# A Multi-step Prediction Method of Urban Air Quality Index Based on Meteorological Factors Analysis

*Yu* Zhang[1], *Mingxiang* Yang[2,*], *Fengyu* Yang[1], and *Ningpeng* Dong[2]

[1]School of Software, Nanchang Hangkong University, Nanchang 330063, China
[2]China Institute of Water Resources and Hydropower Research, Beijing 100038, China

**Abstract.** With the development of science and technology, Industry, transportation and other industries used to discharge a large number of pollutants into the atmosphere, which results in air pollution. When air pollution become serious, it will do great harm to human health. High-precision Air Quality Index(AQI) prediction is as important as weather prediction. People could arrange traveling and their life according to the highly precise prediction results, so as to better protect their own health. Considering a lot of complex factors, we choose several potential meteorological factors and historical data to precisely forecast AQI. The principal component analysis (PCA) is introduced in our method to conduct dimension reduction on nine meteorological factors, in order to reduce noise of data and the complexity of the model calculation, which improves the accuracy of AQI prediction as a result. Then the data of meteorological factors after PCA and historical AQI are input into the multi-step prediction model based on LSSVM to train and refine it. Finally, we set up the experiment with data of meteorological factors and AQI. Experimental results show that the method proposed in this paper has better prediction accuracy over classical ARIMA method and has better generalization than ARIMA method as well.

## 1 Introduction

In recent years, both domestic and foreign countries have paid great attention to air pollution and air quality. With the continuous development of industries such as industry and transportation, the amount of pollution source emissions has gradually increased. Automobile emissions, industrial emissions and residential emissions have become normal, making the normal composition of air occur. Change, which in turn produces air pollution. In 2012, my country officially used AQI (Air Quality Index, AQI) to quantitatively describe air quality. The better the air quality, the smaller the AQI; conversely, the worse the air quality, the greater the AQI.

According to the National Air Quality Status in 2020 issued by the Ministry of Environmental Protection, although the concentration of pollutants has declined by 2020, heavy air pollution has still occurred on a large scale, and the annual average AQI in some cities is as high as 110. The AQI is not only a collection of local evaluation indicators, but also affected by many other factors. Among them, meteorological factors will have a direct or indirect impact on the AQI index, such as pollution sources in surrounding cities on air pollution. Through air circulation, it will directly affect the local AQI value. AQI is a parameter that directly reflects the air quality, so it is necessary to predict the city's AQI with high accuracy, and the multi-step prediction method of AQI can effectively avoid the cumulative error of the system in the single-step prediction, and the prediction effect is better. In this way, people can arrange travel matters in advance based on the AQI forecast results and can take precautionary measures in advance to reduce or even eliminate the harm of air pollution to the human.

Domestic and foreign experts have been studying the impact of meteorological factors on air quality. Research has shown that air quality is closely related to meteorological factors. Sheng Chen and Guangyuan Kan [1] used the one-year data of the six pollutants time series of air pollution and the Air Quality Index (AQI) as model inputs to analyze the impact of air pollutants on the changes in AQI.

More and more scholars take meteorological factors as the main consideration to predict the air quality index. For example: Xin Gao and Xuan Wang [2] analyze and forecast based on PEK and air quality index (AQI) based on machine learning. Jian Xue and Yan Xu [3] use the time series of the O-U model to improve the accuracy of predicting AQI changes. Qunli Wu and Huaxing Lin [4] proposed an optimal hybrid model that combines the advantages of secondary decomposition (SD), artificial intelligence methods and optimization algorithms for AQI prediction. Cyuan-Heng Luo and Hsuan Yang [5] proposed a prediction method called adaptive iterative forecasting (AIF), which can predict the PM2.5 value in the next few hours based on the trend of historical data. In 2018, Yi Lin and Long Zhao [6] proposed a new air quality prediction algorithm based on cloud model granulation (CMG). The cloud model is extracted for each particle from the original data space to the feature space, which can simplify the uncertainty Time series modeling process. In the same year, Yuzhe Yang and Zijie Zheng [7] proposed a Gaussian plume model based on neural network (GPM-NN) to physically characterize

* Corresponding author: yangmx@iwhr.com

the particle diffusion in the air and an adaptive monitoring algorithm, constructing an accurate AQI mapping can provide more High AQI prediction accuracy. The distribution of pollutant concentration is affected by factors such as historical pollution data and weather conditions. Ghaemi Z and Alimohammadi A[8] use the perspective that the distribution of pollutant concentration is affected by factors such as historical pollution data and weather conditions, and design an online algorithm based on LaSVM. Yes, the speed has increased significantly. In 2019, Xinghan Xu and Weijie Ren [9] proposed the SLI-ESN model in order to accurately predict the PM2.5 time series in monitoring and predicting the air quality index (AQI), which improved the impact of historical evolution on the current state. The mRMR feature selection method is introduced to reduce redundant information and improve efficiency. In 2021, Wang Z, Li J [10] believe that accurate forecast and early warning of air pollution episodes is crucial for the air pollution control. Therefore, this article selects meteorological factors and historical AQI as the influencing factors of AQI to predict AQI. AQI multi-step forecasting is the same as weather forecasting, multi-step forecasting has also become a key research object. In 2016, Yang, Xiaoping et al. [11] integrated natural factors, humanistic factors, and self-evolution factors into a vector autoregressive model to predict Beijing air quality in multiple steps. The experimental results show that the model is robust and capable. It can predict haze conditions more accurately and is sensitive to sudden changes in air quality.

Based on the current research status, although there are methods to predict AQI based on meteorological factors, there are few studies used for multi-step prediction, and there are not many analytical studies based on meteorological factors. This article will analyze in detail the impact of 9 meteorological factors on AQI. And based on the analysis results, multi-step prediction of AQI is carried out.

## 2 Materials and Methods

### 2.1 The predictive AQI method proposed in this paper

There are many meteorological factors that affect AQI, including wind speed, weather, temperature, humidity, etc. This paper first analyzes the correlation between various meteorological factors and AQI. Considering the mutual influence of various meteorological factors, the data may contain overlapping feature information. Using PCA to remove overlapping feature information can effectively reduce the computational complexity of the model and improve the performance of the model. In multi-step forecasting, in order to avoid accumulated errors, direct multi-step forecasting methods are used. Since LSSVM can overcome the shortcomings of long training time, randomness of training results and over-learning, and has been applied in many fields such as pattern recognition, the PCA-based LSSVM method is selected to predict AQI. The forecast scheme is shown in Figure 1.

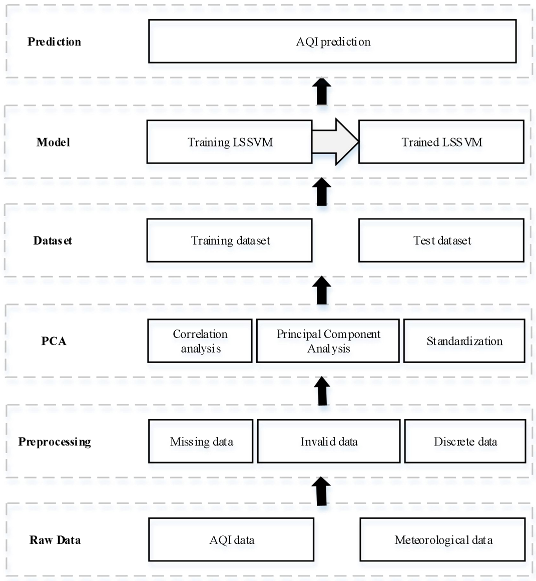


**Fig. 1.** General prediction process.

#### *2.1.1 Preprocess the sample data*

Some of the acquired sample data is missing, and there are some illogical data, which need to be preprocessed. The preprocessed data is $x^{'} = \{x_1^{'}, x_2^{'}, x_3^{'}, ..., x_m^{'}\}$.

#### *2.1.2 Normalize $x^{'}$*

Since the sample data is not the same dimension data, it needs to be standardized and converted to dimensionless data to perform correlation analysis and principal component analysis. The standardization method used in this paper is 0-1 standardization, and the standardization equation is:

$$x = \frac{x-\mu}{\sigma} \quad (1)$$

μ is the mean and σ is the standard deviation. Standardize x^' to convert the sample data into $x^{''}=\{x_1^{''}, x_2^{''}, x_3^{''}, ..., x_m^{''}\}$.

#### *2.1.3 Correlation analysis of AQI and meteorological factor data*

Perform correlation analysis on meteorological factor data and AQI. If there is no correlation or weak correlation, the feature will be directly removed. Combine the data with strong correlation with AQI into a new matrix $Y=\{y_1, y_2, ...y_i,\}$ where l≤m-1 and the remaining data are combined into $Y^{'} = \{Y_1^{'}, ..., Y_{m-i}^{'}\}$.

#### *2.1.4 Principal component analysis*

Principal component analysis is performed on the sample data set $Y = \{Y_1, Y_2, ...Y_i\}$, and the correlation matrix $Y^{'}$ of Y is obtained by eigenvalue $\lambda = [\lambda_1, \lambda_2, ..., \lambda_i]$ and the contribution rate of the eigenvalue. $c = [c_1, c_2, ..., c_i]$ . Extract the components with a cumulative contribution rate of 85% as the main component; according to the eigenvalue λ and the component α eigenvalue, the corresponding eigenvector$d = [d_1, d_2, ..., d_k]$ is obtained. The dimensionality reduction data y of the sample data set Y after principal component analysis can be obtained from the sample data set Y and $D^T = [d_1, d_2, ..., d_k]$.

$$y = [y_1, ..., y_k] = D^T \quad (2)$$

#### *2.1.5 Train the LSSVM model*

Combine the matrix y obtained by principal component analysis and the historical AQI of the previous α day into the input data of the training set $x^{train}$ ,the target AQI data is used as the output of the LSSVM training set to train the LSSVM model.

#### *2.1.6 AQI Forecast*

Use the trained LSSVM model to predict the target AQI of the test set.

## 3 Results & Discussion

### 3.1 Data interpretation

This paper selects three cities of Nanchang, Beijing and Nanjing as the research objects to predict its AQI. The data obtained is from 2018.12.21-2020.06.08, a total of 2703 items, including meteorological factor data and AQI data; the meteorological factor data includes average humidity, rainfall rate, maximum temperature, rainfall, sunny rate, average wind speed, minimum temperature There are nine types of data including, wind direction and average air pressure. The 2703 pieces of data are divided into training data for training models and test data for testing models. 90% are training data and 10% are test data. Some of the obtained data is missing, and some of the data is unreasonable. Preprocessing is needed to prevent it from having a greater impact on the prediction results.

The missing values of the eight types of meteorological factors: average humidity, rainfall rate, maximum temperature, sunny rate, average wind speed, minimum temperature, wind direction, and average pressure, and missing values of AQI are directly taken as the average value of the previous day and the next day. The replacement value of the missing value; for the processing of the missing value of precipitation, all the data consistent with the rainfall rate and sunny rate of the day are selected, and the average rainfall of these data is used as the replacement value of the missing value.

Regarding the wind direction factor, the processing method adopted in this paper is to directly digitize the wind direction. The unsustainable wind direction is digitized as 0, the southeast wind is digitized as 1, the northeast wind is digitized as 2, the southwest wind is digitized as 3, and the northwest wind is digitized as 4.

### 3.2 Correlation analysis

Analyze the correlation between AQI and meteorological factors and the correlation between meteorological factors in the selected three cities, using standard T test for analysis, and the analysis results are shown in Table 1.

**Table 1.** Correlation coefficient between AQI and meteorological factors.

| Meteorological factors | Nanchang | Beijing | Nanjing |
|---|---|---|---|
| Humidity | -0.311 | -0.244 | -0.238 |
| Rain percentage | -0.271 | -0.280 | -0.311 |
| Maximum temperature | -0.236 | -0.283 | -0.306 |
| Rainfall | -0.158 | -0.118 | -0.180 |
| Sunny percentage | 0.252 | 0.229 | 0.261 |
| Wind speed | -0.309 | -0.205 | -0.225 |
| Minimum temperature | -0.361 | -0.376 | -0.415 |
| Atmospheric pressure | 0.369 | 0.164 | 0.115 |
| Wind direction | -0.071** | -0.174 | -0.099 |

(Note: No asterisk indicates significant correlation p⩽0.01 ** indicates significant correlation p⩽0.05)

It can be seen from Table 1 that the selected nine types of meteorological factors are all significantly related to AQI; the correlation between the nine types of meteorological factors and AQI in the three cities is consistent, and the air pressure and weather rate are positively correlated with AQI, that is, as the air pressure rises The high AQI also increases, and the other eight types of meteorological factors are negatively correlated with the AQI, that is, the AQI will decrease with the increase of the value of these eight types of meteorological factors. Therefore, it is feasible to use meteorological factors as the influencing factors of AQI to predict AQI.

### 3.3 Principal component analysis

Principal component analysis is divided into two steps. First, KMO and Chi-square test are used to verify the feasibility of principal component analysis in the city, and then principal component analysis is performed.

**Table 2.** KMO and chi-square test

| Factor analysis | Nanchang | Beijing | Nanjing |
|---|---|---|---|
| KMO test | 0.715 | 0.602 | 0.569 |
| Bartlett test | 6071.24 | 4322.83 | 4341.36 |

It can be seen from Table 2 that the KMO statistic values of the three cities are 0.715, 0.602, 0.569, which are significantly greater than 0.5; the approximate chi-square values are 6071, 4322, and 4341, which are far greater than the critical value of 18.3; and the significance is less than 0.05 These all further verify that the principal component analysis of meteorological data can be carried out.

The principal component analysis method is used to conduct principal component analysis on the experimental data of the three cities of Nanchang, Beijing, and Nanjing. The results of the analysis are shown in Table 3.

**Table 3.** Principal component analysis results of three cities

| Component | Nanchang | | Beijing | | Nanjing | |
|---|---|---|---|---|---|---|
| | Variance % | Cum. Variance % | Variance % | Cum. Variance % | Variance % | Cum. Variance % |
| 1 | 32.34 | 32.34 | 30.35 | 30.35 | 30.73 | 30.73 |
| 2 | 29.14 | 61.48 | 23.95 | 54.304 | 22.41 | 53.14 |
| 3 | 11.67 | 73.15 | 11.98 | 66.29 | 12.82 | 65.96 |
| 4 | 10.07 | 83.22 | 10.41 | 76.70 | 10.60 | 76.56 |
| 5 | 9.39 | 92.61 | 9.57 | 86.262 | 8.92 | 85.48 |
| 6 | 4.16 | 96.77 | 7.03 | 93.290 | 8.44 | 93.91 |
| 7 | 1.94 | 98.70 | 4.27 | 97.56 | 3.94 | 97.85 |
| 8 | 0.91 | 99.61 | 2.05 | 99.61 | 1.72 | 99.57 |
| 9 | 0.39 | 100.00 | 0.39 | 100.00 | 0.43 | 100.00 |

Table 3 shows the principal component analysis results of the three cities. The leftmost column is the 9 components arranged in ascending order of variance. The data of each city consists of 2 columns. The first column "Variance%" is the variance. It also shows the contribution rate of the component, and the second column is the cumulative contribution rate. The first column of each city is the corresponding variance. From Figure 3-2, we can see that the second column is nine eigenvalues arranged by size.

According to the principle of the principal component analysis method, the main components can be selected according to the cumulative contribution rate, and several principal components with a cumulative contribution rate greater than 85% are taken to replace all the components to reduce the dimension of calculation. It can be seen from Table 3 that the cumulative contribution rates of the first five components of the three cities are 92.61%, 86.26%, and 85.48%, which are all greater than 85%. Therefore, the first five components can be selected as the principal components for AQI prediction.

The PCA method performs dimensionality reduction and denoising processing on the meteorological factor data of the three cities of Nanchang, Beijing, and Nanjing. The meteorological factor data of the three cities are all reduced from 9-dimensional data to 5-dimensional data. While removing redundant information, it can greatly Simplify the training and prediction process of LSSVM and improve the training speed.

## 3.4 LSSVM model based on PCA

This paper selects the three-day historical AQI and the meteorological factors of the previous day in the three cities of Nanchang, Beijing, and Nanjing as input data, and the AQI of the day as the output data to verify the feasibility of the PCA method. The prediction results are shown in Table 4.

**Table 4.** Comparison of LSSVM prediction results with or without PCA in three cities

| | Nanchang | | Beijing | | Nanjing | |
|---|---|---|---|---|---|---|
| | MAPE | RMSE | MAPE | RMSE | MAPE | RMSE |
| Without PCA | 0.400 | 24.362 | 0.391 | **24.542** | 0.391 | 25.663 |
| With PCA | **0.395** | **21.182** | **0.342** | 26.693 | **0.271** | **23.070** |

It can be seen from Table 4 that the prediction results of the LSSVM method with principal component analysis in the AQI prediction results of Nanchang and Nanjing are slightly smaller than the prediction results of the LSSVM method without principal component analysis in both MAPE and RMSE; there is no principal component in the prediction results of Beijing AQI The RMSE is small in the predicted results of the analysis, but the MAPE is small in the predicted results of the principal component analysis.

Because the RBF kernel function is suitable for large samples as well as small sample data, suitable for high-dimensional space and low-dimensional space, as well as the advantages of low parameters and low computational complexity, the RBF kernel function is selected as the LSSVM model. The kernel function of, then the regression equation of LSSVM is:

$$f(x) = \sum_{i=1}^{n} \alpha_i exp\left(-\frac{||x-x_i||^2}{2\sigma^2}\right) + b \qquad (3)$$

σ is the parameter in the RBF kernel function.

### *3.4.1 Parameter setting*

Because in all the method models in this paper, there are a total of days of historical AQI, and the penalty parameter C in the LSSVM model needs to be determined in the kernel function parameter σ.

(1) Determination of historical AQI days

The most recent historical AQI has the greatest impact on the current AQI, and the longer the time, the smaller the impact until the impact on the current AQI is negligible. Therefore, this paper selects the days with the best effect in 1-4 days according to the experimental results. The experimental analysis is carried out on three cities of Nanchang, Nanjing, and Beijing. The historical weather factors of the previous day are selected as the influencing factors, and the number of days of the historical AQI is from 1-4 to predict the current AQI.

**Table 5.** The impact of historical AQI days in the three cities a

Nanchang Beijing Nanjing

| | Nanchang | | Beijing | | Nanjing | |
|---|---|---|---|---|---|---|
| | MAPE | RMSE | MAPE | RMSE | MAPE | RMSE |
| 1 | 0.390 | 21.133 | 0.260 | 22.059 | 0.265 | 22.902 |
| 2 | 0.383 | 20.676 | 0.266 | 22.176 | 0.268 | 22.831 |
| 3 | 0.395 | 21.182 | 0.301 | 27.372 | 0.271 | 23.070 |
| 4 | 0.407 | 21.510 | 0.302 | 25.427 | 0.275 | 23.313 |

It can be seen from the table that among the prediction results of the three cities with different historical AQI days as the influencing factor, the MAPE and RMSE of the Nanchang prediction result are the smallest; the MAPE and RMSE of the Beijing prediction result are both the smallest when a=2, but a= The MAPE and RMSE at 2 o'clock are very close to the values at a=2; the MAPE of Nanjing prediction results is the smallest at a=1, the RMSE at a=2 is the smallest, and the MAPE at a=2 is the value at a=1 very close. In summary: when a=2, the overall prediction effect is the best, so a=2.
(2) Determination of the parameters in the penalty parameter C and the kernel function σ

In this paper, the grid search method is used to optimize C and σ, that is, to try all possible C and σ, and then use cross-validation to verify, and find the C and σ with the highest accuracy. Cross-validation is an exhaustive search method. Although it is time-consuming compared to other methods, the complexity is not too high when there are fewer parameters, and the grid search method can find the global optimal solution, so this paper selects cross Verification to verify the parameters.

### *3.4.2 Analysis of multi-step forecast results*

The prediction results of the three cities of Nanchang, Beijing and Nanjing based on the PCA+LSSVM method are shown in Table 6, Table 7, and Table 8, respectively.

**Table 6.** Error analysis of Nanchang AQI multi-step prediction

| Days | MAPE | MAPE error increased by % | RMSE | RMSE error increased by % |
|---|---|---|---|---|
| 1 | 0.384 | | 21.182 | |
| 2 | 0.490 | 27.6% | 24.641 | 16.3% |
| 3 | 0.589 | 20.2% | 27.637 | 12.2% |
| 4 | 0.604 | 2.5% | 28.824 | 4.3% |

It can be seen from Table 6 that as the number of forecast days increases, the accuracy rate gradually decreases. It is predicted that the MAPE and RMSE of the AQI on the second and third days will increase more than the previous day, and the MAPE will increase by 27.6% and 20% respectively; the RMSE will increase by 16.3% and 12.2% compared with the previous day. Although MAPE and RMSE have increased a lot, MAPE and RMSE themselves are relatively small.

**Table 7.** Error analysis of multi-step prediction of Beijing AQI

| Days | MAPE | MAPE error increased by % | RMSE | RMSE error increased by % |
|---|---|---|---|---|
| 1 | 0.267 | | 22.536 | |
| 2 | 0.288 | 7.9% | 24.894 | 10.5% |
| 3 | 0.297 | 3.1% | 25.279 | 1.5% |
| 4 | 0.317 | 6.7% | 26.279 | 4.0% |

It can be seen from Table 7 that as the number of forecast days increases, the accuracy rate gradually decreases. It is predicted that the MAPE and RMSE of the second day will increase more than the previous day. The MAPE will increase by 7.9% and the RMSE will increase by 10.5%.

**Table 8.** Error analysis of Nanjing AQI multi-step prediction

| Days | MAPE | MAPE error increased by % | RMSE | RMSE error increased by % |
|---|---|---|---|---|
| 1 | 0.285 | | 24.094 | |
| 2 | 0.300 | 5.3% | 24.000 | -0.4% |
| 3 | 0.328 | 9.3% | 25.731 | 7.2% |
| 4 | 0.338 | 3.0% | 26.341 | 2.4% |

It can be seen from the table that as the number of forecast days increases, the accuracy rate gradually decreases. It is predicted that the MAPE and RMSE of the AQI on the second day will increase more than the previous day, and the MAPE will increase by 79.3% compared with the previous day; the RMSE will increase by 7.2% compared with the previous day. Since the MAPE and RMSE of the first day's prediction results are small, the accuracy rate decreases with the increase of the number of days is small, so the overall prediction effect is better.

### *3.4.3 Comparison results with ARIMA method*

ARIMA is a classic time series multi-step forecasting method. On the experimental data of three cities, the results of the algorithm in this paper and the ARIMA algorithm are compared. The results of MAPE are shown in Table 9, and the results of RMSE are shown in Table 10.

**Table 9.** MAPE between ARIMA and Ours

| days | Nanchang | | | Beijing | | | Nanjing | | |
|---|---|---|---|---|---|---|---|---|---|
| | ARIMA | Ours | Improve by % | ARIMA | Ours | Improve by % | ARIMA | Ours | Improve by % |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.474 | **0.384** | -20.0 | 0.305 | **0.267** | -12.5 | 0.311 | **0.285** | -8.4 |
| 2 | 0.638 | **0.490** | -23.2 | 0.346 | **0.288** | -16.8 | 0.346 | **0.300** | -13.3 |
| 3 | 0.679 | **0.589** | -13.3 | 0.354 | **0.297** | -16.1 | 0.366 | **0.328** | -10.4 |
| 4 | 0.691 | **0.604** | -12.6 | 0.362 | **0.317** | -12.4 | 0.357 | **0.338** | -5.3 |

**Table 10.** RMSE between ARIMA and Ours

| days | ARIMA | Ours | Improve by % | ARIMA | Ours | Improve by % | ARIMA | Ours | Improve by % |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 23.34 | **21.18** | -9.2 | 24.63 | **24.09** | -2.2 | 25.83 | **25.11** | -2.8 |
| 2 | 30.38 | **24.64** | -18.9 | 27.03 | **24.00** | -11.2 | 28.05 | **23.82** | -15.1 |
| 3 | 31.42 | **27.64** | -12.0 | 27.46 | **25.73** | -6.3 | 28.25 | **25.95** | -8.1 |
| 4 | 31.89 | **28.82** | -9.6 | 27.08 | **26.34** | -2.7 | 27.00 | **24.66** | -8.7 |

It can be seen from Table 9 that the method proposed in this paper predicts that the MAPE of AQI for 1-4 days in Nanchang, Beijing, and Nanjing is less than the MAPE predicted by the ARIMA method, and the error reduction ranges from 5.3% to 23.2%. Most of the MAPE errors are reduced. Not less than 12%. Overall, the forecast results have been greatly improved.

It can be seen from Table 10 that the method proposed in this paper predicts that the RMSE of the 1-4 days AQI of the three cities is less than that of the ARIMA method. The error reduction range ranges from 2.2% to 18.9%, and the average error is about 10%. The overall prediction result is relatively good. Big improvement. However, the improvement of RMSE relative to MAPE is not stable enough. The reason is that MAPE is a relative error, and RMSE is an absolute error. The value of AQI is inherently high and low, and the accuracy of prediction is more reasonable to be measured by MAPE. In general, the PCA+LSSVM method proposed in this paper is more accurate than the ARIMA method in predicting results.

## 4 Conclusions

This paper focuses on analyzing the impact of meteorological factors on AQI, combining meteorological factors and historical AQI as the influencing factors of AQI, and establishing a model to predict AQI. The main conclusions are as follows:

(1) Consider the influence of multiple meteorological factors on AQI. Since multiple meteorological factors influence each other, there is likely to be overlapping feature information. Analyzing the correlation between various meteorological factors and AQI, the results show that all the nine meteorological factors selected in this paper have a significant relationship with AQI, and most of the meteorological factors also have a significant correlation relationship. The influence of AQI is feasible, and the principal component analysis of meteorological factors is also feasible.

(2) Using PCA to reduce the dimensionality and denoising of meteorological factor data will not have a negative impact on the results of LSSVM model predicting AQI. On the contrary, due to its denoising characteristics, it can improve the accuracy of LSSVM prediction to a certain extent.

(3) AQI prediction method based on PCA and LSSVM. The experimental results show that the prediction effect is best when the historical AQI as an impact factor is 2 days.

(4) AQI multi-step forecasting method based on PCA and LSSVM, as the number of forecast days increases, the forecast error will gradually increase, and the error increase will also decrease. Compared with the ARIMA method, the prediction results of the multi-step prediction method based on PCA and LSSVM are smaller in MAPE and RMSE. The method proposed in this paper has better prediction effect and more applicability than the ARIMA method.

This work is supported by the National Natural Science Foundation of China (U1865102) and Beijing Science and Technology Plan Project (Z201100001820022).

## References

1. S. Chen, G. Kan, J. Li, et al, Investigating china's urban air quality using big data, information theory, and machine learning. Polish Journal of Environmental Studies, **27(2)**, 565-578 (2018)
2. X. Gao, X. Wang, H. Sh, Multifractal cascade analysis on the nature of air pollutants concentration time series over china. Aerosol and Air Quality Research, **19(9)**. (2019)
3. J. Xue, Y. Xu, L. Zhao, et al, Air pollution option pricing model based on AQI. Atmospheric Pollution Research, **10(3)** 665-674. (2019)
4. Q. Wu, H. Lin, A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. Science of The Total Environment, **683** 808-821. (2019)
5. C. Luo, H. Yang, L.P. Huang, et al. A Fast PM2.5 Forecast Approach Based on Time-Series Data Analysis, Regression and Regularization. 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI). (2018)
6. Y. Lin, L. Zhao, H. Li and Y. Sun, Air quality forecasting based on cloud model granulation. Eurasip Journal on Wireless Communications & Networking, **2018(1)** 106. (2018)
7. Y. Yang, Z. Zheng, K. Bian, L. Song, H. Zhu, Realtime profiling of fine-grained air quality index distribution using uav sensing. IEEE Internet of Things Journal, **(99)** 1-1 (2017)
8. Z. Ghaemi, A Alimohammadi, M Farnaghi. Lasvm-based big data learning system for dynamic prediction of air pollution in tehran. Environmental Monitoring and Assessment: An International Journal. (2018).
9. X. Xu, W. Ren, Prediction of Air Pollution Concentration Based on mRMR and Echo State Network. Applied Sciences, **9(9)** 1811 (2019)

10. Z. Wang, J. Li, Establishment and operational application of china national air quality forecast and early warning device. (2021)
11. X. Yang, Z. Zhang, Z. Zhang, L. Sun, C. Xu, L. Yu, A long-term prediction model of beijing haze episodes using time series analysis. Computational Intelligence & Neuroscience, 2016, 17. (2016)