Beijing Air Quality Analysis

Anonymous Author(s)

ABSTRACT

Studies showed that air quality data included many potential patterns and can be used to predict future air quality. However, apart from daily broadcasting information, air quality data usually had an extremely large size which required plenty of data mining and visualization techniques to fully analyze. Therefore, our big data analysis project on air quality provided a comprehensive analysis of data management, visualization, and mining. Ethical and legal considerations around the air quality analysis were also presented according to the ACM code of Ethics and United Nations Framework Convention on Climate Change.

In this work, air quality data of air pollutants and meteorological parameters in the Beijing area were collected from 2013 to 2017 along with several supporting economic data from the same time period. The attributes resolved in an entity-relationship model and a database were built in SQL. SQLALchemy was used to manage the connection between python and SQL. Moreover, several novel techniques including scroll detection of outliers were introduced and applied for data preprocessing and cleaning. For data mining analysis, a clustering experiment demonstrated a hidden relationship between weather conditions and air quality via several visualizations. Association analysis revealed another relationship between air quality and human activity. In general, this work suggested a few original analyses on air quality and proposed a potential air quality forecast idea.

KEYWORDS

Air quality, Air pollutant, Outlier detection, Data Management, Clustering, Association

1 INTRODUCTION

As an expanding concept, big data and its analysis have shown a strong capability to benefit many fields including economic and social networking research [7, 21]. As a matter of fact, the data generated by everyone in a single day can reach 2.5 exabytes (1018 bytes), and every single person can generate about 1.7 megabytes in one second [15, 19]. Not only personal data but also companies and organizations collect and analyze a massive amount of data. And among these data, there are plenty of undiscovered opportunities that can be studied. Apart from those well-known and popularly used personalized social media that based on individual favor, economic and environmental data has attracted our interest and inspired our project. And our motivation was directly coming from this.

Dealing with the environmental pollution issue by almost every county has been coming together with its economic development ever since the industrial revolution. Although the seriousness and the control standard can be very different across countries or regions, there are many international consensus and criteria nowadays to indicate and scale the environmental problems, respectively. In fact, sustainable development has been a universal and urgent concept for everyone in the world ever since the U.N. generated it. However, during the last decades, excessive development has bruised the environment in China due to a lack of awareness about environmental protection, even though the Chinese government has planned to control environmental pollution to establish an eco-friendly society. A lot of indices including air quality index (AQI), comprehensive pollution index of water quality, and forest area were raised to reflect different aspects of the environment. Among them, the AQI is one of the most significant and evident factors that not only indicate environmental pollution but also affect the daily life of everyone. Generally, AQI can be displayed as simply as just one number which usually ranges from 1 to 10. Besides this straightforward mode, AQI also contains the records of several common pollutants like sulfur dioxide, nitrogen oxides, and particulate concentrations in the air. Because of the difference in production, concentrations of pollutants can vary in many tendencies. In some cases, pollutants like sulfur dioxide and nitrogen oxide usually come from industry emissions rather than natural reasons. Therefore, it can be useful to associate with both industrial activity and human activity, which can be obtained from some specific organizations. And it is actually one important part of our analysis. In addition, some climate factors like rain and wind can affect the pollutants spreading especially when the data was collected in a relatively small region, like a city, so the weather data collected by monitor sites is another important part to help us analyze the air quality.

In this project, our goal is to analyze the environmental problems displayed by air quality as well as the economic development influence in the capital of China, Beijing. Thus, the Beijing Multi-Site Air-Quality Data dataset [17] was selected to study the variation of air pollution in the Beijing area. This dataset, partially shown in Table 1, contains time-series data from 2013 to 2017, recording the air quality and weather conditions collected from 12 nationally controlled air-quality monitoring sites in Beijing. This dataset measures the air quality through 12 attributes: the concentration of six common pollutants (particulate matter 2.5, particulate matter 10, sulfur dioxide, nitrogen dioxide, carbon monoxide, and trioxygen) and six regular meteorological parameters (such as temperature, pressure). Apart from the air pollutants concentration values and meteorological parameters, some sub-datasets which contains passenger cars in use in China [3], passenger cars production [2] (from NationMaster), gross domestic product (GDP), and primary industry value [4] (from NBS) was also introduced as a part of the supporting economic analysis. This series of data can supply the economic situation varying in the desired time period, which is highly beneficial in this project. Particularly, the correlation between the economic data and the pollution data can tell us the effect of development on air pollution.

The entire Beijing Multi-Site Air-Quality Data dataset contains over 40 million sets of data currently. Considering these meteorological data are collected hourly in multi-sites, both the volume and velocity are matched with the description of big data. Because multiple sub-datasets mentioned previously are introduced in this

project, there are many formats and quality of data with different structures that fit the variety of big data analyses. Since all data were collected from official organizations, we consider them to be highly reliable, which fits the veracity characteristic. A large amount of original data, especially for air pollutants and meteorological parameters, makes the process of finding useful information challenging. In order to find relevant pieces of information and prevent the influence of extreme values, several data cleaning techniques and preprocessing are applied to modify our datasets. In general, we managed to keep the characteristics of the original data and minimized the unnecessary modifications. Specifically, we first examined the null value via three different conditions, and we then filtered outliers by comparing the positive difference between each value and the average with the standard deviation. Moreover, several preprocessing methods were applied to improve the efficiency of our analysis.

With the help of the data cleaning, we simply took the advantage of pandas library for python to build our database by transforming the original CSV file into data frames in pandas and followed by the implementation to a SQL database. For the evaluation, we performed a clustering to explore the relationship between fluctuations of air pollutants and the residents' daily routine. Actually, we processed the data within different time periods and investigated the relations with human activity like traffic influences. Another strategy is to find the spreading of air pollutants in geography since we have 12 monitoring stations across the Beijing area. And we applied association analysis such as how wind direction can affect the air quality as well. There are many works [17, 18, 20] have been done to exhibit the air quality, and they focused on many different investigations like focusing on a specific city, influences of time and space factor, and the contributions to AQI of different air pollutants. In this work, we considered many social factors including industry, construction, people's living conditions, and economic development, and combined them with the air quality to make a comprehensive analysis in this project.

The organization of our report states as follows. In section 2, we will evaluate the intention and inspiration of some valuable work related to analyzing air pollutants in Beijing. The design of the relational database model for the datasets and the data mining algorithm implementation will be discussed in depth in section 3. The analysis of the patterns extracted from datasets will be presented in section 4. Section 5 will state the ethical impact and legal concerns explicitly relating to the air pollution problem, and section 6 will discuss the lessons learned. In section 7, we will talk about the current status and future work. We will draw our final conclusion in section 8.

2 RELATED WORK

Previous works have inspired us from different approaches and data mining techniques to analyze air quality. The approaches for air quality analysis and the data mining techniques pave important roles of our study. We compared the works and modified their approaches to finding different tendencies and making reasonable analyses from our existing data.

For the approaches, Yu et.al. [20] studied the air quality in Wuhan, another metropolis in the center of China. They present dynamic research on different periods. Similarly, they also made use of AQI and air pollutants concentrations to detect the trends in Wuhan via some visual methods. By the means of dynamic analysis, they applied different time intervals ranging from hourly to annual to implement their works. As a result, they exhibited some patterns they found related to periods.

Song et.al. [18] studied the air quality in Jinan, another big city in the north of China. They particularly focus on the contribution of each air pollutant to the AQI. Their research summarized the correlation of each air pollutant to the badness of air quality, and a calculation of each coefficient was present as well. Furthermore, they also provided some possible reasons for this phenomenon, which indicated human factors like the increase of fossil fuel consumption.

For the data mining techniques, we exploit the methods of k means clustering, k-map based positive association mining technique. Baruri et al. [5] introduced a greedy k means algorithm by utilizing layer optimization technique to reduce the limitations. The n nearest clusters combination is performed in the initialization step. Then, for the sake of lowering the computational cost, cluster pruning is used. Finally, the algorithm used an optimized updating method. Dhanasekaran et al. [9] presented a k means clustering algorithm by utilizing an enhanced map reduce technique. The technique can be used to reduce unneeded data, as well as to optimize data storage and accomplish data privacy outsourcing.

Ravi and Khare [14] proposed an Efficient and Optimized Positive-Negative Association Rule Mining algorithm (EO-ARM) to solve the problem that many negative Association Rule Algorithms scan the dataset multiple times to identify frequent item groups and do not ensure that all derived rules are useful. It uses a two-dimensional matrix to scan the dataset only once to find frequent itemsets. It further improves the association rules by pruning less interesting rules.

3 OVERALL APPROACH

First of all, our database management is described as follows. Based on the natural characteristics of Air Pollutants and Weather, two representing tables were created though they came together in the original data, as shown in Figure 1. Moreover, both of them simply took the time factor in Air Pollutants as their primary key because they were recorded at a specific time. Since all different parameters are atomic values which used to be measured by instruments and there was no direct relationship between them obviously, these two tables can be managed to match the first normal form. Beside that, we attempted to decompose the attributes monitor site and time from the previous tables. Each monitor site table took their name as the primary key and each time table took the year and month values as the primary key. And all these tables were connected by a relationship set Record. For the Vehicle dataset, we considered the volumn of the vehicle and the year factor as the primary keys. Similarly, the year and month are settled as the primary keys for the electricity entity dataset as a representative of industry activity. In addition, we use SQLALchemy toolkit to manage the database from python.

Second, we applied several data cleaning and preprocessing methods before actually building the database. To improve the efficiency

NO	YEAR	MONTH	DAY	HOUR	PM2.5	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	WD	WSPM	STATION
1	2013	3	1	0	4	4	4	7	300	77	-0.7	1023	-18.8	0	NNW	4.4	Aotizhongxin
2	2013	3	1	1	8	8	4	7	300	77	-1.1	1023.2	-18.2	0	N	4.7	Aotizhongxin
3	2013	3	1	2	7	7	5	10	300	73	-1.1	1023.5	-18.2	0	NNW	5.6	Aotizhongxin
4	2013	3	1	3	6	6	11	11	300	72	-1.4	1024.5	-19.4	0	NW	3.1	Aotizhongxin
5	2013	3	1	4	3	3	12	12	300	72	-2	1025.2	-19.5	0	N	2	Aotizhongxin

Table 1: The first five entries in the Air-Quality Data. It has 18 attributes. The six attributes (PM2.5, PM10, SO2, NO2, CO, O3) are about the concentration of air pollutants. The other six attributes (TEMP, PRES, DEWP, RAIN, WD, and WSPM) records the weather conditions.

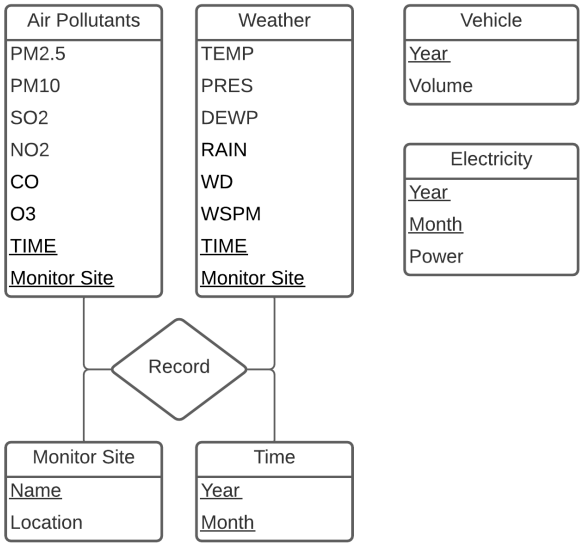

Figure 1: ER model built for Beijing Multi-Site Air-Quality Data dataset. The record relation represents the dataset about the air quality. Air pollutants and weather are weak entities. They are dependent on the monitor site and time table. Monitor site and time are strong entities. Vehicle and electricity are two tables recording the vehicle sale and electricity generating data in Beijing.

of our code and evaluate our approach dynamically, all data cleaning and preprocessing methods were performed in data frames provided by pandas library. Since most of the pollution data [17] are collected automatically from air quality monitor sites, they contain many null and noise values due to the nature of monitoring devices. To better analyse the data, we have to handle null and noise values properly in case that they may reduce the reliability of our analysis results (null details). Null values were usually caused by missing values, while noise data always appear as outliers which are significantly distinguishable from normal data. A general way to detect such outliers is to exploit the standard deviation since 99.7% data always lies within 3 standard deviations of the mean [1]. We calculated the standard deviation for all the attributes from the dataset and found outliers following the Empirical Rule. Specifically, the filtering rule is $|data - mean| > 3 \cdot stdev$. We further calculated the absolute difference between each value and the average value. And we filtered out those values with absolute differences larger than 3 times of the standard deviation with the replacement of the average value.

At the beginning, we tried a naive approach which simply used the entire dataset for the previously described calculation. However, this global approach can lead to a severe error when there was any tendency in the original dataset. For example, if the original data was progressively increased, the global method can cause a sharp fluctuation and make the downstream analysis even harder. Thus, a local approach was introduced after careful consideration. By local we manually separated the original datasets into a lot of continuous overlapping sublists, and performed the previously described calculation within them. Fortunately, this can be accomplished by the rolling method in pandas. And the data frame in pandas also ignores null values automatically, so the outlier filtering can be done before handling null values. Besides, considering these parameters were collected hourly, we limited the size of local sublists to 20. By performing this cleaning, we can exclude most of the outliers but maintain the original tendency most possibly.

In fact, we first replaced the outliers as null values intentionally, and they were replaced by the calculated average value along with the original null values. By doing this we can handle the outliers and null values in a single step, which significantly improved our code efficiency. However, there were still some drawbacks that may cause failure in our preprocessing. For some cases that a continuous series of null values appeared, we had to choose a backward rolling to fill those null values. And to improve the performance of our code and match with the selected dataset, we kept using the backward method.

However, this method cannot apply to the entire dataset since there are still some non-numerical data. For example, a string type of wind direction that could not perform any calculation on mean or standard deviation also contains null values. Our first strategy to handle such attributes was similar to our previous local approach that replace those null values with the local mode value. However, it turned out that the pandas library did not support abusing rolling objects with mode. As a matter of fact, we took another simpler strategy that directly used the backward filling with the first encountered valid value after the null value.

Eventually, we created the database for our work and transported the processed data from pandas data frames into SQL database directly. All previous data cleaning and preprocessing methods already guarantee the data integrity, and the analyses can be done straightforwardly. The general explanation of program diagram is displayed in Figure 2.

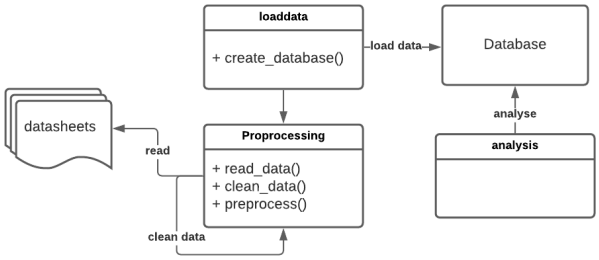

Figure 2: The program diagram of the database implementation. Loaddata module uses read_data function and clean_data function from preprocessing module to read the data from datasheets, clean the data and load the data into database. Analysis module reads data from the database and do further analysis.

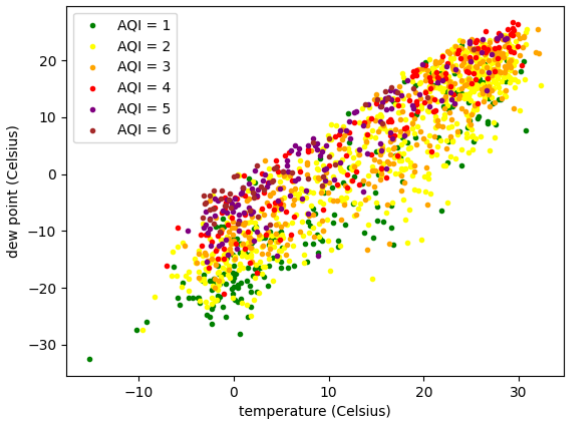

Figure 3: The influence of temperature (Celsius) and dew point (Celsius) on air quality (AQI). Meteorological and air pollutants data was recorded in Aotizhongxin monitor site. The AQI in Beijing was scattered daily from 2013-3-1 to 2017-2-28 and grouped in different color.

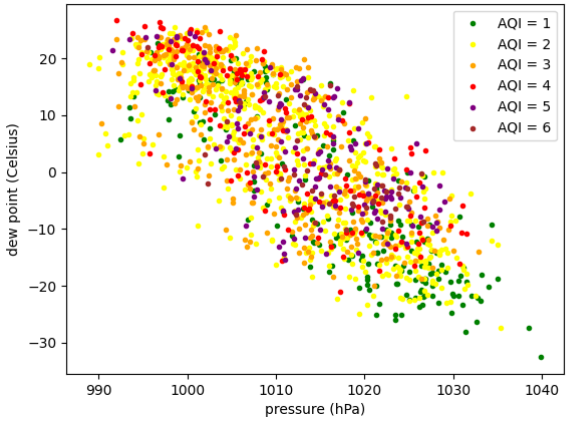

Figure 4: The influence of pressure (hPa) and dew point (Celsius) on air quality (AQI). Meteorological and air pollutants data was recorded in Aotizhongxin monitor site. The AQI in Beijing was scattered daily from 2013-3-1 to 2017-2-28 and grouped in different color.

4 ANALYSIS

In the air quality dataset: the date (year-month-day-hour), the concentration of six pollutants, temperature, pressure, dew point, rain precipitation, and wind speed can all be treated as interval type. The concentration of six pollutants, rain precipitation, and wind speed can be considered with a literally "zero point" though their units are different. The temperature, pressure, and dew point do not have a "zero point," but they are all settled in a specific range. Finally, the wind direction can be treated as an ordinal type where there are 16 different values. Since many continuous intervals and ratio attributes can be directly extracted from these datasets without too much delicate modification, our first data mining attempt is considered clustering.

As one of the most widely used data mining techniques, clustering is frequently used to discover exciting information from the dataset [5, 9, 13] such as the chronological and geographical pattern of the concentrations of air pollutants. We attempted to use clustering to explore the relationship between concentrations of air pollutants and the weather conditions. After calculating the daily air pollutants average, we can group the time period with the calculated AQI. Then, comparing the AQI groups with the typical weather conditions like temperature or pressure can help us understand how the weather could affect air quality. We will also utilize clustering to find out the geographical pattern of the spreading of air pollution. Grouping dates with the same AQI with different weather conditions can expose the influence of weather and geographical factor on air quality in Beijing. Such pollution patterns can expose how regional factors such as natural parks or power plants may affect air pollution. We specifically displayed one relatively good and one relatively bad cluster examples on weather conditions. The good example in Figure 3 showed a blurred separation between air quality on temperature and dew point. The patterns of good air quality data (low AQI) and bad air quality data (high AQI) had an interval difference which can be separated linearly. However, the similar pattern was not observed in the bad example in Figure 4, and all data were squeezed together in the quadrant. Thus, we found one of the possible clustering factors between temperature and dew point, which may be used to forecast air quality.

Association, an essential data mining technique, depicts hidden relations between different data attributes [11, 14]. Here we are going to show an example of the data application using association. Since the wind condition has a crucial influence on air pollutants [6] and is the most apparent attribute in our original dataset, our first

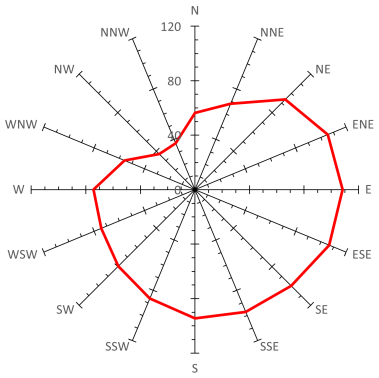

Figure 5: An example exploratory data analysis such that displaying the average PM 2.5 concentration (ug/m^3) measured from 2013 to 2017 in each wind direction from Aotizhongxin monitoring site

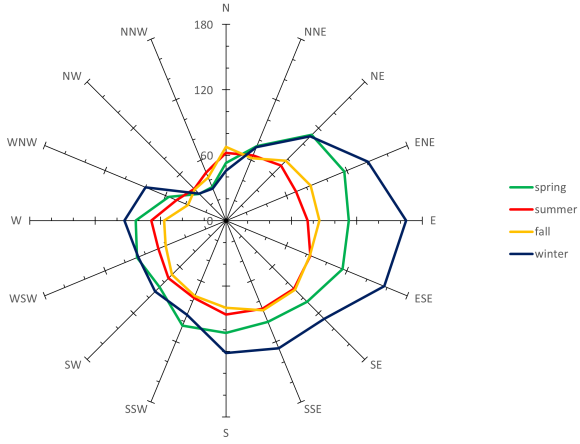

Figure 6: An example exploratory data analysis such that displaying the seasonal average PM 2.5 concentration (ug/m^3) measured from 2013 to 2017 in each wind direction from Aotizhongxin monitoring site

exploratory data analysis attempt focused on the correlations between PM 2.5 concentration (ug/m^3) d the wind direction. Figure 5 shows the average PM 2.5 concentration varying from 16 different wind directions. The average concentration was calculated from 2013 to 2017 in the Aotizhongxin monitoring site while disregarding those not applicable values. Figure 5 indicates that the average PM 2.5 concentration is relatively low when the wind comes from the northwest direction. This phenomenon may require further investigation about Beijing's industrial and business layout and an additional association with traffic or data.

In addition, another assumption was made that the pollution condition may fluctuate by season based on Figure 5. Thus, Figure 6 shows the average PM 2.5 concentration separated into four seasons. Figure 6 does not provide a significant contrast to Figure 5. The northwest wind direction still causes the lowest pollution, and the average PM 2.5 concentration does not differ in most wind directions either. As a short conclusion, the previous assumption may not be a positive inspiration to our mining.

5 LEGAL AND ETHICAL CONSIDERATIONS

As a glance on our project, we accomplished analyses on air quality and the economic development in Beijing using data mining techniques. To elaborate its legal and ethical influence, we would like to raise these following points. First, environmental protection as a global issue has attracted plenty of attention from governments and organizations. Our works provided a signal on the influence of economic and industry development on air quality so that they are social needed and broadly accessible. Thus, they can contribute to the society and to human beings which meet the first regulation in ACM Code (principle 1.1). In the analysis, we tried our best to be unbiased by using the dataset from the government official website and to exploit reasonable algorithms to reduce discriminations (principle 1.4). We appreciate and credit the new ideas and creative work by other computational professionals (principle 1.5). In addition, since we are only professionals in data analyzing fields, we processed, investigated, and visualized our research to the best of our knowledge (principle 2.6). However, we did not stand for any position or propose any opinion on the countermeasures. In fact, we hope this research could bring attention and inspire other researchers on solving environmental issues, since public good is our major concern for this project (principle 3.1) [10].

In addition, our project demonstrated respect for national and international laws. As one of the most important guidance, United Nations Framework Convention on Climate Change (UNFCCC) [12] set most regulations of environmental protection. Guiding by International Environmental Law (IEL) [16], we hope the Chinese government can improve air quality by restricting factory and car emissions. In fact, most countries have their own laws and regulations of environmental protection including air quality control. For example, the USA published Clean Air Act (CAA) [8] as a federal law back in the last century to regulate emissions of hazardous air pollutants. Our project elaborated the agreement of these laws as well as in the data analysis field.

6 LESSONS LEARNED

Although it is the first time we implement a big data analysis, we achieved a lot of objectives and learned a lot of lessons. In the data management component, we kept our eyes on the temporary results and made successive progress to improve our data integrity. In the data mining component, we also learn and exploit data mining techniques including association, clustering, outlier detection, statistical techniques, and visualization.

We made a mistake in data preprocessing, but we realized and corrected it. We try to use the global approach to use the entire data set for the calculations described above. However, this global approach may cause serious errors when there are any trends in the original data set. For example, if the raw data continues to increase, the global approach will cause drastic fluctuations, making downstream analysis impossible. Therefore, after careful consideration, we decided to use a local method , which was done through the

rolling method in Pandas. Locally, we manually divided the original data set into many consecutive partially overlapping sublists, and perform the calculations described earlier in them. The data frame in Pandas also automatically ignores null values, so we can handle outliers before dealing with null values. By performing this cleanup, we can eliminate most outliers and maintain the original trend.

7 CURRENT STATUS & FUTURE WORK

Considering that we are currently in phase 2 of our project, we have ensured our dataset and the analysis goal that focused on Beijing air quality ranged from 2013 to 2017. For the data management component, we designed our ER model based on our air quality data as well as the supporting data. Furthermore, we preprocessed our data using different techniques and successfully built our database in SQL within python. Code and instruction in this component were accomplished and can be seen in the supplement file. For the data mining component, we proposed three different techniques and displayed some primary visualizations and mining as well.

For future work, we plan to perform some querying within SQL to locate the specific data we are interested in, and further solve the data mining component by finding either the hidden patterns with clustering or the relationship between attributes with association techniques. In addition, we are also looking for more visualization methods to display our work more distinctly.

8 CONCLUSIONS

In conclusion, we analyze the Beijing air quality dataset ranged from 2013 to 2017 by using our ER model and database. Specifically, we took the advantage of SQLAlchemy to build and manage the database in SQL from python code. In data preprocessing and cleaning, we proposed and compared several strategies to smoothen the data and detect the outliers for downstream analysis. Then by using several data mining techniques such as clustering and association, we made reasonable analysis for the dataset. First of all, we picked out the two most related weather factors that influence air quality. Specifically, we displayed one good and one bad cluster relations of one monitor site, but the tendency is popular in all 12 monitor sites. Second, we tried to associate the weather and economic factors with the air quality and make assumptions on observed phenomenons. Eventually, we explain the legal and ethical considerations for the analysis. We demonstrated our work guiding by the principles of the ACM code of ethics and the United Nations Framework Convention on Climate Change.

In future works, we could certainly practice SQL querying to improve the analysis efficiency. And more visualization techniques may also benefit our project. In addition, we could focus on how the cluster can be applied to air quality forecasting since several significant patterns were found in the clustering result.

REFERENCES

[1] B.G. Amidan, T.A. Ferryman, and S.K. Cooley. 2005. Data outlier detection using the Chebyshev theorem. In *2005 IEEE Aerospace Conference*. 3814–3819. https://doi.org/10.1109/AERO.2005.1559688

[2] anonymous. 2020. Passenger Cars Production in China. https://www.nationmaster.com/nmx/timeseries/china-passenger-cars-production

[3] anonymous. 2020. Vehicles in Use in China. https://www.nationmaster.com/nmx/timeseries/china-vehicles-in-use

[4] anonymous. 2021. Quarterly National Accounts, NBS. https://data.stats.gov.cn/english/easyquery.htm?cn=B01

[5] Rajdeep Baruri, Anannya Ghosh, Ranjan Banerjee, Anindya Das, Arindam Mandal, and Tapas Halder. 2019. An Empirical Evaluation of k-Means Clustering Technique and Comparison. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. 470–475. https://doi.org/10.1109/COMITCon.2019.8862215

[6] Bert Brunekreef and Stephen T Holgate. 2002. Air pollution and health. *The Lancet* 360, 9341 (2002), 1233–1242. https://doi.org/10.1016/S0140-6736(02)11274-8

[7] Xiong Chenran and Wu Youde. 2015. The Geographic Environment Analysis of Regional Economic Development of Yunnan Province of China Based on the Big Data Technology. In *2015 International Conference on Intelligent Transportation, Big Data and Smart City*. 869–872. https://doi.org/10.1109/ICITBS.2015.220

[8] United States Supreme Court. 1963. Clean Air Act.

[9] S. Dhanasekaran, R. Sundarrajan, B. S. Murugan, S. Kalaivani, and V. Vasudevan. 2019. Enhanced Map Reduce Techniques for Big Data Analytics based on K-Means Clustering. In *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*. 1–5. https://doi.org/10.1109/INCOS45849.2019.8951368

[10] D. Gotterbarn, B. Brinkman, C. Flick, M. S. Kirkpatrick, K. Miller, K. Vazansky, and M. J. Wolf. 2018. ACM Code of Ethics and Professional Conduct. https://www.acm.org/code-of-ethics

[11] Gunhee Kim and Eric P. Xing. 2014. Visualizing Brand Associations from Web Community Photos. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (New York, New York, USA) *(WSDM '14)*. Association for Computing Machinery, New York, NY, USA, 623–632. https://doi.org/10.1145/2556195.2556212

[12] United Nations. 1992. United Nations Framework Convention on Climate Change.

[13] T Parimalam and K Meenakshi Sundaram. 2017. Efficient Clustering Techniques for Web Services Clustering. In *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. 1–4. https://doi.org/10.1109/ICCIC.2017.8524480

[14] Chandrasekar Ravi and Neelu Khare. 2014. EO-ARM: An efficient and optimized k-map based positive-negative association rule mining technique. In *2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014]*. 1723–1727. https://doi.org/10.1109/ICCPCT.2014.7054871

[15] Seref Sagiroglu and Duygu Sinanc. 2013. Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*. 42–47. https://doi.org/10.1109/CTS.2013.6567202

[16] Philippe Sands, Jacqueline Peel, Adriana Fabra, and Ruth MacKenzie. 2018. *Principles of International Environmental Law* (4 ed.). Cambridge University Press. https://doi.org/10.1017/9781108355728

[17] Zhang Shuyi, Guo Bin, Dong Anlan, He Jing, Xu Ziping, and Chen Song. Xi. 2017. Cautionary tales on air-quality improvement in Beijing. *Proc. R. Soc. A.* 473 (2017). https://doi.org/doi.org/10.1098/rspa.2017.0457

[18] Liye Song. 2017. Impact Analysis of Air Pollutants on the Air Quality Index in Jinan Winter. In *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, Vol. 1. 471–474. https://doi.org/10.1109/CSE-EUC.2017.89

[19] Colin Tankard. 2012. Big data security. *Network Security* 2012, 7 (2012), 5–8. https://doi.org/10.1016/S1353-4858(12)70063-6

[20] Changhui Yu. 2016. Research of time series air quality data based on exploratory data analysis and representation. In *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*. 1–5. https://doi.org/10.1109/Agro-Geoinformatics.2016.7577697

[21] Paul Zikopoulos, Chris Eaton, and IBM. 2011. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* (1st ed.). McGraw-Hill Osborne Media.