

# 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

空气质量预报二次建模

#### 摘 要

本文针对空气质量分类和二次预报建模问题，基于 TSI 天气分型法、多元回归、神经网络、高斯扩散等理论及算法，建立二次预测模型，运用 Excel、MATLAB、SPSS 等软件求解出合理的气象类别和二次预测数据。

本文的特色主要有：1. 在问题一中，将 AQI 纳入聚类指标中，顺应按“气象条件对 AQI 的影响程度”为分类依据的逻辑；2. 建立基于多元线性回归与基于神经网络的两种预测模型，通过比较两模型预测结果的误差值，选取更适用的模型。

首先对各附件数据的缺失值、异常值做均值填充处理，并对逐时数据求日平均值。

针对问题一，直接根据题中已知的 AQI 计算公式和附件一数据，计算得出指定的四天的 AQI 分别为 60、46、109、138，首要污染物均为  $O_3$ ，具体结果见正文表 4。

针对问题二，首先由 6 种污染物浓度的季均值判断其具有季节特征，由 5 个气象指标与污染物间的 Pearson 系数判断其相关性显著；然后基于 TSI 天气分型法，对四个季节的 5 个气象指标分别做主成分分析，对春、夏、秋季提取 3 个主成分，对冬季提取 2 个主成分，再对主成分与 AQI 共同做 K-means 聚类分析，最终将春、冬季的气象条件各分为 6 类，将夏、秋季的气象条件各分为 5 类，按各类的 AQI 均值从小到大的顺序进行排序，例如春季第 1 类到第 6 类气象条件的 AQI 均值分别为 13、35、67、115、183、191，并根据类别均值讨论各类气象条件的特征，具体分类及特征分析见正文表 7。

针对问题三，首先基于多元线性回归理论，将变量筛选后的 12 个气象因子作为自变量，实测与一次预报数据差值作为因变量进行回归分析；然后基于 BP 神经网络理论，将一次预报的 15 个气象因子和污染物浓度作为输入层，将实测污染物浓度作为输出层进行训练。选取三个监测站的有效数据，分别做回归分析和样本训练，建立两种二次预测模型，并得到预测结果。通过比较得出：基于 BP 神经网络的二次预测模型的预报值误差更小，结果显示所有首要污染物均为  $O_3$ ，具体数值见表 8。

针对问题四，由于污染物浓度符合大气扩散理论，故基于高斯扩散理论的高架点源扩散情况，根据四个监测站的坐标算得各自得扩散系数，进而得到扩散模式效果。最后，在基于 IOWA 算子的协同组合预测模型下，利用 MATLAB 计算出相应结果，具体数值见表 10。

最后，本文对模型误差的影响因素进行分析：通过调整 BP 神经网络中的参数，对污染物浓度结果进行了灵敏度分析；并对使用的模型进行优缺点评价，对使用到的神经网络算法以及高斯扩散等技术推广应用到大气、AI 等领域。

**关键词：**AQI；K-means 聚类；BP 神经网络；二次预测模型；高斯扩散

# 一、问题重述

## 1.1 问题背景

空气质量预报模型，是一个提前获知可能发生的大气污染过程的重要模型，人们能够据此采取相应控制措施，减少大气污染对人体健康和生态环境造成的危害，提高环境空气质量。

WRF-CMAQ 模型是一个常用的空气质量预报模型，它能够根据地形、气象和污染物排放清单等信息得到污染物浓度和气象数据的预报数据（称其为“一次预报数据”）。由于污染物浓度实测数据、实际气象条件的变化会影响空气质量预报，故在一次预报数据的基础上再结合实测数据进行二次建模，能够优化预报模型，使得二次建模得到的数据（称其为“二次预报数据”）和实际数据更加接近。

## 1.2 相关数据

题目提供了监测点长期空气质量预报基础数据，包括逐日与逐小时的污染物浓度、气象的一次预报数据和实测数据。其中，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，实测数据的时间跨度为 2019 年 4 月 16 日到 2021 年 7 月 13 日；附件一为监测点 A 的数据，附件二为监测点 B、C 的数据，附件三为监测点 A1、A2、A3 的数据。

并且，题目给出六种常规大气污染物、AQI、污染天气、一次污染物与二次污染物、一种近地面臭氧污染形成机制、臭氧最大 8 小时滑动平均、空气质量指数、空气质量等级及首要污染物、预测时间、检测时间、风向、比湿、边界层高度、长波辐射、短波辐射、地面太阳能、感热通量、潜热通量等重要概念的释意。

最后，题目给出数据异常的几种情形，并指出本题提供了温度、湿度、气压、风向、风速共计五项监测气象指标。

值得注意的是，二次污染物臭氧是在大气中经过一系列化学及光化学反应生成的，这导致精确预测臭氧浓度变化较难。因此，需要重点考虑如何利用现有的实测数据和一次预报数据来建立二次预测模型，以提高臭氧预报的准确度。

## 1.3 具体问题

**问题一：**根据附件一中的数据，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。

**问题二：**根据附件一中的数据，基于气象条件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。注意，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 会下降或上升。

**问题三：**根据附件一、二中的数据，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 $>100\text{km}$ ，忽略相互影响）的二次预报数学模型（要求二次预报模型预测结果中 AQI 预报值的最大相对误差要尽量小，且首要污染物预测值的准确度尽量高）。

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。

**问题四：**如图 1 所示，监测点 A 的临近区域内存在监测点 A1、A2、A3，根据附件一、三中的数据，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。

并讨论：与问题三的二次预报数学模型相比，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？说明原因。

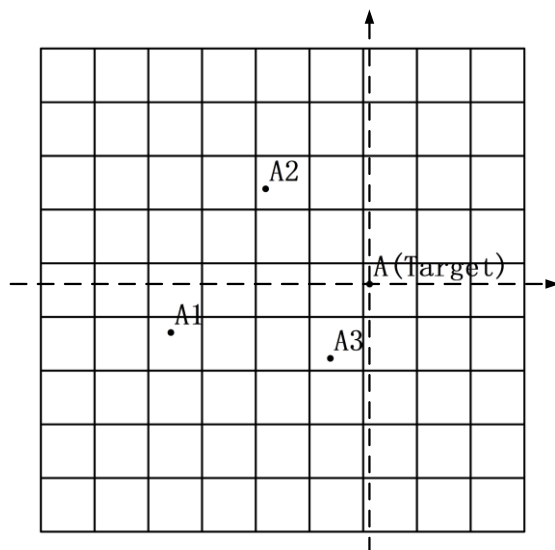


图 1 各监测站点相对位置示意图，正东方向为 x 轴，正北方向为 y 轴，单位：km

A (0, 0) A1 (-14.4846, -1.9699) A2 (-6.6716, 7.5953) A3 (-3.3543, -5.0138)

## 二、问题分析

由于附件中的数据存在缺失、异常等情况，首先应对数据进行预处理，再进行各个问题的分析讨论。

### 2.1 问题一分析

对于问题一，需要计算指定四天的空气质量指数（AQI）和首要污染物。

题目中已经给出 AQI 值的计算方法，即先根据每种污染物  $P$  的当日平均浓度值计算其空气质量分指数（ $IAQI_P$ ），六种污染物的  $IAQI$  值的最大值即为当日 AQI 值，并通过比较  $IAQI$  值的大小可以确定当日首要污染物。根据其计算公式，可以利用 MATLAB 软件编写代码进行求解。

### 2.2 问题二分析

对于问题二，需要根据气象条件对污染物的影响程度，对气象条件进行合理分类。

首先判断污染物浓度是否随季节变化有着明显的变化特征，再决定是否对不同季节的气象条件采取不同的分类。对于气象条件中的温度、湿度、气压、风向、风速共计 5 个指标，先将其与 6 种大气污染物做相关性分析，简单探究其对污染物浓度的影响。然后，基于 TSI 天气分型法，对 5 个气象指标做主成分分析，再将主成分与 AQI 共同做聚类分析，

得出气象条件的分类（这样的聚类方式考虑了气象条件对 AQI 即污染物浓度的影响，而不是单一地根据气象条件来直接分类），最后根据类别均值来分析气象条件的特征。

## 2.3 问题三分析

对于问题三，需要建立一个同时适用于监测站 A、B、C 的二次预报模型，并计算三个监测点在未来三天的常规污染物浓度值、AQI 和首要污染物。二次预报模型应当是以实测数据和一次预报数据相结合为基础。

首先考虑利用多元线性回归做预测模型，将实测数据和一次预报数据的污染物浓度差值作为回归目标，以筛选后的多个气象因子作为因变量，导入三个监测站的历史样本数据，采用最小二乘法求出回归系数，进而得出实测污染物浓度关于一次预报污染物浓度和气象因子的线性组合。再考虑利用 BP 神经网络做预测模型，将一次预报的气象因子和污染物浓度作为输入层，将实测污染物浓度作为输出层，导入三个监测站的历史样本数据，训练出隐含层的预测模型。

两种预测模型都能够根据一次预报数据求得二次预报数据，对其 AQI 最大误差值和污染物浓度预测准确率做讨论，选取误差最小、准确率最高的模型来做预测。

## 2.4 问题四分析

对于问题四，需要建立包含 A、A1、A2、A3 四个监测点的协同预报模型，并计算四个监测点在未来三天的常规污染物浓度值、AQI 和首要污染物。

为更好地构建协同预报模型，先假设污染物浓度在空间中连续。由于涉及大气扩散理论，于是采用高斯扩散模型，主要探讨大空间点源扩散和高架点源扩散两种情况。可以根据四个监测点的坐标，计算得出其扩散系数，再得到四个监测点的扩散模式效果。最后，在基于 IOWA 算子的协同组合预测模型下，可以利用 MATLAB 计算出相应结果。

# 三、模型假设和符号说明

## 3.1 模型假设

- （1）在监测点的检测区域内，每天的污染物排放浓度的静态分布都相同。
- （2）在一天中，不同时刻的污染物浓度变化与 AQI 的高低仅来源于气象条件的影响。
- （3）由于样本数据足够多，故以均值填充法对缺失值做处理后，不影响结果准确性。
- （4）污染物浓度分布在三维空间中是连续的。
- （5）在所有空间中，风速均匀且风向平行于地面。
- （6）污染物之间不发生相互转化。

### 3.2 符号说明

表 1 符号及其说明

符号	说明
$IAQI_P$	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
$BP_{Hi}, BP_{Lo}$	与 $C_P$ 相近的污染物浓度限值的高位值与低位值
$IAQI_{Hi}, IAQI_{Lo}$	与 $BP_{Hi}, BP_{Lo}$ 对应的空气质量分指数
$r$	Pearson 相关系数
$E$	AQI 预报值的相对误差最大值
$S$	首要污染物的预测得分率
$C(x, y, z)$	污染物浓度在坐标 $(x, y, z)$ 点的分布函数
$u$	平均风速（单位 $m/s$ ）
$q$	源强，即单位时间内的排放物（单位 $\mu g/s$ ）
$\sigma_y, \sigma_z$	水平和垂直方向的扩散系数
$H$	污染源的有效高度（单位 $m$ ）

## 四、问题一的求解

利用题中已知的  $AQI$  计算公式，对附件一中数据进行直接计算求解。流程图如下：

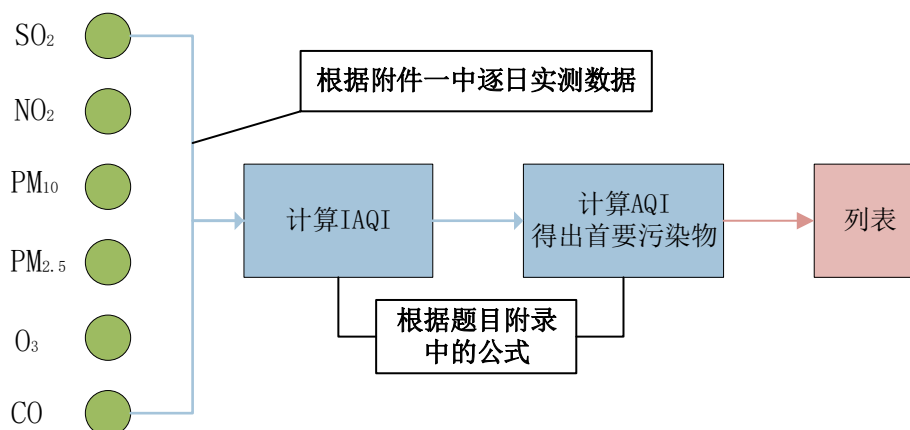


图 2 求解问题一的流程图

#### 4.1 AQI 及首要污染物的计算方法

首先需要得到各项污染物  $P$  的空气质量分指数 ( $IAQI_P$ )，其计算公式如下：

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

其中， $IAQI_P$  的结果进位取整数， $C_P$  指污染物  $P$  的质量浓度值， $BP_{Hi}$ 、 $BP_{Lo}$  指与  $C_P$  相近的污染物浓度限值的高位值与低位值， $IAQI_{Hi}$ 、 $IAQI_{Lo}$  指与  $BP_{Hi}$ 、 $BP_{Lo}$  对应的空气质量分指数。

各项污染物项目浓度限值及对应的空气质量分指数级别如下表所示：

表 2 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	$ug/m^3$
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 $10um$ 颗粒物 ( $PM_{10}$ ) 24 小 时平均	0	50	150	250	350	420	500	600	
6	粒径小于等 $2.5um$ 颗粒物 ( $PM_{2.5}$ ) 24 小 时平均	0	35	75	115	150	250	350	500	

注：（1）臭氧 ( $O_3$ ) 最大 8 小时滑动平均浓度值高于  $800 ug/m^3$  时，不再进行其空气质量分指数计算。（2）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。

然后可以得到空气质量指数 ( $AQI$ )，其计算公式如下：

$$AQI = \max \{ IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{PM_{10}}, IAQI_{PM_{2.5}}, IAQI_{O_3}, IAQI_{CO} \} \quad (2)$$

当  $AQI \leq 50$  时，称当天无首要污染物；当  $AQI > 50$  时，称  $IAQI$  值最大的污染物为首要污染物；若存在两项或两项以上的污染物的  $IAQI$  值均为最大，则它们并列为首要污染物。

## 4.2 AQI 及首要污染物的计算结果

根据上述计算公式（1）和（2），利用 MATLAB 软件编写程序（见附件材料程序一），将附件一中监测点 A 的 2020 年 8 月 25 日到 2020 年 8 月 28 日的逐日污染物浓度实测数据导入 MATLAB 软件中，计算得出结果如表 3 所示：

表 3 六种污染物的  $IAQI$  值及每日  $AQI$  值

	$IAQI$						$AQI$
	$SO_2$	$NO_2$	$PM_{10}$	$PM_{2.5}$	$O_3$	$CO$	
2020/8/25	8	15	27	16	60	13	60
2020/8/26	7	20	24	15	46	13	46
2020/8/27	7	39	37	33	109	15	109
2020/8/28	8	38	47	48	138	18	138

则可得监测点 A 从 2020 年 8 月 25 日到 8 月 28 日每天实测的  $AQI$  和首要污染物如表 4 所示：

表 4 每天实测的  $AQI$  和首要污染物

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	$O_3$

## 五、问题二的建模与求解

本问题主要采用 TSI 天气分型法结合  $AQI$  对气象条件进行分类。

再对问题进行求解之前，先对数据做一系列预处理。

首先，根据污染物浓度季均值分析出其有明显的季节性变化规律，利用 Pearson 相关系数计算得出 5 个气象指标和 6 种污染物浓度之间具有显著相关性；然后，对不同季节的 5 个气象指标的实测逐时数据分别做主成分分析，提取 2 个或 3 个主成分，再结合  $AQI$  值做聚类分析，得到 5 类或 6 类气象条件；最后，根据聚类得到的均值数据分析气象特征。具体流程如下图所示：

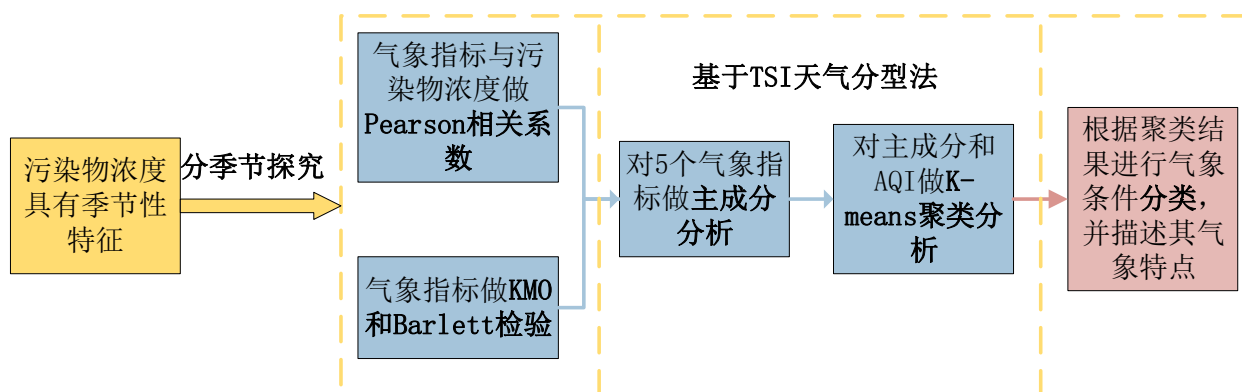


图 3 求解问题二的流程图

## 5.1 数据预处理

### (1) 质量检查

对附件一、二、三中的数据进行质量检查，发现存在部分数据错误、数据缺失等问题，这些问题大多由预报服务器、监测站等设备发生停电、调试、维护等情况引起。

为提高数据的准确性和完整性，利用 MATLAB 软件编写程序（见附件材料程序二），对其做出如下处理：

（a）将逐时污染物浓度、气象指标的缺失值和异常值（例如浓度为负数）进行均值填充，例如 2019 年 4 月 30 日 10:00 的  $PM_{10}$  的浓度值显示缺失，而 9:00 和 11:00 时的  $PM_{10}$  的浓度值分别为  $50\mu g/m^3$  和  $12\mu g/m^3$ ，故 10:00 的  $PM_{10}$  的浓度值被填充为  $31\mu g/m^3$ ；以附件一中实测数据为例，共填充数据 1886 项。

（b）利用箱位图，找出实测数据中偏离数据正常分布的值，并用相应均值替代这些值；以附件一中实测数据为例，共替代数据 32 项。

### (2) 分量扩充

在原数据集中，一次预报数据只有逐时的并且存在多次预报，逐日实测数据无气象条件分量，不方便进行题目求解。

故利用 MATLAB 软件编写程序（见附件材料程序三），对附件一、二、三中各监测点的一次预报数据和实测数据做如下处理，得到新的逐日一次预报数据（共 21 个分量）和逐日实测数据（共 11 个分量）：

（a）选取模型运行当日的当日预报数据为该日的一次预报数据，并对每天逐时的 15 个气象因子和 5 种一次污染物数据分别求该日均值，对二次污染物臭氧求一天中 8 小时滑动平均值的均值，得到逐日的一次预报数据。

（b）由于逐日实测数据中无气象条件数据，故对每天逐时实测数据中的 5 个气象指标分别求该日均值，以均值代表该日实测气象数据，结合逐日实测的 6 种污染物数据得到新的逐日实测数据。

数据预处理的简单流程图如下所示：



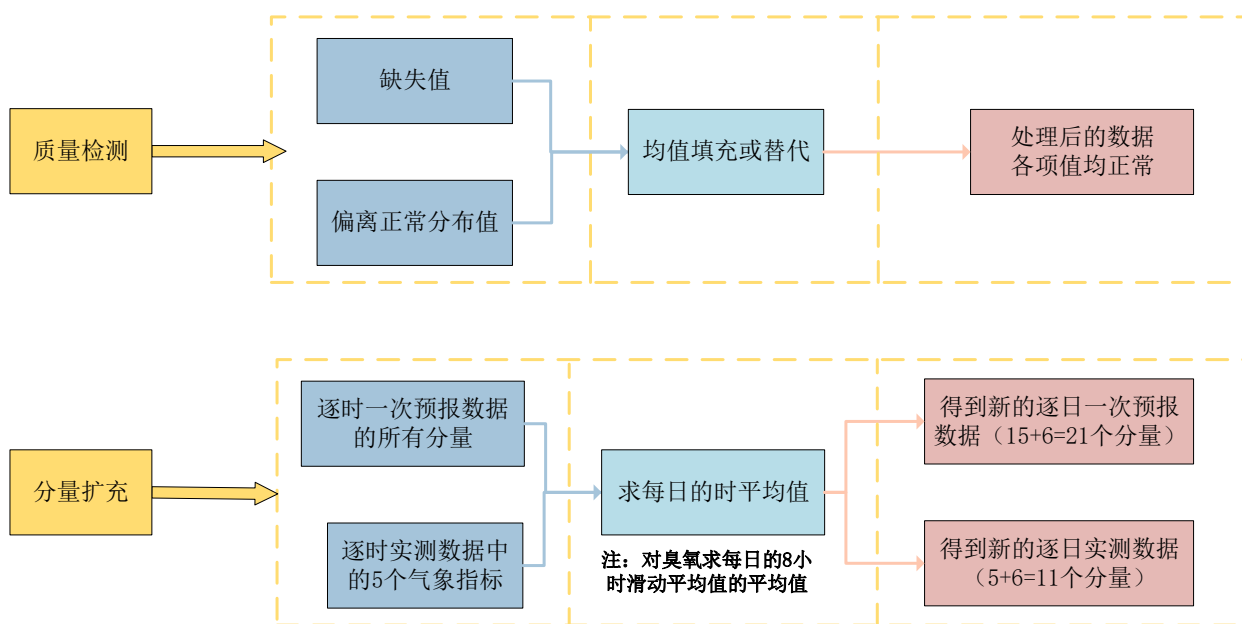


图 4 数据预处理的流程图

在对问题二、三、四的求解中，均以预处理后的数据进行操作。

## 5.2 污染物浓度的季节变化分析

分析题目要求可知，需要根据温度、湿度、气压、风向、风速共计 5 个气象指标对污染物浓度的影响程度，对气象条件进行合理分类。在实际情况中，大气对流、降水量等其他气象条件与这 5 个指标间也存在相互影响，进而影响大气污染物浓度的升高或者降低。

由于气象条件在不同季节中有着显著差异，故先探究污染物浓度是否随着季节变化而有着变化，进而决定是否对不同季节中的气象条件采取不同的分类标准。

将四个季节按照 3-5 月为春季、6-8 月为夏季、9-11 月为秋季、12 月-次年 2 月为冬季的规则进行划分。对预处理后的附件一中逐日实测的 6 种大气污染物（数据跨度为 2019 年 4 月 16 日到 2021 年 7 月 13 日），分别求十二个月和四个季节的每日浓度算术平均值  $\bar{x}$ ，其计算公式如下：

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3)$$

利用 Excel 软件，得到污染物浓度的月平均值和季平均值（见附件材料数据一），并作出可视化折线图和柱状图。

监测点 A 的 6 种大气污染物的月平均浓度值如图 2 与图 3 所示：

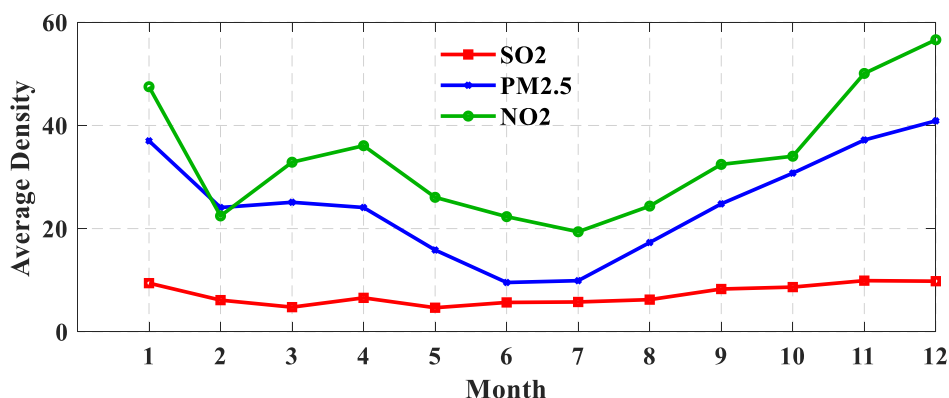


图 5 监测点 A 实测的 SO<sub>2</sub> 、PM<sub>2.5</sub>、NO<sub>2</sub> 的月均浓度变化

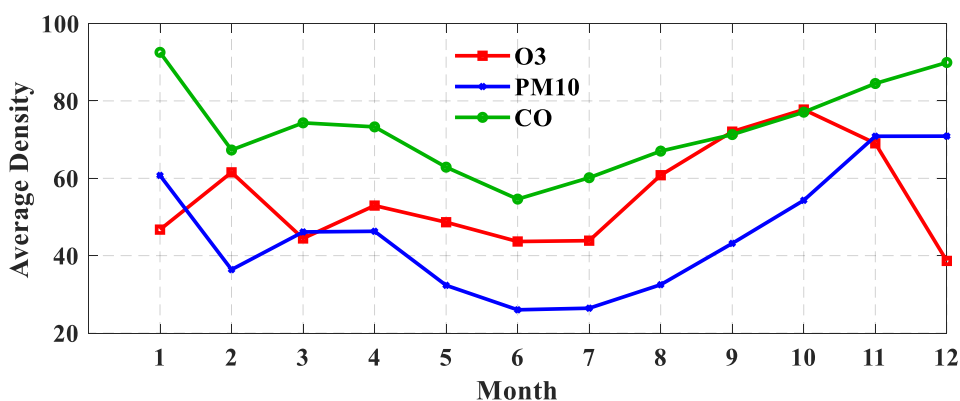


图 6 监测点 A 实测的 O<sub>3</sub> 、PM<sub>10</sub>、CO 的月均浓度变化

监测点 A 的 6 种大气污染物的季平均浓度值如图 5 和图 6 所示：

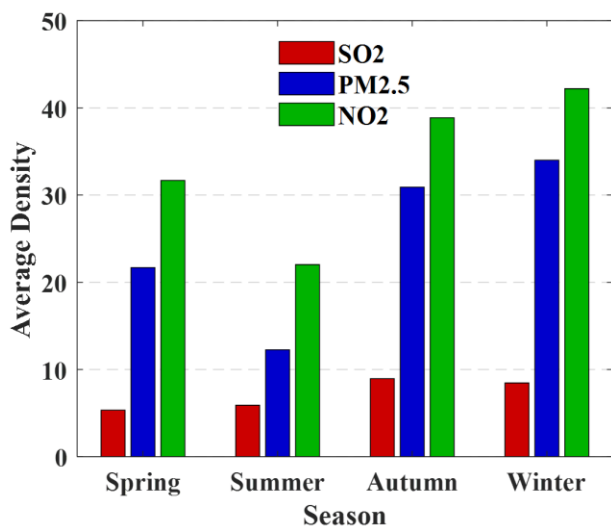


图 7 实测的 SO<sub>2</sub> 、PM<sub>2.5</sub>、NO<sub>2</sub> 季均浓度

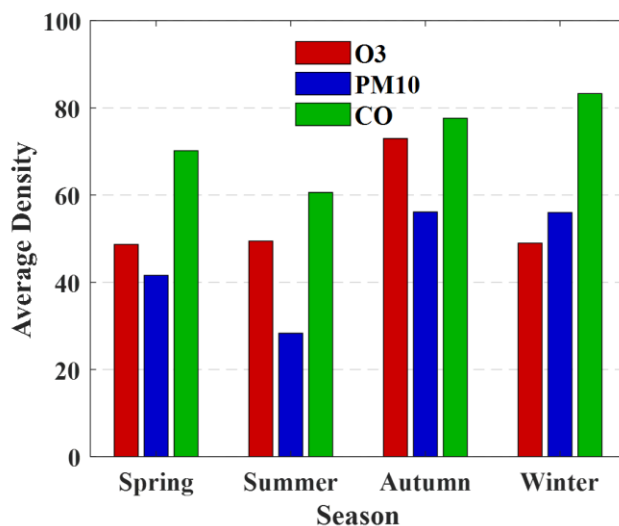


图 8 实测的 O<sub>3</sub> 、PM<sub>10</sub>、CO 的季均浓度

其中，SO<sub>2</sub>、NO<sub>2</sub>、PM<sub>10</sub>、PM<sub>2.5</sub>、O<sub>3</sub> 的浓度单位为 ug/m<sup>3</sup>，CO 的浓度单位为 100 ug/m<sup>3</sup>。

分析数据与图片能够得出：监测点 A 的大气污染物浓度具有明显的季节性变化规律。

PM<sub>10</sub>、PM<sub>2.5</sub>、NO<sub>2</sub>、CO 的浓度冬季明显高于夏季，最高值出现在 12 月或 1 月，最低值出现在

6 月或 7 月，呈现单峰“U”特征；O<sub>3</sub> 的浓度在秋季达到顶峰，呈现单峰“倒 U”型；SO<sub>2</sub> 的浓度整年变化幅度较小，秋冬季略高于春夏季。

由于大气污染物浓度随季节变化有着显著变化，故而在下文的讨论中，将气象指标和大气污染物在春、夏、秋、冬四个季度进行分别探究。

### 5.3 做 TSI 分型前的准备工作

TSI 天气分型法是指选取某一地区和某一段时期的气象要素，通过主成分分析得到能够客观表达这些气象要素的主成分和其得分矩阵，再通过聚类分析得出该地区的主要天气类型。TSI 天气分型方法具有综合性、客观性和确定性等特点[1][2]。

本题要求根据气象条件对污染物的影响程度来进行合理分类，若在 TSI 分型步骤的聚类分析中，将 AQI 与主成分一起做聚类分析，便能够很好地反映不同的气象类别对污染物的影响程度的不同深浅。故而基于 TSI 天气分型法，利用主成分分析和聚类分析对气象条件进行分类。

在对 5 个气象指标做主成分分析之前，需要检测其对污染物浓度的相关性，并检验其相互间的相关关系，以确保 5 个气象指标适合做因子分析。

#### 5.3.1 做 Pearson 相关系数检验

对于监测点 A 的每日实测数据，将温度、湿度、气压、风向、风速视为 5 个自变量（ $X$ ），将 SO<sub>2</sub>、NO<sub>2</sub>、PM<sub>10</sub>、PM<sub>2.5</sub>、O<sub>3</sub>、CO 的浓度视为 6 个因变量（ $Y$ ）。由于气象指标和污染物浓度的量纲不同，故采用能够消除量纲差异的 Person 相关系数对自变量和因变量的相关程度进行度量。Pearson 相关系数计算公式如下：

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (4)$$

其中， $\bar{X}$ 、 $\bar{Y}$  分别代表变量  $X$ 、 $Y$  的  $n$  个数据的算术平均值； $r \in [-1, 1]$ ；当  $r > 0$  时，两个变量是正相关，且  $r$  越接近于 1，代表两个变量的相关程度越强；当  $r < 0$  时，两个变量是负相关，且  $r$  越接近于 -1，相关程度越强。

将预处理后的附件一中逐日实测数据导入 SPSS 软件，得到 5 个气象指标和 6 种大气污染物在不同季节的 Pearson 相关系数如下表所示：

表 5 2019-2021 年监测点 A 实测 5 个气象指标和 6 种大气污染物的 Pearson 相关系数

相关系数	大气污染物	温度 (°C)	湿度 (%)	气压 (MBar)	风速 (m/s)	风向 (°)
春季	SO <sub>2</sub>	-0.092	-0.343	0.256	-0.158	-0.019
	NO <sub>2</sub>	-0.424	0.220	0.297	-0.504	0.224
	PM <sub>10</sub>	-0.094	-0.389	0.318	-0.227	-0.078
	PM <sub>2.5</sub>	-0.226	-0.181	0.359	-0.343	-0.108
	O <sub>3</sub>	0.369	-0.659	-0.029	0.250	-0.022
	CO	-0.397	0.119	0.385	-0.368	-0.010
夏季	SO <sub>2</sub>	0.170	-0.210	-0.106	-0.085	0.187

	NO <sub>2</sub>	-0.462	0.437	-0.095	-0.468	-0.328
	PM <sub>10</sub>	0.379	-0.341	-0.282	-0.051	0.044
	PM <sub>2.5</sub>	0.219	-0.202	-0.342	-0.175	0.038
	O <sub>3</sub>	0.673	-0.713	-0.319	0.211	0.088
	CO	-0.210	0.230	-0.312	-0.309	-0.100
秋季	SO <sub>2</sub>	-0.021	-0.445	0.196	-0.005	0.041
	NO <sub>2</sub>	-0.351	0.011	0.247	-0.471	0.082
	PM <sub>10</sub>	-0.082	-0.425	0.288	-0.205	0.122
	PM <sub>2.5</sub>	-0.003	-0.233	0.198	-0.309	0.175
	O <sub>3</sub>	0.525	-0.600	-0.130	0.227	-0.006
	CO	-0.166	-0.033	0.292	-0.359	0.217
冬季	SO <sub>2</sub>	0.176	-0.330	0.007	-0.087	-0.015
	NO <sub>2</sub>	0.054	-0.033	-0.129	-0.481	0.093
	PM <sub>10</sub>	0.183	-0.158	-0.118	-0.349	0.087
	PM <sub>2.5</sub>	0.163	-0.049	-0.148	-0.411	0.087
	O <sub>3</sub>	0.482	-0.427	-0.098	0.183	-0.115
	CO	-0.051	0.106	-0.202	-0.275	0.176

注：SPSS 结果显示以上 Pearson 系数均通过 0.01 或 0.05 显著水平 t 检验。

根据上表结果，能够对气象指标对污染物浓度的影响程度做出简单分析。以温度为例：对于 SO<sub>2</sub>、PM<sub>10</sub>、PM<sub>2.5</sub>，其浓度与温度在春秋季节呈负相关，在夏冬季呈正相关；对于 NO<sub>2</sub>，其浓度与温度在春、夏、秋季呈负相关，在冬季呈正相关；对于 O<sub>3</sub>，其浓度与温度在全年呈正相关，温度升高会加速臭氧浓度增加；对于 CO，其浓度与温度在全年呈负相关。

### 5.3.2 做 KMO 和 Bartlett 检验

利用 SPSS 软件，对四季的 5 个气象指标分别做 KMO 和 Bartlett 检验（具体结果见本文附录一），得出四个季节的 KMO 值分别为 0.383、0.601、0.435、0.591，Bartlett 检验中 sig 值均为 0.000（小于显著水平 0.05），说明 5 个变量之间存在相关关系，适合做因子分析。

## 5.4 TSI 分型法结合 AQI 进行分类

基于 TSI 分型法，分季节对 5 个气象指标提取主成分，再将主成分与 AQI 做聚类分析。

### 5.4.1 做主成分分析

主成分分析法是一种客观的降维方法，它在保证信息损失尽可能少的前提下，能够将反应样本某项特征的多个指标变量转化为少数综合变量。具体步骤如下[3]：

【Step1】构建样本矩阵。

针对每个季节，对  $x_j (j=1,2,\cdots,5)$  分别表示 5 个气象指标；用  $i=1,2,\cdots,n$  分别表示第 1 天、第 2 天…第  $n$  天，不同季节的  $n$  值不同；第  $i$  天  $x_j$  的取值记作  $(a_{i1}, a_{i2}, \cdots, a_{in})$ 。构造矩阵  $A = (a_{ij})_{n \times 5}$ 。

【Step2】对原始数据进行标准化处理。

将各指标值  $a_{ij}$  转化为标准化指标  $\bar{a}_{ij}$ ，公式如下：

$$\bar{a}_{ij} = \frac{a_{ij} - u_j}{s_j} (i=1, 2, \dots, n; j=1, 2, \dots, 5) \quad (5)$$

其中， $u_j = \frac{1}{n} \sum_{i=1}^n a_{ij}$ ,  $s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (a_{ij} - u_j)^2}$ 。

【Step3】计算相关系数矩阵。

记相关系数矩阵为  $R = (r_{ij})_{s \times s}$ ，且

$$r_{ij} = \frac{\sum_{k=1}^n \bar{a}_{ki} \cdot \bar{a}_{kj}}{n-1} (i, j=1, 2, \dots, 5) \quad (6)$$

其中， $r_{ii} = 1$ ;  $r_{ij} = r_{ji}$  表示第  $i$  个气象指标与第  $j$  个气象指标的相关系数。

【Step4】计算特征值和特征向量，得出主成分。

计算相关系数矩阵  $R$  的特征值  $\lambda_j \geq 0 (j=1, 2, \dots, 5)$  及对应的标准化特征向量

$u_j = (u_{1j}, u_{2j}, \dots, u_{5j})^T (j=1, 2, \dots, 5)$ 。由特征向量组成 5 个新的指标变量如下：

$$\begin{cases} y_1 = u_{11}\bar{x}_1 + u_{21}\bar{x}_2 + \dots + u_{51}\bar{x}_5 \\ y_2 = u_{12}\bar{x}_1 + u_{22}\bar{x}_2 + \dots + u_{52}\bar{x}_5 \\ \dots\dots\dots \\ y_5 = u_{15}\bar{x}_1 + u_{25}\bar{x}_2 + \dots + u_{55}\bar{x}_5 \end{cases} \quad (7)$$

其中， $y_j$  分别代表第  $j$  主成分。

【Step5】选取  $p (p \leq 5)$  个主成分。

主成分  $y_j$  的信息贡献率公式如下：

$$b_j = \frac{\lambda_j}{\sum_{k=1}^5 \lambda_k} (i=1, 2, \dots, 5) \quad (8)$$

主成分  $y_1$ 、 $y_2$ 、 $\dots$ 、 $y_p$  的累计贡献率公式如下：

$$\alpha_p = \frac{\sum_{k=1}^p \lambda_k}{\sum_{k=1}^5 \lambda_k} \quad (9)$$

当 $\alpha_p$ 接近于 1 时, 选择前  $p$  个气象指标  $y_1$ 、 $y_2$ 、 $\cdots$ 、 $y_p$  作为  $p$  个主成分, 代替原来的 5 个气象指标, 进而可对  $p$  个主成分进行综合分析。

对预处理后的附件一中监测点 A 逐时实测的 5 个气象指标数据日均值导入 SPSS 软件, 对其做主成分分析, 得到不同主成分的初始特征值及其贡献率如下表所示:

表 6 不同主成分的初始特征值及其贡献率

春季				夏季			
成分	初始 特征值	贡献率		成分	初始 特征值	贡献率	
		方差的 %	累积 %			方差的 %	累积 %
1	1.978	39.555	39.555	1	2.356	47.124	47.124
2	1.266	25.327	64.882	2	.954	19.080	66.204
3	.923	18.460	83.342	3	.903	18.068	84.272
4	.694	13.889	97.231	4	.717	14.339	98.611
5	.138	2.769	100.000	5	.069	1.389	100.000
秋季				冬季			
成分	初始 特征值	贡献率		成分	初始 特征值	贡献率	
		方差的 %	累积 %			方差的 %	累积 %
1	1.953	39.063	39.063	1	2.125	42.503	42.503
2	1.269	25.380	64.443	2	.998	19.970	62.473
3	.958	19.158	83.601	3	.859	17.170	79.643
4	.629	12.587	96.188	4	.702	14.031	93.674
5	.191	3.812	100.000	5	.316	6.326	100.000

为使每个主成分能够代表的信息量更多, 主成分的选取标准是特征值需要大于 0.95、累计方差贡献率尽可能为 80%以上。观察上表可知, 对春季、夏季和秋季均提取 3 个主成分, 对冬季提取 2 个主成分。

#### 5.4.2 做 K-means 聚类分析

将 6 种污染物的浓度简化成为单一的指数值——AQI, 为使气象条件的分类标准与其对污染物扩散或沉降的影响程度相关, 故对不同季节的 2 个或 3 个主成分及 AQI 共同进行 K-means 聚类分析。

K-means 聚类是一类非监督学习聚类方法, 分类容易且可解释性较强。以春季数据为例, 基本步骤如下[4][5]:

【Step1】对于得到的主成分和 AQI 样本集  $\{X_i\}$  进行标准化处理, 其中

$X_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$  代表第  $i$  天的数据,  $x_{ij} (j=1,2,3)$  代表第  $i$  天的第  $j$  个主成分值,  $x_{i4}$  代表第  $i$  天的 AQI 值。将数值  $x_{ij}$  转化为标准化数值  $\hat{x}_{ij}$ , 公式如下:

$$\hat{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} (i=1,2,\dots,m; j=1,2,3,4) \quad (10)$$

$$\text{其中, } \bar{x}_j = \frac{1}{n} \sum_{i=1}^n a_{ij}, s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (a_{ij} - \bar{x}_j)^2}。$$

【Step2】选取  $k$  个数据点  $C_l = (\hat{x}_{i1}, \hat{x}_{i2}, \hat{x}_{i3}, \hat{x}_{i4})$  作为聚类的中心, 其中  $l=1,2,\dots,k$ 。

【Step3】对样本集中所有数据点求出与  $k$  个中心点  $C_l$  的距离, 并将每个数据点都归到与之距离最近的中心点类。此处采用欧式距离对变量相似度进行度量。

【Step4】将所有数据点归类后, 共得到  $k$  类。再重新计算每类数据的中心, 例如若第

$$l \text{ 类中共有 } m \text{ 个数据点, 则新中心为 } C_{l1} = \left( \frac{\sum_{i=1}^m \hat{x}_{i1}}{m}, \frac{\sum_{i=1}^m \hat{x}_{i2}}{m}, \frac{\sum_{i=1}^m \hat{x}_{i3}}{m}, \frac{\sum_{i=1}^m \hat{x}_{i4}}{m} \right)。$$

【Step5】若新中心与上一次的中心的欧式距离小于某一规定阈值, 则说明分类区域稳定, 该类里的点均收敛至中心; 若距离超过该阈值, 说明该分类不稳定, 还需要对样本集从[Step2]开始迭代计算。

注: K-means 聚类法采用肘部法则来确定初始  $k$  值, 此项工作为 SPSS 软件内部自行运行计算, 此处不进行赘述。

通过 K-means 聚类法不断迭代, 得出聚类结果 (具体结果见本文附录二)。

## 5.5 气象条件的分类结果

根据聚类结果, 将春季和冬季分为 6 类气象条件, 将夏季和秋季分为 5 类气象条件。对每个季节, 将气象条件按其对 AQI 扩散的有效程度进行排类, 即第一类气象条件最有利于污染物的扩散, 此时 AQI 最低、空气质量最好; 第五类或第六类气象条件会加速污染物的沉降, 此时 AQI 最高、空气质量最差。

各季节不同类别气象条件的 AQI 均值如下图所示:

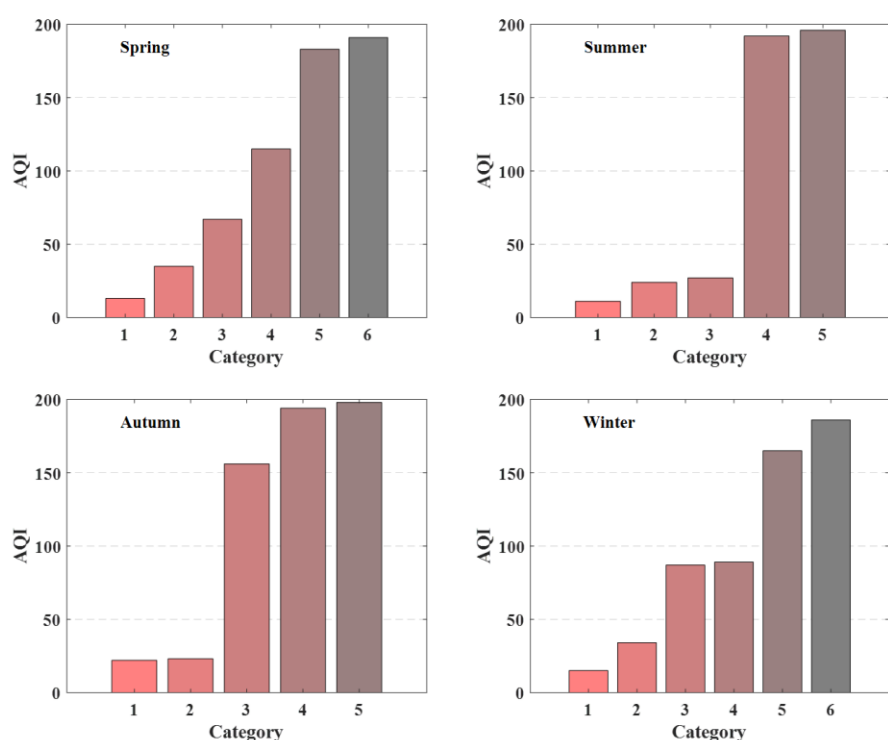


图 9 不同类别气象条件的 AQI 值可视化图

各季节气象条件的具体分类及其特征如下表所示：

表 7 各季节的气象条件类别及其特征

季节	类别	AQI 均值	主要特征
春季	1	13	温度适中，湿度适中，气压很低，风向为 W-WN 区域，风速适中
	2	35	温度适中，湿度很高，气压较低，风向为 EN-E 区域，风速较低
	3	67	温度较低，湿度很低，气压很高，风向为 ES-S 区域，风速很大
	4	115	温度很低，湿度较高，气压较高，风向为 S-WS 区域，风速很小
	5	183	温度很高，湿度适中，气压较低，风向为 W-WN 区域，风速适中
	6	191	温度较高，湿度适中，气压适中，风向为 E-ES 区域，风速较小
夏季	1	11	温度适中，湿度适中，气压很高，风向为 ES-S 区域，风速很小
	2	24	温度很低，湿度很高，气压较高，风向为 S-WS 区域，风速较小
	3	27	温度较低，湿度较高，气压适中，风向为 S-WS 区域，风速适中
	4	192	温度较高，湿度较低，气压很低，风向为 WS-W 区域，风速较大
	5	196	温度很高，湿度很低，气压很低，风向为 E-ES 区域，风速适中
秋季	1	22	温度很低，湿度适中，气压很高，风向为 ES-S 区域，风速较大
	2	23	温度适中，湿度很高，气压适中，风向为 WS-W 区域，风速很大
	3	156	温度适中，湿度适中，气压较高，风向为 N-EN 区域，风速适中
	4	194	温度较高，湿度很低，气压很低，风向为 ES-S 区域，风速适中
	5	198	温度很高，湿度较低，气压很低，风向为 S-WS 区域，风速很小
冬	1	15	温度较低，湿度很高，气压较高，风向为 S-WS 区域，风速适中



季	2	34	温度很低，湿度很低，气压很高，风向为 ES-S 区域，风速很大
	3	87	温度很高，湿度很低，气压较低，风向为 W-WN 区域，风速适中
	4	89	温度适中，湿度较高，气压很低，风向为 EN-E 区域，风速很小
	5	165	温度较低，湿度适中，气压适中，风向为 S-WS 区域，风速很小
	6	186	温度较低，湿度适中，气压适中，风向为 ES-S 区域，风速很小

其中，风向根据如下图形和表格为标准进行分类：

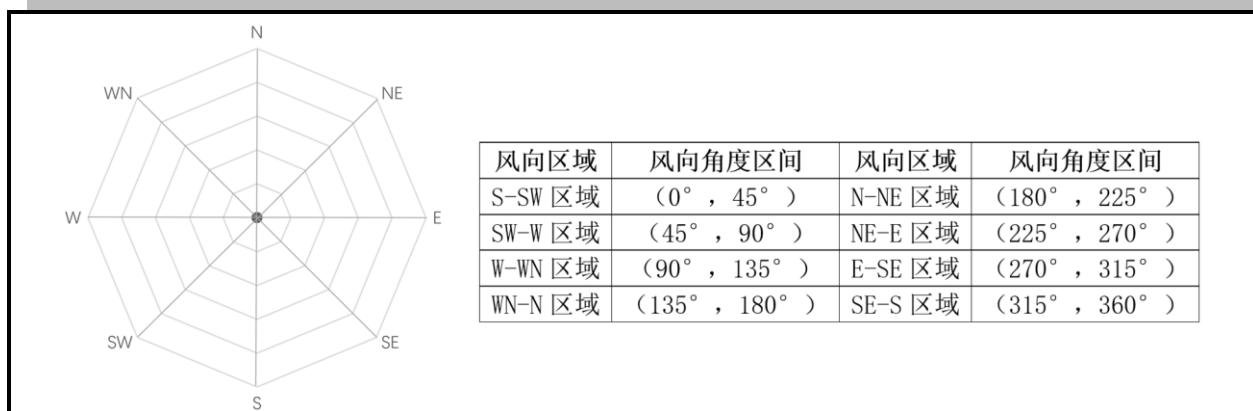


图 10 风向玫瑰图和风向区域分类标准

## 六、问题三的建模和求解

本题要求根据一次预报数据结合实测数据，建立预测模型，得到更加接近准确值的二次预测数据，并且此模型同时适用于 A、B、C 三个监测点。

基于多元线性回归和 BP 神经网络理论，分别建立两种二次预测模型，并对其二次预报值的 AQI 误差和首要污染物准确率做出比较，结论是基于 BP 神经网络的预测模型误差更小、准确率更高，故根据该模型计算出二次预报数据。具体流程图如下：

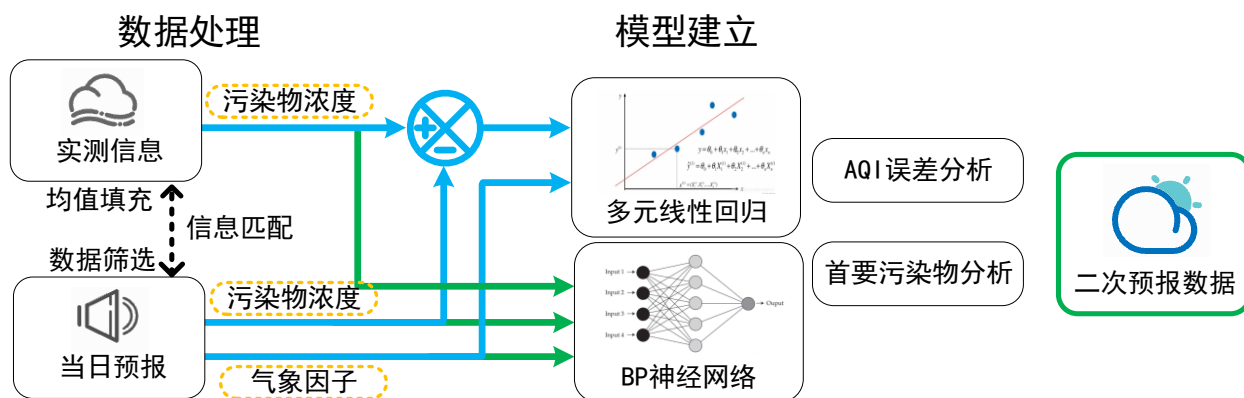


图 11 问题三流程图

### 6.1 数据处理与选取

由于预测模型具有明显时间序列性，故选取临近需预测天的样本数据进行分析即可。

为预测 2021 年 7 月 13 日至 15 日的 6 种污染物浓度值，并且预测模型要同时适用于监测点 A、B、C，故选取三个监测站在 5 月 1 日至 7 月 12 日的共 219 组样本数据进行预测模型的建立。

经对比分析，发现部分组数据的一次预报数据全部缺失。为满足实测数据与一次预报数据的一一对应性，选取有效样本共 214 组。

每组样本数据应包括逐日实测的 6 种污染物浓度值、逐日一次预报的 6 种污染物浓度值和 15 种气象因子（分别为近地 2 米温度、地表温度、比湿、湿度、近地 10 米风速、近地 10 米风向、雨量、云量、边界层高度、大气压、感热通量、潜热通量、长波辐射、短波辐射、地面太阳能辐射）。

## 6.2 基于多元线性回归分析的预测模型

回归分析是一种预测性的建模方法，它能够探究出因变量和自变量间的关系[7]。本题根据具有多个分量的实测数据和一次预报数据，基于多元线性回归分析，构建预测模型进行二次预报，流程示意图如下：

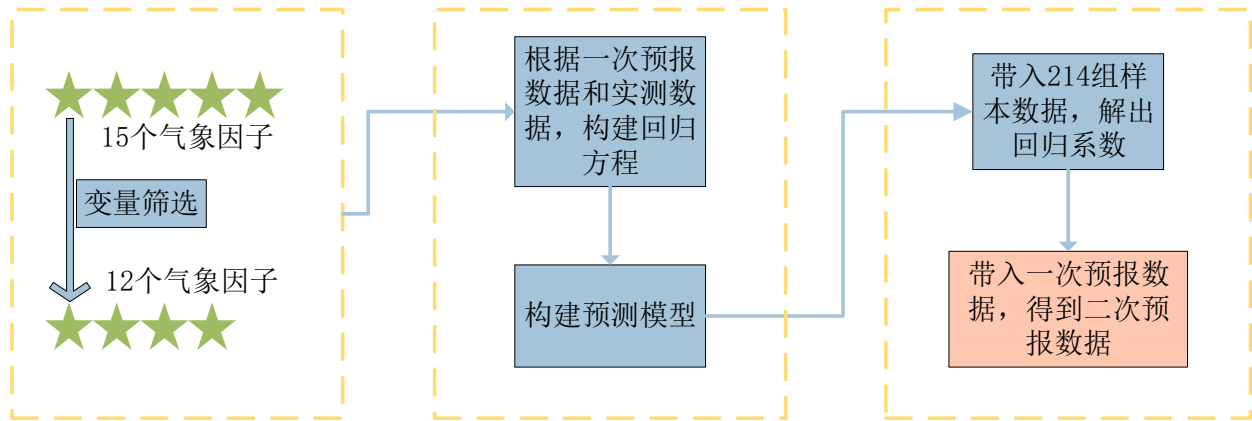


图 12 构建基于多元线性回归分析的预测模型求解问题三

### 6.2.1 建立预测模型

由于 15 个气象因子之间具有一定相关性，既存在信息冗余导致模型准确度降低，又加大计算复杂度，故在进行回归分析前对 15 个气象因子进行变量筛选。将监测点 A、B、C 的所有逐日一次预报的 15 个气象因子数据导入 SPSS 软件，采用逐步回归法进行筛选，得到结果显示短波辐射、雨量、云量这 3 个气象因子的显著性大于 0.05，故将其剔除。

根据逐日一次预报数据和逐日实测数据建立回归方程如下：

$$z = y - y' = b_0 + b_1x_1 + b_2x_2 + \cdots + b_{12}x_{12} + u \quad (11)$$

其中， $y$  代表实测的污染物浓度， $y'$  代表一次预报的污染物浓度， $x_1$ 、 $x_2$ 、 $\cdots$ 、 $x_{12}$  分别代表一次预报的 12 个气象因子， $u$  代表误差。

使用最小二乘法进行参数估计。

将计算出的 13 个系数代回方程（11）中，则可得到回归方程。

则基于多元线性回归的每种污染物浓度的预测模型均如下所示：

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_{12}x_{12} + y'$$

$$\begin{cases} \sum z = nb_0 + b_1 \sum x_1 + b_2 \sum x_2 + \cdots + b_{12} \sum x_{12} \\ \sum x_1 z = b_0 \sum x_1 + b_1 \sum x_1^2 + \cdots + b_{12} \sum x_1 x_{12} \\ \cdots \cdots \\ \sum x_{12} z = b_0 \sum x_{12} + b_1 \sum x_1 x_{12} + \cdots + b_{12} \sum x_{12}^2 \end{cases} \quad (12)$$

对于 6 种污染物，分别将 214 组样本数据中对应的一次预报浓度、实测浓度以及 12 个气象因子数据导入 MATLAB 软件中，编写程序（见附件材料程序四），解出其回归系数（见附录三），进而得到 6 个预测方程。

### 6.2.2 预测结果分析

根据上述基于多元线性回归的预测模型，计算出 6 种污染物在 214 组样本对应天的二次预报浓度，并将其与对应天的一次预报浓度、实测浓度做对比如下图所示：

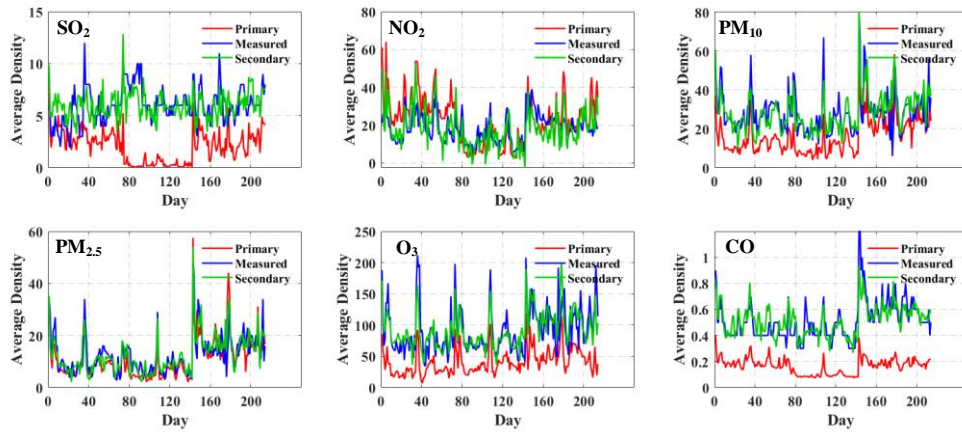


图 13 一次预报浓度、实测浓度与多元线性回归模型的二次预报浓度对比（多元线性回归）

214 组样本对应天的一次预报 AQI、实测 AQI、二次预报 AQI 对比图如下：

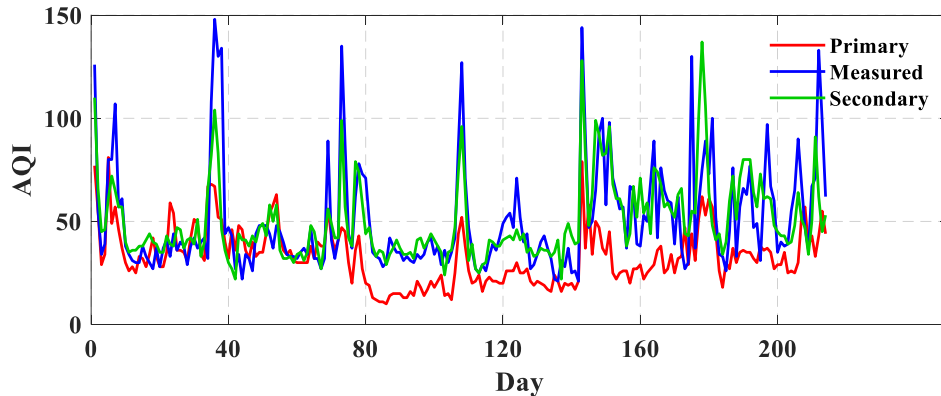


图 14 一次预报 AQI、实测 AQI、二次预报 AQI 对比图（多元线性回归）

214 组样本对应天的一次预报首要污染物、实测首要污染物、二次预报首要污染物对比图如下，其中 1、2、3、4、5、6 分别代表 SO<sub>2</sub>、NO<sub>2</sub>、PM<sub>10</sub>、PM<sub>2.5</sub>、O<sub>3</sub> 和 CO。

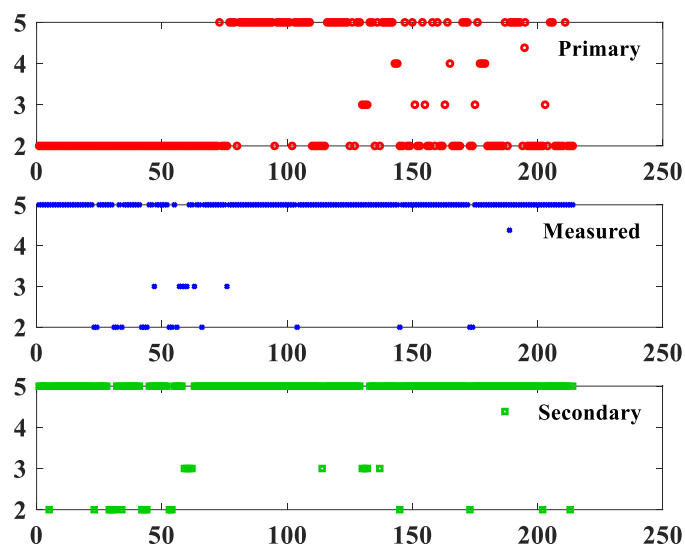


图 15 一次预报首要污染物、实测首要污染物、二次预报首要污染物对比图（多元线性回归）

根据上述三张对比图可以明显看出，无论是污染物浓度、AQI 还是首要污染物，一次预报数据都与实测数据存在较大误差，二次预报数据都与实测数据的吻合性较好，说明二次预报数据更加贴合实际数据，基于多元线性回归分析的二次预测模型结果较为理想。

### 6.3 基于 BP 神经网络的预测模型

BP 神经网络能够基于样本数据集做数据预测，其基本思想和优点如下：无需事先确定输入和输出之间映射关系的数学方程，仅通过自身系统对样本集的训练，学习某种规则，就能在给定输入值时得到最接近期望输出值的结果。

#### 6.3.1 建立预测模型

采取 16-12-1 型结构建立 BP 神经网络预测模型，即对输入层设置一次预报的 15 个气象因子和 1 个污染物浓度值共计 16 个输入变量，对内部隐含层设置 12 个神经元，对输出层设置实测的对应污染物浓度值为 1 个输出变量，具体结构图如下所示：

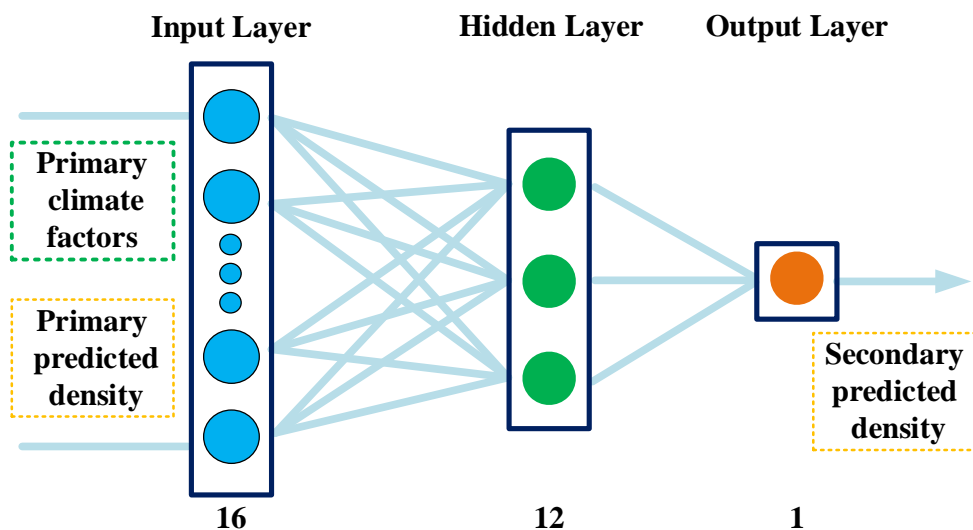


图 16 BP 神经网络结构图

本预测模型包含输入节点数 16 的输入层、节点数为 12 的隐层数、节点数为 1 的输出层 ( $M=16, K=12, N=1$ )，输入层到隐含层的权重为  $w_{ij}$ ，隐含层到输出层的权重为  $w_{jk}$ ，输入层到隐含层的偏置为  $a_j$ ，隐含层到输出层的偏置为  $b_k$ ，学习效率为  $\eta$ ，激励函数取  $\sigma$  函数  $g(x) = \frac{1}{1+e^{-x}}$ ，各层的输出定义等式如下：

$$(1) \text{ 输入层输出: } O_m = I_m (m=1, 2, \dots, 16);$$

$$(2) \text{ 隐含层输出: } H_j = g\left(\sum_{m=1}^M w_{jm} x_m + a_j\right);$$

$$(3) \text{ 输出层输出: } O_k = \sum_{j=1}^K w_{jk} H_j + b_k.$$

然后参数调整以及实现如下计算：

$$(1) \text{ 误差计算: } E = \frac{1}{2} \sum_{k=1}^N (Y_k - O_k)^2 (i=1, 2, \dots, M; j=1, 2, \dots, K; k=1, 2, \dots, N);$$

$$(2) \text{ 权值更新: } w_{ij} = w_{ij} + \eta H_j (1 - H_j) x_i \sum_{k=1}^N w_{jk} e_k;$$

$$(3) \text{ 偏置更新: } a_j = a_j + \eta H_j (1 - H_j) \sum_{k=1}^N w_{jk} e_k.$$

对前文选取的 214 组有效样本，选取 154 个样本作为训练集，剩余 60 个样本作为校验集。将数据导入 MATLAB 软件，编写程序（见附件材料程序五），可得到 6 种污染物浓度分别对应的二次预测模型即隐含层方程。

对 6 种污染物的隐含层方程的训练结果显示其 R 值在 0.8 或 0.9 量级，说明拟合效果好。例如，针对 CO 浓度的训练过程数据如下：

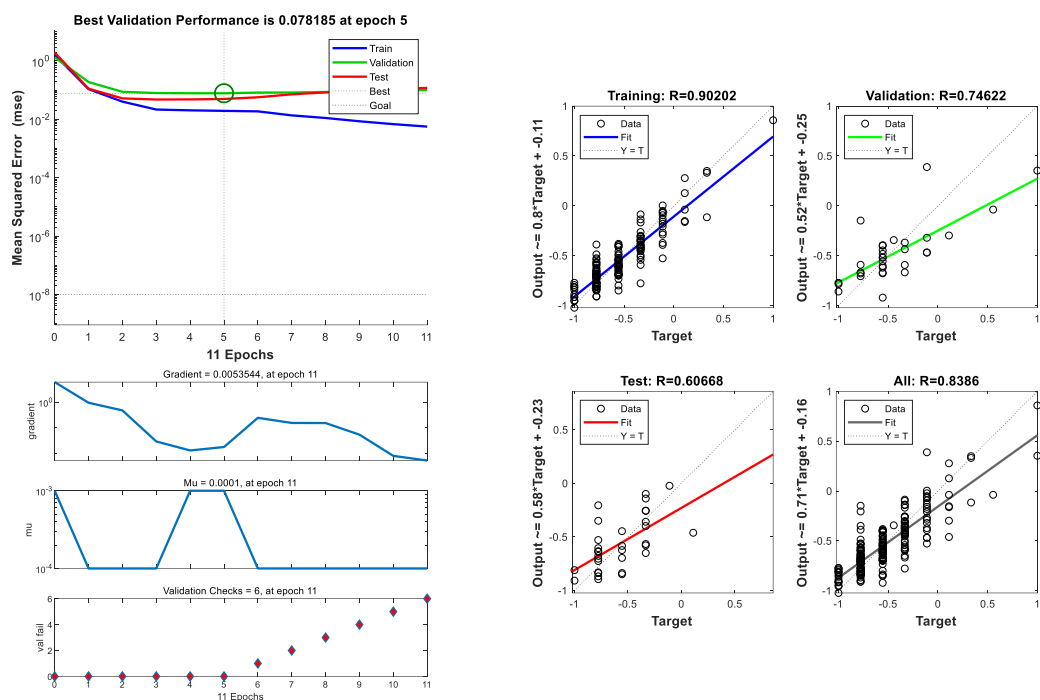


图 17 对 CO 浓度的训练过程数据

### 6.3.2 预测结果分析

根据上述已经构造出隐含层方程的 BP 神经网络，将 60 个校验样本输入其中，得出 6 种污染物在 60 组样本对应天的二次预报浓度、二次预报 AQI、二次预报首要污染物。做出与多元线性回归模型中类似的对比图如以下三张图片所示：

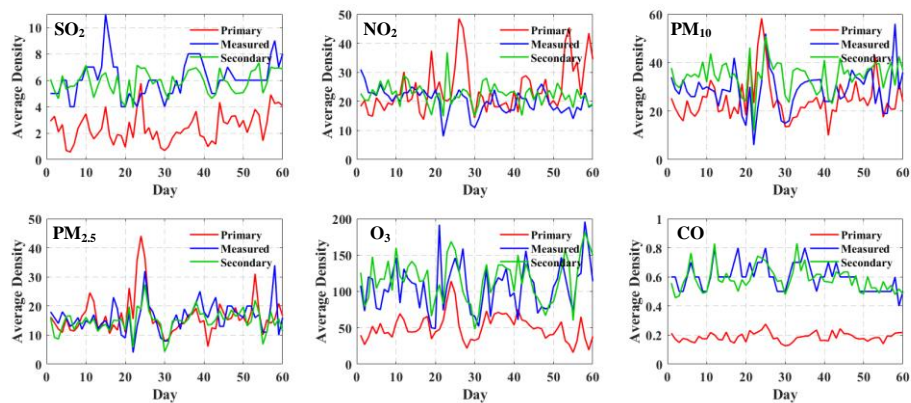


图 18 一次预报浓度、实测浓度与多元线性回归模型的二次预报浓度对比（60 组校验样本）

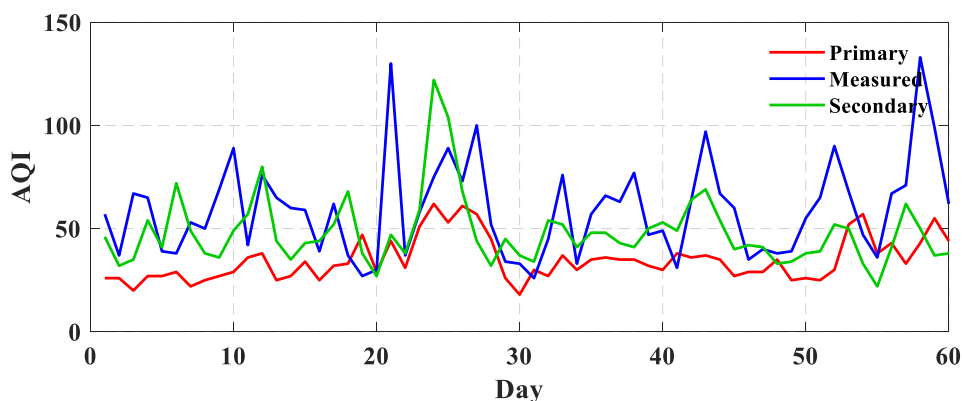


图 19 一次预报 AQI、实测 AQI、二次预报 AQI 对比图（60 组校验样本）

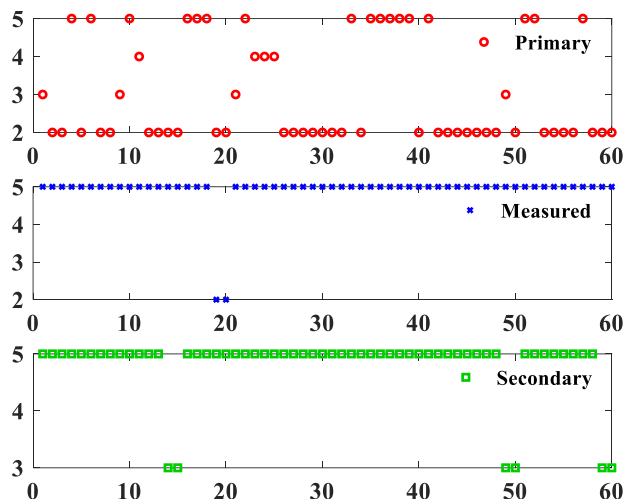


图 20 一次预报首要污染物、实测首要污染物、二次预报首要污染物对比图（60 组校验样本）

根据上述三张对比图可以看出，二次预报数据的污染物浓度、AQI 和首要污染物都与实测数据的吻合性较好，且远远超过一次预报数据的吻合度，说明二次预报数据更加贴合实际数据，基于 BP 神经网络的二次预测模型结果较为理想。

## 6.4 其他预测方法

### （1）LSTM 神经网络

假设监测点附近的天气不会发生突变，且污染物浓度不会随气象条件变化而变化，而是跟随历史数据的污染物变化趋势而变化，则可使用 LSTM 预测模型进行预测。

但根据问题二的结果可知，污染物浓度的变化在极大程度上受到气象条件的影响，故 LSTM 预测模型不适用于此题。

### （2）相空间重构法

相空间指一个能够表示出一系统所有可能状态的空间，系统每个可能的状态都有一相对应的相空间的点。基于时间排列，气象条件和污染物浓度可以看作一个确定的动力系统。

相空间重构法用于本预测问题上，基本理论可以简述如下：对于最近的气象条件和污染物浓度，在过去的时间序列上寻找一段与其相似度最高的气象条件和污染物浓度序列，



依据过去此段序列的变化趋势来预测现在的污染物浓度变化趋势。

根据理论，相空间重构法能够基于实测数据和一次预报数据，进行污染物浓度的二次预测，但此处不进行详细讨论。

### 6.5 两种模型的预测结果准确性分析

题目要求二次预报模型的预测结果中 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，下面对两种预报模型的结果进行检验。

#### (1) AQI 预报值的误差分析

AQI 预报值得最大相对误差表达式  $E$  如下所示：

$$E = \max_{i=1,2,\cdots,n} \{E_i\} = \max_{i=1,2,\cdots,n} \left\{ \left| \frac{A_i - a_i}{A_i} \right| \right\} \tag{13}$$

其中， $E_i$  代表第  $i$  天的 AQI 预报值的相对误差， $A_i$  代表第  $i$  天的 AQI 的实际值， $a_i$  代表第  $i$  天的 AQI 的预测值。

计算得  $E_{\text{多元线性回归}} = 1.354 > E_{\text{BP神经网络}} = 0.8947$ ，表明基于 BP 神经网络的预测模型比基于多元线性回归的预测模型的 AQI 最大相对误差更小。

#### (2) 首要污染物预测准确度分析

记  $n$  个检测样本的首要污染物预测得分率为  $S$ ，表达式如下：

$$S = \frac{\sum_{i=1}^n x_i}{n} \tag{14}$$

其中，若第  $i$  天的首要污染物的预测结果和实际的相同，则  $x_i = 1$ ；否则  $x_i = 0$ 。

计算得  $S_{\text{多元线性回归}} = \frac{191}{214} < S_{\text{BP神经网络}} = \frac{11}{12}$ ，表明基于 BP 神经网络的预测模型比基于多元线性回归的预测模型的首要污染物预测准确率更高。

综上所述，基于 BP 神经网络的预测模型比基于多元线性回归的预测模型的预测准确率更高，故采用前者预测模型对问题三进行求解(见附件材料程序六)，得到指定三天的三个监测点的二次预报数据如下表所示：

表 8 基于 BP 神经网络的二次预报数据

预报日期	地点	二次模型日值预测							
		SO <sub>2</sub> (μg/m <sup>3</sup> )	NO <sub>2</sub> (μg/m <sup>3</sup> )	PM <sub>10</sub> (μg/m <sup>3</sup> )	PM <sub>2.5</sub> (μg/m <sup>3</sup> )	O <sub>3</sub> 最大八 小时滑动 平均值 (μg/m <sup>3</sup> )	CO (mg/m <sup>3</sup> )	AQI	首要污染物
2021/7/13	监测点 A	5.99	3.75	22.14	6.66	105.38	0.43	55	O <sub>3</sub>
2021/7/14	监测点 A	7.69	12.99	33.54	13.82	126.74	0.47	73	O <sub>3</sub>
2021/7/15	监测点 A	6.74	11.63	27.51	9.97	121.75	0.57	69	O <sub>3</sub>



2021/7/13	监测点 B	5.77	12.19	20.62	6.52	70.63	0.47	36	O <sub>3</sub>
2021/7/14	监测点 B	5.92	10.73	20.50	6.61	68.12	0.45	35	O <sub>3</sub>
2021/7/15	监测点 B	6.19	7.79	19.78	7.23	82.48	0.51	42	O <sub>3</sub>
2021/7/13	监测点 C	7.65	17.75	25.92	11.64	116.82	0.56	65	O <sub>3</sub>
2021/7/14	监测点 C	7.46	17.39	32.51	15.60	126.15	0.58	72	O <sub>3</sub>
2021/7/15	监测点 C	8.89	41.29	38.02	19.39	112.67	0.62	61	O <sub>3</sub>

## 七、问题四的建模和求解

由于污染物浓度在三维空间中呈连续状态，相邻区域的污染物浓度往往具有一定的相关性。视 A、A1、A2、A3 四个监测站处于同一个小气象场中，随着大气的扩散，污染物浓度的变化会受到相互影响，则四个监测点附近的污染物浓度具有一定的分布函数关系。

针对物质在大气中的扩散性质，根据不同的研究对象、原理及前提假设，已形成多种形式的扩散模型。对于污染物浓度扩散的情形，应用较多的是基于湍流统计理论体系的高斯扩散模式。

### 一、高斯扩散模型：

模型适用于平稳的气象条件及地面开阔平坦的地区，有着点源的扩散模式。排放污染物的源头不论大小，都能将其视为点源。该模型可以用来观测某个监测点对其他监测点的影响，即污染物浓度的分布关系。

首先提出模型假设：

- (1) 污染物在扩散过程中不发生转化且质量守恒；
- (2) 污染物在 y 轴和 z 轴上分布呈高斯正太分布；
- (3) 污染物源强连续均匀；
- (4) 在所有空间中，风速均匀、风向平直。

本文主要探讨两种高斯扩散模式：

- (1) 在大空间点源扩散情况下，对于原点源下风向任一点，污染物浓度分布函数为：

$$C(x, y, z) = \frac{q}{2\pi u \sigma_y \sigma_z} \exp\left[-\frac{1}{2}\left(\frac{y^2}{\sigma_y^2} + \frac{z^2}{\sigma_z^2}\right)\right] \quad (14)$$

其中， $C(x, y, z)$  为空间点  $(x, y, z)$  处的污染物浓度（单位  $mg/m^3$ ）； $u$  为平均风速（单位  $m/s$ ）； $q = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} u C dy dz$  为源强，即单位时间内排放的污染物（单位  $\mu g/s$ ）； $\sigma_y$ 、 $\sigma_z$  分别为水平和垂直方向的标准差，即 y、z 方向的扩散系数。

- (2) 在高架点源扩散情况下（高斯羽烟模型），实源与像源之和（像源法）即 K 点实际污染物浓度为：

$$C(x, y, z, H) = \frac{q}{2\pi u \sigma_y \sigma_z} \exp\left(-\frac{1}{2} \frac{y^2}{\sigma_y^2}\right) \left\{ \exp\left[-\frac{(z-H)^2}{2\sigma_z^2}\right] + \exp\left[-\frac{(z+H)^2}{2\sigma_z^2}\right] \right\} \quad (15)$$

其中， $H$  为污染源的有效高度（单位  $m$ ）； $x$  为污染源排放点至下风向上任一点的距

离（单位  $m$ ）；  $y$  为排放源中心轴在直角水平方向上到任意点的距离（单位  $m$ ）；  $z$  为任意点到地面的高度（单位  $m$ ）； 其余同上。

基于 A、A1、A2、A3 四个监测点的坐标 A（0，0，0）、A1（-14.4846，-1.9699，0）、A2（-6.6716，7.5953，0）、A3（-3.3543，-5.0138，0），按照高斯羽烟模型方程，设置水平扩散系数为  $\sigma_y = ax^b$ ，垂直扩散系数为  $\sigma_z = cx^d$ 。

计算出 A 与 A1、A2、A3 的欧氏距离如下：

$$\begin{aligned} |AA_1| &= \sqrt{14.4846^2 + 1.9699^2} \approx 12.0911 \\ |AA_2| &= \sqrt{6.6716^2 + 7.5953^2} \approx 10.1093 \\ |AA_3| &= \sqrt{3.3543^2 + 5.0138^2} \approx 6.0324 \end{aligned}$$

设 A 为中心点，则其稳定度最高，扩散系数最大；A1 距离 A 最远，稳定度最低，扩散系数最小。则四个监测站分别对应下表中的水平和垂直扩散系数。

表 9  $y$ 、 $z$  方向扩散系数的系数

监测站	a	b	c	d
A	0.527	0.865	0.28	0.90
A3	0.371	0.866	0.23	0.85
A2	0.209	0.897	0.22	0.80
A1	0.123	0.905	0.20	0.76

根据由 MATLAB 仿真拟合可得以下四个扩散效果：

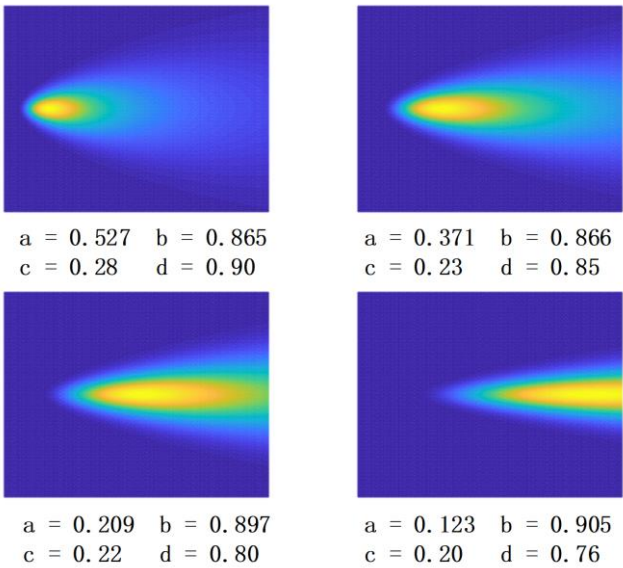


图 21 监测站 A、A1、A2、A3 的扩散模式效果俯视图

由此可见，四个监测站点污染物扩散效果明显，其扩散程度沿中心依次递增；中心点扩散浓度相对集中，A1、A2、A3 之间的污染物可能会相互扩散，且其具有相关性。

建立四个监测点，对污染物浓度进行协同预测，能够尽可能消除它们间的误差部分，从而更准确的预报协同的中心即监测站 A 的污染物浓度[8]。

数值模拟的方法被使用在很多城市气象的模拟和协同预报工具中，国内外也推出了很多模式，包括中尺度空气质量模型、综合空气质量模型、以及气象系统耦合平台等等，应用广泛。而多模式集合预报是现在较为主流的空气环境质量数值模拟的研究方向。

根据高斯扩散的污染物浓度分布函数  $C(x,y,z)$  能够确定空间中一点处的污染物浓度值（在水平和垂直的扩散系数均可算的情况下），再结合监测点 A、A1、A2、A3 逐日的一次预报数据和实测数据，这里以误差平方和为基准的基于 IOWA 算子的协同预测模型表示如下：

$$\begin{aligned} \min S(L) &= \sum_{i=1}^n \sum_{j=1}^n w_i w_j (\sum_{i=1}^n e_{a-index(it)} e_{a-index(jt)}) \\ s.t. &\begin{cases} \sum_{i=1}^n w_i = 1 \\ w_i \geq 0, i = 1, 2, \cdots, n \end{cases} \\ \bar{a}_t(T) &= \frac{1}{T} \sum_{t=1}^{T-1} a_1(N-t), T = 1, 2, \cdots, N \end{aligned}$$

通过 MATLAB 计算（具体代码见附件材料程序七），使用模型预测出污染物浓度及 AQI 如下表所示：

表 10 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							
		SO <sub>2</sub> (μg/m <sup>3</sup> )	NO <sub>2</sub> (μg/m <sup>3</sup> )	PM <sub>10</sub> (μg/m <sup>3</sup> )	PM <sub>2.5</sub> (μg/m <sup>3</sup> )	O <sub>3</sub> 最大八 小时滑动 平均 (μg/m <sup>3</sup> )	CO (mg/m <sup>3</sup> )	AQI	首要污染物
2021/7/13	监测点 A	6.21	4.22	22.24	7.06	103.98	0.42	54	O <sub>3</sub>
2021/7/14	监测点 A	8.12	13.21	33.10	13.90	124.04	0.46	71	O <sub>3</sub>
2021/7/15	监测点 A	6.95	12.86	26.99	9.91	119.92	0.56	67	O <sub>3</sub>
2021/7/13	监测点 A1	9.67	8.60	26.93	12.71	93.96	0.39	47	O <sub>3</sub>
2021/7/14	监测点 A2	15.52	18.26	34.97	19.60	126.92	0.38	73	O <sub>3</sub>
2021/7/15	监测点 A3	11.66	21.91	26.95	13.57	109.02	0.45	58	O <sub>3</sub>
2021/7/13	监测点 A1	7.08	8.76	27.89	9.95	106.94	0.43	56	O <sub>3</sub>
2021/7/14	监测点 A2	9.30	19.36	35.63	13.68	110.89	0.38	60	O <sub>3</sub>
2021/7/15	监测点 A3	6.49	21.47	30.06	9.89	119.72	0.48	67	O <sub>3</sub>
2021/7/13	监测点 A1	5.47	3.18	13.52	5.39	87.24	0.44	44	O <sub>3</sub>
2021/7/14	监测点 A2	6.84	5.85	21.21	9.76	88.33	0.59	45	O <sub>3</sub>
2021/7/15	监测点 A3	6.43	16.15	15.25	5.24	100.05	0.52	51	O <sub>3</sub>

理论上而言，协同预报模型相较于二次预测模型，能够提升监测点 A 的污染物浓度预报准确率。因为前者模型能够根据某监测站周围的污染物浓度情况，对其预测过程中的误差做出一定矫正。

## 八、误差分析和灵敏度分析

### 8.1 误差分析

(1) 由于表格中的数据存在显性错误以及专业相关的隐性错误，处理数据时难免存在人为操作的误差以及数据本身造成的误差；

(2) 机器仿真学习模型中分别考虑偏差和方差，却很少谈及偏差和方差的权衡问题，参数的设置也难免会有缺陷之处。

(3) 原始实测的数据都是取整化的，这会对后来预测计算造成一定的计算干扰，或多或少也会影响到计算结果。

### 8.2 灵敏度分析

在问题三的求解过程中，利用 BP 神经网络进行构建模型的过程中，我们往往会设置误差阈值，来使隐含层的方程达到最优。对此过程，我们可以选择误差精度的不同取值，来对六种污染物浓度进行求解，然后运用 MATLAB 软件来进行灵敏度分析。

根据 MATLAB 软件简单作图，由灵敏度分析可知，可得到误差精度的参数值的分析结果。在参数可变化范围内，污染物浓度的变化区间也存在一定的规律， $\alpha$  越大，污染物的浓度波动的范围没有很大的变化趋势，上下限基本在 10% 左右，说明模型可靠稳定性较好，造成的误差较小，也从另一方面说明神经网络模型具有良好的鲁棒性。

表 11 不同  $\alpha$  值对应的各污染物的值

$\alpha$	SO <sub>2</sub>	NO <sub>2</sub>	PM <sub>10</sub>	PM <sub>2.5</sub>	O <sub>3</sub>	CO
0.01	6.28	12.92	16.90	4.68	90.08	0.4
0.02	5.82	11.48	17.1	4.04	93.22	0.38
0.03	5.96	12.17	16.9	4.46	92.70	0.40
0.04	5.78	10.04	17.2	3.375	92.56	0.36
0.05	5.92	10.38	19.54	3.33	88.46	0.33

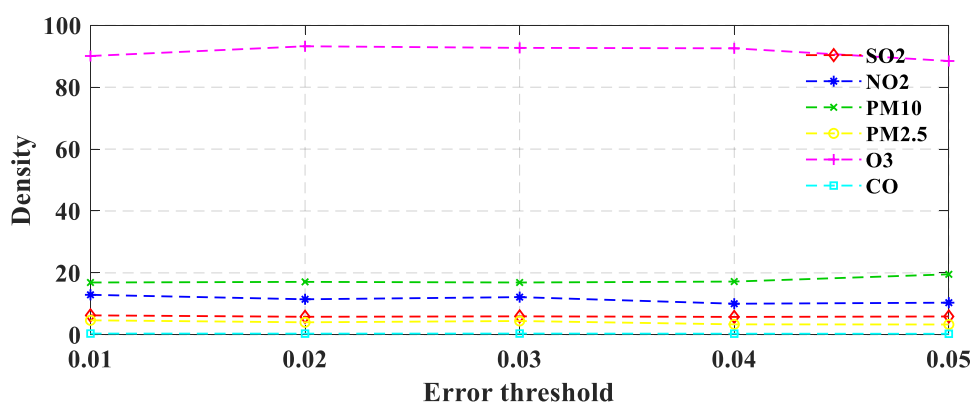


图 22 灵敏度分析

## 九、模型的评价与推广

### 9.1 模型的评价

#### 9.1.1 模型的优点

(1) K-means 算法可解释性较强, 实现相对容易, 随机选取  $K$  个对象作为初始的  $k$  个质心, 运算相对高效, 结果适用性好。

(2) 神经网络具备较强非线性映射能力和自学能力, 并且其容错能力较好: BP 神经网络在其局部的或者部分的神经元受到破坏后对全局的训练结果不会造成很大的影响。

(3) 高斯扩散模型可以用来某个监测点对其他监测点的影响, 同时在具备水平垂直扩散系数的情况下, 能求得空间任意点  $(x, y, z)$  的污染物浓度。模型通用性广、可解释性强、大气扩散情况下结果准确。

#### 9.1.2 模型的缺点

(1) 为使得模型计算更加方便, 会采用一些假设使得模型构造更加理想化, 因此会忽略了一些误差影响因素, 导致模型与实际情况出现一定的偏差。

(2) 在构造二次预测模型时, 考虑因素、带入样本数据较少, 在处理问题时可能存在一些误差, 另外求解时可能带有一定的局限性。

(3) K-means 算法容易导致局部最优解, 它对噪音和异常值比较敏感。


### 9.2 模型的推广

(1) 神经网络是一项机器学习技术, 它不能代替已有的优化技术, 它有自己的适用范围, 可推广到以下领域进行使用: 人工智能, 复杂问题优化, 结构优化设计。BP 神经网络本文中主要运用于仿真预测系统, 具有一定的实践意义, 除了应用于本文污染物浓度预测外, 我们可以对模型采用反向传播算法, 逆向推导出影响的气象因子。

(2) 基于 BP 神经网络的二次预测模型可用于机器学习和人工神经网络的权重系数调整与网络构造问题。

(3) 传统的组合预测模型在各个预测点精度忽高忽低, 稳定性较低, 基于 IOWA 算子的协同预测模型能很好地克服这个问题, 它也可以推广用于固定资产投资组合, 是很多决策者使用的选方案的方法。

## 十、参考文献

- [1] 张潇文, 曾胜兰, 2014~2016 年成都地区空气污染气象要素分型研究, 四川环境, 40(01): 53-58, 2021。
- [2] 田宏伟, 谈建国, 杜子璇, 用 TSI 天气分型方法分析上海环境空气质量, 气象与环境科学, (01): 51-55, 2008。
- [3] 司守奎, 孙兆亮, 数学建模算法与应用, 北京: 国防工业出版社: 387-390, 2009。
- [4] 徐爱兰, 朱晏民, 孙强, 於香湘, 彭小燕, 基于 K-means 划分区域的深度学习空气质量预报, 南通大学学报(自然科学版), 20(03): 49-56。
- [5] 张浩, 合肥市大气能见度与相对湿度、PM<sub>10</sub> 及 PM<sub>2.5</sub> 的关系, 安徽: 安徽省气象科学研究所, 2018-10-19。
- [6] 张宸赫, 赵天良, 陆忠艳, 王东东, 陈煜升, 杨瑞雯, 王富, 沈阳大气污染物浓度变化及气象因素影响分析, 环境科学与技术, 43(S2): 39-46, 2020。
- [7] 王自发, 王威, 区域大气污染预报预警和协同控制, 科学与社会, 4(02): 31-41。
- [8] 张耀, 王湛, NU YU, 修正高斯扩散模型对机场污染物浓度预测影响, 环境保护科学, 47(03): 106-112, 2021。
- [9] 胡玉筱, 段显明, 基于高斯烟羽和多元线性回归模型的 PM<sub>2.5</sub> 扩散和预测研究, 干旱区资源与环境, 29(06): 86-92, 2015。
- [10] 王茜, 吴剑斌, 林燕芬, CMAQ 模式及其修正技术在上海市 PM<sub>2.5</sub> 预报中的应用检验, 环境科学学报, 35(06): 1651-1656, 2015。
- [11] jadefan, 高斯扩散模型 - 高斯烟羽大气污染扩散模型,  <https://www.jianshu.com/p/5cf580af2def>, 2021-10-17。

## 附录

### 附录一

表 A-1 5 个气象指标的 KMO 和 Bartlett 检验

春季	取 样 足 够 度 的 Kaiser-Meyer-Olkin 度量。		.383	夏季	取 样 足 够 度 的 Kaiser-Meyer-Olkin 度量。		.601
	Bartlett 的球 形度检验	近似 卡方	8147.923		Bartlett 的球 形度检验	近似 卡方	12203.349
		df	10			df	10
		Sig.	.000			Sig.	.000
秋季	取 样 足 够 度 的 Kaiser-Meyer-Olkin 度量。		.435	冬季	取 样 足 够 度 的 Kaiser-Meyer-Olkin 度量。		.591
	Bartlett 的球 形度检验	近似 卡方	5447.666		Bartlett 的球 形度检验	近似 卡方	3888.870
		df	10			df	10
		Sig.	.000			Sig.	.000

### 附录二

表 A-2 春季聚类结果

指标	聚类（春）					
	1	2	3	4	5	6
温度	28.30	23.40	30.90	26.40	20.70	27.70
湿度	59.00	24.00	57.00	95.00	69.00	78.00
气压	1008.60	1015.20	1007.60	1006.30	1014.20	1005.50
风向	283.50	358.00	125.30	238.40	3.10	98.90
风速	.80	2.50	1.30	.50	.30	1.30
AQI	191.00	67.00	183.00	35.00	115.00	13.00

表 A-3 夏季聚类结果

指标	聚类（夏）				
	1	2	3	4	5
温度	36.90	28.80	33.80	30.80	31.10
湿度	51.00	81.00	56.00	75.00	70.00
气压	1000.10	1009.40	999.50	1005.70	1009.60
风向	285.90	.80	87.20	179.40	359.80
风速	.90	.90	1.40	1.20	.80
AQI	196.00	24.00	192.00	27.00	11.00

表 A-4 秋季聚类结果

指标	聚类（秋）				
	1	2	3	4	5
温度	34.00	26.50	28.20	33.70	22.50
湿度	49.00	52.00	86.00	31.00	68.00
气压	1007.60	1013.70	1008.20	1007.40	1014.30
风向	7.00	185.20	52.90	354.50	355.90
风速	.60	1.00	2.30	1.50	2.20
AQI	198.00	156.00	23.00	194.00	22.00

表 A-5 冬季聚类结果

指标	聚类（冬）					
	1	2	3	4	5	6
温度	13.50	19.20	12.70	9.90	12.80	26.90
湿度	88.00	82.00	56.00	24.00	60.00	22.00
气压	1020.20	1011.80	1015.50	1023.40	1015.10	1013.50
风向	2.70	235.20	351.10	354.40	.40	109.00
风速	1.20	.50	.30	2.30	.30	1.30
AQI	15.00	89.00	186.00	34.00	165.00	87.00

## 附录三

表 A-6 6 个线性回归方程的 13 个系数

大气污染物	b0	b1	b2	b3	b4	b5
So2	88.73171	0.623967	-0.1233	-362.249	0.114917	-0.40281
No2	4105.664	14.56016	-14.3833	-2881.75	0.641565	-5.98149
PM10	-2842.09	-3.222	8.617121	-4094.25	1.658393	5.796078
PM2.5	-1432.34	3.838076	3.552309	-6561.21	2.198238	1.973352
O3	-3543.51	9.749275	9.608205	-19411.1	4.882901	-2.02993
CO	15.42503	-0.0878	-0.05685	71.60918	-0.02437	-0.06952
大气污染物	b6	b7	b8	b9	b10	b11
So2	-0.00942	-0.00279	-0.81479	0.049902	-0.53307	0.042628
No2	0.058489	-0.02312	-3.39347	0.287272	-0.78216	0.402158
PM10	0.109995	0.000383	3.844562	-0.23189	-0.19319	-0.29543
PM2.5	0.048848	-0.0007	2.204087	-0.07744	-0.04243	-0.05903
O3	0.003006	0.02755	5.067889	-0.2262	1.172975	-0.26231
CO	0.000368	4.17E-05	0.028068	0.001677	0.005005	0.004991