全国研究生工业与经济金融

大数据建模与计算大赛

赛　　题	□ A　　☑ B
队伍编号	2024062
队长姓名	刘钢
队长学校	上海工程技术大学
队员姓名	1. 陈家涛
	2. 鄢浩龙
	3. 李建蕾

全国研究生工业与经济金融

大数据建模与计算大赛

题 目 ***PM***$_{2.5}$浓度影响因素及时空规律探索

摘 要

随着全球工业化和城市化的发展，$PM_{2.5}$污染问题日益严峻，对公共健康和经济发展造成了重大影响。本研究基于 2017 年 1 月 1 日至 2021 年 12 月 31 日期间北京、天津及河北周边 13 个城市的 $PM_{2.5}$ 数据，通过数据预处理，使用 DBSCAN 聚类和四分位距法分别处理重复值和异常值，并采用 Z-Score 进行标准化。

针对问题一（时空相关性分析），研究发现 $PM_{2.5}$ 污染具有弱空间正相关性（Moran's I=0.116），表明地理位置相近的城市污染水平存在一定关联。在时间维度上，表现出显著的连续性特征：日内相关性最强（0.971），年际相关性较高（0.805），季节相关性最弱（-0.182）。时序分析显示 $PM_{2.5}$ 浓度整体呈现逐年下降趋势，年均降幅达 11.21%。

针对问题二（气象因素分析），采用多元线性回归模型（R^2=0.844）量化气象因素的影响程度。结果表明，气压（标准化系数 10.90）和温度（标准化系数-10.28）是影响 $PM_{2.5}$ 浓度最显著的因素，分别表现为强正相关和强负相关。季节性分析发现冬季污染最为严重（均值 74.82μg/m³），夏季最轻（均值 33.65μg/m³），验证了温度对污染物浓度的重要影响。

针对问题三（浓度预测），研究比较了线性回归、XGBoost 和 LightGBM 三种模型在不同时间尺度上的预测性能。在 3 小时短期预测中，LightGBM 模型表现最优（RMSE=17.84，R^2=0.876）；在 24 小时长期预测中，所有模型的性能随预测时长增加而下降，但 LightGBM 和 XGBoost 的预测效果仍明显优于线性回归模型，最后一个预测步（24 小时）的 RMSE 分别为 32.97、32.87 和 36.99。

研究结果不仅揭示了 $PM_{2.5}$ 污染的时空分布规律和主要影响因素，还为污染物浓度的短期和长期预测提供了可靠的模型选择参考，可为区域大气污染防控决策提供科学依据。

关键词：$PM_{2.5}$ 浓度；时空相关性；气象因素；机器学习；预测模型

目　　录

1. 引言

1.1 背景

近年来，随着全球工业化与城市化的迅猛发展，能源消耗不断加剧，空气污染问题日益严峻。在众多空气污染问题中，$PM_{2.5}$ 的威胁备受关注。$PM_{2.5}$ 是指直径小于或等于 2.5 微米的细小颗粒物，因其粒径小、活性强、毒性大，同时易随有害物质进入人体后，可能会深入肺部甚至血液循环系统，从而对呼吸系统与心血管系统造成严重的健康威胁。研究表明，严重的 $PM_{2.5}$ 空气污染与心脏病、肺癌等多种慢性疾病的发病率有较高的正相关性[1]，会提高其发病率与死亡率。

此外，$PM_{2.5}$ 污染不仅会直接对于公共健康造成严重的影响，还会给污染地区带来巨大的经济负担。相关研究表面，在部分地区，$PM_{2.5}$ 污染造成的医疗费用与生产力损失可以达到地区 GDP 的 3.8%以上[2]。因此，对于 $PM_{2.5}$ 的相关治理研究亦具有相当高的经济价值。

综上所述，深入研究 $PM_{2.5}$ 污染的健康与经济影响，特别是针对不同地区的特定特征进行分析研究，不仅对于改善公共健康具有重要意义，也为经济可持续发展和环境政策优化提供重要依据。

1.2 问题重述

本次研究的主要目的是为了更好地探究 $PM_{2.5}$ 的污染特性、主要影响因素以及预测方法，将会从三个方面对其展开研究：$PM_{2.5}$ 的时间和空间相关性分析、气象因素对其的影响机制探索以及未来浓度的精准预测。

研究首先对于 2017 年 1 月 1 日至 2021 年 12 月 31 日间北京市、天津市以及河北省周边共计 13 个城市的 $PM_{2.5}$ 浓度进行数据采样分析，数据的采样频率为每三小时一次，采样内容包含了详细的时空与浓度信息。在时间维度上，可以通过时间序列分解技术提取季节性和周期性变化趋势，分析 $PM_{2.5}$ 浓度随时间的变化规律。在空间维度上，借助地理信息系统和空间统计方法，评估 $PM_{2.5}$ 在不通过城市间的空间分布及其相关性。同时，利用热力图、等值线等可视化工具直观展示污染程度的空间分布特征。通过初步的研究分析，可以识别地区污染演变的时空模式，为后续研究提供基础。

相关研究表面，气象因素会显著影响 $PM_{2.5}$ 的浓度变化[3]。为探究其具体的变化影响机制，通过已有 13 个城市的数据集内容进行解构研究，包括温度、降水量、边界层高度、相对湿度、地表气压以及风速（东西风和南北风）。通过构建多元模型，定量分析气象因素对 $PM_{2.5}$ 浓度的影响。本次研究还特别关注了在冬季时北向风的驱散效应，结合分季节数据进一步验证气象因素对于改善空气污染的作用机制。

最后，为了提高污染物的监管和治理能力，本研究利用所提供的数据，构建了预测模型，以期实现未来 3 小时和 24 小时的 $PM_{2.5}$ 浓度预测。同时选择构建不同的模型进行对比分析，以权衡 $PM_{2.5}$ 浓度的时间和空间相关性。每种模型的预测结果通过均方误差进行量化评估，并结合预测误差分布的分析，优化模型性能。最终的目标是提供精准的短期与中期 $PM_{2.5}$ 浓度预测，为污染控制措施的制定提供科学依据。

2. 数据处理

2.1 数据特征描述

本次题目所提供的数据总共 189905 条，每一条数据具有 11 个属性维度，其中温度、边界层高度、地表气压、降水量、相对湿度、水平风速、方向风速、$PM_{2.5}$ 浓度均为连续性的数值变量，城市为分类变量，为了更好地分析 $PM_{2.5}$ 浓度随着时间增长而产生的变化，我们将把日期和时间视为连续性的变量。

以 A 城市为例，在原始数据集中，A 城市所拥有的数据一共有 14608 条，其中 $PM_{2.5}$ 浓度在大部分的时间区间内集中于$0\mu g/m^3$至$100\mu g/m^3$，但少许 $PM_{2.5}$ 浓度随着时间离散地分布在$200\mu g/m^3$至$600\mu g/m^3$，可以发现数据存在密集部分和稀疏部分。再将 A 城市的数据进行 T-SNE 可视化，结果如图 1.2 所示，其中 $PM_{2.5}$ 浓度位于$200\mu g/m^3$至$600\mu g/m^3$的数据点相对集中形成了一个簇，但属于这个区间范围内的部分数据点也以簇的形式分布于其它区域。从整体分析，在整个区域内，数据点所占范围较大，并且图中没有较大的空白区域，表明数据整体是密集的；从局部分析，存在多个聚集的簇，两个不同簇之间的数据点密度比簇本身的数据点密度存在明显降低。综上，数据集的整体是密集的、范围较大的，存在局部数据组成簇，但簇与簇之间的界限不明显。其中存在对密集部分，我们需要对数据集的重复值进行处理，防止大量重复值存在而导致的结果扭曲；对稀疏部分，我们需要对数据集的离散点进行处理，判断其是否为异常点。

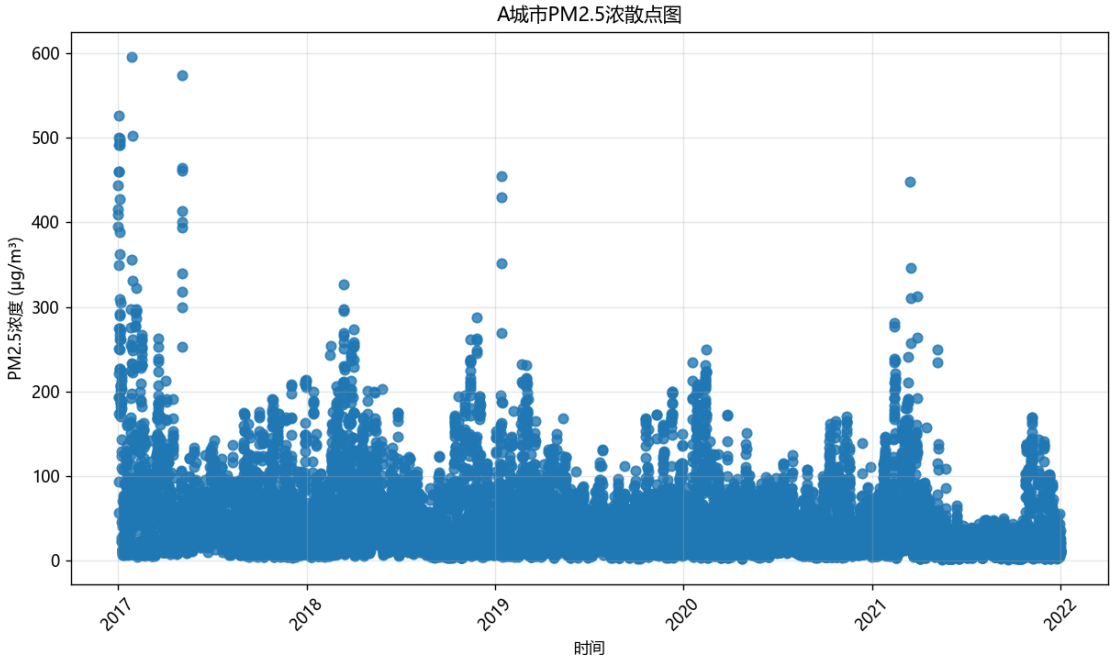

图 1.1 A 城市 $PM_{2.5}$ 浓度散点图

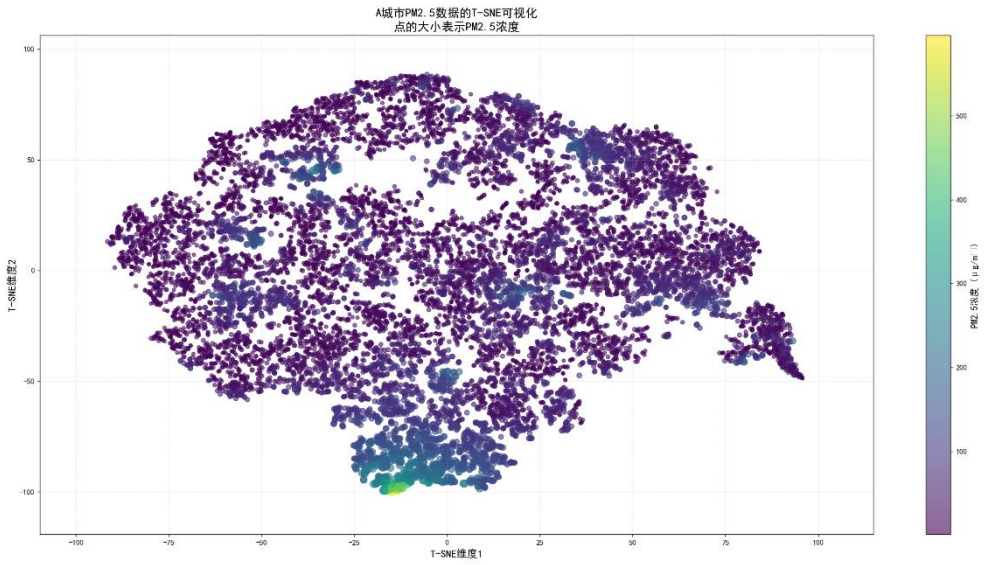

图 1.2 A 城市的 $PM_{2.5}$ 浓度可视化

2.2 特征增维和简单处理

特征增维是数据分析中的一种技术，指的是通过某些方法将原始数据的维度（特征）增加，从而改善模型的性能、提高数据的可分性或优化算法的表现。

为了实现特征增维的目的，可以将时间拆分为 year、month、day、hour，并且增加维度 season，将日期拆分成独立的组件便于分析时间模式，并更容易地研究季节性变化、月度趋势和日内变化；将经纬度信息加入维度，更便于进行空间分析，可以研究 $PM_{2.5}$ 浓度的空间相关性；可通过式(1.1)和式(1.2)从 U 水平风速和 V 方向风速计算得到风速和风向，可以更直观地表示风力特征，有助于分析风向对污染物的影响。

$$wind\_speed = \sqrt{U^2 + V^2} \tag{1.1}$$

$$wind\_direction = \begin{cases} arctan2(V,U) \times \dfrac{180}{\pi} + 360^\circ & \text{if } \alpha < 0^\circ \\ arctan2(V,U) \times \dfrac{180}{\pi} & \text{if } \alpha \geq 0^\circ \end{cases} \tag{1.2}$$

最后得到数据属性有 city、precipitation、humidity、$PM_{2.5}$concentration、year、month、day、hour、season、longitude、latitude、wind_speed、wind_direction、temperature、Boundary_height、pressure、humidity。

2.3 数据清洗

数据清洗在数据质量控制中有着极其重要的作用，足够的数据清洗对于任何数据分析都是至关重要的，并且对最后进行预测结果和分类结果的好坏具有不可忽视的影响。对于数据清洗的具体研究主要在以下三个方面：缺失值的处理、异常值的处理、近似重复值的检测和消除[4]。

考虑到传感器在工作中可能发生故障，以及诸多环境因素，进而导致数据的采集发生错误，该部分通过对初始数据中的各个标签的内容进行统计，观察数据具体情况，是否存在重复值、缺失值、异常值，为之后数据标准化的做好准备。

2.3.1 重复值处理

重复值数据概念包括完全相同数据，相似数据，缩写和全拼数据等等情况，在不同的文献中，重复值也被称为“重复数据”、“相似数据”等等[5]。重复值可能由于数据录入错误、系统错误、数据合并等原因而产生，会对数据分析产生干扰，尤其是计算均值、标准差、回归分析等统计分析时，重复值会扭曲结果。从数据分析的角度来看，重复值数据带来的关联关系破坏是致命的，有可能会导致数据分析过程中抽取错误的模式、获得有偏差的规则、 最终导致数据模型的异常，使得统计分析结果大打折扣， 甚至导致决策支持系统分析出错误 的决策和结果，降低服务质量。因此，首先去除重复值可以确保数据集的质量。

考虑到本次所使用数据按照每三个小时记录各个城市的各类信息，因此在城市相同的情况下，相邻的多个记录时间点所记录的各个数值属性维度的信息可能存在数值变化非常小的情况，而非完全相同，并且数据集中存在多个不同的簇，故我们采用聚类的方法对数据的重复值进行检测。

聚类算法可以根据数据的相似性将近似重复的项聚集在一起，从而实现去重。常见的聚类算法包括 K-means 聚类[6]、DBSCAN[7]。以上两种聚类算法都能够处理具有多种复杂数据类型的数据，拥有自动化处理大规模数据的能力，并且自适应能力强，可以根据数据分布和数据密度的不同调整参数。但 K-means 聚类在很大程度上依赖于初始中心的位置，而随机选择初始中心可能会导致聚类质量低下，簇过于松散或过于紧密，无法有效聚类数据[8]。DBSCAN 需要用户输入指定执行算法的参数值；它在从密度不同的数据集中确定有意义的聚类时容易陷入两难境地；它产生一定的计算复杂度[9]。尽管 DBSCAN 存在以上问题，但是 K-means 需设置簇的个数，而每一个城市的数据本身所拥有的潜在的簇的数量不相同，两个不同簇之间的数据点密度存在差距，并且 DSBSCAN 通过密度的不同确定有意义的聚类，因此受到噪声的影响较小，相比 K-means 聚类更加适合该数据集。

按照城市对数据进行分类，一共可分为 13 类，每一类数据按照日期和时间的先后顺序进行排列，分别对每一类的数据使用基于 DBSCAN 的聚类方法对重复值进行处理。表一中，该算法首先通过 DBSCAN 方法对数据集进行聚类，将数据集中的数据分为多个不同的簇，再对每一个簇中的各个元素进行相似度分析，对那些相似度超过阈值的元素进行删除或者整合。

表 1.1 基于 DBSCAN 的重复值处理算法

基于 DBSCAN 的重复值处理算法
输入：X：数据集 eps：半径参数 MinPts：领域密度阈值
输出：进行了重复值处理后的数据集合 X

labels = DBSCAN(X,eps,MinPts)
for cluster_id in 唯一的聚类标签(labels):
if cluster_id != -1:
簇内数据 = 获取当前簇的数据点(X, labels, cluster_id)
end if
for i in range(len(簇内数据)):
for j in range(i + 1, len(簇内数据)):
相似度 = 计算相似度(簇内数据[i], 簇内数据[j])
if 相似度 > 设定的相似度阈值:
处理重复数据(簇内数据[i], 簇内数据[j])

```
            end if
        end for
    end for
end for
return X
```

在重复值处理中所使用的数据集中的数据全部为数值型数据，因此可以采用欧式距离或者曼哈顿距离衡量两个不同数据之间的相似度，由于原始数据密集，并且不是连续的均匀分布，经过 DBSCAN 的聚类操作之后，每一个簇中的各个数据之间依旧保持着原本的密集，若使用曼哈顿距离进行相似度的计算，可能会将原本距离簇中心较远的元素也纳入重复数据的处理中，导致数据丧失原本的特性。对于数据集中的两个数据元素 $x=(x_1,x_2,...,x_n)$ 和 $y=(y_1,y_2,...,y_n)$，其欧式距离公式为：

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1.3}$$

我们采用删除和整合的方式对超过相似度阈值的数据进行处理，对于那些完全相同的两组数据，直接删除其中一组；对于剩余相似度小于阈值的数据，则采用平均化合并的方法，将两组数据取平均值，合并为一个新的数据。

2.3.2 重复值处理结果

通过该部分对数据集中的重复值进行处理，使得城市 A 的数据总数从 14608 条降低到 2621 条，减少率为82.05%，考虑到原始数据中存在着大量重复数据，这样的结果是可以接受的。在图 1.3 中，数据点的数量得到了明显下降，且整体的数据分布与原始数据相似，表明了在保持数据原本特性的基础上，实现了对数据数量的减少，成功处理了原始数据中存在的大量重复值。

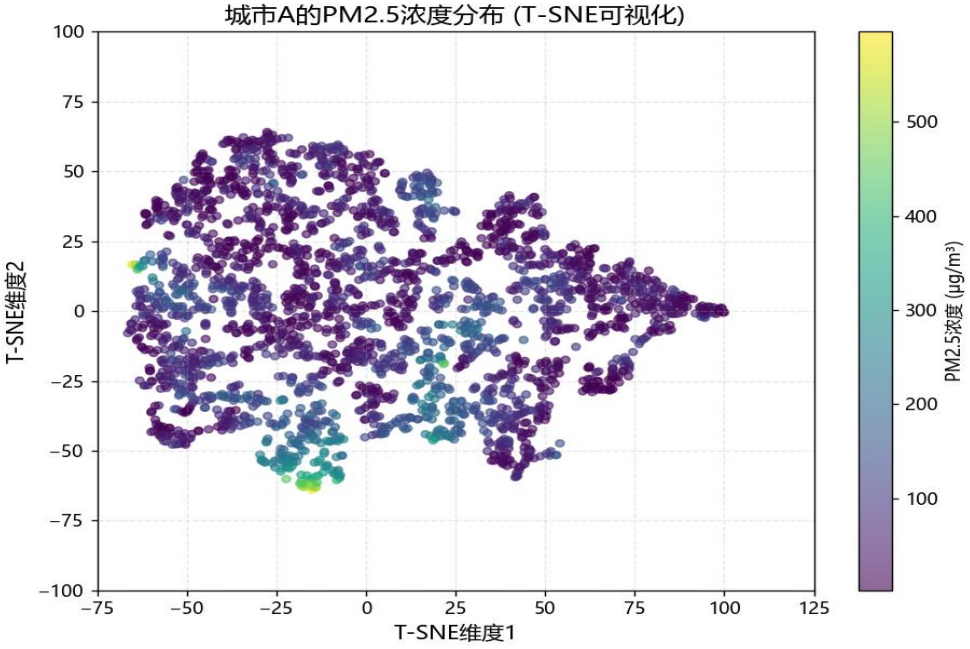

图 1.3 重复值处理后的城市 A 的 $PM_{2.5}$ 浓度数据可视化

2.3.3 缺失值处理

经过统计，本次所使用的数据集中并未存在缺失值，因此不进行对缺失值的处理。

2.3.4 异常值处理

Silvia 认为数据集中的异常值是数据集中的离群点，其中离群点是一个明显偏离其他观测数据的数据点，往往不同于当前数据的产生方式[10]。异常值可能会显著影响数据的基本统计特征，如均值、标准差、方差、相关性等，进而扭曲数据的分布和关系，对模型的训练和预测结果造成严重的影响。

由于所使用的数据集的维度只有 11 个属性维度，将城市类别、时间和日期排除后，总共具有 8 个属性维度，因此对于此类低维数据，异常值检测的常用方法有标准化分数方法、四分位距法、K-近邻法。其中标准化分数方法通过标准化数据点的值来识别异常值，适用于数据近似正态分布的情况，但由于所使用的数据集不符合正态分布，因此不应使用该方法进行异常值的检测。K-近邻法通过计算每个数据点到其 K 个最近邻的平均距离，如果某个数据点的距离明显大于其他数据点的距离，它就有可能是异常值，但该方法对 K 值和距离及算法敏感，即便通过了 DSBSCAN 进行了重复值的处理，此时的数据集依旧较大，并且不同城市的数据集中各个数据分布不同，导致选取合适的 K 值变得困难。四分位距法通过数据的分位数来确定异常值，适用于非正态分布数据，并且由于各个城市的数据本身没有出现分布极端的数据点，使得四分位距法的异常值检测的准确度更高。

表 1.2. 基于四分位距法的异常值检测与处理

基于四分位距法的异常值检测与处理
输入：X：数据集
输出：删除了异常值的数据集 X
1：获取数据集的维度，特征数为 n，数据点数为 m
2：创建空列表 outliers，用于存储异常值的索引
3：for i in range(n):
4：　　feature_i = 第 i 个特征的列
5：　　sort(feature_i)
6：　　Q1_i = calculate_Q1(feature_i)
7:　　Q3_i = calculate_Q3(feature_i)
8:　　IQR_i = Q3_i - Q1_i
9:　　lower_boundary_i = Q1_i – 1.5 * Q1_i
10:　　upper_boundary_i = Q3_i + 1.5 * IQR_i
11:　　for j in range(m):
12:　　　　if X[j][i] < lower_boundary_i or X[j][i] > upper_boundary_i
13:　　　　　　将索引 j 添加到 outliers 中
14：　　　　end if
16：　　end for

17：end for

18：X 根据 outliers 删除相应的数据行

19：return X

2.3.5 异常值处理结果

通过该部分对数据集中的异常值进行处理，使得城市 A 的数据总数从 2611 条降低到 1799 条，减少率为31.09%，考虑到原始数据中存在着大量重复数据，这样的结果是可以接受的。在图 1.4 中，数据点的数量得到了明显下降，实现了对局部异常值的处理，并且清理了远离数据中心的异常值。

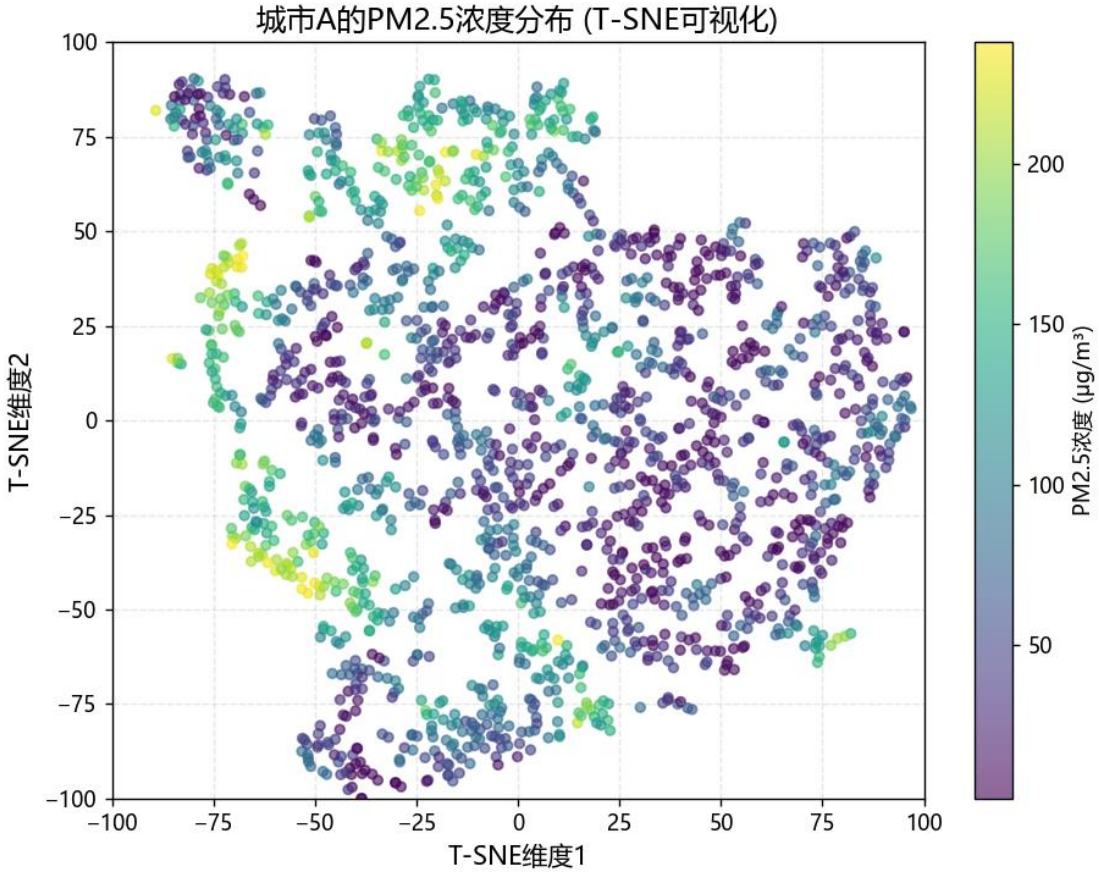

图 1.4 重复值处理后的城市 A 的 $PM_{2.5}$ 浓度数据可视化

2.4 数据标准化

数据标准化的目的是将不同量纲（单位）的数据转化为统一的标准尺度，以消除特征之间的量纲影响，提高模型的稳定性和收敛速度，以便更好地进行模型训练和分析[11]。常见的数据标准化方法分别是 Z-Score 标准化、Min-Max 标准化和绝对值标准化。其中 Min-Max 标准化将数据线性映射到指定的最小值和最大值之间，但这种方法对异常值敏感，并且会导致数据压缩失真；最大绝对值标准化将数据按其最大绝对值进行缩放，使数据的范围在 [-1, 1] 之间，该方法适用于稀疏数据，尤其是当数据包含大量零值时；Z-Score 标准化基于均值和标准差，不易被极端值影响，且适用于大多数数据分布，特别是当数据大致符合正态分布时，它表现得尤为有效，并且该标准化方法适用于大多数机器学习算法、线

性模型和基于距离的模型。

经过数据清洗后的数据，已经消除了大部分的重复值和异常值，但整体的特征间分布差异很大，并且局部的数据点之间依旧是密集的，不存在大量零值，为了使标准化后的数据能够更好地与之后的问题求解方法匹配，这里选择Z-Score标准化实现对数据的标准化。

设 x 为原始数据，μ 为均值，σ 为标准差，z 为标准化结果，则可通过式 1.2 实现对数据的标准化。

$$z = \frac{x - \mu}{\sigma} \tag{1.4}$$

3. 问题一.时空相关性分析

3.1 问题分析

问题一要求对给出的北京、天津及河北周边 13 个城市在 2017 年 1 月 1 日至 2021 年 12 月 31 日期间的 $PM_{2.5}$ 浓度数据进行时间和空间相关性分析。空间相关性分析的分析目标为探究 13 个城市 $PM_{2.5}$ 浓度的空间分布特征及其城市间的关联性，并量化整体空间自相关性。而对于时间相关性分析，主要分析目标为识别 $PM_{2.5}$ 浓度的时间变化特征，包括日内变化、季节性变化和年际变化，并将其量化时间自相关性。

3.2 空间相关性分析

在原始数据中分别提取 13 个城市的 $PM_{2.5}$ 指数，并取均值，用作 $PM_{2.5}$ 污染指数的空间性分析，初步得出的数据如表 3.1 所示。从经纬度数据来看，这 13 个城市大致位于我国华北平原地区，均有不同程度的 $PM_{2.5}$ 污染，因此具有充分的研究价值。

城市	经度	纬度	$PM_{2.5}$ 均值
A	116.3983	40.04598	44.85481
B	117.3222	39.07768	50.97406
C	114.4932	38.03363	65.21598
D	118.1829	39.64495	55.20263
E	119.6069	39.93625	38.68243
F	114.5139	36.60786	65.13712
G	115.4852	38.87627	61.00034
H	114.9009	40.80275	27.91128
I	117.9277	40.96416	31.10722
J	116.7151	39.52605	47.90416
K	116.8716	38.31577	52.46310
L	115.6761	37.7448	57.78656
M	114.5067	37.0771	62.55326

表 3.1 空间性分析数据

从粗处理后的数据来看，C、F、G 和 M 市的 $PM_{2.5}$ 污染指数均值最高，受污染程度最为严重，均达到了 60 以上的污染指数均值，而 H 市受污染较少。将其绘成经纬度散点图，

即图 3.1，同时使用颜色变化表示各城市的 $PM_{2.5}$ 浓度水平，观察散点图可以直观地得到以上结果。

在散点图中，由颜色表示 $PM_{2.5}$ 的浓度，从低到高渐变，即从蓝色向红色渐变，横坐标表示经度，纵坐标表示维度，因此该散点图可以大致表示出这 13 个城市的地理位置分布情况。观察散点图可以发现，$PM_{2.5}$ 的污染浓度在地理上整体呈现南北梯度分布，即越往南方，污染浓度越高。结合地理知识来说，在我国北部地区，城市 C、F、M 的污染浓度较高，可能与当地的地形、气象或者工业发展和密集生产活动相关，而城市 H、I 的浓度最低，这有可能是因风向或地势更有利于污染扩散。

当然，仅观察原始数据或者散点图得到的结论是不具说服力的，因此需要对 13 个城市的 $PM_{2.5}$ 污染程度做空间相关性的量化。首先需要计算城市间的欧几里得距离，两城市间的距离公式如式(3.1)所示。然后计算各城市的皮尔逊相关系数，其计算公式如式(3.2)所示。最后使用城市的 $PM_{2.5}$ 浓度均值及空间权重矩阵，计算 Moran's I，评估污染在空间上的聚集性，其计算公式如式(3.3)所示。其中 n 表示城市观测点数量，W 是空间权重总和，w_{ij} 是空间权重矩阵元素，x_i, x_j 是城市观测值，μ 是总的 $PM_{2.5}$ 污染浓度平均值。最后绘出的城市间距离矩阵热力图与 $PM_{2.5}$ 污染程度相关矩阵热力图如图 3.2 所示。

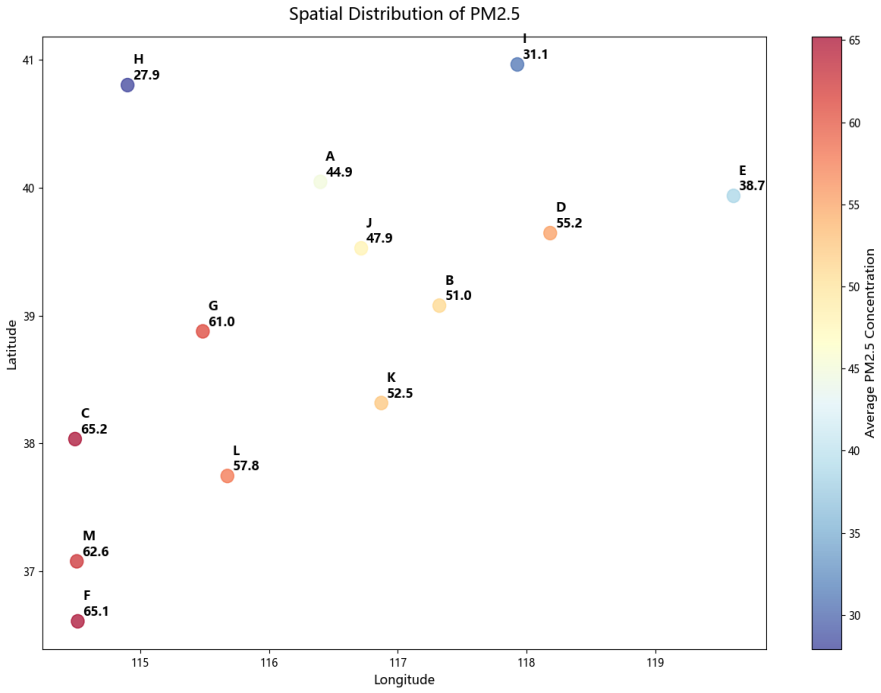

图 3.1 经纬度散点图

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3.1}$$

$$r = \frac{\mathrm{cov}(X, Y)}{(\sigma_X \cdot \sigma_Y)} \tag{3.2}$$

$$I = \frac{n}{W} \cdot \frac{\sum_i \sum_j w_{ij}(x_i - \mu)(x_j - \mu)}{\sum_i (x_i - \mu)^2} \tag{3.3}$$

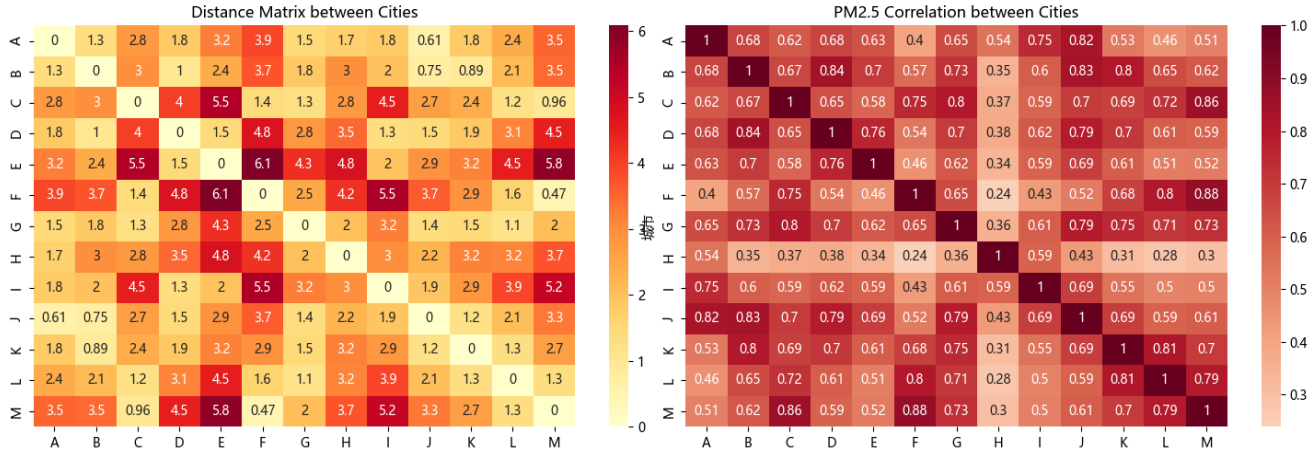

图 3.2 距离矩阵热力图与相关性矩阵热力图

由此可计算出空间自相关指数(Moran's I)为 0.116，表明在数据集采集地区，$PM_{2.5}$ 的污染程度存在较弱的空间正相关，虽然在地理上邻近的城市，在污染程度上有一定的相似性，但是关联性较弱。另外，亦存在城市间的污染程度相关性强弱不一的情况，这可能是由于地理距离与相似的气候条件产生的，因此在后续的研究中将对此进行详细的探讨。

3.3 时间相关性分析

对于时间相关性分析的部分，主要目的是为了探究 $PM_{2.5}$ 的时间变化特性，包括日内变化、季节性变化和年纪变化，并量化其时间相关性。首先对时刻变化进行分析，利用 hour(分时)数据聚合 $PM_{2.5}$ 浓度，然后计算均值和标准差，描绘一天中不同时间点的 $PM_{2.5}$ 污染程度，寻找日内污染峰值。均值与标准差的计算公式如(3.4)和(3.5)所示。接着将月份映射到季节，同样按照 season(季节)分组聚合，计算 $PM_{2.5}$ 浓度的均值，分析季节趋势。再按年纪分组，即按 year(年际)分组聚合，计算每年的 $PM_{2.5}$ 浓度均值和标准差，观察年际变化趋势。最后利用 $PM_{2.5}$ 的时间序列数据，通过滞后相关计算自相关函数 ACF(Auto-Correlation Function)，量化短期（3 小时间隔）污染浓度的时间依赖性，ACF 的计算公式如式(3.6)所示。其中 k 是时间滞后量。

$$\mu = \frac{\sum x}{n} \tag{3.4}$$

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}} \tag{3.5}$$

$$\rho(k) = \frac{\mathrm{cov}(X_t, X_{t+k})}{(\sigma_{X_t} \cdot \sigma_{X_{t+k}})} \tag{3.6}$$

将计算所得数据绘制成折线图或柱状图，共绘制了三张统计图。先观察日内变化的折线图，即图 3.3。可以发现，日内 $PM_{2.5}$ 的峰值时刻出现在 0 时，低谷时刻出现在 15 时。且整体上在夜间时 $PM_{2.5}$ 的浓度较高。这可能是因为夜间大气扩散条件较差，虽然交通活

动减少但持续存在以及工业排放稳定等问题，而日渐阳光照射增强，大气层对流条件较好，部分污染物可能被光化学反应消耗。从折线图来看，变化波动较大，标准差也较大，表明不同时间段内 $PM_{2.5}$ 浓度的波动性较强。

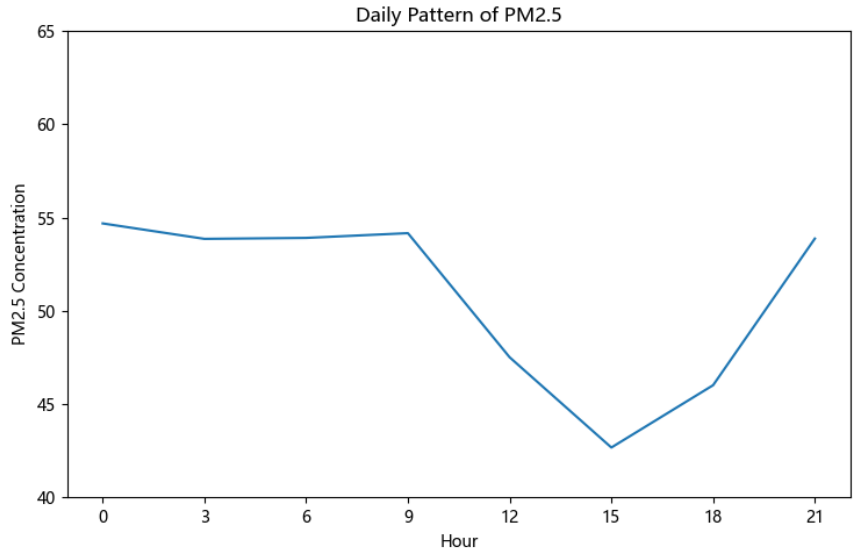

图 3.3 $PM_{2.5}$ 浓度日内变化折线图

而从图 3.4 的 $PM_{2.5}$ 季节性变化柱状图来看，$PM_{2.5}$ 浓度在冬季时，普遍达到当年最高，而夏季时达到最低。这种变化规律与季节性气候密切相关。冬季气温较低，逆温层现象常见，导致污染物累积[12]。夏季降雨量大，空气湿润且风力条件较好，有助于稀释和清除空气中的污染物。此外，可以较为明显的观察到，$PM_{2.5}$ 的污染程度逐年有较为明显的下降趋势。图 3.5 的年际变化折线图进一步验证了这一观点，从图中可以看出，2017 年至 2021 年，$PM_{2.5}$ 浓度呈现显著下降趋势，年均下降幅度可以达到 11.21%。这显然表明在过去的几年中，空气质量得到了明显的改善，同时年际污染度变化的标准差也在逐年减小，说明 $PM_{2.5}$ 浓度的年间波动性减弱。

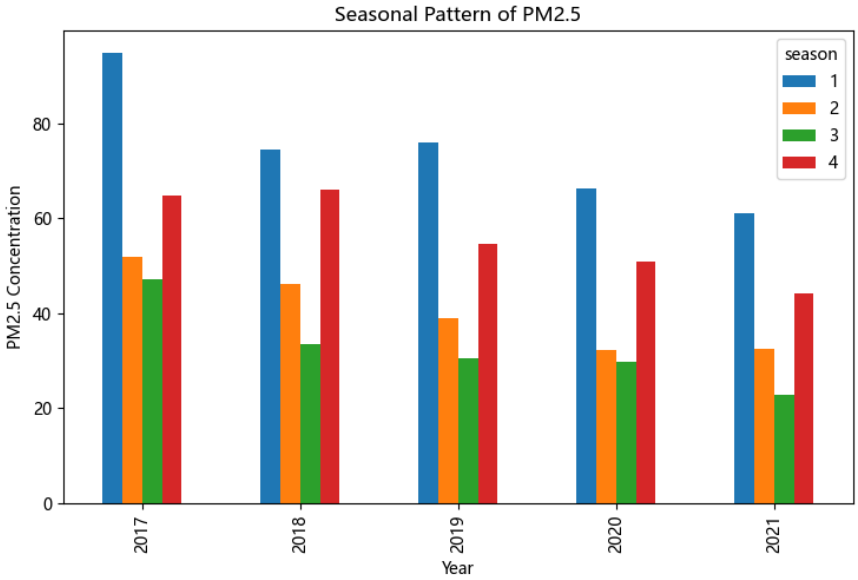

图 3.4 $PM_{2.5}$ 季节性变化柱状图

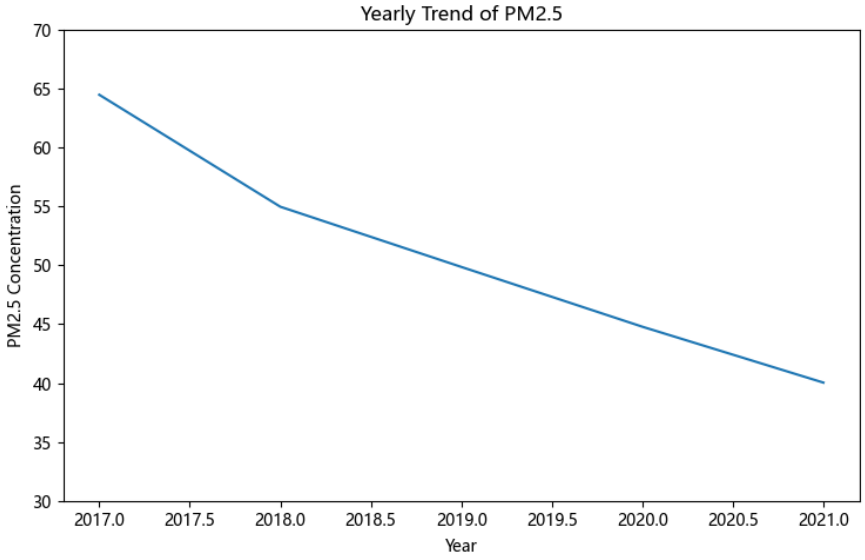

图 3.5 $PM_{2.5}$ 年际变化折线图

最后，将 $PM_{2.5}$ 浓度分项的时间变化数据，进行自相关计算，表 3.2 是汇总的各项自相关系数。从表中数据可以得知，日内和日间的相关系数较高，反映了污染浓度的短期连续性。污染源的累积和扩散在这些时间尺度上较为明显，短时预测模型可以很好地捕捉这些变化规律，并做出较好的预测反馈。而月间相关性也较高，说明在相邻月份内，$PM_{2.5}$ 的变化趋势可能与当地气象条件的延续或者社会活动模式变化相关联。而季节相关性系数为负值，表明不同季节间的污染浓度反差较大。特别是在冬季和夏季之间，$PM_{2.5}$ 浓度的差距明显，这与大气扩散条件和污染源活动的季节性变化密切相关。冬季浓度高的主要原因是逆温层效应以及煤炭供暖[13]，而夏季浓度低则得益于高降水频率等季节性气候变化。年际相关性系数也较高，这反映出长期污染治理措施的显著成效，与前文得出的 $PM_{2.5}$ 浓度下降趋势相一致，表面空气治理的改善工作具有较强的可持续性和连贯性。

表 3.2 时间相关性

时间单位	相关系数
日内	0.971
日间	0.641
月间	0.660
季节	-0.182
年间	0.805

4. 问题二.气象因素分析

4.1 问题分析

第二问的问题核心在于分析多种气象因素与 $PM_{2.5}$ 浓度的相关性，计算诸如温度、风速、气压等与 $PM_{2.5}$ 浓度之间的相关系数[14]，分析这些气象变量对于 $PM_{2.5}$ 浓度的线性关系的强弱[15]。分析得出的相关性结果不仅能够反映变量直接的正相关性和负相关性，还能为后续建模和决策提供数据支持。

4.2 气象因素基本信息分析

针对所给的数据集进行标准化后，验证数据的合理性与准确性是实验开始前的必要步骤。利用数学方法对标准化后的数据进行基本的统计分析，提取各项指标后如表 4.1 所示。其中 Mean 表示均值，Std 表示标准差，Min 表示最小值，25%表示下四分之位数，50%表示中位数，75%表示上四分之数，Max 表示最大值。

表 4. 1 气象因素信息统计

Statistic	Temperature	Boundary Height	Pressure	Humidity	Wind Speed	$PM_{2.5}$ Concentration
Mean	2.001E-13	-1.84306E-13	-1.1E-12	1.61E-12	-4.26519E-13	50.83023
Std	1	1	1	1	1	48.60458
Min	-3.381491	-0.7779643	-3.33318	-1.92685	-1.47865	1
25%	-0.850408	-0.6847516	-0.03542	-0.85106	-0.7868601	20
50%	0.0930469	-0.42978	0.286405	-0.14917	-0.2135633	36
75%	0.8424821	0.3509404	0.580776	0.802263	0.5681939	62
Max	2.316686	7.291196	1.445725	2.373279	5.603423	1083

从表中数据可以得知，所有的气象因素的均值都接近 0，而标准差为 1，这符合标准化后的数据分布特征，说明未出现数据丢失或错误的情况。而 $PM_{2.5}$ 浓度的均值为 50.83，标准差则达到了 48.60，这说明浓度的数据分布存在较大的离散型，从极值、中位数和四分位数的分布中也可以看出，$PM_{2.5}$ 浓度的数值可能会存在偏态分布的现象[16]。

根据标准化后的数据均值，使用皮尔逊相关系数(Pearson Correlation)可计算各气候因素与 $PM_{2.5}$ 浓度的相关性。皮尔逊相关系数计算公式如式(4.1)所示，计算结果如表 4.2 所示。皮尔逊相关系数 r 的数值可以衡量变量间的线性关系，r 的取值范围为[-1,1]，正值为正相关，负值为负相关。其中绝对值小于 0.3 的称为弱相关，大于 0.3 小于 0.7 的称为中等相关，大于 0.7 的称为强相关。了解皮尔逊相关系数的划分规则之后，分析所得结果可知，温度(temperature)和风速(wind_speed)两个因素对于 $PM_{2.5}$ 浓度的线性影响较弱，表面这二者可能通过其他非线性机制的因素间接影响 $PM_{2.5}$ 的浓度。而气压(pressure)对于 $PM_{2.5}$ 的浓度具有弱相关性，可能由于高气压天气通常伴随较弱的水平和垂直气流，导致污染物难以扩散累积。

$$\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_x\sigma_y}=\frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2\sum_{i=1}^{n}(y_i-\mu_y)^2}} \tag{4.2}$$

表 4.2 气候因素与 $PM_{2.5}$ 浓度相关性

气候因素	相关系数
pressure	0.217587
humidity	0.047742
wind_speed	0.037131
Boundary_height	-0.079307
temperature	-0.183420

从整体的相关性系数来看，五个气候因素与 $PM_{2.5}$ 浓度变化都是弱相关及以下，这表明 $PM_{2.5}$ 浓度的变化可能无法简单地通过单一变量的解释[17]，需要设计多个因素的复杂交互作用，或者气象因素与 $PM_{2.5}$ 的浓度变化可能存在非线性效应的影响。因而，接下来

将对其他各方面分项对气候因素进行分析探讨。

4.3 季节因素影响分析

$PM_{2.5}$ 浓度受时间周期影响较大，尤其是在季节更替时会有较大的变化[18]。因而需要进行季节性分析，揭示污染物浓度的变化规律。需将数据按季节(season)分组，然后计算不同季节的均值浓度，并分析差异。计算结果如表 4.3 所示。从表中数据来看，季节平均值呈现明显的冬季大于春季大于秋季大于夏季的分布规律，为了对季节因素进行详细分析，按季节对不同的气候因素进行皮尔逊相关系数的计算分析，结果如图 4.1 的热力图所示。

表 4.3 各季节 $PM_{2.5}$ 平均浓度

季节	相关系数
冬季	74.817308
春季	48.549728
夏季	46.730368
秋季	33.648244

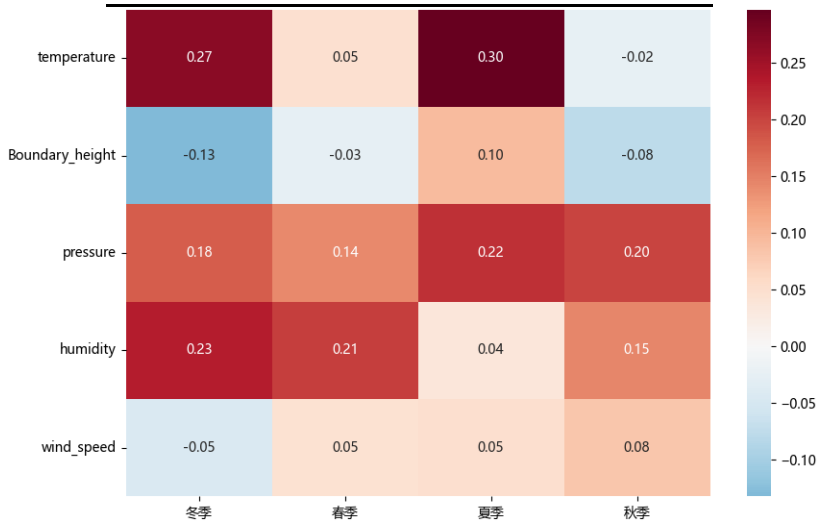

图 4.1 分季节气象因素相关性

从热力图上可以看出，在 $PM_{2.5}$ 浓度的峰值与低谷，即冬季与夏季时，各气候因素的相关性系数明显要高于其他两季的相关性系数。特别是对于温度的相关性系数表现上，尤为明显。而春秋两季是气候的过渡季节，气象条件较为多变，气象因素的作用相对复杂且不稳定。

4.4 风速与风向因素分析

在前文的气候因素分析中，风速与 $PM_{2.5}$ 浓度的变化呈弱相关，具体的相关性分析如图 4.2 所示。在该散点图中，风速值范围为[-2,5]，$PM_{2.5}$ 浓度分布范围为[0，500]，从图中可以观察到，随着风速的增加，$PM_{2.5}$ 浓度整体呈现减少的趋势。这表明，风速对 $PM_{2.5}$ 浓度呈负相关，即风速越大，浓度越低。这说明较大的风速可能对 $PM_{2.5}$ 浓度有稀释作用，风速越大，空气流动性越强，污染物浓度可能会有所降低。而在风速为[-1,1]这一区间内，即无风或微风状态时，$PM_{2.5}$ 的浓度点状分布明显更为密集，且普遍浓度较高，这也进一步说明了空气流通性低时，空气污染物更容易累积。此外，散点图中还出现了一些 $PM_{2.5}$ 浓度超过 300 的极端值，这可能是由于短时间内的特殊污染事件，如大规模的工业排放等。

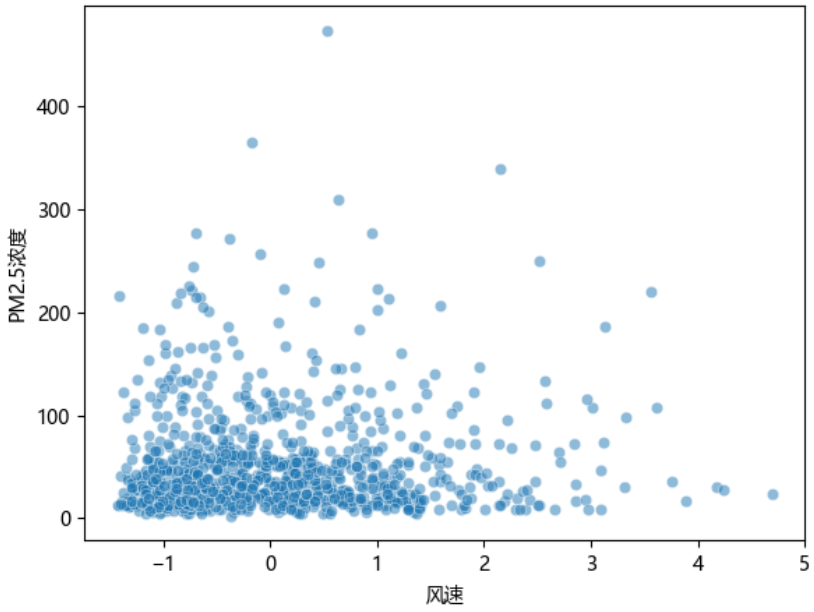

图 4.2 风速与 $PM_{2.5}$ 浓度散点图

从整体上看，风速与 $PM_{2.5}$ 浓度的相关性较弱，这与上文中得出的 0.037 的皮尔逊相关性系数相吻合。但从整体的趋势来看，风速对于 $PM_{2.5}$ 的稀释效应在风速为[1,3]的区间上变化尤为明显，可以明显的观察到随着风速的增长，$PM_{2.5}$ 浓度的散点愈加稀疏。

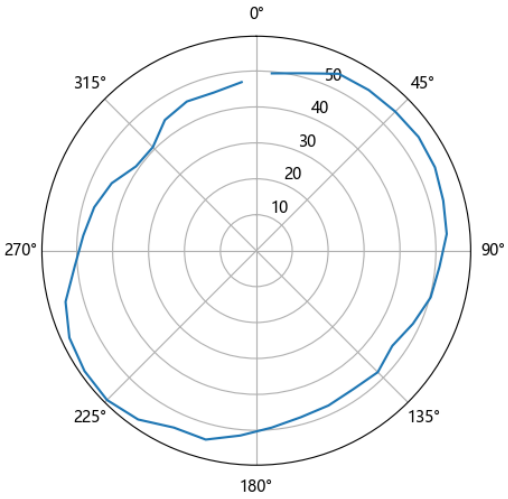

图 4.3 $PM_{2.5}$ 浓度与风向玫瑰图

风向问题也可能是间接影响 $PM_{2.5}$ 浓度的气象因素之一，图 4.3 的风向玫瑰图可以直观地展示风向与 $PM_{2.5}$ 浓度的变化关系。从图中可以发现，西南风与东北风方向时，$PM_{2.5}$ 的浓度明显较高，而风向为西北风时 $PM_{2.5}$ 浓度偏低。由于我国是季风气候，所以风向可能存在与季节因素相关的问题。另外不同风向的 $PM_{2.5}$ 浓度差异也有可能与当地的污染源分布有关。例如某市的西南角为当地的工业区，在风向为西南风时自然会使当地的空气污染加剧。因而，风向因素可能不仅与风速相关联，它与 $PM_{2.5}$ 浓度变化也存在一定的间接相关性。

4.5 多元线性回归分析

经过上文的分析，可以得出结论，$PM_{2.5}$ 浓度的变化不是单一因素决定的，而是由多因素联合作用导致的结果。因此需要使用多元线性回归模型将多个变量整合到一个框架中，

从而量化每个因素的独立作用。多元线性回归方程如式(4.3)所示，其中 β_0 表示截距，β_i 表示第 i 个气象因素的系数，ϵ 为误差项，它表示模型无法解释的部分，是随机误差或噪声的集合。构建多元线性回归模型需要先将气象因素系数进行标准化处理，标准化公式如式(4.4)所示，其中 β 表示原始回归系数[19]，σ_x 表示自变量标准差，即各气象因素的标准差，σ_y 表示因变量标准差，即 $PM_{2.5}$ 浓度的标准差。通过标准化系数，可以在不同量纲的变量间进行比较，衡量其对因标量的实际影响强度[20]。气象因素标准化系数与 $PM_{2.5}$ 浓度的相关性如图 4.4 所示。

$$PM_{2.5} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon \tag{4.3}$$

$$\beta_{std} = \beta \frac{\sigma_x}{\sigma_y} \tag{4.4}$$

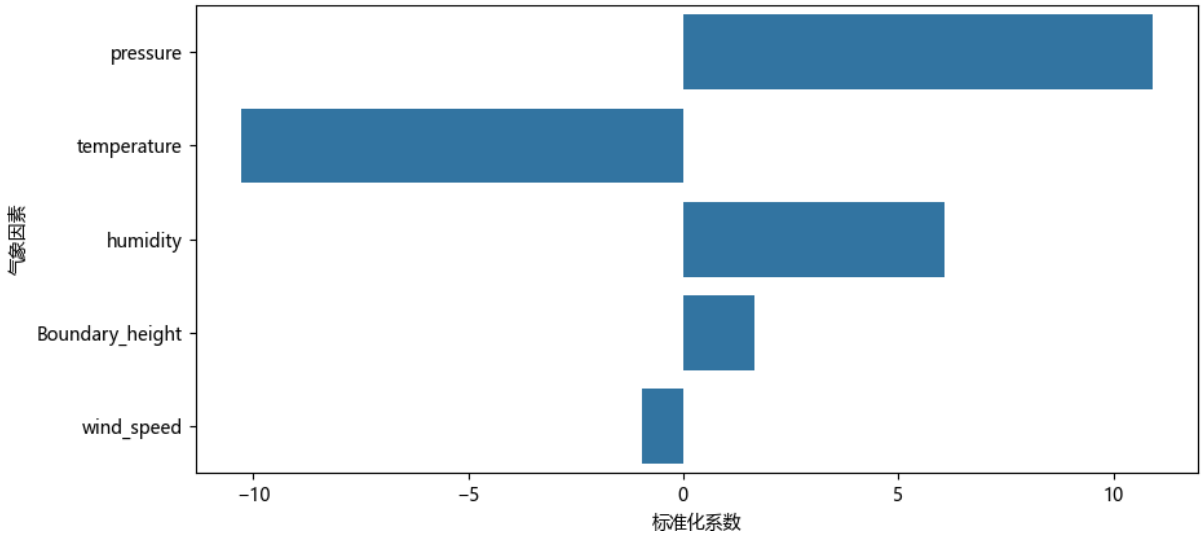

图 4.4 气象因素对 $PM_{2.5}$ 的影响程度

根据得出的标准化系数结果，可以分析不同变量对因变量影响的强度和方向。首先是压力(pressure)，标准化系数有 10.90，对于 $PM_{2.5}$ 浓度属于强正向影响，这意味着随着压力的增大，$PM_{2.5}$ 浓度也会得到显著的提高。这可能是由于压力通过影响气流模式从而直接影响污染物的聚集或扩散。然后是温度(temperature)，标准化系数有-10.28，对于 $PM_{2.5}$ 属于强负向影响，这意味温度越高，$PM_{2.5}$ 浓度越低，反之亦然。这也与前文得出的 $PM_{2.5}$ 浓度冬高夏低的结论相吻合。原因在于高温可能有助于气流的循环和扩散，从而减少空气中污染物的浓度。接着是湿度(humidity)，标准化系数有 6.06，对于 $PM_{2.5}$ 浓度属于中等正向影响。说明湿度增大时，$PM_{2.5}$ 浓度也趋于增加[21]。高湿度环境可能有助于颗粒物的聚集，从而导致污染物的浓度上升。其次是边界层高度(Boundary_height)，标准化系数仅有 1.65，说明边界层高度的影响较弱，且为正向影响，说明当边界层高度增加时，$PM_{2.5}$ 浓度会有所上升，但是影响程度较小。这可能是因为边界层高度会一定程度上影响大气的稳定性，低边界层会限制污染物的垂直扩散。最后是风速(wind_speed)，标准化系数仅有-0.96，呈较弱的负向影响，表明风速增大时，$PM_{2.5}$ 浓度会略微降低，但是整体的影响相较于其他几个因素来说较弱。

综上所述，压力和温度是对 $PM_{2.5}$ 浓度的影响最大因素，分别表现为强正向和强负向影响，其次是湿度，表现为中等正向影响，最次是边界层高度和风速，分别表现为弱正向和弱负向影响。从结果来看，温度和压力在 $PM_{2.5}$ 浓度的控制中起着重要作用，可以作为空气质量预测和污染控制的关键变量。

最后，再计算模型解释率(R^2)以验证结论的可信度，其计算公式如式(4.5)所示，其中 SS_{res} 表示残差平方和(Residual Sum of Squares),由真实值(y_i)和测值($\hat{y}_i$)计算而来，SS_{tot} 表示总平方和(Total Sum of Squares)，亦有真实值和变量均值及因变量的均值($\bar{y}$)计算而来，具体计算公式与式(4.5)后一步中的分子分母对应。

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.5}$$

R^2 的取值范围为[0,1]，理想值为 1，表示模型完美地拟合了数据，所有预测值均与实际值一致。因而，R^2 的取值越接近与 1，表示模型的拟合效果越好。在上文得出的结果中，经计算所得模型解释率 R^2 值为 0.844，说明以上实验对数据的拟合效果较好，但仍有部分变异因素未被捕捉到。主要的自变量如压力与温度等，已可以从标准化系数显示它们对 $PM_{2.5}$ 的重要影响，较高的模型解释率亦表面这些变量确实是影响 $PM_{2.5}$ 浓度的关键因素，模型也已抓住了大部分的主要因素。但是，仍有 15.6%的变异因素未得到解释，这可能是因为模型中有一些未涉及到的简介变量与非线性关系。虽然目前已可以有效地进行简单的应用，但后续仍可以尝试通过添加更多的变量或使用其他改进模型来进一步提升解释率和预测能力。

5. 问题三.对 $PM_{2.5}$ 浓度进行预测

5.1 问题分析

问题三旨在利用所给数据对 $PM_{2.5}$ 浓度进行向前 1 期（3 小时）和向前 8 期（1 天）的预测，并通过均方误差（RMSE）和决定系数（R2）对模型进行评估。所给数据先去除异常值与缺失值，然后进行标准化处理，使它们具有相同的量纲，便于模型的训练和计算。本文使用滑动窗口法处理时间序列，采用线性回归、XGBoost、LightGBM 对数据进行预测，并对每个模型的评估指标进行对比。

5.2 数据预处理

数据集包含了 2017 年至 2021 年的气象和 $PM_{2.5}$ 浓度数据。使用滑动窗口法处理时间序列数据，构建输入窗口为 8 个时间步，输出窗口分别为 1 步（3 小时预测）和 8 步（24 小时预测）的数据集。对数据集进行归一化处理。

预测策略针对不同时间跨度采用差异化方式，在短期预测（3 小时）时，直接对下一个时间步的 $PM_{2.5}$ 浓度进行预测；而对于长期预测（24 小时），则采用分步预测的方法，每 3 小时作为一个预测步长，累计进行 8 步预测，以此来实现对 $PM_{2.5}$ 浓度在不同时间周期的有效预估。

5.3 模型分析

5.3.1 线性回归模型

线性回归作为一种简单且基础的机器学习模型，通常被用作基准模型。使用线性回归建立一个相对简单的基线，初步了解数据中变量之间的关系，并且与更复杂的模型相比，评估复杂模型是否真的有必要以及其性能提升的程度。

在线性回归模型中，令 $PM_{2.5}$ 浓度作为目标变量，它可以被表示为温度、降水量、相对湿度等一系列特征的线性组合。线性回归的模型公式如式(3.1)所示，其中 y 是预测的 $PM_{2.5}$ 浓度，$x_1, x_2, \ldots, x_n$ 是城市、降水、湿度等输入特征的归一化值，β_0 是所有 $x_1, x_2, \ldots, x_n$ 为零时 y 的期望值，$\beta_1, \beta_2, \ldots, \beta_n$ 是需要设计的模型参数，ϵ 是模型暂时不能解释的随机误差。

此时我们使用最小二乘法的估计方法估计参数 $\beta_1, \beta_2, \ldots, \beta_n$ ，它通过最小化误差项的平方和来估计系数，使用公式如式(3.2)所示。其中 m 是样本数量，y_i 和 x_{ij} 分别是第 i 个观测的因变量和第 j 个观测的自变量的值。

最后由式(3.3)矩阵运算可得参数的值。其中 $\hat{\beta}$ 是系数的估计向量，X 是自行设计的矩阵，y 是因变量的值向量。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \tag{5.1}$$

$$\sum_{i=1}^{m} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{ij}))^2 \tag{5.2}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{5.3}$$

5.3.2XGBoost 模型

XGBoost 是一种高效的梯度提升决策树算法，它通过软件和硬件优化减少了计算资源消耗，并提供了优异的预测性能。该算法利用 L1 和 L2 正则化来防止过拟合，并且能够处理稀疏数据和自动处理缺失值。XGBoost 通过并行化树构建和硬件优化，进一步提高了计算效率，所以在线性模型后考虑使用 XGBoost 模型进行更好的预测。

XGBoost 的核心在于构建一个加法模型，该模型的预测值是多个决策树的和。加法模型的公式如式(3.4)所示。对于第 i 个样本来说，它的预测值 $\hat{y}_i$ 是由 K 个决策树 f_k 的预测值累加得到的，其中每个决策树 f_k 都是从所有可能的决策树集合 $\mathcal{F}$ 中选取的，而 x_i 代表样本的特征向量。

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{5.4}$$

在训练过程中，XGBoost 的目标函数由两部分组成：损失项和正则项[22]。损失项衡量的是模型预测值与实际值之间的差异，而正则项则用于控制模型的复杂度，防止过拟合，公式如式(3.5)所示。具体地，目标函数Obj包含了对所有样本的损失函数l的和，以及对新加入的决策树f_t的正则化项$\Omega(f_t)$。损失函数l通常是均方误差，而正则化项Ω包括了树的结构复杂度和叶子节点的 L2 正则化[23]。

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{5.5}$$

XGBoost 还采用了二阶泰勒展开来近似损失函数，这种方法可以在当前模型预测值附近对损失函数进行线性和二次近似。其公式如式(3.6)所示，通过这种方式，可以计算出每个样本在新加入决策树f_t时的梯度g_i和二阶导数h_i，这有助于优化决策树的构建。

$$l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) \approx l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \tag{5.6}$$

最后，XGBoost 定义了树的复杂度，公式如式(3.7)所示，其中包括了叶子结点总数的正则化项γ和叶子结点权重的 L2 正则化项λ。这种定义有助于控制每棵树的复杂度，进一步防止过拟合，同时提高模型的泛化能力。通过这些关键公式和构建方法，XGBoost 能够在保持高效率的同时，提供优秀的预测性能。

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| \| w \| \|^2 \tag{5.7}$$

5.3.3LightGBM 模型

LightGBM 的核心机制基于梯度提升框架，其目标函数与公式(3.5)相同，但为了提高训练效率和减少内存使用，LightGBM 采用了直方图基学习。这种方法通过构建直方图来代替对连续特征值的直接排序，从而加速了模型的训练过程。直方图的使用不仅减少了数据扫描的次数，还降低了对内存的需求，使得 LightGBM 在处理大规模数据集时更加高效。总的来说，LightGBM 通过结合梯度提升的目标函数和直方图基学习，实现了在保持模型预测准确性的同时，显著提升了训练速度和处理大数据的能力。

5.4 预测结果与模型评估

5.4.1 评估指标

（1）均方根误差（RMSE）

RMSE 是一种衡量预测模型准确性的指标，公式如式(3.8)所示[24]，它通过计算预测值与实际值之差的平方的平均值的平方根来得出。这个指标能够量化预测误差的大小，RMSE

值越低，意味着模型的预测结果越接近实际值，即模型的预测精度越高。RMSE 的一个显著优点是它对较大的误差给予更大的惩罚，因为误差在计算过程中被平方，这使得模型在训练时更加注重减少较大的预测偏差。

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{5.8}$$

（2）平均绝对误差（MAE）

MAE 也是一种评估预测模型准确性的指标，公式如式(3.9)所示，它通过计算预测值与实际值之差的绝对值的平均值来衡量预测误差的平均大小。MAE 的一个关键特点是它对所有误差都给予相同的权重，即不论误差的大小，每个误差都被视为同等重要。这与 RMSE 不同，后者通过对误差进行平方来对较大的误差施加更大的惩罚。因此，MAE 提供了一个更为平衡的误差度量，使得模型的预测精度更容易被直观理解：MAE 值越小，表示模型的预测结果越接近实际值，即模型的预测精度越高[25]。

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{5.9}$$

（3）决定系数（R2）

R^2，也称为决定系数，是一个衡量模型预测能力的重要统计指标，公式如式(3.10)所示。它表示模型解释的变异性占总变异性的比例，其值介于 0 到 1 之间。R^2值越接近 1，意味着模型能够解释更多的数据变异性，从而表明模型的解释能力和预测精度越高。R^2的一个显著优点是它提供了一个直观的比例值，使得我们能够清晰地理解模型对数据的拟合程度。

$$R^2=1-\frac{\sum_{i=1}^{n}(y-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{5.10}$$

5.4.2 短期预测（3 小时）结果

在分析时间序列数据时，我们通过样本索引（横轴）和 $PM_{2.5}$ 浓度（纵轴）来观察空气污染程度，短期预测结果对比图和散点图如图 5.1 所示。其中，黑色实线代表实际观测的 $PM_{2.5}$ 浓度，蓝色、橙色和绿色虚线分别代表线性回归、XGBoost 和 LightGBM 模型的预测值。整体来看，三种模型的 3 小时预测曲线在大部分区域与实际观测值紧密跟随，但在样本索引 20 附近的峰值和谷值处，预测值与实际值之间存在一定偏差。

在散点图中，横轴代表实际观测值，纵轴代表模型预测值，每个散点对应一个样本的实际值与预测值的关系。红色虚线表示理想预测线，即预测值与实际值完全吻合的理想状态。分析显示，大多数散点集中在红色虚线附近，这表明模型的预测结果与实际观测值大致相符。尽管如此，仍有部分散点偏离了这条理想线，特别是在 $PM_{2.5}$ 浓度较高的区域。

5.4.3 长期预测（24 小时）结果

对 $PM_{2.5}$ 浓度进行预测时，预测精度和模型表现受到时间尺度的影响[26]。在 3 小时预测中，模型的预测值与实际值的吻合度较高，特别是在 $PM_{2.5}$ 浓度快速变化的区域；在 24 小时预测中，模型的预测值与实际值的偏差相对较大，尤其是在浓度的峰值和谷值处。在 3 小时预测中，三种模型的预测曲线与实际值的波动趋势较为一致，尤其是在浓度变化剧

烈的区域。相比之下，在 24 小时预测中，XGBoost 和 LightGBM 的预测曲线在某些区域与实际值的偏差较大，而线性回归的预测曲线在某些区域与实际值的偏差较小。

预测曲线的平滑度方面，3 小时预测的曲线整体上更为平滑，表明模型能够较好地捕捉到浓度的快速变化。相反，24 小时预测的曲线在某些区域显示出较大的波动，并且对比 3 小时预测时的散点图，可以发现 24 小时预测的点分布更为分散，尤其是在 $PM_{2.5}$ 的高浓度区域，这表明随着预测时间尺度的延长，模型的准确性有所下降。

在低浓度区域（0-200），线性回归、XGBoost 和 LightGBM 的预测结果仍然较为接近，但在高浓度区域，三种模型的预测结果分散程度增加，显示出较大的差异。此外，存在一些远离红色虚线的点，这些点代表了模型预测与实际值之间的较大误差。综合分析表明，从 3 小时到 24 小时的预测，模型的预测准确性在高浓度区域有所下降，而在低浓度区域，三种模型的预测结果较为一致，但在高浓度区域，预测结果的分散程度增加。

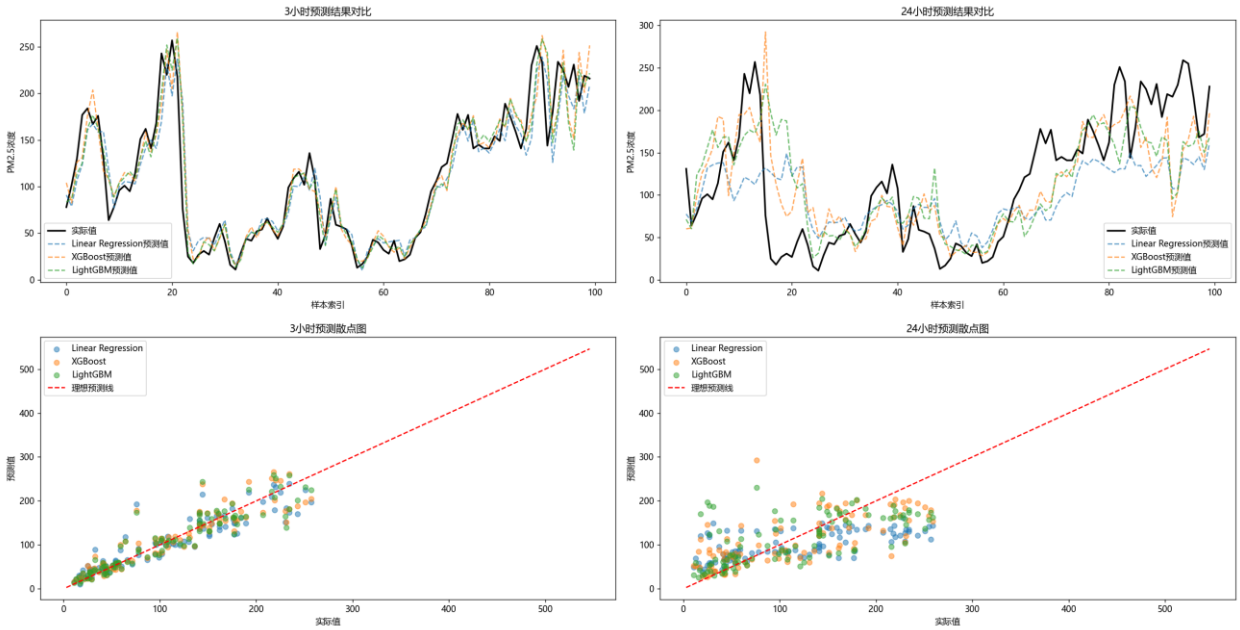

图 5.1 模型预测结果

5.4.4 模型评估

（1）短期预测模型评估

在对 $PM_{2.5}$ 浓度的短期预测中利用上述评估指标对模型进行评价。LightGBM 模型表现最优，其均方误差最小，平均绝对误差最小，决定系数最接近 1；XGBoost 模型表现稍次；线性回归模型表现最差。模型具体评估指标数据如表 3.1 所示，RMSE 和 R2 柱状图如图 5.2 所示，总的来说短期预测效果较好，R^2值都在 0.85 以上。

表 5.1 短期预测（3 小时）评估指标

模型	RMSE	MAE	R^2
线性回归	19.24	11.26	0.855
XGBoost	18.23	10.48	0.870
LightGBM:	17.84	10.31	0.876(最佳)

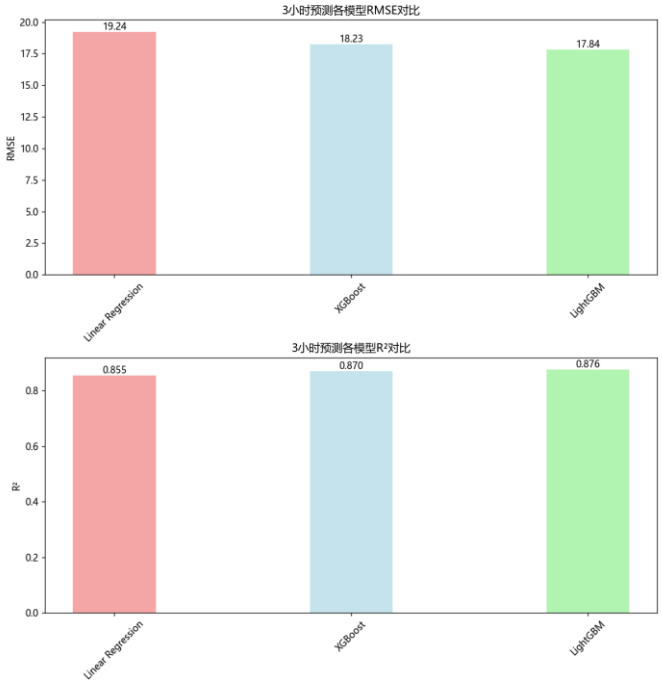

图 5.2 3 小时预测 RMSE、R^2 变化

（2）长期预测模型评估

从表 5.2 中的数据可以看出，线性回归模型在所有时间步的预测中，虽然 R^2值相对较高，但其 RMSE 和 MAE 值显著高于 XGBoost 和 LightGBM 模型。这表明线性回归模型在预测 $PM_{2.5}$ 浓度时，误差较大，且对数据的解释能力有限。相比之下，XGBoost 和 LightGBM 模型在预测精度和解释能力上均优于线性回归模型。特别是在 RMSE 和 MAE 这两个误差指标上，LightGBM 模型通常略低于 XGBoost，显示出更优的预测性能。

表 5.2 长期预测（24 小时）评估指标

模型时间步	RMSE	MAE	R^2
线性回归 1	19.239822	11.260021	0.855356
XGBoost1	18.090509	10.426188	0.872121
LightGBM:1	17.823524	10.299977	0.875867
线性回归 2	26.380987	16.447757	0.727956
XGBoost2	23.455181	14.243194	0.784952
LightGBM:2	23.274937	14.197571	0.788245
线性回归 3	30.053794	19.286654	0.646827
XGBoost3	25.816274	16.068613	0.739399
LightGBM:3	25.808831	16.104126	0.739550
线性回归 4	32.248398	20.946981	0.593243
XGBoost4	27.625879	17.366644	0.701495
LightGBM:4	27.734184	17.417391	0.699150
线性回归 5	33.733947	22.022089	0.554726
XGBoost5	28.958986	18.443226	0.671860
LightGBM:5	29.257705	18.522881	0.665055
线性回归 6	34.848217	22.783575	0.524176
XGBoost6	30.306139	19.279734	0.640130
LightGBM:6	30.460907	19.388552	0.636445

线性回归 7	35.942191	23.554385	0.493045
XGBoost7	31.493796	20.203112	0.610766
LightGBM:7	31.594261	20.237386	0.608279
线性回归 8	36.985197	24.290139	0.462022
XGBoost8	32.867404	21.135533	0.575146
LightGBM:8	32.965502	21.246936	0.572607

随着预测时间步的增加，所有模型的 RMSE 值均呈现上升趋势，预测效果随时间步的增加而降低，RMSE、R2 变化图如图 5.5 所示，这可能是由于长期预测受到更多不确定因素的影响。MAE 变化图如图 5.6 所示，XGBoost 和 LightGBM 模型的误差增长相对较为平稳，而线性回归模型的误差增长速度较快，这进一步证实了线性模型在长期预测中的局限性。

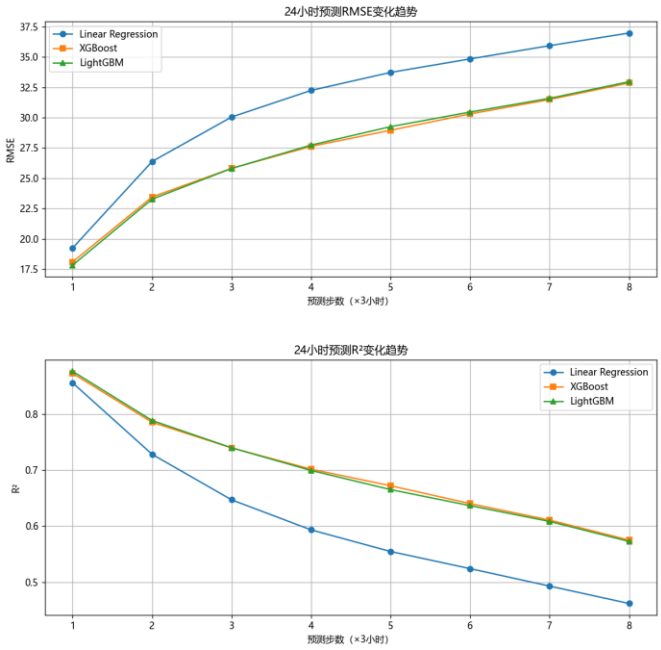

图 5.5 24 小时预测 RMSE、R^2 变化

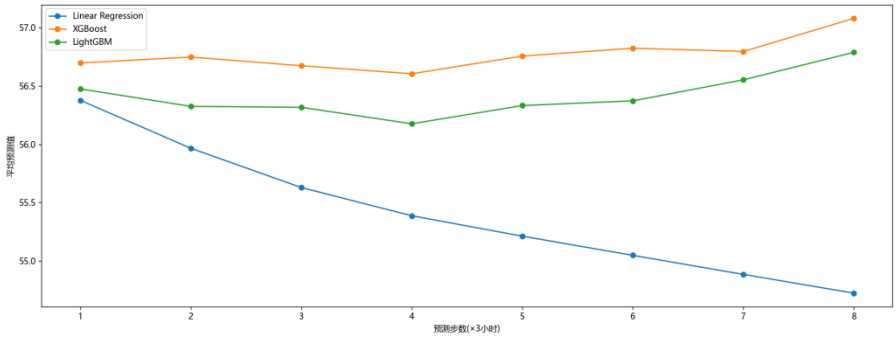

图 5.6 24 小时预测 MAE 变化

综合考虑 RMSE、MAE 和 R^2三个评估指标，对于长期预测 $PM_{2.5}$ 浓度的任务，XGBoost 和 LightGBM 模型相比于线性回归模型提供了更好的预测性能。在这两种模型中，LightGBM 模型在大多数情况下表现略优于 XGBoost.尤其是在 RMSE 和 R2 上，三个模型的直观对比如图 5.7 所示，因此，建议在实际应用中优先考虑使用 LightGBM 模型进行 $PM_{2.5}$ 浓度的长期预测。

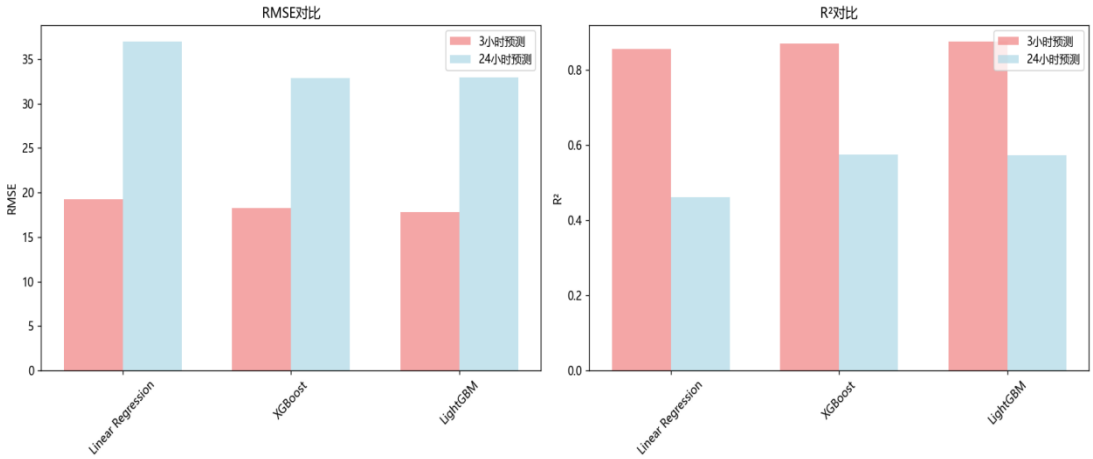

图 5.7 不同模型预测 RMSE、R^2 变化

（3）数据集划分

在本文中，对数据集进行的划需确保模型训练和评估的有效性。于是遵循时间序列的自然顺序，避免随机打乱数据，保持时间连续性，并采用滑动窗口方法导致相邻样本数据重叠，同时确保训练集和测试集之间无重叠，以维护样本独立性[27]。选择 80/20 的数据分割比例是基于其在提供充足训练数据、保留足够测试数据以及符合广泛接受的经验法则方面的优势。此外，考虑到数据总量、时间跨度和季节代表性，以确保测试集能够全面反映不同环境下的模型表现。这些细致的数据划分策略为模型的稳健性和可靠性提供了坚实的基础，确保了模型评估的公正性和泛化能力。

5.5 未来优化方向

通过对比 LightGBM、XGBoost 和线性回归模型在短期（≤6 小时）、中期（6-12 小时）和长期（>12 小时）预测 $PM_{2.5}$ 浓度的性能，据此提出未来优化方向。

为了提升 $PM_{2.5}$ 浓度预测的准确性和可靠性，未来的研究将探索增加历史数据特征[28]，以增强模型对时间序列趋势的捕捉能力，这对于理解长期趋势至关重要。同时，进行时间特征工程，以提升模型对时间依赖性的理解，这对于捕捉周期性和季节性变化非常关键。此外，尝试应用深度学习模型来处理长期预测中的复杂非线性关系，这可能对提高预测精度有显著效果。实施上，将采用滚动预测方法减少时间延迟对预测结果的影响，提高预测的准确性和及时性。增加预测的频率将使我们能够更快地响应环境变化，并及时调整预测策略。通过整合多个模型的预测结果，可以增强预测的稳定性和精确度。这些措施将共同提升 $PM_{2.5}$ 浓度预测的效果，为环境管理和决策提供更有力的支持。

在短期预测（≤6 小时）中，LightGBM 模型表现最佳，预期 R^2大于 0.78，RMSE 小于 24。这表明对于短期预测，LightGBM 因其高效的处理速度和较低的内存消耗，是一个理想的选择。

在短中期预测（6-12 小时）中，建议使用集成模型，预期 R^2大于 0.70，RMSE 小于 28。集成模型能够结合多个模型的优点，提高预测的稳定性和准确性。

在长期预测（>12 小时）中，建议每 12 小时更新一次预测或采用滑动窗口方法，以减少误差累积，预期 RMSE 可能超过 30。这种方法有助于捕捉时间序列数据中的变化趋势。

6. 参考文献

[1]师玉玉,邵奇,王雪茜.不同剂量下 PM2.5 介导的肺组织氧化损伤与上皮屏障破坏[J/OL].中国实验动物学报,1-7[2024-11-25]

[2]王蒙,吕曼青,周雪冬,等.2014-2022 年青岛大气污染健康损害经济价值评估[J].环境科学与技术,2023,46(12):203-211.

[3]梅承志,李斌,何友江,等.2015—2023 年天山北坡区域 PM2.5 与 O3-8 h 变化趋势及影响因素分析[J/OL].环境科学学报,1-12[2024-11-25].

[4]宫璇.基于 ERP 系统的 H 企业物资管理应用研究[D].贵州大学,2022.

[5]张培根.近邻排序算法研究及在中文数据清洗中的应用[D].江苏科技大学,2018.

[6]MacQueen, J. Some methods for classification and analysis of multivariate observations.1967.

[7]Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu.A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 226–231. 1996.

[8]J. Qi, Y. Yu, L. Wang and J. Liu, "K*-Means: An Effective and Efficient K-Means Clustering Algorithm," 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), pp. 242-249, 2016.

[9]K. Khan, S. U. Rehman, K. Aziz, S. Fong and S. Sarasvady. "DBSCAN: Past, present and future," The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014), pp. 232-238, 2014.

[10]CATENI S, COLLA V, VANNUCCI M, et al. Outlier detection methods for industrial applications[J]. Advances in Robotics, Automation and Control, 2008: 265–282.

[11]朱箫,沈晓菁.基于 MATLAB 机器学习的数据预处理研究[J].科技资讯,2024,22(17):19-22.[8]朱箫,沈晓菁.基于 MATLAB 机器学习的数据预处理研究[J].科技资讯,2024,22(17):19-22.

[12]郭凤娟,贾超,窦春苓.克拉玛依 PM2.5 浓度分析及逆温层对其影响研究[J].地理空间信息,2022,20(10):61-64.

[13]谢文豪,张强,杨方社,等.运城市 2020—2021 年采暖季 PM2.5 污染成因分析[J].环境科学学报,2023,43(11).

[14]黄枭. 基于地理加权回归模型的湖北省 PM2.5 遥感反演方法研究[D]. 武汉大学, 2019. DOI:10.27379/d.cnki.gwhdu.2019.001220.

[15]李松洲. 城市空气质量的 PM2.5 浓度预测及监测网络优化研究[D]. 太原理工大学, 2020. DOI:10.27352/d.cnki.gylgu.2020.001279.

[16]B. Qi, Y. Jiang, H. Wang and J. Jin, "Multi-Source PM2.5 Prediction Model Based on Fusion of Graph Attention Networks and Multiple Time Periods," in IEEE Access, vol. 12, pp. 57603-57612, 2024, doi: 10.1109/ACCESS.2024.3390934.

[17]罗奥荣.基于支持向量回归机的大气 PM2.5 浓度预测模型研究[D].北京工业大学,2018.

[18]赵乾.基于气象参数西安城市热岛与空气质量特征及关联性研究[D].西安工程大学,2018.

[19]张帆.气象条件影响下县域苹果产量预测及其趋势分析[D].山东农业大学,2022.DOI:10.27277/d.cnki.gsdnu.2022.000920.

[20]L. Lin et al., "Using Machine Learning Approach to Evaluate the PM2.5 Concentrations in China from 1998 to 2016," 2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics), Hangzhou, China, 2018, pp. 1-5, doi: 10.1109/Agro-Geoinformatics.2018.8475987.
[21]梁丽思.东南沿海经济区 PM2.5 浓度的时空变化、影响因素及反演方法研究[D].桂林理工大学,2020.DOI:10.27050/d.cnki.gglgc.2020.000609.
[22]刘沐阳.基于监管数据和 XGBoost 模型的建设工程质量评价方法研究[J].项目管理技术,2020,18(11):56-62.
[23]G. Thangarasu and K. R. Alla, "Optimization of Livestock Farming Using Multi-Tasking Regularisation Model," 2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon), Singapore, Singapore, 2023, pp. 942-945, doi: 10.1109/SmartTechCon57526.2023.10391686.
[24]郑堪尹.考虑协同效应的项目全生命周期服务商组合选择研究[D].长安大学,2022.DOI:10.26976/d.cnki.gchau.2022.001375.
[25]R. Murugan and N. Palanichamy, "Smart City Air Quality Prediction using Machine Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1048-1054, doi: 10.1109/ICICCS51141.2021.9432074.
[26]秦喜文,刘媛媛,王新民,等.基于整体经验模态分解和支持向量回归的北京市 PM2.5 预测[J].吉林大学学报(地球科学版),2016,46(02):563-568.DOI:10.13278/j.cnki.jjuese.201602206.
[27]G. C. Claasen, P. Martin and K. Graichen, "Error growth due to noise during occlusions in inertially-aided tracking systems," 2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, 2013, pp. 5781-5787, doi: 10.1109/ICRA.2013.6631408.
[28]J. Cai, C. Gu, K. Fang, L. Wang and M. Lv, "Long-term PM2.5 Concentration Prediction Using Hybrid Deep Learning Model," 2023 4th International Conference on Computers and Artificial Intelligence Technology (CAIT), Macau, Macao, 2023, pp. 12-16, doi: 10.1109/CAIT59945.2023.10469109.

7. 附录

1、数据预处理采用编程软件及工具箱为：

编程软件：

PyCharm Community Edition 2024.2.1（IDE 开发环境）

Python 编程语言

主要工具箱/库：

pandas (pd)：用于数据处理和分析

numpy (np)：用于数值计算

scikit-learn (sklearn)：用于机器学习和数据处理

StandardScaler：数据标准化

DBSCAN：密度聚类

NearestNeighbors：寻找最近邻

TSNE：降维可视化

matplotlib (plt)：用于数据可视化

seaborn (sns)：用于统计数据可视化

2、问题一采用编程软件及工具箱为：

编程软件：

PyCharm Community Edition 2024.2.1（IDE 开发环境）

Python 编程语言

主要工具箱/库：

pandas (pd)：数据处理和分析

numpy (np)：数值计算

matplotlib (plt)：数据可视化

seaborn (sns)：统计数据可视化

scipy：科学计算

scipy.stats：统计分析

scipy.spatial.distance：空间距离计算（pdist, squareform）

3、问题二采用编程软件及工具箱为：

编程软件：

PyCharm Community Edition 2024.2.1（IDE 开发环境）

Python 编程语言

主要工具箱/库：

pandas (pd)：数据处理和分析

numpy (np)：数值计算

matplotlib (plt)：数据可视化

seaborn (sns)：统计数据可视化

scipy：科学计算

scipy.stats：统计分析

scikit-learn (sklearn)：机器学习和统计分析

sklearn.linear_model：线性回归模型

sklearn.metrics：模型评估

sklearn.preprocessing：数据预处理（StandardScaler, OneHotEncoder）

4、问题三采用编程软件及工具箱为：

编程软件：

PyCharm Community Edition 2024.2.1（IDE 开发环境）

Python 编程语言

主要工具箱/库：

pandas (pd)：数据处理和分析

numpy (np)：数值计算

matplotlib (plt)：数据可视化

seaborn (sns)：统计数据可视化

scikit-learn (sklearn)：机器学习工具

preprocessing：数据预处理(StandardScaler)

metrics：模型评估(mean_squared_error, mean_absolute_error, r2_score)

model_selection：数据集划分(TimeSeriesSplit)

linear_model：线性回归模型

XGBoost (xgb)：梯度提升树模型

LightGBM (lgb)：轻量级梯度提升框架