Pol. J. Environ. Stud. Vol. 27, No. 2 (2018), 565-578

DOI: *10.15244/pjoes/75159* ONLINE PUBLICATION DATE: 2017-12-28

Original Research

Investigating China's Urban Air Quality Using Big Data, Information Theory, and Machine Learning

Sheng Chen[1, 2], Guangyuan Kan[1, 3]*, Jiren Li[1], Ke Liang[2], Yang Hong[3, 4]

[1]State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, Research Center on Flood and Drought Disaster Reduction of the Ministry of Water Resources, China Institute of Water Resources and Hydropower Research, Beijing 100038, P.R. China

[2]College of Hydrology and Water Resources, Hohai University, Nanjing 210098, P.R. China

[3]State Key Laboratory of Hydroscience and Engineering, Department of Hydraulic Engineering, Tsinghua University, Beijing 100084, P.R. China

[4]Department of Civil Engineering and Environmental Science, University of Oklahoma, Norman, OK, USA

Received: 8 June 2017
Accepted: 20 June 2017

Abstract

With the development of the economy and industrial construction, air quality deteriorates dramatically in China and seriously threatens people's health. To investigate which factors most affect air quality and provide a useful tool to assist the prediction and early warning of air pollution in urban areas, we applied a sensor that observed air quality big data, information theory-based predictor significance identification, and PEK-based machine learning to air quality index (AQI) analysis and prediction in this paper. We found that the stability of air quality has a high relationship with absolute air quality, and that improvement of air quality can also improve stability. Air quality in southern and western cities is better than that of northern and eastern cities. AQI time series of cities with closer geophysical locations have a closer relationship with others. PM2.5, PM10, and SO_2 are the most important impact factors. The machine learning-based prediction is useful for AQI prediction and early warning. This tool could be applied to other city's air quality monitoring and early warning to further verify its effectiveness and robustness. Finally, we suggested the use of a training data sample with better quality and representatives to further improve AQI prediction model performance in future research.

Keywords: urban air quality, analysis, prediction, information theory, machine learning

Introduction

With the continued rapid development of the global economy – especially industrialization – environmental pollution issues such as urban air pollution have become more and more serious [1-7]. Urban areas and the rapid progress of industrialization and technology have led to serious air pollution in urban areas. In particular, human beings trying to maintain urban life harm health. In Europe, more than two-thirds of the total population lives in cities. Population growth and industrialization have

*e-mail: kanguangyuan@126.com

led to air pollution in some cities that reaches levels that threaten human health. This has become one of the most important topics of our day. Human health is affected by all air pollution, but some emissions have more severe atmospheric qualities. In particular, carbon dioxide (CO_2) and other pollutants, which provide global warming, have recently attracted attention as CO_2 is one of the most researched gases. Recent studies have shown PM10 and PM2.5 CO_2 air quality indices [8-19].

Urban air pollution not only affects people's health but also dramatically threatens human lives. During 5-8 December 1952, the "London smog incident" killed 4,000 people in only four days. Two months later, 8,000 more victims died. The world health organization (WHO) estimates that 2 million minors die of air pollution each year. China is one of the countries that suffers from severe air pollution problems and that faces dramatically serious urban air pollution. According to the 2007 report of China's Environmental Protection Department, the air quality of China's prefecture levels and certain cities (including the capitals of the states and unions) are not good. The percentages of cities with first, second, third, and fourth classes are 2.4%, 58.1%, 36.1%, and 3.4%, respectively [20-21]. It was reported by China's Ministry of Health in 2007 that the nation's death toll caused by urban air pollution was 178,000 annually. Respiratory clinic cases caused by poor urban air quality amount to 350,000, while emergency cases are 6.8 million. Air pollution brings about environmental and health damage and causes a 7% loss of China's GDP. Therefore, air quality issues are no longer accidental natural disasters and have become dramatic problems that must be coped with in everyday life.

The Chinese economy has been developing quickly and the urbanization process is accelerating. The urban population, city scale, number of motor vehicles, and industrialization are all sharply increasing. Therefore, the air pollution caused by inhalable particulate matter, sulfur dioxide, and nitrogen oxide have become increasingly serious. These urgent situations have threatened the sustainable development of the national economy and health. The requirement of improving urban air quality and carrying out a strategy of sustainable development of modern cities is increasingly prominent. The principle of environmental protection of China is setting air environment monitoring as a guide and providing scientific support for air environmental protection and management through a real-time monitoring site network, which is a device-based system of automatic monitoring sets. It is responsible for the online air pollution monitoring, telnet, pollution warning, and visualized monitoring, etc. These kinds of systems can quickly report urban air quality through a wired or wireless transmission network. Apart from the monitoring technology, China has conducted weekly and daily air quality reports and predictions, which are urgently needed. Under this situation, it is of great significance to carry out urban air quality prediction, which is in great need for better reflecting urban air pollution variations. Precise and reliable air quality prediction can contribute greatly to preventing the occurrence of serious pollution, enhancing public awareness of environment protection, and improving public life quality.

The objective of air quality evaluation is to assess environment quality and reflect human requirements for the environment [22]. Air quality prediction is usually based on the air quality evaluation criteria. In China, in 2012 a new air quality evaluation criterion named air quality index (AQI) was proposed to replace the previously adopted air pollution index (API). Different from the API, the newly proposed AQI is based on the new environmental air quality standard (GB3095-2012) and considers pollutants such as SO_2, NO_2, PM10, PM2.5, O_3, and CO, etc. On the other hand, the older API criterion is based on the old environmental air quality standard (GB3095-1996) and only assesses SO_2, NO_2, and PM10. Furthermore, the AQI adopts a stricter classification standard than the API and therefore AQI criterion is much more objective and precise.

The AQI is a dimensionless index used to describe air quality. The AQI considers pollutants including fine particulate matter, inhalable particulate matter, SO_2, NO_2, O_3, and CO, etc. Air quality can be classified into six levels according to AQI. Higher AQI indicates a worse air pollution quality and a more severe threat to human health. The AQI levels of 1 to 6 indicate excellent, good, mild pollution, moderate pollution, heavy pollution, and serious pollution, respectively. According to AQI technical specification HJ633-2012 (for trial implementation), the AQIs and corresponding air quality levels are as follows:

- **AQI 0 to 50:** air quality is **level 1** and excellent, with satisfactory quality and almost no pollution; all people can perform normal activities.
- **51 to 100:** air quality is **level 2** and good, with acceptable quality but some pollutants that may have a weak effect on the health of a few people with abnormal sensitivity; a small number of extremely sensitive people should reduce outdoor activities.
- **101 to 150:** air quality is **level 3** with mild pollution; discomfort symptoms happened to susceptible people to a slight degree and irritative symptoms can appear in healthy people; it is recommended that children, the elderly, and patients with heart or respiratory diseases should reduce prolonged and intense outdoor activities.
- **151 to 200:** air quality is **level 4** with moderate pollution; discomfort symptoms are aggravated in susceptible people and the pollution may have an adverse effect on healthy hearts and respiratory systems; it is recommended that patients with chronic diseases should avoid prolonged and intense outdoor activities, and that healthy people should reduce outdoor sports.
- **201 to 300:** air quality is **level 5** with heavy pollution; symptoms in patients with heart and lung diseases significantly increase, while healthy people can feel ill and their exercise tolerances decrease; it is recommended that children, the elderly, and heart

and lung disease patients should stay indoors and stop outdoor activities, and that the general population should reduce outdoor sports.

- **301 and above:** air quality is **level 6** with serious pollution; exercise tolerance of healthy people decreases significantly and strong uncomfortable symptoms appear; it is suggested that children, the elderly, and patients should stay indoors and avoid physical exertion.

The variation of urban air quality is not an untraceable random change process and is affected by some specific impact factors. Urban air quality is mainly influenced by impact factors such as meteorological factors, pollutants, and terrain factors, etc. Therefore, urban air quality is predictable and we can construct an urban air quality prediction method based on these impact factors and provide a useful tool for environmental quality monitoring and protection.

Air quality prediction can be traced back to the 1960s in western developed countries. The real-world application of modern urban air quality prediction is beginning from Japan's nitrogen oxide prediction for Osaka and Tokyo in December 1988. The Netherlands started to carry out nationwide air pollution prediction in 1989 and finished the daily mean pollutant concentration prediction in three years. They implemented predictions with lead time from several hours to three days based on various types of forecast methods. The Danish National Institute for Environmental Studies proposed an air pollution forecast system, the THOR model, which is based on numerical simulations. The model is composed by the urban background and meteorological model, and can predict air pollution for Denmark and Europe three days into the future.

Air pollution prediction research in China started later compared with other countries. In the 1980s, Beijing carried out two potential SO_2 pollution forecasts. Then Shenyang and Shanghai introduced daily air quality prediction in their everyday work. In 1997 the Chinese Academy of Meteorological Sciences (CAMS) developed a City Air Pollution Numerical Prediction System (CAPPS) to forecast pollution quality of the next day according to the previous day's SO_2, NO_2, and PM10 values. CAMS introduced the CAPPS to 47 Chinese cities and provided new technology to the air quality prediction in 2000.

Lots of methods have been applied to air quality prediction. Among these methods, data-driven methods have been recognized as useful tools. Data-driven methods include multiple linear regression, the grey forecasting method, and artificial neural networks, etc. Regression analysis describes the relationship between variables based on possibility and statistical average. Regression methods can be categorized into two types, including linear and nonlinear methods. Air quality prediction usually adopts linear regression methods. The grey system theory was proposed by Deng in 1982. It is a new information theory which is used to study the uncertainty problems with a small data set. Grey forecast has several kinds of models, such as sequence grey prediction, disaster grey prediction, seasonal catastrophic grey prediction, topology grey prediction, system grey prediction, and envelope grey prediction. An artificial neural network (ANN) is a mathematical network composed of a huge number of artificial neurons and is good at modelling, pattern recognition, signal processing and control, and predictions. Until now, the ANN has been widely applied for predictions related to the economy, energy, industry, agriculture, the environment, transportation, and water conservancy.

In this research we studied the big data-based air quality of 16 large cities in China: Beijing, Changchun, Changsha, Guangzhou, Harbin, Hefei, Hohhot, Jinan,

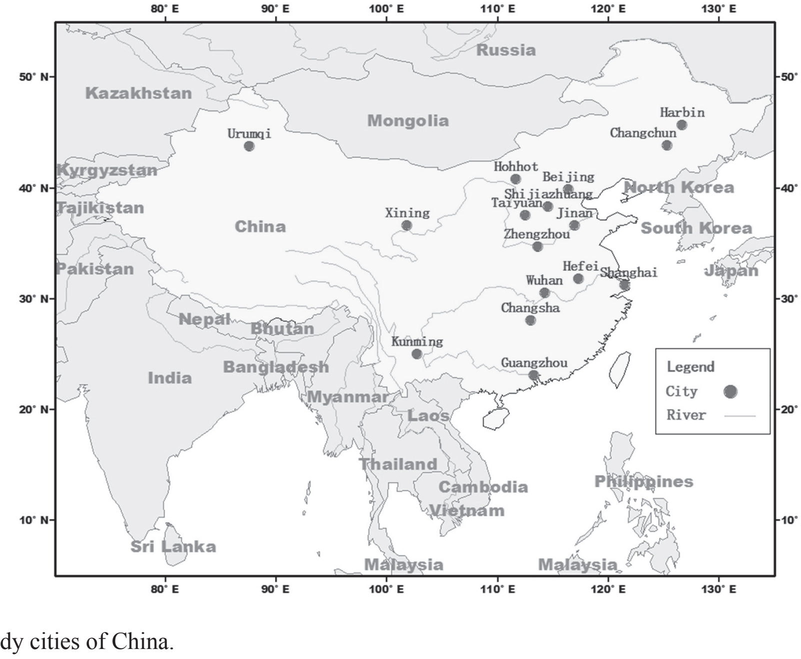


Fig. 1. Location map of study cities of China.

Kunming, Shanghai, Shijiazhuang, Taiyuan, Urumqi, Wuhan, Xining, and Zhengzhou. Urban air quality was analyzed and predicted using big data, information theory, and machine learning-based methods to investigate which factors impact urban air quality significantly and provide a way to monitor pollution and warn the population. This article aims to arouse public attention to urban air pollution big data-based sensor monitoring and protection in order to reduce environmental pollution and provide new methods for analyzing and predicting air quality.

Material and Methods

Study Area and Data Description

The 16 large cities we chose are provincial capitals and are distributed throughout China (Fig. 1). The cities' daily AQIs are adopted as urban air quality indicators and evaluation criteria. The AQI is predicted according to at most 22 potential predictors collected from the Meteorological Administration Department. The potential predictors contain meteorological information such as precipitation, atmospheric pressure, and humidity. The predictors also contain information of pollutants such as PM2.5, PM10, and SO_2. All 22 potential predictors are listed in Table 1. The data range of the daily AQI and potential predictors is from 1 January 2014 to 30 November 2016, with a total of 1,065 data points. In this research 2014 and 2015 data are used as a calibration set while 2016 data are used as a validation set.

Table 1. Predictors of urban air quality.

Physical meaning	Notation	Index
Precipitation (mm)	prec	1
Average atmospheric pressure (hpa)	prsavg	2
Maximum atmospheric pressure (hpa)	prsmax	3
Minimum atmospheric pressure (hpa)	prsmin	4
Relative humidity (%)	rhu	5
Sunshine duration (h)	ssd	6
Average temperature (ºC)	temavg	7
Maximum temperature (ºC)	temmax	8
Minimum temperature (ºC)	temmin	9
Average windspeed (m/s)	windavg	10
Maximum windspeed (m/s)	windmax	11
Wind direction	winddir	12
Average ground temperature (ºC)	gtemavg	13
Maximum ground temperature (ºC)	gtemmax	14
Minimum ground temperature (ºC)	gtemmin	15
Evapatranspiration (mm)	evap	16
PM2.5 (µg/m^3)	pm25	17
PM10 (µg/m^3)	pm10	18
SO_2 (µg/m^3)	so_2	19
NO_2 (µg/m^3)	no_2	20
CO (mg/m^3)	co	21
Ozone (µg/m^3)	o_3	22

Information Theory-Based Predictor Significance Analysis and Selection

Mutual Information

Mutual information (MI) is the fundamental information measure in the area of information theory. This criterion is generally considered a measure of nonlinear dependence between two variables. It can also be considered a measure of the stored information in one variable about another, or the measure of the degree of predictability of the output variable knowing the input variable.

Given two discrete random variables X and Y, the MI between these two variables is defined as follows:

$$MI = \frac{1}{N}\sum_{i=1}^{N} \ln\left[\frac{f_{X,Y}\left(x_i, y_i\right)}{f_X\left(x_i\right) f_Y\left(y_i\right)}\right]$$

…where (x_i, y_i) is the ith bivariate sample pair with i = 1, 2, …, N; $f_{X,Y}(x_i, y_i)$ is the joint probability density function between x_i and y_i; and $f_X(x_i)$ and $f_Y(y_i)$ are the univariate probability densities estimated for each sample point. If X and Y are independent, the joint probability density will be equal to the product of the marginal densities and therefore the MI between them will be 0. If the two variables are strongly dependent, then the joint density will be greater than the product of the marginal probabilities and the MI will be larger than 0.

As can be seen, MI is a good measurement for selecting significant attributes as an attempt to model the system under study. However, this criterion is not able to deal with redundant inputs. For example, if there is a third variable Z highly correlated to X (such as Z = 5X), it will also have a high MI value, and both X and Z would be selected as significant input features according to the MI criterion. In this case, Z would be a redundant variable since it can be completely described by X.

Partial Mutual Information

In order to overcome the disadvantage of MI criterion in predictor significance selection, Sharma [23-24] proposed the partial mutual information (PMI) criterion, which is an extension of MI. Moreover, it is able to capture all dependence between two variables and as it is a model-free strategy, it is not necessary to define a model structure a priori. PMI is a measure of the partial or additional dependence that a new input can add to the existing prediction model [25-26]. Given a dependent discrete variable Y (the output of the model), and an input X (the independent discrete variable), for a set of pre-existing inputs Z, the discrete version of the PMI criterion is defined as:

$$PMI = \frac{1}{N}\sum_{i=1}^{N} \ln\left[\frac{f_{X',Y'}\left(x_i',y_i'\right)}{f_{X'}\left(x_i'\right)f_{Y'}\left(y_i'\right)}\right]$$

…where:

$$x_i' = x_i - E\left(x_i|Z\right)$$

and:

$$y_i' = y_i - E\left(y_i|Z\right)$$

…where $E(\cdot)$ denotes the expectation function; x_i' and y_i' represent the residual components corresponding to the ith data pair sample, i = 1, 2, …, N; and $f_{X'}(x_i')$ $f_{Y'}(y_i')$ and $f_{X',Y'}(x_i', y_i')$ are the respective marginal and joint probability densities. The resulting variables X' and Y' represent only the residual information in variables X and Y once the effects of the existing predictors in Z have been taken into account.

A good estimate of the expectation function is necessary for calculating the PMI criterion. Sharma used an approximation based on Gaussian kernel functions (assuming a normal distribution of the samples) and Bowden [27-28] proposed a general regression neural network. In this research, the expected values are approximated via the Nadaraya-Watson Estimator and the city block kernel function, instead of Gaussian functions. Using the city block kernel function and the kernel estimators, it is not necessary to assume any particular form for the regression function as considered in previous literature. Indeed, computational complexity diminishes once it is not necessary to calculate covariance matrix and its respective inverse, reducing the processing time required. Let $r(x) = E\ (Y|X = x)$. An approximation of $r(x)$ is defined by:

$$\hat{r}\left(x\right) = \sum_{i=1}^{N} \omega_{\lambda_x}\left(x,x_i\right) y_i$$

…where

$$\omega_{\lambda_x}\left(x,x_i\right) = \frac{K_{\lambda_x}\left(x-x_i\right)}{\sum_{j=1}^{N} K_{\lambda_x}\left(x-x_j\right)}$$

…and K_{λ_x} $(x - x_i)$ is the kernel function for x, which in this work is represented by the city block function.

There are different ways for approximating both marginal and joint probabilities. This research approximates marginal and joint probability functions via kernel functions once it is an efficient and robust tool, as shown by Sharma and Moon. Let be N pairs of data sample [x(k), y(k)], with x(k) being an input variable and y(k) being, without loss of generality, its corresponding scalar output of k = 1, 2, …, N. The classic multivariate probability density estimator is given by:

$$f\left(x\right) = \frac{1}{N\lambda}\sum_{i=1}^{N} K\left(\frac{x-x_i}{\lambda}\right) = \frac{1}{N}\sum_{i=1}^{N} K_\lambda\left(x-x_i\right)$$

… where K_λ (t) is the kernel function, and λ is the bandwidth parameter. The kernel function is required to be a valid probability density function. Using kernel functions based on the city block distance, the above equation may be rewritten as follows:

$$f_x\left(x\right) = \frac{1}{N\left(2\lambda\right)^p}\sum_{i=1}^{N}\prod_{j=1}^{p} \exp^{-\left|x_j - x_{ij}\right|/\lambda}$$

…that is,

$$f_x\left(x\right) = \frac{1}{N\left(2\lambda\right)^p}\sum_{i=1}^{N} \exp\left(-\frac{1}{\lambda}\sum_{j=1}^{p}\left|x_j - x_{ij}\right|\right)$$

…where p is the dimension of each xi and i = 1, 2, …, N. The bandwidth parameter λ is calculated by the equation:

$$\lambda = \left(\frac{4}{p+2}\right)^{1/(p+4)} N^{-1/(p+4)}$$

…even though the above equation was estimated assuming Gaussian distributions. The estimation of the above equation has been widely applied in previous literature because of its efficiency and simplicity, and for that reason it will be applied in this research. Finally, joint probability density estimation of x given y is defined by:

$$f_{xy}\left(x,y\right) = \frac{1}{N}\sum_{i=1}^{N} K_\lambda\left(\frac{x-x_j}{\lambda}\right) K_{\lambda_y}\left(\frac{y-y_j}{\lambda_y}\right)$$

…where λ_y is the bandwidth parameter associated with the kernel function for y.

The Predictor Significance Analysis and Selection Algorithm

In this research, the PMI method was used to identify the significance of all predictors. The algorithm is stopped using an AIC criterion. The algorithm for identifying the significance of model inputs (i.e., the predictors) and selecting input variables can be summarized as follows, and detailed descriptions of this algorithm can be found in relevant literature:

1) Build an initial set containing all possible predictors for the model, called z*. Define a set of selected

predictors denoted by z; this is a null vector at the beginning of the algorithm.

2) Evaluate the PMI between each plausible predictor in z* and the dependent variable y, taking into account the selected predictors in z.
3) Identify the variable with the highest PMI in the previous step.
4) Temporarily include the identified predictor into z, and compute AIC using z and y; check whether the AIC is decreased compared with the AIC of the last iteration; if the AIC is decreased, add the identified predictor into z and go to step 5; otherwise, go to step 6.
5) Repeat steps 2-4.
6) End of the algorithm.

PEK-Based Machine Learning Method

The hybrid machine learning method, PEK approximator, is an ensemble artificial neural network-based (ENN) general purpose function approximator. The PEK approximator can be used to simulate the mapping relationship of the multi-input single-output (MISO) system and was proposed by Kan [29-30]. The output is simulated according to the selected input variables, which are selected by the PMI-based separate input variable selection (IVS) scheme. The output is first estimated by the ENN, and then the output error is estimated by the KNN regression. The final simulated output is the sum of the estimated output and the estimated output error. The structure of the PEK approximator [31-40] is shown in Fig. 2, where n1, n2, …, nc denote the number of candidate input variables for Class 1, 2, …, nc, respectively; weight 1, weight 2, …, weight n denote combination weights for component networks; O(s), O(e), and E(e) denote simulated output, estimated output, and estimated output error, respectively. Detailed descriptions and the calibration method of the PEK approximator can be found in literature.

Urban Air Quality Prediction Based on Big Data, Information Theory, and Machine Learning

In this study, the daily AQI value is predicted using the previous day's meteorological and pollutant qualities, which include at most 22 predictors. Therefore, the AQI is predicted one day ahead using the PMI-based IVS method and the PEK approximator. The prediction approach is as follows:

$$AQI_i = \mathrm{F}_{\mathrm{PEK}}\left[\mathrm{F}_{\mathrm{PMIIVS}}\left[\begin{pmatrix} prec_{i-1}, prsavg_{i-1}, prsmax_{i-1}, prsmin_{i-1}, \\ rhu_{i-1}, ssd_{i-1}, temavg_{i-1}, temmax_{i-1}, \\ temmin_{i-1}, windavg_{i-1}, windmax_{i-1}, \\ winddir_{i-1}, gtemavg_{i-1}, gtemmax_{i-1}, \\ gtemmin_{i-1}, evap_{i-1}, pm25_{i-1}, pm10_{i-1}, \\ so_{2i-1}, no_{2i-1}, co_{i-1}, o_{3i-1} \end{pmatrix}\right]\right]$$

…where AQIi is the ith day's predicted air quality index, and FPEK and FPMIIVS denote the PEK-based approximator and the PMI-based IVS methods, respectively. The variables to the right of the equal sign denote AQIi predictors. The meaning of the predictors can be found in Table 1 in previous paragraphs. To make the calibration (or training) more efficient, the input-output samples of the PEK approximator should be scaled so that they always fall within a specified range. In this

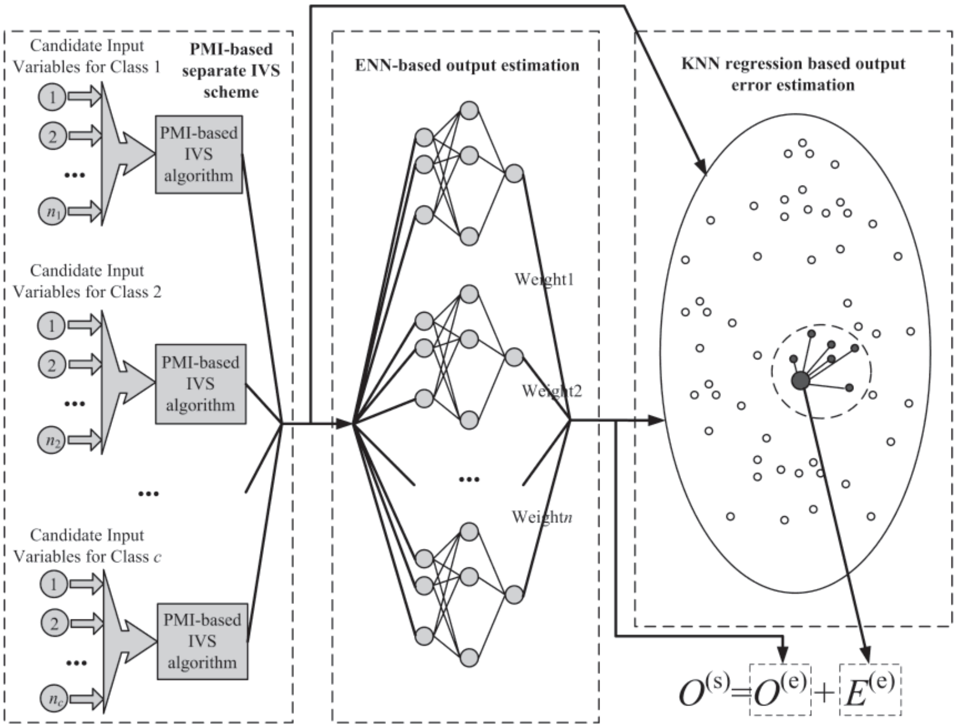
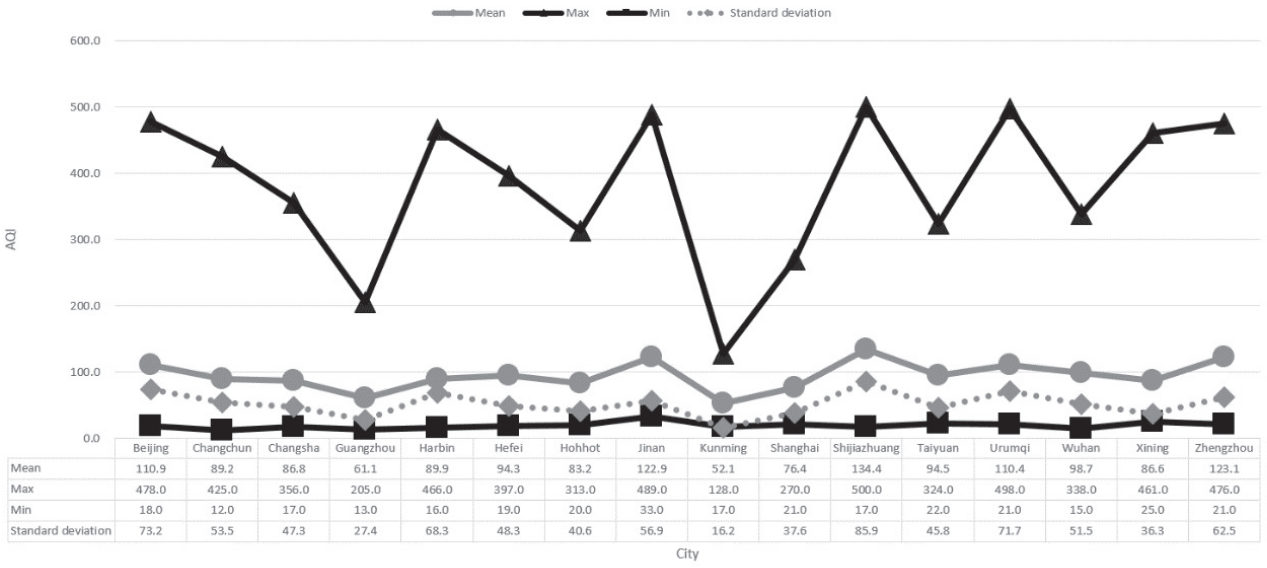

Candidate Input Variables for Class 1
PMI-based separate IVS scheme
1
2
•••
n_1
PMI-based IVS algorithm
Candidate Input Variables for Class 2
1
2
•••
n_2
PMI-based IVS algorithm
•••
Candidate Input Variables for Class c
1
2
•••
n_c
PMI-based IVS algorithm
ENN-based output estimation
Weight1
Weight2
•••
Weight$n$
KNN regression based output error estimation
$O^{(s)} = O^{(e)} + E^{(e)}$

Fig. 2. Structure of the PEK approximator.



	Beijing	Changchun	Changsha	Guangzhou	Harbin	Hefei	Hohhot	Jinan	Kunming	Shanghai	Shijiazhuang	Taiyuan	Urumqi	Wuhan	Xining	Zhengzhou
Mean	110.9	89.2	86.8	61.1	89.9	94.3	83.2	122.9	52.1	76.4	134.4	94.5	110.4	98.7	86.6	123.1
Max	478.0	425.0	356.0	205.0	466.0	397.0	313.0	489.0	128.0	270.0	500.0	324.0	498.0	338.0	461.0	476.0
Min	18.0	12.0	17.0	13.0	16.0	19.0	20.0	33.0	17.0	21.0	17.0	22.0	21.0	15.0	25.0	21.0
Standard deviation	73.2	53.5	47.3	27.4	68.3	48.3	40.6	56.9	16.2	37.6	85.9	45.8	71.7	51.5	36.3	62.5

Fig. 3. Mean, maximum, minimum, and standard deviation of AQI of 16 cities.

research, we scale the samples and make them range from -1 to 1.

Results and Discussion

Urban Air Quality Analysis

Urban Air Quality Analysis of 16 Cities

For each of the 16 studied cities, we computed the mean, maximum, minimum, and standard deviation of the daily AQI time series and drew a plot in Fig. 3.

According to the mean AQI, Shijiazhuang (mean AQI = 134.4), Zhengzhou (mean = 123.1), Jinan (mean = 122.9), Beijing (mean = 110.9), and Urumqi (mean = 110.4) are the top five cities whose mean AQI values are higher than 100. The mean AQI of Kunming (mean = 52.1) is the smallest among the 16 cities. It can be inferred that industrial cities such as Shijiazhuang, Zhengzhou, and Jinan, etc. suffer from severe air pollution issues, while the air quality of such tourist cities as Kunming and Guangzhou are better.

It can be observed that the variation trend of the standard deviation of AQI is same as that of the mean AQI. This indicates that the air quality of a city with higher mean AQI fluctuates more than a city with lower mean AQI. The stability of air quality has an important relationship with absolute air quality. When multi-year average air quality gets worse, the stability of air quality will also deteriorate. This may enlighten us into protecting absolute air quality and, as a result, air quality stability may improve.

As shown in Fig. 3, the maximum AQI of some cities such as Harbin (max AQI = 466.0), Xining (461.0) and Changchun (425) are very high even though the mean AQIs of these cities are not the worst. This indicates that mean AQI is not enough to reflect the air quality of a city and we should also consider the maximum daily AQI to pay more attention to extreme air pollution events. We can see that the difference between the maximum and minimum AQIs of Kunming and Guangzhou is much smaller compared with other cities. This confirms again that air quality and stability in Kunming and Guangzhou are much better than in the other 14 cities.

Urban Air Quality Comparison of Northern and Southern Cities

The 16 studied cities are classified into northern and southern. Northern cities include Beijing, Changchun, Harbin, Hohhot, Jinan, Shijiazhuang, Taiyuan, Urumqi, Xining, and Zhengzhou; southern cities include Changsha, Guangzhou, Hefei, Kunming, Shanghai, and Wuhan. The process graph of AQI of northern and southern cities are shown in Fig. 4. The mean, maximum, minimum, and standard deviation of daily AQI of northern and southern cities are listed in Table 2.

We can observe in Fig. 4 that the AQIs of northern cities are mostly higher than for the southern cities. Air quality of northern and southern cities usually gets worse during November to April for each year. Air pollution between 2015 and 2016 was the worst during 2014 to 2016 for northern cities. The pollution of 2014 was worst during 2014 to 2016 for southern cities. The air quality of the northern cities is getting worse considering that the AQI extreme value of each year is increasing. The air quality of the southern cities is getting slightly better considering that the AQI extreme value of each year is slightly decreasing. Table 2 notes that the mean, maximum, minimum, and standard deviation of southern cities are all smaller than those of northern cities. This indicates that air quality of northern cities is worse than that of southern cities.

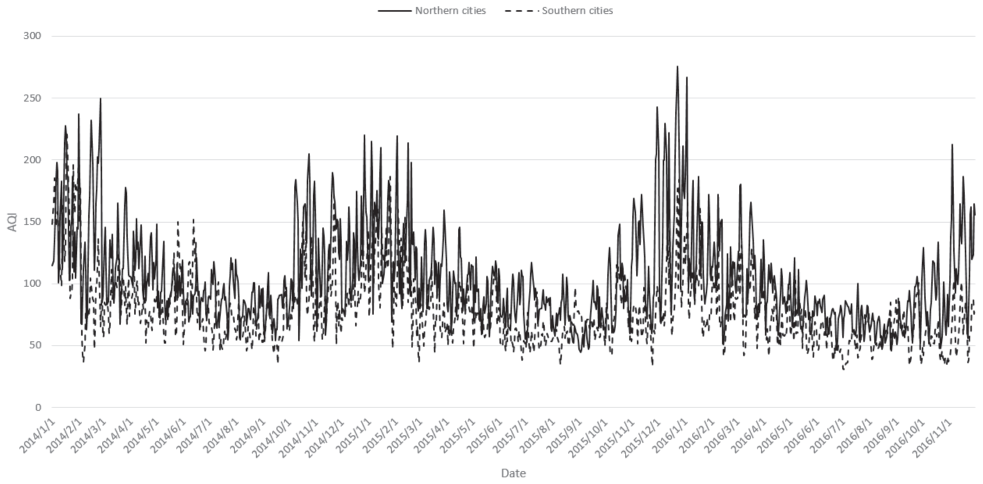


Fig. 4. AQI of northern and southern cities.

Table 2. Mean, maximum, minimum, and standard deviation of AQI of northern and southern cities.

	Northern cities	Southern cities
Mean	104.5	78.2
Maximum	275.6	220.5
Minimum	44.8	30.5
Standard deviation	39.1	28.8

Urban Air Quality Comparison of Eastern and Western Cities

The 16 study cities are classified into eastern and western. Eastern cities include Beijing, Changchun, Changsha, Guangzhou, Harbin, Hefei, Jinan, Shanghai, Shijiazhuang, Wuhan, and Zhengzhou; western cities include Hohhot, Kunming, Taiyuan, Urumqi, and Xining.

Table 3. Mean, maximum, minimum, and standard deviation of AQI of eastern and western cities.

	Eastern cities	Western cities
Mean	98.9	85.4
Maximum	256.6	201.8
Minimum	40.2	40.0
Standard deviation	37.1	26.5

The AQI process graphs for eastern and western cities are shown in Fig. 5, and daily AQI mean, maximum, minimum, and standard deviation are listed in Table 3.

Fig. 5 shows that the AQIs of eastern cities are mostly higher than that of the western cities. Air quality of eastern and western cities usually gets worse during November to April each year. Air pollution in 2014 and 2016 was the worst for eastern cities, while 2016 pollution

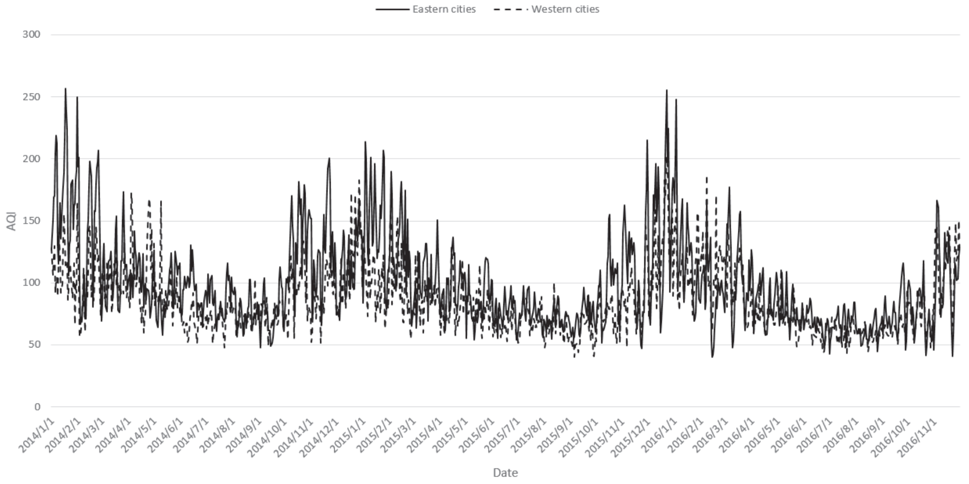


Fig. 5. AQI of eastern and western cities.

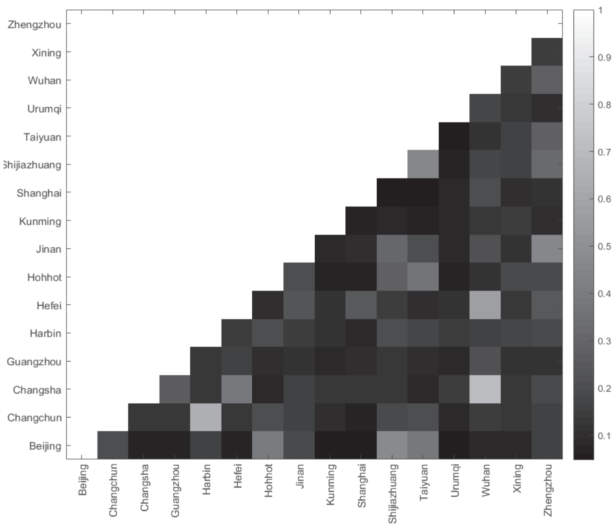


Fig. 6. MI of daily AQI time series of 16 cities.

was the worst for western cities. Air quality of eastern cities fluctuated during 2014 to 2016. The air quality of the western cities is getting slightly worse considering that the AQI extreme value of each year is slightly increasing. Table 3 shows that the mean, maximum, minimum, and standard deviation of western cities are smaller than that of the eastern cities. This indicates that air quality of eastern cities is worse compared to western cities.

MI Analysis of AQI for 16 Cities

To analyze the correlation relationship of daily AQI time series between different cities we drew a two-dimentional MI map in Fig. 6. Each grey-colored square represents an MI value between two different cities' daily AQI time series. The x-axis and y-axis list the 16 study cities, respectively. All MI values fall into the range of 0 and 1. A brighter colored square represents a higher MI value and indicates that the correlation relationship

Table 4. Selected predictors and their significance sequence (SS).

SS	Beijing	Changchun	Changsha	Guangzhou	Harbin	Hefei	Hohhot	Jinan	Kunming	Shanghai	Shijiazhuang	Taiyuan	Urumqi	Wuhan	Xining	Zhengzhou
1	**17**	**17**	**17**	**18**	**17**	**17**	**17**	**18**	**18**	**14**	**18**	**18**	**17**	**17**	**18**	**17**
2	**19**	**5**	**18**	**22**	**20**	**19**	**19**	**17**	**22**	**19**	**19**	**19**	**18**	**18**	**9**	**20**
3	**18**	**21**	**5**	**15**	**6**	**10**	**18**	**19**	**9**	**12**	**17**	**17**	**4**	**20**	**17**	**11**
4	12	19	9	11	18	18	6	8	19	22	12	11	6	19		18
5	20	18	22	17	15	9	21	15	6	9	21	5	19	15		15
6	22	20	19	20	9	21	20	11	17	17	11	9	21	22		19
7	15	2	21	19	19	22	11	22	4	10	20	20	20	3		22
8	11	9	2	9	22	6	16	20	20	21	22	12	8	7		21
9	8	15	7	3	2	4	14	7	5	17	15	15	9	21		8
10	9	11	15	14	21	3	3	4	11	7	7	8	7	9		7
11	4	3	3	2	8	7	9	21	3	13	8	7	14	2		9
12	13	13	4	7	7	16	13	9	8	2	13	13	2	14		16
13	16	4	20	21	10	14	2	16	16	20	3	14	3	8		4
14	6	22	16	13	11	12	4	3	10	3	4	10	11	13		13
15	3	7	8	8	14		8	5	21	4	2	3	12	10		5
16	2	14	10	10	12		10	6	13	5	5	21	10	6		2
17	7	8	13	16	16		15	2	14		14	4	5	5		3
18	14	10	14	4			5	13	2		9	2	15	4		10
19	5	6	6	12			7	14	15		16	6	16	16		14
20	10	16		6			22	10	7		10	22	13	12		12
21	21	12		5			12	12	12		6	16	22	11		6
22				1												

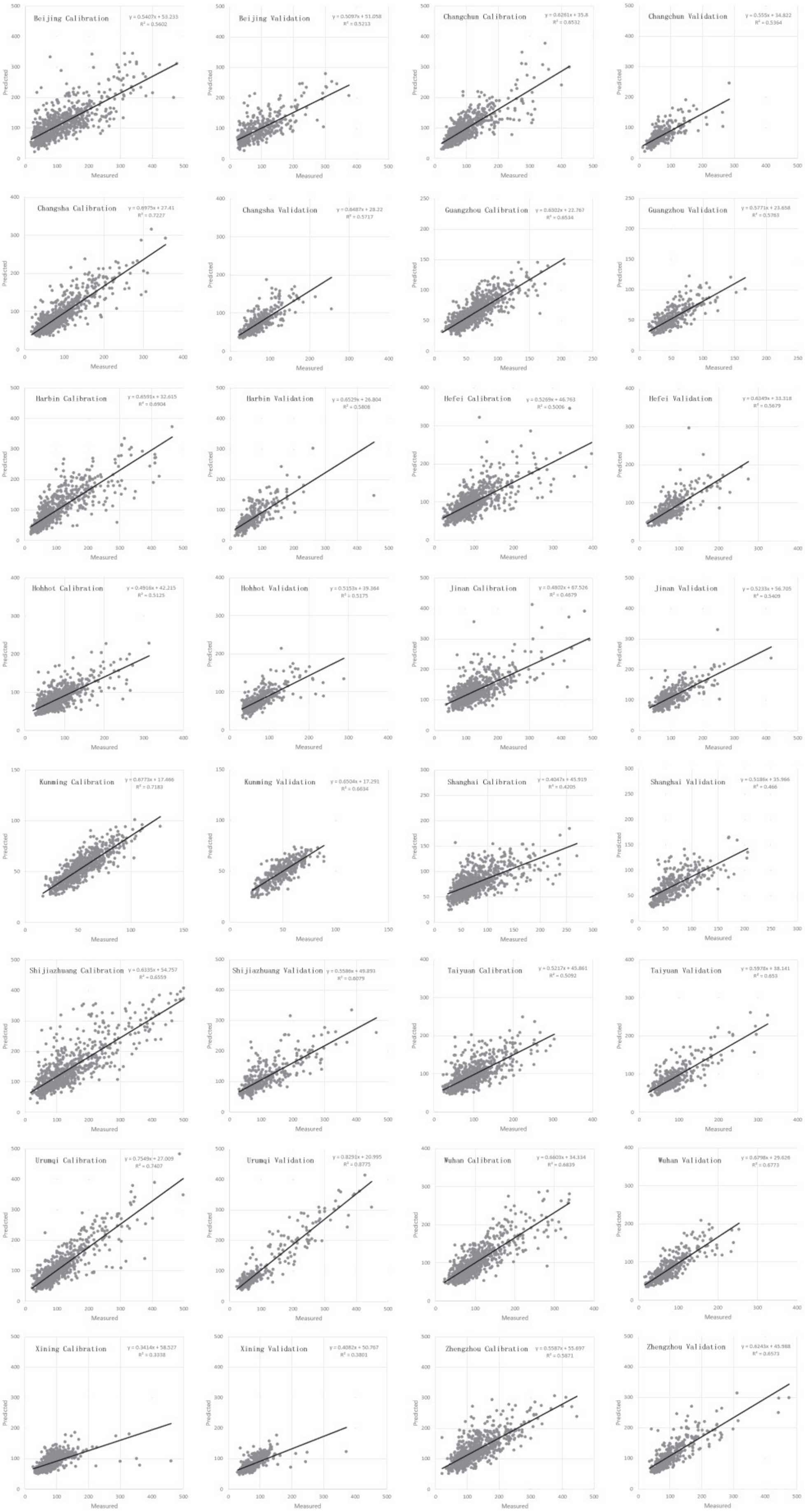
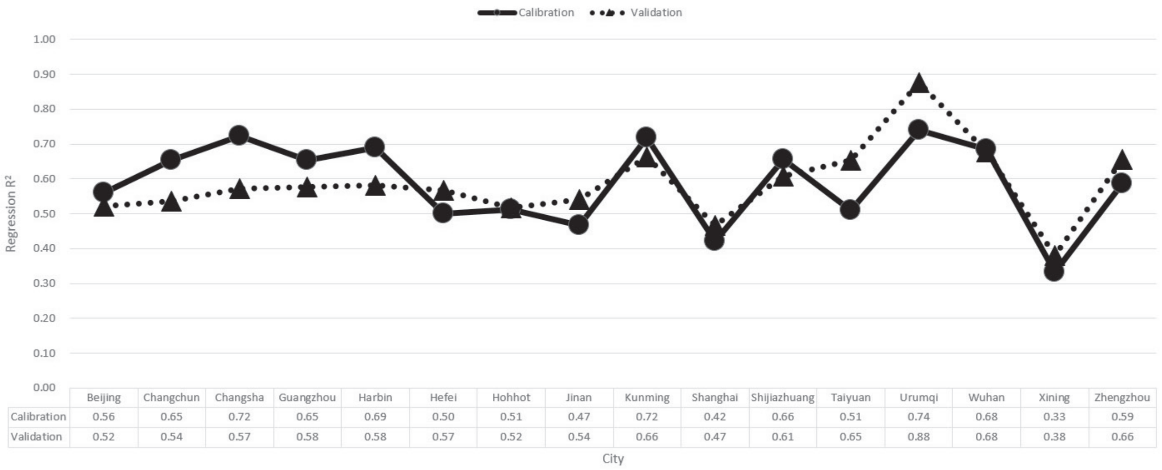


Fig. 7. Scatter plots of measured and predicted AQIs of 16 cities.



	Beijing	Changchun	Changsha	Guangzhou	Harbin	Hefei	Hohhot	Jinan	Kunming	Shanghai	Shijiazhuang	Taiyuan	Urumqi	Wuhan	Xining	Zhengzhou
Calibration	0.56	0.65	0.72	0.65	0.69	0.50	0.51	0.47	0.72	0.42	0.66	0.51	0.74	0.68	0.33	0.59
Validation	0.52	0.54	0.57	0.58	0.58	0.57	0.52	0.54	0.66	0.47	0.61	0.65	0.88	0.68	0.38	0.66

Fig. 8. Regression R^2 values of calibrated and validated AQIs of 16 cities.

between the two cities are higher. We can see that the top three higher MIs are Wuhan-Changsha, Harbin-Changchun, and Wuhan-Hefei. It also indicates that MIs of Hefei-Changsha, Hohhot-Beijing, Shijiazhuang-Beijing, Taiyuan-Beijing, Taiyuan-Hohhot, Taiyuan-Shijiazhuang, and Zhengzhou-Jinan are also high. We can see that these cities are close in geographical position. This indicates that geographical position can impact air quality. Improved air quality of a city can influence the air quality of nearby geography and vice versa.

Significance Analysis and Selection of Urban Air Quality Predictors

Table 4 lists the selected predictors and their significance sequence of the 16 study cities by using the PMI-based IVS method. Table 4 lists the indexes of the selected predictors. The relationship between index and predictor name can be found in Table 1. The top three predictors for each urban area are:

– Beijing: PM25, SO_2, and PM10.
– Changchun: PM25, rhu, and CO.
– Changsha: PM25, PM10, and rhu.
– Guangzhou: PM10, O_3, and gtemmin.
– Harbin: PM25, NO_2, and ssd.
– Hefei: PM25, SO_2, and windavg.
– Hohhot: PM25, SO_2, and PM10.
– Jinan: PM10, PM25, and SO_2.
– Kunming: PM10, O_3, and temmin.
– Shanghai: gtemmax, SO_2, and O_3.
– Shijiazhuang: PM10, SO_2, and PM25.
– Taiyuan: pm10, SO_2, and pm25.
– Urumqi: PM25, PM10, prsmin.
– Wuhan: PM25, PM10, and NO_2.
– Xining: PM10, temmin, and PM25
– Zhengzhou: PM25, NO_2, and windmax.

We can observe that the most important predictors and impact factors of AQI are pm25 (index = 17), pm10 (index = 18), and SO_2 (index = 19). This indicates that the air pollution of China is mainly caused by fine particulate matter. The future action of air protection should pay more attention to pm2.5, pm10, and SO_2 pollutants.

Prediction Error Statistics Based on Scatter Plot

This study carried out urban air quality prediction one day ahead using the PMI-based IVS and PEK-based machine learning method. The scatter plots of measured and predicted AQI of 16 study cities for calibration and validation periods are shown in Fig. 7 (which aslo demonstrates regression R2 values). We can see that the predicted and measured AQIs are close to each other for both calibration and validation periods. To investigate the calibration and validation performance more closely, we draw regression R2 values of calibrated and validated AQIs of 16 cities in Fig. 8, where we observe that seven cities' validation R2 values are worse than those of the calibration R2 values and other 9 cities' validation R2 values are better than those of the calibration R2 values. This indicates that prediction model performance is satisfactory and the overfitting phenomenon is not serious. The PMI-based IVS and PEK-based machine learning method is effective and robust for air quality prediction of 16 study cities in China.

Prediction Error Statistics Based on Air Quality Process Graph

Fig. 9 shows the process graphs of measured and predicted AQIs of 16 study cities. The simulation results are smooth and accurate for most AQI values. This indicates the successful application of the PMI-based IVS and PEK-based machine learning prediction method. It is observed that the simulation of the extreme AQI values are not satisfactory. Although the PMI-based IVS and PEK-based machine leaning prediction method can simulate the medium and small AQI values with very good precision, it cannot simulate well for large AQI values.

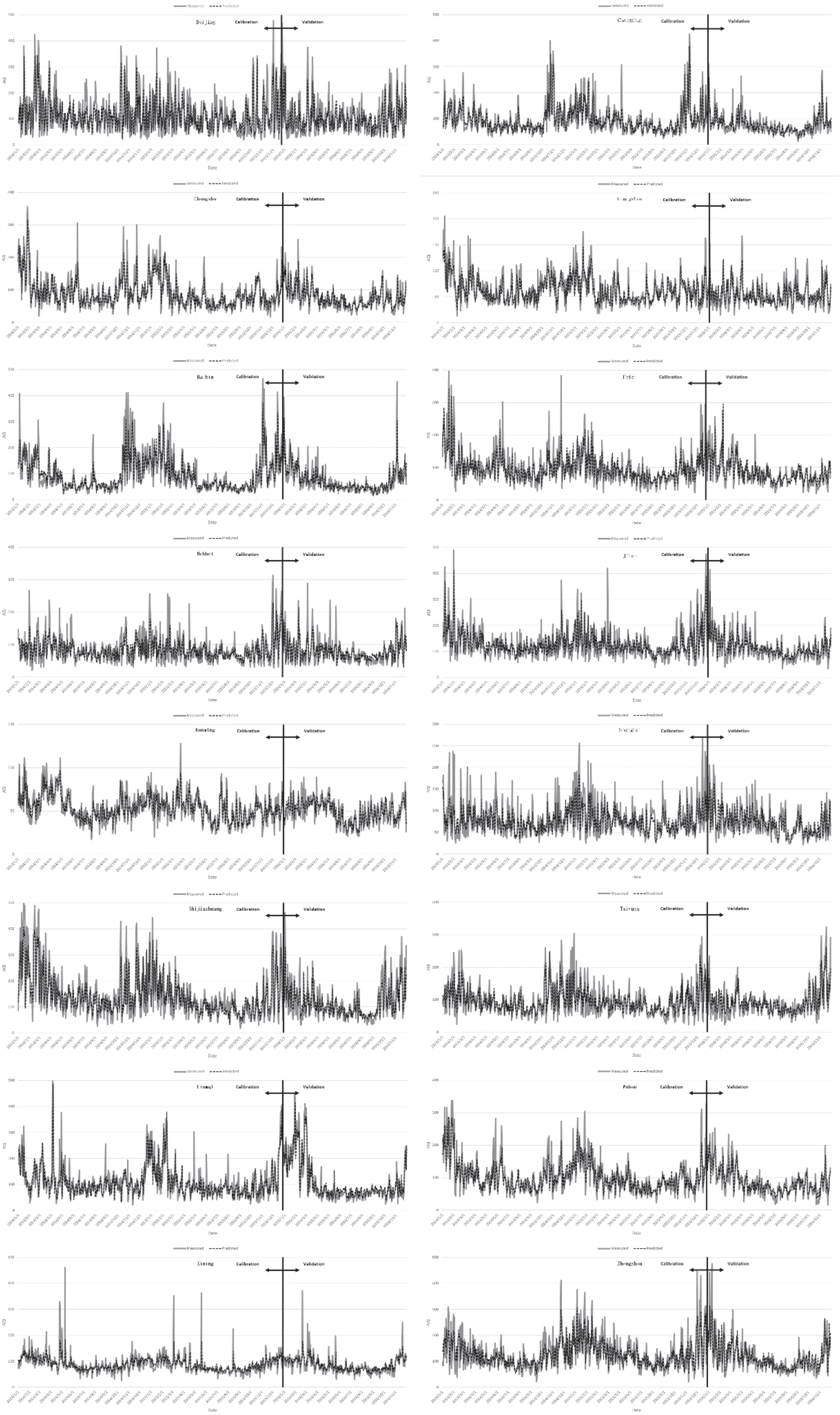


Fig. 9. Process graphs of measured and predicted AQIs of 16 cities.

Conclusions

This paper studies the air quality of 16 large Chinese cities. We carried out two angles of research, including AQI analysis and AQI prediction. The AQI predictors include at most 22 factors. The significances of these predictors are analyzed by using big data and the PMI-based IVS method. By utilizing these predictors, one-day-ahead AQI prediction is carried out through the PEK-based machine learning method. The following conclusions can be drawn:

- The stability of air quality has a strong relationship with absolute air quality. Improvement of air quality can also improve the stability of the quality. The air qualities of southern and western cities are better than those of northern and eastern cities. This indicates that air quality may have a relationship with the development of the economy. The AQI time series of cities with closer geophysical locations have a closer relationship with each other and we can predict the air quality by considering this phenomenon.
- The PMI-based IVS method is useful in the significance sorting and selection of the predictors. The sorting and selection results indicate that pm25, pm10, and SO_2 play important roles in the AQI prediction task. They are the most important predictors. We advise environmental protection organizations, and governments protect and control air quality by controlling and adjusting these factors.
- The AQI prediction result is stable, reliable, and satisfactory. The PBK-based machine learning prediction method is very useful for AQI prediction and early warning of air pollution qualities. This tool could be applied in other cities' air quality monitoring and early warning tasks to verify its effectiveness and robustness. We observed that some extreme peak AQI simulations are not as good as the medium and small AQI values. The poor accuracy of some peak value simulations in the validation period may be because some validation peak values are larger than the calibration peak value and out of the range of the training data set. We advise that the low accuracy of validation peak value simulations can be improved by adopting training data with better quality and representatives. This result indicates that the prediction capability of the PBK-based machine learning method has a close relationship with quality and is representative of the training data sample.

Acknowledgements

This research was funded by the China Postdoctoral Science Foundation (2016M600096), the Major International (Regional) Joint Research Project – China's Water and Food Security under Extreme Climate Change Impact: Risk Assessment and Resilience (G0305, 7141101024), International Project (71461010701), a study of distributed flood risk forecast model and technology based on multi-source data integration and hydro meteorological coupling system (2013CB036406), the Chinese National Flash Flood Disaster Prevention and Control Project (126301001000150068), and the National Key Research and Development Plan (2016YFC0803107). We gratefully acknowledge the support of NVIDIA Corporation with a donation of Tesla K40 GPU used for this research. Sheng Chen, Guangyuan Kan, and Ke Liang contributed equally to this work.

References

1. LEI T., PANG Z., WANG X., LI L., FU J., KAN G., ZHANG X., DING L., LI J, HUANG S., SHAO C. Drought and carbon cycling of grassland ecosystems under global change: a review. Water, **8**, 460, doi: 10.3390/w8100460, **2016**.
2. LI C., CHENG X., LI N., DU X., YU Q., KAN G. A framework for flood risk analysis and benefit assessment of flood control measures in urban areas. International Journal of Environmental Research and Public Health, **13**, 787, doi:10.3390/ijerph13080787, **2016**.
3. LI W., LU C., DING Y. A Systematic Simulating Assessment within Reach Greenhouse Gas Target by Reducing PM2.5 Concentrations in China. Polish Journal of Environmental Studies, **26** (2), 683, **2017**.
4. CHRABASZCZ M., MROZ L. Tree Bark, a Valuable Source of Information on Air Quality. Polish Journal of Environmental Studies, **26** (2), 453, **2017**.
5. FILIPIAK-FLORKIEWICZ A., TOPOLSKA K., FLORKIEWICZ A., CIESLIK E. Are Environmental Contaminants Responsible for 'Globesity'? Polish Journal of Environmental Studies, **26** (2), 467, **2017**
6. MAHMOOD S., ALI S., QAMAR M.A., ASHRAF M.R., ATIF M., IQBAL M., HUSSAIN T. Hard Water and Dyeing Properties: Effect of Pre- and Post-Mordanting on Dyeing Using Eucalyptus globulus and Curcuma longa Extracts. Polish Journal of Environmental Studies, **26** (2), 747, **2017**.
7. SULYMAN M., NAMIESNIK J., GIERAK A. Low-cost Adsorbents Derived from Agricultural By-products/Wastes for Enhancing Contaminant Uptakes from Wastewater: A Review. Polish Journal of Environmental Studies, **26** (2), 479, **2017**.
8. SEVIK H., CETIN M., GUNEY K., BELKAYALI N. The Influence of House Plants on Indoor CO_2. Polish Journal of Environmental Studies. **26** (4), DOI:10.15244/pjoes/68875, **2017**.
9. CETIN M., SEVIK H. Change of air quality in Kastamonu city in terms of particulate matter and CO_2 amount. Oxidation Communications **39** (4), 3394, **2016**
10. SEVIK H., CETIN M., BELKAYALI N., GUNEY K. Chapter 8: The Effect of Plants on Indoor Air Quality, Environmental Sustainability and Landscape Management, ST. Kliment Ohridski University Press, Eds: Recep Efe, Isa Curebal, Abdalla Gad, Brigitta Tóth, p:760, ISBN:978-954-07-4140-6, chapter page: 138, **2016**.
11. CETIN M. Changes in the amount of chlorophyll in some plants of landscape studies. Kastamonu University Journal of Forestry Faculty, **16** (1), 239, **2016**.
12. CETIN M. A Change in the Amount of CO_2 at the Center of the Examination Halls: Case Study of Turkey. Studies on Ethno-Medicine **10** (2), 146, **2016**.

13. CETIN M. Sustainability of urban coastal area management: A case study on Cide. Journal of Sustainable Forestry **35** (7), 527, **2016**.
14. SEVIK H., AHMAIDA E.A., CETIN M. Chapter 31: Change of the Air Quality in the Urban Open and Green Spaces: Kastamonu Sample. Ecology, Planning and Design. Eds: Irina Koleva, Ulku Duman Yuksel, Lahcen Benaabidate, St. Kliment Ohridski University Press, ISBN: 978-954-07-4270-0, 409, **2017**.
15. SEVIK H., CETIN M. Effects of Water Stress on Seed Germination for Select Landscape Plants. Polish Journal of Environmental Studies, **24** (2), 689, **2015**.
16. SEVIK H., CETIN M., BELKAYALI N. Effects of Forests on Amounts of CO_2: Case Study of Kastamonu and Ilgaz Mountain National Parks. Polish Journal of Environmental Studies, **24** (1), 253, **2015**.
17. CETIN M., SEVIK H. Measuring the Impact of Selected Plants on Indoor CO_2 Concentrations. Polish Journal of Environmental Studies, **25** (3), 973, **2016**.
18. CETIN M. Consideration of Permeable Pavement in Landscape Architecture. Journal of Environmental Protection and Ecology, **16** (1), 385, **2015**.
19. CETIN M. Determination of bioclimatic comfort areas in landscape planning: A case study of Cide Coastline. Turkish Journal of Agriculture-Food Science and Technology **4** (9), 800-804, **2016**
20. Environmental Protection Department. China's 2007 Environment Bullitin, **2007**.
21. LIU J., HOU K.P., WANG X.D., YANG P. Temporal-Spatial Variations of Concentrations of PM10 and PM2.5 in Ambient Air. Polish Journal of Environmental Studies, **25** (6), 2435, **2017**.
22. ZUO D., CAI S., XU Z., LI F., SUN W., YANG X., KAN G., LIU P. Spatiotemporal patterns of drought at various time scales in Shandong Province of Eastern China. Theoretical and Applied Climatology, doi: 10.1007/s00704-016-1969-5, **2016**.
23. SHARMA A. Seasonal to internannual rainfall probabilistic forecasts for improved water supply management: Part 1-A strategy for system predictor identification. Journal of Hydrology, **239**, 232-239, **2000**.
24. SHARMA A. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3-A nonparametric probabilistic forecast model. Journal of Hydrology, **239**, 249, **2000**.
25. MAY R.J., DANDY G.C., MAIER H.R., NIXON J.B. Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. Environmental Modelling & Software, **23**, 1289, **2008**.
26. MAY R.J., MAIER H.R., DANDY G.C., GAYANI FERNANDO T.M.K. Non-linear variable selection for artificial neural networks using partial mutual information. Environmental Modelling & Software, **23**, 1312, **2008**.
27. BOWDEN G.J., MAIER H.R., DANDY G.C. Input determination for neural network models in water resources applications. Part 1-background and methodology. Journal of Hydrology, **301**, 75, **2005**.
28. BOWDEN G.J., MAIER H.R., DANDY G.C. Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river. Journal of Hydrology, **301**, 93, **2005**.
29. KAN G., YAO C., LI Q., LI Z., YU Z., LIU Z., DING L., HE X., LIANG K. Improving event-based rainfall-runoff simulation using an ensemble artificial neural network based hybrid data-driven model. Stochastic Environmental Research and Risk Assessment, **29**, 1345, **2015**.
30. KAN G., LI J., ZHANG X., DING L., HE X., LIANG K., JIANG X., REN M., LI H., WANG F., ZHANG Z., HU Y. A new hybrid data-driven model for event-based rainfall-runoff simulation. Neural Computing & Applications, DOI: 10.1007/s00521-016-2200-4, **2015**.
31. DONG J., ZHENG C., KAN G., WEN J., ZHAO M., YU J. Applying the ensemble artificial neural network-based hybrid data-driven model to daily total load forecasting. Neural Computing & Applications, **26** (3), 603, **2015**.
32. KAN G., HE X., DING L., LI J., LEI T., LIANG K., HONG Y. An improved hybrid data-driven model and its application in daily rainfall-runoff simulation. IOP Conference Series: Earth and Environmental Science **46** (2016), 012029 (6th Digital Earth Summit), doi: 10.1088/1755-1315/46/1/012029, **2016**.
33. KAN G., HE X., LI J., DING L., ZHANG D., LEI T., HONG Y., LIANG K., ZUO D., BAO Z., ZHANG M. A novel hybrid data-driven model for multi-input single-output system simulation. Neural Computing & Applications, doi:10.1007/s00521-016-2534-y, **2016**.
34. KAN G., LIANG K., LI J., DING L., HE X., HU Y., AMO-BOATENG M. Accelerating the SCE-UA global optimization method based on multi-core CPU and many-core GPU. Advances in Meteorology, **8483728**, 10 pages, http://dx.doi.org/10.1155/2016/8483728, **2016**.
35. KAN G., LEI T., LIANG K., LI J., DING L., HE X., YU H., ZHANG D., ZUO D., BAO Z., MARK AMO-BOATENG, HU Y., ZHANG M. A multi-core CPU and many-core GPU based fast parallel shuffled complex evolution global optimization approach. IEEE Transactions on Parallel and Distributed Systems. DOI: 10.1109/TPDS.2016.2575822, **2016**.
36. KAN G., ZHANG M., LIANG K., WANG H., JIANG Y., LI J., DING L., HE X., HONG Y., ZUO D., BAO Z., LI C. Improving water quantity simulation & forecasting to solve the energy-water-food nexus issue by using heterogeneous computing accelerated global optimization method. Applied Energy, http://dx.doi.org/10.1016/j.apenergy.2016.08.017, **2016**.
37. KAN G., HE X., DING L., LI J., HONG Y., ZUO D., REN M., LEI T., LIANG K. Fast hydrological model calibration based on heterogeneous parallel computing accelerated shuffled complex evolution method. Engineering Optimization, **2017**.
38. KAN G., HE X., DING L., LI J., HONG Y., REN M., LEI T., LIANG K., ZUO D., HUANG P. Daily streamflow simulation based on improved machine learning method. Tecnologia y Ciencias del Agua, **VIII** (2), 51, **2017**.
39. LI Z., KAN G., YAO C., LIU Z., LI Q., YU S. An improved neural network model and its application in hydrological simulation. Journal of Hydrologic Engineering, **19** (10), 04014019-1 – 04014019-17, **2014**.
40. YIN X.X., WANG L.H., YU X.J., DU S.Y., ZHANG H.C., ZHANG Z.C. Arsenic Accumulation and Speciation of PM2.5 and elevant Health Risk Assessment in Allan, China. Polish Journal of Environmental Studies, **26** (2), 949, **2017**.