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

题 目　　空气质量预报的二次建模及相关问题研究

摘　　要：

在大气环境污染对人体及生态环境有重要影响的今天，采取相应污染防治控制措施十分必要。通过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造成的危害、提高环境空气质量的有效方法之一。我国目前常用的是 WRF-CMAQ 模拟体系对空气质量进行预报，但受制于模拟的气象场以及排放清单的不确定性，以及对包括臭氧在内的污染物生成机理的不完全明晰，大气物理化学机理分析还存在缺陷，使得首要污染物、空气质量等级命中率低，WRF-CMAQ 预报模型的结果并不理想。因此在 WRF-CMAQ 等一次预报模型模拟结果的基础上，结合更多的数据源进行再建模，以提高预报的准确性。

对于问题 1，监测点 A 从 2020 年 8 月 25 日到 8 月 28 日逐日实测的数据中并无异常数据，可直接按照附录中的方法计算监测点 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，结果如表 3-4,3-6 所示。

对于问题 2，将附件 1 中的数据采集记录并针对记录进行**数据预处理**，主要是针对数据异常情形，采取如下处理方式：实测数据部分或全部缺失的情况：变量值缺失值取为前后时刻的平均值；受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：若超出正常变量分布范围则取为对应的最小值或最大值；利用 3σ 准则将异常数据剔除。基于数据预处理的变量值，根据公式计算每天对应的 AQI 值，根据其属性利用基于划分的聚类方法 **k-means 聚类**根据对污染物浓度的影响程度，对气象条件进行聚类，对聚类结果进行可视化；在 AQI 结果排序的基础上对气象条件进行合理分类，并阐述各类气象条件的特征；并对整体数据进行多次聚类，最终根据聚类结果，选择分为三类，聚类中心点如表 4-1 所示，阐述各类气象条件的特征。

对于问题 3，根据前问气象条件对污染物浓度的影响筛选出来的具有代表性及独立性的气象条件，建立了基于 BP 神经网络预测框架。该模型搭建神经网络框架对一次预报数据及实时监测数据输入完成空气质量预报的预测，并在加入更新天气质量数据的过程中不断调整，该模型适用于数据更新的预测建模。经过一系列测试，二次预报数学模型同时适用于 A、B、C 三个监测点，并且建立的 **BP 神经网络二次预报模型**预测结果中 AQI 预报值比一次建模的预报值的误差更小，首要污染物预测准确度更高。此外，对空气质量预测的二次建模模型结果与真实值对比并于其他模型相比较验证了模型的有效性。并使用该模型预测监测点 A、B、C 在 2021 年 7 月 13 日至 7 月 15 日 6 种常规污染物的单日浓度值，给出相应的 AQI 和首要污染物的结果，如表 5-3 所示。

对于问题 4，根据问题 3 中建立的 BP 神经网络预测框架，加入考虑风向及监测点 A、A1、A2、A3 之间的位置，改进模型构造**协同预报模型**，利用**遗传算法**采用实数编码、模拟二进制交叉和多项式变异，对 BP 神经网络预测模型进行行改进，对在测试集的使用上也进行更改，取相同时间段内四个不同监测点测试的污染物浓度和气象信息作为训练集，训练神经网络，结果如表 6-1 所示。

关键词：空气质量预报，首要污染物，AQI，k-means 聚类，二次建模，BP 神经网络，遗传算法

目录

1. 问题重述

空气污染物排放的时空变化和相应的气象条件的变化能够对人体健康产生短期和长期影响，污染物浓度超标对居民健康有着巨大的危害，对生态、环境和交通等人类社会活动等产生巨大的影响。一个有效的空气质量预报系统有助于人类掌握污染物未来浓度信息，制定相应的防治策略，对空气中污染物的浓度水平提前给出精确的预报，使因污染物浓度超标所造成空气污染物排放的时空变化和相应的气象条件的变化能够对人体健康产生短期和长期影响，污染物浓度超标对居民健康有着巨大的危害，对生态、环境和交通等人类社会活动等产生巨大的影响[2]。

目前常用 WRF-CMAQ 模拟体系(以下简称 WRF-CMAQ 模型)对空气质量进行预报。WRF-CMAQ 模型主要包括 WRF 和 CMAQ 两部分：WRF 是一种中尺度数值天气预报系统，用于为 CMAQ 提供所需的气象场数据；CMAQ 是一种三维欧拉大气化学与传输模拟系统，其根据来自 WRF 的气象信息及场域内的污染排放清单，基于物理和化学反应原理模拟污染物等的变化过程，继而得到具体时间点或时间段的预报结果。WRF 和 CMAQ 的结构如图 1-1、图 1-2 所示：

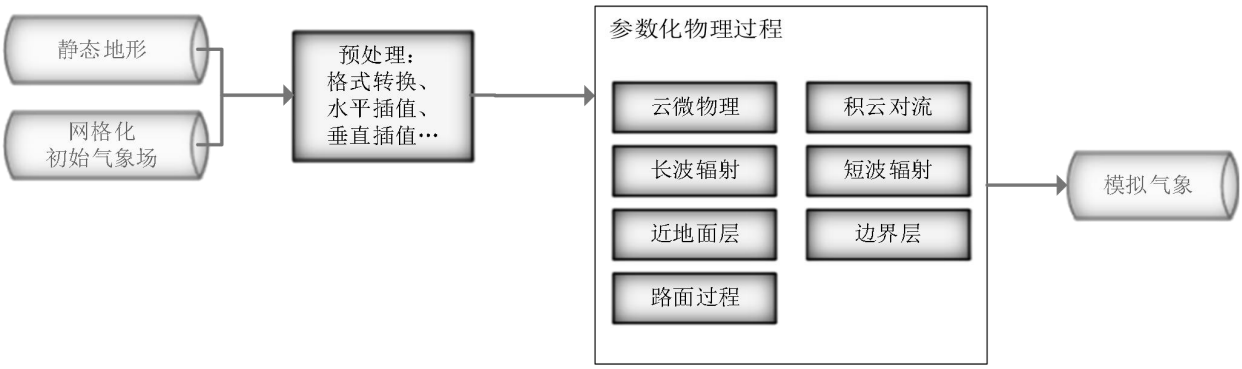


图 1-1 中尺度数值天气预报系统 WRF 结构[1]

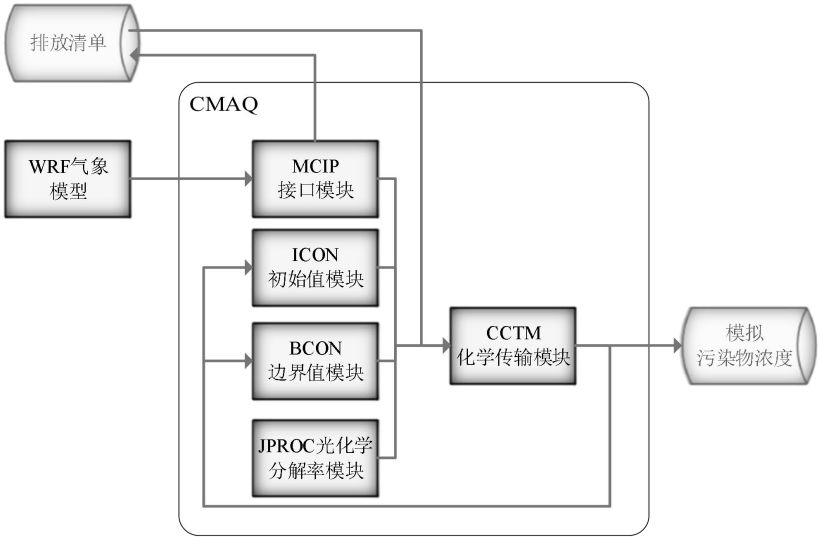


图 1-2 空气质量预测与评估系统 CMAQ 结构

但受制于模拟的气象场以及排放清单的不确定性，以及对包括臭氧在内的污染物生成机理的不完全明晰，WRF-CMAQ 预报模型的结果并不理想。故题目提出二次建模概念：即指在 WRF-CMAQ 等一次预报模型模拟结果的基础上，结合更多的数据源进行再建模，以提高预报的准确性。其中，由于实际气象条件对空气质量影响很大(例如湿度降低有利于臭氧的生成)，且污染物浓度实测数据的变化情况对空气质量预报具有一定参考价值，故目前会参考空气质量监测点获得的气象与污染物数据进行二次建模，以优化预报模型。

二次模型与 WRF-CMAQ 模型关系如图 1-3 所示。将 WRF-CMAQ 模型运行产生的数据简称为“一次预报数据”，将空气质量监测站点实际监测得到的数据简称为“实测数据”。一般来说，一次预报数据与实测数据相关性不高，但预报过程中常会使用实测数据对一次预报数据进行修正以达到更好的效果。

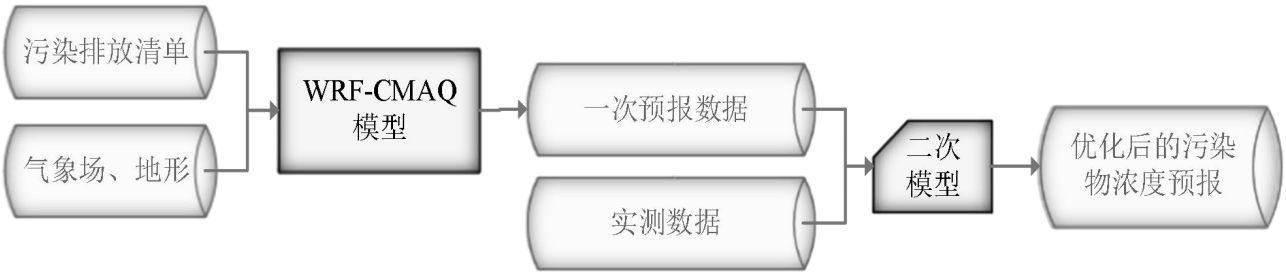


图 1-3 二次模型优化的 WRF-CMAQ 空气质量预报过程

为进行二次建模以预测给定监测点未来三天的空气质量情况，题目提供了监测点长期空气质量预报基础数据，包括污染物浓度一次预报数据、气象一次预报数据、气象实测数据和污染物浓度实测数据，其中，所有一次预报数据的时间跨度为 2020 年 7 月 23 日到 2021 年 7 月 13 日，所有实测数据的时间跨度为 2019 年 4 月 16 日 到 2021 年 7 月 13 日，数据总量在十万量级(详见附件 1—3)。

需要注意的是：(1)每日预报的时间固定为早晨 7 点，此时可以获得当日 7 时及之前时刻的实测数据，以及运行日期在当日及之前日期的一次预报数据(预报时间范围截至第三日 23 时)。监测时间在当日 7 时以后的逐小时实测数据和运行日期在次日及以后的一次预报数据都是无法获得的，例如：在 2021 年 7 月 13 日晨间对 7 月 13 日至 7 月 15 日的空气质量进行预报过程中，可供参考的实测数据时间范围为(2019 年 4 月 16 日 00：00 至 2021 年 7 月 13 日 7：00)，模型运行日期范围为(2020 年 7 月 23 日至 2021 年 7 月 13 日)。(2)受监测数据权限及相应监测设备功能等的限制，部分气象指标的实测数据无法获得。(3)由于一次预报对邻近日期的准确度较高，故理论上二次预报对邻近日期的准确度也较高。

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)。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。臭氧浓度预报也是六项污染物预报中较难的一项，其原因在于：作为六项污染物中唯一的二次污染物，臭氧并非来自污染源的直接排放，而是在大气中经过一系列化学及光化学反应生成的(参考附录一种近地面臭氧污染形成机制部分)，这导致用 WRF-CMAQ 模型精确预测臭氧浓度变化的难度很高；同时，国内外已有的研究工作尚未得出臭氧生成机理的一般结论[2]。

综上所述，根据问题要求，基于一次预报数据及实测数据(见附件)进行空气质量预报二次数学建模，建立一个具有一定的鲁棒性的模型，完成以下四个问题。请注意，实际工作

中会遇到数据为空值或异常值的情况(见附录)，故要求建立的模型具有一定的鲁棒性。

问题 1. 使用附件 1 中的数据，按照附录中的方法计算监测点 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，按照附录“AQI 计算结果表”的格式给出结果。

问题 2. 在污染物排放情况不变的条件下，某一地区的气象条件有利于污染物扩散或沉降时，该地区的 AQI 会下降，反之会上升。使用附件 1 中的数据，根据对污染物浓度的影响程度，对气象条件进行合理分类，并阐述各类气象条件的特征。

问题 3. 使用附件 1、2 中的数据，建立一个同时适用于 A、B、C 三个监测点(监测点两两间直线距离>100km，忽略相互影响）的二次预报数学模型，用来预测未来三天 6 种常规污染物单日浓度值，要求二次预报模型预测结果中 AQI 预报值的最大相对误差应尽量小，且首要污染物预测准确度尽量高。并使用该模型预测监测点 A、B、C 在 2021 年 7 月 13 日至 7 月 15 日 6 种常规污染物的单日浓度值，计算相应的 AQI 和首要污染物，依照附录“污染物浓度及 AQI 预测结果表”的格式给出结果。

问题 4. 相邻区域的污染物浓度往往具有一定的相关性，区域协同预报可能会提升空气质量预报的准确度。如图 4，监测点 A 的临近区域内存在监测点 A1、A2、A3，使用附件 1、3 中的数据，建立包含 A、A1、A2、A3 四个监测点的协同预报模型，要求二次模型预测结果中 AQI 预报值的最大相对误差应尽量小，且首要污染物预测准确度尽量高。使用该模型预测监测点 A、A1、A2、A3 在 2021 年 7 月 13 日至 7 月 15 日 6 种常规污染物的单日浓度值，计算相应的 AQI 和首要污染物，依照附录“污染物浓度及 AQI 预测结果表”的格式在文中给出结果。并讨论：与问题 3 的模型相比，协同预报模型能否提升针对监测点 A 的污染物浓度预报准确度？并说明原因。

2. 模型假设及关键性符号说明

(1)假设空气质量检测站点实时监测所采集数据可以良好反映此时刻的空气质量情况；
(2)假设设备检测过程中所采集的正常数据都是准确的；
(3)假设通过数据找出的首要污染物浓度可以反映出空气质量等级；
(4)假设每改变一次首要污染物的值，空气质量情况和情况都会相应及时的发生变化。

符号说明：

符号	意义
AQI	空气质量指数；
C_{O_3}	臭氧（O_3）最大 8 小时滑动平均；
$IAQI_P$	污染物 P 的空气质量分指数，结果进位取整数；
C_P	污染物 P 的质量浓度值；
BP_{Hi}	与C_P相近的污染物浓度限值的高位值；
BP_{Lo}	与C_P相近的污染物浓度限值的低位值；
$IAQI_{Hi}$	与BP_{Hi}对应的空气质量分指数；
$IAQI_{Lo}$	与BP_{Lo}对应的空气质量分指数。

3. 问题一分析与求解

3.1 问题一分析

题目提供了监测点长期空气质量预报基础数据，包括污染物浓度一次预报数据、气象一次预报数据、气象实测数据和污染物浓度实测数据，其中，所有一次预报数据的时间跨度为2020年7月23日到2021年7月13日，所有实测数据的时间跨度为2019年4月16 日到2021年7月13日，一次预报数据包括了近地2米温度(℃)、地表温度(K)、湿度(%)、近地10米风速(m/s)、大气压(Kpa)的相应气象参数共 15 个，及用于衡量空气质量的六种常规大气污染物的小时平均浓度，常规大气污染物分别为一氧化碳(CO)、二氧化硫(SO_2)、氮氧化物(NOx)、臭氧(O_3)等气体污染物和可吸入颗粒物(PM_{10})、细颗粒物($PM_{2.5}$) 等颗粒态污染物，这些大气污染物对公众生活具有潜在的负面影响，甚至会引发一系列健康问题[1]。

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。臭氧浓度预报也是六项污染物预报中较难的一项，其原因在于：作为六项污染物中唯一的二次污染物，臭氧并非来自污染源的直接排放，而是在大气中经过一系列化学及光化学反应生成的[3]。问题要求使用附件1中的数据，按照附录中的方法计算监测点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，并将结果按照附录“AQI计算结果表”的格式放在正文中。根据观察发现，需要用到的参数在2020年8月25日到8月28日中不存在空值缺失等问题，因此问题1可以直接通过公式求出相应的结果。

3.2 问题一求解

首先要计算臭氧(O_3)最大8小时滑动平均，因为当臭氧(O_3)最大8小时滑动平均浓度值高于800 $\mu g/m^3$ 时，或其余污染物浓度高于IAQI=500对应限值时，不再进行其空气质量分指数计算。

臭氧(O_3)最大8小时滑动平均是指一个自然日内8时至24时的所有8小时滑动平均浓度中的最大值，其中8小时滑动平均值指连续8小时平均浓度的算术平均值。其计算公式如下：

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

其中c_i为臭氧在某日 $t-1$ 时至 t 时的平均污染物浓度。

各项污染物的空气质量分指数(IAQI)，其计算公式如下：

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

各项污染物项目浓度限值及对应的空气质量分指数级别见表3-1。

表 3-1 空气质量分指数(IAQI)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	300	400	500	-

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	$\mu g/m^3$
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 8 小时滑动平均	0	100	160	215	265	800	-	-	
5	粒径小于等于 10μm 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 2.5μm 颗粒物($PM_{2.5}$)24 小时平均	0	35	75	115	150	250	350	500	

空气质量指数(AQI)取各分指数中的最大值，即

$$AQI = \max\{IAQI_1, IAQI_2, IAQI_3, ..., IAQI_n\} \tag{3-3}$$

式中，$IAQI_1, IAQI_2, IAQI_3, ..., IAQI_n$为各污染物项目的分指数。在问题中，对于 AQI 的计算仅涉及表 1 提供的六种污染物，因此计算公式如下：

$$AQI = \max\{IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{PM_{10}}, IAQI_{PM_{2.5}}, IAQI_{O_3}, IAQI_{CO}\} \tag{3-4}$$

空气质量等级范围根据 AQI 数值划分，等级对应的 AQI 范围见表 3-2。

表 3-2 空气质量等级及对应空气质量指数(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,+∞)

当 AQI 小于或等于 50（即空气质量评价为“优”)时，称当天无首要污染物；当 AQI 大于 50 时，IAQI 最大的污染物为首要污染物，若 IAQI 最大的污染物为两项或两项以上时，并列为首要污染物；IAQI 大于 100 的污染物为超标污染物。

通过 Matlab 编程，监测点 A 从 2020 年 8 月 25 日到 8 月 28 日根据逐日实测数据计算出的各项污染物的 IAQI 如表 3-3 所示，

表 3-3 监测点 A 从 2020 年 8 月 25 日到 8 月 28 日每天实测的 IAQI 和空气质量等级(逐日实测数据计算)

监测日期	地点	IAQI 计算						空气质量等级
		$IAQI_{CO}$	$IAQI_{SO_2}$	$IAQI_{NO_2}$	$IAQI_{O_3}$	$IAQI_{PM_{10}}$	$IAQI_{PM_{2.5}}$	
2020/8/25	监测点 A	13	8	15	60	27	16	良
2020/8/26	监测点 A	13	7	20	46	24	15	优
2020/8/27	监测点 A	15	7	39	109	37	33	轻度污染
2020/8/28	监测点 A	18	8	38	138	47	48	轻度污染

则监测点 A 从 2020 年 8 月 25 日到 8 月 28 日根据逐日实测数据计算出的 AQI 和首要污染物，结果如表 3-4 所示。

表 3-4 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(逐日实测数据计算)

监测日期	地点	AQI 计算	
		AQI	首要污染物
2020/8/25	监测点 A	60	O_3
2020/8/26	监测点 A	46	O_3
2020/8/27	监测点 A	109	O_3
2020/8/28	监测点 A	138	O_3

监测点 A 从 2020 年 8 月 25 日到 8 月 28 日每日根据逐小时实测数据的平均值计算出的各项污染物的 IAQI 如表 3-5 所示，

表 3-5 监测点 A 从 2020 年 8 月 25 日到 8 月 28 日每天实测的 IAQI 和空气质量等级(逐小时实测数据计算)

监测日期	地点	IAQI 计算						空气质量等级
		$IAQI_{CO}$	$IAQI_{SO_2}$	$IAQI_{NO_2}$	$IAQI_{O_3}$	$IAQI_{PM_{10}}$	$IAQI_{PM_{2.5}}$	
2020/8/25	监测点 A	13	8	15	61	27	16	良
2020/8/26	监测点 A	13	8	21	47	25	15	优
2020/8/27	监测点 A	17	7	39	109	38	33	轻度污染
2020/8/28	监测点 A	19	9	38	138	48	47	轻度污染

则监测点 A 从 2020 年 8 月 25 日到 8 月 28 日根据逐小时实测数据的平均值根据逐小时实测数据的平均值的 AQI 和首要污染物，结果如表 3-6 所示。

表 3-6 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(逐小时实测数据计算)

监测日期	地点	AQI 计算	
		AQI	首要污染物
2020/8/25	监测点 A	61	O_3
2020/8/26	监测点 A	47	O_3
2020/8/27	监测点 A	109	O_3
2020/8/28	监测点 A	138	O_3

根据结果可以看出：根据逐日实测数据计算出的 AQI 和根据逐小时实测数据的平均值根据根据逐日实测数据计算出的 AQI 值十分接近，且首要污染物相同。

4. 问题二分析与求解

4.1 问题二分析

空气质量预报模型的构建需要符合实际情况且误差控制在一定范围内的、相对准确的相关变量测量数据作为支撑。但是在实际的监测站点设备在进行数据测量、数据记录、数据导出等过程中难免会存在一定问题，使得最终得到的数据（原始数据）与希望得到的良好数据存在一定差距。例如，监测站点设备在进行数据测量过程中会遇到各种不同条件的较为复杂的实际工况，最终会导致采集的原始数据中存在着或多或少的不良数据，包括连续或间断性的数值缺失、数值漂移（偏大或偏小）等情况。因此，对不良数据进行科学有

效地预处理，对于空气质量预报模型的构建有着决定性意义。

对于实时数据库采集的不同位点的数据来说，采集数据的监测站点设备的数据均有部分站点存在问题，即部分站点只含有部分时间段的数据，部分站点的数据全部为空值或部分数据为空值，同时存在部分站点的数据超出了限值，因此对原始数据进行处理后才可以使用；对于实际的监测站点设备采集的近地 2 米温度（℃）、地表温度（K）、湿度（%）、近地 10 米风速（m/s）、大气压（Kpa）的相应气象数据来说，由于这些气象参数通常情况下是波动的，但在较短时间范围内可以认为不发生变化。

4.2 问题二求解

4.2.1 数据预处理

将一氧化碳（CO）、二氧化硫（SO_2）、二氧化氮（NO_2）、臭氧（O_3）、粒径小于 10μm 的颗粒物（PM_{10}）、粒径小于 2.5μm 的颗粒物（$PM_{2.5}$）的浓度的原始数据进行可视化，如图 4-1 所示。

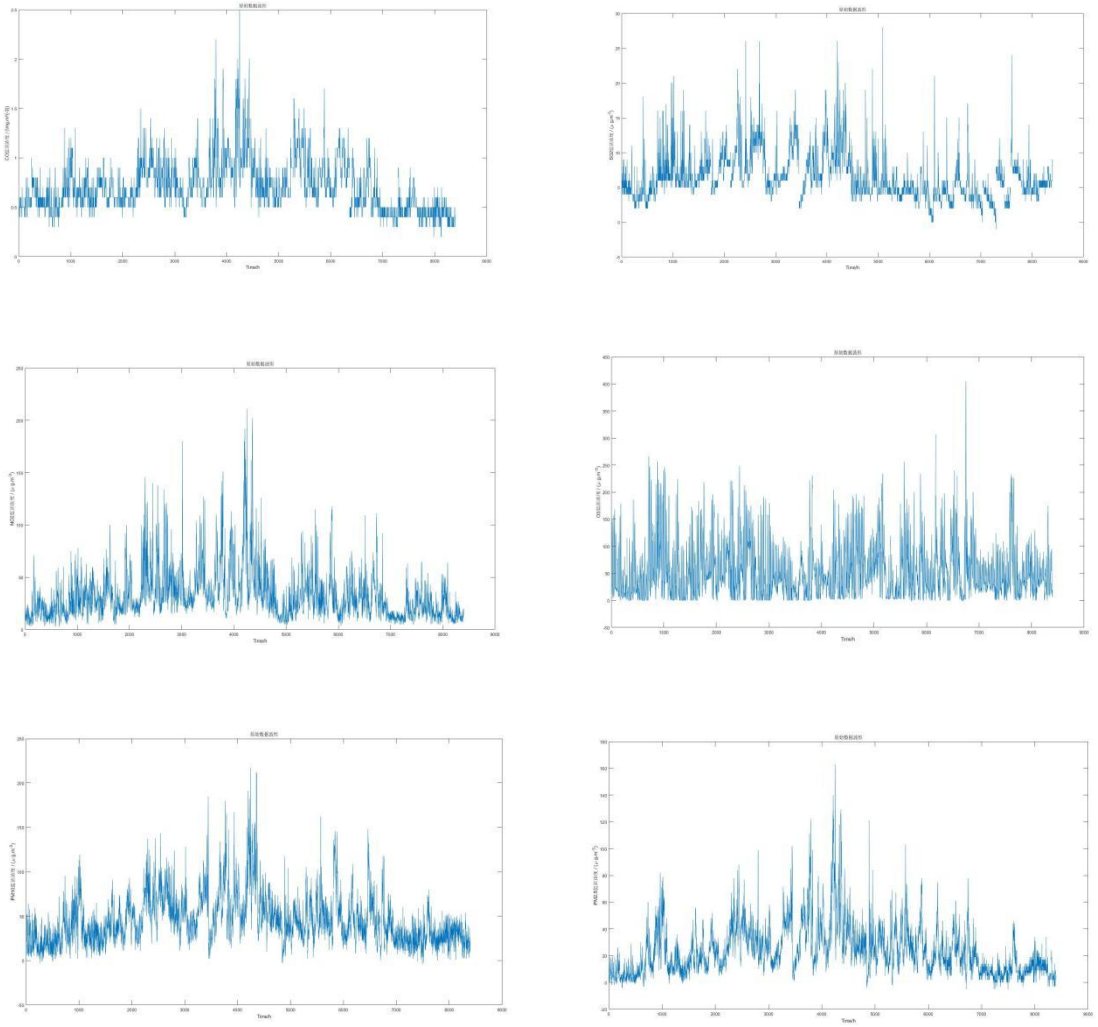


图 4-1 污染物浓度原始数据可视图

数据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；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；题目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。

综合以上分析，现针对附件 1 中的原始数据中各种类型的不良数据分别进行相应分析及处理。对于附件 1 中的原始数据，应将其中的异常值剔除掉，剔除的标准则可以采用拉依达准则(3 σ 准则)。综上，针对该原始数据可能存在的异常情况作如下方面的预处理：

(1) 变量值缺失为空值

监测站点设备数据采集过程中导致部分点位出现连续的或者间断性的变量值缺失为空值，对于这种情况主要采取利用缺失值前后一小时内的数值取平均的方式来处理。

(2) 变量值超出限值范围(超出最小最大值)

在实际操作过程中，可能会使得操作变量的实际控制超出附录所要求的最小最大值范围，对这种情况则考虑将超出范围区间的值剔除出去，即在对不同站点的数据取平均得出最终的确定值过程中，对这些超出最小最大值范围的数据取最小值或最大值。

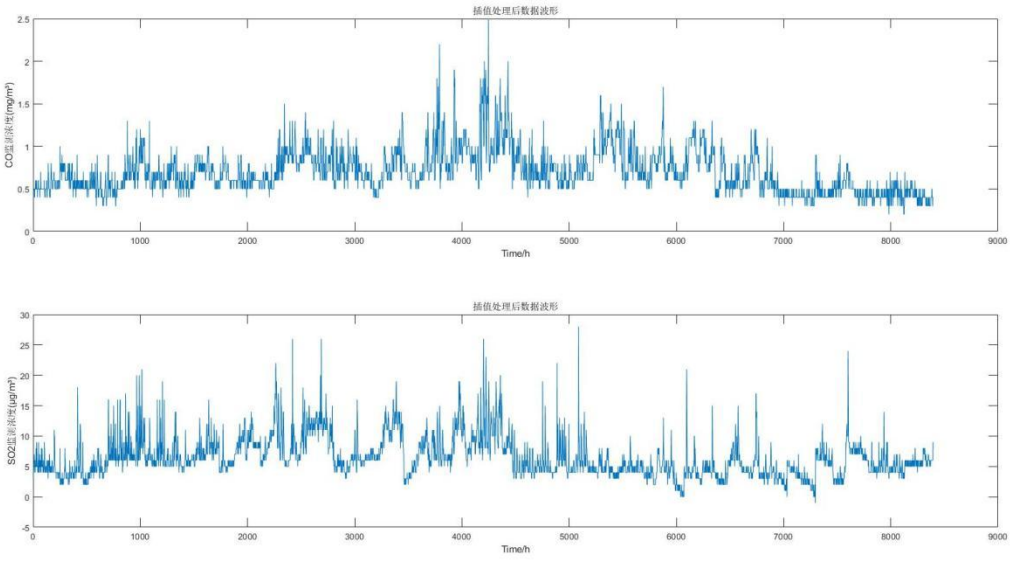
(3) 变量值超出 3 σ 区间范围

有理由认为监测站点设备变量的过程为连续的，不会出现跳跃的情况。但是可能由于操作装置采集数据过程可能存在问题，在实际的原始数据中，可以发现该类型的“跳跃”。因此对于站点采集到的数据，考虑将这样的跳跃值剔除出去，即在对站点的变量数据取平均得出最终的变量确定值过程中，对这些在 3 σ 范围外的数据不予考虑。

（4）变量值存在单位不一致的情况

变量值在附件中存在单位不一致的情况，例如在一次预测数据中气压单位为 Kpa，而在实测数据中气压单位为 MBar，这两者之间的单位换算等价式为 1Kpa=10MBar。

处理后的数据及每日 AQI 值在附件 1、2 中给出，并且给出数据可视化的结果，如图 4-2 所示。



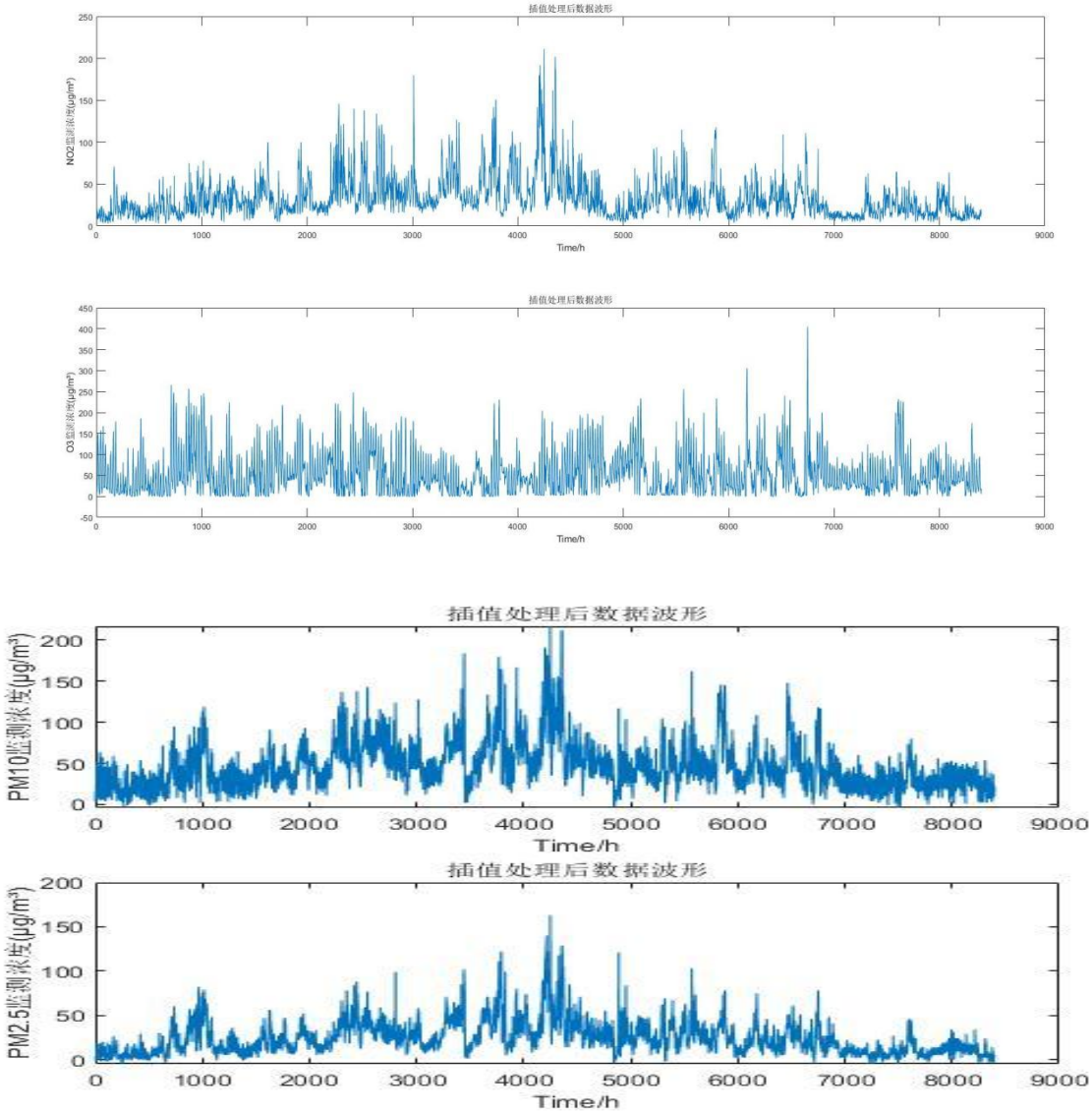


图 4-2 污染物浓度处理后数据可视图

4.2.2 模型原理及框架

k-means 聚类算法模型

对于给定的一个包含 n 个数据点的数据集$X = \{X_1, X_2, \ldots, X_n\}$，其中$\mathrm{X_i} \in \mathrm{R_d}$，以及要生成数据子集的数目 K，K-Means 聚类算法将数据对象组织为 K 个划分 $\mathrm{C} = \{\mathrm{c_k}, \mathrm{i} = 1,2,\ldots,\mathrm{K}\}$。每个划分代表一个类$\mathrm{c_k}$，每个类$c_k$有一个类别中心$\mu_i$。选取欧氏距离作为相似性和距离判别准则，计算该类内各点到聚类中心μ_i的距离平方和

$$J(c_k) = \sum_{x_i \in C_K} \|X_i - \mu_K\|^2 \tag{4-1}$$

聚类目标是使各类总的距离平方和 $\mathrm{J(C)} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{J(c_k)}$最小。

$$J(C) = \sum_{k=1}^{K} J(c_k) = \sum_{k=1}^{K} \sum_{x_i \in C_K} \|X_i - \mu_K\|^2 = \sum_{k=1}^{K} \sum_{i=1}^{n} d_{ki} \|X_i - \mu_K\|^2 \tag{4-2}$$

其中，$d_{ki}=\begin{cases}1，X_i \in c_i\\0，X_i \notin c_i\end{cases}$

显然根据最小二乘法和拉格朗日原理，聚类中心 μ_k 应该取为类别 c_k 类各数据点的平均值。

K-Means 聚类算法从一个初始的 K 类别划分开始，然后将各数据点指派到各个类别中，以减少总的距离平方和。因为 K-Means 聚类算法中总的距离平方和随着类别个数 K 的增加而趋向于减少。因此，总的距离平方和只能在某个确定的类别个数 K 下，取得最小值。

定义：

（1）两个数据对象间的距离：

我们采用欧式距离（Euclidean Distance）进行计算，计算公式为

$$d(x_i,x_j)=(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^2)^{1/2} \tag{4-3}$$

（2）准则函数 E

对于 K-means 算法，通常使用准则函数 E，也就是误差平方和（Sum of Squared Error，SSE）作为度量聚类质量的目标函数。

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}d^2(C_i,x) \tag{4-4}$$

其中，d( )表示两个对象之间的距离。

对于相同的 k 值，更小的 SSE 说明簇中对象越集中。对于不同的 k 值，越大的 k 值应该越小的 SSE。

k-means 聚类算法实现步骤

首先，随机选择 k 个对象，每个对象代表一个簇的初始均值或中心；对剩余的每个对象，根据其与各簇中心的距离，将它指派到最近（或最相似）的簇，然后计算每个簇的新均值，得到更新后的簇中心；不断重复，直到准则函数收敛。通常，采用平方误差准则，即对于每个簇中的每个对象，求对象到其中心距离的平方和，这个准则试图生成的 k 个结果簇尽可能地紧凑和独立。

步骤：

输入：聚类个数 k，以及包含 n 个数据对象的数据库 X；

输出：满足方差最小标准的 k 个聚类。

处理流程：

步骤 1　从 n 个数据对象任意选择 k 个对象作为初始聚类中心；

步骤 2　根据簇中对象的平均值，将每个对象重新赋给最类似的簇；

步骤 3　更新簇的平均值，即计算每个簇中对象的平均值；

步骤 4　循环 Step2 到 Step3 直到每个聚类不再发生变化为止[4]。

4.2.3 聚类结果

用 Matlab 通过 k-means 算法聚类得到的结果如下：

首先，由于在污染物排放情况不变的条件下，某一地区的气象条件有利于污染物扩散或沉降时，该地区的 AQI 会下降，反之会上升。气象因素中高温、低压、低湿、高风速、都有利于污染物浓度的清除和扩散，任何一个因素的变化均会引起环境空气质量的变化，而不同污染物受气象条件影响程度不同，故根据每个气象条件对污染物浓度的影响程度，

对气象条件和计算所得对应 AQI 值分别进行聚类分析，对气象条件进行划分，聚类中心结果保留 3 位小数，

1.温度分成两类：高温（聚类中心为 27.691）、低温（聚类中心为 17.921）

2.湿度分成两类：高湿（聚类中心为 72.988）、低湿（聚类中心为 45.970）

3.气压分成两类：高压（聚类中心为 1016.784）、低压（聚类中心为 1006.346）

4.风速分成三类：微风（聚类中心为 2.226）、轻风（聚类中心为 1.530）、软风（聚类中心为 0.994）

通过对聚类结果进行分析，可以将气象进行如下分类：

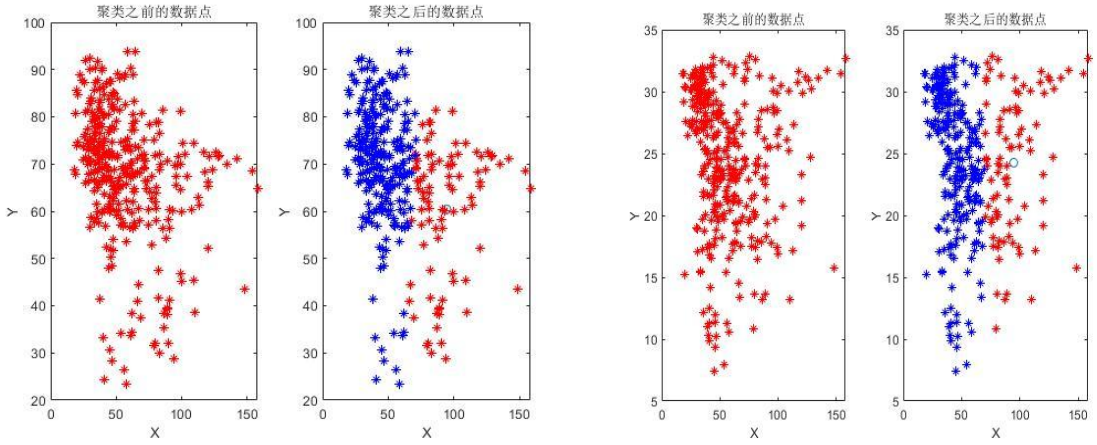
1：按温度分为两类：第一类高温，平均温度 27.691℃，有利于污染物的扩散，下层气温高，使空气流动剧烈，底层乱流、湍流比较旺盛，有利于污染物向高空传输，从而使得近地面污染物浓度降低，使得空气质量指数降低，AQI 取值范围在（0，60）之间；第二类低温，平均温度 17.921℃，不利于污染物扩散，即污染物浓度随环境温度的降低而增高，AQI 取值范围（60，150）；

2：按湿度分为两类：第一类，高湿，平均湿度 72.988%，湿度高，水汽对污染物有吸附作用，特别是降水的时候，空气中的水汽含量高，会使 PM_{10}、$PM_{2.5}$ 质量增加而使浮悬颗粒物沉降到地面，降低 PM10、PM2.5 的浓度，故高湿环境能够促使污染物的扩散，降低污染物浓度，AQI 取值范围（0，65）；第二类低湿，平均湿度 45.970%，不利于污染物扩散，污染物浓度较高，AQI 取值范围（65，150）；

3.按气压分为两类：第一类，高压，平均气压 1016.784MBar，大气压与污染物浓度正相关，高压时，大气层结构稳定，气流下沉，不利于污染物的垂直扩散，污染物浓度累积增加，AQI 取值范围（55，150）；第二类，低压，平均气压 1006.346MBar，由于气流上升，有利于污染物扩散，AQI 取值范围（0，55）；

4.按风速分成三类：第一类，微风，平均风速 2.226m/s，风有利于污染物的水平扩散，风速越大，污染物水平扩散能力越大，降低污染物浓度，AQI 取值范围（0，50）；第二类，轻风，平均风速 1.530m/s，污染物浓度较高，AQI 取值范围（50，100）；第三类，软风，平均风速 0.994m/s，风速过低，混合作用强于扩散作用不利于污染物扩散，污染物浓度高，相应 AQI 值也较高，取值范围（100，150）。

温度、湿度、气压、风速的聚类结果图分别如图 4-3 所示。



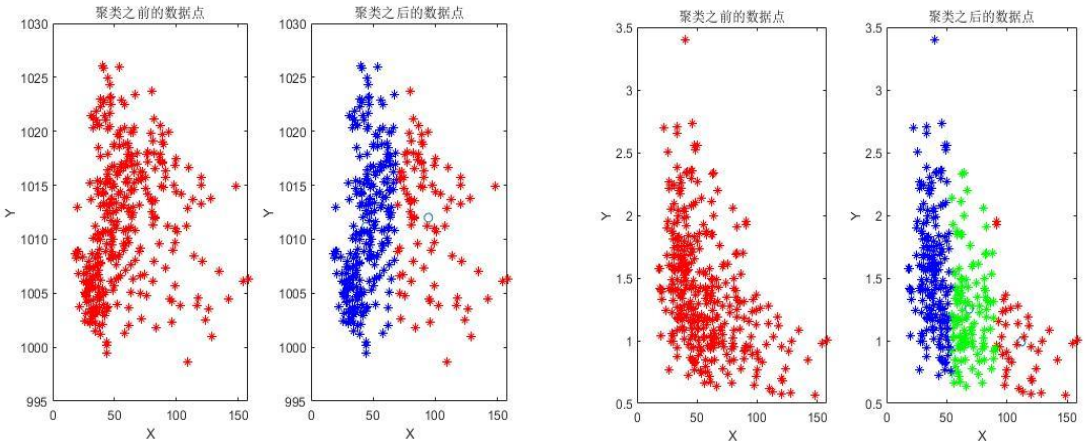


图 4-3 温度、湿度、气压、风速的聚类结果图

在依据 AQI 值对不同气象进行分别聚类之后。对整体也进行了聚类，因 k-means 聚类的先验性条件，需要预先给定聚类簇数，但我们不能预知分为多少簇合适，故设置了一个最大聚类簇数 13，进行多次聚类，依据聚类有效性指标 DB 和 SSE 对聚类结果进行评价，最终选择簇数为 3，将整体分为三类，聚类中心如表 4-1 所示（保留三位小数）：

表 4-1 聚类中心点

聚类中心	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(°)
第一类	0.644	7.240	24.545	122.812	54.732	27.778	26.660	56.879	1011.303	1.575	79.178
第二类	0.707	6.283	31.822	48.908	45.798	24.578	25.100	67.665	1010.3	1.343	288.373
第三类	0.712	6.257	37.400	30.752	41.785	21.729	23.081	72.649	1012.385	1.360	50.188

通过聚类结果分析，可以将气象分为以下三类：

第一类，平均温度 26.60℃，平均湿度 56.876%，平均气压 1011.303MBar，平均风速 1.575m/s，平均风向 79.178°，在此气象条件下 SO_2、NO_2、PM10 的平均污染浓度低，但是 $PM_{2.5}$、O_3 和 CO 的污染浓度在三类中最高；

第二类，平均温度 25.100℃，平均湿度 67.665%，气压 1010.3MBar，平均风速 1.343m/s，平均风向 288.373°，在此气象条件下除 $PM_{2.5}$ 的平均浓度是三类中最高的，其它污染物浓度都居中；

第三类，平均温度 23.081℃，平均湿度 72.649，气压 1012.385MBar，平均风速 1.360m/s，平均风向 50.188°，在此气象条件下除 PM_{10} 的浓度在三类中最高外，其它污染物浓度均处于最低水平。

5. 问题三分析与求解

5.1 问题三分析

一个有效的空气质量预报系统有助于人类掌握污染物未来浓度信息，制定相应的防治

策略，对空气中污染物的浓度水平提前给出精确的预报，使因污染物浓度超标所造成的非线性智能统计模型，目前常用 WRF-CMAQ 模拟体系对空气质量进行预报。WRF-CMAQ 模型主要包括 WRF 和 CMAQ 两部分：WRF 是一种中尺度数值天气预报系统，用于为 CMAQ 提供所需的气象场数据；CMAQ 是一种三维欧拉大气化学与传输模拟系统，其根据来自 WRF 的气象信息及场域内的污染排放清单，基于物理和化学反应原理模拟污染物等的变化过程，继而得到具体时间点或时间段的预报结果[6]。

但受制于模拟的气象场以及排放清单的不确定性，以及对包括臭氧在内的污染物生成机理的不完全明晰，WRF-CMAQ 预报模型的结果并不理想[7]。故提出二次建模概念：即指在 WRF-CMAQ 等一次预报模型模拟结果的基础上，结合更多的数据源进行再建模，以提高预报的准确性。其中，由于实际气象条件对空气质量影响很大（例如湿度降低有利于臭氧的生成），且污染物浓度实测数据的变化情况对空气质量预报具有一定参考价值，故目前会参考空气质量监测点获得的气象与污染物数据进行二次建模，以优化预报模型。

如何预测空气中污染物浓度是一个复杂的问题。目前我国的城市环境空气质量预报的主要模型包括多元线性回归、人工神经网络、NAQPMS、CAMx、WRF-Chem 及多模式集合预报体系等[8]，国内外的研究表明神经网络能够比回归模型更好地预报空气污染物的浓度随时间的变化趋势，得出更加理想的预报效果。神经网络空气质量预报模型适用于当前难以开展空气质量数值预报，并且回归统计模型的预报精度无法满足要求的城市开展空气质量预报工作及相关研究。因此本文在 WRF-CMAQ 的一次预报模型模拟结果的基础上二次建模建立 BP 神经网络空气质量预测模型。

5.2 问题三求解

5.2.1 模型原理及框架

BP 神经网络

BP 算法由数据流的前向计算（正向传播）和误差信号的反向传播两个过程构成。正向传播时，传播方向为输入层→隐层→输出层，每层神经元的状态只影响下一层神经元。若在输出层得不到期望的输出，则转向误差信号的反向传播流程。通过这两个过程的交替进行，在权向量空间执行误差函数梯度下降策略，动态迭代搜索一组权向量，使网络误差函数达到最小值，从而完成信息提取和记忆过程。

神经网络模型原理描述：

（1）正向传播

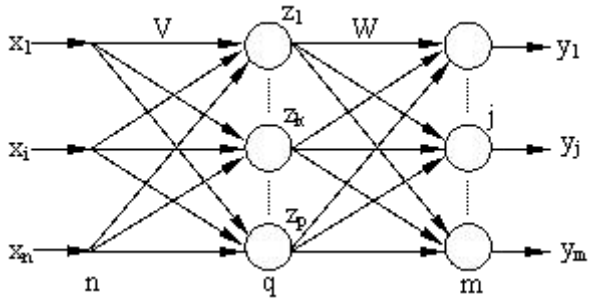


图 5-1　三层神经网络的拓扑结构

设 BP 网络的输入层有 n 个节点，隐层有 q 个节点，输出层有 m 个节点，输入层与隐层之间的权值为v_{ki}，隐层与输出层之间的权值为w_{jk}，如图 5-1 所示。隐层的传递函数为 f1(•)，

输出层的传递函数为 f2(·)，则隐层节点的输出为（将阈值写入求和项中）：

$$z_k = f_1（\sum_{i=0}^{n} v_{ki}\, x_i）\qquad k=1,2,...,q \tag{5-1}$$

输出层节点的输出为：

$$y_j = f_2（\sum_{k=0}^{q} w_{jk}\, z_k）\qquad j=1,2,...,m \tag{5-2}$$

至此 BP 网络就完成了 n 维空间向量对 m 维空间的近似映射[10]。

反向传播

(1) 定义误差函数

输入 P 个学习样本，用$x^1, x^2, \ldots, x^p, \ldots, x^P$来表示。第 p 个样本输入到网络后得到输出$y_j^p$（j=1,2,…m）。采用平方型误差函数，于是得到第 p 个样本的误差E_p：

$$E_p = \frac{1}{2}\sum_{j=1}^{m} (t_j^p - y_j^p)^2 \tag{5-3}$$

式中：t_j^p为期望输出。

对于 P 个样本，全局误差为：

$$E=\frac{1}{2}\sum_{p=1}^{P}\sum_{j=1}^{m} (t_j^p - y_j^p) = \sum_{p=1}^{P} E_p \tag{5-4}$$

采用累计误差 BP 算法调整w_{jk}，使全局误差 E 变小，即

$$\Delta w_{jk}=-\eta \frac{\partial E}{\partial w_{jk}}=-\eta \frac{\partial}{\partial w_{jk}}(\sum_{p=1}^{P} E_p)=\sum_{p=1}^{P} -\eta \frac{\partial E_p}{\partial w_{jk}} \tag{5-5}$$

式中：η—学习率

(2)定义误差信号为：

$$\delta_{yj}=-\frac{\partial E_p}{\partial S_j}=-\frac{\partial E_p}{\partial y_j}\cdot\frac{\partial y_j}{\partial S_j} \tag{5-6}$$

其中第一项：

$$\frac{\partial E_p}{\partial y_j}=\frac{\partial}{\partial y_j}\left[\frac{1}{2}\sum_{j=1}^{m} (t_j^p - y_j^p)^2\right]=-\sum_{j=1}^{m} (t_j^p - y_j^p) \tag{5-7}$$

第二项：

$$\frac{\partial y_j}{\partial S_j}=f_2'(S_j) \tag{5-8}$$

是输出层传递函数的偏微分。于是：

$$\delta_{yj}=\sum_{j=1}^{m} (t_j^p - y_j^p)\, f_2'(S_j) \tag{5-9}$$

由链定理得：

$$\frac{\partial E_p}{\partial w_{jk}}=\frac{\partial E_p}{\partial S_j}\cdot\frac{\partial S_j}{\partial w_{jk}}=-\delta_{yj}z_k=-\sum_{j=1}^{m}(t_j^p-y_j^p)f_2'(S_j)\cdot z_k \tag{5-10}$$

输出层各神经元的权值调整公式为：

$$\Delta w_{jk}=\sum_{p=1}^{P}\sum_{j=1}^{m}\eta(t_j^p-y_j^p)f_2'(S_j)Z_k \tag{5-11}$$

(3)隐层权值的变化

$$\Delta_{v_{ki}}=-\eta\frac{\partial E}{\partial v_{ki}}=-\eta\frac{\partial}{\partial v_{ki}}(\sum_{p=1}^{P}E_p)=\sum_{p=1}^{P}(-\eta\frac{\partial E_p}{\partial v_{ki}}) \tag{5-12}$$

定义误差信号为：

$$\delta_{zk}=\frac{\partial E_p}{\partial S_k}=-\frac{\partial E_p}{\partial z_k}\cdot\frac{\partial z_k}{\partial S_k} \tag{5-13}$$

其中第一项：

$$\frac{\partial E_p}{\partial z_k}=\frac{\partial}{\partial z_k}[\frac{1}{2}\sum_{j=1}^{m}(t_j^p-y_j^p)^2]=-\sum_{j=1}^{m}(t_j^p-y_j^p)\frac{\partial y_j}{\partial z_k} \tag{5-14}$$

依链定理有：

$$\frac{\partial y_j}{\partial z_k}=\frac{\partial y_j}{\partial S_j}\cdot\frac{\partial S_j}{\partial z_k}=f_2'(S_j)w_{jk} \tag{5-15}$$

第二项是隐层传递函数的偏微分：

$$\frac{\partial z_k}{\partial S_k}=f_1'(S_k) \tag{5-16}$$

于是：

$$\delta_{zk}=\sum_{j=1}^{m}t_j^p-y_j^p)f_2'(S_j)w_{jk}f_1'(S_k) \tag{5-17}$$

由链定理得：

$$\frac{\partial E_F}{\partial v_{ki}}=\frac{\partial E_F}{\partial S_k}\cdot\frac{\partial S_k}{\partial v_{ki}}=-\delta_{zk}x_i=-\sum_{j=1}^{m}(t_j^p-y_j^p)f_2'(S_j)w_{jk}f_1'(S_k)\cdot x_1 \tag{5-18}$$

从而得到隐层各神经元的权值调整公式为：

$$\Delta v_{k_i}=\sum_{p=1}^{P}\sum_{i=1}^{m}\eta(t_j^p-y_j^p)f_2'(S_j)w_{jk}f_1'(S_k)x_i \tag{5-19}$$

5.2.2 模型结果

BP 神经网络预测模型是基于统计预报方法的空气质量预测模型，是以分析历史空气质量数据和气象条件为基础，找出内在的发展规律从而完成预测[9]。相比于数值预测模型方法，该方法主要基于历史气象数据以及污染物监测浓度的规律性分析，利用气象条件预报

产品开展污染物浓度预测，分析方法较为灵活多样，具有较好的适用性。 BP 神经网络预测模型结构如图 5-2 所示。

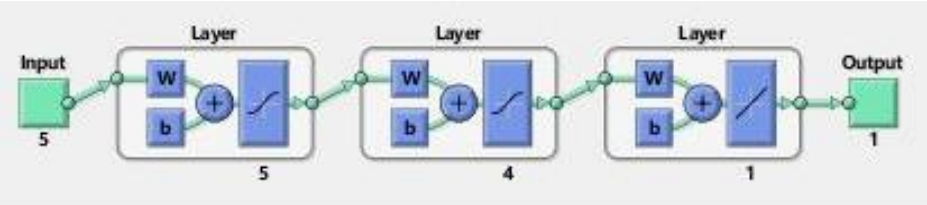


图 5-2　BP 神经网络预测模型结构

根据问题 3 的要求，需要建立一个满足两个目标的预测模型，极小化 AQI 预报值的相对误差，极大化首要污染物精确度，运用 BP 神经网络进行空气中 6 种污染物浓度进行预测,通过输入 6 种污染物浓度及温度、湿度、气压、风速、风向等实测数据，他们对应的评价量化值为 AQI，设置隐层元为 4，输出层为 1.在训练过程中不断修正神经网络，把步数设置为 10000 进行训练。BP 神经网络的学习训练算法流程图如图 5-3 所示。基于问题三的 BP 神经网络模型如图 5-4 所示。

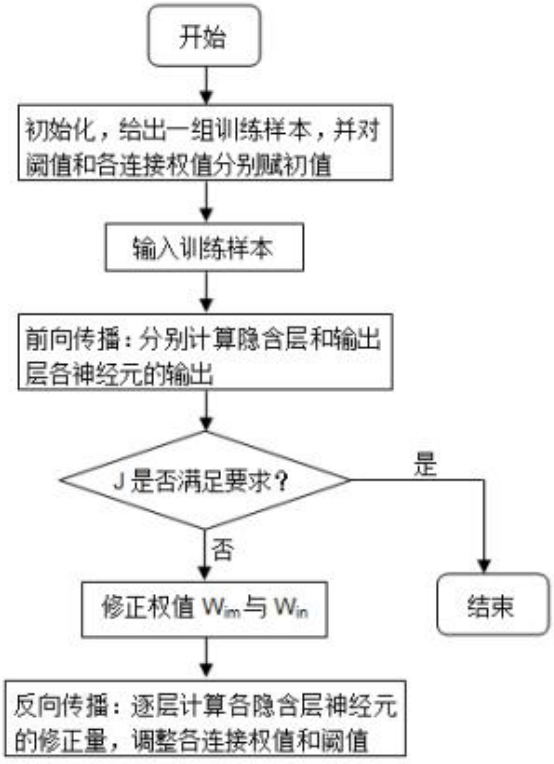


图 5-3　BP 神经网络学习训练算法流程图

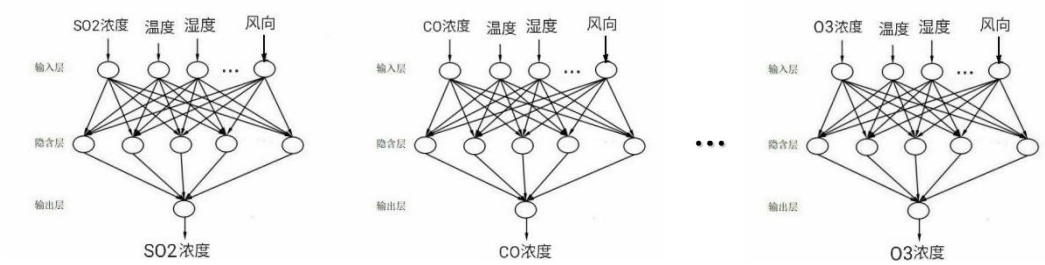


图 5-4　基于问题三的 BP 神经网络模型

借助于 Matlab 编程来实现训练神经网络，对所建立的模型进行多次测试，下面给出了选取七个月的数据对 2021 年 1 月 1 日到 2021 年月 3 日的空气中六种污染物的浓度进行预测的结果，并计算了相应的 AQI 值，与一次预报数据和实测数据进行对比分析，发现建立的二次模型的预测 AQI 值的相对误差小于且优于一次预测结果，首要污染物预测结果与实测数据吻合。训练结果如表 5-1 所示。详细数据见附件 3~5。

表 5-1 污染物的浓度预测测试结果

数据类型	监测时间	地点	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	首要污染物
实测数据	2021/1/1	监测点 A	0.758333333	5.958333333	33.45833333	63.5	41.47222222	19.41666667	42	NO_2
	2021/1/2	监测点 A	0.9	8.833333333	56.54166667	63.0313	54.25	27.33333333	71	NO_2
	2021/1/3	监测点 A	1.270833333	11.25	80.91666667	59.75	99.79166667	53.625	101	NO_2
一次预测数据	2021/1/1	监测点 A	0.250053083	11.40092708	89.6932875	44.9809	57.90251667	48.6755375	105	NO_2
	2021/1/2	监测点 A	0.464492833	16.50580083	112.7520708	29.0943	61.90714167	48.0132375	117	NO_2
	2021/1/3	监测点 A	0.379046917	13.84409125	107.232	40.2365	71.17022917	57.745025	114	NO_2
二次预测数据	2021/1/1	监测点 A	0.985947658	7.032658524	56.241007064	53.3643	60.74665683	28.344585	71	NO_2
	2021/1/2	监测点 A	0.773677718	7.483870105	48.510409485	52.1166	53.95765104	28.59299089	61	NO_2
	2021/1/3	监测点 A	0.821506546	10.86527288	60.32801011	54.4457	69.71932926	44.67635322	96	NO_2

采用相对误差来分析模型的预测效果，相对误差的计算公式为：

$$\text{相对误差}=\frac{|\text{监测 AQI}-\text{预测 AQI}|}{\text{监测 AQI}} \tag{5-18}$$

表 5-2 AQI 相对误差

AQI 相对误差	一次预测	二次预测
2021/1/1	1.5	0.69047619
2021/1/2	0.647887324	0.14084507
2021/1/3	0.128712871	0.04950495

从表中可以看出通过二次建模预测的空气质量指数(AQI)比一次预测所得到的数据更加接近实测的数据，因此相对误差更小，说明建立的二次模型—BP 神经网络的预测模型能够较好地实现污染物浓度的预测，BP 神经网络预测模型比 WRF-CMAQ 预测模型在污染物的预测精度上都有了明显的提高，证明 BP 神经网络预测模型比一般的 WRF-CMAQ 预测模型具有更好的性能。

输入 5 种类型影响污染物浓度的气象指标，通过训练好的神经网络进行预测监测点 A、B、C 在 2021 年 7 月 13 日至 7 月 15 日 6 种常规污染物的单日浓度值，计算相应的 AQI 和首要污染物，结果“污染物浓度及 AQI 预测结果表”如表 5-3 所示。

表 5-3　污染物浓度及 AQI 预测结果表

预报日期	地点	二次模型日值预测							
		SO2 (μ g/m³)	NO2 (μ g/m³)	PM10 (μ g/m³)	PM2.5 (μ g/m³)	O3 最大八小时滑动平均 (μ g/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. 问题四分析与求解

6.1 问题四分析

相邻区域的污染物浓度往往具有一定的相关性，区域协同预报可能会提升空气质量预报的准确度。需要进一步考虑方位和距离对空气质量的影响。需建立监测点 A 及其临近区域内存在监测点 A1、A2、A3，四个监测点的协同预报模型，要求二次模型预测结果中 AQI 预报值的最大相对误差应尽量小，且首要污染物预测准确度尽量高，模型评价指标和问题 3 相同，故可以在问题 3 的基础上进行考虑。需要对建立的 BP 神经网络二次预测模型进行改进，用遗传算法改进 BP 神经网络，并根据点 A 和 A1、A2、A3 的方位和坐标信息，在进行训练集的选取进行改进，选取同一时间段，4 个不同监测点监测的污染物浓度和气象数据，作为训练集[10]。在进行测试前需按照前面对数据的处理方式对表格数据进行预处理，将缺失值、负值和异常值进行处理，并对气压数据进行单位换算。

6.2 问题四求解

6.2.1 模型原理及框架

遗传算法原理

遗传算法（GA）是进化计算的一部分，是模拟达尔文的遗传选择和自然淘汰的生物进

化过程的计算模型，是一种通过模拟自然进化过程搜索最优解的方法。算法简单、通用、鲁棒性强，适于并行处理。遗传算法包括选择、交叉和变异三个算子。

(1)选择算子

通过选择算子模拟“优胜劣汰”，适应度高的个体被遗传到下一代的概率大，适应度低的个体，被遗传到下一代的概率低。常用的选择算子：轮盘赌选择法，假设f_i表示种群中第 i 个个体的适应度函数，n 是种群中个体数量，$\sum f_i$;就是群里适应度的和，产生后代的能力如式(3)所示，其中 k=1,2..。

$$P_i=\frac{f_i^k}{\sum_{i=1}^{n} f_i^k} \tag{6-1}$$

(2) 交叉算子

交叉算子是指对两个相互配对的染色体按某种方式相互交换其部分基因，从而形成新的个体，是遗传算法区别于其它算法的重要特征，是产生新个体的主要方法。常用的交叉方式有单点交叉、双点交叉、多点交叉、均匀交叉、算术交叉等。

(3) 变异算子

遗传算法中的变异运算是指将个体染色体编码串中的某些基因座上的基因值用该基因座的其他等位基因来替换，从而形成一个新的个体。遗传算法中的变异操作是为了保持种群的多样性，防止基因丢失。

遗传算法基本步骤

(1)染色体编码。

(2)初始化种群。

(3)设计适应度函数，确定个体的环境适应能力，适应度值越高，适应能力越强，存活的几率就越大。

(4)基于适应度函数，从父代种群中选择优胜个体。

(5)根据交叉率执行交叉操作。

(6)根据变异率执行变异操作。

(7)判断是否达到终止条件，是，转向(8),否，转向(3)。

(8)结束迭代，选出适应度函数值最大的个体作为问题的最优解[8]。

遗传算法的基本步骤如图 6-1 所示。

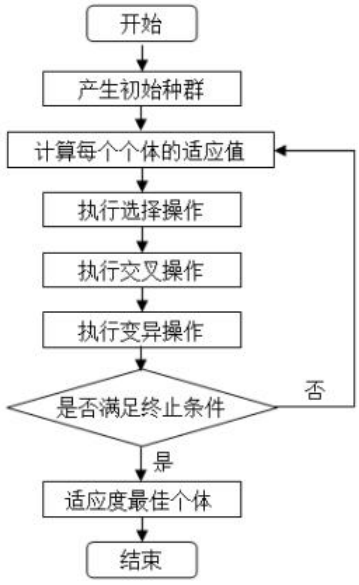


图 6-1　遗传算法基本步骤

改进遗传 BP 神经网络空气质量预报模型

(1)确定 BP 神经网络的结构

以 A、A1、A2、A3 四监测点检测的污染物浓度和气象数据为基础，构建 BP 神经网络的结构，设计输人层、输出层变量，确定隐含层的节点数和层数。

(2)优化 BP 神经网络

设计适应度函数，计算每个个体的适应度值，通过遗传算法中的选择(二进制锦标赛选择)、交叉(模拟二进制交叉)、变异(多项式变异)等相关操作，确定优胜个体，优化 BP 神经网络中每一层的权重和阙值。

基于遗传算法的 BP 神经空气质量预测模型流程图见图 6-2.

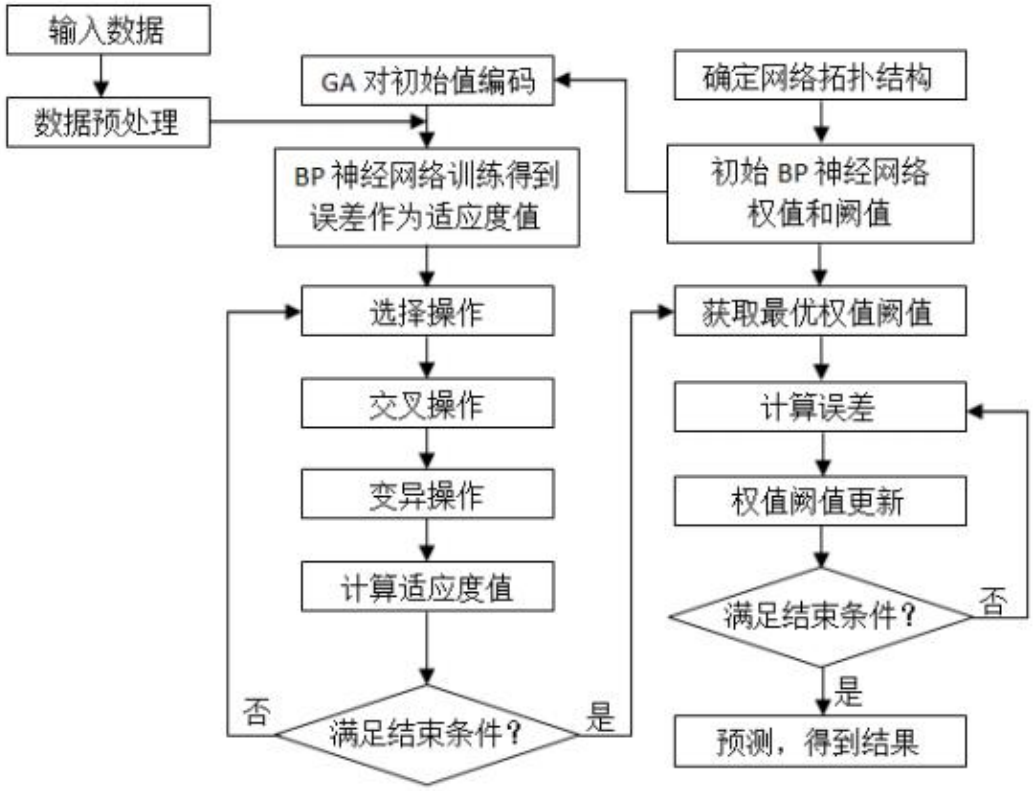


图 6-2　基于遗传算法的 BP 神经空气质量预测模型流程图

6.2.2 模型结果

使用该模型预测监测点 A、A1、A2、A3 在 2021 年 7 月 13 日到 7 月 15 日 6 种常规污染物的单日浓度值，计算相应的 AQI 和首要污染物预测结果(保留六位小数)如表 6-1 所示。处理后的数据及预测结果详细内容见附件 6~9。

表 6-1 二次模型 AQI 和首要污染物预测结果

预报日期	地点	二次模型日值预测							
		SO2 (μg/m³)	NO2 (μg/m³)	PM10 (μg/m³)	PM2.5 (μg/m³)	O3 最大八小时滑动平均 (μg/m³)	CO (mg/m³)	AQI	首要污染物

2021/7/13	监测点 A	6.026498	12.754470	23.135363	6.526287	43.911091	0.396785	24	PM10
2021/7/14	监测点 A	5.981969	12.971135	21.672379	6.321114	44.094971	0.395751	23	03
2021/7/15	监测点 A	5.880314	12.881624	21.495475	6.382446	25.802753	0.396683	22	PM10
2021/7/13	监测点 A1	7.027309	15.332967	29.684920	9.177502	46.661877	0.451259	30	PM10
2021/7/14	监测点 A1	7.281718	14.683218	29.130652	9.039957	46.462130	0.639348	30	PM10
2021/7/15	监测点 A1	7.181598	14.934196	30.305947	9.019191	25.802753	0.961902	31	PM10
2021/7/13	监测点 A2	4.850706	16.518653	27.087839	8.169413	58.828821	0.443849	71	NO2
2021/7/14	监测点 A2	4.894688	16.241063	26.025836	8.814907	58.144192	0.445929	61	NO2
2021/7/15	监测点 A2	4.913525	16.284831	26.080662	8.513877	25.802753	0.466164	76	NO2
2021/7/13	监测点 A3	3.849022	10.188080	17.846609	6.898581	58.828821	0.307802	30	03
2021/7/14	监测点 A3	3.852278	10.017528	17.209906	25.802753	58.144192	0.307467	37	PM2.5
2021/7/15	监测点 A3	7.344998	9.8094291	16.454645	11.317941	25.802753	0.308424	17	NO2

与问题 3 的模型相比，新建立的 GA-BP 神经网络区域协同预报模型提升了监测点 A 的污染物浓度预报准确度，降低了空气质量指标 AQI 预测值的相对误差，预测首要污染物的精准性提高。经过测试发现优化的 GA-BP 神经网络区域协同预报模型对污染物浓度的的预测误差明显小于 BP 网络预测模型，AQI 预测值的相对误差更小，预测的首要污染物更加准确说明 GA-BP 预测模型预测精度更高。通过对两种模型的输出结果与期望输出进行分析，发现 GA-BP 模型的相关系数更高，说明优化的预测模型有更强的泛化能力。

7. 结论与模型评价

7.1 结论

(1)对于问题 1，问题要求使用附件 1 中的数据，按照附录中的方法计算监测点 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，并将结果按照附录“AQI 计算结果表”的格式放在正文中。根据观察发现，需要用到的参数在 2020 年 8 月 25 日到 8 月 28 日中不存在空值缺失等问题，因此问题 1 可以直接通过附录中提供的公式求出空气质量指数(AQI)的值。

(2)利用多种数学原理和计算软件，完成了数据处理

预处理提供了所有时间跨度为 2020 年 7 月 23 日到 2021 年 7 月 13 日的一次预报数据，时间跨度为从 2019 年 4 月 16 日到 2021 年 7 月 13 日的实测数据的采集记录，同时还提供了所有气象条件的相关信息。对于提供的采集数据中存在的各种不良数据或错误数据，先后采用了相应的处理手段：对于由于设备、外界条件导致的变量值缺失为空值的异常数据作了取前后一段时间内平均的补全处理；对于变量值超出可调控最小或最大范围的数据，根据该变量的可调控范围，直接将其取为相应的最小值和最大值；对于变量值超出 3σ 区间范围的数据，考虑利用 3σ 准则将每个站点采集到的跳跃值数据剔除出去。

利用 k-means 聚类，根据每个气象条件对污染物浓度的影响程度，对气象条件和计算所得对应 AQI 值分别进行聚类分析，对气象条件进行划分，1.温度分成两类：高温、低温；2.湿度分成两类：高湿、低湿；3.气压分成两类：高压、低压；4.风速分成三类：微风、轻风、软风。并讨论了每种气象条件的特征，相对应的污染物浓度的变化，空气质量指标 AQI 值的变化范围。并利用 k-means 聚类算法对处理后的数据整体进行聚类，通过选取不同的簇数进行多次聚类，采用聚类有效性指标 DB、SSE，进行聚类结果评价，选取最佳的聚类方案，最终将其分为 3 类，并对聚类结果进行可视化；另外，根据对聚类结果进行分析，

阐述相应的气象特征和污染物浓度变化。

(3)建立了基于 BP 神经网络的空气质量二次预测模型。该模型，首先对 5 个主要的气象特征进行了分析筛选出 4 维特征温度、湿度、气压、风速，其次搭建 BP 神经网络模型，并搭建即时学习框架，根据时间序列，和空气质量监测的时序性，选取 7 个月的数据对未来 3 天的污染物浓度进行预测。此外，BP 神经网络预测模型结果与真实值和一次预测结果进行对比。 为保证预测模型满足两个目标，通过随机选取时间(3 天)对模型进行测试，并与一次预测结果进行比较，结果表明建立的二次预测模型优于一次预测结果。

(4)通过数据处理和对 A、A1、A2、A3 的坐标、方位、距离进行分析，为建立一个协同预测模型，在特征中加入气象特征风向，改进预测模型。

7.2 模型评价

7.2.1 模型优点

(1)设计了较为合理的方法对原始数据进行预处理，同时还考虑到监测站点设备在进行数据测量过程中会遇到各种不同条件的较为复杂的实际工况，使用 3σ准则对数据进行了处理，使原始数据更为精确。

(2)在对数据进行分析时，采用 k-means 聚类方法，对不同的气象特征进行聚类分析，并进行可视化处理，能直观的看到分类的气象特征及对应污染物浓度、空气质量指标 AQI 值的变化。在不确定聚类簇数时采用多组聚类对数据进行聚类，通过指标评价聚类效果，选取最优聚类方案。

(3)现有的空气质量预报模式受到气象(气温、湿度、气压、风速、风向等)和污染源排放的影响，对数据的分析不够深入，无法获取到数据更深层次的联系，而所采用的 BP 神经网络的方法，能通过大量的数据集进行训练，挖掘、捕捉数据直接的深层联系，提高预测结果的精准性，随着训练样本的增长，由于模型具有动态偏差修正功能，能对结果进行实时修正，模型的性能也优于一般方法。BP 神经网络，简单有效，容易实现、训练学习时间快、对基础数据时间长度要求不高。

(4)改进的 GA-BP 神经网络协同预报模型，能更精确的预测污染物浓度、首要污染物、AQI 预测值的相对误差也更小。改进后的 GA-BP 神经网络区域协同预报模型，使得网络收敛速度、预测精度、拟合度以及泛化能力都有所加强。

7.2.2 模型缺点

(1)本文中空气质量预报模型及优化方案是基于 BP 神经网络的空气质量预测模型预测性能有待提高，为之后利用机器学习和深度学习模型进行预测提供了一定的理论基础。同时 BP 神经网络中参数的初始化也是一个问题，网络层数和每层的神经节点个数都需要不断的尝试来确定最优值，网络隐含层神经元数据选择和连接权重初值选取没有理论依据，凭借经验，全局最优收敛没法保证，复杂函数容易落入局部最优解，此外模型的性能对与训练集的要求较高，在进行空气质量预测时，要选取合适的空气质量数据，从而能使得 AQI 预测值的相对误差小，首要污染物的预测精准度高。

(2)建立多种模型相互结合的预测模型。加入影响因子(权重)，包括风速、湿度和降水等因素，挖掘空气质量特征、天气特征、天气预报特征、历史突变相关特征以及外部数据(空间点、污染源清单、文本图像数据)特征，从时间和空间维度提高模型的预测精度。

(3)因遗传算法在初始化种群时采用的是随机初始化种群，有可能陷入局部最优，可以改进初始种群方案；可以改进遗传算子，交叉、变异，在循环过程中可以通过改进遗传算子保留父代种群中的优良解。能加快算法收敛速度。

(4)继续优化空气质量预测模型。针对短时与长时的空气质量预测，训练与测试相对应

的预测模型，达到不仅能准确预测未来一天内的短时间空气质量，更能预测未来长达七天甚至更长时间的空气质量的目的，同时能保证预测精度维持在较高的水平。此外，还可以针对重大的节日活动训练出特定的预测模型，达到特定时间段预测精度提升的效果。对于模型参数的初始化问题，可以针对这些参数再建立一个初始值模型，发掘不同预测场景下初始值之间的内在关联，从而达到预测效果最优化的目的。

(5)建立空气质量与污染源排放之间的深层次关联。针对污染源排放污染物对空气质量具有负面影响这一问题，可以利用深度学习模型，对输入的气象数据、空气质量指数以及污染源排放量、排放地点等参数进行深层次分析，在保证空气质量的前提下，得出可以接受的污染物排放数量，从而将排放量控制在合理范围内，为政府部门进行科学决策提供参考。

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.