Forecast Model of Urban Stagnant Water Based on Logistic Regression

Pan LIU[1,a], Jian-Zhuo YAN[1], Miao-Wen JIANG[1], Mei LIU[2],Xiao-Lan YIN[2] and Xiao-Juan ZHANG[2]

[1]*Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*
[2]*Beijing Water Affairs Information Management Center, Beijing 100038, China*

Abstract. With the development of information technology, the construction of water resource system has been gradually carried out. In the background of big data, the work of water information needs to carry out the process of quantitative to qualitative change. Analyzing the correlation of data and exploring the deep value of data which are the key of water information's research. On the basis of the research on the water big data and the traditional data warehouse architecture, we try to find out the connection of different data source. According to the temporal and spatial correlation of stagnant water and rainfall, we use spatial interpolation to integrate data of stagnant water and rainfall which are from different data source and different sensors, then use logistic regression to find out the relationship between them.

1 Introduction

Big Data Analytics (BDA) is the core of big data concepts and methods. It analyzes the massive, diverse, fast-growing, and content-rich data (big data) to find out the hidden patterns, unknown correlation and other useful information [1]. But the very key of this whole process is to prepare the data which is suitable for data mining, which means we need to do the data fusion job to combine the data from different source.

Data fusion is different from the traditional data integration or knowledge database technology, it requires large span, deep and comprehensive research methods. Data fusion is the process of integrating two or more data. The purpose of the process is to generate an improved data set, which can be superior to the source dataset or the input dataset both in geospatial and attribute traits. Big Data is a complex set of data with large, diverse, high-speed variations [2-3], in these data, spatial data accounted for the vast majority, about 80% of the data has spatial location [4-5]. So in this paper we combine stagnant water data and rainfall data which are two kinds of data source into one dataset by their geographic Information Attributes and their time attribute, but then we may face a problem which is we cannot find these two properties on the same point at the same time. For example, we can get the data of stagnant water at one point and we assume its coordinate is (x,y), but we cannot find the rainfall data in database at the same point, but we can get other rainfall data around this point. So based on the geographical coordinates of the stagnant water, we calculate the value of the rainfall data of this point by the method of spatial interpolation.

[a] Corresponding author: 2795611970@qq.com

In this paper, we will use the method of Inverse Distance to a Power (IDW) to get the data of rainfall, then we use rainfall data as inputs, stagnant water data as output, and logistic regression as a method for machine learning to find the potential relationship between these two.

2 Data preparation

At present, the water data fusion between stagnant water and rainfall is only limited to transferring the data from each sensor into the data warehouse, because the different functions of the sensors are different, the correlation between the tables and the tables is transmitted, weak. However, because each sensor has its own unique spatial information, then I introduced IDW-based spatial interpolation method to integrate the data from different sensors, which means I calculate the rainfall data of this area which has stagnant water data. At last, we get a one point's data of stagnant water, while also be able to get this point's data of rainfall.

2.1 Preprocessing the data of rainfall

First of all, the current structure of the data itself which we can see it in table1.

Table 1. The current structure of data.

ID	Varchar(32)
Code	Varchar(16)
Time	Date
Z	Number(10,2)

The data stored in database comes with a unique code, and this code represents the spatial attribute of the data. In order to calculate the value by IDW, we need to change the structure of the table which is picking every code as a column and make Z as its value. For example, in a table we get 3 kinds of code which are "xxx", "yyy" and "zzz", then we change the structure of this table as Table2.

Table 2. The changed structure of data.

Time	Date
XXX-Z	Number(10,2)
YYY-Z	Number(10,2)
ZZZ-Z	Number(10,2)

Above all, the steps are:
(1) Make the raw data arranged by every hour per day;
(2) Change the table's structure to fit the IDW method just as we talked about above;
(3) Calculate the value by IDW.

2.1.1 Changing the structure of the table by SPSS Modeler

SPSS Modeler is a set of data mining tools that enable you to quickly build predictive models using business technology and apply them to business activities to improve your decision-making process. Designed with reference to the industry standard CRISP-DM model, SPSS Modeler supports the entire data mining process from data to better business results. Flow chart shown in Figure 1.

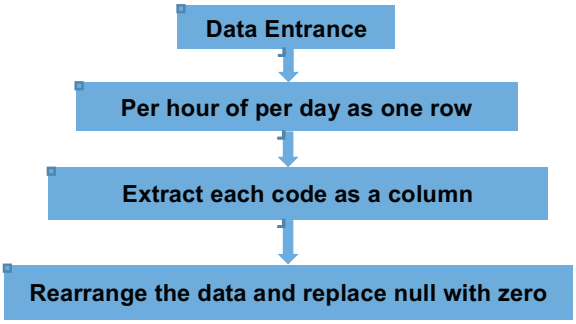

Figure 1. Flow chart of the rainfall data process by spss

SPSS provides “data disaggregation” components through which a given time field can be divided into two columns which are day and hour. At the meantime, SPSS also provides “data reconstruction” components through which each code can be extracted as one column. As shown in Figure 2 is the rainfall data processing flow chart.

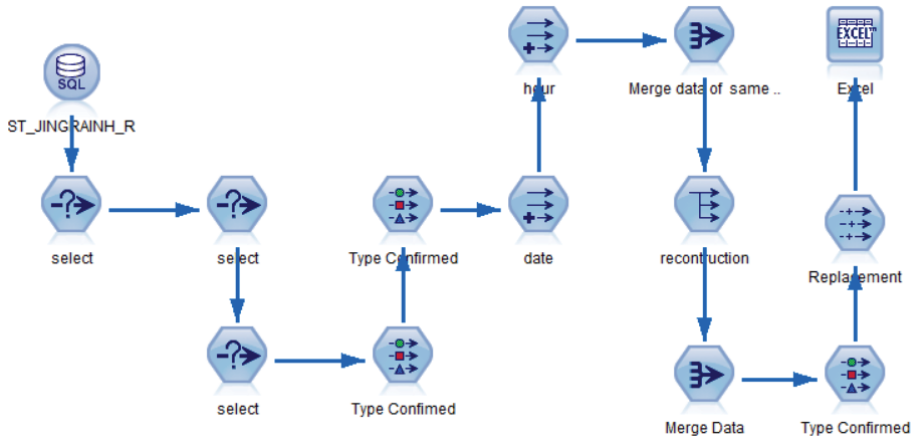

Figure 2. Flow chart of using SPSS to do data processing

As you can see, the flow chart upper there have components named date and hour of which used to change the time dimension to be hour, and the first merge data component is to get the average value of rainfall from same hour same day and record the total number of merge items, and the reason why we want to know how much rows we have merged is to determine the duration of the rainfall, part of the result is stored in excel which we can see in Figure 3. In Figure 3, each column in the first row represents the meaning of the data from this column. The first two columns are well understood, and we start with the third column. The third column represents the amount of rain collected for one hour, excluding the rainfall is zero, which is used to determine the length of time since data is collected every 15 minutes. After the third column represents the rainfall of each point at a certain time

DATE	HOUR	RECORD	30521900_	30522000_	30522850_	30523900_	30523960_	30524200_	30524300_	30526500_
2012/7/22	6	4	226	171	152	197	184	172	172	136
2012/7/22	4	4	226	171	151	197	184	172	171	136
2012/7/22	3	4	226	171	149	197	184	172	171	136
2012/7/22	2	4	226	170	147	195	182	168	170	136
2012/7/22	1	4	225	159	119	191	177	164	166	135
2012/7/22	0	4	222	151	119	188.5	176	162	161	132
2012/7/21	23	1	222	151	119	184	170	152	152	121
2012/7/21	22	1	222	149	117	176	160	145	128.5	109.5
2012/7/21	21	1	220	146	101.5	168	150	138	101	88.5
2012/7/21	20	1	219	142.5	71	154.5	129	104.5	36.5	62
2012/7/21	19	3	212	135	42	104.5	75	50	24.5	43.5
2012/7/21	18	1	208	126.5	26.5	42	30	21	21	24.5
2012/6/25	6	1	34	51	20	67	52	51	28	39

Figure 3. The result set for IDW interpolation

2.1.2 Rainfall data interpolation based on IDW

Inverse Distance to a Power interpolation is first proposed by meteorologists and geologists. Inverse Distance to a Power is the earliest computer interpolation method and is still widely used. Its basic principle is to distribute a series of discrete points on the plane, if we know some point's position coordinates $(x_i, y_i)$and the attribute valueZ_i, according to the attribute values of the surrounding discrete points, the P-point attribute value is interpolated by the Inverse Distance to a Power[6,7]. If there are N data points around, the attribute value of point P is:

$$P(Z)=\sum_{i=1}^{N} \frac{Z_i}{[d_i(x,y)]^u} \Big/ \left(\sum_{i=1}^{N} \frac{1}{[d_i(x,y)]^u}\right), i = 1,2, \dots N \tag{1}$$

And $d_i(x,y)=\sqrt{(x-x_i)^2+(y-y_i)^2}$ means the distance from the i-th data point to the p-point. In this paper, we only need to know the gis coordinates of all the rainfall points in the urban area and the targeted coordinates , so we can calculate the distance between each rainfall point and the targeted point. According to IDW interpolation method, we can get the influence factors of each rainfall point which we name it λ_i.

$$\lambda_i = \frac{\frac{1}{d_i}}{\left(\sum_{i=1}^{n} \frac{1}{d_i}\right)} \tag{2}$$

Then put each point's rainfall into the formula (3),we can get the value of targeted point.

$$\hat{Z}(x,y) = \sum_{i=1}^{n} \lambda_i Z(x_i, y_i) \tag{3}$$

According to figure 3, we can easily calculate the rainfall of targeted point at every hour every day.

2.2 Integrate the data of the stagnant water and the rainfall

We need to determine the possibility of stagnant water happened in this area according to its rainfall, so we plan to use the logistic regression to find out the relation between stagnant water data and rainfall data, we use rainfall data as input and stagnant water data as output. The accumulation area, ground structure, and drainage of the individual water points are different. Therefore, different models should be established for different water accumulation points, in this paper, we choose the point of the bridge of huaxiang as an example.

At first, we need to calculate the rainfall of one stagnant water point by IDW and change the value of stagnant water data into 0 or 1 based on whether stagnant water happened in this area, then we can combine the stagnant water data of this point with the rainfall data through the time dimension. There are three combinations of ways which are external connections, internal connections and partly

external connections. In this paper, we choose partly external connection and take the time of stagnant water as reference. The result set are stored in excel which we can see in figure 4. And in figure 4, the first two columns are time which is based on the original stagnant water, the column called rainfall is calculated by the method of IDW.

A	B	C	D
DATE	HOUR	Rainfall	Stagnant water
2012/7/22	6	199.0206	1
2012/7/22	4	198.9362	1
2012/7/22	3	198.8662	1
2012/7/22	2	196.5751	1
2012/7/22	1	188.27075	1
2012/7/22	0	183.6454	1
2012/7/21	23	178.9515	1
2012/7/21	22	171.33905	1
2013/7/31	23	12.1026	0
2013/7/31	6	15.2719	0
2013/7/31	2	14.858	0
2013/7/31	1	12.4662	0
2013/7/16	6	26.2502	0
2013/7/8	12	15.8921	0

Figure 4. The result set for logistic regression

2.2.1 Filling in the vacancy value of the rainfall data

In this paper, we only fill the data of its stagnant water data is 1 and delete the rest unfilled data, because most of the data to be analyzed has a stagnant water property of zero, and in order to balance the data set, we need to reduce it. We choose the method of "linear regression" which is we consider the linear correlation between rainfall and time. The specific means is "if there are missing value in the same day, we calculate the data of the rainfall by the other data in the same day to fill, if not, we delete this row".

2.2.2 Data grouping based on the rainfall

Because the distribution of the entire dataset is not even, we cannot use single formula to describe the whole situation. By querying the Bureau of Meteorology on the definition of rainfall levels, we divide our data into five parts which we can see it in Table3.

Table 3. Data grouping based on rainfall levels.

Rainfall levels	**Rainfall**
Light rain	0.1-4.9mm
Medium rain	5.0-9.9mm
Heavy rain	10.0-29.9mm
Rainstorm	30.0-69.9mm
Super rainstorm	>70mm

3 Logistic regression analysis of rainfall and stagnant water

In the stagnant water analysis, rainfall can be used as independent variables, and the occurrence of stagnant water can be used as binary variables (0 on behalf of the stagnant water does not occur, and 1 represents stagnant does occur,). When a dependent variable is a bivariate variable, a multivariate logistic regression model is used to generate the regression coefficients for the respective variables based on the sample data and to discuss the relationship between the dependent and independent variables in the model. Let p be the probability of occurrence of the event, in the range of 0 to 1, then

1-p is the probability that the event does not occur, this probability can be calculated using logistic function, the expression is [8]:

$$p = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{k}\beta_i x_i)}} \tag{4}$$

Logistic function is a nonlinear function of covariance, in order to obtain the regression coefficient, the logit transformation of (4), to obtain a linear formula :

$$ln\,(^{p}/_{1-p}) = \beta_0 + \sum_{i=1}^{k}\beta_i \tag{5}$$

In this formula,X_i are the independent variables, and β are the coefficients for the variables, formula(5) is also called the odds ratio (OR). Since the OR has some good properties in measuring the association [9], it can be used to describe the effect of the independent variables on the event probability in the logistic regression model, Therefore, it is often used to interpret the regression coefficients of the logistic regression model [10].

3.1 Data preprocessing for logistic regression

3.1.1 Determine the data range based on the data distribution

Before data analysis, we need to look at the quality of the data of each group. We describe the quality of each group by stagnant water distribution ratio. Details can be seen in Table 4.

Table 4. Stagnant water distribution ratio of each data group.

Rainfall levels	Stagnant water(1)	Not stagnant water(0)
Light rain	3%	97%
Medium rain	14%	86%
Heavy rain	27%	73%
Rainstorm	72%	28%
Super rainstorm	92%	8%

As you can see, light rain only has the possibility of 3% to cause stagnant water and super rainstorm has the possibility of 92% to cause stagnant water, so we deem light rain with a very low possibility to cause stagnant water and rainstorm with a very high possibility to cause stagnant water.

As we know, not necessarily large rainfall will cause stagnant water, stagnant water is related with the duration of rain and urban drainage capacity. In all case, the urban drainage capacity is changeless and that makes the duration of rain to be our only concern. According to the data, almost 90% of the rainstorm lasts less than one hour so that we can take an hour as a unit analyse the relationship between the stagnant water and rainfall, and also we believe that the rainstorm can produce stagnant water more suddenly, and the other groups, we cannot simple assume the duration of the rainfall is less than one, so in this paper, we choose the group of rainstorm to analyse.

3.1.2 logistic regression between rainstorm and stagnant water

Before we put the group of rainstorm into logistic regression training, we only have the data of the average of the rainfall as one input and it's not enough, so we think about the length of the rainfall and we can get the number of the records of the rainfall through the section 2.1.1, and we can determine whether the duration of the rainfall is less than half hour through the record, if the record is less than 3

means the length of the rainfall is less than half hour, The data for logistic regression is below in Figure5.

Rainfall	rainfallPeriod	Stagnant water
69.7047	0	1
63.9312	1	1
61.8981	0	1
57.76515	1	1
54.7541	1	1
54.6983	0	1
54.3044	1	1
54.2553	1	1
53.8186	0	1
53.1894	1	1
53.1522	1	1
53.0445	0	1
52.9581	1	1
52.3235	1	1
51.7775	1	1
50.7895	0	1
50.6901	1	1
50.0265	1	1
49.4876	1	1
48.8222	1	1
48.5478	1	0

Figure 5. The data set of rainstorm and stagnant water for logistic regression

In Figure 5, there are 3 columns, representing rainfall, the record of the rainfall, the existence of stagnant water, we use rainfall and continuous as input and stagnant water as output to perform logistic regression train.

After that, we find out the coefficient of the rainfall corresponding to the significant level of Sig less than 0.05, the regression results can be considered by 5% significance level test. So the formula of logistic regression is

$$p = \frac{1}{1+e^{-(-11.239+0.240x+0.761y_1)}} \tag{5}$$

In those formula, x stands for rainfall, y1 stands for the rainstorm whether lasts half an hour.

3.2 The test of regression equation

In order to test this model, we randomly selected some of the rainfall points which can be thought as rainstorm in the year of 2016 and the year of 2015, and the result is in Figure 6.

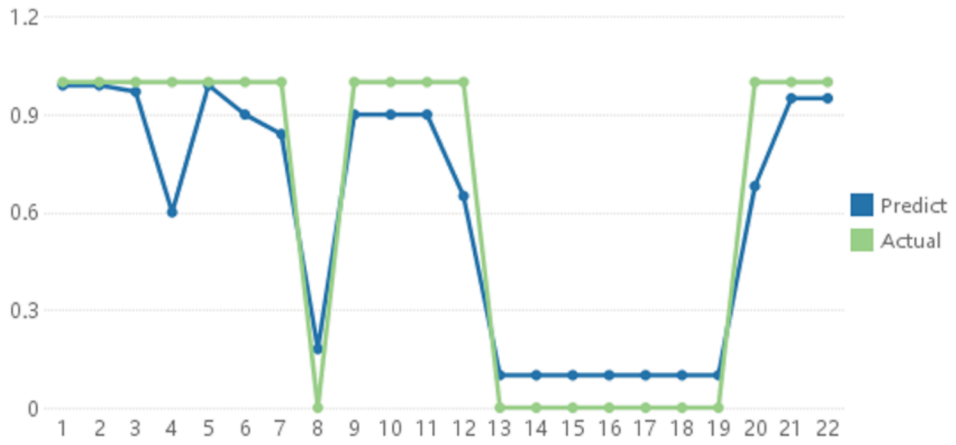

Figure 6. Comparison of Predictive Probability and Actual Value

As you can see, the green part represents the actual stagnant water status, and the blue part represents the predict stagnant water status. The value of 1 is the occurrence of water and the value of

0 represents no stagnant water occur. So according to this figure, we can see when there is actually stagnant water, the probability of the forecast mostly more than 80%, while there is no stagnant water, the probability of forecasts mostly less than 20%.

Conclusion

In this paper, we use the idea of big data analysis as the core, make the different sources of data together according to their spatial and temporal attributes by spatial interpolation, make the previous independent one-dimensional data into two-dimensional data. In the past studies, most of the scholars are through the analysis of drainage measures and other factors to determine whether the stagnant water exist, or according to the stagnant water to analyze the trend of stagnant water, but in this paper, we use logistic regression to get the relation of stagnant water and rainstorm. Through rainfall and rain duration to predict whether the water, which is based on a data mining ideas. With the formula we get we can get the possibility of the existence of stagnant water. In this paper, we only consider the amount of the rainfall and we hadn't considered the randomness of the rainfall, so if we want to predict stagnant water risk precisely, we may need to consider the randomness of rainfall on the basis of the logistic regression we established.

References

1. LAPKIN A. Hype Cycle for Big Data[R]. Gartner, Inc.G00235042,2002.
2. DENSHAMPJ, GOODCHILD M F. Spatial Decision Sup-port Systems: A Research Agenda [C]/ / Proceedings GIS / LIS 89, Orlando, FL, 1989: 707-716.
3. SHEKA R S, XIONG H(Eds). Encyclopedia of GIS[M]. New York: Springer,2007.
4. MILLER.H.J,HAN.J. Geographic-Data-Mining and Knowledge Discovery[M].2nd edition. London: Taylorand Francis, 2009.
5. ESTEM, et.al. Spatial Data Mining: Databases Primitives, algorithms and efficientDBMS support [J].Data-Mining and Knowledge Discovery,2000(4): 193-216.
6. Wu Lun, Liu Yu, Zhang Jin. Geographic Information System Principles and Applications[M]. Beijing Science Press, 2001:184-185.
7. Wu xincai. Geographic Information System[M] Electronic Industry Press,2002.172-173
8. Wang jichuan, Guo zhigang. The logistic regression model: methods and application.[M] Beijing: Higher Education Press,2001.
9. Bui D T, Lofman O, Revhaug I, et al. Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression[J].Natural Hazards, 2011, 59(3): 1413-1444.
10. Xie Hualin. Analysis of regionally ecological land useand its influencing factors based on a logistic regressionmodel in the Beijing-Tianjin-Hebei region, China[J].Resources Science, 33(11): 2063－2070. (in Chinese with English abstract)