Pol. J. Environ. Stud. Vol. 28, No. 5 (2019), 3877-3885
DOI: *10.15244/pjoes/99639* ONLINE PUBLICATION DATE: 2019-05-29

Original Research

Applying Machine-Learning Methods Based on Causality Analysis to Determine Air Quality in China

Bocheng Wang*

Communication University of Zhejiang, Hangzhou, China

Received: 29 August 2018
Accepted: 7 November 2018

Abstract

A novel method was proposed for identifying air quality in China. Causality analysis-based significance tests combined with different machine-learning algorithms were carried out to achieve an automated and accurate classification. To this end, the most developed 100 cities in China were selected as study areas. We analyzed meteorological factors such as temperature, humidity, precipitation, wind speed, air pressure, sunshine duration, evaporation and grand surface temperature, and the individual industrial pollutants of NO_2, SO_2, CO and O_3 by means of time series from a large amount of air monitoring data, and focused on the causality influence of the accumulative process of each pollution ingredient on $PM_{2.5}$. In order to better clarify the formation of haze, joint regression models were established to quantify the influence degree of different factors on the cause of $PM_{2.5}$. Different classification models, including KNN, SVM, ensemble and decision tree were trained and tested to predict air quality. An accuracy of 90.2% with the ensemble (boosted trees) classifier was obtained in this study. Results of feature selection and classification both indicated that NO_2 took an important role in the contribution of $PM_{2.5}$ concentrations during 2015-2017 in China.

Keywords: air quality; classification; $PM_{2.5}$ concentrations; feature selection; causality test

Introduction

The fog and haze weather that occurs year by year in China seriously affects people's life and health. Pollution weather often sweeps through most parts of the country, especially in northern China, the Yangtze River Delta and central China. These areas are densely populated and economically developed, and their demand for natural resources is far higher than those in other parts of China. With the increasing consumption of fossil fuels from factories and private cars, SO_2 and NO_x emitted to the air not only do direct harm to human beings and plants, but also cause secondary pollution such as acid rain, haze, greenhouse effect and photochemical smog. Severe smog pollution has also occurred in many developed countries, such as the photochemical smog events in Los Angeles in 1955 and 1970, and the smog events in London in 1952, which resulted in a large number of deaths. As the culprit of air pollutants, $PM_{2.5}$ concentration increases the mortality of respiratory and cardiovascular diseases.

*e-mail: wangboc@zjicm.edu.cn

In order to monitor air quality, many countries have set up monitoring stations. The air components are recorded and analyzed in real time. However, for those areas without air monitoring stations, how to accurately predict the air quality and timely report to the public is still a difficult problem. The factors affecting air quality are very complex, including natural and human factors such as temperature, humidity, atmospheric pressure, and fossil fuel combustion. These factors together lead to the non-linear distribution of air quality in space. Therefore, air quality of a certain area cannot be judged well by monitoring data in nearby city stations. Distance is not a good means of evaluation.

Air quality forecasting methods can be classified into three kinds of categories, physical characteristics-based, statistical characteristics-based and hybrid methods. Pollutant diffusion model is one of the physical characteristics-based methods. It establishes a mathematical formula, such as Gaussian plume models, which brings meteorological data, street structure, traffic flow and then evaluates air quality. Toja et al. [1] demonstrated that the average height of the buildings showed a clear influence on the vertical profile of carbonic oxide concentration. Super et al. [2] proposed a multi-model approach based on the combination of Eulerian model and Gaussian plume model to monitor emissions of CO from the urban-industrial complex. For statistical characteristics-based methods, time series analysis and significance test are often used to evaluate air quality. Linear and non-linear regression models in statistics reflect the intrinsic property of different air components. Chang WY et al. [3] evaluated the long-term historical records for 1970-2010 in eastern China, and a significant relationship was found between $PM_{2.5}$ concentrations and SO_2. Liu et al. [4] recommended a mixed forecast strategy ARIMAX for values of $PM_{2.5}$, NO_2 and O_3 based on daily and hourly records. Chen et al. [5] suggested that individual meteorological factors can influence local $PM_{2.5}$ concentrations indirectly in interacting with other meteorological factors. They tested the convergent cross-mapping (CCM) causality relationship between different meteorological factors and $PM_{2.5}$, and it was proven that $PM_{2.5}$ concentrations in winter were notably higher than that in other seasons, meaning that temperature took significant impacts on air quality. Moreover, positive bidirectional coupling between humidity and $PM_{2.5}$ concentrations, and negative bidirectional coupling between wind, solar radiation and $PM_{2.5}$ concentrations were explained by comparing the causality direction results. On the other hand, it is not convincing to affirm that air pollution is only caused by natural factors. The influence of human activities and industrial production is also crucial. Kolluru et al. [6] discussed the contribution of different travel modes to passengers' pollutant exposure for long-distance travel on a national highway in India. The concentrations of CO, CO_2, and $PM_{2.5}$ were studied by the analysis of variance (ANOVA) method and it was concluded that avoiding national highways passing through cities can reduce up to 25% $PM_{2.5}$ and 50% CO mass exposures. Zhou et al. [7] indicated that population density, industrial structure, industrial soot (dust) emissions, and road density had a significantly positive effect on $PM_{2.5}$ concentrations, with a significantly negative influence exerted only by economic growth in China.

For hybrid models, most of the air quality evaluation methods combine the advantages of physical and statistical methods, and predict the air quality by artificial intelligence algorithms. This includes types of applications with machine learning. Cordero et al. [8] measured NO_2 concentrations using multivariate linear regression, random forests and artificial neural networks. Zhu et al. [9] achieved high classification accuracy in predicting the haze in China based on a selective ensemble algorithm. However, feature selection in machine learning is a difficult problem. There is no universal criterion to determine whether a feature is suitable or not until the prediction results come out after iterations of solving.

In this study, a novel method was proposed for identifying air quality in China. Causality analysis-based significance tests combined with different machine learning algorithms were carried out to achieve an automated and accurate classification. To this end, the most developed 100 cities in China were selected as study areas. We analyzed types of meteorological factors and the individual industrial pollutants of NO_2, SO_2, CO and O_3 by means of time series from a large number of air monitoring data, and focused on the causality influence of the accumulative process of each pollution ingredient on $PM_{2.5}$. In order to better clarify the formation of haze, joint regression models were established to quantify the influence degree of different factors on the cause of $PM_{2.5}$. Features selected by filter and wrapper methods were used to train and test the classification model. An accuracy of 90.2% with the ensemble (boosted trees) classifier was obtained in the comparison of different algorithms.

Material and Methods

Data Acquisition and Preprocessing

The first part of the data used in this paper comes from the China Meteorological Data Service Center (http://data.cma.cn), an authoritative and unified shared service platform for the China Meteorological Administration, and its meteorological data resources are open to domestic and global researchers. Various factors under record have been continuously updated from 1951 to three months lagging before current in the database of "Daily data set of surface climate in China" in version 3.0. In this paper, we analyzed natural factors composed of average station air pressure (PRS, 0.1 hPa), average air temperature (TEM, 0.1ºC), average relative humidity (RHU, 1%), average wind speed

(WIN, 0.1m/s), daily total precipitation (PRE, 0.1 mm), evaporation (EVP, 0.1 mm), sunshine duration (SSD, 0.1 h) and ground surface temperature (GST, 0.1ºC).

The analyzed daily data in this study covered the duration from January 1, 2015 to December 31, 2017. Data of each factor referred to different types of record structures, so preprocessing was carried out for normalization. First, we filtered the most developed 100 cities in China. Each city contains various monitoring stations, so normalization is crucial to the comparison among regions. The records of a city with incomplete information were excluded in this paper. As a result, there were 93 cities left, including Shanghai, Guangzhou, Shenzhen, Chengdu, Hangzhou, Tianjin, Nanjing, Chongqing, Xi'an, Qingdao, Dalian, Xiamen, Ningbo, Hefei, Zhengzhou, Ha'erbin, Kunming, Taiyuan, Nanchang, Nanning, Wenzhou, Shijiazhuang, Changchun, Quanzhou, Guiyang, Changzhou, Zhuhai, Jinhua, Yantai, Haikou, Huizhou, Wulumuqi, Xuzhou, Jiaxing, Weifang, Luoyang, Nantong, Yangzhou, Shantou, Lanzhou, Guilin, Sanya, Huhehaote, Shaoxing, Yinchuan, Zhoushan, Xining, Wuhu, Ganzhou, Jinyang, Zhangzhou, Linyi, Tangshan, Taizhou, Yichang, Huzhou, Baotou, Jining, Yancheng, Langfang, Hengyang, Qinhuangdao, Daqing, Huaian, Linjiang, Jinzhou, Lianyungang, Zhangjiakou, Zunyi, Shangrao, Longyan, Quzhou, Chifeng, Yuncheng, E'erduosi, Yueyang, Anyang, Zhuzhou, Zibo, Qizhou, Nanping, Qiqihaer, Changde, Liuzhou, Nanchong, Luzhou, Bengbu, Baoji, Yibin, Yichun, Huaihua, Yulin and Meizhou.

The other part of the data used in this study were records of air pollutants. Since 2012, the Chinese government has put more effort to monitoring air quality, and has released the air quality index (AQI) information of each city in the country hourly, which can be acquired from the Ministry of Environmental Protection of the People's Republic of China (http://datacenter.mep.gov.cn). We focused on the major pollutants, including $PM_{2.5}$ (μg/m^3), CO (mg/m^3), NO_2 (μg/m^3), O_3 (μg/m^3) and SO_2 (μg/m^3). Records of pollutants from the 93 cities above also cover the period from January 1 2015 to December 31 2017. All time series were normalized.

Finally, there was a matrix of 93 x 1096 x 13 datasets, where 93 represented the number of cities, 1096 days from 2015 to 2017, and 13 included 8 meteorological factors and 5 pollutants.

Proportion-Based Causality Test

Proportion-based causality (PBC) [10] used in this study is briefly described below. As a statistical theory, it evolves from the granger causality (GC) [11], which is widely used in the field of economics. PBC avoids the fatal drawback in GC, which only takes the error term into consideration to calculate the log value and neglects the role of independent variable terms in regression. To deal with time series, we estimate the current value of the variable through its past values. The joint auto regression model is applied to measure the regression characteristics between multi-variables. In this study, we consider the bi-variable situation and assume the lagged length to be m. This means that the current value is linearly related to the time series of the preceding m moments. The autoregressive representations and their joint representations are described respectively in the following Eq. (1) and Eq. (2):

$$\begin{cases} X_{1,t} = \sum_{j=1}^{m} a_{11,j} X_{1,t-j} + \varepsilon_{1,t} \\ X_{2,t} = \sum_{j=1}^{m} a_{22,j} X_{2,t-j} + \varepsilon_{2,t} \end{cases} \tag{1}$$

$$\begin{cases} X_{1,t} = \sum_{j=1}^{m} a_{11,j} X_{1,t-j} + \sum_{j=1}^{m} a_{12,j} X_{2,t-j} + \eta_{1,t} \\ X_{2,t} = \sum_{j=1}^{m} a_{21,j} X_{1,t-j} + \sum_{j=1}^{m} a_{22,j} X_{2,t-j} + \eta_{2,t} \end{cases} \tag{2}$$

…in which ε and η are uncorrelated noise terms, and a is the coefficient of variables in the regression model. In the autoregressive model of Eq. (1), the current value of X depends linearly on its own previous values and on the stochastic term ε. Covariance between η_1 and η_2 is defined by $\sigma_{\eta_1\eta_2} = \mathrm{cov}(\eta_1\eta_2)$. If the past values of variable X_2 make the estimation of X_1 to be more accurate, the noise term of $\sigma^2{}_{\eta_1}$ should be less than $\sigma^2{}_{\varepsilon_1}$. In this case, X_2 is said to have a causal influence on X_1. While if $\sigma^2{}_{\varepsilon_1} = \sigma^2{}_{\eta_1}$, X_2 has no causal impact on X_1. In the GC method, causality value from X_2 to X_1 is defined as Eq. (3).

$$F_{X_2 \to X_1} = \ln \frac{\sigma^2_{\varepsilon_1}}{\sigma^2_{\varepsilon_2}} \tag{3}$$

There is no causal influence from X_2 to X_1 when $F_{X_2 \to X_2} = 0$, and if $F_{X_2 \to X_2} > 0$, X_2 takes causal impact on the value of X_1. For long-term empirical research, the vector of past values in X_1 or X_2 will be too large to build the regressive model. A general approach for determining the lagged order m is AIC-Akaike Information Criterion (AIC) [12]. It was proposed by Akaike to evaluate the behavior of a statistical model. In general, the lagged order m can be defined as Eq. (4).

$$\mathrm{AIC(m)} = \mathrm{N} \log(\det(\sum \sigma_{\eta_1\eta_2})) + 2\,\mathrm{mn}^2 \tag{4}$$

…in which N is the length of sampled time series data, m represents the lagged order, and n is the number of variables used in Eq. (2). The appropriate m will minimize AIC value.

In Eq. (2), past values of $X_{1,t-j}$ and $X_{2,t-j}$ occupy a large portion among the three contributors to $X_{1,t}$ or $X_{2,t}$. Based on this, a more appropriate form of causality for multivariate interactions is defined as Eq. (5) and Eq. (6). Fig. 1 explains how to derivate Eq. (6). Firstly, $n_{X_i \to X_k}$ represented the causality from X_i to X_k. Terms of X_i and X_k were expanded within Eq. (5), and the lagged order m was highlighted in green in Fig. 1. Thus, from X_k to $X_{(k-m+1)}$ there were totally m expanded equations that were considered as taking influence on the current value of X_k. All the X_i related terms, which were delineated with a circle in Fig. 1, were summed to be the divisor in Eq. (6). The summation of X_k in Eq. (5) was the dividend in Eq. (6). It describes what proportion X_i occupies among all the contributions of X_k. More details can be referred to the definition of PBC, which is also named as a new causality method proposed by Hu [10].

$$\begin{cases} X_{1,t} = \sum_{j=1}^{m} a_{11,j} X_{1,t-j} + \cdots + \sum_{j=1}^{m} a_{1n,j} X_{n,t-j} + \eta_{1,t} \\ X_{2,t} = \sum_{j=1}^{m} a_{21,j} X_{1,t-j} + \cdots + \sum_{j=1}^{m} a_{2n,j} X_{n,t-j} + \eta_{2,t} \\ \vdots \\ X_{n,t} = \sum_{j=1}^{m} a_{n1,j} X_{1,t-j} + \cdots + \sum_{j=1}^{m} a_{nn,j} X_{n,t-j} + \eta_{n,t} \end{cases} \tag{5}$$

$$n_{X_i \to X_k} = \frac{\sum_{t=m}^{N} (\sum_{j=1}^{m} a_{ki,j} X_{i,t-j})^2}{\sum_{h=1}^{n} \sum_{t=m}^{N} (\sum_{j=1}^{m} a_{kh,j} X_{h,t-j})^2 + \sum_{t=m}^{N} \eta_{k,t}^2)} \tag{6}$$

In this paper causality relationship between different factors and $PM_{2.5}$ concentrations was tested, and based on Eq. (7), models were built to describe the influence of each component contributing to haze.

$$PM_{2.5_t} = \sum_{j=1}^{m} a_{11,j} PM_{2.5t-j} + \sum_{j=1}^{m} a_{12,j} F_{t-j} + \eta_{1,t} \tag{7}$$

…in which F represented time series of different factors related to $PM_{2.5}$.

Significance Test

In order to determine whether the causality test results were accidental, we repeated the whole process one hundred times to verify the effectiveness of time series. Besides the normal time series aligned in the duration from January 1, 2015 to December 31, 2017, sequence of variables in the other processes was resampled and shuffled into disorder. It was analogous to the bootstrap function widely used in statistical analysis.

Feature Selection

An efficient feature selection strategy is a crucial part in classification, especially in the case of a high dimensional dataset. In this paper, the dataset was 93 x 1096 x 13, where 13 was the number of candidate features including meteorological factors and pollutants, and 93 x 1096 = 105216 was the number of instances used for training and testing the classification model. The number of instances was much greater than that of the feature, so the choice of feature selection was simpler compared with other pattern recognition

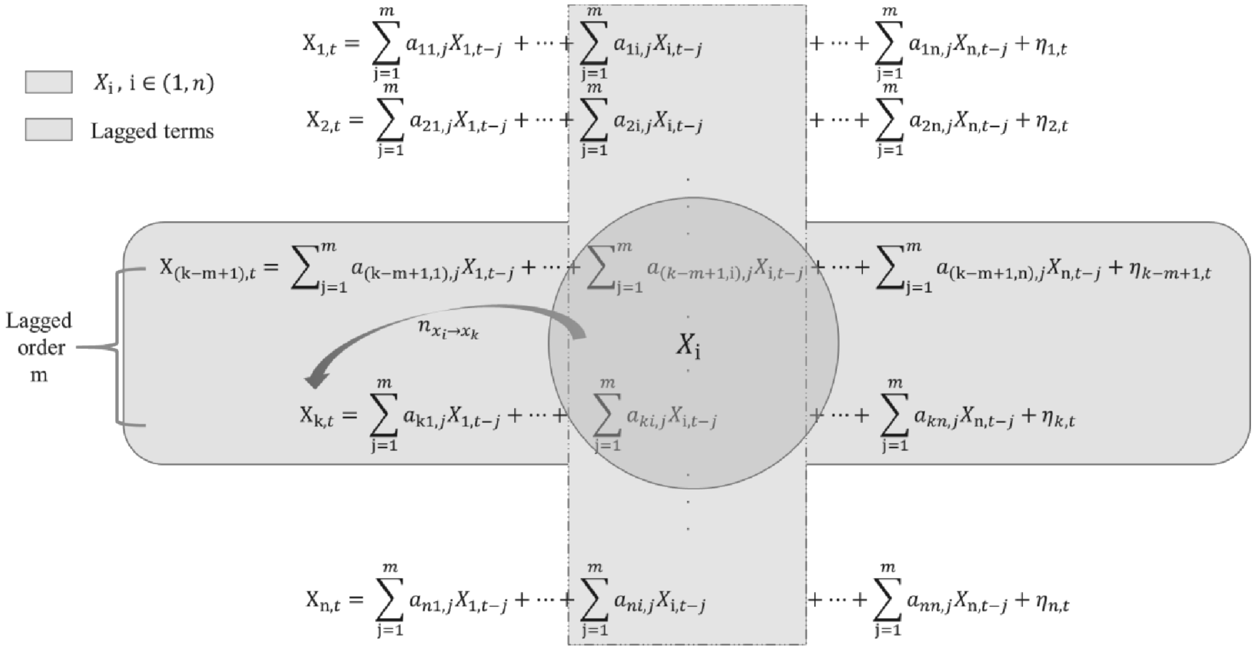
X_i, i ∈ (1, n)
Lagged terms
$X_{1,t} = \sum_{j=1}^{m} a_{11,j}X_{1,t-j} + \cdots + \sum_{j=1}^{m} a_{1i,j}X_{i,t-j} + \cdots + \sum_{j=1}^{m} a_{1n,j}X_{n,t-j} + \eta_{1,t}$
$X_{2,t} = \sum_{j=1}^{m} a_{21,j}X_{1,t-j} + \cdots + \sum_{j=1}^{m} a_{2i,j}X_{i,t-j} + \cdots + \sum_{j=1}^{m} a_{2n,j}X_{n,t-j} + \eta_{2,t}$
$X_{(k-m+1),t} = \sum_{j=1}^{m} a_{(k-m+1,1),j}X_{1,t-j} + \cdots + \sum_{j=1}^{m} a_{(k-m+1,i),j}X_{i,t-j} + \cdots + \sum_{j=1}^{m} a_{(k-m+1,n),j}X_{n,t-j} + \eta_{k-m+1,t}$
Lagged order m
$n_{x_i \to x_k}$
X_i
$X_{k,t} = \sum_{j=1}^{m} a_{k1,j}X_{1,t-j} + \cdots + \sum_{j=1}^{m} a_{ki,j}X_{i,t-j} + \cdots + \sum_{j=1}^{m} a_{kn,j}X_{n,t-j} + \eta_{k,t}$
$X_{n,t} = \sum_{j=1}^{m} a_{n1,j}X_{1,t-j} + \cdots + \sum_{j=1}^{m} a_{ni,j}X_{i,t-j} + \cdots + \sum_{j=1}^{m} a_{nn,j}X_{n,t-j} + \eta_{n,t}$

Fig. 1. Illustration of how Eq. 6 was derived.

problems. Two kinds of algorithms were adopted to rearrange features to train the model, combining filter and wrapper methods. For the filter feature selection, we performed the so-called "importance ranking" using the ReliefF [13] script implemented in MATLAB. Features were passed to that function, and the ranks and weights of predictors would be returned for the training dataset. Here we labeled a daily record in each city with one of five symbols that represented five levels of air quality, including excellent ("00001"), good ("00010"), slightly polluted ("00100"), polluted ("01000") and seriously polluted ("10000") according to their scores of AQI. Therefore, "labels" stood for response vector and features were predictors. Based on the rank and weight of each feature, the sequence of features during training were rearranged.

In the second wrapper feature selection method [14], customized objective function was designed to evaluate the performance of each feature. In the subsequent classification procedure, we tested plenty of classifiers to test which one was more suitable for air quality recognition. Thus the customized objective function should be implemented by integrating the corresponding classifiers. We started the wrapper feature selection with an empty feature sequence, which was also called Forward sequential feature selection (FSFS) [15], and each feature that helped improve prediction accuracy was kept in the queue. Feature selection is useful for reducing noise features and training time while maintaining the high performance of classification.

Classification

Supervised machine-learning methods were used to evaluate the behavior of different features and achieve high accuracy of air quality recognition. In the feature selection section, we labeled each instance with one of five levels, which was in accordance with air quality. Classifiers including KNN, tree, ensemble and SVM, with different kernel functions under test to train the model and test data. All the classification algorithms were implemented in the MATLAB and LIBSVM [16] software package (http://www.csie.ntu.edu.tw/~cjlin/libsvm). 5-cross validation was used to assess the performance of each classifier. It is not enough to conclude the fitness of a classifier based on single testing dataset or prediction results. K-cross validation method divides the whole raw training dataset into K equal parts. For each part, it is used to test the model trained by the remaining K – 1 parts. As a result, each instance in the dataset would be used for training and testing. The original SVM algorithm only realized binary classification, while LIBSVM extended SVM and achieved multi-class recognition. Five statistical indicators were introduced to evaluate performance of classifiers: accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). TP, TN, FP, and FN respectively correspond to the terms of true positive, true negative, false positive and false negative. Eq. (9) to Eq. (13) describe these statistical indicators.

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN) \quad (9)$$

$$\text{Sensitivity} = TP / (TP + FN) \quad (10)$$

$$\text{Specificity} = TN / (TN + FP) \quad (11)$$

$$PPV = TP / (TP + FP) \quad (12)$$

$$NPV = TN / (TN + FN) \quad (13)$$

Results and Discussion

The data processing and machine learning work stream described in this study are shown in Fig. 2. The raw dataset acquired was normalized and labeled. Causality and significance tests were carried out to coarsely determine which meteorological factor or pollutant impacted air quality. The filtered factors were treated as features used to train the classification model. Starting from an empty set of features, the FSFS algorithm sequentially added features that resulted in the highest objective function when combined with the features that have already been selected. Using

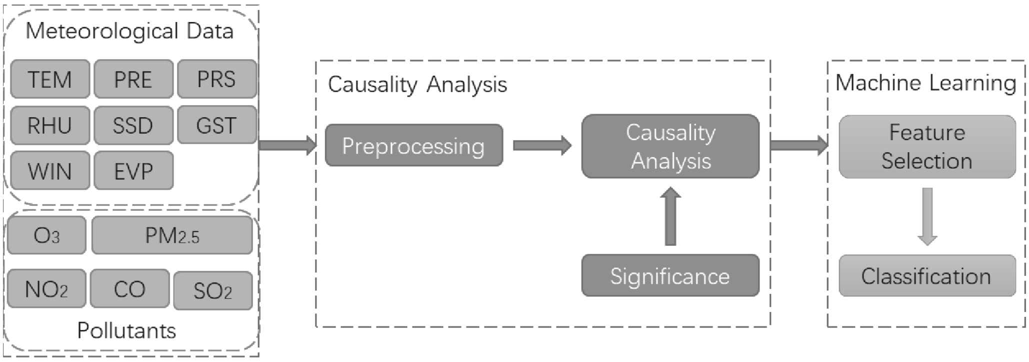

Fig. 2. Work stream proposed in this study.

Table 1. Rank of predictor and significant causality results; the number of cities increased when the influence from different factors to $PM_{2.5}$ was greater than that of inverse direction.

Rank of Predictor	Number of City	Percentage (%)	AC (10^{-2})
$NO_2 \rightarrow PM_{2.5}$	85	0.91	5.6521
$PRS \rightarrow PM_{2.5}$	86	0.92	4.3418
$SO_2 \rightarrow PM_{2.5}$	77	0.83	4.3012
$CO \rightarrow PM_{2.5}$	67	0.72	4.2179
$TEM \rightarrow PM_{2.5}$	88	0.95	4.0612
$O_3 \rightarrow PM_{2.5}$	92	0.99	3.2139
$RHU \rightarrow PM_{2.5}$	74	0.80	2.4122
$WIN \rightarrow PM_{2.5}$	55	0.59	2.0669

AC Average Causality test value

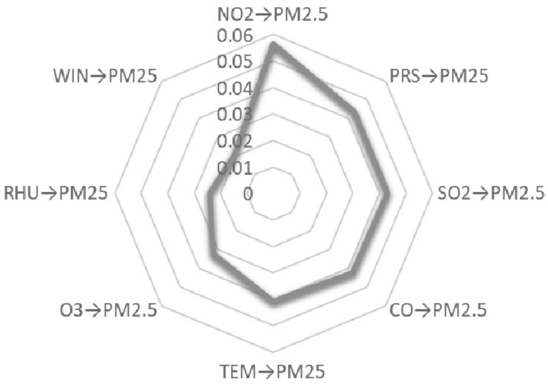

Fig. 4. Impacts from different factors on $PM_{2.5}$ calculated by causality test.

the MATLAB machine learning toolbox, six classifiers were trained and tested to achieve high accuracy of identification of air quality.

There were no significant results in testing the causality from SSD, PRE, GST and EVP to $PM_{2.5}$. Causality value of the normal time series from January 1, 2015 to December 31, 2017 was smaller than that of the other disordered sequences, which can be observed in Fig. 3. As a result, SSD, PRE, GST and EVP could not be considered as candidate features used to predict the emergence of $PM_{2.5}$. The remaining eight factors were proved to have significant influence on air quality. In Table 1, rows were sorted by the average causality test values from 93 cities. The ranking was in line with the relief function in MATLAB. NO_2 impacted $PM_{2.5}$ with the most degree. Next were PRS, SO_2, and CO. Three of four pollutants occupied the upper part of Table 1. More than half of the studied regions showed consistent results. Fig. 4 showed the impacts from different factors to $PM_{2.5}$ calculated by causality test. The distribution of studied areas was drawn in Fig. 5. Influence degree from different features was represented by a heat map. The deeper the color, the greater the degree.

After causality analysis, we investigated whether these meteorological factors combining pollutants can differentiate air quality. To this end, we started the feature selection produced by an empty set of candidate features, and a model was trained by first using the top ranking predictor in Table 1. By comparing with the following predictors, we obtained the best feature with greatest classification accuracy. Then we trained and tested the classifier by searching the second feature. The number of the features set increased in each stage

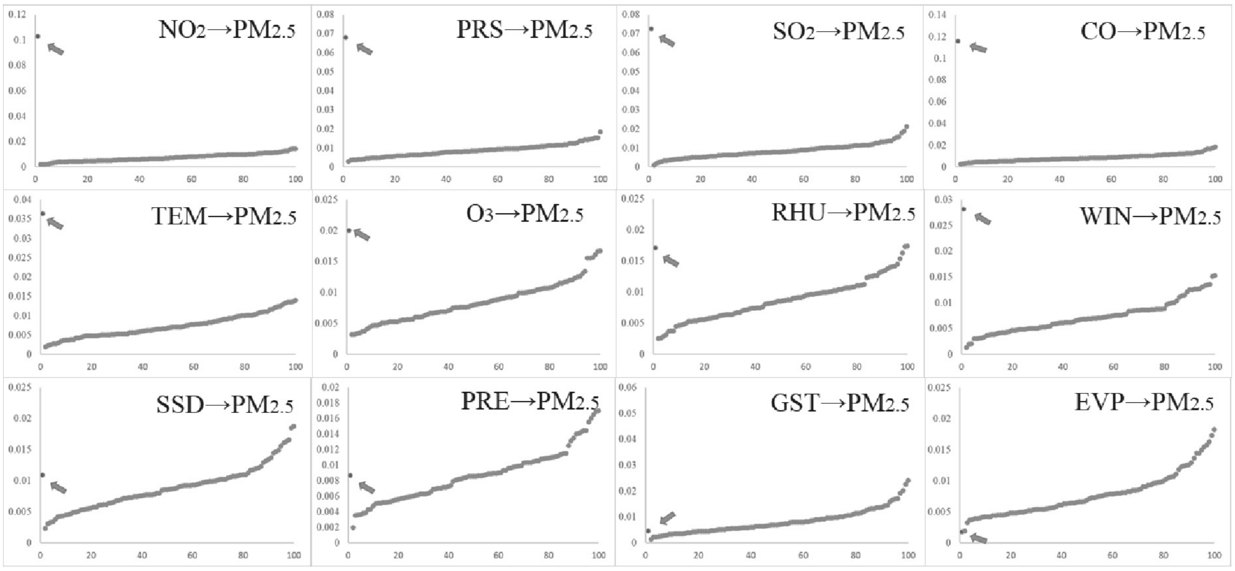

Fig. 3. Significance test. In each sub-figure, the first point represented the normal time series sequence of causality test. The following permutated 100 times of test were used to evaluate the robustness and stability of the normal sequence causality result. If the first value was greater than 95% of the 100 subsequent tests, the result was significant. In this study, causality results of SSD, PRE, GST and EVP were of no significance and were excluded from the feature selection procedure.

Table 2. Performance comparison of different classifiers in identifying air quality.

(%)	Accuracy	Sensitivity	Specificity	PPV	NPV
Tree	75.1	83.99	64.43	73.92	77.02
SVM (Linear)	82.3	91.59	69.50	80.53	85.72
SVM (Gaussian)	85.6	92.08	75.55	85.39	86.01
KNN	80.2	77.52	83.06	83.64	76.80
Ensemble (Boosted Trees)	90.2	92.50	83.36	94.30	78.89
Ensemble (Bagged Trees)	89.1	94.34	78.41	89.91	87.17

PPV Positive predictive value, *NPV* Negative predictive value, *SVM* Support vector machine, *KNN* K-nearest neighbor

of searching until iteration was terminated by certain criterion. The criterion could be the situation that all features were used to train and test the model, or that prediction accuracy was no longer a significant change by adding any features.

Different classifiers were trained and tested to recognize the labeled air quality. Accuracy, sensitivity, specificity, PPV and NPV were calculated during each procedure of feature selection. Table 2 listed the performance of each classification algorithm, and it could be observed that the highest accuracy of 90.2% appeared when the ensemble method with boosted trees was used to differentiate patterns. The other classifiers (except the decision tree algorithm) also achieved high accuracies. The decision tree algorithm has a good performance in solving linear classification, while air quality is often associated with many factors and influence by comprehensive situations, the recognition and prediction of that are usually non-linear problems. Therefore, the other non-linear classifiers, like SVM or ensemble methods, are more capable of achieving better classification.

Long-term record-based recognition of air quality with high classification accuracy was achieved in this study. 93 cities in China with the most developed economy were studied. As shown in Fig. 5, most of the selected cities were basically located in the eastern region, and the westerns are less significant. It is consistent with China's economy that the eastern regions are more developed than the western, and serious polluted air often occurs in eastern cities in China, such as Beijing, Shanghai and Guangzhou.

Causality and significance tests were used to evaluate the influence from meteorological factors and pollutants to $PM_{2.5}$. This would be useful for excluding noisy features and reducing time consumption. It is a common method in machine learning to perform filter feature selection by ranking the importance of candidate features, especially in a high-dimension training set. In this study, NO_2 was found to be the most effective predictor to train and test classifiers, and the ensemble method with boosted trees fitted the recognition best. Zhe et al. [17] and Zhai et al. [18] studied the contributions of $PM_{2.5}$ emission sources in northern China, and proposed approaches measuring the impact of pollutants on haze. While regional records of meteorological data and pollutants often cannot reflect the overall source of air pollution, we know that haze frequently occurs in the Jing-Jin-Ji region in China, and may be caused by wind direction, wind speed, precipitation and other factors in the surrounding areas. Haze pollution is a national phenomenon in China, especially with the more developed areas, which have the most serious pollution. Evidence in recent years

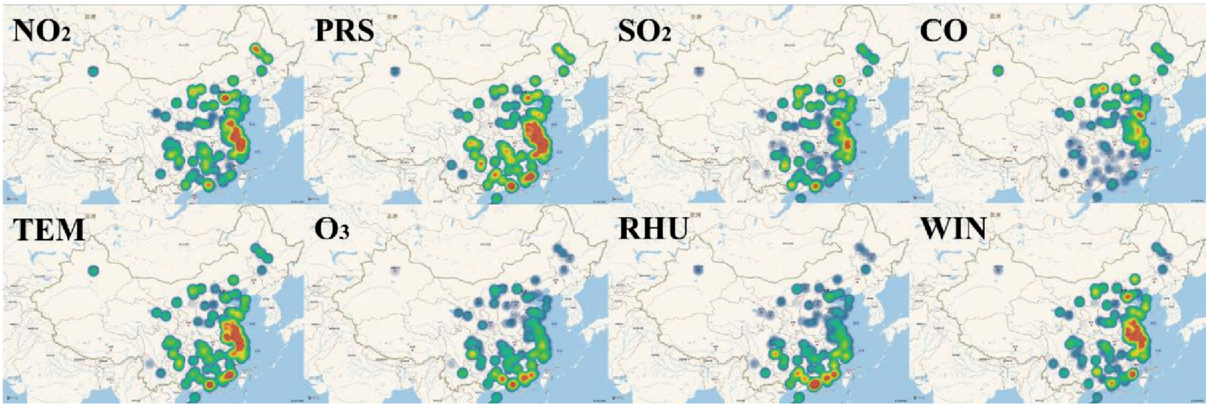

Fig. 5. Causality test results of impacts from different factors to $PM_{2.5}$; the studied areas covered the most developed 93 cities in China (most located in the southeast).

shows that haze in Japan [19] and Korea [20, 21] are aggravated by China's industrialization process. Therefore, by analyzing the meteorological data and air pollution in different latitudes and longitudes, the causes of high $PM_{2.5}$ concentrations in China can be better explained and further forecasted.

In this study, results of feature selection and classification both indicate that NO_2 takes an important role in the contribution of $PM_{2.5}$ concentrations. This is consistent with the previously reported conclusions that gas-phase and heterogeneous reactions with other gaseous contaminants and organic matter during the formation of nitrate and with the transformation of secondary aerosols enhance $PM_{2.5}$ pollution levels. Binxu Zhai et al. [22] demonstrated that NO_2 and CO concentrations measured from the city of Zhangjiakou were taken as the most important elements of pollutants for $PM_{2.5}$, with the overall classification accuracy level of 73.93%. As to meteorological factor, SSD was considered a key feature for classification. However, causality test in this study showed that there were 33 of 93 cities exhibiting evident impact from SSD on air quality. As shown in Fig. 3, the influence from SSD to $PM_{2.5}$ was of no significance. Therefore, SSD was excluded for the subsequent feature selection procedure and classification. It can be explained that high $PM_{2.5}$ concentrations affected the duration of sunshine rather than the duration of sunshine aggravated air pollution. For meteorological factors, air pressure, relative humidity and temperature were found to have the most significant impact on the prediction of air quality.

More aspects can be improved to further reach higher classification accuracy in the future. First, more detailed characteristics should be added to the candidate features set. The cause of haze is closely related to natural factors such as seasonal variations or area locations. Quantitative research of different factors such as economic development, population density, and traffic circumstance will greatly provide more clues to the formation of $PM_{2.5}$. Statistical methods can also be beneficial for revealing internal relationships between factors. Second, meteorological factors and pollutants in different regions interact with each other, forming a network structure that affects air quality. Thus, a graph-based investigation like weighted, directed network measures can be used as features to train and test the prediction model. It has been an effective feature extraction method in other machine-learning applications. Besides, the unsupervised neural network machine-learning methods are more suitable for classifying large datasets without explicit feature extraction.

Conclusions

A classification model with high accuracy was trained and tested based on different features and causality tests. The studied cities covered the most developed areas in China, and 93 of 100 with significance were selected to be analyzed. Historical daily records of pollutants and meteorological data from January 1, 2015 to December 31, 2017 were collected to fit the model. Our model evaluation demonstrates that NO_2 plays a crucial role in identifying air quality, and the ensemble method with boosted trees performs better in classifying the air quality in China with the highest accuracy of 90.2%.

Acknowledgements

This research was supported by Zhejiang Provincial Natural Science Foundation of China under grant No. LGF18A010001.

Conflict of Interest

The authors declare no conflict of interest.

References

1. TOJA-SILVA F., PREGEL-HODERLEIN C., CHEN J. On the urban geometry generalization for CFD simulation of gas dispersion from chimneys: Comparison with Gaussian plume model. Journal of Wind Engineering and Industrial Aerodynamics, **177**, 1, **2018**.
2. SUPER I., VAN DER GON H.A., VAN DER MOLEN M.K., STERK H.A., HENSEN A., PETERS W. A multi-model approach to monitor emissions of CO_2 and CO from an urban-industrial complex. Atmospheric Chemistry and Physics, **17** (21), 13297, **2017**.
3. LI Y., JIANG P., SHE Q., LIN G. Research on air pollutant concentration prediction method based on self-adaptive neuro-fuzzy weighted extreme learning machine. Environmental Pollution, **241**, 1115, **2018**.
4. LIU T., LAU A.K., SANDBRINK K., FUNG J.C. Time Series Forecasting of Air Quality Based On Regional Numerical Modeling in Hong Kong. Journal of Geophysical Research: Atmospheres, **123** (8), 4175, **2018**.
5. CHEN Z., CAI J., GAO B., XU B., DAI S., HE B., XIE X. Detecting the causality influence of individual meteorological factors on local PM 2.5 concentration in the Jing-Jin-Ji region. Scientific Reports, **7**, 40735, **2017**.
6. KOLLURU S.S., PATRA A.K., SAHU S.P. A comparison of personal exposure to air pollutants in different travel modes on national highways in India. Science of The Total Environment, **619**, 155, **2018**.
7. ZHOU C., CHEN J., WANG S. Examining the effects of socioeconomic development on fine particulate matter ($PM_{2.5}$) in China's cities using spatial regression and the geographical detector technique. Science of The Total Environment, **619**, 436, **2018**.
8. CORDERO J.M., BORGE R., NARROS A. Using statistical methods to carry out in field calibrations of low cost air quality sensors. Sensors and Actuators B: Chemical, **267**, 245, **2018**.
9. ZHU X., NI Z., CHENG M., JIN F., LI J., WECKMAN G. Selective ensemble based on extreme learning machine

and improved discrete artificial fish swarm algorithm for haze forecast. Applied Intelligence, **1**, **2017**.

10. HU S., CAO Y., ZHANG J., KONG W., YANG K., ZHANG Y., LI X. More discussions for granger causality and new causality measures. Cognitive neurodynamics, **6** (1), 33, **2012**.
11. COX JR L.A., POPKEN D.A., SUN R.X. Evaluation Analytics for Public Health: Has Reducing Air Pollution Reduced Death Rates in the United States? In Causal Analytics for Applied Risk Analysis 417. Springer, **2018**.
12. ZHAI L., LI S., ZOU B., SANG H., FANG X., XU S. An improved geographically weighted regression model for $PM_{2.5}$ concentration estimation in large areas. Atmospheric Environment, **181**, 145, **2018**.
13. ROBNIK-ŠIKONJA M., KONONENKO I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, **53** (1-2), 23, **2003**.
14. TORIJA A.J., RUIZ D.P. A general procedure to generate models for urban environmental-noise pollution using feature selection and machine learning methods. Science of The Total Environment, **505**, 680, **2015**.
15. PECLI A., CAVALCANTI M.C., GOLDSCHMIDT R. Automatic feature selection for supervised learning in link prediction applications: a comparative study. Knowledge and Information Systems, **56** (1), 85, **2018**.
16. CHANG C.C., LIN C.J. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), **2** (3), 27, **2011**.
17. ZHU W., XU X., ZHENG J., YAN P., WANG Y., CAI W. The characteristics of abnormal wintertime pollution events in the Jing-Jin-Ji region and its relationships with meteorological factors. Science of The Total Environment, **626**, 887, **2018**.
18. ZHAI S., AN X., ZHAO T., SUN Z., WANG W., HOU Q., GUO Z., WANG C. Detection of critical $PM_{2.5}$ emission sources and their contributions to a heavy haze episode in Beijing, China, using an adjoint model. Atmospheric Chemistry and Physics, **18** (9), 6241, **2018**.
19. LI P., SATO K., HASEGAWA H., HUO M., MINOURA H., INOMATA Y., TAKE N., YUBA A., FUTAMI M., TAKAHASHI T., KOTAKE Y. Chemical Characteristics and Source Apportionment of $PM_{2.5}$ and Long-Range Transport from Northeast Asia Continent to Niigata in Eastern Japan. Aerosol and Air Quality Research, **18** (4), 938, **2018**.
20. HEO J., KIM S.W., MANN KIM B., KIM J.Y. Chemical composition and source apportionment of $PM_{2.5}$ in Seoul, Korea during 2012-2013. Presented at the EGU General Assembly Conference Abstracts, **19**, 5940, **2017**.
21. LEE K., KIM Y.J., KANG C.H., KIM J.S., CHANG L.S., PARK K. Chemical characteristics of long-range-transported fine particulate matter at Gosan, Jeju Island, in the spring and fall of 2008, 2009, 2011, and 2012. Journal of the Air & Waste Management Association, **65** (4), 445, **2015**.
22. ZHAI B., CHEN J. Development of a stacked ensemble model for forecasting and analyzing daily average $PM_{2.5}$ concentrations in Beijing, China. Science of The Total Environment, **635**, 644, **2018**.