Secondary Modeling of Air Quality Forecast Based on the Transformer Model

November 13, 2023

Abstract

The concentration of atmospheric pollutants is influenced by meteorological elements such as temperature, pressure, humidity, and wind speed, as well as the concentration of other pollutants. Accurately predicting pollutant concentrations using a variety of feature data combined with primary forecast data from the **CMAQ** model is a complex and highly uncertain research topic. Due to uncertainties in the simulated meteorological fields and emission inventories, as well as incomplete understanding of the generation mechanisms of pollutants including ozone, the **WRF-CMAQ** forecast model results are not ideal. Successfully predicting pollutant concentrations is of significant importance for public health and the formulation of government policies.

For Problem 1, calculate the **AQI** at monitoring point A from August 25 to 28, 2020, using the method in the appendix. The daily measured **AQIs** were 60, 46, 109, and 138, respectively, with ozone (O_3) being the primary pollutant each day. By applying existing formulas to calculate the air quality index for six pollutants and taking the maximum value, the primary pollutant can be determined.

For Problem 2, analyze the impact of various meteorological conditions on pollutant concentrations based on the correlation coefficients between 21 elements and their physical causes.

The meteorological conditions are categorized into four types:

1. High-impact radiation factors, including specific humidity, atmospheric pressure, longwave radiation, and 2m temperature;

2. Humidity factors, including surface temperature, wind direction, rainfall, and relative humidity;

3. Thermodynamic factors, including 10m wind speed, sensible heat, latent heat, and boundary layer height;

4. Cloud cover, shortwave radiation, and surface radiation.

For Problem 3, we established a neural network based on the **Transformer model** to predict pollutant concentrations 72 hours in advance. The **Encoder model** inputs include temperature, pressure, humidity, wind speed, wind direction, and the concentrations of six pollutants for the previous 840 hours, while the **Decoder** inputs the predicted values of pollutants for the next 72 hours. For the periods required by the topic, the Decoder inputs the processed primary forecast data predictions. On the test set, for point A, the

model's training effect in predicting O_3 reached an error of only 0.667. We obtained **AQIs** of 23, 24, and 30 for point A on the 13th, 14th, and 15th, respectively, with NO_2 as the primary pollutant. For point B, the **AQI** was 26 for all three days, with NO_2 as the primary pollutant. For point C, the **AQIs** were 40, 39, and 41, respectively, with NO_2 as the primary pollutant.

For Problem 4, based on the model from Problem 3, we input meteorological elements and pollutant concentrations for points A, A1, A2, A3, and A4 into the encoder, with primary forecast data as the decoder input. The model structure is the same as in Problem 3. The results show that for point A, the error in O_3 prediction is only 0.675, with no significant change in error. We obtained **AQIs** of 32, 35, and 32 for point A on the 13th, 14th, and 15th, respectively, with NO_2 as the primary pollutant.

Keywords: Transformer model, STL, Multiple Imputation, Correlation Coefficient, Pollution Concentration Prediction

Contents

1 Problem Restatement

1.1 Problem Background

Environmental air quality is a hot topic of concern for people's daily travel and living, as it is closely related to our health and sense of well-being. Studying the variation and trend of pollutant concentrations in the air and establishing accurate air quality prediction models can help in pollution control and analysis of environmental air quality. The Air Quality Index (**AQI**) is an important indicator for determining air quality levels. Currently, six common atmospheric pollutants are used to measure air quality: sulfur dioxide (SO_2), nitrogen dioxide (NO_2), particulate matter less than 10 micrometers (PM_{10}), particulate matter less than 2.5 micrometers ($PM_{2.5}$), ozone (O_3), and carbon monoxide (CO). In recent years, research on air quality prediction using machine learning theory has become a hot topic. Many studies have focused on constructing prediction and forecasting models based on deep learning and BP neural networks. Some research has been conducted using the **LIBSVM** method to establish prediction models for meteorological conditions and pollutant concentrations, focusing on accuracy. By utilizing time-domain convolutional networks (**TCN**), multi-scale bilinear weather prediction models have been developed, extracting features from historical observation data to further uncover potential correlations within the data.

This paper aims to predict the Air Quality Index as accurately as possible, using hourly pollutant concentration data, meteorological primary forecast data, and actual meteorological data, as well as daily pollutant concentration data. Based on the primary model **WRF-CMAQ**, we use actual pollutant concentration data and primary model data to establish a method for predicting air quality, with the goal of minimizing the relative error of AQI predictions and maximizing the accuracy of primary pollutant predictions. Models are established for different regional monitoring points and adjacent monitoring points within the same region.

Finally, using the pollutant concentration data and primary model data from July 13 to 15, 2021, we demonstrate the effectiveness and robustness of the model in improving AQI prediction accuracy.

1.2 Problems to be Solved

Based on the above research background, the problems to be solved in this paper are as follows:

Problem 1: Calculation of **AQI** and Primary Pollutant

Given the measured pollutant concentration data from August 25 to 28, 2020 (Annex 1), and the AQI calculation method (Appendix), calculate the AQI and primary pollutants for the 25th to 28th using the measured data.

Problem 2: Classification and Feature Description of Meteorological Conditions

It is known that when pollutant concentrations are constant, the more favorable the meteorological conditions for pollutant dispersion, the lower the AQI

value; conversely, the more favorable the conditions for pollutant deposition, the higher the AQI value. Based on hourly pollutant concentration data, meteorological primary forecast data, actual meteorological data, and daily pollutant concentration data (Annex 1), classify the meteorological conditions and describe the features of each type.

Problem 3: Establishing a Secondary Forecast Model for Pollutant Concentrations

Based on the hourly pollutant concentration data and meteorological primary forecast data, actual meteorological data, and daily pollutant concentration data for monitoring points A, B, and C, establish a secondary forecast mathematical model to predict the daily concentration values of the six common pollutants for the next three days. Use the AQI formula in the appendix to solve for the daily pollutant concentrations and calculate the AQI prediction results and primary pollutants for July 13 to 15, 2021, at monitoring points A, B, and C.

Problem 4: Establishing a Collaborative Forecast Model

Based on the principle that regional collaborative forecasting can help improve accuracy, fully utilize the relationship between pollutant concentration changes in adjacent areas. Considering the location relationship of monitoring points A, A1, A2, and A3, use the hourly pollutant concentration data, meteorological primary forecast data, and actual data, as well as daily pollutant concentration data (Annexes 1 and 3), to establish a collaborative forecast model for monitoring points A, A1, A2, and A3. This model should reduce the relative error of AQI and improve the accuracy of primary pollutants compared to primary forecast data. Based on the daily concentration values of the six common pollutants at monitoring points A, A1, A2, and A3 from July 13 to 15, 2021, calculate the AQI prediction results and primary pollutants. Discuss whether the model results for different monitoring points A, B, and C, compared to the model established using adjacent monitoring points A, A1, A2, and A3, help improve forecasting accuracy, and provide reasons.

2 Model Assumptions and Key Symbol Descriptions

2.1 Model Assumptions

(1) It is assumed that there will be no occurrences during the research period.
(2) It is assumed that computation costs are not considered.

Symbol	Description
δ	Relative error of AQI
σ	Prediction accuracy
$IAQI_p$	Air Quality Sub-Index for pollutant P
Cp	Mass concentration value of pollutant P
AQI	Air Quality Index
BP_{hi}	Breakpoint concentration higher limit close to C_p
BP_{lo}	Breakpoint concentration lower limit close to C_p
$IAQI_{hi}$	IAQI value for BP_{hi}
$IAQI_{lo}$	IAQI value for BP_{lo}
x_i	Original x position in the sorted list
y_i	Original y position in the sorted list
x_i'	Rank of original x in the sorted list
y_i'	Rank of original y in the sorted list
d_i	Rank difference

Table 1: Air Quality Index (AQI) Calculation Parameters

2.2 Key Symbol Descriptions

3 Problem One: Calculation of AQI and Primary Pollutant

3.1 Problem Analysis

For Problem One, it is necessary to calculate the daily measured AQI and primary pollutants from August 25 to August 28, 2020. As shown in Figure 1, first, according to Table 1, determine the high and low limit values for the concentration of six pollutants for each day, as well as the corresponding high and low sub-index values for air quality. Next, calculate the air quality sub-index (IAQI) for each pollutant using the formula, and finally, determine the day's AQI value and corresponding primary pollutant by comparing the IAQIs.

No.	Index or Pollutant Item	Air Quality Sub-Index and Corresponding 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/m^3
2	Sulfur Dioxide (SO_2) 24-hour average	0	50	150	475	800	1600	2100	2620	-
3	Nitrogen Dioxide (NO_2) 24-hour average	0	40	80	180	280	565	750	940	-
4	Ozone (O_3) Max 8-hour moving average	0	100	160	215	265	800	-	-	-
5	Particulate $Matter \leq 10\mu m$ (PM_{10}) 24-hour average	0	50	150	250	350	420	500	600	$\mu g/m^3$
6	Particulate $Matter \leq 2.5\mu m$ ($PM_{2.5}$) 24-hour average	0	35	75	115	150	250	350	500	$\mu g/m^3$

Table 2: Air Quality Index (IAQI) Calculation Parameters and Pollutant Concentration Limits

Note: (1) For the maximum 8-hour moving average concentration of ozone (O_3) exceeding $800\mu g/m^3$, no further air quality sub-index calculation is conducted. (2) For other pollutants, when concentrations exceed the limit value corresponding to IAQI=500, no further air quality sub-index calculation is conducted.

3.2 Numerical Solution

According to the "Technical Regulation for Environmental Air Quality Index (AQI) (Trial)", the calculation of the AQI includes more types and reflects the degree of pollution and primary pollutants. The calculation formula is as follows:

$$IAQI_P = \frac{IAQI_{Hi} - IAQI_{LO}}{BP_{Hi} - BP_{LO}} (C_P - BP_{LO}) + IAQI_{LO} \tag{1}$$

$$AQI = \max \{IAQI_1, IAQI_2, IAQI_3, \ldots, IAQI_n\} \tag{2}$$

Where: $IAQI_P$ is the air quality sub-index for pollutant P, rounded up to the nearest integer; C_P is the mass concentration value of pollutant P; BP_Hi and BP_LO are the high and low limit values of pollutant concentration close to C_P;$IAQI_Hi$ and $IAQI_LO$ are the high and low sub-index values for air quality corresponding to BP_Hi and BP_Lo, respectively. AQI is the Air Quality Index.

Using MATLAB programming, the algorithm for solving AQI is as follows:

Table 3: Steps of the AQI Calculation Algorithm

AQI Calculation Algorithm Steps	
Step 1:	Select the daily concentration of each pollutant for the specified date.
Step 2:	Determine the corresponding concentration of the pollutant for the day and the high and low limits of the air quality sub-index.
Step 3:	Calculate the IAQI for each pollutant concentration on the day.
Step 4:	Solve for AQI and primary pollutant.

3.3 Result Conclusion

By referring to Table 1 and using formulas (1) and (2), as well as the solution algorithm, the AQIs and primary pollutants for each day from August 25 to August 28, 2020, are calculated as shown in Table 3:

Table 4: AQI Calculation Results

Monitoring Date	Location	AQI Calculation	
		AQI	Primary Pollutant
2020/8/25	Monitoring Point A	60	Ozone (O_3)
2020/8/26	Monitoring Point A	46	Ozone (O_3)
2020/8/27	Monitoring Point A	108	Ozone (O_3)
2020/8/28	Monitoring Point A	137	Ozone (O_3)

In summary, the AQIs for August 25 to August 28, 2020, are 60, 46, 108, and 137, respectively. The primary pollutant for each day is ozone (O_3).

4 Problem Two: Classification of Meteorological Conditions Affecting Pollutant Concentration

4.1 Analysis of Problem Two

Generally, sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO) primarily originate from automobile exhaust and factory fuel combustion emissions and are primary pollutants, which are directly emitted into the atmosphere from pollution sources. $PM_{10.0}$ and $PM_{2.5}$ are particulate pollutants, while O_3 is the only secondary pollutant among the six pollutants, with a unique generation process. Its accumulation process is not only related to meteorological factors such as NO_2 photolysis but also to various primary pollutants. Meteorological conditions influence air quality by affecting the dispersion, dilution, and migration of pollutants. The impact of meteorological factors on atmospheric pollution is not singular; instead, it is formed through the coordination and interaction of various meteorological elements, creating special meteorological conditions conducive to the generation, stagnation, or dispersion of pollutants, thereby collectively affecting the concentration of atmospheric pollutants. Therefore, meteorological conditions can be reasonably classified based on their synergistic impact on atmospheric pollutant concentrations.

Considering that the CMAQ model can forecast three days' worth of hourly data at once, this section first classifies the 15 meteorological elements according to the forecast periods of $0 \sim 23$ hours, $24 \sim 47$ hours, and $47 \sim 72$ hours based on primary forecast data. The forecast results closest in time (i.e.,$0 \sim 23$ data) are analyzed, followed by the construction of a 21-dimensional **Spearman** correlation coefficient matrix based on the same-time 8396-row samples in Annex 1. These 21 dimensions include 15 meteorological elements output by the WRF model and the measured hourly concentrations of six pollutants. Finally, the final classification of meteorological conditions is determined based on the correlation coefficients between each meteorological element and pollutant concentrations, as well as among the meteorological elements themselves.

4.2 Correlation Analysis between Meteorological Element Forecast Values and Pollutant

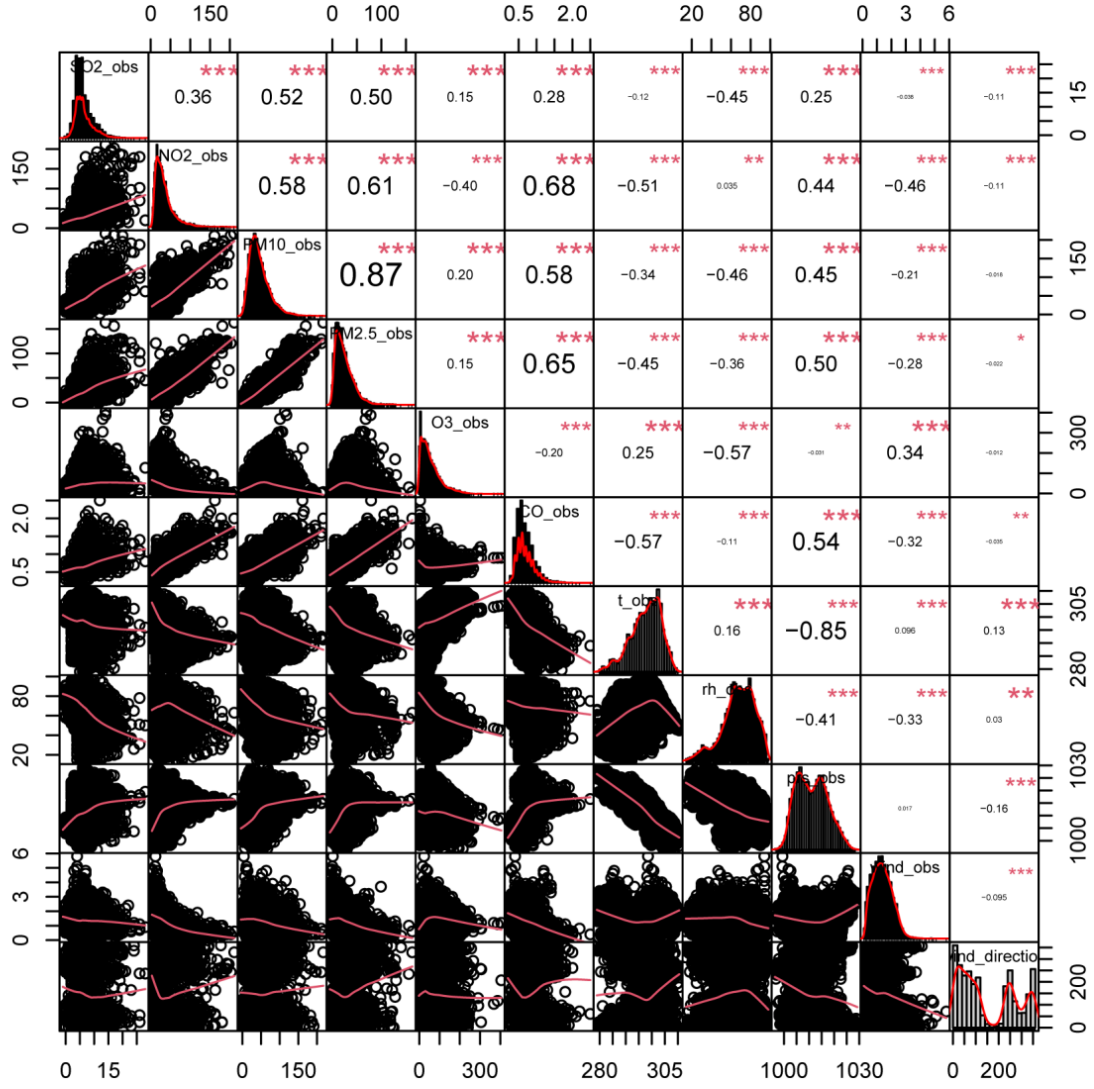
Considering that different meteorological elements may not necessarily follow a normal distribution (such as precipitation and wind direction data), this paper uses non-parametric (distribution-independent) methods to calculate the correlation coefficient matrix. The Spearman rank correlation coefficient is one such method, used to measure the strength of the relationship between variables. In cases without repeated data, where one variable is a strict monotonic function of another, the Spearman rank correlation coefficient is either +1 or -1, indicating complete Spearman rank correlation. Meteorological elements and pollutant concentration data x_i and y_i are sorted from highest to lowest, and the original ranks of x_i' and y_i' in the sorted list, x_i and y_i, are noted as the ranks. Rank difference

indicating the distance measure between two ranks, is expressed as:

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n\left(n^2 - 1\right)} \tag{3}$$

Based on existing data, the Spearman rank correlation coefficient matrix and the interrelationships between elements are calculated.

Figure 1: Correlation Coefficient Matrix of Measured Concentrations of Six Types of Pollutants with Five Types of Meteorological Factors



4.3 Classification of Meteorological Conditions Based on Their Impact on Pollutant Concentrations

Based on the correlation coefficients among the 21 elements and their physical causes, the characteristics of various meteorological conditions and their impact on pollutant concentrations are analyzed.

The first category of meteorological conditions includes high-impact radiation factors, such as specific humidity, atmospheric pressure, longwave radiation, and 2m temperature.

The second category includes humidity factors, such as surface temperature, wind direction, rainfall, and relative humidity.

The third category consists of thermodynamic factors, including 10-meter wind size, sensible heat, latent heat, and boundary layer height.

The fourth category includes cloud cover, shortwave radiation, and surface radiation.

5 Problem Three: Prediction of Pollutant Concentrations and AQI Based on the Transformer Model

5.1 Analysis of Problem Three

CMAQ is a first-principle scientific computer model that comprehensively represents the most important processes affecting air quality and atmospheric chemistry, including emissions from various sources and transport by wind and precipitation events. CMAQ uses an extensive database of atmospheric chemical reactions to predict the chemical production and loss of hundreds of pollutants as they are transported downwind from their sources. In addition to gaseous substances, many pollutants are partially or entirely present in airborne particles, allowing them to interact with incoming solar radiation and clouds in complex ways.

A major challenge faced by models like CMAQ is capturing changes in anthropogenic emissions and weather patterns in the United States or globally. In addition to spatial variations, the CMAQ method must also handle emission and weather changes occurring over decades, entire seasons, or even within a single day. Despite these challenges, accurately representing the atmosphere and all its complexities in CMAQ is crucial, as the gases and particles included in the model have significant impacts on public health and the environment.

5.2 Preprocessing of Measured Data and Forecast Data from Meteorological Stations

5.2.1 Methods for Handling Missing Data

The missing data situation at Point A (Figure 7) reveals continuous missing humidity data at Point C for over half a year, as well as continuous missing values for certain periods and random missing values for individual time points. The proportions of missing data are 0.3%, 1.4%, and 3%, respectively.For these situations, this article uses meteorological factor scaling, simultaneous period ratio for pollutant concentrations, and multiple imputation to fill in missing values in the measured data, laying an ideal data foundation for the subsequent forecasting model. The methods for supplementing various missing values are explained

as follows:

(1) Continuous Long Sequence Missing Data: Data scaling based on the ratio of pollutant concentrations to climate in the same period.

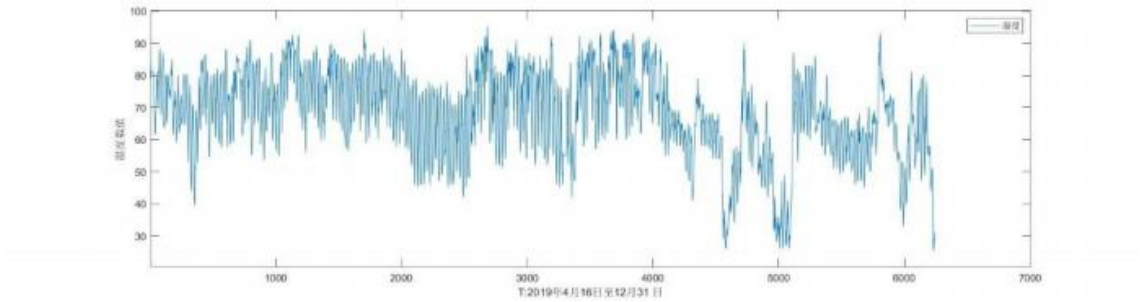
For long-time continuous missing sequences as shown in Figure 4, consider using data from the same period in previous or subsequent years to estimate the missing sequence. Specific method:

Due to reasons such as equipment calibration and maintenance, measured data may be partially or entirely missing. Meteorological indicators vary across monitoring stations due to equipment differences, and certain meteorological indicators cannot be obtained at some stations. As shown in Figure **??**, since the measured data are recorded hourly, this article supplements missing values using data from the same period in previous years and data before and after the missing values.

Figure 2: First Category of Missing Data, Horizontally Representing Time Dimensions, with Blanks Indicating Large Continuous Gaps

To exemplify using supplementary data C to compensate for missing humidity data and missing pollutant concentration and meteorological data at point A.

Figure 3: Supplementation Results for Missing Data at Point C

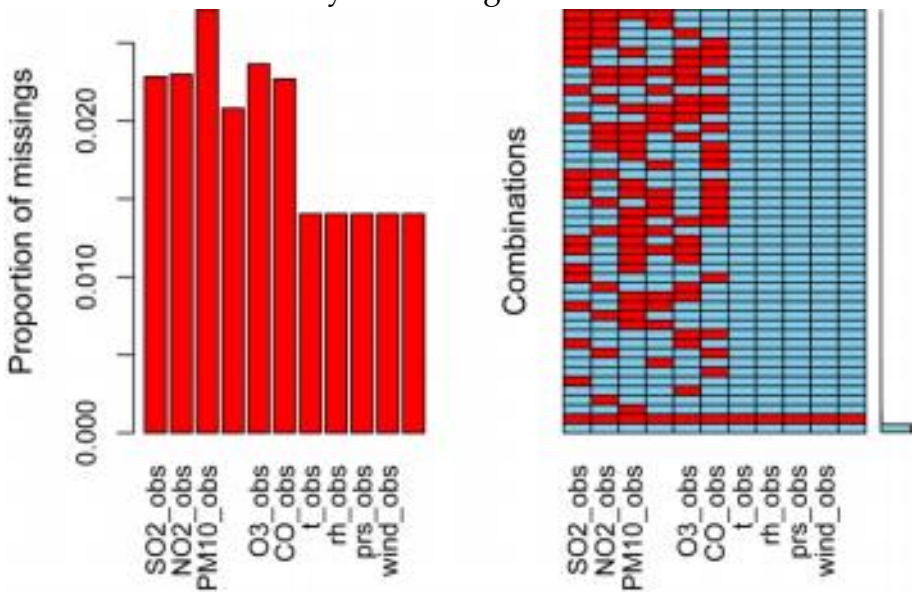


(2) Multiple Imputation For scattered missing data, as shown in Figure 6, this article adopts the multiple imputation method to fill in missing data. Multiple imputation is a method based on repeated simulation for handling missing values, generating a set of complete datasets from a dataset with missing values, where these missing values are replaced using the Monte Carlo method.

Figure 4: Second Category of Missing Data, Horizontally Representing Time Dimensions, with Blanks Indicating Scattered Gaps

Using the **mice** package in R, the data missing situation can be clearly analyzed, and random missing values can be identified and imputed. After supplementing the continuous missing sequences based on historical same-period ratio scaling, the article uses multiple imputation, adopting the **pmm** (predictive mean matching) method for variables with missing values. Preliminary trials showed that the pmm method is more efficient than methods like random forests, thus it was chosen to reduce computation time. For example, at Point A, the data missing situation is as shown in Figure 7, with a total missing data proportion of 2.28%.

Figure 5: Statistical Summary of Missing Data in Measured Data at Point A



5.2.2 Anomaly Detection and Handling

The filled data are then subjected to anomaly detection, anomaly filtering, and anomaly processing. The method for anomaly detection uses the **STL** time series anomaly detection method and **IQR** (Interquartile Range) for anomaly filtering, setting the maximum allowable proportion of anomalies at 2.5%. An example of anomaly filtering for SO_2 at Point A is shown in Figure, where red dots represent detected anomalies. Through these operations, robust data unaffected by anomalies are obtained.

(1) Time Series Decomposition Algorithm (STL) The Seasonal-Trend decomposition procedure based on Loess (STL) is a common algorithm in time series decomposition. Based on **LOESS**, it decomposes data at a certain time Y_v into a trend component, periodic component, and residual:

$$Y_v = T_v + S_v + R_v \quad v = 1, \cdots, N \tag{4}$$

STL is divided into inner and outer loops. The inner loop is mainly used for trend fitting and calculation of periodic components, with steps as follows:

The outer loop is mainly used to adjust robustness weights. If there are outliers in the data series, the residuals will be larger. Definition, for the data point at

position v, its robustness weight is, where the B function is the bisquare function:

$$B(u) = \begin{cases} (1-u^2)^2 & \text{for } 0 \le u \le 1 \\ 0 & \text{for } u \ge 1 \end{cases} \tag{5}$$

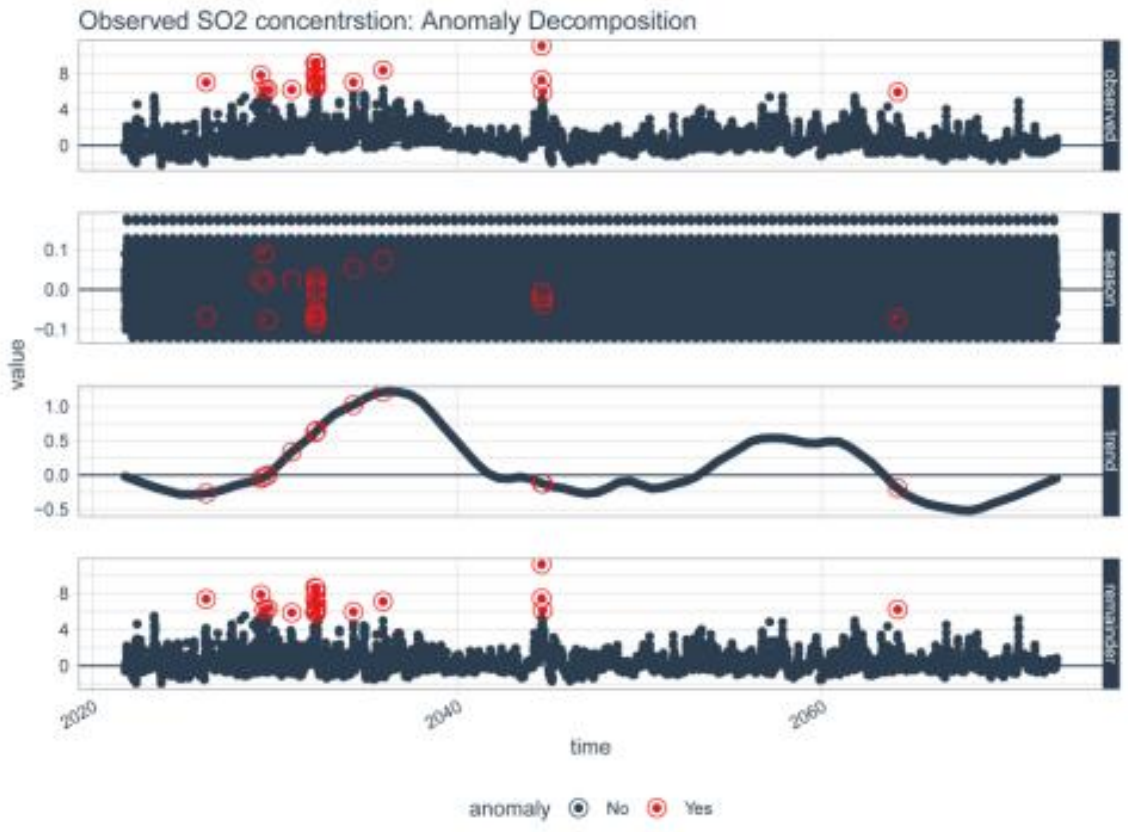
Then, in each iteration of the inner loop, when performing LOESS regression in Steps 2 and 6, the neighborhood weight needs to be multiplied by, to reduce the impact of outliers on the regression.

(2) Interquartile Range (**IQR**)

If the data are arranged in ascending order, the midpoint is called the median. Dividing the location of the median into two parts, the midpoints in these two segments are called quartiles. Thus, a set of data can be divided into three quartiles: the first, second (median), and third. The Interquartile Range (IQR) is the distance between the first and third quartiles.

The theoretical basis for using the IQR method for detection: If a data point is too far from the first and third quartiles, it may be an outlier. The method for handling anomalies in this article is to use the upper quartile limit to replace abnormal data. Using the anomalize package in R, meteorological element time series can undergo anomaly detection, filtering, and replacement. The effectiveness of anomaly detection for SO2 is shown in Figure.

Figure 6: Effectiveness of SO2 Anomaly Detection at Point A



5.3 Prediction of Pollutant Concentrations and Primary Pollutants at Point A Based on the Transformer Model

Based on primary forecast data and measured data, we calculate the relative error **MAPE** and **bias** values to assess the accuracy of the 72-hour forecasts in primary forecast data, aiming to identify error patterns in primary forecast data. This serves as a foundation for establishing a reasonable secondary forecast model (adaptively correcting regions with significant errors).

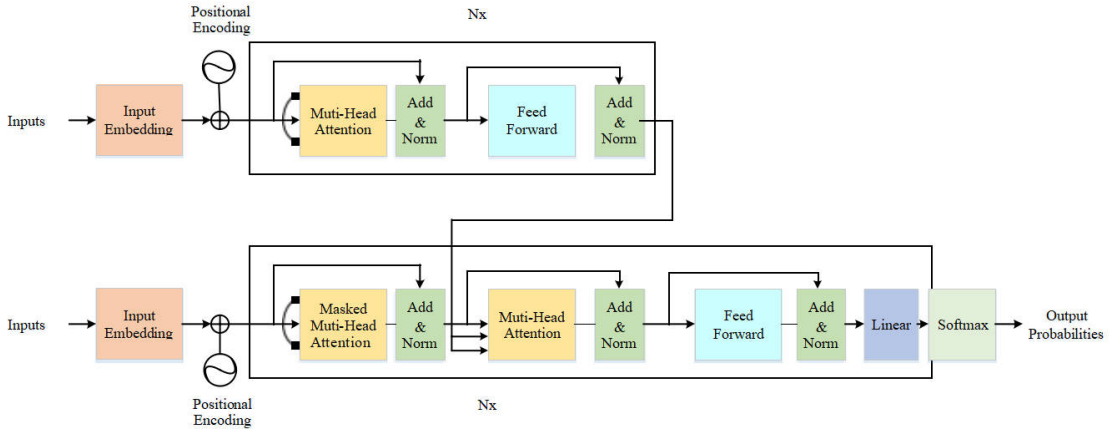
Pollutant concentration observation data are time series data. Hence, we consider using time series forecasting methods to predict future pollutant concentrations. Traditional time series forecasting models include **AR**, **MA**, **ARMA**, etc., which are more suitable for small-scale, single-variable forecasts. In data mining scenarios, due to the need for extensive parametric modeling, these methods are less applicable. In recent years, with the popularity of machine learning, applying machine learning methods to time series forecasting has achieved good results in many scenarios. Therefore, we consider using the **Transformer model** for pollutant concentration prediction.

5.3.1 Main Framework of the Transformer

The Transformer model was initially proposed by the Google team in 2017 and applied in machine translation. This model abandons the traditional recurrent neural network approach to extracting sequential information, innovatively introduces the attention mechanism, and achieves rapid parallelization, overcoming the slow training of recurrent neural networks. Essentially, the Transformer is an **Encoder-Decoder structure**. The Encoder part (as shown in the red box in the figure) consists of a series of encoders. The Decoder part (as shown in the blue box in the figure) is also made up of the same number of decoders. Each encoder has the same structure (but they do not share weights), with two sublayers: a self-attention layer and a fully connected feed-forward network. The input vectors to the encoder first pass through a self-attention layer, which helps the encoder focus on other features of the input vector while encoding each feature. The output of the self-attention layer is then passed to the feed-forward neural network.

The complete structure of the Encoder and Decoder is shown in the figure:

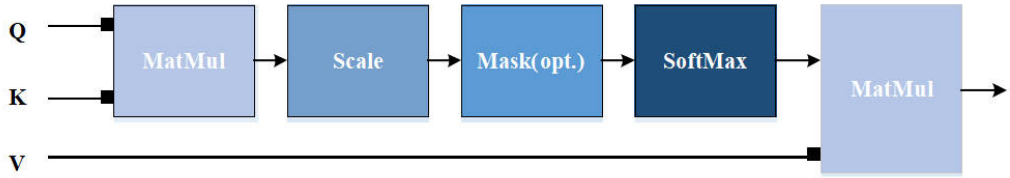
Figure 7: Structure of the Transformer Model



(1) Self-Attention Mechanism

The self-attention mechanism was first proposed in the field of natural language processing. The self-attention function can be described as mapping a query and a set of key-value pairs to output, where **Query**, **Key**,**Value** , and **output** are all vectors. The output is computed as a weighted sum of the values, with the weights assigned to each value determined by a compatibility function of the query with the corresponding key value.

Figure 8: Self-Attention Mechanism



The formulaic definition of Attention is:

$$\text{Attention}(Q, K, V) = \text{soft}\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Where $Q \subseteq R^{n\times d_k}, K \subseteq R^{m\times d_v}, V \subseteq R^{m\times d_v}$. Ignoring the activation function $\ell softmax$, it is essentially the multiplication of three matrices $n*d_k$, d_k*m,$m*d_v$, resulting in an $n*d_k$ matrix. This shows that it is an Attention layer, encoding the $n*d_k$ sequence Q into a new $n*d_v$ sequence.

5.3.2 Model Training

(1) Data Preparation

We used hourly pollutant concentration and meteorological measured data from points A, B, C to train the model. Due to a large amount of missing data

in the dataset, we preprocessed the data, applying XXX method for data supplementation. We extracted 70% of the data as the training set and the remaining 30% as the validation set.

(2) Model Training

During the training process, we used the preceding 840 hours of time-series meteorological elements (temperature, humidity, pressure, wind speed, wind direction) and time-series pollutant concentrations (SO_2 monitoring concentration, NO_2 monitoring concentration, PM_{10} monitoring concentration, $PM_{2.5}$ monitoring concentration, O_3 monitoring concentration, CO monitoring concentration) as feature inputs to the Encoder. The subsequent 72 hours of time-series pollutant concentrations were input into the Decoder, and the model output the predicted results for the next 72 hours of time-series pollutant concentrations. After training the model, we used the validation set to verify the effectiveness of the model, and the results are as follows.

To complete the prediction of pollutant concentrations for July 13 to July 15, 2021, we selected the previous 840 hours of time-series meteorological elements and time-series pollutant concentrations from the measured data as feature inputs to the Encoder. Lacking time-series pollutant concentration values for these three days, we assumed that the primary forecast data obtained from the WRF-CMAQ model has similar distribution characteristics to the measured data. Therefore, we used the predicted pollutant concentration values for July 13 to July 15 from the primary forecast data as feature inputs to the Decoder, resulting in the model prediction data, which is our required secondary.

5.3.3 Result Analysis

As shown in Figure 12, we present the error analysis graph for different pollutant concentrations at Point A. The blue curve represents the actual observed values, and the red curve represents the predicted values of our secondary forecasting model. It can be seen that our model's predictions match well with the actual observations, indicating that our model has good predictive ability.

Combining the analysis of Figure 12, we believe that our model performs better in predicting data with higher variance, while its performance is relatively weaker for more stable data. For this problem, we consider that our model is particularly effective in predicting SO_2, PM_{10}, and O_3 concentrations.

Figure 9: Error Analysis of Different Pollutant Concentrations at Point A

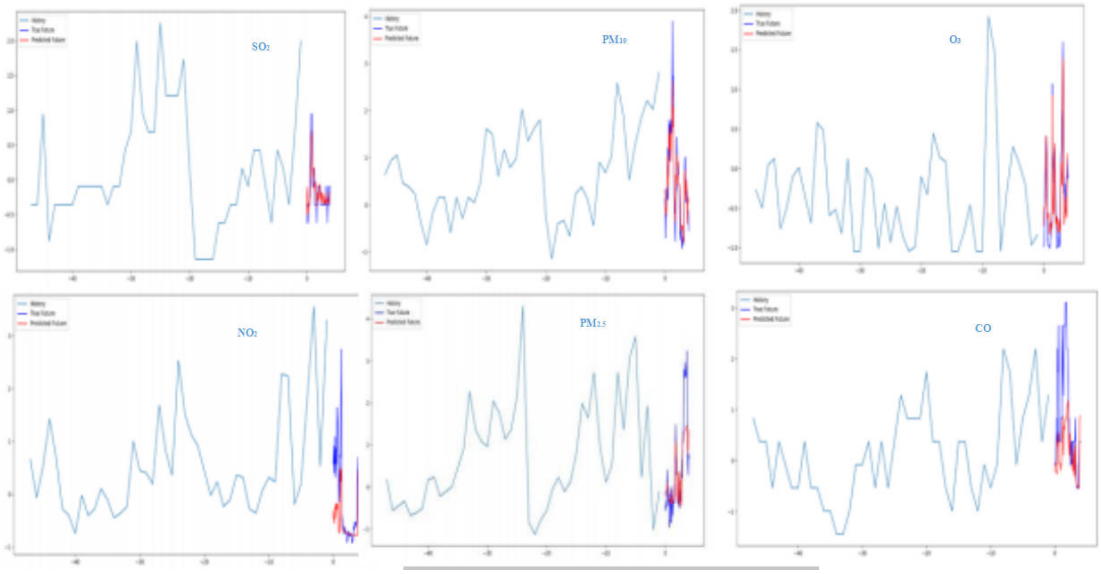


Figure 13 presents the error analysis graph for different pollutant concentrations at Point B. It can be seen from the figure that the model has significant errors, indicating that the model's predictive performance at Point B is average. However, the model still performs well in predicting O_3 concentrations.

Figure 10: Error Analysis of Different Pollutant Concentrations at Point B

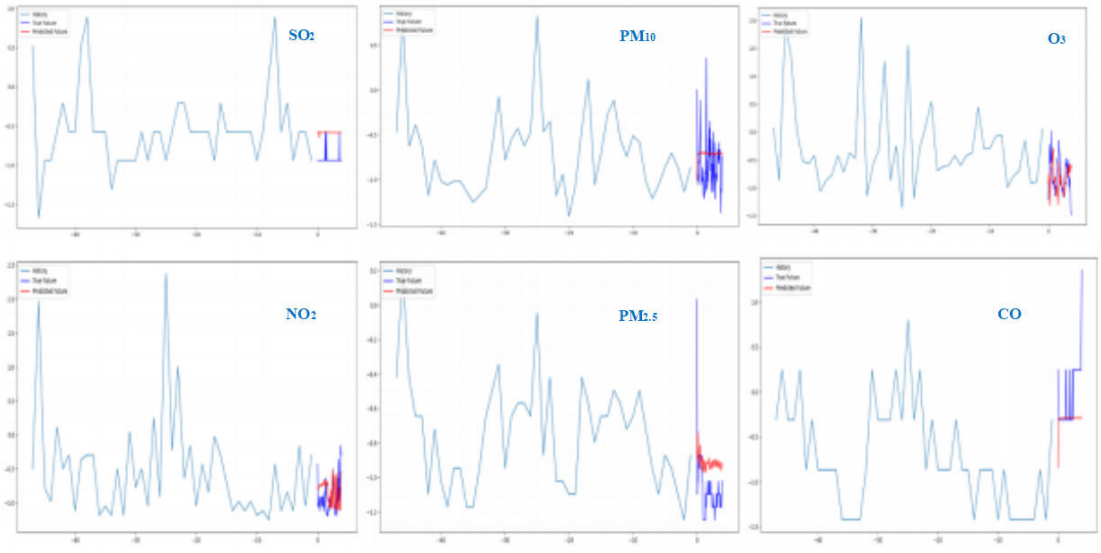
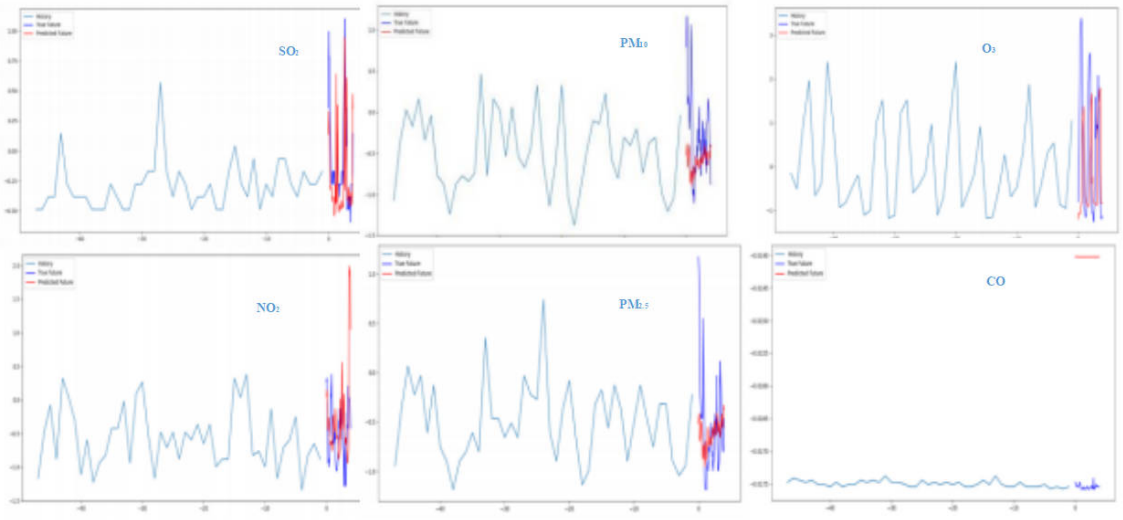


Figure 14 provides the error analysis graph for different pollutant concentrations at Point C. The figure shows that the model predicts well, demonstrating a certain robustness of our model.

Figure 11: Error Analysis of Different Pollutant Concentrations at Point C



In conclusion, we believe that our model has good robustness and high accuracy in predicting O_3 concentrations, making it a highly usable air quality forecasting model.

Forecast Date	Location	SO_2 ($\mu g/m^3$)	NO_2 ($\mu g/m^3$)	PM10 ($\mu g/m^3$)	PM2.5 ($\mu g/m^3$)	O_3 8hr max ($\mu g/m^3$)	CO (mg/m^3)	AQI	Primary Pollutant
2021/7/13	Monitoring Point A	4.07	13.97	22.48	9.13	13.95	0.53	23	NO_2
2021/7/14	Monitoring Point A	4.19	14.70	23.31	9.69	15.50	0.54	24	NO_2
2021/7/15	Monitoring Point A	4.97	19.76	29.10	13.58	26.30	0.59	30	NO_2

Table 5: Table 3: Daily Forecast of Secondary Model Values

Forecast Date	Location	SO_2 ($\mu g/m^3$)	NO_2 ($\mu g/m^3$)	PM10 ($\mu g/m^3$)	PM2.5 ($\mu g/m^3$)	O_3 8hr max ($\mu g/m^3$)	CO (mg/m^3)	AQI	Primary Pollutant
2021/7/13	Monitoring Point B	5.97	12.79	25.07	10.81	43.45	0.45	26	NO_2
2021/7/14	Monitoring Point B	5.98	12.82	25.11	10.84	43.52	0.45	26	NO_2
2021/7/15	Monitoring Point B	5.97	12.79	25.07	10.82	43.45	0.45	26	NO_2

Table 6: Table 6: Pollutant Concentration and Primary Pollutant Forecast Results at Point B from July 13 to July 15, 2021

Forecast Date	Location	SO_2 ($\mu g/m^3$)	NO_2 ($\mu g/m^3$)	PM10 ($\mu g/m^3$)	PM2.5 ($\mu g/m^3$)	O_3 8hr max ($\mu g/m^3$)	CO (mg/m^3)	AQI	Primary Pollutant
2021/7/13	Monitoring Point C	6.42	25.30	39.17	23.80	50.50	0.21	40	NO_2
2021/7/14	Monitoring Point C	6.29	25.06	38.76	23.51	49.89	0.20	39	NO_2
2021/7/15	Monitoring Point C	6.86	26.11	40.57	24.77	52.60	0.25	41	NO_2

Table 7: Table 7: Pollutant Concentration and Primary Pollutant Forecast Results at Point C from July 13 to July 15, 2021

6 Problem Four: Collaborative Forecasting Model Incorporating Nearby Stations

6.1 Analysis of Problem Four

For this problem, we continued to use the model from the previous problem. Considering that the pollutant concentrations in adjacent areas often have a certain correlation, and deep learning is driven by large data, we hypothesized that increasing data dimensions might improve the model's effectiveness. Therefore, we used the time-series meteorological elements (temperature, humidity, air pressure, wind speed, wind direction) and time-series pollutant concentrations (SO_2 monitoring concentration, NO_2 monitoring concentration, PM_{10} monitoring concentration, $PM_{2.5}$ monitoring concentration, O_3 monitoring concentration, CO monitoring concentration) from points A, A1, A2, and A3 as features input into the model. This model was then used to predict pollutant concentrations at Point A for July 13 to July 15, 2021.

6.2 Model Processing

In this model, our input features are 43-dimensional, and the output features are 72-dimensional. To simplify calculations, we randomly selected three sets of data to calculate the prediction errors for Point A's individual forecast and the collaborative forecast of Points A, A1, A2, and A3.

It can be seen that the collaborative forecasting model performs better than the individual point forecasting model. As we mentioned in the problem analysis, we suspect two reasons for this:

1. The meteorological elements input to our model include wind direction and wind speed data. If the wind directions at A1, A2, and A3 generally point towards Point A, and the wind speed at Point A is relatively low, forming a convergence trend, then pollutants are likely to accumulate at Point A.

2. From the model's perspective, the Transformer is a deep learning model that relies heavily on data. Increasing data volume and feature dimensions can potentially improve the model's prediction accuracy to some extent.

6.3 Forecast Effectiveness Evaluation

To evaluate the model's forecasting effectiveness, we selected the Mean Squared Error (**MSE**) as the test indicator. The Mean Absolute Percentage Error (**MAPE**) is commonly used to measure prediction accuracy, such as in time series forecasting.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{7}$$

We calculated the daily values of pollutant concentrations at Point A from July 13 to July 15, 2021. Finally, considering the model's structural errors, using

real-time observation values to correct the model's predictions helped improve the model's prediction accuracy.

7 Model Evaluation

7.1 Model Evaluation

7.1.1 Advantages of the Model

(1) Each layer's computational complexity is lower than that of recurrent neural networks.

(2) It can be calculated in parallel.

(3) Looking at the path length needed to compute information of a sequence of length n, convolutional neural networks need to increase convolutional layers to expand their field of view. Recurrent neural networks compute from 1 to n sequentially, while self-attention mechanisms only need one step of matrix computation. Self-attention mechanisms can solve long-term dependency problems better than recurrent neural networks.

(4) The attention model is more interpretable. The distribution of attention results indicates that the model has learned some syntactic and semantic information.

7.1.2 Disadvantages of the Model

(1) Practically: Some problems easily solved by recurrent neural networks cannot be handled by transformers.

(2) Theoretically: Transformers are not Turing-complete. These non-recurrent neural network models are not Turing-complete and cannot independently perform computational tasks such as reasoning and decision-making in NLP.

References

[1] Zhang Jinting, Zhao Yudan, Tian Yangge, He Qingqing, Zhuang Yanhua, Peng Yunxi, Hong Song. Study on the spatial non-synergistic coupling of atmospheric pollutant emissions and particulate matter environmental air quality - A case study of Wuhan City. Progress in Geography Science, 2019, 38(04): 612-624.

[2] Liu Yuan. Study on Air Pollution Index Forecast Based on BP Neural Network Model and ARMA-BP Combined Model [D]. Nanjing Normal University, 2018.

[3] Chen Shile, Wang Xiao, Zhou Changjun. Multi-factor Stock Forecasting Based on GA-Transformer Model. Journal of Guangzhou University (Natural Science Edition), 2021, 20(01): 44-55.

[4] Qian Weimiao, Chen Jing, Han Juncai, Cheng Xinghong, Wang Xiaomin. Study on Nonlinear Forecasting Model Based on WRF-CMAQ. Environmental Science and Technology, 2017, 40(08): 106-114.

[5] Sun Gao, Ning Ping, Shi Jianwu, Zhang Chaoneng, Zhong Yaoyan, Sun Baolei. Air Quality Forecast Based on Improved Time Series Statistical Model. Journal of Kunming University of Science and Technology (Natural Science Edition), 2017, 42(01): 91-97.

[6] Liu Qing, Li Dian, Wang Liwei, Xu Yaqi, Wu Yutong. Study on the Correlation Modeling of Air Pollutant Concentration Variation Characteristics and Meteorological Factors. Environmental Science and Management, 2021, 46(04): 136-140.

[7] Li Yun, Zhang Ying, Xu Jinping, Jiang Xiaomei, Yang Dan, Liang Mingzhu. Study on Heavy Pollution Weather Forecast Model and Circulation Characteristics in Tianmu Mountain Area. Meteorological and Environmental Sciences, 2021, 44(03): 54-60.

[8] Zhang Ying. Study on the Characteristics of Air Pollution and Its Health Impacts and Forecast in Typical Chinese Cities [D]. Lanzhou University, 2016.

[9] Zhao Yibing, Zhao Kaihui, Wang Ying, Dong Lifan, Wang Jinting. Relationship between Particulate Pollutant Concentration Distribution Characteristics and Meteorological Conditions in Xi'an City Area. Journal of Northwest Norma