# K-clustering methods for investigating social-environmental and natural-environmental features based on Air Quality Index

Victor Chang[1], Pin Ni[2] and Yuming Li[2]

1. Teesside University, UK
2. University of Liverpool, UK

**Abstract**

Air pollution has caused environmental and health hazards across the globe, particularly in emerging countries such as China. In this paper, we propose the use of Air Quality Index (AQI) and the development of advanced data processing, analysis and visualization techniques based on the AI-based k-clustering method. We analyze the air quality data based on seven key attributes and discuss its implications. Our results provide meaningful values and contributions to the current research. Our future work will include the use of advanced AI algorithms and big data techniques to ensure better performance, accuracy and real-time checks.

Keywords: Air pollution analysis; Air Quality Index (AQI); K-clustering; K-clustering algorithm; seven key attributes for AQI.

## 1 Introduction

### 1.1 Motivation

Maintaining the quality of the air has become more challenging due to the economic development and industrialization. Pollution is a common problem in emerging countries [1]. Therefore, the quality of air has been down because of more toxic and harmful gases due to burning, heavy uses of air-conditioning systems, traffic congestion and industrial activities [2]. The quality of the air is essential to our day-to-day lives and businesses. Without the guarantee for clean and refreshing air, the health conditions of the population may get worse, with higher rates of getting lungs and pollution-related diseases and cancers. Maintaining the quality of the air has become a challenging and urgent task for emerging countries such as China.

With the rapid development of data mining techniques and substantial growth of environmental data volume, it is worth exploring natural and artificial attributes related to AQI (Air Quality Index). The aim is to provide scientific guidance to regulate the air environment. AQI can be discussed in detail and has been summed up as an effective way to reduce carbon and toxic gas emission, as well as improving the awareness of maintaining good air quality [2].

In our project, we have collected ten natural and social-economic features among 323 cities in China and the AQI from China National Bureau of Statistics and related database websites. As for geographic data, software like Google Earth could help acquire latitude, longitude and altitude, as well as coastal. This paper is presented as follows. Section 2 describes the related work for this research and Section 3 illustrates the methodology in place. Section 4 offers results and analysis, including the most significant seven key features. Section 5 discusses topics of relevance and importance and Section 6 sums up this paper with future work.

## 2. Related Work

### 2.1 Literature Review

This section provides a literature review for air quality research and its status, as well as the methods to conduct air quality research. Di et al. [3] used the two-pollutant Cox proportional hazards model to assess mortality associated with exposure to PM2.5 and ozone. The results of the study indicate that exposure to PM2.5 and ozone can have adverse effects, which is most pronounced among ethnic minorities and low-income people. Raaschou-Nielsen et al. [4] estimated the association between the components of a 'particular matter' and lung cancer incidence. They suggested that the effect on lung cancer depends on the composition of a particular matter. For example, inhalation of contaminating particles containing Ni and S elements can have a more adverse impact.

Inhaled PM may induce adverse cardiovascular reactions through three potential mediators. It is impossible to avoid air pollution altogether, but the current restrictive standards should be guaranteed to reduce the source of potentially polluting air particles [5]. Shah et al. [6] suggested that particulate air pollutants have a significant relationship with stroke mortality and environmental health policies that reduce air pollution can reduce the risk of stroke.

Concerning the methods to conduct air

quality research, AI-based algorithms with big data mining and analysis can excavate valuable data, and further discover related correlation and decision-based knowledge. This method has been devoted to environmental data research, especially for the study of our atmospheres. Excavation around one city is the most common research [7]. Besides, the use of geographical and meteorological data has been matured during the last two decades. Furthermore, the socio-economic factors are gradually considered in the mechanism analysis, such as Gross Domestic Product and rapid population expansion. The broadest range of research in China is in 87 cities [8].

There were unexplored research gaps. Firstly, with the rapid urbanization in China and the advance of regional communication, it is essential to include enough cities to handle with PM2.5. Secondly, environmental and socio-economic features are analyzed separately. It ignores the relation between them with air quality.

Our project would investigate both environmental and social-economic features to provide valuable insights and analysis useful for top decision-makers in mainland China. Additionally, this research can enhance the awareness of having a good quality of life since the quality of the air can directly influence it. Everyone has the right to know the air quality at any time, particularly about when the haze will happen and their impacts on our health. Therefore, this research can potentially provide Chinese cities with the first-hand analysis and predicted outcomes, facilitating them to manage air quality issue corporately by using mutual patterns and intelligent algorithms.

### 2.2 Theories and selected methods

Many available models can perform data analysis and simulations. Our previous work has included various portfolios and works examples in different disciplines. Our previous work has adopted different models and blend them successfully with visualization and analytics. However, they are not entirely suitable for AQI analysis. The reason is AQI may collect datasets with huge variance. It needs a model that can perform categorization of datasets well, and focus on computational analysis for each of the categories. Hence, the chosen model should fulfill these criteria.

K-clustering is a vector quantization-based method often used for cluster analysis in data mining [9, 10]. Often there are different types of data and emphasis, and it is crucial to categorize different types of data into different groups. Sometimes it is unclear at the beginning before handling unprocessed data. Steps are used to identify the correlation between datasets and understand their similarities so that they can be classified into different groups for analysis [10]. K-clustering is a very useful method when we have a large quantity of data that requires categorization and understand any possible direct correlation between datasets. Therefore, the k-clustering method has been used in our analysis.

## 3 Methodology

We have used advanced data extraction and analysis. The critical aspect of the code is presented in the Appendix, with full details on data processing, analysis, results and interpretations. In our case, the raw data of each attribute has been integrated into a complete data set. It was written into a CSV file to storage. Pandas Library for Python provides

perfect data management and abundant analysis methods. We then prepare raw data before computational analysis.

### 3.1 Data Pre-Processing and Fixing

The way to deal with the problem is grouping the data sets into clusters by clustering algorithms, and then filling in the missing records with its group means. As two of the most classical and effective methods in clustering methods [9, 10], K-Means and DBSCAN are chosen to cluster data sets.

Sugar and James propose that a fundamental problem in cluster analysis is to determine the optimal number of groups [11]. To determine the modeling parameters, Calinski and Harabaz indices are chosen, which are based on recommendations of other researchers' studies. According to Milligan and Cooper in the 1985 study, Calinski and Harabaz indices are considered to be the robust indices that define the number of clusters [12].

The project finally finds that K-Means' overall score was better than DBSCAN under different parameters. Therefore, the research decides to conduct a more in-depth parameter for the K-Means algorithm. Considering the small number of cities in the dataset and a substantial K value will generate many clusters deviating the focus, we choose K = 5 as our clustering method. We compute for Calinski-Harabaz Score, DBSCAN and Calinski-Harabaz Score for different k-means values.

In the process of filling in missing data, it was found that different methods have a noticeable influence on the degree of correlation between various features. Meanwhile, through the benefits of visualization, the result showed that clustering did not work well in serious missing data such as Precipitation, Green Coverage Rate, Incineration, etc. The clustering method will generate the same results to fill in similar types of cities, see Figure 1.

Figure 1 Filling for the incomplete data using DBSCAN clustering method


Hence, the research built several machine learning models with existing data to predict the missing part to obtain more realistic filling results and features correlations [13]. The research tested and compared other eight classical machine learning methods including Bayesian Linear Regression (BLR), K-NN, Support Vector Regression (SVR), Logistic Regression (LR), Decision Tree Classifier (DTC), Linear Discriminant Analysis (LDA), Support Vector Classification (SVC) and Gaussian Naive Bayes (GNB) [12-15] for predicting and filling missing data. Based on observations, filling the missing data with SVR methods may cause over uniform for filling results. Furthermore, regression type methods may lead to the negative values of the filling data etc. Therefore, we chose the model based on the decision tree algorithm, whose cross-validation evaluating performance is better than other learning models. Besides, it only needs a small extent of computing resources.

### 3.2 Correlation analysis and Discretization

For analyzing the importance of different features, the research calculated features covariance matric to compare correlation with AQI. The marked red column is the correlation with AQI, the higher the absolute value, the more relative to the corresponding feature. A positive value means a positive correlation, and a negative value stands for the negative correlation. This step provides a solid basis for the next analysis step because it could facilitate judge the influence weight of attributes.

Our work can split the constant features data into 4-grade bins (dividing by 25%/50%/75% values) so that the data become organized and more natural to compare and analyze. The grade bin makes it more intuitive than a cluttered point set. More visualization results will be presented in “Results and Analysis” and “Conclusion and 247 Future Work” sections.

## 4 Results and Analysis

We also analyze the correlation of eight factors with AQI focusing on latitude, precipitation, temperature and altitude, GDP, coastal, incineration and longitude. The analysis of each factor will be explained. According to the correlation coefficient calculating, the two features owning the highest correlation with AQI are latitude 0.55 and precipitation -0.4. Besides, there is no obvious difference between temperature -0.22, altitude -0.2, GDP 0.16, and coastal -0.15.

There is no obvious evidence to indicate that population density, vegetation coverage rate and longitude are key attributes to influence AQI. It can be seen from our computational analysis, the correlation of precipitation – temperature 0.64, incineration – population density 0.9, and temperature – latitude -0.74 is high.

To improve accuracy, we check the predicted and actual values all the time. We can identify weaknesses and improve our computational analysis. We can use algorithms developed in [16] to ensure smaller differences between predicted and actual values within 5% most of the time and also within 10% when great fluctuations happen due to strict traffic control or coal-burning.

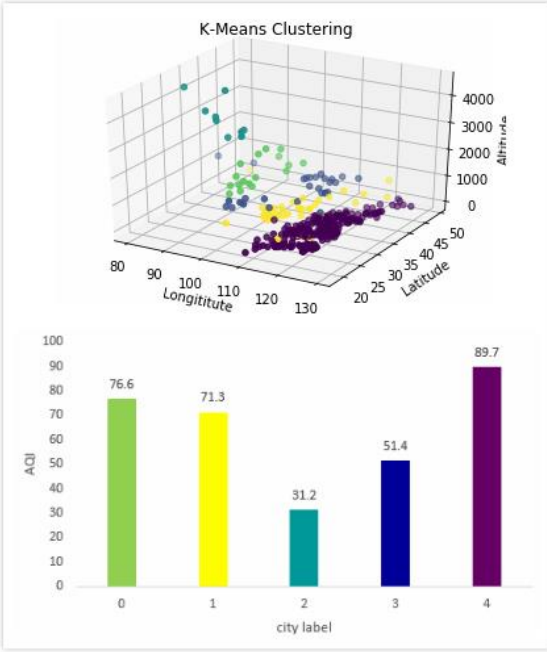
The following would try to elaborate in detail the mechanism of key features affecting AQI based on fitting, regression, binning and clustering results.

### 4.1 Clustering

In Figure 2, Column 4 cities with Violet locate in the east coastal area. Their value is highest, 89.7. They are the most densely developed area in China, including the northeast area, the Bohai gulf area, North China Plain, the East China area, and the Pearl River Delta region.

The worst air environment for Column 4 cities is the consequence of overexploitation. One reason for reliance on this region is the dominant position of the marine economy stimulated by economic globalization. Another reason is that they own the vastest plain productive land.

Figure 2 AQI of clustered cites




Colum 1 cities with yellow locate in the middle area in terms of longitude. Their value is third highest, 71.3. They are mainly in Loess Plateau and Central China area.

Except for column 4 cities, column 1 cities own the largest portion of heavy industry. The heavy industry is the main emission source of air pollutants. Restricted by insufficient capital and a weak sense of preventing and treating pollution, air pollution from heavy industry cannot be handled well.

Column 3 cities with dark blue, concentrate on Guanzhong Plain (the central Shaanxi Plain) and Chengdu Plain. Their value is fourth highest, 51.4.

Compared with the east and middle region, there exists fewer cities, population and industry. Besides, the abundant unexploited natural environment could provide a buffer to pollution impact.

Column 0 cities with light green are mainly located in the northwest area. Their value is the second highest, 76.6. They are the underdeveloped area in China and few populations live here. The vegetation is rare. The issue of soil-erosion and desensitization is serious.

Column 2, with blue-green, is the northwest frontier region of China, Xinjiang. It owns the best air quality with 31.2 AQI.

The population number and density here are nearly least in China. Therefore, their activity could not result in a serious influence on air.

### 4.2 Key feature – temperature

We undertake temperature analysis. The main concentration of cites is from 14 degrees to 28 degrees. Humans do prefer living in cities with annual average temperatures from 14 to 28. AQI decreases with the increase in annual city temperature. The higher the temperature, the stronger the convection activity near the ground, the more unstable the structure of the atmospheric layer. It is easier for air pollutants to diffuse. The number of cities with annual temperatures in the range of 0 to 10 is rare. It is not suitable enough to research the recent interaction issue of society and the environment.

### 4.3 Key feature – latitude

We analyze regression between the latitude and AQI and find there is a positive correlation between both. This means that in China, generally, with the increase of latitude (from 20 degrees north to 50 degrees north), the value of AQI gradually increases (from 20 to 180), and air quality deteriorates. Generally, the temperature of the south is higher than in the north. Stronger air convection activity due to hot air accelerates the diffusion rate of pollutants.

Therefore, it is difficult for haze to stay above the city for a long time. Besides that, when winter comes, there are coal-fired power plants in the north for heating. Coal-fired power is a kind of high-polluted energy for air. Another consideration is that the vegetation coverage in the north is low. It could not prevent soil loss. With wind, sand storms often outbreak in the north in spring.

### 4.4 Key feature – precipitation

We analyze regression between the precipitation and AQI and find there is a negative correlation between precipitation and AQI. The greater the rainfall is, the lower the AQI becomes, and the better the air quality is. Rainfall has the capability of wash and dissolution to airborne particles. These actions are beneficial to the wet deposition of particles. Therefore, rainfall could reduce the concentration of pollutants in near-surface air.

### 4.5 Key feature – GDP (Gross Domestic Product)

Cities are clustered into four groups based on GDP index: low, middle, high and very high. From low to very high, the relation between GDP and AQI is roughly proportional. Our analysis shows that the higher the GDP, the higher the AQI. It's worth noting that the AQIs of middle and high cities are almost equivalent. Further GDP composition should be researched to investigate this state.

### 4.6 Key feature – Coastal

Cites are divided into two clusters: coastal and non- coastal. An analysis is taken. The average of AQI of coastal cities is 63.2, 16 less than 79.1 of non-coastal cities. Air quality in coastal areas is better than non-coastal regions by an average of 15%.

Firstly, the wind is much stronger in coastal cities because of the sea. It promotes ventilation. Secondly, the sea could absorb particles in the air. It is a natural absorbent. Finally, due to the nourishment of the sea, vegetation could purify the air.

### 4.7 Key feature – Altitude

We perform an analysis of attitude and AQI. While city altitude increases, air quality gradually improves. With altitude increases, the pressure gradually decreases. The structure of the atmospheric layer is comparably stable for high pressure, since it is difficult for pollutants to diffuse. While it turns to low pressure, pollutants near ground rise with the convergence of air. It is easy for diffusion. The concentration of particles would decrease.

## 5 Discussion

Combining the inner correlation with impact factors, the polluted atmosphere can break stability and accelerate deposition is the main method to decrease air particle concentration. Therefore, either for the individual pursuit for good air quality or for industry layout with air pollutant emission, controlling natural features should be given precedence. Restricted by intrinsic property and capacity of nature, human beings should carry out activities aligned to natural conditions to achieve sustainability. Our analysis can provide useful insights to understand the status of air quality in China. Hazy smog is common in many parts of China

in winter, due to the emission of burning fuels and coals to keep the household warm. Our analysis was based on air quality before the COVID-19 outbreak. Real-time analysis can ensure the research outputs can be more meaningful and beneficial to the public. For our research contributions, we demonstrated the effective use of k-clustering methods to analyze weather data and understand correlations between different key attributes.

## 6. Conclusion and Future Work

In light of the pollution analysis, AQI key feature identification was a complicated process. Restricted to the initial stage of AQI big data exploration, plus the incompleteness of dataset among 300 cites in China, it was difficult to contribute a full AQI to represent entire China. We managed to demonstrate that our k-clustering methods could analyze AQI. Further targeted clustering research should consider key attributes, so that the prediction model could be built to crackdown air pollution. We analyzed seven key attributes based on AQI and presented its insights to report essential lessons learned and decisions to be taken.

Considering the robust execution of China's policy, the distribution of regional population and industry layout could be controlled to improve cities' structure. Thus, a better buffering could mitigate the air pollutants impact. For example, temperature adjust means could reduce the urban heat island effect to accelerate airflow. Proper manual intervention to rainfall could facilitate wet deposition.

Our future work will include the blending of big data techniques and intelligent AI algorithms to provide analysis with better performance and accuracy. We will predict the air pollution status in different cities of China and the other parts of the world, such as the UK.

## Acknowledgment

We are grateful to VC Research funding: VCR 000011 to support.

## References

[1] Martínez-Zarzoso, I., & Maruotti, A. (2011). The impact of urbanization on CO2 emissions: evidence from developing countries. Ecological Economics, 70(7), 1344-1353.

[2] Roy, S., Bose, R., & Sarddar, D. (2017). Smart and healthy city protecting from carcinogenic pollutants. Int. J. Appl. Environ. Sci, 12, 1661-1692.

[3] Di, Q. et al. (2017). Air pollution and mortality in the Medicare population. New England Journal of Medicine, 376(26), 2513-2522.

[4] Raaschou-Nielsen, O. et al. (2016). Particulate matter air pollution components and risk for lung cancer. Environment international, 87, 66-73.

[5] Shen, F., Ge, X., Hu, J., Nie, D., Tian, L., & Chen, M. (2017). Air pollution characteristics and health risks in Henan Province, China. Environmental research, 156, 625-634.

[6] Shah, A. S., et al. (2015). Short term exposure to air pollution and stroke: systematic review and meta-analysis. bmj, 350, h1295.

[7] Lin, G. et al. (2014). Spatio-temporal variation of PM2. 5 concentrations and their relationship with geographic and socio-economic factors in China. *International journal of environmental research and public*

*health*, *11*(1), 173-186.

[8] Hao, Y., & Liu, Y. M. (2016). The influential factors of urban PM2. 5 concentrations in China: a spatial econometric analysis. *Journal of Cleaner Production*, *112*, 1443-1453.

[9] Chakraborty, S., Nagwani, N. K., & Dey, L. Performance comparison of incremental k-means and incremental dbscan algorithms. arXiv preprint arXiv:1406.4751(2014).

[10] Schubert, E. Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3), 1-21.

[11] Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association*, *98*(463), 750-763.

[12] Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159-179.

[13] Lakshminarayan, K., Harp, S. A., Goldman, R. P., & Samad, T. (1996, August). Imputation of Missing Data Using Machine Learning Techniques. In *KDD* (pp. 140-145).

[14] Izenman, A. J. (2013). Linear discriminant analysis. Modern multivariate statistical techniques. Springer, New York, NY, 2013. 237-280.

[15] Ullrich, Christian. Support vector classification. *Forecasting and Hedging in the Foreign Exchange Markets*. Springer, Berlin, Heidelberg, 2009. 65-82.

[16] Chang, V., Li, T., & Zeng, Z. (2019). Towards an improved Adaboost algorithmic method for computational financial analysis. *Journal of Parallel and Distributed Computing*, *134*, 219-232.

**Victor Chang** is with Teesside University, Middlesbrough, U.K. He is the corresponding author of this article. Contact the author at: ic.victor.chang@gmail.com.

**Pin Ni** is with University of Liverpool, Liverpool, U.K. Contact the author at: maxwellnee@gmail.com.

**Yuming Li** is with University of Liverpool, Liverpool, U.K. Contact the author at: yumingli1996@gmail.com