Regional Collaborative Prediction of Air Pollutants Based on CNN-BiLSTM Model

Yanan Lu (yananlu_nan@163.com)
Shanghai University of Finance and Economics  https://orcid.org/0000-0002-5066-2198

Kun Li
Tiangong University

Research Article

Keywords: deep learning, air pollutant prediction, long short-term memory (LSMT), collaborative prediction, Bayesian optimization, data-driven model

Posted Date: February 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2537110/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Regional Collaborative Prediction of Air Pollutants Based on CNN-BiLSTM Model

Yanan Lu¹,*  Kun Li²

1. School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China
2. School of Economics and Management, Tiangong University, Tianjin 300387, China
* Corresponding author: yananlu_nan@163.com

Abstract The development of industry has brought serious air pollution problems. It is very important to establish a high-precision and high-performance air quality prediction model and take corresponding control measures. In this paper, based on four years of air quality and meteorological data in Tianjin, China, the relationship between various meteorological factors and air pollutant concentrations are analyzed, the abnormal data are detected and preprocessed. A hybrid deep learning model consisting of convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) is proposed to predict pollutant concentrations, and the effects of three different database input modes are compared. In addition, the Bayesian optimization algorithm is applied to obtain the optimal combination of hyper-parameters for the proposed deep learning model, which makes the model have higher generalization ability. Furthermore, based on the air quality data of multi stations in the region, a regional collaborative prediction method is designed, the concept of strongly-correlated station (SCS) is defined, and the results of collaborative prediction are modified using the idea of SCS, effectively improving the accuracy of prediction.

Keyword deep learning; air pollutant prediction; long short-term memory (LSMT); collaborative prediction; Bayesian optimization; data-driven model

Introduction

With the continuous advancement of industrialization and urbanization, air pollution has become an increasingly serious problem and has caused widespread concerns around the world (Castells-Quintana et al. 2021). Establishing air quality prediction model and taking corresponding control measures in industrial production and daily life can effectively reduce the harm of air pollution to human health and environment. According to China’s ambient air quality standards (GB3095-2012), there are six kinds of conventional air pollutants used to measure air quality, namely fine particulate matter (PM2.5), ozone (O3), sulfur dioxides (SO2), inhalable particles (PM10), nitrogen dioxides (NO2), and carbon monoxide (CO) (2012).

Up to now, many studies have been carried out on the air quality prediction. Because the air quality data are more related to their historical data and meteorological data, which have a lot of uncertainties, and the relationships among the data are also more complex, the traditional model-based methods have some limitations in solving these problems, and the data-driven technology is used to effectively analyze the air quality prediction. Among these data-driven methods, machine learning (ML) and artificial neural
network (ANN) models have become the popular approach for air pollution modeling, such as the support vector regression (SVR) model (Sanchez et al. 2011; Liu et al. 2019), linear regression (LR) and random forest (RF) model (Gu et al. 2021), BP neural networks (Ding et al. 2016), multilayer perceptron (MLP) model (Stadlober et al. 2008), RF and extreme gradient boosting (Joharestani et al. 2019), etc.

However, these models depend on predefined parameters and nonlinear mapping, it is difficult to capture the true underlying nonlinear relationship between inputs and target values. Recently, deep-learning models, which are advanced forms of traditional machine learning techniques, are becoming very popular in some field, such as computer vision (Dourado et al. 2021), natural language processing (Popel et al. 2020), navigation (Long et al. 2017), etc. The prediction of urban air pollutant concentration based on deep-learning has also received corresponding attention (Li et al. 2016; Zaini et al. 2022). Through effective training of a large number of data, deep-learning can well extract the temporal and spatial correlation among data. Bai et al. (2019) proposed a seasonal stacked autoencoder model combining seasonal analysis and deep feature learning for forecasting the hourly PM2.5 concentration. Qi et al. (2018) proposed a deep air learning, which utilized the information pertaining to the unlabeled spatiotemporal data to improve the performace of the interpolation and the prediction, and performs feature selection and association analysis to reveal the main relevant features to the variation of the air quality.

Among those deep learning-based prediction models, the long short-term memory (LSTM) (Hochreiter S. 1997) model can capture temporal autocorrelations and can be widely used in prediction tasks by configuring the corresponding mechanisms for historical time series data. More importantly, the LSTM model can overcome gradient explosion and disappearance in error backpropagation and is much better at capturing long short-term information. LSTM autoencoder multitask learning model was used to predict PM2.5 time series in multiple locations (Xu and Yoneda 2021). Wang et al. (2021) used chi-square test LSTM (CT-LSTM) to improve the prediction accuracy of AQI and compared it with other four methods. Li et al. (2017) proposed a LSTM extended model (LSTME) to predict hourly PM2.5 pollutant concentrations. LSTM layers were used to extract inherent features from historical air pollutant data, and auxiliary data to enhance the performance. Nath et al. (2021) presented a comparative study of various statistical and deep learning methods to forecast long-term pollution trends for PM2.5 and PM10 concentrations in the upcoming months. Qi et al. (2019) used graph convolutional LSTM (GC-LSTM) to model and forecast the spatiotemporal variation of PM2.5 concentration. Espinosa et al. (2021) completed a methodology to build deep learning and machine learning models for the forecast of contaminants NO2 and NOx with and without O3, and also considered sliding window transformation with different window sizes chosen appropriately in the context of the problem. Wardana et al. (2021) predicted air quality in the following 24 hours with a CNN-LSTM, the model was optimized for edge devices, which reduced file size to a quarter of the original. Wang et al. (2021) proposed a spatiotemporal convolutional recursive long short-term memory (CR-LSTM) model for predicting the PM2.5 concentration in long-term prediction tasks by combining a convolutional long short-term memory (ConvLSTM) neural network and a recursive strategy. Du et al. (2021) presented the base modules include CNNs and Bi-directional LSTM (Bi-LSTM), and two different real-world databases have been used in experimental evaluations. Chang et al. (2020) proposed a hybrid model that exploits stacking ensemble learning model to integrate various machine learning models for improving the air pollution forecasting accuracy.

These methods based on deep learning can better predict and analyze the concentration of air pollutants, but most of them do not consider the correlation between pollutants. On the other hand, the prediction is basically based on independent station, few models show and emphasize the importance of regional air
quality analysis and prediction. In recent years, some scholars have established a spatiotemporal model
based on joint prediction (Qi et al. 2018; Xu and Yoneda 2021; Wang et al. 2021; Zhang et al. 2022).
Zhao et al. (2019) constructed a three-dimensional data structure called relevance data cube and utilized
a clustering algorithm, time sliding windows and correlation analysis of factors, so as to systematically
analyze closely connected regions and predict multiple locations at the same time, but few studies have
conducted in-depth analysis on the relationship among different stations in the region.

The main contributions of this paper are as follows:
1. The influence of meteorological factors on pollutants and the correlation between pollutants are fully
analyzed. A deep learning model based on CNN-BiLSTM is established to predict the concentration of
air pollutants. Feature maps are generated sequentially from the input time series data using a sliding
window pattern, and different size feature maps based on different input databases are discussed.
2. The model is optimized by using the Bayesian optimization algorithm to obtain the optimal
combination of hyper-parameters, which effectively improves the accuracy of prediction and
generalization ability.
3. Fully considering the relationship between stations and meteorological data, especially the influence
of wind direction on regional prediction data, a new multi stations collaborative prediction scheme is
proposed to further improve the prediction effect.

Data Description

In this section, the datasets used in the article are introduced and the data processing is described.

Data Source

Our study collected the hourly historical air pollutants data of five air quality monitoring stations in
Tianjin, China, from 1 January 2019 to 30 June 2022. Air quality data comes from the national urban air
quality real-time publishing platform of China Environmental Monitoring Center. The air quality data
include six pollutant concentrations (µg/m³), namely PM2.5, PM10, O₃, SO₂, NO₂, and CO. We selected
the five monitoring stations around the city, numbered 1015A, 1017A, 1018A, 1019A and 1021A.
Meteorological data comes from the public FTP server of NCDC (National Climatic Data Center), which
include five factors: air temperature (TEM; °C), dew point temperature (DPT, °C), sea level pressure
(SLP; hPa), wind direction (WD, °) and wind speed (WS; m/s).

Data Processing

Meteorological Data

1) Meteorological data comes from seven data resources at 0:00, 3:00, 6:00, 9:00, 12:00, 18:00 and
21:00 every day. The time interval is basically three hours, but the interval between 12:00 and 18:00 is
six hours, so 15:00 data is filled with the average value of 12:00 and 18:00 data.
2) Some meteorological data is missing. In 2019, for example, there are 365 days. After completing
the first step, a total of 2920 sets of data are required, but only 2614 sets of data are provided. These data
need to be processed, otherwise it will affect the accuracy of the model established later. To solve this
problem, weighted sum data is used to fill in the vacancy. The calculation method is as follows:
\[
data(n) = 0.7 \times \frac{1}{2+k} \sum_{i=\max(n-2)}^{n+2} data(i) + 0.3 \times data(n-8)
\]  
(1)

where \( data(n) \) is the sampling data at the \( n \)th time, \( data(n-8) \) is the data at the same time of the previous day, and \( k \) is the number of non-null values in \( data(n+1) \) and \( data(n+2) \).

3) Abnormal data processing. According to the characteristics of meteorological data, the change of short-term data is relatively stable and has a certain range. For example, the temperature in this region is usually between -200 and 400 Degrees Celsius. If the data value at a certain time deviates from the normal distribution of the data, that is, it differs greatly from most of the observed data, the value will be determined as abnormal. Isolated Forest Algorithm (IF) is used as a method to process abnormal data (Ding and Fei 2013), and the correction method of abnormal data is the same as formula (1) in the previous step.

Air Quality Monitoring Data

1) Some or all of the observed data is missing for a period of time due to equipment debugging and maintenance. Data processing is the same as 1) in meteorological data processing.

2) Affected by some accidental factors near the monitoring station, some observed values have large deviations. In this case, refer to 3) in meteorological data processing, use the IF to detect abnormal data and complete data processing.

3) Since the meteorological data is monitored every three hours, while the air quality data is hourly, the data do not match and cannot be analyzed in the next step. Therefore, the average value of air quality data is obtained every three hours. The subsequent prediction analysis is also carried out on a 3-hour time scale.

CNN-BiLSTM Model for Air Pollutants Prediction

CNN Model

Convolutional neural network (CNN) is a kind of deep feedforward neural network with convolution operation as its core. Complete two-dimensional convolutional neural network usually includes input layer, convolutional layer, pooling layer, full connection layer, output layer and other modules (Yamashita et al. 2018). The feature extraction of convolutional neural network mainly depends on the convolutional kernel in the convolutional layer. Mathematically, the feature value at location \((i, j)\) in the \(k\)th convolutional kernel of \(l\)th layer is calculated by:

\[
y'_{i,j,k} = \sigma \left( W'_k \otimes x'_{i,j} + b'_k \right)
\]  
(2)

Where, \( x'_{i,j} \) is the input at location \((i, j)\) of the \(l\)th layer, \( \otimes \) is the convolutional operation, \( W'_k \) and \( b'_k \) are respectively the weight coefficient and bias parameter of the \(k\)th convolutional kernel of the \(l\)th layer, \( \sigma \) is the activation function of neurons (Gu et al. 2018).

The convolutional kernel (also known as filter) scans two-dimensional data and performs convolutional operation by means of gradual translation to extract data features. As shown in Figure 1, the blue area is
the calculation area of the convolutional kernel, the green part is the convolution kernel, and the red part
is the calculation result of the convolutional kernel.

![Fig.1 Convolutional kernel calculation process](image)

The pooling layer used in this paper is the maximum pooling layer. The pooling operation reduces the
dimensions of the extracted features to avoid over fitting the model and retain the main features. The
maximum pooling layer takes the characteristic matrix obtained from the previous convolution layer as
the input, slides a pooling window on this matrix, takes the maximum value of the pooling window in
each slide, and outputs a more expressive characteristic matrix.

**BiLSTM Model**

Long short-term memory (LSTM)(He et al. 2019) is a kind of recurrent neural network (RNN), which
solves the problem of gradient disappearance during RNN training, and has good performance in dealing
with long time series. LSTM is composed of multiple identical cell structures, and each cell structure is
composed of forgetting gate, input gate and output gate. Its architecture is shown in Figure 2.

The functions of each door of LSTM unit are as follows. Forgetting gate: control the degree of
information forgetting. Input gate: control the amount of information that input into memory unit. Output
Gate: controls the weight of the output unit.

![Fig. 2 The architecture of LSTM](image)

The conventional LSTM network guided the information transmission through the forget gate $f_t$, the
input gate $i_t$, and the output gate $o_t$. $x_t$, $S_t$, and $h_t$ represent the input, internal and external
states at time $t$, respectively. The detailed information transformation in the LSTM block are as
follows:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$  \hspace{1cm} (3)
\[ i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (4) \]
\[ g_t = \tanh(W_g[h_{t-1}, x_t] + b_g) \quad (5) \]
\[ S_t = f_t \cdot S_{t-1} + i_t \cdot g_t \quad (6) \]
\[ o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (7) \]
\[ h_t = o_t \cdot \tanh(S_t) \quad (8) \]

where \( W \) and \( b \) represent the weights and bias between the transmission layers of the LSTM block, \( \sigma \) (Sigmoid) and \( \tanh \) are the activation functions, and the formulas are as follows:

\[ \sigma = \frac{1}{1 + e^{-x}} \quad (9) \]
\[ \tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (10) \]

The standard LSTM network structure only has forward propagation operation. Lack of logic between sequences. The bidirectional long short memory (BiLSTM) network is based on the LSTM (Xu et al. 2019), using the known time series and reverse position series, through forward and back propagation bidirectional operations, strengthen the ability to extract the original sequence features and improve the accuracy of the model output results. The final output of the BiLSTM network is the sum of the forward and back propagation LSTM outputs. The structure of the BiLSTM network can be seen in Figure 4.

**CNN-BiLSTM Model**

The CNN-BiLSTM network hybrid model proposed in this paper takes the time series as the input of the network. The two independent time series of historical air quality data (pollutant concentration) and meteorological data are combined into a new time series datasets. Each time series has \( m \) variables (see Section 4 for the value of \( m \)). Feature maps are generated sequentially from the input time series datasets using a sliding window pattern. The sliding window width is set to 8 records (i.e., one day’s data) and the step size is 1, so the size of the feature graph is \( m \times 8 \). The feature maps are also arranged in time series, as shown in Figure 3. Where \( t \) is the time scale of the axis, \( T \) is a specific moment, and \( T+n \) refers to the time 3xn hours after time \( T \).

![Fig. 3 Structure of input data in CNN-biLSTM net model](image)

The structure of the CNN-BiLSTM network hybrid model proposed is shown in Figure 4, where ReLU is the activation function, and the expression is \( f(x) = \max(0, x) \).

The model is mainly composed of two parts, the two-layer CNN is mainly responsible for feature extraction, while the LSTM network for prediction. The input data is extracted by convolutional layer and pooling layer. After the output results are flattened, further feature learning is carried out through
three layers of LSTM layer in turn. Finally, the full connection layer outputs the value of the \( k \) predictive value in the future, and \( k \) is the step of backward prediction.

![CNN-BiLSTM model structure](image)

**Bayesian Optimization**

Generally, in the process of model training, it is necessary to optimize the hyperparameters and select a group of optimal hyperparameters to improve the performance and effectiveness of prediction. In random search, the combination of random parameters is sampled according to the statistical distribution given by the user, which may not find the optimal hyperparameter. Bayesian optimization can make full use of historical optimization information, reduce unnecessary objective function evaluation, and improve the efficiency of parameter search (Abbasimehr and Paki 2021).

In the Bayesian optimization of CNN-BiLSTM, the network with different parameter combinations are taken as independent variables, and mean error (MAE) is taken as the output of the Bayesian framework. Taking the number of input types \( m=9 \) as an example, the eight parameters of the model are optimized using Bayesian optimization method. Table 1 shows the scope of parameter search and the optimal set of values after applying Bayesian optimization.

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Search Space</th>
<th>Optimal Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FilterSize1</td>
<td>[2,2], [3,3], [4,4]</td>
<td>[4,4]</td>
</tr>
<tr>
<td>Numfliters1</td>
<td>8–64</td>
<td>48</td>
</tr>
<tr>
<td>FilterSize2</td>
<td>[2,2], [3,3], [4,4]</td>
<td>[3,3]</td>
</tr>
<tr>
<td>Numfliters2</td>
<td>8–64</td>
<td>56</td>
</tr>
<tr>
<td>Learning rate</td>
<td>[1e-5, 1e-3, 1e-1]</td>
<td>1e-1</td>
</tr>
<tr>
<td>NumHiddenUnit1</td>
<td>[50,100,200]</td>
<td>200</td>
</tr>
<tr>
<td>NumHiddenUnit2</td>
<td>[50,100,200]</td>
<td>100</td>
</tr>
<tr>
<td>NumHiddenUnit2</td>
<td>[50,100,200]</td>
<td>100</td>
</tr>
</tbody>
</table>

**Prediction Results and Analysis**
Selection of Training Set

As mentioned earlier, the data are collected from 1 January 2019 to 30 June 2022, eight times a day, so a total of 10,216 sets of data for 1277 days. A total of 8,768 datasets for three years from 2018 to 2021 are used for training, and 1,448 datasets for six months in 2022 are used for testing. First, we select one of the stations, 1017A, for the prediction study. As mentioned in the structure of the CNN-BiLSTM network, the input sliding window width is set to 8 records (i.e. data of a day), the size of the feature graph is $m \times 8$, where $m$ is the type number of input data. In order to get better prediction results, we study different types of input datasets.

1) Meteorological data and historical data of predicted pollutant are chosen as input. In this case, the number of input types $m=6$. For example, if the concentration of PM2.5 is predicted, TEM, DPT, SLP, WD, WS and PM2.5 are used as input, and if SO2 is predicted, then the historical data of SO2 will be used as input.

2) All pollutants and the meteorological data are chosen as input. Pearson correlation coefficient was used to analyze the correlation among the features, and the correlation degree between the features was obtained, as shown in Figure 5.

![Fig.5 Pearson correlation coefficient between meteorological data and pollutants](image)

From the correlation analysis of each factor, it can be seen that the pollutant concentrations have greater correlation with air temperature, dew point temperature and air pressure, but a smaller correlation with wind speed and direction. On the other hand, there are strong correlations between different pollutants, which are basically positive, only O$_3$ has a negative correlation with other pollutants. According to the analysis results, wind speed and wind direction are ignored when the prediction model is established, and the other 9 variables are taken as inputs, thus reducing the input dimension, and $m=9$ in this case.

In order to effectively evaluate the prediction results, coefficient of determination ($R^2$), root mean square error (RMSE), and mean absolute error (MAE) are used as the evaluation indexes of the model. The calculation formula are as follows:
\[ R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \]  

(11)

\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \]  

(12)

\[ \text{MAE} = \frac{1}{n} \sum_{i=1}^{n}|y_i - \hat{y}_i| \]  

(13)

Where \( y_i \) and \( \hat{y}_i \) represent the true value and the predicted value of the \( i \)th sample respectively, \( n \) represents the number of samples, and \( \bar{y} \) is the mean of the true values of all samples.

The prediction results of 1,448 data from January to June 2022 are shown in Figure 6. According to the evaluation indicators, the combination of meteorological data and all historical pollutant data (\( m=9 \)) has a better prediction effect than the combination of meteorological data and predicted pollutant data (\( m=6 \)). On the other hand, the prediction effect of each pollutant is obviously different. The prediction accuracy of \( O_3 \) is the highest (\( R^2 \) is 0.85939 and 0.86082 respectively), followed by PM2.5 (\( R^2 \) is 0.81532 and 0.83761 respectively), and the worst is SO\(_2\) (\( R^2 \) is 0.57323 and 0.61518 respectively). Figure 7 shows the three-year variation of six pollutant concentrations. It can be seen that the concentrations of all pollutants change with seasons, among which \( O_3 \) shows a more significant seasonal effect than other, while SO\(_2\) and CO are weaker, which is consistent with our prediction results.
Fig. 6 Prediction results of CNN-BiLSTM model under different input types (m=6 and m=9)

Fig. 7 Three years data graph of pollutant concentration

Regional Collaborative Prediction

It should be noted that meteorological factors, geographical factors and their interaction form the complex relationship and constraint of spatial and temporal characteristics of regional air quality. Regional analysis allows for full consideration of the interaction of air quality systems.

Multi-station Collaborative Prediction Model
Figure 8 shows the location of the 5 air quality monitoring stations. Their location coordinates are: 1015A (117°1444'E, 39°1661'N), 1017A (117°327'E, 39°1082'N), 1018A (117°202'E, 39°297'N), 1019A (117°1837'E, 39°213'N), 1021A (117°307'E, 39°0877'N). For station 1017A, the distances between 1015A, 1018A, 1019A and 1021A are 8.8 km, 3.6 km, 12.7 km and 6.5 km respectively. Through Pearson correlation analysis, the five sites are highly correlated with each other (R^2 > 0.95).

If the observation data of the remaining four stations are added to the input data, taking PM2.5 of 1017A as an example, add the historical observation data of PM2.5 of the other four stations into the input datasets, then the size of the input data feature graph becomes 13 * 8, in which the number of input data types m = 13. In this model, ten predictions were made for each pollutant, and the mean values were recorded in Table 4.

Collaborative Prediction Correction Model

Further analyze the correlation between stations. The wind direction is divided into 8 areas ranging from 0° to 360° per 45°. That is, 0°~45° is area 1, 45°~90° is area 2, … 15°~360° is area 8.

We study the effect of wind direction on 1017A station from 2019 to 2021, and here we define the concept of strongly correlated station (SCS). At each sampling time, the station whose pollutant concentration is closest to 1017A station is taken as SCS at that time. For example, Table 2 shows the observed data of SO_2 at the five stations on January 1, 2019. According to the definition, SCS of SO_2 at each moment can be obtained, as shown in the last column.

<table>
<thead>
<tr>
<th>time</th>
<th>1015A</th>
<th>1018A</th>
<th>1019A</th>
<th>1021A</th>
<th>1017A</th>
<th>SCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:00</td>
<td>28</td>
<td>44</td>
<td>29</td>
<td>35</td>
<td>36</td>
<td>1021A</td>
</tr>
<tr>
<td>3:00</td>
<td>44.67</td>
<td>38.67</td>
<td>32.33</td>
<td>33.67</td>
<td>33.67</td>
<td>1021A</td>
</tr>
<tr>
<td>6:00</td>
<td>36.67</td>
<td>34</td>
<td>31.67</td>
<td>30.33</td>
<td>33</td>
<td>1018A</td>
</tr>
<tr>
<td>9:00</td>
<td>41.33</td>
<td>36</td>
<td>33.33</td>
<td>31</td>
<td>27</td>
<td>1021A</td>
</tr>
<tr>
<td>12:00</td>
<td>70</td>
<td>27.33</td>
<td>23</td>
<td>32</td>
<td>23.33</td>
<td>1019A</td>
</tr>
<tr>
<td>15:00</td>
<td>41.67</td>
<td>21</td>
<td>27</td>
<td>23.5</td>
<td>16.33</td>
<td>1018A</td>
</tr>
<tr>
<td>18:00</td>
<td>13.33</td>
<td>14.67</td>
<td>31</td>
<td>15</td>
<td>9.33</td>
<td>1015A</td>
</tr>
<tr>
<td>21:00</td>
<td>17.33</td>
<td>11.33</td>
<td>11</td>
<td>12</td>
<td>5</td>
<td>1019A</td>
</tr>
</tbody>
</table>
Further statistics are made on the wind direction at each time and the SCS at the next time. We count the data of all pollutants in three years. Figure 9 shows the data trend of different wind directions and the SCS at the next moment.

![Fig. 9 Relationship between wind direction and SCS](image)

Figure 10 shows the proportion of SCS for four stations in different wind directions, and Table 3 shows the sum of SCS data of all pollutants in different wind directions.

![Fig.10 Proportion of SCS for four stations](image)

<table>
<thead>
<tr>
<th>stations</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1015A</td>
<td>1862</td>
<td>565</td>
<td>1608</td>
<td>1686</td>
<td>2294</td>
<td>1059</td>
<td>964</td>
<td>1643</td>
<td>11681</td>
</tr>
<tr>
<td>1018A</td>
<td>2582</td>
<td>822</td>
<td>2280</td>
<td>2702</td>
<td>3196</td>
<td>1756</td>
<td>1260</td>
<td>2600</td>
<td>17198</td>
</tr>
<tr>
<td>1019A</td>
<td>1415</td>
<td>381</td>
<td>1241</td>
<td>1279</td>
<td>1707</td>
<td>769</td>
<td>529</td>
<td>2032</td>
<td>9353</td>
</tr>
<tr>
<td>1021A</td>
<td>2072</td>
<td>620</td>
<td>2377</td>
<td>2532</td>
<td>2575</td>
<td>1242</td>
<td>887</td>
<td>2071</td>
<td>14376</td>
</tr>
</tbody>
</table>

According to Table 3, 1018A is the station with the largest number of SCS, with 17198 times, followed by 1021A, 14376 times, and 1019A is the least, with only 9353 times. Further consider the relationship between stations, the nearest to 1017A is 1018A, and the farthest is 1019A. Obviously, the closer the station is, the stronger the correlation and the more obvious the interaction between stations is.
We further analyze the effect of wind direction. In Figure 9, it can be seen that the wind direction in Tianjin is more from the south and north, and less from the east and west. In addition, combined with the proportion of stations under each wind direction in Figure 10, it can be seen that in the area 3, the value of 1021A is better than that of other stations, and in area 5 and 6, 1018A had a clear advantage. In area 7 and 8, the wind direction is northwest, and the values of 1015A and 1019A have relatively large increases. Therefore, we believe that there is a correlation between current wind source direction and pollutant data at the next moment. Based on this, we construct an improved collaborative prediction model, as shown in the following:

\[
\text{predict}(i) = \left(1 - \frac{1}{(\text{dis}(WSS))^q}\right) \text{predict}(i) + \frac{1}{(\text{dis}(WSS))^q} \times \text{observe}_{WSS}(i-1) \tag{14}
\]

Where \(\text{predict}(i)\) is the predicted value of station 1017A at time \(i\), \(WSS\) is the wind source station, and \(\text{observe}_{WSS}(i-1)\) is the observed value of the wind direction source station at the previous time. \(\text{dis}(WSS)\) is the distance from the wind direction source station to 1017A. \(q\) is to weight the wind direction source data, with a lot if debugging, we take \(q = 3/2\).

So far, this paper has studied three different input modes for the proposed CNN-BiLSTM network. Mode1: prediction based on reduced dimension meteorological and air quality data (\(m=9\)). Mode2: collaborative prediction (\(m=13\)). Mode3: collaborative prediction correction model. Prediction research has been carried out for different modes, as shown in Table 4. Each mode has been run for ten times, and all evaluation indicators in the table are the average of the results of ten times of operation.

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Index</th>
<th>Mode1</th>
<th>Mode2</th>
<th>Mode3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM2.5</td>
<td>R²</td>
<td>0.8276</td>
<td>0.8436</td>
<td>0.8494</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>12.523</td>
<td>11.9978</td>
<td>11.7669</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>8.5870</td>
<td>7.7165</td>
<td>7.6503</td>
</tr>
<tr>
<td>PM10</td>
<td>R²</td>
<td>0.8245</td>
<td>0.8428</td>
<td>0.8411</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>19.9586</td>
<td>19.1638</td>
<td>19.3332</td>
</tr>
<tr>
<td>SO2</td>
<td>R²</td>
<td>0.6152</td>
<td>0.6433</td>
<td>0.6474</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>2.0691</td>
<td>2.0186</td>
<td>1.9981</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>1.4998</td>
<td>1.4187</td>
<td>1.4161</td>
</tr>
<tr>
<td>NO2</td>
<td>R²</td>
<td>0.7674</td>
<td>0.7766</td>
<td>0.7781</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>9.7363</td>
<td>9.5347</td>
<td>9.4853</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>7.2048</td>
<td>6.9586</td>
<td>6.811</td>
</tr>
<tr>
<td>O3</td>
<td>R²</td>
<td>0.8731</td>
<td>0.8811</td>
<td>0.8849</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>11.0456</td>
<td>14.1436</td>
<td>10.9097</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>14.6135</td>
<td>10.3872</td>
<td>10.4909</td>
</tr>
<tr>
<td>CO</td>
<td>R²</td>
<td>0.7982</td>
<td>0.7919</td>
<td>0.7894</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1515</td>
<td>0.1528</td>
<td>0.1531</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>0.1023</td>
<td>0.1065</td>
<td>0.1087</td>
</tr>
</tbody>
</table>
Further Discussion

Analyze the performance indicators in Table 4. Compared with mode 1 and mode 2, the prediction effect of most pollutants is better in mode 2, only the prediction effect of CO is worse, $R^2$ is decreased from 0.7982 to 0.7919, RMSE and MAE are increased. Therefore, it can be concluded that the regional collaborative prediction can further improve the prediction accuracy. As for CO, more satisfactory results may be obtained by increasing the number of tests.

Further comparison between mode 2 and mode 3 shows that among the six groups of pollutants, the prediction effect of PM2.5, SO$_2$, NO$_2$ and O$_3$ is better in mode 3, while the evaluation indicators of PM10 and CO are more advantageous in mode 2. Different from the above analysis, increasing the number of tests will not change the results, but the results can be changed by adjusting the parameters of equation (14). The parameters given are the best ones we have adjusted many times. We further analyze the main reasons why the improved model has not achieved better expected results. There are some other factors that we haven't taken into account in our correction model which affect pollutant concentrations, such as wind speed and unknown industrial gas emissions, and so on. at the same time, more random interferences are also the limitation of model-driven in dealing with big data problems.

Conclusion

In this paper, a Bayesian optimization-based depth learning model for spatiotemporal tasks is proposed for air quality prediction. By learning meteorological data, historical data of air quality and historical data of multiple stations in the region, this method can extract relevant pollutant time series characteristics, and predict the pollutant concentration at the subsequent time. Through the analysis of historical data and its correlation, three different datasets input modes are proposed. The simulation results show that the model can implicitly learn the common rules of historical data and multiple site data, and achieve better prediction results. At the same time, we also analyze the advantages and disadvantages of different modes.

As for future work, we would like to obtain more auxiliary information, such as industrial gas emissions and other factors that never considered in traditional prediction models. At the same time, the regional collaborative prediction model will be further discussed and studied. In addition, the prediction effect of SO$_2$ is the worst among pollutants, which can further study the influencing factors of the pollutants and improve the prediction accuracy.

Statements and Declarations

Ethical Approval: Not applicable

Consent to Participate: Not applicable

Consent to Publish: Not applicable

Funding: This research is partly funded by the National Natural Science Foundation of China (No. 71602143).

Conflicts of Interest: The authors declare no conflict of interest.
Author Contributions: Conceptualization: [Yanan Lu, Kun Li]; Methodology: [Yanan Lu, Kun Li]; Software: [Yanan Lu]; Validation: [Yanan Lu, Kun Li]; Writing—original draft preparation: [Yanan Lu]; Writing—review and editing [Kun Li]. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Availability of data and materials: The hourly historical air quality data came from the National Meteorological Science Data Center (国家气象信息中心-中国气象数据网 (cma.cn)) (accessed on 4 October 2022)

References


