Anaylsis of Air Quality in Guangzhou, China

Patrick Zhang
Department of Computer Science
City, University of London
London, United Kingdom
Patrick.Zhang@city.ac.uk

***Abstract*— Measuring and monitoring air quality is important because air pollution is harmful to people's health and the environment. In this report, we will investigate an air quality dataset of Guangzhou, a city located in southern China, and analyze and discover features that contribute to either excellent or poor air quality. Afterwards, we will investigate which districts are the most polluted and identify any correlation between features such as the month of the monitoring date or the monitoring point that determine air quality in general. A random forest model was created to see how reliable it is to predict the air quality category given the features.**

Keywords—air quality, air pollution, classification, correlation, random forest

I. Introduction

Air pollution is one of the largest issues China is facing to this present day. The Beijing-Tianjin-Hebei region, Central China, and the Yangtze River Delta are usually the areas of interest due to the environmental issues present. Several studies have shown that urbanization is a negative driving factor to China's air quality [1]. There have been efforts to reduce air pollution. For example, in 2014, the Chinese government announced a war with pollution and policies like the *Environmental Protection Law* went into effect [2]. As a result of the policies made, there was significant improvement to the air quality. Many studies said that urbanization had a negative impact. In the case of China, there has been both negative and positive effects to the air quality and that urbanization [3].

Guangzhou is the capital and largest city of Guangdong Province in China. Located in the Pearl River Delta and approximately 120 km north-northwest of Hong Kong, it's the 5th populous city in China. Air quality in the city was heading in the right direction. In 2019, the annual average of particulate matter (PM2.5) was approximately 30 μg/m^3, which met a secondary national standard. Despite that, there were seasonal fluctuations. For example, during the winter, the highest concentration of particulate matter (PM2.5) was 75 μg/m^3 and half of the winter days reached the national air quality standard [4]. Overall, this report aims to provide insights about the city's air quality and the factors that contribute to it.

II. Research Questions and Analytical Approach

2.1 Data and Domain

The dataset used in this report was acquired from the Guangzhou Municipal Government Data Unified Open Platform and was prepared by the Guangzhou Municipal Ecological Environment Bureau. Air quality data was collected from April 2022 to July 2022 and there are 10,000 rows and 8 columns. Columns included AQI, createTime, id, updateTime, monitoring point name, air quality category, primary pollutant, and monitoring date.

There were some limitations that came up while observing the dataset and during the analysis. There was the absence of coordinates for all monitoring points that would be beneficial in the analysis, such displaying on the map which monitoring points had the highest average AQI. In addition, there was a column titled *primary pollutant*, but no columns that listed the concentrations of each individual pollutant such as nitrogen dioxide and fine particulate matter.

2.2 Research Questions

The first step taken will be to present insights via with visuals to detect any patterns that contribute to Guangzhou's air quality. Afterwards, a correction analysis and modeling will be conducted to figure out which are the most important features that affect the air quality. We present some questions that are the focus of this report:

1. Which districts are the most polluted during the period?
2. Does the month of the year affect the air quality index (AQI)?
3. What pollutants are present in the most polluted areas in each district?
4. What is the most important feature to determine the air quality category?

2.3 Analytical Process

In order to investigate the research questions, a plan was drafted. Here are the steps taken:

1. Download the dataset
2. Clean the dataset
3. Visualize the data
4. Analyze patterns from the data
5. Correlation analysis and modeling
6. Results and conclusions

III. Analysis

3.1 Data Preparation and Derivation

After examining the dataset, the first step taken is to drop the columns that are irrelevant to the analysis. The createTime, id, and updateTime columns were dropped as they weren't useful for the analysis.

A closer examination of the dataset shows that there are two measurements taken at most of the monitoring points at a given day between April 17th and June 30th. Prior to dropping the irrelevant columns, each measurement is assigned a unique ID number. For example, on June 7, 2022, two measurements were taken at Baiyun Mountain, and both had an AQI of 217. It's worth noting that one measurement's ID is 432A0CF9-A96C-4F0D-2AA5-876878381801 while the other is 5E3614D3-CD6A-8822-1F7D-84109425CC05. It's possible that two measurements were taken to validate the measurement. However, having duplicate measurements can skew the AQI distribution. To ensure there are unique values in the data, the duplicates will need to be removed.

The dataset is in Chinese. To get an understanding of what the remaining column names and row values are in English, the dataset will need to be translated. The following table shows each column name in Chinese and English.

TABLE I. COLUMN NAMES IN CHINESE AND ENGLISH

Chinese	***English***
AQI	AQI
监测点位名称	monitoring point name
空气质量类别	air quality category
首要污染物	primary pollutant
监测日期	monitoring date

After the translation is complete, final checks for the dataset were conducted. Tasks included checking data types and removing NaN values and outliers. There were some rows where the AQI was -99, meaning that there was a possible error that occurred when taking the measurement. These were removed to improve the distribution of the dataset, as shown in Figure 1.

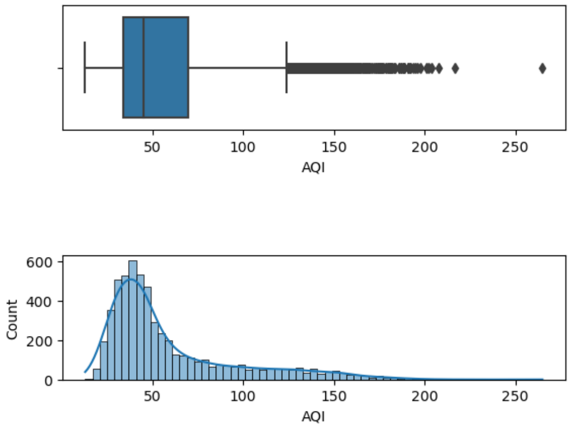


Fig. 1. Distribution of AQI after cleaning the data

Each measuring point will be categorized by district as this will be crucial to the analysis, leaving to creating a new column called *District*. Some of the monitoring points have the district name in them and a web search was conducted to find out the location of the remaining monitoring points. City records were separated from the data frame and placed into a new dataframe as the analysis focused mostly on the districts. That way, we can view any differences in statistics between the districts and the city like the average AQI. Another dataframe will be derived from the original translated dataframe for correlation analysis and modeling, where the month and day of the monitoring date will be taken into consideration along with the other features. These were chosen instead of the whole monitoring date because all measurements were taken in 2022, so the year is irrelevant in the analysis. We wanted to see if both the month and day has any impact to the other features because different conditions at a given day can impact the air quality. Before the prepared dataset was used for correlation analysis, NaN values had to be dropped and that affected the rows, where the air quality category is marked as excellent. As a result, there were only 4 of the air quality categories will be predicting.

3.2 *Methodology*

Majority of the data analysis phase included graphic visualizations. This allowed clear representations between selected features such as district vs. AQI and monitoring point name vs. primary pollutant. A correlation matrix was created to see how strong the relationship between two features is. Finally, a machine learning model was applied to the prepared dataset. Random Forest was selected due to its simplicity and ability to train faster.

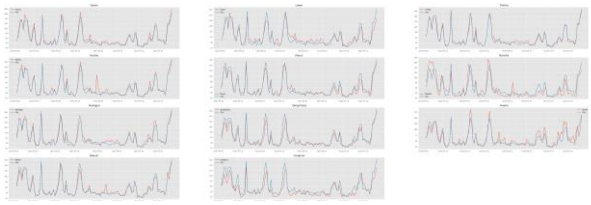


Fig. 2. Plot of average AQI of each district vs the city average AQI over time

In Figure 2 above, the average AQI of each district was plotted against the city's average AQI over time. As seen in the plots, most of the high AQI averages occurred during May and July. A closer inspection shows that Nansha, Huadu, and Yuexiu had averages higher than the city's average in May. In June, all districts and the city had steady averages until the end of the month.

3.3 *Correlation Analysis*

We want to observe if there was any correlation between the features presented and any possibility to predict the air quality category with a machine learning model. A random forest model was chosen for this study. It was chosen due to its simplicity and ability to train faster than other methods. RAQ is a method based off random forest and was developed to predict air quality in urban settings. Yu *et al.* completed a study with RAQ with real time air quality data and concluded that its performance is better that other machine learning methods.

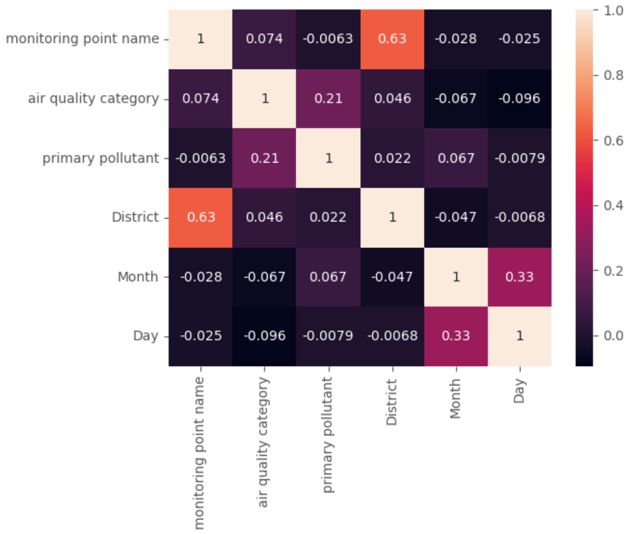


Fig. 3. Correlation matrix between features

The above matrix in Figure 3 shows the correlation between the features. AQI wasn’t considered in the correlation analysis after a few trial runs generating the confusion matrix. Regardless of the test size when splitting the dataset into training and testing data, the accuracy score came out to either 100% or close to 100%, which is unlikely in any setting. Aside from a perfect accuracy score, the feature score for AQI was higher than the other features.

There was little correlation between monitoring point name and air quality category, month and primary pollutant, and district and air quality category. Higher values of correlation only existed between monitoring point name and district, day and month, and air quality category and primary pollutant.

IV. Findings and Discussion

4.1 Findings from Analysis

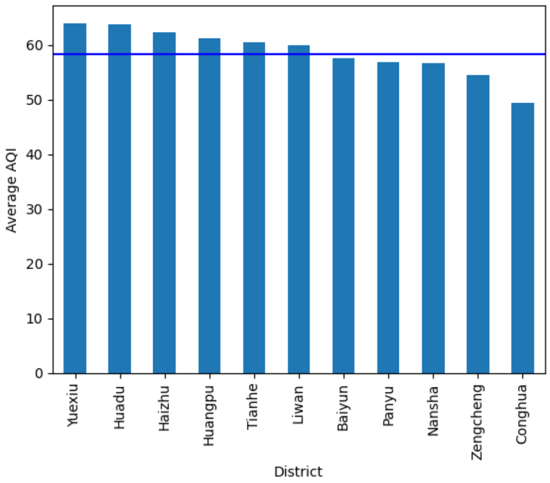


Fig. 4. Bar chart showing overall average AQI of each district compared with city average

The average AQI was calculated and then compared with the city average, as shown in Figure 4. Yuexiu, Huadu, Haizhu, Huangpu, Tianhe, and Liwan had AQI averages higher than the city. Further examination looked into each of the monitoring points to find out which ones contributed to the higher average. 28 of the 53 monitoring points had averages above the city’s. All 28 points were discovered in all districts except Conghua. Huangpu, Tianhe, and Panyu each had 4 measuring points with high averages. Huadu had 2 measure points that came out on top.

The analysis focused mostly on measuring points with pollution. The monitoring points considered for analysis had either light, moderate, or heavy pollution (Figure 5). 6 entries came up for heavy pollution and there were two entries for Huadu while Nansha, Zengcheng, Baiyun, and Haizhu had only one entry each.

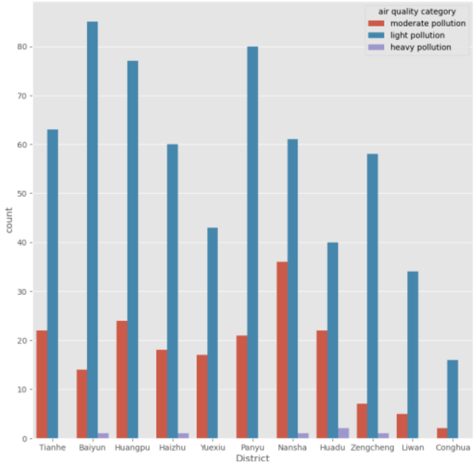


Fig. 5. Grouped bar chart showing count of monitoring points that reported pollution

No pollutants are present at the monitoring points where there was excellent air quality. 14 entries showed that there were two pollutants present at the time and most of them were in Huangpu District. The air quality was categorized as “good”. The Yanji Roadside Station in Yuexiu District and Huangsha Roadside Station in Liwan District appeared twice, but the measurements are taken at different dates. Closer examination showed that between measurements, the Yanji Roadside Station AQI remained constant while the Huangsha Roadside Station AQI increasde by 13. The heavy pollutated monitoring points had Ozone 8h as the primary pollutant.

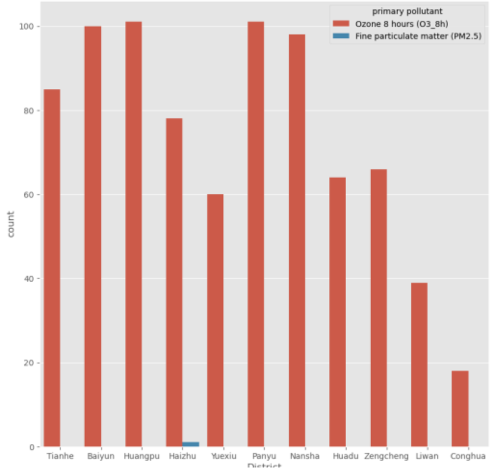


Fig. 6. Group bar chart showing count of monitoring points and the primary pollutant present.

Figure 6 shows how many monitoring points with pollution discovered a pollutant. All the points had either Ozone 8h or fine particulate matter (PM2.5) reported. Haizhu Lake in Haizhu District is the only monitoring point to have fine particulate matter (PM2.5) in the air at the time of measuring.

4.2 *Modeling Outcomes*

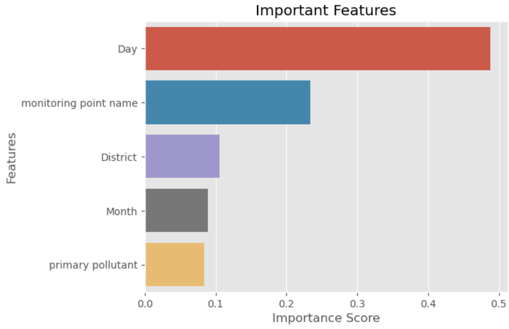
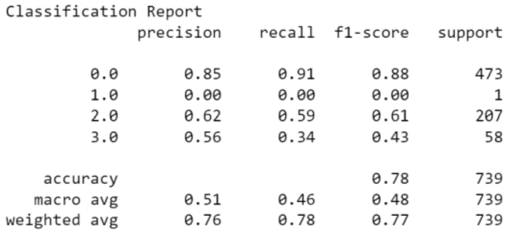


Fig. 7. Graph showing importance score of each feature from random tree model

Figure 7 shows the importance score of each of the features. The day came out on top, which answered our fourth research question.



```
Classification Report
              precision    recall  f1-score   support

         0.0       0.85      0.91      0.88       473
         1.0       0.00      0.00      0.00         1
         2.0       0.62      0.59      0.61       207
         3.0       0.56      0.34      0.43        58

    accuracy                           0.78       739
   macro avg       0.51      0.46      0.48       739
weighted avg       0.76      0.78      0.77       739
```

Fig. 8. Classification report for random tree model

As a result of fitting the dataset into the model, the overall accuracy score came out to 78%. There were zeroes for one label, meaning the model couldn't provide predictions for it.

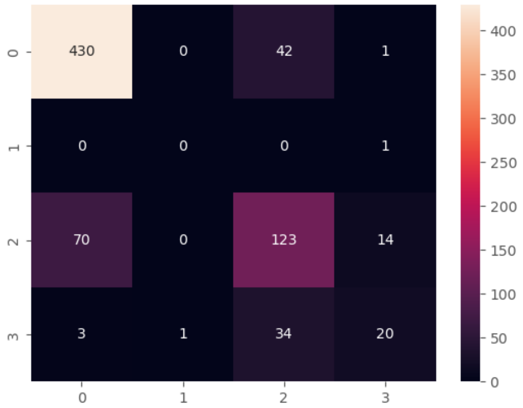


Fig. 9. Confusion matrix for random tree model.

V. Relections and Further Work

With an average AQI of approximately 58.3, Guangzhou's air quality is classified as good. After the visual analysis, we can conclude that Huadu and Yuexiu are the most polluted districts. Ozone 8h was the primary pollutant detected in monitoring points where heavy pollution was reported.

The model used in this report was able to predict and classify 78% of the measurements correctly. The model did perform decently and with fine tuning and improved feature selection, the accuracy score can improve. This can be taken a further step by comparing the accuracy score of this model with other machine learning models.

Further investigation can be done by considering the concentration of different pollutants at each different monitoring point and what other factors contribute to the air quality such as the number of factories operating in each district or how many vehicles are on the road at a given day. An extra step is to compare Guangzhou to other Chinese cities like Beijing or Shanghai.

References

[1] Bao, R. and Liu, T. (2022). How does government attention matter in air pollution control? Evidence from government annual reports. *Resources, Conservation and Recycling*, 185, p.106435. doi:10.1016/j.resconrec.2022.106435.

[2] Greenstone, M., He, G., Li, S. and Zou, E.Y. (2021). China's War on Pollution: Evidence from the First 5 Years. Review of Environmental Economics and Policy, 15(2), pp.281–299. doi:10.1086/715550.

[3] Liu, H., Cui, W. and Zhang, M. (2022). Exploring the causal relationship between urbanization and air pollution: Evidence from China. *Sustainable Cities and Society*, 80, p.103783. doi:10.1016/j.scs.2022.103783.

[4] Yao, Y., Wang, Y., Ni, Z., Chen, S. and Xia, B. (2022). Improving air quality in Guangzhou with urban green infrastructure planning: An i-Tree Eco model study. *Journal of Cleaner Production*, 369, p.133372. doi:10.1016/j.jclepro.2022.133372.

[5] Yu, R., Yang, Y., Yang, L., Han, G. and Move, O. (2016). RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. *Sensors*, 16(1), p.86. doi:10.3390/s16010086.

Section	***Word Count***
Abstract	105
Introduction	241
Data	294
Analysis	828
Findings, reflections, and further work	512