# Weather Pollution Prediction Based on XGBoost and ARIMA: Evidence from Shanghai

Jingkun Yu

Joint Institute, Shanghai Jiao Tong University, Shanghai, China

yujingkun@sjtu.edu.cn

**Abstract.** As a matter of fact, the weather pollution prediction is a pretty important issue in environment analysis. With the rapid development of machine learning techniques and bigdata analysis, it is available to realize accurate prediction based on the state-of-art models. In recent years, the weather condition is pretty worse in Shanghai. With this in mind, the general situation of air quality in Shanghai was statistically analyzed using air pollution detection data such as API in Shanghai from 2014 to 2023, and the relationship between air quality influencing factors and the air quality index API was investigated based on the construction of the machine learning model and scenarios. With the analysis and the usage of XGBoost and ARIMA, one can get the future trends and prediction. According to the analysis, the impact factors are discussed as well as the current limitations are demonstrated. At the same time, the future prospects are discussed.

**Keywords:** Air quality, XGBoost, ARIMA, prediction.

## 1. Introduction

Air quality, which is a social indicator of human survival and development, has attracted particular attention in recent years. China used to face severe air pollution, but with the effort of air pollution control and environmental protection, the air quality has been improving [1]. Shanghai, as a cosmopolitan city and one of the largest cities in China, has an extremely important economic position and deserves a good environment. In fact, the impact of air quality on all aspects of a city is much higher than one might think. Air quality has a seemingly unrelated but actually huge systemic impact not only on residents' quality of life, but also on transportation, industry and even housing prices [2].

In order to rate air quality, the scientific community has introduced a measure, i.e., Air Quality Index(AQI) [3, 4]. Hence, studying the AQI of Shanghai is of great significance to learn the overall change of air quality. Prior to this, there has been a great deal of statistical and research on air quality in China, especially on macro-regional studies. However, there is not much research on the trend of air quality change in Shanghai, especially on the prediction of future trend. Besides, XGBoost, as a kind of AI analysis and prediction model emerging in recent years, is more and more widely used.

To predict AQI changes, one needs to introduce some AI models. Moreover, the XGBoost model and ARIMA model are two typical models of predicting This research is based on the air quality data published on the Shanghai Real-Time Air Quality Website and the Shanghai Weather Historical Database. The study will look at indicators such as the AQI (Ambient Air Quality Index) to reflect the air quality situation in Shanghai.

## 2. Data and Method

The data are taken from Shanghai's monthly air quality data from 2014 to 2023, including AQI, PM2.5, PM10, NO2, SO2, ozone and other indicators, with the first 7 months of data taken in 2023. The average AQI of air quality in Shanghai between 2014-2022 is 75.23, which is good, and 102 out of 108 months have an average AQI between 50-100, which is good quality, and four months have an average API greater than 100, which is mildly polluted, while two months have an average API less than 50, which is excellent quality:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{113} X_i = 75.23 \tag{1}$$

And the variance of the monthly average AQI is 260.76.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{113}(X_i - \bar{X})^2 \approx 260.76 \tag{2}$$

By the law of large numbers, the sample moments converge in probability to the overall moments, and the sample moments can be used as an estimate of the corresponding overall moments when the sample size is large. That is, the kth-order origin moments of the sample are used to estimate the overall first-order origin moments, and the second-order center moments of the sample can also be used to estimate the overall second-order center moments, thus, the overall parameters obtained by the method of moment estimation can be used. The moment estimates were 75.23 for the mean and 260.76 for the variance $\hat{\sigma}^2 = \sigma^2 = 260.76$. The sample variance was 263.09.

This study examines air quality in Shanghai for four quarters and individual months from 2014 to 2022. The air quality was found to be poor in the months of December, January, April and May, while the air quality was better from August to November. Analyzing the reasons, there are reasons for polluted air from December to January due to heating in the north [6], and also due to increased electricity consumption and number of people using electric heating in Shanghai [7]. While April and May had poorer air quality mainly due to dust storms, the air quality was better on dates without dust storms [8]. The relationship between daily PM2.5 content and temperature, humidity and wind magnitude in Shanghai in 2021-2023 was obtained from the data platform and made into graphs with the following results:

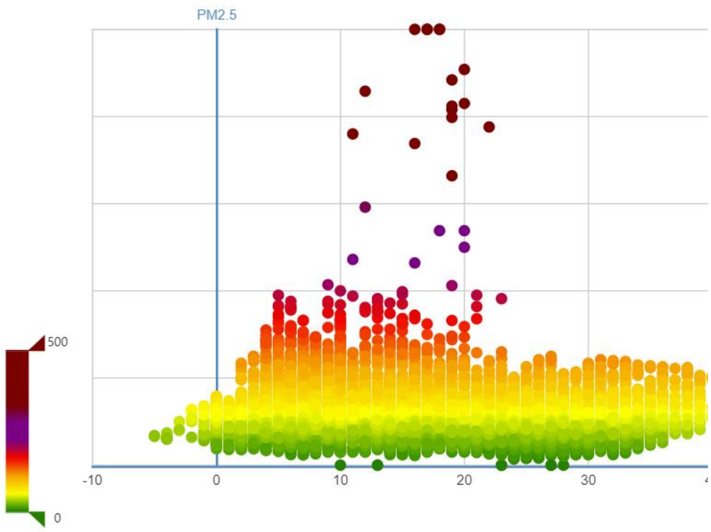


**Figure 1.** PM2.5 and temperature

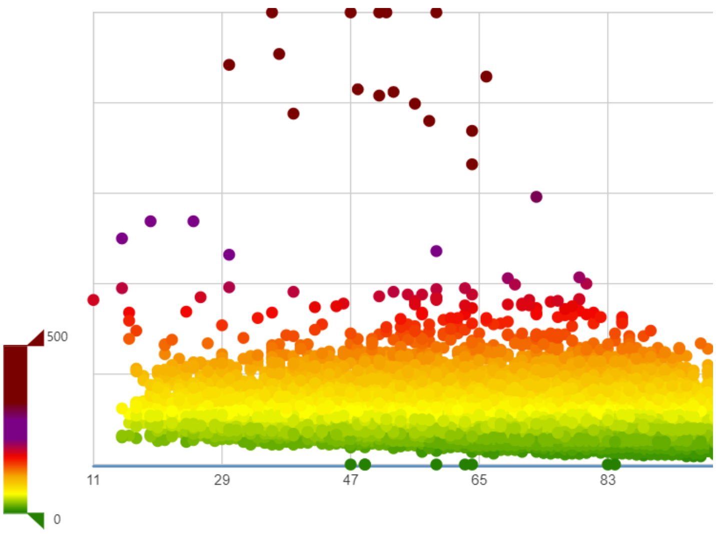


**Figure 2.** PM2.5 and Humidity

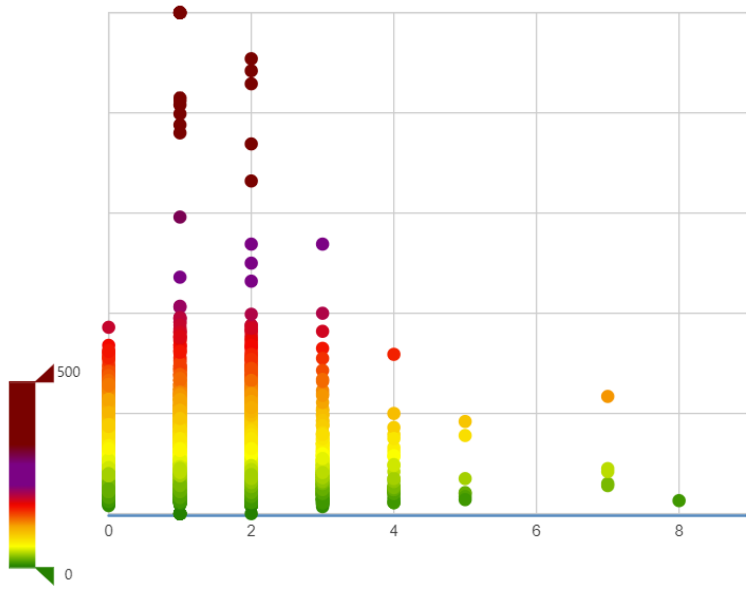


**Figure 3.** PM2.5 and wind power

It can be inferred from the Fig 1, Fig 2 and Fig3 that extreme heavy pollution (PM2.5 levels >200) is less likely to occur when the mean daily temperature is greater than 30°C or less than 10°C. After an independence test, this conclusion has >99% confidence. Humidity >80, heavy pollution is difficult to occur, and reasonable by independent tests. With winds greater than 4, heavy pollution is difficult to occur, and the independence test yields a confidence level of >99%. Averaging Shanghai's air pollution from 2014 to 2022, the average AQI (ambient air quality index) for Shanghai over the 10-year period is 75.2, compared to the average AQI over the years (seen from Table 1).

**Table 1.** Summary of AQI

| particular year | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|---|---|---|---|
| AQI | 80.3 | 88.5 | 80.8 | 83.7 | 70.2 | 72.5 | 67.9 | 63.8 | 66.0 |
| quality level | very much | very much | very much | very much | very much | very much | very much | very much | very much |

It can be noticed that there is a significant drop in API after 2018, from an average of around 80 to an average of around 70, while the average API for all three years from 2020 to 2022 is below 70. Counting the average API for each month from 2014 to 2022, the vast majority of the quality ratings are good, while the five months of January 2014, December 2015, January 2015, January 2016, and July 2017 have an average API of over 100, which is considered to be mildly polluted; and the two months of 2022, October and November, have excellent air quality.

The relationship between AQI and each indicator was taken for the calendar years 2014-2022 and linear correlation was analyzed in origins softwarewith IC 0.79671. The AQI is calculated by calculating the IAQI (Converted Air Quality Index) for each of these six indicators and taking the maximum value, so the linear fit is calculated to reflect the degree of correlation between each indicator and the overall pollution. Through calculations, it is found that although all coefficients are positively correlated, the linear correlation coefficients of PM2.5, PM10, CO, and SO2 with API are higher, above 0.7, while the linear correlation coefficients of NO2 are not very strong, around 0.5, and the linear correlation of O3 indicator is not very strong, only around 0.2.

The models used in this study are mainly XGBoost and ARIMA, two efficient models that are widely applied in all sorts of prediction, such as time-series prediction. GBDT models, including XGBoost, are considered advanced and practical in many occasions, and works quite well in many online competitions to bring competitors to the top of the table [9]. Meanwhile, ARIMA is a popular and widely-used time series forecasting and analysis technique used in statistics. It is a powerful and flexible method for modeling and predicting time series data, which are sequences of data points measured at successive points in time [10]. Therefore, one considered to analyze the air quality of Shanghai separately, to better reflect the overall trend of air quality in Shanghai, as well as compare the applicability of the models.

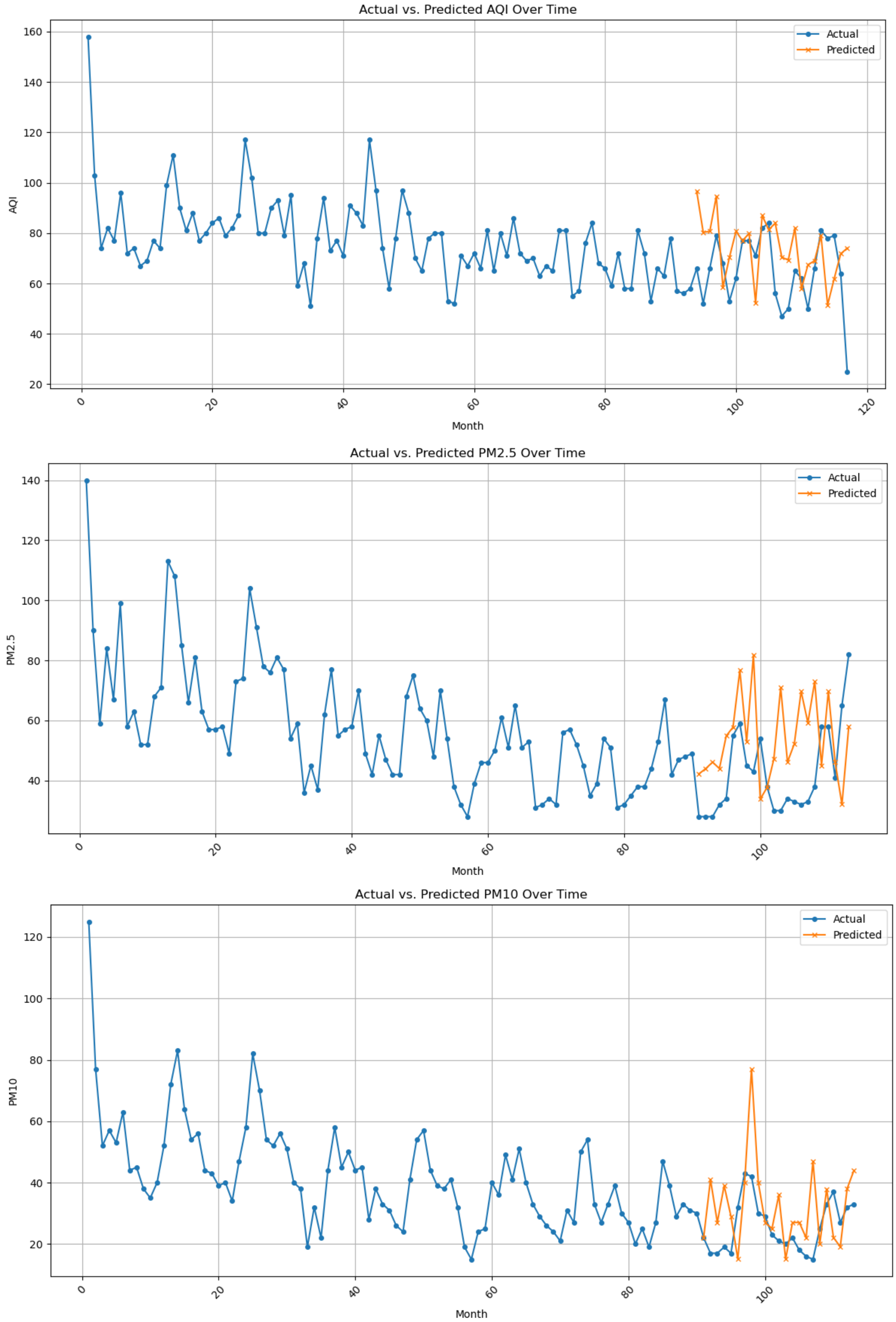


**Figure 4.** Prediction based on XGBoost

## 3. Results and Discussion

One important thing about XGBoost is that it generates random numbers and split buckets, so one can test the effect of individual seeds. Fig. 4 shows the results for the better fitting seeds over a range.

For the ARIMA model, the important point is to adjust its functional form and parameters so that a better fit can be achieved (seen from Fig. 5). Although one have a rich and complete variety of samples, the sample size is still insufficient and is only accurate to the monthly average; and there are still questionable adjustments to the model, especially the division of the buckets for XGBoost.

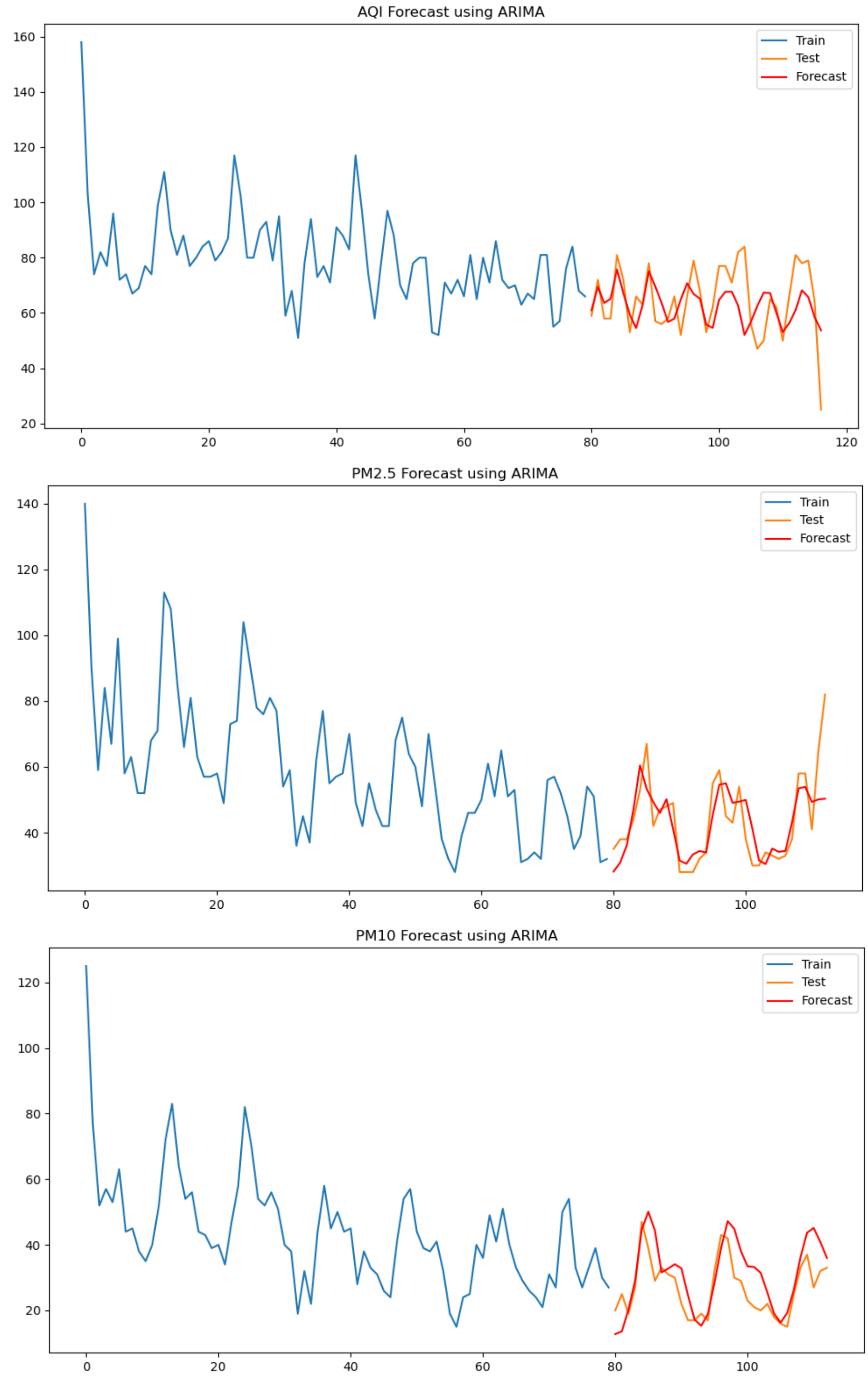


**Figure 4.** Prediction based on ARIMA

## 4. Conclusion

To sum up, according to the analysis, the average AQI for air pollution in Shanghai is around 75, with the vast majority of cases being good. Shanghai air quality is worse in winter and spring than in summer and fall, presumably due to northern heating and dust storms. Serious air pollution is difficult to occur when the temperature is higher than 30°C or lower than 10°C, when the humidity reaches 80 or more, or when the wind reaches level 4 or more. Severe air pollution is likely to occur when the temperature is moderate, the humidity is not too high, and the wind is not strong. Shanghai's air pollution has been improving year by year, with significant improvement after 2018 and a significant reduction in API. Among the indicators affecting API in Shanghai, PM2.5, PM10, CO, and $SO_2$ have high linear correlation coefficients with API and strong correlation, while $NO_2$ and ozone do not have high correlation coefficients with API. The ARIMA model still has a very strong advantage in predicting timescales, especially regarding samples with distinct patterns and small sample sizes, whereas the XGBoost model relies more on adjustments to the model, such as dividing the buckets, and is not as adaptable to small samples.

## References

[1] Yuan G, Yang W. Evaluating China's air pollution control policy with extended AQI indicator system: Example of the Beijing-Tianjin-Hebei Region. Sustainability, 2019, 11 (3): 939.

[2] Wu J. Air pollution on real estate price. Shanghai University of Finance and Economics, 2022.

[3] Zhang X. Does asymmetric persistence in convergence of the air quality index (AQI) exist in China. Environmental Science and Pollution Research, 2020, 27: 36541-36569.

[4] Duangsuwan S. 3D AQI mapping data assessment of low-altitude drone real-time air pollution monitoring. Drones, 2022, 6 (8): 191.

[5] Wan Q, Luo X, Pan F, et al. Evolution of China city air quality and convergence trend. Geography Science, 2022, 42 (11): 1943-1953.

[6] Chen R, Ma J, Qu Y, Chang F. Characterization of air quality change and analysis of meteorological influences in Shanghai region in 2019-2020. Meteorological Research and Application, 2022, 43 (01): 59-65.

[7] Zhang T. Shanghai's air quality index excellence rate will stabilize at 85% in 2025. Wen Wei Po, 2021, 8 (19): 007.

[8] Li J. Measurement and analysis of spatio-temporal correlation of air quality in China. Shanghai Normal University, 2018.

[9] Sagi O, Lior R. Approximating XGBoost with an interpretable decision tree. Information Sciences, 2021, 572: 522-542.

[10] Benvenuto D. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief, 2020, 29: 105340.