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

学　　校	西安科技大学
参赛队号	**21107040042**
队员姓名	**1.李金龙**
	2.易小虎
	3.丁楠

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：

本文旨在通过对 WRF-CMAQ 模型的一次预报数据和监测点的实测数据进行二次建模，得到较为精准的二次建模模型。在构建二次预测模型过程中，首先进行了 AQI 的计算和对气象条件与各污染气体的相关关系，其次进行了两种二次建模方式的尝试，一是通过对多个远距离监测点的一次预报数据和实测数据进行二次建模；二是通过对近距离（天气具有相关性）多个监测点的一次预报数据和实测数据进行二次建模。最后比较两种方法的优劣性。

针对问题一，在所有公式都已经给定的情况下，计算得到了监测点 A 从 2020 年 8 月 25 日至 8 月 28 日每天实测的 AQI 和首要污染物。计算发现，首要污染物都是 O_3。

针对问题二，首先对给出的数据进行预处理，在处理过程中发现，实测数据中存在较多数据空缺和时间的空缺，此外还有异常值的处理，主要是存在气体浓度为负数和某一时刻数据过大的情况。对数据预处理后，通过对数据进行分析，发现单一气象条件和单一污染气体浓度没有较强的相关性，随后通过多元线性回归对数据进行拟合，通过 F 检验和 P 值检验，验证了模型的正确性，并将多元线性回归后得到的各变量的显著性参数作为新的变量，进行系统聚类。研究发现了各气象条件的变化对某一污染气体浓度下降起促进还是抑制作用，并将气象条件分为了两类，一类为风速，其对六种污染气体都有显著作用；其他气象条件（温度、湿度、气压和风向）为另一类，此类气象条件与风速相比，影响较不明显。

针对问题三，对数据进行预处理后，本文对监测点 A、B、C 污染气体的一次预报模型和实测模型数据进行比对，分析发现，难以直接使用单个污染气体的一次预报数据和实测数据进行二次建模。由于监测点 A、B、C 距离很远，可以直接将三个监测点的数据放到同一样本集中进行分析，根据三个监测点的数据，提出了向后逐步回归模型和 BP 神经网络模型两种二次建模方法，使用 2020 年 7 月 23 至 2021 年 7 月 2 日数据作为训练集进行拟合，并对监测点 A、B、C 的 2021 年 7 月 8 日至 10 日的数据作为验证集。对比发现，BP 神经网络的回归值 R 和预测效果都明显优于多元线性回归，监测点 A 的 AQI 二次预测平均正确率为 83.88%，因此，本文使用 BP 神经网络对 2021 年 7 月 13 日的预报结果进行二次建模，并计算出了各污染气体浓度预测结果和 AQI 值。

针对问题四，与问题三相比，由于四个检测点距离较近，因此需要考虑临近点污染气体和监测点污染气体的相关性，不能简单地将监测点 A、A1、A2、A3 放到同一样本集中

进行预测，因此本文使用一种地质学中较为经典的 Kriging 方法来讨论四个监测点污染物浓度的相关性。在使用 Kriging 方法对某一监测点进行插值预测后，将插值预测值与一次预报值进行灰色关联度分析，求解污染气体插值预测数据和一次预报数据对污染气体实测数据的权重。计算发现，插值预测值和一次预报值的权重基本相同，在 0.6-0.81 之间。研究证明，区域协同预报可以提升空气质量预报的准确度。最后，将插值预测结果与一次预报结果数据通过权重相加，使用 BP 神经网络进行二次建模，并以 2021 年 7 月 7 日至 10 日作为验证集验证，结果表明该模型预测效果好，平均正确率为 88.05%。通过该二次预测模型计算出了 2021 年 7 月 13 日各污染气体浓度二次预测结果和 AQI 值。问题三、四验证结果对比发现，与问题三相比，区域协同预测更能提升预报准确度。

关键词：多元线性回归 系统聚类 灰色关联度分析 Kriging 插值 BP 神经网络

目录

1 问题重述

1.1 研究背景

大气污染是由于自然发展及人类活动在大气中产生的某些物质，经过一定时间、累计到一定浓度后所产生的环境现象。常见的大气污染物有小颗粒状污染物如：烟、雾烟、粉尘等，也有气态污染物，包括 CO_2、CO、和 SO_2 等。目前已知的大气污染物约有一百多种，不仅对生态环境造成巨大的破坏，对人类正常生存及发展也产生极大威胁。建立准确可靠的空气质量预测模型、合理选择预防控制措施是预防大气污染、改善环境质量的一种有效方法。

空气质量预测目前多采用 WRF-CMAQ 模型。该模型由 WRF 和 CMAQ 两部分组成。首先采用 WRF 预报气象数据，CMAQ 基于 WRF 提供的数据，模拟污染物理化学过程来得到具体时间的预报结果。但由于污染物生成过程及排放物的不确定性，WRF-CMAQ 一次预报模型的结果并不完全准确[1-5]。本文采用二次建模概念来提高预报的准确性。在 WRF-CMAQ 模型一次预报的基础上，采用实测数据对其进行修正，通过建立二次模型来提高预报的准确度。

1.2 待解决问题

问题 1. 通过附件 1 已知数据及 AQI 计算方法，计算监测点 A 从 2020 年 8 月 25-28 日每天实测的 AQI 和首要污染物。

问题 2. 气象条件会对污染物浓度产生影响。首先判断气象条件对各污染物浓度的影响程度，然后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，忽略监测点相互影响，建立一个同时适用于 A、B、C 三个监测的二次预报模型。对 2021 年 7 月 13 日至 7 月 15 日 6 种常规污染物的单日浓度值进行预测。其中要求 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。

问题 4. 在问题 3 的基础上，考虑相邻监测点的影响，建立二次预报模型。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。与问题 3 结果进行比对，分析考虑相邻监测点的协同预报模型是否能够提升该监测点预报准确度。

2 问题假设

为简化问题，做出如下合理假设：

（1）同一时间，自建点与国控点所处的客观环境是一致的，即污染物浓度和天气因素的实际值是完全相同的。

（2）不考虑监测数据的存储、传输、分发等环节的随机噪声所导致的数据误差。

3 问题一

3.1 问题分析

问题一可拆分为两个子问题：（监测点 A 于 2020 年 8 月 25-28 日）实测 AQI 计算、首要污染物计算。根据附件 1 中监测点 A 逐日污染物浓度实测数据，与附录中 AQI 及首要污染物的计算公式，基于 matlab 编制相关程序即可对本问题进行求解。

3.2 模型建立与求解

本题考虑 SO_2、NO_2、PM10、PM2.5、O_3 和 CO 六种大气污染物来衡量空气质量，AQI 计算公式如下：

$$\mathrm{AQI} = \max\left\{\mathrm{IAQI}_{\mathrm{SO_2}}, \mathrm{IAQI}_{\mathrm{NO_2}}, \mathrm{IAQI}_{\mathrm{PM_{10}}}, \mathrm{IAQI}_{\mathrm{PM_{2.5}}}, \mathrm{IAQI}_{\mathrm{O_3}}, \mathrm{IAQI}_{\mathrm{CO}}\right\} \quad (3\text{-}1)$$

其中，IAQI 为各污染物的空气质量分指数，计算如下：

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

其中，$\mathrm{IAQI_P}$为污染物 P 的空气质量分指数；$\mathrm{C_P}$为污染物 P 的质量浓度值；$\mathrm{BP_{Hi}}/\mathrm{BP_{Lo}}$为与$\mathrm{C_P}$相近的污染物浓度限值的高位值与低位值；$\mathrm{IAQI_{Hi}}/\mathrm{IAQI_{Lo}}$为与 $\mathrm{BP_{Hi}}/\mathrm{BP_{Lo}}$对应的空气质量分指数。各污染物浓度限值可参考附录表一。基于式（1）、（2）进行计算，即可得到 AQI 值。

当 AQI 小于或等于 50 时，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。表 3-1 给出了 AQI 计算结果及首要污染物。

表 3-1 AQI 计算结果

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	无
2020/8/27	监测点 A	109	O_3
2020/8/28	监测点 A	138	O_3

3.3 结果分析

由表 1 可知，监测点 A 在 25、27 和 28 日 AQI 均大于 50，首要污染物为 O_3；26 日 AQI 小于 50，故当日无首要污染物。

4 问题二

4.1 问题分析

问题二需要在假定污染物排放量一定的情况下研究气象条件对污染物浓度的影响程度，因此需要找到气象条件与各污染物的关系。在对数据进行预处理之后，分别对气象条件与污染物进行相关性分析，判断各气象条件与各污染物浓度之间是否存在单纯的线性关系，若无直接线性关系，需要对数据之间的关系进行进一步探讨。随后建立气象条件与污染物浓度的多元线性回归模型，识别重要变量，计算其回归系数拟合值，并进行显著性检验。如果其显著性概率 0<P<0.05，则认为在 95%的置信区间内认为回归效果显著。如果 P>0.5，则认为其拟合系数为 0，变量对浓度变化不显著将其剔除再进行回归。最终得到回归系数估计值、置信区间、检验统计量。最后对回归系数拟合值进行聚类分析来得到气象条件与污染物浓度之间的关系，从而达到对气象条件分类的目的。

该问题的求解思路如下图所示：

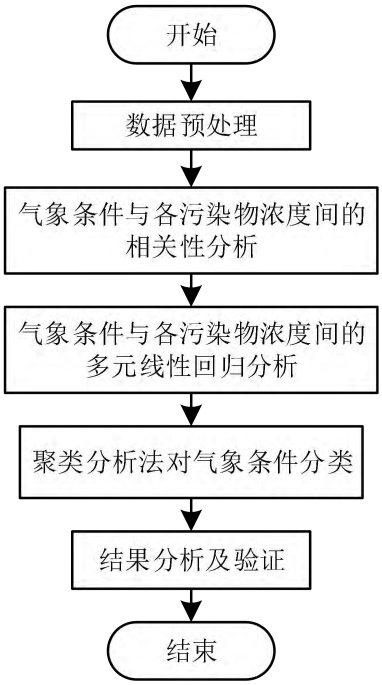


图 4-1 问题 2 求解流程图

4.2 模型的建立

4.2.1 相关性分析模型

相关系数法[6]假设有随机变量X和 Y，对其进行 n 次随机试验，得到其观测值$(\mathrm{X_i},\mathrm{Y_i})(i=1,2,\cdots n)$，$\overline{X}$、$\overline{Y}$为随机变量的期望值。$\sqrt{DX}$，$\sqrt{DY}$为$X$、$Y$ 的方差，$\mathrm{Cov}\,(X,Y)$ 为X、Y 的协方差，r 为相关系数，R 为X和 Y 据对于样本$(\mathrm{X_i},\mathrm{Y_i})(i=1,2,\cdots n)$的相关性系数，也被称为样本相关系数。在实际分析过程中，常采用 R 作为 r 的估计值。

计算公式如下：

$$\begin{cases} r = \frac{\mathrm{Cov}\,(X,Y)}{\sqrt{DX}\sqrt{DY}} \\ R = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \\ \mathrm{Cov}\,(X,Y) = E[(X-\bar{X})(Y-\bar{Y})] \\ \bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i \\ \bar{Y} = \frac{1}{n}\sum_{i=1}^{n}Y_i \end{cases} \tag{4-1}$$

4.2.2 多元线性回归分析模型

多元线性回归的一般性模型的表达式为[6]：

$$Y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p + \varepsilon\sigma^2 \tag{4-2}$$

其中，自变量 $x_1, x_2, \cdots, x_p$ 为预测因子，Y为因变量，ε 是服从正态分布 $N(0,\sigma^2)$ 的随机变量， $b_0, b_1, \cdots, b_p$ 为待估计参数，该参数的计算采用最小二乘估计法，为了方便，引入矩阵记号：

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} x_1(u_1) & x_2(u_2) & \cdots & x_m(u_m) \\ x_2(u_1) & x_2(u_2) & \cdots & x_2(u_m) \\ & & & \cdots \\ x_n(u_1) & x_n(u_2) & \cdots & x_n(u_m) \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_m \end{bmatrix} \tag{4-3}$$

选取 β 的一个估计值 $\hat{\beta}$，使得随机误差 ε 的平方和达到最小，即

$$\mathrm{m}\ \ \boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} = \mathrm{m}\ \ (\boldsymbol{Y}-X\boldsymbol{\beta})^T(\boldsymbol{Y}-X\boldsymbol{\beta}) \tag{4-4}$$

写成分量形式为：

$$Q(\beta_1, \beta_3, ..., \beta_m) = \sum_{i=1}^{n}\ [y_i - \beta_1x_1(u_i) - \beta_2x_2(u_i) - \cdots - \beta_mx_m(u_i)]^2 \tag{4-5}$$

则

$$Q(\hat{\beta}_1, \hat{\beta}_2 ... \hat{\beta}_m) = \mathrm{m}\ \ Q(\beta_1, \beta_2, ..., \beta_m) \tag{4-6}$$

注意到 $Q(\beta_1, \beta_3, ..., \beta_m)$ 是非负二次式，由多元函数取得极值的必要条件，可得 $\frac{\partial Q}{\partial \beta_j} = 0(j = 1,2, ..., m)$，即

$$\sum_{i=1}^{n}\ \left[y_i - \hat{\beta}_1x_1(u_i) - \hat{\beta}_2x_2(u_i) - \cdots - \hat{\beta}_mx_m(u_i)\right]x_j(u_i) = 0\ j = 1,2, ...m \tag{4-7}$$

$$\begin{cases} \left[\sum_{i=1}^{n}x_1^2(u_i)\right]\hat{\beta}_1 + \left[\sum_{i=1}^{n}x_1(u_i)x_2(u_i)\right]\hat{\beta}_2 + \cdots + \left[\sum_{i=1}^{n}x_1(u_i)x_m(u_i)\right]\hat{\beta}_m \\ = \sum_{i=1}^{n}x_1(u_i)y_i \\ \left[\sum_{i=1}^{n}x_1(u_i)x_m(u_i)\right]\hat{\beta}_1 + \left[\sum_{i=1}^{n}x_2(u_i)x_m(u_i)\right]\hat{\beta}_2 + \cdots + \left[\sum_{i=1}^{n}x_m^2(u_i)\right]\hat{\beta}_m \\ = \sum_{i=1}^{n}x_m(u_i)y_i \end{cases} \tag{4-8}$$

将 $\hat{\beta}$ 带入模型（2-3）中，得模型的估计：$\hat{Y} = X^T \hat{\beta}$ 。

一般线性回归方程利用模型拟合程度R^2、F 检验来对结果进行分析验证。

（1）预测值与观测值的相关系数 R，表示模型对观测数据的解释程度，一般其值大于 0.85 认为拟合成功，其表达式如下：

$$R = \frac{\sum (y_i - Y)(Y_i - Y)}{\sqrt{\sum (y_i - Y)^2 (Y_i - y)^2}} \quad (4\text{-}9)$$

（2）F 检验可检验线性回归的整体显著性，即检验假设 H_0：$b_0 = b_1 = ... = 0$ 是否为真，由于：

$$F = \frac{\sum (Y_i - Y)^2 / P}{\sum (y_i - Y_i)^2 / (n - P - 1)} \sim F(n, n - P - 1) \quad (4\text{-}10)$$

若$P\{F \geq F(p, n - p - 1)\} \leq a$, 其中$a$为显著性水平（一般选择 0.05），则拒绝联合显著性假设 H_0，即认为 $b_0 \sim b_p$不全为 0，回归方程有效。

4.2.3 聚类分析模型

系统聚类算法是通过计算不同属性数据点之间的距离，将最为接近的几类数据进行组合，进行反复迭代到将所有数据合成一类，最终生成聚类谱系图。

聚类分析是一种简化数据的方法，具有简单、直观的特点。其主要应用于探索性的研究，通过分析结果提供多个可能的解，最终研究者通过主观判断和分析选择最终的答案。传统的统计聚类分析方法包括系统聚类法、分解法、加入法、动态聚类法和模糊聚类等。许多著名的统计分析软件包如 SPSS、SAS 已经加入了 k-均值、k-中心点等聚类分析算法。

聚类分析法对样本进行聚类分析，就是对数据集中的数据应用某种方法进行分组，将样本按相似程度或距离远近来划分类别，使得同一类中的样本之间的相似性比其他类的样本的相似性更强。由于聚类数的设定对聚类结果会有一定的影响，我们首先通过系统聚类，确定分类数。在进行聚类时, 每个个体自成一类, 然后根据距离最小原则，将最近的两类合成一类，并计算当前类与新类的距离，直至仅有一类。在分析中，本文选用欧氏距离法衡量样本之间的距离，即：

$$\rho = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (4\text{-}11)$$

$$|X| = \sqrt{x_2^2 + y_2^2} \quad (4\text{-}12)$$

其中，ρ 为点 (x_2, y_2) 与点 (x_1, y_1) 之间的欧氏距离，$|X|$ 为点 (x_2, y_2) 到原点的欧氏距离。

4.3 模型的求解

4.3.1 数据预处理

附件 1 中给出了监测点 A 逐日气象条件及污染物的实测数据。对数据进行初步分析可知，所提供的数据主要存在三类问题（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；（3）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。对这三种情况进行分析后，对数据的预处理方式如下。

1）缺失数据数目

本文中数据缺失数据主要分为两种，一为已知日期，数据缺失；二为日期缺失。对监测点 A 实测数据中气象条件及各污染物浓度的数据缺失数目如图 4-2 所示。

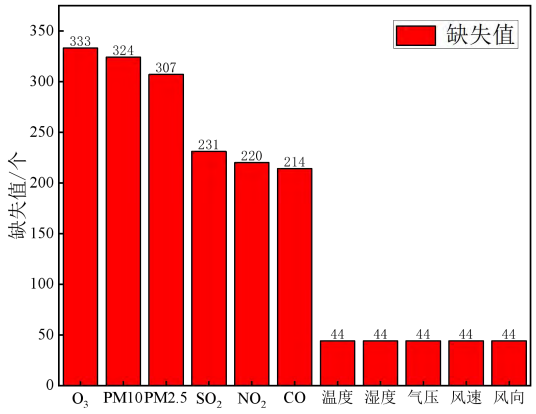


图 4-2 数据缺失示意图

数据缺失可分为三大类：完全随机缺失（MCAR）、随机缺失（MAR）和完全非随机缺失（MNAR）。在 MCAR 假设中，数据缺失不依赖任何变量；MAR 中，数据缺失依赖于完全变量;MNAR 依赖于其自身不完全变量。经分析本文数据缺失属于随机缺失（MAR），采用均值插补法进行数据的填充。采用存在值相邻数据的平均值来插补缺失值。为了使数据具有更好的连续性，一般先取缺失值相邻的四个未缺失值数据的平均值作来替代缺失值，如果仍有数据缺失，则增大求平均值的相邻数据的个数，知道缺失值完全消失。

2）异常数据修正

本文实测数据中数据存在负值、0 值和数据偏离值，在实际大气模型中，气体的浓度不可能为负值或者零值，对出现负值的情况，将此类数据作为缺失值进行处理。当污染气体浓度为零值时，加一个无穷小数进行修正。

除数据可能为零值或负数外，实测数据中还有一些数据出现严重偏离，若不对此类数据进行处理，在下一步进行归一化或标准化时，数据会大量集中在前半部分，数据分配极不合理。为了保证数据的完整性和合理性，本文先将异常数据剔除后再对其进行填充。拉依达准则（3σ准则）是剔除异常数据的一种常用方法，其适用于测量次数充分大的情况，与本题相适应，因此本文采用 3σ准则对异常数据进行剔除，再采用均值插补法对其进行填充，从而完成对异常数据的修正。

3σ准则定义如下：

设测量值为 $x_1, x_2 \cdots, x_n$ ，算出其算术平均值 x 及剩余误差$v_i = x_i - x\ (i = 1,2, \cdots, n)$，依据贝塞尔公式算出标准偏差$\sigma$，当某数据$x_b$的剩余误差 $v_b(1 <= \mathrm{b} <= \mathrm{n})$，满足下式

$$|v_b| = |x_b - x| > 3\sigma \qquad (4\text{-}13)$$

则认为x_b应予剔除。

3）数据归一化、标准化

分析附件 1 污染物浓度及气象数据可以发现，不同变量之间量级存在较大差异。因此，本文采用归一化和标准化两种方法对数据进行处理，消除不同变量之间的量纲差异。

数据归一化：

$$x_k = \frac{x_k - x_{min}}{x_{max} - x_{min}} \tag{4-14}$$

数据标准化：

$$x_k = \frac{x_k - x_{ave}}{\sigma} \tag{4-15}$$

4）AQI 首要污染物占比分析

在对 AQI 进行计算过程中，由于 AQI 的值为六种污染气体 IAQI 的最大值，因此可以对 AQI 的首要污染物占比进行分析，各污染性气体的占比如图 4-3 所示。

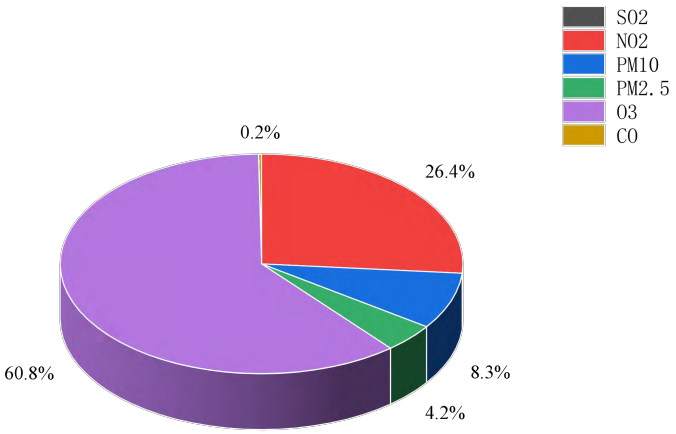


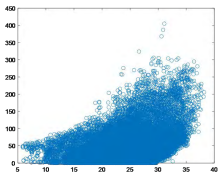
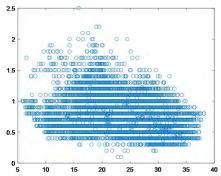
图 4-3 首要污染物占比

由图 4-3 可知。大气中的首要污染物主要为 O_3，其占比高达 60.8%。排名前三位污染物分别为 O_3、NO_2、PM10。因此在后续的问题分析过程中，O_3 的拟合正确率尤为重要。

4.3.2 相关性分析

基于上述理论，分别将温度、湿度、气压、风速和风向 5 种气象条件设为随机变量 X，将 SO_2、NO_2、PM10、PM2.5、O_3 和 CO 设为随机变量 Y，利用 matlab 编程来计算气象条件与污染物浓度之间相关关系，5 种气象条件与各污染物浓度及 AQI 指标的散点图见附录 A。选取 5 种气象条件与 O_3 及 CO 监测浓度散点图如表 4-1 所示。

表 4-1 气象条件与污染物浓度散点图

气象条件	污染物浓度	
	O_3	CO
温度		

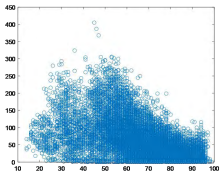
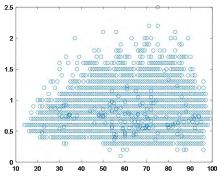
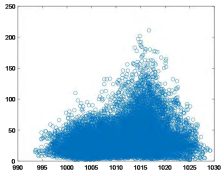
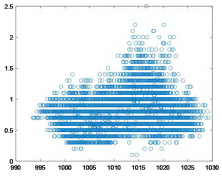
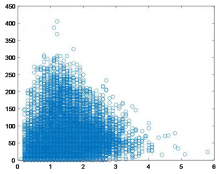
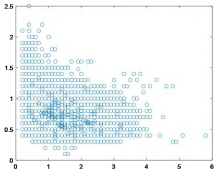
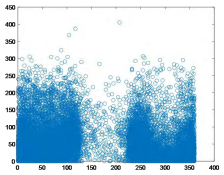
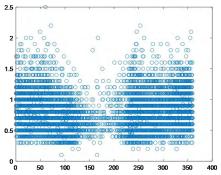
湿度		
气压		
风速		
风向		

表 4-1 给出了与各类气象条件相关性最强的 O_3 浓度与各类气象条件的散点图和气象条件相关性最弱的 CO 浓度与各类气象条件的散点图。图中发现 O_3 浓度与温度和气压成正相关性，与湿度和风速呈负相关。但相关性不强。其他污染物气体与气象条件散点图同 CO 散点图大致一致，因此得出气象条件与各污染物浓度不存在的单一的线性关系。基于此，下文采用多元线性回归法再进行研究。

在搭建多元线性回归模型中，利用 SPSS 软件求解多元回归分析模型的拟合系数，进而对气象条件与各污染物浓度之间的关系进行研究。首先以不同气象条件为预测变量，各种污染物浓度为响应变量建立多元线性线性回归方程。在 F 检验的成功的基础上基于最小二乘法得到回归系数。

表 4-2、4-3、4-4 分别为 F 检验结果（P>|t|）、显著性拟合结果（Beta）、回归系数拟合结果（coef）。

表 4-2 F 检验结果

气象条件	P>|t|					
	SO_2	NO_2	PM10	PM2.5	O_3	CO
温度	0.000	0.000	0.000	0.424	0.000	0.000
湿度	0.000	0.000	0.000	0.000	0.000	0.000
气压	0.000	0.000	0.000	0.000	0.000	0.001
风速	0.000	0.000	0.000	0.000	0.006	0.000
风向	0.478	0.000	0.103	0.000	0.000	0.000

表 4-3 显著性拟合结果

气象条件	Beta					
	SO_2	NO_2	PM10	PM2.5	O_3	CO

温度	0.062	-0.339	0.077	-0.009	0.519	-0.305
湿度	-0.449	-0.152	-0.415	-0.302	-0.557	-0.094
气压	0.159	-0.055	0.259	0.255	0.141	0.043
风速	-0.217	-0.480	-0.343	-0.390	0.015	-0.323
风向	-0.004	-0.062	0.010	0.021	-0.073	0.067

表 4-4 不同气象条件与各污染物浓度间回归系数拟合结果

气象条件	SO2	NO2	PM10	PM2.5	O3	CO
温度	0.040067	-1.413208	0.3680358	0	4.606274	-0.0126675
湿度	-0.102737	-0.225195	-0.7033594	-0.3434324	-1.753809	-0.0013874
气压	0.09043	-0.204475	1.09563	0.72156	1.10786	0.00157
风速	-1.192491	-17.09738	-14.00453	-10.64181	1.126921	-0.1143232
风向	0	-0.01263	0	0.00329	-0.03176	0.0001364

由表 4-2 F 检验结果可知，除了风向与 SO_2、PM10，温度与 PM2.5 之外，其余气象条件与各污染物浓度的 P 值均小于 0.05，可以认为在 95%的置信区间下拒绝联合显著性检验，各拟合系数显著不等于零，回归效果显著。分析表 4-2 可得到以下结论：风向对 SO_2、PM10 浓度影响较弱，温度对 PM2.5 浓度影响较弱。

在 F 检验成功的基础上得到表 4-3 标准回归系数拟合结果。当 Beta 绝对值越大时变量间相互影响越强。分析表 4-3 可知：

（1）各气象条件对 SO2 的影响程度排序为：湿度>风速>气压>温度>风向

（2）各气象条件对 NO2 的影响程度排序为：风速>温度>湿度>气压>风向

（3）各气象条件对 PM10 的影响程度排序为：湿度>风速>气压>温度>风向

（4）各气象条件对 PM2.5 的影响程度排序为：风速>湿度>气压>温度>风向

（5）各气象条件对 O3 的影响程度排序为：湿度>温度>气压>风向>风速

（6）各气象条件对 CO 的影响程度排序为：风速>温度>湿度>风向>气压

分析表 4-4 回归系数拟合结果可知：

温度与 O_3、NO_2、PM10 相关性较强，与 SO_2、CO 相关性较弱。在相关性较强的因素中与 O_3、PM10 呈正相关关系，与 NO_2 与呈负相关关系。

湿度与 O_3、PM10、PM2.5 相关性较强，与 NO_2、SO_2、CO 相关性较弱。湿度与 6 种污染物均呈负相关关系，即湿度升高会引起各污染物浓度降低。

气压与 O_3、PM10、PM2.5 相关性较强，与 NO_2、SO_2、CO 相关性较弱。气压与与其相关性较强的污染物浓度均呈正相关关系。

风速与 SO_2、O_3、NO_2、PM10、PM2.5 相关性较强，与 CO 相关性较弱。在相关性强

的因素中，除与 O_3 与呈正相关关系外，风速与其他污染物均呈负相关关系。

风向与各污染物浓度未呈现明显的相关性。

4.3.3 聚类分析分类

系统聚类算法是通过计算不同属性数据点之间的距离，将最为接近的几类数据进行组合，进行反复迭代到将所有数据合成一类，最终生成聚类谱系图。

本文根据上文计算出的回归系数的拟合结果，将每一种气象条件看作一个类别，采用系统聚类方法对气象条件进行分类。算法流程图如图 4-4 所示。

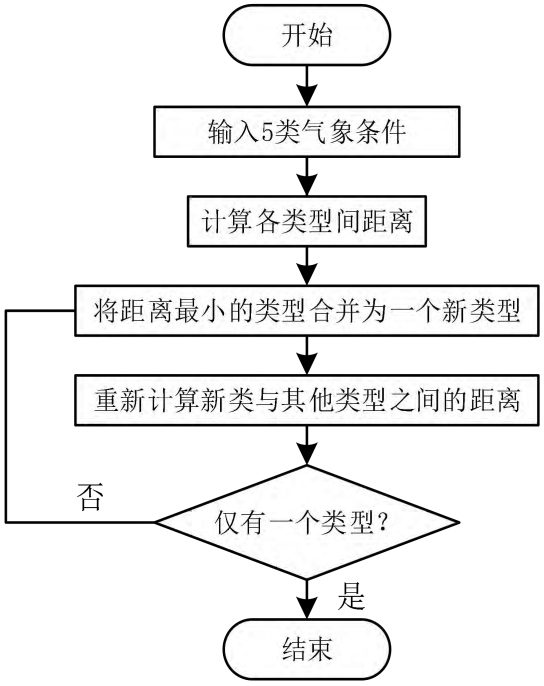


图 4-4 系统聚类算法流程图

根据上述流程图，基于 SPSS 软件对气象条件进行聚类分析，图 4-5 给出了计算结果。

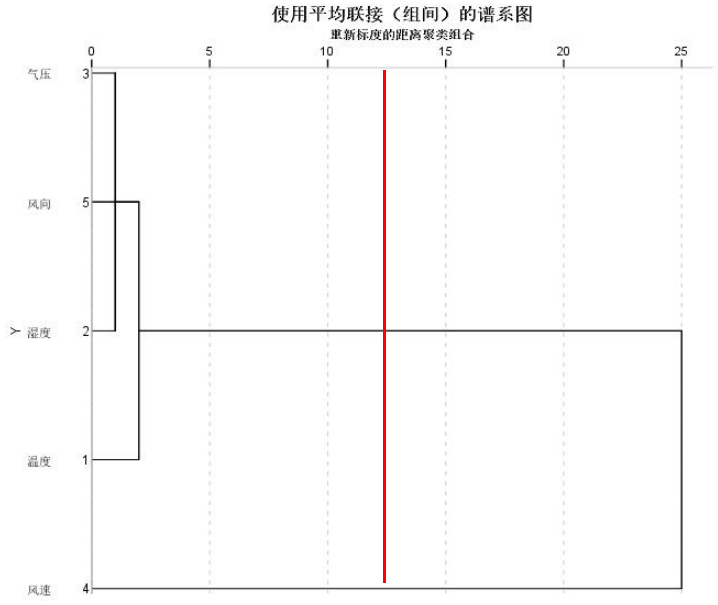


图 4-5 系统聚类谱系图（树状图）

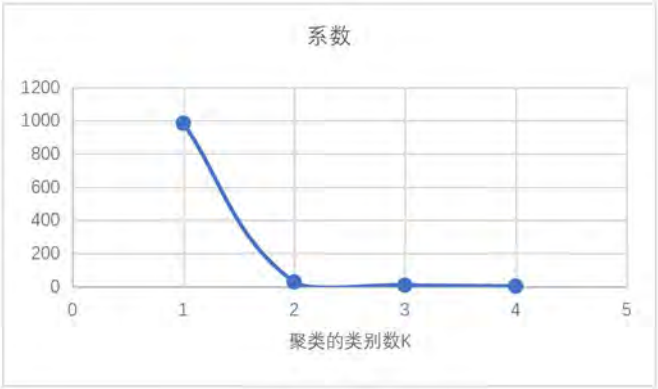


图 4-6 分类图

由图 4-6 可知，当分类个数为 2 时聚类系数曲线趋于平滑，所以聚类个数为 2。

4.4 结果分析

根据上文所述方法进行建模计算，最终得到气象条件的分类示意图 4-7。

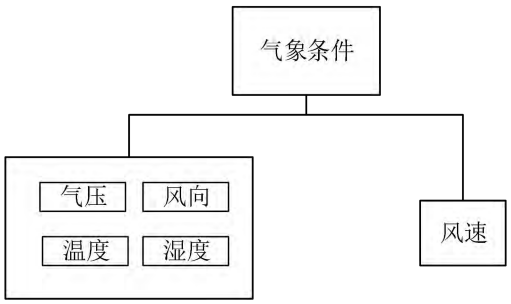


图 4-7 气象分类示意图

结合表 4-2、4-3、4-4 及图 4-7 对气象条件分类展开分析，由图 4-7 可知，应当将聚类结果分为两类。首先将气压、风向、温度与湿度划为一类，风速自成一类。这是由于在 6 中污染物中，风速与 SO_2、NO_2、PM10、PM2.5 这 4 中污染物都呈明显的负相关关系，相关系数分别为-1.192491、-17.09738、-14.00453、-10.64181，其值远大于其他相关系数值，即风速对各污染物浓度影响很大。随着风速的增大，除 O_3 外各污染物浓度均会降低。

5 问题三

5.1 问题分析

问题三旨在对 A、B、C 三个监测点一次预报数据与实测数据的基础上建立二次预报数学模型，来预测 6 种常规污染物单日浓度值，并保证二次预测的 AQI 值与 O_3 的气体浓度尽可能地接近实测值。本文首先对附件 2 中数据进行预处理及分析（如附件 2 中监测点 C 逐小时污染物浓度与气象实测数据的第 2284-2289 行数据严重偏离正常值范围）。接着采用多元线性回归和 BP 神经网络两种方法建立二次预测模型。

对修正的一次预报数据采用多元线性回归法逐步回归时，识别 21 个变量中的重要变量，计算其回归系数拟合值，并进行显著性检验。将预测值与实测值进行对比，根据拟合度判断预测是否准确。其次，根据一次预报搭建的多元线性回归模型显著性结果，构建 BP 神经网络预测模型，最终得到污染物单日浓度二次预测模型。通过对比分析 AQI 和首要污染物 O3 的拟合预测结果选择最优的方法。

该问题求解思路如下图所示：

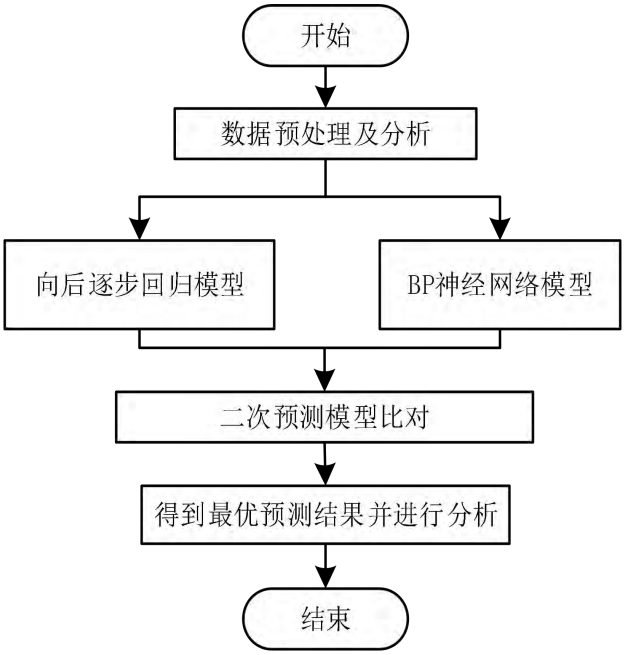


图 5-1 问题三求解流程图

5.2 模型建立与求解

5.2.1 数据预处理及分析

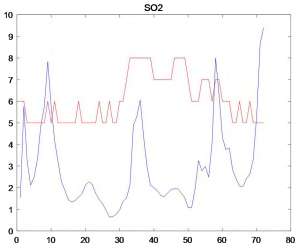
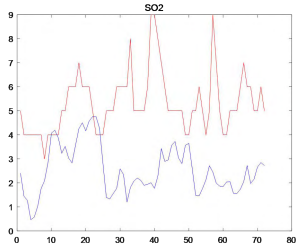
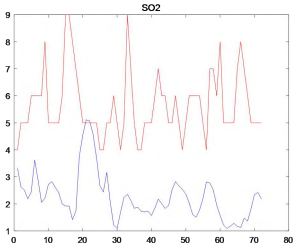
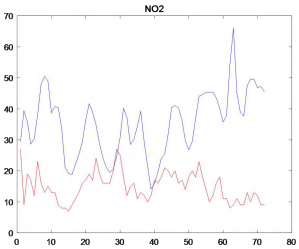
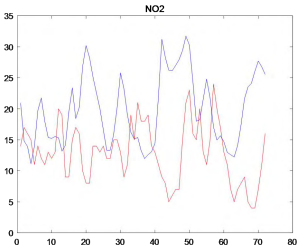
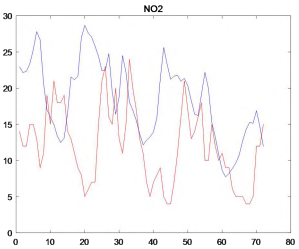
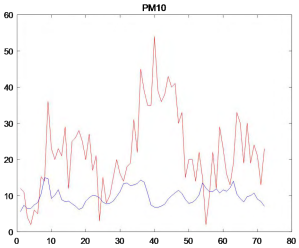
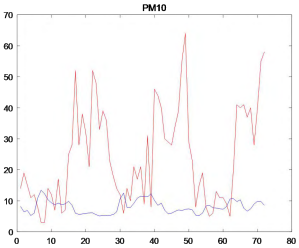
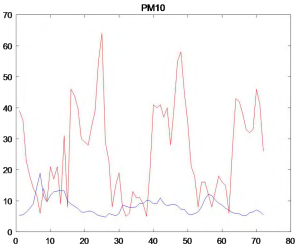
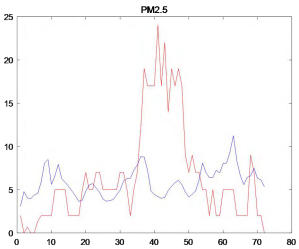
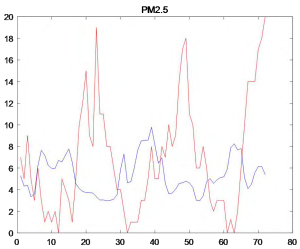
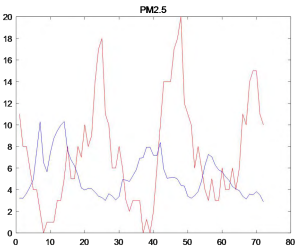
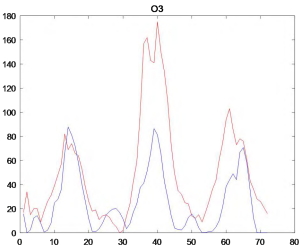
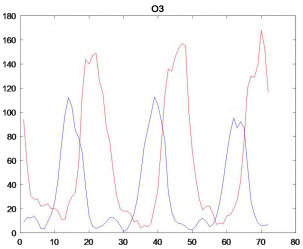
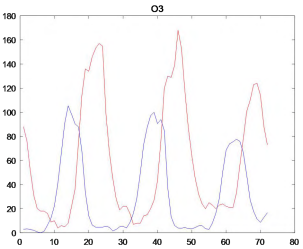
1）数据预处理

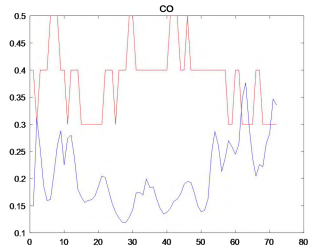
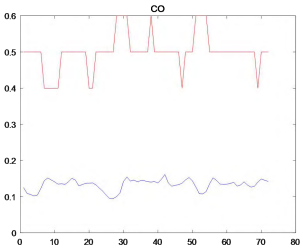
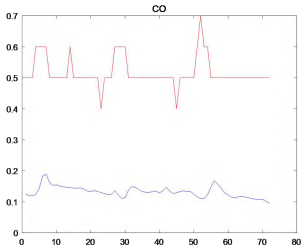
附件 2 给出了 A、B、C 三个监测点的预报数据，首先对数据进行预处理。同问题二步骤相同，完成对缺失数据的填充、异常数据修正和数据归一化、标准化。

2）数据分析（一次预报数据与实际数据间的关系）

根据附件 2 所给数据，画出 2021 年监测点 A 于 2020 年 7 月 23 日、24 日和 2021 年 7 月 10 日的各污染物浓度预测结果与实测结果对比图，如表 5-1 所示。

表 5-1 污染物浓度实测数据与预测数据对比图

气象条件	日期		
	2020 年 7 月 23 日	2020 年 7 月 24 日	2021 年 7 月 10 日
SO_2			
NO_2			
PM10			
PM2.5			
O_3			

CO			

注：红色为实测数据，蓝色为预测数据。

分析表 5-1 可知，23、24 日各污染物浓度实测数据较为相似，与 10 日实测数据存在一定差异。由图也可以看出，一次预报模型与实际测量模型都很难看出周期性，很难使用灰色预测模型和时间序列分析，且一次预报模型与实测数据没有强相关性，对比预测曲线与实测曲线，发现在 6 种污染物中，一次预报对 O_3 浓度预测最为准确，其预测值与实测数据在趋势上基本一致，但仍有很大误差；对 NO_2、SO_2、PM2.5 浓度预测较不准确，预测曲线与实际数据趋势很难保持一致；对 PM10、CO 预测准确度不高，预测曲线与实际曲线存在一定差异，无法反映其浓度变化趋势。由此发现，仅采用一次预报并不能实现对天气情况的准确预报，需要在一次预报的基础上进行二次预测以达到准确预测的目的。

5.2.2 二次数据预测模型

5.2.2.1 基于向后逐步回归法建模预测

针对一次预报数据与实测数据存在较大差异的问题，本节先采用多元线性回归法（逐步向后回归）对附件 2 一次预报数据的 21 个变量建模，识别其中重要变量，计算拟合回归系数，并进行显著性检验。将多元回归后预测值与实测值进行对比，根据拟合度判断预测是否准确。

1）逐步回归法

逐步回归法分为向前逐步回归和向后逐步回归。其基本思想是通过剔除变量中不重要且和其他变量高度相关的变量，降低多重共线性程度。本文选用向后逐步回归，首先将所有变量均放入模型，之后尝试将其中一个自变量从模型中剔除，看整个模型解释因变量的变异是否有显著变化，之后将最没有解释力的那个自变量剔除；此过程不断迭代，直到没有自变量符合剔除的条件。具体流程图如图 5-2 所示：

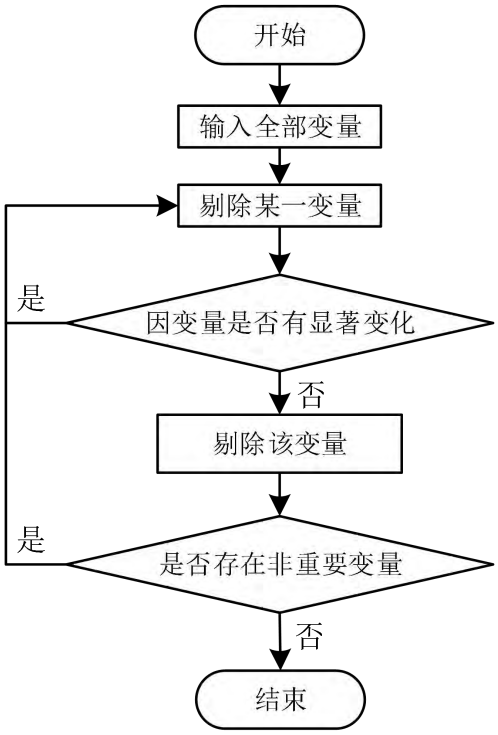


图 5-2 向后逐步回归流程图

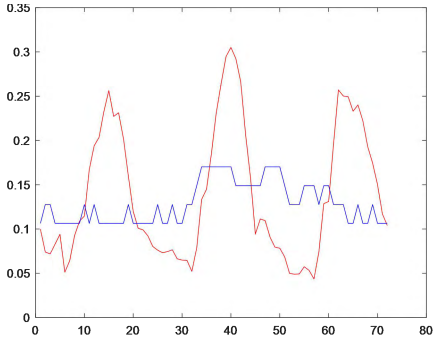
2）向后逐步回归建模预测

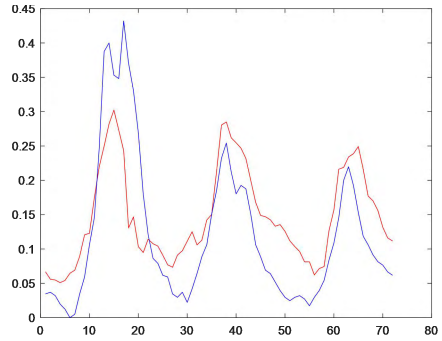
将附件 2 一次预报数据中气象条件及各污染物浓度等 21 个指标为因变量，实测数据中各污染物浓度为因变量，建立向后回归预测模型。通过 stata 软件计算发现存在多重共线性问题，经过分析一次预报数据后得知，短波辐射与地面太阳能存在多重共线性问题，剔除短波辐射变量后，多元线性回归的 F 检验成立，可以使用多元线性回归求解。以 O_3 为例，使用向后逐步回归模型剔除变量为：短波辐射、大气压、CO 浓度和近地两米温度。拟合优度 R=0.432，其结果较低，预测结果可能较差。

表 5-2 线性回归拟合优度结果

	SO2	NO_2	PM_{10}	$PM_{2.5}$	O_3	CO
拟合优度 R	0.3154	0.1941	0.2287	0.1653	0.3766	0.3530

利用上述回归模型对 7 月 8 日至 10 日 3 天的各污染物浓度进行预测，模型所得结果与实测数据对比图见附录 A。图 5-3 给出了 O_3 浓度模型预测数据与实测数据对比图，图 5-4 为拟合效果图。





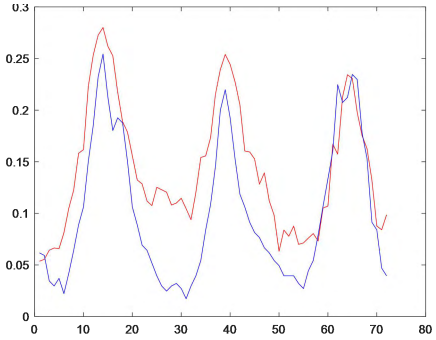
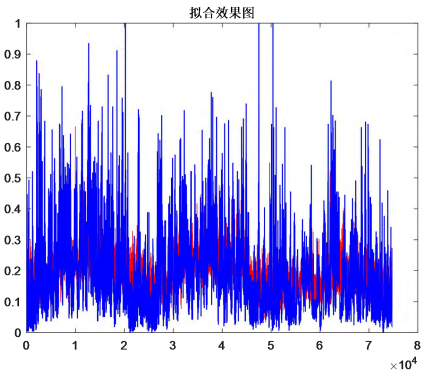
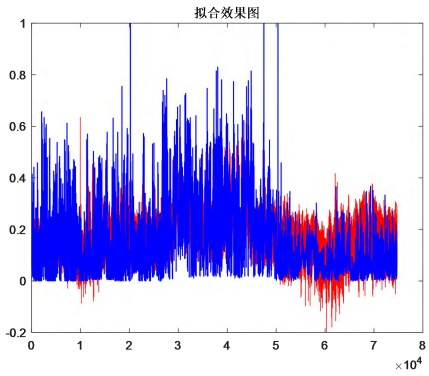


图 5-3 向后逐步回归法预测对比图





AQI O_3

图 5-4 拟合效果图

21 个变量的多元线性回归拟合系数值见表 5-3。

表 5-3 多元线性回归拟合系数值

参数	SO_2	NO_2	PM10	PM2.5	O_3	CO
近地 2 米温度	-0.0736	0.0815	0.0978	0.0000	0.0396	0.2500
地表温度	-0.4819	0.6080	0.4300	0.4404	0.4922	0.1082
比湿	0.3163	-0.4533	-0.4973	-0.5174	-0.3273	-0.2941
湿度	-0.2482	0.2026	0.1019	0.2065	0.0344	0.1725
近地 10 米风速	0.0286	0.0000	0.0249	-0.0197	-0.0345	-0.0126
近地 10 米风向	0.0324	0.0397	0.0312	0.0544	-0.0225	0.0369

雨量	0.0635	-0.0312	0.0000	0.0000	0.0180	0.0152
云量	-0.0134	0.0444	0.0000	0.0102	-0.0456	0.0493
边界层高度	0.0000	-0.0658	-0.0828	-0.0400	-0.0712	0.0000
大气压	0.0000	0.0000	-0.0841	-0.0610	0.0192	0.2397
感热通量	-0.2194	-0.0381	-0.1316	-0.0260	0.2633	0.0000
潜热通量	-0.0731	-0.1569	-0.2456	-0.0820	0.1559	-0.0376
长波辐射	0.1057	-0.0889	0.0000	-0.0490	-0.0257	-0.0516
短波辐射	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
地面太阳能辐射	0.3286	-0.1052	0.1876	-0.0639	-0.2687	-0.0519
SO_2 小时平均浓度	0.1322	0.0451	0.1042	0.0898	-0.0333	-0.0213
NO_2 小时平均浓度	0.0911	0.0893	0.1546	0.1319	0.1725	0.0886
PM10 小时平均浓度	0.7694	0.0569	0.3315	0.2561	0.1157	0.0309
PM2.5 小时平均浓度	-0.5319	0.0000	-0.2293	0.0000	-0.1798	0.0746
O_3 小时平均浓度	0.2230	-0.0518	0.1169	0.1338	0.4045	0.0670
CO 小时平均浓度	-0.0901	0.0963	0.0177	0.0000	-0.0456	-0.0218
CONS	0.4131	0.1129	0.2356	0.1955	0.0841	0.1748

可以直观地看出来，相较于表 5-1 中一次预报数据与实测数据的拟合程度，采用逐步向后回归模型对一次预报数据进行再预测后，其预测值与实际值拟合度有所提高，但结果仍不理想，存在着一定误差。基于上述研究内容，下文结合 BP 神经网络建立预测模型。

5.2.2.1 BP 神经网络

BP（Back Propagation）神经网络是一种按照误差逆向传播算法训练的多层前馈神经网络，也是应用最广泛的神经网络[7]。其不需要确定输入与输出之间的数学方程，通过自身不断训练，学习给定的规则，当给定输入值后即可得到输出值。BP 算法的基本思想是梯度下降法，使输入和输出误差均方差为最小。基本 BP 算法由信号的前向传播和误差的反向传播两个过程组成。经过反复学习训练，选择最小误差相对应的网络参数，来使得训练停止。这时经过训练的神经网络即能对类似样本的输入来输出误差最小的输出值。

BP 神经网络模型拓扑结构包括输入层、隐层和输出层。网络结构如下图：

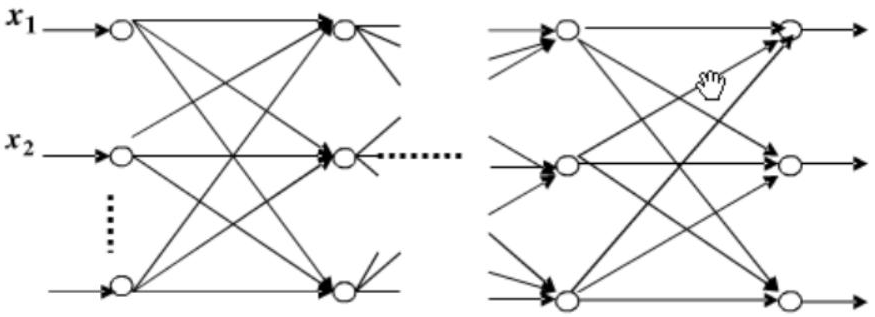


图 5-5　BP 网络结构

BP 神经网络整个学习过程如下图所示：

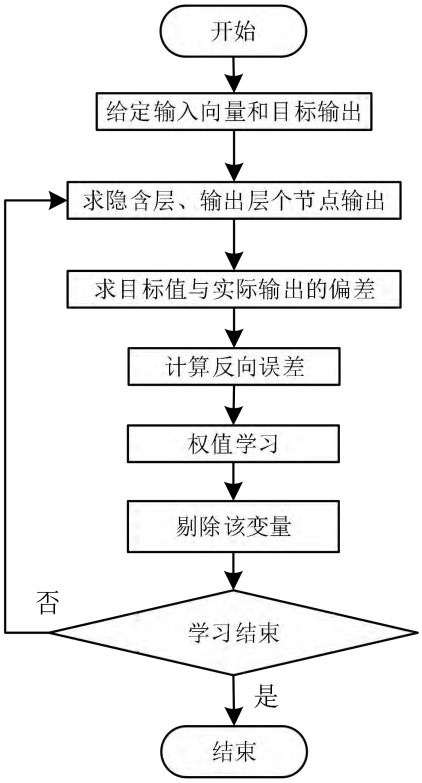


图 5-6　学习过程

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_{kj}，隐层与输出层之间的权值为 w_{jk}。隐层的传递函数为 $f_1(\cdot)$，输出层的传递函数为 $f_2(\cdot)$， 则隐层节点的输出为：

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

输出层节点的输出为：

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

根据以上原理，可以将人工神经网络计算过程归纳为如下几步：

（1）初始值选择 $w(0)$；

（2）前向计算, 求出所有神经元的输出： $a^k(t)$；

（3）对输出层计算 $\delta: \delta_j = (t_j - a_j)a_j(1 - a_j)$；

（4）从后向前计算各隐层 $\delta: \delta_j = a_j(1 - a_j)\sum_i w_{ji}\delta_i$；

（5）计算并保存各权值修正量: $\Delta w_{ij} = -\eta\delta_j a_i$；

（6）修正权值: $w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}$；

（7）判断是否收敛，如果收敛则结束，不收敛则转至步骤（2)。

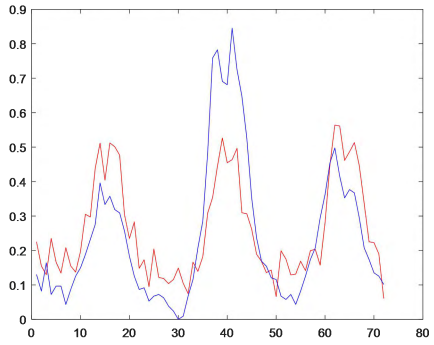
训练方法采用 trainlm (Levenberg-Marquardt 方法)，权值函数设置为 negdist (归一化列权值初始化函数)，学习函数设为 learngdm (附加动量因子的梯度下降权值/阈值学习函数)，初始化参数采用 initwb 函数。图 5-7 给出了 341 个样本（2020 年 7 月 23—2021 年 7 月 2

日）对样本数据进行 BP 神经网络训练后的回归曲线图。

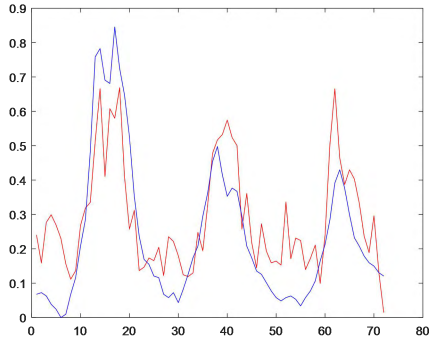
表 5-4 BP 神经网络拟合度表

	SO2	NO_2	PM_{10}	$PM_{2.5}$	O_3	CO
回归值 R	0.804	0.765	0.746	0.769	0.854	0.789

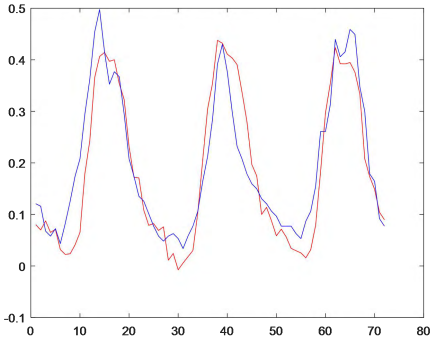
为了验证预测模型的准确性，根据题目的所给要求，在 AQI 最大相对误差尽量小、首要污染物预测准确度尽量高的情况下，对 2021 年监测点 A：7 月 8 日至 10 日的 AQI 及 O_3 浓度进行预测。预测值与实际值对比如下图所示。



7 月 8 日



7 月 9 日



7 月 10 日

图 5-7　O_3 预测值与实测值对比图

通过图 5-7 可以发现，O_3 预测值与实际值在数值及趋势上都呈现出比多元线性回归更

为良好的拟合度和预测结果，本文最终决定利用 BP 神经网络预测模型得到二次预测值。

5.3 结果分析

基于上述方法，对 2021 年 7 月 13 日至 15 日各污染物浓度进行预测，预测结果如下表所示：

表 5-5 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_2 ($\mu g/m^3$)	NO_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	5.7228	16.9041	23.1995	10.6749	109.8581	0.3957	59	O_3
2021/7/14	监测点 A	5.7840	16.2567	21.0790	10.2133	108.6080	0.3933	58	O_3
2021/7/15	监测点 A	5.6664	16.2423	21.3304	9.2234	113.7558	0.3882	62	O_3
2021/7/13	监测点 B	6.4427	8.7157	13.4350	5.6570	92.2311	0.3126	47	无
2021/7/14	监测点 B	6.4442	8.3435	13.1052	5.1888	91.9425	0.3081	46	无
2021/7/15	监测点 B	6.4450	8.6088	12.2888	4.7073	90.8675	0.3060	46	无
2021/7/13	监测点 C	15.6343	25.9811	37.5084	23.6116	126.2348	0.5182	72	O_3
2021/7/14	监测点 C	15.8407	26.9351	38.4564	23.7146	138.7639	0.5074	83	O_3
2021/7/15	监测点 C	15.6749	25.6906	39.1785	23.8042	159.9454	0.5133	100	O_3

分析表 5-5 可得如下结论：

（1）2021 年监测点 A：7 月 13 日的 AQI 为 59，首要污染物为 O_3；7 月 14 日的 AQI 为 58，首要污染物为 O_3；7 月 15 日的 AQI 为 62，首要污染物为 O_3。

（2）2021 年监测点 B：7 月 13 日的 AQI 为 47，无首要污染物；7 月 14 日的 AQI 为 46，无首要污染物；7 月 15 日的 AQI 为 46，无首要污染物。

（3）2021 年监测点 C：7 月 13 日的 AQI 为 72，首要污染物为 O_3；7 月 14 日的 AQI 为 83，首要污染物为 O_3；7 月 15 日的 AQI 为 100，首要污染物为 O_3。

6 问题四

6.1 问题分析

与问题三相比，问题四需要考虑临近点污染气体和监测点污染气体的相关性，因此不能简单地将 A、A1、A2、A3 放到同一样本集中进行预测，本文首先用在地质学中较为经典的 Kriging 插值法来进行临近点的预测，即选择三个监测点作为插值点，对剩余的一个点进行插值预测，将 Kriging 插值预测结果与一次预报结果进行对比分析，并将插值预测结果与一次预报结果作为因变量，对实测的污染气体浓度进行灰色关联度分析，得到插值预测结果和一次预报结果的权重，随后将根据权重的得到的变量作为自变量，实测数据作为因变量进行 BP 神经网络拟合，得到二次预测模型，最终得到污染物单日浓度预测值。根据计算公式得到相应 AQI 和首要污染物。该问题求解思路如下图所示：

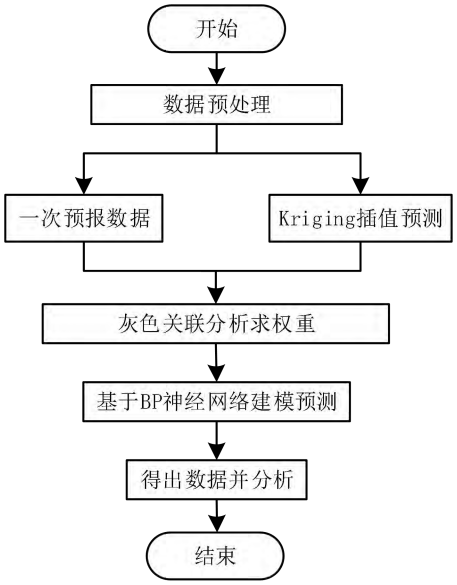


图 6-1 问题四求解流程图

6.2 模型建立

6.2.1 Kriging 算法

Kriging[8]代理模型是一种基于统计理论的插值技术，由一个参数模型和非参数随机过程联合构成的。其不仅比单个参数化模型更具有灵活性、有更强的预测能力，还克服了非参数化模型处理高维数据存在的局限性。相比于其它传统的插值技术，kriging 模型有如下优点：1）采用已知信息的动态构造，不需要所有的信息，只使用估计点附近的某些信息即可进行模拟。2）具有局部、全局的统计特性。该性质使其可以对已知信息的趋势和动态进行分析。

Kriging 作为线性回归分析的一种改进的技术，包含了线性回归部分和非参数部分，其中的非参数部分被视作随机分布的实现。Kriging 模型在周围的已知变量的信息基础上对某一点进行预测，也就是说通过对某一点、在一定范围内其他信息加权的线性组合来估计这一点的未知信息，通过最小化估计值的误差方差来选择加权，基于此，Kriging 模型被视为最优的线性无偏估计。

Kriging 插值方法是为了建立一个替代模型。从 Kriging 的角度来看，$f(x)$ 应该是一个高斯过程，$f(Xt)$ 服从一个联合高斯分布，其中 Xt 是 n 维训练集的输入。高斯过程完全是由它的平均函数 $m(x)$ 和协方差函数 $k\left(x,x'\right)$ 指定的。它可表示为：

$$f(x) \sim GP\left(m(x), k\left(x, x'\right)\right) \quad (6\text{-}1)$$

平均函数 m(x) 表示 kriging 的趋势，通常定义为：

$$m(x) = g^T(x)\beta \quad (6\text{-}2)$$

其中 $g^T(x) = [g1(x), ..., gp(x)]$ 是一个阻遏物向量，而 β 是一个回归参数的 p 维向量。

协方差函数 $k\left(x, x'\right)$模拟了高斯过程 Z(x) 的不同值之间的依赖关系。它可表示为：

$$k\left(x, x'\right) = \sigma^2 r\left(x, x'; \theta\right) \quad (6\text{-}3)$$

其中 σ^2 和 θ 为协方差函数的超参数；$r(x，x；\theta)$为仅依赖于输入样本之间的距离，即当距离为 0 时，相关函数等于 1，当距离接近无穷时，相关函数等于 0。

事实上，协方差函数是高斯过程中最重要的组成部分，并且在所考虑的高斯过程的正则性与协方差函数的正则性之间存在着紧密的关系。平方指数协方差函数、 v-Matern 协方差函数和 γ-指数协方差函数是三个广泛使用的协方差函数，本文采用的是平方指数协方差函数，其相应公式如下所示：

$$k\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \sigma^2 \exp\left(-\frac{\|x - x'\|^2}{2\theta^2}\right) \quad (6\text{-}4)$$

Kriging 均值$\hat{m}\left(x_p\right)$是用来近似目标函数$f\left(\boldsymbol{x}_p\right)$的替代模型，Kriging 方差$\hat{s}^2\left(\boldsymbol{x}_p\right)$表示模型的均方误差。从（）中我们可以看出，平均函数和方差函数不仅取决于预测点与训练点之间的距离，而且还取决于训练点本身之间的距离。为了估计参数 $(\beta, \sigma^2, \theta)$，一般采用最大似然估计。即[8]：

$$\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\boldsymbol{\theta}} = \arg\ \mathrm{m}\ \ \log P(\boldsymbol{Y}_t | \boldsymbol{X}_t, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) \quad (6\text{-}5)$$

$$P(\boldsymbol{Y}_t | \boldsymbol{X}_t, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = N(\boldsymbol{Y}_t | \boldsymbol{g}\boldsymbol{\beta}, R(\boldsymbol{X}_t, \boldsymbol{X}_t | \boldsymbol{\theta}) + \sigma^2 \boldsymbol{I}_n) \quad (6\text{-}6)$$

为了提高 Kriging 模型的准确性和鲁棒性，在估计超参数之前，应将训练样本进行归一化。

6.2.2 灰色关联分析法

灰色关联分析[9-11]可以对系统发展变化态势进行定量描述和比较。其通过确定参考数据和比较数据列的几何形状的相似程度来判断其联系的紧密程度，反映出了曲线间的关联程度。若两个因素变化的趋势一致，则认为二者关联程度较高，否则则较低。灰色关联度可作为衡量因素间关联程度的一种方法。通常采用此方法来分析各个因素对于结果的影响程度，其核心是根据规则通过时间关系确立母序列，估计对象作为子序列，求子序列与母序列的相关程度，根据其相关性得出结论。灰色关联分析方法也适用无规律数据，不会出现量化结果与定性分析结果不符的情况。其首先将估计指标和原始数据进行去量纲化处理，通过计算关联系数、比较关联度大小对预估指标进行排序。灰色关联度应用范围极广，在社会科学和自然科学社会经济领域都取得较好的应用效果。

灰色关联分析的具体计算步骤如下：

1）确定分析数列

2）对变量进行去量纲化

进行灰色关联度分析前，要进行去量纲化，将系统中各数据量统一，便于比较。

$$x_i(k)=\frac{X_i(k)}{X_i(l)},k=1,2,\Lambda,n;i=0,1,2,\Lambda,m \quad (6\text{-}7)$$

3）计算关联系数

$x_0(k)$ 与 $x_i(k)$ 的关联系数为：

$$\xi_i(k)=\frac{\min\limits_i|y(k)-x_i(k)|+\rho\max\limits_i\max\limits_\ell|y(k)-x_i(k)|}{|y(k)-x_i(k)|+\rho\max\limits_i\max\limits_\ell|y(k)-x_i(k)|} \quad (6\text{-}8)$$

$$\triangle_i(k)=|y(k)-x_i(k)|_i \quad (6\text{-}9)$$

$$\xi_i(k)=\frac{\min\limits_i\min\limits_k\Delta_i(k)+\rho\max\limits_i\Delta_i(k)}{\Delta_i(k)+\rho\min\limits_i\Delta_i(k)} \quad (6\text{-}10)$$

$\rho\in(0,\infty)$称为分辨系数。一般 ρ 的取值区间为 $(0,1)$。当 ρ 越小，分辨力越大，$\rho\le0.5463$ 时，分辨力最好，通常取 $\rho=0.5$ 。

4）计算关联度 关联度 r_i 公式如下:

$$r_i=\frac{1}{n}\sum_{k=1}^{n}\xi_i(k),k=1,2,\Lambda,n \quad (6\text{-}11)$$

6.3 模型求解

6.3.1 数据预处理

（1）缺失值处理

在对 A、A1、A2、A3 数据进行预处理时，除已经标记的缺失值外，发现四个点的实测数据还有少量未标记日期的缺失值，由于这些缺失日期时间较长，使用 3σ准则难以保证数据的准确度，因此，直接将这些点剔除。剔除后，问题四的样本数据为四个监测点在 2020 年 7 月 27 日至 2021 年 7 月 2 日的一次预报和实测数据，并用 2021 年 7 月 7 日至 2021 年 7 月 10 日的数据作为验证数据，验证预测结果的正确性。其余缺失值和异常值的处理过程与问题 2、3 相同。

（2）变量处理

由于本题使用了 Kriging 插值法和灰色关联度分析，其实测数据和一次预报数据中，气象条件对实测的污染气体浓度的相关性的影响本题暂未考虑，只考虑相邻区域的污染物浓度的相关性。

表 6-1 问题 4 剔除变量

剔除变量		
感热通量	近地 2 米温度	近地 10 米风向
潜热通量	地表温度	雨量
长波辐射	比湿	云量
短波辐射	湿度	边界层高度
地面太阳能辐射	近地 10 米风速	大气压

6.3.2 Kriging 插值预测

在本文中，以 A、A1、A2、A3 四个点的坐标作为 x 值，使用 matlab 中的 DACE 工具箱计算其插值预测值。计算得到的四个检测点的插值预测结果如表 6-2 所示，其每个污染气体在每个监测点的插值预测结果样本值都有两万多个，由于篇幅有限，只截取 A 监测点 SO_2 部分进行展示

表 6-2 Kriging 插值预测结果

O_3 浓度

A	A1	A2	A3
31.82766667	27.26796667	28.59736667	30.1142
40.15776667	35.67136667	36.48243333	38.31433333
49.39236667	47.2938	47.6688	47.90053333
52.18676667	53.42736667	52.45483333	54.22413333
45.9086	42.6405	41.63406667	44.99033333
45.5484	42.54033333	40.6916	44.37976667
40.3308	36.3747	35.95063333	38.17766667
31.82766667	27.26796667	28.59736667	30.1142
40.15776667	35.67136667	36.48243333	38.31433333
⋮	⋮	⋮	⋮
51.77616667	50.39833333	51.18	49.2718
42.64836667	42.19876667	41.79413333	40.14193333
30.7832	28.5929	27.78066667	27.95103333
15.90816667	14.41943667	13.46730333	12.90660333
51.77616667	50.39833333	51.18	49.2718
42.64836667	42.19876667	41.79413333	40.14193333

对结果进行分析可知，Kriging 插值预测的结果与四个监测点的一次预报结果在数值上有较大差异，需对其关系进行进一步探讨。

6.3.3 灰色关联度分析

灰色理论应用最广泛的是关联分析方法，其关联系数反映各评价对象对理想（标准）对象的接近程度，关联系数越大则评价指标越优。其计算过程及结果如下：

（1）选择参考数列

在灰色关联度分析中，设 i 为第 i 个时刻，$i=1，2，…，n$；k 为第 k 个指标，在本文中以某监测点污染气体一次预报数据和其他监测点污染气体 Kriging 插值预测数据作为指标，取各时刻的插值预测数据和一次预报数据的最大值u_{0k}为参考数列，则有：

$$U_0 = (c, u_{02}, \cdots u_{0m}) \tag{6-12}$$

（2）指标值的规范化

由于评价指标间通常是有不同的量纲和数量级，因此归一化对数据进行处理，其处理过程与问题二的预处理过程相同。

（3）计算关联系数

根据灰色系统理论，将规范化后的数列 U0k 作为参考数列， 将x_i作为比较数列，带入公式（6-12）求解得，不同污染气体在不同观测点处的关联系数值如表 6-3 所示。

表 6-3 关联系数值

监测点	权重	SO_2	NO_2	PM10	PM2.5	O_3	CO
A	插值预测权重	0.8023	0.7782	0.7412	0.7818	0.7939	0.6153
	一次预报变量权重	0.7933	0.7466	0.7368	0.7773	0.7860	0.6272
A1	插值预测权重	0.8129	0.7542	0.7429	0.7796	0.7928	0.6231

	一次预报变量权重	0.7882	0.7714	0.7324	0.7767	0.7922	0.6056
A2	插值预测权重权重	0.7878	0.7633	0.7405	0.7779	0.7912	0.6221
	一次预报变量权重	0.8044	0.7649	0.7356	0.7782	0.7948	0.6104
A3	插值预测权重	0.7998	0.7645	0.7366	0.7846	0.7946	0.6148
	一次预报变量权重	0.7867	0.7592	0.7441	0.7816	0.7861	0.6284

由关联系数表我们可以得知，六种污染气体插值预测和一次预报数据得权重值大体相同，这说明两种预测得值虽然存在较大的差异，但图形大体相似，因此，可以用相邻观测站得一次预报数据协同预测，可以建立包含 A、A1、A2、A3 四个监测点的协同预报模型。

6.3.4 BP 神经网络构建二次预测模型

将通过灰色关联度计算得到的权值进行污染气体的数据拟合，其公式如（6-13）所示：

$$F = w_1 * F_1 + w_2 * F_2 \quad (6\text{-}13)$$

式中，F 为拟合值，F1、F2 分别为插值和一次预报值，w1、w2 为其对应的权值。

将 F 作为自变量，对应时刻的实测数据作为因变量，将一次预报模型运行的第 *i* 天作为其第 *i* 个样本点对第 *i* 天对应日期的 72 个时刻点进行神经网络模型的预测，其神经网络结构图如图 6-2 所示。

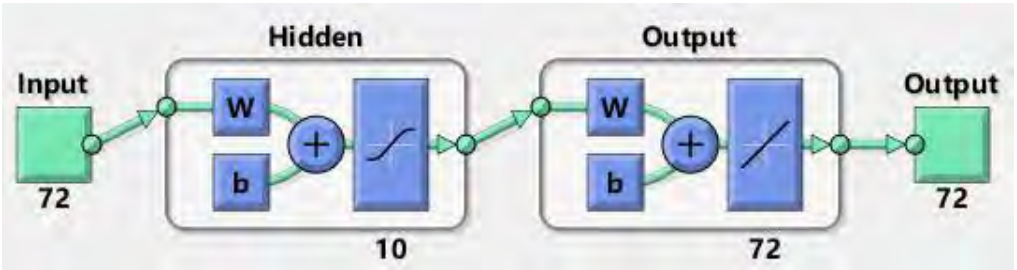
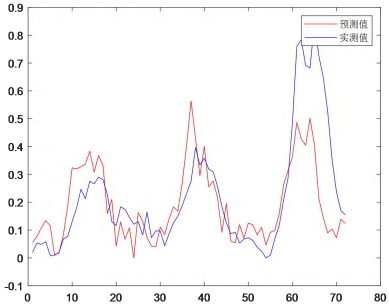


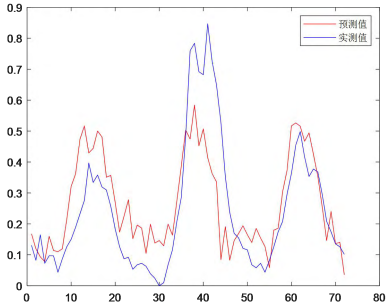
图 6-2 神经网络结构图

6.4 结果分析

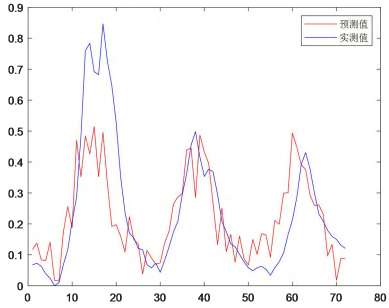
在使用 BP 神经网络得到二次预测模型后，使用 2021 年 7 月 7 日至 7 月 10 日结果进行验证，将二次预测结果与监测点实际测量结果进行比对，由于 6 种污染气体在四个监测点三天的预测结果图需要 72 个，会占用大量篇幅，因此本文仅将首要污染物 O_3 在四个监测点的拟合结果放在正文中。O_3 浓度拟合结果如图 6-3 所示。



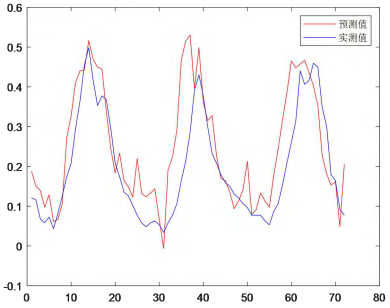
2021 年 7 月 7 日



2021 年 7 月 8 日

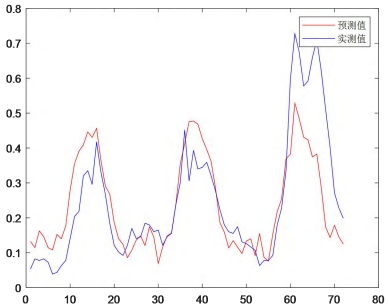


2021 年 7 月 9 日

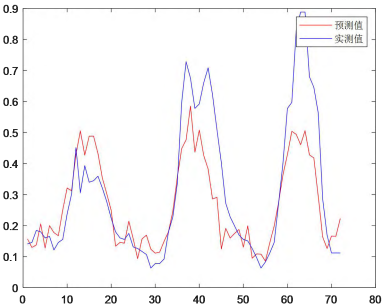


2021 年 7 月 10 日

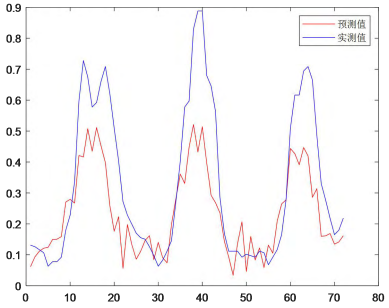
（a）监测点 A 拟合图



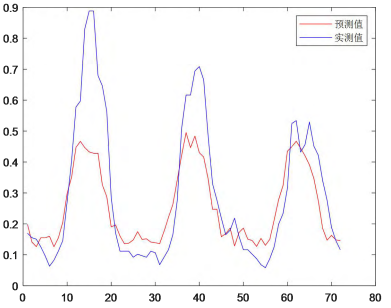
2021 年 7 月 7 日



2021 年 7 月 8 日

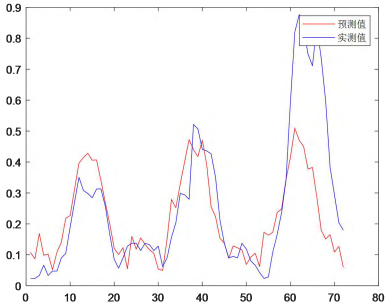


2021 年 7 月 9 日

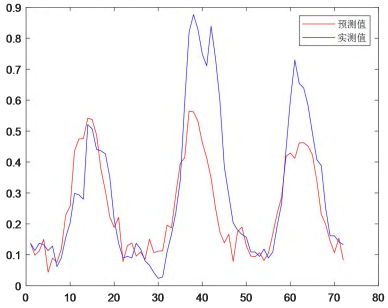


2021 年 7 月 10 日

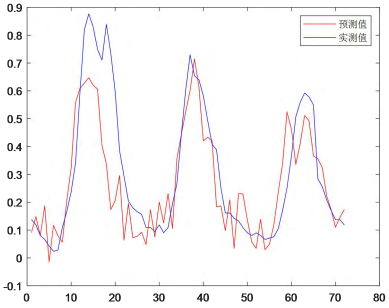
（b）监测点 A1 拟合图



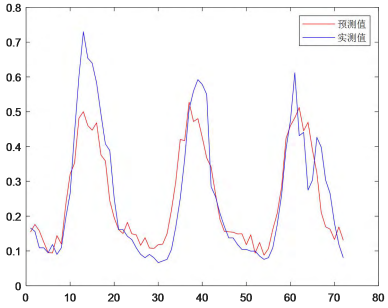
2021 年 7 月 7 日



2021 年 7 月 8 日



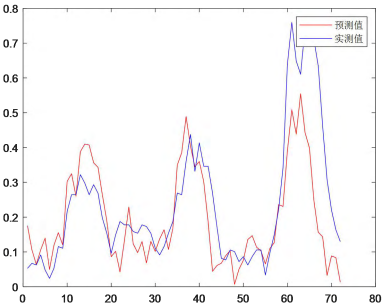
2021 年 7 月 9 日



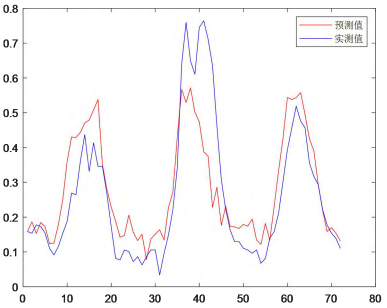
2021 年 7 月 10 日

（c）监测点 A2 拟合图

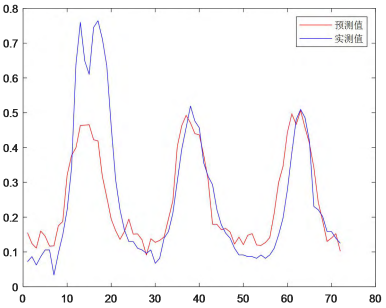
图 6-3 O_3 浓度拟合结果



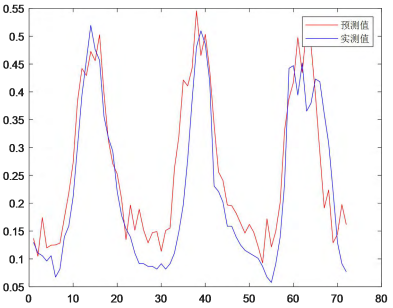
2021 年 7 月 7 日



2021 年 7 月 8 日



2021 年 7 月 9 日



2021 年 7 月 10 日

（d）监测点 A2 拟合图

由图 6-3 可以得出，本次建立的二次预测模型对 O_3 浓度拟合程度高，准确性较好。2021 年 7 月 13 日预测结果如表 6-4 所示。

表 6-4 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_2 ($\mu g/m^3$)	NO_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	6.6614	16.0742	23.6818	5.9854	90.0670	0.4105	46	无
2021/7/14	监测点A	6.6371	15.7309	22.4409	4.9186	91.4836	0.3893	46	无
2021/7/15	监测点A	6.6176	15.1669	21.4572	5.1709	95.2891	0.3869	48	无
2021/7/13	监测点A1	7.7158	16.1117	27.9728	10.1842	100.7164	0.3604	51	O_3
2021/7/14	监测点A1	7.5887	14.9890	26.9157	9.5931	102.4186	0.3500	53	O_3
2021/7/15	监测点A1	7.6746	15.3091	26.9227	8.9776	97.7607	0.3171	49	无
2021/7/13	监测点A2	5.1836	18.4170	24.1885	7.8572	94.4586	0.4446	48	无
2021/7/14	监测点A2	5.1048	17.3578	22.2639	6.6343	96.2254	0.4146	49	无
2021/7/15	监测点A2	5.0637	16.1468	27.0544	6.9223	100.2603	0.3997	51	O_3
2021/7/13	监测点A3	4.4649	8.9861	11.4886	6.1651	99.7503	0.3755	50	无
2021/7/14	监测点A3	4.4192	7.6394	12.6952	4.4308	114.6731	0.3585	63	O_3
2021/7/15	监测点A3	4.4344	8.6350	11.0058	5.7372	108.1219	0.3451	57	O_3

将问题 3 与问题 4 构建模型在 2021 年 7 月 7 日至 12 日的 AQI 计算结果分析可知，问题四的计算结果更接近于于实测值，其计算结果分别如表 6-5 所示。

表 6-5 两种二次预测 AQI 对比

	AQI 实测值	问题 3 AQI 预测值	问题 4 AQI 预测值
2021 年 7 月 8 日	32	33	34
2021 年 7 月 9 日	89	47	60
2021 年 7 月 10 日	41	42	37
2021 年 7 月 11 日	32	37	34
2021 年 7 月 12 日	41	46	43

将问题三四构建的模型预测的在 2021 年 7 月 8 日至 2021 年 7 月 12 日的 AQI 预测结果如表 6-5 所示，其绝对误差的和为 54 和 39，平均正确率分别为 83.88%和 88.05%，结果表明，问题四构建模型，即区域协同预测模型二次预测效果更好。

7 总结

针对提出的四个问题，本文通过计算分析，总结如下：

(1)计算了监测点 A 从 2020 年 8 月 25 日至 8 月 28 日每天实测的 AQI 和首要污染物。

(2)首先对数据进行预处理及分析。由图 4-3 可知，大气中的首要污染物主要为 O_3，其占比高达 60.8%。排名前三位污染物分别为 O_3、NO_2、PM10。分析了各气象条件与各污染气体的相关性，并对气象条件和污染气体进行了多元线性回归，并用多元线性回归的显著值进行系统聚类，研究发现，温度与 O_3、NO_2、PM10 相关性较强，与 SO_2、CO 相关性较弱。在相关性较强的因素中与 O_3、PM10 呈正相关关系，与 NO_2 与呈负相关关系。湿度与 O_3、PM10、PM2.5 相关性较强，与 NO_2、SO_2、CO 相关性较弱。湿度与 6 种污染物均呈负相关关系，即湿度升高会引起各污染物浓度降低。气压与 O_3、PM10、PM2.5 相关性较强，与 NO_2、SO_2、CO 相关性较弱。气压与与其相关性较强的污染物浓度均呈正相关关系。风速与 SO_2、O_3、NO_2、PM10、PM2.5 相关性较强，与 CO 相关性较弱。在相关性强的因素中，除与 O_3 与呈正相关关系外，风速与其他污染物均呈负相关关系。风向与各污染物浓度未呈现明显的相关性。对聚类结果进行分析后，将聚类结果分为两类。首先将气压、风向、温度与湿度划为一类，风速自成一类。这是由于在 6 中污染物中，风速与 SO_2、NO_2、PM10、PM2.5 这 4 中污染物都呈明显的负相关关系，相关系数分别为-1.192491、-17.09738、-14.00453、-10.64181，其值远大于其他相关系数值，即风速对各污染物浓度影响很大。随着风速的增大，除 O_3 外各污染物浓度均会降低。

(3)将监测点 A、B、C 的一次预报和实测数据放到同一样本集中进行分析，发现一次预报结果与实测数据无太大相关性，分两种情况进行预测：一是对一次预报数据的 21 个变量与实测污染气体浓度进行多元线性回归，进行 F 检验和 P 值检验，筛选出不显著变量并求出拟合系数，求出 2021 年 7 月 8 日至 10 日的预测结果与实测数据进行比对；二是对一次预报数据 21 个变量和污染气体进行 BP 神经网络预测，求出 2021 年 7 月 8 日至 10 日的预测结果与实测数据进行比对。并将两种方法的预测结果相比，对比发现，BP 神经网络预测效果更好。最后使用 BP 神经网络对 2021 年 7 月 13 日的一次预报结果进行未来三天的二次预测。

(4)考虑到 A、A1、A2、A3 监测点的空间相关性，本文首先用在地质学中较为经典的 Kriging 插值法来进行监测点的预测，即选择三个监测点作为插值点，对剩余的一个点进行插值预测，将 Kriging 插值预测结果与一次预测结果进行对比分析，并将插值预测结果与一次预测结果作为因变量，对实测的污染气体浓度进行灰色关联度分析，得到插值预测结果和一次预测结果的权重，随后将根据权重的得到的变量作为自变量，实测数据作为因变量进行 BP 神经网络拟合，得到二次预测模型，最终得到污染物单日浓度预测值。最后，将问题三四搭建模型在监测点 A 处 2021 年 7 月 8 日至 12 日 AQI 二次预测情况与实测值进行比对，发现问题四中的区域协同预测能更好的对数据进行二次预测。

参考文献

[1] 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.

[2] 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.

[3] 戴树桂. 环境化学 [M]. 北京: 高等教育出版社, 1997.

[4] 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, 2018, 31(05): 72-76.

[5] 陈敏东. 大气臭氧污染形成机制及研究进展 [J/OL] 2018, https://max.book118.com/html/2018/0201/151478594.shtm.

[6]程开明.结构方程模型的特点及应用[J].统计与决策,2006(10):22-25.

[7]杨淑娥,黄礼.基于 BP 神经网络的上市公司财务预警模型[J].系统工程理论与实践，2005(01):12-18+26.

[8] Kleijnen,J.P.C. Kriging Metamodeling in Simulation: A Review [J].Operations research. 2009,192 ,707-716.

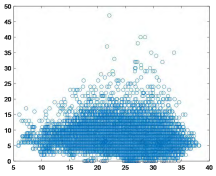
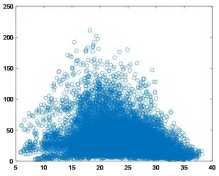
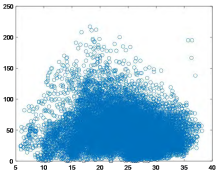
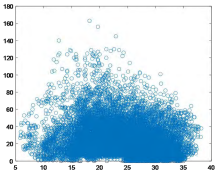
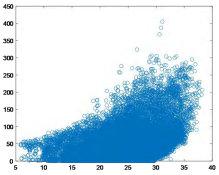
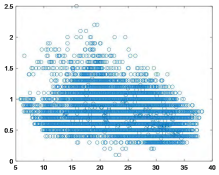
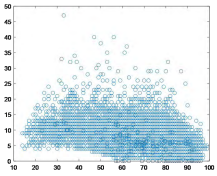
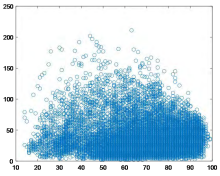
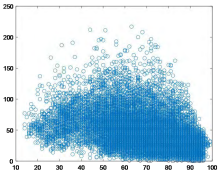
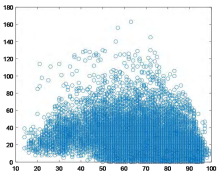
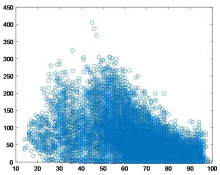
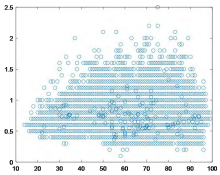
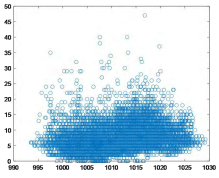
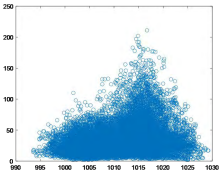
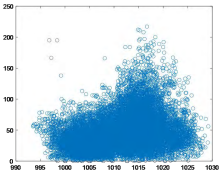
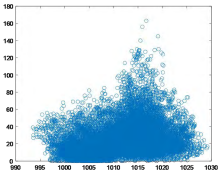
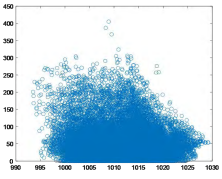
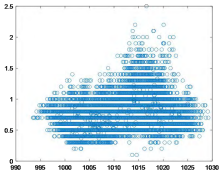
[9]罗毅,李昱龙.基于熵权法和灰色关联分析法的输电网规划方案综合决策[J].电网技术,2013,37(01):77-81.

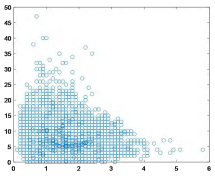
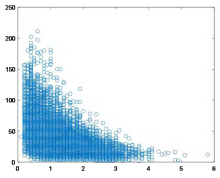
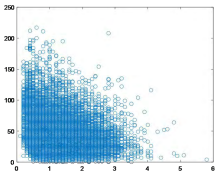
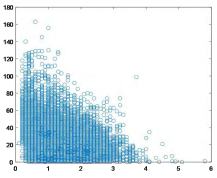
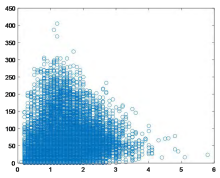
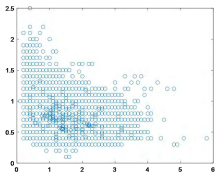
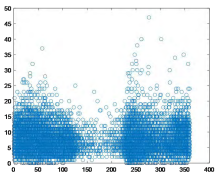
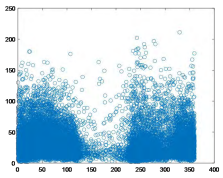
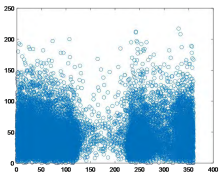
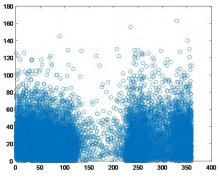
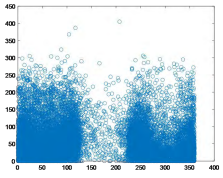
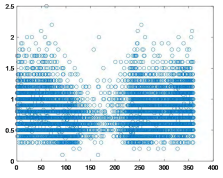
[10]张雄君,程林松,李春兰.灰色关联分析法在产量递减率影响因素分析中的应用[J].油气地质与采收率,2004(06):48-50+84.

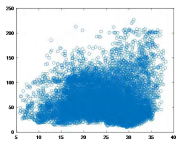
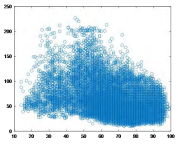
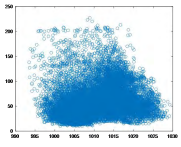
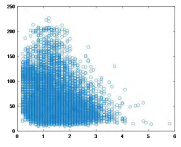
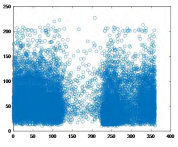
[11]付会,孙英兰,孙磊,王翠.灰色关联分析法在海洋环境质量评价中的应用[J].海洋湖沼通报,2007(03):127-131.

附录

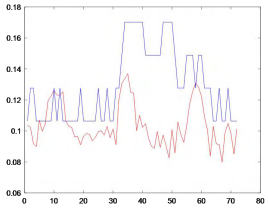
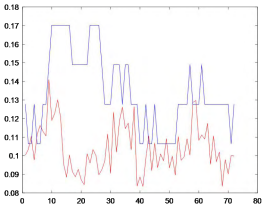
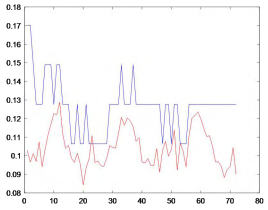
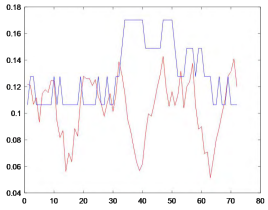
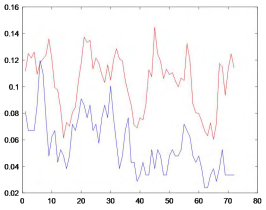
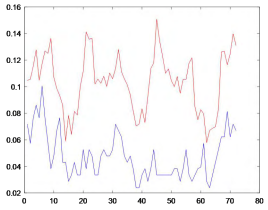
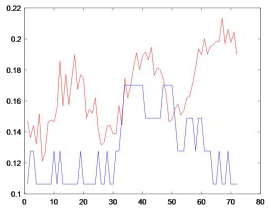
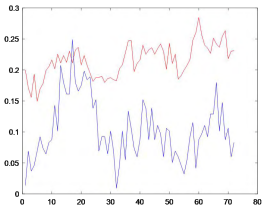
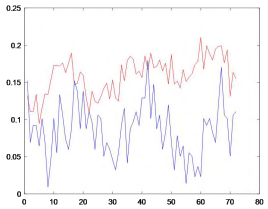
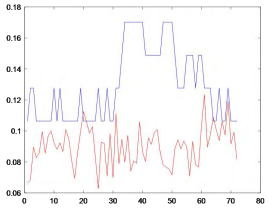
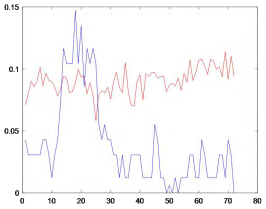
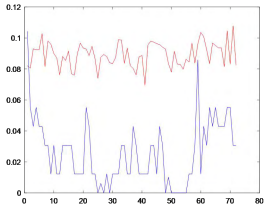
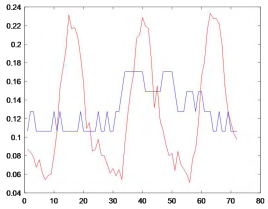
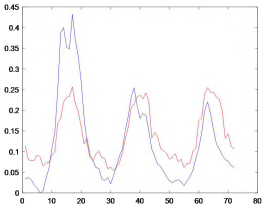
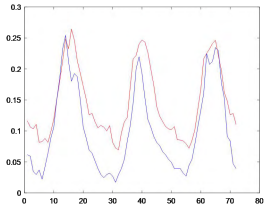
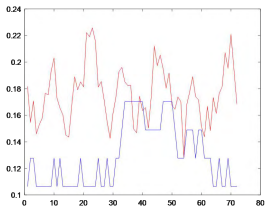
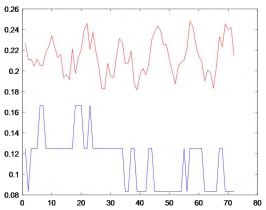
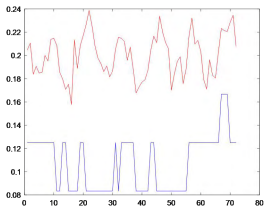
附录 A 图表

气象条件	污染物浓度		
温度	 SO_2	 NO_2	 PM10
	 PM2.5	 O_3	 CO
湿度	 SO_2	 NO_2	 PM10
	 PM2.5	 O_3	 CO
气压	 SO_2	 NO_2	 PM10
	 PM2.5	 O_3	 CO

风速	 SO_2	 NO_2	 PM10
	 PM2.5	 O_3	 CO
风向	 SO_2	 NO_2	 PM10
	 PM2.5	 O_3	 CO

	温度	湿度	气压	风速	风向
AQI 指标					

7 月 8 日至 10 日各污染物浓度预测与实际对比图

	7 月 8 日	7 月 9 日	7 月 10 日
SO_2			
NO_2			
PM 10			
PM 2.5			
O_3			
CO			

附录 B 程序

问题 1：

```
clear,clc
%导入 8.25-8.28 分小时各浓度
fujian1=xlsread('附件 1  监测点 A 空气质量预报基础数据.xlsx','监测点 A 逐日污染物浓度实测数据','C499:H502');
judge=xlsread('质量分指数表.xlsx');
%计算某一变量某一天 AQHI
rowb=size(fujian1,1);
for i=1:rowb
    for ii=1:6
        cp=fujian1(i,ii);
        HI=find(cp<=judge(ii,:),1);
        lo=find(cp>=judge(ii,:));
        LO=lo(end);
        BPHI=judge(ii,HI); BPLO=judge(ii,LO);
        IAQIHI=judge(7,HI);IAQILO=judge(7,LO);
        IAQI(i,ii)=((IAQIHI-IAQILO)/(BPHI-BPLO))*(cp-BPLO)+IAQILO;
        IAQI=ceil(IAQI);
        if ii==5&&cp>800
            disp('第' ,num2str(i), '天不再进行其空气质量分指数计算。');
        end
    end
    if max(IAQI(i,:))>=500
        disp('第', num2str(i) ,'天不再进行其空气质量分指数计算。');
    end
    AQI(i)=max(IAQI(i,:));
    max_AQI(i)=find(IAQI(i,:)==max(IAQI(i,:)));
    disp(['第',num2str(i),'天第',num2str(max_AQI(i)),'项为主要污染源']);
end
```

问题 2：

```
clear,clc
A=xlsread('附件 1 监测点 A 空气质量预报基础数据.xlsx','监测点 A 逐小时污染物浓度与气象实测数据','C2:M19433');
%% 缺失比例分析(NaN 与负数)
[row,col]=size(A);
for i=1:col
    a=0;
    for ii=1:row
    if isnan(A(ii,i))
        a=a+1;
    elseif A(ii,i)<0
        a=a+1;
    end
    end
    aa(i)=a;
    alfa(i)=a/row;
    disp(['第',num2str(i),'个元素得缺失比率为',num2str(alfa(i))])
end
disp('**************************')
%% 填补缺失值
[i1,j1]=find(A<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);
A(i2,j2)=NaN;
end
for i=1:col
    A1(:,i)= fillmissing(A(:,i),'movmean',5);
    A2(:,i)= fillmissing(A1(:,i),'movmean',15);
end
A3=A2;
% 异常值处理 3 方差准则
mean_A=mean(A3);var_A=std(A3);
A4=zeros(row,col);num=0;A5=A3;
for i=1:col
    for ii=1:row
        DA=abs(A3(ii,i)-mean_A(i));
        if DA>3*var_A(i)
            A4(ii,i)=A3(ii,i);
            num=num+1;
            A5(ii,i)=mean_A(i);
        end
    end
end
%对有害气体 0 值进行处理：加一个无穷小数
```

```
[i1,j1]=find(A5(:,1:6)==0);
A5(i1,j1)=eps+A5(i1,j1);
%% 数据归一化
max_A=max(A5);min_A=min(A5);max_A=repmat(max_A,row,1);min_A=repmat(min_A,row,1);
A61=(A5-min_A)./(max_A-min_A);
A6=[A5(:,1:6),A61(:,7:11)];

%% 对附件 1 表 1 进行相关性分析
%首先计算 AQI
judge=xlsread('质量分指数表.xlsx');
max_AQI=zeros(row,6);
%计算某一变量某一天 AQHI
for i=1:row
    for ii=1:6
        cp=A6(i,ii);
        HI=find(cp<=judge(ii,:),1);
        LO=HI-1;
%           lo=find(cp>=judge(ii,:));
%           LO=lo(end);
        BPHI=judge(ii,HI); BPLO=judge(ii,LO);
        IAQIHI=judge(7,HI);IAQILO=judge(7,LO);
        IAQI(i,ii)=((IAQIHI-IAQILO)/(BPHI-BPLO))*(cp-BPLO)+IAQILO;
        IAQI=ceil(IAQI);
        if ii==5&&cp>800
            disp(['第' ,num2str(i), '天不再进行其空气质量分指数计算。']);
        end
    end
    if max(IAQI(i,:))>=500'
        disp(['第', num2str(i) ,'天不再进行其空气质量分指数计算。']);
    end
    AQI(i)=max(IAQI(i,:));
    max_AQI1=find(IAQI(i,:)==max(IAQI(i,:)));
    for iii=1:length(max_AQI1)
        max_AQI(i,iii)=max_AQI1(iii);
    end
%       disp(['第',num2str(i),'天第',num2str(max_AQI(i)),'项为主要污染源']);
end
%% AQI 中主要影响因素分析
aaa=unique(max_AQI(:,1));
for i=1:6
    [i1,j1]=find(max_AQI==i);[i2,j2]=find(max_AQI~=0);
    bizhi(i)=length(i1)/length(i2);
    disp(['第',num2str(i),'个污染气体作为主要污染源的占比为',num2str(bizhi(i))]);
    disp('**********************************************************')
```

```
end
%% 相关性分析 散点图
%由图形判断是否有相关性，如果有相关性则查看 xg2 矩阵中的相关系数
X_A=A5(:,7:11);Y_A=A5(:,1:6);Y_A=[Y_A,AQI'];
N_X=size(X_A,2);N_Y=size(Y_A,2);
j=1;j1=1;
for i=1:N_Y
% j=1;
% figure(j1)
for ii=1:N_X
% subplot(1,5,j);
figure(j)
plot(X_A(:,ii),Y_A(:,i),'o')
% hold on
% box off
xg=cov(X_A(:,ii),Y_A(:,i));xg1=xg(1,2);
xg2(ii,i)=xg1;
j=j+1;
end
j1=1+j1;
end
%% 灰色关联度分析各变量的权重
Mean = mean(A5);  % 求出每一列的均值以供后续的数据预处理
A5 = A5 ./ repmat(Mean,size(A5,1),1);  %size(gdp,1)=6, repmat(Mean,6,1)可以将矩阵进行复制，复制为和 gdp 同等大小，然后使用点除（对应元素相除），这些在第一讲层次分析法都讲过
disp('预处理后的矩阵为：'); disp(A5)
Y = A5(:,1:6);  % 母序列
X = X_A; % 子序列
for i=1:7
    Y=Y_A(:,i);
    absX0_Xi = abs(X - repmat(Y,1,size(X,2)));  % 计算|X0-Xi|矩阵(在这里我们把 X0 定义为了 Y)
    a = min(min(absX0_Xi));      % 计算两级最小差 a
    b = max(max(absX0_Xi));  % 计算两级最大差 b
    rho = 0.5; % 分辨系数取 0.5
    gamma(:,:,i) = (a+rho*b) ./ (absX0_Xi   + rho*b);  % 计算子序列中各个指标与母序列的关联系数
    disp(['子序列对第',num2str(i),'个指标的灰色关联度分别为：'])
    disp(mean(gamma(:,:,i)))
    disp('**********************************************************')

end
%% stata 软件进行标准回归

%% spss 软件聚类
```

问题 3：

```
%数据预处理
clear,clc
A1=xlsread('附件 1.xlsx','监测点 A 逐小时污染物浓度与气象实测数据','C2:M19479');
A2=xlsread('附件 1.xlsx','监测点 A 逐小时污染物浓度与气象一次预报数据','D2:X25417');
B1=xlsread('附件 2.xlsx','监测点 B 逐小时污染物浓度与气象实测数据','C2:M19599');
B2=xlsread('附件 2.xlsx','监测点 B 逐小时污染物浓度与气象一次预报数据','D2:X25345');
C1=xlsread('附件 2.xlsx','监测点 C 逐小时污染物浓度与气象实测数据','C2:M19565');
C2=xlsread('附件 2.xlsx','监测点 C 逐小时污染物浓度与气象一次预报数据','D2:X25345');
C1(2280:2288,6)=NaN;
%% 数据预处理
%缺失值 最后 C12 中仍有大量缺失值，直接剔除
[rowA1,colA1]=size(A1);[rowA2,colA2]=size(A2);
[rowB1,colB1]=size(B1);[rowB2,colB2]=size(B2);
[rowC1,colC1]=size(C1);[rowC2,colC2]=size(C2);
%浓度为负值时，先处理为缺失值在统一处理
[i1,j1]=find(A1(:,1:6)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);A1(i2,j2)=NaN;
end
[i1,j1]=find(B1(:,1:6)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);B1(i2,j2)=NaN;
end
[i1,j1]=find(C1(:,1:6)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);C1(i2,j2)=NaN;
end
[i1,j1]=find(A2(:,end-5:end)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);A2(i2,j2)=NaN;
end
[i1,j1]=find(B2(:,end-5:end)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);B2(i2,j2)=NaN;
end
[i1,j1]=find(C2(:,end-5:end)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);C2(i2,j2)=NaN;
end
for i=1:colA1
    A11(:,i)= fillmissing(A1(:,i),'nearest'); A12(:,i)= fillmissing(A11(:,i),'nearest');
    B11(:,i)= fillmissing(B1(:,i),'nearest'); B12(:,i)= fillmissing(B11(:,i),'nearest');
    C11(:,i)= fillmissing(C1(:,i),'nearest'); C12(:,i)= fillmissing(C11(:,i),'nearest');
```

```
end
for i=1:colA2
    A21(:,i)= fillmissing(A2(:,i),'nearest'); A22(:,i)= fillmissing(A21(:,i),'nearest');
    B21(:,i)= fillmissing(B2(:,i),'nearest'); B22(:,i)= fillmissing(B21(:,i),'nearest');
    C21(:,i)= fillmissing(C2(:,i),'nearest'); C22(:,i)= fillmissing(C21(:,i),'nearest');
end
%0 值处理
[i1,j1]=find(A12(:,1:6)==0);A12(i1,j1)=eps+A12(i1,j1);[i1,j1]=find(B12(:,1:6)==0);B12(i1,j1)=eps+B12(i1,j1);[i1,j1]=find(C12(:
,1:6)==0);C12(i1,j1)=eps+C12(i1,j1);
[i1,j1]=find(A22(:,end-5:end)==0);A22(i1,j1)=eps+A22(i1,j1);[i1,j1]=find(B22(:,end-5:end)==0);B22(i1,j1)=eps+B22(i1,j1);[i1,
j1]=find(C22(:,end-5:end)==0);C12(21,j1)=eps+C22(i1,j1);

%% 3$准则——异常值处理
a22=mean(A22);a12=mean(A12);b22=mean(B22);b12=mean(B12);c22=mean(C22);c12=mean(C12);
sa22=std(A22);sa12=std(A12);sb22=std(B22);sb12=std(B12);sc22=std(C22);sc12=std(C12);
% A 点处理
for i=1:19478
    for j=1:11
ma12(i,j)=abs(A12(i,j)-a12(1,j));
 if ma12(i,j)>=3*sa12(1,j)
        A12(i,j)=a12(1,j);
 end
    end
end
% B 点处理
for i=1:19598
    for j=1:11
mb12(i,j)=abs(B12(i,j)-b12(1,j));
 if mb12(i,j)>=3*sb12(1,j)
        B12(i,j)=b12(1,j);
 end
    end
end
% C 点处理
for i=1:19564
    for j=1:11
mc12(i,j)=abs(C12(i,j)-c12(1,j));
 if mc12(i,j)>=3*sc12(1,j)
        C12(i,j)=c12(1,j);
 end
    end
end
for i=1:25416
    for j=1:21
```

```
ma22(i,j)=abs(A22(i,j)-a22(1,j));
 if ma22(i,j)>=3*sa22(1,j)
        A22(i,j)=a22(1,j);
 end
     end
end
for i=1:25344
     for j=1:21
mb22(i,j)=abs(B22(i,j)-b22(1,j));mc22(i,j)=abs(C22(i,j)-c22(1,j));
 if mb22(i,j)>=3*sb22(1,j)
        B22(i,j)=b22(1,j);
 elseif mc22(i,j)>=3*sc22(1,j)
        C22(i,j)=c22(1,j);
 end
     end
end

%检验预处理结果
[i1,j1]=find(A12(:,1:6)<=0);
num=length(i1);
if num
     disp('预处理失败')
end
[i1,j1]=find(B12(:,1:6)<=0);
num=length(i1);
if num
     disp('预处理失败')
end
[i1,j1]=find(C12(:,1:6)<=0);
num=length(i1);
if num
     disp('预处理失败')
end
[i1,j1]=find(A22(:,end-5:end)<=0);
num=length(i1);
if num
     disp('预处理失败')
end
[i1,j1]=find(B22(:,end-5:end)<=0);
num=length(i1);
if num
     disp('预处理失败')
end
[i1,j1]=find(C22(:,end-5:end)<=0);
```

```
num=length(i1);
if num
    disp('预处理失败')
end
D=sum(sum(isnan(A12)));
D=sum(sum(isnan(A12)))+sum(sum(isnan(B12)))+sum(sum(isnan(C12)))+sum(sum(isnan(A22)))+sum(sum(isnan(B22)))+sum(
sum(isnan(C22)))

%数据归一化
clear,clc
close all
load('shuju1.mat')%C1 中 PM10 由大量空缺 12676-12802
A1=A12;A2=A22;B1=B12;B2=B22;C1=C12;C2=C22;
%对 A
maxA1=max(A1);minA1=min(A1);maxA2=max(A2);minA2=min(A2);
A1=(A1-minA1)./(maxA1-minA1);A2=(A2-minA2)./(maxA2-minA2);
A11=A1;A11(19323:end,:)=[];A11(1:11042,:)=[];%从开始预测的时间开始
A21=A2(1:24624,:);
hour1=size(A11,1);day1=hour1/24;
hour2=size(A21,1);day2=hour2/24;

%对 B 进行归一化
maxB1=max(B1);minB1=min(B1);maxB2=max(B2);minB2=min(B2);
B1=(B1-minB1)./(maxB1-minB1);B2=(B2-minB2)./(maxB2-minB2);

B11=B1;B11(19560:end,:)=[];B11(1:11087,:)=[];%从 2020.7.23 开始到 2021 年 7.10
B21=B2(1:25056,:);
hour1=size(B11,1);day1=hour1/24;
hour2=size(B21,1);day2=hour2/24;
%对 C 进行归一化
maxC1=max(C1);minC1=min(C1);maxC2=max(C2);minC2=min(C2);
C1=(C1-minC1)./(maxC1-minC1);C2=(C2-minC2)./(maxC2-minC2);
C11=C1;C11(19557:end,:)=[];C11(1:11060,:)=[];%从 2020.7.23 开始到 2021 年 7.9
C21=C2(1:25056,:);
hour1=size(C11,1);day1=hour1/24;
hour2=size(C21,1);day2=hour2/24;

%   BP 神经网络训练
clear,clc
close all
load('shuju01.mat')%归一化后的结果
A1=A11;A2=A21;B1=B11;B2=B21;C1=C11;C2=C21;
load('nihexishu1.mat')%最后一项为常数项
%1 代表实测，2 代表一次预报值
```

```
iy=5;                                %修改因变量改气体
rela=A1(:,iy);prea=A2;
relb=B1(:,iy);preb=B2;
relc=C1(:,iy);prec=C2;
j=1;
for i=1:(size(rela,1)/24)-3
for ii=1:72
    YA(j,1)=rela(ii+24*(i-1));j=j+1;
end
end
k=1;l=1;
for i=1:(size(preb,1)/24)/3
for ii=1:72
    YB(k,1)=relb(ii+24*(i-1));k=k+1;
    YC(l,1)=relc(ii+24*(i-1));l=l+1;
end
end
rel1=[YA;YB;YC];
pre1=[prea;preb;prec];
pre=pre1*xishu(1:21,iy)+xishu(22,iy);
%% 生成 XY 变量，##*72
j=1;
for i=1:(size(rel1,1)/72)
for ii=1:72
    X(i,ii)=pre(j);
    Y(i,ii)=rel1(j);j=j+1;
end
end
%% 神经网络工具箱 自变量是 X，因变量是 Y

%% 拟合结果分析
t=1:size(pre,1);
plot(t,pre,'r',t,rel1,'b')
```

问题 4：

```
clear,clc
A01=xlsread('附件 1.xlsx','监测点 A 逐小时污染物浓度与气象实测数据','C2:H19479');
A02=xlsread('附件 1.xlsx','监测点 A 逐小时污染物浓度与气象一次预报数据','D2:X25417');
A11=xlsread('附件 3.xlsx','监测点 A1 逐小时污染物浓度与气象实测数据','C2:H19665');
A12=xlsread('附件 3.xlsx','监测点 A1 逐小时污染物浓度与气象一次预报数据','D2:X25417');
A21=xlsread('附件 3.xlsx','监测点 A2 逐小时污染物浓度与气象实测数据','C2:H19662');
A22=xlsread('附件 3.xlsx','监测点 A2 逐小时污染物浓度与气象一次预报数据','D2:X25417');
A31=xlsread('附件 3.xlsx','监测点 A3 逐小时污染物浓度与气象实测数据','C2:H19266');
A32=xlsread('附件 3.xlsx','监测点 A3 逐小时污染物浓度与气象一次预报数据','D2:X25417');
%% 数据预处理
%缺失值 最后 C12 中仍有大量缺失值，直接剔除
[rowA01,colA01]=size(A01);[rowA02,colA02]=size(A02);
[rowA11,colA11]=size(A11);[rowA12,colA12]=size(A12);
[rowA21,colC1]=size(A21);[rowA22,colA22]=size(A22);
%浓度为负值时，先处理为缺失值在统一处理
[i1,j1]=find(A01(:,1:6)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);A01(i2,j2)=NaN;
end
[i1,j1]=find(A11(:,1:6)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);A11(i2,j2)=NaN;
end
[i1,j1]=find(A21(:,1:6)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);A21(i2,j2)=NaN;
end
[i1,j1]=find(A31(:,1:6)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);A31(i2,j2)=NaN;
end
[i1,j1]=find(A02(:,end-5:end)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);A02(i2,j2)=NaN;
end
[i1,j1]=find(A12(:,end-5:end)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);A12(i2,j2)=NaN;
end
[i1,j1]=find(A22(:,end-5:end)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);A22(i2,j2)=NaN;
end
```

```
[i1,j1]=find(A32(:,end-5:end)<0);
for i=1:length(i1)
i2=i1(i);j2=j1(i);A32(i2,j2)=NaN;
end
for i=1:6
    A11(:,i)= fillmissing(A11(:,i),'movmean',5); A11(:,i)= fillmissing(A11(:,i),'movmean',15);A11(:,i)=
fillmissing(A11(:,i),'movmean',100);
    A01(:,i)= fillmissing(A01(:,i),'movmean',5); A01(:,i)= fillmissing(A01(:,i),'movmean',15);A01(:,i)=
fillmissing(A01(:,i),'movmean',110);
    A21(:,i)= fillmissing(A21(:,i),'movmean',5); A21(:,i)= fillmissing(A21(:,i),'movmean',15);A21(:,i)=
fillmissing(A21(:,i),'movmean',200);
    A31(:,i)= fillmissing(A31(:,i),'movmean',5); A31(:,i)= fillmissing(A31(:,i),'movmean',15);A31(:,i)=
fillmissing(A31(:,i),'movmean',200);
end
junzhi=mean(A01);var=std(A01);[i,j]=find(abs(A01-junzhi)>3*var);
dd=length(j);
for ii=1:dd
A01(i(ii),j(ii))=junzhi(j(ii));
end
junzhi=mean(A02);var=std(A02);[i,j]=find(abs(A02-junzhi)>3*var);
dd=length(j);
for ii=1:dd
A02(i(ii),j(ii))=junzhi(j(ii));
end
junzhi=mean(A11);var=std(A11);[i,j]=find(abs(A11-junzhi)>3*var);
dd=length(j);
for ii=1:dd
A11(i(ii),j(ii))=junzhi(j(ii));
end
junzhi=mean(A12);var=std(A12);[i,j]=find(abs(A12-junzhi)>3*var);
dd=length(j);
for ii=1:dd
A12(i(ii),j(ii))=junzhi(j(ii));
end
junzhi=mean(A21);var=std(A21);[i,j]=find(abs(A21-junzhi)>3*var);
dd=length(j);
for ii=1:dd
A21(i(ii),j(ii))=junzhi(j(ii));
end
junzhi=mean(A22);var=std(A22);[i,j]=find(abs(A22-junzhi)>3*var);
dd=length(j);
for ii=1:dd
A22(i(ii),j(ii))=junzhi(j(ii));
end
```

```
junzhi=mean(A31);var=std(A31);[i,j]=find(abs(A31-junzhi)>3*var);
dd=length(j);
for ii=1:dd
A31(i(ii),j(ii))=junzhi(j(ii));
end
junzhi=mean(A32);var=std(A32);[i,j]=find(abs(A32-junzhi)>3*var);
dd=length(j);
for ii=1:dd
A32(i(ii),j(ii))=junzhi(j(ii));
end
save('shuju09.mat','A01','A11','A21','A02','A12','A22','A32');
max01=max(A01);max11=max(A11);max21=max(A21);max31=max(A31);
max02=max(A02);max12=max(A12);max22=max(A22);max32=max(A32);
min01=min(A01);min11=min(A11);min21=min(A21);min31=min(A31);
min02=min(A02);min12=min(A12);min22=min(A22);min32=min(A32);

A01=(A01-min(A01))./(max(A01)-min(A01));
A11=(A11-min(A11))./(max(A11)-min(A11));
A21=(A21-min(A21))./(max(A21)-min(A21));
A31=(A31-min(A31))./(max(A31)-min(A31));

A02=(A02-min(A02))./(max(A02)-min(A02));
A12=(A12-min(A12))./(max(A12)-min(A12));
A22=(A22-min(A22))./(max(A22)-min(A22));
A32=(A32-min(A32))./(max(A32)-min(A32));
save('shuju01.mat','A01','A11','A21','A02','A12','A22','A32');
A01(19323:end,:)=[];A01(1:11139,:)=[];A02(24481:end,:)=[];A02(1:288,:)=[];
A11(19417:end,:)=[];A11(1:11233,:)=[];A12(24481:end,:)=[];A12(1:288,:)=[];
A21(19414:end,:)=[];A21(1:11230,:)=[];A22(24481:end,:)=[];A22(1:288,:)=[];
A31(19018:end,:)=[];A31(1:10834,:)=[];A32(24481:end,:)=[];A32(1:288,:)=[];
%% iy 选择气体 Ai 选择监测点
iy=5;
Ai=1;
%% kriging 方法 3 点生成 1 点的插值数据
%生成数据,实测和预测的维度保持一致
shice=[A01(:,iy),A11(:,iy),A21(:,iy),A31(:,iy)];
j=1;
for i=1:(size(shice,1)/24)-3
for ii=1:72
    Yshice(j,1)=shice(ii+24*(i-1),Ai);j=j+1;
end
end
Yshice(24193:end,:)=[];
%选择预测点并计算
```

```
yuce1=[A02(:,iy+15),A12(:,iy+15),A22(:,iy+15),A32(:,iy+15),];
for i=1:4
yuce=[A02(:,iy+15),A12(:,iy+15),A22(:,iy+15),A32(:,iy+15),]';
xy=[0 0;-14.4846 -1.9699;-6.6716 7.5953;-3.3543 -5.0138];
xy=(xy-min(xy))./(max(xy)-min(xy));
% for ii=1:4
% xy(ii,:)=norm(xy(ii,:));
% end
% xy=xy(:,1);
% A (0, 0)       A1 (-14.4846, -1.9699)       A2 (-6.6716, 7.5953)       A3 (-3.3543, -5.0138)
h1=xy(i,:);
xy(i,:)=[];
yuce(i,:)=[];
% for ii=1:size(yuce,2)
% lb=1e-2;ub=10000;theta1=1;
lb=[1e-2,1e-2];ub=[10000,10000];theta1=[1,1];
% [dmodel1, perf1] = dacefit(xy,yuce(:,ii), @regpoly0, @corrgauss,theta1,lb,ub);
[dmodel1, perf1] = dacefit(xy,yuce, @regpoly0, @corrgauss,theta1,lb,ub);
theta(i,:)=dmodel1.theta;
[YX1,MSE1] = predictor(h1, dmodel1);
pre(:,i)=YX1;
end
% end
%% 求灰色关联度
for i=1:4
yhui=Yshice;
xhui=[pre(:,i),yuce1(:,i)];
absX0_Xi = abs(xhui - repmat(yhui,1,size(xhui,2)));  % 计算|X0-Xi|矩阵(在这里我们把 X0 定义为了 Y)
a = min(min(absX0_Xi));      % 计算两级最小差 a
b = max(max(absX0_Xi));  % 计算两级最大差 b
rho = 0.5; % 分辨系数取 0.5
gamma = (a+rho*b) ./ (absX0_Xi    + rho*b);  % 计算各个指标与母序列的关联系数
disp(mean(gamma));
w(:,i)=mean(gamma);%列分别表示求解 A,A1,A2,A3 时的权重
end
w1=w;
for i=1:8
      shuchu(i,1)=w1(i)
end
%% 生成 XY 修改 Ai 可以分别求解 A,A1,A2,A3
xhui=[pre(:,Ai),yuce1(:,Ai)];
pre=xhui*w(:,Ai);
rel=Yshice;
j=1;
```

```
for i=1:(size(rel,1)/72)
for ii=1:72
    X(i,ii)=pre(j);
    Y(i,ii)=rel(j);j=j+1;
end
end
```