PLOS ONE

RESEARCH ARTICLE

Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang

Bing-Chun Liu[1], Arihant Binaykia[2], Pei-Chann Chang[3,4]*, Manoj Kumar Tiwari[2], Cheng-Chin Tsao[3]

1 Research Institute of Circular Economy, Tianjin University of Technology, Tianjin, P.R. China, **2** Department of Industrial and Systems Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India, **3** Department of Information Management, Yuan Ze University, Taoyuan, Taiwan, ROC, **4** Office of Academic Affairs, Zhuhai College of Beijing Institute of Technology, Zhuhai, China

* iepchang@saturn.yzu.edu.tw

Citation: Liu B-C, Binaykia A, Chang P-C, Tiwari MK, Tsao C-C (2017) Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. PLoS ONE 12(7): e0179763. https://doi.org/10.1371/journal.pone.0179763

Editor: Chon-Lin Lee, National Sun Yat-sen University, TAIWAN

Received: December 3, 2016

Accepted: June 2, 2017

Published: July 14, 2017

Data Availability Statement: Air Quality Datasets used in this study are from third parties and available from the URLs and locations below. Daily Air Pollutant Concentration Data download links for Beijing, Tianjin and Shijiazhuang: https://www.aqistudy.cn/historydata/monthdata.php?city=%E5%8C%97%E4%BA%AC https://www.aqistudy.cn/historydata/monthdata.php?city=%E5%A4%A9%E6%B4%A5 https://www.aqistudy.cn/historydata/monthdata.php?city=%E7%9F%B3%E5%AE%B6%E5%BA%84 Meteorological data download links for Beijing, Tianjin and

Abstract

Today, China is facing a very serious issue of Air Pollution due to its dreadful impact on the human health as well as the environment. The urban cities in China are the most affected due to their rapid industrial and economic growth. Therefore, it is of extreme importance to come up with new, better and more reliable forecasting models to accurately predict the air quality. This paper selected Beijing, Tianjin and Shijiazhuang as three cities from the Jingjinji Region for the study to come up with a new model of collaborative forecasting using Support Vector Regression (SVR) for Urban Air Quality Index (AQI) prediction in China. The present study is aimed to improve the forecasting results by minimizing the prediction error of present machine learning algorithms by taking into account multiple city multi-dimensional air quality information and weather conditions as input. The results show that there is a decrease in MAPE in case of multiple city multi-dimensional regression when there is a strong interaction and correlation of the air quality characteristic attributes with AQI. Also, the geographical location is found to play a significant role in Beijing, Tianjin and Shijiazhuang AQI prediction.

Introduction

Air quality has a huge impact on the quality of living, the well-being of the population as well as the image of the city. With the increase in population in urban areas, there has been an increase in the development of the industries as well as the consumption of fossil fuels. This has led to the increase in air pollution in China [1]. The dreadful effects of air quality in the capital city of China—Beijing due to increase in population which has led to increase in number of vehicles as well as increase in fuel consumption has been the point of discussion for many researchers [2]. It can be said that the progress of the human society is at the expense of our lives as well as the environment [3]. The goal of environment sustainability is difficult to

Shijiazhuang: http://lishi.tianqi.com/beijing/index.html http://lishi.tianqi.com/tianjin/index.html http://lishi.tianqi.com/shijiazhuang/index.html Latitude and Longitude Values: Air Quality Monitoring Stations in Beijing, Tianjin and Shijiazhuang Beijing: According to the address of the monitoring stations, the Latitude and Longitude Values of these stations can be found using MyPositions software. These data are included in the Supporting Information files.

Funding: The authors received funding support from the National Natural Science Foundation of China (71503180).

Competing interests: The authors have declared that no competing interests exist.

achieve due to excessive air pollution in urban cities of China including Beijing, Tianjin and Shijiazhuang [4]. Airborne particulate matter (PM) is especially detrimental to health and has been estimated to cause between 3 and 7 million deaths every year, primarily causing cardiorespiratory disease [5]. NO and NO_2 are the two air pollutants which also cause respiratory diseases proven by using multiple linear regression to analysis the correlation coefficient between the outpatient visits and air pollution [6].

Beijing-Tianjin-Hebei economic zone is the capital of Northern China's largest urban agglomeration that has witnessed rapid economic and population growth. Shijiazhuang is the capital and largest city of North China's Hebei Province. Due to the rapid economic and population growth it has encountered a series of environment protection and sustainable development related issues. In particular air pollution has a direct impact on the health of the residents as well as the quality of living and image of the city. The region faces industrial emissions, large scale urban construction and other characteristics of a fast growing region which are dangerous environmental threats of air pollution to the public.

Air Quality Index (AQI) is a widely used index for public understanding and for evaluation of air pollution on human health indicators [7]. In 1972, the US Environmental Protection Agency (EPA) first proposed the Pollution Standard Index (PSI). In 1999, EPA proposed the Air Quality Index (AQI) which is now used worldwide. In China, the AQI is measured through real time monitoring of Air Quality Data obtained through the conversion process which is very important for future AQI forecast. AQI forecast research is currently focused on the use of statistical and machine learning models in order to predict the future AQI values. McKeen [8] and Chuang [9] worked on real time air quality forecasting and developed online meteorological models to predict air quality. XU Xiaofeng [10] applied a method to determine that the type of pollution in Beijing is an ongoing process of research and found that low wind speed and air layer structural stability is the main cause of air pollution. Anikender Kumar [11] used the principal component regression technique in order to forecast the Air Quality Index (AQI) values in Delhi, India. Yongtao Hu [12] used synoptic classification for evaluating an operational air quality forecasting system in Atlanta. Computational models use a lot of the historical pollution data to predict the relationship between the input features and the output features by simple regression or complex machine learning methods. The absence of knowledge sources and the physical process do not significantly change the conditions of application of deterministic model [13].

There are a variety of machine learning algorithms in air pollution forecasting applications. Chen Chun Qi [14] applied multiple linear regression model in Wuhan to study the Wuhan Meteorological Environment impact and correlation on air quality. Pérez [15] obtained a model for forecasting PM_{10} values using the neural networks and compared it to the obtained linear model. Li [16] compared various artificial intelligence and machine learning models for air pollution forecasting. Bai Heming [17] used Neural Networks algorithm for forecasting the Air Quality Index. Shad [18] and Alhanafy [19] used the FC (Fuzzy logic) algorithm for predicting air pollution. Zolghadri [20] and Hoi [21] used the KF (Kalman filter) algorithm for predicting air quality parameters. Liu [22] used a Back-Propagation Neural Network and a Selection Sample Rule for forecasting Urban Air Quality. Li Xiang [23] also used GAB and fuzzy BP neural network for Air quality forecasting. Sun [24] used HMM (Hidden Markov Model) algorithm for prediction of $PM_{2.5}$ concentrations in Northern California. Support Vector Regression [25] is a proven and wildly used machine learning algorithm for robust and reliable prediction results. It is also known for handling multi-dimensional data sets [26]. Also, SVR specializes for small number of samples for training [27].

This study is different from the past studies on air quality from start to the changes in input characteristic variables while observing the interaction of different conditions of urban air

pollution. The present study examines the correlation between the same areas of urban air pollution and takes into account the air quality information of several cities together as input variables in order to predict the AQI using the Support Vector Regression (SVR) method with an aim to further improve the AQI forecast accuracy.

The harmful health effects of air pollution are because of multiple pollutants present in the atmosphere that cannot be justified with a single pollutant index. Hence a multiple pollutant AQI model is more effective than the presently utilized single pollutant model in modelling air quality across the pollutant concentrations. We have chosen the SVR algorithm for our study. The reason that SVMs often outperform Artificial Neural Networks (ANNs) in practice is that they deal with the biggest problem with ANNs: overfitting. SVR is less prone to overfitting than the ANNs due to the presence of regularization parameters. In SVR, the basic idea is to map the multi-variate data into higher-dimensional feature space via a nonlinear mapping with the help of a kernel trick and then perform regression in this space that avoids difficulties of using linear functions in the high dimensional feature space and the final optimization problem is transformed into dual convex quadratic programmes.

Data and method

Air quality data

Jingjinji represents the Beijing-Tianjin-Hebei region where Jing means Beijing, Jin means Tianjin and Ji means the Hebei region. According to geographical location characteristics of this capital economic area as well as the increase in urbanisation of Beijing-Tianjin-Hebei region (Fig 1), the paper selected Beijing, Tianjin and Shijiazhuang as the research objects. Fig 2 shows the exact location of the air quality monitoring stations that are used in our study. The pollutant concentrations and the AQI values for each city is the mean of the values obtained from the various monitoring stations around each city.

The data used consists of two parts: the daily air pollutant concentration data in three cities, obtained from the China Environmental Monitoring Center; the other part is the three cities

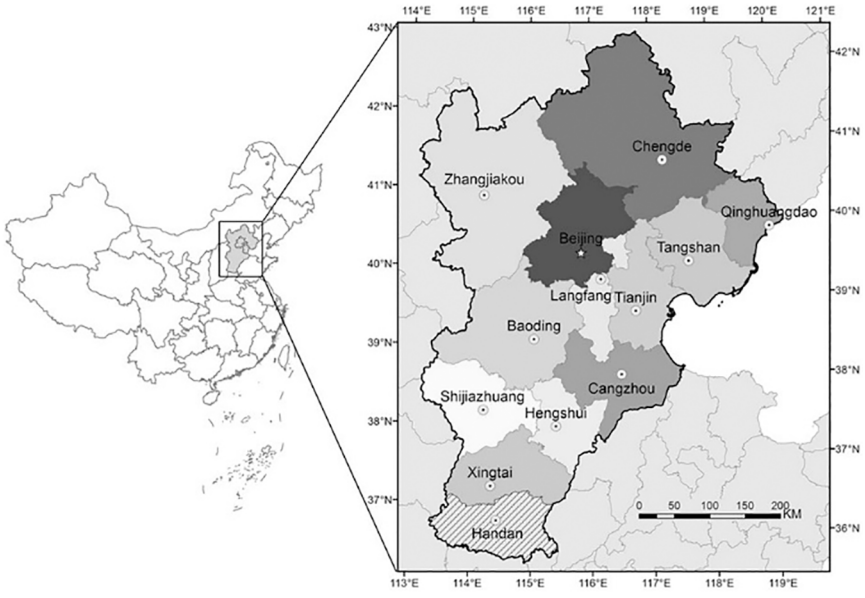

Fig 1. Jingjinji region. Map developed in ArcGIS (www.arcgis.com).

https://doi.org/10.1371/journal.pone.0179763.g001

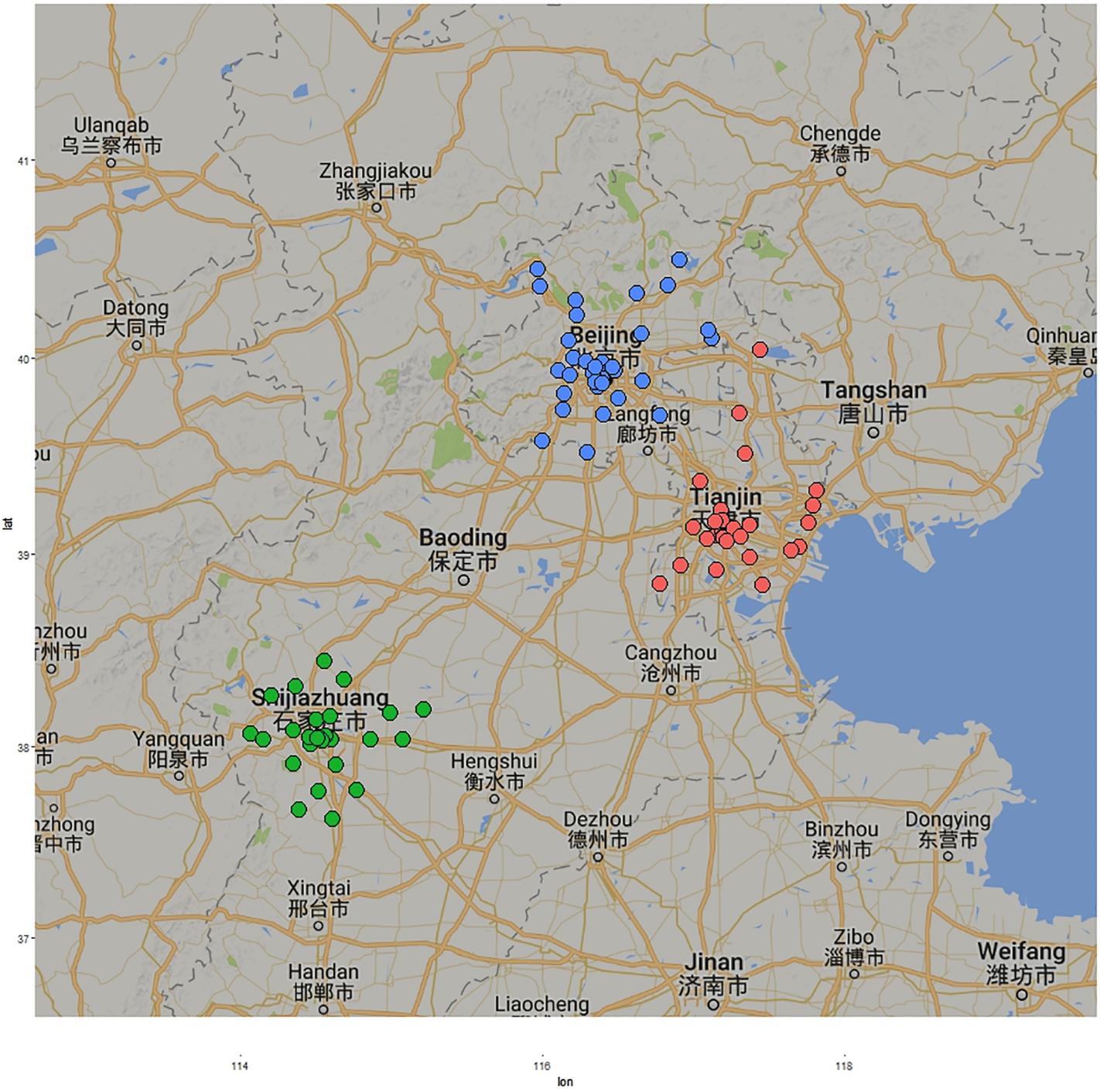

Fig 2. Map of air quality monitoring stations in Beijing (in blue), Tianjin (in red) and Shijiazhuang (in green). X-axis represents longitude and Y-axis represents latitude (Latitude and Longitude data is available in S1, S2 and S3 Tables). Generated using the ggplot2 (http://CRAN.R-project.org/package=ggplot2) and ggmap (https://cran.r-project.org/package=ggmap) packages of R.

https://doi.org/10.1371/journal.pone.0179763.g002

daily weather condition and the meteorological data obtained from the China Meteorological Administration. The daily air pollutant concentration is calculated by selecting the maximum value from the eight-hour average concentration values. Hourly air pollutant concentrations

are calculated using the eight-hour midpoint average concentrations i.e. for calculating the concentration for a particular hour, three consecutive next hours, four consecutive previous hours and that particular hour's raw data is considered. Here we have considered a daily approach for Air Quality forecast primarily due to simplicity, practicality and operational reasons although the approach would have been similar if the hourly air pollutant data was considered. A total of 12 characteristic variables are considered for each city including the six air pollutants concentration namely "$PM_{2.5}$", "PM_{10}", "SO_2", "CO", "NO_2" and "O_3"; five variable weather conditions namely "minimum temperature", " maximum temperature "," weather "," wind direction " and " wind power "; and the last day's observed AQI values. The unit of measurement of air pollution features $PM_{2.5}$, PM_{10}, SO_2, NO_2 and O_3 is $\mu g/m^3$ and for CO is mg/m^3. These are the most harmful pollutants present in large quantities in the atmosphere and therefore have a large amount of feature importance for AQI prediction. Hence the various automated air quality monitoring stations around each city focuses on only these six main pollutants mentioned above.

It is also important here to discuss about the various weather conditions included in our study for Air Quality forecast and modelling. Wind Direction sometimes has a substantial effect on the air quality of a particular city as well as a region. Air quality can either become better or worse based on wind direction. If the wind is coming from an area with extremely less pollution, then the air quality improves a significant amount. But if the wind is coming from a region that is highly polluted it is likely to become worse. Low wind speed for a highly polluted region with multiple sources of pollution is a problem because the pollution stays in the same region rather than blowing away in the direction of the wind [28]. Strong wind speeds generally promote the transport and travel of pollutants rapidly to distant places. High Temperature during the summer days contribute to photochemical reactions especially in the case of particulate matter and ozone. Whereas rain can clean the air but can cause problems of acid rain and soil pollution. In our study of the Jingjinji Region we have taken into account the features corresponding to these weather conditions. The feature weather is classified as partly cloudy, sunny, rainy, cloudy, snow, dust, haze and fog. These are represented by values 0 to 7 respectively shown in Table 1. Wind direction has been classified as north, northeast, east, south-east wind, southerly, north-west, west wind, south-west wind and no sustained wind nine types. These are represented by values 0 to 8 respectively as shown in Table 2. Wind Power has been summarized into 5 levels of wind power density based on the national standard of wind power in China: <3, 3–4, 4–5, 5–6 and 6–7 grade as shown in Table 3.

Support Vector Regression (SVR)

Support Vector Machines (SVM) is a machine learning algorithm that constructs hyperplanes for separating different classes and is generally used for analyzing data that has a categorical output variable. Whereas in case of a continuous numeric output variable we use regression analysis in place of classification called Support vector regression (SVR). An SVR model is used to obtain an approximate function $g(x)$ from a given complex sample data $G = \{(x_i, y_i)\}\binom{N}{i=1}$. The main idea is to first map the non-linearly separable data into a higher dimensional linearly separable feature space and then using this feature space for computation using linear programming [29].

$$f(x) = \sum_{i=1}^{D} w_i \phi_i(x) + b \tag{1}$$

In the Eq (1), $\phi_i(x)$ is characterized by variables b and w_i, and it can be estimated from the data. When the data is non-linearly separable, we need to map the data into richer feature

Table 1. Factors of input feature: Weather.

Weather	Factors
Partly Cloudy	0
Sunny	1
Rainy	2
Cloudy	3
Snow	4
Dust	5
Haze	6
Fog	7

https://doi.org/10.1371/journal.pone.0179763.t001

space where the data is separable. The minimum function coefficients w_i can be obtained by:

$$R[w] = \frac{1}{N}\sum_{i=1}^{N}|f(x_i) - y_i|_\epsilon + \lambda||w||^2 \tag{2}$$

In the Eq (2), λ is a standardization constant and function $|f(x_i)-y_i|_\epsilon$ can be defined as:

$$|f(x_i) - y_i|_\epsilon = \begin{cases} |f(x) - y| - \epsilon, & |f(x_i) - y_i| \geq \epsilon \\ 0, & other \end{cases} \tag{3}$$

The minimizing function can also be expressed in the following form [30]:

$$f(x, \alpha, \alpha^*) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{4}$$

Simultaneously, $\alpha_i\alpha_i^* = 0$, $\alpha_i, \alpha_i^* \geq 0$, $i = 1,..,N$, the inner product kernel function can be expressed as

$$k(x, y) = \sum_{j=1}^{D}\phi_j(x)\phi_j(y) \tag{5}$$

The coefficients α_i *and* α_i^* can be obtained by using the following equation:

$$R(\alpha_i^*, \alpha_i) = -\varepsilon\sum_{i=1}^{N}(\alpha_i^*+\alpha_i) + \sum_{i=1}^{N}y_i(\alpha_i^*-\alpha_i) - \frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i^*+\alpha_i)(\alpha_i^*-\alpha_i)k(x_i, x_j) \tag{6}$$

Constraint to $\sum_{i=1}^{N}(\alpha_i^*-\alpha_i) = 0,\ \alpha_i \geq 0,\ \alpha_i^* \leq C$.

For our present study we have used R programming language [31] for Support Vector Regression in order to forecast a continuous-valued attribute i.e. Air Quality Index (AQI) in our case.

Table 2. Factors of input feature: Wind direction.

Wind Direction	Factors
North wind	0
North-east wind	1
east wind	2
South-east wind	3
south wind	4
North-west wind	5
west wind	6
South-west wind	7
No sustained wind	8

https://doi.org/10.1371/journal.pone.0179763.t002

Table 3. Factors of input feature: Wind power density.

Wind Power Levels	Wind Power Density (W/m^2)	Factors
<3 level	<150	0
3–4 level	150–250	1
4-5level	250–300	2
5–6 level	300–400	3
6–7 level	400–1000	4

https://doi.org/10.1371/journal.pone.0179763.t003

Multi-dimensional collaborative SVR model

The aim of this study is to present a new model for AQI forecasting using collaborative multiple city air quality data as input. The structured flowchart representation of the proposed model is clearly shown below in Fig 3.

In order to forecast the Air Quality Index (AQI) values for city A, we take into account the air quality data of neighboring cities B and C along with city A. We take 3 cases—firstly we take air quality data of City A and City B as input; secondly we take air quality data of City A and City C as input and finally we take air quality data of all 3 cities as input in order to forecast the AQI values for city A. After this we split the three data sets into training and testing data. Then we develop SVR machine learning algorithm on all the three training data sets. Finally, Forecasting is carried out on the testing dataset based on the developed model on the training set. Since the number of input variables or features are increasing when we take into consideration the information from more than one city, the training complexity increases that leads to larger time to train the data. After the training the number of support vectors are selected and at the time of testing the complexity is linear on the number of the support vectors and linear on the number of features. This may vary for different kernels.

Results and analysis

Data distribution

Prediction and forecasting leads us into unknown territory. During the development of a predictive model we must assess its accuracy, reliability and credibility. Hence it is very important to divide the available data into separate partitions, developing our models on one of these partitions and using the other for predictive model assessment and validating it for possibly model refinement. The model development is done on the training set and the prediction is carried on the testing set. For the present study we used 4-fold cross validation technique in order to get accurate and credible results. We split the data into 4 folds with 25% data into each fold. Now we selected 3 folds for training the model and the remaining fold for testing the model. This was done until all the folds for training as well as testing were exhausted in order to avoid any bias in the dataset and also avoid any variations due to different seasons. A total of 851 days daily data is used for our study starting from January 1, 2014 to April 30, 2016. For each experiment the data division into 4 folds of training and testing datasets which is shown below in Table 4.

Beijing-Tianjin-Shijiazhuang air quality index forecast

Beijing air quality index forecast. For Beijing AQI forecast, the experimental analysis is done taking into account the AQI information of different cities in 4 cases. In the first case, individual air pollutants and meteorological data of Beijing namely "$PM_{2.5}$", "PM_{10}", "SO_2", "CO", "NO_2" and "O_3"; five variable weather conditions namely "minimum temperature", " maximum temperature "," weather "," wind direction " and " wind power " along with the last day's observed AQI values are taken as 12 (11+1) input characteristics variables to predict

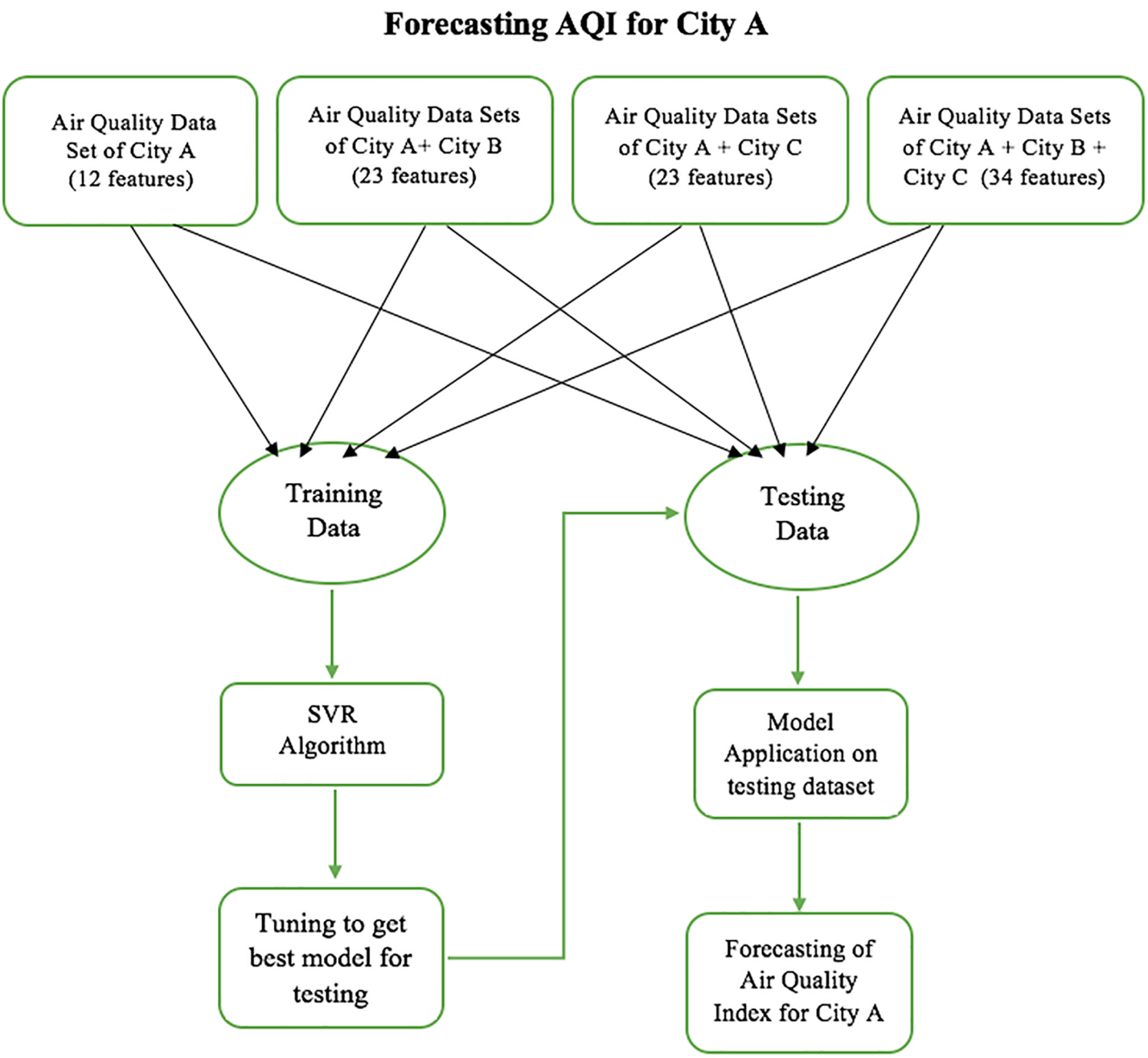

Fig 3. Multi-dimensional codimensional collaborative SVR Model for forecasting AQI values of City A using air quality data sets of city A, city A + city B, city A + city C, city A + city B + city C.

https://doi.org/10.1371/journal.pone.0179763.g003

Beijing's AQI as an output variable using the SVR algorithm. In the second case, the individual air pollutants and meteorological data of Beijing as well as Shijiazhuang along with the last day's observed AQI values are taken as 23 (11+11+1) input characteristics variables to predict Beijing's AQI as an output variable. The third case is exactly similar to the 2nd case except for the fact that we use Tianjin's data instead of Shijiazhuang's data. Finally, in the 4th case we use the individual air pollutants and meteorological data of all the three cities–Beijing, Tianjin and

Table 4. 4-fold cross validation training and testing data division for Beijing, Tianjin and Shijiazhuang.

Sl. No.	Data (training/testing)	Duration	No. of data points (days)
1	Train 1	01.01.2014–30.09.2015	638
	Test1	01.10.2015–30.04.2016	213
2	Train 2	01.08.2014–30.04.2016	639
	Test 2	01.01.2014–31.07.2014	212
3	Train 3	02.03.2015–31.07.2014 *	638
	Test 3	01.08.2014–01.03.2015	213
4	Train 4	30.09.2015–01.03.2015**	639
	Test 4	02.03.2015–29.09.2015	212

* 02.03.2015–31.07.2014 represents data from 2nd March 2015 to 30th April 2016 and from 1st January 2014 to 31st July 2014.

** 30.09.2015–01.03.2015 represents data from 30th September 2015 to 30th April 2016 and from 1st January 2014 to 1st March 2015.

https://doi.org/10.1371/journal.pone.0179763.t004

Shijiazhuang along with the last day's observed AQI values are taken as 34 (11+11+11+1) input characteristics variables to predict Beijing's AQI as an output variable using SVR.

We have used 4-fold cross validation for estimating the error of our models. The 4-fold cross-validation estimator has a lower variance than a single set estimator. If we take a single set, where 75% of data are used for training and 25% used for testing, the test set is very small, hence there will be a lot of variation in the performance estimate for different samples or different partitions of the data to form training and test sets. 4-fold validation reduces this variance by averaging over 4 different partitions, so the performance estimate is less sensitive to the partitioning.

We select MSE (mean square error), the experimental prediction RMSE (root mean square error), MAE (mean absolute error) and MAPE (mean absolute percentage error) to calculate the prediction errors in all the 4 cases and finally to judge the performance of the different model. MSE, RMSE and MAE are scale dependent measures based on squared and absolute error values. MAE is generally smaller than the RMSE and is less sensitive to the large error values as it takes the absolute values and does not square the error value. But still MAE is a common and popular measure for comparing different methods on the same data set due to its simplicity and ease of calculation. On the other hand, MAPE is based on percentage error values and have the advantage of being scale independent. Hence it is the best measure to compare the forecast performance of the Support Vector Regression model between different datasets and needs to be minimum possible. The two most important measures are RMSE and MAPE. The RMSE values of the training and the testing datasets should be less than 12 and almost similar for the training and the testing datasets to conclude that the SVR model is strong and reliable. The MAPE for all the cases should falls between 0.05 ~ 0.09 to indicate accurate prediction results. The two significant digits for RMSE and four significant digits for MAPE are used in Table 5 to clearly show the difference in error values for different air quality information. From Table 5, the prediction error is minimized in the 4th case when we use the Beijing, Tianjin and Shijiazhuang three cities information to predict the AQI values. Fig 4 shows the comparison of the predicted and actual values when training was done on Train 3 dataset and testing was done on Test 3 dataset for Beijing AQI forecast using all 3 city information. Fig 5 shows the comparison of actual and predict values for Beijing AQI forecast based on three city information for the complete dataset.

Table 5. Beijing AQI forecast results comparison based on different urban air quality information.

City Information	MSE	RMSE	MAE	MAPE
Beijing (1 city)	124.11	11.07	7.55	0.0914
Beijing + Shijiazhuang (2 City)	115.39	10.67	7.44	0.0911
Beijing, Tianjin (2 city)	107.20	10.33	7.42	0.0855
Beijing, Tianjin, Shijiazhuang (3 city)	87.69	9.35	6.66	0.0844

https://doi.org/10.1371/journal.pone.0179763.t005

Tianjin and Shijiazhuang air quality index forecast. Tianjin's and Shijiazhuang's AQI forecast is done in a similar way as described for Beijing above. Similar to Beijing, in Table 6, the prediction error is minimized in the fourth case when we use the Beijing, Tianjin and Shijiazhuang three cities information to predict the AQI values. Fig 6 shows the comparison of the predicted and actual values when training was done on Train 3 dataset and testing was done on Test 3 dataset for Tianjin AQI forecast using all 3 city information. Fig 7 shows the comparison of actual and predict values for Tianjin AQI forecast based on three city information for the complete dataset.

In the case of Shijiazhuang, we see in Table 7 that the prediction error is minimized in the first case when we use only the Shijiazhuang information to predict the AQI values. Fig 8 shows the comparison of the predicted and actual values when training was done on Train 3 dataset and testing was done on Test 3 dataset for Shijiazhuang AQI forecast using only one city information. Fig 9 shows the comparison of actual and predict values for Shijiazhuang AQI forecast based on its own city information for the complete dataset.

Discussion

For our model, we have a single algorithm and we are working with different datasets for forecasting and establishing geographical relations. Hence we required a scale independent performance measure for comparing the same algorithm on various datasets. MAPE solves this purpose. From the three cities AQI forecast results it can be observed that surrounding cities air quality information helps in improving forecasting results when there is strong interaction and correlation between the input and the output features. It is also important to note the existence of differences in air quality interaction when we take multiple city air quality information as input. It can be observed that the in order to improve the prediction accuracy, Beijing's AQI forecast needs the help of the air quality information of Tianjin and Shijiazhuang (Table 5). Table 5 clearly shows the reduction in prediction error (MAPE) for our proposed model compared to the single city air quality prediction. Similarly, in the case of Tianjin's AQI forecast, Beijing, Shijiazhuang and Tianjin's air quality information gave the best forecasting results (Table 6). Table 6 clearly shows the reduction in prediction error (MAPE) for the proposed model compared to the single city air quality prediction. Finally, in the case of Shijiazhuang AQI forecast, MAPE is minimum when we use only Shijiazhuang information to predict the AQI values (Table 7). Shijiazhuang air quality is not affected by air quality information of Beijing and Tianjin.

The results can be explained and some important conclusions can be made from the air pollution sources and geographical point of view. Beijing is a social services oriented political capital with multiple high-tech industries. Tianjin is an important economic center as well as an advanced manufacturing center in electronic information, aerospace and automobile manufacturing. Shijiazhuang is the leading heavy industry base based on steel, pharmaceutical, coal and chemical industries. Due to the different cities functional orientations, the sources and amount of air pollution are also different. It can be clearly seen that in general the AQI values observed in Shijiazhuang is more than Beijing and Tianjin (on comparing Figs 5, 7 and 9). Due to the more air pollution in

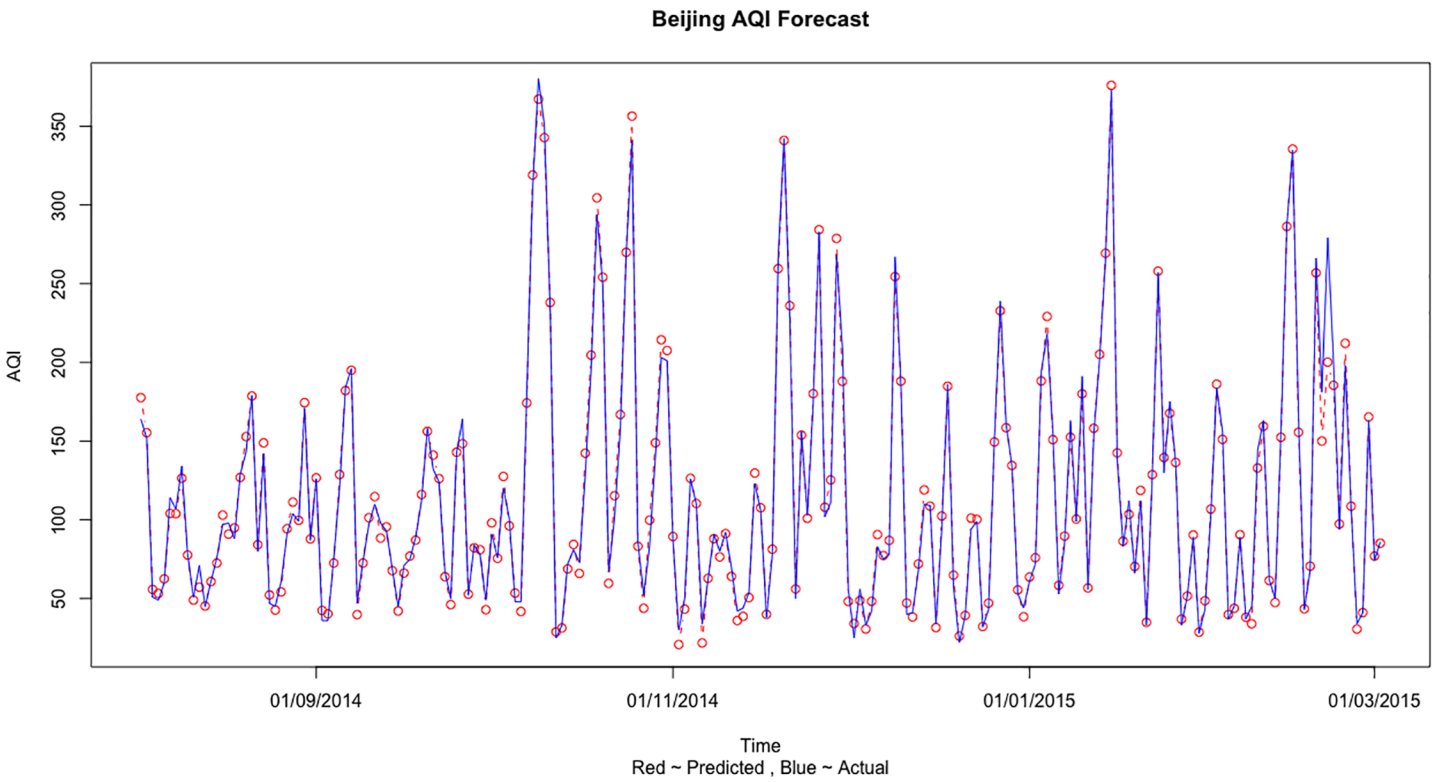

Fig 4. The comparison of actual and predict values for Beijing AQI forecast based on three city information for Test 3 dataset.

https://doi.org/10.1371/journal.pone.0179763.g004

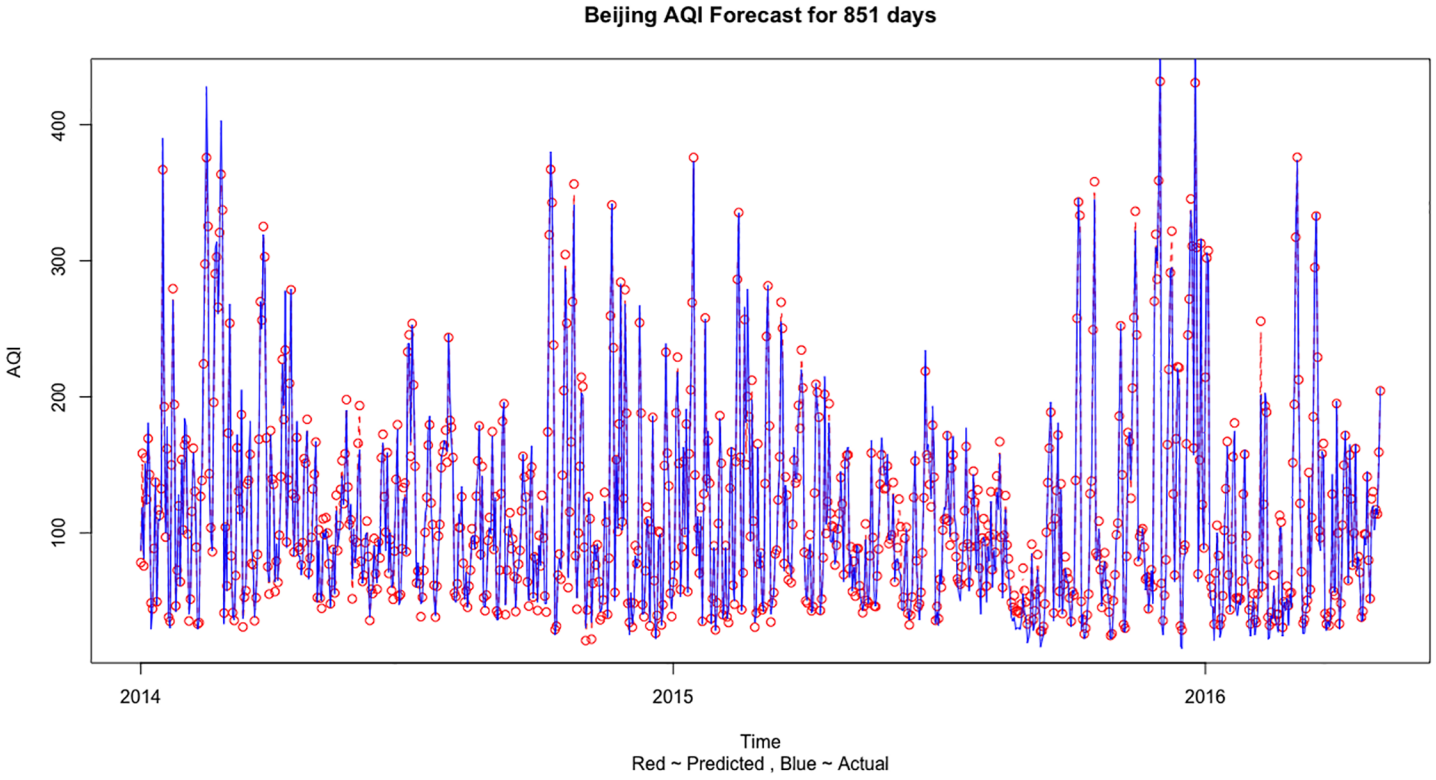

Fig 5. The comparison of actual and predict values for Beijing AQI forecast based on three city information for the complete dataset.

https://doi.org/10.1371/journal.pone.0179763.g005

Table 6. Tianjin AQI forecast results comparison based on different urban air quality information.

City Information	MSE	RMSE	MAE	MAPE
Tianjin (1 city)	63.75	7.97	5.85	0.0597
Tianjin, Shijiazhuang (2 city)	61.01	7.78	5.75	0.0584
Tianjin, Beijing (2 city)	56.26	7.44	5.53	0.0579
Beijing, Tianjin, Shijiazhuang (3 city)	42.78	6.54	4.90	0.0534

https://doi.org/10.1371/journal.pone.0179763.t006

Shijiazhuang, the surrounding areas of Tianjin and Beijing are also affected. So, Shijiazhuang air quality information will play an important role in Beijing, Tianjin and Shijiazhuang AQI forecast. It is important to look at the geographical locations of the three cities (see Fig 1). It can be seen that Beijing and Tianjin are located very close and they also share common boundaries. Therefore, we saw a similar forecasting result for both the cities. Whereas Shijiazhuang is a bit farther away from both Beijing and Tianjin and hence we see a different result for Shijiazhuang AQI forecast. Also, on comparing the 2-city and 3-city results for Beijing AQI forecast in Table 5, the decrease of MSE is more evident when Tianjin is introduced in the model (115.39 to 87.69) than Shijiazhuang (107.20 to 87.69). Thus it can be inferred from our observations that the geographical location plays an important role in AQI forecast.

In this paper we have used U.S. AQI based on daily values of several real time pollutant concentrations and the decision on using this indicator scheme was based on simplicity rather than on exact scientific reasoning. In order to further improve the air quality forecast in the future it is advisable to adopt the Air quality indicators that are close to the atmospheric reality

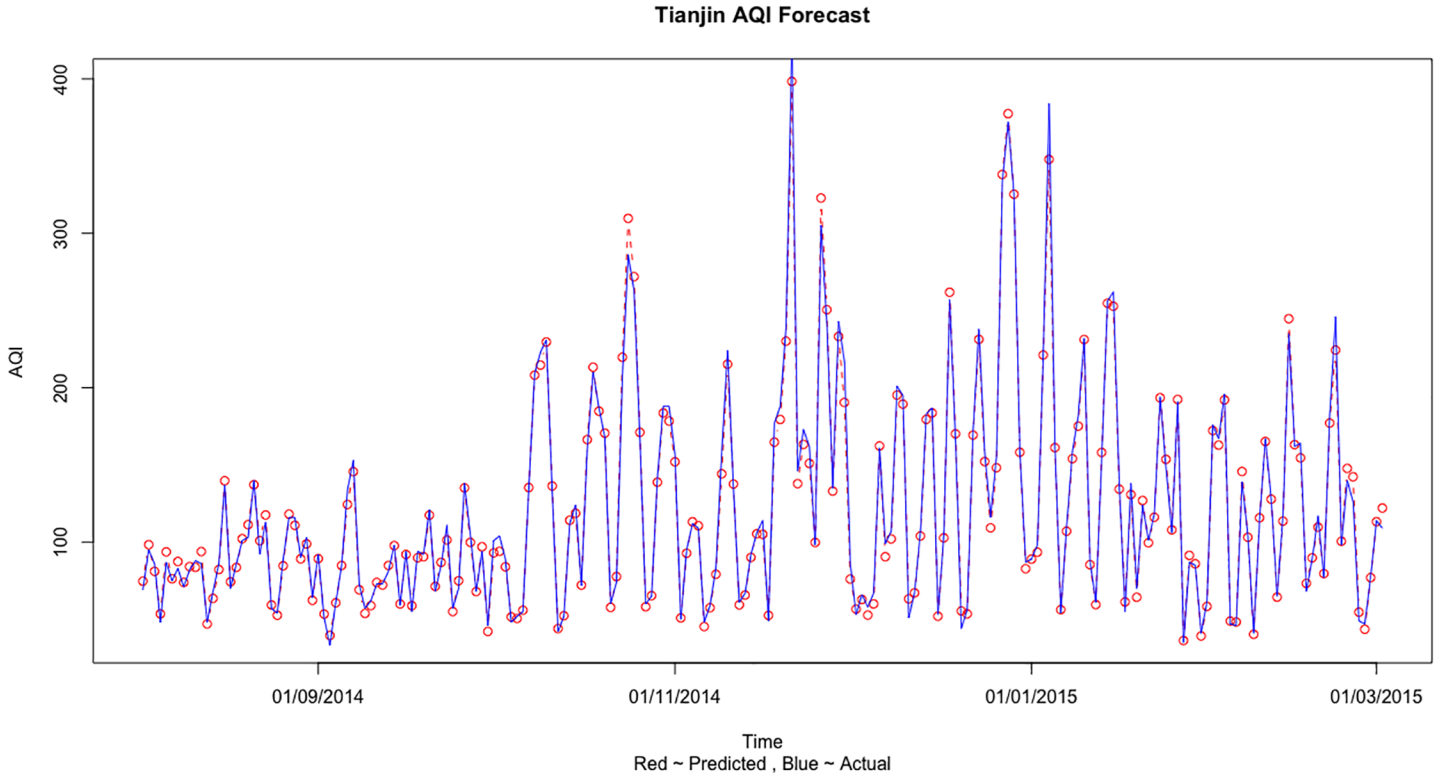

Fig 6. The comparison of actual and predicted values for Tianjin AQI forecast based on 3 city information for Test 3 dataset.

https://doi.org/10.1371/journal.pone.0179763.g006

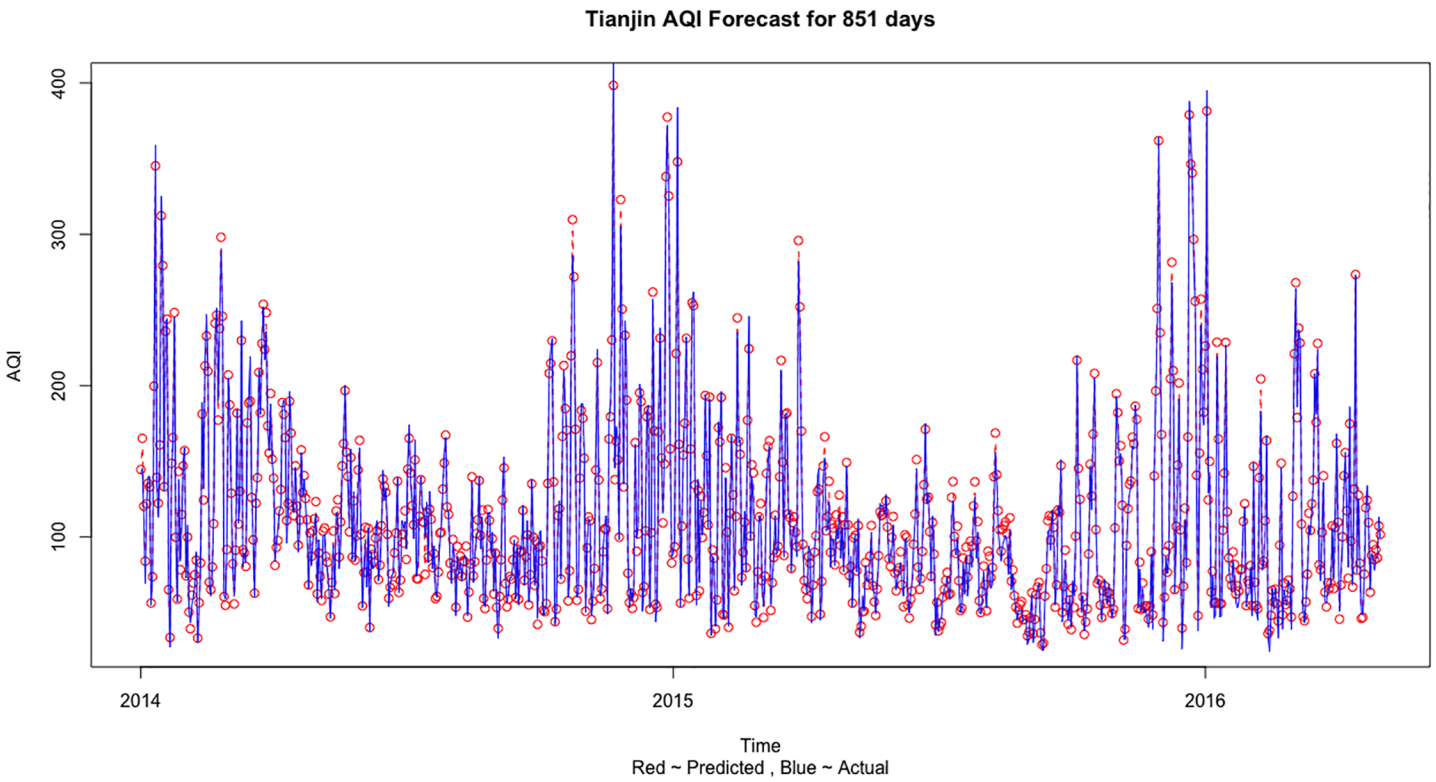

Fig 7. The comparison of actual and predict values for Tianjin AQI forecast based on three city information for the complete dataset.

https://doi.org/10.1371/journal.pone.0179763.g007

[32]. These air quality indicators when used with hourly air pollutant concentration data could help in dealing with the auto-cancelling effects and rapid chemical transformation of some of the pollutants after they are released in the atmosphere.

Conclusion

For the present study we selected different urban air quality information of Beijing, Tianjin and Shijiazhuang for AQI forecasting. By conducting the three cities AQI forecasting experiments we have reached to the following conclusions: The RMSE values of the training and the testing datasets are <12 and are almost similar for the training and the testing datasets for most of the cases, hence we can conclude that the support vector regression model is strong and reliable for predicting the AQI values. If the RMSE values for the test set is very higher than that of the training set then there is a problem of overfitting the data i.e. the model performs well on the training set but fails to give good predictions on the test set. Also the MAPE for all the cases falls between 0.05 ~ 0.09 and indicates highly accurate prediction result.The

Table 7. Shijiazhuang AQI forecast results comparison based on different urban air quality Information.

City Information	MSE	RMSE	MAE	MAPE
Shijiazhuang (1 city)	106.22	9.66	6.77	0.0590
Shijiazhuang, Tianjin (2 city)	112.82	10.25	7.29	0.0620
Shijiazhuang + Beijing (2 city)	122.76	10.70	7.48	0.0663
Beijing, Tianjin, Shijiazhuang (3 city)	128.70	11.01	7.72	0.0731

https://doi.org/10.1371/journal.pone.0179763.t007

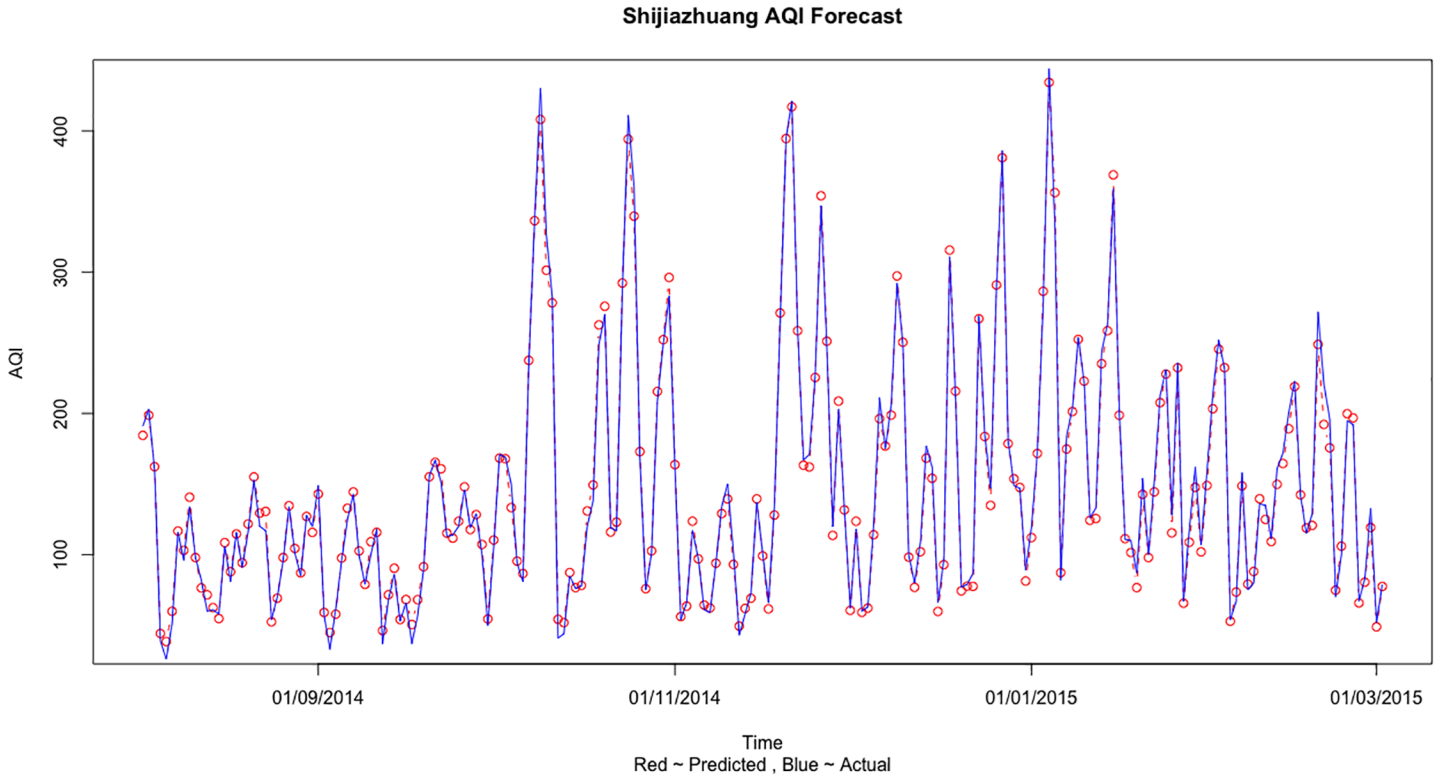

Fig 8. The comparison of actual and predicted values for Shijiazhuang AQI forecast based on its own city information for Test 3 dataset.

https://doi.org/10.1371/journal.pone.0179763.g008

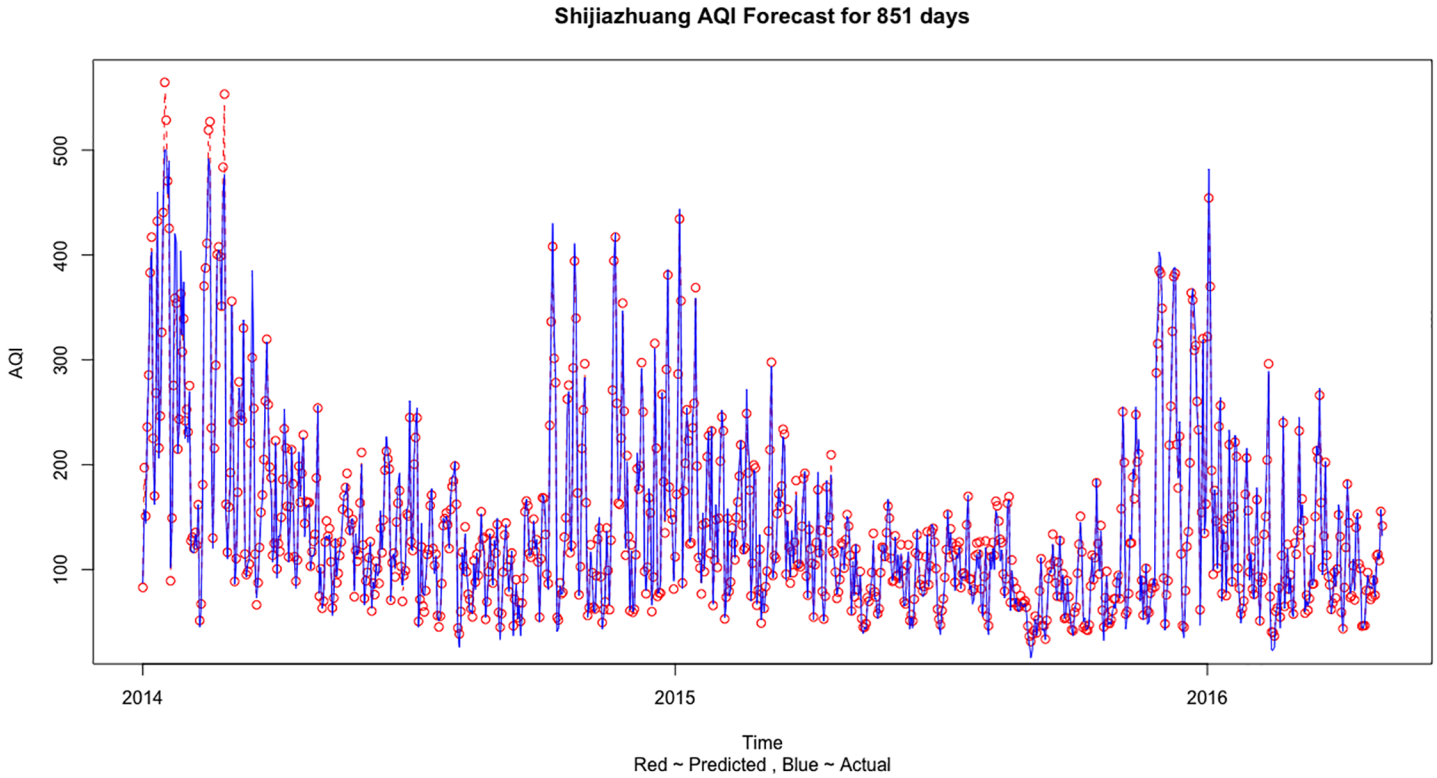

Fig 9. The comparison of actual and predict values for Shijiazhuang AQI forecast based on its own city information for the complete dataset.

https://doi.org/10.1371/journal.pone.0179763.g009

analysis shows that Shijiazhuang suffers a more serious air pollution because of its heavy industry base and needs adjustments and industrial upgrading to improve its present conditions. This new prediction method for air quality forecasting could provide better scientific basis for this kind of research work. This study was done on a particular region for a small period of time but it still shows that by taking careful advantage of the multiple city information based on the geographical location as well as using the weather conditions, we can expect a significant improvement in the forecasting results. We understand that in order to generalise the results elsewhere or for a different temporal series will require future readjustments and some modifications. This paper also opens the doors for further research and exploration of different forecasting methods and machine learning techniques in order to achieve better accuracy of air quality prediction in regression mode. Further research work could also be to use deep learning models along with multi city forecasting which can automatically extract the useful information required to forecast.

Supporting information

S1 Table. Latitude and longitude values of air quality monitoring stations in Beijing.
(DOCX)

S2 Table. Latitude and longitude values of air quality monitoring stations in Tianjin.
(DOCX)

S3 Table. Latitude and longitude values of air quality monitoring stations in Shijiazhuang.
(DOCX)

Acknowledgments

The authors would like to acknowledge the funding support from the National Natural Science Foundation of China. The authors would also like to thank Anurag Tiwari for his valuable feedback.

Author Contributions

Conceptualization: BCL PCC.

Data curation: BCL AB PCC MKT CCT.

Formal analysis: BCL AB PCC MKT CCT.

Funding acquisition: BCL.

Investigation: BCL AB PCC MKT CCT.

Methodology: BCL AB PCC MKT CCT.

Project administration: BCL AB PCC MKT CCT.

Resources: BCL AB PCC MKT CCT.

Software: BCL AB CCT.

Supervision: BCL PCC MKT.

Validation: BCL AB PCC MKT CCT.

Visualization: BCL AB PCC MKT CCT.

Writing – original draft: BCL AB PCC MKT CCT.

Writing – review & editing: BCL AB PCC MKT CCT.

References

1. Liu DJ, Li L. Application Study of Comprehensive Forecasting Model Based on Entropy Weighting Method on Trend of PM2.5 Concentration in Guangzhou, China. Int J Environ Res Public Health. 2015; 12(6):7085–99. https://doi.org/10.3390/ijerph120607085 PMID: 26110332
2. Chen W, Wang F, Xiao G, Wu K, Zhang S. Air Quality of Beijing and Impacts of the New Ambient Air Quality Standard. Atmosphere. 2015; 6(8):1243–1258. https://doi.org/10.3390/atmos6081243
3. Li L, Liu DJ. Study on an Air Quality Evaluation Model for Beijing City under Haze-Fog Pollution Based on New Ambient Air Quality Standards. Int J Environ Res Public Health. 2014; 11(9):8909–8923. https://doi.org/10.3390/ijerph110908909 PMID: 25170682
4. Hu L, Liu J, He Z. Self-Adaptive Revised Land Use Regression Models for Estimating PM2.5 Concentrations in Beijing. China Sustainability. 2016; 8(8):786. https://doi.org/10.3390/su8080786
5. Rohde RA, Muller RA. Air Pollution in China: Mapping of Concentrations and Sources. PLoS ONE. 2015; 10(8): e0135749. https://doi.org/10.1371/journal.pone.0135749 PMID: 26291610
6. Wang K-Y, Chau T-T. An Association between Air Pollution and Daily Outpatient Visits for Respiratory Disease in a Heavy Industry Area. PLoS ONE. 2013; 8(10): e75220. https://doi.org/10.1371/journal.pone.0075220 PMID: 24204573
7. Plain A, Ruggieri M. Air quality indices: a review. Rev Environ Sci Biotechnol. 2011; 10:165–179. https://doi.org/10.1007/s11157-010-9227-2
8. McKeen S, Grell G, Peckham S, Wilczak J, Djalalova I, Hsie EY et al. An evaluation of real-time air quality forecasts and their urban emissions over eastern Texas during the summer of 2006 Second Texas Air Quality Study field study. Journal of Geophysical Research. 2009; 114: D00F11. https://doi.org/10.1029/2008JD011697
9. Chuang M-T, Zhang Y, Kang D. Application of WRF / Chem-MADRID for real-time air quality forecasting over the southeastern United States. Atmos.Environ. 2011; 45:6241–6250. https://doi.org/10.1016/j.atmosenv.2011.06.071
10. Xiaofeng XU, Qingchun L, Xiaoling Z. Beijing time local weather conditions of heavy pollution process analysis. Meteorological Science and Technology. 33 2005; 6:543–547.
11. Kumar A, Goyal P. Forecasting of air quality in Delhi using principal component regression technique. Atmospheric Pollution Research. 2011; 2:436–444. https://doi.org/10.5094/APR.2011.050
12. Hu Y, Chang ME, Russell AG, Odman MT. Using synoptic classification to evaluate an operational air quality forecasting system in Atlanta. Atmospheric Pollution Research. 2010; 1:280–287. https://doi.org/10.5094/APR.2010.035
13. Stern R, Builtjes P, Schaap M, Timmermans R, Vautard R, Hodzic A et al. A model inter-comparison study focusing on episodes with elevated PM10 concentrations. Atmos.Environ. 2008; 42:4567–4588. https://doi.org/10.1016/j.atmosenv.2008.01.068
14. Chen C-Q, Chen Y-Q, Tang S-J, WU S-J. Analysis of Effect of Meteorological Factor on Air Quality of Wuhan in 2001–2010. Environmental Science & Technology. 2013; 36(5):130–133.
15. Perez P, Reyes J. An integrated neural network model for PM10 forecasting. Atmos.Environ. 2006; 40:2845–2851. https://doi.org/10.1016/j.atmosenv.2006.01.010
16. Li M, Hassan R. Urban Air Pollution Forecasting Using Artificial Intelligence-Based Tools. Air Pollution. Vanda Villanyi (Ed.). ISBN: 978-953-307-143-5, InTech. 9:195–219.
17. Bai H-M, Shen R-P, Shi H-D, Dong Y-C. Forecasting model of air pollution index based on BP neural network. Environmental Science & Technology. 2013; 36(3) 186–189.
18. Shad R, Mesgari MS, Abkar A, Shad A. Predicting air pollution using fuzzy genetic linear membership Kriging in GIS. Comput.Environ.Urban Syst. 2009; 33:472–481. https://doi.org/10.1016/j.compenvurbsys.2009.10.004
19. Alhanafy TE, Zaghool F, Moustafa ASED. Neuro fuzzy modeling scheme for the prediction of air pollution. Journal of American Science. 2010; 6:605–616.
20. Zaharim A, Shaharuddin M, Nor MJM, Karim OA, Sopian K. Relationships between airborne particulate matter and meteorological variables using non-decimated wavelet transform. European Journal of Scientific Research. 2009; 27(2):308–312.
21. Hoi KI, Yuen KV, Mok KM. Kalman filter based prediction system for wintertime PM10 concentrations in Macau. Global NEST Journal. 2008; 10:140–150.
22. Liu Y, Zhu Q, Yao D, Xu W. Forecasting Urban Air Quality via a Back-Propagation Neural Network and a Selection Sample Rule. Atmosphere. 2015; 6(7):891–907. https://doi.org/10.3390/atmos6070891

23. Xiang L. Air quality forecasting based on GAB and fuzzy BP neural network. Journal of Huazhong University of Science and Technology(Nature Science). 2013; 41(1):63–69.

24. Sun W, Zhang H, Palazoglu A, Singh A, Zhang W, Liu S. Prediction of 24-hour-average PM2.5 concentration using a hidden Markov model with different emission distributions in Northern California. Science of the Total Environment. 2013; 443:93–103. https://doi.org/10.1016/j.scitotenv.2012.10.070 PMID: 23178893

25. Schölkopf B, Smola AJ, Williamson RC, Bartlett PL. New support vector algorithms. Neural Computation. 2000; 12:1207–1245. https://doi.org/10.1162/089976600300015565 PMID: 10905814

26. Schwieder M, Leitão PJ, Suess S, Senf C, Hostert P. Estimating Fractional Shrub Cover Using Simulated EnMAP Data: A Comparison of Three Machine Learning Regression Techniques. Remote Sensing. 2014; 6(4):3427–3445. https://doi.org/10.3390/rs6043427

27. Konar P, Chattopadhyay P. Bearing fault detection of induction motor using wavelet and Support Vector Machines (SVMs). Applied Soft Computing. 2011; 11:4203–4211. https://doi.org/10.1016/j.asoc.2011.03.014

28. Queensland Government. Influence of Meteorology on Air Quality. March 2017. Available from: https://www.qld.gov.au/environment/pollution/monitoring/air-monitoring/meteorology-factors/

29. Müller KR, Smola AJ, Rätsch G, Schölkopf B, Kohlmorgen J, Vapnik V. Predicting time series with support vector machines. Artificial Neural Networks ICANN'97; 999–1004. https://doi.org/10.1007/BFb0020283

30. Vapnik V, Golowich S, Smola A. Support vector method for function approximation, regression estimation, and signal processing. Advances in Neural Information Processing Systems. 1997; 9:281–287.

31. R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available URL: https://www.R-project.org/.

32. Kassomenos P, Skouloudis AN, Lykoudis SP, Flocas HA, "Air-quality indicators" for uniform indexing of atmospheric pollution over large metropolitan areas. Atmospheric Environment June 1999; 33 (12):1861–1879.