**2024年扬州大学第三届研究生数学建模竞赛**

　　我们完全明白，在竞赛开始后参赛队员不能以任何方式与对外的任何人研究、讨论与赛题有关的问题。

　　我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用出和参考文献中明确列出。

　　我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **队员姓名** | **年级** | **学院** | **公选课(填是或否)** |
| **1.** | **刘尊续** | **研1** | **信息工程学院** | **否** |
| **2.** | **郑亚澎** | **研1** | **信息工程学院** | **否** |
| **3.** | **李鹏海** | **研1** | **信息工程学院** | **是** |

**注：关于“公选课”栏的填写，请本学期选修研究生《数学建模》公选课程的同学填“是”，其他同学填“否” 。**



**2024年扬州大学第三届研究生数学建模竞赛**

**题 目 空气质量预测建模**

摘 要：

随着可持续发展战略的推进和持续深入，空气质量日益成为人们生活关注的重点。建立准确的空气质量预测模型可以为政府决策的制定和居民活动提供有用的信息。本文以提高预测准确率为目标，结合实测数据和不同问题需求分别建立多种空气质量预测模型，分析比较不同模型性能。

针对问题一，在**不考虑气象条件**的基础上，计算监测点从2022年8月25日到8月28日每天实测的空气质量和首要污染物。首先，我们**考虑了污染物浓度数据在实际采集过程中的可能误差**，因此采用**模型**，对2020年8月25日至8月28日监测点的主要污染物浓度进行了预测。模型预测基于已知的历史数据，成功预测了、、等污染物的浓度。通过计算每种污染物的空气质量分指数，并选取这些值中的最高者作为整体的空气质量指数，结果如表5所示。

针对问题二，我们分析了气象条件如温度、湿度、风速和降雨对空气质量指标（）及主要空气污染物（、、、、和）浓度的影响。首先，我们对数据进行了预处理，包括**填充缺失值**、**转换数据格式**和**去除异常值**，确保数据质量满足建模要求。之后我们以温度、湿度、风速和气压作为自变量，以为因变量，建立**多元线性回归方法**，以研究它们对空气质量的具体影响。其中，为了减轻模型的过拟合问题并提高预测的可解释性，采用**主成分分析（）法**来减少维度。最后，通过**值**来评价模型对数据变异性的解释能力，使用**检验**来评估模型中各变量的系数是否显著不同于零。

问题三的目的是建立一个通用的空气质量预测模型，适用于多个监测点。为此，我们采用了**随机森林算法**，由于其在处理高维数据和非线性问题上的优势，特别适合复杂的环境数据集。首先，我们整合了所有监测点的空气质量和气象数据，包括污染物浓度和天气条件等。随后，数据经过**标准化处理**，确保输入特征在模型中保持一致性。在模型建立过程中，随机森林被用来预测不同污染物的浓度，其输出帮助判断首要污染物。模型的预测准确性通过、与**精确度**三个指标进行了评估，结果表明模型具有良好的泛化能力，问题三预测结果如表7所示。

**关键字**：空气质量预测，，模型，，多元线性回归**，**随机森林算法

目录

[一、问题背景与重述 4](#_Toc166483489)

[1.1问题背景 4](#_Toc166483490)

[二、模型假设 5](#_Toc166483491)

[三、问题一的建模与求解 6](#_Toc166483492)

[3.1问题分析 6](#_Toc166483493)

[3.2数据处理 6](#_Toc166483494)

[3.3建立模型 6](#_Toc166483495)

[3.4结果 8](#_Toc166483496)

[四、问题二的建模与求解 10](#_Toc166483497)

[4.1问题分析 10](#_Toc166483498)

[4.2数据处理 10](#_Toc166483499)

[4.3建立模型 11](#_Toc166483500)

[4.4结果 11](#_Toc166483501)

[五、问题三的建模与求解 13](#_Toc166483502)

[5.1问题分析 13](#_Toc166483503)

[5.2数据处理 13](#_Toc166483504)

[5.3建立模型 13](#_Toc166483505)

[5.4模型评价 14](#_Toc166483506)

[5.5结果 14](#_Toc166483507)

[六、参考文献 16](#_Toc166483508)

# 一、问题背景与重述

## 1.1问题背景

空气质量的监测和预测是现代环境科学中一个至关重要的研究领域。随着工业化进程的加速和城市化水平的提高，空气污染已成为影响公众健康和生态安全的一个主要问题。据《环境空气质量标准》（GB3095-2012）规定，空气质量的评价主要依据六种常规大气污染物的浓度，包括二氧化硫（）、二氧化氮（）、可吸入颗粒物（）、细颗粒物（）、臭氧（）和一氧化碳（）。其中，、和被认为是一次污染物，直接从源头释放到大气中；而则是典型的二次污染物，通过复杂的化学和光化学反应在大气中形成。

大气污染不仅由人类活动产生的排放所驱动，气象条件也在污染物扩散和沉降过程中发挥了关键作用。例如，低温和高湿条件可能促进某些污染物的生成，而适宜的风速和风向可以帮助污染物扩散，从而降低某地区的污染水平。因此，了解和预测气象条件对空气质量的具体影响是制定有效空气质量管理策略的关键。

1.2问题重述

结合气象条件和已知污染物浓度数据，建立数学模型以解决一下问题：

问题一：使用过去一年的数据，预测给定城市的2023年9月1日至9月3日的日均值和首要污染物。

问题二：根据历史气象和污染物数据，预测在气象条件显著变化时（如温度、风速的快速变化），特定日子的空气质量变化趋势。

问题三：为了进一步验证模型的可靠性和精确性，选择三个不同的城市（城市、城市、城市）作为案例研究，使用相同的数学模型预测这些城市在不同气象条件下的变化。

问题四：考虑到实际操作中可能出现的数据缺失或异常值问题，要求模型具有一定的鲁棒性，能够在数据不完整或部分指标不可用的情况下，依然保持预测准确性。

# 二、模型假设

1. 假设污染物的来源和类型在预测期间内相对稳定，不会因突发事件或季节性变化而有显著的改变。
2. 假设数据中的缺失值和异常值已通过适当的数据处理方法进行处理，确保模型输入的数据质量和完整性。
3. 假设、、三个监测点间的污染物扩散和交互作用可以忽略，每个监测点的空气质量独立于其他监测点。
4. 由于部分污染物如臭氧，其生成和消亡速度与气象条件如温度和日照强度有直接关联，假设这些关系在预测期间内保持不变，以便能够准确估算这些污染物的浓度变化。

# 三、问题一的建模与求解

## 3.1问题分析

对于问题一，关键在于使用历史污染物浓度和数据来预测未来特定日期的空气质量。这涉及到采用时间序列分析或机器学习方法来识别和预测污染物浓度及其对的影响。挑战主要包括如何精确地预测首要污染物，以及如何处理季节性变化和突发事件对空气质量的影响。

## 3.2数据处理

为了预测特定日期的空气质量指数（）和首要污染物，我们首先对历史污染物浓度数据和记录进行了彻底的清洗和预处理。这包括处理缺失值，通过线性插值法补全缺失数据，以及识别并处理异常值，通过箱线图和标准差方法剔除数据集中的离群点。

线性插值法：假设和是已知的数据点，并且缺失的数据点，则

此外，我们采用最小-最大缩放法对数据进行正规化处理，将所有特征缩放到同一量级，以消除不同测量单位带来的影响，确保模型的准确性和一致性。

正规化处理：假设为原始数据，为正规化后的数据，则

## 3.3建立模型

由于不同污染物的生成、扩散和化学性质可能会有很大的差异，这些差异会影响它们各自的时间序列特性，因此对于复杂的空气质量监测数据，需要对, , , ,, 分别单独建立模型。

下以为例，进行建模。

1.平稳性检验：

使用测试确定时间序列的平稳性。如果序列非平稳，执行差分操作。

结果显示，，时间序列非平稳，需要进行一次差分操作。

进行一次差分操作后数据的时间序列图如图1所示，已趋于平稳。

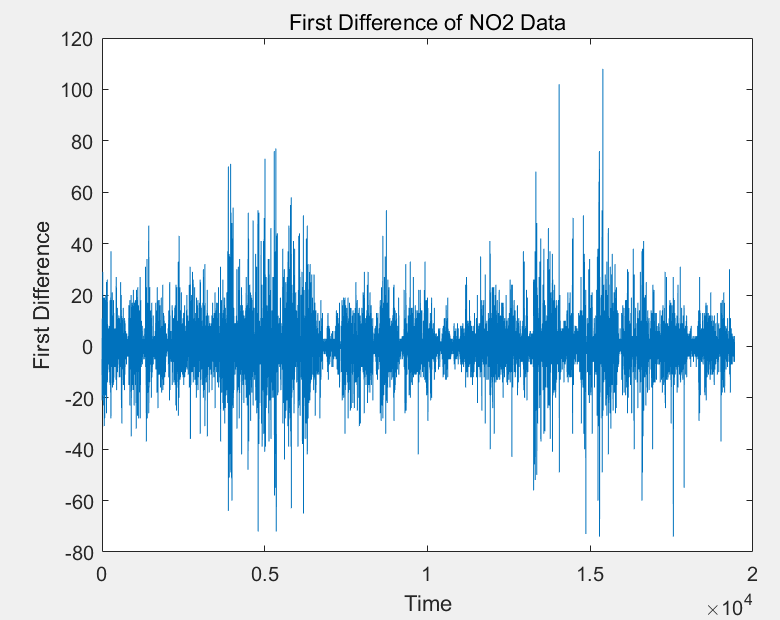


图1 差分后的时间序列图

2.模型参数选择：

对差分后的时间序列数据进行和分析，以确定模型的参数、和。

由差分操作仅进行了1次，得=1。

的和如图2、3所示。从图可以看出，序列在进行一次差分后已经显示出接近白噪声的性质，但第一个滞后的显著性表明移动平均部分应该设置为1。而图展示了第一个滞后有一个显著的尖峰，随后的所有滞后值都在置信区间内并且接近零，这表明自回归部分应该设置为1。

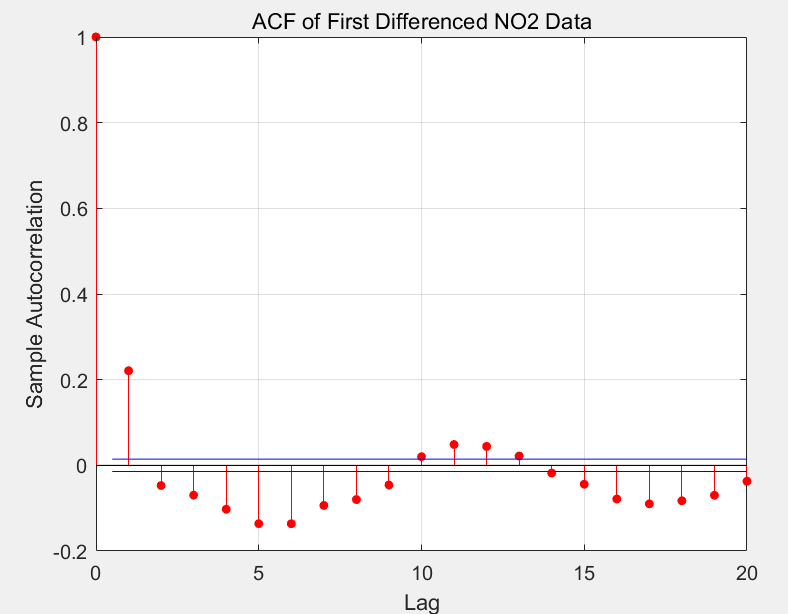
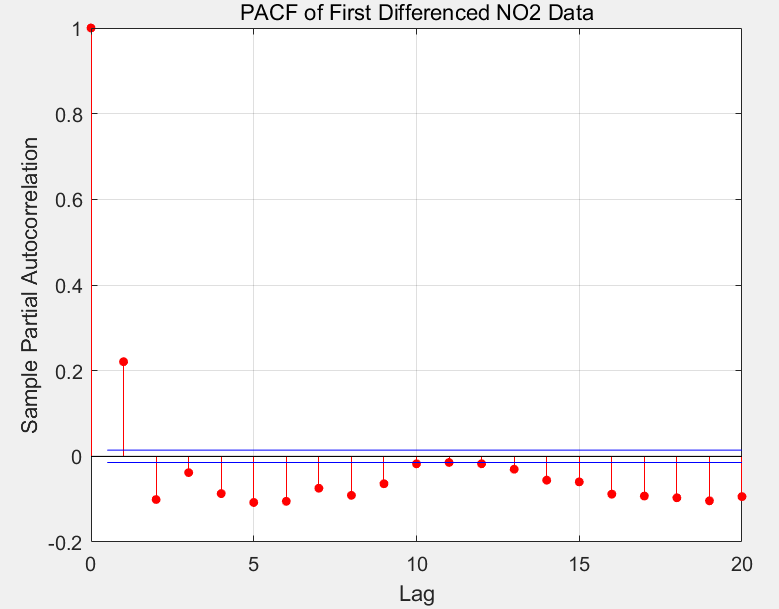
 

图2 ACF 图3 PACF

因此，针对污染物浓度预测，采用模型。

3.模型拟合与验证

使用实测参数对模型进行拟合和残差诊断，检验其有效性。

拟合结果如表1所示。

表1 拟合结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

项和项的显著性说明模型捕获了数据的主要动态。常数项的不显著可能意味着数据中没有必要的非零均值调整或模型已经通过差分处理有效地消除了数据的非平稳性。

残差诊断结果如图4所示。理想情况下，如果模型适配良好，残差的ACF应该显示所有滞后的自相关都在蓝线的置信区间内，接近于零。从图中可以看出，所有的自相关系数几乎都在置信区间内，表明残差接近白噪声。这是模型拟合得当的一个重要表现。

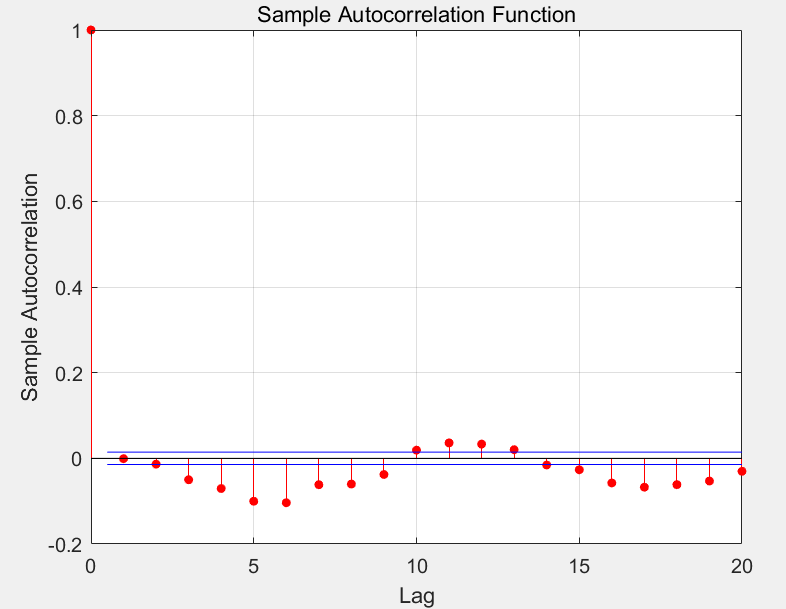


图4 残差诊断结果

由此可以得出，模型非常适合污染气体浓度的预测，能够有效捕捉数据时间序列的结构。

对, , ,, 污染气体浓度实测数据进行分析之后发现，模型依旧合适。

## 3.4结果

通过模型预测了, , , ,, 气体在2022年8月25日到8月28日的逐小时检测污染物浓度，再通过计算得到对应气体逐日监测浓度如表2所示。

表2 2022年8月25日到8月28日逐日污染物浓度预测

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | （单位：） |
| （单位：） | | | | |
| 2022/8/25 | 8 | 12 | 27 | 11 | 112 | 0.5 |
| 2022/8/26 | 7 | 16 | 24 | 10 | 92 | 0.5 |
| 2022/8/27 | 7 | 31 | 37 | 23 | 169 | 0.6 |
| 2022/8/28 | 8 | 30 | 47 | 33 | 201 | 0.7 |

查阅文献[1]，各项污染物项目浓度限值及对应的空气质量分指数级别如表3所示。取值对应的空气质量等级及首要污染物如表4所示。

表3 及对应污染物浓度限制

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 污染物 | 空气质量分指数及对应污染物浓度限制 | | | | | | | | 单位 |
|  | 0 | 50 | 100 | 150 | 200 | 300 | 400 | 500 | - |
|  | 0 | 2 | 4 | 14 | 24 | 36 | 48 | 60 |  |
|  | 0 | 50 | 150 | 475 | 800 | 1600 | 2100 | 2620 |  |
|  | 0 | 40 | 80 | 180 | 280 | 565 | 750 | 940 |
|  | 0 | 100 | 160 | 215 | 265 | 800 | - | - |
|  | 0 | 50 | 150 | 250 | 350 | 420 | 500 | 600 |
|  | 0 | 35 | 75 | 115 | 150 | 250 | 350 | 500 |

表4 取值对应空气质量等级及首要污染物

|  |  |  |
| --- | --- | --- |
| 范围 | 空气质量等级 | 首要污染物 |
| [0,50] | 优 | 无 |
| [51,100] | 良 | IAQI最大污染物 |
| [101,150] | 轻度污染 |
| [151,200] | 中度污染 |
| [201,300] | 重度污染 |
| [301,+∞] | 严重污染 |

是根据单个污染物浓度对应的空气质量分指数计算的，计算公式为：

其中，是污染物的空气质量分指数；是污染物的实际浓度；和 是污染物浓度的分界点，在这两个值之间；和是对应于和的空气质量分指数。

总AQI反映了空气质量的整体情况，以最不健康的单项污染物为准，即：

因此，结合表2、表3和表4计算可得问题一结果如表5所示。

表5 监测点的及首要污染物

|  |  |  |  |
| --- | --- | --- | --- |
| 监测日期 | 地点 | 计算 | |
|  | 首要污染物 |
| 2022/8/25 | 监测点 | 61 |  |
| 2022/8/26 | 监测点 | 45 | 无 |
| 2022/8/27 | 监测点 | 110 |  |
| 2022/8/28 | 监测点 | 136 |  |

# 四、问题二的建模与求解

## 4.1问题分析

问题二专注于探索不同气象条件对的影响。此问题要求分析温度、湿度、风速等气象因素与之间的关联，通常通过建立多变量统计模型来实施。这要求高度的数据完整性和处理技巧，特别是在处理气象数据的多重共线性问题时。

## 4.2数据处理

针对问题二，我们专注于分析气象条件如温度、湿度、风速等因素对的影响。我们首先集成并同步了气象数据与污染物浓度数据的时间戳，确保数据对齐精确。随后，使用多变量分析之前，同样采用最小-最大缩放法对数据进行了归一化处理，以平衡各特征的影响力。对于包含多重共线性的气象变量，我们应用了主成分分析（）来减少维度，并提取最有影响的特征，这有助于减轻模型的过拟合问题并提高预测的可解释性。

主成分分析（）[2]是一种常用的数据降维技术，其目的是在减少数据集中的变量数量的同时，尽可能保留原始数据中的信息。操作步骤如下

主成分分析（）是一种常用的数据降维技术，其目的是在减少数据集中的变量数量的同时，尽可能保留原始数据中的信息。这种方法通过以下步骤实现：

①标准化数据：

首先，对原始数据集中的每个特征进行标准化处理，以消除不同特征之间的量纲影响。标准化公式为：

其中，是第个样本的第个特征值，是第个特征的均值，是标准差。

②构造协方差矩阵：

接着，使用标准化后的数据构造协方差矩阵。协方差矩阵的元素由以下公式计算：

这里，表示特征和特征之间的协方差。

③计算特征值和特征向量：

计算协方差矩阵的特征值和特征向量。特征向量表示数据的主成分方向，而特征值则表示在这些方向上的方差大小。将特征值从大到小排序，选择前个最大的特征值对应的特征向量，这些向量构成了主成分。

④转换数据到新的特征空间：

使用选定的特征向量，将原始数据转换到由这些主成分定义的新特征空间。转换公式为：

其中，是选定的特征向量，是转换后的数据。

⑤解释方差比：

为了评估保留了多少原始数据的信息，可以计算保留的特征值所占的比例：

其中，表示第个主成分解释的方差比，表示前个主成分总共解释的方差比。

## 4.3建立模型

我们以为因变量，各气象因素为自变量，构建多元线性回归模型。模型形式如下：

其中，为截距项，、、、为各自变量的回归系数，。

采用检验来评估各回归系数的显著性，计算每个系数的值，确定其统计显著性。

计算调整后的值，评估模型解释变量对变化的解释程度。调整后的计算公式为：

其中，是样本大小，是模型中解释变量的数量。

## 4.4结果

根据实测逐小时气象数据、、各气象因素所建立的多元线性回归模型结果如表6所示。

表6 各气象因素与及污染物浓度的多元回归分析

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 温度 | 湿度 | 气压 | 风速 | 常数 |  |
|  | 35.446 | -103.767 | -0.253 | -122.7554 | 149.021 | 0.445 |
|  | -0.211 | -11.046 | -0.918 | -6.067 | 16.662 | 0.434 |
|  | -44.424 | -26.884 | -16.457 | -61.306 | 107.227 | 0.468 |
|  | -4.852 | -64.076 | 15.699 | -66.047 | 101.441 | 0.531 |
|  | -9.504 | -32.456 | 11.304 | -51.246 | 61.814 | 0.245 |
|  | 163.468 | -167.843 | 46.935 | -128.751 | 114.105 | 0.478 |
|  | -0.474 | -0.1776 | -0.126 | -0.5124 | 1.3676 | 0.328 |

通过表6可以发现：

1.温度:对的影响显著，系数为35.446，表明温度上升与增高正相关。这可能是由于高温加速了某些化学反应的速率，导致臭氧等二次污染物的增加。

2.湿度:对多数污染物的影响较小，但对的系数为11.304，这表明湿度增加可能会轻微增加的浓度。湿度可能影响颗粒物的水溶性和聚集性。

3.风速:对有显著正影响，系数为46.935，这可能因为风速的增加有助于扩散，从而减少了地面近处的污染物浓度。同时，风速的增加可能带来更多的清洁空气和加速污染物的稀释。

4.降雨:对有负影响，系数为-122.754，说明降雨有助于清除大气中的污染物，尤其是颗粒物。

5.模型的解释能力：的为0.445，这意味着模型能够解释大约44.5%的变异性。而对于，其值仅为0.245，表明模型对的解释能力较弱。

# 五、问题三的建模与求解

## 5.1问题分析

问题三要求开发一个能够适用于多个监测点的通用空气质量预测模型。这涉及到构建一个足够灵活以适应不同地理和环境条件的模型，同时保持预测的准确性和可靠性。这通常需要利用集成学习或多模型策略来增强模型的泛化能力和鲁棒性。

## 5.2数据处理

在构建适用于多个监测点的通用空气质量预测模型中，进行了彻底的数据标准化处理来确保数据在不同监测点间的一致性和可比性。标准化处理主要通过标准化实现，其公式如下：

其中，表示第个样本的第个特征值，是第个特征的均值，是对应的标准差。这种标准化方法使得每个特征的均值为0，标准差为1，从而消除了不同测量尺度带来的影响，并使模型更专注于数据的结构和模式。

通过这种方法，我们确保了模型能够有效地处理来自不同监测点的数据，而不受原始测量尺度的干扰，从而提高了预测模型的准确性和可靠性。

## 5.3建立模型

对处理过后的数据采用随机森林算法来建立空气质量预测模型。

随机森林[3]是一个强大的机器学习模型，通常用于处理分类和回归问题。它通过构建多棵决策树并整合它们的结果来提高预测准确性和防止过拟合。其执行步骤如下所示：

1. 决策树的构建

随机森林通过构建多棵决策树来工作，每棵树都是在数据集的一个随机子集上训练得到。每个树节点的分裂基于以下几种标准之一：

①信息增益：用于选择使信息增益最大化的分裂属性，公式为：

其中熵定义为：

②基尼不纯度：另一种选择分裂属性的方法，测量的是随机分类时的错误概率：

2. 随机森林的预测

在所有决策树建立完成后，随机森林的预测通过以下方式进行：

①回归问题：输出是所有决策树预测的平均值：

②分类问题：使用多数投票法来确定最终的类别：

3. 特征重要性

随机森林还可以评估各个特征对模型预测性能的影响，这是通过计算特征在所有树中减少不纯度的平均贡献来估计的：

随机森林模型不仅预测性能优秀，还可以处理大量数据中的复杂非线性关系，并提供对哪些变量最重要的直观理解，这使其在现代数据科学应用中非常有价值。

## 5.4模型评价

模型训练完成之后，采用均方误差（）、决定系数（）和精确度来评价模型的预测能力。

是衡量模型预测值与实际值之间差异的常用指标。它计算了每个预测值与实际值差的平方的平均值，公式如下：

其中，是模型对第个观察值的预测，是实际观察值，是观察值的总数。的值越小，表示模型的预测值与实际值之间的差异越小，模型的预测性能越好。

，又称拟合优度[4]，是反映模型解释变量变异性的比例。它表明了因变量的变异中有多少能够通过自变量来解释。其公式为：

其中，是 的均值。的值通常介于0和1之间，值越接近1，说明模型的解释能力越强，预测的准确性越高。

精确度通常用于分类模型，定义为正确预测的观察值与总观察值之比：

精确度是评价分类器性能的直观方式，它简单地衡量了分类器正确分类的比例。精确度高意味着分类器在预测正确类别方面表现良好。

测验结果显示，本随机森林模型在与精确度指标上均表现良好，且在性能上显著表明模型的解释能力良好。

## 5.5结果

采用所建立模型预测监测点A、B、C在2023年7月13日至7月15日6种常规污染物的单日浓度值，并计算相应的AQI和首要污染物，结果如表7所示。

表7 监测点A、B、C污染物浓度预测

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 预报日期 | 地点 | 模型日值预测 | | | | | | | |
| SO2  (μg/m³) | NO2  (μg/m³) | PM10  (μg/m³) | PM2.5  (μg/m³) | O3最大8小时滑动平均  (μg/m³) | CO  (mg/m³) | AQI | 首要污染物 |
| 2023/7/13 | 监测点A | 6 | 13 | 20 | 5 | 81 | 0.4 | 40.5 | O3 |
| 2023/7/14 | 监测点A | 6 | 11 | 20 | 3 | 63 | 0.3 | 31.5 | O3 |
| 2023/7/15 | 监测点A | 6 | 11 | 17 | 5 | 81 | 0.4 | 40.5 | O3 |
| 2023/7/13 | 监测点B | 5 | 9 | 19 | 7 | 49 | 0.5 | 24.5 | O3 |
| 2023/7/14 | 监测点B | 5 | 6 | 17 | 4 | 51 | 0.5 | 25.5 | O3 |
| 2023/7/15 | 监测点B | 5 | 8 | 18 | 3 | 42 | 0.6 | 21.0 | O3 |
| 2023/7/13 | 监测点C | 9 | 23 | 56 | 34 | 196 | 0.6 | 53.0 | PM10 |
| 2023/7/14 | 监测点C | 7 | 18 | 29 | 10 | 158 | 0.4 | 98.333 | O3 |
| 2023/7/15 | 监测点C | 8 | 19 | 36 | 16 | 114 | 0.5 | 61.667 | O3 |

# 六、参考文献

[1] [想养只大黄](https://blog.csdn.net/m0_58032776), 中国空气质量指数（AQI）及其计算方式, [中国空气质量指数（AQI）及其计算方式\_aqi计算-CSDN博客](https://blog.csdn.net/m0_58032776/article/details/117912575),2024.5.13

[2]李铮铮,贾金娜,刘蓓蕾,等.基于PCA与K-均值聚类的学习者特征识别研究[J].现代信息科技,2023,7(22):142-145.

[3]刘婕,郝舒欣,万红燕,等.三种机器学习模型用于空气质量等级预测的比较研究——以保定市为例[J].环境卫生学杂志,2024,14(03):264-269.

[4]张希萌. Copula模型下二元区间删失数据的回归分析和拟合优度检验[D].吉林大学,2023.