RESEARCH PAPER

Association discovery and outlier detection of air pollution emissions from industrial enterprises driven by big data

Zhen Peng[1†], Yunxiao Zhang[2], Yunchong Wang[2], Tianle Tang[2]

[1]China University of Geosciences Beijing, School of Economics and Management, Beijing 100038, China

[2]Beijing Institute of Petrochemical Technology, School of Economics and Management, Beijing 102617, China

Keywords: Air Pollution Emissions of Enterprises; Outlier detection based on clustering; Association Rule Mining; Grey Association Analysis; Big data

Citation: Peng, Z., Zhang, Y.X., Wang, Y.C., Tang, T.L.: Association discovery and outlier detection of air pollution emissions from industrial enterprises driven by big data. Data Intelligence 5(2), 438-456 (2023). doi: https://doi.org/10.1162/dint_a_00205
Submitted: December 5, 2022; Revised: February 2, 2023; Accepted: February 15, 2023

ABSTRACT

Air pollution is a major issue related to national economy and people's livelihood. At present, the researches on air pollution mostly focus on the pollutant emissions in a specific industry or region as a whole, and is a lack of attention to enterprise pollutant emissions from the micro level. Limited by the amount and time granularity of data from enterprises, enterprise pollutant emissions are still understudied. Driven by big data of air pollution emissions of industrial enterprises monitored in Beijing-Tianjin-Hebei, the data mining of enterprises pollution emissions is carried out in the paper, including the association analysis between different features based on grey association, the association mining between different data based on association rule and the outlier detection based on clustering. The results show that: (1) The industries affecting NOx and SO2 mainly are electric power, heat production and supply industry, metal smelting and processing industries in Beijing-Tianjin-Hebei; (2) These districts nearby Hengshui and Shijiazhuang city in Hebei province form strong association rules; (3) The industrial enterprises in Beijing-Tianjin-Hebei are divided into six clusters, of which three categories belong to outliers with excessive emissions of total VOCs, PM and NH3 respectively.

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

[†] Corresponding author: Zhen Peng, China University of Geosciences Beijing (E-mail: pengzhen@cugb.edu.cn; ORCID: 0000-0003-3907-7200).

© 2023 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

1. INTRODUCTION

Air pollution is a major environmental problem that seriously threatens human life, health, production and living. The World Health Organization and the United Nations Environment Organization state that air pollution has become an unavoidable fact of life in cities around the world.[①] Over the past ten years, along with rapid economic development and urbanization, the intensification of industrial development and energy consumption have increased greatly, and air environment situation is particularly severe. According to the Bulletin of China's Ecological Situation released by the Ministry of Ecology and Environment, the average concentration of PM2.5 in Beijing-Tianjin-Hebei and its surrounding areas in 2019 was about 1.6 times higher than China's second-level standard for atmospheric environment, and heavy atmospheric pollution in regional areas occurred from time to time. Air pollution emissions from industrial enterprises have become the main influencing factors of regional air pollution, among which SO_2, NO_x, etc. are the important sources of the air pollution [1, 2].

In 2013, China introduced the 'Atmosphere Ten' Five-Year Action Plan to comprehensively launch the tough battle of pollution prevention and control focusing on tackling haze. In 2018, the Three-Year Action Plan of "Blue Sky Defense War" was launched successively, and a decisive battle against PM2.5 is underway in key areas such as Beijing-Tianjin-Hebei. Then, The State Council issued the "Opinions on Deepening the Battle against Pollution" in November 2021. It has set a goal of continuously improving the ecological environment and decreasing the total emission of major pollutants by 2025. It is clear that the current air pollution prevention and control is a long-term and persistent battle. Although the Beautiful China has achieved preliminary goals, the air quality of Chinese cities is still not out of the "meteorological influence type" on the whole, and there is still a long way to go in atmospheric governance.

The main cause of air pollution is man-made pollutants, which mainly come from fuel combustion and large-scale production emissions of industrial enterprises. SO_2, NO_x and so on emitted by industrial enterprises are important sources of air pollution, which are the inevitable products of energy consumption. Only the pollution source is located in the most fundamental unit of the enterprise, can air pollution control be provided a reasonable direction for energy efficiency improvement and industrial structure adjustment from the view of micro, which make it possible that urban air quality can be improved fundamentally.

However, the existing researches mainly focus on the characteristics and impacts of air pollution emissions in industries or regions, and there are few researches on the view of enterprise to discover the internal rules of air pollution emissions of industrial enterprises.

With the strengthening of national environmental supervision, the development of Internet of Things technology and the digitalization of enterprise emissions, it is possible to explore the rules of air pollution emissions from industrial enterprises, and make air pollution control more precise and reliable. Therefore, based on big data and information technology, this paper will analyze and discuss the rules of air pollution

[①] https://www.unep.org/news-and-stories/press-release/un-environment-and-world-health-organization-agree-major

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

emission of enterprises and micro-regions. The significance and value of this article is mainly reflected in the following two points:

(1) It provides an important data base for air pollution control in Beijing-Tianjin-Hebei from the micro enterprise level.
(2) It provides information methods and decision-making suggestions for micro-region and enterprise air pollution control from the perspective of data.

The remainder of this paper is organized as follows: Section 2 reviews the relevant literature. Section 3 and Section 4 present theoretical basis, data and methods. Section 5 analyzes the results and discussion. Finally, Section 5 reveals the conclusions and provides policy suggestions.

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

2. LITERATURE REVIEW

In view of the air pollution emissions, most studies have focused on the characteristics of pollutants and the influence of pollutants on the atmosphere in different industries or regions by using different methods.

2.1 Specific Industry-Oriented Air Pollution Analysis

There are many industry-specific studies focusing on the impact of industrial production on air pollution, the assessment of emissions, spatial and temporal distribution of air pollution. The industry involves the construction, transportation, or steel industry accompanied by serious pollution.

For the construction-related industry, considering the cross-regional transfer and capacity replacement of the cement industry, Li et al. used spatial econometric methods to explore the impact of China's cement production on air pollution and the spatial spillover effect [3]; Zhang et al. evaluated the energy saving and emission reduction potential of CO_2 and air pollutants in cement industry of Jiangsu by using the energy model based on Geographic Information System [4]; Qu et al. analyzed the spatial distribution of the potential maximum emissions of VOCs and other pollutants during the whole life cycle of asphalt mixture by establishing appropriate laboratory testing conditions and methods [5]; Giunta integrated the CALPUFF model of air quality and the meteorological CALMET model to study PM_{10} concentration prediction and diffusion near the construction site [6].

In the transportation industry, Guo et al. analyzed the technological development path, emission path and energy structure adjustment path of the transportation industry, as well as the synergistic benefits of pollutants and carbon emission reduction, based on a multi-stage bottom-up mathematical model of automobile development [7]; Masiol et al. analyzed various sources of ground pollution discharged by civil aviation industry and their contributions to ground air quality [8].

For the steel industry, Tang et al. used the comprehensive air quality model (CAMx) to estimate the unit level of pollution sources in China's iron and steel industry and explored their contributions to air quality [9]; Gao et al. analyzed the spatial and temporal characteristics of air emissions from the steel industry and the emission trends of different pollutants based on the air inventory pollutants from China's steel industry [10].

In short, the research on industrial air pollution is mainly oriented towards a specific industry, which indicates that there is a certain relationship between air pollution and industry, but the relationship is still unclear.

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

2.2 Region-Oriented Air Pollution Analysis

For a specific region, the studies focus on the impact of regional industrial structure and regional industrial emissions.

The impact of industrial structure on air quality was usually studied from a national macro perspective. For example, Fan et al. adopted the Dynamic Spatial Durbin Model (DSDM) to find that industrial structure upgrading, clean energy promotion and technological innovation were the driving forces for haze reduction in China [11]. Zheng et al. constructed a panel threshold model between air pollution and industrial structure in China, and found that industrial structure could change the impact of economic development on air pollution [12].

There are studies on the impact of industrial pollution emissions of administrative provinces and cities. Based on the pollution panel data of the Beijing-Tianjin-Hebei region from 2013 to 2017, Meng et al. used Stochastic Impacts by Regression on Population, Affluence, and Technology (STIRPAT) to fit the relationship between pollution emission levels and relevant socio-economic indicators, and found that industrial exhaust emissions played a decisive role in the level of atmospheric pollution [13]. Sun et al. used Thermal Anomaly Radiative Power (CTRP) to evaluate the spatio-temporal pattern of industrial pollution emission and its impact on air quality in Beijing-Tianjin-Hebei region from 2012 to 2018. They observed that the spatial distribution of CTRP was unbalanced in the Beijing-Tianjin-Hebei region, and Tangshan, Tianjin, Handan, Xingtai and Shijiazhuang were high-density pollution areas, the CTRP of Tangshan and Handan was higher than that of other cities, and the CTRP was positively correlated with industrial energy consumption [2].

It can be seen that the industrial pollution emissions are the main sources affecting regional air quality and all these regions are provinces or cities. The certain correlation between different micro-regions needs further study.

2.3 Industrial Enterprises-Oriented Air Pollution Analysis

The research on atmospheric emissions from enterprises has been emerging. Wu et al. analysed the correlation between the power consumption of industrial enterprises and the total industrial output and between the total industrial output and the direct emissions of major pollutants, estimated and predicted the direct pollution emissions of industrial enterprises using power grid big data [14]. In addition, Xiao et al.

constructed three spatial emission characteristic indexes of Industrial Pollutant Emission Intensity (IPEI), Industrial Pollution Concentration Emission Intensity (IPCEI) and Density of Waste Gas Monitoring Enterprise (DWGME) based on the emission monitoring data of 37.123 million enterprises in 31 cities in China [15]. And based on the data the correlation degrees between different pollutants and between cities and industrial air pollution in Beijing-Tianjin-Hebei are found.

In summary, limited by the availability, amount or time granularity of enterprise data, there are a few researches on atmospheric emission of enterprises, which are not detailed enough. Therefore, it is necessary to deeply study the rules of enterprise pollution emission, districts or counties, pollutants and industries from enterprise emission data.

Taking this as the entry point, this paper collects the unstructured data of atmospheric emissions of monitored enterprises in Beijing-Tianjin-Hebei in the past seven years. On the basis of data processing, the information of industrial enterprises and their air pollution is integrated, clustered, and associated and so on. The relationship between pollutants and their industries, as well as the relationship between districts or counties region is analyzed. The abnormal clusters are mined, which seriously impact on air quality. It not only has important reference significance for the emission reduction of enterprises in Beijing-Tianjin-Hebei, but also provides research approach for other regional pollution, which has important reference and practical significance.

3. THEORETICAL BASIS

The theoretical basis of this research is the theory of corporate social responsibility, air basin theory and environmental informatics.

The theory of corporate social responsibility focuses that enterprise is the main body of emissions and should bear the social responsibility of environment [16]. The air basin theory points out that the atmosphere is a whole and there are no boundaries that prevent the flow of air. However, pollutants discharged to the atmosphere from a local source do not immediately mix uniformly around the globe and generally only pollute the air in local areas [17]. The environmental informatics means that the collection and effective processing of environmental information is the basis of environmental management, planning and emergency plan decision-making [18].

Therefore, from the perspective of data, it is necessary and of great significance to study the rules of pollution emissions in enterprises and different micro-regions through the collection, processing and mining of enterprise data for air pollution decision-making and management of regions or enterprises.

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

4. DATA AND METHODS

4.1 Research Framework

This paper constructs an association discovery and outlier detection research framework, as shown in Figure 1. The specific steps are as follows:

(1) To acquire and pre-process the big data of air pollution of industrial enterprises. It mainly includes two kinds of data. One is the characteristic data of the enterprise itself, and the other is the data of the enterprise's air pollution emissions. Data pre-processing refers to unstructured processing, integration, unification, classification, and standardization, etc.

(2) To conduct data mining on the processed data. It mainly includes outlier detection based on clustering, association mining between data and between features, and evaluation.

(3) To analyse the results of (2). It includes the relationship analysis between air pollution and the industries, the relationship analysis between pollution districts or counties, the abnormal analysis of enterprise air pollution emissions, and put forward control measures and suggestions.

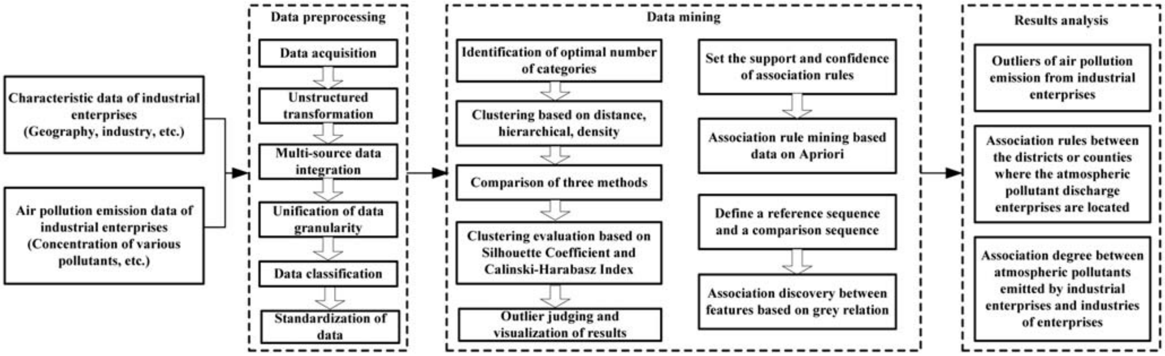

Characteristic data of industrial enterprises (Geography, industry, etc.)
Air pollution emission data of industrial enterprises (Concentration of various pollutants, etc.)
Data preprocessing
Data acquisition
Unstructured transformation
Multi-source data integration
Unification of data granularity
Data classification
Standardization of data
Data mining
Identification of optimal number of categories
Clustering based on distance, hierarchical, density
Comparison of three methods
Clustering evaluation based on Silhouette Coefficient and Calinski-Harabasz Index
Outlier judging and visualization of results
Set the support and confidence of association rules
Association rule mining based data on Apriori
Define a reference sequence and a comparison sequence
Association discovery between features based on grey relation
Results analysis
Outliers of air pollution emission from industrial enterprises
Association rules between the districts or counties where the atmospheric pollutant discharge enterprises are located
Association degree between atmospheric pollutants emitted by industrial enterprises and industries of enterprises

Figure 1. Research framework.

4.2 Data Sources and Data Processing

The data source is the monitoring information disclosure platform of ecological environment departments and government, which includes Beijing Municipal Bureau of ecological environment, Tianjin Ecological Environment Bureau, Hebei Provincial Department of ecological environment, National pollution source monitoring information management and sharing platform.

From the above official platforms, the emission data of monitored enterprises in Beijing-Tianjin-Hebei from 2013 to 2019 were collected and collated. The main types of data are Web text data, PDF data and other unstructured forms. The collected data includes 35 enterprises in Beijing, 52 enterprises in Tianjin, and 1349 enterprises in Hebei Province, totaling nearly 300,000 pieces. The data obtained include two categories of documents. One is the description of the characteristics of the industrial enterprise itself including the enterprise name, industry, region, and county name, etc. The other is pollution data. The fields

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

of pollutant discharge data include enterprise name, monitored year, monitored quarter, monitored date, pollution type, monitored point, emission concentration, emission unit, etc.

Then, the data are cleaned and processed through the following steps.

(1) Unstructured transformation. The government publishes the pollutant discharge monitored data of enterprises in PDF format. Since it is difficult to extract information automatically from PDF files, it is converted to CSV format by some technical means.

(2) Multi-source data integration. It is the integration of the characteristic data of the enterprise and the air pollution emission data of the enterprise. The integration of the two is based on the same enterprise name. However, there are some cases of enterprises changing their names. Whether an enterprise has changed its name and industry can be queried through Tianyan website (https://www.tianyancha.com).

(3) Unity of feature names and merging of similar features. The unification of feature names is to standardize the features with different names but the same meaning, such as ammonia, ammonia gas and ammonia (escape), which are uniformly named ammonia. Jinghai County and Jinghai District are unified as Jinghai District according to the latest administrative planning. Some pollutants are merged into one category. For example, Volatile organic compounds include TVOC, benzene, benzene series, etc., and heavy metals and their compounds include lead, mercury, chromium, etc.

(4) Unification of data granularity. At present, there are semi-annual emission data, quarterly emission data, and air pollution emission data of enterprises with the specific time. If we unify these data into semi-annual data, the amount of data is too few to data mining. If we unify these data into data with specific time, the accuracy of the data will be greatly reduced. So, the "middle" principle is the best way to deal with them. In order to facilitate the subsequent data mining and maintain the accuracy, all data are refined or expanded into quarterly data. The semi-annual emission data are converted to quarterly data according to its change rate of pollution emission. The time-specific data of air pollution are averaged to represent quarterly data.

(5) Data classification. There are 492 enterprise types in the data. For the convenience of research, it can be summarized into the following ten types based on industry tags from enterprise search website and expert knowledge. They are power and heat production and supply, nonmetal mineral manufacturing, electrical machinery and equipment manufacturing, ferrous metal smelting and rolling processing, chemical raw materials and chemical products, metal products, science and technology promotion and application service, crude oil processing and petroleum products manufacturing, nonferrous metal smelting and rolling processing, other manufacturing and processing.

(6) Standardization of data. Due to the inconsistent of different magnitude, the differences between values of different features may be large. In order to eliminate the impact of different magnitude, it is necessary to standardize data into the same interval.

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

4.3 Grey Association Mining Between Different Features

Grey correlation analysis is a method used to quantitatively describe the degree of correlation between different features according to the similarity between their development trends [19, 20]. In the process of system development, if the change trend of two features is consistent, that is, the degree of synchronous change is high, the degree of correlation between them is high.

In this study, industry of enterprise and pollutants emitted by enterprise are different features, so grey correlation analysis can be used for the association analysis between the industry and one pollutant according to the data change trend of these two features. The data of the industry refer to the average concentration of different pollutants in the industry at different times, which are taken as the reference sequence $\boldsymbol{a}_0$. The average value of one pollutant emission concentration at the different times are taken as the comparison sequence $\boldsymbol{a}_i$.

The correlation coefficient between the industry and the pollutant at the k^{th} time can be got according to Eq. (1).

$$\delta_i(k) = \frac{\min\left|\boldsymbol{a}_0 - \boldsymbol{a}_i\right| + \max\left|\boldsymbol{a}_0 - \boldsymbol{a}_i\right| * \rho}{\left|\boldsymbol{a}_0(k) - \boldsymbol{a}_i(k)\right| + \max\left|\boldsymbol{a}_0 - \boldsymbol{a}_i\right| * \rho}, \tag{1}$$

where $\delta_i(k)$ represents the grey correlation coefficient, ρ is the resolution coefficient, $0 < \rho < 1$.

The comprehensive correlation degree r_i from the k^{th} time to the y^{th} time is calculated according to each correlation coefficient as shown in Eq. (2):

$$r_i = \frac{1}{y-k}\sum_k^y \delta_i(k) \tag{2}$$

The associations between one comparison sequence and the reference sequence are arranged in descending order reflecting the degree of their relevance.

4.4 Association Rule Mining Between Different Data

Association rule mining is usually used to discover the association between different data [21, 22], which is suitable for association discovery between different districts or counties with industrial pollution emissions.

For any $A \rightarrow B$, A is the antecedent, B is the consequent, and A and B are mutually exclusive. It forms an association rule, which means that A will lead to B, if and only if Eq. (3) and (4) are true.

$$sup(A \rightarrow B) = P(A \cup B) = \frac{\sigma(A \cup B)}{N} > min_{sup}, \tag{3}$$

$$conf(A \rightarrow B) = P(B|A) = \frac{P(A \cup B)}{P(A)} = \frac{\sigma(A \cup B)}{\sigma(A)} > min_{conf}, \tag{4}$$

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

where the support of *sup*($A \rightarrow B$) is the probability of $A \cup B$ occurring at the same time, representing the frequency of both occurrences; the confidence of *conf*($A \rightarrow B$) refers to the probability of the occurrence of the consequent B under the condition of the antecedent A. Only when the support is greater than the min_{sup} threshold and the confidence is greater than the min_{conf} threshold, $A \rightarrow B$ can be called an association rule. Moreover, for any rule, the confidence must be greater than the support.

4.5 Outlier Detection Based on Clustering

Outlier detection based on clustering uses outlier factor to determine outliers on the basis of clustering results [23, 24]. The clustering mainly includes three steps [20]. The first is to determine the optimal number of clusters; the second is to cluster data; the third is to evaluate the clustering results.

Elbow method based on *SSE* (Sum of Square Error) as Eq. (5) is often used to determine the optimal number of clusters. When the number of clusters k is less than the real number of clusters, the increase of k will greatly increase the degree of aggregation in each class, that is, the decline range of *SSE* will be large. After k reaches the real number of clusters, the decline of *SSE* will be slower. Therefore, the relationship between *SSE* value and cluster number k forms an elbow shape, and the k of this elbow is the real cluster number.

$$SSE = \sum_{i=1}^{k} \sum_{u \in C_i} \left(m_i - x_u\right)^2, \tag{5}$$

where k represents the number of clusters, C_i represents the i^{th} cluster, u represents a sample point in C_i, x_u represents the value of the sample u, and m_i represents the centroid of the cluster C_i. How to determine the centroid is the key. The best centroid is to minimize the partial derivative of *SSE* with respect to the centroid m_i as Eq. (6). The partial derivative is set to 0 as Eq. (7), where the best center of mass is the mean value of all points in the cluster.

$$\frac{\partial SSE}{\partial m_i} = \frac{\partial}{\partial m_i} \sum_{i=1}^{K} \sum_{u \in C_i} \left(m_i - x_u\right)^2 = \sum_{i=1}^{K} \sum_{u \in C_i} \frac{\partial}{\partial m_i} \left(m_i - x_u\right)^2 = \sum_{u \in C_i} 2\left(m_i - x_u\right), \tag{6}$$

$$\sum_{u \in C_i} 2\left(m_i - x_u\right) = 0 \Rightarrow m_i \mid C_i \mid = \sum_{u \in C_i} x_u \Rightarrow m_i = \frac{1}{\mid C_i \mid} \sum_{u \in C_i} x_u \tag{7}$$

Where $|C_i|$ represents the number of all sample points in C_i.

This paper uses multiple clustering methods (based on distance, hierarchy and density) to cluster industrial enterprises according to their emitted pollutants, compares these clusters and analyzes the characteristics of the clusters.

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

In this study, Silhouette Coefficient and Calinski-Harabasz Index are applied to estimate the clustering effect. $a(p)$ is how similar an object p is to its own cluster. $b(p)$ represents the similarity between an object p and other clusters. The Silhouette Coefficient takes both cohesion $a(p)$ and separation $b(p)$ into account. The value range is [-1, 1] as Eq. (8). The larger the value, the better the clustering effect.

$$S(p) = \frac{b(p) - a(p)}{\max\{a(p), b(p)\}} \tag{8}$$

Calinski-Harabasz Index is another clustering evaluation method, which is much faster than Silhouette Coefficient.

$$CH(k) = \frac{trB(k)/(k-1)}{trW(k)/(n-k)} \tag{9}$$

Where n represents the number of training sample sets, $B(k)$ is the covariance matrix between clusters, $W(k)$ is the covariance matrix of data within clusters, and tr represents the trace of the matrix as Eq. (9). For clustering, smaller within-cluster variance is better, and larger between-cluster variance is better. So the higher Calinski-Harabasz index indicates the better clustering effect.

Finally, the outlier factor $OF1(p)$ of object p is defined as the weighted average of the spacing between all clusters as Eq. (10).

$$OF1(p) = \sum_{i=1}^{k} \frac{|C_i|}{|D|} d(p, C_i) \tag{10}$$

Where $|D|$ is the total amount of samples, $d(p, C_j)$ is the distance between the sample p and the cluster C_i.

When Eq. (11) is satisfied, the object p is an outlier.

$$OF1(p) \geq Ave\_OF + \beta * Dev\_OF \quad (1 \leq \beta \leq 2) \tag{11}$$

where $OF1(p)$ is the outlier factor, Ave_OF is the average, and Dev_OF is the standard deviation of all in dataset D.

5. RESULTS AND DISCUSSION

5.1 Association Between Air Pollution and Industries of Enterprises

Based on the grey correlation model, we study the correlation between ten kinds of industries and the emissions of NO_x and SO_2 pollutants of Beijing-Tianjin-Hebei region. The results are shown in Table 1.

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

Table 1. Association between industries and pollutants.

Industries	NO_x	SO_2
Crude oil processing and petroleum products manufacturing	0.61	0.58
Nonferrous metal smelting and rolling processing	0.69	0.63
Chemical raw materials and chemical products manufacturing	0.64	0.64
Power and heat production and supply	0.69	0.69
Nonmetal mineral manufacturing	0.6	0.62
Science and technology promotion and application service	0.62	0.61
Electrical machinery and electric equipment manufacturing	0.6	0.65
Other manufacturing	0.58	0.65
Ferrous metal smelting and rolling processing	0.69	0.66
Metal products	0.69	0.64

From the association ranking of industries, power and heat production and supply, ferrous metal smelting and rolling processing, metal products and nonferrous metal smelting and rolling processing are high correlation with NO_x emissions, and power and heat production and supply is most closely related to SO_2 emissions in Beijing-Tianjin-Hebei.

Therefore, the power and heat production and supply industry have the highest association degree with NO_x and SO_2. And the association degrees of ferrous metal smelting and rolling processing industries are generally higher. These results indicate that the reduction of emissions from these industries will effectively impact on the overall emission trend of Beijing-Tianjin-Hebei.

5.2 Association Rules of Districts or Counties With the Industry Enterprises

Taking the air pollutant NO_x with the highest emission as the object, based on the association rule mining, in the frequent item sets the high confidence between districts with NO_x emission is shown in Figure 2. Among them, Taocheng District→Hengshui Economic Development Zone, and Jingxing county →Xingtang County form strong association rules.

Taocheng District is the urban area of Hengshui City. Hengshui economic development zone is located in the suburb of Hengshui City, which is geographically close to Taocheng District, indicating that for Hengshui City, the air pollution in the urban area has a great impact on the suburbs. Xingtang County is located in the north of Shijiazhuang city and Jingxing County is located in the west of Shijiazhuang city. Interestingly, the two counties are not adjacent, but there is a strong correlation. The strong correlation between the two counties shows that there is similar NO_x emission trend in the industrial enterprises of the two counties. The possible reason is that it is influenced by other factors, such as meteorological conditions.

In addition, this study finds that the areas with high confidence are Shijiazhuang, Tangshan, which indicates that there is more NO_x emitted by industrial enterprises in these areas, and the frequency and activity of the production in these areas are also relatively high. So for Beijing, we should pay attention to the influence of these industrial enterprises in surrounding areas and strengthen their management.

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

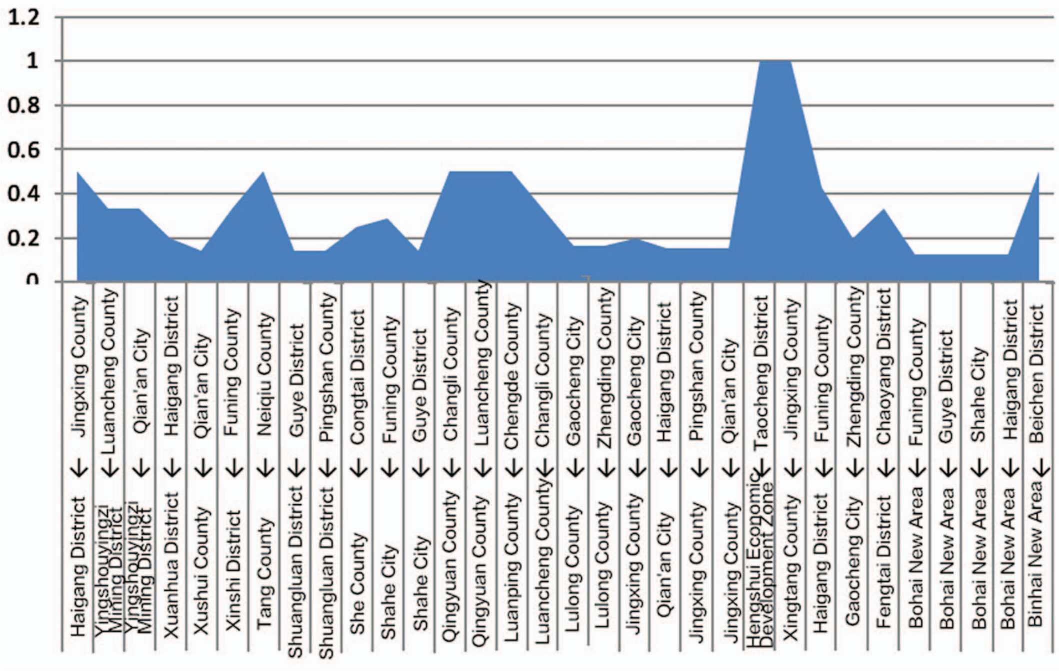


Figure 2. High confidence between districts of NO_x emission.

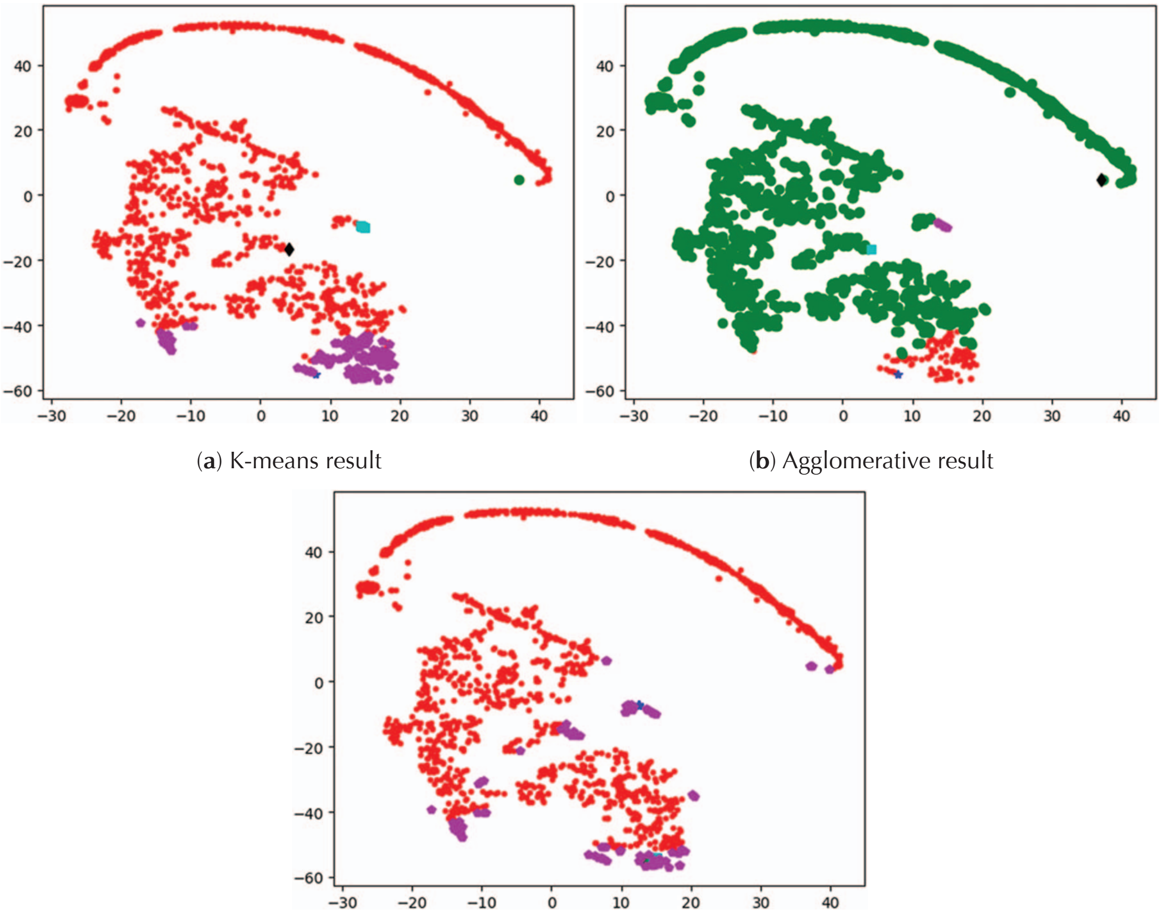
5.3 Outliers of Air Pollution Emission from Industrial Enterprises

Through the statistical analysis of air pollutants emitted by industrial enterprises in Beijing-Tianjin-Hebei, it is found that six kinds of pollutants including NO_x, SO_2, VOCs, PM, heavy metals and their compounds, and NH_3 occur the most frequently in the enterprises.

According to the elbow rule in the clustering principle, it can be judged that the optimal number of clustering for the data set is 6. Taking $c = 6$ as the parameter, three clustering methods are implemented including *K*-means based on distance, Agglomerative based on hierarchy and DBSCAN based on density.

TSNE dimensionality reduction method is used to visualize the clustering results, which reduces the distribution dimension of the original data in high-dimensional space to two-dimensional space, as shown in Figure 3 (a), (b), and (c).

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023



(**a**) K-means result

(**b**) Agglomerative result

(**c**) DBSCAN result

Figure 3. The results of three clustering algorithms.

The evaluation results are shown in Table 2. It can be seen that K-means algorithm has the best clustering results for industrial pollution emissions in Beijing-Tianjin-Hebei.

Table 2. Evaluation results of three clustering methods.

Clustering Method	Silhouette Coefficient	Calinski-Harabasz Index
K-means	0.6725	566.065
Agglomerative	0.6561	358.386
DBSCAN(the best: eps=0.5, min_Samples=3)	0.5195	46.300

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

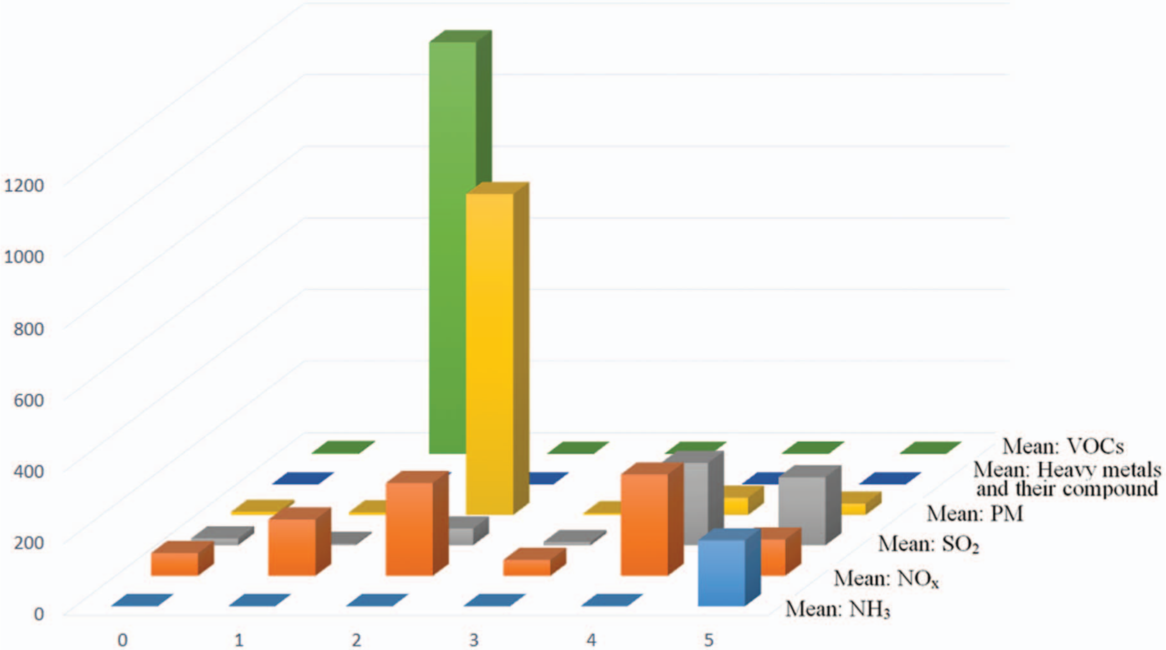


Figure 4. Clustering results of enterprisepollution emissions.

As shown in Figure 4, there are six clusters in industrial enterprise pollution emission based on *K*-means. According to Eq. (11) with β = 1.5, the Cluster 1, Cluster 2 and Cluster 5 are determined as outliers.

Cluster 1 is Hebei ** Energy Technology Development Co., Ltd in Shijiazhuang City, Hebei Province, belonging to abnormal cluster. The enterprise belongs to the crude oil processing and petroleum products manufacturing industry, and the VOCs emission concentration seriously exceeds the standard.

Cluster 2 is Hebei ** bio Power Generation Co., Ltd. in Shijiazhuang City, Hebei Province, which belongs to the power and heat production and supply industry. The emission concentration of NO_x is high and the emission concentration of PM seriously exceeds the standard. It is also an abnormal cluster.

The enterprise in Cluster 5 is Hebei ** Chemical Co., Ltd. in Cangzhou city, Hebei province, which belongs to the manufacturing industry of chemical raw materials and chemical products. The NH_3 emission concentration of the enterprise seriously exceeds the standard and the SO_2 emission concentration is too high.

6. CONCLUSIONS AND SUGGESTIONS

It can be seen that the remarkable achievements have been attained in atmospheric pollution control of Beijing-Tianjin-Hebei in recent years, but it still needs improvement in the following aspects.

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

(1) In Beijing-Tianjin-Hebei, on the whole, the power and heat production and supply industries contribute the most to the main pollutants NO_x and SO_2, followed by smelting and rolling processing industries.

These industrial enterprises need strengthen the desulfurization and denitrification treatment of waste gas, especially in urban heating, accelerate the construction of regional central heating, improve efficiency and reduce the emission of pollutants.

(2) The main air pollutants emitted by all types of enterprises in the production and manufacturing process are NO_x, SO_2, PM and smoke. The industrial enterprises with high association degree are located in Hebei, and the association between counties near Hengshui and Shijiazhuang is greater. These results show that the air pollutant in these places is not spread diffusely so that there exists a serious accumulation.

It is suggested that the geographical distribution of industrial enterprises should be adjusted avoiding the aggregation of enterprises in Hebei Province if possible, especially in Hengshui and Shijiazhuang.

(3) There are a large number of enterprises in Beijing-Tianjin-Hebei, especially in Hengshui, Tangshan, Baoding, Xingtai et al, where there are three outliers. The first outlier is the emission concentration of VOCs. Most of VOCs not only have pungent smell, but also are key substances that produce ozone. The second outlier has high NO_x and PM emission concentration. These two pollutants are the main pollutants in air quality. The NH_3 and SO_2 emission concentration is the third anomaly. NH_3 will react with SO_2 in the air to form granular substances.

The regulatory authorities should strengthen the management of these areas and urge their enterprises to upgrade and transform waste gas treatment facilities. The above three abnormal enterprises should be ordered to stop production.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China [grant number 72271033]; the Beijing Municipal Education Commission and Beijing Natural Science Foundation [grant number KZ202110017025]; the National Undergraduate Innovation and Entrepreneurship Plan Project (2022J00244).

COMPETING INTERESTS

We declared that we have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

AUTHOR CONTRIBUTIONS

Zhen Peng (pengzhen@cugb.edu.cn, 0000-0003-3907-7200): Conceptualization, Methodology, Validation, Writing, Review & editing. Yunxiao Zhang (yunxiaozhang2022@163.com): Conceptualization, Methodology, Validation, Writing. Yunchong Wang (yunchongwang2022@163.com): Data Curation, Methodology, Validation, Writing. Tianle Tang (tangtianle20@163.com): Data Curation, Methodology, Validation.

DATA AVAILABILITY

Please refer to Beijing Municipal Bureau of ecological environment http://sthjj.beijing.gov.cn/, Tianjin Ecological Environment Bureau http://sthj.tj.gov.cn/, Hebei Provincial Department of ecological environment http://hbepb.hebei.gov.cn/, National pollution source monitoring information management and sharing platform http://123.127.175.61:6375/.

REFERENCES

[1] Wang, Y., Wen, Z., Dong, J.: The city-level precision industrial emission reduction management based on enterprise performance evaluation and path design: A case of Changzhi, China. Science of The Total Environment 734, 139350 (2020)

[2] Sun, S., Li, L., W, Z., Gautam, A., Li, J., Zhao, W.: Variation of industrial air pollution emissions based on VIIRS thermal anomaly data. Atmospheric Research 244, 105021 (2020)

[3] Li, M., Zhang, M., Du, C., Chen, Y.: Study on the spatial spillover effects of cement production on air pollution in China. Science of The Total Environment 748, 141421 (2020)

[4] Zhang, S., Ren, H., Zhou, W., Yu, Y., Chen, C.: Assessing air pollution abatement co-benefits of energy efficiency improvement in cement industry: A city level analysis. Journal of Cleaner Production 185, 761–771 (2018)

[5] Qu, S., Fan, S., Wang, G., He, W., Xu, K., Nie, L., Zhao,Y., Zhu, Q., Li, T., Li, G.: Air pollutant emissions from the asphalt industry in Beijing, China. Journal of Environmental Sciences 109, 57–65 (2021)

[6] Giunta, M.: Assessment of the environmental impact of road construction: Modelling and prediction of fine particulate matter emissions. Building and Environment 176, 106865 (2020)

[7] Guo, J., Zeng, Y., Zhu, K., Tan, X.: Vehicle mix evaluation in Beijing's passenger-car sector: From air pollution control perspective. Science of The Total Environment 785, 147264 (2021)

[8] Masiol, M., Harrison, R.M.: Aircraft engine exhaust emissions and other airport-related contributions to ambient air pollution: A review. Atmospheric Environment 95, 409–455 (2014)

[9] Tang, L., Xue, X., Jia, M., Jing, H., Wang, S.: Iron and steel industry emissions and contribution to the air quality in China. Atmospheric Environment 237(8),117668 (2020). https://doi.org/10.1016/j.atmosenv.2020.117668.

[10] Gao, C., Gao, W., Song, K., Na, H., Tian, F., Zhang, S.: Spatial and temporal dynamics of air-pollutant emission inventory of steel industry in China: A bottom-up approach. Resources, Conservation and Recycling 143, 184–200 (2019)

[11] Fan, X., Xu, Y.: Convergence on the haze pollution: City-level evidence from China. Atmospheric Pollution Research 11(6), 141–152 (2020)

[12] Zheng, Y., Peng, J., Xiao, J., Su, P., Li, S.: Industrial structure transformation and provincial heterogeneity characteristics evolution of air pollution: evidence of a threshold effect from China. Atmospheric Pollution Research 11(3), 598–609 (2019)

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

[13] Meng, M., Zhou, J.: Has air pollution emission level in the Beijing-Tianjin-Hebei region peaked? A panel data analysis. Ecological Indicators 119, 106875 (2020)

[14] Wu, L., Zhou, Y., Chen, H., Yang Z., University F.: Emission Characteristics of Industrial Air Pollution by Using Smart-Grid Big Data. Chinese Journal of Environment Management 8(04), 37–42 (2016)

[15] Xiao, C., Chang, M., Guo, P., Yuan, M., Xu, C., Song, X., Xiong X., Li, Y., Li, Z.: Characteristics analysis of industrial atmospheric emission sources in Beijing-Tianjin-Hebei and Surrounding Areas using data mining and statistics on different time scales. Atmospheric Pollution Research 11(1), 11–26 (2020)

[16] Wang, W., Guo H.: A Summary of the Research on the Environmental Responsibility of Chinese Enterprises. Economic Research Guide 34, 21–23 (2021)

[17] Wang, J., Ning, M., Sun, Y.: Study on Theory and Methodology about Joint Prevention and Control of Regional Air Pollution. Environment and Sustainable Development 5, 5–10 (2012)

[18] James, E.F., Jeff, D.: Environmental Informatics. Annual Review of Environment and Resources 37, 449–472 (2012)

[19] Yin, M.S.: Fifteen years of grey system theory research: A historical review and bibliometric analysis. Expert Systems with Applications 40(7), 2767–2775 (2013)

[20] Zheng, S., Shi, J., Luo, D.: A Grey correlational analysis method based on cross-correlation time-delay. The Journal of Grey System 32, 104–118 (2020)

[21] Altay, E.V., Alatas, B: Differential Evolution and Sine Cosine Algorithm Based Novel Hybrid Multi-Objective Approaches for Numerical Association Rule Mining. Information Sciences 554(10), 198–221 (2020)

[22] Huang, M.J., Sung, H.S., Hsieh, T.J., Wu, M.C., Chung, S.H.: Applying data-mining techniques for discovering association rules. Soft Computing 24, 8069–8075 (2019)

[23] Krleza, D., Vrdoljak, B., Brcic, M.: Statistical hierarchical clustering algorithm for outlier detection in evolving data streams. Machine Learning 110(1), 139–184 (2021)

[24] Balusamy, B., Abirami, R.N., Kadry, S., Gandomi, A.H.: Big Data: Concepts, Technology, and Architecture. John Wiley & Sons, Inc. (2021)

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023

AUTHOR BIOGRAPHY

Zhen Peng received respectively her bachelor degree and master degree in computer science and computer application technology from Shandong University, China, her Ph.D in computer application technology from Beijing University of Science and Technology, China. Now she is a professor in department of management science and engineering, China University of Geosciences Beijing. Her research interests include data mining, game theory and environmental information.

E-mail: pengzhen@cugb.edu.cn, zhen_peng1981@163.com

ORCID: 0000-0003-3907-7200

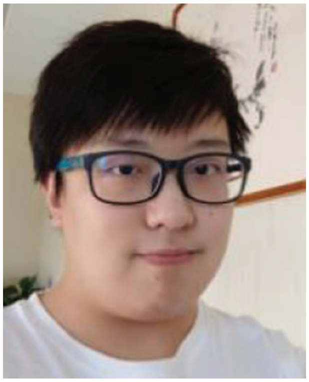
Yunxiao Zhang has obtained engineering degree of engineering management from Shandong Jianzhu University in 2019. Starting in 2021, she is currently studying at school of economics and management of Beijing Petrochemical Institute and specializes. Her research interests include environmental governance and carbon emissions.

E-mail: yunxiaozhang2022@163.com

Yunchong Wang is an undergraduate majoring in management and application of big data at Beijing Institute of Petrochemical Technology. His interests include big data, environmental information, IT, AI.

E-mail: 13241840707@163.com, yunchongwang2022@163.com

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023



Tianle Tang received his bachelor degree in information management and information system from Beijing Institute of Petrochemical Technology. His research interests include big data processing and analysis, data. And he is good at writing programs.
E-mail: tangtianle20@163.com

Downloaded from http://direct.mit.edu/dint/article-pdf/5/2/438/2089779/dint_a_00205.pdf by guest on 07 September 2023