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	浙江工商大学
参赛队号	21103530001
队员姓名	**1.黄诗婷**
	2.傅苏婷
	3.朱静波

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　　二次优化的 WRF-CMAQ 空气质量预报模型

摘　　要：

当前，空气质量对人类生存环境、人体健康的危害衍生为不可回避的话题。开展空气质量监测、预测、数据分析及可视化的研究可以全面把控城市空气污染源的排放数据及不同空间区域内的浓度数据，掌握空气质量在时间和空间维的变化发展趋势，对城市规划与建设、污染控制、环境管理、公共事业发展等意义重大。基于此，本文为进一步**优化**依据大气污染防治体制机制建立的 WRF-CMAQ 预报模型， 引入空气质量监测点获得的实际气象条件和污染物浓度实测数据进行空气质量预报模型修正。

本文首先对各监测点空气质量预报数据进行数据初处理，避免因数据缺失、异常、重复或出现野值等造成研究误差。具体流程：①考虑到污染物数据及气象数据具有较强的时序性，因此本文运用均值法进行插值补全。**特别处理：补全监测点逐小时**污染物浓度与气象数据中的**缺失值，**用均值插补和**趋势外推**结合进行预测；依据监测点 A 与邻近监测点 A1、A2 和 A3 的**相对地理位置**，对附件 3 中“监测点 A1 逐小时污染物浓度与气象实测数据”中**缺失的“气压(MBar)”指标**进行全列补充；针对附件 3 中“监测点 A3 逐小时污染物浓度与气象实测数据” 2020/12/31 23:00- 2021/4/23 15:00 的五个气象指标（温度、湿度、气压、近地风速、风向）大段连续缺失值，依据气象指标和相对地理位置的关系，使用临近点 A1 和 A2 的气象指标进行填补。②根据拉依达准则（3 σ 准则）去除非操作变量异常值；③采用最值归一化算法保证数据在同一尺度下。

针对问题一：本文依据《环境空气质量指数（AQI）技术规定（试行）》（HJ633-2012）的方法，计算得到监测点 A 从 2020 年 8 月 25 日到 8 月 28 日每天实测的 AQI 和首要污染物**结果**如下：8 月 25 日 AQI=60，首要污染物为臭氧 O_3；8 月 26 日 AQI=46，首要污染物无；8 月 27 日 AQI=109，首要污染物为臭氧 O_3；8 月 28 日 AQI=138，首要污染物为臭氧 O_3。

针对问题二：要求在污染排放清单不变的情况下，利用气象条件对大气污染物浓度的影响程度，对气象条件进行分类。为此，基于监测点实测数据和一次模型预报数据，寻求气象因子及污染物浓度的变化规律，提出**周期循环**概念。在此基础上利用 Pearson 相关系数探究两者之间的相关性。进一步地，在科学对比 K-means、AP、GMM、凝聚层次聚类后，选取 K-means 进行无监督聚类，将气象条件聚成 8 类。为进一步优化分类，选用泛化性强、对噪声不敏感的随机森林分类算法将类别调整为 6 类，删除两个错误率较高的类别。同时为提高二次预测模型的准确度及增强类别的可解释性，选用**熵权法**对气象因子的重要程度进行排序，最终将气象条件分为 6 类进行特征总结。

针对问题三：要求建立一个适用于多个点的二次预报模型，预测未来三天的 6 种常规污染物浓度值。为此，基于前述处理数据集，综合考虑一次模型预测数据、监测站实际监测气象数据、污染物浓度及**臭氧等污染物的生成机理**，结合问题二的分类结果及**信息增益**

特征筛选结果，筛选出特征显著的变量，包括**时序因子、气候因子和历史数据因子**。将上述因子作为输入变量，选用 RMSE、MAE 和 IOA 作为评价指标，对比原始 CMAQ 一次预报模型、SVR 预测模型、Elman 预测模型和 DBN-BP 预测模型的预报效果，最终建立**基于 Elman 神经网络的优化算法**进行空气质量二次优化，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，并依据问题一的计算方法得到污染物浓度及 AQI 预测结果表。

针对问题四：要求建立**邻近区域的协同预报模型**，因为相邻区域间污染物浓度具有一定关联性。首先基于 **BMA 法**确定各邻近区域监测点的数据的权重，其中利用 EM 算法确定模型的参数。然后，分别对各个气象因子进行气象协同预报，对协同预报成员的结果加权处理。最后，在此数据基础上，加入**风速和风向**两个参数，利用 **wiener** 对预报值进行修正。选用 RMSE、MAE 和 IOA 三个评价指标进行效果评价，结果表明邻近区域协同模型的预测效果优于问题三的空气质量二次预测模型，也就是说邻近区域协同模型**能提升**针对监测点 A 的污染物浓度预报值。

最后对本文的模型进行了评价与推广，可以不断优化模型，提升空气质量预报模型预测精度。

关键词：空气质量二次预报；信息增益特征筛选；Elman 神经网络；Wiener 模型；区域协同预报

目录

二次优化的 WRF-CMAQ 空气质量预报模型

1 问题重述

1.1研究背景

空气质量攸关人民幸福生活、社会长远发展。据 2018 年《中国生态环境状况公报》显示，在全国地级及以上城市中，环境空气质量达标城市占比 35.8%，较上一年增长，但仍未超过半数，空气质量依旧十分严峻。在此形式下，习近平同志在十九大会议中指出要建设美丽中国，为人民创造良好生产生活环境，同时十九大文件起草组成员杨伟民解读十九大报告中污染防治：像对待生命一样对待生态环境。因此，建立科学有效的空气质量预测模型尤为重要。

区域空气质量的预测研究，借助统计科学方法，依托空气质量历史数据，对未来的空气质量信息和变化趋势提供较为精准的预测，有助于洞悉空气污染背后成因、提供健康出行的有效参考、拓宽空气质量了解途径。对于政府部门而言，空气质量预测研究能提前预防可能发生的污染事故，避免严重污染事件的发生或降低污染事件的危害，同时生成的辅助材料为环境相关部门的规划管理措施存在不可或缺的指导意义[1]。

1.2 问题分析

当前常见的空气质量预报模型为 WRF-CMAQ 模拟体系（简称 WRF-CMAQ 模型），其中 WRF 是一种中尺度数值天气预报系统，输入静态地形和初始气象指标，在对数据进行格式转换、水平插值、垂直插值等预处理后，为后续 CMAQ 分模型提供模拟气象；CMAQ 分模型是一种三维欧拉大气化学与传输模拟系统，在 WRF 气象模型基础上，加入区域范围内的污染排放清单，输出污染物浓度的模拟值。

空气质量模型模拟的愿景分析对政府制定保障空气质量的相关政策有重要意义。但是，空气质量模型存在一定缺陷，预测的精度极大程度依赖于污染排放清单和模拟出的气象场精度，甚至如接口模块、初始值模块、化学传输模块等内置模块也会对其产生影响，从而造成木桶原理和蝴蝶效应。也就是说，模型的精度是由最不精准的反应所决定的。此外，污染物生成机理的不明晰性，同样会导致 WRF-CMAQ 模型预测结果的不准确性。

因此，本文的**重点目标**是在 WRF-CMAQ 模型一次预报的基础上，参考空气质量监测点获得的气象因子、地形特征和污染物实测数据对一次预报数据进行修正，以**优化模型**，输出优化后的污染物浓度。

具体包括以下四个问题：

问题一：空气质量指数（AQI）和首要污染物计算

从监测点 A 污染物浓度与气象实测数据中，按照《环境空气质量指数（AQI）技术规定（试行）》（HJ633-2012）中的所述方法计算该点从 2020 年 8 月 25 日到 8 月 28 日每天实测的 AQI 和首要污染物，并呈现计算结果表。

具体步骤：

步骤一：首先对附录 1 数据进行预处理，利用均值法进行缺失值处理、补全监测点 A 逐小时污染物浓度与气象实测数据中的缺失值、3 σ 原则和协同过滤推荐算法进行异常值处理，并对数据进行归一化处理。

步骤二：在上述基础上，根据《环境空气质量指数（AQI）技术规定（试行）》（HJ633-2012）

中的所述方法，利用给出的六种常规污染物的实时浓度，首先计算出污染物的平均浓度，得到空气质量分指数（IAQI）。

步骤三：取空气质量分指数（IAQI）的最大值为 AQI，空气质量分指标数值最大的为首要污染物，同时可对空气质量进行等级划分。

问题二：气象条件分类

地区的气象条件对于大气污染物的产生及流动密切相关[2]，且气象条件的变化对污染物的扩散和输送也有着较大的影响。现要求根据相关污染物浓度数据在相关性分析的基础上，进行无监督聚类（选取效果最好的 K-means），然后利用随机森林进行气象条件分类。接着为提高二次预测模型的准确性，使用熵权法调整气象因子的影响系数后，对监测点 A 的气象特征进行分类，并阐述。最后返回来对应每个类别的气象状况下的污染物指数。

具体步骤：

步骤 1：使用进行数据预处理后的附件 1 中数据进行气象条件及污染物浓度的描述性分析，观察得到“监测点 A 逐小时污染物浓度与气象一次预报数据”表中数据具有很强的周期性，而“监测点 A 逐小时污染物浓度与气象实测数据”中的数据周期性不是很强。

步骤 2：利用数据预处理后的附件 1 数据进行相关性分析，观察各变量间的相关关系，以相关系数表及热力图的形式展现，得到污染物浓度及气象因子变化的规律。

步骤 3：在根据对污染物浓度的影响程度，将气象条件进行合理分类时，考虑到合理的时间间隔，设定每三个周期一循环，得到污染物与气象条件的初始变量值及循环周期内变量变化值共 22 个指标，运用聚类分析模型进行无监督分类。由于现有聚类算法的广泛性，我们选取了当下较为热门的四种聚类方法：K-means 聚类、AP（Affinity Propagation）聚类、高斯混合模型（GMM）、凝聚层次聚类（HAC），对数据分别进行建模分析，科学对比四种模型效果并选取最优结果进行后续分析。

步骤 4：Breiman 提出的随机森林算法具有泛化性强、稳健性、对噪声不敏感、能处理连续属性等特点，并且被广泛运用于分类模型中。因此在得到根据对污染物浓度的影响程度，将气象条件进行分类结果后，我们选用随机森林算法构建模型，对类别变量进行验证调整。

步骤 5：分析以上结果，为使后续预测模型更准确，本文利用熵权法对影响大气污染物浓度的各气象条件进行权重计算，依据各气象因子的权重系数，对上述分类进行调整。

步骤 6：分析上述分类结果，具体说明各类气象条件的特征。

步骤 7：根据得到的气象类别情况，返回得到每个类别的气象分类下的污染指标范围。

问题三：建立具有适用性的空气质量二次预报模型

建立一个同时适用于 A、B、C 监测点的二次预报模型。首先，本文选取的是经过题二中预处理的，一次模型预测和监测站实际监测的数据。然后根据问题二中的相关和聚类结果，筛选出特征显著的变量，分别为时序因子、气候因子和历史数据因子。最后对数据使用 3 种不同的模型进行预测，对比 RMSE、MAE、IOA 值后选择了 Elman 模型，并根据公式计算出最终的结果。

具体步骤：

步骤 1：本文使用的是经过题二中预处理的，一次模型预测和监测站实际监测的数据。首先将一次模型预测结果与监测站实际监测结果进行对比分析，发现一次模型结果在一定程度上是反应了实际的空气质量的，但是模拟的并不是很好，还可以进一步的去优化。

步骤 2：问题二中给出了数据之间的相关性和气候的聚类结果，本文据此数据，提取出了 6 项污染物的特征显著变量，主要分为时序因子、气候因子和历史数据因子。污染物

浓度数据会随着小时的变动呈现规律变动，本文以 6 个小时为一周期，将 24 个小时划分为 4 个周期，分别为 7：00-12：00，12：00-19：00，19：00-24：00，0：00-7：00。气候因子中归纳出影响力较大的因子，其中湿度>风向>气压>温度>风速。污染物之间也是存在相互影响的，因此，本文还新增了过去 24 个小时的污染物实际监测数据。CMAQ 一次模型中，O3 的预测偏差较大，对比，本文创新性的新增了 NO2 和 CO 作为模型输入变量，并根据问题二的结果对其赋予权重，进行修正。

步骤 3：最后，本文利用了三种不同的模型，分别为 SVR 模型、DBN-BP 模型、Elman 模型，分别对 6 种不同的污染物浓度进行预测，并综合对比模拟结果，结果发现这三种模型相较于 CMAQ 一次模型，对污染物浓度的预测都有了一定的提升，其中 Elman 模型预测的最为准确，其 RMSE 值、MAE 值也是最小的，因此最终对 13-15 号的预测选用了 Elman 模型的预测结果，并使用题干中给出的公式，计算出 6 种污染物的单日浓度值和 AQI 结果。

问题四：建立区域协同预报模型

监测点与其附近区域的污染物浓度具有一定的关联性，为此建立一个邻近区域协同预报模型，有可能会提高空气质量预报的准确性。这其中需要综合考虑邻近区域监测点的地理位置、气象条件、实测污染物浓度等，为此建立了基于 BMA 的的空气质量邻近区域协同预报模型。

具体步骤：

步骤 1：本文首先对数据进行了描述性分析，探索 A、A1、A2、A3 四个临近区域的气象因子、污染浓度值的相关关系，结果发现四处的气象因子存在一定的相似性，污染物浓度变化存在相似的波动。

步骤 2：将 A、A1、A2、A3 的数据输入到题目三的模型得到输出结果，然后利用贝叶斯模型平均法（BMA）和 EM 算法，协同预报气象因子，并对不同的监测点赋予权重，加权处理题目三的输出结果。

步骤 3：对于区域相邻的监测站而言，风速和风向因子的重要性不言而喻。本文在此，使用了 wiener 模型，总结了风速、风向和实际值之间的非线性关系，并对结果进行修正，利用公式法计算出对应的 AQI 的值。最后，对比了题目三和题目四得出的 A 的结果值，并与实际监测结果对比，发现区域协同模型的 RMSE、MAE 值都更加小，表明邻近区域协同模型的预测结果比问题 3 的模型结果更为精确。

1.3 本文的研究思路

本文基于有关气象条件及主要大气污染物的一次预报数据和实测数据，进行了空气质量二次预报，主要解决 WRF-CMAQ 模型预测结果的不准确性问题，全文研究思路图如下图 1-1。

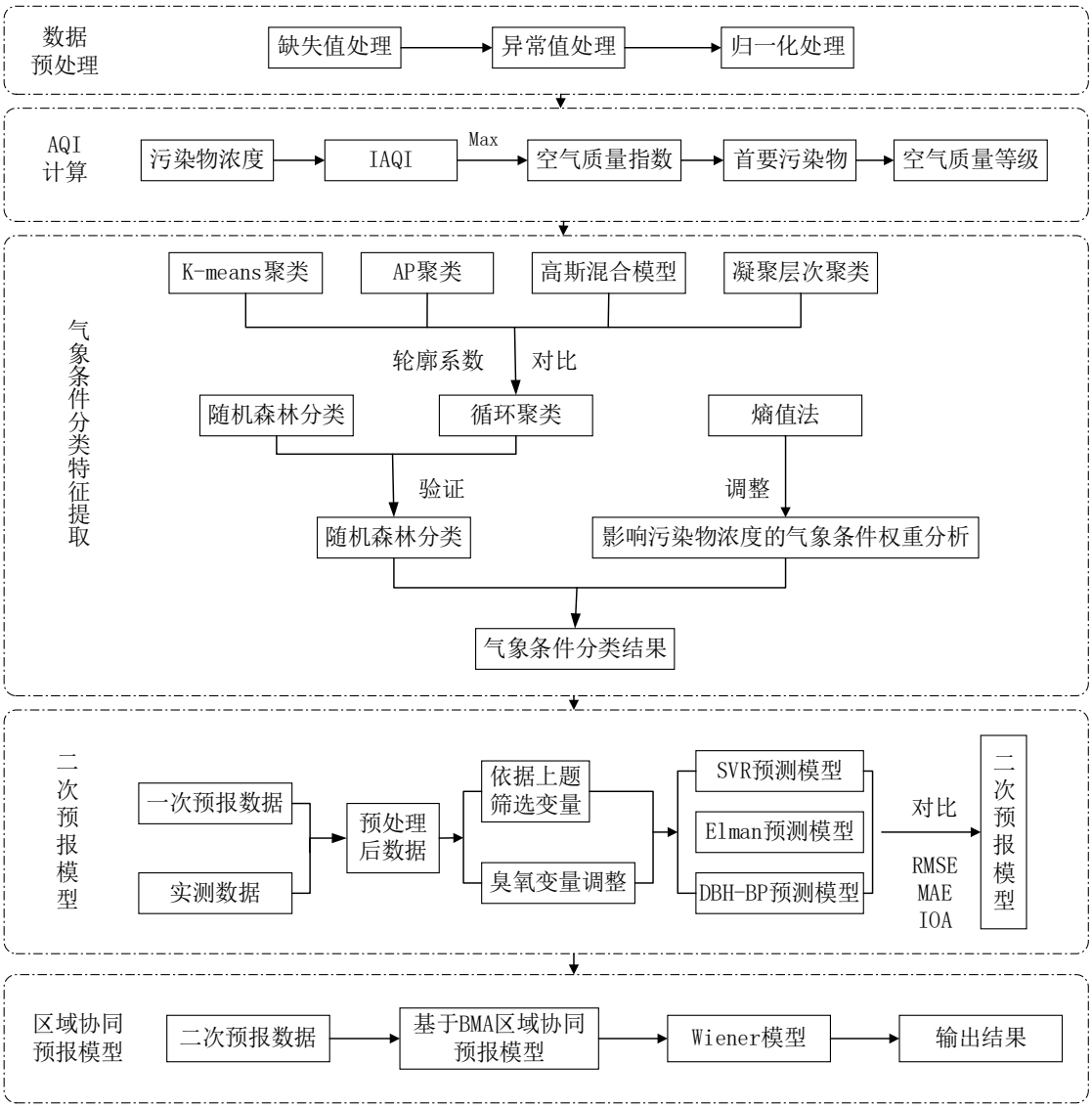


图 1-1 全文研究思路图

2 模型假设与符号设定

为了便于问题的研究，对题目中某些条件进行简化及合理的假设。

2.1模型假设

1. 假设本文的数据真实有效，不存在错误和虚假数据，符合数据统计分析的基本要求，能准确反应污染物浓度和气象条件变化的基本规律。
2. 假设各监测点的污染物排放情况不发生变化。
3. 假设同一气象特征对污染物浓度的影响程度不会发生变化，同时该特征也不会因为时间的变动而变动。
4. 假设模型建立过程中，仅考虑问题中的核心因素，不考虑次要因素的影响。
5. 假设每改变一次主要操作变量的值，输出的污染物浓度都会相应及时的发生变化。
6. 假设一次和二次预报模型中，对邻近日期的准确度都较高。

2.2符号说明

表 2-1　符号说明

变量（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
AQI	空气质量指数

3 数据预处理

3.1数据集描述

为对 WRF-CMAQ 一次预报模型进行优化，以预测给定监测点未来三天的空气质量情况，题目提供的数据集为监测点长期空气质量预报基础数据，包括污染物浓度一次预报数据、气象一次预报数据、气象实测数据和污染物浓度数据，其中所有一次预报数据的时间跨度为 2020-7-23～2021-7-13，所有实测数据的时间跨度为 2019-4-16～2021-7-13。

具体来说，附件数据集所包含的指标如下表 3-1 所示，监测点分别为 A、A1、A2、A3、B 和 C。

表 3-1　附件数据集指标情况

数据集名称	包含指标（单位）	包含指标（单位）
监测点逐小时污染物浓度与气象一次预报数据	模型运行日期 预测时间 地点 近地 2 米温度（℃） 地表温度（K） 比湿（kg/kg） 湿度（%） 近地 10 米风速（m/s） 近地 10 米风向（°） 雨量（mm） 云量 边界层高度（m）	大气压（Kpa） 感热通量（W/m²） 潜热通量（W/m²） 长波辐射（W/m²） 短波辐射（W/m²） 地面太阳能辐射（W/m²） SO2 小时平均浓度(μg/m³) NO2 小时平均浓度(μg/m³) PM10 小时平均浓度(μg/m³) PM2.5 小时平均浓度(μg/m³) O3 小时平均浓度(μg/m³) CO 小时平均浓度(mg/m³)
监测点 A 逐小时污染物浓度与气象实测数据	SO2 监测浓度(μg/m³) NO2 监测浓度(μg/m³) PM10 监测浓度(μg/m³) PM2.5 监测浓度(μg/m³) O3 监测浓度(μg/m³) CO 监测浓度(mg/m³)	温度(℃) 湿度(%) 气压(MBar) 风速(m/s) 风向(°)

监测点 A 逐日污染物浓度实测数据	监测日期 地点 SO2 监测浓度(μ g/m³) NO2 监测浓度(μ g/m³) PM10 监测浓度(μ g/m³)	PM2.5 监测浓度(μ g/m³) O3 最大八小时滑动平均监测浓度(μ g/m³) CO 监测浓度(mg/m³)

3.2 缺失数据和异常数据

构建预报模型时，低质量数据属于训练的噪声数据，会提高模型预测结果的误差率[3]。

3.2.1 缺失值形成原因

污染物浓度数据和气象预报数据是通过服务器获取，实测数据是监测站点采集。服务器在获取预报数据过程中，由于服务器受外接电源长时间停电等情况影响，导致部分运行日期的一次预报数据缺失。而在实测数据的采集过程中，可能会存在以下情况导致数据缺失：

1.偶然因素影响，如因监测站点设备调试、维护等原因，导致实测数据在连续时间内存在部分或全部缺失的情况；

2.监测点不同，使用设备存在差异，可能会导致部分监测点温度、湿度、气压、风向和风速五个指标中的部分缺失。

除此以外，数据传输时，可能由于网络中断，存储空间不足等，造成数据出现错漏。

3.2.2 异常数据形成原因

采集数据过程中，由于传感器故障等原因会导致数据出现异常或突变的情况，或者说监测站点及其附近某些偶然因素影响，数据录入错误等，使实测数据在某天或者某个小时的数据偏离数据正常分布，应将其作为异常值剔除掉。

3.2.3 不良数据处理

结合以上气象指标和污染物浓度指标的具体不良数据情况（见下表 3-2），对原始数据可能存在的异常情况作如下方面的预处理，处理程序如下图 3-1。

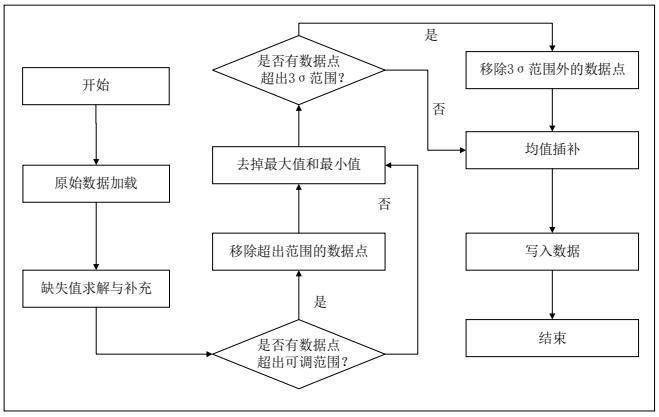


图 3-1 数据预处理流程图

1. 变量值缺失为空值

（1）平均值填补（Mean/Mode Completer）常规缺失值

考虑到大气污染数据及气象数据具有较强的时序性，为非离散数值，不会随时间出现大幅度的变高或变低，变化趋势较缓，于是本文采取**均值法**对常规缺失数据进行插值处理，即通过缺失值前后时间范围内的均值数据进行插值补全。

特别关注：通过观测各监测点逐小时污染物浓度与气象数据，发现**实测数据一共缺失231个小时数据**，例如2019/5/6缺少15:00:00、2020/1/1缺少0:00:00和3:00:00，2021/7/3的0:00:00到2021/7/6 23:00:00连续缺少92个小时数据，具体缺失时点见附录1。

对于单点缺失的小时数据，使用均值插补法。对于连续缺失的小时数据，本文采用前72小时数据进行**趋势外推**预测进行填补。

（2）相对地理位置系数调整连续大段缺失值及指标列缺失

通过观测发现，附录3“监测点A1逐小时污染物浓度与气象实测数据”的“气压(MBar)”**指标数据整列缺失**，“监测点A3逐小时污染物浓度与气象实测数据”的2020/12/31 23:00-2021/4/23 15:00的五个气象指标（温度、湿度、气压、近地风速、风向）**全部缺失**。

对此，本文结合监测点A与监测点A1、A2和A3相对位置关系，结合气压、温度、湿度、近地风速和风向与位置见的关系，对以上缺失数据进行了填补。

2. 变量值超出可调控范围

结合实际情况考虑出现的所有指标取值情况，得出各污染物（CO、NO2、SO2、O_3、PM10、PM2.5）监测**浓度存在负值**为异常情况，需要对其进行处理。处理方法为：先将其填充为缺失值，再进行均值插补。

3. 变量值超出3σ区间范围

3σ原则（又称拉依达准则），具体来说是先假设一组检测数据只含有随机误差，对原始数据进行计算处理得到标准差，然后按一定的概率确定一个区间，认为误差超过这个区间的就属于异常值。具体处理方法为：异常值3σ检验后，进行均值填补。

3.3归一化处理

在对数据进行训练前，需将数据集分为训练集和测试集，对此需保证所有数据在同一尺度下，否则神经网络为适应各个数据特征维度需进行额外的学习，收敛速度会因此下降，延长神经网络的训练时间。例如CO浓度数值明显小于其他浓度，且大约相差两个量级，因此，本文对数据进行归一化处理，也就是将所有数据映射到[0,1]之间，避免上述问题产生，提高算法的学习训练速度，最值归一算法如下公式1所示。

$$y=\frac{x-\min}{\max-\min} \qquad \text{公式 1}$$

其中y表示归一化后的数据，x表示原数据，max/min表示指标中最大/小值。

处理时，正向性指标包括：温度、湿度、风速、雨量、近地10米风速、近地2米温度、地表温度、感热通量、潜热通量、比湿、边界层高度、云量、地面太阳能辐射、短波辐射、长波辐射、；负向性指标包括：六种污染物浓度、气压、风向、大气压、近地10米风向。

表 3-2　数据集缺失及异常情况

附件	数据集	缺失值（个）	浓度负值（个）
附件 1	监测点 A 逐小时污染物浓度与气象一次预报数据	-	-
	监测点 A 逐小时污染物浓度与气象实测数据	1582	267
	监测点 A 逐日污染物浓度实测数据	39	-
附件 2	监测点 B 逐小时污染物浓度与气象一次预报数据	-	-
	监测点 C 逐小时污染物浓度与气象一次预报数据	-	-
	监测点 B 逐小时污染物浓度与气象实测数据	1837	41
	监测点 C 逐小时污染物浓度与气象实测数据	8034	8
	监测点 B 逐日污染物浓度实测数据	26	-
	监测点 C 逐日污染物浓度实测数据	135	
附件 3	监测点 A1 逐小时污染物浓度与气象一次预报数据	-	-
	监测点 A2 逐小时污染物浓度与气象一次预报数据	-	-
	监测点 A3 逐小时污染物浓度与气象一次预报数据	-	-
	监测点 A1 逐小时污染物浓度与气象实测数据	2796	-
	监测点 A2 逐小时污染物浓度与气象实测数据	1042	255
	监测点 A3 逐小时污染物浓度与气象实测数据	14380	127
	监测点 A1 逐日污染物浓度实测数据	8	-
	监测点 A2 逐日污染物浓度实测数据	47	-
	监测点 A3 逐日污染物浓度实测数据	55	-

4 问题一：AQI 及首要污染物计算

4.1 问题提出

问题一要求利用附录中的 AQI 计算方法，即《环境空气质量指数（AQI）技术规定（试行）》（HJ633-2012）中的所述方法，利用给出的六种常规污染物的实时浓度，首先计算出污染物的平均浓度，得到空气质量分指数（IAQI）；紧接着取空气质量分指数（IAQI）的最大值为 AQI，空气质量分指标数值最大的为首要污染物，同时可对空气质量进行等级划分。

使用的数据集：附件 1“监测点 A 逐日污染物浓度实测数据”的六种污染物监测浓度。

4.2 AQI 计算方法

空气质量指数（Air Quality Index，AQI)是定量描述空气质量状况的无量纲指数，AQI 的大小直观反映了空气的污染程度。空气质量分指数（Individual Air Quality Index，IAQI）是单项污染物的空气质量指数，IAQI 的计算公式 2 如下：

$$IAQI_p = \frac{IAQI_{Hi} - IAQI_{LO}}{BP_{Hi} - BP_{LO}}(C_P - BP_{LO}) + IAQI_{LO} \qquad \text{公式 2}$$

其中，各项污染物项目浓度限值及对应的空气质量分指数级别见表 4-1。

表 4-1　空气质量分指数及对应的污染物浓度限值

序号	指数或污染物项目	空气质量分指数及对应污染物浓度限值								单位
0	空气质量分指数（IAQI）	0	50	100	150	200	300	400	500	-
1	一氧化碳（CO）24 小时平均	0	2	4	14	24	36	48	60	mg/m^3
2	二氧化硫（SO_2）24 小时平均	0	50	150	475	800	1600	2100	2620	$\mu g/m^3$
3	二氧化氮（NO_2）24 小时平均	0	40	80	180	280	565	750	940	
4	臭氧（O_3）最大 8 小时滑动平均	0	100	160	215	265	800	-	-	
5	粒径小于等于 10μm颗粒物（PM_{10}）24 小时平均	0	50	150	250	350	420	500	600	
6	粒径小于等于 2.5μm颗粒物（$PM_{2.5}$）24 小时平均	0	35	75	115	150	250	350	500	

注：(1) 臭氧（O3）最大 8 小时滑动平均浓度值高于 800 μg／m^3 的，不再进行其 IAQI 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。

空气质量指数（AQI）取各分指数中的最大值，即

$$AQI = \max\{IAQI_1, IAQI_2, IAQI_3, ..., IAQI_n\} \quad \text{公式 3}$$

在该题中，由于仅涉及表 4-1 中的六种污染物（CO、SO_2、NO_2、O_3、PM_{10}、$PM_{2.5}$），因此 IAQI 计算公式变为：

$$AQI = \max\{IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{PM_{10}}, IAQI_{PM_{2.5}}, IAQI_{O_3}, IAQI_{CO}\} \quad \text{公式 4}$$

首要污染物是指 AQI 的值大于 50 时，IAQI 最大值对应的空气污染物。如果 IAQI 最大的污染物有两项或两项以上，并列为首要污染物。污染等级是根据 AQI 的大小来划分的，划分方式如表 4-2 所示。

表 4-2　空气质量等级划分说明

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

4.3 问题一分析过程及结果

表 4-3　2020/8/25-2020/8/28 六项污染物监测浓度

时间	SO2 监测浓度(μg/m³)	NO2 监测浓度(μg/m³)	PM10 监测浓度(μg/m³)	PM2.5 监测浓度(μg/m³)	O3 最大八小时滑动平均监测浓度(μg/m³)	CO 监测浓度(mg/m³)
2020/8/25	8	12	27	11	112	0.5
2020/8/26	7	16	24	10	92	0.5
2020/8/27	7	31	37	23	169	0.6
2020/8/28	8	30	47	33	201	0.7

根据上述方法，先计算得到六个污染物的 IAQI，取其最大值为 AQI，进而依据对照表得到，如表 4-4 所示的 AQI 计算结果。

表 4-4 AQI 计算结果表

监测日期	地点	AQI 计算	
		AQI	首要污染物
2020/8/25	监测点 A	60	臭氧（O3）
2020/8/26	监测点 A	46	-
2020/8/27	监测点 A	109	臭氧（O3）
2020/8/28	监测点 A	138	臭氧（O3）

4.4 总结

目前，环境空气质量指数能直观反映空气中污染物的基本情况，环境空气质量指数（AQI）是定量描述空气质量状况的无量纲指数，数值越小环境空气质量越好，污染物浓度越低，数值越大环境空气质量越差，污染物浓度越高。随着空气污染加重，人的健康遭受严重威胁，为此空气质量监测意义深远，让人们及时了解生活的空气质量状况的同时，为空气污染治理提供依据，改善居住环境，提高生活品质。

表 4-5 空气质量指数影响

空气质量指数（AQI）范围	空气质量等级	对人身体健康的影响	建设采取措施
[0,50]	优	正常活动，健康未受到影响	不需要采取措施
[51,100]	良	正常活动，健康未受到影响	不需要采取措施
[101,150]	轻度污染	易感染人群症状加剧，健康人群会发生刺激症状	患有呼吸疾病和心脏病人群应减少户外运动
[151,200]	中度污染	肺病或心脏病患者症状显著，人体耐受力降低，健康人员也会出现不良症状 肺病、心脏病患者、老年人应停留在户内，不外出，	肺病、心脏病患者、老年人应停留在户内，不外出， 减少体力活动
[201,300]	重度污染	健康人员耐力降低，会提前出现一些疾病	一般人员要尽量避免户外运动
[301,+∞)	严重污染	健康人员耐力受损，会出现一些疾病	尽量避免户外运动

5 问题二：基于循环聚类及随机森林的气象条件特征提取模型

5.1 问题提出与解决思路

在污染物排放情况不变的条件下，某一地区的气象条件有利于污染物扩散或沉降时，该地区的 AQI 会下降，反之会上升。例如风向影响污染物的水平迁移扩散方向，总是不断将污染物向下风向输送，污染区总是分布在下风向上，高污染浓度常出现在大污染源的下风向。为确切了解气象条件对污染物浓度的影响机制，使用附件 1 中的数据，根据对污染物浓度的影响程度，对气象条件进行合理分类，追寻各类气象条件的特征。问题二具体解决思路如下图 5-1 所示。

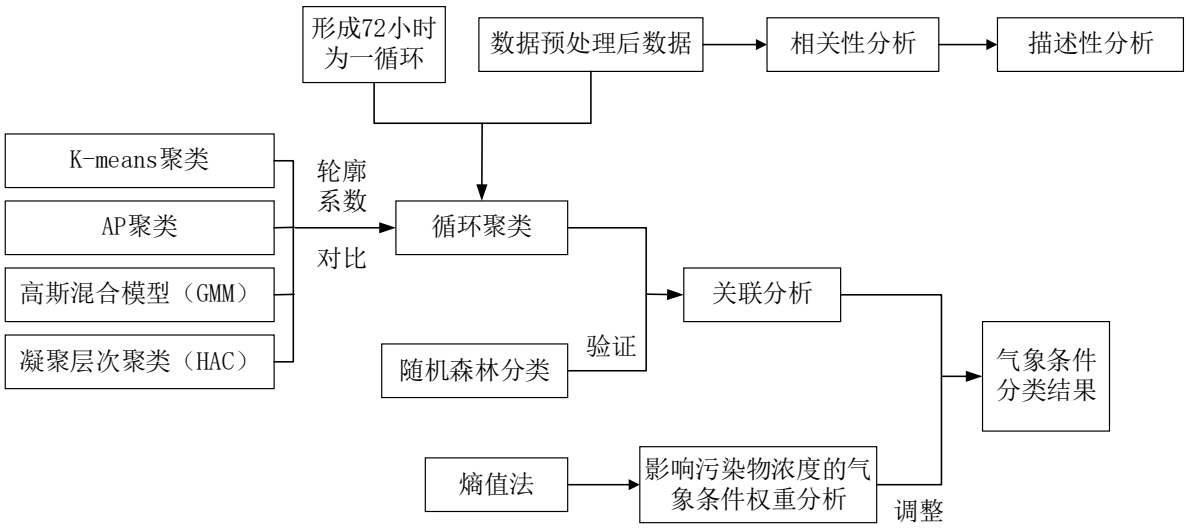


图 5-1　问题二研究思路图

5.2 数据集描述

1. 三天循环说明

对附件 1 中的数据进行描述性分析时发现，“监测点 A 逐小时污染物浓度与气象一次预报数据”中的预报数据有明显周期性，在进一步观察“监测点 A 逐小时污染物浓度与气象实测数据”时发现，实测数据的周期性不是很强。因此在结合“监测点 A 逐小时污染物浓度与气象实测数据”中的数据进行分析时，需要考虑数据周期性问题。同时需要注意的是考虑到每天的数据存在一定的周期性尤其是温度这一变量，我们要有一个周期以上的时间间隔变化来进行反映。在本次建模过程中，我们参考一次预报数据跨度为三天，我们设定三个周期为一循环，也就是三天一循环。举例说明，即 2021-7-12 7:00、2021-7-15 7:00 时刻的温度、湿度、气压、风速、风向，影响的结果为 2021-7-12 7:00、2021-7-15 7:00 两个时刻的各污染物变化量。

2. 变量说明

第二题选用的变量是涵盖了空气污染物浓度、气象条件及周期性的。

第一个是循环初始时刻气象条件(温度、湿度、气压、风速、风向)，第二个是到第三个周期同时刻即循环结束时刻的气象变化量，第三个是污染物循环初始时刻的浓度，第四个是污染物第三个周期同时刻即循环结束时刻的污染物浓度变化量，共有 22 个指标，且每三天为一循环。

5.3 大气污染物及气象条件特征描述性分析

在附件 1“监测点 A 逐小时污染物浓度与气象一次预报数据”中，将时间按季节切分，共分出 5 个季节，在每个季节的数据中，随机挑选一天，绘制出气象条件三天预测变化图和污染物小时平均浓度三天变化趋势图。

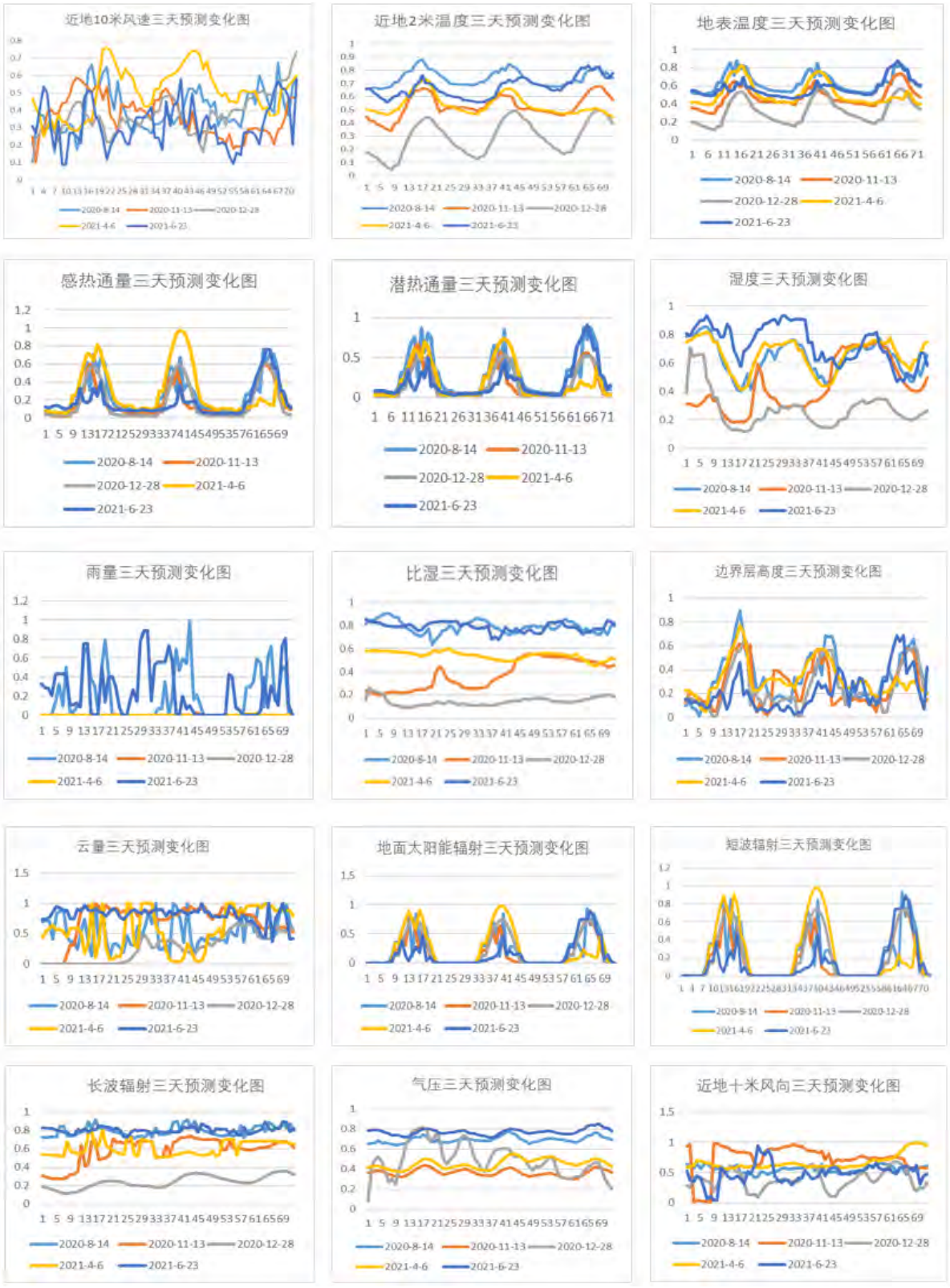


图 5-2　气象条件三天预测变化图

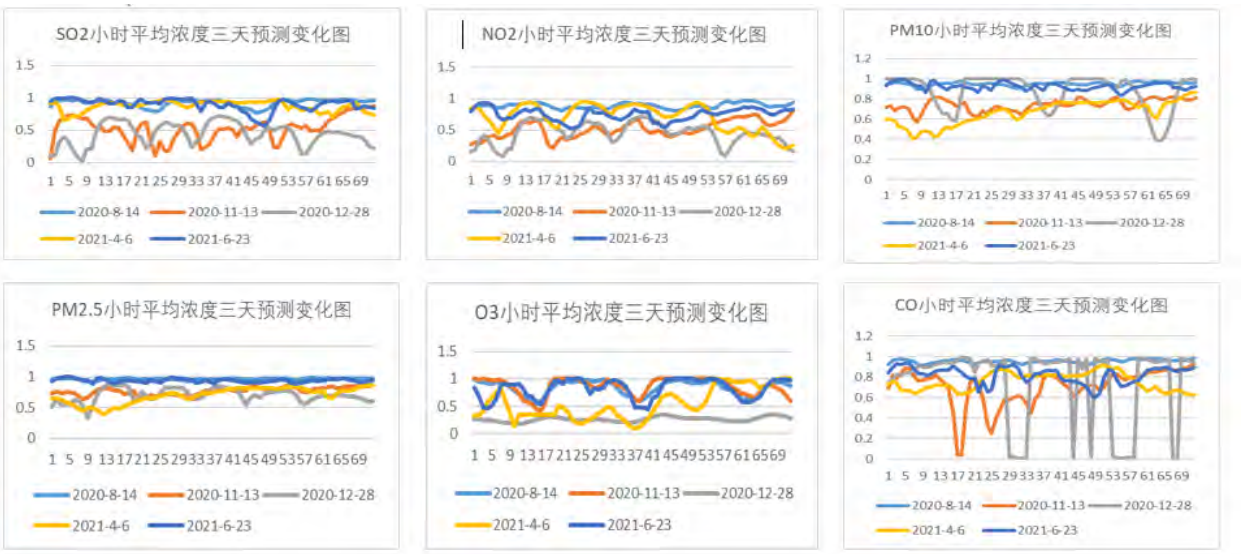


图 5-3 污染物小时平均浓度三天变化趋势图

通过分析，可以发现气象条件和污染物浓度存在周期性变化，变化周期为一天。尤其是温度、地面太阳辐射、短波辐射、感热通量和潜热通量指标，周期性十分明显，这是由于地球自然条件所致。

在污染物小时平均浓度变化趋势中，PM2.5 受时间的影响最小，而 CO、O3 受时间影响较大。并且大致可以看出，环境污染有变好趋势，但仍然有待改善。

5.4 主要大气污染物与气象条件相关性分析

5.4.1 主要大气污染物与气象条件相关性分析

空气质量的好坏受气象条件、污染源地理位置、排放强度等多因素的影响，但首要因素为气象条件，其输送、扩散或沉降会对大气污染物浓度产生明显影响，同时也是变化速度最显著、变化规律最复杂的影响因子[4]。为保证人民居住环境，提高空气质量气象预报条件水平，开展监测点大气污染物与气象相关性分析迫在眉睫，能为后续开展空气质量预警提供科学依据。

本文采用 Pearson 相关系数法计算大气污染物浓度与气象条件的相关关系，公式如下所示：

$$r = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \quad \text{公式 5}$$

其中 E 表示数学期望。相关系数范围在[-1,1]之间，$r=0$ 表示不相关，$0<r<1$ 表示正相关，$-1<r<0$ 表示负相关。

表 5-1　监测点 A 大气污染物与气象条件相关系数

	相关	温度	湿度	风速	气压	风向
PM_{10}	相关系数	0.213**	**0.434****	0.238**	**0.379****	-0.028**
	显著性	.000	.000	.000	.000	.000
$PM_{2.5}$	相关系数	0.289**	**0.316****	**0.325****	**0.405****	-0.028**
	显著性	.000	.000	.000	.000	.000
O_3	相关系数	**-0.327****	**0.566****	-0.229**	-0.049**	-0.051**
	显著性	.000	.000	.000	.000	.000
CO	相关系数	**0.374****	0.063**	**0.331****	**0.337****	0.032**
	显著性	.000	.000	.000	.000	.000
NO_2	相关系数	**0.378****	0.025**	**0.485****	**0.322****	-0.099**
	显著性	.000	.001	.000	.000	.000
SO_2	相关系数	0.156**	**0.468****	0.100**	**0.316****	-0.036**
	显著性	.000	.000	.000	.000	.000

注：**. 在 0.01 级别（双尾），相关性显著。

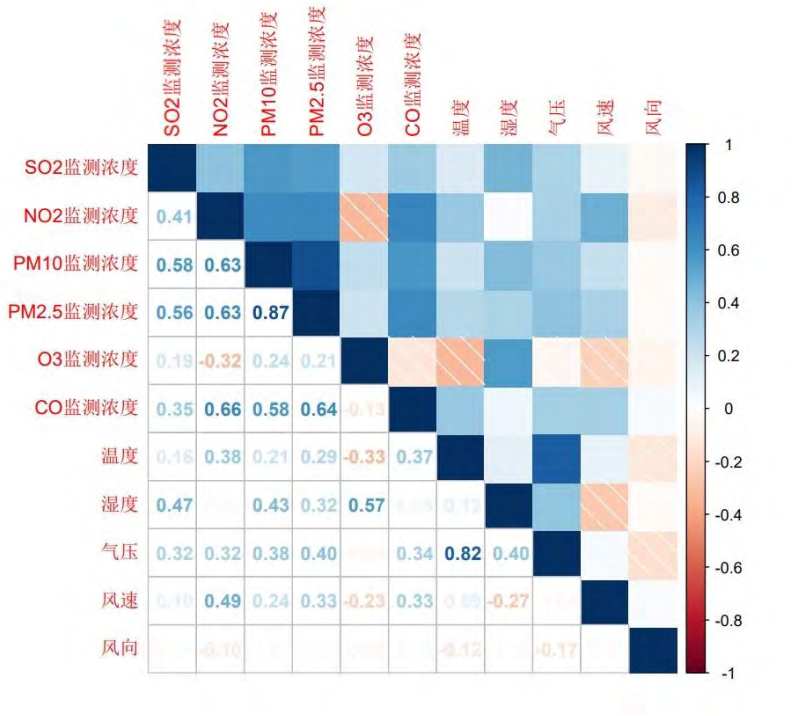


图 5-4　污染物与气象条件相关热力图

从上表 5-1 可知，通过对监测点 A 主要污染物浓度实测数据与气象要素的相关性分析，得到结果如下：（1）各项污染物浓度与气象条件均通过显著性水平为 0.05 的显著性检验。（2）通过显著性检验的各要素中，PM_{10}、$PM_{2.5}$、O_3、SO_2 与湿度呈显著正相关性；NO_2、CO 与温度呈正相关，O_3 与温度呈负相关；$PM_{2.5}$、CO、NO_2 与风速呈正相关；PM_{10}、$PM_{2.5}$、CO、NO2、SO2 与气压呈显著正相关；（3）虽然风向与大气污染物的相关性通过了显著性检验，但其相关性较弱。

5.4.2 与气温的相关性分析

在春季，温度与 O_3 浓度为负相关，与 CO 和 NO_2 呈正相关；O_3 浓度随温度变化规律性较强，温度升高有利于 O_3 浓度增加，与温度在春夏秋冬均为中度负相关；CO 与温度在春季呈现为正相关，相关系数为 0.402，在其他季节，相关性极弱。

表 5-2 四季污染物浓度与温度的相关系数

	春季	夏季	秋季	冬季
PM_{10}	0.081**	-0.379**	0.087**	-0.199**
$PM_{2.5}$	0.219**	-0.217**	-0.002	-0.192**
O_3	**-0.384****	**-0.685****	**-0.510****	**-0.479****
CO	**0.402****	0.214**	0.173**	0.039**
NO_2	**0.423****	**0.463****	**0.375****	-0.059**
SO_2	0.087**	-0.190**	0.039*	-0.181**

注：**. 在 0.01 级别（双尾），相关性显著。

5.4.3 与湿度的相关性分析

监测点 A 的 PM_{10}、$PM_{2.5}$、O_3、SO_2 浓度在四个季节中湿度呈现正相关关系，其中臭氧的浓度与湿度呈现较强的相关性。此外，湿度对 PM_{10} 的影响总体大于对 $PM_{2.5}$ 的影响。CO、NO 的浓度在四季基本与湿度呈现负相关关系。

表 5-3 四季污染物浓度与湿度的相关系数

	春季	夏季	秋季	冬季
PM10	**0.399****	**0.340****	**0.443****	0.179**
PM2.5	0.184**	0.200**	0.254**	0.052**
O3	**0.671****	**0.725****	**0.608****	**0.428****
CO	-0.123**	-0.231**	0.032*	-0.092**
NO2	-0.226**	**-0.436****	-0.036*	0.024
SO2	**0.347****	0.226**	**0.490****	**0.333****

注：**. 在 0.01 级别（双尾），相关性显著；*. 在 0.05 级别（双尾），相关性显著。

5.4.4 与风速的相关性分析

风靠空气中的水平运动形成，其中风速是其中很重要的因素之一。很大程度上，空气中的污染物浓度靠风来扩散[5]。相关研究表明：风速越大，大气对污染物的输送能力越差，尤其是静风情况下，污染物会在一定范围内形成堆积。

表 5-4　四季污染物浓度与风速的相关系数

	春季	夏季	秋季	冬季
PM10	0.233**	0.044**	0.200**	0.357**
PM2.5	0.350**	0.176**	0.310**	0.429**
O3	**-0.265****	**-0.240****	**-0.256****	**-0.192****
CO	0.364**	0.324**	0.362**	0.256**
NO2	**0.506****	**0.483****	**0.479****	**0.503****
SO2	0.158**	0.077**	-0.001	0.091**

注：**. 在 0.01 级别（双尾），相关性显著。

由表 5-4 可知，臭氧浓度在四季均与风速呈现负相关关系；二氧化氮浓度在四季均与风速正相关，且相关性程度较高；同时 PM10 和 PM2.5 与风速也均为正相关。

5.4.5 与气压的相关性分析

气压与大气环境的质量关系密切。同一季节下，气压变大使空气下沉，气流凝聚，阻碍污染物在同一空间下的扩散速度，从而与污染物浓度呈现正相关。

表 5-5　四季污染物浓度与气压的相关系数

	春季	夏季	秋季	冬季
PM10	0.312**	-0.268**	0.294**	-0.116**
PM2.5	**0.353****	**-0.344****	0.198**	-0.145**
O3	-0.039**	**-0.320****	-0.119**	-0.091**
CO	**0.392****	**-0.310****	0.298**	-0.215**
NO2	0.293**	-0.095**	0.256**	-0.120**
SO2	0.252**	-0.096**	0.231**	0.009

注：**. 在 0.01 级别（双尾），相关性显著。

5.4.6 与风向的相关性分析

风向影响污染物的水平迁移扩散方向，总是不断将污染物向下风向输送，污染区总是分布在下风向。根据表 5-6 的相关系数可知，四季污染物浓度与风向的相关性较弱。其中 NO2 在春季和夏季与风向呈现中度负相关。

表 5-6　四季污染物浓度与风向的相关系数

	春季	夏季	秋季	冬季
PM10	-0.084**	0.047**	0.119**	0.092**
PM2.5	-0.112**	0.039**	0.165**	0.085**
O3	-0.026	0.090**	-0.013	-0.116**
CO	-0.004	-0.108**	0.218**	0.203**
NO2	**-0.228****	**-0.332****	0.063**	0.098**
SO2	-0.025	0.184**	0.024	-0.015

注：**. 在 0.01 级别（双尾），相关性显著。

5.5基于循环聚类的气象条件特征提取模型

5.5.1 算法简介

1.K-means 聚类

K-means（K 均值聚类算法）是通过迭代方法求解的分析算法，作用机理是把一堆数据点分成若干类。在指定的 K 个簇中，聚类效果与簇内的数据样本相似度成正比。基于以上思想，该聚类的目标函数如公式 5 所示:

$$J(C_1,C_2,...,C_k)=\sum_{j=1}^{k}\sum_{i}^{n_j}(x_i-c_j)^2 \qquad \text{公式 6}$$

其中公式中各符号含义如下表 5-7 所示：

表 5-7　K-means 聚类目标函数公式符号含义

符号	解释
c_j	第 j 个簇的簇中心
x_i	第 j 个簇的样本 i
n_j	第 j 个簇的样本总量

对于该目标函数而言，c_j 为未知的参数。只有事先知道 c_j 的值，才能求得目标函数的最小值。为达到目标，K-means 的步骤阐述如下：

步骤一：确定聚类数据 K，在特征空间中选择 K 个点作为初始的聚类中心点。

步骤二：计算每个特征到聚类中心点的距离，并分配给距离最短的中心点。

步骤三：得到每个聚类中心点对应的簇，从而完成第一轮聚类。

步骤四：每轮结束后，需结合各个簇所对应聚类点的特征坐标，根据设定方法更新聚类中心，以新的中心重新聚类，选出新的簇。不断重复上述过程，知道所有的簇聚成一类为止。

2.AP（Affinity Propagation）聚类

2007 年 Frey 等人在著名科学杂志 science 上提出了 AP（Affinity Propagation）聚类算法。AP 算法根据 N 个数据点之间的相似度进行聚类，不需要事先指定聚类数目，而是将所有的数据点都作为潜在的聚类中心。

N 个数据点之间的相似度，就组成一个 $N\times N$ 的相似度矩阵 S，并以对角线上的值 $S(i,i)$

即参考度 p（preference）作为第 i 个数据点能否成为聚类中心 k 的评判依据，该值越大，表明该数据点成为聚类中心的可能性也就越大。

表 5-8 AP 算法中传递两种类型的信息表

信息	表达形式	表示内容	反映内容
吸引度	r(i,k)	从点 i 发送到聚类中心 k 的数值信息	k 点作为 i 点聚类中心的合适程度
归属度	a(i,k)	从聚类中心 k 发送到 i 的数值信息	i 点选择 k 点作为聚类中心的合适程度

可看出，吸引度和归属度越强，k 点作为 i 点聚类中心的可能性越大。AP 算法就是通过多次迭代，更新每一个点的吸引度和归属度信息。当已经历最大迭代次数，或数值收敛，则对于任意点 i，计算它与所有样本的吸引度与之和，那么 i 点的聚类中心 k 点如下式选择：

$$\mathrm{k}=\mathrm{argmax}\left(\mathrm{a}\left(\mathrm{i},\mathrm{k}\right)+r\left(i,k\right)\right) \qquad \text{公式 7}$$

AP 聚类具有如下优势：

不需要事先指定聚类的数量。聚类的数量，由参考度(preference)S(i,i)的初始值与数据的分布共同决定；

聚类的结果不会多次运行而随机变化。这比通用的 k-means 聚类更加稳定；

适用于非对称与稀疏的相似性矩阵。

3.高斯混合模型（GMM）

高斯混合模型可以看作是由 K 个单高斯模型组合而成的模型，这 K 个子模型是混合模型的隐变量（Hidden variable），K 的取值需要事先确定，具体的形式化定义如下：

$$\mathrm{P}\left(y\,|\,\theta\right)=\sum_{k=1}^{K}\partial_k\phi\left(y\,|\,\theta_k\right) \qquad \text{公式 8}$$

其中 ∂_k 是样本集合中 k 类被选中的概率：$\partial_k=P\left(z=k\,|\,\theta\right)$，其中 z=k 指的是样本属于 k 类，那么 $\phi\left(y\,|\,\theta_k\right)$ 可以表示为 $\phi\left(y\,|\,\theta_k\right)=P\left(y\,|\,z=k,\theta\right)$，很显然 $\partial_k\geq 0,\sum_{k=1}^{K}\partial_k=1$，y 是观测数据。

首先可以先假设聚成 k 类，然后选择参数的初始值 θ_0（总共 2K 个变量），这里需要引进一个变量 γ_{jk}，表示的是第 j 个观测来自第 k 个 component 的概率，即数据 j 由第 k 个 component 生成的概率，根据后验概率计算得到：

$$\gamma_{jk}=P\left(z=k\,|\,y_j,\theta\right)=\frac{\partial_k\phi\left(y_j\,|\,\theta_k\right)}{\sum_{k=1}^{K}\partial_k\phi\left(y_j\,|\,\theta_k\right)} \qquad \text{公式 9}$$

注：这个与 ∂_k 的区别，∂_k 指的是第 k 个 component 被选中的概率，需要 γ_{jk} 对所有的数据进行累加。

上面是根据数据 j 计算各个 component 的生成概率，而现在根据每个 component 生成了 1,2,…,N 点数据，每个 component 又是一个高斯分布，那么根据 ∂,μ,σ^2 的定义又可以直观地得出如下式子：

$$\partial_k=\frac{\sum_{j=1}^{N}\gamma_{jk}}{N} \qquad \text{公式 10}$$

$$\mu_k = \frac{\sum_{j=1}^{N} \gamma_{jk} y_j}{\sum_{j=1}^{N} \gamma_{jk}} \qquad \text{公式 11}$$

$$\sigma_k^2 = \frac{\sum_{j=1}^{N} \gamma_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^{N} \gamma_{jk}} \qquad \text{公式 12}$$

这样其实只是把原本样本一定属于某一类改成了样本属于某类的概率，而 k 类样本数量 N_k 变成了概率相加，$N_k = \sum_{j=1}^{N} \gamma_{jk}$，就可以直接得出上述公式，进行相互迭代，直到收敛，高斯混合模型就聚类完成。

4.凝聚层次聚类（HAC）

层次聚类算法属于无监督学习的一种聚类算法，分为 2 种聚类方式：凝聚式层次聚类（Hierarchical Agglomerative Clustering），简称 HAC 和分裂式层次聚类（DHC）。

凝聚式层次聚类的思想是：首先将每个数据单独成簇，之后按照相似性度量标准将相似性最高的数据先进行合并，依照数据相似度从高到低的顺序依次合并成簇，簇间的相似度随着簇的合并而降低，直到达到给定的相似度阈值才会停止。

首先将原始样本$(F_1, F_2, \ldots, F_n)$中每个样本自成一类，原始的类中心为$(C_1, C_2, \ldots, C_n)$，然后根据相似性度量标准将类中心最近的 2 个文本合并，根据聚类收敛的条件不断重复这一过程直至所有可合并的类合并完成。本文采用的相似性度量标准为欧式距离法，为

$$d(F_i, F_j) = \sqrt{\sum_{m=1}^{n} (w_{m,i} - w_{m,j})^2} \qquad \text{公式 13}$$

式中：F_i、F_j 分别表示文本集合中第 i 个和第 j 个文本；$w_{m,i}$、$w_{m,j}$ 分别表示文本 F_i 和 F_j 的第 m 个特征项的权重；n 表示文本特征项的数目。

该方法的优点在于：K-Means 等方法需要事先设定初始聚类中心和聚类个数，导致不同的聚类中心和聚类种类产生不同的聚类结果不尽相同，且产生局部最优解的可能性较大。对于 AHC 而言，因这种方法不需要事先设置数据的聚类中心，而是以整个输入数据为聚类中心，不断进行融合凝聚，最后达到聚类效果，不会陷入局部最优解，因此该方法能在一定程度上增加聚类的有效性。

5.5.2 模型选择与建立

本文选用 **K-means 聚类、AP 聚类、高斯混合聚类（GMM）、凝聚层次**四种不同的方法，依据气象因子对污染物浓度的影响程度，对气象因子进行聚类。采用**轮廓系数**评估聚类质量。轮廓系数取值在[-1,1]，越接近于 1，所包含的簇越紧凑，并且远离其他簇，这是可取的情况。反之，当轮廓系数为负值时，这种情况不可取。

四种方法的聚类结果如下表显示，从轮廓系数可以看出 K-means 聚类的效果最优，并结合各聚类算法的特点，下文采用 **K-means** 进行聚类。

表 5-9　四种聚类算法效果评估

聚类方法	K-means	AP	GMM	HAC
轮廓系数	0.338	0.231	0.194	0.207

5.5.3 结果分析

通过 K-means 聚类分析，得到如下图 5-5 所示的陡坡图，依据选取规则，可以发现聚为 8 类最合适，当然选取规则不唯一，具体应根据实际情况而定。从指标上看，选择坡度变化不明显的点最为最佳聚类数目。

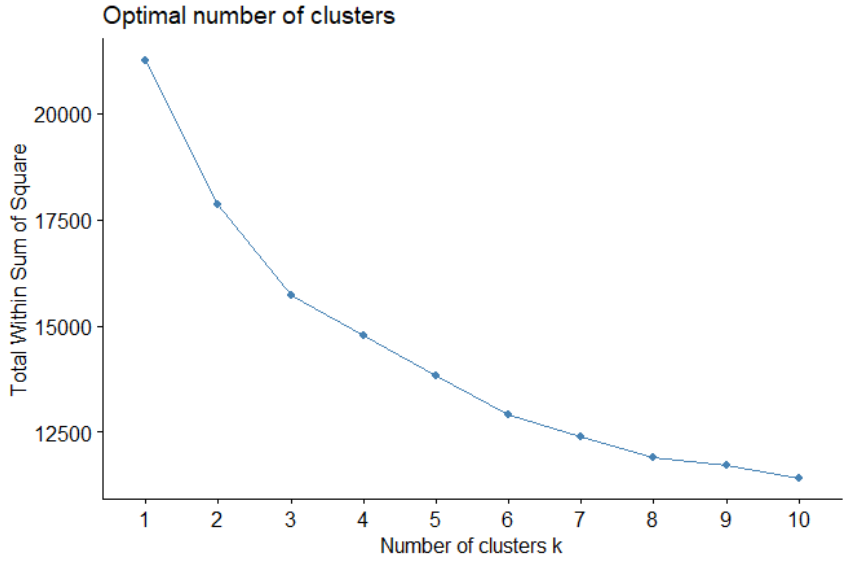


图 5-5　K-means 聚类陡坡图

聚类结果显示具有 8 个大小簇的 K 均值聚类：3389, 1621, 1572, 2365, 4027, 2423, 2542, 1652。聚类中心结果如下表 5-10 所示。

表 5-10 聚类中心结果表

	1	2	3	4	5	6	7	8
SO2 监测浓度(μg/m³)	0.686	0.424	0.458	0.513	0.700	0.659	0.703	0.560
NO2 监测浓度(μg/m³)	0.689	0.328	0.461	0.770	0.878	0.759	0.702	0.749
PM10 监测浓度(μg/m³)	0.741	0.340	0.400	0.518	0.786	0.732	0.742	0.652
PM2.5 监测浓度(μg/m³)	0.774	0.373	0.415	0.568	0.871	0.767	0.789	0.705
O3 监测浓度(μg/m³)	0.850	0.867	0.773	0.343	0.756	0.800	0.826	0.698
CO 监测浓度(mg/m³)	0.540	0.307	0.313	0.534	0.669	0.505	0.567	0.546
温度(℃)	0.542	0.435	0.492	0.658	0.739	0.481	0.575	0.351
湿度(%)	0.765	0.567	0.559	0.431	0.643	0.656	0.744	0.400
气压(MBar)	0.531	0.384	0.448	0.523	0.686	0.483	0.563	0.297
风速(m/s)	0.336	0.234	0.217	0.417	0.510	0.386	0.333	0.556
风向(°)	0.795	0.785	0.155	0.693	0.359	0.128	0.828	0.873
SO2 变化量	-0.024	0.126	0.116	0.088	0.004	-0.062	-0.063	-0.130

NO2 变化量	0.036	0.304	0.182	-0.012	-0.065	-0.115	0.076	-0.300
PM10 变化量	-0.032	0.248	0.196	0.115	-0.015	-0.135	-0.025	-0.247
PM2.5 变化量	-0.019	0.261	0.216	0.132	-0.028	-0.140	-0.025	-0.282
O3 变化量	-0.077	-0.085	-0.020	0.245	0.009	-0.057	-0.059	0.066
CO 变化量	0.020	0.134	0.156	0.002	-0.042	-0.012	-0.020	-0.157
温度变化量	0.009	-0.008	-0.014	-0.045	-0.028	0.022	0.032	0.067
湿度变化量	-0.081	0.007	0.027	0.101	0.038	-0.050	-0.087	0.097
气压变化量	0.007	0.008	0.016	0.002	0.012	0.015	0.015	0.065
风速变化量	0.063	0.164	0.155	0.031	-0.072	-0.050	0.048	-0.305
风向变化量	0.003	0.180	0.523	0.081	0.103	0.563	0.627	0.319

聚类后，各个类别的样本数如下表 5-11 所示，同时聚类结果如下图 5-6 所示。

表 5-11 聚类类别样本数

类别	1	2	3	4	5	6	7	8
样本数	3389	1621	1572	2365	4027	2423	2542	1652

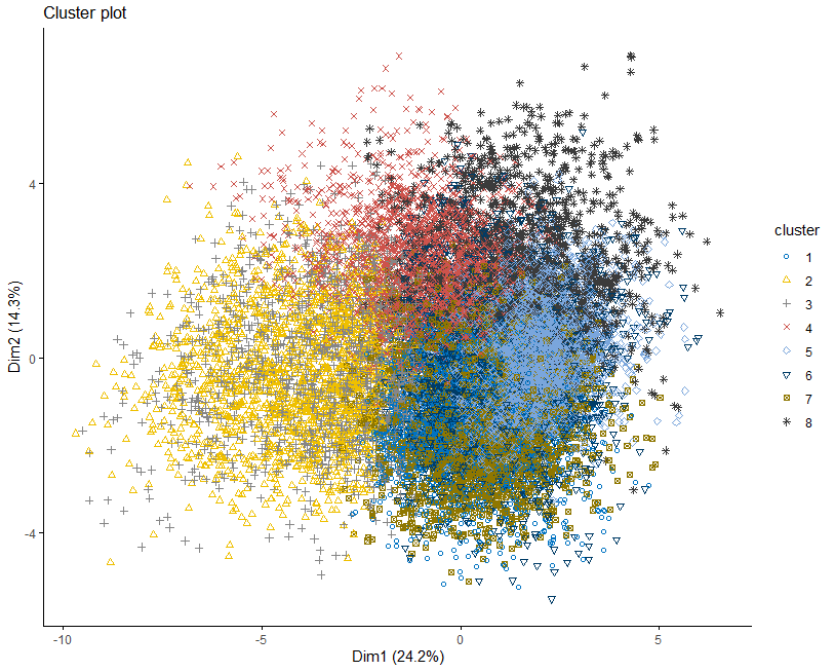


图 5-6 聚类结果图

5.6基于随机森林的分类模型

利用 K-means 进行聚类，效果最佳的 K 值为 8，但是由于影响气象条件的污染物浓度包括 SO2 监测浓度(μg/m³)、NO2 监测浓度(μg/m³)、PM10 监测浓度(μg/m³)、PM2.5 监测浓度(μg/m³)、O3 最大八小时滑动平均监测浓度(μg/m³)、CO 监测浓度(mg/m³)这 6 个指标，聚类的团簇个数理应与气象条件类别数相同，因此本文利用具有泛化性强、稳健性、对噪声不敏感等特点的随机森林分类算法进行调整。

5.6.1 算法简介

随机森林是一种有监督的集成学习方法，由于其不存在过拟合问题并具有良好的分类性能，因此被广泛应用于各种分类问题中。采用 CART[3]作为随机森林元分类器，用 Bagging 方法[4]生成每棵决策树的随机训练样本集。当构造一棵树时，随机选取训练样本中的特征来确定决策树的节点分裂。Bagging 方法和 CART 的结合，并且随机选择特征进行属性分割，使得 RF 能较好的容忍噪声，并具有较好的分类性能。

随机森林是由多个决策树{h(x, θ k),k=1,2,…,n}组成的组合分类器，其中{ θ k}是一个独立同分布的随机向量，通过对所有决策树结果的投票产生输出结果。一个随机森林由 N 棵决策树构成，所有决策树（如决策树 T1,T2,…,TN）是一个分类器，随机森林的决策结果由所有决策树分类结果的组合策略得出。

5.6.2 模型建立

随机森林模型有 3 个重要的可调参数：

表 5-12 随机森林模型可调参数

参数	含义
nodesize	包含样本的叶节点数
ntree	森林中树的数目
mtry	每个节点的候选特征数

一般而言，节点大小为 1 表示分类，5 表示回归。因此，本文取 nodesize=1。

研究表明，mtry 是影响随机森林模型性能最明显的参数，要求 mtry<<M，它表示每个分段中随机选择的候选变量数。在分类中，mtry 的建议值是整个变量个数的均方根；在回归中，mtry 的建议值是变量总数的 1/3。通过实证，发现 mtry 对模型的分类性能影响不大，因此本文取 mtry=3。

ntree 的设置相对简单，只要随机森林的总体误差率趋于稳定。Breiman 定义了随机森林的间隔函数，并根据大数定律证明了 RF 的泛化误差随着林中树数的增加趋于有限上界。因此只要 ntree 值足够大，就可以保证 RF 收敛。实证表明，当 ntree=79 时，分类效果最好，准确率达到 0.8235。

因此，随机森林的最优模型参数如下，达到了最好的分类效果：

表 5-13 随机森林最优模型参数

参数	含义
nodesize	1
ntree	79
mtry	3

5.6.3 结果分析

使用袋外错误率（out of bag，error rate）作为随机森林的评价指标，oob 误分率是随机森林泛化误差的一个无偏估计，它的结果近似于需要大量计算的 k 折交叉验证。泛化误差代表的是在测试集上表现的好坏。随机森林调整分类评价效果如下表 5-14 所示。

表 5-14 随机森林分类效果

	1	2	3	4	5	6	7	8	错误率	正确率
1	2431	172	22	91	84	82	416	91	0.282679	0.717321
2	315	904	98	119	1	7	84	93	0.44232	0.55768
3	7	31	1161	33	116	218	0	6	0.26145	0.73855
4	102	88	90	1655	227	41	62	102	0.300803	0.699197
5	108	1	36	124	3455	196	102	5	0.142041	0.857959
6	5	3	262	39	340	1757	0	7	0.271861	0.728139
7	743	148	0	161	112	2	1289	87	0.492919	0.507081
8	117	83	9	146	19	43	27	1208	0.268765	0.731235

由随机森林分类结果可知，第 2、7 条聚类得到规则误差率极高，故删除。

5.7 基于信息熵的权重调整

根据相关研究表明，在污染物排放总量不变的前提下，气象条件的变动是致使空气质量改变的重要因素。因此，为了提高 WRF-CMAQ 模型后续二次预报的准确率，本文选用熵权法找出影响大气污染物浓度的主要气象因素，即对气象因子重要程度进行排序。

在权重调整后，将六个污染物及五个气象条件因子的变化量返回到因子本身，剩余得到 11 个指标。

5.7.1 算法简介

信息量是一个很难描述的概念，直到 1948 年信息论之父 Shannon 提出信息熵的概念，信息度量问题才得以解决[6]。信息熵的概念是指热力学中热熵的概念，用来衡量信息的不确定性、稳定性和信息量。信息熵越大，信息越无序，信息熵越小，信息越有序。由于信息中必然存在冗余，因此冗余的大小与信息组成要素出现的概率有关。组成元素出现的单词越多，概率越大，信息的不确定性越小，信息量越大，反之亦然[7]。因此，信息熵是剔除冗余后的平均信息量，具体数学公式如下：

$$H(x) = -\sum_{i=1}^{m} p(x_i) \log p(x_i)$$ 公式 14

5.7.2 结果分析

空气质量指数（AQI）的大小直观反映了空气的污染程度，则只要找出影响空气质量指数的主要气象因素，就可以找到影响污染物浓度产生影响的气象因子。通过分析计算得出信息熵见表 5-15。

表 5-15 影响污染物浓度的主要气象因子及其信息熵

SO2 监测浓度	NO2 监测浓度	PM10 监测浓度	PM2.5 监测浓度	O3 监测浓度	CO 监测浓度	温度	湿度	气压	风速	风向
0.091986	0.084877	0.090991	0.091517	0.101422	0.086628	0.087713	0.096736	0.090352	0.086028	0.091748

可以看出，6 项污染物浓度的特征重要度情况排序：O3> SO2> PM2.5> PM10> CO> NO2。5 项气象因子的特征重要度排序：湿度>风向>气压>温度>风速。

5.8气象条件特征分析

由以上分析可知，监测点 A 大气污染物随季节变化存在明显的规律性，除 O_3 外，其余几种污染物的浓度基本呈现出冬高夏低的变化趋势，而 **O_3 浓度则为冬低夏高**。对污染物浓度与气象条件的相关性分析可知：

（1）春季，温度与 O3、NO2 和 CO 的相关相对较强，湿度与 O3 相关，风速与 NO2 相关性较其他污染物强，其余无明显相关性；

（2）夏季，温度与 O3、NO2 的相关相对较强，湿度与 O3 和 NO2 相关，风速与 NO2 相关性较其他污染物强，气压与 PM2.5 的相关性相对较强，风向与 NO2 的相关性比其他污染物强；

（3）秋季，温度与 O3、NO2 的相关相对较强，湿度与 O3、PM10 和 SO2 的浓度相关比其他污染物强，风速与 NO2 相关性较其他污染物强，其他无明显相关性。

（4）冬季，温度与 O3 的相关相对较强，湿度与 O3 和 SO2 相关，风速与 NO2 相关性较其他污染物强，其他无明显相关性。

紧接着选用了聚类效果最好的 **K-means 将气象因子聚为 8 类**，继而选用**随机森林算法构建模型，将类别调整为 6 类**。为了进一步探究气象因子对污染物浓度的重要程度，利用**熵权法**进行权重计算，调整上述分类。

最后先将污染物浓度的聚类分类对应气象分类，再反过来对应每个类别气象分类下的污染物浓度。

表 5-16 污染物浓度——气象条件特征分类表

气象类别 / 范围	1	2	3	4	5	6
SO2 监测浓度 (μg/m³)	8-9	13-26	**27-47**	0-5	9-12	6-7
NO2 监测浓度 (μg/m³)	2-23	91-153	**154-211**	38-59	24-37	60-90
PM10 监测浓度 (μg/m³)	**164-217**	108-163	69-107	0-26	50-68	27-49
PM2.5 监测浓度 (μg/m³)	14-24	25-43	0-13	44-63	64-109	**110-163**
O3 监测浓度	37-78	79-126	0-36	127-165	166-237	**238-405**

(μg/m³)						
CO 监测浓度 (mg/m³)	0.5-0.8	0.1-0.4	0.9-1.1	1.2-1.6	1.7-2.1	**2.1-2.5**
温度(℃)	**30.8-38.2**	28.9-30.7	22.6-25.8	5.8-14.4	14.5-18.2	18.3-22.5
湿度(%)	73-84	**85-99**	61-72	14-21	22-44	45-60
气压(MBar)	**1019.0-2029.2**	1003.4-1008.2	1008.3-1013.5	1013.6-1018.9	993.5-999.6	999.7-1003.3
风速(m/s)	**3.6-5.8**	2.5-3.5	1.5-2.0	0.9-1.4	2.1-2.4	0.1-0.8
风向(°)	141-198	199-253	0-140	254-288	322-360	289-321

据此，归纳出 **6 类气象条件特征：**

表 5-17 气象条件—污染物浓度特征表

类别	气象条件特征	污染物浓度特征
第一类	气压高，终年炎热，多雨，湿度高，年温差小，偏南风，风力强	PM10 监测浓度高 SO2 监测浓度低
第二类	高温多雨，降水充沛，偏南风，风力强	除 CO 外，其他污染物浓度均较高
第三类	偏东风，风速中等，温湿度适宜	NO2 监测浓度高 PM2.5、O3 监测浓度低
第四类	气温低，湿度小，气压高，风速偏小，风向多西风	PM10、SO2 监测浓度低 O3 监测浓度高
第五类	气温偏低，湿度和气压较小，风速中等，偏北风	PM2.5、O3、CO 监测浓度高 NO2、监测浓度
第六类	温湿度适宜，微风，风向偏西北风	CO、O3、PM2.5 监测浓度高 NO2 监测浓度较高

6 问题三：基于 Elman 的空气质量二次预报模型

6.1 问题提出与解决思路

由于 WRF 模型得到的模拟气象场数据的不准确性、排放清单的不确定性、臭氧与氮氧化物之间的转化反应等因素，使得 WRF-CMAQ 预报模型的结果不理想，为此需结合实际气象条件、污染物实测数据进行二次建模，以得到优化后的污染物浓度预报。本题的研究思路图如下。

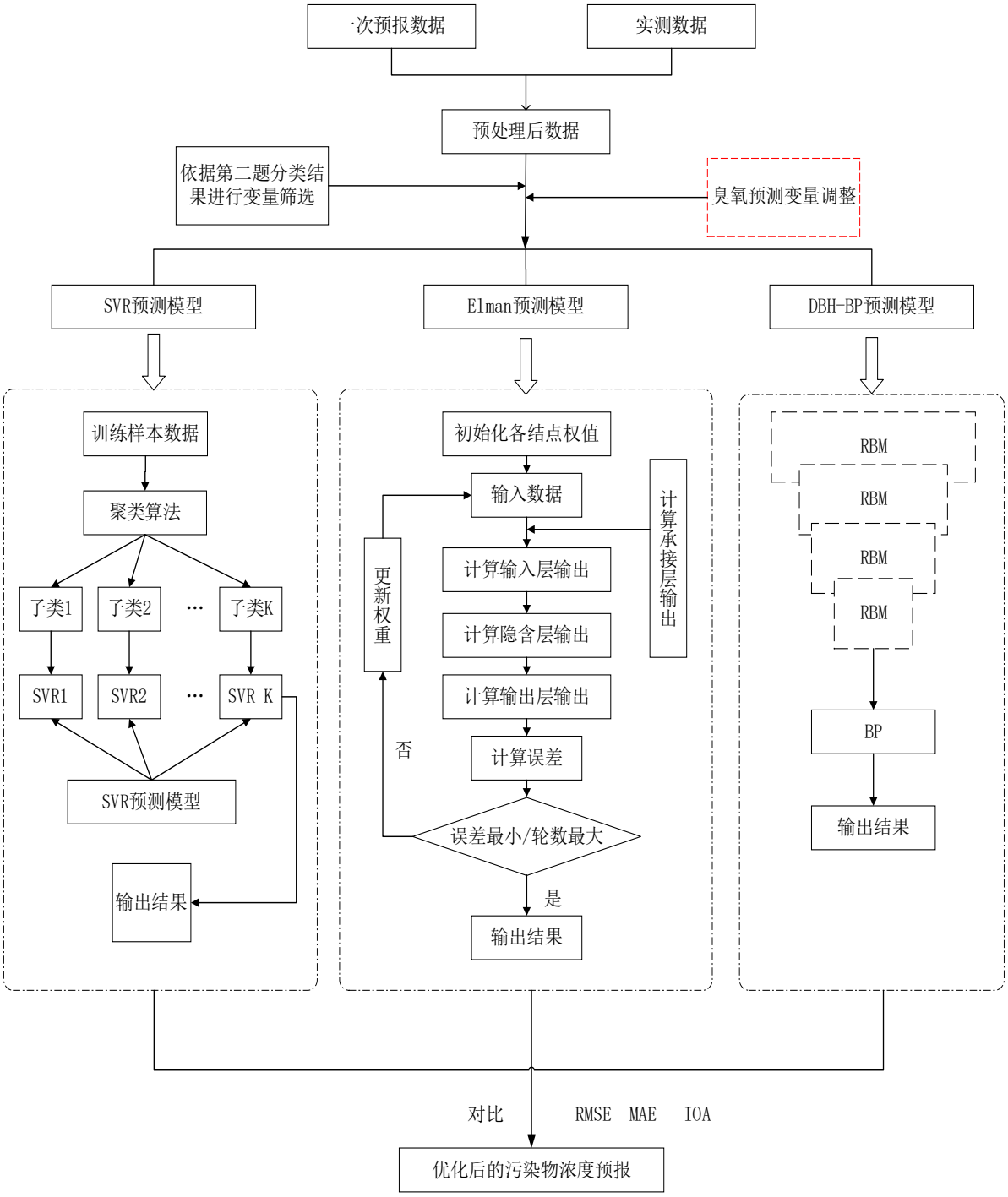


图 6-1 问题三研究思路图

6.2 二次模型的设计

6.2.1 数据获取

本文使用了附件 1、2 中的预测和实测数据，附近 1、2 中已包含一次模型后预报的数据，在对比预报数据和真实实测数据后发现，预报数据在一定程度上拟合了真实数据，但还是存在着较大的差别，**尤其是在对 O3 的预测**，因此，本文有必要在一次预报模型的基础上，建立二次模型，对一次预报模型的结果进行优化。

6.2.2 输入变量处理

实测数据的五个气象因子和模拟气象场的 15 个气象因子之间的相关关系如下图 6-2 所示。

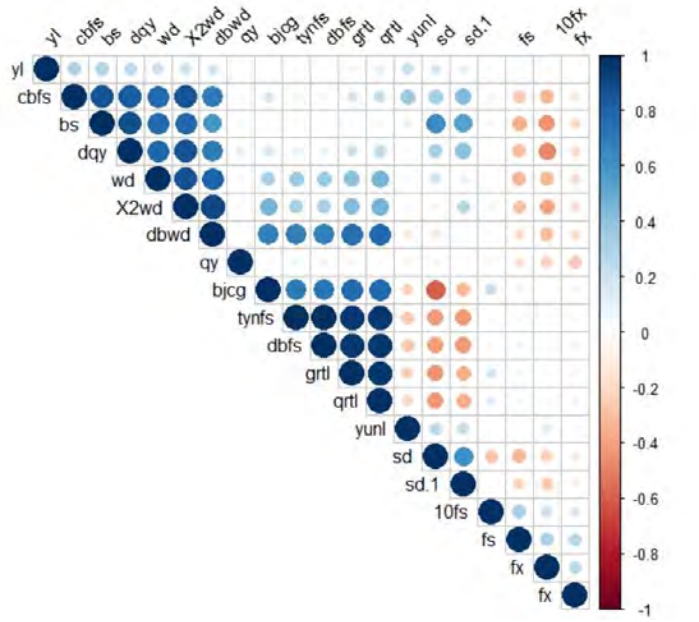


图 6-2 模拟气象场与实测气象指标间关系

污染物浓度与气象之间大部分呈现非线性关系，因此前文中采用**相关矩阵求逆法及熵权法**探究其关系，结果显示 **6 项污染物浓度的特征重要度**情况排序：O_3> SO_2> $PM_{2.5}$> PM_{10}> CO> NO_2。

在排放源维持稳定的情况下，SO_2 的浓度与 NO_2 的浓度主要是受到气象因子的影响，但由于 NO_2 与气象因子会发生化学反应，从而对**臭氧**（O_3）造成很大影响，因此本文选择**气温、大气压、风速、风向、温度露点差、相对湿度等气象因素作为 SO_2 浓度的变量，为 SO_2 额外增加一个 O_3 的变量。**CO 在气候因子催化下产生的化学反应，不会大比例的影响其他污染物的浓度，因此，仅用气象条件模拟预测 CO。O_3 与 NO_2 和 CO 都易发生化学反应，影响到浓度变化，因此，将气象因子和 NO_2 和 CO 作为输入变量。$PM_{2.5}$ 和 PM_{10} 的输入变量相同，可以选择 SO_2、NO_2 和 CO 作为输入变量。

表 6-1　二次空气质量预报模型输入变量

输出结果	变量
So2 实测浓度预测	气象因子、SO2 预测时点前 24 小时实测浓度 SO2 预测时点一次预告浓度
NO2 实测浓度预测	气象因子、O3 预测时点前 24 小时实测浓度 NO2 预测时点一次预告浓度
CO 实测浓度预测	气象因子、CO 预测时点一次预告浓度
O3 实测浓度预测	气象因子、NO2、CO 预测时点前 24 小时实测浓度 O3 预测时点一次预告浓度
PM2.5 实测浓度预测	气象因子、SO2、NO2、CO 预测时点前 24 小时实测浓度 PM2.5 预测时点一次预告浓度
PM10 实测浓度预测	气象因子、SO2、NO2、CO 预测时点前 24 小时实测浓度 PM10 预测时点一次预告浓度

对 6 项污染物**分别建立 6 个模型**，输入变量为当前预测时刻污染物和气象的预报数据、预测时点前 24 小时污染物实测浓度，对其进行训练后，得出二次预报数据。

6.2.3 二次空气质量预报模型

（1）SVR 预测模型[8]

支持向量机回归（SVR）利用了核函数和线性回归的思想，寻求最优分类面使得所有训练样本离该最优分类面误差最小。SVR 的主要求解方法与 SVM 相似，数据描述如图 6-4 所示。

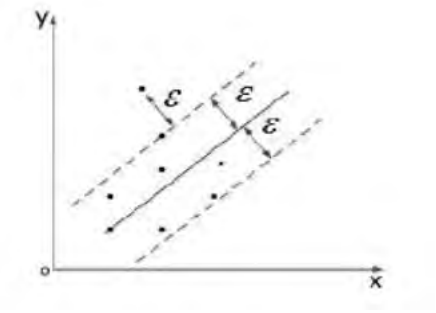


图 6-3　SVR 基本思想示意图

$$\begin{cases} \min \dfrac{1}{2}\|\mathbf{w}\|^2 + \overline{C}\sum_{i=1}^{L}(\varepsilon_i + \varepsilon_i^*) \\ s.t.\begin{cases} \mathbf{y}_i - \mathbf{w}\phi(\mathbf{x}_i) - \mathbf{b} \le \varepsilon + \xi_i \\ -\mathbf{y}_i + \mathbf{w}\phi(\mathbf{x}_i) + \mathbf{b} \le \varepsilon + \xi_i^* \\ \xi_i \ge 0, \xi_i^* \ge 0 \end{cases} \end{cases}$$

图 6-4　SVR 数据描述公式

引入 Lagrange 函数转换成对偶形式进行求解：

$$\begin{cases} \max\limits_{\partial,\partial^*}[-\sum\limits_{i=1}^{L}(\partial_i+\partial_i^*)\varepsilon+\sum\limits_{i=1}^{L}(\partial_i-\partial_i^*)\mathbf{y}_i-\frac{1}{2}\sum\limits_{i=1}^{L}\sum\limits_{j=1}^{L}(\partial_i-\partial_i^*)(\partial_j-\partial_j^*)K(\mathbf{x}_i,\mathbf{x}_j)] \\ s.t.\begin{cases}\sum\limits_{i=1}^{L}(\partial_i-\partial_i^*)=0 \\ 0\le\partial_i\le\overline{C} \\ 0\le\partial_i^*\le\overline{C}\end{cases} \end{cases}$$

公式 15

设最优解为$\partial=[\partial_1,\partial_2,...,\partial_L]$，$\partial^*=[\partial_1^*,\partial_2^*,...,\partial_L^*]$，得出的回归函数如下公式所示。

$$f(\mathrm{x})=\mathrm{w}^*\phi(\mathrm{x})+\mathrm{b}^*=\sum_{\mathrm{i}=1}^{L}(\partial_i-\partial_i^*)K(\mathrm{x_i},\mathrm{x})+\mathrm{b}^* \qquad \text{公式 16}$$

（2）DBN-BP 预测模型[9]

基于 DBN-BP 的空气质量预报修正模型是对当前 WRF-CMAQ 模型一次预报结果的修正[10]，通过结合时间、空间、历史监测的污染物数据及 WRF-CMAQ 一次预报模型的污染物数据及预报气象数据训练建立模型，实现对一次预报模型的修正，建立一个四层 RBM 的深度信念网络（DBN），并在 DBN 的最后一层加入一个 BP 神经网络，DBN-BP 的结构图如下**错误!未找到引用源。**

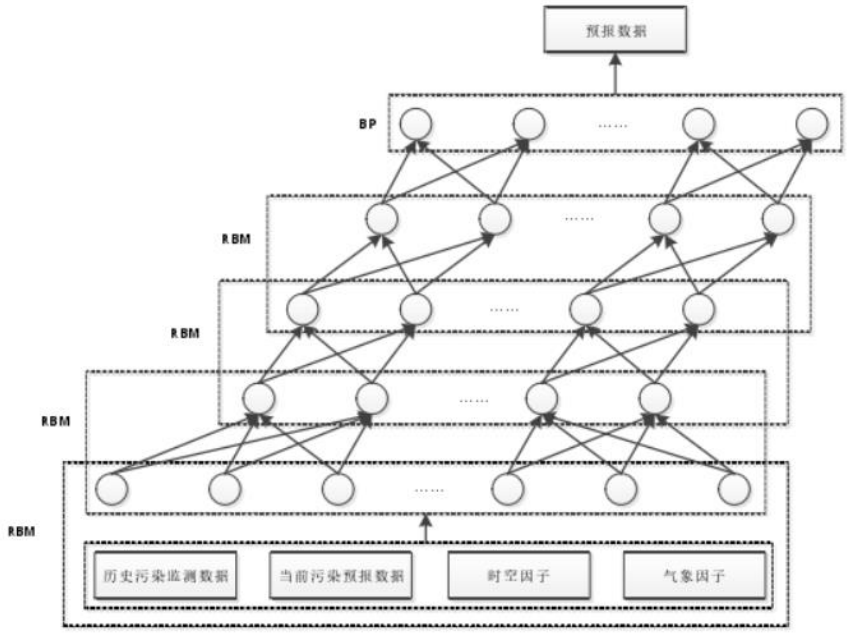


图 6-5 DBN-BP 训练模型架构

DBN-BP 模型训练过程：（1）预训练：采用对比散度算法 CD—K 逐层对 RBM 训练，反复无监督训练，更新权值；（2）微调过程：使用反向 BP 传播算法对预测过程中进行微调，反向传递误差，达到全局最优收敛效果。（3）隐藏层确定：$l=\sqrt{m+n}+\alpha$，其中 m 表示输入数据个数；n 表示输出数据个数；α 为正常数，取值在（0，10）之间，l 表示隐层神经元数。

（3）Elman 预测模型

Elman 神经网络是一种动态局部反馈型神经网络[11]，由输入层、隐含层、承接层和输出层四层构成，在传统神经网络的基础上增加了承接层，记忆隐含层的输出，保留了更完整的信息。Elman 神经网络结构如图 6-6 所示。

Elman 神经网络的流程：

①初始化各层节点的权值；

②输入训练数据，计算各层输入、输出值；

③承接层处理；

④最小误差后输出，否则更新权值。

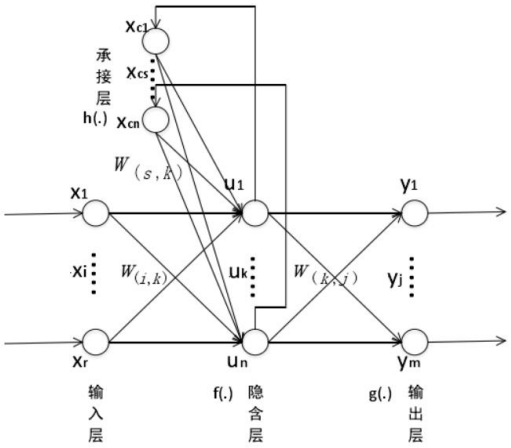


图 6-6 Elman 神经网络结构示意图

输出层输出公式： $y(t) = g(f(t)\,w_{(k,j)})$ 公式 17

隐含层输出公式： $u(t) = f(x(t)\,w_{(i,k)} + x_c(t-1)\,w_{(s,k)})$ 公式 18

承接层输出公式： $x_c = h(u(t-1))$ 公式 19

误差函数计算公式： $E(t) = \frac{1}{2}\left(y(t) - y_a(t)\right)^T (y(t) - y_a(t))$ 公式 20

根据误差逆向传播算法，可得： $\Delta w_{(k,j)} = \eta_3 \delta_j u_k(t)$ 公式 21

$$\Delta w_{(i,k)} = \eta_2 \delta_k x_{cs}(t)$$ 公式 22

$$\Delta w_{(s,k)} = \eta_1 \sum_{k=1}^{n} (\delta_j w_{(k,j)}) \frac{\partial x_{cs}(t)}{\partial w_{(s,k)}}$$ 公式 23

其中，

$$\delta_j = (y_j(t) - y_{aj}(t))\, g_j'(.)$$ 公式 24

$$\delta_k = \sum_{j=1}^{m} (\delta_j w_{(k,j)})\, f_k'(.)$$ 公式 25

$$\frac{\partial x_{cs}(t)}{\partial w_{(s,k)}} = f_k'(.)\, x_{cs}(t-1) + \partial \frac{\partial x_{cs}(t-1)}{\partial w_{(s,k)}}$$ 公式 26

表 6-2 Elman 预测模型变量说明

$w_{(i,k)}$	输入层到隐含层的权重矩阵
$w_{(k,j)}$	隐含层到输出层权重矩阵
$w_{(s,k)}$	承接层到隐含层的权重矩阵
$f(.)$	隐含层的激活函数
g(.)	输出层的激活函数
h(.)	接层的激活函数
x_c	承接层的输出
t	时间步长
η_i	$w_{(i,j)}$ 的学习率
δ_j	输出层神经元的梯度项
δ_k	隐含层神经元的梯度项
x_{cs}	承接层第 s 维输出
u_k	隐含层第 k 维输出
$y_j(t)$	第 j 个结点第 t 轮的输出值
$y_{aj}(t)$	第 t 轮第 j 个结点的标准输出值
$g_j^{'}(.)$	输出层的导数
$f_k^{'}(.)$	隐含层的导数

本文采用的是一次预报模型和实测的数据，并对其进行了预处理，填补了缺失值，处理了异常值，并进行了归一化。然后利用一次预报模型中的气候因子、一次预报模型中预测的 6 项污染物的浓度值、实测数据中的气候因子、实测数据中 24 小时前的污染物浓度历史数据作为输入变量，输入到 Elman 的空气质量二次预报模型中，输入变量如表 6-4。二次预报模型的参数如表 6-3，其中输入层结点数为 2，输出层结点个数设置为 1，隐含层个数为 10，最小误差为 0.01，最大训练轮数 10000，隐含层激活函数使用了 sigmoid 函数，输出层激活函数使用了 Purclin 函数。

表 6-3 Elman 的空气质量二次预报模型参数

输入层结点数	2
输出层结点数	1
隐含层个数	10
最小误差	0.01
最大训练轮数	10000
隐含层激活函数	sigmoid 函数
输出层激活函数	Purclin 函数

表 6-4 基于 Elman 的小时值二次预报模型输入值

影响因子	小时预报模型
时间因子	1h（小时）
空间因子	A、B、C 三个监测站
预报气象因子	近地 2 米温度(℃)、地表温度(K)、比湿(kg/kg)、湿度(%)、近地 10 米风速(m/s)、近地 10 米风向(°)、雨量(mm)、云量、边界层高度(m)、大气压(Kpa)、感热通量(W/m^2)、潜热通量(W/m^2)、长波辐射(W/m^2)、短波辐射(W/m^2)、地面太阳能辐射(W/m^2)
实测气象因子	温度(℃)、湿度(%) 、气压(MBar)、风速(m/s)、风向(°)
预报污染物	SO2 预报浓度($\mu g/m^3$)、NO2 预报浓度($\mu g/m^3$)、PM10 预报浓度($\mu g/m^3$) 、PM2.5 预报浓度($\mu g/m^3$)、O3 预报浓度($\mu g/m^3$)、CO 预报浓度(mg/m^3)
历史监测污染物	当前预报时刻前 24 小时的 SO2 监测浓度($\mu g/m^3$)、NO2 监测浓度($\mu g/m^3$)、PM10 监测浓度($\mu g/m^3$)、PM2.5 监测浓度($\mu g/m^3$)、O3 监测浓度($\mu g/m^3$)、CO 监测浓度(mg/m^3)

Elman 神经网络模型的算法学习流程伪代码如下表 6-5 所示：

表 6-5 Elman 神经网络模型的算法学习流程伪代码

训练集 $D=\{(x_t, y_t)\}$
1: 初始化 Elman 神经网络中所有连接权重和阈值。
2：for all $(x_t, y_t)\in D$ do
3： 根据公式计算每一层的输出值；
4: 根据公式计算输出层神经元梯度项；
5: 根据公式计算隐含层神经元梯度项；
6: 根据公式更新权值；
7: if(达到停止条件)
8: break；
9: end if
10:end for

6.2.4 模型对比结果

本文采用均方根误差（RMSE）、平均绝对误差（MAE）和一致性指数（IOA）来分析评价模型的准确性，指标计算公式如下：

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i-Q_i)^2}, i=1,...,N \qquad \text{公式 27}$$

$$MAE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}|\mathrm{P_i}-Q_i|},i=1,...,N \quad \text{公式 28}$$

$$IOA=1-\frac{\sum_{i=1}^{N}(\mathrm{P_i}-Q_i)^2}{\sum_{i=1}^{N}(|\mathrm{Q_i}-\bar{Q}|+|\mathrm{P_i}-\bar{Q}|)^2}\ i=1,...,\mathrm{N} \quad \text{公式 29}$$

其中 P_i 为模型预测值；Q_i 为实际观测值；$\bar{Q}$ 为实际观测值的平均值。

利用 RMSE、MAE 和 IOA 对比原始 CMAQ 一次预报模型、SVR 预测模型、Elman 预测模型和 DBN-BP 预测模型的预报效果，结果如下所示，可以看出添加了实测数据的二次预报模型效果**均好于** CMAQ 一次预报模型。

在 SVR、DBN-BP、Elman 三个二次预报模型中，**Elman 预测模型效果最佳**。

表 6-6　四种预测模型效果对比

	PM2.5				**PM10**			
评价函数	CMAQ	SVR	Elman	DBN-BP	CMAQ	SVR	Elman	DBN-BP
RMSE	36.39	28.18	**6.68**	11.18	65.85	44.79	**18.78**	22.25
MAE	25.63	21.38	**6.2**	8.22	50.66	19.76	**11**	14.63
IOA	0.69	0.77	**0.89**	0.87	0.54	0.67	**0.90**	0.82

	SO2				**CO**			
评价函数	CMAQ	SVR	Elman	DBN-BP	CMAQ	SVR	Elman	DBN-BP
RMSE	41.50	30.75	**8.97**	10.18	0.56	0.33	**0.201**	0.25
MAE	29.55	18.36	**5.31**	6.95	0.42	0.30	**0.08**	0.19
IOA	0.54	0.72	**0.91**	0.89	0.72	0.82	**0.94**	0.93

	NO2				**O3**			
评价函数	CMAQ	SVR	Elman	DBN-BP	CMAQ	SVR	Elman	DBN-BP
RMSE	24.19	20.46	**12.33**	14.63	60.58	50.88	**15.46**	17.13
MAE	19.07	18.87	**7**	11.63	37.78	34.67	**10.34**	12.86
IOA	0.58	0.63	**0.88**	0.86	0.40	0.78	**0.92**	0.90

从下述拟合图可以看出，二次预报模型预测的污染物浓度值比 CMAQ 一次预报更贴近实测值，同时 Elman 二次预测模型效果优于 DBN-BP 预测模型和 SVR 预测模型。因此，本文建立基于 Elman 的空气质量二次预报模型，对监测点 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。

图 6-7 四种预测模型拟合图

6.3 基于 Elman 的空气质量二次预报模型结果分析

根据 Elman 空气质量二次预报模型，预测出监测点 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，从而依据前述计算 AQI 的方法，得到如下表 6-7 结果。

表 6-7 问题三空气质量二次预报模型结果

预报日期	地点	二次模型日值预测 （单位：μg/m³ mg/m³）							
		SO_2	NO_2	PM_{10}	$PM_{2.5}$	O_3最大八小时滑动平均	CO	AQI	首要污染物
2021/7/13	监测点 A	5.02	10.86	18.86	4.68	77.99	0.39	39	无
2021/7/14	监测点 A	4.92	9.22	17.81	3.86	78.25	0.38	40	无
2021/7/15	监测点 A	4.46	8.56	16.69	3.70	78.74	0.37	40	无
2021/7/13	监测点 B	4.99	7.49	17.15	3.48	43.66	0.55	22	无
2021/7/14	监测点 B	4.64	6.49	17.35	3.63	43.91	0.49	22	无
2021/7/15	监测点 B	3.45	5.08	17.00	3.44	44.97	0.46	23	无
2021/7/13	监测点 C	9.08	18.50	32.26	13.57	116.90	0.53	65	O3
2021/7/14	监测点 C	8.05	17.48	30.40	12.55	117.89	0.49	65	O3
2021/7/15	监测点 C	7.54	17.32	20.34	11.08	118.37	0.48	66	O3

7 问题四：基于 BMA 的空气质量邻近区域协同预报模型

7.1 问题提出与解决思路

本文认为区域相邻的监测站之间，气候因子是大致相似的，对于各个监测站预报出的不同数值，可以使用贝叶斯模型平均法 BMA（EM 算法）综合各个监测站的预报值，得出一个更为精确的，有代表性的值。从空间角度来看，临近监测站之间的污染物会由于风的因素，被从一个监测站带到另外一个邻近区域。因此，可以利用 wiener 修正模型，在预报结果上加入风速和风向影响因素，总结出风速、风向和实际值之间的非线性关系，对预报结果进行修正。本题的研究思路如下图 7-1 所示。

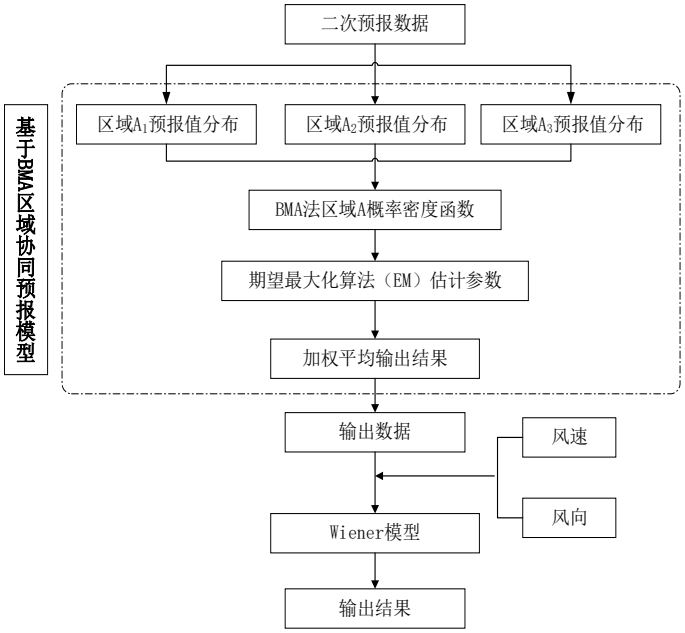


图 7-1 问题四研究思路图

7.2 研究数据处理

7.2.1 探索性数据分析

本文使用了附件一和附件三中的数据，包含了 A、A1、A2、A3 四个临近区域的数据，在对四个临近区域的真实监测数据，做 6 项污染物浓度和气候因子的相关性分析时，可以发现，A、A1、A2、A3 四个临近区域之间的气候因子和污染物浓度变化之间有着相似的趋势，如图所示（以 O3 为例）。因此，本文认为用临近区域，例如：用 A1、A2、A3 的数据预测 A 的数据、用 A、A2、A3 的数据去预测 A1 的数据，是存在一定的合理性的。

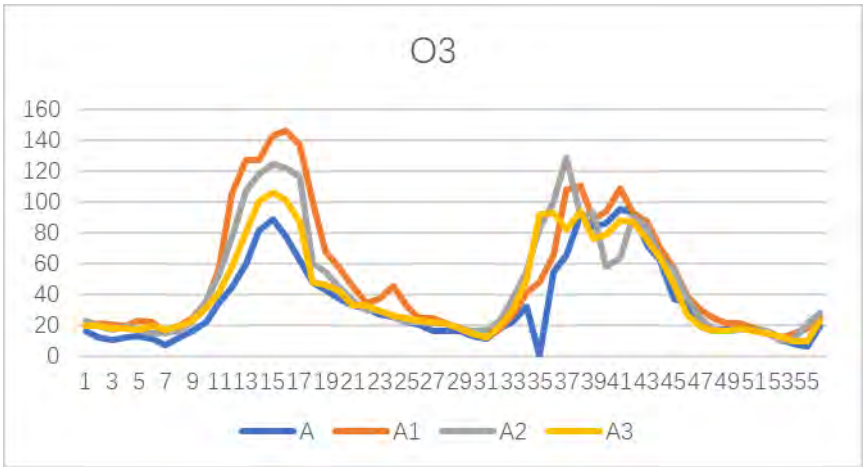


图 7-2 邻近区域 03 浓度值变化

7.2.2 风向指标划分

从空间角度来看，临近监测站之间的污染物会由于风的因素，被从一个监测站带到另外一个邻近区域。本文按照 360 度划分风向，划分为 4 类风向。

表 7-1 风向划分

风向	角度范围
东风	45-135
南风	135-225
西风	225-315
北风	315-360

由表 7-2 可知，除臭氧（O_3）的污染物与风速呈负相关，风速越大，浓度越低，臭氧与之相反。此外，偏西风时，风速越低时，污染物浓度越高。因此，本文利用 wiener 修正模型，在出来的结果上加入风速和风向影响因素，总结出风速、风向和实际值之间的非线性关系，对结果进行修正。

表 7-2　不同风向下各污染物浓度

	SO_2	NO_2	PM_{10}	$PM_{2.5}$	O_3	CO	风速
北风	5.25	20.14666387	18.90	3.93	80.17	0.33	1.51
东风	4.01	17.75190009	18.51	3.05	95.59	0.31	1.76
南风	4.62	29.96305414	19.59	4.56	74.56	0.34	1.42
西风	5.61	36.44269774	19.63	7.48	45.79	0.38	1.31

7.3 领域空气质量协同预报模型

7.3.1 贝叶斯模型平均法 BMA

BMA 采用贝叶斯公式，结合了先验分布与似然函数[12]，通过分析各预报模型的后验分布，调整权重，提供一个比单个模型预报更准确的预测值。

$$p(y|D)=\sum_{j=1}^{K}P(M_j|D)P(y|M_j,D) \qquad \text{公式 30}$$

其中 $P(M_j|D)$ 是模型 M_j 的后验概率即模型 M_j 的权重，$P(y|M_j,D)$ 为模型 M_j 的后验密度分布，y 为预测变量，D 为实测样本数据，$p(y|D)$ 为预测变量 y 的后验密度分布。后验概率 $P(M_j|D)$ 公式如下所示：

$$P(M_j|D)=\frac{P(D|M_j)P(M_j)}{\sum_{h=1}^{K}P(D|M_h)P(M_h)},P(D|M_j)=\int P(D|\theta_j,M_j)P(\theta_j|M_j)d_{\theta_j} \qquad \text{公式 31}$$

上述公式中 $P(M_j)$ 为模型 M_j 对应的参数向量。对公式 30 两边同时取期望可得到预测变量 y 的综合预测值，式中 $P(M_j|D)$ 为权重即 $w_j=P(M_j|D)$，f_j 为模型 M_j 的预测值。

$$E(y|D)=\sum_{j=1}^{K}P(M_j|D)E\left[(y|M_j,D)\right]=\sum_{j=1}^{K}w_jf_j \qquad \text{公式 32}$$

表 7-3　气象因子权重值（BMA 法）

BMA-weight	A	A1	A2	A3
湿度	0.281705	0.261805	0.246197	0.210283
气压	0.295479	0.283352	0.281055	0.142808
温度	0.277738	0.265597	0.245762	0.208803

BMA 采用贝叶斯公式，结合了先验分布与似然函数，通过分析各预报模型的后验分布，对气象因子重新调整了权重。

7.3.2 Wiener 模型

Wiener 模型是由静态和动态两部分模块组成的非线性模型[13]，通过串联相连。如下图 7-3 所示。

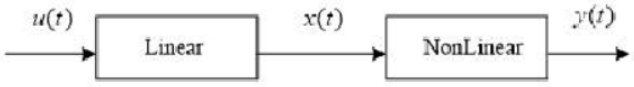


图 7-3 wiener 模型

定义$\theta=\left[a_1,\cdots,a_n,b_0,\cdots,b_m,D,b\right]$为 Wiener 模型的参数向量。辨识的目的是根据系统的输入$u(k)$和输出$y(k)$估计模型的参数向量θ。设$\hat{\theta}=\left[\hat{a}_1,\cdots,\hat{a}_n,\hat{b}_0,\cdots,\hat{b}_m,\hat{D},\hat{b}\right]$是参数向量$\theta$的估计值。假设 Wiener 模型的结构是已知的，即$A\left(q^{-1}\right)$和$B\left(q^{-1}\right)$的阶次 n，m 以及延迟时间 d 都是已知的。辨识的 Wiener 模型可表示为：

$$\begin{cases}\hat{A}\left(q^{-1}\right)\hat{x}(k)=q^{-d}\hat{B}\left(q^{-1}\right)u(k)\\ \hat{y}(k)=f\left(\hat{X}(K)\right)+e(k)\end{cases}$$

公式 33

其中，

$$\hat{A}\left(q^{-1}\right)=1+\hat{a}_1q^{-1}+\cdots+\hat{a}_nq^{-n}$$

公式 34

$$\hat{B}\left(q^{-1}\right)=1+\hat{b}_1q^{-1}+\cdots+\hat{b}_mq^{-m}$$

公式 35

$\hat{y}(k)$是辨识模型的输出。为了估计参数向量θ，我们定义预测误差函数为：

$$\hat{J}(q)=\sum_{k=1}^{j}\left[y(k)-\hat{y}(k)\right]^2$$

公式 36

其中，L 为用于辨识的数据点个数，$y(k)$和$\hat{y}(k)$分别是实际系统和辨识模型的输出。Wiener 模型的参数估计就是利用最小化预测误差函数来完成的，即

$$\hat{\theta}^*=\min_{\theta\in T}J\left(\hat{\theta}\right)$$

公式 37

由问题二的相关性分析，已知 6 种污染物的浓度与风速和风向是息息相关的，因此先利用模型三得出的结果，并且通过了 BMA 处理，在变量不变的基础上，**新增了风速和风向变量**，对预报值继续进行修正。

7.4模型结果分析

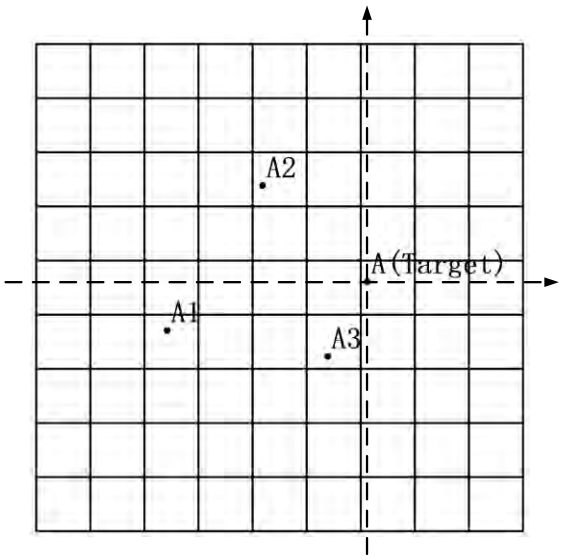
7.4.1 邻近区域影响

由第四题题干中的信息，已知临近的四个监测点站 A、A1、A2、A3 的具体的地理位置信息。根据他们彼此的地理位置，利用 wiener 模型对会产生影响的风向进行了归纳总结。

表 7-4　各监测点风向归纳表

预测站点	A	A1	A2	A3
A	-	西北	西北	西南
A1	东北	-	东北	东南
A2	东南	西南	-	东南
A3	东北	西北	西北	-

本文认为，风向和风速会导致相邻区域污染物的传输和扩散。当 A 监测点为西风时，各污染物浓度较高，当 A1 监测点为东风时，各污染物浓度较高，当 A2 监测点为南风时，各污染物浓度较高，当 A3 监测点为北风时，各污染物浓度较高。



注：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)

图 7-4 各监测站点相对位置示意图

7.4.2 修正预报模型对比

本文利用 wiener 模型将风速和风向两个参数单独拿出来，对预报结果进行修正。其中 wiener 模型中的输入变量为经过 BMA 调整后的 elman 模型预报值和预测时点的前 24 小时的风速和风向，对应的是可以输出当天的预测时点的污染物浓度修正值。

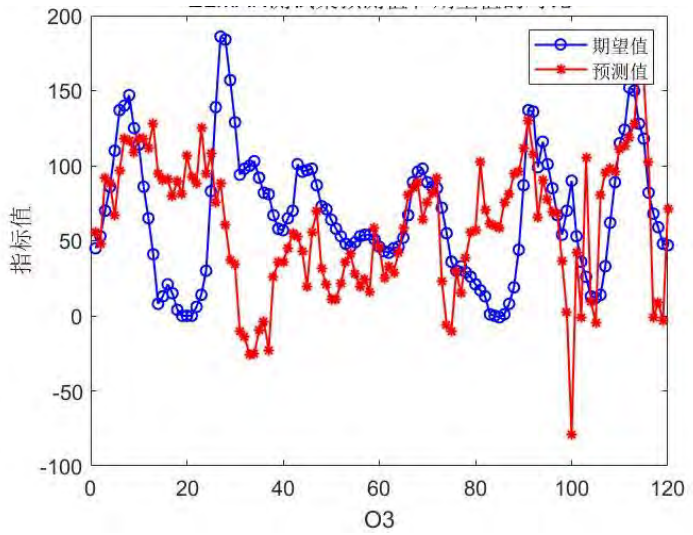


图 7-5　预报结果对比（以 03 为例）

根据基于 wiener 模型的空气质量协同预报模型，预测出监测点 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，从而依据前述计算 AQI 的方法，得到如下表 7-5 结果。

表 7-5　基于 wiener 模型的空气质量协同预报模型预测结果

预报日期	地点	二次模型日值预测 （单位：μg/m³　mg/m³）							
		SO_2	NO_2	PM_{10}	$PM_{2.5}$	O_3 最大八小时滑动平均	CO	AQI	首要污染物
2021/7/13	监测点 A	5.85	18.23	24.6	11.13	89.9	0.47	45	无
2021/7/14	监测点 A	6.24	16.98	23.75	11.25	97.4	0.41	49	无
2021/7/15	监测点 A	5.91	18.46	26.3	11.20	107.21	0.43	57	O3
2021/7/13	监测点 A1	7.21	19.37	28.31	16.44	121.5	0.53	68	O3
2021/7/14	监测点 A1	7.82	18.93	29.46	17.31	132.16	0.55	77	O3
2021/7/15	监测点 A1	7.53	20.16	30.21	16.93	128.72	0.53	74	O3
2021/7/13	监测点 A2	6.41	17.63	26.59	13.82	98.62	0.48	49	无
2021/7/14	监测点 A2	6.83	17.71	24.57	14.19	109.51	0.44	58	O3
2021/7/15	监测点 A2	6.27	19.64	28.06	14.08	110.34	0.51	59	O3

7.5 区域协同预报模型优化？

针对监测点 A 的污染物浓度预测，本文建立的基于 BMA 模型的空气质量协同预报模型，**综合考虑了临近点 A1、A2、A3 的相关数据**，以及在预报结果的基础上添加**风向和风速**对区域协同预报模型进行修正，使用的修正模型为 **wiener 模型**。将风向划分为东南西北四个风向，建立各个方向的修正模型，发现**能有效提高预测的精度**，相关结果如下所示。

表 7-6　不同修正模型效果对比

	PM2.5			PM10		
评价函数	CMAQ	Elman	wiener	CMAQ	Elman	wiener
RMSE	36.39	6.68	**6.57**	65.85	18.78	**15.83**
MAE	25.63	6.2	**5.32**	50.66	11	**9.35**
IOA	0.69	0.89	**0.93**	0.54	0.90	**0.98**

	SO2			CO		
评价函数	CMAQ	Elman	wiener	CMAQ	Elman	wiener
RMSE	41.50	8.97	**7.31**	0.56	0.201	**0.17**
MAE	29.55	5.31	**4.37**	0.42	0.08	**0.06**
IOA	0.54	0.91	**0.94**	0.72	0.94	**0.97**

	NO2			O3		
评价函数	CMAQ	Elman	wiener	CMAQ	Elman	wiener
RMSE	24.19	12.33	**11.48**	60.58	15.46	**13.73**
MAE	19.07	7	**6.36**	37.78	10.34	**9.15**
IOA	0.58	0.88	**0.91**	0.40	0.92	**0.95**

区域协同预报模型能优化问题 3 建立的基于 Elman 的空气质量二次预报模型预测精度，主要原因是**各气象因素是相互影响和制约的**，综合考虑附近区域的气象条件对测算区域的污染物影响，可以使得预测数据更接近真实值，容错率会降低。

8 模型评价与推广

8.1 模型优点

1. 对原始数据进行了较为合理且充分的预处理，同时还考虑到气象监测数据采集设备自身存在的白噪声误差，使用异常值 3σ 准则对数据进行了处理，提高模型精度。

2. 针对问题二，聚类时选取 K-means、GMM、HAC、AP 聚类进行科学对比，为后续进行气象特征分类打下夯实基础。同时在分类过后，利用熵权法对气象因子影响度进行调整，进一步增强分类的合理性。

3. 通过支持向量机回归（SVR）、Elman 模型、DBN-BP 模型对污染物浓度的预测结果对比，用 RMSE、MAE 和 IOA 这三个评价指标，选取出预测精度最高的模型。该预测模型精度高、收敛速度快并且**鲁棒性强**。

4. 本文使用的基于 Elman 神经网络的预报结果优化算法，结合了对数值模式的预报和人工神经网络预报模型的优点，能提高预报结果的准确性。

8.2 模型不足

在变量特征提取筛选的过程中，通过不同变量簇之间的聚类及计算信息增益值可得到相关结果，根据特定规则进行变量数筛选，后续需选取更加科学的筛选指标构建适用于气

象特征的变量筛选体系。

8.3 模型改进

总体而言，找到了合适且有价值的问题去分析研究，但是附件中仍有大部分数据的价值没有被开发，后期或许可以从更细节的角度去发掘，例如更加细致的考虑包括臭氧在内的污染物生成机理、进一步提升模拟气象场的真实性。

为了使模型的鲁棒性能强，则需在模型的训练过程中应用大量的数据集。

8.4 模型推广

在后续研究中，为保证同时兼顾数据的深度与广度，首先需提升硬件性能，对数据集进行多次训练，不断优化模型，提升空气质量预报模型预测精度。其次改进之处在于，对污染物浓度的数值以及变化趋势进行分地域研究，或进一步考虑更多的空间分布特征及地形特性，提高区域协同预报的精度。

本文在研究空气质量建模过程中，并未使用传统的统计预测模型，但是模型的准确率较高，因此在气象学领域的从业人员可以考虑使用本文建立的模型与传统气象模型相互验证。此外，本文使用的 wiener 模型在化学、通信等众多领域中都能得到运用[13]。

参考文献

[1]熊帅晨. 基于分形插值的空气质量指数混合预测模型[D].上海师范大学,2021.

[2]魏丽欣, 张良玉, 王欢. 保定市大气污染与气象条件的影响分析和模拟研究[J]. 环境与发展，2018, 30(08):162-163.

[3]夏起铁.基于机器学习技术的城市空气质量预测研究[J].信息记录材料,2020,21(12):89-90.

[4]祁栋林,张加昆,李晓东,魏鸿业,王力,马明亮,孔维强,肖宏斌,张娟.2001—2011 年西宁市空气质量特征及其与气象条件的关系[J].气象与环境学报,2014,30(02):51-59.

[5]王祥炳,黄晓容,邓茂,张永江,李清芳.重庆市黔江区春季大气污染物浓度变化与气象因素关系研究[J].环境科学与管理,2016,41(06):77-81.

[6] Shann once. A mathematical theory of communication[J]BellSystemTech,1948,27(3):379-423.

[7]王莉亚,张志强.基于信息熵的信息整合主题演化研究[J].图书情报工作,2012,(06):102-106.

[8]熊帅晨. 基于分形插值的空气质量指数混合预测模型[D].上海师范大学,2021.

[9]马元婧. 基于深度学习的大气环境监测系统关键技术研究[D].中国科学院大学(中国科学院沈阳计算技术研究所),2021.

[10] Guo R F, Ma Y J, Wang S, et al. Establishment of Air Quality Forecast Model Based on Deep Learning [C]. 2020 IEEE 6th International Conference on Computer and Communications.2020:1500-1504.

[11]于海飞. 基于多模式预报的空气质量预警系统设计与实现[D].中国科学院大学(中国科学院沈阳计算技术研究所),2020.

[12]文琴. 基于集合预报的成都市空气质量指数预测研究[D].成都信息工程大学,2018.

[13]李海琴. 基于时空数据驱动的宁波市空气质量预警[D].宁波大学,2017.

附录 实测数据缺失小时明细表

2019/5/6 15:00	2020/1/14 3:00	2020/4/24 23:00	2020/4/26 5:00	2021/5/2 0:00	2021/7/3 15:00	2021/7/4 21:00	2021/7/6 3:00
2019/10/11:00	2020/1/15 2:00	2020/4/25 0:00	2020/4/26 6:00	2021/5/2 1:00	2021/7/3 16:00	2021/7/4 22:00	2021/7/6 4:00
2019/11/21 3:00	2020/1/16 2:00	2020/4/25 1:00	2020/4/26 7:00	2021/5/2 2:00	2021/7/3 17:00	2021/7/4 23:00	2021/7/6 5:00
2019/11/22 2:00	2020/1/18 2:00	2020/4/25 2:00	2020/4/26 8:00	2021/5/2 3:00	2021/7/3 18:00	2021/7/5 0:00	2021/7/6 6:00
2019/11/23 2:00	2020/1/19 2:00	2020/4/25 3:00	2020/4/26 9:00	2021/5/2 4:00	2021/7/3 19:00	2021/7/5 1:00	2021/7/6 7:00
2019/11/24 2:00	2020/1/22 2:00	2020/4/25 4:00	2020/4/26 10:00	2021/5/2 5:00	2021/7/3 20:00	2021/7/5 2:00	2021/7/6 8:00
2019/11/26 2:00	2020/1/31 2:00	2020/4/25 5:00	2020/5/17 5:00	2021/5/2 6:00	2021/7/3 21:00	2021/7/5 3:00	2021/7/6 9:00
2019/11/27 2:00	2020/2/2 2:00	2020/4/25 6:00	2020/5/17 6:00	2021/5/2 7:00	2021/7/3 22:00	2021/7/5 4:00	2021/7/6 10:00
2019/11/28 2:00	2020/2/5 2:00	2020/4/25 7:00	2020/5/17 7:00	2021/5/2 8:00	2021/7/3 23:00	2021/7/5 5:00	2021/7/6 11:00
2019/12/8 10:00	2020/2/6 2:00	2020/4/25 8:00	2020/8/2 11:00	2021/5/2 9:00	2021/7/4 0:00	2021/7/5 6:00	2021/7/6 12:00
2019/12/8 11:00	2020/2/15 1:00	2020/4/25 9:00	2020/8/2 12:00	2021/5/2 10:00	2021/7/4 1:00	2021/7/5 7:00	2021/7/6 13:00
2019/12/8 12:00	2020/2/16 1:00	2020/4/25 10:00	2020/8/2 13:00	2021/5/2 11:00	2021/7/4 2:00	2021/7/5 8:00	2021/7/6 14:00
2019/12/16 3:00	2020/2/17 1:00	2020/4/25 11:00	2020/11/21 11:00	2021/5/2 12:00	2021/7/4 3:00	2021/7/5 9:00	2021/7/6 15:00
2019/12/17 3:00	2020/2/19 1:00	2020/4/25 12:00	2020/11/21 12:00	2021/5/2 13:00	2021/7/4 4:00	2021/7/5 10:00	2021/7/6 16:00
2019/12/18 3:00	2020/2/25 3:00	2020/4/25 13:00	2020/11/21 13:00	2021/5/2 14:00	2021/7/4 5:00	2021/7/5 11:00	2021/7/6 17:00
2019/12/21 3:00	2020/2/26 3:00	2020/4/25 14:00	2020/11/21 14:00	2021/5/2 15:00	2021/7/4 6:00	2021/7/5 12:00	2021/7/6 18:00
2019/12/22 3:00	2020/3/13:00	2020/4/25 15:00	2020/11/21 15:00	2021/5/2 16:00	2021/7/4 7:00	2021/7/5 13:00	2021/7/6 19:00
2019/12/23 3:00	2020/3/23:00	2020/4/25 16:00	2020/11/21 16:00	2021/5/2 17:00	2021/7/4 8:00	2021/7/5 14:00	2021/7/6 20:00
2019/12/24 3:00	2020/3/63:00	2020/4/25 17:00	2020/11/21 17:00	2021/5/2 18:00	2021/7/4 9:00	2021/7/5 15:00	2021/7/6 21:00
2019/12/28 3:00	2020/3/83:00	2020/4/25 18:00	2021/1/10:00	2021/7/3 4:00	2021/7/4 10:00	2021/7/5 16:00	2021/7/6 22:00
2019/12/29 3:00	2020/3/27 2:00	2020/4/25 19:00	2021/5/114:00	2021/7/3 5:00	2021/7/4 11:00	2021/7/5 17:00	2021/7/6 23:00

2019/12/30 3:00	2020/3/29 1:00	2020/4/25 20:00	2021/5/1 15:00	2021/7/3 6:00	2021/7/4 12:00	2021/7/5 18:00	
2019/12/31 3:00	2020/3/30 1:00	2020/4/25 21:00	2021/5/1 16:00	2021/7/3 7:00	2021/7/4 13:00	2021/7/5 19:00	
2020/1/1 0:00	2020/4/24 16:00	2020/4/25 22:00	2021/5/1 17:00	2021/7/3 8:00	2021/7/4 14:00	2021/7/5 20:00	
2020/1/1 3:00	2020/4/24 17:00	2020/4/25 23:00	2021/5/1 18:00	2021/7/3 9:00	2021/7/4 15:00	2021/7/5 21:00	
2020/1/4 3:00	2020/4/24 18:00	2020/4/26 0:00	2021/5/1 19:00	2021/7/3 10:00	2021/7/4 16:00	2021/7/5 22:00	
2020/1/5 3:00	2020/4/24 19:00	2020/4/26 1:00	2021/5/1 20:00	2021/7/3 11:00	2021/7/4 17:00	2021/7/5 23:00	
2020/1/6 3:00	2020/4/24 20:00	2020/4/26 2:00	2021/5/1 21:00	2021/7/3 12:00	2021/7/4 18:00	2021/7/6 0:00	
2020/1/8 3:00	2020/4/24 21:00	2020/4/26 3:00	2021/5/1 22:00	2021/7/3 13:00	2021/7/4 19:00	2021/7/6 1:00	
2020/1/12 3:00	2020/4/24 22:00	2020/4/26 4:00	2021/5/1 23:00	2021/7/3 14:00	2021/7/4 20:00	2021/7/6 2:00	