Cluster analysis of $PM_{2.5}$ pollution in China using the frequent itemset clustering approach

Liankui Zhang, Guangfei Yang

PII: S0013-9351(21)01304-9

DOI: https://doi.org/10.1016/j.envres.2021.112009

Reference: YENRS 112009

To appear in: *Environmental Research*

Received Date: 10 May 2021

Revised Date: 1 September 2021

Accepted Date: 2 September 2021

Cluster analysis of $PM_{2.5}$ pollution in China using the frequent itemset clustering approach

Liankui Zhang, Guangfei Yang*

Institute of Systems Engineering, Dalian University of Technology, Dalian, China

*Corresponding author at: Institute of Systems Engineering, Dalian University of Technology, No. 2 Linggong Road, Ganjingzi District, Dalian City, Liaoning Province, 116024, PR China

E-mail address: gfyang@dlut.edu.cn (G.Yang)

Abstract

In recent years, severe air pollution has frequently occurred in China at the regional scale. The clustering method to define joint control regions is an effective approach to address severe regional air pollution. However, current cluster analysis research on the determination of joint control areas relies on the Pearson correlation coefficient as a similarity measure. Due to nonlinearity and outliers in air pollution data, the correlation coefficient cannot accurately reveal the similarity in air quality between different cities. To bridge this gap, we proposed a method to delineate spatial patterns of $PM_{2.5}$ pollution and regional boundaries of polluted areas using the frequent itemset clustering approach. The frequent itemsets between cities were first mined, and the support values were employed as interestingness metrics to describe the significance of similar variation patterns between cities. Then, the hierarchical clustering method was applied to identify appropriate areas for joint pollution control. The proposed clustering algorithm exhibits the advantages of not requiring model assumptions and a robustness to the outliers, which is a cost-effective approach to define joint control regions. By analysing urban $PM_{2.5}$ pollution in China from 2015 to 2018, we obtained results demonstrating that the frequent itemset clustering approach can efficiently determine pollution patterns and can effectively identify regional divisions. The clustering approach could facilitate a greater understanding of $PM_{2.5}$ spatiotemporal aggregation to design joint control measures among areas. The findings and methodology of this research have important implications for the formulation of clean air policies in China.

Keywords

Urban air pollution; $PM_{2.5}$; cluster analysis; pollution control

1. Introduction

In recent years, severe $PM_{2.5}$-dominated air pollution attributed to significant economic prosperity and urbanization has remained a persistent issue in China (Fontes et al., 2017). Long-term exposure to $PM_{2.5}$ causes many health problems, such as cardiovascular disease, stroke, respiratory diseases, and lung cancer (Al-Hemoud et al., 2019; Lu et al., 2015). Air quality deterioration has become a notable environmental and social problem in China (Zhang et al., 2012). As a pollutant, $PM_{2.5}$ is prone to transboundary transport via atmospheric circulation, resulting in regional air pollution (Khuzestani et al., 2017; Yu et al., 2021). Hence, the air quality in one city could be highly influenced by neighbouring cities (Chen et al., 2016; Zhang et al., 2020). A strong temporal correlation of $PM_{2.5}$ pollution between cities within 250 km has frequently been observed (Hu et al., 2014).

To strengthen programme planning for the joint prevention and control of atmospheric pollution, China has formulated and implemented a series of relevant policies. In 2012, China designated 13 key areas for joint prevention and control of air pollution based on the level of economic development and the severity of air pollution (Ministry, 2012). To ensure good air quality in Beijing during the Asia-Pacific Economic Cooperation (APEC) period, Beijing and five neighbouring provinces implemented a rigorous joint atmospheric pollution prevention and control programme. This joint policy achieved excellent results and led to the APEC blue phenomenon (Wang et al., 2016a). In 2017, to improve the ambient air quality in the Beijing-Tianjin-Hebei (BTH) region and surrounding areas, the Chinese government formulated a joint control policy involving 28 cities (including 2 municipalities and 26 prefecture-level cities, referred to as the policy of the “2+26” cities). Given the natural and socioeconomic conditions of the different areas in China, there are substantial regional variations in the emissions, transformation, and diffusion of $PM_{2.5}$ pollution, resulting in notable spatiotemporal variations and agglomeration characteristics of the distribution of the $PM_{2.5}$ concentration (Timmermans et al., 2017). However, current regional environmental management cooperation practices in China are mainly formulated for China’s well-developed urban agglomerations, thereby ignoring the spatiotemporal heterogeneity of air pollution (Yao et al., 2020).

Since joint air pollution control in regions is more effective and efficient than a single pollution control strategy applied in respective cities (Wu et al., 2015), the scientific formulation of joint control regional divisions is a reliable strategy to improve both local and regional air quality levels. To address this problem, studies have applied clustering methods to define atmospheric pollution joint control regions. Zhang et al. (2018a) established a network correlation model to demarcate highly intercorrelated regions within China. Based on spatiotemporal clustering of the $PM_{2.5}$ concentration, Chen et al. (2019) divided China into six areas for joint pollution control. Wang and Zhao (2018) defined regional divisions within the BTH region through cluster analysis of the correlation between $PM_{2.5}$ and PM_{10} pollutants. The key to clustering techniques is how similarity is measured, with similarity measures commonly based on the distance. Different distance measurement methods are suitable for data with different characteristics. To date, however, most studies have relied on the Pearson correlation coefficient as a similarity measure (Wang and Zhao, 2018; Wang et al., 2016a; Zhang et al., 2018a). The Pearson correlation coefficient was used to measure the simple linear relationship between two continuous variables (Mukaka, 2012). This coefficient usually requires data to conform to normal distribution and is vulnerable to extreme values. More importantly, this approach cannot assess nonlinear relationships between variables. Since the generation, transformation, and diffusion processes of $PM_{2.5}$ are affected by multiple factors, $PM_{2.5}$ data of a city include complex nonlinear time series data. Due to nonlinearity and outliers in air pollution data, the correlation coefficient cannot accurately reflect the similarity in air quality between different cities. Therefore, the current clustering algorithms cannot accurately determine joint control areas. A scientific, reasonable, and effective method urgently should be adopted to mitigate this problem.

To bridge this gap, in this paper, we propose a novel framework to delineate regional boundaries of $PM_{2.5}$ pollution using the frequent itemset clustering approach. Regional environmental issues are characterized by the fact that the patterns of the variation in air quality among different cities within a region are consistent. Therefore, a region containing cities with similar patterns of the variation in air quality is the optimal joint control region. The similarity in patterns of the variation in air quality among cities can be used to measure the similarity in regional urban air pollution. The proposed method first applies a frequent itemset mining technique to mine similar patterns of variation between cities as a similarity measure and then applies a hierarchical clustering algorithm to define the scope of joint control regions. Frequent itemset mining is a data mining technique that can effectively discover cities with consistent patterns of air pollution variation. The method is quite robust to the presence of outliers because outlier values do not exceed the minimum support threshold during the pattern generation process. Furthermore, the method does not require model assumptions and yields the advantage of easy interpretability, making it more practical for domain experts to better understand and implement the results. This approach is more suitable in complex air pollution domains than traditional methods in terms of effectiveness and interpretability.

The rest of this paper is organized as follows: the data and methods used in this research are introduced in Section 2. Section 3 provides a detailed explanation of the experimental results. A discussion of the research is contained in Section 4. Finally, the conclusion of this study is presented in Section 5.

2. Data and Methods

2.1. Data

The Ministry of Ecology and Environment of China started to establish a national ambient air quality monitoring network in 2013, and the monitoring network has covered all the major cities in China since 2015 (Li et al., 2019; Ye et al., 2018). The original data of $PM_{2.5}$ pollutant concentration is released publicly online by the China National Environmental Monitoring Platform (http://106.37.208.233:20035/) on an hourly basis. We collected the hourly $PM_{2.5}$ pollution data from the National Environmental Monitoring Platform, covering four municipalities and 334 prefecture-level cities across China from 2015 to 2018; and then, the daily, monthly and annual average $PM_{2.5}$ concentration data for each city can be calculated through arithmetic averages method.

2.2. Methods

This paper presents a novel method based on the integration of frequent itemset mining and agglomerative hierarchical clustering to identify potential patterns of the variation in air pollution among cities.

Specifically, the proposed approach first identifies cities with consistent patterns of air pollution variation based on the frequent itemset mining algorithm. Then, the mined frequent itemsets are applied as a similarity measure among cities, and a hierarchical clustering algorithm is applied to define the scope of joint control regions. Finally, the q-statistic test is performed to verify the effectiveness of the proposed approach. The principle of our method is inspired by the divide-and-conquer paradigm widely applied in data mining and machine learning (Witten et al., 2016). This paradigm first breaks down a complex problem into subproblems and then combines the answers to the individual subproblems to appropriately generate the final solution. The frequent itemset mining method reveals the relationships between pairs of cities, and local patterns can be efficiently identified in a dividing manner. After the subproblems (i.e., local cities) are solved, solutions to the whole problem (i.e., the nation as a whole) are considered in a systematic way based on the agglomerative hierarchical clustering method. The clustering approach merges all the determined local patterns and provides a comprehensive view of pollution in cities across China. Previous studies have usually solved this problem in either a piecemeal or integral manner, but our approach gathers abundant knowledge from partitioned problems to tackle the overall problem by combining partial patterns.

2.2.1. Frequent itemset mining

Frequent itemset mining is a data mining technique to find groups of items (i.e., itemsets) that appear frequently together and reveal interesting associations hidden in a database. Frequent itemset mining was first proposed by Agrawal in the context of a database of customer transactions to determine the patterns of purchasing behaviours (Agrawal et al., 1993). Given a customer transaction database, the task of frequent itemset mining is to discover itemsets that are frequently purchased together by customers (Fournier-Viger et al., 2017). In frequent itemset mining, the frequency (or interestingness) of an itemset in a database is usually evaluated by the *support* parameter, which corresponds to the number of transactions where the items in the itemset co-occur divided by the total number of transactions in the database. The support of an itemset *X* in a database D can be estimated using Eq. 1.

$$sup(X) = \frac{\left|\{T \left| T \supseteq X, T \in \mathrm{D}\}\right.\right|}{\left|\mathrm{D}\right|} \qquad (1)$$

where *T* is a set of distinct items that make up database D, and |D| denotes the total number of transactions in D. If an itemset is a frequent itemset, its support value is no less than a user-specified minimum support threshold called *minsup*.

To efficiently mine frequent itemsets, the FP-growth algorithm (Han et al., 2004) was proposed. The algorithm does not need to scan the database repeatedly or to generate a large set of candidates, which greatly reduces the time and space requirements compared to traditional algorithms. The core concepts of the FP-growth algorithm are illustrated by the pseudocode in Fig. S1 in the appendix. Although itemset mining was originally designed for the market basket database, its application can be further widened more generally to discover groups of attribute values frequently co-occurring in various databases. It is profitably exploited in different domains, such as malware process identification (Duan et al., 2015), image classification (Fernando et al., 2012), and bioinformatics data analysis (Naulaerts et al., 2015).

Regional air pollution suggests that cities within a specific region exhibit similar patterns of the spatiotemporal variation in air pollutants. The similarity in patterns of the variation in air quality among cities can be used to measure the similarity in regional urban air pollution to delineate regions of joint prevention and control. Therefore, we mine the frequent itemsets of common $PM_{2.5}$ concentration growth (co-growth) on the same day between two cities to represent similar patterns of spatiotemporal variation. Co-growth of $PM_{2.5}$ in any two cities indicates consistent pollution patterns, which can be caused by their energy/industry structures or geo-climatic environments (Zong et al., 2018). These patterns provide useful information for joint prevention and control of pollution.

2.2.2. Agglomerative hierarchical clustering

The goal of a clustering algorithm is to divide a dataset into different groups, where the objects have higher similarity in the same group and higher dissimilarity with other groups (Peng et al., 2011). In this paper, we apply frequent itemset patterns to measure similarity and to use a hierarchical clustering algorithm to determine groups (Murtagh and Contreras, 2012).

The hierarchical clustering algorithm regards each sample in the dataset as an initial cluster and then combines the two nearest clusters in each step of the algorithm. The algorithm repeats the above steps until the termination condition is reached. The pseudocode of the hierarchical clustering algorithm is displayed in Fig. S2 in the appendix.

To cluster data, we need the distance measure to calculate the similarity between different objects. In this study, the distance measure is expressed as Eq. 2.

$$\text{Distance}(city_1, city_2) = 1 - \text{Support}(city_1, city_2) \tag{2}$$

where Distance ($city_1$, $city_2$) is the distance value between the two cities, and Support ($city_1$, $city_2$) is the support value of the frequent itemset (city_1, city_2). When the co-growth patterns of $PM_{2.5}$ concentrations in two cities are more similar, the intercity support of the two cities is closer to one, and the distance will be close to 0. The two cities with shorter distances should be clustered together. If there are no frequent itemsets between two cities, the distance is set to 1, and they would not be clustered into one group.

According to distances between pairs of clusters, hierarchical clustering algorithms are divided into three categories: single-link, complete-link, and average-link algorithms (Cohen-addad et al., 2019). To calculate the similarity between groups, the average-link algorithm is used in this study as Eq. 3.

$$\text{D}_{avg}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{city_1 \in C_i} \sum_{city_2 \in C_j} D(city_1, city_2) \tag{3}$$

where D_{avg} (C_i, C_j) is the distance of groups C_i and C_j. $|C_i|$ and $|C_j|$ are the number of cities in C_i and C_j, respectively.

2.2.3. Assessment of clustering validity

The q-statistic is widely applied to measure the degree of spatial heterogeneity in ecological phenomena (Wang et al., 2016b). The technique can estimate differences among clusters and the similarity within clusters and the q-statistic is a reliable metric to assess the spatiotemporal clustering validity of $PM_{2.5}$ pollution (Chen et al., 2019). The q-statistic is calculated with Eq. 4.

$$q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2} \tag{4}$$

where q is the value of the q-statistic index and L denotes the number of clusters. N_h and N are the number of cities in cluster h and all clusters, respectively. σ_h^2 and σ^2 denote the variance in the urban $PM_{2.5}$ concentration in cluster h and all clusters, respectively. The value of the q-statistic occurs in [0,1], and a larger value indicates a more efficient regional division of $PM_{2.5}$ pollution.

Transformation of the q-statistic index can satisfy a non-central F-distribution, as expressed in Eqs. 5 and 6, and the F-test can thus be applied to assess the significance of the q-statistic index.

$$F = \frac{N-L}{L-1} \times \frac{q}{1-q} \sim F(L-1, N-L; \lambda) \tag{5}$$

$$\lambda = \frac{1}{\sigma^2}[\sum_{h=1}^{L} \mu_h^2 - \frac{1}{N}(\sum_{h=1}^{L} \sqrt{N_h} \mu_h)^2] \tag{6}$$

where λ denotes the non-centrality parameter and μ_h is the urban average $PM_{2.5}$ concentration in cluster h.

3. Results

In this section, we investigate the patterns of urban $PM_{2.5}$ pollution in Chinese cities by using the FP-growth algorithm and hierarchical clustering algorithm.

3.1. Overview of the $PM_{2.5}$ concentration

From the perspective of temporal variations, Fig. 1 shows the daily average $PM_{2.5}$ concentration in Chinese cities from 2015 to 2018, which ranged from 20 to 135 μg/m^3 across all four years. During these four years, the overall pollution level exhibited a downward trend. According to China's national standards, as listed in Table S1 (HJ633, 2012), the air quality is classified as excellent when the $PM_{2.5}$ concentration is lower than the first level threshold (35 μg/m^3) and polluted when the concentration is higher than the second level threshold (75 μg/m^3). As shown in Fig. 1, the pollution threshold is exceeded mostly in winter (December, January, and February). This indicates that cities in China most commonly suffer from serious $PM_{2.5}$ pollution in winter. On average, nearly 38% of the days in the four years met the first level of the air quality standard, and these conditions usually occurred in summer (June, July, and August).

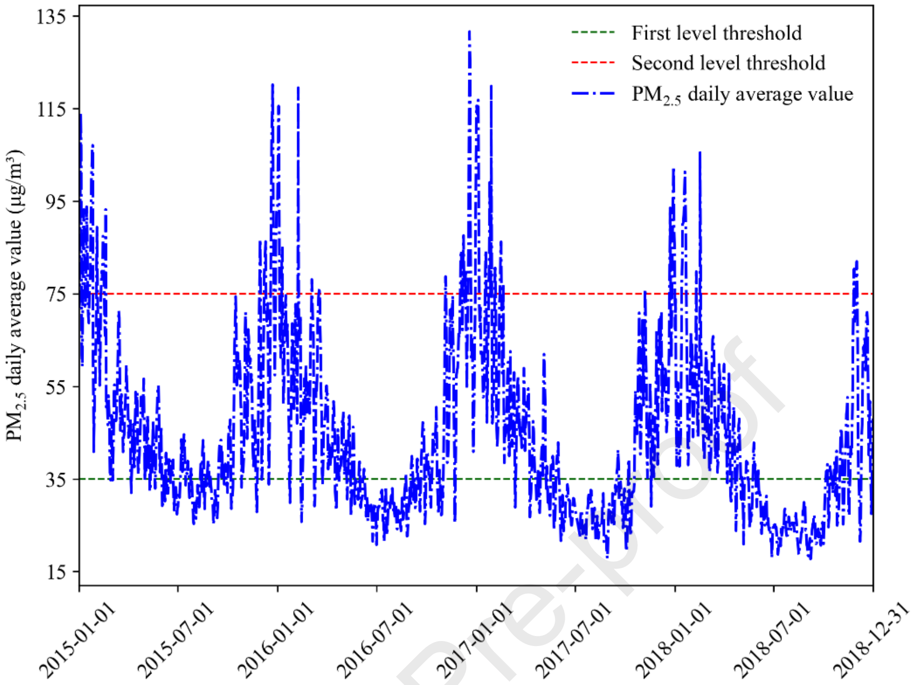

Fig. 1. Daily average of urban $PM_{2.5}$ concentrations in China.

The monthly average $PM_{2.5}$ concentration is shown in Fig. 2, and during most months, the values exhibited a general downward trend from 2015 to 2018. In 2018, the monthly average $PM_{2.5}$ concentration was 20.40% lower than that in 2015. Moreover, the largest and smallest declines in the monthly average $PM_{2.5}$ concentration were 33.95% and 2.64%, respectively, which occurred in September and March, respectively. Compared to 2015, the maximum monthly average $PM_{2.5}$ concentration in 2018 decreased by 9.24%. The maximum and upper-quartile values of the monthly average $PM_{2.5}$ pollution level are distinct, which indicates the unbalanced distribution of the $PM_{2.5}$ concentration in China. Moreover, the largest quartile deviation occurred in July and August, and the smallest deviation occurred in January and December among these four years. This indicates that the air quality imbalance phenomenon among Chinese cities is the most serious in winter and the least severe in summer.

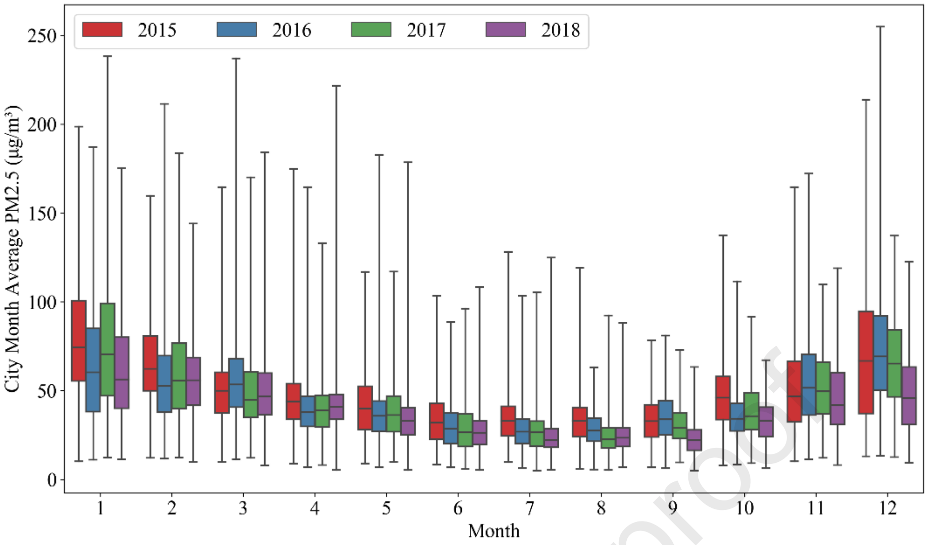

Fig. 2. Monthly average of urban $PM_{2.5}$ concentrations in China.

The cities in this study are shown in Fig. 3, and the annual average $PM_{2.5}$ concentrations from 2015 to 2018 are represented by different colours, where red, yellow, and green denote serious, medium, and slight pollution, respectively. Significant spatial differences in the $PM_{2.5}$ concentration among cities were observed. The highest annual average $PM_{2.5}$ concentrations mainly occurred in two areas: southwestern Xinjiang and the BTH region. Existing research suggests that energy consumption, transport, industry, and population growth are the major contributors to serious $PM_{2.5}$ pollution in the BTH region (Zou and Shi, 2020). In contrast to the BTH region, the air quality in Xinjiang Province is largely affected by the Taklimakan Desert (Geng et al., 2015).

According to the standards of the World Health Organization listed in Table S2 (WHO, 2006), the lowest standard threshold for the annual average concentration is 35 μg/m³. As shown in Fig. 3, most cities in China do not yet reach this standard, which indicates that $PM_{2.5}$ pollution control remains an urgent problem.

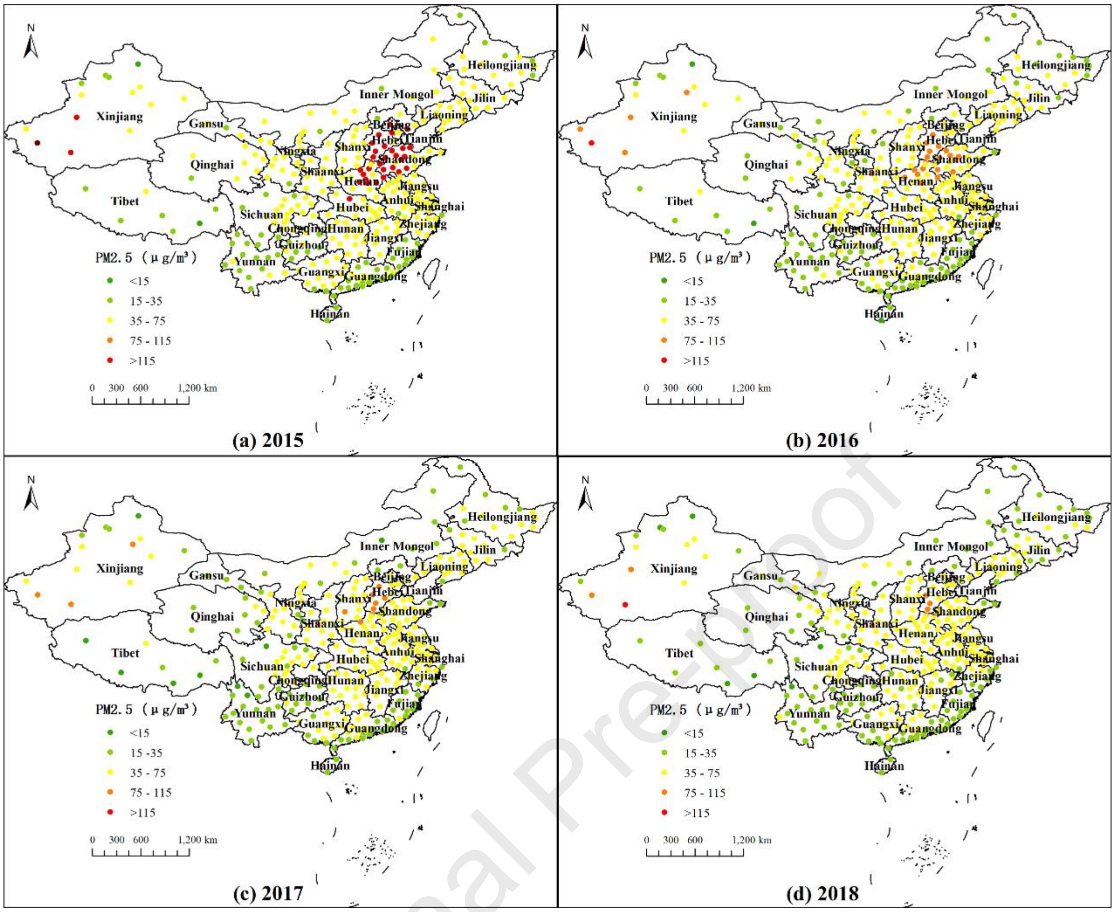

Fig. 3. Annual average of urban $PM_{2.5}$ concentrations in China.

3.2. *Frequent itemset mining of urban $PM_{2.5}$ pollution in China*

Based on daily average pollution data for each city, we mined associative relationships by the frequent itemsets between each pair of cities. By comparing the $PM_{2.5}$ concentration value on the previous day, we first calculated the daily variation value of the $PM_{2.5}$ concentration in each city from 2015 to 2018. Then, all the cities where the daily $PM_{2.5}$ concentration variation status increased on the same day were collected as itemset candidates. Finally, we applied a mining algorithm to efficiently search the frequent itemsets. The threshold *minsup* was set to 0.1 in this study. The identification of frequent itemsets indicated that the $PM_{2.5}$ concentration in two cities increased on the same day with a certain degree of probability. In other words, the cities among the frequent itemsets exhibited a certain consistency or similarity in terms of their $PM_{2.5}$ pollution patterns, where the support values of the frequent itemsets could be adopted as a similarity measure. To visualize the frequent itemsets, the cities in this study were assigned IDs from 1 to 338, and these cities are located in seven geographic regions of China, as shown in Fig. S3 (North China, Central China, East China, South China, Northeast China, Southwest China, and Northwest China, abbreviated as N, C, E, S, NE, SW, and NW, respectively). The IDs for each region were numbered as follows: N (1-35), C (36-78), E (79-156), S (157-195), NE (196-232), SW (233-286), and NW (287-338).

The frequent itemsets obtained through the mining process are visualized on a map with 338×338 grids (please refer to Fig. 4). The left and bottom coordinates of the grid indicate the IDs of the cities, while the right and top coordinates indicate the regions. The colours in each grid cell indicate the support values, where red indicates a high support value and blue indicates a low support value.

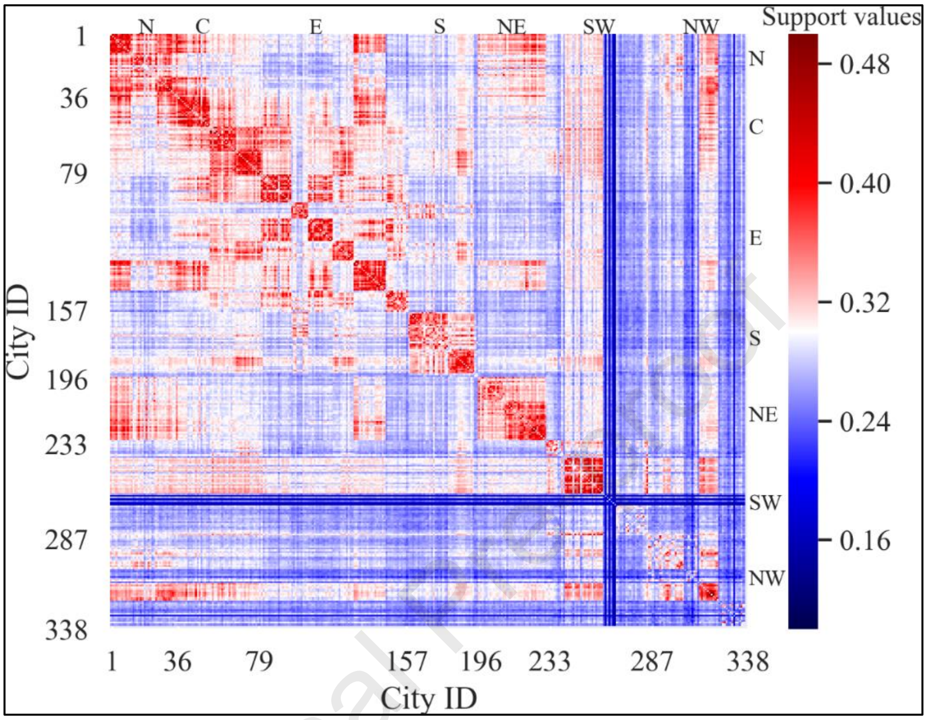

Fig. 4. Frequent itemsets of urban $PM_{2.5}$ pollution in China.

Regional air pollution exhibits similar spatiotemporal features of air pollutant concentrations (Yao et al., 2020), and these features can be used to identify joint control areas. By mining long-term and massive data on the urban $PM_{2.5}$ concentration in China, similar patterns in the variation in air pollutants were found. Fig. 4 shows a comprehensive description of the similarity in patterns of the spatiotemporal variation between cities across China. The air pollution patterns indicate associative relationships between urban air pollution and exhibit obvious regional characteristics. As shown in Fig. 4, the frequent pollution patterns among cities are aggregated, especially near the diagonal region. This result reveals significant mutual effects of $PM_{2.5}$ pollution between adjacent cities, which is consistent with the findings of previous studies (Lu et al., 2019; Zhao et al., 2013). In addition to the cities within a short distance of each other, our results support associations between certain cities within a relatively long distance from each other, which complements previous findings.

The results in Fig. 4 demonstrate that there are frequent patterns widely distributed in many cities, indicating the features of spatial associative relationships and, in particular, adjacent agglomerations of urban $PM_{2.5}$ pollution. Upon close examination of the seven geographic regions of China, as shown in Fig. 4, we observe that most of the pairs of cities within the same region attain higher support values than do those in different regions and that the relationships exhibit distinct features in the different regions. For example, weak frequent itemsets occurred in the SW and NW regions, which suggests that the association of $PM_{2.5}$

pollution between cities in these two regions was poor. This phenomenon could be explained by the fact that there are basins in these regions surrounded by mountains, which may reduce pollution dispersion and weaken the relationships of $PM_{2.5}$ pollution in these regions (Geng et al., 2015).

Regionally, air environmental issues are characterized by the fact that the patterns of the variation in air quality among the different cities within a region are consistent. Interaction areas with frequent pollution patterns can reveal the underlying associations of $PM_{2.5}$ pollution and capture the regional characteristics of $PM_{2.5}$ pollution. Hence, consistent patterns of air quality variation among cities can be used to identify the regional scope of joint pollution control. The relationships among cities enable delineation of the boundaries of regional $PM_{2.5}$ pollution, facilitating the grouping of related cities with frequent patterns into a joint control region and amplification of the effects of pollution prevention policies. Achieving this goal calls for a clustering approach to help analyse all the relationships from an overall perspective, where cities within the same cluster have strong associations, while cities in different clusters have weak associations. This issue is examined in the following subsection.

3.3. Clustering results of urban $PM_{2.5}$ pollution in China

To determine the number of clusters, a distance threshold is defined to measure the dissimilarity between objects in the same cluster. If the distance between two objects is below the threshold, these objects belong to the same cluster. The threshold is the key parameter of hierarchical cluster analysis, which determines how many clusters are generated. Parameter selection depends on the target problem within the context of the data under investigation (Estivill-Castro, 2002). In this study, the distance parameter of the hierarchical cluster is set to 0.7, which indicates that two objects (i.e., cities) are grouped into one cluster only when the similarity in their pollution patterns is greater than 0.3.

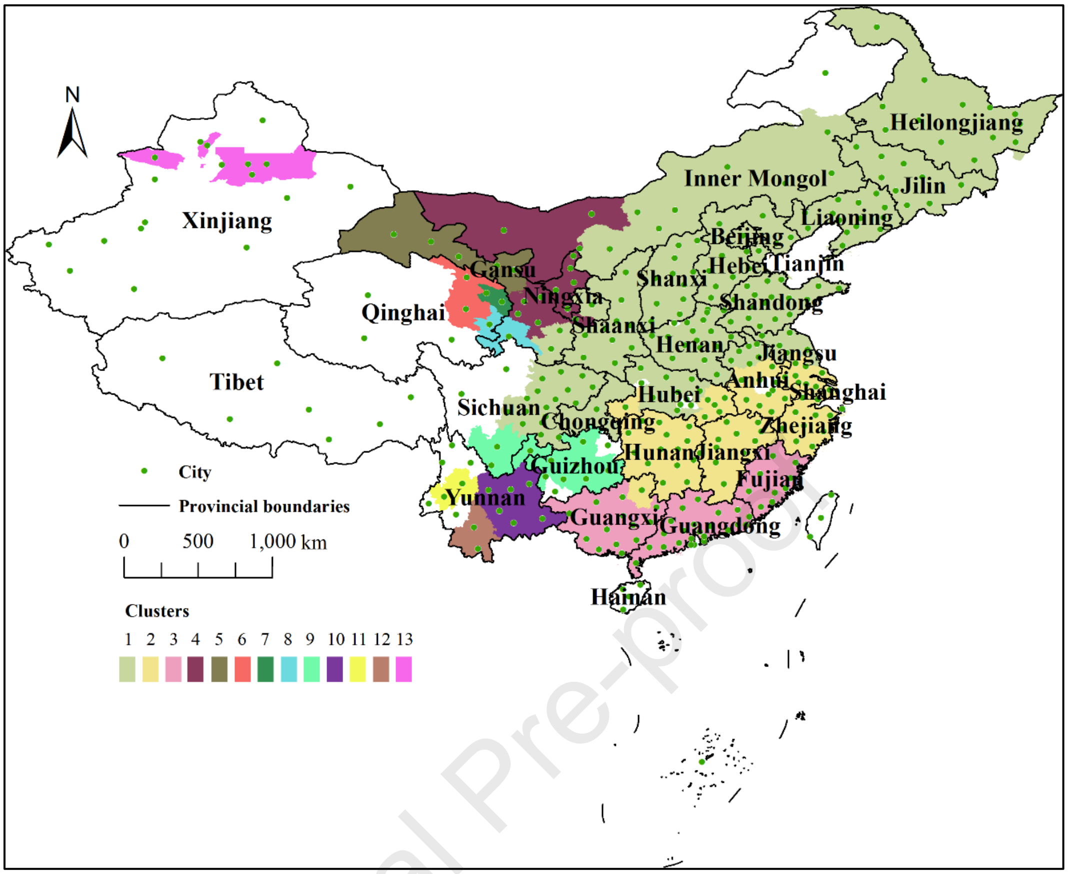

Fig. 5. Clusters of urban $PM_{2.5}$ pollution in China.

Through the clustering algorithm, the cities in China can be divided into 13 clusters (more detailed information is provided in Table S3 in the appendix). The value of the q-statistic is affected by hierarchical differentiation in the data sample and the interpretability of clustering methods regarding the spatial heterogeneity among clusters. To evaluate the validity of the clustering results, the q-statistic threshold was set to 0.27 by referring to related study on $PM_{2.5}$ pollution division in China (Chen et al., 2019). Our results demonstrate that the q-statistic of the proposed method is 0.29 and P-value is significant at the 0.01 level. Compared to the baseline model, the relatively large q-statistic index value indicates that the frequent itemset clustering approach is an effective method to properly identify areas for joint pollution control.

Fig. 5 shows the results of the clustering algorithm where the different colours indicate the different clustering partitions, and the results generally agree with Tobler's first law of geography (Tobler, 1970), namely, pollution interdependence and aggregation effects are observed for closer cities. Many of the cities in southwestern and northwestern China cannot be clustered because of independence of their pollution patterns, and these cities are usually isolated by geographical features such as mountains, where pollution is mainly caused by local sources. $PM_{2.5}$ pollution in these regions is less influenced by transmission from cities outside these areas (Ye et al., 2018).

Based on the observations in Fig. 5, the 13 clusters could be roughly categorized into three divisions in general.

1) The first division consists of cluster 1, including Heilongjiang, Jilin, Shanxi, Hebei, Henan, and

Shandon provinces, and parts of Anhui, Hubei, Jiangsu, Sichuan, the eastern part of Inner Mongol, Beijing, Tianjin, and Chongqing. This region comprises North China and Central China, which are the most highly polluted, rapidly developing, and densely populated areas (Song et al., 2016). Due to the corresponding meteorological conditions and flat terrain, most cities in these areas share similar pollution patterns. For example, it has been reported that dust storms stemming from Inner Mongol are one of the major reasons for the formation of $PM_{2.5}$ pollution in these regions (Cao et al., 2014; Tan et al., 2012).

2) The second division comprises clusters 2, 3, 9, 11 and 12, which are located to the south of the first division, including Hunan, Jiangxi, and Zhejiang, and parts of Anhui, Hubei, Jiangsu and Shanghai, Guangxi, Guangdong, Yunnan, Guizhou and Fujian. Mountains, lakes and rivers are distributed throughout this division, and the weather is usually humid and rainy. Compared with the first division, the agglomeration effect is less obvious. Most of the cities in this division feature a well-developed economy and dense population. Although there are developed industries, they are mainly light industries with less severe pollution.

3) The third division contains the remaining clusters and is located west of the first division, including the western part of Inner Mongol, Ningxia, Gansu, Qinghai, and Xinjiang. In these areas, there are many mountains, plateaus and deserts, which may limit the large-scale transmission of air pollution between the different cities. The cities in this division are usually less developed, and the population density is not very high.

In summary, there are obvious geoclimatic differences among these three divisions. The line of the Qinling Mountains-Huaihe River, the geographical boundary of China, roughly separates the first and second divisions. There occurs a temperate continental and monsoon climate to the north and a subtropical monsoon climate to the south in China. There are obvious differences in climate, geography, economic development, and lifestyles between both sides of the geographical boundary, resulting in spatiotemporal heterogeneity in air pollution across China. The line of the Qinling Mountains-Huaihe River is also a boundary of China's heating policy, which exerts a notable impact on air pollution (Chen et al., 2013). The third division hosts unique natural conditions such as mountains, plateaus and deserts, and its economy is relatively underdeveloped. These factors limit large-scale transmission of air pollution and separate the third division from the other two divisions.

Specifically, there are three important urban agglomerations in China characterized by a high population density, prosperous economy and heavy pollution, namely, the BTH region, the Yangtze River Delta (YRD), and the Pearl River Delta (PRD). Clusters 1, 2 and 3 among the results are mainly composed of these three urban agglomerations. As national political and economic centres, these three regions have always been the key areas to be considered when designing joint control policies to mitigate pollution. For example, the Air Pollution Prevention and Control Action Plan (APPCAP) was released by the Chinese government in 2013, and this policy mainly focused on the abovementioned three regions. In addition to these three important urban agglomerations, the other cluster areas require more attention and support. The fourth, fifth, sixth, seventh, and eighth cluster regions consist of 24 cities located in western Inner Mongol, Gansu, Ningxia, and Qinghai. These cities exhibit complex patterns of air pollution due to their industry, geographical conditions, and climate. The air conditions in the above regions are seriously impacted by two sand-dust storm areas (the Tarim basin region and the Hexi Corridor region) (Wang et al., 2009). Combined with the existing administrative divisions, these five cluster regions could formulate joint management strategies. Most of the cities in Guizhou Province and two neighbouring cities in Sichuan Province constitute the ninth cluster region. Yunnan Province is affected by the airflow across the Yunnan-Guizhou Plateau and the Bangladesh Plateau, leading to a complex pattern of air pollution (Teng et al., 2018). Yunnan Province is divided into cluster regions 10, 11, and 12. The thirteenth cluster region is located in northwestern Xinjiang, which is one of the most serious air pollution areas in China. This region is the most highly industrialized area in Xinjiang (Turap et al., 2018). Air pollution in Western China is concentrated in a few cities in Xinjiang Province, and these cities should be treated centrally based on their unique local geographical and economic conditions.

It should be noted that Fig. 5 reveals $PM_{2.5}$ pollution patterns in China based on agglomerations, which are not simply divided by the administrative areas commonly adopted to design pollution control policies. The findings suggest that we should pay more attention to pollution patterns beyond administrative divisions and that singular management inefficiently controls air pollution in China. Refined joint control policies should be systematically formulated based on the clusters and divisions shown in Fig. 5. The existing control measures in the different regions should be adapted according to the pollution patterns. Cities within one cluster/division displaying similar air pollution patterns should be subject to unified management measures, and specific emission-reduction strategies should be accordingly implemented. This finding can be a reference for regional control of $PM_{2.5}$ pollution in a more integrated way.

The fact that a city is not located in a given cluster does not suggest that there occurs no air pollution. This indicates that the association between its pollution and that in other cities is weak, so joint treatment may not be appropriate.

3.4. Implications of the clusters and divisions

The clustering results indicate that $PM_{2.5}$ pollution in China is mainly divided into three divisions based on 13 clusters, which has practical implications for pollution control in China.

First, the current joint prevention and control areas should be expanded based on the three urban agglomerations to manifest their radiation effect. The cities involved in the current "2+26" joint control plan are route cities of BTH region air pollution transmission channels and are included in the first division of our study. The winter monsoon originating in Inner Mongol exerts an important impact on the air quality in downwind areas, i.e., Inner Mongol, the BTH region, and the Huaihe River region (Yang et al., 2018). Based on the scope of the first division, the "2+26" plan ignores the transfer of pollutants from Inner Mongol to the BTH region. The air quality in the area of the Huaihe River Basin (Hubei, Chongqing, and Sichuan) is influenced by North China. These areas should consider entering the above joint control plan in the future. The Yangtze River Delta urban agglomeration is a densely populated and well-developed economic centre, mainly near Shanghai, with specific cities in Anhui, Jiangsu, and Zhejiang provinces. Affected by emissions, geographical features, and climate, air pollution exhibits the characteristics of spatial aggregation in the YRD region. Regional joint control of air pollution in the Yangtze River Delta urban agglomeration is of great concern (Ma et al., 2019). However, cluster 2 reveals that it is not enough for joint control areas to be confined to the interior of the YRD. Central China and Jiangxi Province are the major potential atmospheric transport source regions of the YRD region (Ming et al., 2017), and the cities located in Hubei, Hunan, and Jiangxi provinces should be combined with the Yangtze River Delta. The cluster 3 regions encompass South

China, except for Hainan Island, which is one of the seven geographical regions of China. Local emissions and meteorological conditions lead to similar pollution patterns in South China (Wu, 2014). The key areas of joint control should be extended from the PRD region towards the whole inland area of South China. Joint control strategies should stress transboundary air pollutant transport from Southeast Asia (Ma et al., 2019).

Then, the core of joint prevention and control policy is to formulate integrated development plans and build regional coordination mechanisms based on the characteristics of the divisions. To improve the air quality in the whole division, these three urban agglomerations should provide financial and technical support to other cities within their division. Adjustment of the energy structure and industrial structure of the division is the key to achieving sustainable development and air quality enhancement. For example, the establishment of ultra-low emission standards for the steel industry and coal-fired power plants and the implementation of coal-to-gas policies are necessary for the first division region (Geng et al., 2021; Li et al., 2020). To enhance transboundary cooperation, industrial emission permits, emission information collection systems, air pollutant emission standards and treatment technical standards should be jointly established and implemented within a cluster or a division. Cities within the same division should pay close attention to the influence of meteorological factors and emissions on the air quality. Hence, each division should create unified action plans considering air pollution patterns that frequently occur to refine the emission limits of key industries and strengthen joint law enforcement.

It should be noted that the pollution patterns in the third division are relatively independent and that the cluster regions are also relatively scattered. Hence, this region is suitable for the development of individual and customized joint control strategies.

3.5. $PM_{2.5}$ pollution joint control in the BTH region

One of the advantages of the frequent itemset clustering approach in this paper is the hierarchy generated to elaborate the associative relationships between cities, which could provide different granularity levels for scenario analysis. This approach is convenient for both macro-scale analysis of national scenarios and micro-scale analysis of local scenarios. Since the BTH region is the most important political and economic centre in China and greatly suffers from heavy pollution (Wu et al., 2018), it is worth closely examining this region and revealing the refined relationships between cities based on above the hierarchy. Many factors, such as topography, emission sources, and climate conditions, lead to the interconnection of $PM_{2.5}$ among different cities. Extracting pollution patterns in the BTH region could provide useful information towards joint pollution control, which could be optimized from a fine-grained perspective.

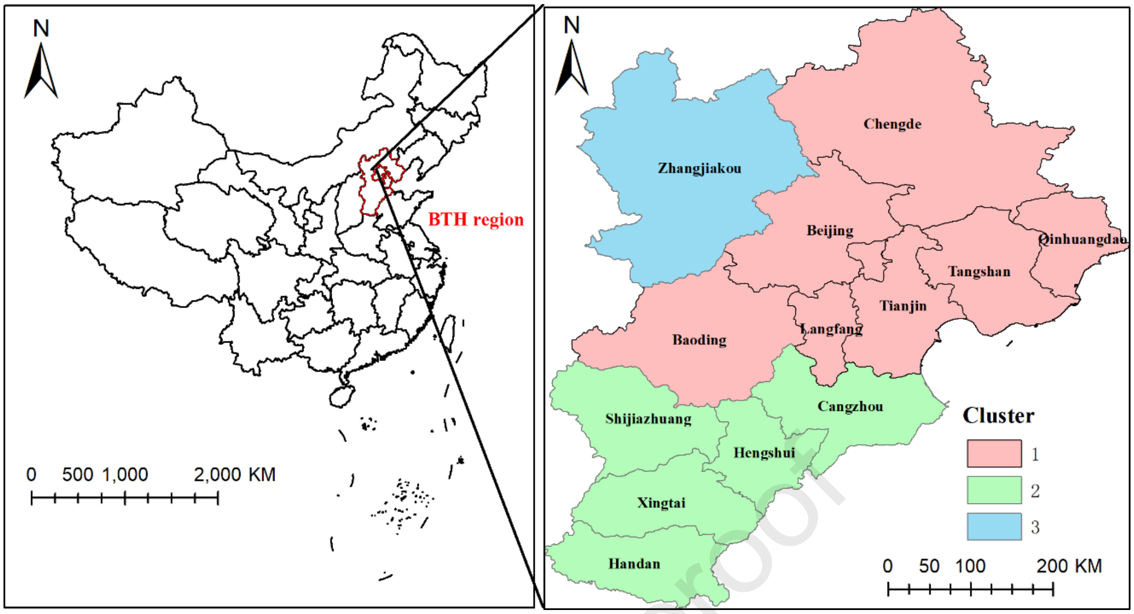

Fig. 6. Clusters of urban PM2.5 pollution in BTH region.

The 13 cities in the BTH region can be roughly divided into three clusters (please refer to Fig. S4 for more details on the hierarchy). As shown in Fig. 6, the cities of Chengde, Beijing, Langfang, Tangshan, Qinhuangdao, Baoding, and Tianjin occur in the first cluster, Shijiazhuang, Hengshui, Cangzhou, Xingtai, and Handan are located in the second cluster; and Zhangjiakou is separate as the third cluster, mainly because the city is isolated by the Yanshan Mountains.

As shown in Fig. S5, Zhangjiakou city is located between the Taihang Mountains and the Yanshan Mountains, where the geography and climate conditions are different from those in the other cities of BTH (Wang and Zhao, 2018). Based on the three clusters generated with the frequent itemset clustering approach, more targeted joint control policies could be designed. More specifically, there are major differences in pollution patterns between the northern and southern BTH regions, which has also been revealed in other studies (Wang et al., 2012; Yan et al., 2018).

The seven cities in the first cluster are west of the Yanshan Mountains and Taihang Mountains. The terrain causes pollution to accumulate more easily (Wang et al., 2016c). The five cities in the second cluster are located in the lowlands of the Taihang Mountains in the west and near the border of Shandong Province (also a heavily polluted area) in the North China Plain, which results in different patterns from those in the first cluster. The three cities of Xingtai, Hengshui, and Handan in the second cluster exhibit high $PM_{2.5}$ concentrations and are affected mainly by their industry structure, energy consumption, and civilian combustion (Wang et al., 2019). In contrast to the cities of Beijing and Tianjin in the first cluster, they contain fewer pollution sources associated with traffic (Zhang et al., 2018b).

At the BTH regional level, industrial structure adjustment and industrial planning considering environmental effects are necessary. To improve the overall air quality, equal attention should be paid to both the southern BTH region and the northern region where the capital is located. Air pollution control policies in the different cluster cities should be based on their characteristics and should be combined with the control strategy of the BTH region. Vehicle emission control and new energy vehicle support are appropriate in the first cluster. The cities in the second cluster should limit industrial emissions and upgrade heavy industries. Because of the better air quality, the third cluster city is not a focus area of pollution prevention and control.

4. Discussion

The objective of this study is to identify pollution areas for joint control with the frequent itemset clustering approach. By accessing a large amount of air pollution data, the spatiotemporal associations of $PM_{2.5}$ among the different urban areas in China were mined to identify joint control areas. The support threshold and clustering parameters were adjusted to optimize the resolution of clustering division. This technique is an effective and feasible method to identify areas for joint control of $PM_{2.5}$ pollution.

To mitigate $PM_{2.5}$ pollution, the following measures and suggestions are proposed based on our research. Due to transboundary transport of $PM_{2.5}$ pollution, joint control of air pollution in China should be extended beyond the limitations of provincial-level administration units, which could be achieved through regulation by the central government. The atmospheric environmental capacity of a region should be scientifically evaluated as natural capital, which is the cornerstone of effective regional air pollution joint control policies (Jin et al., 2016). Total emission quota control within a joint control area could be a feasible strategy to achieve national emission reduction targets (Stranlund and Moffitt, 2014). Reduction targets for air pollution control within a joint region should be defined systematically, and assessment targets should include a reduction target for the whole region and reduction targets for individual cities. Top-level design of regional joint control measures should be strengthened and a professional cross-district governance agency should be established to monitor control efforts, conduct associated law enforcement and formulate specific joint control measures for joint control regions. Due to the economic ties among cities within a region, the industrial chain should be upgraded to reduce air pollution based on the characteristics of the regional industrial structure. Imposing stricter emission standards and switching to cleaner energy sources are effective policies. Because of the negative externalities of air pollution, the rights and obligations of each city within a joint control region must be clear, and an ecological compensation system should be established. According to the so-called polluter pays principle, identifying the pollution contribution of each city in a given region and determining its responsibility for pollution control are important to accomplish effective cooperation. To control haze more accurately, more fine-grained regionalization and measures should be implemented in joint regions.

Due to the complexity of $PM_{2.5}$ generation and transmission, it is difficult to determine joint pollution control areas from the perspective of atmospheric dynamics theory (Zhang et al., 2018a). This paper proposes an effective approach to solve this challenging problem from a data-driven perspective. The data-driven method does not rely on complicated processes, such as the enumeration of external influencing factors and analysis of their relationships. By discovering the underlying patterns in actual data, the clustering approach revealed spatiotemporal relationships to assist the construction of a proper division of $PM_{2.5}$ pollution. The air pollution relationships between cities were determined by mining co-growth pollution patterns. In general, the proposed approach does not require auxiliary data, such as geographical or meteorological data. The approach is insensitive to the data distribution and robust to outliers, and it contains only a few parameters, which makes it easy to implement. In summary, the proposed approach can delineate the regional boundaries of joint air pollution control regions in a cost-effective way.

5. Conclusion

The task of air pollution mitigation has attracted broad attention, and joint air pollution control is a promising strategy. Identifying cities with associated pollution patterns can advance our understanding of the characteristics of $PM_{2.5}$ pollution and guide the formulation of joint control policies. To assist in regional air pollution control, we employed the frequent itemset clustering approach and mined the co-growth pollution patterns between cities. Then, we conducted hierarchical clustering to delineate and optimize regional boundaries for $PM_{2.5}$ joint control in China. The clustering results indicated that $PM_{2.5}$ pollution in China is mainly divided into three divisions based on 13 clusters, and these results could provide a useful reference for regional air pollution control in China. The proposed method was then applied in the BTH region of China to specifically analyse joint pollution control. This study demonstrated that the proposed method can reveal $PM_{2.5}$ pollution patterns and identify $PM_{2.5}$ pollution joint control areas, thereby providing important implications for the design of future clean air policies in China. This study applied a data-driven approach to delineate regional joint control areas based on consistent patterns of air pollution variation. This data mining-based method is highly interpretable and can provide a reference in the field of joint air pollution control. Furthermore, the method has no restrictions regarding the data distribution and is quite robust to the presence of outliers. The technique is practical and can also be applied to the joint control field considering other regional atmospheric pollutants.

CRediT authorship contribution statement

Liankui Zhang: Conceptualization, Software, Data curation, Writing – original draft, Writing - review & editing. **Guangfei Yang:** Conceptualization, Methodology, Supervision, Project administration, Resources, Data curation, Writing - review & editing.

Declaration of competing interest

The authors declared that they have no conflicts of interest to this manuscript.

Acknowledgments

This work was supported by National Natural Science Foundation of China (71671024, 71601028, 71421001).

Appendix

The following is the supplementary information related to this paper:

Supplement figures. Fig. S1 and Fig. S2 show the pseudocode of the FP-growth algorithm and the hierarchical cluster, respectively. Fig. S3 shows the distribution of 338 cities. Fig. S4 shows a hierarchical clustering dendrogram for the BTH region. Fig. S5 shows the BTH region geography. Supplement tables. Table S1 and Table S2 show China's national ambient air quality standards and WHO air quality guidelines and interim targets for $PM_{2.5}$ concentrations. Table S3 shows detailed information on the 13 clusters.

Table S1. China's national ambient air quality standards for $PM_{2.5}$ concentrations.

$PM_{2.5}$ 24-h mean concentrations (μg/m^3)	Level of air quality
≤35	Excellent
(35,75]	Good
>75	Polluted

Table S2. World Health Organization ambient air quality standards for $PM_{2.5}$ concentrations.

Items	Annual concentrations (μg/m^3)
Air quality guideline	10
Interim target-3	15
Interim target-2	25
Interim target-1	35

```
Input: database D and threshold minsup
FP-tree construction
scan D once, get frequent items F, the list of frequent items FL
create the root of an FP-tree, label it as "null"
FOR itemset in D
    sort frequent-item list [p | P]
    call Insert_tree ([p | P], FP-tree)
ENDFOR
Insert_tree ([p | P], null)
IF N.item-name=p.item-name
    N's count =  N's count +1
ELSE
    create node N, N's count=1
IF P is nonempty
    call Insert_tree (P,N)
FP-growth(Tree, α)
IF Tree contains a single prefix path P
    FOR each combination of the nodes in P, labelled $\beta$
        Pattern=$\beta \cup$ null, support = minimum support of nodes in $\beta$
    ENDFOR
ELSE
    FOR each item $\alpha_i$ in FL
        pattern $\beta = \alpha_i \cup \alpha$, support = $\alpha_i$.support
        construct $\beta$'s conditional pattern-base and then $\beta$'s conditional FP-tree Tree$\beta$
        IF Tree$\beta \neq \emptyset$
            call FP-growth (Tree$\beta$, $\beta$)
    ENDFOR
```

Fig. S1. The pseudocode of the FP-growth algorithm.

The FP-growth algorithm uses the divide-and-conquer strategy to compress the database whose itemset satisfies the minimum support degree into a frequent pattern tree (FP-tree). This process maintains the association relation according to each frequent itemset in the item header table, which determines the corresponding conditional FP-tree, and mines the frequent itemset until all the FP-trees are mined. The algorithm comprises two main steps:

(1) Compression of the dataset by the construction of FP-trees.

The database D is scanned once to collect the frequent items of length 1 (i.e., F) and their supports; then F is sorted in support descending order as FL, the list of frequent items. The root of an FP tree is created and labelled "null". For each itemset in D, the frequent items are selected and sorted according to the order of the FL list. The sorted frequent-item list is then [p | P], where p is the first element, and P is the remaining list. Then, insert tree function is called to construct an FP-tree.

(2) Mining of frequent itemsets based on FP-trees.

For each frequent item in FL, its conditional projection database and FP-tree are constructed. The conditional FP-tree construction process is recursively performed until the constructed new FP-tree is empty or contains only one path. When the constructed FP tree is empty, its prefix is frequent itemsets; when only one path is included, frequent itemsets can be obtained by enumerating all possible combinations and appending the prefix of the tree.

```
Input: data set $D = \{c_1, c_2 \ldots c_m\}$
       distance measure method d
       the number of clustering k
Process:
FOR $i$ in (1, m) // Each sample in the data set as an initial cluster
    $C_i$ = (Huang et al.)
ENDFOR
FOR $i$ in (1, m) // Initialize the distance matrix M
    FOR $j$ in ($i$+1, m)
        $M(i, j) = d(C_i, C_j)$
        $M(i, j) = M(j, i)$
    ENDFOR
ENDFOR
Initialize: current clusters number q = m
WHILE q>k:
    SELECT $\min(M(C_{ii}, C_{jj}), ii < jj)$
    $C_{ii} = C_{ii} \cup C_{jj}$ // Combine the two nearest clusters
    DELETE $C_j$
    FOR $j$ in ($jj$+1, $m$)
        $C_j = C_{j-1}$
    ENDFOR
    FOR $i$ in (1, q-1) // Update the distance matrix M
        $M(ii, j) = d(C_{ii}, C_j)$
        $M(j, ii) = M(j, ii)$
    ENDFOR
    q = q-1
ENDWHILE
Output: $C = \{C_1, C_2 \ldots, C_k\}$
```

Fig. S2. The pseudocode of the hierarchical cluster.

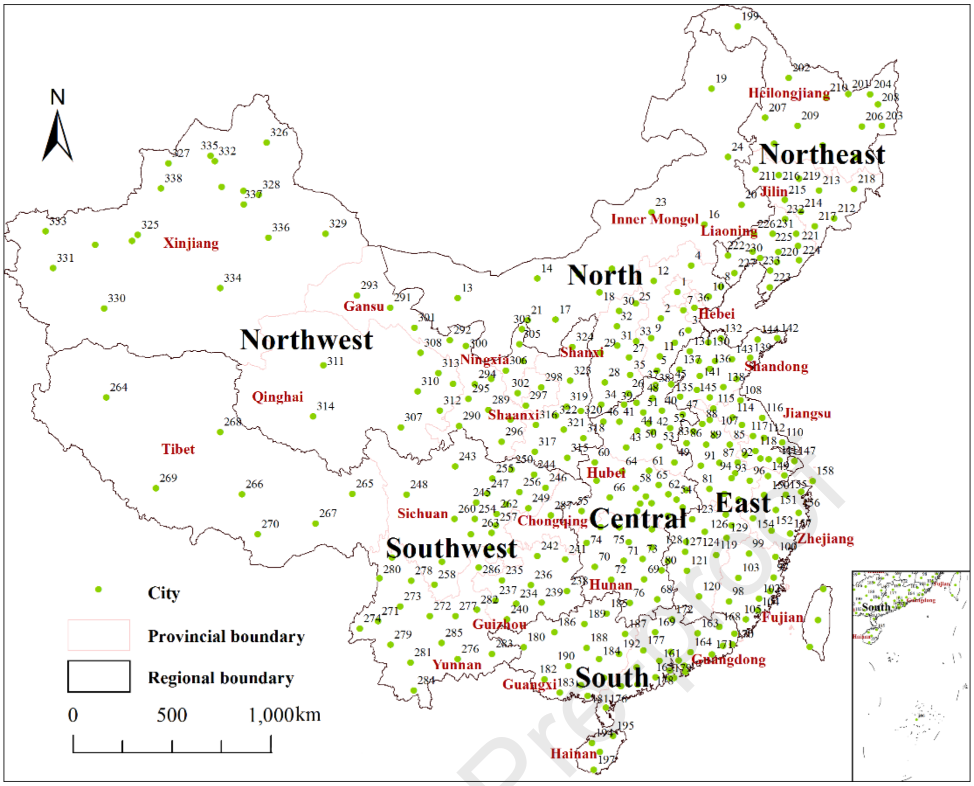

Fig. S3. Distribution of the 338 cities.

Table S3. Specific information on the 13 clusters

Cluster	Province	City Name
1	Anhui	Bangbu, Bozhou, Fuyang, Huaibei, Huainan, Suzhou
	Beijing	Beijing
	Chongqing	Chongqing
	Gansu	Longnan, Qingyang, Tianshui
	Hebei	Baoding, Cangzhou, Chengde, Handan, Hengshui, Langfang, Qinhuangdao, Shijiazhuang, Tangshan, Xingtai, Zhangjiakou
	Heilongjiang	Daqing, Daxinganlingdiqu, Hadongbin, Hegang, Heihe, Jiamusi, Jixi, Mudanjiang, Qiqihadong, Qitaihe, Shuangyashan, Suihua, Yichun
	Henan	Anyang, Hebi, Jiaozuo, Kaifeng, Luohe, Luoyang, Nanyang, Pingdingshan, Puyang, Sanmenxia, Shangqiu, Xinxiang, Xinyang, Xuchang, Zhengzhou, Zhoukou, Zhumadian
	Hubei	Jingmen, Jingzhou, Shiyan, Suizhou, Xiangyang, Xiaogan, Yichang
	Jiangsu	Huaian, Lianyungang, Suqian, Xuzhou, Yancheng
	Jilin	Baicheng, Baishan, Changchun, Jilin, Liaoyuan, Siping, Songyuan, Tonghua, Yanbianzhou
	Liaoning	Anshan, Benxi, Chaoyang, Dalian, Dandong, Fushun, Fuxin, Huludao, Jinzhou, Liaoyang, Panjin, Shenyang, Tieling

	Inner Mongol	Baotou, Chifeng, Dongdongdongsi, Huhehaote, Tongliao, Wulanchabu, Xilinguolemeng, Xinganmeng
	Shaanxi	Ankang, Baoji, Hanzhong, Shangluo, Tongchuan, Weinan, Xian, Xianyang, Yanan, Yulin
	Shandong	Binzhou, Dezhou, Dongying, Heze, Jinan, Jining, Laiwu, Liaocheng, Linyi, Qingdao, Rizhao, Taian, Weifang, Weihai, Yantai, Zaozhuang, Zibo
	Shanxi	Changzhi, Datong, Jincheng, Jinzhong, Linfen, Lvliang, Shuozhou, Taiyuan, Xinzhou, Yangquan, Yuncheng
	Sichuan	Bazhong, Chengdong, Dazhou, Deyang, Guangan, Guangyuan, Leshan, Luzhou, Meishan, Mianyang, Nanchong, Neijiang, Neijiang, Suining, Yaan, Yibin, Zigong, Ziyang
	Tianjin	Tianjin
2	Anhui	Anqing, Chizhou, Chuzhou, Hefei, Huangshan, Liuan, Maanshan, Tongling, Wuhu, Xuancheng
	Guangxi	Guilin
	Hubei	Dongshizhou, Dongzhou, Huanggang, Huangshi, Wuhan, Xianning
	Hunan	Changde, Changsha, Chenzhou, Hengyang, Huaihua, Loudi, Shaoyang, Xiangtan, Xiangxizhou, Yiyang, Yongzhou, Yueyang, Zhangjiajie, Zhuzhou
	Jiangsu	Changzhou, Nanjing, Nantong, Suzhou, Taizhou, Wuxi, Yangzhou, Zhenjiang
	Jiangxi	Fuzhou, Ganzhou, Jian, Jingdezhen, Jiujiang, Nanchang, Pingxiang, Shangrao, Xinyu, Yichun, Yingtan
	Shanghai	Shanghai
	Zhejiang	Hangzhou, Huzhou, Jiaxing, Jinhua, Lishui, Ningbo, Quzhou, Shaoxing, Taizhou, Wenzhou, Zhoushan
3	Fujian	Fuzhou, Longyan, Nanping, Ningde, Putian, Quanzhou, Sanming, Xiamen, Zhangzhou
	Guangdong	Chaozhou, Dongwan, Foshan, Guangzhou, Heyuan, Huizhou, Jiangmen, Jieyang, Maoming, Meizhou, Qingyuan, Shantou, Shanwei, Shaoguan, Shenzhen, Yangjiang, Yunfu, Zhanjiang, Zhaoqing, Zhongshan, Zhuhai
	Guangxi	Baise, Beihai, Chongzuo, Fangchenggang, Guigang, Hechi, Hezhou, Laibin, Liuzhou, Nanning, Qinzhou, Wuzhou, Yulin
4	Gansu	Baiyin, Dingxi, Lanzhou, Linxiazhou, Pingliang
	Inner Mongol	Alashanmeng, Bayannaodong, Wuhai
	Ningxia	Guyuan, Shizuishan, Wuzhong, Yinchuan, Zhongwei
5	Gansu	Jiayuguan, Jinchang, Jiuquan, Wuwei, Zhangye
6	Qinghai	Haibeizhou, Hainanzhou
7		Haidongdiqu, Xining
8	Gansu	Gannanzhou
	Qinghai	Huangnanzhou

9	Guizhou	Anshun, Bijie, Guiyang, Liupanshui, Qiandongnanzhou, Qiannanzhou, Zunyi
	Sichuan	Liangshanzhou, Panzhihua
	Yunnan	Zhaotong
10		Chuxiongzhou, Honghezhou, Kunming, Qujing, Wenshanzhou, Yuxi
11		Baoshan, Dalizhou
12		Pudong, Xishuangbannazhou
13	Xinjiang	Bozhou, Changjizhou, Kelamayi, Wulumuqi

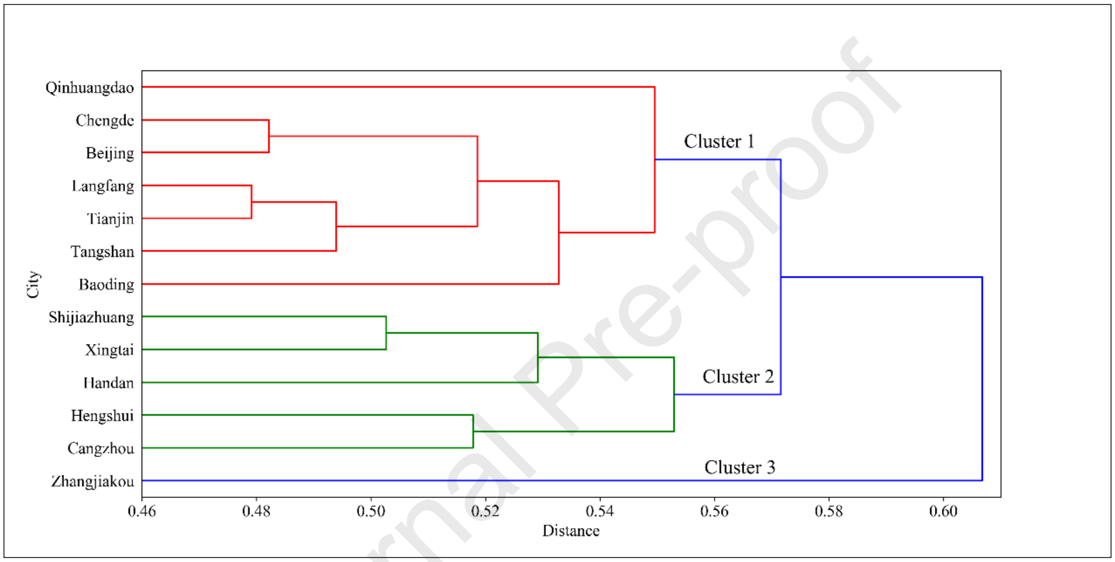

Fig. S4. Hierarchical clustering dendrogram for the BTH region

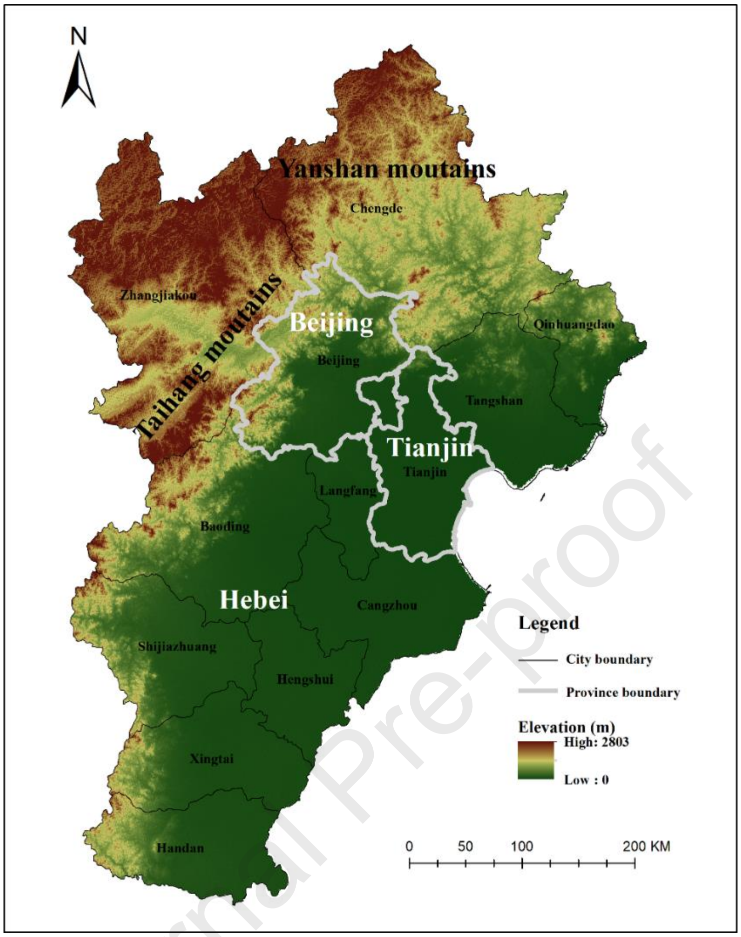

Fig. S5. BTH region geography

References

Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases, Proceedings of the 1993 ACM SIGMOD international conference on Management of data. Association for Computing Machinery, Washington, D.C., USA, pp. 207–216. https://doi.org/10.1145/170035.170072.

Al-Hemoud, A., Gasana, J., Al-Dabbous, A., et al., 2019. Exposure levels of air pollution (PM2.5) and associated health risk in Kuwait. Environ. Res. 179, 108730. https://doi.org/10.1016/j.envres.2019.108730.

Cao, C., Zheng, S., Singh, R.P., 2014. Characteristics of aerosol optical properties and meteorological parameters during three major dust events (2005–2010) over Beijing, China. Atmos. Res. 150, 129-142. https://doi.org/10.1016/j.atmosres.2014.07.022.

Chen, Y., Ebenstein, A., Greenstone, M., et al., 2013. Evidence on the impact of sustained exposure to air pollution on life expectancy from China’s Huai River policy. Proc. Natl. Acad. Sci. U.S.A. 110, 12936-12941. https://doi.org/10.1073/pnas.1300018110.

Chen, Z., Chen, D., Xie, X., et al., 2019. Spatial self-aggregation effects and national division of city-level PM2.5 concentrations in China based on spatio-temporal clustering. J. Cleaner Prod. 207, 875-881.

https://doi.org/10.1016/j.jclepro.2018.10.080.

Chen, Z., Xu, B., Cai, J., et al., 2016. Understanding temporal patterns and characteristics of air quality in Beijing: A local and regional perspective. Atmos. Environ. 127, 303-315. https://doi.org/10.1016/j.atmosenv.2015.12.011.

Cohen-addad, V., Kanade, V., Mallmann-Trenn, F., et al., 2019. Hierarchical Clustering: Objective Functions and Algorithms. J. ACM. 66, 1-42. https://doi.org/10.1145/3321386.

Duan, Y., Fu, X., Luo, B., et al., 2015. Detective: Automatically identify and analyze malware processes in forensic scenarios via DLLs, 2015 IEEE International Conference on Communications (ICC), pp. 5691-5696. https://doi.org/10.1109/ICC.2015.7249229

Estivill-Castro, V., 2002. Why so many clustering algorithms: a position paper. SIGKDD Explor. Newslett. 4, 65–75. https://doi.org/10.1145/568574.568575.

Fernando, B., Fromont, E., Tuytelaars, T., 2012. Effective Use of Frequent Itemset Mining for Image Classification, Computer Vision – ECCV 2012. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 214-227. https://doi.org/10.1007/978-3-642-33718-5_16

Fontes, T., Li, P., Barros, N., et al., 2017. Trends of PM2.5 concentrations in China: A long term approach. J. Environ. Manage. 196, 719-732. https://doi.org/10.1016/j.jenvman.2017.03.074.

Fournier-Viger, P., Lin, J.C.-W., Vo, B., et al., 2017. A survey of itemset mining. WIREs Data Min. Knowl. Discovery. 7, e1207. https://doi.org/10.1002/widm.1207.

Geng, G., Zhang, Q., Martin, R.V., et al., 2015. Estimating long-term PM2.5 concentrations in China using satellite-based aerosol optical depth and a chemical transport model. Remote Sens. Environ. 166, 262-270. https://doi.org/10.1016/j.rse.2015.05.016.

Geng, G., Zheng, Y., Zhang, Q., et al., 2021. Drivers of PM2.5 air pollution deaths in China 2002–2017. Nat. Geosci. 10.1038/s41561-021-00792-3.

Han, J., Pei, J., Yin, Y., et al., 2004. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Min Knowl Disc. 8, 53-87. https://doi.org/10.1023/B:DAMI.0000005258.31418.83.

Hu, J., Wang, Y., Ying, Q., et al., 2014. Spatial and temporal variability of PM2.5 and PM10 over the North China Plain and the Yangtze River Delta, China. Atmos. Environ. 95, 598-609. https://doi.org/10.1016/j.atmosenv.2014.07.019.

Huang, R.-J., Zhang, Y., Bozzetti, C., et al., 2014. High secondary aerosol contribution to particulate pollution during haze events in China. Nature. 514, 218-222. https://doi.org/10.1038/nature13774.

Jin, Y., Andersson, H., Zhang, S., 2016. Air Pollution Control Policies in China: A Retrospective and Prospects. Int. J. Environ. Res. Public Health. 13, 1219. https://doi.org/10.3390/ijerph13121219.

Khuzestani, R.B., Schauer, J.J., Wei, Y., et al., 2017. Quantification of the sources of long-range transport of PM2.5 pollution in the Ordos region, Inner Mongolia, China. Environ. Pollut. 229, 1019-1031. https://doi.org/10.1016/j.envpol.2017.07.093.

Li, J., Zhou, S., Wei, W., et al., 2020. China's retrofitting measures in coal-fired power plants bring significant mercury-related health benefits. One Earth. 3, 777-787. https://doi.org/10.1016/j.oneear.2020.11.012.

Li, R., Wang, Z., Cui, L., et al., 2019. Air pollution characteristics in China during 2015–2016: Spatiotemporal variations and key meteorological factors. Sci. Total Environ. 648, 902-915.

https://doi.org/10.1016/j.scitotenv.2018.08.181.

Lu, F., Xu, D., Cheng, Y., et al., 2015. Systematic review and meta-analysis of the adverse health effects of ambient PM2.5 and PM10 pollution in the Chinese population. Environ. Res. 136, 196-204. https://doi.org/10.1016/j.envres.2014.06.029.

Lu, Y., Wang, Y., Wang, L., et al., 2019. Provincial analysis and zoning of atmospheric pollution in China from the atmospheric transmission and the trade transfer perspective. J. Environ. Manage. 249, 109377. https://doi.org/10.1016/j.jenvman.2019.109377.

Ma, T., Duan, F., He, K., et al., 2019. Air pollution characteristics and their relationship with emissions and meteorology in the Yangtze River Delta region during 2014–2016. J. Environ. Sci. 83, 8-20. https://doi.org/10.1016/j.jes.2019.02.031.

Ming, L., Jin, L., Li, J., et al., 2017. PM2.5 in the Yangtze River Delta, China: Chemical compositions, seasonal variations, and regional pollution events. Environ. Pollut. 223, 200-212. https://doi.org/10.1016/j.envpol.2017.01.013.

Ministry, 2012. The Twelfth Five-year Plan for Prevention and Control of Atmospheric Pollution in Key Regions. Ministry of Environmental Protection, Ministry of Environmental Protection

Mukaka, M.M., 2012. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal. 24, 69-71. https://doi.org/10.2166/wh.2012.000.

Murtagh, F., Contreras, P., 2012. Algorithms for hierarchical clustering: an overview. WIREs Data Min. Knowl. Discovery. 2, 86-97. https://doi.org/10.1002/widm.53.

Naulaerts, S., Meysman, P., Bittremieux, W., et al., 2015. A primer to frequent itemset mining for bioinformatics. Briefings Bioinf. 16, 216-231. https://doi.org/10.1093/bib/bbt074.

Peng, Y., Wang, G., Kou, G., et al., 2011. An empirical study of classification algorithm evaluation for financial risk prediction. Appl. Soft Comput. 11, 2906-2915. https://doi.org/10.1016/j.asoc.2010.11.028.

Song, Y., Wang, X., Maher, B.A., et al., 2016. The spatial-temporal characteristics and health impacts of ambient fine particulate matter in China. J. Cleaner Prod. 112, 1312-1318. https://doi.org/10.1016/j.jclepro.2015.05.006.

Stranlund, J.K., Moffitt, L.J., 2014. Enforcement and price controls in emissions trading. J Environ Econ Manage. 67, 20-38. https://doi.org/10.1016/j.jeem.2013.10.001.

Tan, S.-C., Shi, G.-Y., Wang, H., 2012. Long-range transport of spring dust storms in Inner Mongolia and impact on the China seas. Atmos. Environ. 46, 299-308. https://doi.org/10.1016/j.atmosenv.2011.09.058.

Teng, M., Yang, K., Shi, Y., et al., 2018. Study on the Temporal and Spatial Variation of PM2.5 in Eight Main Cities of Yunnan Province, 26th International Conference on Geoinformatics, pp. 1-7. https://doi.org/10.1109/GEOINFORMATICS.2018.8557198

Timmermans, R., Kranenburg, R., Manders, A., et al., 2017. Source apportionment of PM2.5 across China using LOTOS-EUROS. Atmos. Environ. 164, 370-386. https://doi.org/10.1016/j.atmosenv.2017.06.003.

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography. 46, 234-240. https://doi.org/10.2307/143141.

Turap, Y., Talifu, D., Wang, X., et al., 2018. Concentration characteristics, source apportionment, and oxidative damage of PM2.5-bound PAHs in petrochemical region in Xinjiang, NW China. Environ. Sci. Pollut. Res. 25, 22629-22640. https://doi.org/10.1007/s11356-018-2082-3.

Wang, H., Zhao, L., 2018. A joint prevention and control mechanism for air pollution in the Beijing-Tianjin-Hebei region in china based on long-term and massive data mining of pollutant concentration. Atmos. Environ. 174, 25-42. https://doi.org/10.1016/j.atmosenv.2017.11.027.

Wang, H., Zhao, L., Xie, Y., et al., 2016a. "APEC blue"—The effects and implications of joint pollution prevention and control program. Sci. Total Environ. 553, 429-438. https://doi.org/10.1016/j.scitotenv.2016.02.122.

Wang, J.-F., Zhang, T.-L., Fu, B.-J., 2016b. A measure of spatial stratified heterogeneity. Ecol. Indic. 67, 250-256. https://doi.org/10.1016/j.ecolind.2016.02.052.

Wang, L., Xiong, Q., Wu, G., et al., 2019. Spatio-Temporal Variation Characteristics of PM2.5 in the Beijing–Tianjin–Hebei Region, China, from 2013 to 2018. Int. J. Environ. Res. Public Health. 16, 4276. https://doi.org/10.3390/ijerph16214276.

Wang, L., Xu, J., Yang, J., et al., 2012. Understanding haze pollution over the southern Hebei area of China using the CMAQ model. Atmos. Environ. 56, 69-79. https://doi.org/10.1016/j.atmosenv.2012.04.013.

Wang, S., Feng, X., Zeng, X., et al., 2009. A study on variations of concentrations of particulate matter with different sizes in Lanzhou, China. Atmos. Environ. 43, 2823-2828. https://doi.org/10.1016/j.atmosenv.2009.02.021.

Wang, Y., Jiang, H., Zhang, S., et al., 2016c. Estimating and source analysis of surface PM2.5 concentration in the Beijing–Tianjin–Hebei region based on MODIS data and air trajectories. Int. J. Remote Sens. 37, 4799-4817. https://doi.org/10.1080/01431161.2016.1220031.

Witten, I., Frank, E., Hall, M.A., et al., 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Wu, D., Xu, Y., Zhang, S., 2015. Will joint regional air pollution control be more cost-effective? An empirical study of China's Beijing–Tianjin–Hebei region. J. Environ. Manage. 149, 27-36. https://doi.org/10.1016/j.jenvman.2014.09.032.

Wu, R., 2014. Seasonal dependence of factors of year-to-year variations in South China AOD and Hong Kong air quality. Int. J. Climatol. 34, 3204-3220. https://doi.org/10.1002/joc.3905.

Wu, X., Ding, Y., Zhou, S., et al., 2018. Temporal characteristic and source analysis of PM2.5 in the most polluted city agglomeration of China. Atmos. Pollut. Res. 9, 1221-1230. https://doi.org/10.1016/j.apr.2018.05.008.

Yan, D., Lei, Y., Shi, Y., et al., 2018. Evolution of the spatiotemporal pattern of PM2.5 concentrations in China – A case study from the Beijing-Tianjin-Hebei region. Atmos. Environ. 183, 225-233. https://doi.org/10.1016/j.atmosenv.2018.03.041.

Yang, G., Huang, J., Li, X., 2018. Mining sequential patterns of PM2.5 pollution in three zones in China. J. Cleaner Prod. 170, 388-398. https://doi.org/10.1016/j.jclepro.2017.09.162.

Yao, X., Ge, B., Yang, W., et al., 2020. Affinity zone identification approach for joint control of PM2.5 pollution over China. Environ. Pollut. 265, 115086. https://doi.org/10.1016/j.envpol.2020.115086.

Ye, W.-F., Ma, Z.-Y., Ha, X.-Z., 2018. Spatial-temporal patterns of PM2.5 concentrations for 338 Chinese cities. Sci. Total Environ. 631-632, 524-533. https://doi.org/10.1016/j.scitotenv.2018.03.057.

Yu, Y., Xu, H., Jiang, Y., et al., 2021. A modeling study of PM2.5 transboundary transport during a winter severe haze episode in southern Yangtze River Delta, China. Atmos. Res. 248, 105159. https://doi.org/10.1016/j.atmosres.2020.105159.

Zhang, L., Yang, G., Li, X., 2020. Mining sequential patterns of PM2.5 pollution between 338 cities in China. J. Environ. Manage. 262, 110341. https://doi.org/10.1016/j.jenvman.2020.110341.

Zhang, N.-N., Ma, F., Qin, C.-B., et al., 2018a. Spatiotemporal trends in PM2.5 levels from 2013 to 2017 and regional demarcations for joint prevention and control of atmospheric pollution in China. Chemosphere. 210, 1176-1184. https://doi.org/10.1016/j.chemosphere.2018.07.142.

Zhang, Q., He, K., Huo, H., 2012. Cleaning China's air. Nature. 484, 161-162. https://doi.org/10.1038/484161a.

Zhang, X., Shi, M., Li, Y., et al., 2018b. Correlating PM2.5 concentrations with air pollutant emissions: A longitudinal study of the Beijing-Tianjin-Hebei region. J. Cleaner Prod. 179, 103-113. https://doi.org/10.1016/j.jclepro.2018.01.072.

Zhao, X.J., Zhao, P.S., Xu, J., et al., 2013. Analysis of a winter regional haze event and its formation mechanism in the North China Plain. Atmos. Chem. Phys. 13, 5685-5696. https://doi.org/10.5194/acp-13-5685-2013.

Zong, Z., Wang, X., Tian, C., et al., 2018. PMF and PSCF based source apportionment of PM2.5 at a regional background site in North China. Atmos. Res. 203, 207-215. https://doi.org/10.1016/j.atmosres.2017.12.013.

Zou, Q., Shi, J., 2020. The heterogeneous effect of socioeconomic driving factors on PM2.5 in China’s 30 province-level administrative regions: Evidence from Bayesian hierarchical spatial quantile regression. Environ. Pollut. 264, 114690. https://doi.org/10.1016/j.envpol.2020.114690.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: