Heavy Metal Pollution

Data Analysis and an Approach to Predict the location of the contaminant source

2011/9/12

Summary Sheet

In this paper, we use statistical, mathematical and programming methods to analyse the heavy metals pollution and construct a mathematical model to predict the location of the contaminant source.

In chapter 1, we use Mathematica to plot spatial distributions of the eight kinds of heavy metals and use Nemerow composite index method to analyse the pollution degree in different areas. Then, we use Chi-Square Test to determine main reasons for the heavy metals pollution.

In chapter 2, we construct a mathematical model and use computer programming (Matlab) to estimate the location of the contaminant source.

- In the model, we first assume one of the sites to be the contaminant source, and then simulate how the pollution level of the entire region will be under this assumption.
- Normal distribution is used to approximate the distribution of heavy metals in the region. Linear Regression is used to examine the degree of smoothness of the land path between the contaminant source and the site. Several correction factors, such as the altitude difference between the site and the contaminant source, the direction and strength of wind, are also taken into consideration (using vectors).
- Then, we compare the simulated pollution level of the region under the assumption with the real situation provided in the excel data to test for degree of similarity (using Mean Square Error).
- We iterate the above procedure for other sites (assume them to be the contaminant source) and obtain a degree of similarity for each of the site. By comparing all the assumptions, we will get a site which has the greatest degree of similarity. We conclude that this site is the location of the contaminant source.
- Finally, we further generalise our model to predict the locations of the contaminant sources when there are more than one pollutant.

In the final chapter, we examine strength and weakness of our model and provide possible ways to improve. We also discussed how we can study evolution models of geological environment of the city by collecting additional information.

Contents

1. Introduction

Nowadays, pollution becomes increasingly prevalent in our daily life. In some areas, it has a fairly negative impact on our health. Among all the pollutions, heavy metal is an important part which should never be neglected. To improve the environment we reside requires reducing the pollution; to reduce the pollution demands for having a good knowledge of the pollution degree in the areas we live; to acquire the situation of pollution, we need to collect the pollution concentrations, and analyse the collected data.

However, since it is extremely hard and tedious to obtain every single data from a big city, thus a sample survey is the only easy approach. However, another issue occurs, how to accurately interpret these sample data and how to estimate the contaminant source? In this paper, we present a possible way to interpret these data through a statistical way and construct a mathematical model to investigate in the possible location and number of contaminant sources.

Since we need to regard all the pollution of heavy metal as a whole, we cannot analyse the data of each heavy metal separately. We choose Single Contamination Index method and Nemerow Pollution Index method to calculate the comprehensive pollution, and analyse the generated dataset of the integration.

To locate the contaminant source, we divide our model into two sub-models according to two different ways of the diffusion of heavy model. In each model, we use the normal distribution and take into account the factors, such as terrain of the location and the distance, to calculate the assumed pollution concentrations, provided a certain point is the source. Then we compare the generated concentrations with the real concentrations, and pick up the assumption with the smallest difference related to the real values to be most possible source.

This model offers a pragmatic way to analyse the pollution data, and predict the location of contaminant source.

1.1 Outline of Our Paper

The beginning of the paper will be devoted to analyzing the original data (answering the first two questions of the problem). Then we will present the theoretical framework of our model. The later sections will be devoted to applying our models to predict the location of contaminant source and analysis of some disadvantages and some further improvement of the model.

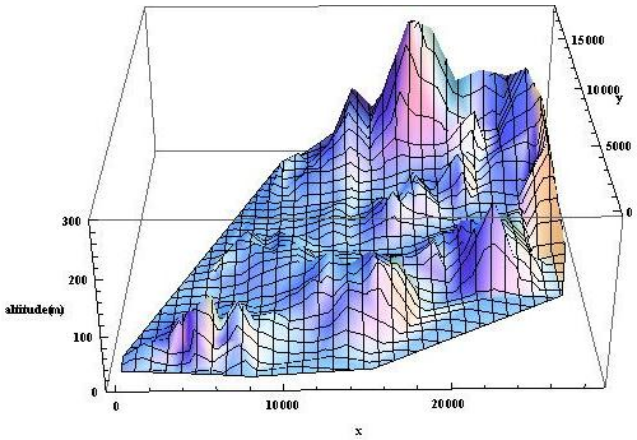
1.2 General Assumptions

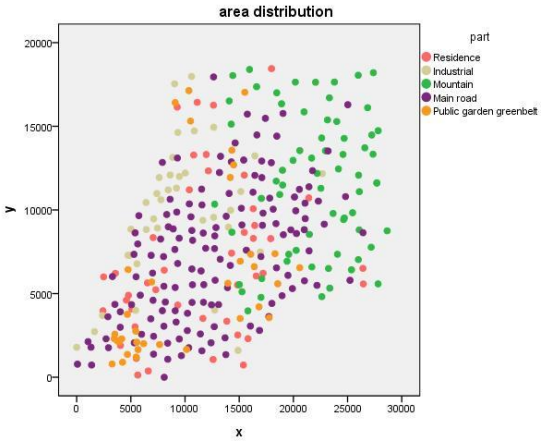
- The density of the sample represents the average density of the square kilometre;
- The land shape can be approximated by the smooth surface connecting all the sample sites;
- Only diffusion through soil and diffusion through air are considered in our model. Other type of propagation may exist, but are not statistically significant;;
- The source of contaminant is located at one of the sample sites.

2. Data Analysis

2.1 Data Analysis of the eight kinds of heavy metals

2.1.1 Space distributions of the eight kinds of heavy metals in the urban area





The graph on the left is a 3D plot of the terrain, where x, y and z axes represent the coordinates of the sample site. Each of the grids represents one kilometre square which is the sample area. The chart on the right denotes the distribution of the five functional areas.

Now, we move forward to plot the space distributions of the eight kinds of heavy metals.

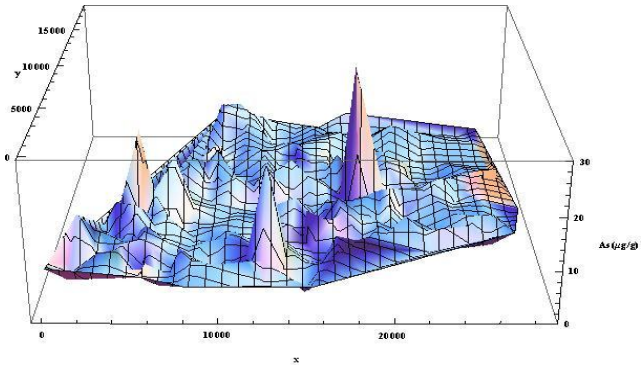


Figure 1 Concentration Distribution for As

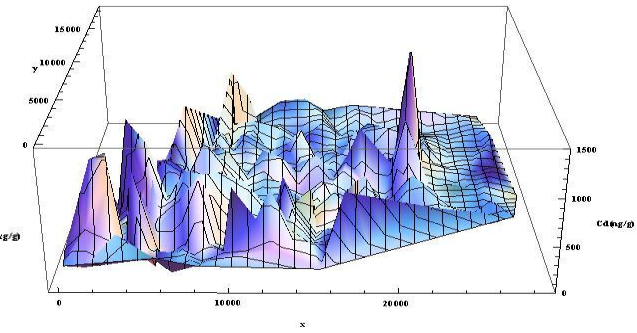


Figure 2 Concentration Distribution for Cd

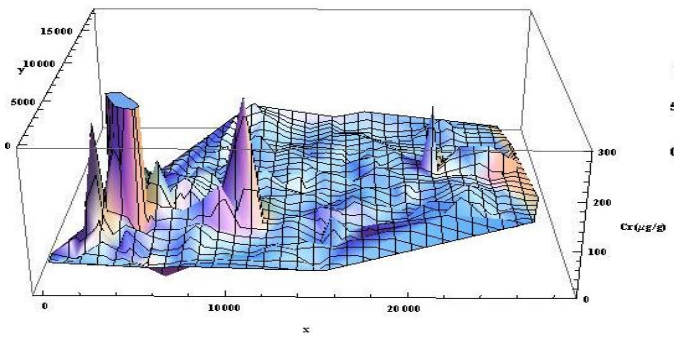


Figure 3 Concentration Distribution for Cr

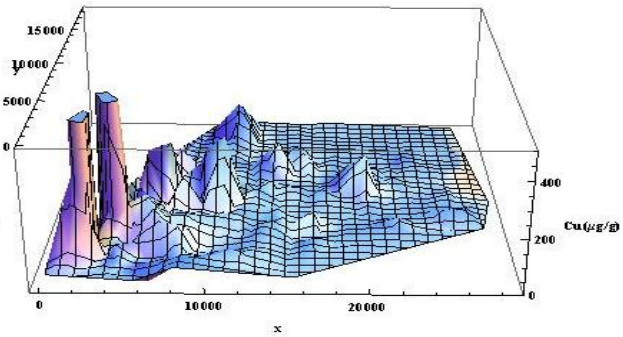


Figure 4 Concentration Distribution for Cu

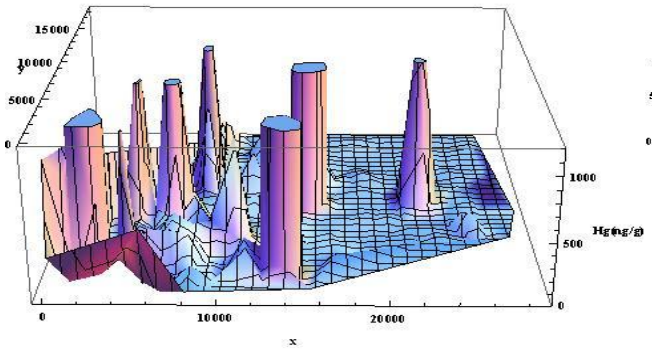


Figure 5 Concentration Distribution for Hg

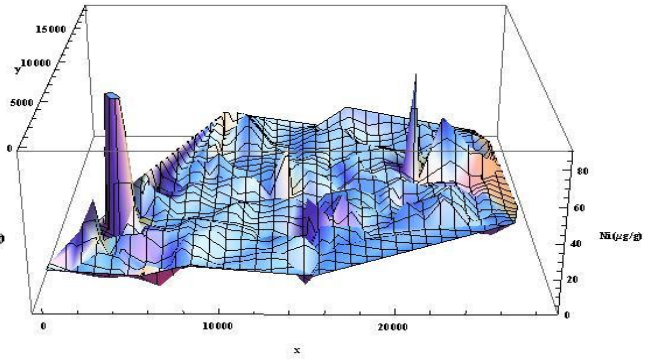


Figure 6 Concentration Distribution for Ni

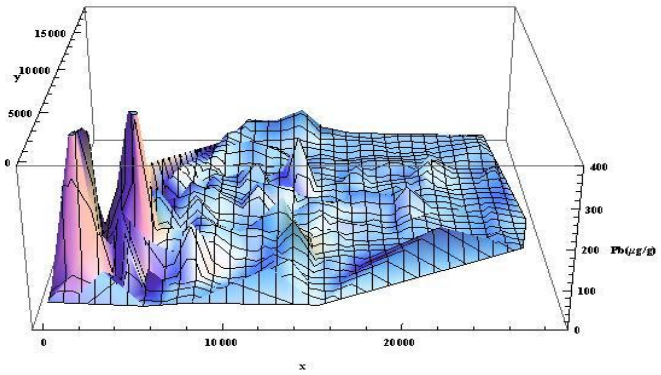


Figure 2 Concentration Distribution for Pb

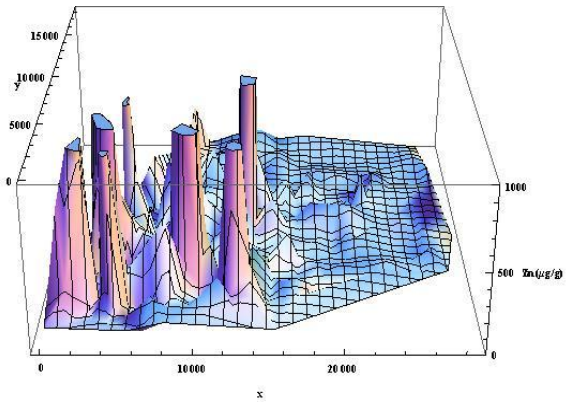


Figure 3 Concentration Distribution for Zn

	As	Cd	Cr	Cu	Hg	Ni	Pb	Zn
Mean	1.576803	2.326125	1.726118	4.167935	8.56318	1.403402	1.991643	2.91598
Variance	0.705736	2.995232	5.099116	152.3263	2167.673	0.653261	2.607471	24.17112
Max	8.369444	12.46	29.70452	191.5515	457.1429	11.58537	15.24129	54.50464

Comparing the graphs shown above with the functional area distribution, we can see that there is rare pollution in mountain area. Comparatively, when it comes to main road area and industrial area, the concentration of heavy metal is higher.

Now, we analyse the distribution of each heavy metal separately. In this section, we will only analyse in a qualitative way. A mathematical approach to analyse the distribution will come later in this paper.

For As, since a large amount of As comes from industry waste, it has a slightly higher concentration along the industry area.

For Cd, according to historical data, a large amount of it comes from the industry. This matches our plot.

For Cr, it mainly comes from industry waste and auto-cars exhaust, hence it is mainly near the industry area and concentrated in this area.

For Cu, it is quite similar to Cr except that it has more concentration near coordinate (0,0) and average of the concentration is comparatively high (rank the second). This indicates that Cu has a quite large amount of contribution to this city's pollution. Moreover, the variance of Cu also ranks the second.

For Hg, the variance is extremely large and it obtains a lot of peak points. This shows that in this city, a large amount of the pollution is due to the pollution of Hg (the average also ranks the first) and it has a tendency that it can still spread out.

For Ni, the graph is a bit flat except for two peak points.

For Pb, it mainly comes from the auto-cars exhaust. However, except for the points near (0,0), the other points' concentration is not that high.

For Zn, the average concentration and the variance both rank the third. By common knowledge, Zn comes from exhausts and it spreads out in the city mainly in the main road area.

2.1.2 Analysis of pollution degrees of the heavy metals in different areas.

To analyse pollution degrees of the heavy metals in different areas, we use two formulae to calculate the comprehensive pollution index (Lian Feng Wang, 2011):

Single Contamination Index method is:

$$P_i = C_i / S_i$$

Where $\mathrm{P_i}$ is the Single Contamination Index of heavy metal pollutant i, $\mathrm{C_i}$ is its real concentration, and $\mathrm{S_i}$ is its regional background value is:

$$P_c = [(P_{max}^2 + P_{ave}^2)/2]^{1/2} \text{ (Jiang Y, 2011)}$$

Where $\mathrm{P_i}$ is the Nemerow Pollution Index of heavy metal pollutant i, $\mathrm{P_{max}}$ is the highest Single Contamination Index in the area, and $\mathrm{P_{ave}}$ is the average Single Contamination Index in the area.

We standardize the heavy metal pollution according to the (Environmental quality standard for soils, 1995-7-13).

Classification standards of soil pollution evaluation

Classification	P	Pollution grades
I	$P \leq 1$	Very low
II	$1 < P \leq 2$	Low
III	$2 < P \leq 3$	Medium
IV	$3 < P \leq 5$	High
V	$5 < P \leq 10$	Very high
VI	$P > 10$	Super high

To analyse the data, we use the descriptive statistics as follow:

$$median = \begin{cases} p_{(n+1)/2}, & n\ is\ odd \\ p_{n/2} + p_{n/2+1}, & n\ is\ even \end{cases}$$

$$mean = \frac{\sum_{k=1}^{n} p_k}{n}$$

$$variance = \frac{\sum_{k=1}^{n} (p_k - mean)^2}{n}$$

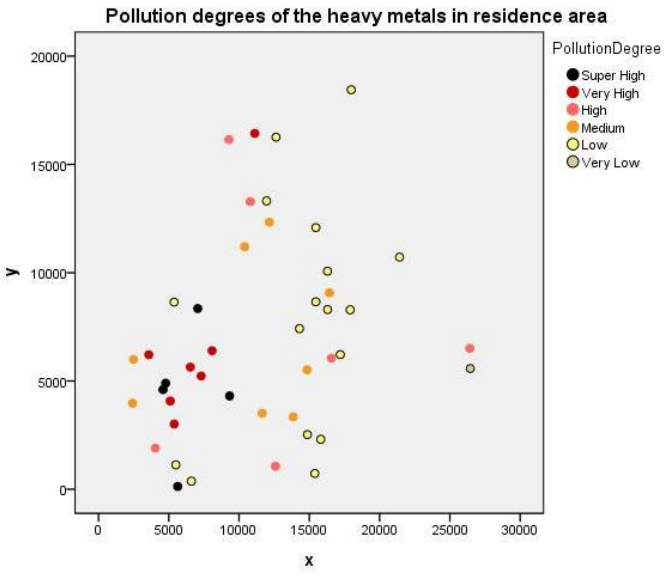
$$std\ deviation = \sqrt{variance}$$

$$range = maximum - minimum$$

$$percentile\ t = p_{n*t\%}$$

- Residence area

In this area, the pollution degree of heavy metal and the descriptive statistics of the pollution index data are shown in the following graph:



Statistics of pollution in residence area

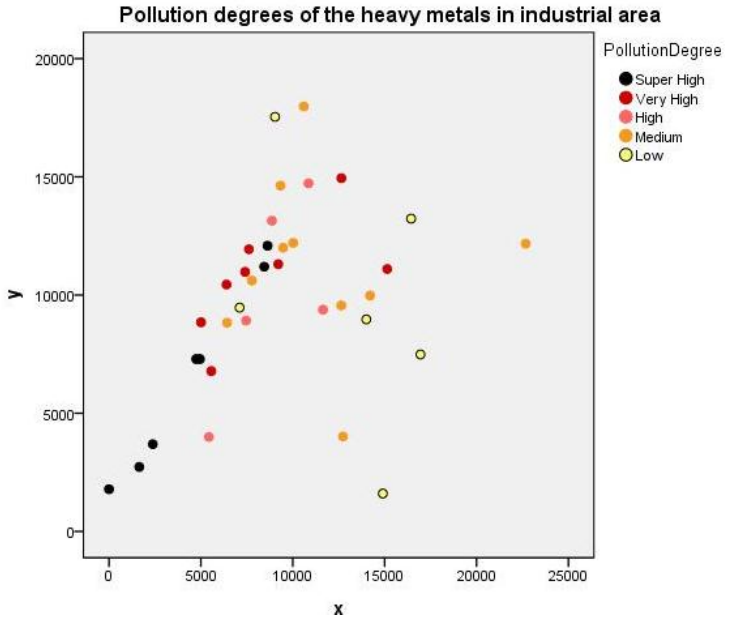
N	Valid	44
	missing	0
Mean		4.55863122039517
Median		2.63766584646765
Std. Deviation		5.424862250896100
Variance		29.429
Range		29.476003426519
Minimum		.721991608639
Maximum		30.197995035158
Percentile	25	1.65296015917159
	50	2.63766584646765
	75	5.16641283188787

And the first 5 highest pollution indexes are: 30.197995035158, 18.078460014980, 13.714554002858, 11.866921320781, 11.396202111091.

The mean is 4.55863122039517 which is inside the interval [3,5], showing that the average pollution in residence area is high, while the median and the third quartile are respectively 2.63766584646765 and 5.16641283188787, which means that though the mean is high, the most part of the area is not that polluted.

- Industrial area

In this area, the pollution degree of heavy metal is as the following graph:



The descriptive statistics of the pollution index data is as follow:

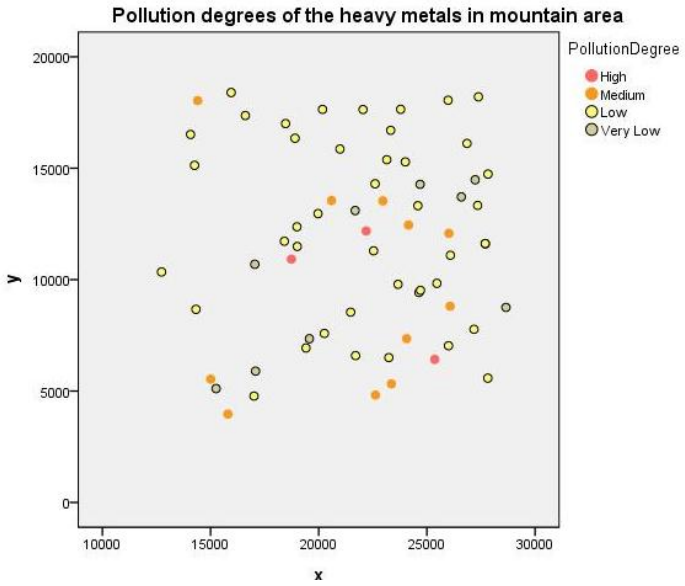
Statistics of pollution in industrial area		
N	Valid	36
	Missing	0
Mean		14.68164899010807
Median		4.07573764881697
Std. Deviation		45.977173611562485
Variance		2113.900
Range		276.883739615803
Minimum		1.547746029193
Maximum		278.431485644995
Percentile	25	2.47816212126414
	50	4.07573764881697
	75	7.41827668852050

And the first 5 highest pollution indexes are: 278.431485644995, 36.898348163311, 31.277250875575, 22.724400402542, 19.482765297168.

The mean and the median of the dataset are both above 4, which illustrates the pollution in this area is really high. And the maximum of the dataset is extremely high, from which we can say the pollution in the industrial area is really heavy.

- Mountain area

In this area, the pollution degree of heavy metal is as the following graph:



The descriptive statistics of the pollution index data is as follow:

Statistics of pollution in mountain area		
N	Valid	66
	missing	0
Mean		1.65285948202346
Median		1.50390827338595
Std. Deviation		.790888016727264
Variance		.626
Range		4.002587041842

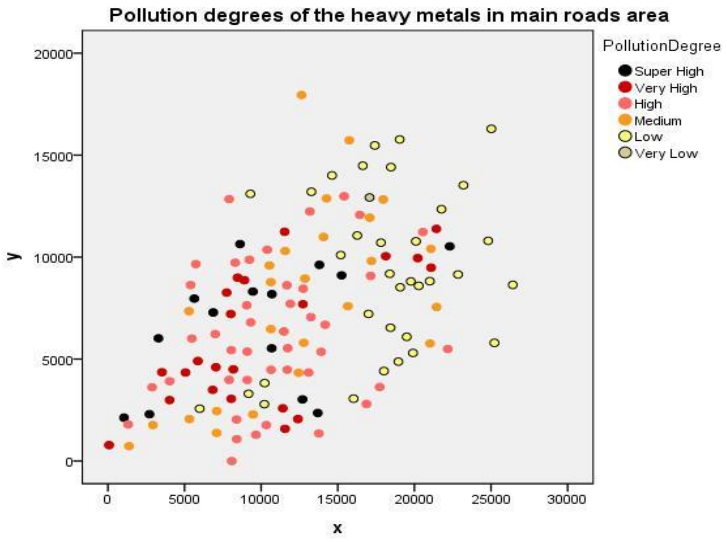
Minimum		.680564271380
Maximum		4.683151313222
Percentile	25	1.13193284362476
	50	1.50390827338595
	75	1.90009690256432

And the first 5 highest pollution indexes are: 4.683151313222, 4.430537788796, 3.975748198589, 2.935557000654, 2.924402194566

From the mean, the median and even the first 5 largest indexes of the dataset of this area show that the average pollution degree in mountain area is very low. The small std. deviation shows that the pollution degree in this area is very uniform.

- Main roads area

In this area, the pollution degree of heavy metal is as the following graph:



The descriptive statistics of the pollution index data is as follow:

Statistics of pollution in main roads area		
N	Valid	138
	missing	0

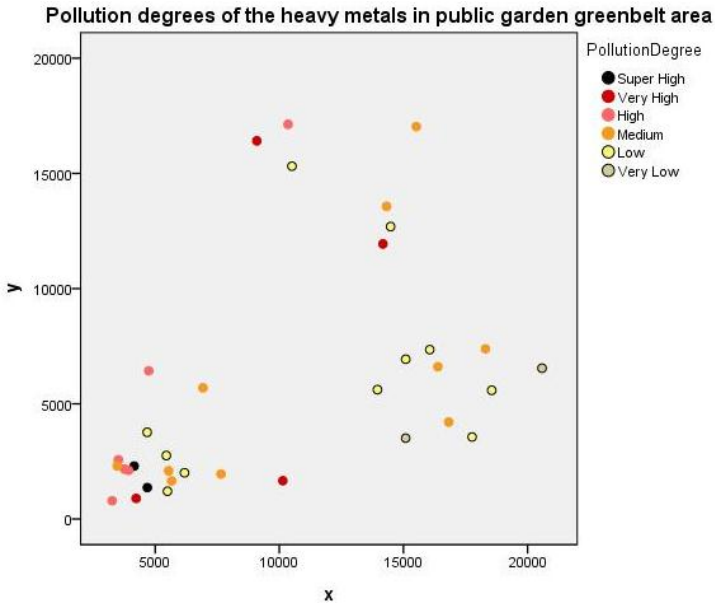
Mean		11.90536449813606
Median		3.40167757947323
Std. Deviation		44.545325741672590
Variance		1984.286
Range		325.445185520462
Minimum		.849988961415
Maximum		326.295174481878
Percentile	25	1.96826978726696
	50	3.40167757947323
	75	5.19791475908830

And the first 5 highest pollution indexes are: 326.295174481878, 303.591749248731, 281.167826088115, 75.061946429160, 39.076789250677.

The mean of the dataset is above 10, and even the median is above 4, which shows that the pollution of this area is high. But the large std. deviation indicates that the pollution in the main roads area is not even, it may vary largely across this area.

- Public garden greenbelt area

In this area, the pollution degree of heavy metal is as the following graph:



The descriptive statistics of the pollution index data is as follow:

Statistics of pollution in public garden greenbelt area		
N	Valid	35
	missing	0
Mean		3.74416744756902
Median		2.08087939039525
Std. Deviation		4.894780640629676
Variance		23.959
Range		26.634769702893
Minimum		.820072898169
Maximum		27.454842601062
Percentile	25	1.54062408546135
	50	2.08087939039525
	75	3.36068390258410

And the first 5 highest pollution indexes are: 27.454842601062, 14.932896440396, 8.235493594804, 6.829487504058, 5.463305948626.

From the median and the third quartile of the dataset we can tell that the pollution of most part of this area is not high. But the relatively large mean and the large std. deviation illustrate that some part, maybe not very large, of the area is still heavily polluted.

2.2 Examining the main reasons for heavy metals pollution by Chi-Square Test

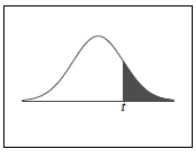
Pollution level of different types of functional Areas

Frequency

		Pollution Level						
		Super High	Very High	High	Medium	Low	Very Low	Total
Type of Functional Area	Residential Area	5	7	6	8	17	1	44
	Industrial Area	7	8	5	10	6	0	36
	Mountain Area	0	0	3	11	43	9	66
	Main Roads Area	14	22	42	25	34	1	138
	Public Garden Greenbelt Area	2	4	6	10	11	2	35
Total		28	41	62	64	111	13	319

From the table, we can see that the pollution level in residential area, industrial area and main road area are clearly higher than the other two. Therefore, we make a hypothesis that the heavy metal pollution are mainly caused by industrial wastes, residential wastes and car emission. To test whether these hypothesis are correct, we conduct a categorical data analysis and use the chi-square test with confidence level α =0.1 to test whether the type of region (e.g. industrial area) will affect the pollution level.

t-Distribution Table



The shaded area is equal to α for $t = t_\alpha$.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925

From the t-Distribution Table, we find out that $t_{\frac{\alpha}{2}} = 6.314$.

There are r = 2 rows and c = 2 columns in the table, the "theoretical frequency" for a cell, given the hypothesis of independence, is

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N},$$

and fitting the model of "independence" reduces the number of degrees of freedom by $p = r + c - 1$. The value of the test-statistic is

$$\mathrm{X}^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

The number of degrees of freedom is equal to the number of cells rc, minus the reduction in degrees of freedom, p, which reduces to $(r - 1)(c - 1)$.

- Hypothesis 1: Residential Waste (Such as used Battery) causes heavy metal pollution.

We first construct a 2×2 frequency table:

Contingency Table

Frequency

		Highly Polluted?		
		P>5	P<5	Total
Residential Area?	Yes	27	53	80
	No	42	197	239
Total		69	250	319

Chi-Square Test

	Value	Df	Asymp. Sig. (2-sided)	Exact Sig.(2-sided)	Exact Sig.(1-sided)
Pearson Chi-Square	9.253a	1	.002		
Continuity Correctionb	8.323	1	.004		
Likelihood Ratio	8.654	1	.003		
Fisher's Exact Test				.004	.003
McNemar Test				.305c	
N of Valid Cases	319				

a. 0 cells(.0%) have expected count less than 5 ◦ The minimum expected count is 17.30.

Since Chi-Square (with d.f=1) = 9.253>6.314, the p-value< α =0.1. Therefore, we can conclude that the row variable and the column variable are dependent, which means residential activity accompanies higher heavy metal concentration. So the hypothesis is true.

- Hypothesis 2: Industrial waste causes heavy metal pollution.

We construct a 2×2 frequency table:

Contingency Table

Frequency

		Highly polluted?		
		Yes	No	Total
Industrial Area?	Yes	15	21	36
	No	54	229	283
Total		69	250	319

Chi-Square Test

	Value	Df	Asymp. Sig. (2-sided)	Exact Sig.(2-sided)	Exact Sig.(1-sided)
Pearson Chi-Square	9.610a	1	.002		
Continuity Correctionb	8.324	1	.004		
Likelihood Ratio	8.385	1	.004		
Fisher's Exact Test				.004	.003
McNemar Test				.000c	
N of Valid Cases	319				

a. 0 cells(.0%) have expected count less than 5。 The minimum expected count is 7.79.

Since Chi-Square (with d.f=1) = 9.610>6.314, the p-value< α =0.1. Therefore, we can conclude that the row variable and the column variable are not independent, which means industrial activities accompanies higher heavy metal concentration. So the hypothesis is true.

- Hypothesis 3: Car emission on the main road causes heavy metal pollution.

We construct a 2×2 frequency table:

Contingency Table

Frequency

		Highly Polluted?		
		Yes	No	Total
Main Road Area?	Yes	36	102	138
	No	33	148	181
Total		69	250	319

Chi-Square Test

	Value	Df	Asymp. Sig. (2-sided)	Exact Sig.(2-sided)	Exact Sig.(1-sided)
Pearson Chi-Square	2.850a	1	.091		
Continuity Correctionb	2.405	1	.121		
Likelihood Ratio	2.829	1	.093		
Fisher's Exact Test				.101	.061
McNemar Test				.000c	
N of Valid Cases	319				

a. 0 cells(.0%) have expected count less than 5。 The minimum expected count is 29.85.

Since Chi-Square (with d.f=1) = 2.850<6.314, the p-value> α =0.1. Therefore, there is no enough evidence for us to say that the row variable and the column variable are dependent, which means that being in the main road area does not necessarily mean the pollution level will be higher. So the hypothesis is false.

We may also test for the other two types of regions: mountain area and public garden and greenbelt area. These two tests are omitted in our paper for simplicity.

In conclusion, the main causes of heavy metal pollution are residential wastes and industrial wastes.

3. Mathematical Model to predict the location of contaminant source

3.1 The big picture

3.1.1 Background information of the propagation characteristics of heavy metals

According to Wikipedia, heavy metals propagate mainly through soil system and air. Hence, we construct two sub-models to simulate the two ways of propagation. These two models will be illustrated in later chapters of this paper.

3.1.2 Brief introduction to our model

In order to predict the location of the pollutant, we first assume one of the sampling sites (any site is ok.) is the pollutant. Then by simulating the propagation of heavy metals (using the two sub-models: soil propagation and air propagation), we calculate a relative level of pollution for each of the other sites (besides the site that is assumed to be the pollutant). Hence, we compare the simulated level of pollution with the real data, and use Mean Square Error to describe the difference between the estimated situation and the real situation.

After that, we iterate the above procedure by assuming each of the other sites to be the pollutant, and obtain a Mean Square Error (the difference between the real situation and the simulated situation) for each of them.

Thus, by comparing the mean square errors of all the sites, we will find the site with the least square error, which means that the simulated situation of this site is the least different from the real situation.

Therefore, we conclude that this site is our predicted location of pollutant.

3.2 Our algorithm

Since the simulation is done by Matlab, we also give a pseudo-code here to illustrate how the program runs.

```
Beginning of the program: import data, import necessary libraries.
   Loop: for i = 1 to 319, assume site i to be the pollutant.
   {
   Inner loop: for j = 1 to 318
         /*site j's are all the other sites except the assumed pollutant site i  */
      {
         Pollution by soil = function(D,L,A);
```

```
        Pollution by air = function(D,W)
        Relative pollution level of site j = pollution by soil + pollution by air;
          /* D = distance from the pollutant;
               L = landform between site j and the pollutant;
               A = difference in altitude between site j and the pollutant;
               W = Wind Strength and direction */

      }
  Obtain the relative pollution level from the original data set;

  Calculate the Mean Square Error between the simulated pollution levels of the sample sites
    with the original ones;

  /* the Mean Square Error of the two data set shows the degree of difference between
          simulated situation and the real situation of site I */

  Return (Mean Square Error of site i);

 }

Find the site with the least mean square error; denote it by Site (LS)

Conclude that Site (LS) is the predicted pollutant.
```

3.3 Sub-Model One: Diffusion through soil

To develop the model of propagation of heavy metal pollution, we use the normal distribution to simulate the pollution distribution. As the real world operates, normal distribution is the most general object distribution law. Besides, no matter what distribution the diffusion of the pollution follows, according to the central limit theory, it converges to normal distribution when the amount of the pollution is very large.

Therefore, we assume that the final distribution of the pollution is just like a normal distribution density curve (the same method applies to sub-model two), and we use the normal distribution density function to denote the pollution at a certain point where the terrain among A and B is a straight line:

$$f_A(d_{A,B}) = \frac{1}{\sqrt{2\pi}\sigma_B} e^{-\frac{{d_{A,B}}^2}{2{\sigma_B}^2}}$$

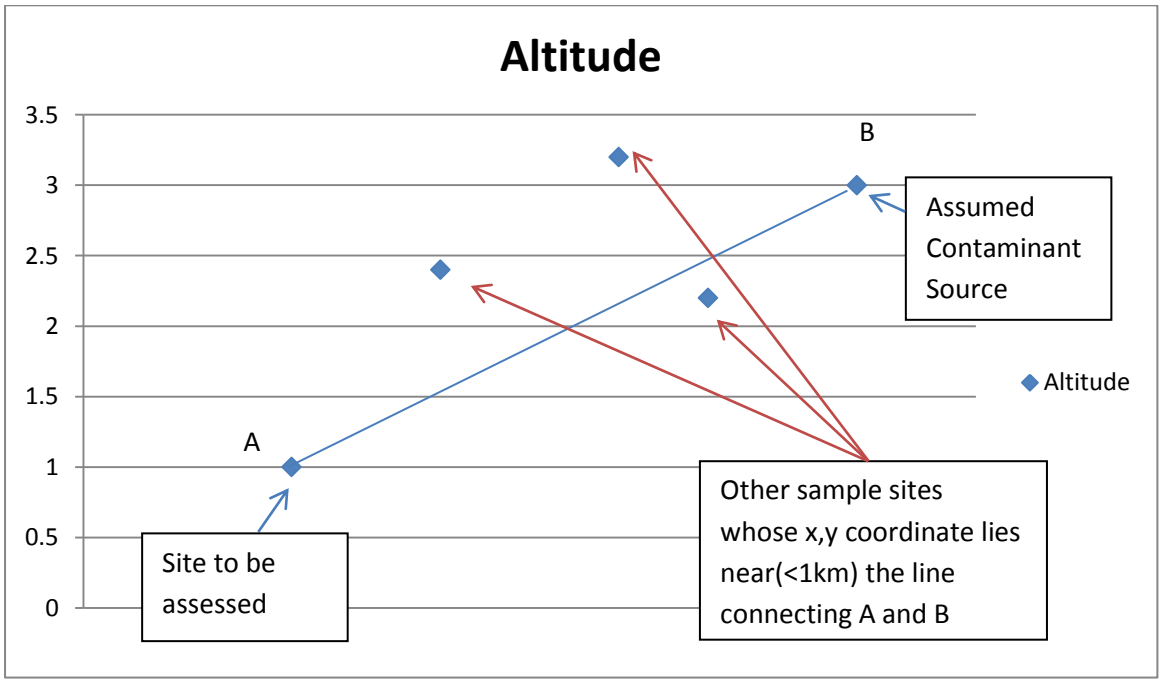
Here we assume A is the contaminant source, and $f_A(d_{A,B})$ denotes the pollution concentration at B given the distance between A and B, i.e. $d_{A,B}$, and $f_A(d_{A,B})$ is a normal distribution density function given the standard deviation of the distribution is σ_B, which indicates that 95% of the pollutions are in the distance of $2\sigma_B$ away from A.

But to obtain the distribution that is closer to reality, we need to take into account the terrain among A and B. So we use a line across A and B to simulate the terrain. As it appears, it's not so reasonable to just ignore the fluctuation of the road connecting A and B. Thus, we use the correlation coefficient r to modify the assumption.

$$F_A(d_{A,B}) = r \times f_A(d_{A,B})$$

The correlation coefficient is calculated using the site point, the contaminant source point and all other points whose x,y coordinate lies near the line connecting the two points:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$



Besides, the above modification is just about the terrain assumption. We still need to consider the gap in the altitude of two points and the slope of the line we mentioned above. Here, we use slope to modify the standard deviation of normal distribution, so as to make the distribution more realistic.

$$k = \frac{h_A - h_B}{d_{A,B}}$$

$$\theta = arctan(k)$$

Here k θ denotes the oblique angle. Then we can modify σ_B byθ.

$$\sigma_B = \begin{cases} \sigma \cos\theta, & \theta \geq 0 \\ \dfrac{\sigma}{\cos\theta}, & \theta < 0 \end{cases}$$

Here σ is the standard deviation of the normal distribution where the land is plain (the height of the area is a constant).

Since we just compare the final difference between the generated pollution and real pollution to predict the location of the contaminant source, we can just use an assumed function that has the same properties of the real distribution function to calculate.

As far as it's concerned, when the terrain among A and B is a straight line and A is above B, the steeper the slope is, the wider the pollutants are able to spread. Accordingly, the larger θ is, the larger 95%-pollution-distributing area is, thus the greater σ_B will be. And when$\theta = 0$, $\sigma_B = \sigma$ according to the definition.

Similarly, when A is lower than B, the steeper the slope is, the harder pollutants are able to spread from A to B. Accordingly, the larger θ is, the smaller 95%-pollution-distributing area is, thus the less σ_B will be. And when $\theta = 0$, $\sigma_B = \sigma$ according to the definition, while when θ goes to infinity, the pollution will not spread upward, thus $\sigma_B = 0$.

The σ_B function we define obeys all these properties mentioned above. So the σ_B function is fairly reasonable and well-defined.

Therefore, we obtain the final distribution function of the heavy metal pollution propagated via soil.

$$F_A(d_{A,B}) = r \times \frac{1}{\sqrt{2\pi}\sigma_B} e^{-\frac{{d_{A,B}}^2}{2{\sigma_B}^2}}$$

3.4 Sub-Model Two: Diffusion through air

In this sub-model, we still assume that the distribution of pollution level can be approximated by a normal distribution with centre at the contaminant source. However, we need to take into consideration some other influencing factors: wind strength and wind direction, which may change the shape of the distribution.

First, assume that the direction of wind is completely random, then the distributions of the concentrations of heavy metals will follow normal distribution which centres at the original point (contaminant source), thus $x \sim N(0, \delta^2)$.

Then based on analysis of the 8 spatial distribution graphs in the second chapter, we find out that the distribution of heavy metals is severely skewed towards the left-bottom corner of the graph. Thus, it is reasonable for us to assume that the direction of wind follows the direction of the vector <-1, -1> and since the strength of the wind is completely random (especially over a long period), the wind will shift the normal distribution curve. Hence we define that the mean of the final normal distribution $\mu \sim U(0, L_{average\_strength})$, where $L_{average\_strength}$ is the distance that the wind with average strength can blow the substances contain heavy metals to in the direction of the wind. Within this distance, since the wind is completely random, it is reasonable to assume that it follows the uniform distribution. Thus we can calculate the distribution along the direction <-1,-1> by using stochastic analysis. We already know that the density distribution of a normal curve is:

$$f(x) = \frac{1}{\sqrt{2\pi\delta^2}} e^{\frac{-(x-\mu)^2}{2\delta^2}}$$

Since $\mu \sim U(0, L_{average\_strength})$, we can change the density function of normal distribution to:

$$f(x|\mu) = \frac{1}{\sqrt{2\pi\delta^2}} e^{\frac{-(x-\mu)^2}{2\delta^2}}$$

Since:

$$f(\mu) = \frac{1}{L_{average\_strength}}$$

Then:

$$f(x) = \int_0^{L_{average\_strengt\hbar}} f(x,\mu)d\mu = \int_0^{L_{average\_strengt\hbar}} f(x|\mu)f(\mu)\,d\mu$$
$$= \frac{\sqrt{\pi}\delta(Erf[\frac{L_{average\_strengt\hbar} - x}{\sqrt{2}\delta}] + Erf[\frac{x}{\sqrt{2}\delta}])}{2\sqrt{\pi\delta^2}\ L_{a\quad\ age\_strengt\hbar}}$$

Where $Erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt.$

Then we need to know how it will change the normal distribution's mean and variance, we can proceed this way:

$$E(X|\mu) = E(N(\mu,\delta^2)|\mu)$$

$$E(N(\mu=\mu_0,\delta^2)|\mu=\mu_0) = \mu_0$$

Hence

$$E(X) = E\big(E(X|\mu)\big) = E\big(E(N(\mu,\delta^2)|\mu)\big) = E(\mu) = \frac{L_{average\_strengt\hbar}}{2}$$

Then

$$Var(X) = E\big(Var(X|\mu)\big) + Var\big(E(X|\mu)\big)$$

$$Var(X|\mu=\mu_0) = Var(N(\mu=\mu_0,\delta^2)|\mu=\mu_0) = \delta^2$$

Hence

$$Var(X) = E\big(Var(X|\mu)\big) + Var\big(E(X|\mu)\big) = E(\delta^2) + Var(\mu) = \delta^2 + \frac{L_{average\_strengt\hbar}}{12}$$

The above analysis is based on two points which lies exactly in the same direction as the wind vector, namely **e**. Then we will generalize our case for any two points. Suppose that the contaminant point is A(xa, ya), another point is B(xb, yb), then there are two vectors, namely:

$$\boldsymbol{e} = < -1, -1 >$$

$$\boldsymbol{AB} = < x_b - x_a, y_b - y_a >$$

Then we can calculate the angle between these two vectors:

$$cos\alpha = \frac{\boldsymbol{AB}\cdot\boldsymbol{e}}{|\boldsymbol{AB}||\boldsymbol{e}|}$$

Knowing the angle between these two vectors, we can obtain a new random variable which suits the general cases by changing the uniform distribution $\mu \sim U(0, \cos(\theta)\, L_{average\_strength})$ and we plug it into the previous equations, we can obtain the following:

$$f(x) = \frac{\sqrt{\pi}\delta(Erf[\frac{L_{average\_strength} - x}{\sqrt{2}\delta}] + Erf[\frac{x}{\sqrt{2}\delta}])}{2\ L_{average\_strength} cos\alpha \sqrt{\pi\delta^2}}$$

$$E(X) = \frac{L_{average\_strength}}{2} cos\alpha$$

$$Var(X) = E\big(Var(X|\mu)\big) + Var\big(E(X|\mu)\big) = E(\delta^2) + Var(\mu) = \delta^2 + \frac{L_{average\_strength}}{12} cos\alpha$$

By using this method, we can calculate the concentration of the ending point by knowing the contaminant point given that it is only diffused through air.

3.5 Combination of the sub-models

From the above two sub-model, we obtain the distribution of the heavy metal pollution propagated via soil and the distribution of the pollution diffused via air. Now we combine these two distributions to get the final distribution of the pollution.

To calculate the final distribution, we assume a fixed proportion τ of the pollution, which is propagated via soil. Accordingly, $1 - \tau$ of the pollution is diffused by air.

Assuming A is the source of pollution, we want to get the concentration of pollution at B. Firstly, we need to obtain the ratio of pollution concentration at B to concentration at A diffused via soil and air respectively by the two sub-model.

We use C_{A1}, C_{A2}, C_{B1}, C_{B2} to denote the pollution concentration diffused via soil and air at A and B respectively.

According to the sub-models,

$$\frac{C_{B1}}{C_{A1}} = \frac{r \times \frac{1}{\sqrt{2\pi}\sigma_B} e^{-\frac{{d_{A,B}}^2}{2{\sigma_B}^2}}}{\frac{1}{\sqrt{2\pi}\sigma}} = r \times \frac{\sigma}{\sigma_B} e^{-\frac{{d_{A,B}}^2}{2{\sigma_B}^2}}$$

$$\frac{C_{B2}}{C_{A2}} = \frac{\frac{\sqrt{\pi}\delta(Erf[\frac{L_{average\_strength} - d_{A,B}}{\sqrt{2}\delta}] + Erf[\frac{x}{\sqrt{2}\delta}])}{2\ L_{average\_strength} cos\alpha\sqrt{\pi\delta^2}}}{\frac{\sqrt{\pi}\delta Erf[\frac{L_{average\_strength}}{\sqrt{2}\delta}]}{2\ L_{average\_strength}\sqrt{\pi\delta^2}}}$$

$$= \frac{Erf[\frac{L_{average\_strength} - d_{A,B}}{\sqrt{2}\delta}] + Erf[\frac{x}{\sqrt{2}\delta}]}{cos\alpha Erf[\frac{L_{average\_strength}}{\sqrt{2}\delta}]}$$

Then, the final ratio of the total pollution concentration at B given A is the contaminant source is

$$\rho = \frac{C_B^A}{C_A} = \tau\frac{C_{B1}}{C_{A1}} + (1-\tau)\frac{C_{B2}}{C_{A2}}$$

Hence, the estimated pollution concentration at B is

$$C_B^A = \rho \times C_A$$

Where C_A is the real pollution concentration we collect at A.

By the method above, we can get one dataset of the estimated pollution concentrations when assuming a certain point is the contaminant source. For instance, we assume point i is contaminant source, we get C_k^i which denotes the estimated pollution concentration at point k given i is the source. So we can get 319 such C_k^is.

To compare the estimated value with the real value, we use the following formula:

$$MSE_i = \frac{\sum_{k=1}^{319}(C_k^i - C_k)^2}{319}$$

Where C_k is the real pollution concentration at point k.

Applying the same method to every point (changing i we mentioned in the last paragraph), we can get 319 datasets of the estimated values, and we also get 319 MSE_is.

If there's only one contaminant source, we shall find out $\min_i\{MSE_i\}$ when i=l, then point l is the contaminant source.

If there're two contaminant sources, we use $\mathrm{C_k^i}\ (\mathrm{i,k} = 1,2,\ldots,319)$ generated above to calculate the sum pollution concentration caused by two contaminant sources. For example, we assume point 1,2 are two contaminant sources, then we get the estimated pollution concentration at point k:

$$C_k^{1,2} = C_k^1 + C_k^2$$

Similarly, we get

$$MSE_{i,j} = \frac{\sum_{k=1}^{319}(C_k^{i,j} - C_k)^2}{319}$$

And find out $\mathrm{min_{i,j}\{MSE_{i,j}\}}$ when i=m, j=n, and point m and n are the most possible two contaminant sources.

By the same method we could get the most possible k contaminant sources (k=1,2,3,…). But we cannot increase k to 319, which will be too large for programming. So we just increase k until the minimum MSE of k-1 assumed sources is smaller than the minimum MSE of k assumed sources. Thus, we get the most possible number of the contaminant sources and the location of them.

By combining the two sub-models, we can predict the location of the contaminant sources.

3.6 Programming for the multiple contaminant sources model

At first look, it seems unrealistic to use programming to carry out our model when the number of contaminant sources increases to greater than 5:

The number of iterations of our program (when the number of contaminant sources is 5) is $\binom{319}{5} = 2.67 \times 10^{10}$. This seems too large for computer to run in realistic time duration.

To solve this problem, we implemented the algorithm of dynamic programming (Farmer, 2007): store all the $\binom{319}{n}$ cases for n contaminants scenario, and use the stored data to calculate the data for the $\binom{319}{n+1}$ cases of the n+1 contaminants scenario. Hence, we significantly improved the programming efficiency and make it possible to carry out the calculation.

4. Results & Interpreting the Results

4.1 Results

According to 'Evaluation Soil Contamination' published by U.S. Department of Interior (Evaluation Soil Contamination, 1990), it is reasonable for us to assume that τ_{As}=0.38, τ_{Cd}=0.18, τ_{Cr}=0.64, τ_{Cu}=0.42, τ_{Hg}=0.21, τ_{Ni}=0.17, τ_{Pb}=0.08, τ_{Zn}=0.35; and we assume that σ = 1000, δ = 5000;

One contaminant source Location	minimum mean square error
(2383, 3617)	4694.064
(15248,9106)	6027.641
(2708,2295)	6538.082
(13694, 2357)	8469.937

Two contaminant sources	Minimum mean square error
(2383, 3617) (13694, 2357)	2505.354
(2708,2295) (15248,9106)	3054.389
(15248,9106) (13694, 2357)	4326.549
(2383, 3617) (2708,2295)	6343.890

Three contaminant sources	Minimum mean square error
(3299, 6018) (13694, 2357) (15248,9106)	1942.031
(2383, 3617) (15248,9106) (13694, 2357)	2589.376
(2708,2295) (13694, 2357) (13694, 2357)	3892.432
(2383, 3617) (13694, 2357) (2708,2295)	6243.146

Four contaminant sources	Minimum mean square error
(3299, 6018) (2383, 3617) (13694, 2357) (15248,9106)	1892.324
(2383, 3617) (15248,9106) (13694, 2357) (13797,9621)	2498.879
(15248,9106) (13694, 2357) (9319,6799) (2383, 3617)	2754.345
(13694, 2357) (13797,9621) (2383, 3617) (3299, 6018)	3564.231

Five contaminant sources	Minimum mean square error
(2383, 3617) (15248,9106) (13694, 2357) (2708,2295) (3299, 6018)	1094.358
(2383, 3617) (15248,9106) (13694, 2357) (2708,2295) (13797,9621)	1732.398

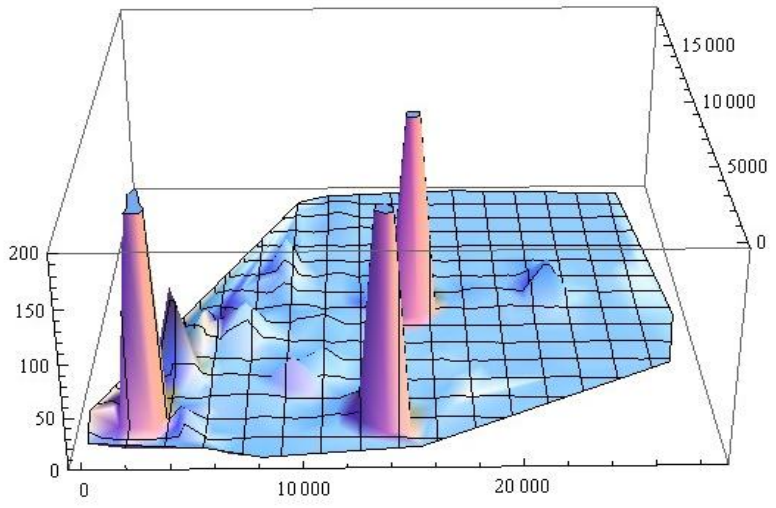
(2383, 3617) (15248,9106) (13694, 2357) (2708,2295) (9319,6799)	1921.301
(2383, 3617) (15248,9106) (13694, 2357) (3299, 6018) (13797,9621)	2487.628

Six contaminant sources	Minimum mean square error
(2383, 3617) (15248,9106) (13694, 2357) (2708,2295) (3299, 6018) (13797,9621)	1108.231
(2383, 3617) (15248,9106) (13694, 2357) (2708,2295) (3299, 6018) (9319,6799)	1324.387
(2383, 3617) (15248,9106) (13694, 2357) (2708,2295) (3299, 6018) (8629,12086)	1431.216
(2383, 3617) (15248,9106) (13694, 2357) (2708,2295) (13797,9621) (9319,6799)	1442.418

4.2 Interpreting the Result

From the mean square error, we can see that when there are five contaminant sources, it can achieve the minimum mean square error and their locations are (2383, 3617) (15248, 9106) (13694, 2357) (2708, 2295) (3299, 6018). This is reasonable because all of these five points frequently shown in the above data, which indicates they have a quite large probability to be the contaminant points.

However, since the mean square error is still very large, we have to suspect our five contaminant points. Because we have made an assumption which states that the only possible contaminant points are those points shown in the data file, but in reality, it is only partly correct. If we revise it a bit and try the middle point between (2383, 3617) and (2708, 2295) and also substitute the middle point of their heavy metals' concentrations, we can obtain a reasonably good mean square error which is 190.134, 5 times smaller than the previous mean square error when we try 3 contaminant points. These three contaminant points are (15248, 9106) (13694, 2357) (2545.5, 2956). What is more, we can take a look at the following graph whose z axe is the Nemerow Pollution Index number, which gives us the contaminant level of a particular point. The most important point is our three points fit the three peak points incredible well which partially prove that our model is correct and can predict the contaminant locations quite accurate!



5. Assessment of our Model

Our model takes into consideration a lot of factors which can affect our determination of contaminant points, such as considering the slope and distance's effect in diffusion through soil model, the random behavior of the wind in diffusion-through-air model, different mobility of different heavy metals in propagation and etc.

Just as every coin has two sides, because of taking into consideration a lot of various factors, there are several main constraints in our models with respect to these variables. Firstly, some of the factors are difficult to obtain, for example, the factor that characterizes the mobility of each heavy metal, propagating either in soil or in air. This factor is hard to determine since each metal has its own characteristic in propagating and it is also deterministic by matters around that particular particle. Due to the time constraint, we do not have enough time to model each of the metal for each scenario. Hence, we directly use some historical data, but may lead to some bias when applying to our case study. Secondly, the algorithm needs to be improved. Some of the codes need a long time to run, due to a large number of for-loop. In future, we need to use more dynamic programming to come up with better algorithm so as to save our precious time.

6. Additional information & Evolution Model

To study evolution models of geological environment of the city better, we need to take into consideration the time factor. Hence we need to collect the concentrations of heavy metals at a certain interval at the same points as we choose previously. As time goes by, we are also able to collect the data of weather of the city, so as to analyse the impact on pollution distribution by rainfalls, the wind direction and the information about the rivers across the city.

Given the information mentioned above, we could modify our model by correcting the pollution extent and wind direction in diffusion by air sub-model, adding the factor of the rivers and rainfalls. Furthermore, we can develop a new model to study evolution models of geological environment of the city better on the basis of the current model. In the new model, we could analyse the pollution distribution of a certain point as time goes on, instead of the pollution distribution in a large area at a certain time. And this new model will help us predict the trend of pollution distribution as time goes on. What's more, we could predict what the environment we reside will be in future, and the significance of environment protection will become a highlight of our life. Therefore, we are able to avoid the trend of environment deterioration, and make our environment better and better.

References

(1990). *Evaluation Soil Contamination.* U.S.Department of Interior.

(1995-7-13). *Environmental quality standard for soils.* National Bureau of Environmental Protection.

Farmer, J. E. (2007, 5 8). *Introduction to Dynamic Programming*. Retrieved 9 12, 2011, from http://20bits.com/articles/introduction-to-dynamic-programming/

Jiang Y, W. X. (2011, 3 1). Contamination, source identification, and risk assessment of polycyclic aromatic hydrocarbons in agricultural soil of Shanghai, China. *SpringerLink*.

Lian Feng Wang, Y. X. (2011, July). Frontiers of Green Building, Materials and Civil Engineering. *Applied Mechanics and Materials*.