Location of atmospheric pollutant source based on Weighted Location Set Cover Model and Data Mining

Weikang Wang[1, a]

[1]No.689, Huadian Road, School of Computer Science and Technology, North China Electric Power University, Baoding, China

[a]wwk55551111@hotmail.com

Keywords: Cluster analysis, Cover model, Superposition Principle, Data Mining

Abstract. The PM2.5 data in March, 2014, Beijing is researched to locate the source of atmospheric pollutants. Since the data is highly related, this article use cluster analysis to reduce the dimension of the data to make them irrelevant. A weighted Location Set Cover Model, collaborated with superposition principle, is built to predict the atmosphere pollutant sources.

Introduction

Location Set Cover problem is proposed by *Toregas* in 1971 aiming to establish emergency service facilities which can satisfied all of the emergency location with least cost [1]. It has been studied and used in the fields of data mining like location and prediction. *C.S.Revene* and *H.A.Eiseit* suggested the service radius be standardized [1]. *Wenqian Zhu* developed an AOM cover model to locate the source of the heavy metal pollutants [2]. *Daijun Yin* developed a generalized maximum covering model to locate the emergency facilities [3]. The research on Location Set Cover problem has developed in many fields. However, in this problem, different point's contribution to a whole two-dimensional plane can be various depending on the attributes of the center point like position and strength. Moreover, it demand the center points be irrelevant when they are analysed. Therefore, to optimize Location Set Cover problem by measuring different irrelevant point's attributes has become the key of research.

Weighted Location Set Covering Model

Weighted Location Set Cover Model (WLSCM). Weighted Location Set Cover Model also build a two-dimensional plane coordinate system. However, it is optimized by introducing weight in the calculation of distances between points. The following are the steps of Weighted Location Set Cover Model.

Step 1. Select a series of points marked as $C_1, C_2, ..., \mathrm{C}_n$.

Step 2. Construct a distance matrix L that $L = \left(L_{ij\,4\times4}\right), i = 1, 2, ..., n,\ j = 1, 2, ..., n.$

Step 3. Calculate the average distance between C_i and other points with the Eq.2

$$L_i = \frac{1}{\sum_{1}^{m} n_p - 1} \sum_{j\neq1} n_j L_{ij} \qquad i = 1, 2, ..., \mathrm{m}. \tag{2}$$

Here we define that L_{ij} is the distance from C_i to C_j and n_p is the weight of C_p .

Step 4. Draw circles which have the center C_i with radius L_i .

Step 5. Circle overlap to form an area that is the best prediction of source or the best location expected.

Location of atmospheric pollutant source

Classification of observation point based on cluster analysis. Take pm2.5 as an example, the observation point is relatively close to each other which cause linear correlation among points. However, Superposition Principle require the data be orthogonal and uncorrelated so that they can be analyzed to create a more accurate model that can reflect the value in a two-dimensional plane. Cluster analysis is used to reduce the dimension of data and make it orthogonal and uncorrelated.

Step 1. The calculation between points.

The distance between points were calculated by the Eq.3

$$d(x,y)=\sqrt{(x-y)^T\Sigma^{-1}(x-y)} \tag{3}$$

Put the 24hours' PM2.5 data collected by 35 observation points into matlab as original data and calculate the distance matrix.

Step 2. Draw the result of cluster analysis.

Cluster analysis showed the data can be clustered to 9 classes. However, it is possible that some remote points are classified into one class because of coincidentally same value. For the purpose of getting more accurate equivalent pollution areas, the result of cluster analysis is discussed based on the position of the observation points and is finally refined into 6 classes of equivalent area marked A_1 to A_6 . Table.1 shows the final result of cluster analysis.

Table 1 Cluster analysis result of PM2.5

Area	Observation points number included
1	1,2,3,4,5,6,7,8,9,10,11,12,14,15,16,17,19,28,31,32,33,3 4,35
2	18,24
3	23,25
4	13,29,30
5	20,27
6	21,22,26

Positioning the source of pollutant based on WLSCM. The intensity of equivalent observation area are different. Muller accumulation index [4] is introduced to qualify the classes that can be used for the WLSCM. Muller accumulation index measure the index of the equivalent observation points as Eq.4

$$Igeo=\log_2\left(C_n/1.5B\right) \tag{4}$$

Here we define that $Igeo$ is Muller index of an equivalent observation area which may contain several observation points. C_n is the average pollution value of equivalent area A_n. B is a standard.

Table 2 Muller accumulation index and Pollution degree

degree	EL	L-EL	L	M-L	M	S-M	S	NONE
index	7	6	5	4	3	2	1	0

The Muller accumulation index and corresponding degree are showed in Table 2. Calculate the average PM2.5 value and equivalent points of areas came from cluster analysis and mark the equivalent point as $P_i, i=1,2,3,...,n$. Give a proper value of B and calculate the $Igeo$ of C_n. Pick out the points whose $Igeo$ are above 1.7 which is the S-M degree in Table 2. Take the $Igeo$ value as their weights. Distances between equivalent areas were calculated as Eq. 5

$$L_i = \frac{1}{\sum_{1}^{m} n_p - 1} \sum_{j \neq i} n_j L_{ij} \quad , \quad i = 1, 2, 3, \ldots, m$$

(5)

Here we define that n_i is the weight of C_i, L_{ij} is the distance between A_i and A_j. L_i is the average distance from other areas to A_i. Draw circles with the center of P_i and the radius of L_i. Fig. 1 shows the graph according to the WLSCM. The red area is the predicted source of atmospheric pollutant.

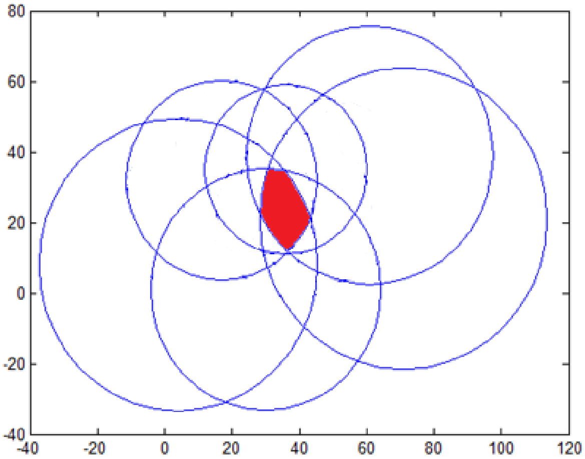

Fig. 1 Cover model graph of predicted atmospheric pollutant source

Conclusion

So far, there are a lot of ways to predict the source of pollutants. However, some models didn't take the correlation of data collected by observation points that are close to each other into consideration. This model assumes that the pollution level of a two-dimensional plane is the superposition result of irrelevant pollution areas. Superposition requires the data be irrelevant. Therefore, cluster analysis is used to reduce the dimension of data. Then, we used Weighted Location Set Cover Model to predict the source of the atmospheric pollutant. From the result of the prediction, it can be concluded that the combination of cluster analysis and cover model has great value of applying in the fields of data analysis and positioning.

References

[1] C. Toregas, R. Swain, C. ReVelle, L. Bergman, The location of emergency service facilities, Operations Research 19(1971)1363-1373.

[2] W. Zhu, The location of the pollutants source based on AOM cover model, Journal of Shangqiu Vocational Technical College, 12(2013)103-104. (In Chinese)

[3] D. Yin, The research of general maximun covering model location application in emergency facilities location (2007). (In Chinese)

[4] G.R. Hu, R.L. Yu, Evaluation of the heavy metal pollutants at Tangtou part of No.324 national highway based on accumulation index method and enrichment factor method, China Mining Magazine 4(2008) 17-19. (In Chinese)