Pollution analysis and prediction based on Voronoi Maps and neural networks

Zejia Zheng
School of Mathematics
Fudan University
Shanghai, China

Liwen Zhou
Department of Computer Science
Fudan University
Shanghai,China

Longfei Ren
School of Economics
Fudan University
Shanghai,China

***Abstract*—Attention of this paper focuses on the evaluation of pollution condition of a city and finding the pollution sources using data gathered from observation stations all over the city. We assume every observation data represents the condition of pollution in a certain area, which was given by formulating a Voronoi map of the region with the center points of each Voronoi area representing the observation station. By calculating the area of each Voronoi area, we assigned weight to each observation station, thus modifying the traditional method to evaluate the pollution condition. We further assume that pollutants will only be transferred between adjacent Voronoi areas. So the existence of pollution transfer can be represented by drawing a line between adjacent observation stations, thus giving us the Delaunay Triangulation of the city. We substitute the observation station as neurons, the edge of the Delaunay Triangulation as synapses, thus getting a neural network. According to our assumption, the transfer rate of pollution between two neurons, or the weight of the neural network, depends on the difference of their density of pollution, altitude, and the distance between the two observation stations. We show that prediction and back-track pollutant transfer can be realized naturally from our network.**

Keywords-Neural Networks, Voronoi Map, Pollutant

I. Introduction

We are given current observation data (geographical location, altitude, soil concentration of 8 heavy metal pollutants) from 319 observation stations all over the city. We were also given background statistics gathered from unpolluted areas in the city.

Our model achieves the following goals:

i) Find the spatial distribution of pollutants. Evaluate the seriousness of pollution.

ii) Find the location of the source of pollution.

iii) Predict and backtrack the flow of pollutants.

II. Mathematical Models

A. Analyzing data

319 soil samples to be studied come from five different categories of districts, namely, living district, industrial district, mountainous district, transportation district, and park district. Descriptive statistics of measured results of heavy metal concentration is shown in Table 1a-e.

B. Evaluating the curent condition

Traditional pollution evaluation methods are Single Factor Index Method (SFIM) and Comprehensive Index Method (CIM). [1]

1) *SFIM*

Evaluation is given by the following formula:

$$P_i = \frac{C_i - b_i}{Std_i - b_i}$$

Where P_i is the pollution index of the ith heavy metal pollutant C_i is the measured concentration of the ith heavy metal.

Std_i is the national secondary standard of the ith heavy metal concentration in soil (see Appendix II), and b_i is the background value from the unpolluted area.

Traditional methods evaluate the whole area by calculating the total average of P_i.

$$\mathrm{Eval} = \frac{1}{n}\sum P_i$$

However, observation stations in this particular problem are not evenly distributed. Thus we improve SFIM by giving each P_i a weight. P_i's weight shows the area that the ith observation station can represent.

We derived weight from Voronoi map of the observation stations. That is, P_i's weight W_i is the area of P_i's Voronoi region. Thus, the improved formula is:

$$\mathrm{Eval\_vor} = \frac{\sum(P_i * W_i)}{\sum W_i}$$

The result of the improved evaluation method is shown in Table 2.

2) *CIM*

One disadvantage of SFIM is that SFIM only evaluates on specific heavy metal. CIM, on the other hand, is able to highlight role of several heavy metal pollutants.

	As (µg/g)	Cd (ng/g)	Cr (µg/g)	Cu (µg/g)	Hg (ng/g)	Ni (µg/g)	Pb (µg/g)	Zn (µg/g)
Mean value	6.27	289.96	69.02	49.40	93.04	18.34	69.11	237.01
Standard deviation	2.15	183.68	107.89	47.16	102.90	5.66	72.33	443.64
Maximum	11.45	1044.50	744.46	248.85	550.00	32.80	472.48	2893.47
Minimum	2.34	86.80	18.46	9.73	12.00	8.89	24.43	43.37
Coefficient of variation	0.34	0.63	1.56	0.95	1.11	0.31	1.05	1.87
Sample size	44	44	44	44	44	44	44	44

Table 1a Descriptive statistics of soil heavy metals concentration in living districts

	As (µg/g)	Cd (ng/g)	Cr (µg/g)	Cu (µg/g)	Hg (ng/g)	Ni (µg/g)	Pb (µg/g)	Zn (µg/g)
Mean value	7.25	393.11	53.41	127.54	642.36	19.81	93.04	277.93
Standard deviation	4.24	237.58	44.00	414.94	2244.07	8.37	85.37	350.83
Maximum	21.87	1092.90	285.58	2528.48	13500.00	41.70	434.80	1626.02
Minimum	1.61	114.50	15.40	12.70	11.79	4.27	31.24	56.33
Coefficient of variation	0.59	0.60	0.82	3.25	3.49	0.42	0.92	1.26
Sample size	36	36	36	36	36	36	36	36

Table 1b Descriptive statistics of soil heavy metal concentration in industrial districts

	As (µg/g)	Cd (ng/g)	Cr (µg/g)	Cu (µg/g)	Hg (ng/g)	Ni (µg/g)	Pb (µg/g)	Zn (µg/g)
Mean value	4.04	152.32	38.96	17.32	40.96	15.45	36.56	73.29
Standard deviation	1.80	78.38	24.59	10.73	27.85	10.43	17.73	30.94
Maximum	10.99	407.60	173.34	69.06	206.79	74.03	113.84	229.80
Minimum	1.77	40.00	16.20	2.29	9.64	5.51	19.68	32.86
Coefficient of variation	0.44	0.51	0.63	0.62	0.68	0.67	0.49	0.42
Sample size	66	66	66	66	66	66	66	66

Table 1c Descriptive statistics of soil heavy metal concentration in mountainous districts

	As (µg/g)	Cd (ng/g)	Cr (µg/g)	Cu (µg/g)	Hg (ng/g)	Ni (µg/g)	Pb (µg/g)	Zn (µg/g)
Mean value	5.71	360.01	58.05	62.21	446.82	17.62	63.53	242.85
Standard deviation	3.24	243.39	81.61	120.22	2180.27	11.79	32.53	384.78
Maximum	30.13	1619.80	920.84	1364.85	16000.0	142.50	181.48	3760.82
Minimum	1.61	50.10	15.32	12.34	8. 7	6.19	22.01	40.92
Coefficient of variation	0.57	0.68	1.41	1.93	4.88	0.67	0.51	1.58
Sample size	138	138	138	138	138	138	138	138

Table 1d Descriptive statistics of soil heavy metal concentration in transportation districts

	As (µg/g)	Cd (ng/g)	Cr (µg/g)	Cu (µg/g)	Hg (ng/g)	Ni (µg/g)	Pb (µg/g)	Zn (µg/g)
Mean value	6.26	280.54	43.64	30.19	114.99	15.29	60.71	154.24
Standard deviation	2.02	235.84	14.84	22.68	224.28	4.97	45.84	230.92
Maximum	11.68	1024.90	96.28	143.31	1339.29	29.10	227.40	1389.39
Minimum	2.77	97.20	16.31	9.04	10.00	7.60	26.89	37.14
Coefficient of variation	0.32	0.84	0.34	0.75	1.95	0.33	0.76	1.50
Sample size	35	35	35	35	35	35	35	35

Table 1e Descriptive statistics of soil heavy metal concentration in park districts

District	As	Cd	Cr	Cu	Hg	Ni	Pb	Zn
Living	0.12	0.94	0.14	0.42	0.12	0.16	0.14	0.93
Industrial	0.17	1.55	0.08	1.32	1.31	0.20	0.23	1.15
Mountainous	0.02	0.13	0.03	0.05	0.01	0.08	0.02	0.02
Transport	0.10	1.35	0.10	0.56	0.89	0.14	0.12	0.96
Park	0.12	0.89	0.05	0.20	0.17	0.08	0.11	0.47

Table 2 Improved Evaluation

CIM gets one evaluation result by putting weights on P_i of different pollutants:

$$\overline{P} = \frac{\sum_{i=1}^{n} \lambda_i P_i}{\sum_{i=1}^{n} \lambda_i}$$

According to reference[1], pollutants are categorized into 3 different levels by decreasing significance. In our problem, the levels of the pollutants are shown in the table 3a.

	As	Cd	Cr	Cu	Hg	Ni	Pb	Zn
Level	I	I	II	II	I	II	I	II
Weight	3.0	3.00	2.00	2.00	3.00	2.00	3.00	2.00

Table 3a Levels of pollutants

Thus, CIM results are shown in table 3b.

Distruct	Living	Industrial	Mountainous	Transportation	Park
CIM	1.07	1.85	0.26	1.75	0.79

Table 3b CIM results

3) *Conclusion of evaluation*

From previous results, we can see that living district suffers from serious Zn pollution, while Cd, Cu, Hg, Zn pollution affects the industrial districts most.

Detailed results can be seen in table 4.

C. *Voronoi Map and Neural Networks*

1) *Voronoi Map*

By plotting every observation station on a 2 dimensional plane, we get figure 1.

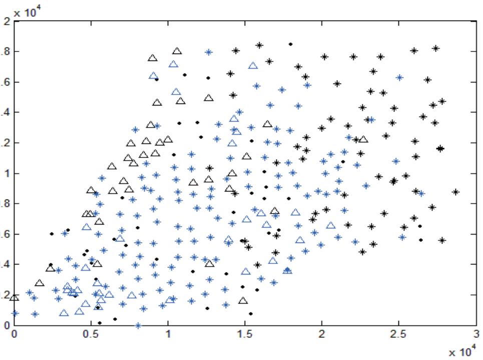

Figure 1 Two-dimensional diagram of 319 observation stations

We assume that each observation station is able to represent its adjacent area. All we have to do is to define the word ‘adjacent’.

Consider the simplest scenario: there are only 2 observation stations. We can divide the whole city by drawing a midnormal of the two stations. Each station represents the area it is located in. See figure 2a .

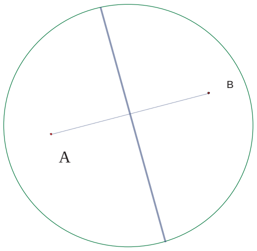

Figure 2a two observation stations

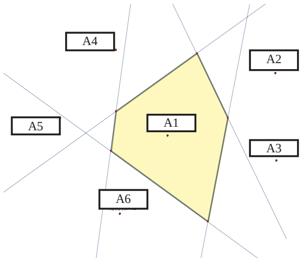

Figure 2b Multiple observation stations

When there are multiple stations, the area of one particular station can be given by the intersection of all its midnormals with other stations. See figure 2b.

This coincides with the definition of Voronoi Map (figure 3).

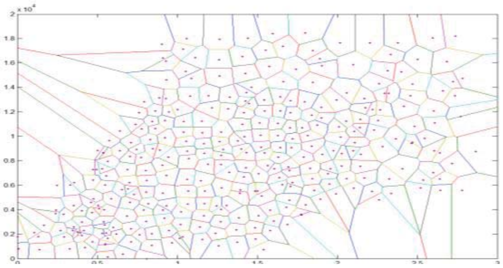

Figure 3 Voronoi Map of the observation stations

2) *Delaunay Map*

It can easily be seen from the Voronoi Map that soil pollutants can only transfer between adjacent Voronoi areas. We connect two observation stations if and only if their Voronoi areas area adjacent. Thus we get the Delaunay Map of the observation stations. Figure 4a(2D) and 4b(3D) shows the result of Delaunay Triangulation.

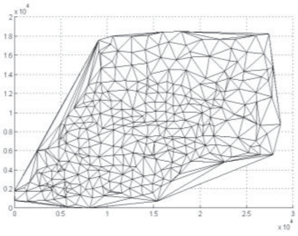

Figure 4a 2D Delaunay Triangulation

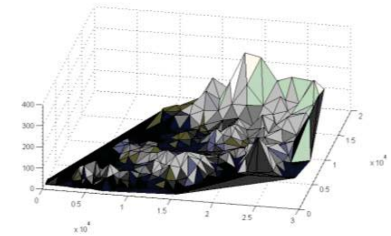

Figure 4b 3D Delaunay Triangulation

3) *Construction of Neural Network*

If we view the edges in the Delaunay Map as synapses and the observation stations as neurons, we will get a neural

Districts	Comprehensive pollution level	Heavy metal pollution	Main heavy metal pollutants
Living	Mild pollution	Cd > Zn> Cu > Ni > Pb > Cr > As= Hg	Cd, Zn
Industrial	Moderate	Cd > Cu > Hg > Zn > Pb > Ni > As > Cr	Cd、Cu、Hg、Zn
Mountainous	Non-pollution	Cd > Ni > Cu > Cr > Pb = Zn = As > Hg	None
Transportation	Moderate	Cd > Zn > Hg > Cu > Ni > Pb > As = Cr	Cd
Park area	Non-pollution	Cd > Zn > Cu > Hg > As > Pb > Ni > Cr	None

Table 4 Detailed results of evaluation

network, which depicts the flow and transportation of pollutants between adjacent Voronoi areas. There exist several pollution sources in the network which can be viewed as outside stimulus.

Define $X_{i,t}$ as the observation result of the ith station at time t. Thus in the network,

$$X_{i,t+1} = f(X_{i1,t}\,, X_{i2,t}\,, X_{i3,t}\,, \dots, X_{in,t+1})$$

i1, i2, i3, … , in are the stations connected with the ith station.

Next section will be focused on finding the recurrence formula f.

4) Recurrence formula f

We need several assumptions to find the suitable f to this network.

Assumption I: the rate of pollutant transportation between two adjacent Voronoi areas is determined by the difference of the areas' altitude, the areas' pollutant soil concentration, and the distance between the two stations.

Assumption II: pollutants only flow from high soil concentration areas to low soil concentration areas.

Assumption III: it is easier for pollutants to flow from high altitude areas to low altitude areas than to flow from low altitude areas to high altitude areas.

Thus, we establish the following difference equation:

$$x_{i,t+1} = \alpha_i(\omega_{i,1}, \omega_{i,2}, \dots, \omega_{i,m}) \begin{pmatrix} x_{1,t} - x_{i,t} \\ x_{2,t} - x_{i,t} \\ \vdots \\ x_{m,t} - x_{i,t} \end{pmatrix} + x_{i,t}$$

α_i is the coefficient represent the property of soil. It shows that different soil in different location may affect the rate of pollutant transportation. m is the number of observation stations, or in this particular problem, 319.

$\omega_{i,j}$ is the weight representing the influence of the jth station on the ith station. According to assumption I, it depends on by the difference of the areas' altitude and the distance between the two stations.

We define $\omega_{i,j}$ as:

$$\omega_{i,j} = \begin{cases} g(dist(i,j), h_j - h_i) & if\ i\ is\ connected\, to\ j \\ 0 & else \end{cases}$$

We use the sigmod function to represent g.

$$g(dist(i,j), h_j - h_i) = \frac{1}{\exp\left[k_{i,j}\,(h_i - h_j) + dist(i,j)\right]}$$

Define

$$\delta_{i,j} = \begin{cases} 1 & if\ i\, is\ connected\ with\ j \\ 0 & else \end{cases}$$

Then

$$\omega_{i,j} = g(i,j) * \delta_{i,j}$$

Until now we have only considered how the flow of pollutants under current circumstances. We now take the factor of pollution source into consideration. Define

$$S_i = \begin{cases} Pi & \text{If i is pollution source} \\ 0 & \text{else} \end{cases}$$

P_i is the amount of pollution the pollution source produces from t to t+1.

To sum up, we have the following recurrence equations:

$$\begin{cases} x_{i,t+1} = \alpha_i(\omega_{i,1}, \omega_{i,2}, \dots, \omega_{i,m}) \begin{pmatrix} x_{1,t} - x_{i,t} \\ x_{2,t} - x_{i,t} \\ \vdots \\ x_{m,t} - x_{i,t} \end{pmatrix} + x_{i,t} + S_i \\ \omega_{i,j} = \dfrac{1}{\exp\left[k_{i,j}\,(h_i - h_j) + dist(i,j)\right]} * \delta_{i,j} \\ \delta_{i,j} = \begin{cases} 1 & if\ i\, is\ connected\ with\ j \\ 0 & else \end{cases} \\ S_i = \begin{cases} Pi & if\ i\ is\ pollution\ source \\ 0 & else \end{cases} \end{cases}$$

5) Locating the pollution sources

a) Intuitive Deduction

We plotted the spatial distribution of the eight pollutants by MATLAB.

Apparently, the most polluted area should be the source of pollution. Table 5 shows the most polluted observation stations according to the given data. The number in the table is the number of the observation stations. These stations should be the potential pollution sources.

Heavy metal pollutant			Number of observation stations		
As	84	178	29	30	41
Cd	95	22	9	6	8
Cr	22	20	49	8	14
Cu	8	22	6	54	42
Hg	9	182	257	8	41
Ni	22	135	128	8	61
Pb	16	6	8	20	143
Zn	61	36	22	178	30

Table 5 most polluted observation stations

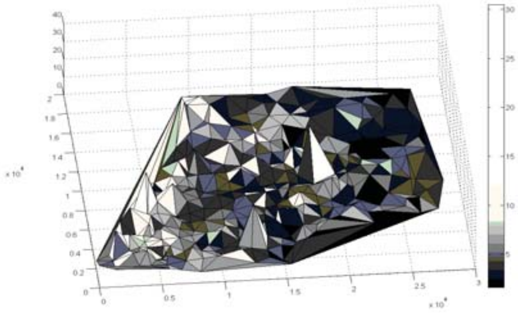

Figure 5a As contamination after 30 steps of backtracking

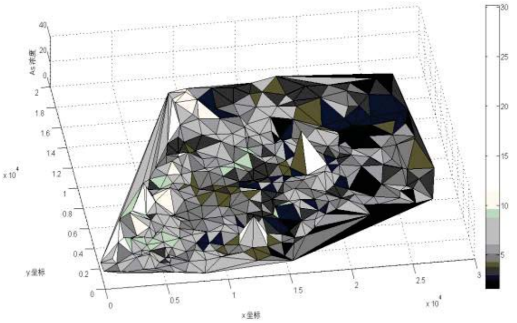

Figure 5b As contamination after 100 steps of backtracking

Heavy metal pollutant			**Pollutant converge point No.**		
As	84	29	178	6	30
Cd	22	223	95	16	6
Cr	22	20	8	49	36
Cu	8	22	42	54	45
Hg	9	182	8	257	232
Ni	22	135	8	128	231
Pb	16	6	31	8	221
Zn	61	22	36	30	8

Table 6   Backtracking results

b) Neural Network backtracking

We have already established the difference equation:

$$x_{i,t+1}=\alpha_i(\omega_{i,1},\omega_{i,2},\ldots,\omega_{i,m})\begin{pmatrix}x_{1,t}-x_{i,t}\\ x_{2,t}-x_{i,t}\\ \vdots\\ x_{m,t}-x_{i,t}\end{pmatrix}+x_{i,t}$$

Set a time t. Define $X_t = X_0$. Our goal is to use the equation above to find $X_{-1}, X_{-2}, X_{-3\ldots}$ Thus the pollutants will converge to its sources.

There are two underlying ideas in the following backtracking algorithm :

Pollutants flow from low concentration areas to high concentration areas.

If the soil concentration in one area is lower than the background statistics, which means the area is not polluted, the pollutants in the area will not flow to adjacent areas.

Then the algorithm is :

Step1 : B1=(b11,b12,b13,……b1m) X0=(x11, x12, x13,……,x1n)

Step2 : Initialization.

$$\omega_{i,j}=\frac{1}{\exp[k_{ij}*(h_j-h_i)+dist(i,j)]}*\delta_{i,j}$$

$$\delta_{i,j}=\begin{cases}1 & if\ i\,is\ connected\ with\ j\\ 0 & else\end{cases}$$

Step3 : while (loop_num < Defined_loop_num)

do(

find the stations with b_{ij} lower than background statistics. Set all their weights to zero.

$$B_{i,t+1}=\alpha_i(\omega_{i,1},\omega_{i,2},\ldots,\omega_{i,m})\begin{pmatrix}B_{it}-B_{1t}\\ B_{it}-B_{2t}\\ \vdots\\ B_{it}-B_{mt}\end{pmatrix}+B_{it}$$

t = t+1;

)

End of algorithm

c) Computer Simulation results

We backtrack the distribution of pollutants according to the above algorithm. The pollutants converge to certain points, which must be the location of pollution source.

During the simulation, we set α = 1 and k = 0.5. Our simulation ended after 100 steps.

Element As is taken as an example. Figure 5a and Figure 5b show the distribution after 30 steps of backtracking and the distribution after 100 steps of backtracking respectively.

We found the convergence points of each heavy metal pollutants, as shown in Table 6.

d) *Comparisons and Conclusions*

Putting Table 5 and Table 6 together, we will find that the listed potential pollution sources falls into the following categories:

- Exists in both tables. Then this station is definitely a pollution source.
- Only exists in Table 6. This shows the inaccuracy of intuitive methods. We only pick five stations with the highest soil concentration, thus it is natural that we left some pollution sources out.
- Only exists in Table 5. It is possible that this is a new pollution source. Another explanation is that the pollutants just happen to converge at this point at this time.

Table 7 shows the definite pollution sources.

Pollutant	**Definite Pollution Sources**
As	84,29,178,30
Cd	22,6
Cr	22,20,8,49
Cu	8,22,42,54
Hg	9,182,8,257
Ni	22,135,8,128
Pb	16,6,8
Zn	61,22,36,30

Table 7 Definite Pollution Sources

III. More Considerations

A. *Model Extensions*

Our model can also be used to predict the flow of pollutants. Just by following the recurrence formula, we can get a prediction of how the pollutants will distribute after a certain amount of time.

B. *Disadvantages*

During the backtracking procedure, we eliminate the factor that the pollution sources are producing pollutants. So in the long term, backtracking result can be not so accurate. That is why we only backtracked 100 steps.

We arbitrarily set $\alpha = 1$ and $k = 0.5$. We only got limited details of the data, so this is just a demonstration of what our model is capable to do. We will be able to find the ideal parameters as soon as we get further detailed data.

IV. Appendix and Supplementary material

The data we use is provided by the CUMCM committee.

It can be found online at the following URL address:

http://www.mcm.edu.cn/html_cn/block/c61dfec317d7a5bd9b2b8efed81c8af3.html

References

[1] Zheng Hailong , Chen Jie , Deng Wenjing , et al . Assessment of soil heavy metals pollution in the chemical industrial areas of Nanjing peri2urban zone . Acta Scientiae Circumstantiae ,2005 ,25 (9) :1182 - 1188

[2] Yan Lin， Soil Heavy metal pollution evaluation and prediction based on soil statistics and GIS. MS degree graduation thesis， 1-74， 2009

[3] Mark de Berg、Otfried Cheong, etc. Algorithms and Applications，Qinghua University Press，2009