Air Pollutants Prediction in Shenzhen Based on ARIMA and Prophet Method

Ziyuan Ye

Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, Guangzhou, 518000, China

Abstract. In order to improve the accuracy of predicting the air pollutants in Shenzhen, a hybrid model based on ARIMA (Autoregressive Integrated Moving Average model) and prophet for mixing time and space relationships was proposed. First, ARIMA and Prophet method were applied to train the data from 11 air quality monitoring stations and gave them different weights. Then, finished the calculation about weight of impact in each air quality monitoring station to final results. Finally, built up the hybrid model and did the error evaluation. The result of the experiments illustrated that this hybrid method can improve the air pollutants prediction in Shenzhen.

1 Introduction

These Due to the rapid development of the city, air quality problem has become an increasingly basic issue in urban sustainable development. The polluted air contains many toxic or harmful substances including PM_{2.5}, PM₁₀, NO₂, SO₂, O₃, CO and so on [1] which may give rise to illness, allergies and even death [2]. Poor air quality in urban city increases the risk of respiratory disease and cardiovascular disease [3]. According to an air pollution and mortality research, the mortality rate has a great relationship with the increase of and (see Table 1 [4]). Air pollution not only brings disease and death, it also adds a heavy economic burden to society. A research, estimated the indirect costs associated with lost productivity caused by premature death associated with the 2016 Colombian air quality risk factor have found that the economic burden caused by premature death from the disease being

analyzed was \$845,967,999. From this burden, air risk quality factors accounted for 17.8%, corresponding to US \$150,585,143 [5].

Monitoring and predicting air quality are very important to control and prevent the bad effects of air pollution [6]. Thus, if the forecasting air quality result can be swift and accurate, it will not only reduce a large economic loss, but also contribute to help people with preventing bad air in advance and reducing the occurrence of related diseases.

The Pearl River Delta is one of the regions where China's population and economy are highly concentrated. Environmental issues are highlighted and urban air quality problem is receiving much attention. The Pearl River Delta is known as one of the three major acid rain regions in the world, and Shenzhen, as part of it, naturally faces severe urban air quality problems [7].

Table 1. Risk of Death Associated with an Increase of 10 µg per Cubic Meter in PM_{2.5} or an Increase of 10 ppb in O₃ Concentration

Type	Model	PM _{2.5}	O ₃
Two-pollutant analysis	Main analysis	1.73(1.071–1.075)	1.011(1.010–1.012)
	Low-exposure analysis	1.136(1.131–1.141)	1.010(1.009–1.011)
	Analysis based on data from nearest monitoring site (nearest-monitor analysis)	1.061(1.059–1.063)	1.001(1.000–1.002)
Single-pollutant analysis	Not mention	1.084(1.081–1.086)	1.023(1.022–1.024)

2 Related Work

The classic time series feature representation model is the primary method for analysing and predicting air quality. In the prediction of single air pollutants, Hasan and Zhao use gradient boosting, artificial neural networks model and k-nearest-neighbors model to predict the

concentration of PM_{2.5} in a rural area of Hong Kong. The result illustrates that gradient boosting has higher prediction accuracy and stability [8]. Granular transformation can change the solution space of air quality prediction problem. Wang used a dynamic granular wavelet neural network model to predict air quality and got high accuracy [9]. In the prediction of multiple air

*Corresponding author's e-mail: chunchenxiao@cas-harbour.org

pollutants, Goyal and Kumar use neural network model based on principal Component Analysis (PCA) to predict air quality index in Delhi [10]. As the result shows in their article, the combination of PCR and ARIMA model can successfully cope with the issue of exploiting the autocorrelation and collinearity in the variables.

The traditional predicting air quality method almost all based on air pollutants or together with local meteorological data [11][12][13] (temperature, humidity, wind, weather, etc). The models based on mathematical statistics or machine learning or deep learning have been done comprehensively in time series. However, these methods analyse the time series directly but ignore the effects of spatial factors. We human live in a space rather than a time line which means that local air quality is not only related to historical air quality and historical meteorological data. It is also related to the air quality and meteorological data of nearby cities. If the method only includes time series data, predict data will highly depends on the stability and regularity of real data. Besides normal fluctuations, air quality can also cause dramatic changes due to certain space factors. For instance, if there is a sudden dust explosion in Zhongshan city, the air quality in Shenzhen will be affected definitely. Although this spatial impact is not common, it still brings great difficulties to the prediction of air quality. Using the RNN and LSTM model with 3-Nearest area's PM_{2.5} and weather data, the air quality prediction discover that DRNN pairs based on forward complement has the best prediction ability for time series and have good predictive and generalization abilities [14]. Based on the air quality data and meteorological data obtained by air quality monitoring stations and meteorological data monitoring stations in Beijing, Guo[15] obtained a method to predict the PM_{2.5} concentration of each air quality monitoring station in the next 48 hours by using the Tensorflow framework together with RNN, LSTM and GRU network and integrating the data of surrounding monitoring stations. At present, there are few ways to predict the air quality by adding space factors. It still needs a long way to go.

This project aim at finding the relationship between spatial factors and air quality and constructing a reliable model based on time series data and spatial data to predict air quality. Additionally, according to the result of air quality predict model, the weight in different models will be analysed. For example, the AQI model and PCR model have difference attribute and both can predict the air quality. Figuring out the suitable weights in the mixed model can make the result has better stability, robustness and anti-disturbance than any single system.

3 Proposed Method

During the research, the ARIMA model and prophet model have been implemented. Two algorithms both improve the predicting results. However, the two models have different prediction accuracy for different air pollution particles. Thus, in my final model, I use a mixture of the two to make weather forecasts. In the following subsections, these two methods will be briefly

introduced. The framework of my model will be explained also.

3.1 ARIMA method

The Autoregressive Integrated Moving Average model (ARIMA) was proposed by Box and Jenkins in the 1970s and is a well-known time series prediction model. It is based on Autoregressive Moving Average model (ARMA). In ARMA model, the current value of the time series is linearly represented as its previous value as well as the current value and the previous residual sequence [16]. For non-stationary time series, the sequence must first be converted to a stationary random sequence. The differential process effectively converts non-stationary data into smooth data. The meaning of ARIMA is the single product autoregressive moving average process, which means: A stochastic process is said to be a single product autoregressive moving average process if it has d unit roots and can be transformed into a stationary autoregressive moving average process after d - difference. Then the stochastic process is called the simple product autoregressive moving average process. The whole process of the ARIMA can be divided into the following steps:

(a) Judging whether the stochastic process converges after d-difference.

(b) After finding a suitable d, X_t is transformed into a stationary stochastic process $\Delta^d X_t$.

(c) Constructing $\Delta^d X_t$ as autoregressive moving average process.

For I(d) is to get the stationary sequence by the difference of the nonstationary sequence.

For AR(p), in general, time series variables have temporal correlation so the next few time points can be predicted based on the observations of the last few points in time which is called AR(p) model.

$$Y_t = \epsilon_t + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} \quad (1)$$

For MA(q) model, the Lower order MA(q) model can replace higher order AR model.

$$Y_t = \epsilon_t + \beta \epsilon_{t-1} \quad (2)$$

In air pollutants prediction, the key point is to figure out the optimal ARIMA order (p, d, q) and seasonal ARIMA order (P, D, Q, S). The grid search method is implemented in order to explore different combinations of parameters iteratively.

The AIC is used to assess the complexity of the accounting model and to measure the good fit of the statistical model. After doing the iteration for PM_{2.5}, PM₁₀, NO₂, SO₂, O₃ and CO separately, the optimal order of (p, d, q) and (P, D, Q, S) can be obtained with lowest AIC value. Once get the optimal sequence of (p, d, q) and (P, D, Q, S) the model can be built.

3.2 Prophet

The package of fbprophet is a time series prediction algorithm that Facebook introduced and opened up to the public in 2017.

Prophet algorithm can not only handle the outliers of time series but also some missing values. Most importantly, it can almost automatically forecast the future trend of time series [17]. In the prophet, there are two important function which are logistic function and piecewise linear function.

The traditional logistic function is as follows:

$$f(x) = \frac{C}{1 + e^{-k(x-m)}} \quad (3)$$

However, this function is too ideal. In the real world, C , k and m are definitely not constant. They are likely to update over time. Thus, in prophet the function becomes like follows:

$$f(x) = \frac{C(t)}{1 + e^{-k(t)(x-m(t))}} \quad (4)$$

After considering the change point detection, the trend of the function will be change with the S change point:

$$f(x) = \frac{C(t)}{1 + e^{-(k+\alpha(t)\delta) \cdot (t-(m+\alpha(t)^T\gamma))}} \quad (5)$$

where the α, γ and δ are like follows:

$$\alpha(t) = (\alpha_1(t), \dots, \alpha_s(t))^T \quad (6)$$

$$\gamma = (\gamma_1, \dots, \gamma_s)^T \quad (7)$$

$$\delta = (\delta_1, \dots, \delta_s)^T \quad (8)$$

For piecewise linear function the model is as follows:

$$f(t) = (k + \alpha(t)\delta) \cdot t + (m + \alpha(t)^T\gamma) \quad (9)$$

where k is the growth rate, the change of growth rate and m represents offset parameter.

3.3 The model building

Shenzhen has 11 air quality monitoring stations, so the models were trained on each station separately. During the training section, the accuracy of predicting result of CO and SO₂ with prophet method is better than ARIMA model.

Based on training, I gave different weights to the prediction results from the two models to get the final prediction results.

$$M_{\text{final}}(t) = \alpha M_{\text{ARIMA}} + (1 - \alpha) M_{\text{prophet}} \quad (10)$$

For the final result of the Shenzhen air pollutants are calculated by different stations. The distribution of every air quality monitoring station in Shenzhen is shown in the Fig. 1. The blue dot is an air quality monitoring station and the red dot is an average value computed from the longitude and latitude of all monitoring stations.

Figure 1. Distribution of every air quality monitoring station in Shenzhen

The closer the distance between the air detection station and the centre point, the higher the influence of the station on the centre point. The distance D_i between the station _{i} and center station can be calculated with sum of squares of difference of longitude and dimension under square sign. Thus, the proportion of weights ω_i of station _{i} at the centre point can be defined as follows:

$$D_i = \sqrt{(\text{longitude}_{e_i} - \text{longitude}_{e_{\text{center}}})^2 + (\text{latitude}_{e_i} - \text{latitude}_{e_{\text{center}}})^2} \quad (11)$$

$$\alpha \left(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_i} \right) = 1 \quad (12)$$

$$\omega_i = \alpha \frac{1}{x_i} \quad (13)$$

In this equation, ω_i will has a larger value if the station _{i} and center are closer.

The final prediction of air pollution X is based on the following formula:

$$X_{\text{final}} = \omega_1 X_1 + \dots + \omega_i X_i \quad (14)$$

4 Results and Discussion

4.1 Datasets

The datasets this research used are the air quality information in Shenzhen from 2016/01/01 to 2017/01/31. The datasets include data from 11 air quality monitoring stations in Shenzhen. Each station has monitored PM_{2.5}, PM₁₀, NO₂, SO₂, O₃, CO data.

4.2 Result

In the **Table 2**, the three-days air pollutants prediction mean square error (MSE) and mean absolute error (MAE) are as follows:

Table 2. Three -day air pollutants prediction error.

	PM _{2.5}	PM ₁₀	NO ₂	SO ₂	O ₃	CO
MSE	107.83	240.73	138.46	4.35	3.85	0.01
MAE	23.75	19.50	24.68	18.58	2.36	11.77

Figure 2 represents the final forecast result of 6 different air pollutants. As we can see, the difference between forecast trend and true trend of PM_{2.5}, PM₁₀, NO₂, SO₂ and O₃ are small, and the difference between

forecast trend and true trend of CO is obvious. The orange lines illustrate the real data while blue lines are the predict data.

Figure 2. The three-days air pollutants factor prediction result

4.3 Limitations of propose method

Although the forecast results seem acceptable, there still exist some limitations, such as:

- ARIMA is more accurate than Prophet in some factors, but it takes about ten times as long as B. The main calculation time of ARIMA is spent looking for the optimal sequence of (p, d, q) and (P, D, Q, S). However, the time it takes to find the optimal sequence is difficult to reduce.
- When models are used to predict long-term air pollution conditions, their errors can be significant.

5 Conclusion and future work

In this paper, I have proposed a new method for forecasting air pollutants based on sequential data and spatial factors. On the temporal dimension, the main framework is based on ARIMA and Prophet. Given different weight to these two results respectively to get the final sequential result. On the spatial dimension, I propose a weighted method if an air quality monitoring station closer to the centre point then it will cause greater impact on the final forecast data. According to the final result, this hybrid method can achieve good results in the

prediction of short-term air pollutants concentration. The future work will focus on utilizing different method to find out optimal sequence and optimizing the existing spatial model.

References

1. Brunekreef, B. (2007) Health effects of air pollution observed in cohort studies in Europe. *Journal of Exposure Science & Environmental Epidemiology*, 17: S61–S65.
2. Cao, J., Yang, C. X., Li, J. X., Chen, R. J., Chen, B. H., Gu, D. F., et al. (2011) Association between long-term exposure to outdoor air pollution and mortality in China: A cohort study. *Journal of Hazardous Materials*, 186(2-3): 1594–1600.
3. Mirabelli, M. C., Boehmer, T. K., Damon, S. A., Sircar, K. D., Wall, H. K., & Yip, F. Y., et al. (2018) Air quality awareness among u.s. adults with respiratory and heart disease. *American Journal of Preventive Medicine*, S0749379718300692.
4. Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., & Choirat, C., et al. (2017) Air pollution and mortality in the medicare population.

- New England Journal of Medicine, 376(26), 2513-2522.
5. Alvis-Zakzuk, Nelson J. & Jiménez, Diana & Hilarión-Gaitón, L. & Valencia, S. & Gutierrez-Clavijo, J. & Castañeda-Orjuela, Carlos. (2019). PMU35 ECONOMIC BURDEN OF AIR POLLUTION IN COLOMBIA. *Value in Health*. 22: S255.
6. Ni, X. Y., Huang, H., & Du, W. P. (2017) Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data. *Atmospheric Environment*, 150: 146–161.
7. Liu J, Yuan L, Liu Y, et al. (2009) Analysis on Spatial-temporal Characteristics of Air Pollution during a Regional Haze Episode over Pearl River Delta [J]. *Journal of environmental science*, 28(4):78-82.
8. Zhao, Y., & Hasan, Y. A. (2013) Comparison of three classification algorithms for predicting pm2.5 in Hong Kong rural area. *Asian Journal of Scientific Research*, 3(7):715-728.
9. WANG Xiaohan, ZHANG Yanping, ZHAO Shu, et al. (2013) Air quality forecasting based on dynamic granular wavelet neural network. *Computer Engineering and Applications*, 49 (6) :221-224.
10. Kumar, A. , & Goyal, P. . (2011) Forecasting of daily air quality index in delhi. *Science of the Total Environment*, 409(24): 5517-5523.
11. K. Bashir Shaban, A. Kadri and E. Rezk. (2016) Urban Air Pollution Monitoring System with Forecasting Models, *IEEE Sensors Journal*, vol. 16, no. 8, pp. 2598-2606.
12. Contreras-Ochando, L. , & Ferri, C. . (2016) airVLC: An Application for Visualizing Wind-Sensitive Interpolation of Urban Air Pollution Forecasts. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). IEEE.
13. Zhao, C. , & Song, G. . (2017) Application of data mining to the analysis of meteorological data for air quality prediction: A case study in Shenyang. *Iop Conference Series: Earth & Environmental Science*. IOP Conference Series: Earth and Environmental Science.
14. Fan, J., Li, Q., Hou, J., Feng, X., Karimian, H. & Lin, S. (2017) A Spatiotemporal Prediction Framework for Air Pollution Based on Deep RNN. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:15.
15. Guo, H. (2019). Research and implementation on air quality prediction method deep learning based [D]. Beijing University of Posts and Telecommunications.
16. Bartholomew, D. , Box, G. E. P. , & Jenkins, G. M. . (1971) Time series analysis: forecasting and control. *Journal of the Operational Research Society*, 22(2), 199-201.
17. Taylor SJ, Letham B. (2017) Forecasting at scale. <https://doi.org/10.7287/peerj.preprints.3190v2>