INFLUENCE ANALYSIS OF WATERLOGGING BASED ON DEEP LEARNING MODEL IN WUHAN

Yin Pan [a], Zhenfeng Shao [a, *], Tao Cheng [a], Zhiqiang Wang [b], Zhaoyu Zhang [b]

[a] State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[b] School of Remote Sensing Information Engineering, Wuhan University, Wuhan 430079, China
Email: shaozhenfeng@whu.edu.cn

Commission VI, WG VI/4

KEY WORDS: Deep Learning, Waterlogging Influence, Stacked Autoencoder, Analytic Hierarchy Process (AHP)

ABSTRACT:

This paper analyses a large number of factors related to the influence degree of urban waterlogging in depth, and constructs the Stack Autoencoder model to explore the relationship between the waterlogging points' influence degree and their surrounding spatial data, which will be used to realize the comprehensive analysis in the waterlogging influence on the work and life of residents. According to the data of rainstorm waterlogging in 2016 July in Wuhan, the model is validated. The experimental results show that the model has higher accuracy than the traditional linear regression model. Based on the experimental model and waterlogging points distribution information in Wuhan over the years, the influence degree of different waterlogging points can be quantitatively described, which will be beneficial to the formulation of urban flood control measures and provide a reference for the design of city drainage pipe network.

1. INTRODUCTION

The rapid development of the city and the dramatic changes in the global climate has led to the increased risk of sudden rainstorms in urban areas, and their impacts have gradually expanded. The waterlogging caused by heavy rainstorms in the city always makes public traffic paralysis, public travelling inconvenience, or brings a serious threat to the safety of people's life and property.

At present, the most of researches on urban waterlogging are mainly aimed at the problems of waterlogging reasons, model prediction, risk assessment, prevention and control decisions, etc. However, the architecture methods of waterlogging prediction model are complicated, such as Flood Area Model (Xue et al., 2016), Storm Water Management Model (Liao et al., 2014; Chen et al., 2010). Otherwise, the related researches in risk assessment usually establish a system or model to evaluate the threat of waterlogging, such as Simplified Urban Waterlogging Model (Quan et al., 2010), Network Information System (Rahadianto et al., 2015). The realization of these complicated models or systems often requires longer time and more funds, so they are more suitable for long-term planning in urban waterlogging disaster prevention and control work.

To solve the problem of urban waterlogging disaster, the most direct method is to replace the city drainage system. However, in the urban short-term waterlogging disaster prevention and control work, the overall change of the city drainage system is not realistic. Therefore, this paper makes a statistical analysis of a large number of relevant data around the waterlogging points, and combined with the principle of the Stacked Autoencoder network, to dig out a hidden deep relationship between these

numerous data and realize the quantitative analysis of the waterlogging influence on the residents' work and life. The analysis results of the waterlogging influence degree can be used as a reference for whether this region is given a priority in the waterlogging prevention and control work.

2. FRAMEWORK

In order to analyse the influence degree of waterlogging in the urban areas, this paper quantifies the waterlogging impact. Based on the actual data acquired from the urban management department, the spatial data related to waterlogging influence degree will be comprehensively analysed. As the living density, educational facilities, transportation network, service facilities and other spatial data around the waterlogging point are closely related to the influence degree, the statistical analysis function of ArcGIS software can be used to calculate the relevant data within a certain buffer range. After the statistical data is normalized, a data series related to influence degree of each waterlogging point will be obtained.

In the study of the waterlogging influence analysis, the data volume is large, and the relationship between these spatial data and waterlogging influence degree is complex, which is difficult to describe with simple linear function. Therefore, based on the principle of deep learning model, the Stacked Autoencoder network constructed by multi-layer Sparse Autoencoder can be adopted to excavate the complex relationship, so as to realize quantitative analysis of urban waterlogging effect.

After determining the algorithm model and input data, the labels need to be required for the process of model's parameters training. At present, the relevant index system comprehensively evaluating waterlogging impact has not yet formed, so the labels of waterlogging influence degree cannot be obtained directly. In

---

* Corresponding author: Shao Zhenfeng, professor, Ph.D supervisor. E-mail: shaozhenfeng@whu.edu.cn

this paper, the spatial data associated with the waterlogging impact has a clear hierarchical structure, so we can obtain the relative weight of each type of data by using AHP method, finally, to obtain the influence degree label of each waterlogging point.

Acquiring the related data series and the influence degree of every waterlogging point, the Autoencoder algorithm combined with fine-tuning algorithm will be adopted to train the Stacked Autoencoder model, and some waterlogging points will be used as test samples to evaluate the model accuracy. At the same time, through the waterlogging distribution information over the years, some waterlogging points can be applied to do model prediction. After analysing their influence degree, the prediction results will provide a reference for the design of pipe network and scheme of flood control planning. The specific research framework is shown in Figure 1.
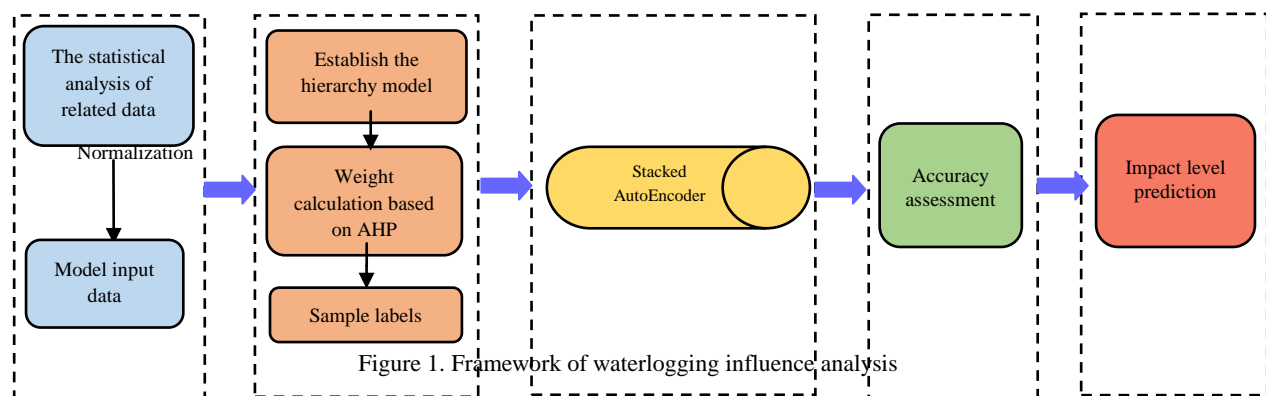

Figure 1. Framework of waterlogging influence analysis

3. ANALYSIS PROCESS OF WATERLOGGING INFLUENCE

3.1 Statistical Analysis of Data

Considering the spatial data around the waterlogging point is closely related to the influence degree, therefore, by using the statistical analysis function of ArcGIS software to set reasonable buffer (the experiment set a unified buffer of 500m), for each waterlogging point, the same type of data within the scope of buffer will be summed up by statistics. At the same time, through the screening and merging in the database, the spatial data types to be considered in this paper will be determined. These spatial data types are broadly divided into 7 categories, involving a total of 21 types, as shown in the Table 1.

It is because of the data types of basic geographic information database including points, lines, surfaces, the method of statistical analysis is different. For the point data, the number of same type of data within the scope of buffer is considered. For example, within the buffer of No. 1 waterlogging point, what the number of bus stations is. For the linear data, the length of a certain type of road in the buffer is considered. For example, within the buffer of No. 1 waterlogging point, what the total length of all main roads is. For the surface data, the area of such data in the buffer is considered. For example, within the buffer of No. 1 waterlogging point, what the total area of all primary schools is.

Table 1. Spatial data categories related to waterlogging impacts

Major categories	Spatial data types					
Residential related B1	Housing area C11					
Educational assistance B2	Library number C21	Museum number C22	Cultural relics protection unit C23			
School related B3	Primary school area C31	Secondary school area C32				
Station number B4	Bus station C41	Coach station C42	Railway station C43	Subway station C44		
Traffic roads B5	Bridges C51	Main roads C52	Secondary roads C53	Village internal roads C54	Branch roads C55	Tunnels C56
Number of service facilities B6	Service center C61	Parking lot C62	Express outlets C63	Medical facilities C64		
Enterprise related B7	Enterprise number C71					

As the statistical analysis process contains different dimensions of the data, such as length, area, quantity, the statistical spatial data need to be normalized. Eventually, the spatial data series related to the influence degree of each waterlogging point can be obtained.

3.2 Label Acquisition Based on AHP

It is also necessary to use the analytic hierarchy process to obtain the label value of influence degree. Firstly, the hierarchical model is established through the classification results of all spatial data (as shown in Table 1). Then, according to the relative importance of each sub-element in hierarchical model, the judgment matrix will be constructed. By calculating the eigenvector of the matrix, we can get the weight of each sub-element relative to the upper-element. Finally, the weight of each sub-element and its specific value can be weighted sum, and the calculation results will be mapped to waterlogging influence labels. The specific steps are as follows:

(1) According to the classification results in Table 1, all the spatial data related to the impact of waterlogging in the statistical analysis will be divided into seven categories, involving 21 types of data, which corresponds 8 hierarchical models, the detailed models are shown in Figure 2;

(2) Each hierarchical model corresponds to a judgment matrix whose values are obtained by comparing the relative importance of the two sub-elements, generally expressed by 1-9 and the reciprocals. Through solving the eigenvectors of the judgment matrix, and normalizing the eigenvectors corresponding to the largest eigenvalues, the relative weight of each sub-element to its upper-element will be calculated;

(3) In order to ensure the rationality of the judgment matrix, it is necessary to introduce the consistency index CI:

$$CI = (\lambda_{max} - n) / (n-1) \qquad (1)$$

where λ_{max} = largest eigenvalue of the judgment matrix
  n = judgment matrix dimension

Since the CI value will change with the order of the judgment matrix, it is necessary to introduce the average random consistency index RI. For the matrix with fixed dimension, RI value can be obtained by look-up table, as shown in Table 2:

Table 2. RI value

Dimension	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.9	1.12	1.24	1.38	1.41	1.46

After obtaining the CI and RI values of the judgment matrix, the stochastic consistency ratio CR can be calculated by the ratio of CI and RI. When CR is less than 0.1, it is considered that the

judgment matrix is reasonable and passes the consistency test;

(4) For each waterlogging point, with the relative weight of each type of spatial data, we can calculate the weighted summation combined with their normalized values. Considering the classification performance of the classifier, the numerical value of the calculation is transformed into a natural number of 1 to 10, which will be regarded as the label value of influence degree. For example, if the weighted results is 0 to 0.1, the label is 1; if the weighted results is 0.9 to 1, then the label is 10.

$$label = ceil(10*W*X) \qquad (2)$$

where label = label value of waterlogging influence degree
  ceil = rounding up function
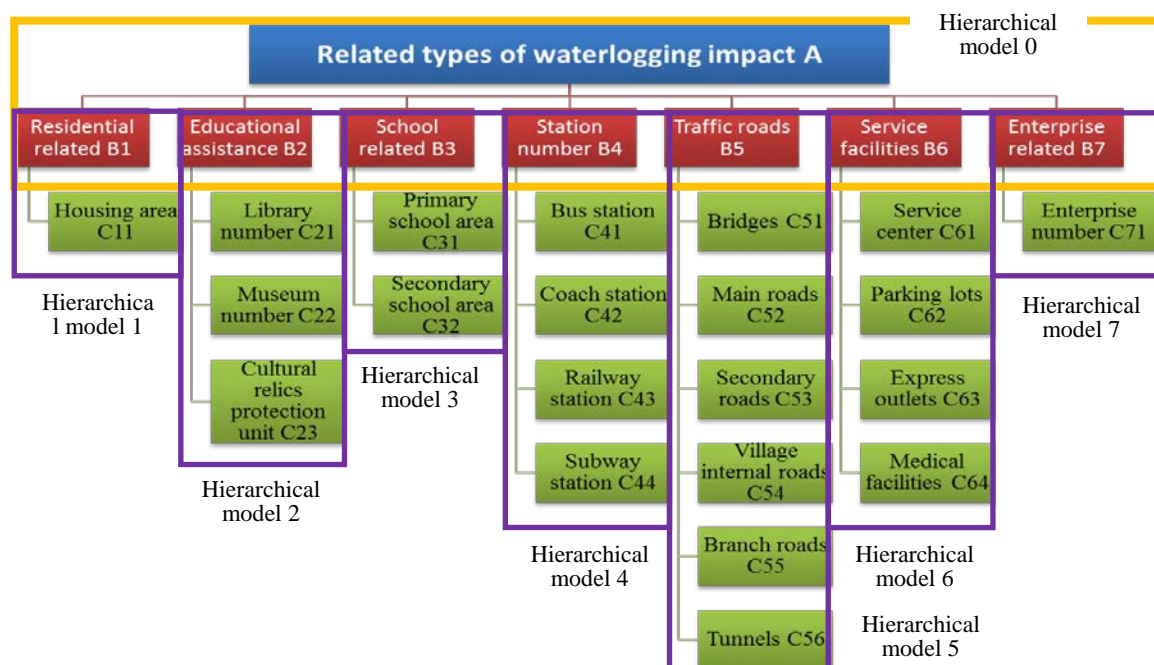 W = weight vector
  X = normalized spatial data series

Figure 2. Hierarchical models

3.3 Stacked Autoencoder Model Training

In this paper, two Sparse Autoencoders are combined to form a Stacked Autoencoder neural network, which includes two hidden layers. At the same time, taking into account the need to achieve the quantitative analysis in waterlogging influence on the residents' work and life, it is necessary to construct a complete neutral network framework with softmax classifier.

In the process of training, we take the method of greedy layer-wise pre-training to train each layer of the network in turn, that means the output of the former layer will be used as the input of the next layer. Through this layer-by-layer training, the characteristics and relevance of the data are sequentially mapped to each hidden layer network, and eventually abstracted to the deepest hidden unit. The deepest hidden layer output is regarded as the input of the softmax classifier, and it will be mapped to a numeric label by supervised training. Comparing the classification output value with the label value, the gradient of the classification error will be propagated back to the coding

layer, and then the parameters of the whole network will be iteratively optimized. Finally, a Stacked Autoencoder model with the ability of classification in waterlogging influence analysis will be established, and by the use of it, the influence degree of other waterlogging points will be predicted accurately.

After the statistical analysis of the data is completed, each sample includes the serial number, the label and its 21 kinds of spatial data. According to the principle of Stacked Autoencoder, the input data series and the corresponding labels of the training samples are input, and the number of neurons in two hidden layers is given to realize the training process of the neural network. Then, combined the input data series of the test samples with the training parameters of the model, the output values will be obtained. Finally, the ratio of the number of test samples whose prediction results are in good agreement with their labels and the total number of the test samples, is used as an index to evaluate the accuracy of the model.

4. EXPERIMENT OF WUHAN CITY

4.1 Experimentation Regional Overview

Wuhan is located at 113°41′E to 115°05′E, 29°58′N to 31°22′N, in the middle and lower reaches of the Yangtze River, the east of Jianghan Plain. The climate of Wuhan belongs to the subtropical monsoon climate. The rainfall is unevenly distributed in the season, and the annual rainfall can reach up to 1205 mm. In spring and summer, the rainfall is large. Especially in the summer, heavy rain sudden, it is easy to cause the waterlogging.

Wuhan is one of the most serious waterlogging disasters. Every summer, Wuhan will experience several heavy rains, which will form a unique scene of "see the sea in summer". During the heavy rainstorms from June to July 2016, the cumulative number of affected people was as high as 757328, involving 12 different regions. Waterlogging disasters caused by heavy rain resulted in more than 230 traffic lines outages, multiple walls and dams collapsed. At the same time, the waterlogging in the campus led to the majority of teaching buildings and dormitories without water, power outages and network stopped.

4.2 Statistical Results of Samples

According to the relevant data provided by the basic geographic information database in Wuhan, this experiment selects 135 waterlogging points in 2016 July as the research objects, among which the top 100 waterlogging points are used as training samples and 35 points are used as test samples. The distribution of these points is shown in Figure 3. According to the statistical methods in Section 3.1, the spatial data series of each waterlogging point is counted, and the 21 types of normalized data of training and test samples are shown in Table 3.
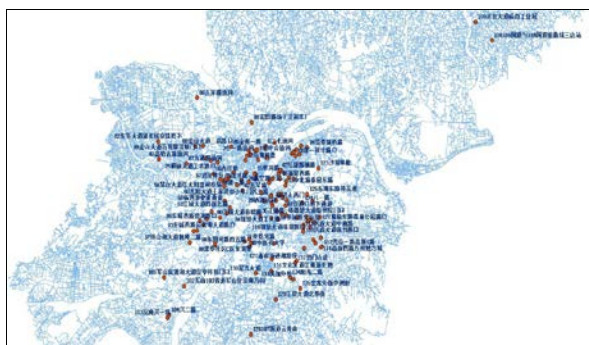

Figure 3. Distribution of sample waterlogging points

Table 3. Input data of samples

Number	0	1	2	...	45	46	...	133	134
Housing area	0.14	0.42	0.41		0.51	0.20		0.37	0.27
Library number	0.00	0.00	0.00		0.00	0.00		0.00	0.00
Museum number	0.00	0.00	0.00		0.25	0.00		0.00	0.00
Cultural relics protection unit	0.00	0.00	0.00		0.00	0.00		0.00	0.00
Primary school area	0.00	0.00	0.14		0.22	0.00		0.00	0.00
Secondary school area	0.00	0.006	0.00		0.009	0.00		0.47	0.00
Bus station	0.00	0.35	0.35		0.52	0.00		0.17	0.009
......									
Enterprise number	0.00	0.02	0.27		0.27	0.00		0.00	0.00

4.3 Label Results Based on AHP

According to the principle of analytic hierarchy process (AHP), the judgment matrix is constructed for eight hierarchical models. In this experiment, the consistency judgment of each judgment matrix is carried out, and the eight judgment matrices in this experiment are consistent. Then, we calculate the eigenvector of each judgment matrix, the weight of the sub-element relative to the upper-element in each hierarchical model is obtained. The weight calculation results are shown in Table 4.

For the quantitative analysis of the waterlogging impact, the influence degree is divided into ten levels, respectively, 1 to 10, and each value represents the size of the influence degree. After obtaining the relative weight of each spatial data, combining with its numerical value, the weighted sum will be calculated, and the calculation results can be mapped into the influence degree. The label results are shown in the Table 5.

Table 4. Weight calculation results

B-A		C-B		C-A	
B1-A	0.0932	C11-B1	1.0000	C11-A	0.0932
B2-A	0.0244	C21-B2	0.6480	C21-A	0.0158
		C22-B2	0.2299	C22-A	0.0056
		C23-B2	0.1222	C23-A	0.0030
B3-A	0.0554	C31-B3	0.7501	C31-A	0.0416
		C32-B3	0.2500	C32-A	0.0139
B4-A	0.1547	C41-B4	0.2277	C41-A	0.0352
		C42-B4	0.0897	C42-A	0.0139
		C43-B4	0.0452	C43-A	0.0070
		C44-B4	0.6373	C44-A	0.0986
B5-A	0.2553	C51-B5	0.0317	C51-A	0.0081
		C52-B5	0.2427	C52-A	0.0620
		C53-B5	0.1449	C53-A	0.0370
		C54-B5	0.0597	C54-A	0.0152
		C55-B5	0.0949	C55-A	0.0242
		C56-B5	0.4261	C56-A	0.1088
B6-A	0.0352	C61-B6	0.1022	C61-A	0.0036
		C62-B6	0.1757	C62-A	0.0062
		C63-B6	0.0475	C63-A	0.0017
		C64-B6	0.6745	C64-A	0.0237
B7-A	0.3818	C71-B7	1.0000	C71-A	0.3818

where A is the influence degree of waterlogging, B_i represents the spatial data of the i-th category associated with the waterlogging impact, and C_{ij} is the j-th spatial data of the i-th category, the specific meaning can refer to the hierarchical model in Figure 2. B_i-A denotes the relative weight of the i-th category to the waterlogging impact (A), and C_{ij}-B_i denotes the relative weight of the j-th subelement to the i-th category. The above two relative weights are multiplied to obtain the relative weight C_{ij}-A, which denotes the relative weight of each type of spatial data to waterlogging impact.

Table 5. Labels of influence degree

Number	Weighted sum	Label	...	Number	Weighted sum	Label
0	0.03	1	...	120	0.09	1
1	0.18	2	...	121	0.07	1
2	0.26	3	...	122	0.10	2
3	0.11	2	...	123	0.08	1
4	0.14	2	...	124	0.06	1

5	0.16	2	…	125	0.03	1
6	0.17	2	…	126	0.05	1
7	0.31	4	…	127	0.03	1
8	0.49	5	…	128	0.04	1
9	0.17	2	…	129	0.07	1
10	0.17	2	…	130	0.03	1
11	0.30	3	…	131	0.07	1
12	0.25	3	…	132	0.04	1
13	0.77	8	…	133	0.08	1

4.4 Model Accuracy Evaluation

As for the traditional linear regression model, above statistical training samples and their labels are used to train the model parameters, and the accuracy of the model is evaluated by the test samples. As shown in Figure 4, there is a big difference between label values and output values of the linear regression model. Therefore, the linear regression model can't meet the requirements of practical application.
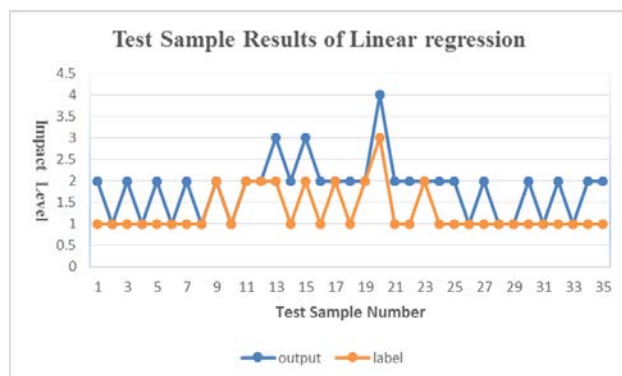

Figure 4. Accuracy evaluation of linear regression model

In terms of Stacked Autoencoder model, the same training samples and test samples are used to train the experimental model. When the number of neurons in the first hidden layer is set to 27 and the second layer is set to 15, the precision of the model is high, which can reach up to 94.286%. Comparing the label values of the test samples with the classification values (as shown in Figure 5), it can be seen that the classification results of the No. 18 (National Road South Road) and No. 34 (Sunshine Road) test samples are second level, but their label values are 1. The output results of the other 33 test samples are completely coincident with their labels, so the accuracy of the model is up to 94.286%. Therefore, the trained Stacked Autoencoder model has the capability of realizing the prediction of waterlogging influence degree, and which can be used to provide reference for the waterlogging prevention and control work of the management department.
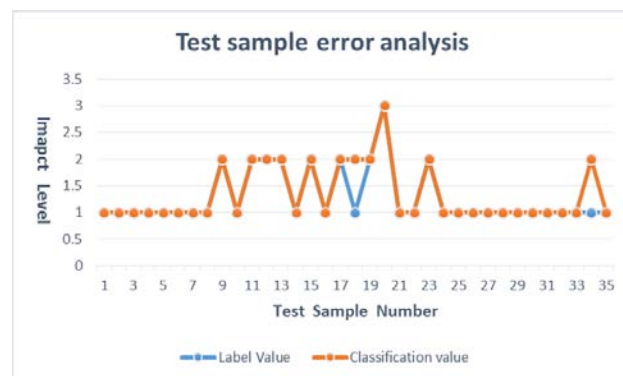

Figure 5. Analysis of test sample results

4.5 Prediction Results of Waterlogging Influence Degree

With high precision waterlogging impact analysis model, this paper searches 30 waterlogging points based on the information of waterlogging location from over the years in Wuhan. The distribution information of these points is shown in Figure 6. According to the statistical extraction method of waterlogging points, the spatial data related to the influence degree of 30 waterlogging points are extracted, and the statistical results are shown in the Table 6.
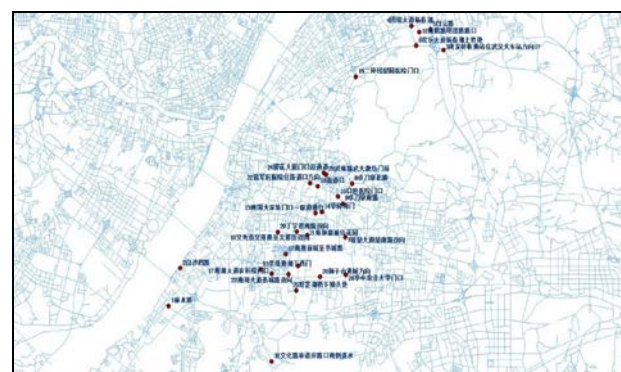

Figure 6. Distribution of waterlogging points over the years

Table 6. Spatial data table of waterlogging points over the years

Number	1	2	3	…	15	16	…	29	30
Housing area	0.208	0.030	0.134		0.363	0.185		0.327	0.342
Library number	0.000	0.000	0.000		0.000	0.000		0.000	1.000
Museum number	0.000	0.000	0.000		0.000	0.000		0.000	0.000
Cultural relics protection unit	0.000	0.000	0.000		0.000	0.000		0.200	0.000
Primary school area	0.000	0.000	0.199		0.129	0.000		0.494	0.000
Secondary school area	0.000	0.000	0.000		0.154	0.000		0.243	0.000
Bus station	0.174	0.000	0.261		0.217	0.435		0.435	0.348
……									
Enterprise number	0.000	0.000	0.000		0.544	0.500		0.012	0.003

The above data of the waterlogging points are used as the input data of the trained model, and the high precision model will predict their influence degrees on the work and life of the residents. The concrete results are shown in Figure 7.
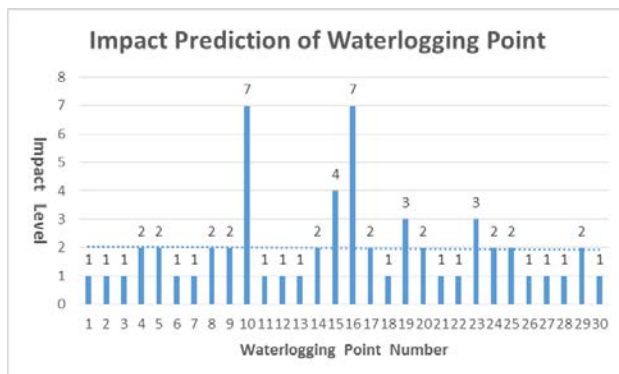
Fig.7. Influence degree prediction map of waterlogging points

From the results of model prediction, the influence degree of waterlogging points are generally distributed in the first and second level, as shown in dotted line. The individual prediction value of waterlogging influence is far more than other waterlogging points, such as No. 10 (Wenxiu Street Wenxiang Road to Wenxin Street two-way), No. 16 (Huang Lu Road, Ming Road junction), whose waterlogging influence degree are 7. The part of waterlogging points' prediction values are slightly higher than the others', such as No. 15 (Second Ring Road, Liyuan Hospital entrance), No. 19 (Stomatological Hospital entrance) and No. 23 (Southland Installed Gate entrance), whose influence degree are between 3 to 4.

Therefore, these waterlogging points should be regarded as the key areas of waterlogging disaster prevention and control in Wuhan. The management departments can analyse their regional impervious rates, topography, drainage capacity, so as to find out the causes of waterlogging disaster and give priority of planning governance to these areas.

5. CONCLUSION

This paper puts forward a kind of quantitative analysis method on influence analysis of waterlogging based on deep learning model. The statistical analysis of the spatial data around 135 waterlogging points, combined with Stacked Autoencoder neural network, the paper trained a waterlogging influence prediction model with high precision, so as to complete the quantitative analysis in waterlogging influence on residents' work and life. At the same time, the relationship between the spatial data series and the influence degree of the waterlogging was excavated, so the influence degree of the other waterlogging points can be predicted accurately. These prediction results can provide a guidance for the prevention and control of waterlogging disaster.

ACKNOWLEDGEMENTS

REFERENCES

Xia, W. (2013). The cause analysis and countermeasures of waterlogging in wuhan. *Applied Mechanics & Materials, 357-360*, 1962-1965.

Xue, F., Huang, M., Wang, W., & Zou, L. (2016). Numerical simulation of urban waterlogging based on floodarea model. *, 2016*(1), 1-9.

Liu, Y., Du, M., Jing, C., & Cai, G. (2014). *Design and Implementation of Monitoring and Early Warning System for Urban Roads Waterlogging. Computer and Computing Technologies in Agriculture VIII.* Springer International Publishing.

Liao, W., Wang, C., Zhou, X., & Wang, Z. (2014). Simulation and application on storm flood in dongguan city based on swmm. *Journal of Mass Spectrometry, 43*(1), 42-52.

Chen, X., Deng, H., Ma, X., & He, Y. (2010). Research on the matched recurrence interval relationship between urban waterlogging control and drainage system based on storm water management model. *International Symposium on Aqua Science, 1251*(1), 141-144.

Quan, R. S., Liu, M., Lu, M., Zhang, L. J., Wang, J. J., & Xu, S. Y. (2010). Waterlogging risk assessment based on land use/cover change: a case study in pudong new area, shanghai. *Environmental Earth Sciences, 61*(6), 1113-1121.

Jin, J. L., Fu, J., Wei, Y. M., Jiang, S. M., Zhou, Y. L., & Liu, L., et al. (2015). Integrated risk assessment method of waterlog disaster in huaihe river basin of china. *Natural Hazards, 75*(2), 155-178.

Rahadianto, H., Fariza, A., & Hasim, J. A. N. (2015). Risk-level assessment system on Bengawan Solo River basin flood prone areas using analytic hierarchy process and natural breaks: Study case: East Java. International Conference on Data and Software Engineering(pp.195-200). IEEE.

Li, W., Fu, H., Yu, L., Gong, P., Feng, D., & Li, C., et al. (2016). Stacked autoencoder-based deep learning for remote-sensing image classification: a case study of african land-cover mapping. *, 37*(23), 5632-5646.

Zabalza, J., Ren, J., Zheng, J., Zhao, H., Qing, C., & Yang, Z., et al. (2016). Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. *Neurocomputing, 185*, 1-10.

Lu, S., Liu, H., & Li, C. (2015). Manifold Regularized Stacked Autoencoder for Feature Learning. *IEEE International Conference on Systems, Man, and Cybernetics* (pp.2950-2955). IEEE.

Budiman, A., Fanany, M. I., & Basaruddin, T. (2015). Online Marginalized Linear Stacked Denoising Autoencoders for Learning from Big Data Stream. *ICACSIS 2015 at PSJ UI.*

Fariza, A., Hasim, J. A. N., & Rahadianto, H. (2015). Flood risk mapping of Bengawan Solo river in east java using the analytic hierarchy process and head/tails break classification. *International Conference on Science in Information Technology* (pp.331-336). IEEE.