for 20%. Gridsearch was deployed for detecting the best parameters. Accuracy Score performed as the benchmark.

3.5.3 Artificial Neural Networks

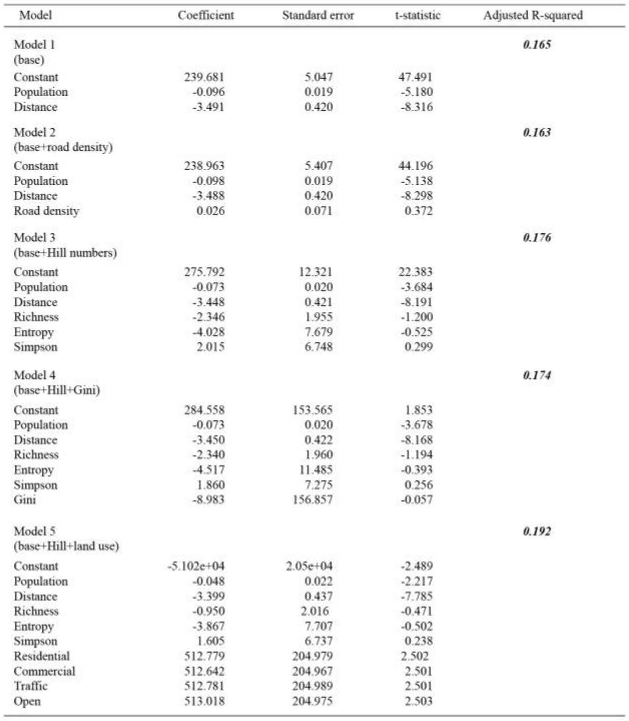
The more generalized regression model was adopted for ANN. The basic structure was with 8 input nodes, x number of hidden layers and y number of nodes in each hidden layer, and 1 output node. We adjusted the value of x and y to seek for the best predicting result. The validation split was 20% with batch size of 100. All models run with 1000 epochs. Adam optimizer assisted the training process. Sigmoid and MSE served as the activation and loss function respectively.

4 RESULTS

4.1 Results of Correlation Analysis

In the contrast experiment, the relative effect of input indicators are shown in Table 3. We selected four metrics to evaluate the effects of the 5 models, including adjusted R2, coefficient, standard error and t-statistic.

Model 1 as the basic model, achieved 0.165 R-squared, with the effects of distance (0.420 std err) and population density (0.019 std err). Model 2 added road density as the variable while achieved lower R-squared (0.163), and no influence on the effects of distance and population density, which indicates the road density as a minor factor. Model 3, considering Hill numbers, approached to a higher R-squared (0.176) compared to the basic model, while at the same time elevated the effects of distance and population density. Also, the Hill numbers had significantly higher effects on the risk level. Model 4 added Gini-coefficient, which elevated the effects of POI diversity but lowered the overall R-squared (0.174). Model 5 supplemented four land use type. This model achieved the best performance with R-squared of 0.192, and relatively higher input effects. Among these indicators, we finally determined eight variables (including distance, population density, POI richness, entropy, and land use type) as the input of machine learning models according to the evaluation metrics.



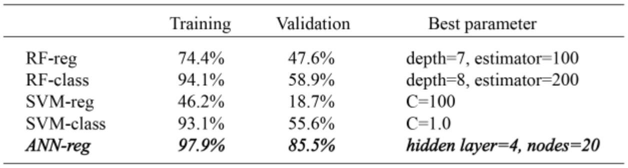
Model	Coefficient	Standard error	t-statistic	Adjusted R-squared
Model 1 (base)				***0.165***
Constant	239.681	5.047	47.491	
Population	-0.096	0.019	-5.180	
Distance	-3.491	0.420	-8.316	
Model 2 (base+road density)				***0.163***
Constant	238.963	5.407	44.196	
Population	-0.098	0.019	-5.138	
Distance	-3.488	0.420	-8.298	
Road density	0.026	0.071	0.372	
Model 3 (base+Hill numbers)				***0.176***
Constant	275.792	12.321	22.383	
Population	-0.073	0.020	-3.684	
Distance	-3.448	0.421	-8.191	
Richness	-2.346	1.955	-1.200	
Entropy	-4.028	7.679	-0.525	
Simpson	2.015	6.748	0.299	
Model 4 (base+Hill+Gini)				***0.174***
Constant	284.558	153.565	1.853	
Population	-0.073	0.020	-3.678	
Distance	-3.450	0.422	-8.168	
Richness	-2.340	1.960	-1.194	
Entropy	-4.517	11.485	-0.393	
Simpson	1.860	7.275	0.256	
Gini	-8.983	156.857	-0.057	
Model 5 (base+Hill+land use)				***0.192***
Constant	-5.102e+04	2.05e+04	-2.489	
Population	-0.048	0.022	-2.217	
Distance	-3.399	0.437	-7.785	
Richness	-0.950	2.016	-0.471	
Entropy	-3.867	7.707	-0.502	
Simpson	1.605	6.737	0.238	
Residential	512.779	204.979	2.502	
Commercial	512.642	204.967	2.501	
Traffic	512.781	204.989	2.501	
Open	513.018	204.975	2.503	

Table 3. Regression results on the relationship between spatial elements and COVID-19 risk level.

4.2 Results of Machine Learning Prediction

Table 4 illustrates the results of different machine learning and deep learning models. The two machine learning models (RF and SVM) had relatively lower predicting accuracy compared to deep learning models (ANN). For the Random Forest regression model, it achieved 74.4% on training dataset and 47.6% on validation dataset. The best parameters by gridsearch was 7 for the model depth and 100 for the model estimator. With the classification model, it achieved 94.1% on training dataset and 58.9% on validation dataset. The best parameters by gridsearch was 8 for the model depth and 200 for the model estimator. For the Support Vector Machine regression model, it achieved 46.2% on training dataset and 18.7% on validation dataset. The best parameters by gridsearch was 100 for C in the model. With the classification model, it achieved 93.1% on training dataset and 55.6% on validation dataset. The best parameters by gridsearch was 1.0 for C in the model.

As shown in the results, Artificial Neural Networks had better predicting performance. For the regression model, it achieved 97.9% on training dataset and 85.5% on validation dataset. The best parameters were with 4 hidden layers and 20 nodes in each hidden layer.



	Training	Validation	Best parameter
RF-reg	74.4%	47.6%	depth=7, estimator=100
RF-class	94.1%	58.9%	depth=8, estimator=200
SVM-reg	46.2%	18.7%	C=100
SVM-class	93.1%	55.6%	C=1.0
ANN-reg	***97.9%***	***85.5%***	***hidden layer=4, nodes=20***

Table 4. Predicting performance of different machine learning and deep learning models.

5 DISCUSSION

5.1 Correlation Analysis

In the correlation analysis, population density is one of the vital factors related to COVID-19 risk level. Relevant researches show that the infection rate is strongly correlated with the population density, by conducting spatial regression models [14]. The distance to the breakout location also determines the exposure to the disease. The number of infection cases decrease with the distance growing. Thus social distancing and travel restriction support the prevention of COVID-19 spread.

POI diversity (richness, entropy, simpson), as a measurement of urban land mixed use, advocates a balanced mode of public realm in the city. Mixed use, encouraged by theories such as sustainable development, regains concern due to its restoration of economic vitality, social equity and environmental quality [15]. Within a city neighborhood, the level of mixed use triggers human mobility with geographic centrality [3]. While the human mobility is one of the largest driven factors correlated with disease spread. As the results suggest, POI richness and entropy (orderliness) have a negative correlation with COVID-19 risk level. In other words, the neighborhoods with better functional quality possess lower potential of infection. This may be partially attributed to the less demand of cross-region mobility, which alleviates the human-to-human transmission. The main lesson from this result is the decentralization approach. It is highlighted that smaller urban units should be reasonably distributed and the local urban centers should be strengthened [16]. The decentralization approach could accelerate horizontal expansion, by maintaining sustainable development and resilient city planning in the future [17].

Land use type (residential, commercial, traffic, and open space), also plays an important role on the risk level of COVID-19. Since the pandemic, many policymakers and planners try to increase the protection and defense system by avoiding high density and overcrowding. Especially for the residential areas, conventional housing with high density severely exacerbates the unhygienic conditions and the spread of communicable diseases. Under this context, green spaces such as urban parks serve as the essential intermediate buffers in the built environment. It affirms that urban parks and large open spaces can provide residents with safe outdoor activities and social interaction in a green environment, while at the same time maintain the health and quality of life [18].

5.2 Machine Learning Prediction

Among the machine learning and deep learning models, ANN achieved better results compared to RF and SVM. Relevant researches on using these models for predictions in urban studies compare and analyze their characteristics. Random Forest involves a lesser difficulty in training, and the prospective models are with greater ensembles [19]. However, ANN and SVM are more complex. The combination of parameters of kernel types in SVM is different for optimization. And the accuracy of ANN increases as the networks become more complex, i.e., increase of nodes in each hidden layer. Due to the performances, RF model outstands. There is a well-round discussion of the comparison between Random Forest and ANN, while the option depends [20]. The key criteria include performance, robustness, comprehensibility, cost and time expenditure. Although Random Forest is a better option in many practical applications, deep learning methods are still able to improve the quality given a large amount of data or complex situations.

ANN has been used to predict the prevalence of COVID-19 in some researches. The proposed model [21] can perform multi-step forecasts for further days with a reasonable absolute percentage error less than 5%. And the experiments can be transferred from Egypt to other countries. Also, the study presents a new method of intelligent curve fitting and forecasting for different non-linear models, by showing the results that ANN can efficiently train any set of country's data trend for future cases [22].

5.3 Future Research

5.3.1 Scenario Planning

The ANN model in the experiment achieved the ultimate accuracy of 85.5% which can be used to conduct the scenario planning research [10] in the following steps. Specifically, it can be observed how the risk level distribution in the study area would change based on the variations of the input spatial elements (land use, POI diversity, population, distance). Compared with doing risk level prediction with the existing urban data, the scenario planning research allows the planners and decision-makers to simulate the effect of future urban design on the potential risk level of pandemic. And this planning mode can also be transferred to other urban areas to assist intelligent decisions on smart cities.

5.3.2 Model Optimization

For the variable input, another two essential factors could be added to enrich the dataset of spatial elements. The first one is the place connectivity [23]. It delineates the relationship between neighborhoods by encoding place characteristics as node features and representing place connections as edge features. This variable can be taken into the current ANN model for better detection of spatial elements. The second one is the time variation [24]. A citywide spatio-temporal flow volume is predicted by dynamically learning the temporal dependency via the feedback connections. In our ANN model, we could adopt the GPS data, i.e. location-based service (LBS) data from the mobile phones of the citizens, in order to demonstrate the real-time mobility in the city for predicting the risk level.

For the model construction, some advanced version of deep learning models for predicting COVID-19 has been put forward since the pandemic. And these models can inspire us on optimizing the current ANN model. For instance, a novel framework of COVID-19Net is proposed, which combines CNN and GRUs to accurately predict the accumulated number of confirmed cases and serve as a crucial reference for devising public health strategies [25]. Also, agent-based simulations and deep-learning techniques are utilized to predict transportation trends in COVID-19, as well as the impact of proposed phased reopening strategies [26].

5.3.3 Other Latent Factors

The main purpose of this research is to detect the correlation between urban spatial elements and the risk level of COVID-19, while other latent factors might have influence on the prediction. The natural environment such as the climate factors [5]; and the human environment such as the living quality of the old neighborhoods compared to the new neighborhoods [12], travel destinations and intensity [27], as well the government intervention, risk perception and the adoption of protective action recommendations [28], etc. can all contribute to the spread and risk level of COVID-19. How these latent factors are related to the current prediction process still deserves to be explored.

6 CONCLUSION

In this paper, we perform a series of experiment to analyze the relationship between urban spatial elements and the risk level of COVID-19 from a geographical perspective by employing multifaceted data. Land use, POI diversity, population density and distance to the breakout location are major indicator to the risk level based on correlation analysis and serve as the input for further prediction model. Different machine learning methods are conducted and compared including Random Forest, Support Vector Machine, and Artificial Neural Network. In our study, ANN achieves the best performance among the three models. Future research will focus on: scenario planning to better support the urban planners and decision-makers; model optimization to complement spatial elements such as place connectivity and time variation, as well as more advanced model construction for prediction; and adoption of other latent factors containing both natural environment (i.e. climate factors) and human environment (i.e. living quality, travel demands, and government intervention).

We believe this project represents an important advancement of depicting the risk distribution of pandemic in urban areas from an intelligent view, thus further exploration is promising and encouraging for AI-assisted decision-making process for the risk-resilient system of future cities.

ACKNOWLEDGMENTS

We thank all publications support and staff, who wrote and provided helpful comments on previous versions of this document. As well authors 1, & 2 gratefully acknowledge the grant from the authorities.

REFERENCES

[1] Shi, Qi, Liu, Niu, and Zhang, "Urban Land Use and Land Cover Classification Using Multisource Remote Sensing Images and Social Media Data," *Remote Sens.*, vol. 11, no. 22, p. 2719, Nov. 2019, doi: 10.3390/rs11222719.

[2] Y. Yue, Y. Zhuang, A. G. O. Yeh, J.-Y. Xie, C.-L. Ma, and Q.-Q. Li, "Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy," *Int. J. Geogr. Inf. Sci.*, vol. 31, no. 4, pp. 658–675, Apr. 2017, doi: 10.1080/13658816.2016.1220561.

[3] S. K. Chong, M. Bahrami, H. Chen, S. balcisoy, B. Bozkaya, and A. "Sandy" Pentland, "Economic outcomes predicted by diversity in cities," Open Science Framework, preprint, Feb. 2020. doi: 10.31219/osf.io/j59u3.

[4] N. A. Megahed and E. M. Ghoneim, "Antivirus-built environment: Lessons learned from Covid-19 pandemic," *Sustain. Cities Soc.*, vol. 61, p. 102350, Oct. 2020, doi: 10.1016/j.scs.2020.102350.

[5] M. M. Hoque, U. Saima, and S. S. Shoshi, "Correlation of Climate Factors with the COVID-19 Pandemic in USA," p. 5.

[6] Y. Qiu, X. Chen, and W. Shi, "Impacts of social and economic factors on the transmission of coronavirus disease (COVID-19) in China," p. 27.

[7] I. E. Ruiz Hernandez and W. Shi, "A Random Forests classification method for urban land-use mapping integrating spatial metrics and texture analysis," *Int. J. Remote Sens.*, vol. 39, no. 4, pp. 1175–1198, Feb. 2018, doi: 10.1080/01431161.2017.1395968.

[8] J. Zhao and Z. Jin, "Predict Coordinated Development Degree of County Eco-Environment System Using GA-SVM: A Case Study of Guanzhong Urban Agglomeration," *J. Glob. Inf. Manag.*, vol. 26, pp. 1–10, Jul. 2018, doi: 10.4018/JGIM.2018070101.

[9] F. Li, W. Wang, J. Xu, J. Yi, and Q. Wang, "Comparative study on vulnerability assessment for urban buried gas pipeline network based on SVM and ANN methods," *Process Saf. Environ. Prot.*, vol. 122, pp. 23–32, Feb. 2019, doi: 10.1016/j.psep.2018.11.014.

[10] B. C. Pijanowski, "A big data urban growth simulation at a national scale: Configuring the GIS and neural network based Land Transformation Model to run in a High Performance Computing (HPC) environment," *Environ. Model.*, p. 19, 2014.

[11] P. Ulmas and I. Liiv, "Segmentation of Satellite Imagery using U-Net Models for Land Cover Classification," *ArXiv200302899 Cs*, Mar. 2020, Accessed:

Aug. 31, 2020. [Online]. Available: http://arxiv.org/abs/2003.02899.

[12] Y. Zhang, Y. Li, B. Yang, X. Zheng, and M. Chen, “Risk Assessment of COVID-19 Based on Multisource Data From a Geographical Viewpoint,” *IEEE Access*, vol. 8, pp. 125702–125713, 2020, doi: 10.1109/ACCESS.2020.3004933.

[13] D. Liu *et al.*, “A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models,” *ArXiv200404019 Cs Q-Bio Stat*, Apr. 2020, Accessed: Nov. 09, 2020. [Online]. Available: http://arxiv.org/abs/2004.04019.

[14] S. Copiello and C. Grillenzoni, “The spread of 2019-nCoV in China was primarily driven by population density. Comment on ‘Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China’ by Zhu et al.,” *Sci. Total Environ.*, 2020, doi: 10.1016/j.scitotenv.2020.141028.

[15] J. Grant, “Mixed Use in Theory and Practice: Canadian Experience with Implementing a Planning Principle,” *J. Am. Plann. Assoc.*, vol. 68, no. 1, pp. 71–84, Mar. 2002, doi: 10.1080/01944360208977192.

[16] J. Aguilar *et al.*, “Impact of urban structure on COVID-19 spread,” *ArXiv200715367 Phys. Q-Bio*, Jul. 2020, Accessed: Nov. 18, 2020. [Online]. Available: http://arxiv.org/abs/2007.15367.

[17] A. Madanipour, “How Relevant Is ‘Planning by Neighbourhoods’ Today?,” *Town Plan. Rev.*, vol. 72, no. 2, pp. 171–191, 2001.

[18] J. Xie, S. Luo, K. Furuya, and D. Sun, “Urban Parks as Green Buffers During the COVID-19 Pandemic,” *Sustainability*, vol. 12, no. 17, p. 6751, Aug. 2020, doi: 10.3390/su12176751.

[19] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, “Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines,” *Ore Geol. Rev.*, vol. 71, pp. 804–818, Dec. 2015, doi: 10.1016/j.oregeorev.2015.01.001.

[20] D. P. Roßbach, “Neural Networks vs. Random Forests – Does it always have to be Deep Learning?,” p. 8.

[21] A. I. Saba and A. H. Elsheikh, “Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks,” *Process Saf. Environ. Prot.*, p. 9, 2020.

[22] S. K. Tamang, P. D. Singh, and B. Datta, “Forecasting of Covid-19 cases based on prediction using artificial neural network curve fitting technique,” *Glob. J. Environ. Sci. Manag.*, vol. 6, no. Special Issue (Covid-19), pp. 53–64, Aug. 2020, doi: 10.22034/GJESM.2019.06.SI.06.

[23] “Understanding Place Characteristics in Geographic Contexts through Graph Convolutional Neural Networks: Annals of the American Association of Geographers: Vol 110, No 2.” https://www.tandfonline.com/doi/abs/10.1080/24694452.2019.1694403 (accessed Nov. 20, 2020).

[24] Y. Ren, H. Chen, Y. Han, T. Cheng, Y. Zhang, and C. Ge, “A hybrid integrated deep learning model for the prediction of citywide spatio-temporal flow volumes,” *Int. J. Geogr. Inf. Sci.*, vol. 34, pp. 1–22, Aug. 2019, doi: 10.1080/13658816.2019.1652303.

[25] C.-J. Huang, Y. Shen, P.-H. Kuo, and Y.-H. Chen, *Novel Spatiotemporal Feature Extraction Parallel Deep Neural Network for Forecasting Confirmed Cases of Coronavirus Disease 2019*. 2020.

[26] D. Wang *et al.*, *Agent-based Simulation Model and Deep Learning Techniques to Evaluate and Predict Transportation Trends around COVID-19*. 2020.

[27] X. Wu, J. Yin, C. Li, H. Xiang, M. Lv, and Z. Guo, “Natural and human environment interactively drive spread pattern of COVID-19: a city-level modeling study in China,” *Sci. Total Environ.*, Oct. 2020, doi: 10.1016/j.scitotenv.2020.143343.

[28] T. Duan, H. Jiang, X. Deng, Q. Zhang, and F. Wang, “Government Intervention, Risk Perception, and the Adoption of Protective Action Recommendations: Evidence from the COVID-19 Prevention and Control Experience of China,” *Int. J. Environ. Res. Public. Health*, vol. 17, no. 10, May 2020, doi: 10.3390/ijerph17103387.