Eurasia Journal of Science and Technology
ISSN: 2663-1016
DOI: https://doi.org/10.61784/ejst3083

AIR QUALITY ANALYSIS BASED ON OUTLIER DETECTION ALGORITHMS

QiZhi Zhang, Nan Jiang*
Northeast Forestry University, Harbin 150040, Heilongjiang, China.
Corresponding Author: Nan Jiang, Email: jngrace@nefu.edu.cn

Abstract: This study employs an outlier detection algorithm based on the distance metric of the Two-Step clustering algorithm to analyze potential untrustworthy data in pollutant concentration records across the region of Beijing-Tianjin-Hebei. By calculating anomaly indices through designated formulas and evaluating variable contribution rates, abnormal data points were identified for each monitoring area. Subsequent analysis of these anomalies provides substantial evidence supporting the existence of unreliable data within the dataset.
Keywords: Data mining; Two-Step Clustering algorithm; Outlier detection algorithm; Air quality analysis

1 INTRODUCTION

In recent years, with the deepening industrialization and urbanization in China, air quality issues arising from economic growth and population concentration have increasingly become a major concern for the public, the government, and relevant authorities[1]. Relevant research indicates that there is a positive correlation between social development and air quality. Whether viewed from the perspective of residents' health or social production, air quality is closely linked to people's lives. Good air quality not only beneficial benefit for health but also enhances physical and mental well-being, enabling individuals to engage more efficiently in life and work[2-3]. In an invisible way, it indirectly promotes comprehensive social development.

The Air Pollution Index (API) is a method that converts the concentrations of several commonly measured air pollutants into numerical values and represents the air pollution status in a graded form. It is an indicator method that reflects the quality of air, and the results of this method are very simple and intuitive. The air quality pollution index used in China can be divided into six levels. If API≤50, it indicates that the air quality in this area is excellent; if 50<API≤100, it indicates that the air quality in this area is good; if 100<API≤150, it indicates that the air quality in this area is generally average, with slight pollution; if 150<API≤200, it indicates that the air quality in this area is relatively poor, with moderate pollution; if 200<API≤300, it indicates that the air quality in this area is extremely poor, with severe pollution; if API>300, it indicates that the air quality in this area is very poor, with serious pollution. The national air quality classification standards can be divided into three levels. The first-level standards are implemented in Class I areas (natural reserves, scenic spots, forest and grass areas, etc.); the second-level standards are implemented in Class II areas (urban residential areas, mixed commercial areas, agricultural areas, general industrial areas, etc.); and the third-level standards are implemented in Class III areas (special industrial areas). The assessment of regional air quality mainly relies on the concentrations of pollutants in the air of the region, such as inhalable particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2)[4].

The evaluation of excellent air quality comes from the results of data collection, but the collected data presents some anomalies, which are caused by various subjective and objective reasons. For example, data anomalies caused by the collection equipments and those caused by human intervention[5]. The region of Beijing-Tianjin-Hebei, Yangtze River Delta region, and Pearl River Delta are among China's five major urban agglomerations, playing a pivotal role in the country's socio-economic development. In recent years, the air quality in these regions has received widespread attention. Based on some air pollution data from the region of Beijing-Tianjin-Hebei, an outlier detection algorithm based on the Two-Step clustering method to measure distance was adopted. By establishing corresponding mathematical models or evaluation indicators, the authenticity of air quality data was analyzed to determine whether there were any data anomalies, thus evaluating the quality of air. The authenticity of air quality was evaluated by establishing corresponding mathematical models or evaluation indicators.

2 OUTLIER DETECTION ALGORITHM BASED ON TWO-STEP CLUSTERING METHOD

Hierarchical clustering method forms a clustering tree according to the data set. According to whether the hierarchical decomposition is bottom-up or top-down, hierarchical clustering methods can be further divided into condensed and split hierarchical clustering. Agglomerative hierarchical clustering is a bottom-up strategy, first takes each sample as a cluster, and then combines these atomic clusters into larger and larger clusters until all samples are in one cluster or a certain termination condition is met. Most hierarchical clustering methods belong to this category, but they are different in the definition of similarity between clusters[6]. The basic method is: given n clusters to be clustered, calculate their $n \times n$ distance matrix, find the closest two classes and merge them into one class, so the total number of classes is less than one, then recalculate the distance between the new class and all old classes, select the closest two classes to merge. This process iterates until either all clusters merge into a single class, or a predefined stopping condition is satisfied.

© By the Author(s) 2025, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

The Two-Step clustering method employs a log-likelihood distance metric to handle both continuous and categorical attributes. Two-Step clustering method takes each entry of each leaf node as an atomic cluster, and uses the clustering method of condensation to continuously merge these clusters[7-9]. First, calculate the log likelihood value, assuming that the values of continuous attributes are normally distributed, and the values of discrete attributes are multinomial distribution. Attributes are assumed to be mutually independent. The distance between cluster i and cluster j is:

$$d(i,j)=\xi_i+\xi_j-\xi_{<i,j>}, \tag{1}$$

where $\xi_v=-N_v\left(\sum_{k=1}^{K^A}\frac{1}{2}\log(\hat{\sigma}_k^2+\hat{\sigma}_{vk}^2)+\sum_{k=1}^{K^B}\hat{E}_{vk}\right), \hat{E}_{vk}=-\sum_{l=1}^{L_k}\frac{N_{vkl}}{N_v}\log\frac{N_{vkl}}{N_v}.$

The Two-Step algorithm adopts a "two-phase" methodology[10].

(1) Phase 1: compute preliminary estimates of cluster numbers. For each candidate clustering scheme, calculate its Bayesian Information Criterion (BIC). For example, the BIC for a partition with J clusters is given by:

$$\mathrm{BIC}(J)=-2\sum_{j=1}^{J}\xi_j+m_J\log(N) \tag{2}$$

Let $d\mathrm{BIC}(J)=\mathrm{BIC}(J)-\mathrm{BIC}(J+1)$, which represents the difference between the scheme of J clusters and the scheme of $J+1$ clusters. Then, calculate $R_1(J)=\frac{d\mathrm{BIC}(J)}{d\mathrm{BIC}(1)}$, which indicates the change degree of $d\mathrm{BIC}(J)$ relative to $d\mathrm{BIC}(1)$.

If $d\mathrm{BIC}(1)<0$, then the order of the clustering tree is set to 1. Otherwise, the preliminary estimate value k of the clustering tree is the minimum J that makes $R_1(J)<0.04$ hold.

(2) Phase 2: determine the number of clusters. According to value k from phase 1, calculate $R_2(k)=\frac{d_{\min}(C_k)}{d_{\min}(C_{k+1})}$.

Where C_k represents a partition scheme with a cluster number of k, $d_{\min}(C_k)$ represents the distance between the two clusters with the smallest distance in the scheme. Similarly C_{k+1}.

Then recalculate the Variable Deviation Index (VDI) and Group Deviation Index (GDI). As the attribute variable X_k is a continuous variable,

$$d_k(h,s)=\frac{1}{2}[-N_h\log(\Delta_k+\hat{\sigma}_{hk}^2)-\log(\Delta_k)], \tag{3}$$

As the attribute variable X_k is a discrete variable,

$$d_k(h,s)=-N_h\hat{E}_{hk}+(N_h+1)\hat{E}_{<h\,s>k}, \tag{4}$$

After calculating the Variable Deviation Index $\{\mathrm{VDI}_k, k=1,2,\cdots,K+1\}$ of all attribute variables, the Group Deviation Index (GDI) of the sample can be calculated,

$$\mathrm{GDI}=d(h,s)=\sum_{k=1}^{K^A+K^B}d_k(h,s). \tag{5}$$

Next, calculate the Anomaly Index (AI) and Variable Contribution (VC). The anomaly index of a sample s is used to measure how abnormal the sample is compared with other samples in cluster h. It is the ratio of the GDI of sample s divided by the average GDI of all samples in cluster h,

$$\mathrm{AnomalyIndex}=\frac{\mathrm{GDI}_s}{\mathrm{mean}(\mathrm{GDI}_h)}, \tag{6}$$

A higher Anomaly Index indicates greater sample anomaly. Generally, the observation value with the anomaly index value less than 1 or even less than 1.5 will not be regarded as an anomaly value, because the deviation is the same as the average value or just a little larger. However, the observation value with index value greater than 2 may be abnormal observation value, because the deviation is at least twice the average value.

The variable contribution rate of an attribute variable X_k is for a single sample s. It is the ratio of VDI_k of attribute variable X_k in sample s divided by GDI, which is,

$$\mathrm{VCM}_k=\frac{\mathrm{VDI}_k}{\mathrm{GDI}_s}. \tag{7}$$

Variable Contribution Rate measures the degree to which an attribute contributes to a sample's anomaly. A higher variable contribution rate indicates the attribute exerts stronger influence in making the sample anomalous. For samples with high anomaly indices, this metric identifies which specific attributes drive their outlier status.

3 APPLICATION AND ANALYSIS OF ALGORITHM IN AIR QUALITY DATA

This study applies an outlier detection algorithm based on distance measurement using the Two-Step clustering method to analyze the authenticity of pollutant concentration data in the Beijing-Tianjin-Hebei region, with a threshold set at 2.

The data was sourced from the official website of the Ministry of Environmental Protection of the People's Republic of China. The Beijing-Tianjin-Hebei regional data was processed through Two-Step clustering, with the clustering results presented in Table 1.

Table 1 AQI Data Clustering of Beijing-Tianjin-Hebei

Beijing-Tianjin-Hebei						
	Peer group 1 records:1795 Exception records: 103		Peer group 1 records: 2066 Exception records:140		Peer group 1 records: 2390 Exception records:125	
contribution	records	average index	records	average index	records	average index
PM2.5	16	0.153	73	0.317	89	0.260
PM10	41	0.179	90	0.269	91	0.258
NO_2	88	0.333	96	0.247	73	0.225
SO_2	94	0.483	83	0.427	52	0.279
CO	70	0.179	78	0.290	70	0.290

From Table 1, it can be observed that the Two-Step clustering algorithm partitioned the Beijing-Tianjin-Hebei regional data into three equivalent groups, comprising a total of 6,251 records. Among these, 368 records were identified as anomalies. For analytical purposes, we selected 20 representative records from these 368 anomalies (the complete dataset is not provided in the paper due to space constraints), with the sampled 20 records displayed in Table 2.

Table 2 Anomalous Data in the Beijing-Tianjin-Hebei Region

PM2.5	PM10	CO	NO_2	SO_2	OAI	OPG	OF1	OFI1	OF2	OFI2	OF3	OFI3
6	34	0.4	6	2	2.31811	1	NO_2	0.48866	SO_2	0.223552	CO	0.11232
5	31	0.3	8	4	2.24192	1	NO_2	0.43011	SO_2	0.202844	CO	0.16164
31	77	0.5	57	8	2.03586	1	NO_2	0.75396	SO_2	0.168452	CO	0.05448
189	111	1.3	39	4	2.26511	2	PM2.5	0.72764	SO_2	0.173162	PM10	0.05600
186	118	1.3	39	4	2.14657	2	PM2.5	0.72726	SO_2	0.182725	PM10	0.04520
202	154	1.4	35	8	2.53839	2	PM2.5	0.80889	SO_2	0.120479	NO_2	0.03625
25	74	0.7	59	5	2.25049	1	NO_2	0.77757	SO2	0.188811	CO	0.01415
193	72	1.6	35	3	2.82689	2	PM2.5	0.62543	SO_2	0.147051	PM10	0.14586
24	71	0.5	60	4	2.46768	1	NO_2	0.75496	SO_2	0.184286	CO	0.04494
72	53	1	50	3	2.17831	1	NO_2	0.41350	PM2.5	0.317269	SO_2	0.22304
38	91	0.6	66	4	3.32109	1	NO_2	0.78891	SO_2	0.136931	PM10	0.04306
50	88	0.7	61	3	2.77356	1	NO_2	0.71382	SO_2	0.175175	PM2.5	0.05791
190	126	1.8	60	6	2.78799	2	PM2.5	0.60177	NO_2	0.155822	SO_2	0.12465
43	90	0.7	61	7	2.58979	1	NO_2	0.76447	SO_2	0.142486	PM10	0.05152
104	160	1.3	88	12	2.65306	2	NO_2	0.83590	SO_2	0.087206	PM2.5	0.04418
5	41	0.2	10	2	2.16899	1	NO_2	0.37381	SO_2	0.238921	CO	0.22249

97	120	2	79	39	2.13157	2	NO_2	0.69612	CO	0.205764	PM10	0.04209
108	126	1.9	78	47	2.05970	2	NO_2	0.68537	CO	0.165766	PM2.5	0.07072
138	20	1.5	40	8	2.06282	2	PM10	0.53686	PM2.5	0.240635	SO_2	0.14825
41	276	0.4	14	4	2.97071	2	PM10	0.39035	NO_2	0.241590	CO	0.15558

Note: O-AnonalyIndex is abbreviated as OAI, O-PeerGroup is abbreviated as OPG, O-Field-1 is abbreviated as OF1, O-FieldImpact-1 is abbreviated as OFI1, O-Field-2 is abbreviated as OF2, O-FieldImpact-2 is abbreviated as OFI2, O-Field-3 is abbreviated as OF3, and O-FieldImpact-3 is abbreviated as OFI3.

Taking the first row of data as an example. The abnormal index O-AnomalyIndex equals 2.31811, which is greater than the abnormal index threshold 2 set by us, so we consider this data to be anomalous Data. We can also see from table 2 that, O-PeerGroup=1, which indicates that the record belongs to peer group 1. In Table 2, we rank the influence factors of the variable deviation index in descending order, so the attribute that has the greatest impact on the variable deviation index in the record is NO_2. The O-FieldImpact equals 0.48866, which is greater than the average index of NO_2 (0.333) in peer group, thus qualifying as anomalous data. Then consider SO_2, whose variable deviation index O-FieldImpac value is 0.22355, which is less than the average index of SO_2 (0.483) in peer group 1, indicating that SO_2 does not affect the variable deviation index and belongs to the normal attribute range in this sample. Similarly, CO is also a normal attribute in this sample.

In summary, the algorithmic model and numerical simulation experiments demonstrate the existence of non-authentic data phenomena.

4 CONCLUSION

Air quality issues are inextricably linked to social production and daily life. Since China's implementation of air quality monitoring, recurring reports have revealed a concerning paradox: monitored data shows improvement while actual environmental conditions continue to deteriorate—a clear discrepancy that warrants serious scrutiny and reflection. The dangers of artificial data manipulation are evident and severe, as it compromises the early-warning function of environmental monitoring systems, and infringes upon citizens' right to information. When severe air pollution occurs without corresponding official alerts, citizens face the gravest risks—being unknowingly exposed to hazardous conditions while deprived of critical health warnings. Empirical studies conducted using such manipulated data may yield erroneous conclusions, consequently generating misleading policy recommendations. The outlier detection algorithm based on the distance metric of the Two-Step clustering method proposed in this paper can effectively analyze and screen abnormal air quality data, assist decision-makers in making informed decisions, and to a certain extent, contribute positively to the improvement of air quality.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work is supported by the Fundamental Research Funds for the Central Universities (Grant No. 2572022BC07).

REFERENCES

[1] Cheng Hanxi, Zhu Hongxia, Wang Jing, et al. Research on the impact of air pollution prevention and control on air quality in the Beijing-Tianjin-Hebei region. Environmental Impact Assessment, 2024, 46(06): 78-85. DOI: 10.14068/j.ceia.2024.06.013.

[2] Song Guojun, Li Honglin. Targeting PM2.5 pollution: updated design of an air quality management policy framework. China Population Resources and Environment, 2023, 33(02): 1-10.

[3] Fu Jianhua, Zhou Fangzhao. Measurement and influencing factors analysis of provincial air quality in China. Urban Problems, 2020(05): 20-27. DOI: 10.13239/j.bjsshkxy.cswt.200503.

[4] Wu EMY, Kuo SL. A study on the use of a statistical analysis model to monitor air pollution status in an air quality total quantity control district. Atmosphere, 2013(04): 349-364.

[5] Zhang Y, Xu L, Lu Z. Synergistic effect of factors influencing urban air quality in China: a hybrid model integrating WGRA and QCA. International Journal of Environmental Science and Technology 2023(11): 12179-12194.

[6] Yu Heng, Hou Xiaolan. Application of hierarchical clustering algorithm in astronomy. Scientia Sinica (Physica, Mechanica & Astronomica), 2022, 52(08): 118-131.

[7] Huang Mengting. A recommendation algorithm based on The combination of two-step clustering and association rules. Information & Computer, 2021, 33(01): 35-37.

[8] Ding Sifang, Wang Shouwei, Zhu William. Density peaks clustering algorithm based on two-step allocation strategy. 2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). IEEE, 2023(02): 946-950.
[9] Hong Xia, Gao Junbin, Wei Hong, et al. Two-step scalable spectral clustering algorithm using landmarks and probability density estimation. Neurocomputing, 2023, 519: 173-186.
[10] Kumar Y, Sahoo G. A two-step artificial bee colony algorithm for clustering. Neural Computing and Applications, 2017, 28: 537-551.