Heliyon 8 (2022) e10691

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Time series-based $PM_{2.5}$ concentration prediction in Jing-Jin-Ji area using machine learning algorithm models

Xin Ma [a,1], Tengfei Chen [a,1], Rubing Ge [b,*], Caocao Cui [a], Fan Xu [a], Qi Lv [a]

[a] School of Management and Economics, North China University of Water Resources and Electric Power, Zhengzhou 450046, China
[b] Environmental Protection Investment Performance Center, Chinese Academy of Environmental Planning, Beijing 100012, China

ARTICLE INFO

Keywords:
Jing-Jin-Ji city group
$PM_{2.5}$ prediction
Lasso regression
Gradient boosting
Linear SVR
K-Nearest Neighbor

ABSTRACT

Globally all countries encounter air pollution problems along their development path. As a significant indicator of air quality, $PM_{2.5}$ concentration has long been proven to be affecting the population's death rate. Machine learning algorithms proven to outperform traditional statistical approaches are widely used in air pollution prediction. However research on the model selection discussion and environmental interpretation of model prediction results is still scarce and urgently needed to lead the policy making on air pollution control. Our research compared four types of machine learning algorisms LinearSVR, K-Nearest Neighbor, Lasso regression, Gradient boosting by looking into their performance in predicting $PM_{2.5}$ concentrations among different cities and seasons. The results show that the machine learning model is able to forecast the next day $PM_{2.5}$ concentration based on the previous five days' data with better accuracy. The comparative experiments show that based on city level the Gradient Boosting prediction model has better prediction performance with mean absolute error (MAE) of 9 ug/m^3 and root mean square error (RMSE) of 10.25–16.76 ug/m^3, lower compared with the other three models, and based on season level four models have the best prediction performances in winter time and the worst in summer time. And more importantly the demonstration of models' different performances in each city and each season is of great significance in environmental policy implications.

1. Introduction

Air pollution has adverse impacts on economic development and human health, therefore, accurate prediction of air pollutant concentration is important for policy making. Globally many countries encounter air pollution problems along their development path. So as to improve citizens' health status and well-being governments worldwide have paid a lot of efforts in tackling their air pollution issue. Researches on the accurate prediction of air pollution play a fundamental role and should get more academic and public attentions. $PM_{2.5}$ is particulate matter with an average aerodynamic diameter of up to 2.5 and it is a significant indicator for air quality assessment. Epidemiological and experimental evidences have proven it to be associated with respiratory and cardiovascular mortality and morbidity rates, life expectancy (Burnett et al., 2014; Xing et al., 2016; Apte et al., 2018; Al-Hemoud et al., 2019; Diao et al., 2020; Bu et al., 2021; Geng et al., 2021), and the threat to public health may remain even when its concentration is at low levels (Feng et al., 2016; Ouyang et al., 2020; Yu et al., 2020).

Traditional statistical models such as partial least squares regression model (Polat and Gunay, 2015), generalized Markov model (Sun et al., 2013; Alyousifi et al., 2019), Bayesian method (Riccio et al., 2006; Liu et al., 2008; Faganeli Pucer et al., 2018), etc., are often used for the prediction of air pollutant concentration on time series. However, because these models all have the shortcoming of over-simplified, they inherently have difficulties in unraveling the nonlinear interaction relationship between multivariate factors and $PM_{2.5}$ concentration, so that the favorable factors for $PM_{2.5}$ prediction cannot be fully utilized (Ni et al., 2017). Time series prediction models such as Autoregressive Integrated Moving Average (ARIMA) which are specifically designed for analyzing time series data have also been used to predict $PM_{2.5}$ concentration (Gocheva-Ilieva et al., 2014; Abhilash et al., 2018; Bhatti et al., 2021). But due to the high complexity, randomness, non-stationarity and nonlinearity property of $PM_{2.5}$ time series data, ideal prediction accuracy may not be obtained solely by ARIMA model (Niu et al., 2016; Yan and Ma, 2016). By reviewing existing researches on statistical tools predicting air pollutant, it was found artificial neural network methods (ANN)

* Corresponding author.
E-mail address: gerb@caep.org.cn (R. Ge).
[1] Xin Ma and Tengfei Chen contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2022.e10691
Received 10 May 2022; Received in revised form 24 July 2022; Accepted 14 September 2022

were preferred when predicting PM and the combination of ANN and statistical models is a hot research trend (Liao et al., 2021).

Benefiting from the rapid development of the artificial intelligence (AI), huge progress in air pollutant concentration prediction has been made. Machine learning which is an important branch of AI has a long history in unraveling the interconnections in a chaotic system. Especially with the assistance from the fast developing data science, the prediction performance of the mass data-driven algorithm has been substantially enhanced (Jordan and Mitchell, 2015; Ma et al., 2020). When solving nonlinear regression problems, machine learning models are proven to have good data fitting and learning capacity. Numerous machine learning models have been widely used to make predictions in various fields, such as image processing (Glowacz, 2021a, 2021b), medical use (Kaplan et al., 2021; Khera et al., 2021), text classification (Luo, 2021), and especially air pollution prediction (Castelli et al., 2020; Choubin et al., 2020; Harishkumar et al., 2020; Liang et al., 2020; Lv et al., 2021). Ensemble algorithms have been developed to further enhance the prediction capacity of AI models, such as Gradient boosting (Shahbazi et al., 2020; Su, 2020) and adaboost model (Liu et al., 2019; Bahad and Saxena, 2020) and bagging model (Khan et al., 2022).

Compared with the chemical transport models (CTM) which forecast air pollution based on atmospheric chemistry simulations (Di et al., 2016; Hu et al., 2019; Zhang et al., 2021), machine learning models have distinct advantages of little computation cost, good learning and fitting ability etc. However, the black box model structure makes it hard to explain pollutant formation mechanism and transporting process. We reckon that considering the environmental meanings at both modeling stage and results interpretation stage is beneficial and can to some extent help overcome machine learning models' shortcomings mentioned above.

In this paper, four typical machine learning models were selected to predict $PM_{2.5}$ time series. Data collected contain air quality, meteorological, time and historical features. Rolling prediction method was applied when feeding the data into models, and optimal step length was determined by experimenting. Four commonly used indicators i.e., mean absolute error (MAE), root mean square error (RMSE), index of agreement (IA) and correlation coefficient (R) were used to evaluate the predictive performance. The comparative experiments show that the Gradient Boosting prediction model had better prediction performance with lower mean absolute error and root mean square error compared with the models such as the Lasso regression, K-Nearest Neighbor and SVR. Moreover, it has better generalization ability, which can predict the $PM_{2.5}$ concentration more accurately. And more importantly the demonstration of models' different performances on each city and each season is of great significance in environmental policy implication.

The rest of this paper is organized as follows and the structure of our research is presented in Figure 1: Linear SVR, K-Nearest Neighbor, Lasso regression, Gradient boosting are introduced in Section 2. In Section 3, prediction results of four models are evaluated. Section 4 draws the final conclusion.

2. Data and model implementation

2.1. Data description

The Jing-Jin-Ji Metropolitan Region also known as Beijing-Tianjin-Hebei, located in northeast of China (Figure 2) is China's capital economic circle. $PM_{2.5}$ pollution has become a thorny problem of environmental control in Jing-Jin-Ji region. Research on $PM_{2.5}$ pollution is of great significance for the prevention and control of urban pollution in the "capital economic circle". In 2017 China's ministry of ecology and environment issued a working plan aiming at reducing the air pollution in Jin-Jin-Ji region. The concept of atmospheric pollution transmission channel cities surrounding Jin-Jin-Ji region which are referred to "2 + 26" cities was first time officially brought up. Since then, stronger environmental policies and stricter inspections have been adopted in this region, therefore it is an ideal experimental place to test how air pollution would be affected under those two factors of economic development pressure and strong environmental protection policies happening simultaneously.

This study was conducted based on the data collected from the following monitoring sites, Beijing, Tianjin, Baoding, Cangzhou, Handan, Hengshui, Langfang, Shijiazhuang, Tangshan, Xingtai, which are the overlapping cities of Jing-Jin-Ji city group and "2 + 26" atmospheric

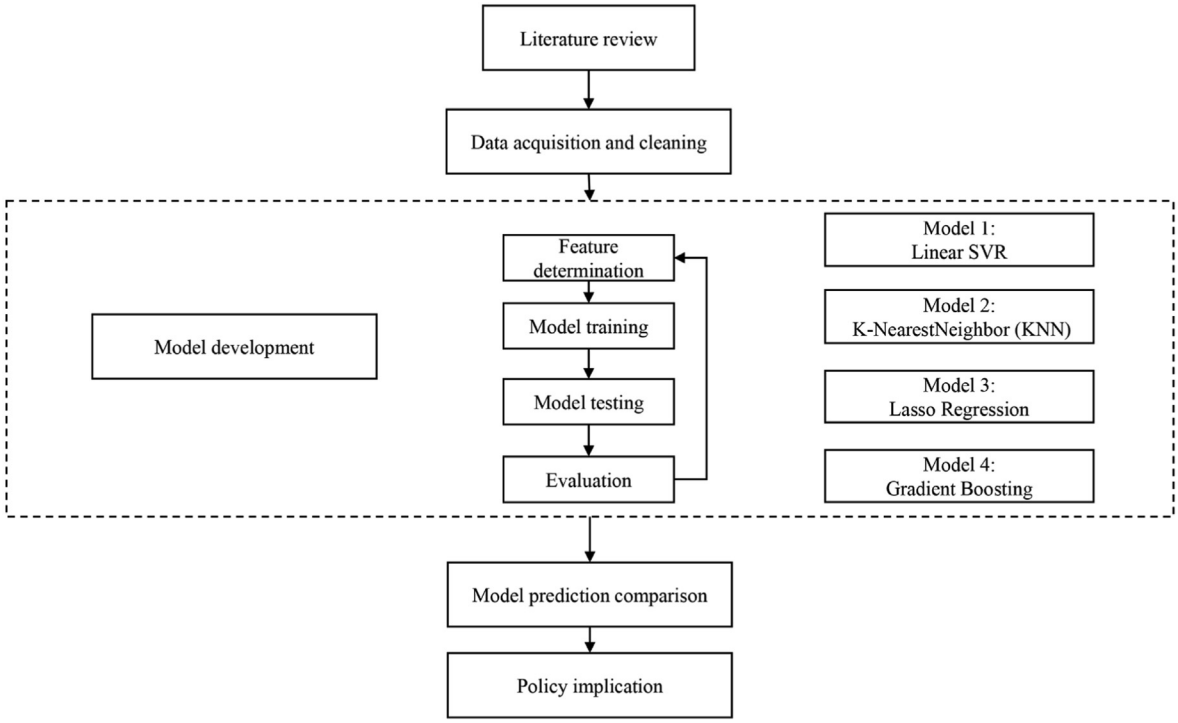

Figure 1. The diagram of the research's processes.

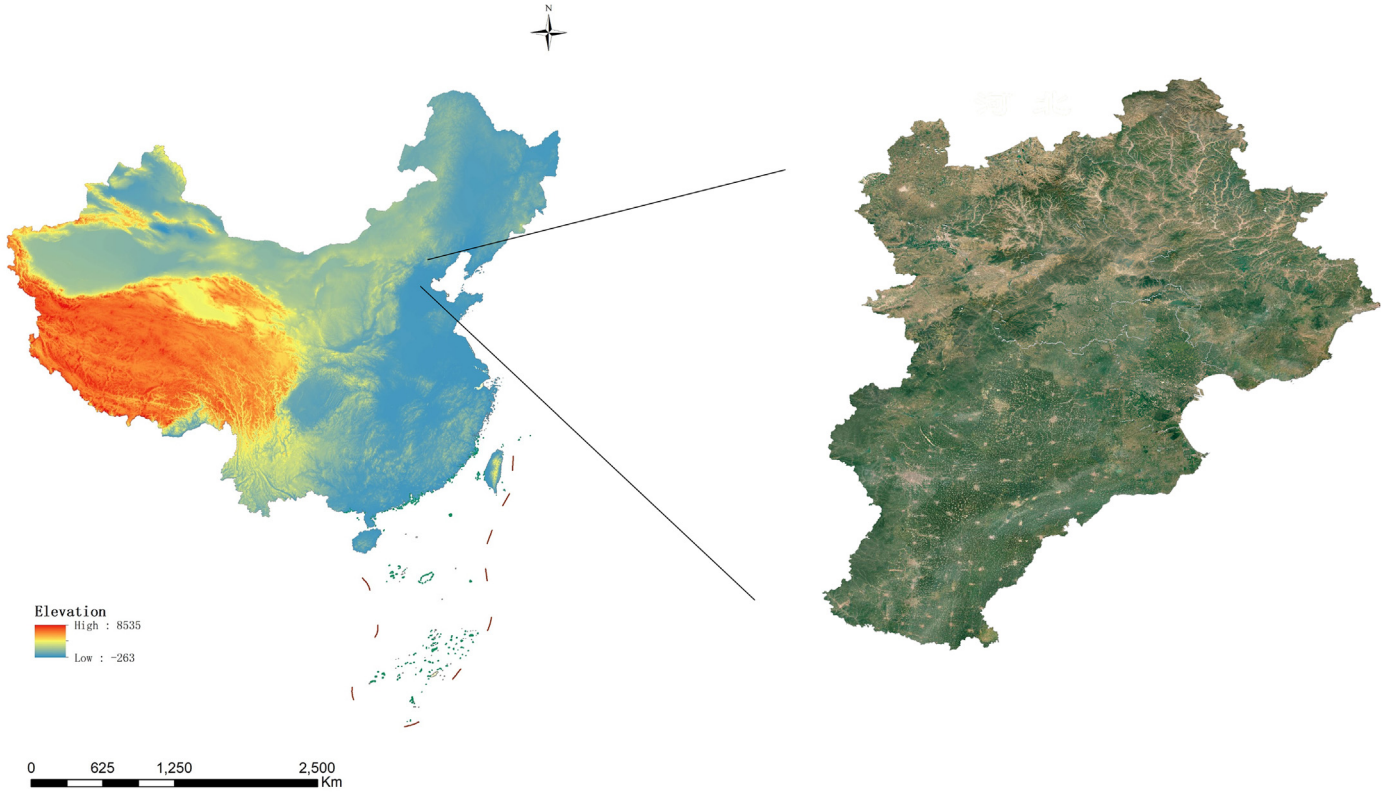

Figure 2. Geographic location of Beijing-Tianjin-Hebei in China.

pollution transmission channel cities (Figure 3). Air quality data were drawn from the China National Environmental Monitoring Centre (CNEMC). Meteorological data were drawn from China Weather (CW). We collected historical 2206 days daily data of the monitoring cites from those ten cities to train models.

Three types of features chosen to train our models are listed in Table 1. We used 1874 days' monitoring data drawn from each city cite as training and validation dataset, and 332 days data as testing dataset.

The data used as training and validation dataset were from 2013 Oct 28th to 2018 Dec 31st, and the testing dataset were formed data from 2019 Jan 1st to 2019 Dec 31st. In order to manifest the data distribution and variation more directly, the various statistical indicators of air quality and meteorological quality are calculated as shown in Table 2. In the selected time period, the $PM_{2.5}$ concentration has a minimum of 0, a maximum of 796 and a variance of 4374.42, and it also has the characteristics of high non-stationary, non-linearity.

Kernel density estimation is used to estimate an unknown density function in probability theory as one of the non-parametric test methods. To further analyze the frequency distribution of the selected variables, the kernel density estimation was applied to determine their density distribution. This method can visually demonstrate the distribution characteristics of the data through the violin graph without making any assumptions nor having any prior knowledge regarding the data

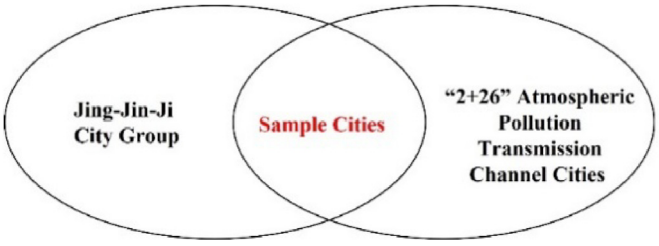

Figure 3. Sample cities selection criteria.

Table 1. Feature selection of dataset.

Indicator type	Indicators
Air quality features	PM_{10}, SO_2, NO_2, CO, O_3, AQI_L5, PM_{10}_L5, SO_2_L5, NO_2_L5, CO_L5, O_3_L5, AQI ranking_L5
Meteorological features	Lowest temperature, highest temperature, wind speed, Lowest temperature_L5, highest temperature_L5, wind speed_L5
Time features	month, year, season, year_L5, month_L5, season_L5
Historical features	$PM_{2.5}$_L1, $PM_{2.5}$_L2, $PM_{2.5}$_L3, $PM_{2.5}$_L4, $PM_{2.5}$_L5

L1 is data of one day before, L5 is data of five days before.

distribution. According to the distribution pattern of each variable (Figure 4), the $PM_{2.5}$ concentration is right skewed and the maximum value exists in the upper quartile. Moreover, the distribution trend of AQI, PM_{10}, SO_2, NO_2, and O_3 is basically consistent with $PM_{2.5}$, while the changing trend of CO, lowest temperature, highest temperature, wind speed is quite different from $PM_{2.5}$ concentration.

The geographical distribution of training and testing samples and their distribution in the four seasons of each city site is shown in Figures 5 and 6 respectively. It shows that the number of samples drawn in each season is relatively uniform. Therefore, the training results in the Jing-Jin-Ji region are representative for spatial and temporal features. The testing datasets can be grouped in two approaches, the first approach is to divide by cities, the second is to divide by seasons. Then the well-trained models were used to predict the $PM_{2.5}$ concentration by being fed with city-based, and season-based testing dataset respectively. Furthermore, empirical analysis was made according to the prediction results.

Rolling forecast method in this study was applied when training the model to enhance the generalization capacity and accuracy. The model was fed with previous five days' monitoring data in a rolling manner to predict $PM_{2.5}$ concentration of the next day. Considering that machine learning model is more applicable for larger dataset, meanwhile due to the periodicity, volatility and integrity nature of the time series data, we

Table 2. Statistical summary of the collected data (2013–2019).

	Unit	Mean	Variance	Minimum	25% quantile	Median	75% quantile	Maximum
AQI	—	116.41	5156.15	16	69	96	138	500
$PM_{2.5}$	ug/m^3	78.66	4374.42	0	36	59	98	796
PM_{10}	ug/m^3	134.76	8786.33	0	72	110	168	937
SO_2	ug/m^3	32.60	1220.53	0	11	21	40	437
NO_2	ug/m^3	48.32	559.85	0	31	44	61	235
CO	mg/m^3	1.40	1.11	0	0.75	1.09	1.7	18.92
O_3	ug/m^3	57.98	1491.04	0	26	51	83	234
Lowest temperature	°C	8.72	117.92	-20	-2	9	19	29
Highest temperature	°C	19.05	123.20	-12	9	20	29	40
Wind speed	Force (Beaufort scale)	2.41	1.00	0	2	3	3	8

applied the rolling prediction method to enlarge our training dataset. As a result, we got a training dataset of which the input data is a matrix of 18,735 × 29 and the output vector is the $PM_{2.5}$ concentration. The training dataset occupies 82.3% of the total sample data volume, therefore it matches the machine learning data slicing rule. K-fold cross validation (K = 5) was adopted to train.

The data matrix of 332 × 30 forms the rest 17.7% testing data's input and output data. Statistical summary results show that the data are sampled evenly from each city site, and season as well.

2.2. Development of predictive models

2.2.1. Linear SVR

Support vector machine (SVM) is a classification algorithm proposed by Vapnik of which the learning strategy is to maximize the interval (Cortes and Vapnik, 1995). Samples are made linearly separable by mapping the samples to a higher-dimensional feature space, and nuclear functions are also introduced to implement nonlinear mapping. Support vector regression (SVR) is developed on the basis of optimization SVM. Compared with neural network, SVR based on deep learning mechanism overcomes the problems of overfitting, underfitting and local optimization using mathematical methods and optimization techniques. As described in Eqs. (1) and (2), this model was performed as follows:

For dataset D:

$$D=\{(x_i,y_i),i=1,2,\cdots,m\}\ \ x_i\in R^N,\ \ y_i\in R^N \tag{1}$$

Separating hyperplane model can be constructed in higher-dimensional space:

$$f(x)=w^T\varphi(x)+\vartheta \tag{2}$$

$f(x)$denotes forecast value, $\varphi(x)$ is nonlinear mapping function, w is weight coefficient, ϑ is intercept, ε is maximum margin. Only when the training samples fell within the maximum margin could the prediction results be considered correct. Therefore, by introducing the relaxation variables ξ_i, $\widehat{\xi}_i$ and the penalty function E, the optimized objective function can be obtained (Eq. (3)):

$$\underset{w,\vartheta,\xi_i,\widehat{\xi}_i}{Min}\frac{1}{2}||w||^2+E\sum_{i=1}^{m}(\xi_i+\widehat{\xi}_i) \tag{3}$$

$$\text{s.t}\ \begin{cases} f(x_i)-y_i\le\varepsilon+\xi_i \\ y_i-f(x_i)\le\varepsilon+\widehat{\xi}_i \\ \xi_i\ge 0,\widehat{\xi}_i\ge 0,i=1,2,\cdots,m \end{cases}$$

The "dual problem" of support vector regression is obtained by using

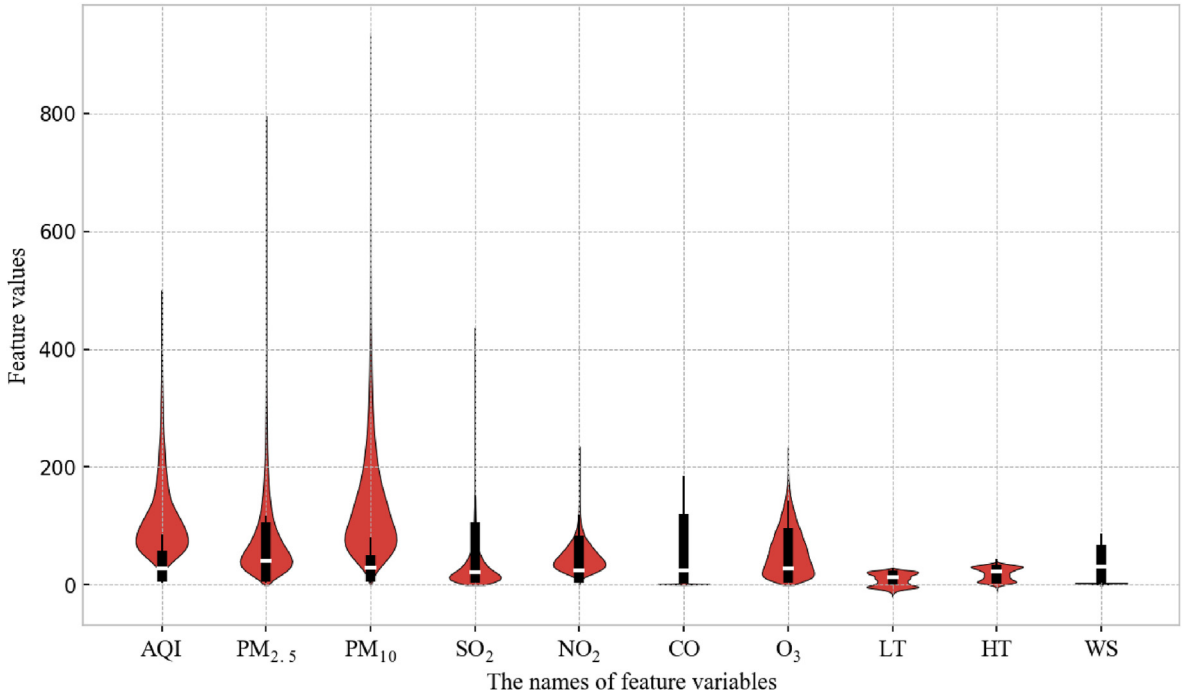

Figure 4. Violin plot of the distribution of the feature value.

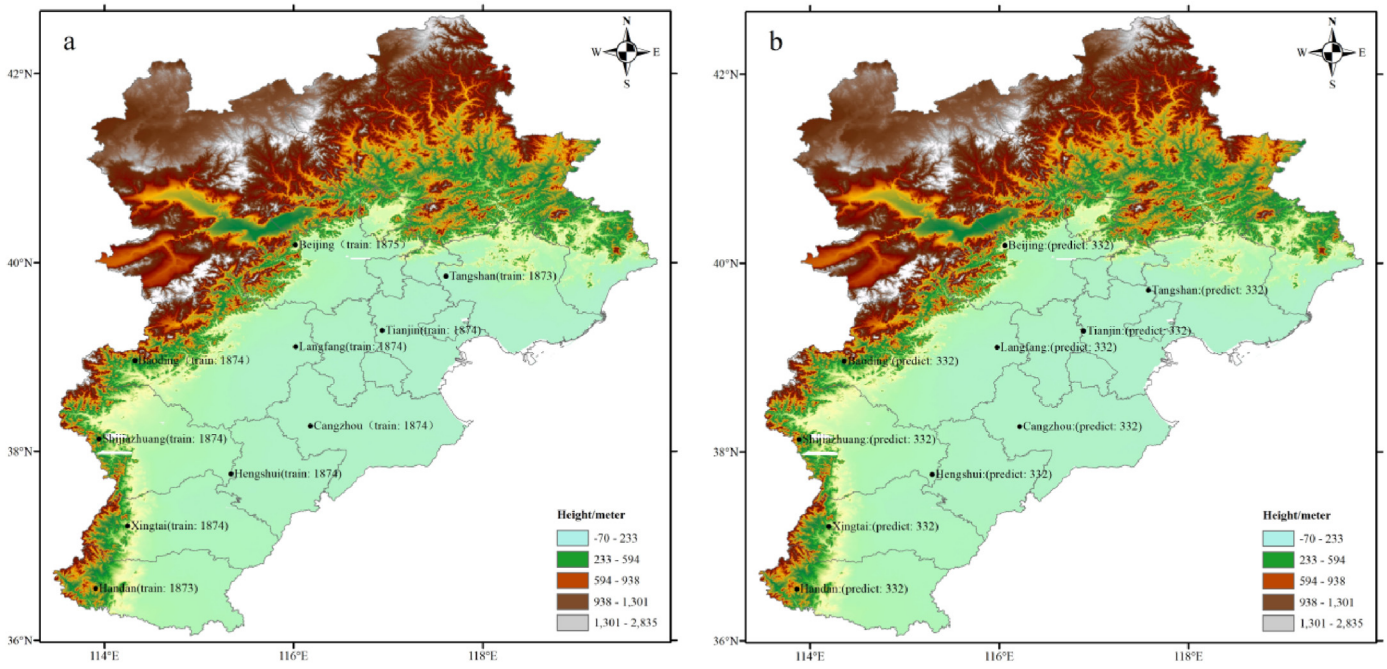

Figure 5. Geographical distribution of the training dataset (a) and the testing dataset (b) (2013–2019).

the Lagrange multiplier method. The optimal Lagrange multiplier can be solved by the sequence minimum optimization (Sequential minimal optimization, SMO) algorithm, and the final solution is obtained under the KKT (Karush-Kuhn-Tucker (KKT)) condition (Eq. (4)):

$$f(x) = \sum_{i=1}^{m} (\widehat{\lambda}_i - \lambda_i) K(x_i^T x_i) + \vartheta \tag{4}$$

λ_i, $\widehat{\lambda}_i$ are Lagrange multiplier, $K(x_i^T x_i)$ is kernel function which is linear type.

2.2.2. K-Nearest Neighbor (KNN)

The k-nearest neighbors algorithm (k-NN) was expanded by Altman (1992) after it was first developed in 1951. This supervised ML algorithm can be used both in classification and regression problems. The following is pseudo-code based on which we implemented KNN model.

2.2.3. Lasso regression

Lasso regression is the Lasso (Least absolute shrinkage and selection operator) method first proposed by Tibshirani (1996). It is a biased estimation method that can be used for feature selection in

Algorithm: KNN

input: $x_i, y_i, i \leftarrow 1...n$
output: $KNN(Xtest, Xtrain, Ytrain, Nneighbors)$

```
function GETDISTANCE(X, Y):
    return ∑((X − Y)^2)^0.5
end function
function KNN(Xtest, Xtrain, Ytrain, k):
    distances ← []
    Ykind ← []
    for i in Xtrain do
        distance.append(getdistance(Xtest, i))
    end for
    tmp ← list(enumerate(distances))
    tmp.sort(key ← lambda x : x[1])
    minkDis ← tmp[: k]
    for j in minkDis do
        tKey ← Ytrain[j[0]]
        if tKey in Ykind.keys() then
            Ykind[tKey] ← Ykind[tKey] + 1
        else
            Ykind.setdefault(tKey, 1)
        end if
    end for
    t ← sorted(Ykind.items(), key ← lambda x : x[1], reverse ← True)
end function
return t[0][0]
```

high-dimensional data. The Lasso method is designed for dealing with data with complex collinearity by constructing a penalty function allowing some coefficient to be minimized to the value 0, thus preserving the characteristics of subset shrinkage. When predicting $PM_{2.5}$, objective function $f(\lambda)$ is constructed as follow (Eq. (5)):

$$f(\lambda) = ||y - \left(\sum \lambda_i x_i + \lambda_0\right)||^2 + \alpha||\lambda|| \tag{5}$$

Here, y stands for observed $PM_{2.5}$, x_i is the value of the feature i in the feature vector of the independent variable. $(\sum \lambda_i x_i + \lambda_0)$ is the $PM_{2.5}$ concentration predicted by a linear combination of 31 features, and λ_i, λ_0 can be obtained by minimizing $f(\lambda)$.

2.2.4. Gradient Boosting

Gradient boosting model is a typical ensemble algorithm which creates model with stronger prediction ability by combining several weak classifiers (Bentéjac et al., 2021). The following is pseudo-code based on which we implemented Gradient boosting model.

Algorithm: Gradient Boosting

input: $x_i, y_i, i \leftarrow 1...n$

output: $F_M(x)$

1: Initialize $F_0(x) \leftarrow arg\,minLoss \sum_{h \epsilon H}(y_i, h(x_i))$
2: **for** $m \leftarrow 1 : M$ **do**
3: $g_m \leftarrow -\frac{\partial Loss(y, F_{m-1}(x))}{\partial F_{m-1}(x)}$ Compute the negative gradient
4: $\sum_{i=1}^{N}(g_m^i - h(x_i))^2$ Fit a weak learner which minimize
5: $F_m \leftarrow F_{m-1} + vh(x)$ Update
6: **end for**
7: **return** $F(x) \leftarrow F_M(x)$

2.3. Evaluation metrics

Evaluation metrics for machine learning models are often used to quantify the performance of predictive model by comparing the prediction values and actual observed values. Four commonly used indicators i.e., mean absolute error (MAE), root mean square error (RMSE), index of agreement (IA) and Pearson's correlation coefficient (described in Eqs. (6), (7), (8), and (9)) were adopted to measure the prediction accuracy here:

$$MAE(y, \hat{y}) = \frac{1}{m}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{7}$$

$$IA = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(|\hat{y}_i - \overline{y}| + |y_i - \overline{y}|)^2} \tag{8}$$

$$R^2(y, \hat{y}) = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \overline{y})^2} \tag{9}$$

where y_i is the prediction of the $PM_{2.5}$ for time i, while $\hat{y}_i$ represents the actual value for time I, and $\overline{y}$ is the observed mean.

3. Experimental results and discussion

3.1. Prediction comparison based on city level

Abovementioned four evaluation indicators were used here to compare the prediction results of 10 surveyed cities' daily $PM_{2.5}$ concentration generated by model Lasso, Gradient Boosting, LinearSVR,

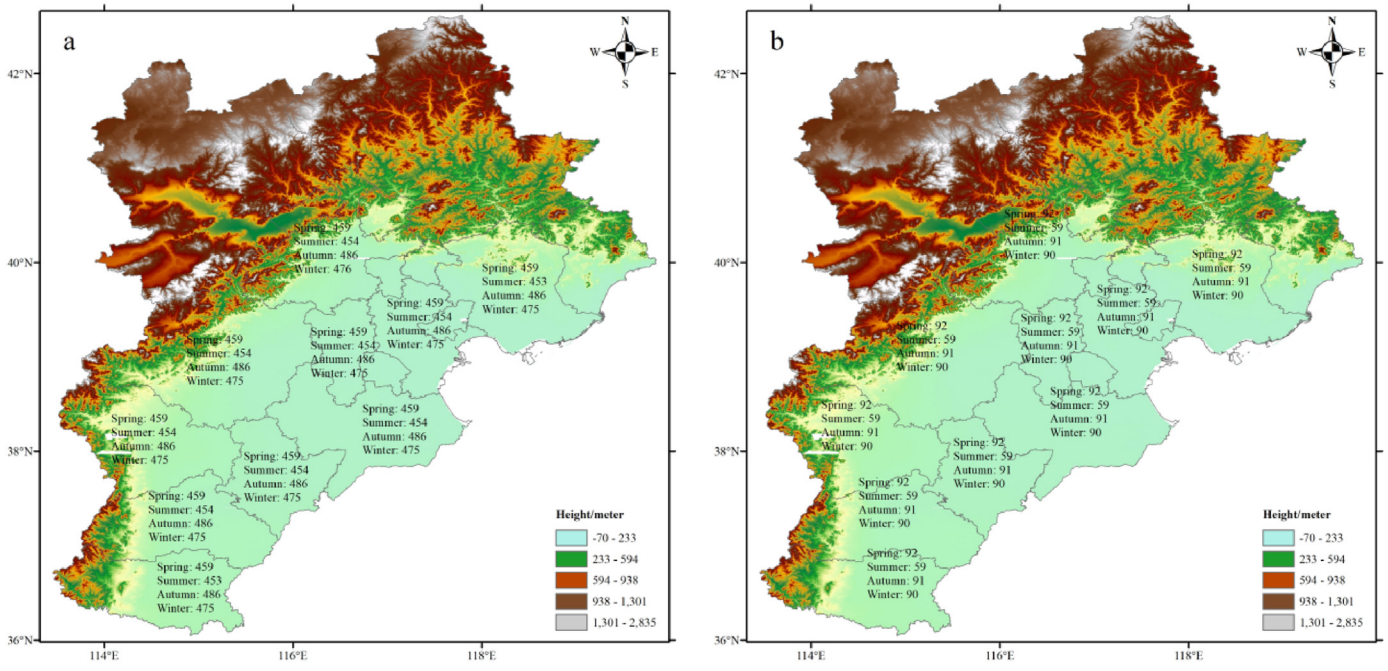

Figure 6. Geographical distribution of training dataset (a) and testing dataset (b) for each season (2013–2019).

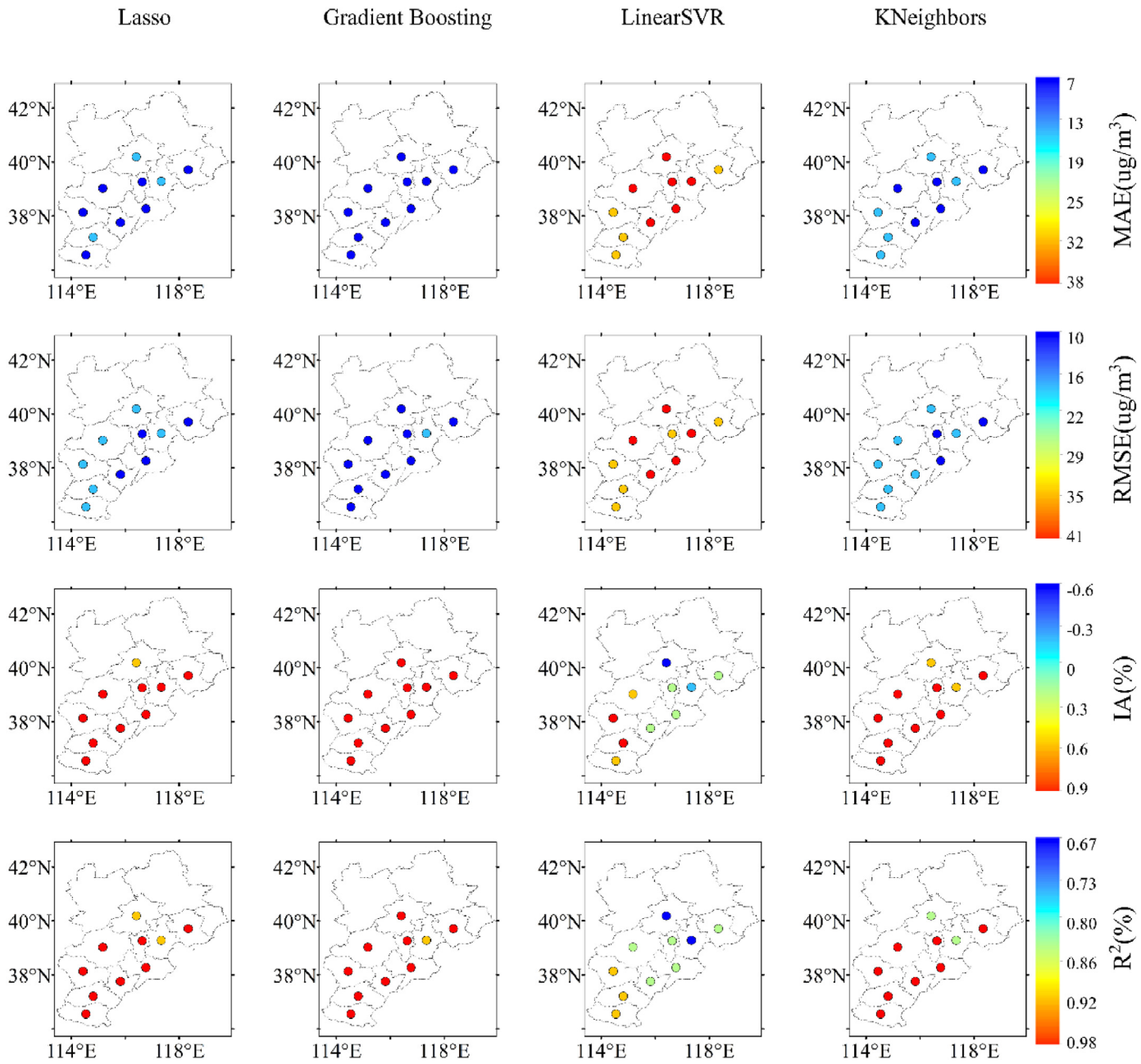

Figure 7. Prediction performance evaluation for four models based on city level.

Table 3. Distribution of the prediction errors of each city.

City	Lasso prediction error		Gradient Boosting prediction error		LinearSVR prediction error		KNeighbors prediction error	
	90%	75%	90%	75%	90%	75%	90%	75%
Beijing	-26–99	-22–13	-19–27	-14–8	-55–61	-52~ -20	-24–81	-18–15
Tianjin	-26–100	-18–16	-21–36	-16–10	-54–70	-45~ -13	-28–67	-20–14
Baoding	-20–89	-15–15	-13–32	-11–9	-52–47	-48~ -21	-20–60	-14–15
Cangzhou	-17–118	-13–14	-15–32	-10–9	-48–78	-45~ -19	-18–100	-14–16
Handan	-17–77	-13–22	-13–50	-8–17	-46–42	-41~ -8	-18–94	-13–23
Hengshui	-23–52	-17–14	-16–33	-13–7	-52–12	-48~ -19	-24–61	-17–14
Langfang	-13–97	-11–15	-9–40	-7–14	-46–63	-43~ -17	-14–72	-11–15
Shijiazhuang	-15–77	-12–22	-10–44	-7–13	-44–47	-39~ -9	-17–90	-13–20
Tangshan	-11–119	-9–21	-8–74	-6–17	-45–87	-43~ -11	-13–64	-9–20
Xingtai	-18–81	-13–24	-15–46	-10–14	-45–47	-41~ -9	-21–103	-14–20

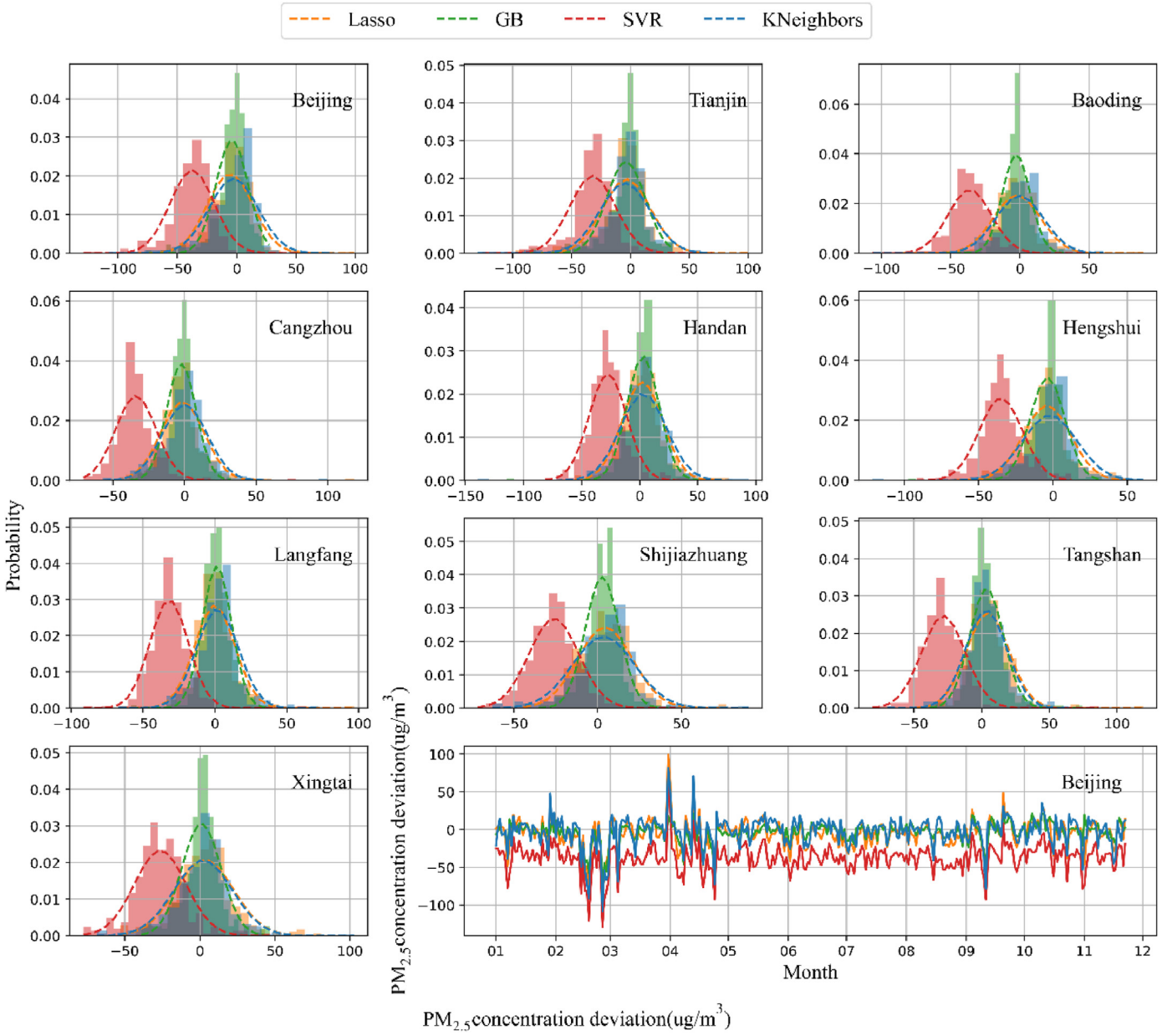

Figure 8. Probability distribution of $PM_{2.5}$ prediction errors for each city.

KNeighbors (Figure 7). The geographical distribution of evaluation results based on city lever is illustrated in Figure 7. Gradient Boosting model has the best MAE result which is highly concentrated around 9 ug/m^3, followed by Lasso and KNeighbors model of which the prediction MAE are both around 12 ug/m^3, in contrast, LinearSVR model prediction MAE values fall into the range of 27.09–38.26 ug/m^3. From the perspective of RMSE, Gradient Boosting is still the best model compared to the rest three, and the RMSE of its predictions and observations is between 10.25 and 16.76 ug/m^3. Lasso and KNeighbors' RMSE are 14.00–20.47 ug/m^3 and 14.67–21.85 ug/m^3 respectively, LinearSVR is ranking the fourth with a poor result of 29.97–41.85 ug/m^3. Gradient Boosting model's prediction IA ranges from 0.92 to 0.99, better than Lasso, KNeighbors and LinearSVR. The performances of the four models in the fourth indicator R^2 are basically similar to their performances in IA. Among the 10 sample cities, apart from Gradient Boosting, all other three models generally cannot achieve ideal prediction results for Beijing and Tianjin like other cities. So on city level, our results show that daily $PM_{2.5}$ concentration prediction generated by Gradient Boosting model outperformed the Lasso and KNeighbors model, and LinearSVR has the poorest capacity.

Prediction error's distribution for each city is summarized in Table 3, and illustrated in Figure 8. More specifically, LinearSVR predictions tend to have negative errors, and Lasso and KNeighbors prediction errors have larger range. Gradient Boosting outperformed the other models, because most of its prediction errors are concentrated evenly around 0 ug/m^3.

3.2. Prediction comparison based on seasons

Scatter plot of the $PM_{2.5}$ predictions and the observations grouped by seasons is shown in Figure 9. Combined with the IA index and MAE of different seasons in Table 4, the overall performance of Lasso, Gradient Boosting, LinearSVR and KNeighbors models on the quarterly $PM_{2.5}$ concentration prediction in the Jing-Jin-Ji region shows that the four models have the best prediction results in winter time, and the worst in summer time. This is likely because in Jing-Jin-Ji region the particulate pollutant is the most serious in winter time, average $PM_{2.5}$ concentration

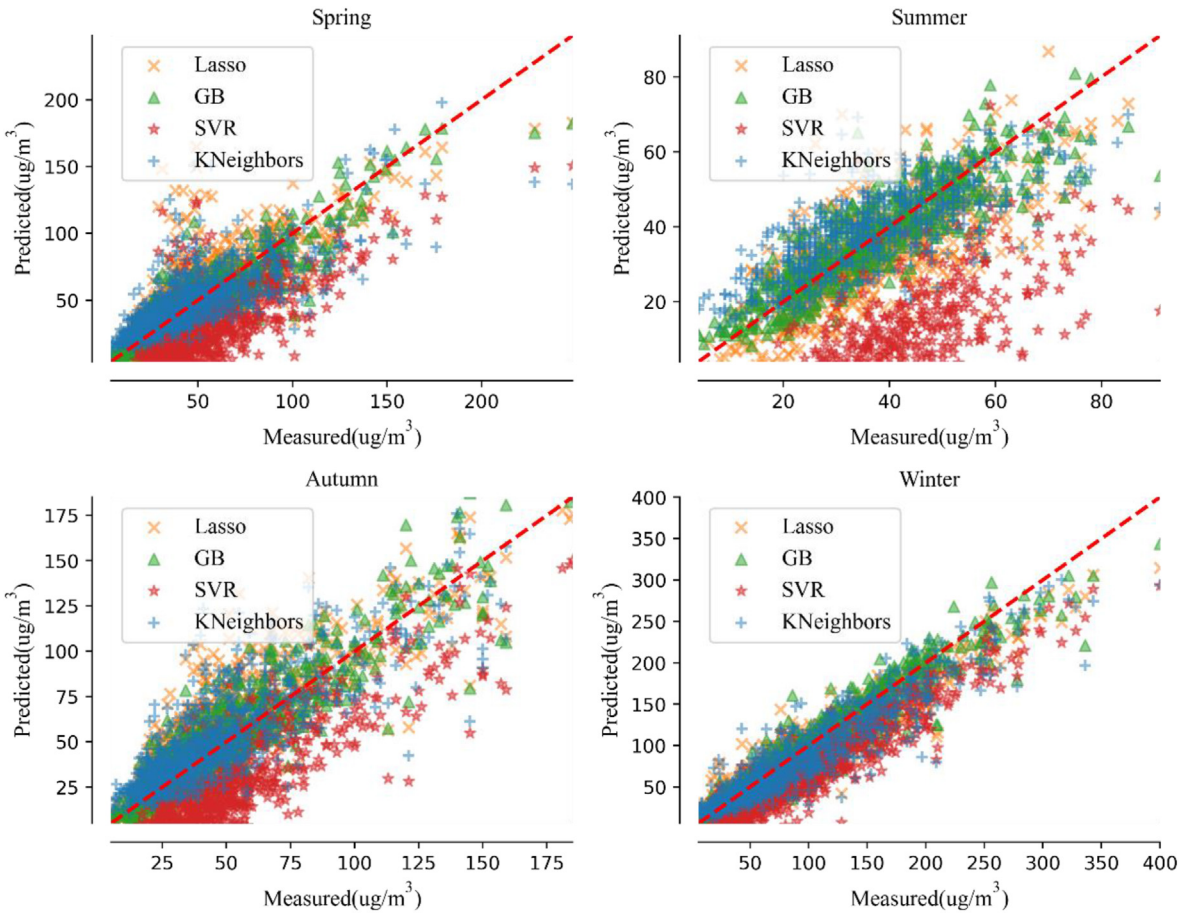

Figure 9. Scattering plot of predictions and observations for each season.

Table 4. Evaluation on the prediction results of four models for each season.

		Lasso	Gradient Boosting	LinearSVR	KNeighbors
Spring	MAE	12.82	8.34	33.02	13.04
	RMSE	18.62	11.92	35.98	18.77
	IA	0.90	0.95	0.72	0.87
	R^2	0.60	0.84	-0.49	0.59
Summer	MAE	9.02	5.47	33.04	7.98
	RMSE	11.80	7.31	34.55	10.20
	IA	0.84	0.93	0.51	0.84
	R^2	0.40	0.77	-4.18	0.55
Autumn	MAE	12.92	8.86	29.58	12.23
	RMSE	17.53	12.33	32.78	17.13
	IA	0.91	0.96	0.76	0.91
	R^2	0.64	0.82	-0.25	0.66
Winter	MAE	14.09	11.27	33.74	16.48
	RMSE	19.79	16.61	38.01	23.64
	IA	0.97	0.98	0.91	0.96
	R^2	0.90	0.93	0.65	0.86

Table 5. Model construction time and occupied memory size.

Model	Model construction (S)	Occupied memory size (KB)
Lasso	1.02	8.78
Gradient Boosting	2.68	378
LinearSVR	8.07	3.81
KNeighbors	5.04	6010.88

is much higher than the other three seasons, in contrast, particulate pollutant in summer has minor effects and ozone pollution is the primary pollution issue.

The distinct performance of different models has practical implications in the selection of models targeting different seasons, thus affecting policy making process. Gradient Boosting has the best results in IA and MAE. Lasson and KNeighbours have the similar results in IA and MAE for

summer and winter time, in spring, Lasso slightly outperformed KNeighbors. Comparatively they are superior than LinearSVR overall.

3.3. Overall evaluation

As Table 5 shows, LinearSVR model had the longest construction time, followed by KNeighbors, and Lasso had the shortest model construction time of 1.02 s. In terms of memory occupancy, KNeighbors occupies the largest memory, the model size is 6010.88 KB, followed by Gradient, Boosting model, the memory occupancy size is 378 KB, the smallest model is LinearSVR, and the memory size is only 3.81 KB.

Researches show that SVM is difficult to implement for large-scale training samples and is sensitive to missing data, and when the sample of the K Nearest Neighbor model is not balanced, the prediction bias is relatively large, and it will likely lead to a curse of dimensionality. Lasso regression is a generalized linear model, it has limitations in handling samples with nonlinearity, randomness, and uncertainty features. The advantage of applying Gradient Boosting model in $PM_{2.5}$ concentration prediction and influencing factor analysis is obvious: on one hand, it can improve the feature selection process, on the other hand, it can reduce the complexity of model construction and the risk of model overfitting.

3.4. Discussion

Industrial structure and regional environment management policies play a crucial role in determining air pollution concentration during certain period (Zheng et al., 2020), thus inevitably creating obstacles for machine learning models to gain ideal prediction performance. For instance, among the 10 cities, Tangshan and Handan's development are largely depending on heavy industries such as steel, coke and cement. Market fluctuation or production regulations due to environmental protection purpose both can lead to air pollution's sudden change in the short term, which will become noise for the model training. In contrast, densely populated cities like Beijing and Tianjin, their pollution emissions are mainly attributed to household and transportation sectors. And also compared to other cities, Beijing, Tianjin and Shijiazhuang are facing more stringent industry development policies. The uncertainties mainly lie in the hourly data, comparatively, good prediction accuracy can be attained for daily, seasonal and annual data with less efforts.

During winter time, despite the stringent ban on scattered coal consumption in rural area, coal burning is still the main approach for heating in most northern cities in China, which puts heavy pressure on regional air quality. Further, though sampled cities are all located in the same climate zone, their locations in the pollution transmission channel are very decisive in pollutants formation process. Additive effects and spillover effects of various air pollution are making the differences of seasonal $PM_{2.5}$ concentrations among cities more significant. All factors combined are making the system more complex.

Our research was conducted with the aim of comparing the models' prediction capacity for complex experimenting conditions with multi-sourced influencing factors, therefore ideally it is expected to make contributions in the following ways: (1) raise the awareness for policy makers of the discrepancy of machine learning models, and consider the regional and seasonal differences when selecting models; (2) selected features for training models ending up with good prediction accuracy can help enlighten the similar work.

4. Conclusion

In this paper, to get models with good $PM_{2.5}$ prediction accuracy, we collected daily monitoring data from 10 atmospheric pollution transmission channel cities located in Jing-Jin-Ji area, the data mainly incorporate air quality features, meteorological features, time features and historical features. Based on the multi-sourced data, four machine learning models, LinearSVR, KNeighbors, Lasso, Gradient Boosting, were used to make predictions of $PM_{2.5}$ concentration on city level and season level.

First, the city level results show that the prediction performance of Gradient Boosting model was significantly better than Lasso and KNeighbors model, and LinearSVR model's performance was comparatively dissatisfying. Two cities (Beijing, Tianjin) had a slightly lower IA value, while the remaining eight cities had a relatively significant IA value, which is consistent with the results of the root mean square error. LinearSVR predictions tend to have negative errors, and Lasso and KNeighbors prediction errors have larger range. Gradient Boosting outperformed the other models, because most of its prediction errors are concentrated evenly around 0 ug/m^3.

Second, the results of seasonal prediction show that the four models had the best prediction performances in winter time and the worst in summer time. This is likely because particulate pollution in the Beijing-Tianjin-Hebei region is generally more serious in winter time, with $PM_{2.5}$ concentration much higher than in other three seasons. In contrast, in summer the particulate pollution level is low while ozone pollution is the primary pollutant.

Lastly, in the model overall evaluation, the Gradient Boosting model had comparatively ideal performance in terms of training time and occupied memory size compared to the other three.

Social and economic and urban development factors have also been proven to affect the $PM_{2.5}$ concentration. For future work, it is meaningful to add these features into the modeling process, thus models untangling interconnections between multi-disciplinary features and air quality can be built, based on which richer environmental implications will be attained. In addition, as the time series dataset is enlarged after gathering more recent data, effects of environmental protection policies issued by central government in recent years, e.g. *Three-year Action Plan to Fight Air Pollution (2018)* and *Further Prevention and Control of Pollution (2021)* on air pollution control can be simulated and evaluated.

Declarations

Author contribution statement

Xin Ma: Conceived and designed the experiments; Wrote the paper.

Tengfei Chen: Performed the experiments; Analyzed and interpreted the data.

Rubing Ge: Analyzed and interpreted the data; Wrote the paper.

Caocao Cui; Fan Xu; Qi Lv: Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by Key Soft Science Projects in Henan Province (222400410010) and Philosophy and Social Science Team Project of North China University of Water Resources and Electric Power (20200704).

Data availability statement

Data will be made available on request.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Al-Hemoud, A., Gasana, J., Al-Dabbous, A., Alajeel, A., Al-Shatti, A., Behbehani, W., Malak, M., 2019. Exposure levels of air pollution (PM2. 5) and associated health risk in Kuwait. Environ. Res. 179, 108730.

Abhilash, M., Thakur, A., Gupta, D., Sreevidya, B., 2018. Time series analysis of air pollution in Bengaluru using ARIMA model. In: Ambient Communications and Computer Systems. Springer, pp. 413–426.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. Am. Statistician 46 (3), 175–185.

Alyousifi, Y., Ibrahim, K., Kang, W., Zin, W.Z.W., 2019. Markov chain modeling for air pollution index based on maximum a posteriori method. Air Qual., Atmos. Health 12 (12), 1521–1531.

Apte, J.S., Brauer, M., Cohen, A.J., Ezzati, M., Pope III, C.A., 2018. Ambient PM2. 5 reduces global and regional life expectancy. Environ. Sci. Technol. Lett. 5 (9), 546–551.

Bahad, P., Saxena, P., 2020. Study of adaboost and gradient boosting algorithms for predictive analytics. In: International Conference on Intelligent Computing and Smart Communication 2019. Springer.

Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev. 54 (3), 1937–1967.

Bhatti, U.A., Yan, Y., Zhou, M., Ali, S., Hussain, A., Qingsong, H., Yu, Z., Yuan, L., 2021. Time series analysis and forecasting of air pollution particulate matter (PM 2.5): an SARIMA and factor analysis approach. IEEE Access 9, 41019–41031.

Bu, X., Xie, Z., Liu, J., Wei, L., Wang, X., Chen, M., Ren, H., 2021. Global PM2.5-attributable Health burden from 1990 to 2017: estimates from the global burden of disease study 2017. Environ. Res. 197, 111123.

Burnett, R.T., Pope III, C.A., Ezzati, M., Olives, C., Lim, S.S., Mehta, S., Shin, H.H., Singh, G., Hubbell, B., Brauer, M., 2014. An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. Environ. Health Perspect. 122 (4), 397–403.

Castelli, M., Clemente, F.M., Popovič, A., Silva, S., Vanneschi, L., 2020. A machine learning approach to predict air quality in California. In: Complexity 2020.

Choubin, B., Abdolshahnejad, M., Moradi, E., Querol, X., Mosavi, A., Shamshirband, S., Ghamisi, P., 2020. Spatial hazard assessment of the PM10 using machine learning models in Barcelona, Spain. Sci. Total Environ. 701, 134474.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Di, Q., Koutrakis, P., Schwartz, J., 2016. A hybrid prediction model for PM2. 5 mass and components using a chemical transport model and land use regression. Atmos. Environ. 131, 390–399.

Diao, B., Ding, L., Zhang, Q., Na, J., Cheng, J., 2020. Impact of urbanization on PM2.5-related health and economic loss in China 338 cities. Int. J. Environ. Res. Publ. Health 17 (3), 990.

Faganeli Pucer, J., Pirš, G., Štrumbelj, E., 2018. A Bayesian approach to forecasting daily air-pollutant levels. Knowl. Inf. Syst. 57 (3), 635–654.

Feng, S., Gao, D., Liao, F., Zhou, F., Wang, X., 2016. The health effects of ambient PM2. 5 and potential mechanisms. Ecotoxicol. Environ. Saf. 128, 67–74.

Geng, G., Zheng, Y., Zhang, Q., Xue, T., Zhao, H., Tong, D., Zheng, B., Li, M., Liu, F., Hong, C., He, K., Davis, S.J., 2021. Drivers of PM2.5 air pollution deaths in China 2002–2017. Nat. Geosci. 14 (9), 645–650.

Glowacz, A., 2021a. Thermographic fault diagnosis of ventilation in BLDC motors. Sensors 21 (21), 7245.

Glowacz, A., 2021b. Ventilation diagnosis of angle grinder using thermal imaging. Sensors (Basel) 21 (8).

Gocheva-Ilieva, S.G., Ivanov, A.V., Voynikova, D.S., Boyadzhiev, D.T., 2014. Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach. Stoch. Environ. Res. Risk Assess. 28 (4), 1045–1060.

Harishkumar, K., Yogesh, K., Gad, I., 2020. Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models. Procedia Comput. Sci. 171, 2057–2066.

Hu, J., Ostro, B., Zhang, H., Ying, Q., Kleeman, M.J., 2019. Using chemical transport model predictions to improve exposure assessment of PM2. 5 constituents. Environ. Sci. Technol. Lett.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349 (6245), 255–260.

Kaplan, A., Cao, H., FitzGerald, J.M., Iannotti, N., Yang, E., Kocks, J.W., Kostikas, K., Price, D., Reddel, H.K., Tsiligianni, I., 2021. Artificial intelligence/machine learning in respiratory medicine and potential role in asthma and COPD diagnosis. J. Allergy Clin. Immunol. Pract. 9 (6), 2255–2261.

Khan, N.U., Shah, M.A., Maple, C., Ahmed, E., Asghar, N., 2022. Traffic flow prediction: an intelligent scheme for forecasting traffic flow using air pollution data in smart cities with bagging ensemble. Sustainability 14 (7), 4164.

Khera, R., Haimovich, J., Hurley, N.C., McNamara, R., Spertus, J.A., Desai, N., Rumsfeld, J.S., Masoudi, F.A., Huang, C., Normand, S.-L., 2021. Use of machine learning models to predict death after acute myocardial infarction. JAMA Cardiol. 6 (6), 633–641.

Liang, Y.-C., Maimury, Y., Chen, A.H.-L., Juarez, J.R.C., 2020. Machine learning-based prediction of air quality. Appl. Sci. 10 (24), 9151.

Liao, K., Huang, X., Dang, H., Ren, Y., Zuo, S., Duan, C., 2021. Statistical approaches for forecasting primary air pollutants: a review. Atmosphere 12 (6), 686.

Liu, Y., Guo, H., Mao, G., Yang, P., 2008. A Bayesian hierarchical model for urban air quality prediction under uncertainty. Atmos. Environ. 42 (36), 8464–8469.

Liu, H., Jin, K., Duan, Z., 2019. Air PM2.5 concentration multi-step forecasting using a new hybrid modeling method: comparing cases for four cities in China. Atmos. Pollut. Res. 10 (5), 1588–1600.

Luo, X., 2021. Efficient English text classification using selected machine learning techniques. Alex. Eng. J. 60 (3), 3401–3409.

Lv, L., Wei, P., Li, J., Hu, J., 2021. Application of machine learning algorithms to improve numerical simulation prediction of PM2.5 and chemical components. Atmos. Pollut. Res. 12 (11), 101211.

Ma, J., Ding, Y., Cheng, J.C.P., Jiang, F., Tan, Y., Gan, V.J.L., Wan, Z., 2020. Identification of high impact factors of air quality on a national scale using big data and machine learning techniques. J. Clean. Prod. 244, 118955.

Ni, X.Y., Huang, H., Du, W.P., 2017. Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data. Atmos. Environ. 150, 146–161.

Niu, M., Wang, Y., Sun, S., Li, Y., 2016. A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5 concentration forecasting. Atmos. Environ. 134, 168–180.

Ouyang, R., Yang, S., Xu, L., 2020. Analysis and risk assessment of PM2.5-bound PAHs in a comparison of indoor and outdoor environments in a middle school: a case study in beijing, China. Atmosphere 11 (9), 904.

Polat, E., Gunay, S., 2015. The comparison of partial least squares regression, principal component regression and ridge regression with multiple linear regression for predicting pm10 concentration level based on meteorological parameters. J. Data Sci. 13 (4), 663–692.

Riccio, A., Barone, G., Chianese, E., Giunta, G., 2006. A hierarchical Bayesian approach to the spatio-temporal modeling of air quality data. Atmos. Environ. 40 (3), 554–566.

Shahbazi, Z., Hazra, D., Park, S., Byun, Y.C., 2020. Toward improving the prediction accuracy of product recommendation system using extreme gradient boosting and encoding approaches. Symmetry 12 (9), 1566.

Su, Y., 2020. Prediction of air quality based on gradient boosting machine method. In: 2020 International Conference on Big Data and Informatization Education (ICBDIE). IEEE.

Sun, W., Zhang, H., Palazoglu, A., Singh, A., Zhang, W., Liu, S., 2013. Prediction of 24-hour-average PM2. 5 concentrations using a hidden Markov model with different emission distributions in Northern California. Sci. Total Environ. 443, 93–103.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B 58 (1), 267–288.

Xing, Y.-F., Xu, Y.-H., Shi, M.-H., Lian, Y.-X., 2016. The impact of PM2. 5 on the human respiratory system. J. Thorac. Dis. 8 (1), E69.

Yan, Q., Ma, C., 2016. Application of integrated ARIMA and RBF network for groundwater level forecasting. Environ. Earth Sci. 75 (5).

Yu, W., Guo, Y., Shi, L., Li, S., 2020. The association between long-term exposure to low-level PM2. 5 and mortality in the state of Queensland, Australia: a modelling study with the difference-in-differences approach. PLoS Med. 17 (6), e1003141.

Zhang, W., Hai, S., Zhao, Y., Sheng, L., Zhou, Y., Wang, W., Li, W., 2021. Numerical modeling of regional transport of PM2.5 during a severe pollution event in the Beijing–Tianjin–Hebei region in November 2015. Atmos. Environ. 254, 118393.

Zheng, Y., Peng, J., Xiao, J., Su, P., Li, S., 2020. Industrial structure transformation and provincial heterogeneity characteristics evolution of air pollution: evidence of a threshold effect from China. Atmos. Pollut. Res. 11 (3), 598–609.