Advances in Applied Mathematics 应用数学进展, 2023, 12(1), 54-61
Published Online January 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.121008

Hans汉斯

应用线性回归优化模型分析空气质量问题

刘　雨，米　娜，孙艺宁，孙菊贺

沈阳航空航天大学理学院，辽宁 沈阳

收稿日期：2022年12月15日；录用日期：2023年1月8日；发布日期：2023年1月17日

摘　要

针对基础的WRF-CMAQ空气预报模型的结果不理想的问题，通过对不同研究地点不同时间段的数据采集，并对官网数据的预处理，研究了在污染物排放情况不变的条件下，采用逐步回归算法分析了各气象条件与空气质量(AQI)之间的关系。根据逐步回归优化方法中需要考虑的因素，按照自变量X (气象条件)对因变量Y (AQI)作用的显著性大小，由大到小逐个引入方程中，最后根据所建立的最优回归方程的相关系数，得出影响地区空气质量的主要气象因素为气压、风速、湿度和温度，且皆与空气质量是正相关的。并且针对不同地点建立学习型线性回归优化模型，去实现对一次预报数据的修正，进而提高预报结果的准确性。

关键词

空气质量，空气预报模型，逐步回归法，学习型线性回归优化模型

Analysis of Air Quality Problems by Linear Regression Optimization Model

Yu Liu, Na Mi, Yining Sun, Juhe Sun

College of Science, Shenyang Aerospace University, Shenyang Liaoning

Received: Dec. 15th, 2022; accepted: Jan. 8th, 2023; published: Jan. 17th, 2023

Abstract

Aiming at the problem that the results of the basic WRF-CMAQ air prediction model are not ideal, the relationship between meteorological conditions and air quality (AQI) is analyzed by stepwise regression algorithm under the condition that the pollutant emission is unchanged through data

文章引用: 刘雨, 米娜, 孙艺宁, 孙菊贺. 应用线性回归优化模型分析空气质量问题[J]. 应用数学进展, 2023, 12(1): 54-61. DOI: 10.12677/aam.2023.121008

collection at different research sites and different time periods, and preprocessing of official website data. According to the factors that need to be considered in the stepwise regression optimization method and the significance of the independent variable (meteorological conditions) on the dependent variable (AQI), the equation is introduced one by one from large to small. Finally, according to the correlation coefficient of the established optimal regression equation, it is concluded that the main meteorological factors affecting the regional air quality are air pressure, wind speed, humidity and temperature, and they are all positively related to air quality. In addition, learning linear regression optimization model is established for different locations to correct the prediction data and improve the accuracy of the prediction results.

Keywords

Air Quality, Air Forecast Model, Stepwise Regression, Learning Linear Regression Optimization Model

1. 引言

随着中国工业化、城市化进程的快速发展，城市人口迅速膨胀，能源、交通规模持续扩大，产业结构不合理，能源消耗量大、利用率低等导致中国空气质量急剧下降。以可吸入颗粒物、二氧化硫、氮氧化物等为主要污染物的大气环境污染问题日趋严重，持续恶化的空气质量状况已经严重威胁到了公众的身体健康和国家经济的可持续发展。对空气质量进行全面、科学、准确的分析和预测，对于公众有效规避大气污染导致的健康损害，政府环保部门加强污染源监管和提高重污染日应急能力等方面都具有重要的理论意义和实用价值。

根据《环境空气质量标准》，用于衡量空气质量的常规大气污染物共有六种，分别为二氧化硫(SO_2)、二氧化氮(NO_2)、粒径小于 10 μm 的颗粒物(PM_{10})、粒径小于 2.5 μm 的颗粒物($PM_{2.5}$)、臭氧(O_3)、一氧化碳(CO)。其中，臭氧污染在全国多地区频发，对臭氧污染的预警与防治是环保部门的工作重点。因此，利用现有的实测数据和一次预报数据建立二次模型以提高臭氧预报的准确度是重难点之一。

全国多地的污染问题日趋严重，关注空气质量的人群越来越多，空气质量预报也显得愈发重要[1]。空气预报可采用的方法有多种，如文献[2]采用了逐步回归法进行了空气预报。文献[3]则利用 CMAQ 方法进行了空气质量的预测，而文献[4]使用了 CMAQ-MOS 空气质量预报模型。当实验数据有缺失时，文献[5]通过深度学习实现了空气质量的预测。由于污染物臭氧特殊性，文献[6]特别地采用了主成分分析法和逐步回归分析等方法确定了影响臭氧浓度的主要气象因素。文献[7]主要研究了 CMAQ 空气质量数值预报模式对区域性空气质量的预报准确度，通过对不同监测点的数据进行家聚类分析，以次进行预报的误差分析。文献[8]也提出了通过遗传算法以此提高空气质量预报的准确性。

但是目前国内常用的 WRF-CMAQ [2]模拟体系对空气质量预报的结果并不理想。本文首先介绍了空气质量指数(AQI)以及空气质量分指数(IAQI)的定义，数值以及计算方法，然后利用最优回归方程确定不同空气污染物的显著性大小，最后基于一次预报数据及实测数据进行空气质量预报二次数学建模，建立了学习型线性回归模型，在预报过程中使用实测数据对一次预报数据进行修正，优化了 WRF-CMAQ 预报模型，提高了空气预报的准确性。

2. 空气质量预报问题研究

2.1. 气象条件特征

空气质量指数 AQI 是目前开展空气质量监测各个国家用于通知群众污染程度的一个指标。本文中，根据空气质量指数对空气质量等级进行划分。如表 1 所示，AQI 的值越大，则对应的等级越高，也就意味着空气污染越严重，对人类自身健康危害程度越高。

Table 1. Air quality grade and range of air quality index
表 1. 空气质量等级及对空气质量指数(AQI)范围

空气质量等级	优	良	轻度污染	中度污染	重度污染	严重污染
空气质量指数(AQI)范围	[0, 50]	[51, 100]	[101, 150]	[151, 200]	[201, 300]	[301, +∞]

下表 2 则表示空气质量分指数(IQAI)不同标准时，所对应不同污染物的浓度限值。

Table 2. Air quality sub index and corresponding pollutant concentration limit
表 2. 空气质量分指数(IQAI)及对应的污染物项目浓度限值

序号	指数或污染物项目	空气质量分指数及对应污染物浓度限值								单位
0	空气质量分指数(IAQI)	0	50	100	150	200	250	300	350	-
1	一氧化碳(CO) 24 小时平均	0	2	4	14	24	36	48	60	mg/m^3
2	二氧化硫(SO_2) 24 小时平均	0	50	150	475	800	1600	2100	2620	
3	二氧化氮(NO_2) 24 小时平均	0	40	80	180	280	565	750	940	
4	臭氧(O_3)最大八小时滑动平均	0	100	160	215	265	800	-	-	$\mu g/m^3$
5	粒径小于等于 10 μm 颗粒物(PM_{10}) 24 小时平均	0	50	150	250	350	420	500	600	
6	粒径小于等于 2.5 μm 颗粒物($PM_{2.5}$) 24 小时平均	0	35	75	115	150	250	350	500	

首先搜集大量相关数据，将地点命名为监测点 A，那么为了计算监测点 A 从 2020 年 8 月 25 日到 8 月 28 日每天实测的 AQI 和首要污染物，我们将使用大气污染 P 的空气质量分指数(IAQI)的计算方法。如式(1)所示。

$$IQAI_{IP} = \frac{IQAI_{Hi} - IQAI_{L0}}{BP_{Hi} - BP_{L0}} \cdot \left(C_p - BP_{L0}\right) + IQAI_{L0}$$

其中，C_p 为污染物 P 的测量浓度值，BP_{Hi} 为与 C_p 相近的污染物浓度限值的高位值；BP_{L0} 为与 C_p 相近的污染物浓度限值的低位值； $IAQI_{Hi}$ 为与 BP_{Hi} 对应的 IAQI； $IAQI_{L0}$ 为与 BP_{L0} 对应的 IAQI。

根据所采集的污染物监测浓度数据，我们通过 $IAQI_{IP}$ 计算公式可以计算出所有监测种类污染物的 IAQI 值，所有 IAQI 值中数值最大的那一个作为 AQI 值，也即

$$AQI = \max\left\{IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{PM_{10}}, IAQI_{PM_{2.5}}, IAQI_{O_3}, IAQI_{CO}\right\}$$

环境空气质量主要取决于环境中污染物的排放情况和环境中大气的扩散能力[3] [4] [5]。其中，社会经济因素或直接或间接影响了所处环境中大气污染物的排放情况，气象因素主要决定了环境中大气的扩散能力。接下来在污染物排放情况不变的情况下分析各类气象条件对污染物浓度的影响程度。因此，在污染物排放相对稳定的情况下，气象条件对空气质量状况发挥主要作用。本文根据搜索的数据资料，从湿度、温度、气压、风速这四个气象条件着手，采用逐步回归算法[6] [7]分析各气象条件与空气质量(AQI)两者之间的相关关系。逐步回归算法是在全部需要考虑的因素中，按照它对因变量 Y 作用显著程度的大小，由大到小逐个引入回归方程。

则空气质量 Y 与预报因子 X 建立的最优回归方程为：

$$Y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_nX_n$$

其中，Y 为空气质量；B_0 是常数项 $X_1, X_2, \cdots, X_n$ 为选入预报因子，$B_1, B_2, \cdots, B_n$ 为选入的预报因子系数。

设当日 24 h 内的平均气压为 x_2 (MBAR)，平均风速 x_2 (m/s)，平均湿度为 x_3 (℃)，平均温度 x_3 (%)随机选取 40 个样本今复兴逐步回归分析，回归结果如表 3 所示。

Table 3. AQI and meteorological correlation coefficient
表 3. AQI 与气象相关系数

模型	未标准化系数		标准化系数	t	显著性
	B	标准误差	Beta		
1 (常量)	102.928	24.772		4.155	0.000
气压	−0.822	0.317	−0.401	−2.592	0.014
2 (常量)	408.687	73.330		5.573	0.000
气压	−2.122	0.396	−1.036	−5.364	0.000
风速	−6.906	1.592	−0.838	−4.337	0.000
3 (常量)	395.705	64.673		6.119	0.000
气压	−2.255	0.3511	−1.101	−6.432	0.000
风速	−5.221	0.4925	−0.633	−3.500	0.001
温度	−18.584	0.635	−0.447	−3.298	0.002
4 (常量)	2078.575	758.814		2.739	0.010
气压	−2.378	0.3361	−1.161	−7.081	0.000
风速	−5.341	0.4115	−0.648	−3.785	0.001
温度	−21.773	0.515	−0.524	−3.948	0.000
湿度	−1.658	0.745	−0.244	−2.225	0.033

从表 3 中可以看出，影响污染物浓度的有 4 个因子，即气压、风速、温度、湿度。sig 均小于 0.05，说明具有显著意义。同时根据表所示得到空气质量的逐步回归方程式为：

$$Y=-2.378x_1-5.341x_2-21.733x_3-1.658x_4+2078.575$$

通过以上方法，已经基本确定了在污染物排放情况不变的条件下，某地区影响空气质量的主要气象因素为气压、风速、湿度、温度，并且皆与空气质量存在正效应作用。

1) 温度对空气质量的影响

某一地区空气温度越高，必将引起空气体积的膨胀，近地面的大气对流增强，空气中的分子必要向周围地区扩散，那么空气质量就越好。反之，气温越低，近地面的大气对流减弱，不利于空气中的分子向四周扩散，随之而来必将导致污染物浓度增加，空气质量变差。

2) 湿度对空气质量的影响

某一区域空气中湿度越大时，空气中的水汽更易于吸附污染物，从而沉降，那么该地区的空气质量就会变好。反之，湿度越小越不利于可吸入污染物颗粒的沉降以净化空气。

3) 风速对空气质量的影响

某一区域风速越大时，空气流动越强，越有利于空气中的分子向四周扩散，那么大气污染物得以稀释扩散，空气质量越好。反之，风速越小，大气流动活动减弱，污染物聚集，那么空气质量必然变差。

4) 气压对空气质量的影响

低气压中心地带常常是阴雨天，当空气温度大降雨量少或不能形成降雨时，空气中的颗粒物几乎不能被净化，甚至可能导致颗粒物的吸温增长，造成污染物堆积，空气质量变差。

2.2. 二次预报模型

目前国内对空气质量预报常用 WRF-CMAQ 模拟系统，但是 WRF-CMAQ 预报模型的预报结果并不理想。该系统往往受制于模拟气象场以及排放清单的不确定性，以及包括臭氧在内的污染物生成机理的不完全明晰。因此，需要利用已搜集的 A，B，C 三个监测点实测数据和一次预报数据进行二次建模。为了二次预报数学模型预测结果中 AQI 预报值的最大相对误差尽量小，首要污染物预测准确度尽量高，本文采用学习型线性回归方法进行修正预报。

二次建模的主要思路为：以起报日前一定历史时期的污染物浓度预报效果作为根据，通过训练样本总结出各污染物浓度预报值与实际监测值之间的函数关系。倘若该函数关系具有一定程度的延续性，那么在日预报时间仅需根据各污染物浓度预报值与相应的函数关系式相结合，我们就可以得出最终的预报效果。模拟中采用一元线性回归的方法，在 MATLAB 中通过输入历史时期各污染物浓度预报值与实际检测值，训练出函数模型。由于污染物浓度的预报性可能随着时间的推移发生变化，因此合理的选择回归样本数且对于修正预报的效果起到至关重要的作用。由于回归时始终选择正确距离预报时刻相近的一段日期内的污染物浓度数据作为样本，因此二次预报对邻近日期的各污染物浓度的预报的准确度较高。

一元线性回归的求解步骤：

第一步：给定一个训练数据集，其中测数据 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，$x_i$ 是输入，表示实际监，y_i 是输出，表示一次预报数据。$(i=1,\cdots,N)$；□

第二步：把训练数据带入构建的模型，即函数 $y=ax+b$；

第三步：求损失函数(用来衡量预测值和真实值差距的公式)；

第四步：采用梯度下降法让损失函数最低；

第五步：对新的输入 x_{N+1}，预测系统根据学习的模型输出 y_{N+1}。

学习型线性回归模型建模思路如下图 1 所示。

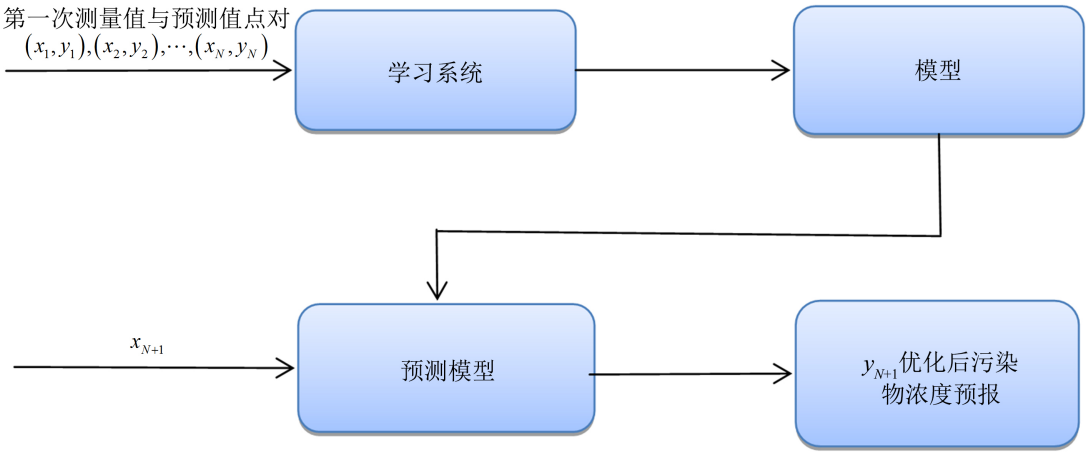

Figure 1. Thinking of learning linear regression model
图 1. 学习型线性回归模型思路

基于学习型线性回归 A 监测点 2021 年 7 月 13 日二氧化硫污染物浓度预测模型如图 2 所示。

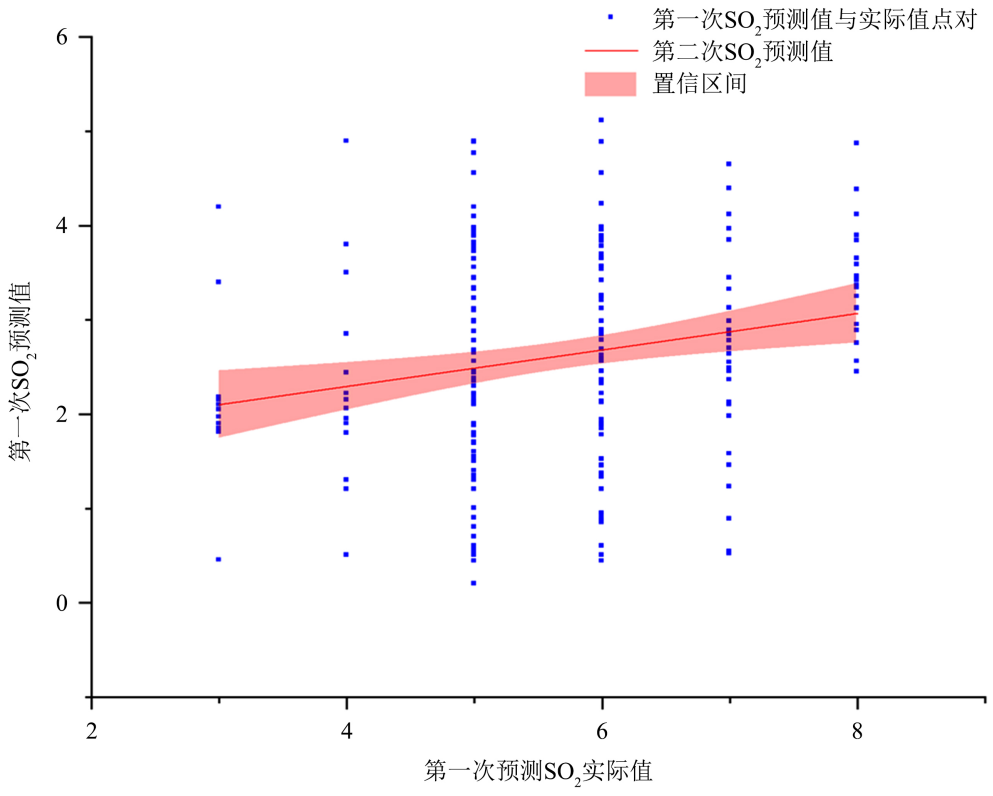

Figure 2. Prediction model of sulfur dioxide pollutant concentration
图 2. 二氧化硫污染物浓度预测模型

基于学习型线性回归 A 监测点 2021 年 7 月 13 日二氧化氮污染物浓度预测模型如图 3 所示。

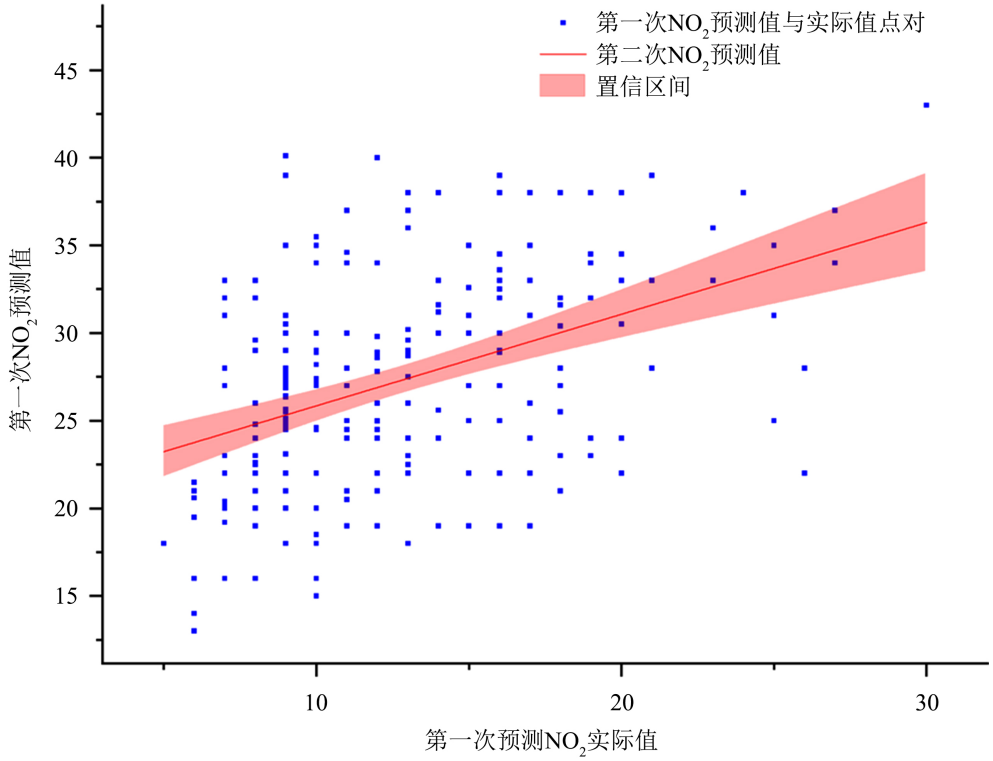

Figure 3. Prediction model of nitrogen dioxide concentration
图 3. 二氧化氮浓度预测模型

则根据训练模型各污染物浓度及 AQI 的预测结果如下表 4 所示。

Table 4. Prediction results of pollutant concentration and AQI
表 4. 各污染物浓度及 AQI 的预测结果

预报日期	地点	二次模型日值预测							
		SO_2 ($\mu g/m^3$)	$N0_2$ ($\mu g/m^3$)	PM_{10} ($\mu g/m^3$)	$PM_{2.5}$ ($\mu g/m^3$)	O_3 最大八小时滑动平均($\mu g/m^3$)	CO (mg/m^3)	AQI	首要污染物
2021/7/13	监测点 A	2.91	30.34	7.95	5.48	62.20	0.19	31.10	优
	监测点 B	0.15	6.67	8.13	4.23	33.40	0.09	17.70	优
	监测点 C	4.06	19.51	17.90	12.80	119.19	0.21	65.99	良
2021/7/14	监测点 A	2.91	28.72	6.13	4.24	62.25	0.20	35.9	优
	监测点 B	0.15	6.12	5.51	3.34	30.24	0.09	15.12	优
	监测点 C	3.54	20.13	19.02	13.89	118.97	0.172	68.81	良
2021/7/15	监测点 A	2.91	28.72	9.48	6.82	89.2	0.29	44.6	优
	监测点 B	0.12	5.52	5.72	3.12	53.8	0.18	26.9	优
	监测点 C	5.02	37.2	26.2	19.74	121.21	0.24	67.67	良

3. 模型评估

3.1. 模型的优缺点

3.1.1. 模型的优点

1) 采用逐步回归法可在所有需要考虑的因素中，按照自变量 X (各气象条件)对因变量 Y (AQI)影响的显著程度的大小，由大到小逐个引入回归方程。若某气象条件影响不显著，则将其从回归方程中删除，以保证在众多预报因子中选出最佳的预报因子组合，并且可根据选入的预报因子的相关系数判断各气象条件对空气质量的影响；

2) 建立的线性回归模型思想易理解，合理的选择回归样本可以训练出一个可靠的函数模型来描述数据之间的关系。那么新的输入出现时，可以得到一个修正后的污染物浓度预报值；

3) 模型比较简单，便于操作计算。

3.1.2. 模型的缺点

1) 本文所建立的模型对数据的依赖性很强，必须保证数据本身的准确性；

2) 本文在预测分析中仅考虑了气象条件的影响，忽略了社会经济因素的影响。以此影响了预测结果的全面性、准确性；

3) 没有讨论六种主要污染物在逐步回归模型中选入的最佳预报因子组合，也没有分析不同污染物的哪些气象条件对其浓度的影响起到了显著性作用；

4) 模型拓展性不高，容易受外界因素的影响。

3.2. 模型的推广与改进

在预测分析中加入社会经济因素，多影响因素模态下的空气质量预测可使预测结果更准确、更全面。

臭氧的预报可使用主成分析法与逐步回归法相结合，除了本文所使用的学习型线性回归模型还可使用 CMAQ-MOS 模型对 CMAQ 模式进行优化，从而提高预报的准确性。

基金项目

本篇文章受辽宁省教育厅面上项目的支持，项目编号为 LJKMZ20220524。

参考文献

[1] 卢亚灵, 李勃, 范朝阳, 王建童, 张鸿宇, 蒋洪强. 空气质量预测模拟技术演变与发展研究[J]. 中国环境管理, 2021, 13(4): 84-92.

[2] 谢敏, 钟流举, 陈焕盛, 陈多宏. CMAQ 模式及其修正预报在珠三角区域的应用检验[J]. 环境科学与技术, 2012, 35(2): 96-101.

[3] 刘闽, 王帅, 林宏, 许荣. 沈阳市冬季环境空气质量统计预报模型建立及应用[J]. 中国环境监测, 2014, 30(4): 10-15.

[4] 许建明, 徐祥德, 刘煜, 丁国安, 陈怀亮, 胡江凯, 张建春, 吴昊, 李维亮, 何金海, 杨元琴, 王佳禾. CMAQ-MOS 区域空气质量统计修正模型预报途径研究[J]. 中国科学(D 辑: 地球科学), 2005(S1): 131-144.

[5] 莫炜聪. 基于深度学习的空气质量预测研究[D]: [硕士学位论文]. 上海: 华东师范大学, 2022.

[6] 张灿, 蒋昌潭, 罗财红, 刘姣姣, 叶堤, 谯捷, 韩世刚. 气象因子对臭氧的影响及其在空气质量预报中的应用[J]. 中国环境监测, 2017, 33(4): 221-228.

[7] 王茜, 剑斌, 林燕芬. CMAQ 模式及其修正技术在上海市 PM_(2.5)预报中的应用检验[J]. 环境科学学报, 2015, 35(6): 1651-1656.

[8] 许洋, 顾海航. 基于遗传算法优化的 ELM 的空气质量预测研究[J]. 计算机时代, 2022(9): 73-77.