Activation Force-based Air Pollution Observation Station Clustering

Di Huang[1], Ni Zhang[2], Hong Yu[1], Huanyu Zhou[1], Zhanyu Ma[1,*], Weisong Hu[2], and Jun Guo[1]

[1]Pattern Recognition and Intelligent System Lab.,
Beijing University of Posts and Telecommunications, Beijing, China
huangdidi@bupt.edu.cn; hongyu@bupt.edu.cn; zhouhuanyu@bupt.edu.cn; mazhanyu@bupt.edu.cn; guojun@bupt.edu.cn
[2]NEC Labs China, Beijing, China
zhangni_nlc@nec.cn; hu_weisong@nec.cn

***Abstract*—With huge amount of observed air quality and components data, it is of great challenge to analyze and trace the pollutant diffusion path. Partitioning the air pollution sources (air quality observation stations) into subnetworks will help a lot in tracing the air pollution diffusion path. Conventional air pollution sources clustering methods, which are based on geography or pollutant levels, present weak correlation with pollution transmission links. In order to overcome such problem, a method of air pollution sources clustering via activation force (AF) model is introduced in this paper. We model the connections of the pollution sources by AF so that the relationship among the observation stations and the coincidence of the transmission links can be modeled effectively. With the affinity matrix obtained via AF modeling, we conduct clustering of the air pollution sources via modularity measurement. Compared to *K*-means clustering method purely, which is based on the air quality index of pollutants, the proposed approach shows several advantages in air pollution network clustering.**

***Keywords*—Air pollution; Subnetworks; Activation Force; Clustering**

I. INTRODUCTION

Considering the serious air pollution in many cities, it is critical to trace the diffusion path of the pollutant and locate the source so that we can prevent and control pollution efficiently. However, the reason of air pollution is unclear and complicated. Many factors, such as emissions, weather conditions, geographic environment, make the trace of pollution source difficult. Some cities in large scale, e.g., Beijing and Shanghai, have many pollution sources located in different areas, which makes it even more difficult in identifying all the pollution sources directly through limited amount of observation stations. To tackle this problem, we need to find the observation stations with similar patterns and close time-spatial connections. Clustering these observation stations and accordingly partitioning the area into subnetworks will make the diffusion path analysis easier.

The most popular model of air pollution source tracing is HYSPLIT-4 [1]. HYSPLIT-4 is a system for computing simple air parcel trajectories to complex dispersion and deposition simulations. This system was developed from a joint project between the National Oceanic and Atmospheric Administration(NOAA) and Australia Bureau of Meteorology. With enhancements provided by different contributors, such as improved advection algorithms, updated stability and dispersion equations, continued improvements to the graphical user interface, and the option to include modules for chemical transformations, HYSPLIT-4 is widely used in research of air pollution transport, diffusion and source tracing. However, HYSPLIT-4 model is overcomplicated, with multiple meteorological elements and physical process involved. Furthermore, HYSPLIT-4 is often used in diffusion path analysis researching among cities, i.e., in large range.

In order to find a simple and convenient method that is suitable for a city size, we focus on data patterns and model the air pollution data of Beijing based on data mining and information processing technology. It is not clear that how many candidate pollution sources are there in the city area and which observation stations have connections with the same air pollution source. Therefore, it is necessary to partition the observation stations correspond to the same pollution source into a cluster so that the pollution source tracing will be more efficient.

Existing popular methods of air pollution cluster mainly consider the geographical locations [2], pollutant components [3], or pollutant levels [4] [5] [6]. All the aforementioned methods have their drawbacks. In the geographical location-based method [2], different sources have different influence regions. Hence, it is difficult to decide boundaries based only on the geography information. In the method based on pollutant components [3], because the components of air pollutants are very complex and unstable, clustering result is obviously unreliable. Besides, in the pollution levels based method [4], with the transport of pollution, the pollutant levels of the area affected by the same source may change a lot. Hence, it is difficult to cluster the observation stations based only on the pollutant levels. The most important concern is that, the observation stations affected by the same source may present the similar links relation, while the aforementioned methods all ignore such information presented by data, especially their inter-relationship.

*Corresponding author: Zhanyu Ma.

QSHINE 2015, August 19-20, Taipei, Taiwan

DOI 10.4108/eai.19-8-2015.2259688

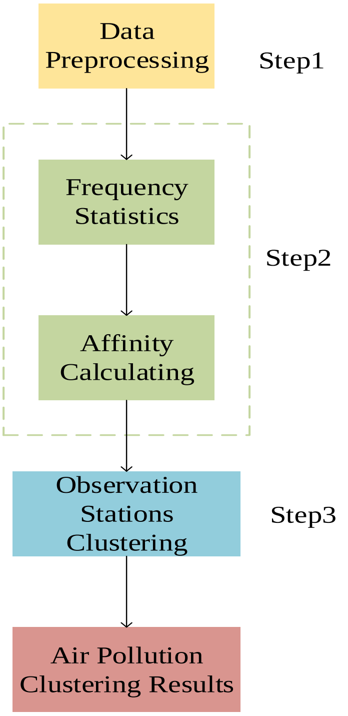


Fig. 1. Air Pollution Observation Stations Clustering Flow Chart.

The purpose of this paper is to present a method for clustering the observation stations, which focuses on data patterns and their internal relationship. The proposed partitioning method is based on the Activation Force (AF) theory [7] which can represent the relations of the observation stations according to their activation orders. With such AFs, an affinity matrix is created to describe the mean overlap rates of the in-links and out-links of the inquired two nodes in a network. Clustering is conducted based on the modularity metric [8]. In contrast to the clustering based on IAQI [9], our approach makes the observation stations in the same cluster reflect the geographical location closeness. This also overcomes the drawbacks in the geographical location-based method. Experimental results demonstrate the good performance of the proposed method.

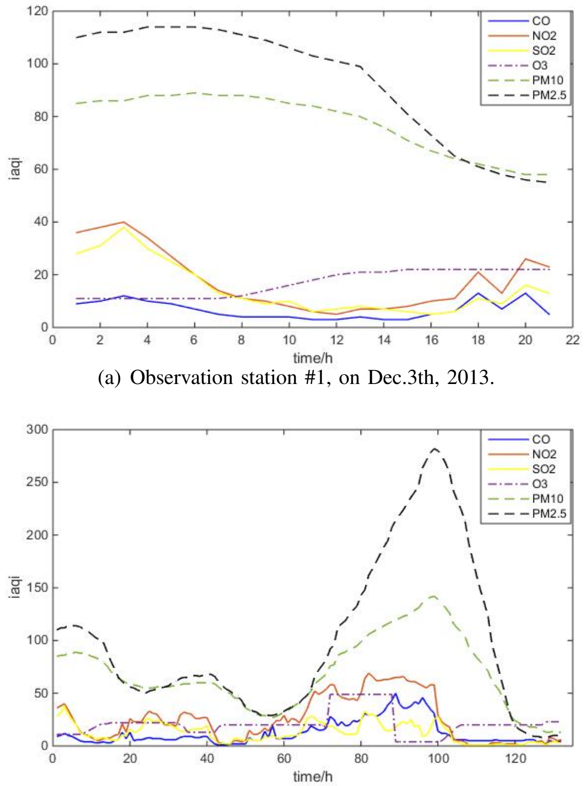
II. CLUSTERING OF AIR POLLUTION OBSERVATION STATIONS

The proposed air pollution observation stations clustering method mainly includes three steps: (1) data preprocessing, (2) modeling data with AF theory, and (3) observation stations clustering based on affinity matrix. The flow chart of this approach is described in Fig.1.

In the following paragraph, we will explain the steps of the proposed method in detail.

A. Data pre-processing

We collected data from 35 observation stations in Beijing area. These data contains six primary air pollutants as CO, NO_2, SO_2, O_3, PM2.5, and PM10. The data were released by Beijing Municipal Environmental Monitoring Center. China posted air quality index (AQI) according to the new Ambient Air Quality Standard (GB3095-2012) [9]. The main pollutants that we analyze are CO, NO_2, SO_2, O_3, PM2.5, and PM10.



(a) Observation station #1, on Dec.3th, 2013.

(b) Observation station #1, from Dec.3th to Dec.9th, 2013.

Fig. 2. IAQI waves per day and per week of the observation station #1.

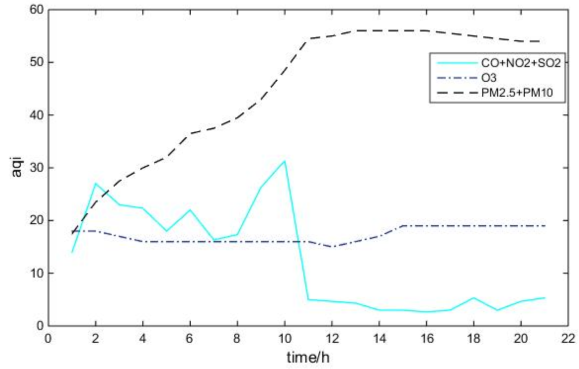


Fig. 3. Trends of the mean IAQI, data collected at observation station #2, on Dec.10th, 2013.

Based on the measured concentration of the pollutants, we can calculate their individual air quality index (IAQI) by equation (1) [10] as

$$IAQI_P = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}}(C_P - BP_{Lo}) + IAQI_{Lo} \quad (1)$$

TABLE I
POLLUTION SPLITTING LEVELS FOR GROUP PM2.5.

IAQI	**1-25**	**26-50**	**51-75**	**76-100**	**101-125**
Level	**1**	**2**	**3**	**4**	**5**
IAQI	**126-150**	**151-200**	**201-250**	**251-300**	**$>$ 300**
Level	**6**	**7**	**8**	**9**	**10**

TABLE II
POLLUTION SPLITTING LEVELS FOR GROUP OXIDE.

IAQI	**1-3**	**4-6**	**7-10**	**11-15**	**16-20**
Level	**1**	**2**	**3**	**4**	**5**
IAQI	**21-25**	**26-30**	**31-40**	**41-50**	**$>$ 50**
Level	**6**	**7**	**8**	**9**	**10**

The overall AQI value is the maximum one in all the IAQIs, as shown in equation (2).

$$AQI = max\{IAQI_1, IAQI_2, IAQI_3, ..., IAQI_n\} \quad (2)$$

In the above equations, $IAQI_P$ represents the IAQI of pollutant P. C_P is the rounded concentration measurement of pollutant P. BP_{Hi} is the breakpoint that is greater than or equal to C_P. BP_{Lo} is the breakpoint that is less than or equal to C_P. $IAQI_{Hi}$ is the AQI value corresponding to BP_{Hi} and $IAQI_{Lo}$ is the AQI value corresponding to BP_{Lo}.

It is difficult to distinguish the pollution components by only using the AQIs. In this paper, our model is based on all the IAQIs of the pollutants. The IAQI values of the six pollutants per day and per week are showed in Fig.2, respectively. From Fig.2, we find that the oxides (CO, NO$_2$, and SO$_2$), the fine particulate matters (PM2.5 and PM10), and O$_3$, have similar distributions respectively. In order to reduce the computational cost, we partition the pollutant into three groups. For the group contains CO, NO$_2$, and SO$_2$, we name it as group oxide. Another group consists of PM2.5 and PM10, and we name it as PM. O$_3$ is assigned to an individual group that contains only itself. The IAQI of each group is defined as the mean value of the IAQIs of the pollutants belong to this group. The trends of three groups are showed in Fig.3.

The pollution levels by Ambient Air Quality Index of Technical Regulations (HJ 633-2012) [11] aims at showing the degree of pollution, which is too rough to show trend variations of different air pollutant. Meanwhile, the data distribution is also not uniform. Therefore, we redefined the pollution levels. To this end, we make histogram equalization to make the data uniformly distributed. Afterwards, we split data into 10 levels for each group. The level division details are shown in TABLE I, II, and III, respectively.

B. Modeling data with AF

1) Activation Force modeling: When analyzing a complex network, efficient clustering will be benefit for exploring the relationship among nodes and the implied structure of the complex network a lot. However, it is difficult to find an effective affinity measure for nodes with the limits of the irregularity information from the given complex network [7]. There are some typical network-weighting schemes, such as independent paths [12], betweenness centrality [13]. However, these work consume much time and request too much information. The observation stations in our case are distributed irregularly, and the AQI collected by the environmental monitor stations can only reflect the concentration information of air pollution data, regardless of the relationship among the observation stations. This makes pollution source tracing extremely difficult. To overcome such problem, we propose a new method of clustering the observation stations by exploiting their affinities through modeling the data by AF. The new clustering method calculates the similarity of link structures between nodes rather than the simple distance between them.

TABLE III
POLLUTION SPLITTING LEVELS FOR GROUP O_3

IAQI	**1-5**	**6-9**	**10-12**	**13-15**	**16-17**
Level	**1**	**2**	**3**	**4**	**5**
IAQI	**18-19**	**20-21**	**22-23**	**24-25**	**$>$ 25**
Level	**6**	**7**	**8**	**9**	**10**

The network is measured by a new type of statistics, named as the AF, from the source data. For a given pair of nodes i and j, the strength of the link from node i to node j, af_{ij}, is defined as

$$af_{ij} = \frac{(\frac{f_{ij}}{f_i})(\frac{f_{ij}}{f_j})}{d_{ij}^2} \quad (3)$$

where f_i is the occurrence frequency of node i, f_{ij} is the co-occurrence frequency of node i and node j, and d_{ij} is a distance between the two nodes in their co-occurrences. The defined strength of the link from node i to node j is called the activation force from node i to node j. Based on the activation forces, any complex network of interest can be represented by a matrix $A\{af_{ij}\}$, where nonzero elements in the i^{th} row provide the out-links of the i^{th} node (from node i to others), while nonzero elements in the i^{th} column provide its in-links (from others to node i).

The affinity of nodes could be calculated from activation force. The affinities between nodes i and j, A_{ij}^{af},is defined as

$$A_{ij}^{af} = [\frac{1}{|K_{ij}|}\sum_{k\in K_{ij}} OR(af_{ki}, af_{kj})\frac{1}{|L_{ij}|}\sum_{l\in L_{ij}} OR(af_{il}, af_{jl})]^{\frac{1}{2}} \quad (4)$$

where $K_{ij} = \{k|af_{ki} > 0 \ or \ af_{kj} > 0\}$, $L_{ij} = \{k|af_{il} > 0 \ or \ af_{jl} > 0\}$, and $OR(x, y) = \frac{min(x,y)}{max(x,y)}$. Readily, K_{ij} is the set of labels of nodes with out-links to node i or node j, while L_{ij} is the set of labels of nodes with in-links from node i or node j. $OR(x, y)$ is an overlap rate function of x and y.

2) Frequency statistics:

a) Co-occurrence based statistics: Our experimental data is IAQI sequences for hours. We firstly count how many times observation station i and observation station j co-occurred in the same level at the same time, during a period (such as a month, or a week). This count is considered as f_{ij} in equation (3). Similarly, we count how many times the observation station i and j occurred in the corresponding co-occurrence level respectively. These counts are treated as as f_i and f_j in equation (3).

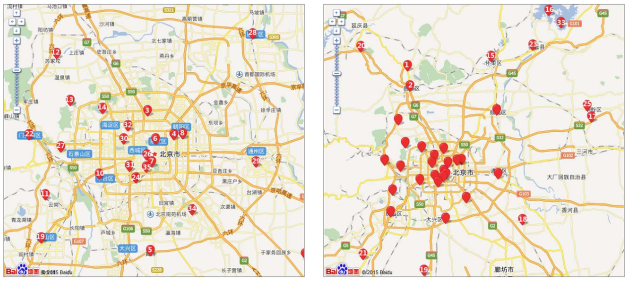

(a) Urban area. (b) Suburban area.

Fig. 4. Distributions of the observation stations. Map provide by Baidu.com, Inc.

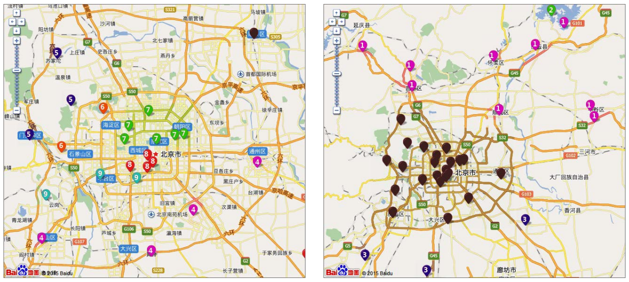

(a) Urban area. (b) Suburban area.

Fig. 5. Sub-netting results, counting frequency by II. B. 2). a), and clustering by K-means based on Affinity. Map provide by Baidu.com, Inc.

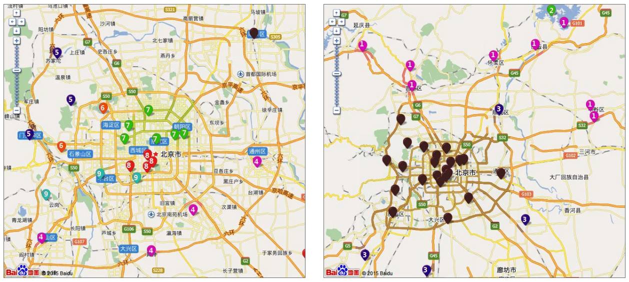

(a) Urban area. (b) Suburban area.

Fig. 6. Clustering results, counting frequency by II. B. 2). b), and clustering by k-means based on Affinity. Map provide by Baidu.com, Inc.

b) Co-grade changing based statistics: In consideration of the co-occurrence based statistics disadvantage in reflecting the relevance of pollution trends, we also make co-grade changing based statistics. In this statistics, we count how many times the observation i and observation j co-occurred in the same level, and also co-occurred in the same level at the next time slot, during a period(such as a month, or a week). We set this count as f_{ij} in equation (3). For the same reasoning, we also count how many times the observation i and j occurred with a similar level changing manner respectively, and set these counts as f_i and f_j in equation (3), respectively.

c) Co-occurrence based statics limited by time-window: Since the above two statistics methods in hours could not show us the order of grade changing, we count frequency during a period defined by a time window. In the time-window, we can get the directional activation force according to the observations order of getting the co-occurrence level, and discover the direction of pollution propagation. We set the time-window as 4 hours, and count how many times the observation station i and observation station j occurrence in the same level, at the same time, in a period (such as a month, or a week). For two observation stations i and j which co-occurred at the same level, if i reaches the co-occurrence level

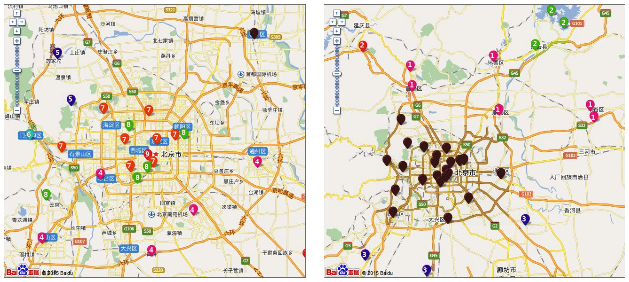

(a) Urban area. (b) Suburban area.

Fig. 7. Clustering results based only on IAQI. Map provide by Baidu.com, Inc.

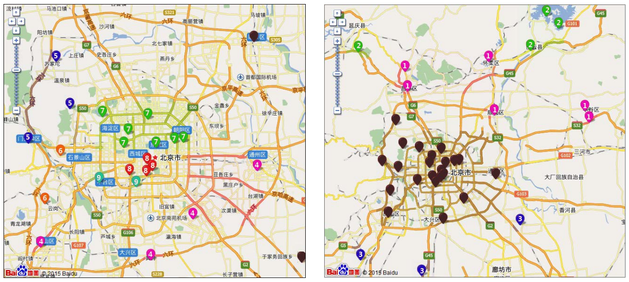

(a) Urban area. (b) Suburban area.

Fig. 8. Clustering results, counting frequency by II. B. 2). a), and clustering by modularity method based on Affinity. Map provide by Baidu.com, Inc.

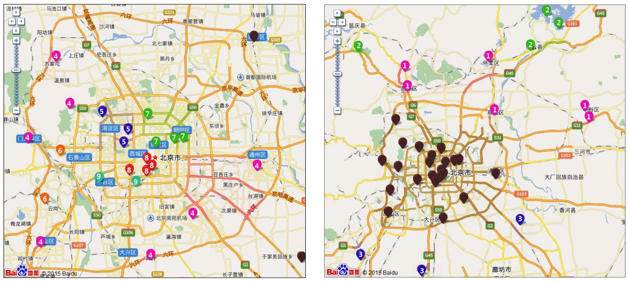

(a) Urban area. (b) Suburban area.

Fig. 9. Clustering results, counting frequency by II. B. 2). b),and clustering by modularity method based on Affinity. Map provide by Baidu.com, Inc.

earlier than j, we add one to f_{ij}. Otherwise, we add one to f_{ji}. And then, similar as the proposed method in section II. B. 2). a), we count how many times the observation station i and j occurrence in the co-occurrence level, and set the counts as f_i and f_j, respectively.

3) Affinity matrix calculation: After the statistics of counts, we can get f_{ij}, f_i and f_j to calculate the activation force, af_{ij}, defined in equation (3). Based on the activation force matrix, we get the affinity matrix, A_{ij}, as shown in equation (4).

C. Subnetwork clustering based on affinity matrix

1) Clustering base on improved K-means with A_{ij}: With the affinity matrix, we cluster the observations by the *K*-means method. Obviously, two observations with large affinity will be closely related. We take the squared root of A_{ij} as the distance between i and j for the purpose of clustering. In addition to this, due to the special properties of the air pollution data, we also make some adjustments to the typical *K*-means method. The typical *K*-means consists four steps [14] [15]:
Step1. Set initial centers of c clusters.
Step2. In the K_{th} iteration, calculate each samples distance to centers, and put the sample into the cluster which has the center closest to the sample.
Step3. Update centers by the method of averaging.
Step4. Keep iterating until the centers converged.

We make improvement for step1 and step3 as follows:
Step1. Initialization. In the consideration of the difference in air pollution between urban and suburban, we split the whole city into two parts, where the Sixth Ring Road is used as the boundary. In this case, we cluster the observations in urban and suburban areas respectively. According to the observation stations geographic position, we chose $c1$ centers inside the Sixth Ring Road (urban area), and $c2$ centers outside the Six Ring Road (suburban area).
Step2. In the K_{th} iteration, calculate the distances of each sample to all centers, and assign the sample to the closest cluster by measuring the distances.
Step3. Centers updating. For each cluster, calculate sum_i, which is the summation of the distances from the i^{th} sample to the rest ones. Set sample j as the new center such that $sum_j = \min_i \ sum_i$.
Step4. Keep iterating until convergence.

2) Clustering based on modularity: Modularity [8] is a popular clustering method in community detection [16]. It is defined as the fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random. The metric was proposed by M. E. J. Newman in 2004, which was originally applied to measure the strength of division of a network into modules (also called groups, clusters or communities). The observation stations and the affinities among these stations in Beijing area can be regarded as a network, with nodes represented by the stations and edges presented by affinity between stations. Clustering by modularity can ensure that the edges inside community have closer relations than those outside the community. This means that, such nodes have high relevance, which meets our requirements. Gephi [17] is a software aims at analyzing complex network via modularity. Gephi provide the function of modularity calculation so that we can cluster the stations by it directly.

III. EXPERIMENTAL RESULTS

Locating the pollution source effectively and efficiently is essential to solve the serious air pollution problems. Clustering observation stations in advance, and tracing pollution source inside subnetworks, rather than the whole city, will make the source tracing more convenient.

Our experiments include two parts.

(1) Network partition with the first two frequency statistics methods in section II. B. 2), and the clustering methods in section II. C. (2) Network partition based on the frequency statistics introduced in section II. B. 2). c) and the modularity based method. We collected data from 35 observations in Beijing, which were released by Beijing Municipal Environmental Monitoring Center. The observation stations are distributed as shown in Fig.4. For the convenience of data processing, we chose only the IAQI sequences of PM2.5, which are representative and relatively complete. The experiment data were collected during seven months, from Dec. 2013 to Jun. 2014. We also make interpolation to fill some missing data. We divide the observation stations in suburban area into 3 clusters and get 6 clusters in urban area. We compare the clustering results obtained by the modularity strategy with the one obtained by the *K*-means method.

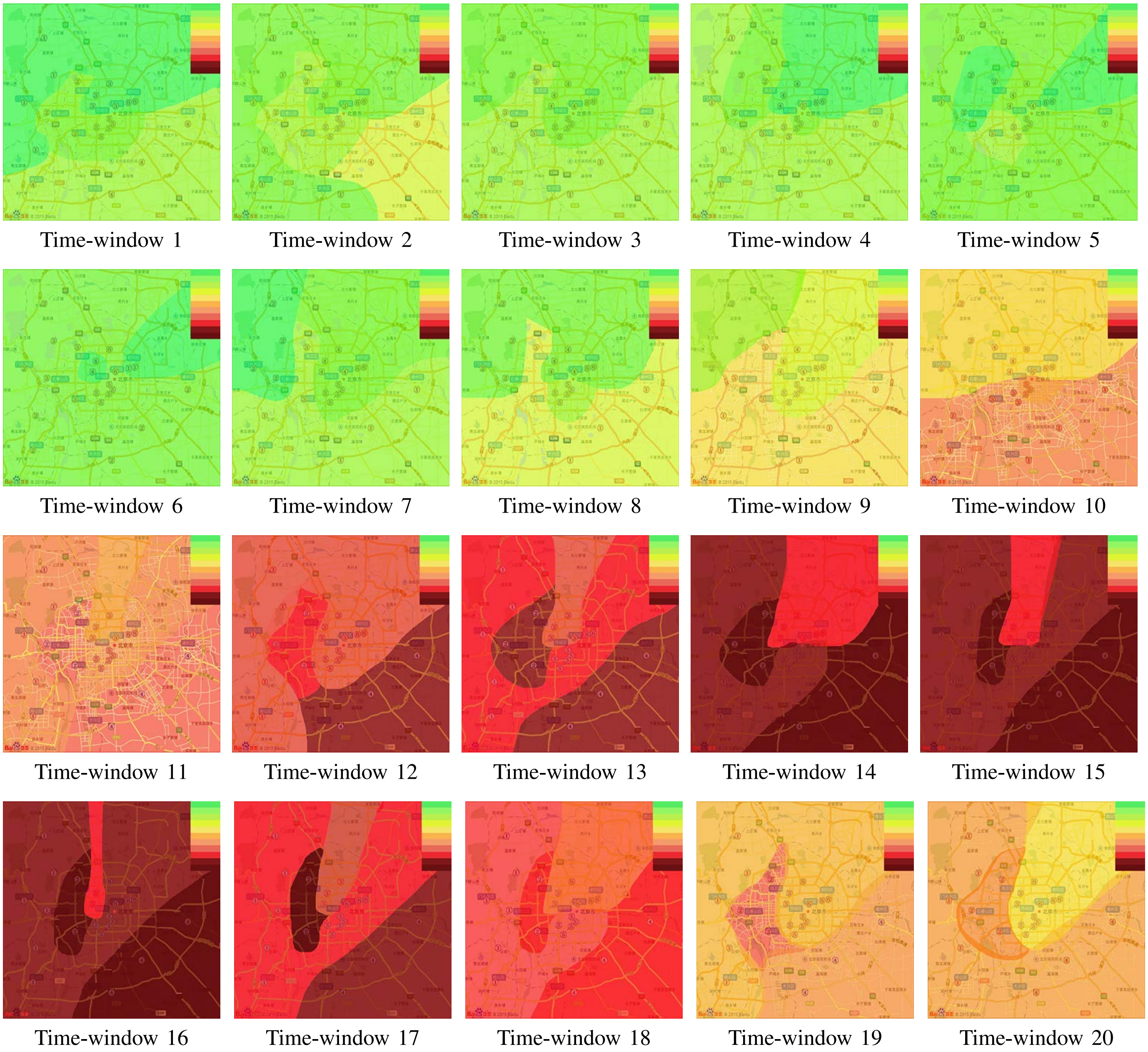
We marked all the results on the map. Labels with same color and same number belong to the same cluster. In this way, we can compare different methods directly. The observation stations affected by the same pollution source will be highly corrected in geographical distribution.

The clustering results are showed in Fig.5-Fig.9. Maps are all provide by Baidu.com, Inc.

We observe that, compared to the clustering based on IAQI and *K*-means method directly(as shown in Fig.7), all the results based on modularity method(as shown in Fig.5, Fig.6, Fig.8, and Fig.9) show higher relevance in distribution. Hence, we can conclude that the proposed clustering method shows significant improvement.

Beside this, by comparing results in Fig.5 and Fig.6 (or Fig.8 and Fig.9), we can find that the two frequency statistics methods don't show significant difference. It indicates that the two methods reflect the same relevance between observation stations.

To illustrate the pollution trends better, we also draw the pollution diffusion trend with several consecutive windows. We used the IAQI frequency of PM2.5 from Feb. 2014 to Mar. 2014, and divide the observation stations in urban area into 6 clusters. We set the time-window as four hours and slide it every ten hours. Colors from green to red denotes the pollution level from the lowest to the highest. 20 pictures in Fig.10 show the pollution diffusion trend clearly.


Time-window 1 Time-window 2 Time-window 3 Time-window 4 Time-window 5

Time-window 6 Time-window 7 Time-window 8 Time-window 9 Time-window 10

Time-window 11 Time-window 12 Time-window 13 Time-window 14 Time-window 15

Time-window 16 Time-window 17 Time-window 18 Time-window 19 Time-window 20

Fig.10. Pollution diffusion trends in 20 time-windows.

This experiment indicates that the directional activation force can reflect the transmission traces of pollution, and can also be used to trace the air pollution path directly. We will do further research about tracing the air pollution path and source by AF in the future.

IV. Conclusions

This paper presented an air pollution observation station partition method based on the activation force. We utilized the observations stations relevance, which is represented by the affinity calculated from the activation force. The clustered observation stations have highly mutual relevance and also show geographical closeness. The proposed method provides a novel way in air pollution network clustering. This can benefit in solving the consequent source tracing problem in our future work.

Acknowledgment

This work was supported by National Natural Science Foundation of China (NSFC) under Grant Nos. 61402047, 61511130081, and 61273217, Scientific Research Foundation for Returned Scholars, Ministry of Education of China, the 111 Project of China under Grant No. B08004, and EU FP7

IRSES Mobile Cloud Project Grant No. 612212.

References

[1] *Technical Assistance Document for the Reporting of Daily Air Quality*, (EPA-454/B-13-001).

[2] S. R. Dorling, T. D. Davies, and C. E. Pierce, *Cluster Analysis: A Technique for Estimating the Synoptic Meteorological Controls on Air and Precipitation ChemistryłMethod and Applications*, Atmospheric Environment. Part A. General Topics, Vol. 26, Issue 14: pp. 2575-2581, 1992.

[3] E. J. Teng, W. Hu, G. P. Wu, and F. S. Wei, *The Composing Characteristics of Elements in Coarse and Fine Particle in Air of the Four Cities in China*, China Environmental Science, Vol.19, No.3: pp. 238-242, 1999. (in Chinese)

[4] B. Wang, *The Spatial and Temporal Variation of Air Pollution Characteristics in China Adopting Air Pollution Index (API) Analysis*, Qingdao: Ocean University of China, 2008. (in Chinese)

[5] *IBM to Acquire SPSS Inc. to Provide Clients Predictive Analytics Capabilities.*

[6] C. T. Robert, *Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality.* Edwards Brothers.

[7] J. Guo, H. Guo, and Z. Wang, *An Activation Force-based Affinity Measure for Analyzing Complex Networks*, Science Reports, pp. 1-113, 2011.

[8] M. E. J. Newman, *Modularity and community structure in networks*, Proceedings of the National Academy of Sciences of the United States of America 103 (23): pp. 8577C8696, 2006.

[9] *Ambient Air Quality Standards* (GB 3095ł2012).

[10] *Technical Assistance Document for the Reporting of Daily Air Quality* (EPA-454/B-13-001).

[11] *Ambient Air Quality Index of Technical Regulations* (HJ 633ł2012).

[12] M. Girvan, and M. E. J. Newman, *Community structure in social and biological networks*, Proc. Natl Acad. Sci. USA 99, pp. 7821C7826, 2002.

[13] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, *Defining and identifying communities in networks*, Proc. Natl Acad. Sci. USA 101, pp. 2658C2663, 2004.

[14] D. S. Modha, and W. S. Spangler, *Feature Weighting in K-means Clustering*, Machine Learning: vol. 47, 2002.

[15] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press. ISBN 0-19-853864-2, 1995.

[16] Clauset, Aaron, M. E. J. Newman, Moore, Cristopher, *Finding community structure in very large networks*, Phys. Rev. E 70 (6): 066111, 2004.

[17] *Fast unfolding of communities in large networks.* J Stat Mech: Theory Exp 2008: P10008. (http://findcommunities.googlepages.com/).