中国研究生创新实践系列大赛

“华为杯”第十八届中国研究生

数学建模竞赛

Research on secondary modeling and related issues of air quality forecasting

Abstract:

Today, when atmospheric environmental pollution has an important impact on the human body and the ecological environment, it is very necessary to take corresponding pollution prevention and control measures. By establishing an air quality forecast model, knowing the possible air pollution processes in advance and taking corresponding control measures is one of the effective methods to reduce the harm caused by air pollution to human health and improve ambient air quality. Currently, the WRF-CMAQ simulation system is commonly used in my country to forecast air quality, but it is limited by the uncertainty of the simulated meteorological fields and emission inventories, as well as the incomplete understanding of the generation mechanism of pollutants including ozone, and the atmospheric physics and chemistry. There are also flaws in the mechanism analysis, which results in a low hit rate for primary pollutants and air quality levels, and the results of the WRF-CMAQ forecast model are not ideal. Therefore, based on the simulation results of primary forecast models such as WRF-CMAQ, more data sources are combined for remodeling to improve the accuracy of forecasts.

For question 1, there is no abnormal data in the daily measured data of monitoring point A from August 25 to August 28, 2020. You can directly calculate the monitoring point A from August 25 to August 2020 according to the method in the appendix. The measured AQI and primary pollutants were measured every day on the 28th. The results are shown in Tables 3-4 and 3-6.

For question 2, collect and record the data in Appendix 1 and perform data preprocessing on the records. Mainly to deal with data anomalies, the following processing methods are adopted: When part or all of the measured data is missing: the missing value of the variable value is taken as the time before and after Average value; affected by some accidental factors, the value of the measured data at a certain hour (a certain day) deviates from the normal distribution of the data: if it exceeds

the normal variable distribution range, it is taken as the corresponding minimum or maximum value; the 3σ criterion is used to remove the abnormality Data culling. Based on the variable values of data preprocessing, the corresponding AQI value for each day is calculated according to the formula. According to its attributes, the clustering method k-means based on division is used to cluster the meteorological conditions according to the degree of influence on the pollutant concentration. Visualize the category results; reasonably classify meteorological conditions based on the ranking of AQI results, and describe the characteristics of various meteorological conditions; cluster the overall data multiple times, and finally choose to divide it into three categories based on the clustering results. The cluster center points are as shown in the table As shown in 4-1, the characteristics of various meteorological conditions are explained.

For question 3, a prediction framework based on BP neural network was established based on the representative and independent meteorological conditions selected based on the impact of previous meteorological conditions on pollutant concentrations. This model builds a neural network framework to input primary forecast data and real-time monitoring data to complete the prediction of air quality forecast, and continuously adjusts it during the process of adding updated weather quality data. This model is suitable for predictive modeling of data updates. After a series of tests, the secondary forecast mathematical model is applicable to the three monitoring points A, B, and C at the same time, and the AQI forecast value in the forecast result of the established BP neural network secondary forecast model has a smaller error than the forecast value of the primary modeling. , the prediction accuracy of primary pollutants is higher. In addition, the results of the secondary modeling model for air quality prediction were compared with the real values and compared with other models to verify the effectiveness of the model. And use this model to predict the single-day concentration values of 6 conventional pollutants at monitoring points A, B, and C from July 13 to July 15, 2021, and provide the corresponding AQI and primary pollutant results, as shown in Table 5 -3 is shown.

For question 4, based on the BP neural network prediction framework established in question 3, the wind direction and the positions between monitoring points A, A1, A2, and A3 were added to improve the model to construct a collaborative forecast model. Genetic algorithms were used to use real number coding and simulated binary Crossover and polynomial mutation were used to improve the BP neural network prediction model, and the use of the test set was also changed. The pollutant concentration and meteorological information tested at four different monitoring points in the same time period were taken as the training set. Train the neural network and the results are shown in Table 6-1.

Keywords: air quality forecast, primary pollutants, AQI, k-means clustering, quadratic modeling, BP neural network, genetic algorithm.

Context

1. Problem Restatement

The spatiotemporal variation of air pollutant emissions and the corresponding changes in meteorological conditions can have both short-term and long-term impacts on human health. Exceeding pollutant concentration limits poses significant harm to public health and has a profound effect on ecosystems, the environment, transportation, and other human societal activities. An effective air quality forecasting system helps humans monitor future pollutant concentration levels, formulate appropriate prevention and control strategies, and provide accurate predictions of pollutant concentration levels in advance. This enables the spatiotemporal variation of pollutant emissions and the corresponding meteorological changes to be anticipated, mitigating their short-term and long-term impacts on human health, and reducing the harmful effects of pollutant concentration excesses on public health, ecosystems, the environment, transportation, and other human activities【2】.

The commonly used WRF-CMAQ modeling system (hereinafter referred to as the WRF-CMAQ model) is employed for air quality forecasting.

The WRF-CMAQ model mainly consists of two components: WRF and CMAQ. WRF is a mesoscale numerical weather prediction system used to provide the necessary meteorological data for CMAQ. CMAQ is a three-dimensional Eulerian atmospheric chemistry and transport modeling system that simulates the changes in pollutants based on meteorological information from WRF and the emission inventory within the domain. It uses the principles of physical and chemical reactions to model the variation processes of pollutants, providing forecast results for specific time points or periods. The structures of WRF and CMAQ are shown in Figures 1-1 and 1-2,

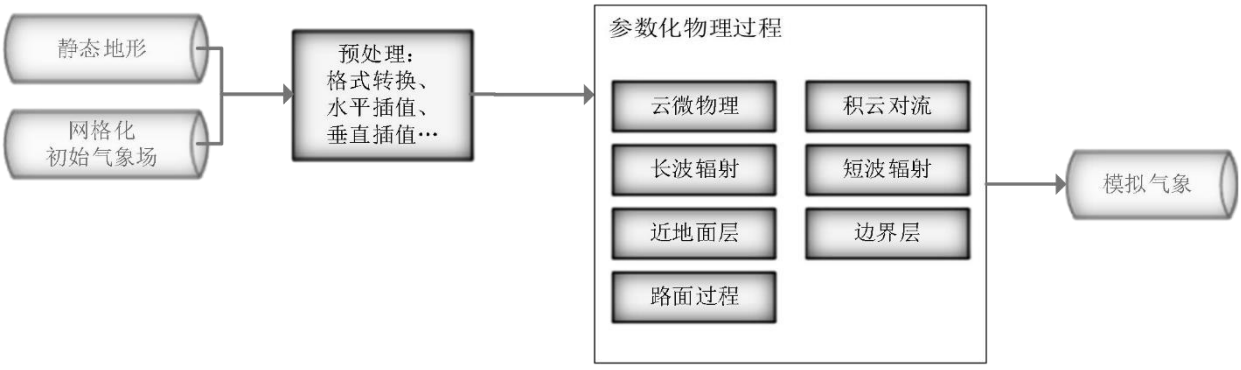


Figure 1-1 Structure of the mesoscale numerical weather prediction system WRF[1]

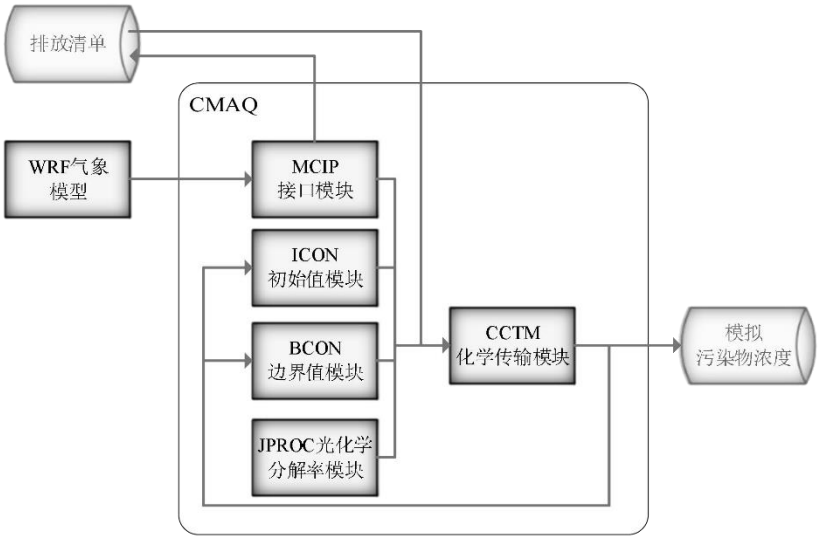


Figure 1-2 Structure of the Air Quality Prediction and Assessment System CMAQ

However, due to the uncertainties in the simulated meteorological fields and emission inventories, as well as the incomplete understanding of the generation mechanisms of pollutants, including ozone, the results of the WRF-CMAQ forecast model are not ideal. Therefore, the concept of secondary modeling is proposed: this refers to re-modeling based on the results of first-stage forecast models such as WRF-CMAQ, incorporating additional data sources to improve forecast accuracy. Since actual meteorological conditions have a significant impact on air quality (e.g., reduced humidity is conducive to ozone formation), and the variation of observed pollutant concentration data provides valuable reference for air quality forecasting, secondary modeling typically refers to integrating meteorological and pollutant data obtained from air quality monitoring stations to optimize the forecast model.

The relationship between the secondary model and the WRF-CMAQ model is shown in Figure 1-3. The data generated by running the WRF-CMAQ model is referred to as "first-stage forecast data," while the data obtained from air quality monitoring stations is referred to as "observed data." Generally speaking, the correlation between the first-stage forecast data and the observed data is not high, but during the forecasting process, observed data is often used to adjust the first-stage forecast data for better results.

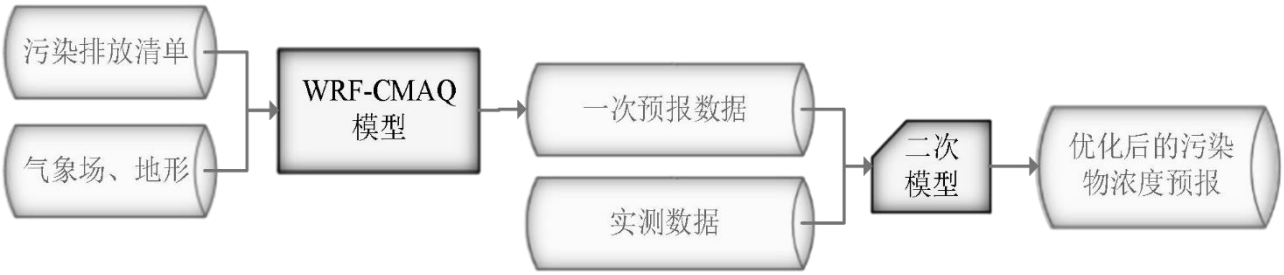


Figure 1-3 WRF-CMAQ Air Quality Forecasting Process Optimized by Secondary Modeling

To perform secondary modeling for predicting the air quality at a given monitoring station for the next three days, the study provides the foundational data for long-term air quality forecasting at the monitoring station, including first-stage forecast data for pollutant concentrations, first-stage forecast data for meteorological conditions, observed meteorological data, and observed pollutant concentration data. The time span for all first-stage forecast data is from July 23, 2020, to July 13, 2021, while the time span for all observed data is from April 16, 2019, to July 13, 2021. The total data volume is in the order of hundreds of thousands (for detailed information, see Appendices 1-3).

It should be noted that:

(1) The daily forecast time is fixed at 7:00 AM, at which point observed data up to and including 7:00 AM of the current day can be obtained, along with first-stage forecast data for the current and prior dates (with the forecast period extending until 11:00 PM on the third day). Observed hourly data after 7:00 AM on the current day, as well as first-stage forecast data for the next day and beyond, are unavailable. For example, in the air quality forecast for July 13 to July 15, 2021, conducted on the morning of July 13, 2021, the available reference observed data range is from 00:00 on April 16, 2019, to 7:00 AM on July 13, 2021, while the model's operational date range is from July 23, 2020, to July 13, 2021.

(2) Due to restrictions on monitoring data access and the functionality of relevant monitoring equipment, some meteorological variables observed data are unavailable.

(3) Since the first-stage forecast tends to be more accurate for dates closer to the forecast date, the secondary forecast is theoretically also more accurate for dates close to the forecast date.

According to the "Ambient Air Quality Standards" (GB3095-2012), there are six common air pollutants used to assess air quality: sulfur dioxide (SO2), nitrogen dioxide (NO2), particulate matter with a diameter less than 10 μm (PM10),

particulate matter with a diameter less than 2.5 μm (PM2.5), ozone (O3), and carbon monoxide (CO). Among them, ozone pollution occurs frequently in many regions across the country, and early warning and prevention of ozone pollution are key tasks for environmental protection agencies. Ozone concentration forecasting is also one of the more difficult tasks among the six pollutants, because: as the only secondary pollutant among the six, ozone is not directly emitted from pollution sources but is generated through a series of chemical and photochemical reactions in the atmosphere (refer to the section on the mechanism of near-surface ozone pollution formation in the appendix). This makes it very challenging to accurately predict ozone concentration changes using the WRF-CMAQ model. Furthermore, existing research both domestically and internationally has not yet reached a general conclusion on the mechanism of ozone formation .

To sum up, according to the requirements of the problem, based on the primary forecast data and measured data (see attachment), the mathematical modeling of air quality forecast was conducted twice, and a model with certain robustness was established to complete the following four problems. Please note that in actual work, the data will be null or outlier (see appendix), so the established model is required to have a certain degree of robustness.

Question 1. Using the data in Annex 1, calculate the daily measured AQI and primary pollutants at monitoring point A from August 25 to August 28, 2020 according to the method in the appendix, and give the results in accordance with the format of the appendix "AQI Calculation result Table".

Question 2. Under the condition of constant discharge of pollutants, when the meteorological conditions of a certain region are conducive to the diffusion or subsidence of pollutants, the AQI of the region will decrease, and vice versa. Using the data in Annex 1, the meteorological conditions are reasonably classified according to the degree of influence on pollutant concentrations, and the characteristics of each type of meteorological conditions are described.

Question 3. Using the data in Attachments 1 and 2, A quadratic prediction mathematical model is established which is applicable to the three monitoring points A, B and C at the same time (the 3linear distance between the two monitoring points is >100km, ignoring mutual influence), and is used to predict the one-day concentration value of six normal pollutants in the next three days. It is required that the maximum relative error of AQI forecast value should be as small as possible, and the prediction accuracy of primary pollutants should be as high as possible. The model is also used to predict the single-day concentration values of six conventional pollutants at the monitoring points A, B and C from July 13 to July 15, 2021, and the corresponding AQI and primary pollutants are calculated. The results are presented according to the format of the appendix "Pollutant Concentration and AQI Prediction Result Table".

Question 4. Pollutant concentrations in adjacent areas are often correlated to a certain extent, and regional collaborative forecasting may improve the accuracy of air quality forecasting. As shown in Figure 4, there are monitoring points A1, A2 and A3 in the area adjacent to monitoring point A. A collaborative prediction model containing four monitoring points A, A1, A2 and A3 is established using the data in Attachments 1 and 3. It is required that the maximum relative error of AQI forecast value in the pre-measurement results of the secondary model should be as small as possible, and the prediction accuracy of primary pollutants should be as high as possible. The model was used to predict the single-day concentration values of six conventional pollutants at the monitoring points A, A1, A2 and A3 from July 13 to July 15, 2021, and the corresponding AQI and primary pollutants were calculated. The results were presented in the paper according to the format of "Pollutant Concentration and AQI Prediction Results Table" in the appendix. And discuss: Compared with the model in question 3, can the collaborative prediction model improve the prediction accuracy of pollutant concentration at monitoring point A? And explain why.

2. Model assumptions and description of key symbols

(1) Assume that the data collected by the real-time monitoring of the air quality detection station can well reflect the air quality situation at this moment;

(2) Assume that the normal data collected during the equipment detection process are accurate;

(3) Assume that the concentration of primary pollutants found through the data can reflect the air quality level;

(4) Assume that every time the value of the primary pollutant is changed, the air quality conditions and conditions will change correspondingly and timely.

Symbol specification:

Symbol specification:

符号	意义
AQI	Air Quality index;
C_{O_3}	Maximum 8-hour moving average for ozone (O_3);
$IAQI_P$	The air quality index of pollutant P, the result is integer;
C_P	The mass concentration of pollutant P;
BP_{Hi}	The high value of the pollutant concentration limit similar to C_P;

BP_{Lo}	The lower limit value of pollutant concentration similar to C_P;
$IAQI_{Hi}$	The air quality sub-index corresponding to BP_{Hi};
$IAQI_{Lo}$	Air quality sub-index corresponding to BP_{Lo}.

3. Analysis and solution of problem 1

3.1 Analysis of problem 1

The topic provides the basic data of long-term air quality forecast of the monitoring points, including the one-time forecast data of pollutant concentration, the one-time forecast data of meteorological data, the measured meteorological data and the measured pollutant concentration data. Among them, the time span of all the one-time forecast data is from July 23, 2020 to July 13, 2021. The time span of all measured data is from April 16, 2019 to July 13, 2021. The first forecast data includes 15 corresponding meteorological parameters of temperature (℃) at 2 meters near the earth, surface temperature (K), humidity (%), wind speed (m/s) at 10 meters near the earth, and atmospheric pressure (Kpa). And the hourly average concentration of six conventional air pollutants used to measure air quality, including carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen oxides (NOx), ozone (O_3) and other gaseous pollutants and particulate pollutants such as inhalable particulate matter (PM_{10}) and fine particulate matter ($PM_{2.5}$). These air pollutants have a potentially negative impact on public life and can even cause a range of health problems [1].

Among them, ozone pollution occurs frequently in many regions of the country, and the early warning and prevention of ozone pollution is the focus of the environmental protection department. The prediction of ozone concentration is also a difficult one among the six pollutants. The reason is that, as the only secondary pollutant among the six pollutants, ozone does not come from the direct emission of pollution sources, but is generated in the atmosphere through a series of chemical and photochemical reactions [3]. The question requires using the data in Annex 1 to calculate the daily measured AQI and primary pollutants at monitoring point A from August 25 to August 28, 2020 according to the method in the Appendix, and the results are presented in the text in the format of the appendix "AQI Calculation Result Table". According to the observation, there are no problems such as missing null value in the parameters needed from August 25 to August 28, 2020, so the corresponding result can be obtained directly through the formula for question 1.

3.2 Solution to problem 1

The maximum 8-hour moving average of ozone (O_3) should be calculated first, because when the maximum 8-hour moving average concentration of ozone (O_3) is higher than 800 μg/m3, or when the concentration of other pollutants is higher than the corresponding limit of IAQI=500, the AQI sub-index will not be calculated.

The maximum 8-hour moving average of ozone (O_3) is the maximum of all 8-hour moving average concentrations between 08:00 and 24:00 in a natural day, where the 8-hour moving average is the arithmetic mean of the average concentration for eight consecutive hours. Its calculation formula is as follows:

$$C_{O_3} = \max_{t=8,9,\ldots,24}\{\frac{1}{8}\sum_{i=t-7}^{t} c_i\} \qquad (3\text{-}1)$$

Where ci is the average pollutant concentration of ozone from $t-1$ to t on a given day. The Air Quality Sub-Index (IAQI) for each pollutant is calculated by the following formula:

$$IAQI_P = \frac{IAQI_{Hi}-IAQI_{Lo}}{BP_{Hi}-BP_{Lo}} \cdot (C_P - BP_{Lo}) + IAQI_{Lo} \qquad (3\text{-}2)$$

The concentration limits of each pollutant item and the corresponding air quality sub-index levels can be seen in table 3-1.

Table 3-1 Air Quality Sub-Index (IAQI) and corresponding pollutant item concentration limits

Index Number	Index or pollutant item	Air quality index and corresponding pollutant concentration limits								Unit
0	Air quality sub-index (IAQI)	0	50	100	150	200	300	400	500	-
1	Carbon monoxide (CO)24 hour average	0	2	4	14	24	36	48	60	mg / m³
2	Sulfur dioxide (SO_2)24 hours average	0	50	150	475	800	1600	2100	2620	μg / m³
3	Nitrogen dioxide (NO_2)24 hour average	0	40	80	180	280	565	750	940	
4	Ozone (O_3) Max. 8 hours sliding flat	0	100	160	215	265	800	-	-	
5	Particle size less than or equal to 10μm(PM_{10}) 24-hour average	0	50	150	250	350	420	500	600	
6	Particle size of 2.5μm or less ($PM_{2.5}$) 24-hour average	0	35	75	115	150	250	350	500	

The Air Quality Index (AQI) is the maximum of each sub-index , i.e

$$\mathrm{AQI} = \max\{\mathrm{IAQI}_1, \mathrm{IAQI}_2, \mathrm{IAQI}_3, \ldots, \mathrm{IAQI}_n\} \quad (3\text{-}3)$$

Where, IAQI1, IAQI2, IAQI3,... IAQIn is the sub-index of each pollutant project. In the question, the calculation of AQI only involves the six pollutants provided in Table 1, so the calculation formula is as follows:

$$\mathrm{AQI} = \max\{\mathrm{IAQI}_{SO_2}, \mathrm{IAQI}_{NO_2}, \mathrm{IAQI}_{PM_{10}}, \mathrm{IAQI}_{PM_{2.5}}, \mathrm{IAQI}_{O_3}, \mathrm{IAQI}_{CO}\} \quad (3\text{-}4)$$

The air quality class range is divided according to the AQI value. The AQI range corresponding to the class is shown in Table 3-2.

Table 3-2 Air quality class and corresponding Air Quality Index (AQI) range

Air quality class	Excellent	Good	Light pollution	Moderate pollution	Heavy pollution	Serious Pollution
Air Quality Index (AQI) range	[0,50]	[51,100]	[101,150]	[151,200]	[201,300]	[301,+∞)

When AQI is less than or equal to 50 (that is, the air quality evaluation is "excellent"), there is no primary pollutant on the day; When AQI is greater than 50, the largest IAQI pollutant is the primary pollutant. If the largest IAQI pollutant is two or more, it is listed as the primary pollutant. A pollutant with IAQI greater than 100 is considered to exceed the standard.

Through Matlab programming, the IAQI of various pollutants calculated by monitoring point A from August 25 to August 28, 2020 according to the daily measured data is shown in Table 3-3.

Table 3-3 IAQI and Air quality levels measured daily at monitoring site A from August 25 to August 28, 2020 (calculated from daily measured data)

Monitoring date	Locations	IAQI computing						Air quality class
		$IAQI_{CO}$	$IAQI_{SO2}$	$IAQI_{NO2}$	$IAQI_{O3}$	$IAQI_{PM10}$	$IAQI_{PM2.5}$	
2020/8/25	Monitoring point A	13	8	15	60	27	16	Good
2020/8/26	Monitoring point A	13	7	20	46	24	15	Excellent
2020/8/27	Monitoring point A	15	7	39	109	37	33	Light pollution
2020/8/28	Monitoring point A	18	8	38	138	47	48	Light pollution

The AQI and primary pollutants calculated by monitoring point A from August 25 to August 28, 2020 based on daily measured data are shown in Table 3-4.

Table 3-4 Daily measured AQI and primary pollutants at monitoring site A from August 25 to August 28, 2020 (calculated from daily measured data)

Monitoring date	Locations	AQI computing	
		AQI	Primary pollutant
2020/8/25	Monitoring point A	60	O_3
2020/8/26	Monitoring point A	46	O_3
2020/8/27	Monitoring point A	109	O_3
2020/8/28	Monitoring point A	138	O_3

The IAQI of various pollutants calculated based on the daily average of hourly measured data for Monitoring Point A from August 25, 2020, to August 28, 2020, is shown in Table 3-5.

Table 3-5: Measured IAQI and Air Quality Levels for Monitoring Point A from August 25, 2020, to August 28, 2020 (calculated from hourly measured data).

Monitoring date	Locations	IAQI computing						Air quality class
		$IAQI_{CO}$	$IAQI_{SO2}$	$IAQI_{NO2}$	$IAQI_{O3}$	$IAQI_{PM10}$	$IAQI_{PM2.5}$	
2020/8/25	Monitoring point A	13	8	15	61	27	16	Good
2020/8/26	Monitoring point A	13	8	21	47	25	15	Excellent
2020/8/27	Monitoring point A	17	7	39	109	38	33	Light pollution
2020/8/28	Monitoring point A	19	9	38	138	48	47	Light pollution

The AQI and primary pollutants of monitoring point A from August 25 to August 28, 2020, based on the average value of hourly measured data and the average value of hourly real measured data, and the results are shown in Table 3-6.

Table 3-6 Monitoring site A Measured AQI and primary pollutants per day from August 25 to August 28, 2020 (hourly measured data calculation)

Monitoring date	Locations	AQI computing	
		AQI	Primary pollutant
2020/8/25	Monitoring point A	61	O_3
2020/8/26	Monitoring point A	47	O_3
2020/8/27	Monitoring point A	109	O_3

2020/8/28	Monitoring point A	138	O_3

According to the results, it can be seen that the AQI calculated according to the daily measured data is very close to the average value of the hourly measured data, and the primary pollutants are the same.

4. Analysis and solution of problem 2

4.1 Analysis of problem two

The construction of air quality prediction model needs the support of relatively accurate measurement data of related variables which conform to the actual situation and the error is controlled within a certain range. However, in the actual monitoring site equipment in the process of data measurement, data recording, data export and so on, there will inevitably be some problems, so that the final data (original data) and the desired good data there is a certain gap. For example, in the process of data measurement, the monitoring site equipment will encounter a variety of complicated actual conditions under different conditions, which will eventually lead to more or less bad data in the original data collected, including continuous or intermittent value missing, value drift (too large or too small) and other situations. Therefore, the scientific and effective preprocessing of bad data is of decisive significance to the construction of air quality prediction model.

For data collected in real time from different sites in the database, some monitoring station devices have issues with data from certain sites. Specifically, some sites only have data for certain time periods, some sites have entirely empty data or partially missing values, and some sites have data exceeding set limits. Thus, it is necessary to process the raw data before using it.

Regarding meteorological data collected by actual monitoring station devices—such as near-surface temperature at 2 meters (° C), surface temperature (K), humidity (%), near-surface wind speed at 10 meters (m/s), and atmospheric pressure (KPa)—these meteorological parameters typically fluctuate. However, within short time frames, they can be considered relatively stable.

4.2 Solution to Question 2

4.2.1 Data preprocessing

The raw concentration data of carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), particulate matter with a diameter of less than 10 μm (PM_{10}), and particulate matter with a diameter of less than 2.5 μm ($PM_{2.5}$) is visualized, as shown in Figure 4-1.

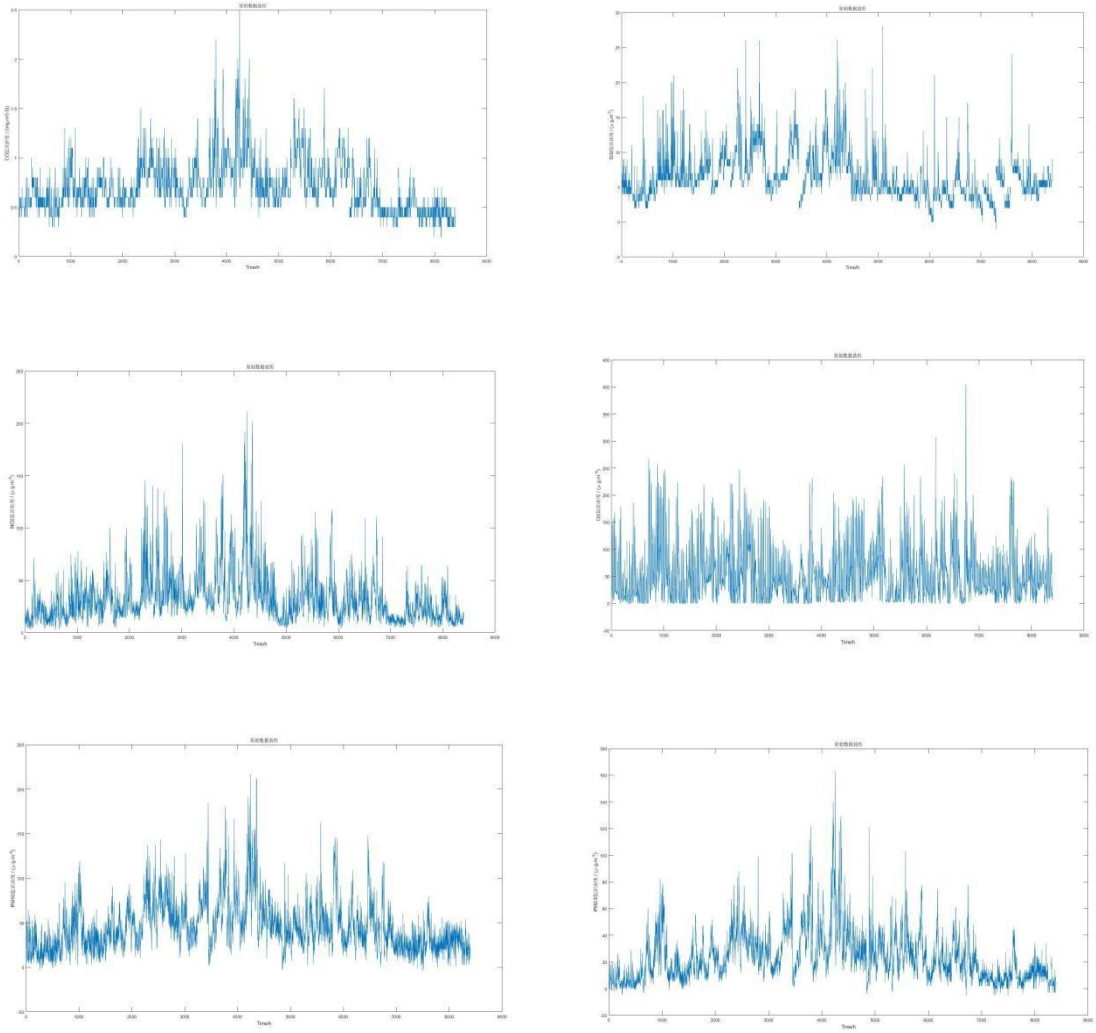


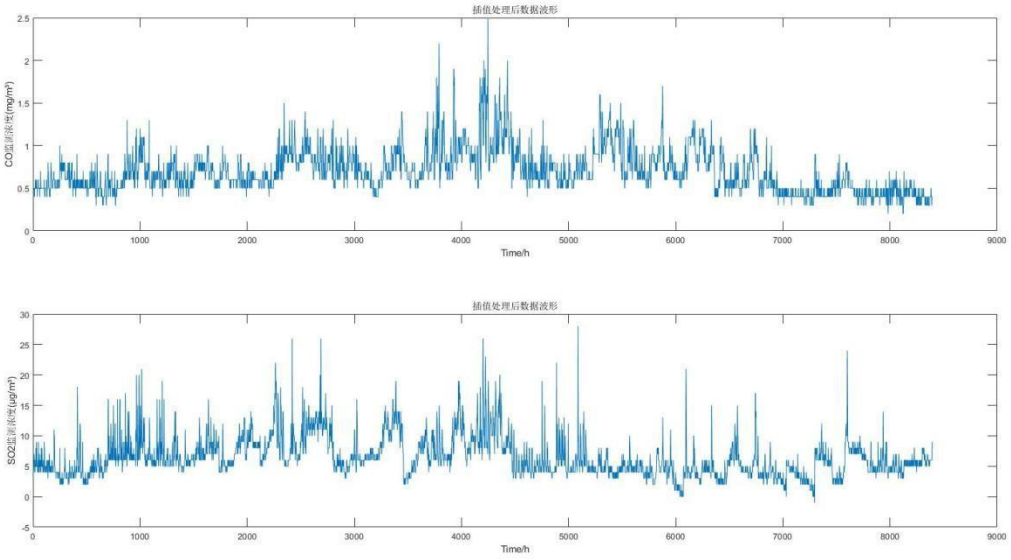
Figure 4-1 The original data of pollutant concentration visualization

Due to factors such as monitoring station equipment adjustments and maintenance, the observed data may have partial or complete gaps over continuous time periods. Additionally, influenced by certain occasional factors around the monitoring site, the observed data may deviate from normal distribution values for a particular hour (or day). The meteorological indicators provided in the project include a total of five items (temperature, humidity, pressure, wind direction, and wind speed), but due to differences in equipment used across different monitoring stations, some indicators may be unavailable at certain sites.

Based on the above analysis, a corresponding analysis and processing will be conducted for each type of problematic data in the raw data from Appendix 1. For the raw data in Appendix 1, outliers should be removed, with the removal criteria set according to the Riad Criterion (3 σ rule). In summary, the following preprocessing steps are proposed for potential anomalies in the raw data:

1. Missing Values:
   During data collection by monitoring station equipment, certain points may experience continuous or intermittent missing values. For such cases, the average of the values within one hour before and after the missing point will be used to fill in the gaps.
2. Values Outside the Range Limits (Exceeding Minimum or Maximum):
   In practical operations, there may be instances where the actual control of variables exceeds the minimum or maximum range specified in the appendix. In these cases, values outside the range will be removed. When calculating the final averaged values across different sites, any data exceeding the limits will be replaced by the minimum or maximum allowable values.
3. Values Outside the 3 σ Range:
   It is reasonable to assume that the variables from monitoring station equipment are continuous and should not exhibit abrupt jumps. However, due to potential issues during the data collection process, certain "jumps" may be observed in the raw data. To address this, these jump values will be excluded when averaging variable data from each station, meaning data points outside the 3 σ range will not be considered in the final averaged value.
4. Inconsistent Units:
   The units of certain variables in the appendix are inconsistent. For example, in one forecast dataset, the unit for pressure is given in KPa, while in the actual data, it is in MBar. The conversion between these units is 1 KPa = 10 MBar.

The processed data and daily AQI values are provided in Appendices 1 and 2, along with the visualized data results, as shown in Figure 4-2.



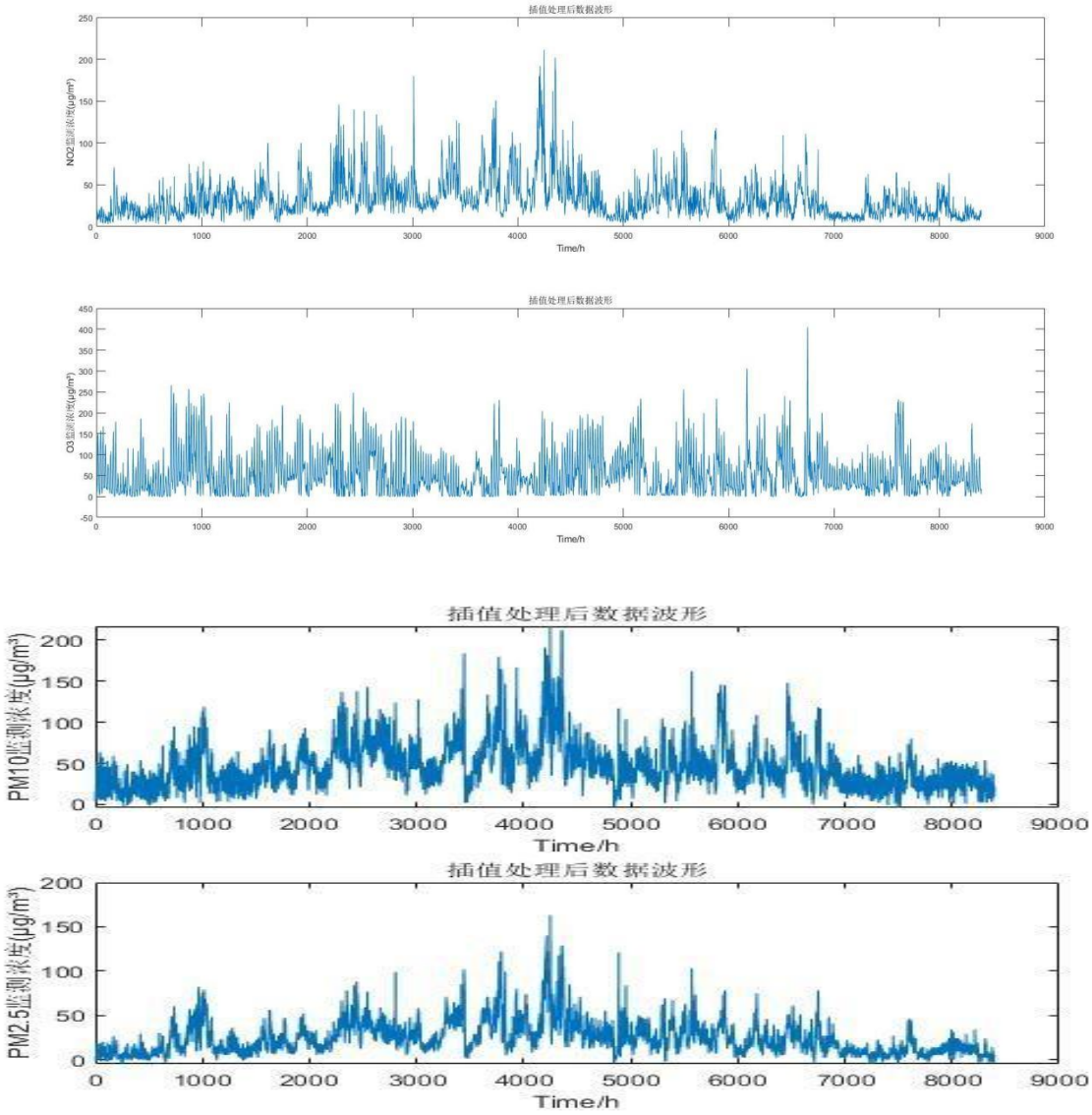


Figure 4-2 Visualization of Pollutant Concentration Data After Processing

4.2.2 Model Principles and Framework

k-means Clustering Algorithm

For a given dataset containing n data points X={X1,X2,…,Xn}, where each Xi∈Rd, and the number of subsets to be generated, the K-Means clustering algorithm organizes the data objects into k partitions C={c1,c2,…,cK}. Each partition represents a cluster with each cluster having a centroid.

The Euclidean distance is used as the similarity and distance criterion. The algorithm calculates the sum of squared distances between the points within each cluster and its centroid.

$$J(c_k) = \sum_{x_{i \in C_k}} \|X_i - \mu_k\|^2 \qquad (4\text{-}1)$$

Clustering is to make $j(c) = \sum_{k=1}^{K} j(C_k)$ smallest

Insides: $d_{ki} = \begin{cases} 1, X_i \in Ci \\ 0, \ X_i \notin Ci \end{cases}$ (4-2)

Clearly, based on the least squares method and the Lagrange principle, the cluster center μk should be the average of the data points in the cluster ck. The K-Means clustering algorithm starts with an initial division into K clusters, then assigns data points to the clusters in a way that minimizes the total squared distance. Since the total squared distance in K-Means clustering tends to decrease as the number of clusters K increases, the total squared distance can only reach its minimum value for a specific number of clusters

（1）Direction of two objects：

We use Euclidean Distance to calculate the formulation is:

$$d(x_i, x_j) = \left(\sum_{K=1}^{P} \left|X_{ik} - X_{jk}\right|^2\right)^{\frac{1}{2}} \quad (4\text{-}3)$$

(2) The criterion function E is commonly used in the K-means algorithm, which is the Sum of Squared Errors (SSE), as the objective function to measure the quality of clustering.

$$\sum_{i=1}^{k} \sum_{x \in C_i} d^2(C_i, x)$$

Here d() represents the distance between two objects.

For the same k, the less SSE represents。The objects within each cluster become more concentrated as k increases. For different values of k, a larger value should result in a smaller SSE

The k-means clustering algorithm involves the following steps:

First, randomly select k objects as the initial cluster centers. Then, for each remaining object, assign it to the cluster whose center is closest (or most similar). Afterward, update the cluster centers by calculating the mean of all objects in each cluster. Repeat the process of assigning objects to the nearest cluster and updating cluster centers until the clusters no longer change. The goal is to minimize the squared error criterion, which ensures that the clusters are as compact and independent as possible.

Input: The number of clusters k and a database X with n data objects.

Output: k clusters that minimize variance.

4.2.3 Clustering Results

The results obtained from clustering using the k-means algorithm in Matlab are as follows:

First, under the condition of constant pollutant emissions, when meteorological conditions in a certain area favor pollutant dispersion or settlement, the AQI in that area decreases, and vice versa, it increases. Among meteorological factors, high temperature, low pressure, low humidity, and high wind speed all promote the removal and dispersion of pollutant concentrations. Any change in these factors can cause fluctuations in ambient air quality, and different pollutants are affected by meteorological conditions to varying degrees. Therefore, based on the impact of each meteorological condition on pollutant concentrations, clustering analysis was performed on meteorological conditions and the calculated corresponding AQI values, with cluster centers rounded to three decimal places

1. Temperature was divided into two categories:
    - High temperature (cluster center: 27.691)
    - Low temperature (cluster center: 17.921)
2. Humidity was divided into two categories:
    - High humidity (cluster center: 72.988)
    - Low humidity (cluster center: 45.970)
3. Pressure was divided into two categories:
    - High pressure (cluster center: 1016.784)
    - Low pressure (cluster center: 1006.346)
4. Wind speed was divided into three categories:
    - Gentle breeze (cluster center: 2.226)
    - Light breeze (cluster center: 1.530)
    - Calm wind (cluster center: 0.994)

Based on an analysis of the clustering results, meteorological conditions can be classified as follows:

1.Temperature Classification:

Category 1: High temperature, with an average temperature of 1.27.691° C, which is conducive to pollutant dispersion. The higher temperature at the lower layer intensifies air movement, increasing turbulence, which aids in the upward transfer of pollutants, thereby reducing pollutant concentration near the surface and improving air quality. The AQI range for this category is (0, 60).

Category 2: Low temperature, with an average temperature of 17.921° C, which is unfavorable for pollutant dispersion. Pollutant concentrations increase as ambient temperature decreases, with an AQI range of (60, 150).

2.Humidity Classification:

Category 1: High humidity, with an average humidity of 72.988%. High humidity helps in pollutant adsorption, especially during precipitation when high water vapor content can increase the mass of PM10 and PM2.5, causing particulate matter to settle. This lowers the concentration of PM10 and PM2.5, aiding pollutant dispersion and reducing concentrations. The AQI range for this category is (0, 65).

Category 2: Low humidity, with an average humidity of 45.970%, which is less conducive to pollutant dispersion, leading to higher pollutant concentrations and an AQI range of (65, 150).

3.Pressure Classification:

Category 1: High pressure, with an average pressure of 1016.784 MBar. Atmospheric pressure is positively correlated with pollutant concentration; under high pressure, the atmospheric layer is stable, airflow is downward, which hinders vertical dispersion of pollutants, leading to an accumulation of pollutant concentration. The AQI range for this category is (55, 150).

Category 2: Low pressure, with an average pressure of 1006.346 MBar, where upward airflow promotes pollutant dispersion, resulting in an AQI range of (0, 55).

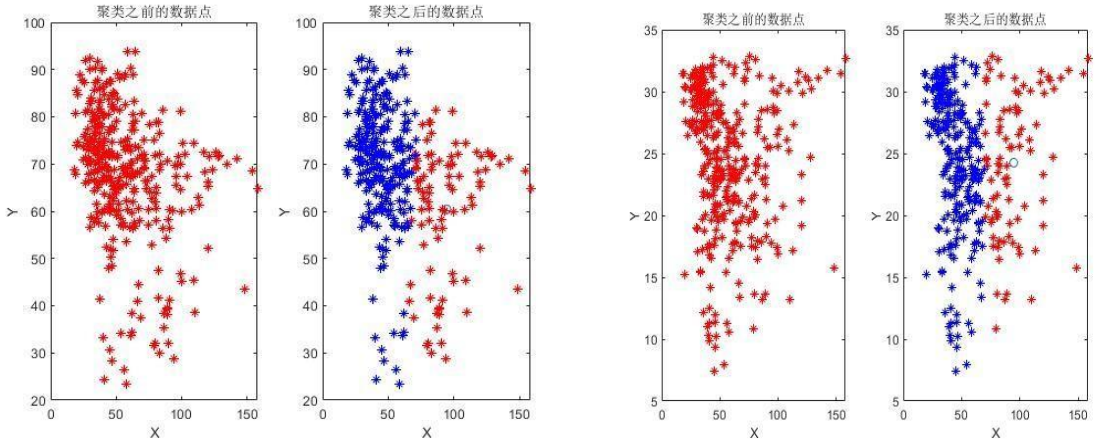
4.Wind Speed Classification:

Category 1: Gentle breeze, with an average wind speed of 2.226 m/s. Wind aids in the horizontal dispersion of pollutants; the stronger the wind, the better the horizontal dispersion capability, which lowers pollutant concentration. The AQI range for this category is (0, 50).

Category 2: Light breeze, with an average wind speed of 1.530 m/s, where pollutant concentration is relatively high, with an AQI range of (50, 100).

Category 3: Calm wind, with an average wind speed of 0.994 m/s. Low wind speed enhances mixing rather than dispersion, which is unfavorable for pollutant dispersion, resulting in high pollutant concentrations and correspondingly high AQI values, ranging from (100, 150).

The clustering result graphs for temperature, humidity, pressure, and wind speed are shown in Figure 4-3



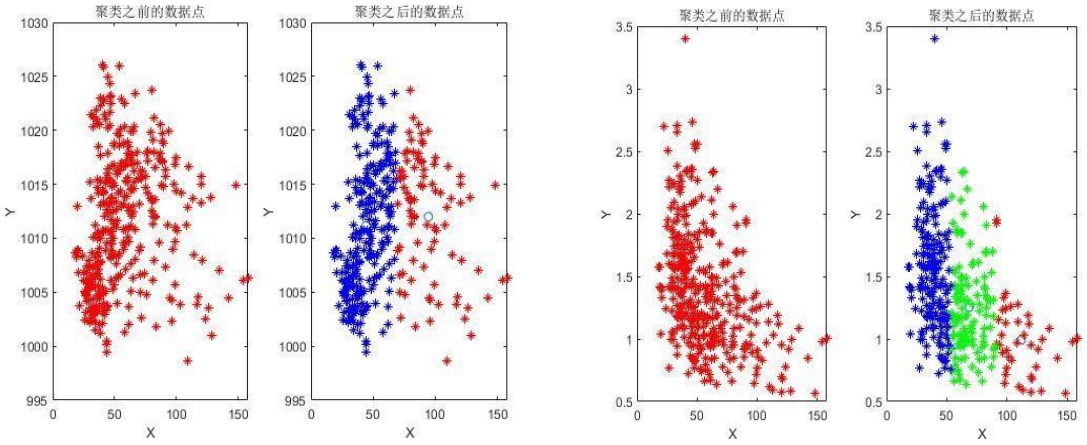


Figure 4-3 Clustering Results of Temperature, Humidity, Pressure, and Wind Speed

After performing clustering based on AQI values for different meteorological conditions, clustering was also conducted for the entire dataset. Since k-means clustering requires prior specification of the number of clusters, but we cannot predict the optimal number of clusters, a maximum of 13 clusters was set for multiple clustering attempts. The clustering results were evaluated using the clustering validity indicators DB (Davies-Bouldin index) and SSE (sum of squared errors). Ultimately, the optimal number of clusters was determined to be 3, dividing the entire dataset into three categories. The cluster centers are shown in Table 4-1 (rounded to three decimal places)：

Table 4-1 cluster centers

Centers	SO_2 (μg/m3)	NO_2(μg/m³)	PM_{10}(μg/m³)	$PM_{2.5}$(μg/m³)	O_3(μg/m³)	CO(mg/m³)	温 度 (℃)	湿 度 (%)	气 压 (MBar)	风速 (m/s)	风 向 (°)
First	0.644	7.240	24.545	122.812	54.732	27.778	26.660	56.879	1011.303	1.575	79.178
Second	0.707	6.283	31.822	48.908	45.798	24.578	25.100	67.665	1010.3	1.343	288.373
Thrid	0.712	6.257	37.400	30.752	41.785	21.729	23.081	72.649	1012.385	1.360	50.188

Based on the clustering results, the meteorological conditions can be divided into three categories:

Category 1 has an average temperature of 26.60° C, average humidity of 56.876%, average pressure of 1011.303 MBar, average wind speed of 1.575 m/s, and an average wind direction of 79.178° . Under these conditions, the average pollution concentrations of SO2, NO2, and PM10 are low, but the pollution concentrations of PM2.5, O3, and CO are the highest among the three categories.

Category 2 has an average temperature of 25.100° C, average humidity of 67.665%, average pressure of 1010.3 MBar, average wind speed of 1.343 m/s, and an average wind direction of 288.373° . Under these conditions, the average concentration of PM2.5 is the highest among the three categories, while the concentrations of other pollutants are moderate.

Category 3 has an average temperature of 23.081° C, average humidity of 72.649%, average pressure of 1012.385 MBar, average wind speed of 1.360 m/s, and an average wind direction of 50.188° . Under these conditions, except for PM10, which has the highest concentration among the three categories, the concentrations of other pollutants are at their lowest levels.

5. Analysis and solution of Question 3

5.1 Analysis of Problem 3

An effective air quality forecasting system helps humans grasp future concentrations of pollutants, formulate corresponding control strategies, and provide accurate forecasts of pollutant concentration levels in the air in advance. This helps mitigate the non-linear impact caused by pollutant concentration exceeding the standard. Currently, the commonly used WRF-CMAQ modeling system is employed for air quality forecasting. The WRF-CMAQ model consists of two parts: WRF and CMAQ. WRF is a mesoscale numerical weather prediction system used to provide the meteorological field data required for CMAQ; CMAQ is a three-dimensional Eulerian atmospheric chemical and transport modeling system that simulates the variation of pollutants based on physical and chemical reaction principles, using meteorological information from WRF and emission inventories, and generates forecast results for specific time points or periods.

However, due to the uncertainties in the simulated meteorological fields and emission inventories, as well as the incomplete understanding of the pollutant

generation mechanisms, including ozone, the results of the WRF-CMAQ forecasting model are not ideal. Therefore, the concept of secondary modeling is proposed, which refers to remapping the forecasting model results from primary models like WRF-CMAQ by incorporating more data sources to improve forecasting accuracy. In this process, actual meteorological conditions significantly impact air quality (e.g., lower humidity favors ozone generation), and the variation of real-time pollutant concentration data also provides valuable reference for air quality forecasting. Thus, air quality monitoring data on meteorology and pollutants are often referenced for secondary modeling to optimize the forecast model.

Predicting the concentration of pollutants in the air is a complex issue. Currently, the main models used for urban air quality forecasting in China include multiple linear regression, artificial neural networks, NAQPMS, CAMx, WRF-Chem, and multi-model forecasting systems. Research both domestically and internationally shows that neural networks can forecast the temporal variation trends of air pollutants better than regression models, yielding more accurate forecasting results. Neural network-based air quality forecasting models are suitable for cities where numerical air quality forecasting is difficult to implement, and where regression statistical models cannot meet the required forecasting accuracy. Therefore, based on the primary forecasting model results from WRF-CMAQ, this paper develops a BP neural network-based air quality prediction model through secondary modeling.

5.2 Analysis and solution of Question 3.

5.2.1 The model principle and framework

BP Neural Network

The BP (Backpropagation) algorithm consists of two main processes: forward computation (forward propagation) of data flow and backward propagation of error signals. During forward propagation, the direction of propagation is from the input layer → hidden layer → output layer, with the state of each neuron only affecting the neurons in the next layer. If the expected output is not obtained at the output layer, the process switches to backward propagation of the error signal. By alternating between these two processes, the gradient descent strategy for the error function is executed in the weight vector space, dynamically iterating to search for a set of weight vectors that minimize the network error function, thus completing the information extraction and memory process.

Neural Network Model Principles Description:

(1) Forward Propagation

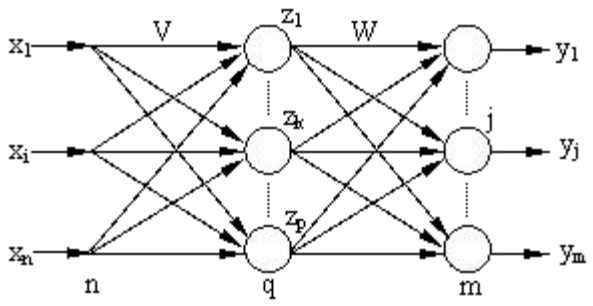


Figure 5-1 The topology of a three-layer neural network

Suppose the BP network has n nodes in the input layer, q nodes in the hidden layer, and m nodes in the output layer. The weights between the input layer and the hidden layer are denoted as v_{ki}, and the weights between the hidden layer and the output layer are denoted as w_{jk} , as shown in Figure 5-1. The transfer function for the hidden layer is f1(·), and the transfer function for the output layer is f2(·). Then, the output of the hidden layer nodes is (including the threshold in the summation term)：

$$\mathrm{zk} = \mathrm{f1}\left(\sum_{i=0}^{n} vki\ xi\right) \qquad \mathrm{k}=1,2,\ldots,\mathrm{q} \tag{5-1}$$

every input point output：

$$\mathrm{y_j} = \mathrm{f_2 y_j} = \mathrm{f_2}\left(\sum_{\mathrm{k}=0}^{\mathrm{q}} \mathrm{w_{jk}\, z_k}\right) \quad \mathrm{j} = 1,2,\ldots,\mathrm{m} \tag{5-2}$$

At this point, the BP network has completed the approximation mapping of an n-dimensional space vector to an m-dimensional space。

Backpropagation

(1) Definition of the Error Function (5-3)

$$\mathrm{E_p} = \frac{1}{2}\sum_{\mathrm{j}=1}^{\mathrm{m}} \left(\mathrm{t_j^p} - \mathrm{y_j^p}\right)^2$$

The global error for P samples is：

$$E = \frac{1}{2}\sum_{p=1}^{p}\sum_{j=1}^{m}\left(t_j^p - y_j^p\right) = \sum_{p=1}^{P} E_p$$

Using the cumulative error BP algorithm to adjust w_{jk}, in order to reduce the global error E, that is:

$$\Delta W_{jk} = -n\frac{\partial E}{\partial w_{jk}} = -n\frac{\partial}{\partial w_{jk}}\left(\sum_{p=1}^{P} E_p\right) = \sum_{p=1}^{P} -n\frac{\partial E_p}{\partial w_{jk}}$$

In the equation: n is the learning rate.

(2)The error signal is defined as:：

$$\delta_{yj} = -\frac{\partial E_p}{\partial S_j} = -\frac{\partial E_p}{\partial y_j}\cdot\frac{\partial y_j}{\partial S_j}$$

In the equation: the first formulation is ：

$$\frac{\partial E_p}{\partial y_j} = \frac{\partial}{\partial y_j}\left[\frac{1}{2}\sum_{j=1}^{m}\left(t_j^p - y_j^p\right)^2\right] = -\sum_{j=1}^{m}\left(t_j^p - y_j^p\right)$$

The second formulation：

$$\frac{\partial y_j}{\partial S_j} f_2'(S_j)$$

It is the partial derivative of the output layer's transfer function.
Therefore:

$$\delta_{yj} = \sum_{j=1}^{m} \left(t_j^p - y_j^p\right)\ \mathrm{f}_2'\left(S_j\right)$$

By the chain rule, we get:

$$\frac{\partial\ \mathrm{E_p}}{\partial\ \mathrm{w_{jk}}} = \frac{\partial\ \mathrm{E_p}}{\partial\ \mathrm{s_j}} \cdot \frac{\partial\ \mathrm{S_j}}{\partial\ \mathrm{w_{jk}}} = -\ \delta_{\mathrm{yj}} \mathrm{z_k} = -\sum_{\mathrm{j=1}}^{\mathrm{m}} \left(\mathrm{t_j^p} - \mathrm{y_j^p}\right) \mathrm{f_2'}\left(\mathrm{S_j}\right) \cdot \mathrm{z_k}$$

The weight adjustment formula for each neuron in the output layer is:

$$\Delta_{\mathrm{w_{jk}}} = \sum_{\mathrm{p=1}}^{\mathrm{P}} \sum_{\mathrm{j=1}}^{\mathrm{m}}\ \mathrm{n}\ \left(\mathrm{t_j^p} - \mathrm{y_j^p}\right) \mathrm{f_2}\left(\mathrm{S_j}\right) \mathrm{Z_k}$$

(3) The change in the weights of the hidden layer is:

$$\Delta_{\mathrm{v_{ki}}} = -\ \eta \frac{\partial\ \mathrm{E}}{\partial\ \mathrm{v_{ki}}} = -\ \mathrm{n}\ \frac{\partial}{\partial\ \mathrm{v_{ki}}} \left(\sum_{\mathrm{p=1}}^{\mathrm{P}} \mathrm{E_p} \right) = \sum_{\mathrm{p=1}}^{\mathrm{P}} \left(-\ \mathrm{n}\ \frac{\partial\ \mathrm{E_p}}{\partial\ \mathrm{v_{ki}}} \right)$$

The error signal is defined as:

$$\delta_{\mathrm{zk}} = \frac{\partial\ \mathrm{E_p}}{\partial\ \mathrm{S_k}} = -\ \frac{\partial\ \mathrm{E_p}}{\partial\ \mathrm{z_k}} \cdot \frac{\partial\ \mathrm{z_k}}{\partial\ \mathrm{S_k}}$$

The first term is:

$$\frac{\partial\ \mathrm{E_p}}{\partial\ z_k} = \frac{\partial}{\partial\ z_k} \left[\frac{1}{2} \sum_{j=1}^{\mathrm{m}}\ \left(\mathrm{t_j^p} - \mathrm{y_j^p}\right)^2 \right] = -\sum_{j=1}^{\mathrm{m}}\ \left(\mathrm{t_j^p} - \mathrm{y_j^p}\right) \frac{\partial\ \mathrm{y_j}}{\partial\ z_k}$$

According to the chain rule, we have:

$$\frac{\partial\, y_{\mathrm{j}}}{\partial\, z_{\mathrm{k}}} = \frac{\partial\, y_{\mathrm{j}}}{\partial\, S_{\mathrm{j}}} \cdot \frac{\partial\, S_{\mathrm{j}}}{\partial\, z_{\mathrm{k}}} = \mathrm{f}_2'(S_{\mathrm{j}}) w_{\mathrm{jk}}$$

The second term is the partial derivative of the hidden layer's transfer function:

$$\frac{\partial\, \mathrm{z_k}}{\partial\, \mathrm{s_k}} = \mathrm{f}_1'(\mathrm{S_k})$$

Therefore：

$$\delta_{\mathrm{zk}} = \sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{t_j^p} - \mathrm{y_j^p}) \mathrm{f}_2'(\mathrm{S_j}) \mathrm{w_{jk}} \mathrm{f}_1'(\mathrm{S_k})$$

According to the chain rule, we have:

$$\frac{\partial\, \mathrm{E_F}}{\partial\, \mathrm{v_{kl}}} = \frac{\partial\, \mathrm{E_F}}{\partial\, \mathrm{S_k}} \cdot \frac{\partial\, \mathrm{S_k}}{\partial\, \mathrm{v_{kl}}} = -\, \delta_{\mathrm{zk}} \mathrm{x_i} = -\sum_{j=1}^{\mathrm{m}} \left(\mathrm{t_j^p} - \mathrm{y_j^p}\right) \mathrm{f}_2'(\mathrm{S_j}) \mathrm{w_{jk}} \mathrm{f}_1'(\mathrm{S_k}) \cdot \mathrm{x_1}$$

Thus, the weight adjustment formula for each neuron in the hidden layer is:

$$\Delta \mathrm{v_{k_i}} = \sum_{\mathrm{p=1}}^{\mathrm{P}} \sum_{\mathrm{i=1}}^{\mathrm{m}} \eta \left(\mathrm{t_j^p} - \mathrm{y_j^p}\right) \mathrm{f}_2'(\mathrm{S_j}) \mathrm{w_{jk}} \mathrm{f}_1'(\mathrm{S_k}) \mathrm{x_i}$$

5.2.2 Model results

The BP neural network prediction model is an air quality prediction model based on statistical forecasting methods. It is based on the analysis of historical air quality data and meteorological conditions to identify the underlying development patterns, thereby making predictions. Compared to numerical forecasting models, this method mainly relies on the regularity analysis of historical meteorological data and pollutant monitoring concentrations. It uses meteorological condition forecast products to predict pollutant concentrations, and the analysis method is more flexible and diverse, with good applicability. The structure of the BP neural network prediction model is shown in Figure 5-2

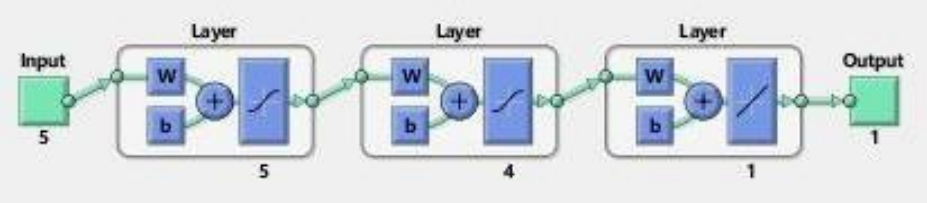


Figure 5-2: Structure of the BP Neural Network Prediction Model

According to the requirements of Problem 3, it is necessary to establish a predictive model that meets two objectives: minimizing the relative error of AQI forecast values and maximizing the accuracy of the primary pollutant. ABP neural network is used to predict the concentrations of six pollutants in the air. By inputting the measured data of the concentrations of six pollutants as well as temperature, humidity, air pressure, wind speed, and wind direction, their corresponding quantitative evaluation values are AQI. The hidden layer is set to have 4 neurons, and the output layer is set to 1. During the training process, the neural network is continuously corrected, with the number of steps set to 10,000 for training. The flowchart of the BP neural network's learning training algorithm is shown in Figure 5-3. The BP neural network model based on Problem 3 is shown in Figure 5-4.

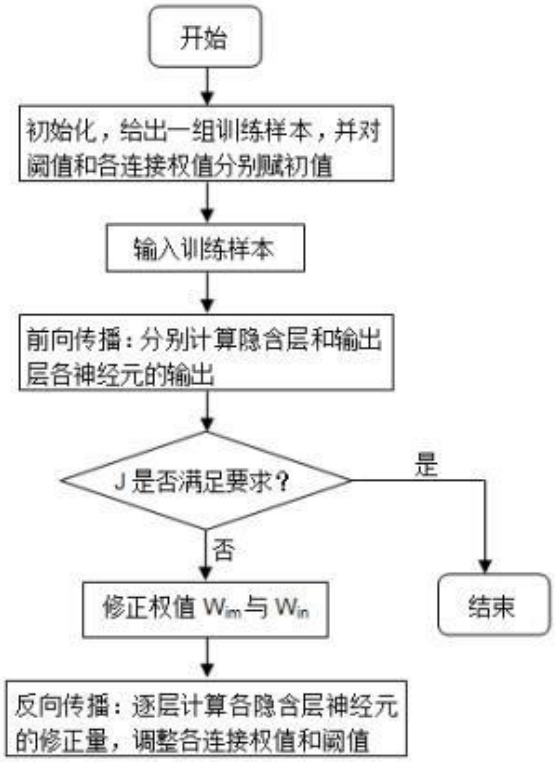


Figure 5-3: Flowchart of the BP Neural Network Learning and Training Algorithm

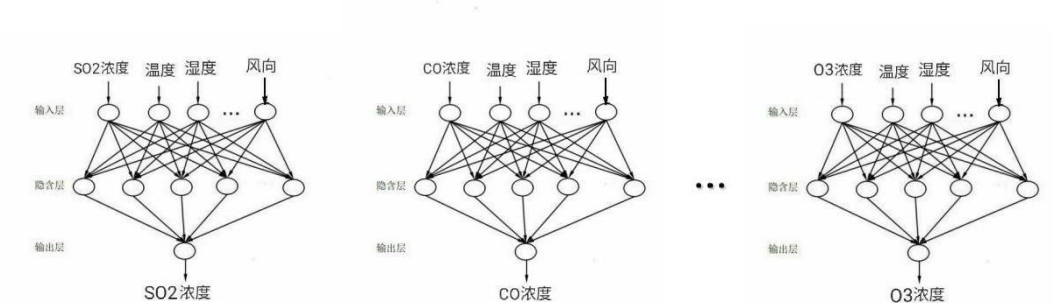


Figure 5-4: BP Neural Network Model Based on Problem Three

Utilizing MATLAB programming to train the neural network, multiple tests were conducted on the established model. The following presents the results of predicting the concentrations of six pollutants in the air from January 1, 2021, to March 3, 2021, using data from seven months, and the corresponding AQI values were calculated. A comparative analysis was made with the forecast data and the actual measured data, and it was found that the relative error of the predicted AQI values from the established secondary model was smaller and better than the primary forecast results, and the predictions for the primary pollutants matched the actual data. The training results are shown in Table 5-1. Detailed data can be found in Attachments 3-5.

Table 5-1 Pollutant concentration prediction test results

数据类型	Monitoring time	Location	CO(mg/m³)	SO2(μg/m³)	NO2(μg/m³)	O3最大八小时滑动平均(μg/m³)	PM10(μg/m³)	PM2.5(μg/m³)	AQI	Main pollutants
实测数据	2021/1/1	Monitoring Point A	0.75833333	5.958333333	33.4583333	63.5	41.4722222	19.41666667	42	NO_2
	2021/1/2	Monitoring Point A	0.9	8.833333333	56.541666667	63.0313	54.25	27.33333333	71	NO_2
	2021/1/3	Monitoring Point A	1.270833333	11.25	80.916666667	59.75	99.791666667	53.625	101	NO_2
一次预测数据	2021/1/1	Monitoring Point A	0.250053083	11.400 92708	89.6932 875	44.9809	57.902516 67	48.6755375	105	NO_2
	2021/1/2	Monitoring Point A	0.46449 2833	16.505 80083	112.752 0708	29.0943	61.907141 67	48.0132375	117	NO_2
	2021/1/3	Monitoring Point A	0.37904 6917	13.844 09125	107.232	40.2365	71.170229 17	57.745025	114	NO_2
二次预测数据	2021/1/1	Monitoring Point A	0.98594 7658	7.0326 58524	56.2410 7064	53.3643	60.746656 83	28.344585	71	NO_2
	2021/1/2	Monitoring Point A	0.77367 7718	7.4838 70105	48.5104 9485	52.1166	53.957651 04	28.592990889	61	NO_2
	2021/1/3	Monitoring Point A	0.82150 6546	10.865 27288	60.3280 1011	54.4457	69.719329 26	44.67635322	96	NO_2

To analyze the predictive performance of the model, the relative error is employed. The formula for calculating the relative error is:

$$\text{Relative error} = \frac{|monitored\ AQI - predicted\ AQI|}{monitored\ AQI}$$

AQI Relative error	first-time forecast data	second-time forecast data
2021/1/1	1.5	0.69047619
2021/1/2	0.647887324	0.14084507
2021/1/3	0.128712871	0.04950495

From the table, it can be observed that the Air Quality Index (AQI) predicted through secondary modeling is closer to the actual measured data compared to the data obtained from primary prediction, hence the relative error is smaller. This indicates that the established secondary model—BP neural network prediction model—can effectively predict the concentration of pollutants. The BP neural network prediction model has significantly improved the prediction accuracy of pollutants compared to the WRF-CMAQ prediction model, proving that the BP neural network prediction model has better performance than the general WRF-CMAQ prediction model.
By inputting 5 types of meteorological indicators that affect pollutant concentrations, the well-trained neural network is used to predict the daily concentration values of 6 routine pollutants at monitoring points A, B, and C from July 13th to July 15th, 2021, and to calculate the corresponding AQI and primary pollutants. The results are shown in the "Pollutant Concentration and AQI Prediction Results Table" as Table 5-3.

Table 5-3 Pollutant concentration and AQI prediction results

Forecast date	Location	Quadratic model daily forecast							
		SO2 (μ g/m³)	NO2 (μ g/m³)	PM10 (μ g/m³)	PM2.5 (μ g/m³)	**O_3 Maximum eight-hour moving average (μg/m³)**	CO (mg/m³)	AQI	首要污染物
2021/7/13	监测点 A	5.1526922	23.122807	27.121428	10.885860	49.14677	0.4228630	71	NO_2
2021/7/14	监测点 A	5.0530244	22.85133	27.502836	10.873185	50.910004	0.4358628	61	NO_2
2021/7/15	监测点 A	5.1166190	21.855769	26.21298	10.820821	45.12730	0.4198489	76	NO_2
2021/7/13	监测点 B	5.8729689	11.324613	23.587731	7.8127631	51.147857	0.4836970	26	O_3
2021/7/14	监测点 B	5.8453821	11.556665	23.588575	7.429732	51.332743	0.4158622	26	O_3
2021/7/15	监测点 B	5.8383749	11.508974	23.404963	7.8730105	51.033290	0.4177488	26	O_3
2021/7/13	监测点 C	6.4421460	18.96087	27.282557	17.81683	56.397678	0.5776881	29	O_3
2021/7/14	监测点 C	6.4111316	18.25703	27.15913	17.081047	55.955754	0.5745183	28	O_3
2021/7/15	监测点 C	6.5140483	17.940326	26.907403	17.039779	58.792392	0.5793493	30	O_3

6. Problem Four Analysis and Solution

6.1 Problem Four Analysis

The pollutant concentrations in adjacent areas often have a certain correlation, and regional collaborative forecasting may improve the accuracy of air quality forecasting. It is necessary to further consider the impact of direction and distance on air quality. A collaborative forecasting model needs to be established for monitoring point A and its nearby monitoring points A1, A2, and A3. The secondary model's prediction results should have a minimal maximum relative error for AQI forecast values, and the accuracy of the primary pollutant prediction should be as high as possible. The model evaluation indicators are the same as those in Problem 3, so it is possible to consider building upon Problem 3. The established BP neural network secondary prediction model needs to be improved using a genetic algorithm. Based on the direction and coordinate information of points A, A1, A2, and A3, the selection of the training set should be improved by choosing the pollutant concentrations and meteorological data monitored by the 4 different monitoring points during the same period as the training set [10]. Before testing, the table data should be preprocessed according to the previous data processing methods, dealing with missing values, negative values, and abnormal values, and converting the unit of pressure data.

6.2 Problem Four Solution

6.2.1 Model Principle and Framework

The genetic algorithm (GA) is part of evolutionary computation, simulating the biological evolution process of Darwinian genetic selection and natural elimination. It is a computational model that searches for the optimal solution by simulating the natural evolution process. The algorithm is simple, universal, robust, and suitable for parallel processing. The genetic algorithm includes three operators: selection, crossover, and mutation.

(1) Selection Operator

The selection operator simulates "survival of the fittest". Individuals with higher fitness are more likely to be inherited to the next generation, while individuals with lower fitness have a lower probability of being inherited to the next generation. A common selection operator is the roulette wheel selection method. Suppose represents the fitness function of the individual in the population, n is the number of individuals in the population, and Σf_i is the sum of the fitness of the individuals in the population. The ability to produce offspring is given by equation (3), where $k = 1,2, ...$

$$P_i = \frac{f_i^k}{\sum_{i=1}^{n} f_i^k}$$

(2) Crossover Operator

The crossover operator refers to the process of exchanging parts of genes between two paired chromosomes in a certain way, thereby creating new individuals. This is an important characteristic that distinguishes genetic algorithms from other algorithms and is the main method for generating new individuals. Common crossover methods include single-point crossover, double-point crossover, multi-point crossover, uniform crossover, and arithmetic crossover.

(3) Mutation Operator

Mutation in genetic algorithms involves replacing the gene values at certain gene loci in an individual's chromosome code with other alleles at those loci, thus forming a new individual. The mutation operation in genetic algorithms is to maintain the diversity of the population and prevent gene loss.

Basic Steps of Genetic Algorithms

(1) Chromosome encoding.
(2) Initialize the population.
(3) Design a fitness function to determine an individual's environmental adaptability; the higher the fitness value, the stronger the adaptability, and the greater the chance of survival.
(4) Based on the fitness function, select superior individuals from the parent population.
(5) Perform crossover operations based on the crossover rate.
(6) Perform mutation operations based on the mutation rate.
(7) Determine whether the termination condition is met; if yes, go to (8); if no, go back to (3).
(8) End the iteration and select the individual with the highest fitness function value as the optimal solution to the problem [8].

The basic steps of genetic algorithms are shown in Figure 6-1.

◦

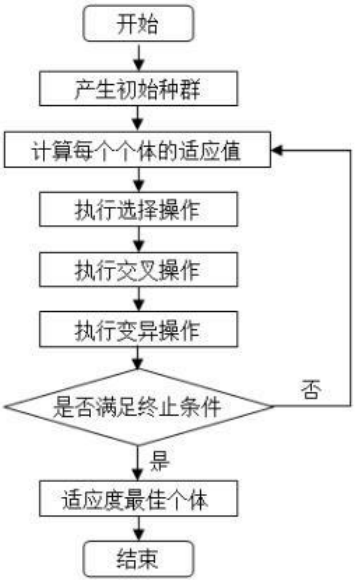


Figure 6-1 Basic steps of genetic algorithm

Improved Genetic BP Neural Network Air Quality Forecasting Model

(1) Determine the Structure of the BP Neural Network

Based on the pollutant concentrations and meteorological data detected by monitoring points A, A1, A2, and A3, construct the structure of the BP neural network, design the input and output layer variables, and determine the number of nodes and layers in the hidden layer.

(2) Optimize the BP Neural Network
Design a fitness function to calculate the fitness value of each individual. Through genetic algorithm operations such as selection (binary tournament selection),

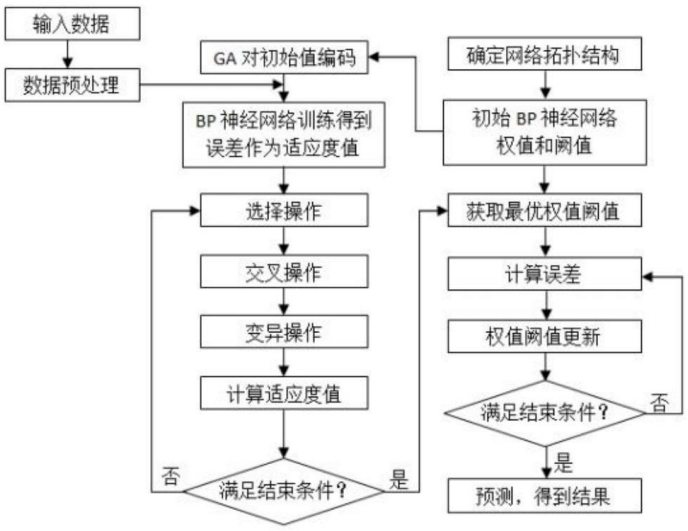


Figure 6-2 Flowchart of BP neural air quality prediction model based on genetic algorithm

crossover (simulated binary crossover),

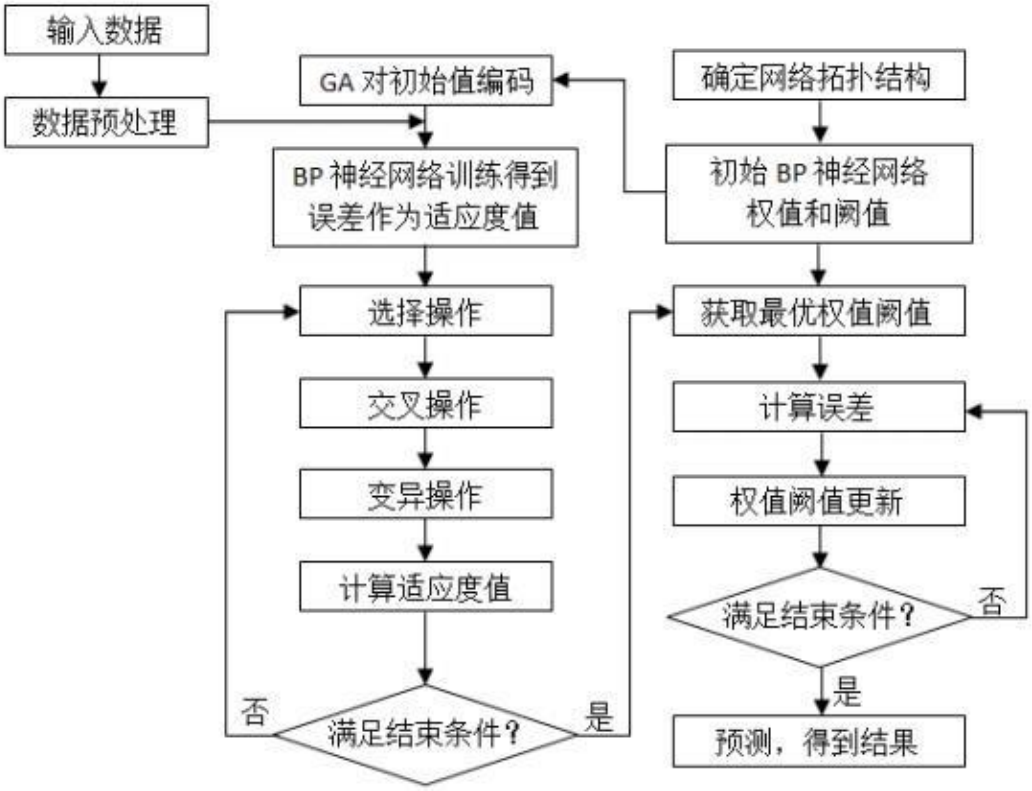


Figure 6-2 Flowchart of BP neural air quality prediction model based on genetic algorithm

6.2.2 Model Results

Using this model, the daily concentration values of 6 routine pollutants at monitoring points A, A1, A2, and A3 from July 13, 2021, to July 15, 2021, were predicted, and the corresponding AQI and primary pollutant prediction results (retaining six decimal places) are shown in Table 6-1. The detailed content of the processed data and prediction results can be found in Attachments 6-9.

Forecast date	Location	Quadratic model daily forecast							
		SO2 (μg/m³)	NO2 (μg/m³)	PM10 (μg/m³)	PM2.5 (μg/m³)	O_3 Maximum eight-hour moving average (μg/m³)	CO (mg/m³)	AQI	Main pollutants
2021/7/13	Monitoring Point A1	6.026498	12.754470	23.135363	6.526287	43.911091	0.396785	24	PM10
2021/7/14	Monitoring Point A1	5.981969	12.971135	21.672379	6.321114	44.094971	0.395751	23	03

2021/7/15	Monitoring Point A1	5.880314	12.881624	21.495475	6.382446	25.802753	0.396683	22	PM10
2021/7/13	Monitoring Point A1	7.027309	15.332967	29.684920	9.177502	46.661877	0.451259	30	PM10
2021/7/14	Monitoring Point A1	7.281718	14.683218	29.130652	9.039957	46.462130	0.639348	30	PM10
2021/7/15	Monitoring Point A1	7.181598	14.934196	30.305947	9.019191	25.802753	0.961902	31	PM10
2021/7/13	Monitoring Point A2	4.850706	16.518653	27.087839	8.169413	58.828821	0.443849	71	NO2
2021/7/14	Monitoring Point A2	4.894688	16.241063	26.025836	8.814907	58.144192	0.445929	61	NO2
2021/7/15	Monitoring Point A2	4.913525	16.284831	26.080662	8.513877	25.802753	0.466164	76	NO2
2021/7/13	Monitoring Point A3	3.849022	10.188080	17.846609	6.898581	58.828821	0.307802	30	O3
2021/7/14	Monitoring Point A3	3.852278	10.017528	17.209906	25.802753	58.144192	0.307467	37	PM2.5
2021/7/15	Monitoring Point A3	7.344998	9.8094291	16.454645	11.317941	25.802753	0.308424	17	NO2

Compared to the model in Problem 3, the newly established GA-BP neural network regional collaborative forecasting model has improved the prediction accuracy of pollutant concentrations at monitoring point A, reduced the relative error of the air quality index AQI forecast values, and enhanced the precision of predicting primary pollutants. Tests have shown that the optimized GA-BP neural network regional collaborative forecasting model has a significantly smaller prediction error for pollutant concentrations than the BP network prediction model, with a smaller relative error in AQI forecast values and more accurate predictions of primary pollutants, indicating that the GA-BP prediction model has higher prediction accuracy. By analyzing the output results of the two models against the expected output, it was found that the GA-BP model has a higher correlation coefficient, indicating that the optimized prediction model has stronger generalization capabilities.

7. Conclusion and Model Evaluation

7.1 **Conclusion**

(1) For Problem 1, the problem requires the use of data in Appendix 1 to calculate the actual measured AQI and primary pollutants at monitoring point A from August 25, 2020, to August 28, 2020, according to the method in the appendix, and to place the results in the text in the format of the "AQI Calculation Result Table" in the appendix. Observations indicate that the required parameters from August 25, 2020, to August 28, 2020, do not have issues such as null values or missing data, so Problem 1 can be directly solved using the formulas provided in the appendix to calculate the Air Quality Index (AQI) values.

(2) Using various mathematical principles and computational software, data processing and preprocessing were completed, providing all the forecast data for the time span from July 23, 2020, to July 13, 2021, and the collection records of actual measured data for the time span from April 16, 2019, to July 13, 2021, along with all relevant meteorological conditions. For various data or erroneous data in the provided collection data, corresponding processing methods were adopted: for abnormal data caused by equipment or external conditions resulting in missing variable values, an average processing was applied over a period of time before and after; for variable values exceeding the adjustable minimum or maximum range, they were directly taken as the corresponding minimum or maximum values based on the adjustable range of the variable; for variable values exceeding the 3σ range, the 3σ criterion was used to remove the jump values collected at each station.

(3) A secondary air quality prediction model based on the BP neural network was established. The model first analyzed and selected 4 main meteorological features: temperature, humidity, air pressure, and wind speed. Then, a BP neural network model was built, along with an instant learning framework. Based on the time series and the sequential nature of air quality monitoring, data from 7 months were used to predict the pollutant concentrations for the next 3 days. Additionally, the BP neural network prediction model results were compared with actual values and primary forecast results. To ensure the prediction model meets two objectives, the model was tested randomly over a period of 3 days and compared with primary forecast results, indicating that the established secondary prediction model is superior to the primary forecast results.

(4) Through data processing and analysis of the coordinates, directions, and distances of A, A1, A2, and A3, a collaborative forecasting model was established, incorporating the meteorological feature wind direction into the features to improve the prediction model.

7.2 **Model Evaluation**

7.2.1 **Model Advantages**

(1) A relatively reasonable method was designed for preprocessing the original data, also considering the various complex actual working conditions that monitoring station equipment may encounter during data measurement. The 3σ criterion was used to process the data, making the original data more accurate.

(2) In the data analysis, the k-means clustering method was used for cluster analysis of different meteorological features and visualization processing, allowing for a direct view of the classified meteorological features and the corresponding changes in pollutant concentrations and Air Quality Index (AQI) values. When the number of clusters is uncertain, multiple groups of clustering were applied to the data, and the clustering effectiveness was evaluated using indicators such as DB and SSE, selecting the optimal clustering scheme.

(3) Existing air quality forecasting models are influenced by meteorological factors (temperature, humidity, air pressure, wind speed, wind direction, etc.) and pollution source emissions, and the analysis of data is not deep enough to obtain deeper connections within the data. The BP neural network method used can train through a large dataset, mine and capture deep connections within the data, improving the accuracy of the forecast results. With the growth of training samples, due to the model's dynamic bias correction function, results can be corrected in real-time, and the model's performance is superior to general methods. The BP neural network is simple and effective, easy to implement, fast in training and learning time, and does not require a long duration of basic data.

(4) The improved GA-BP neural network collaborative forecasting model can more accurately predict pollutant concentrations, primary pollutants, and the relative error of AQI forecast values is also smaller. The improved GA-BP neural network regional collaborative forecasting model enhances the network's convergence speed, prediction accuracy, fitting degree, and generalization capabilities.

7.2.2 Model Disadvantages

(1) The air quality forecasting model and optimization scheme in this paper are based on the BP neural network air quality prediction model, and the prediction performance needs to be improved, providing a theoretical basis for subsequent predictions using machine learning and deep learning models. At the same time, the parameter initialization in the BP neural network is also a problem. The number of network layers and the number of neural nodes per layer need to be continuously tried to determine the optimal value. There is no theoretical basis for the selection of neural data in the

hidden layer and the initial value of connection weights, relying on experience, and the global optimal convergence cannot be guaranteed. Complex functions are prone to falling into local optimal solutions. In addition, the model's performance is highly dependent on the training set. When predicting air quality, appropriate air quality data must be selected to ensure that the relative error of AQI forecast values is small and the prediction of primary pollutants is accurate.

(2) Establish a prediction model that combines multiple models and includes influencing factors (weights), such as wind speed, humidity, and precipitation, to explore air quality characteristics, weather characteristics, weather forecast characteristics, historical mutation-related characteristics, and external data (spatial points, pollution source lists, text image data) characteristics, improving the model's prediction accuracy from temporal and spatial dimensions.

(3) Since the genetic algorithm uses random initialization of the population, it may fall into local optimal solutions. The initial population scheme can be improved; genetic operators, crossover, and mutation can be improved. During the loop process, excellent solutions from the parent population can be retained through improved genetic operators to speed up the algorithm's convergence speed.

(4) Continue to optimize the air quality prediction model. Train and test corresponding prediction models for short-term and long-term air quality predictions to achieve not only accurate predictions of air quality within a short time frame of one day but also predictions of air quality for up to seven days or even longer, while maintaining high prediction accuracy. In addition, specific prediction models can be trained for major festivals and events to improve prediction accuracy during specific time periods. For the initialization problem of model parameters, an initial value model can be established for these parameters to explore the internal connections between initial values under different prediction scenarios, thereby achieving the optimization of prediction effects.

(5) Establish a deep connection between air quality and pollution source emissions. For the issue that pollution source emissions of pollutants have a negative impact on air quality, deep learning models can be used to deeply analyze input meteorological data, Air Quality Index, and pollution source emissions, emission locations, and other parameters. Under the premise of ensuring air quality, acceptable amounts of pollutant emissions can be derived, thereby controlling emissions within a reasonable range and providing a reference for government departments to make scientific decisions.

参考文献

[1]伯鑫等，空气质量模型(SMOKE、WRF、CMAQ 等)操作指南及案例研究[M].北京:中国环境出版集团,2019.
[2]赵秋月,李荔,李慧鹏.国内外近地面臭氧污染研究进展[J].环境科技,31(05):72-76,2018.
[3]陈敏东.大气臭氧污染形成机制及研究进展[J/0L].https://max.book118.com/html/2018/0201/151478594.shtm.2018.
[4] 陶莹 , 杨锋 , 刘洋 , 戴兵 .K 均值聚类算法的研究与优化 [J]. 计算机技术与发展,28(06):90-92,2018.
[5]王晓彦,刘冰,李健军,丁俊男,汪巍,赵熠琳,鲁宁,许荣,朱媛媛,高愈霄,李国刚.区域环境空气质量预报的一般方法和基本原则[J].中国环境监测,31(01):134-138,2015.
[6]郝吉明,马广大,王书肖.大气污染控制工程[M].北京:高等教育出版社,2010.
[7]戴树桂.环境化学[M].北京:高等教育出版社,1997.
[8]宋鹏程,张馨文,黄强,龙平,杜云松.我国城市环境空气质量预报主要模型及应用[J].四川环境,38(03):70-76,2019.
[9]宋宇辰,甄莎.BP 神经网络和时间序列模型在包头市空气质量预测中的应用[J].干旱区资源与环境,27(07):65-70,2013.
[10]牛玉霞.基于遗传算法和 BP 神经网络的空气质量预测模型研究[J].软件,38(12):49-53,2017.