Temporal Patterns in Fine Particulate Matter Time Series in Beijing: A Calendar View

Author: Liu Jianzheng, PhD student

Affiliation: Department of Urban Planning and Design, Faculty of Architecture, The University of Hong Kong, Hong Kong

Address: KB841, Knowles Building 8/F, The University of Hong Kong, Pokfulam Road, Hong Kong

Email: jzliu@hku.hk

Referee: Dr. Weifeng Li

Affiliation: Department of Urban Planning and Design, Faculty of Architecture, The University of Hong Kong, Hong Kong

Email: wfli@hku.hk

Phone: (+852) 2859 2566

THE UNIVERSITY OF HONG KONG 香港大學
faculty of architecture 建築學院
Department of Urban Planning and Design 城市規劃及設計系
Centre of Urban Studies and Urban Planning 城市研究及城市規劃中心

15 December 2015

To UCRN 2016 organizing committee,

Re: Jianzheng Liu

Jianzheng Liu is a PhD student under my supervision in the Faculty of Architecture and I write to recommend him to be considered one of the prestigious UCRN Doctoral Student Paper award.

The paper titled "A Calendar Visualization of Fine Particulate Matter Time-Series in Beijing during 2014" by Mr. Liu addresses one of the most critical challenges confronting urban China: air pollution. This well-written paper adopts an innovative calendar visualization technique and the results reveal very interesting insights that previous studies ignored. I think the paper fits the conference theme well and it would definitely add strength to and help promote the UCRN conference in Xi'an.

I also confirmed that the Department of Urban Planning and Design, The University of Hong Kong will cover Mr. Liu's round-trip transportation between Hong Kong and Xi'an if the paper is selected.

I strongly recommend him for the award. I would be proud for him to receive the award and I have no doubt that he would make UCRN proud too.

Yours sincerely,

Weifeng Li

Dr. Weifeng Li
Assistant Professor
Department of Urban Planning and Design
Faculty of Architecture

8/F, Knowles Building
Pokfulam Road, Hong Kong

Tel/ 852/ 2859 2721
Fax/ 852/ 2559 0468

香港 薄扶林道 鈕魯詩樓 8樓

電話/ 852/ 2859 2721
傳真/ 852/ 2559 0468

http://fac.arch.hku.hk/upad

Abstract

Extremely high fine particulate matter ($PM_{2.5}$) concentration has become synonymous to Beijing, the capital of China, posing critical challenges to its sustainable development and leading to major public health concerns. In order to formulate mitigation measures and policies, knowledge on $PM_{2.5}$ variation patterns should be obtained. While previous studies are limited either because of availability of data, or because of problematic a priori assumptions that $PM_{2.5}$ concentration follows subjective seasonal, monthly, or weekly patterns, our study aims to let data show themselves through visualization rather than imposing arbitrary periodic patterns upon the data. To achieve this, we conduct a time-series cluster analysis on full-year $PM_{2.5}$ data in Beijing in 2014, and provide an innovative calendar visualization of $PM_{2.5}$ measurements throughout the year. Intuitive insights from the analysis on temporal variation of $PM_{2.5}$ concentration show that no weekly patterns are found, and seasonal patterns do exist but do not follow a strict temporal division. These findings contradict current views on temporal patterns in $PM_{2.5}$ data and offer a different perspective which can help with policy formulation on $PM_{2.5}$ mitigation.

Key words: Fine particulate matter, Beijing, Calendar view, Cluster analysis

Introduction

Beijing, the capital of China where more than 20 million people reside, would probably never have considered it would gain the title "Capital of Smog" that was used for London 60 years ago. But now the title seems to fit Beijing appropriately. Clearly, air pollution not only undermines the reputation of Beijing as a historic world-renowned city, but more importantly it poses citizens and the government with a critical challenge for the sustainable development of urbanization that involve major public health concerns. Of all the most common detrimental air pollutants, fine particulate matter ($PM_{2.5}$) is believed to be the most serious pollutant due to its harmful health impact on the cardiovascular, respiratory, and pulmonary functionality in humans[1]. There are increasing evidences that $PM_{2.5}$ is associated with population mortality [2,3], cardiovascular and respiratory diseases mortality [4], and has adverse impacts on growth of new-borns [5], and even on mental health and can cause anxiety [6].

Actions need to be taken to mitigate $PM_{2.5}$ problems in Beijing as well as in other cities of China. For this, we should first measure $PM_{2.5}$ and analyse it to determine inherent variation patterns. Till date, a handle of research efforts have been made to this end [7-11]. While these studies report interesting results about $PM_{2.5}$, they are defected in several ways. Some research are limited because of data availability such as a limited number of $PM_{2.5}$ monitoring stations [7,11], or that data is only available for a limited period [10,12-14]. These studies fail to provide a sufficient overview of $PM_{2.5}$ concentration patterns across the city of Beijing and through a full year. There have been studies that analysed $PM_{2.5}$ measurements data in a full year across Beijing that were provided by the newly launched air pollution monitoring network since later 2012 [8,9,15,16], but these studies, when analysing the temporal variation of $PM_{2.5}$ concentration, used an a priori assumption that $PM_{2.5}$ concentration follows seasonal, monthly, or weekly patterns. The reasoning in these studies is that since the $PM_{2.5}$ concentration probably follows seasonal, monthly, or weekly patterns, the analysis framework could be based on an imposed seasonal, monthly, or weekly profile analysis. We argue that the variation of $PM_{2.5}$ concentration may vary on different time scales other than these predefined scales, and studies using these predefined time scales are likely to be problematic and therefore unconvincing.

Our study, instead of making arbitrary assumptions on weekly, monthly, and seasonal patterns, prefers to let the data show itself. Using a full year $PM_{2.5}$ ground-level measurements from January 2014 to December 2014 in Beijing, our study aims to conduct a time-series clustering for all the daily $PM_{2.5}$ measurements. In this way, our study offers an innovative calendar visualization of $PM_{2.5}$ concentration over the year of 2014, which yields intuitive insights on temporal variation patterns of $PM_{2.5}$ concentration.

The contribution of our study is two-fold. First, our study presents an innovative and straightforward calendar visualization of daily $PM_{2.5}$ time-series in Beijing in the year of 2014. This technique provides a very useful and intuitive tool to visualize and understand the data and can be applied to examine temporal patterns of other air pollutants. Second, the insights generated from the two calendar plots advance our understanding of Beijing's $PM_{2.5}$ concentration. Contrary to conclusions drawn by previous studies on Beijing's $PM_{2.5}$ concentration, our study brings in different yet convincing and detailed insights on $PM_{2.5}$

concentration.

Data and Methods

Data

The $PM_{2.5}$ measurement data in Beijing used in this study were originally obtained from the official hourly air quality reporting platform (http://zx.bjmemc.com.cn/) run by Beijing Environment Protection Agency. This platform is part of the national air quality monitoring network initiated in late 2012. The data is rich, reporting hourly concentrations of six pollutants: particulate matter with aerodynamic diameter no greater than 2.5 microns ($PM_{2.5}$), particulate matter with aerodynamic diameter larger than 2.5 but less than 10 microns (PM_{10}), and sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), and carbon monoxide (CO) in 35 stations across Beijing (Fig. 1). However, the data is not easily accessible because the online reporting system only reports the air quality of the day and does not show historical data and is unavailable to the public. Fortunately, third parties created by civic efforts such as PM25.in, AQISTUDY.cn, and EPMAP.org have been crawling this data since late 2013.

Our study uses one -year air quality monitoring data from 1 January 1 2014 to 31 December 2014 from AQISTUDY.cn, EPMAP.org, and the US Embassy Beijing Air Quality Monitor (Figure 1). We noticed that there are missing hourly measurements in all the three data sources. Therefore, we combined them to get complete $PM_{2.5}$ measurement data covering 24 hours of all the 365 days in 2014. The US Embassy Beijing Air Quality Monitor is operated by the US Department of State. The US Department of State requires that the following disclaimer by included in any publication that uses these data: “State Air observational data are not fully verified or validated; these data are subject to change, error, and correction. The data and information are in no way official.”

A comprehensive data quality check on the raw data is conducted to reduce the impact of problematic data points, including duplicated data records, missing measurements with a placeholder, implausible zeros, etc. After the data quality check, the hourly $PM_{2.5}$ measurement data for all 35 stations are then aggregated into one averaged $PM_{2.5}$ concentration per hour for cluster analysis as explained below.

Method

Since we have 24 hourly $PM_{2.5}$ measurements for each day, it implies we have 365 time-series objects with 24 data points each to analyse. We would like to aggregate together time-series objects with similar variation patterns of $PM_{2.5}$ concentration and separate those with dissimilar patterns into different groups. Thus, we employ time-series clustering technique to mine the data.

In general, there are two essential components in a clustering analysis: clustering algorithm and distance measure [17]. Clustering algorithm controls the procedures on how similar objects are clustered, while distance measures are used to establish the resemblance between two objects. There are several algorithm and distance measures available in the field of cluster analysis but our study employed the most straightforward and suitable clustering method and metrics. Specifically, we use average-linkage agglomerative hierarchical clustering as the clustering method because this method generates repeatable and consistent results and does not require the number of clusters to be specified as compared with K-means [18], and it is usually able to obtain more robust cluster results than other hierarchical clustering methods [19].

Distance measures were selected based on the two basic features of the $PM_{2.5}$ time-series data: level and shape. Level refers to the quantity of $PM_{2.5}$ concentration, and the Euclidean distance is used to identify the level difference between $PM_{2.5}$ time-series. Shape refers to trends in $PM_{2.5}$ concentration variation with respect to time, and we use Pearson's correlation-based distance to capture the shape difference between $PM_{2.5}$ time-series. We derived a generalized correlation-based dissimilarity function from this study[20] by making the coefficient α and power β adjustable (equation (1)).

$$D(x,y) = (\alpha(1-\rho))^{\beta} \qquad (1)$$

Where the correlation coefficient $\rho = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, with\ \alpha,\beta > 0$

This dissimilarity function satisfies all the requirements for dissimilarity measure: the non-negativity, symmetry, and identity [21,22]. When both α and β are set to 1, this dissimilarity function becomes the classic Pearson's correlation coefficient distance that has been used in several studies[23]. In our study, however, we deliberately set α and β to 0.5 and 0.25, respectively, in order to attain a desirable robust cluster result.

We employ the cophenetic correlation coefficient to examine the validity and robustness of the cluster analysis. Cophenetic correlation coefficient is a measure of how faithfully the hierarchical cluster results represent the dissimilarity among observations [24]. It is defined as the linear correlation coefficient between the original pairwise dissimilarities and the cophenetic dissimilarities obtained from the dendrogram. The value of this coefficient varies between 0 and 1. A higher cophenetic correlation coefficient indicates a better cluster solution, and a value of 0.8 or higher is usually regarded as a successful cluster application [25].

It turns out that the cophenetic correlation coefficients for Euclidean-distance-based and correlation-distance-based cluster analyses are 0.86 and 0.81, respectively, suggesting that both cluster results are robust and valid.

We used Python to process and analyse the data, and R to draw the calendar plots.

Results and Discussion

Figure 2 shows two calendar views of the cluster analysis using the correlation distance and Euclidean distance, and two corresponding trend curves of averaged $PM_{2.5}$ concentrations. We obtain three clusters for the analysis based on correlation distance, and each of them has 162, 117, and 86 time-series (days). We named these as S1, S2, and S3, as shown in Fig. 2. For the cluster analysis based on Euclidean distance, nine clusters are formed, each consisting of 255, 82, 15, 5, 2, 2, 2, 1, and 1 time-series. They are named as L1, L2, L3, L4, O1, O2, O3, O4, and O5 (Fig. 2). Those clusters with less than three time-series, namely O1, O2, O3, O4, and O5, are considered as “outliers” that either have extremely high $PM_{2.5}$ concentration or exhibit odd variation patterns. We will discuss these “outliers” later.

Interpretation on calendar visualization

The calendar plot based on correlation distance (top subplot in Fig. 2) and corresponding curve (left bottom in Fig. 2) shows the cluster result based on shape differences among the 365 $PM_{2.5}$ time-series. The result shows that there are about three distinct variation patterns for the $PM_{2.5}$ time-series. An increasing pattern from 0 AM to 11 PM in a day is most likely to be observed from January to March and from September to December (S1 in Fig. 2). For these days that show an increasing $PM_{2.5}$ concentration pattern, the maximum $PM_{2.5}$ concentration of the day usually occurs at night. The decreasing pattern can be observed in all months throughout the year (S2 in Fig. 2) and this pattern attains its minimum value in the afternoon. The third pattern with a shape like an inverted V often take place from April to August (S3 in Fig. 2) and the maximum $PM_{2.5}$ concentrations during these days usually peaks at noon. Different from conclusions drawn in previous studies on diurnal variations of $PM_{2.5}$ concentration [9,26], our findings show that there is more than one diurnal pattern of $PM_{2.5}$ and the diurnal patterns vary considerably through the year.

We can see that not all variation patterns in $PM_{2.5}$ concentration match human activities such as transportation that usually peaks in the morning and afternoon during a full day. The third pattern (S3) is the closest one that possibly matches human activities but this pattern usually happens from April to August. This suggests that human activities may be one of the multiple factors that affect variations of $PM_{2.5}$ concentration but do not play a key role; and their effect may vary at different time periods or under different scenarios. We speculate that from January to March and September to December, weather conditions including wind, temperature, etc., may be the major factors determining variations in $PM_{2.5}$ concentration. However, from April to October, human activities may have larger impact on $PM_{2.5}$ variation than other factors.

The cluster result based on differences in $PM_{2.5}$ concentration levels can be found in the calendar plot based on Euclidean distance (second top subplot in Fig. 2) and the corresponding curve (right bottom in Fig. 2). We can see that a majority of days in the year have an averaged $PM_{2.5}$ concentration of over 50 $\mu g/m^3$ (L1 in Fig. 2), a figure far from the WHO (25 $\mu g/m^3$) and USA air quality standards (15 $\mu g/m^3$). The calendar plot also indicates that high averaged $PM_{2.5}$ concentration around 150 $\mu g/m^3$ (L2 in Fig. 2) are likely to occur in every month throughout the year. Also, extremely high $PM_{2.5}$ concentration above 250 $\mu g/m^3$ (L3, O1, O2, O3, O4, and O5 in Fig. 2) can be usually observed in January, February, March, October, November, and December. This finding is consistent with previous studies in that $PM_{2.5}$ concentration is generally the highest during winter and lowest during summer [15,16].

Outliers

A few “outliers” (O1, O2, O3, O4, and O5 in Fig. 2) can be found in the second calendar view. For example, two notable “outliers” O4 and O5 on January 15 and February 26, 2014, respectively, show quite drastic variations across the day. As we can see, extremely high $PM_{2.5}$ concentrations (O5 has a maximum $PM_{2.5}$ concentration of 534 $\mu g/m^3$) are observed on the two days and the two incidents were reported by the Guardian [27], Time magazine [26], and Financial Times [28].

One event of particular interest is the Asia-Pacific Economic Cooperation (APEC) Summit on 10 and 11 November 2014 in Beijing. It is reported that in order to maintain a blue sky in Beijing during the APEC Summit, coordinated efforts were taken by the governments of Beijing and six surrounding provinces before the summit [29]. Measures included impositions on road traffic and plant operations. The two calendar visualization plots in our study indicate that $PM_{2.5}$ concentration was very high in mid-October before the summit. For example, on October 19, 24, and 25, the $PM_{2.5}$ concentration was over 150 μg/m^3. After the emission control measures were enforced, the $PM_{2.5}$ concentration was greatly reduced on November 1. However, on November 4, a sharp increase in $PM_{2.5}$ concentration was observed, which was around 150 μg/m^3. Fortunately, a significant reduction occurred on November 5 and $PM_{2.5}$ concentration returned to lower level afterwards by November 15, four days after the summit. These interpretations from the two calendar plots can also be obtained from local observations, but here we would like to note that the two calendar visualizations in our study offer a much more intuitive understanding of the whole picture of $PM_{2.5}$ variations over time than using other tools.

Seasonal and weekly patterns?

As we can see from the two cluster results, both shape and level variation do not follow a strict seasonal pattern. Take the level pattern shown in the second top calendar plot for example. Days in L3 cluster usually occur near winter (in February, March, October, November and December but not January) and days in L1 and L2 clusters can be found in any month throughout the full year which doesn't exhibit very clear seasonal pattern. Although there may exist significant differences in $PM_{2.5}$ concentration levels between different seasons [9,15,16], we argue that the arbitrary seasonal division of variation in $PM_{2.5}$ concentration may result in information loss which may conceal potentially important insights. The calendar visualization used in our study, however, provides an informative and straightforward way to look into variation patterns of air pollutants.

Visual inspection over the two calendar plots also show that there is no universal weekly variation pattern in $PM_{2.5}$ concentration. For example, $PM_{2.5}$ concentrations were the lowest on early weekdays and highest on weekends from February 10 to 16. However, from March 24 to 30, the lowest $PM_{2.5}$ concentrations were observed on weekends while the highest were on weekdays. This finding suggests that again the weekly cycle of human activities may not play a key role in determining variations in $PM_{2.5}$ concentration. Our finding contradict previous studies that report weekly patterns in $PM_{2.5}$ concentration in Beijing [8].

References

1 Yuan, Y., Liu, S., Castro, R. & Pan, X. PM2. 5 monitoring and mitigation in the cities of China. *Environmental science & technology* **46**, 3627-3628 (2012).

2 Wong, C. M. *et al.* Satellite-Based Estimates of Long-Term Exposure to Fine Particles and Association with Mortality in Elderly Hong Kong Residents. *Environ Health Perspect* (2015).

3 Samet, J. M., Dominici, F., Curriero, F. C., Coursac, I. & Zeger, S. L. Fine particulate air pollution and mortality in 20 US cities, 1987–1994. *New England journal of medicine* **343**, 1742-1749 (2000).

4 Dominici, F. *et al.* Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA: the journal of the American Medical Association* **295**, 1127-1134 (2006).

5 Jedrychowski, W. *et al.* Gender differences in fetal growth of newborns exposed prenatally to airborne fine particulate matter. *Environ. Res.* **109**, 447-456 (2009).

6 Power, M. C. *et al.* The relation between past exposure to fine particulate air pollution and prevalent anxiety: observational cohort study. *BMJ* **350** (2015).

7 Streets, D. G. *et al.* Air quality during the 2008 Beijing Olympic Games. *Atmospheric Environment* **41**, 480-492, doi:10.1016/j.atmosenv.2006.08.046 (2007).

8 Chen, W., Tang, H. & Zhao, H. Diurnal, weekly and monthly spatial variations of air pollutants and air quality of Beijing. *Atmospheric Environment* **119**, 21-34 (2015).

9 Huang, F. *et al.* PM 2.5 Spatiotemporal Variations and the Relationship with Meteorological Factors during 2013-2014 in Beijing, China. *PLoS One* **10**, e0141642 (2015).

10 Wang, L. S. *et al.* in *Advanced Materials Research.* 874-880 (Trans Tech Publ).

11 Zheng, S., Pozzer, A., Cao, C. & Lelieveld, J. Long-term (2001–2012) concentrations of fine particulate matter (PM 2.5) and the impact on human health in Beijing, China. *Atmospheric Chemistry and Physics* **15**, 5715-5725 (2015).

12 Rohde, R. A. & Muller, R. A. Air Pollution in China: Mapping of Concentrations and Sources. *PLoS One* **10**, e0135749, doi:10.1371/journal.pone.0135749 (2015).

13 Hu, J., Wang, Y., Ying, Q. & Zhang, H. Spatial and temporal variability of PM 2.5 and PM 10 over the North China plain and the Yangtze River delta, China. *Atmospheric Environment* **95**, 598-609 (2014).

14 Chai, F. *et al.* Spatial and temporal variation of particulate matter and gaseous pollutants in 26 cities in China. *Journal of Environmental Sciences* **26**, 75-82 (2014).

15 Wang, Y., Ying, Q., Hu, J. & Zhang, H. Spatial and temporal variations of six criteria air pollutants in 31 provincial capital cities in China during 2013–2014. *Environ. Int.* **73**, 413-422, doi:10.1016/j.envint.2014.08.016 (2014).

16 Zhang, Y.-L. & Cao, F. Fine particulate matter (PM2.5) in China at a city level. *Scientific Reports* **5**, 14884, doi:10.1038/srep14884 (2015).

17 Wang, X., Smith, K. & Hyndman, R. Characteristic-based clustering for time series data. *Data mining and knowledge Discovery* **13**, 335-364 (2006).

18 Aghabozorgi, S., Seyed Shirkhorshidi, A. & Ying Wah, T. Time-series clustering – A decade review. *Information Systems* **53**, 16-38, doi:10.1016/j.is.2015.04.007 (2015).

19 Rodrigues, F. M. & Diniz-Filho, J. A. F. Hierarchical structure of genetic distances: effects of

matrix size, spatial distribution and correlation structure among gene frequencies. *Genetics and Molecular Biology* **21**, 233-240 (1998).

20 Golay, X. *et al.* A new correlation‐based fuzzy logic clustering algorithm for FMRI. *Magnetic Resonance in Medicine* **40**, 249-260 (1998).

21 Veltkamp, R. C. in *Shape Modeling and Applications, SMI 2001 International Conference on.* 188-197 (IEEE).

22 Wang, L., Yang, C. & Feng, J. in *Proceedings of the 24th international conference on Machine learning.* 991-998 (ACM).

23 Jaskowiak, P. A., Campello, R. J. & Costa, I. G. On the selection of appropriate distances for gene expression data clustering. *BMC bioinformatics* **15**, S2 (2014).

24 Sokal, R. R. & Rohlf, F. J. The comparison of dendrograms by objective methods. *Taxon*, 33-40 (1962).

25 Romesburg, C. *Cluster analysis for researchers*. (Lulu. com, 2004).

26 Beech, H. *China's Smog Is So Bad They're Now Calling It a 'Nuclear Winter'*, <http://time.com/9802/beijing-air-pollution-nuclear-winter/> (2014).

27 Gardner, D. *Beijing's smog is increasingly toxic for China's politicians*, <http://www.theguardian.com/environment/2015/jan/20/beijings-smog-increasingly-toxic-chinas-politicians> (2014).

28 Hornby, L. *China pollution: Trouble in the air*, <http://www.ft.com/cms/s/0/c8d06578-98d8-11e3-8503-00144feab7de.html#axzz3tkftQYx5> (2014).

29 Zheng, S. R. *Beijing Smog: The Day After 'APEC Blue'*, <http://thediplomat.com/2014/11/beijing-smog-the-day-after-apec-blue/> (2014).

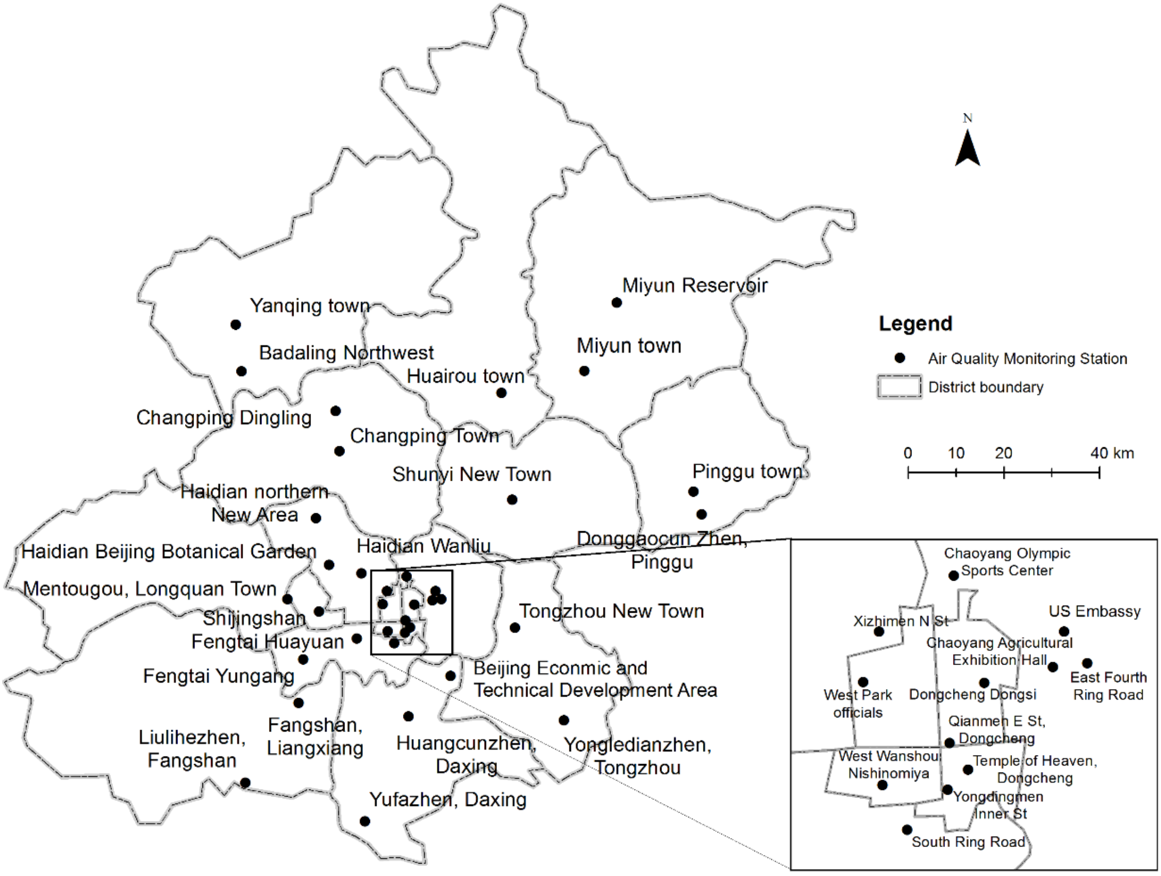

Figure 1 Air Quality Monitoring Stations in Beijing

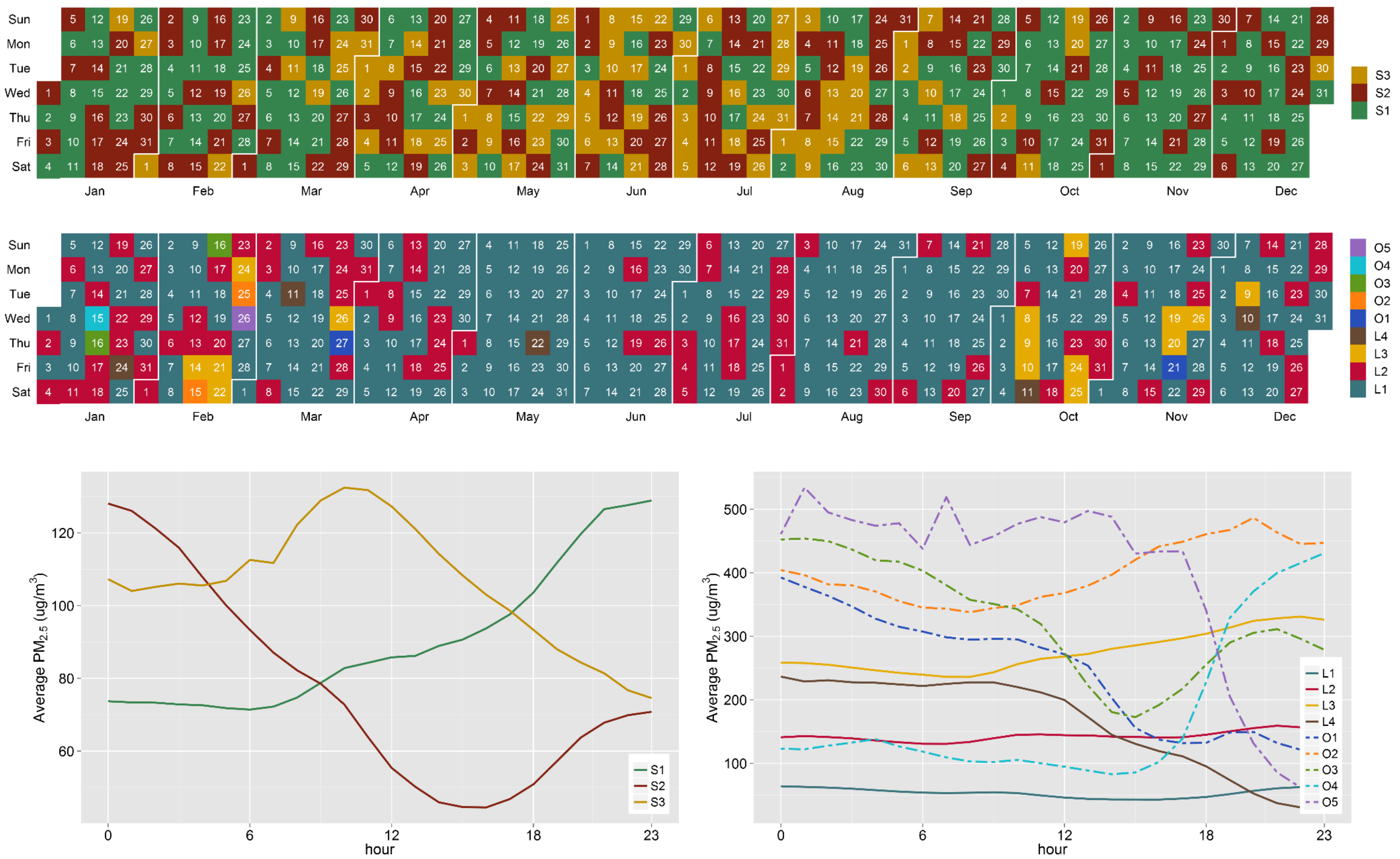

Figure 2 Calendar view of $PM_{2.5}$ concentration clusters in Beijing in the year 2014. (Top) The first calendar shows $PM_{2.5}$ time-series cluster result based on correlation distance, and the letter S denotes "shape"; the second shows the cluster result based on Euclidean distance, L denotes "level" and O refers to "outlier". (Bottom) The left plot shows the averaged $PM_{2.5}$ trend for clusters based on correlation distance, and the right plot shows the averaged $PM_{2.5}$ for clusters based on Euclidean distance. Note that the colours and labels are matched for each cluster for consistency, and the lines for O1, O2, O3, O4, and O5 are set to dash for clear presentation.