sciendo

JAISCR, 2020, Vol. 10, No. 4, pp. 255 – 270
10.2478/jaiscr-2020-0017

DATA-DRIVEN TEMPORAL-SPATIAL MODEL FOR THE PREDICTION OF AQI IN NANJING

Xuan Zhao*, Meichen Song, Anqi Liu, Yiming Wang,
Tong Wang, Jinde Cao*

School of Mathematics, Southeast University, Nanjing 210096, P. R. China

E-mail: xuanzhao11@seu.edu.cn; jdcao@seu.edu.cn

Submitted: 18th March 2020; Accepted: 5th May 2020

Abstract

Air quality data prediction in urban area is of great significance to control air pollution and protect the public health. The prediction of the air quality in the monitoring station is well studied in existing researches. However, air-quality-monitor stations are insufficient in most cities and the air quality varies from one place to another dramatically due to complex factors. A novel model is established in this paper to estimate and predict the Air Quality Index (AQI) of the areas without monitoring stations in Nanjing. The proposed model predicts AQI in a non-monitoring area both in temporal dimension and in spatial dimension respectively. The temporal dimension model is presented at first based on the enhanced k-Nearest Neighbor (KNN) algorithm to predict the AQI values among monitoring stations, the acceptability of the results achieves 92% for one-hour prediction. Meanwhile, in order to forecast the evolution of air quality in the spatial dimension, the method is utilized with the help of Back Propagation neural network (BP), which considers geographical distance. Furthermore, to improve the accuracy and adaptability of the spatial model, the similarity of topological structure is introduced. Especially, the temporal-spatial model is built and its adaptability is tested on a specific non-monitoring site, Jiulonghu Campus of Southeast University. The result demonstrates that the acceptability achieves 73.8% on average. The current paper provides strong evidence suggesting that the proposed non-parametric and data-driven approach for air quality forecasting provides promising results.

Keywords: Air quality prediction, k-Nearest Neighbor, BP neural network, Non-monitoring stations

1 Introduction

With the advancement in technology and the deployment of air quality monitoring stations, the air quality problem has gradually entered the field of vision. While paying attention to the weather forecast, people also care about the air quality today as it has always been a significant problem that concerns the future of humanity. However, the rapid increase in the number of factories and cars leads to a sharp rise in the contents of particulate matter in the air, and the problem of environmental pollution is becoming more and more serious.

Research on air pollution prediction started after the first National Conference on environmental protection in 1973. Before 1980, the weather and meteorological conditions that affected the dilution and diffusion of pollutants were mainly studied. Since the 1980s, the research and prediction of the level of urban air pollution based on SO_2 were

carried out in Beijing, Lanzhou and Shenyang, et al. In the 1990s, China made outstanding achievements in forecasting a condition of air pollution in urban areas [1]. Note that haze can directly enter and adhere to the lower respiratory tract and the lobes of the human body, which harms the health of the human body. Therefore, the early warning and prediction work of the heavily polluted weather is particularly important. It can not only enable the public to arrange production and life in advance, but also allow the relevant government departments to take urgent measures in time [2]. The sources of haze in cities are emissions from various chemical plants, vehicle exhausts and heating coal [3]. Although these factors can not be effectively controlled by individuals, defensive measures can be found through research. However, the majority of the researches on the generation and diffusion of haze are based on the principles of meteorology, and data mining research based on the intelligent algorithm is still rare [4, 8]. It has obvious fuzziness, randomness and incompleteness of information of haze, which is challenging to use the general method to get an accurate prediction. In recent years, the haze monitoring equipment has been gradually improved throughout the country [4]. Up to now, 98 national control stations have been established in Jiangsu Province, and a large number of monitoring data and meteorological data related to haze have been accumulated [42] so that intelligent algorithms will be more widely used in the field of air quality prediction.

The change in the haze is affected by many factors. Moreover, it is difficult to describe the connection between the factors affecting the formation of haze accurately. However, the formation of haze has a nonlinear relationship with the main inhalable particulate matter [5, 6, 7]. In the past, some scientists did research to evaluate the level of particulate matter in cities. Deacon et al. [8] analyzed data from UK monitoring sites in 2 years to describe the PM_{10} level at that time. Especially, the contribution of road traffic to the concentration of PM_{10} was estimated. At the same time, the influence of meteorological factors on the level of PM_{10} was studied by utilizing meteorological data in Edinburgh. Harrison and Deacon [9] used correlation analysis to build a prediction model that consider more occasions on exceedances at only one site. Siting different stations is possible to give an acceptable prediction. Gravas et al. [10] learned about the temporal and spatial variation of PM_{10} volume in Athens by air quality data from four monitoring stations. The concentration of PM_{10} was slightly higher in the cold season. And compared with the weekend, the pollution level was significantly higher on weekdays. Finally, considerable spatial heterogeneity was found in this paper. Kukkonen et al. [11] chose PM_{10} events in four European cities, utilizing PM_{10} data sets and local meteorological data. The conclusion demonstrated that that the event was mainly related to the high pressure area and inversion events. The sources of particulate matter in seven European regions were studied in the work of Quall et al. [12], the measurements of PM_{10} and $PM_{2.5}$ were used to find out the impact of PM_{10} level in Athens in Greece and Bermingham in the UK. The measured data on the regional background and local traffic were provided to reach the average annual levels of $PM_{10}$and $PM_{2.5}$. Statheropoulos et al. [13] used Principal Component Analysis(PCA) applied on five years' meteorological data to catch underlying components and to attribute physical meaning to them. The results presented relationships among the data with physical meanings. An overview of the processes influencing levels of PM is presented in the work of Viana et al. [14]. The results obtained showed that the model could assess the effects of different types of particulate matter on environmental particulate matter levels and particle size fractions.

Methods of existing researches on the haze forecasting mainly include Gray System Theory [15, 16], Fuzzy Theory [17], Artificial Neural Network(ANN) [18, 19, 20] and PCA, etc. Among these methods, ANN has incomparable advantages: the ability to approximate non-linear functions and the process of self-learning to adapt to changes. Zhou et al. [21] provided a new method of haze prediction based on multivariate diagnosis and discussed the influence of variable selection and threshold on haze prediction. Miao et al. [22] established an objective fuzzy logic haze prediction model based on the relationship between the inaccuracy of the predictive factors and the occurrence of haze, then outputted the parameters to a numerical weather prediction model based on high resolution. Wang et al. [23] extracted an Autoregressive Integrated Moving Average Model(ARIMA) model from data of California Air Resources Board to predict $PM_{2.5}$. In particular, The ARIMA model re-

flected the characteristics of the season. Cobourn et al. [24] used nonlinear regression and back-trajectory concentrations to construct an enhanced model. The 24h back-trajectory concentration is calculated by integrating back-air-trajectories and local air quality information. And the advantage of the model is that it can be adjusted easily if the availability of forecasting tools changes over time. Yu et al. [25] integrated three parts of the Bayesian maximum entropy method(BEM) to predict the monthly distribution of $PM_{2.5}$. The result showed that the predictive ability could be improved effectively by incorporating PM_{10} and total suspended particulate(TSP). Sun et al. [26] suggested a hidden Markov model(HMM) with different emission distributions. HMMs with log-normal, Gamma and generalized extreme value distributions were developed in this paper, and the result showed that the closer to the observation sequence the distribution employed in HMM was, the better the model prediction performance was. Jiang et al. [27] investigated the effect of ambient air pollution on premature birth by time series methods. At the same time, the generalized additive model (GAM) was used to make sample curves. The time series approach has also been applied to the urban air quality of Beirut in the research of Farah et al. [28]. The autocorrelation function of time series was introduced to carry out the relationship of factors other than SO_2. Residual analysis of fluctuation of data series of different levels was carried out simultaneously. Kang et al. [29] evaluated the effect of the prediction of $PM_{2.5}$ and air quality index by using real time deviation adjustment methods in the United States. Compared with the traditional kalman filter (KF), real-time deviation adjustment is more significant, which dramatically reduces the false alarm rate. Han et al. [30] used a simplified two-dimensionality heating capacity model, which relied on EnergyPlus simulation results and enlarged the application of the AQI-Heating (A-H) model. After that, bias-adjustment techniques were introduced to decrease systematic biases in the prediction of surface O_3 [31, 32, 33, 34, 35]. The Kalman filter (KF) forecaster dramatically improved the forecasting technology, and the ensemble average method obtained the best overall ozone forecast. Xu et al. [36] aimed to study the impact of trade liberalization on haze pollution in China. Therefore the research adopted the impulse response function and the decomposition method of the variance of prediction error based on the binary vector autoregression model. Finally, the experimental conclusion indicated that trade liberalization significantly alleviates haze pollution. Li er al. [37] constructed the artificial intelligence model by using the enhanced Kolmogorov-Zurbenko (KZ) filter. Results showed that this model recognized nonlinearities and interactive relationships and obtained accurate results. In order to predict $PM_{2.5}$, concentration accurately, Liu et al. [38] proposed a Self-organizing Single Hidden-Layer Long Short-Term Memory Neural Network and employed a self-organizing algorithm which automatically adjusts the number of hidden neurons in the learning stage by using information processing power (IPC). Yusof et al. [39] proposed the utilization of Artificial Neural Networks (ANN) and Multiple Linear Regressions (MLR) to identify the pollution level of PM_{10}. Moreover, sensitivity analysis (SA) as an additional feature was introduced in models to rank the most contributed parameter to PM_{10} variations. Although scientists have suggested models to estimate the relation between air quality and other meteorological factors, these models grounded on experimental assumptions and parameters may not be applicable to all urban environments.

In this paper, influencing factors in time and space dimensions are considered respectively to establish two different models to forecast air quality. The advanced KNN algorithm with the correlation coefficient is used to improve the acceptability of the forecast. Meanwhile, the existing monitoring stations data and spatial features (e.g. relative locations and geographical distance) are taken as the input of the BP neural network to infer the air quality of non-monitoring areas. In addition, this model can be applied to sites with similar topology structure. In order to test the proposed model, the air quality of the Jiulonghu Campus of Southeast University was measured at 0 am, 8 am, and 4 pm every day from October 18, 2016, to November 8, 2017, through experimental equipment. The results of the research show that the final prediction for the area without the monitoring station is accurate. The proportion of the Relative Percentage Error within 20% is about 73.8%.

The remainder of the current paper is organized as follows. Section 2 gives details on the available

methodology of the proposed models, including the k-Nearest Neighbor algorithm and the Back Propagation algorithm. In Section 3, the dataset of air quality in Nanjing used is introduced for the numerical experiments. Notably, the temporal-spatial model is established in this Section. Selection of model parameters and performance evaluation are also presented. The paper ends in Section 4 with some concluding remarks and comments on future work.

2 Methodology

Air prediction is complicated rather than pure linear research since varieties of factors influence the air quality. Considering the meteorological factors, the spatial and temporal relationship of Air Quality Index values also have an essential impact on the model. The spatial and temporal dimension models are proposed to predict air quality for non-monitoring sites and obtain a reliable forecast result. The proposed models are based on the available methodology, including the k-Nearest Neighbor algorithm and the Back Propagation algorithm. The details of the methodology will be introduced in the following.

2.1 Temporal dimension model

The temporal dimension model is built based on the KNN algorithm with the consideration of the similarities between the attributes of historical and current data. Several concepts are introduced as follows before explaining the enhanced K-NN method in detail:

Prediction duration $Y = (y_T, y_{T+1}, ..., y_{T+T_1-1})$ is a vector, in which y_T represents the AQI value at T moment and T_1 represents the length of time that needs to be predicted. In other words, the value of T_1 is set to be 8 indicates that the AQI values of the next 8 hours would be predicted when the time interval is one hour.

Front lag duration $X = (x_{T-p}, x_{T-p+1}, ..., x_{T-1})$ is a state vector where x_{T-p} represents the attribute value at T moment, and p represents the front lag length of the attribute values. That means when $p = 4$, the attribute values of the advanced 4 hours are chosen to determine similarity when the time interval is one hour.

Sample points $Q = (X, Y)$ are known values, in which X and Y are defined above and obtained from historical data referred to the training set.

Sample points set $SP = \{Q | Q = (X, Y)\}$

Target point $G = (X, Y)$, in which Y is an unknown term to be predicted and X is the corresponding front lag duration.

Target points set $TP = \{G | G = (X, Y)\}$

Measures of similarity between attribute vector X of the sample point $Q \in SP$ and it of the target point $G \in TP$ are used to recognize a similar pattern, including Euclidean distance and Pearson correlation coefficient.

Candidate points, also called the K nearest neighbors, represent the K sample points similar to the target point.

KNN is a non-parametric pattern recognition method. For a given target point, the KNN algorithm identifies the K most similar sequences from the sample points and aggregates the corresponding value of Y. To be more specific, the overall flow of the KNN algorithm which has been shown in Figure 1 is: firstly, choose an appropriate method to measure the similarity, including Euclidean distance and Pearson correlation coefficient; secondly, select the optimum number of neighbors K; thirdly, identify K neighbors that are most similar to the target point from the historical data; finally, aggregate the value of K neighbors as the final prediction. KNN needs a large amount of calculation, and there is a sample imbalance problem in this algorithm. However, the KNN algorithm has a promising result in nonlinear regression. Furthermore, it is not so sensitive to outliers.

2.2 Spatial dimension model

With the consideration of the scarcity of monitoring stations in most cities and the air quality varies from one place to another dramatically due to complex factors, a spatial model based on the BP algorithm is established to figure out the prediction of air quality at non-monitoring sites. As depicted in Figure 2, the spatial model treats $\{Q_{(i),t}\}$ as input, which include AQI values and geographical distances between the target site and surrounding monitoring stations. The $Q_{(i),t}$ is defined as fol-

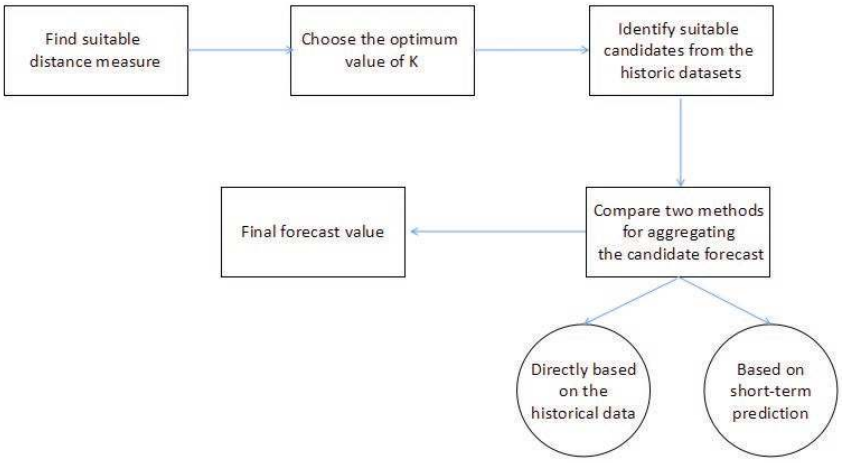

Figure 1. the KNN algorithm

lows

$$Q_{(i),t} = \begin{pmatrix} \mathbf{AQI}_{(i),t-T} \\ \vdots \\ \mathbf{AQI}_{(i),t-1} \\ d_{\widetilde{x}_{(i)},\widetilde{x}_p} \end{pmatrix}, \tag{1}$$

where $\{\widetilde{x}_i | i = 1,2,...,N\}$ denotes the set of locations with monitoring station and $\widetilde{x}_p$ is the target site to be inferred; $\{\widetilde{x}_{(i)} | i = 1,2,...,m\}$ denotes the set of the nearest m locations are known as reference stations with target site; $d_{\widetilde{x}_{(i)},\widetilde{x}_p}$ means the geographical distance between the location $\widetilde{x}_{(i)}$ and $\widetilde{x}_{(p)}$, which are calculated based on the latitude and longitude for the locations and introduced in Section 4.1. For each location, $\widetilde{x}_{(i)}$ $(i = 1,2,...,m)$, there are historical AQI values recorded as $\{AQI_{(i),t-n}\}_{n=1}^{T}$, where T means the number of history AQI data.

Input: Assuming that the number of monitoring stations is N, each monitoring station in turn is treated as the target point to be inferred, namely, $\widetilde{x}_p = \widetilde{x}_i$ $(i = 1,2,...,N)$ respectively. In order to predict the air quality of station $\tilde{x}_p$ at time t, the $Q_{(i),t}$ of the nearest m reference stations are regarded as the input of the neural network shown in Figure 2. This rotation will improve the adaptability to the geographical distance for the spatial model, owing to the expansion of the sample.

Artificial neural network: The simplicity and generality of the BP algorithm can lead to a bright prospect, so the BP algorithm with hidden layers is chosen in this paper. In each layer, a linear function is set for the input data then add the biases. After that, a nonlinear function $\varphi(x)$ is used to activate the neurons, which can get a nonlinear fitting result. Formally defined as follows

$$c^k = \varphi\left\{\sum_r w_r \varphi\left[\sum_q w'_{qr} \cdot \left(\sum_p f_p w_{pq} + b_q\right) + b'_n\right] + b_r\right\}, \tag{2}$$

where f_p is input; b_q, b'_n, and b_r are the biases in different layers to make minor adjustments; w_{pq}, w'_{qr} and w_r denote the different degrees of correlation between neurons.

Because of excellent generalization ability, nonlinear mapping capability and fault tolerance, the BP network has become the most extensively applied artificial neural network in a wide range of areas. The basic BP neural network is a feed-forward neural network with a multi-layer structure, which is trained by the error reverse propagation algorithm and incorporates two processes: the forward propagation of the signal and the back propagation of the error. The superiority of the BP is that it owns a powerful nonlinear mapping ability and network structure with high flexibility. In the light of specific conditions, the number of intermediate layers and the quantity of neurons in each layer are set in the network to change the structure. Hence the performance of BP algorithm will be improved.

In this study, establishing and applying the spatial model can obtain air quality predictions at these non-monitoring sites. In the training process, the following tasks are finished:

1. Modify the parameters involved in the BP neural network step by step, including network functions, learning rate, and other parameters.

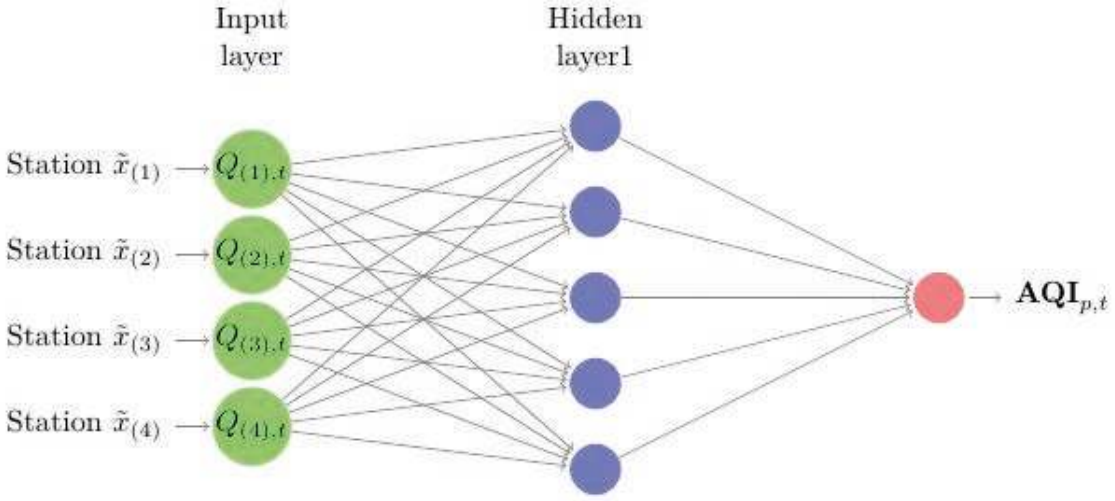

Figure 2. BP neural network

2. Determine the appropriate number of reference stations.

3. Verify the impact of topology structure on the accuracy of the prediction.

3 Experiments

The meteorological data used in this paper is from the website, http://www.pm25.in/, which contains more than 5,000 monitoring stations in China from 0:00 on October 1, 2016, to 10:00 on December 25, 2017, obtained every hour. The rate of missing data is about 24%, and the missing data are imputed using linear interpolation. To evaluate the effectiveness of the spatial model, the AQI values for the first 10 months of 2017 are utilized to construct a case database, while the remaining AQI values of the following two months are served as the test data set for prediction. In addition, the meteorological data from 9 monitoring sites in Nanjing, Jiangsu Province are screened out for the following experiments. Next, through an air-related website, UrbanAir (http://urbanair.msra.cn/En), map information of Nanjing be found as shown in Figure 3, in which the red dots represent the location of the air monitoring stations, and the blue dot means the location of the Jiulonghu Campus, Southeast University.

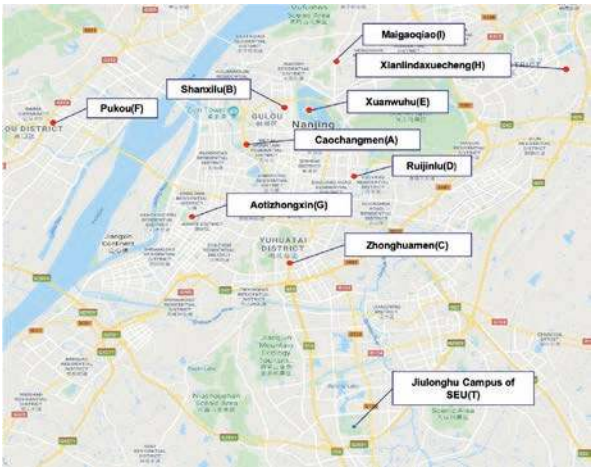

Figure 3. Map-infomation of Nanjing

In order to assess the accuracy and effectiveness of the method proposed in this paper, the parameter Relative Percentage Error(RPE) has been chosen, which is defined as follows

$$\mathrm{RPE} = \frac{|y_T - \hat{y}_T|}{y_T} \times 100\%, \tag{3}$$

where y_T and $\hat{y}_T$ mean that the real value and the estimation of AQI at moment T respectively. A prediction can be considered as an acceptable result, while the RPE is within 20%. Then, the relative accuracy rate could be defined r as follows

$$r = \frac{N_{RPE}}{N}, \tag{4}$$

where N_{RPE} means the amount of forecast data with RPE less than 20%, and N means the amount of the forecast data. The higher the relative accuracy rate, the greater the acceptability of the model performance.

3.1 Experiments of Time Model

The following experiments are performed to determine time model parameters, compare impacts by the two methods of the construction of state vectors proposed to forecast the air quality in a long period of time on prediction accuracy and verify the effect of enhanced KNN algorithm.

3.1.1 Selection of front lag duration and K nearest neighbors

In order to obtain an optimal performance of short-term prediction , a suitable front lag duration p and the number of candidates are essential. Before using cross-validation to determining the lag duration p and the number of candidates K below, a heat map is displayed in Figure 4 to judge the influence of p and K on the acceptability. It can be preliminarily observed that choosing large K and small p lead to great acceptability.

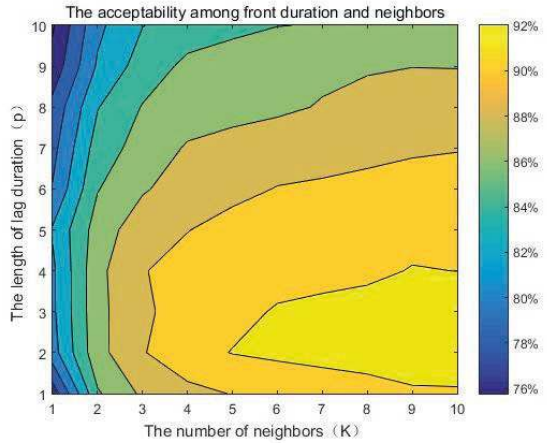

Figure 4. The heat map of K and p

In general, the oversize p makes the front lag duration contains earlier historical data that has a low impact on the current value and increases the time consumed by the algorithm. However, the small p may reduce the prediction accuracy due to the accidental errors. According to Figure 4, the acceptability presents similar trends for any given K between 1 and 10. Therefore, experiments are performed at monitoring sites in Nanjing respectively with different values of p for short-term prediction in the case of $T_1 = 1$ and $K = 1$. The RPEs between the real values and the predicted values are calculated as shown in Figure 5. It can be concluded that the short-term prediction has the best performance at $p = 4$, and the accurate rate r is more than 79%.

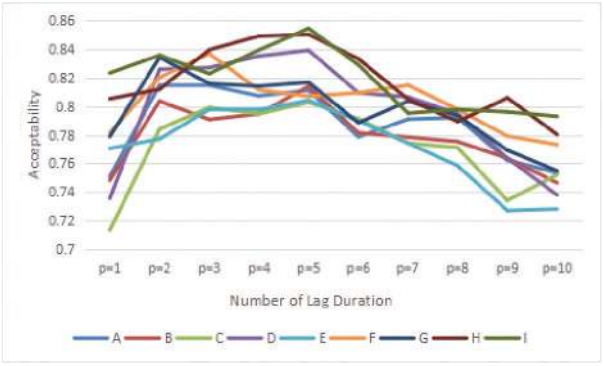

Figure 5. Acceptability of different number of lag duration(p) given K=1

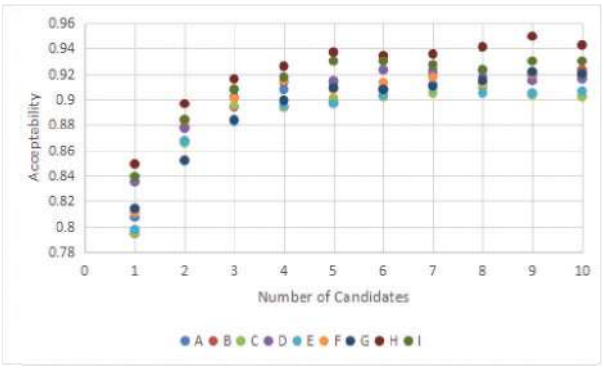

Figure 6. Acceptability of different number of candidates(K) given p=4

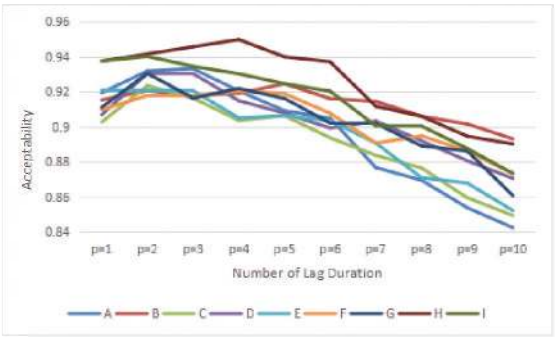

Figure 7. Acceptability of different number of lag duration(p) given K=9

For further removing the influence of random errors, it is of great importance to select a suitable K for the KNN algorithm. According to the above experiments, the parameters p and T_1 have been chosen as $p = 4$, $T_1 = 1$. Simultaneously, the time model acceptability of different values of K are compared, as shown in Figure 6. Through Figure 7, it can be concluded that the prediction acceptability of each monitoring station becomes higher as we choose a bigger K within 10 generally. Moreover,

it is found that the optimal prediction result can be obtained at $K = 9$, and the accurate rate r can reach 95% in Figure 7.

3.1.2 Enhanced KNN algorithm

The enhanced KNN algorithm is proposed for further consideration that there must be certain correlations between the sample points obtained by the KNN algorithm and the target point. Hence, the Pearson correlation coefficient is chosen between the sample points and the target point to measure their similarities in this paper.

The candidate points with a larger Pearson correlation coefficient have a higher similarity degree. Simultaneously, the contribution to the updated value is more significant. In Eq. (5), each candidate point is given a weight according to give weight to candidate points chosen by the KNN algorithm to its corresponding Pearson correlation coefficient.

$$\hat{Y}_T = \frac{\sum_{i=1}^{K} (\gamma_i \cdot Y_i)}{\sum_{i=1}^{K} (\gamma_i)}, \tag{5}$$

where γ_i represents the Pearson correlation coefficient between the ith candidate point and the target point. $\hat{Y}_T$ and Y_i mean the predicted AQI value of the target point at moment T and the AQI value of ith candidate point at the same moment respectively.

In addition, two methods based on the enhanced KNN algorithm are proposed to forecast the air quality in a longer period of time.

Method I: The AQI values of a period of time are forecasted directly based on historical data. In other words, when unknown $Y = (y_T, y_{T+1}, ...y_{T+7})$ need to be forecasted, these 8 values are predicted based on the vector $X = (x_{T-p}, x_{T-p+1}, ...x_{T-1})$ simultaneously.

Method II: Once a new predicted AQI value has been obtained, it will be added to the historical data to update and enlarge the historical data set, which can be used for the prediction of AQI value in the next hour. For example, for $T_1 = 8$, the goal is to forecast $Y = (y_T, y_{T+1}, ...y_{T+7})$. When y_T is predicted by historical data, it can be used to construct the state vector which is applied to predict the value of y_{T+1}. Then this step is repeated until the final result of Y could be obtained.

The experiments are carried out by using the enhanced KNN algorithm and two methods of the construction of the state vector. The performance of the two methods are compared by the relative accuracy rate as shown in Table 1.

Table 1. Acceptability of two methods

	Method I	Method II
1h	92.01%	92.01%
2h	77.76%	78.83%
3h	67.25%	68.48%
4h	59.38%	60.85%
5h	55.16%	55.82%
6h	51.17%	52.75%
7h	47.88%	50.18%
8h	45.53%	48.43%

Comment 3.1. According to Table 1, the acceptability of the results of the method I achieves 92% for one-hour prediction, 78% for two-hour prediction and 60% for four-hour prediction. The acceptability is obviously reduced, but is still relatively high.

Comment 3.2. The accurate rate r of method I about 2% on average is higher than it of method I as shown in Table 1, which indicates a significant improvement in the prediction of AQI values in a long period of time.

3.2 Experiments of Spatial Model

In this Section, a spatial model is established with the data of nine monitoring stations in Nanjing in 2017 to obtain predictions of air quality in areas without monitoring stations. Moreover, the acceptability will be tested with the observing data in one specific area.

3.2.1 Preliminary of the data

In this part, the coordinates of the latitude and the longitude of nine monitoring stations and the Jiulonghu Campus, Southeast University are shown in Table 2. Then, in order to reflect the change of

spatial dimension, the relative distance will be introduced. The calculation process of the distance between two sites according to their latitude and longitude is shown as following [40]. The general equation of Haversine [41] is utilized as follows

$$Haversine(\theta) = \sin\frac{\theta^2}{2}. \tag{6}$$

Let θ_1 and θ_2 be the latitude of the site A and the site B respectively, λ_1 and λ_2 be the longitude of the site A and the site B respectively, the distance between two sites A and B over the radius R of the chosen sphere represented as

$$d = 2R * \arcsin\sqrt{D_1 + D_2}, \tag{7}$$

where $D_1 = \sin((\theta_2 - \theta_1)/2)^2$ and $D_2 = \cos\theta_1 \cos\theta_2 \sin((\lambda_2 - \lambda_1)/2)^2$.

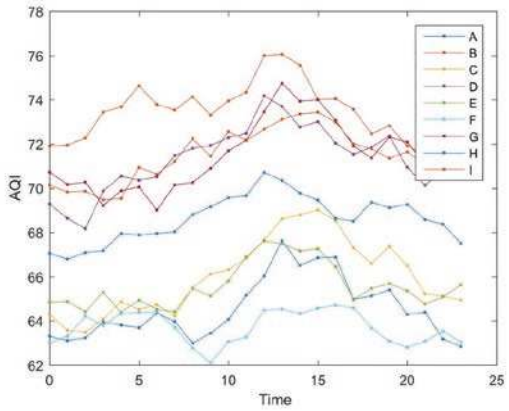

Figure 8. Air quality record from 9 stations

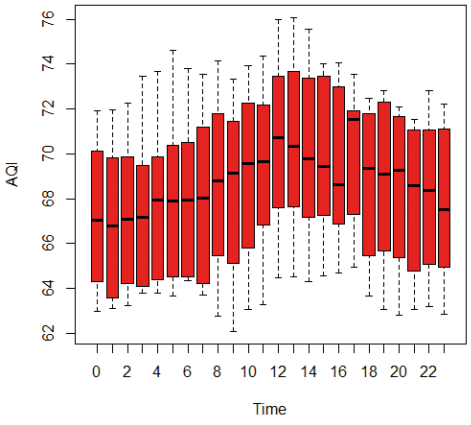

Figure 9. Deviation of AQI values among 9 stations

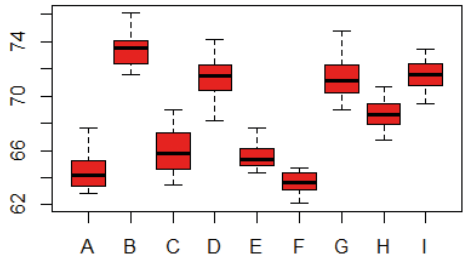

Figure 10. Graphical representation of AQI distribution among 9 stations

In order to mine the distribution feature of the data, Figure 8, Figure 9 and Figure 10 are sketched to show the air quality at nine monitoring stations in Nanjing. The trend and the distribution in 24 hours for each station of the AQI values are shown in Figure 8 and Figure 9. It is shown that the level of the AQI values for stations B, D, G, H and I are high, whereas stations C, D and G have a larger deviation of AQI values in one specific day. The box diagram Figure 10 reflects the trend of the average AQI values per hour in all stations, from which we find that AQI values from 10 am to 6 pm is higher than the rest time domain. It is obvious that human activity and weather conditions have a positive impact on the increase of AQI values.

3.2.2 Determination of the number of reference stations

The number of monitoring stations chosen around the target site P to be predicted is determined by the test, which is essential in the process of establishing the spatial model. On the one hand, if too many stations are added in the process of modeling, poor accuracy will be obtained for the areas far away from each site. On the other hand, it is difficult to describe a suitable relationship between the spatial distributions of various pollutant concentrations if fewer stations are used. An appropriate amount of reference monitoring stations will be explored in this part. Moreover, the spatial dependence between those stations will be observed.

At the same time, in order to establish a universal model, three kinds of test sets are used to investigate our algorithm and synthesize various factors to obtain the final optimal number of stations. In the following, sites A, B and F will be taken as the target site to construct dataset I, II and III, respectively.

Table 2. Latitude and longitude of each site

Label	Monitoring Station	Longitude(east longitude)	Latitude(north latitude)
A	Caochangmen	118.745954	32.062848
B	Shanxilu	118.775771	32.071374
C	Zhonghuamen	118.78829	32.019339
D	Ruijinlu	118.823293	32.038076
E	Xuanwuhu	118.800398	32.078555
F	Pukou	118.650549	32.078911
G	Aotizhongxin	118.731268	32.014363
H	Xianlindaxuecheng	118.919324	32.10507
I	Maigaoqiao	118.816053	32.108986
T	Jiulonghu campus of SEU	118.835302	31.889474

The $\{Q_{(i),t}\}$ of the nearest m reference stations are regarded as the input of the neural network to predict the AQI value of the target site from November to December. Moreover, the acceptability is calculated with the comparison of the prediction and the real data. It is noticed that the model with site B as its center and the model with site A as its center have similar geometric topologies, and both site A and site B are in a relatively central position as shown in Figure 11. Meanwhile, there are different geometric topological structures between the model with site A as the center and the model with site F as the center as shown in Figure 12, which has a similar topology structure with our final prediction target site T as shown in Figure 13. The distribution of the target site and its adjacent sites is similar to the umbrella structure as shown in Figure 12 and Figure 13.

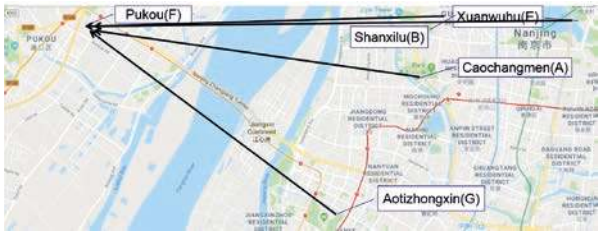

Figure 12. Map of test site F

The spatial experiments are performed on different datasets. Because of the significant diffusion effect of pollutants in the atmosphere, the AQI values of sites that are close to each other naturally have similar performance. Therefore, *m* nearest stations are selected as the reference monitoring stations, and the training matrix will be constructed using the data of those stations. Then the experiments will be repeated 10 times to reduce the random error. The mean values of the acceptability of those results are shown in Table 3.

Table 3. The acceptability of different sets for different number of stations

Stations m	Data set I	Data set II	Data set III
1	88.40%	62.87%	64.60%
2	88.09%	73.09%	74.60%
3	91.85%	73.91%	75.92%
4	91.06%	82.10%	74.21%
5	91.13%	80.36%	75.41%
6	89.82%	81.93%	76.78%
7	91.68%	82.16%	76.01%
8	91.61%	83.97%	76.45%

Considering the above experimental results and the operability in the following experiment process comprehensively, it is concluded that the number of reference monitoring stations between 3 and 5 is suitable. In order to reduce the computation time of the neural network, this paper chooses four closest stations of the target site to build the model. Observing the experimental results, when the dataset I is used as a training set, the model is able to accurately predict the data and the acceptability is as high as 90%. In addition, the dataset II is used as a training set. Although the acceptability is obviously reduced, it can still be 80%. When the training set is changed to the dataset III, the acceptability is reduced to about 75%, but is still relatively high. The experimental results of different data sets show that the choice of four closest stations of the target site is reasonable.

Comment 3.3: Those results prove that the number of reference monitoring stations chosen is suit-

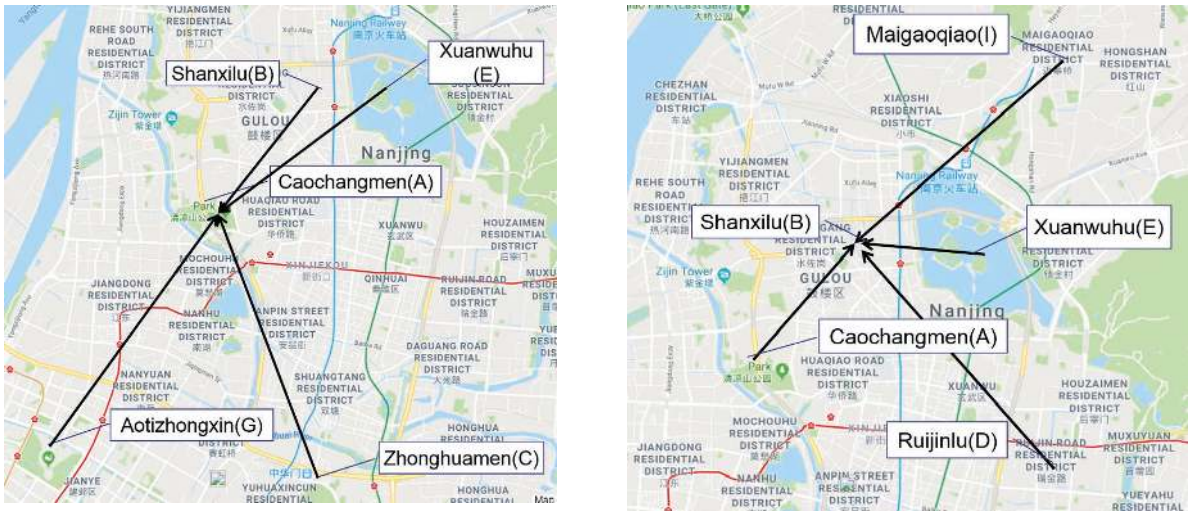

Figure 11. Map of test sites A and B

able and the spatial dimension model has excellent adaptability and stability among data sets with different topology structures. The limitation of this experiment is that the types of data sets with different geographical topologies are relatively few due to the low number of monitoring sites in Nanjing. However, when the spatial model is applied to other cases with more monitoring stations, the method proposed in this part can be transplanted to find the corresponding number of reference stations.

3.2.3 The influence of the distance parameter

The significant influence of the distance parameter on the forecast result will be verified in this part. A control group is set up without considering the distance element as the input parameter when the spatial dimension model is built. Thus, the input of the model also changes from the original $2m$-dimensional vector to the m-dimensional vector where m refers to the number of reference monitoring stations. The test uses site A as a target site P to build a model and applies it to other sites to predict. The experimental results are shown in Table 4. It can be seen from the results that the acceptability is improved when the distance factor is added. Particularly, the acceptable results of Site D and Site E are above 70%, whereas that of Site D and Site E are below 70% without distance factor.

Table 4. The effect of the distance parameter on the acceptability

	Adding distance factor	Ignoring distance factors
Site A	80.94%	79.18%
Site B	74.24%	74.46%
Site C	79.72%	75.17%
Site D	71.35%	68.00%
Site E	72.19%	65.95%
Site F	71.95%	73.29%
Site G	78.86%	73.97%
Site H	72.98%	73.14%
Site I	79.85%	78.52%

3.2.4 The effect of the topology structure

The effect of the topology structure on the acceptability will be explored in this part. In the experiment process, the scale of the training matrix is expanded to make full use of the data set. Instead of just centering on one site, we train many site-centric situations at the same time. However, the experimental results demonstrate that the model does not have a high degree of accuracy. Therefore, the important model parameters such as weights and thresholds obtained after neural network training are highly correlated with the central site. In other words, the training matrix was simply expanded by adding sites, which means ignoring the topological similarities between sites.

So it can be concluded that when a central site is selected for training, the similarities of the topology structure of the sites to be estimated with other

monitoring stations around it will be considered. Next, the influence of the topology structure on the prediction performance will be verified by experiments. Taking site F and site A as centers separately, we establish two models, and use the models to predict AQI values of site B and site H. Among them, the topology structure of site F with other monitoring stations is similar with it of site H, which is the umbrella structure, and the distance between monitoring stations and the target site is relatively long. There is a similar relationship between site A and site B, and the monitoring stations around them are evenly distributed around them.

Training model	Test Station	
	Site B	Site H
Site A	89.29%	82.14%
Site F	79.27%	96.75%

Table 5. The effect of the topology structure on the acceptability

According to the results in Table 5, the model obtained through site F training data is used to predict AQI values of site H significantly better than site B. At the same time, the model obtained from site A is used to predict AQI values of site B, whose result is better than the prediction of site H. Therefore, it can be preliminarily concluded that the effect of the topology structure on the acceptability is essential. Therefore, we choose site F with a similar topology structure to site T as a training set, then predict the AQI value of site T through the spatial model constructed by site F.

Comment 3.4. The limitation of this experiment is that the types of data sets with different geographical topologies are relatively few due to the low number of monitoring sites in Nanjing. Nonetheless, when the spatial dimension model is utilized to other circumstances with various types of topology structures, the method proposed in this part can be used to find a promising result.

3.3 Experiments of the Temporal-Spatial Model

The ultimate goal of our model is to predict the concentration of atmospheric pollutants in a certain period in the future, which is reflected by AQI values, including areas with a monitoring station and areas without monitoring stations. The areas with the monitoring station can be predicted by the time model directly, and the areas without monitoring stations can be predicted through prediction data of monitoring stations and the spatial model. This can be achieved by synthesizing the two models mentioned above. The adaptability of the temporal-spatial model is tested on a specific non-monitoring site, Jiulonghu Campus of Southeast University, as shown in Figure 13.

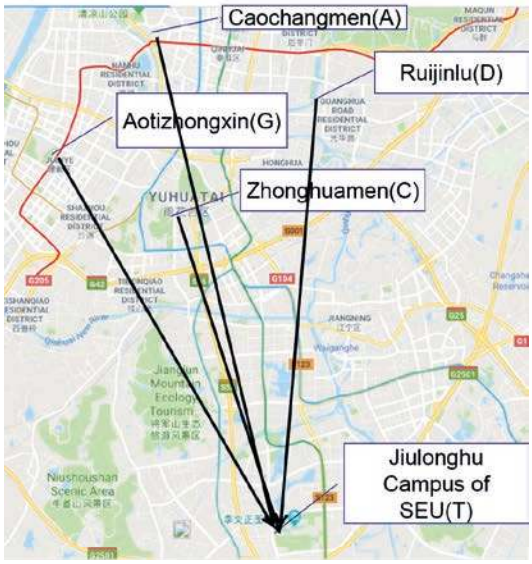

Figure 13. Map of test site T

We measure the AQI values of site T at 0 am, 8 am, and 4 pm every day from October 18 to 31 through experimental equipment as the real value of the comparison. A spatial model is established taking site F as the target site. Then the parameters of the model are provided by the previous article. The prediction data of each monitoring station from October 18 to 31 can be obtained by the time model and are used as input of the spatial model. Finally, the temporal-spatial model prediction results are shown in Figure 14. It can be found that the final prediction for the non-detected area site T is accurate, and the acceptability of the result is about 73.81%.

Comment 3.5. The temporal-spatial model tested in the non-monitoring area T shows excellent performance of the model with high acceptability and accurate AQI variation trend. In order to reduce the influence of the measuring error, more tests could be carried out in different types of target areas without monitoring stations, and more rigorous verification would be obtained.

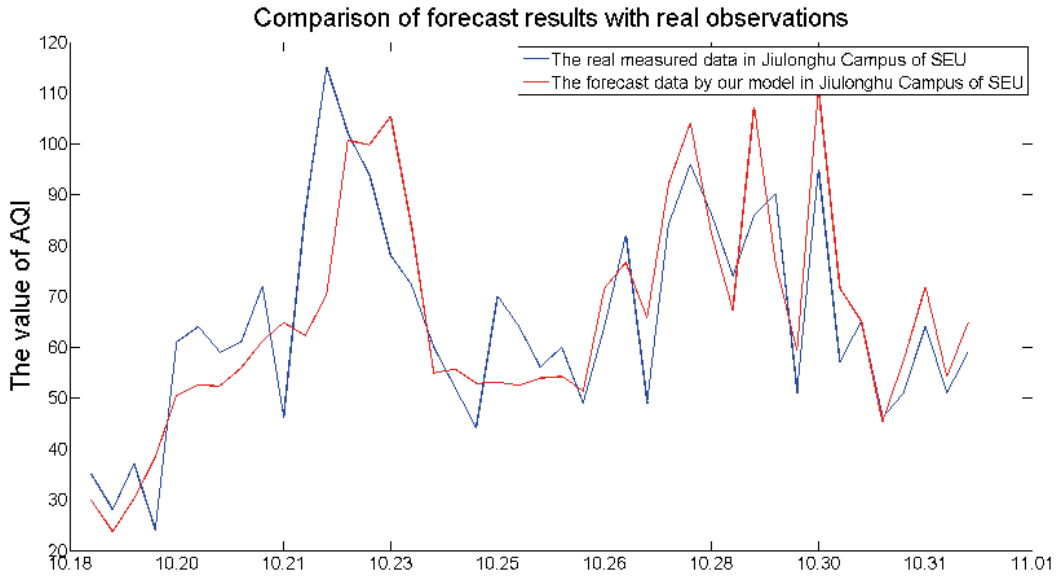

Figure 14. Comparison of forecast results with real observations

4 Concluding Remarks

The capability to timely and spatially predict the variation of air quality in the future is imperative for proactive government strategies and provision of regular outdoor activities time for citizens. In the current paper, the air quality time model based on the enhanced K-nearest neighbors is established. To match similar patterns precisely, the Pearson correlation coefficient is used in the enhanced KNN. Moreover, the spatial model is constructed by the Back-Propagation neural network using AQI values and relative distance as the input to estimate the AQI values among areas without monitoring stations. Finally, the temporal-spatial model is built, and the adaptability is tested on a specific non-monitoring site, Jiulonghu Campus of Southeast University. This research provides evidence that indicates the following main findings: Using the Pearson correlation coefficient and the state vector, which is updated with prediction, in the advanced KNN algorithm, can improve the acceptability of forecast. The prediction for non-monitoring areas can achieve a promising result by the spatial-temporal model using the data in existing stations.

The main limitation of the study is that the acceptability of forecasting is declining as time goes by. Moreover, some dynamic factors, like people's activities, traffic flow, can be considered to promote the practicability of the model. Future works will concentrate on the prediction of AQI in a longer period by incorporating exogenous factors that affect air quality into the forecast model.

Acknowledgment

Xuan Zhao (xuanzhao11@seu.edu.cn) is supported by the National Natural Science Foundation of China (Grant Nos.11701081,11861060), the Fundamental Research Funds for the Central Universities.

References

[1] W. N Deng, PM_{10} pollution forecast based on BP Neural Network and MATLAB implementation in Xi'an City, Xi'an university of science and technology, 2008.

[2] W. Sun, H. Zhang, P. Ahmet et al., Prediction of 24-hour-average $PM_{2.5}$ concentrations using a hidden Markov model with different emission distributions in northern California, Science of the Total Environment, 443, 2013, 93-103.

[3] J. Kukkonen, M. Pohjola, R. S. Sokhi et al., Analysis and evaluation of selected local-scale PM_{10} air pollution episodes in four European cities: Helsinki, London, Milan and Oslo, Atmospheric environment, 39, 2004, 2759-2773.

[4] Y. Zheng, F. Liu and H. P. Hsieh, U-Air: when urban air quality inference meets big data, ACM SIGKDD international conference on knowledge discovery and data mining, 2013, 1436-1444.

[5] Z. Cheng, S. Wang, J. Jiang et al., Long-term trend of haze pollution and impact of particulate matter in the Yangtze River Delta, China, Environmental pollution, 182, 2013, 101-110.

[6] N. H. Hanafi, M. H. Hassim, and Z. Z. Noor, Overview of Health Impacts due to Haze Pollution in Johor, Malaysia, Journal of Engineering and Technological Sciences, 50, 2018, 818-831.

[7] Y. Bian, Z. Huang, J. Ou et al., Evolution of anthropogenic air pollutant emissions in Guangdong Province, China, from 2006 to 2015, Atmospheric Chemistry and Physics, 19, 2019, 11701-11719.

[8] A. R. Deacon, R. G. Derwent, R. M. Harrison et al., Analysis and interpretation of measurements of suspended particulate matter at urban background sites in the United Kingdom, Science of the total environment, 203, 1997, 17–36.

[9] R. M. Harrison and A. R. Deacon, Spatial correlation of automatic air quality monitoring at urban background sites: implications for network design, Environmental technology, 19, 1998, 121-132.

[10] G. Grivas, A. Chaloulakou, C. Samara et al., Spatial and temporal variation of PM_{10} mass concentrations within the greater area of Athens, Greece, Water air and soil pollution, 158, 2004, 357-371.

[11] J. Kukkonen, M. Pohjola, R. S. Sokhi et al., Analysis and evaluation of selected local-scale PM_{10} air pollution episodes in four European cities: Helsinki, London, Milan and Oslo, Atmospheric environment, 39, 2004, 2759-2773.

[12] X. Querol, A. Alastuey, C. R. Ruiz et al., Speciation and origin of PM_{10} and $PM_{2.5}$ in selected European cities, Atmospheric environment, 38, 2004, 6547-6555.

[13] M. Statheropoulos, N. Vassiliadis and A. Pappa, Principal component and canonical correlation analysis for examining air pollution and meteorological data, Atmospheric environment, 32, 1998, 1087-1095.

[14] M. Viana, X. Querol, A. Alastuey et al., PM levels in the Basque Country (Northern Spain): analysis of a 5-year data record and interpretation of seasonal variations, Atmospheric environment, 37, 2003, 2879-2891.

[15] S. W. Jia, X. L. Liu and G. Yan, The dynamic analysis of a vehicle pollutant emission reduction management model under economic means, Clean Technologies and Environmental Policy, 21, 2019, 243-256.

[16] X. N. Yue, Z. J. Meng and Z. H. Yuan, Multiple regression analysis on causes of urban fog-haze in China-based on data mining, The 27th Chinese Control and Decision Conference (2015 CCDC), 2015, 4408-4413.

[17] W. Q. Huang, H. B. Fan and Y. Qian, Modeling and efficient quantified risk assessment of haze causation system in China related to vehicle emissions with uncertainty consideration, Science of the total environment, 668, 2019, 74-83.

[18] Y. Luo, T. Mengfan, Y. Kun et al., Research on $PM_{2.5}$ estimation and prediction method and changing characteristics analysis under long temporal and large spatial scale-A case study in China typical regions, Science of the Total Environment, 696, 2019, 133983.

[19] W. Zhuang, J. Fan, Y. Gao et al., Study on prediction model of space-time distribution of air pollutants based on artificial neural network, Environmental Engineering & Management Journal (EEMJ), 18, 2019.

[20] X. T. Li and X. D. Zhang, Predicting ground-level $PM_{2.5}$ concentrations in the Beijing-Tianjin-Hebei region: A hybrid remote sensing and machine learning approach, Environmental pollution, 249, 2019, 735-749.

[21] B. B. Zhou, J. Du, I. Gultepe et al., Forecast of low visibility and fog from NCEP: current status and efforts, Pure and applied geophysics, 169, 2012, 895-909.

[22] Y. Miao, R. Potts, X. Huang et al., A fuzzy logic fog forecasting model for Perth Airport, Pure and applied geophysics, 169, 2012, 1107-1119.

[23] W. Q. Wang and Y. Guo, Air pollution $PM_{2.5}$ data analysis in Los Angeles long beach with seasonal ARIMA model, 2009 international conference on energy and environment technology, 3, 2009, 7-10.

[24] W. G. Cobourn, An enhanced $PM_{2.5}$ air quality forecast model based on nonlinear regression and back-trajectory concentrations, Atmospheric Environment, 44, 2010, 3015-3023.

[25] H. L. Yu and C. H. Wang, Retrospective prediction of intraurban spatiotemporal distribution of $PM_{2.5}$ in Taipei, Atmospheric Environment, 44, 2010, 3053-3065.

[26] W. Sun, H. Zhang, P. Ahmet et al., Prediction of 24-hour-average $PM_{2.5}$ concentrations using a hidden Markov model with different emission distributions in northern California, Science of the Total Environment, 443, 2013, 93-103.

[27] L. L. Jiang, Y. H. Zhang, G. X. Song et al., A time series analysis of outdoor air pollution and preterm birth in Shanghai, China, Biomedical and Environmental Sciences, 20, 2007, 426.

[28] A. Charbel, C. Carine, B. Agnes et al., SO_2 in Beirut: air quality implication and effects of local emissions and long-range transport, Air Quality Atmosphere and Health, 1, 2008, 167-178.

[29] D. Kang, R. Mathur and S. T. Rao, Real-time bias-adjusted O_3 and $PM_{2.5}$ air quality index forecasts and their performance evaluations over the continental United States, Atmospheric Environment, 44, 2010, 2203-2212.

[30] H. Li, S. You, H. Zhang et al., Modelling of AQI related to building space heating energy demand based on big data analytics, Applied Energy, 203, 2017, 57-71.

[31] L. D. Monache, T. Nipen, X. X. Deng et al., Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction, Journal of geophysical research-atmospheres, 111, 2006, D05308.

[32] S. McKeen, J. Wilczak, G. Grell et al., Assessment of an ensemble of seven real-time ozone forecasts over eastern north America during the summer of 2004, Journal of Geophysical Research: Atmospheres, 110, 2005, D21307.

[33] L. Delle Monache, J. Wilczak, S. Mckeen et al., A Kalman-filter bias correction method applied to deterministic, ensemble averaged and probabilistic forecasts of surface ozone, Tellus Series b-chemical and physical meteorology, 60, 2008, 238-249.

[34] J. Wilczak, S. McKeen, I. Djalalova et al., Bias-corrected ensemble and probabilistic forecasts of surface ozone over eastern North America during the summer of 2004, Journal of geophysical research-atmospheres, 111, 2006, D23S28.

[35] D. W. Kang, R. Mathur, S. T. Rao et al., Bias adjustment techniques for improving ozone air quality forecasts, Journal of geophysical research-atmospheres, 113, 2008, D23308.

[36] Y. Z. Xu, X. M. Fan, Z. Q. Zhang et al., Trade liberalization and haze pollution: Evidence from china, Ecological Indicators, 109, 2020, 105825.

[37] P. Z. Li, Y. Wang and Q. L. Dong, The analysis and application of a new hybrid pollutants forecasting model using modified Kolmogorov–Zurbenko filter, Science of The Total Environment, 583, 2017, 228-240.

[38] X. Liu, Q. Liu, Y. Zou et al., A LSTM-Based Approach to Haze Prediction Using a Self-organizing Single Hidden Layer Scheme, International Conference on Security with Intelligent Computing and Big-data Services, 2018, 701-706.

[39] K. M. K. K. Yusof, A. Azid, M. S. A. Sani et al., The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models over particulate matter (PM_{10}) variability during haze and non-haze episodes: A decade case study, Malaysian Journal of Fundamental and Applied Sciences, 15, 2019, 164-172.

[40] J. Z. Levin, A rational parametric approach to latitude, longitude and altitude, Navigation, 35, 1988, 361–370.

[41] H. Mahmoud and N. Akkari, Shortest path calculation: a comparative study for location-based recommender system, 2016 world symposium on computer applications & research (WSCAR), 2016, 1-5.

[42] https://aqicn.org/city/nanjing/cn/

Xuan Zhao received the B.S. degree from Nanjing Xiaozhuang University, Nanjing, China, the M.S. degree from Southeast University, Nanjing, China, and the Ph.D. degree from Southeast University, all in applied mathematics/computational mathematics, in 2008, 2011, and 2014, respectively. She was also a joint Ph.D. student at Brown University, Providence, USA. She worked as a Postdoc research fellow at Department of Applied Mathematics of Brown University and Beijing Computational Science Research Center from 2014 to 2015. She is an associate professor at School of Mathematics of Southeast University since April 2018, the Secretary of the Jiangsu Provincial Key Laboratory of Networked Collective Intelligence of China. Prof. Zhao was a recipient of ICFDA16 Riemann-Liouville Award: Best FDA Paper (Application). Her current research interests include machine learning, fast solvers and applications of fractional differential equations and short-term data prediction.

Meichen Song received the Bachelor's degree in Statistics from Southeast University, Nanjing, Jiangsu Province. Now, she is a Ph.D. student of Applied Mathematics and Statistics department at Stonybrook University. Her current research interests include machine learning, parallel computing and Monte Carlo Methods, especially research in massive multiscale biomedical modeling and molecular dynamics simulations through machine learning.

Anqi Liu received the Bachelor's degree in Mathematics and Applied Mathematics from Southeast University, Nanjing, Jiangsu Province. She is the co-president of American Statistical Association (ASA) Student Chapter in DC from 2019. She is currently a master student at George Washington University. Her research interests focus on algorithm design, machine learning, social computing and behavioral-cultural modeling & prediction.

Yiming Wang received the Bachelor's degree in Statistics, from Southeast University, Nanjing, Jiangsu Province. Now, she is a Master student in Shanghai Jiao Tong University, Shanghai. Her main research interests include dimension reduction, variation approximation, statistics methods selection and machine learning.

Tong Wang received the Bachelor's degree in economics from Shandong of Technology and Business University, Yantai, China. Currently, she is a master student in the School of Mathematics at Southeast University, Nanjing, China. Her present research interests include machine learning and Artificial Intelligent algorithms and their applications in the shortterm traffic prediction.

Jinde Cao received the B.S. degree from Anhui Normal University, Wuhu, China, the M.S. degree from Yunnan University, Kunming, China, and the Ph.D. degree from Sichuan University, Chengdu, China, all in mathematics/applied mathematics, in 1986, 1989, and 1998, respectively. He is an Endowed Chair Professor, the Dean of the School of Mathematics, the Director of the Jiangsu Provincial Key Laboratory of Networked Collective Intelligence of China and the Director of the Research Center for Complex Systems and Network Sciences at Southeast University. Prof. Cao was a recipient of the National Innovation Award of China, Obada Prize and the Highly Cited Researcher Award in Engineering, Computer Science, and Mathematics by Thomson Reuters/Clarivate Analytics. He is elected as a fellow of IEEE, a member of the Academy of Europe, a member of the European Academy of Sciences and Arts, a fellow of Pakistan Academy of Sciences, an IASCYS academician, and a full member of Sigma X.