Modeling and Simulation 建模与仿真, 2022, 11(3), 648-662
Published Online May 2022 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2022.113061

Hans汉斯

空气质量预报二次建模

薛　浩，岳坤明，魏炳胜

上海理工大学机械工程学院，上海

收稿日期：2022年3月21日；录用日期：2022年5月10日；发布日期：2022年5月16日

摘　要

建立空气质量预报模型提前获知可能发生的大气污染状况，并采取相应控制措施是减少大气污染危害的有效方法之一。本文针对WRF-CMAQ等模型预报结果不理想的问题，提出了基于一次预报数据和实测数据的二次预报模型，提高了一次预报模型的准确性，分析了气象条件对污染物浓度的影响程度，提出了时间序列预测模型和基于粒子群优化的BP神经网络模型，大大地提高了预测的精度。

关键词

大气污染物，灰色关联度，相关分析，时间序列预测，BP神经网络，粒子群算法

Secondary Modeling of Air Quality Forecasting

Hao Xue, Kunming Yue, Bingsheng Wei

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Mar. 21st, 2022; accepted: May 10th, 2022; published: May 16th, 2022

Abstract

One of the effective ways to reduce the harm of air pollution is to establish an air quality forecast model to know the possible air pollution situation in advance and take corresponding control measures. Aiming at the problem that WRF-CMAQ and other models have unsatisfactory forecast results, this paper puts forward a secondary forecast model based on primary forecast data and measured data, which improves the accuracy of the primary forecast model, analyzes the influence of meteorological conditions on pollutant concentration, and puts forward a time series forecast model and a BP neural network model based on particle swarm optimization, which greatly improves the forecast accuracy.

文章引用: 薛浩, 岳坤明, 魏炳胜. 空气质量预报二次建模[J]. 建模与仿真, 2022, 11(3): 648-662.
DOI: 10.12677/mos.2022.113061

Keywords

Atmospheric Pollutants, Grey Correlation Degree, Correlation Analysis, Time Series Prediction, BP Neural Network, Particle Swarm Optimization

1. 引言

随着工业和经济的快速发展以及城市人口数量的增长，对能源的不断消耗使得排放出来的大量污染物严重污染了我们赖以生存的环境。人类对自然环境产生的破坏主要体现在土地、森林、水源、以及空气等方面[1]，其中对人类的生存发展影响范围最广的就是大气污染问题。由于环境质量恶化所造成的危害严重影响着人们的身体健康和生活方式，有研究表明 $PM_{2.5}$ 的污染特征对人类疾病的产生具有较高的相关性[2]。空气污染问题的防治问题迫在眉睫。大气污染问题已越来越引起人们和各国政府的重视[3]，由国务院召开的全国环境保护大会多次在北京召开，汇集众人的智慧为中国的生态环境治理提供解决方案，不断推动生态文明建设，解决生态环境问题，构建美丽家园。

我国的空气质量监测起步较晚，但是发展较为迅速，从环境保护领导小组、环境保护局、国务院环境保护委员会，然后将国家环境保护局作为国务院环境保护委员会的办事机构。国家在保护环境的重任面前，不断地克服困难，坚决打胜这场污染防治攻坚战。想要为空气污染防治提出更为有效的解决办法，对空气质量监测提出更高的要求[4]，为制定解决策略提供理论基础。

根据《环境空气质量标准》(GB3095-2012)，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)等[5]。污染防治实践表明，建立空气质量预报模型，提前获知可能发生的大气污染过程并采取相应控制措施，是减少大气污染对人体健康和环境等造成的危害[6]，提高环境空气质量的有效方法之一。

近些年来，国内外对于空气质量预测的研究多集中于空气污染物的浓度预测。Liu [7]提出了基于回归模型和支持向量机的空气质量预测模型，验证了组合预测模型的的预测效果较好。Zamani [8]利用了深度学习等方法预测来了 $PM_{2.5}$ 的数值，结果表明，相比于深度学习和随机森林方法，XGBoost 的预测误差最小。Bhat [9]提出了多变量回归与双变量线性模型对印度克什米拉阳地区的 $PM_{2.5}$ 数值，经过分析对比，双变量线性模型的预测效果更好。Ma [10]提出了基于网格重要等级的 XGBoost 的空气质量预测模型，预测结果的 R2 分数可以达到 0.9 以上。

目前对于大气的质量预报的研究方法主要分为两种，第一种方法是统计预报的方法，该方法的研究方法是通过大量地监测到的空气质量数据。通过统计学的分析方法，对收集到的数据进行分析，得到污染物浓度与气象条件之间的关系，通过建立预测模型得到空气质量的预报[11]。另外一种方法是研究大气的动力学理论，通过对复杂的大气物理、化学变化的分析，建立大气污染物浓度在空气中的传输扩散数值模型，通过较为复杂的计算过程，借助于计算机完成运算，得到多种大气污染物在空气中的动态分布。

2. 监测点 A 单日 *AQI* 值数学模型计算

2.1. 数据预处理

提取 2020 年 7 月 23 日 0 时至 2021 年 7 月 15 日实测数据进行预处理，异常值采用剔除和拉格朗日插值的方法进行处理，用均值法将样本数据从逐小时处理为逐日数据，

1) 数据异常情形

关于一次预报数据：预报工作中，服务器受外接电源长时间停电等情况影响，导致部分运行日期的一次预报数据缺失。

关于实测数据：1) 因监测站点设备调试、维护等原因，实测数据在连续时间内存在部分或全部缺失的情况；2) 受监测站点及其附近某些偶然因素的影响，实测数据在某个小时(某天)的数值偏离数据正常分布；本题提供的监测气象指标共计五项(温度、湿度、气压、风向、风速)，因不同监测站点使用设备存在差异，部分气象指标在某些监测站点无法获取。

2) 异常数据值处理

异常数据可能有多个来源，如数据本身、数据存储过程或者数据转换过程。由于异常数据会影响特征，也会影响最后的模型结果，因此对数据进行预处理十分必要。异常值的处理方法有多种，如删除记录、视为缺失值、平均值修正、不处理等。异常值如何处理，需要视具体应用背景分析而定。

2.2. 计算空气质量指数

根据《环境空气质量指数(*AQI*)技术规定(试行)》(HJ633-2012)，空气质量指数(*AQI*)可用于判别空气质量等级。

首先需得到各项污染物的空气质量分指数(*IAQI*)，其计算公式如下：

$$IAQI_P = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}} \cdot \left(C_P - BP_{Lo}\right) + IAQI_{Lo} \tag{1}$$

式中 $IAQI_P$ 为污染物 P 的空气质量分指数，C_P 为污染物 P 的质量浓度值，BP_{Hi}、BP_{Lo} 分别为与 C_P 相近的污染物浓度限值的高位值与低位值，$IAQI_{Hi}$、$IAQI_{Lo}$ 分别为与 BP_{Hi}、BP_{Lo} 对应的空气质量分指数。

空气质量指数(*AQI*)取各分指数中的最大值，即

$$AQI = \max\left\{IAQI_1, IAQI_2, IAQI_3, \cdots, IAQI_n\right\} \tag{2}$$

式中，$IAQI_1, IAQI_2, IAQI_3, \cdots, IAQI_n$ 为各污染物项目的分指数。在本节中，对于 AQI 的计算仅涉及六种污染物，因此计算公式如下：

$$AQI = \max\left\{IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{PM_{10}}, IAQI_{PM_{2.5}}, IAQI_{O_3}, IAQI_{CO}\right\} \tag{3}$$

3. 基于污染物浓度影响程度的气象条件分类

3.1. 问题分析

在某一地区，当假设污染量排放的浓度恒定不变的情况下，有些天气情况会直接影响该地区的 AQI 值。利用给定的监测点 A 的数据，分析各天气参数对该地区污染浓度的影响程度，并且对气象条件进行合理的分类，同时阐述各天气条件对该地区污染物浓度的影响特征。

分析天气条件对污染物浓度的影响，本节采用的是偏相关分析以及灰色关联分析，通过偏相关分析的控制变量的方法，不考虑天气条件之间的协同作用，单考虑单一因素对污染物浓度的影响，得到各天气条件与各种污染物的相关性，通过相关性对气象条件进行粗分类，分为对污染物浓度有正相关性或负

相关两类。接下来采用灰色关联度分析模型，通过比较气象条件与各污染物浓度的关联度，以此来比较各气象条件对某一污染物浓度的影响程度强弱。为清晰化表述，设计如图 1 的流程图展示：

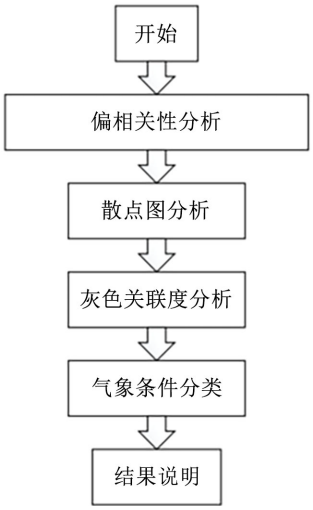

Figure 1. Problem solving flow chart
图 1. 解题流程图

3.2. 偏相关性分析模型

首先利用附件给的数据计算样本的偏相关系数，当控制第三个变量来研究另外的变量之间的相关性时，一阶相关系数的表达式为

$$r_{12(3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} \tag{4}$$

在实际应用中，本节的控制变量有四个，通过迭代法求取偏相关系数，迭代法认为简单的相关系数为 0 阶偏相关系数，任何 n 阶的偏相关系数都可以用 3 个$(n-1)$阶偏相关系数计算得到。

对变量间是否有净相关进行推断，采用 t 统计量作为偏相关分析的检验计量。其数学表达式如下定义：

$$t = \frac{r\sqrt{n-q-2}}{\sqrt{1-r^2}} \tag{5}$$

其中 r 为计算的偏相关系数，n 为样本数量，q 为阶数。

通过对给定的数据进行分析，可以得到如图 2 所示的气象条件与污染物浓度的相关性，部分相关系数如图 2 所示。

通过对数据进行分析，可以得到各气象条件与污染物浓度之间的相关性关系，将各天气条件对污染物浓度的相关性作图，可以得到不同气象条件对于大气污染物浓度的影响程度。

由表可知对二氧化硫而言，大气压、感热通量、潜热通量、长波辐射、近地 2 米温度与二氧化硫浓度成正相关关系，地表温度、比湿、近地 10 米风向、边界层高度与二氧化硫浓度呈负相关关系，且显著性符合要求。对二氧化氮而言，近地两米温度、湿度、云量、感热通量、潜热通量与二氧化氮浓度成正比，比湿、近地 10 米风速与二氧化氮含量浓度呈强负相关性，且显著性复合要求。对 PM_{10} 和 $PM_{2.5}$ 而言，地表温度、比湿、近地 10 米风速、雨量、长波辐射对粒径小的颗粒物有减弱作用，呈强负相关性，

相关性								
污染物	近地2米温度（℃）	地表温度（K）	比湿（kg/kg）	湿度（%）	近地10米风速（m/s）	近地10米风向（°）	雨量（mm）	…
SO2小时平均浓度（μg/m³）	0.181	-0.149	-0.198	-0.042	-0.201	-0.176	-0.011	…
NO2小时平均浓度（μg/m³）	0.071	0.014	-0.296	0.141	-0.349	-0.153	-0.022	
PM10小时平均浓度（μg/m³）	0.143	-0.117	-0.222	0.085	-0.247	-0.070	-0.041	
PM2.5小时平均浓度（μg/m³）	0.091	-0.072	-0.204	0.041	-0.170	-0.105	-0.023	
O3小时平均浓度（μg/m³）	-0.142	0.171	-0.094	0.104	0.144	0.086	-0.012	
CO小时平均浓度（mg/m³）	0.122	-0.069	-0.214	0.122	-0.267	-0.071	-0.022	

注：黄色为在 0.01 水平上显著(双尾)，绿色为在 0.05 水平上显著(双尾)

Figure 2. Correlation coefficient between meteorological conditions and pollutant concentration
图 2. 气象条件和污染物浓度相关性系数

而近地 2 米温度、湿度、大气压 、潜热通量对 PM_{10} 和 $PM_{2.5}$ 有促进作用，这些指数高的情况下会加重污染物浓度。

对臭氧而言，地表温度、湿度、近地 10 米风速、边界层高度、大气压、长波辐射与臭氧含量成正相关，且相关系数较大，而近地 2 米温度、感热通量对臭氧浓度呈负相关。对一氧化碳而言，近地 2 米温度、湿度、感热通量与一氧化碳浓度呈正相关性，且相关系数值较大，相关性显著，比湿、近地 2 米风速与一氧化碳浓度呈强负相关性。

3.3. 散点图分析模型

本节以监测点 A 逐小时污染物浓度与气象一次数据作为样本，将大气污染物浓度作为对比变量，将重要的气象条件变量与空气污染物浓度的关系用散点图进行分析，观察气象条件变量与空气污染物浓度之间是否存在一定关系，对指标进行了相关性分析。如图 3 所示是气象条件与二氧化硫浓度分布的散点图的部分结果。

从图 3 中可以看出，雨量、感热通量与二氧化硫浓度的散点图呈现 y 轴垂直的式样，这表示雨量、感热通量的变化并不能直接影响二氧化硫的浓度变化，从图中其他的点上也不能够观察出指数、对数、幂次的非线性关系。但我们可以看出散点集中于坐标轴左下方，说明二氧化硫浓度的分布集中于在雨量和感热通量较小环境下，而当雨量与感热通量较大时，二氧化硫浓度的散点分布在靠近横轴的区域。

3.4. 灰色关联度分类模型

首先参考数列的选取，以大气污染物浓度作为参考数列，记为 X_0，比较数列的选取，以气象条件作为比较数列，记为 X_i。在附件中给出的气象参数中，由于 CO 的浓度与其他的污染物浓度的单位不同，需要对给出的数据进行无量纲化处理。这里采用极差值法对数据进行无量纲处理，以此来消除量纲给分析结果带来的影响。极差值法消除量纲的方式如下：

$$令\ x'_{ij}=\frac{x_{ij}-m_j}{M_j-m_j}\left(i=1,2,3,\cdots,18411;\ j=1,2,3,4,5\right) \tag{6}$$

接下里利用关联度系数的计算公式计算关联度系数，公式为：

$$\delta_i(k)=\frac{\min_i\min_k\left|x_i(k)-x_0(k)\right|+\rho\max_i\max_k\left|x_i(k)-x_0(k)\right|}{\left|x_i(k)-x_0(k)\right|+\rho\max_i\max_k\left|x_i(k)-x_0(k)\right|} \tag{7}$$

$\left|x_i(k)-x_0(k)\right|$ 为 $x_i(k)$ 和 $x_0(k)$ 在第 k 个点的绝对误差，$\min_i\min_k\left|x_i(k)-x_0(k)\right|$ 为两级最小值，ρ 为分辨率，$0<\rho<1$，在解决该问题时选取 ρ 为 0.5，ρ 与分辨率成反比，ρ 越大分辨率越小，反之则越大。

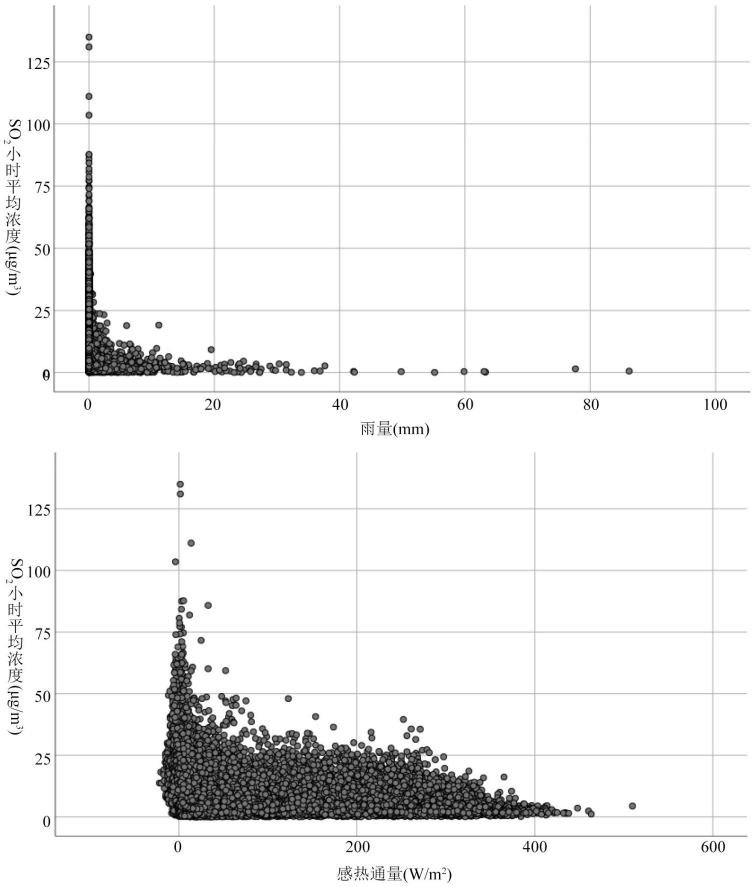

Figure 3. Scatter diagram of partial meteorological conditions and SO_2 concentration distribution
图 3. 部分气象条件与 SO_2 浓度分布散点图

最后再根据关联度计算公式来计算其关联度，关联度计算公式如下所示：

$$r_i = \frac{1}{n}\sum_{k=1}^{n}\delta_i(k) \tag{8}$$

式中 r_i 为 x_i 对 x_0 的关联度。

根据灰色关联度分析，得到相应的关联度数排名结果如图 4 所示。

以气象条件与污染物浓度的灰色关联度数为 0.5 作为刻度，将气象条件分为两类，分别是对二氧化硫影响较大的气象条件和对二氧化硫影响不大的气象条件。各气象因子对污染物浓度的关联程度排序为：雨量 > 感热通量 > 潜热通量 > 短波辐射 > 地面太阳能辐射 > 边界层高度 > 近地 10 米风速 > 云量 > 大气压 > 比湿 > 地表温度 > 湿度 > 长波辐射 > 近地 2 米温度。由图可知大气压、云量、近地 10 米风向、近地 10 米风速、边界层高度、地面太阳能辐射、短波辐射、潜热通量、感热通量、雨量对二氧化硫浓度影响较大，且雨量相较于其余气象条件变量，影响效果更大。近地 2 米温度、长波辐射、湿度、地表温度、比湿对二氧化硫的浓度的影响效果并不明显，所以将其归为第二类。

3.5. 距离相关验证分析

在统计学中，皮尔逊相关系数是用来度量两个变量之间的线性相关，其值在−1 和 1 之间。可以用来判断数据样本是否集中分布在同一条线上，用来衡量定距变量之间是否存在特定的线性关系。而样本

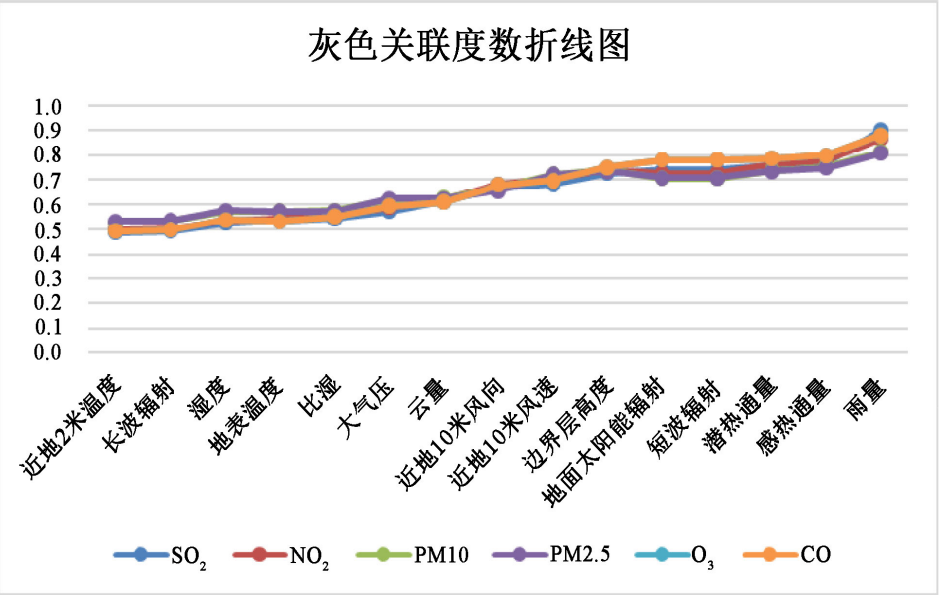

Figure4. Line chart of grey correlation degree
图 4. 灰色关联度折线图

之间线性关系不明显，因此皮尔逊相关性分析并不适合本体的数据样本。距离分析可以分为个案间和变量间举例分析两种，分析的方法有相似性和不相似性分析两种。距离分析客服了传统的相关分析的缺点，能够分析非线性数据变量之间的关系，对于两个随机变量 x，y 而言，距离相关系数可以定义为

$$R^2(x,y)=\frac{v^2(x,y)}{\sqrt{v^2(x,x)v^2(y,y)}} \tag{9}$$

其中：

$$\begin{aligned} v^2(x,y)&=\frac{1}{n^2}\sum_{i,j=1}^{n}A_{i,j}B_{i,j}\\ A_{i,j}&=\left\|x_i-x_j\right\|_2-\frac{1}{n}\sum_{k=1}^{n}\left\|x_k-x_j\right\|_2-\frac{1}{n}\sum_{l=1}^{n}\left\|x_i-x_l\right\|_2+\frac{1}{n^2}\sum_{k,l=1}^{n}\left\|x_k-x_l\right\|_2\\ B_{i,j}&=\left\|y_i-y_j\right\|_2-\frac{1}{n}\sum_{k=1}^{n}\left\|y_k-y_j\right\|_2-\frac{1}{n}\sum_{l=1}^{n}\left\|y_i-y_l\right\|_2+\frac{1}{n^2}\sum_{k,l=1}^{n}\left\|y_k-y_l\right\|_2 \end{aligned} \tag{10}$$

同理：

$$\begin{aligned} v^2(x,x)&=\frac{1}{n}\sum_{i,j=1}^{n}A_{i,j}^2\\ v^2(y,y)&=\frac{1}{n}\sum_{i,j=1}^{n}B_{i,j}^2 \end{aligned} \tag{11}$$

本节所测量的变量如表 1 所示：

最后我们采用距离相关系数对分类效果进行验证，相关系数热力图如图 5 所示，由图可知，近地 2 米温度、长波辐射、湿度、地表温度、比湿这几个变量的相关性均在 0.7 以上，说明对大气污染物浓度影响不明显的一类中变量相关性显著，而其余变量之间的相关性较小，符合上述分类结果，故分类效果较好。

Table 1. Changes measured
表 1. 测量的变量表

排名	名称	排名	名称
1	近地 2 米温度	9	边界层高度
2	地表温度	10	大气压
3	比湿	11	感热通量
4	湿度	12	潜热通量
5	近地 10 米风速	13	长波辐射
6	近地 10 米风向	14	短波辐射
7	雨量	15	地面太阳能辐射
8	云量	/	/

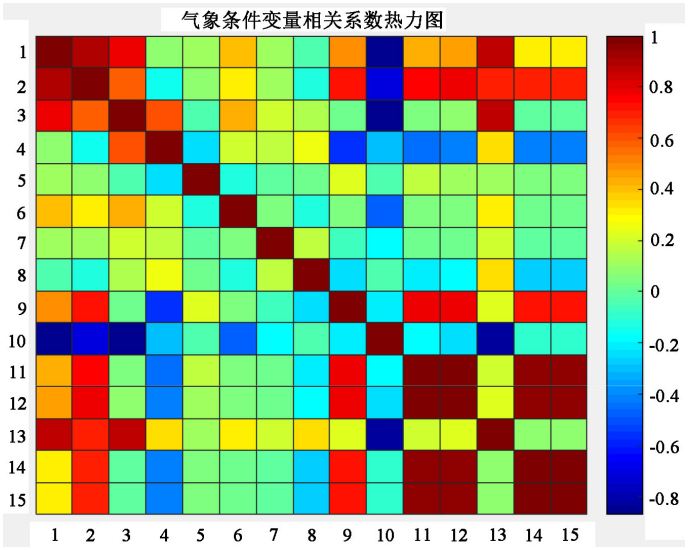

Figure 5. Heat diagram of correlation coefficient of meteorological condition variables
图 5. 气象条件变量相关系数热力图

4. 大气污染物浓度预测

4.1. 大气污染物浓度趋势时间序列分析

ARIMA 模型是差分模型、自回归模型和移动平均模型的结合，差分模型是为了实数剧更加平缓。

1) 自回归模型 AR

自回归模型是利用因素自身的历史数据对自身进行预测，预测只能在满足平稳性要求的情况下才能进行，平稳性检测采取 ADF 指标进行检验。

自回归模型表达式如下：

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{p} \gamma_i y_{t-i} \tag{12}$$

式中，y_t 表示当前值，μ 为常数，p 为阶数，γ_i 为自相关系数，ϵ_t 为误差。

2) 移动平均模型 MA 移动平均模型是为了消除自回归模型的随机波动即误差 ϵ_t，其表达式如下：

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{13}$$

如图 6 所示，时间序列模型的流程可总结如下：

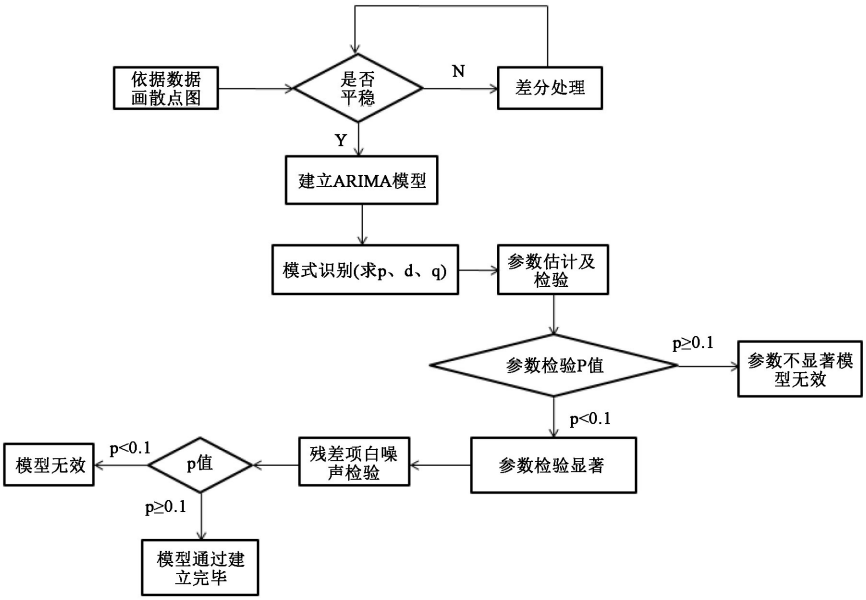

Figure 6. Flow chart of time series model
图 6. 时间序列模型流程图

本文利用 SPSS 软件对数据进行时间序列预测，发现污染物浓度的规律有明显的季节性，根据时间序列预测模型，本文通过两年中每个月的同一天的污染物浓度变化规律设计时间序列模型，得出三天的污染物浓度变化数值，部分数据趋势图如图 7 所示。

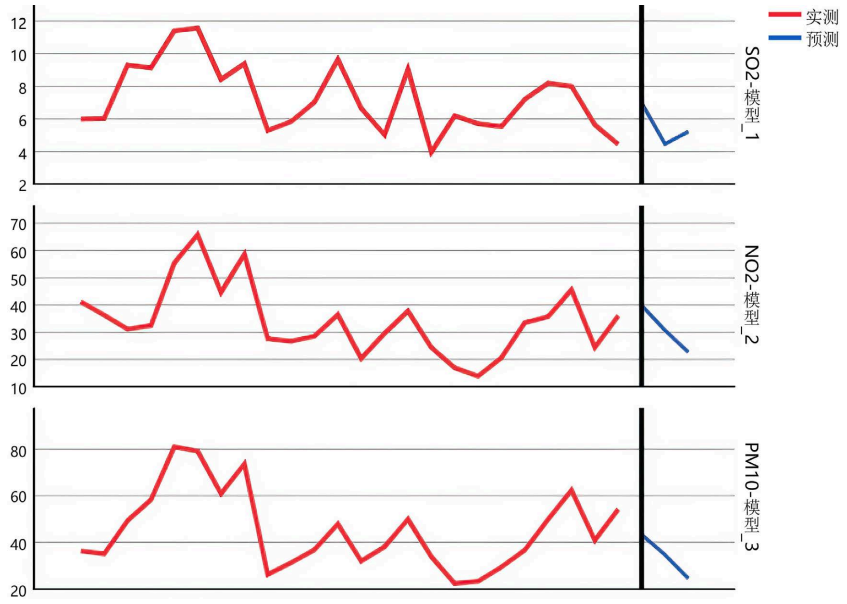

Figure7. 3-day concentration change trend diagram
图 7. 三天浓度变化趋势图

对与大气污染物浓度的时间序列预测，从总体而言建立的时间序列预测模型效果较好，由于预测事件较短，预测出来的结果符合本节关于时间序列图中的大气污染物浓度趋势，在 7 月份夏季时部分污染物浓度处于低值，符合预期，通过观察时间序列预测的结果可以发现预测结果与历史周期规律高度一致，但是严格来说由于大气污染物浓度的变化和气象条件的变化具有时间滞后性，所以在现实应用中参考性不大，无法保证其准确性，因此本文使用 BP 神经网络预测模型加以参考，提高预测结果的可靠性。

4.2. 大气污染物浓度趋势 BP 神经网络分析

二次空气质量的预测模型，采用的是多层感知器的误差反向传播算法，也就是所说的 BP 神经网络，具体的学习步骤如图 8 所示。

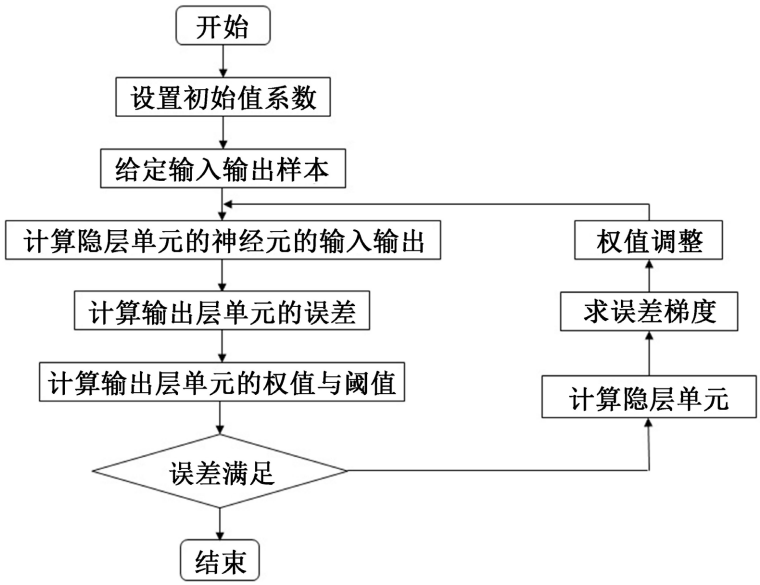

Figure 8. BP neural network learning flow chart
图 8. BP 神经网络学习流程图

建立 BP 神经网络模型时，将样本通常分为训练集和测试集，训练集用来构建 BP 神经网络的模型结构，测试集用来分析和检验模型的实际效果，一般通常以百分之八十和百分之二十分配数据样本个数，通常在设计神经网络结构时，隐含层层数不超过两层，隐含层层数设置过多不仅会减慢其收敛速度还会对模型最后的预测精度造成一定影响，本文采取三层经典神经网路结构：输入层，隐含层，输出层。对于节点个数的选取，本文选取四个节点用来设计神经网络模型。

1) 训练数据和测试数据确定

对于附件一中的数据，取监测点 A 逐小时污染物浓度与气象一次预报数据作为神经网络训练数据，将监测点 A 逐小时污染物浓度与气象实测数据的后 50 行的气象数据作为参数输入到网络中，得出后 50 行的六种污染物的浓度值，并将此数据与 A 点逐小时的实测数据进行误差分析，验证二次模型的可行性。基于此设定，先将数据进行归一化处理：

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{14}$$

2) 参数设定

本研究设计的 BP 神经网络图设计 4 个参数，分别为迭代次数、隐含层元素个数、训练目标最小误差和学习速率。由于设计数据较大，本研究设计的迭代次数为 10,000，训练目标最小误差为 0.000001。然后采用控制变量法分别讨论隐含层元素个数和学习速率的设定，以要求六种污染物浓度的相对误差为最小。首先设定学习速率为 0.1，隐含层个数分别为 20、45、60、90，然后求得在不同隐含层下各污染物的总偏差。通过对比污染物的总偏差大小和偏差均值来确定隐含层元素的个数。

从图 9 可以看出当隐含层元素个数为 20 时总偏差较为平稳且偏差均值最低。因此设置隐含层的个数为 20。在此基础上，分别设定学习速率为 0.1、0.4、0.7、0.9。

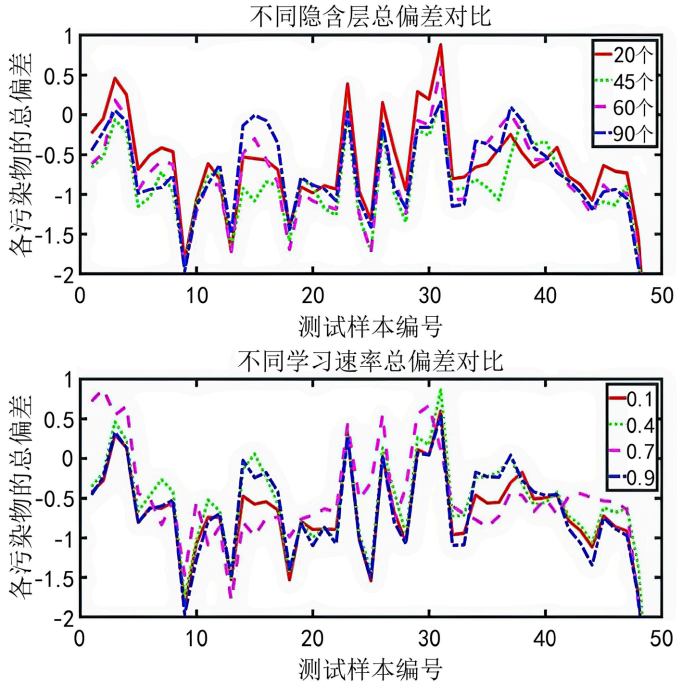

Figure 9. Comparison of total deviation of different hidden layers and total deviation of different learning rates
图 9. 不同隐含层总偏差与不同学习速率总偏差对比图

结合图 9 可以看出，当学习速率为 0.7 时污染物的总偏差平稳且偏差均值为最小。综上所述，该神经网络的迭代次数为 10,000，训练目标最小误差为 0.000001，隐含层元素个数为 20 且学习速率为 0.7。

在设定好的神经网络下，将附件 1 中的监测点 A 逐小时污染物浓度与气象实测 2021/7/13 日之前的数据输入到神经网络中进行训练，将监测点 A 逐小时污染物浓度与气象一次预报 2021/7/13~2021/7/15 的五个气象数据(温度、湿度、气压、风速、风向)作为参数输入到神经网络中，来输出 7/13~7/15 逐小时的六种污染物浓度。最后求取 3 天逐小时的污染物平均浓度作为 3 天逐日的污染物浓度并且算出 *AQI*。

将两种模型得出的结果进行对比可以发现，时间序列模型的短期预测结果一般，最大相对误差较大，BP 神经网络的预测结果较好，最大相对误差较小，但浓度分布均处在正常范围之内，因此可以接受，在条件允许的情况下，可综合考虑时间预测模型和 BP 神经网络预测模型的预测结果。

5. 粒子群算法优化

粒子群优化算法对神经网络的调整过程如下所示：1) 对参数运用网络自身进行二次优化，直到得到

最佳的权值和阈值之后停止搜索。2) 粒子群替代梯度下降法对神经网络权值、阈值进行优化，直到计算出来的适应度值无法继续下降迭代才会停止。神经网络中的权值和阈值数据会被粒子群算法充当粒子的位置向量，通过寻求粒子群的最佳位置同时寻找最佳权值和阈值，再利用 BP 神经网络的正向传播计算粒子群算法的适应度。

算法流程图如图 10 所示。

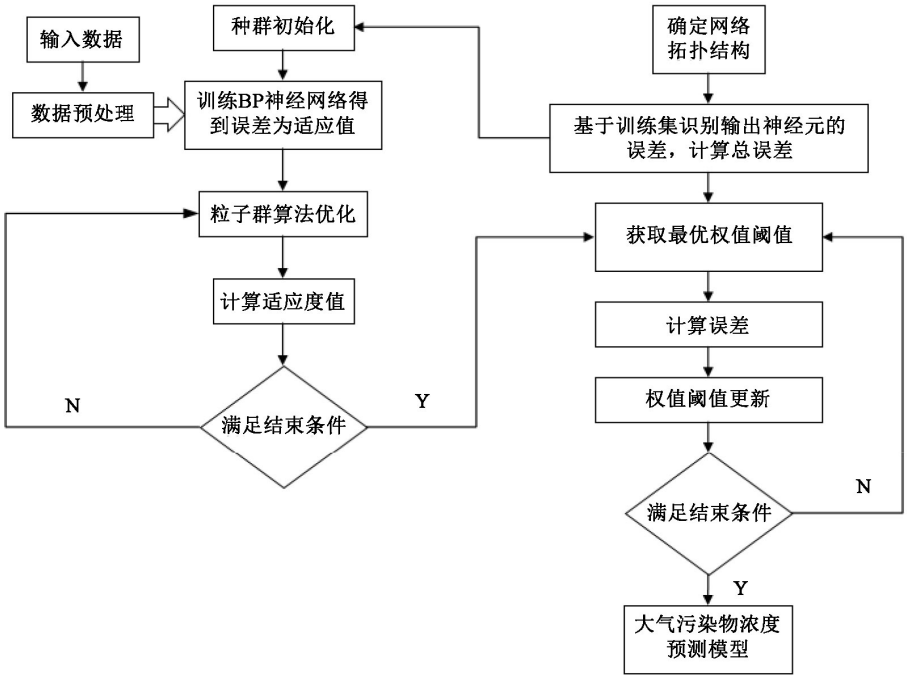

Figure 10. Logical diagram of BP neural network prediction model based on particle swarm optimization
图 10. 基于粒子群优化的 BP 神经网络预测模型逻辑图

将修正好的一次预报数据，作为训练集输入到基于粒子群优化的 BP 神经网络中，将监测点逐小时的实测数据的后 50 行的气象条件作为参数输入到网络中，得到后 50 行的六种污染物浓度的预测值，并与逐小时的实测数据进行对比，如下图 11 所示。

由图 11 可知，由于对一次预报数据进行了修正及对 BP 神经网络进行了粒子群算法的优化，优化后的模型预测精度更高，准确度更好。

6. 模型评价

如图 12 所示为不同模型下六种污染物浓度的总偏差，其中经过粒子群优化的 BP 神经网络总偏差较小且相对未优化的 BP 神经网络更加平稳。

为了便于比较，取均方根(*RMS*)为性能指标：

$$\|e\|_{RMS} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\|e(k)\|^2} \tag{15}$$

由表 2 可知，均方根值随着数据的增大而增大，且已优化的 BP 神经网络预测出的六种浓度的均方根值远远小于未优化的 BP 神经网络预测值。

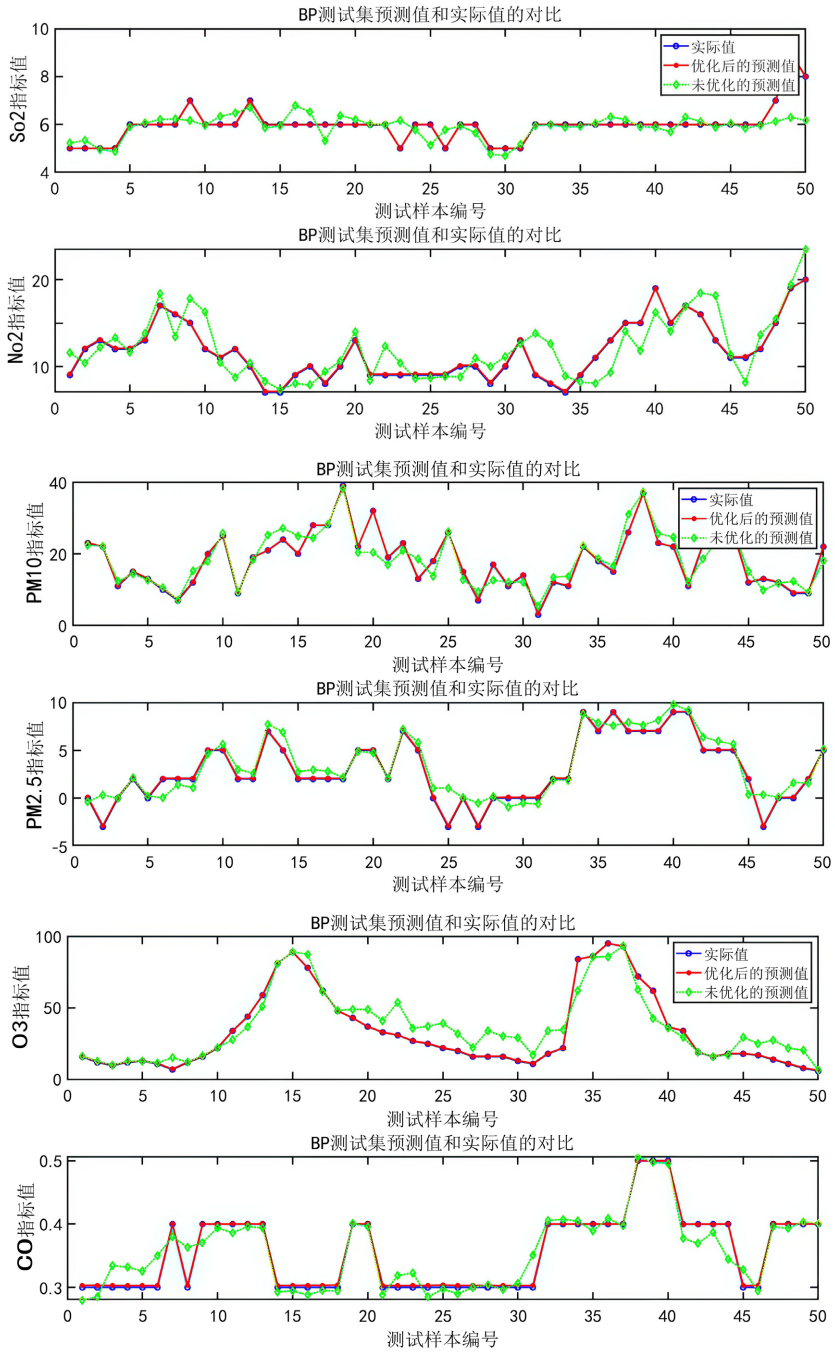

Figure 11. Comparison of pollutant concentration prediction of different models
图 11. 不同模型的污染物浓度预测对比图

7. 结语

1) 根据大气污染物浓度变量的取值范围，剔除不在范围的变量数据样本；对数据缺失值运用拉格朗日插值法进行填充。根据预处理后的数据，通过给定的环境空气质量分指数公式计算监测点 A 处各项污染物的 *IAQI* 值，再取污染物浓度计算得到的最大值作为最终单日的 *AQI* 值，最大值的污染物作为首要污染物。

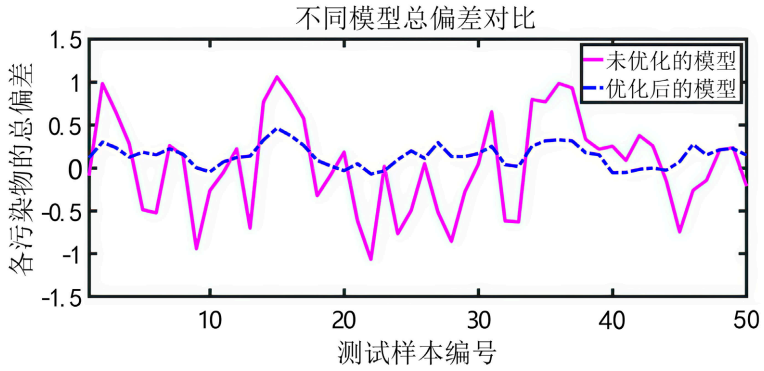

Figure 12. Comparison of total deviation of pollutants in different models
图 12. 不同模型的污染物总偏差对比图

Table 2. Root-mean-square table of pollutant concentration prediction of different models
表 2. 不同模型的污染物浓度预测均方根表

性能指标	未优化的 BP 神经网络	已优化的 BP 神经网络
SO_2	0.3631	0.0104
NO_2	1.1392	0.0919
PM_{10}	2.9766	0.1192
$PM_{2.5}$	1.2462	0.0776
O_3	9.7693	0.1870
CO	0.0209	0.0023

2) 根据偏相关系数分析气象条件之间的相关关系，再利用散点图分析气象条件对污染物浓度的影响规律。利用灰色关联度分析气象条件与污染物浓度之间的关联度，以关联度的强弱将气象条件分成两类。最后通过距离相关得出不同类别气象条件之间的相关系数热力图验证分类的准确度。

3) 本文首先采用时间序列预测模型根据历史数据对 7 月 13 日至 7 月 15 日的数据进行初步预测，再利用 BP 神经网络模型进行污染物浓度的预测。然后将一次预报数据作为训练集输入到经过粒子群算法优化的 BP 神经网络中，得到三点的二次预报数据，利用该数据修正 A 点的二次预报数据，采用 BP 神经网络的预测方法得到预测值和实测值的误差，得出二次模型预测效果良好。

本文在一次预测模型的结果的基础上，研究表明 BP 神经网络的预测效果要大于时间序列预测效果。结合更多的空气数据对预测模型进行二次建模，二次建模的结果准确性大大优于原始的预测结果，利用粒子群算法对 BP 神经网络的预测模型进行优化，能够不断对参数进行更新与修正，从而实现对空气质量的准确预测，预测结果有一定的鲁棒性，对后续空气质量预测有一定的研究意义。

参考文献

[1] 陈啸龙. 浅析环境监测在大气污染治理中的作用[N]. 江苏经济报, 2021-10-14(B03).

[2] 陈道新, 李江敏, 张瑞哲, 欧阳雨婷. 防治大气污染 保卫蓝天白云[N]. 南昌日报, 2021-10-14(006).

[3] 忽建永, 钱雪莹, 殷文涛, 黄奕. 近 3 年西双版纳州勐腊县大气污染基本特征及污染原因分析[J]. 环境科学学报, 2021(11): 4388-4395.

[4] 江思力, 李文学, 步犁, 吕嘉韵, 冯文如, 杨轶戬. 广州市 2020 年 $PM_{2.5}$ 的污染特征与居民循环系统疾病的相关

性[J]. 中国热带医学, 2021(12): 1144-1149.

[5] Zhang, H., Ji, Y.Y., Wu, Z.H., Peng, L., Bao, J.M., Peng, Z.J. and Li, H. (2022) Atmospheric Volatile Halogenated Hydrocarbons in Air Pollution Episodes in an Urban Area of Beijing: Characterization, Health Risk Assessment and Sources Apportionment. *Science of the Total Environment*, **806**, Article No. 150283. https://doi.org/10.1016/j.scitotenv.2021.150283

[6] Ullah, S., Ullah, N., Rajper, S.A., Ahmad, I. and Li, Z.Q. (2021) Air Pollution and Associated Self-Reported Effects on the Exposed Students at Malakand Division, Pakistan. *Environmental Monitoring and Assessment*, **193**, Article No. 708. https://doi.org/10.1007/s10661-021-09484-2

[7] Liu, B., Jin, Y. and Li, C. (2021) Analysis and Prediction of Air Quality in Nanjing from Autumn 2018 to Summer 2019 Using PCR-SVR-ARMA Combined Model. *Scientific Reports*, **11**, Article No. 348. https://doi.org/10.1038/s41598-020-79462-0

[8] Zamani, M. (2019) $PM_{2.5}$ Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere*, **10**, Article No. 373. https://doi.org/10.3390/atmos10070373

[9] Bhat, M.A., Romshoo, S.A. and Beig, G. (2021) Measurement and Modelling of Particulate Pollution over Kashmir Himalaya, India. *Water, Air, & Soil Pollution*, **232**, Article No. 120. https://doi.org/10.1007/s11270-021-05062-x

[10] Ma, J., Cheng, J.C.P., *et al*. (2020) Identification of the Most Influential Areas for Air Pollution Control Using XGBoost and Grid Importance Rank. *Journal of Cleaner Production*, **274**, Article No. 122835. https://doi.org/10.1016/j.jclepro.2020.122835

[11] 闫艳. 完善观测方法提高预报预警能力[N]. 中国环境报, 2013-09-02(002).