参赛承诺书

提交包含此承诺书的 pdf 文件，表明所有此文件的作者共同承诺：

我们完全清楚，在竞赛开始后参赛队员不能以任何方式，包括电话、电子邮件、“贴吧”、QQ 群、微信群等，与队外的任何人（包括指导教师）交流、讨论与赛题有关的问题；无论主动参与讨论还是被动接收讨论信息都是严重违反竞赛纪律的行为。

我们以中国大学生名誉和诚信郑重承诺，严格遵守竞赛章程和参赛规则，以保证竞赛的公正、公平性。如有违反竞赛章程和参赛规则的行为，我们将受到严肃处理。

我们授权北京理工大学数学建模竞赛组织方，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

2025.3

----------本页下方保持空白----------

关于大气边界层高度及其他气象因素的量化模型建立与研究

摘要

在研究空气污染扩散、天气预报、气候变化等问题时，大气边界层高度(ABLH)是一个非常重要的指标。大气边界层高度影响气象要素分布与变化，其高度变化影响湍流活动、热量与水汽交换 ，进而影响气温、湿度、风速风向，准确掌握 ABLH 有助于提高天气预报精度，特别是对强对流天气、局地气象变化的预测。ABLH 与很多因素有关，比如温度、湿度、风速、气压等。位于嘉兴的北京理工大学长三角研究院曾利用激光雷达观测大气边界层高度，并同步获取了当地的温度、湿度、风速、气压等相关数据，我们有效利用相关数据建立正确且合适的数学模型解决对应的问题。

针对问题一，题目要求建立数学模型，分析 ABLH 与温度、湿度、风速、气压等因素之间的量化关系，并评估这些因素对 ABLH 的影响大小。经过讨论与思考，我们使用四种模型（分别是 LinearRegression（线性回归模型），Ridge（岭回归模型），Lasso（套索回归模型），ElasticNet（弹性网络回归模型））对相关数据进行处理、训练，其中，在线性回归模型中，我们对原始特征进行了三阶多项式扩展，以更好地捕捉 ABLH 与气象因素之间的非线性关系。最后我们对这三种模型的效果进行对比，得出效果最佳的模型。

针对问题二，题目要求我们结合当地空气质量数据，选择合理的数学模型分析 ABLH 与空气质量相关指标($PM_{2.5}$、PM_{10}、NO_2、AQI 等)的量化关系。那我们经过分析，选取每日 **24 小时的 ABLH 值作为 1*24 的输入向量，将 AQI 等指标作为输出向量**，决定采用 **GAM（Generalized additive model）**处理 ABLH 与空气质量相关指标之间的关系并进行检验，同时使用**随机森林（Random Forest）算法**与 **XGBoost（eXtreme Gradient Boosting）算法**进行对比得出解答。

针对问题三，题目要求建立一个数学模型，根据天气预报数据来预测次日 ABLH，并说明该模型的合理性。那么我们结合第一问和第二问，分别训练出两组模型，一组使用解决第一问的四种模型（线性回归模型、岭回归模型、套索回归模型、弹性网络回归模型）训练，另一种使用解决第二问的两大模型算法（随机森林算法与 XGBoost 算法）训练，分别在两组里面**选择效果最好的两个模型作为基学习器**，并**以线性回归作为元学习器对基学习器进行融合**。最终，我们在验证集或测试集上对融合模型的预测精度进行对比分析，以检验其在次日 ABLH 预测中的合理性和鲁棒性，然后得出最终模型。

【关键词】大气边界层高度 空气质量指标 线性回归 决策树

一、问题背景

1.1 问题背景

在大气科学领域，大气边界层高度（ABLH）作为关键参数，深刻影响着诸多重要研究方向。对于天气预报而言，精确掌握大气边界层高度是提升预报准确性的关键，大气边界层内的风速、温度、湿度等气象要素的变化，与边界层高度密切相关。为深入探究大气边界层高度及其相关影响因素，北京理工大学长三角研究院于嘉兴开展了相关研究。我们将依据长三角研究院所给出的相关数据对具体问题进行具体研究，并建立相关数学模型进行求解。

1.2 问题要求

问题一：请建立数学模型，分析 ABLH 与温度、湿度、风速、气压等因素之间的量化关系，评估这些因素对 ABLH 的影响大小。

问题二：ABLH 的变化会影响到当地空气质量，请结合当地空气质量数据，选择合理的数学模型分析 ABLH 与空气质量相关指标($PM_{2.5}$、PM_{10}、NO_2、AQI 等)的量化关系，空气质量数据参见附件 2。

问题三：请建立一个数学模型，根据天气预报数据来预测次日 ABLH，并说明该模型的合理性。

二、问题分析

2.1 问题一的分析

题目要求我们建立数学模型，分析 ABLH 与温度、湿度、风速、气压等因素之间的量化关系，评估这些因素对 ABLH 的影响大小。那么我们从北京理工大学长三角研究院在嘉兴获取的数据集中，提取大气边界层高度（ABLH）、温度、湿度、风速、气压等相关变量的数据，先计算各变量之间的相关系数以排除掉重复指标，然后我们对变量进行标准化处理，使得各变量具有相似的尺度，避免因变量尺度差异影响模型性能。接下来，我们采用提及的四种模型（LinearRegression（线性回归模型），Ridge（岭回归模型），Lasso（套索回归模型），ElasticNet（弹性网络回归模型））进行训练：

首先是 LinearRegression（线性回归模型），假设 ABLH 与温度、湿度、风速、气压等因素之间存在线性关系，使用训练数据集，通过最小化残差平方和（RSS）来估计回归系数β_i；

然后是 Ridge（岭回归模型），在线性回归的基础上，引入 L2 正则化项，以防止过拟合。在 scikit-learn 中，使用 Ridge 类，通过调整 α 参数，对模型进行训练。通过交叉验证等方法选择合适的 α 值，以优化模型性能；

再是 Lasso（套索回归模型），Lasso 回归同样是在线性回归基础上，引入 L1 正则化项，能够进行特征选择，使得部分回归系数为零。使用 scikit-learn 中的 Lasso 类，通过调整 α 参数，对模型进行训练，通过交叉验证等方式确定最佳的 α 值，使得模型在训练集和验证集上都有较好的表现；

最后是 ElasticNet（弹性网络回归模型）弹性网络回归结合了 L1 和 L2 正则化，在 scikit-learn 中，使用 ElasticNet 类，通过调整 α 和 β 参数，对模型进行训练。同样通过交叉验证等方法确定最优的 α 和 β 值。最后我们进行对

比分析，得到最优的数学模型。

2.2 问题二的分析

大气边界层高度（ABLH）的变化会对当地空气质量产生影响，而空气质量通常由多个指标来衡量，如 $PM_{2.5}$、PM_{10}、NO_2和 AQI 等。需要选择合适的数学模型来量化 ABLH 与这些空气质量指标之间的关系，而随机森林算法和 XGBoost 算法都是强大的机器学习算法，能够处理复杂的非线性关系，适合用于分析这种多因素之间的相互关系。

我们首先构建特征编码,选取每日 24 小时的 ABLH 值作为 1*24 的输入向量，将 AQI 等指标作为输出向量进行训练。

对于广义可加模型（GAM）的应用，我们同样以每日 24 小时的 ABLH 作为输入特征，将 $PM_{2.5}$、PM_{10}、NO_2、AQI 等空气质量指标作为因变量，首先对原始数据进行随机划分与标准化预处理，并在训练集上拟合模型，以保证数据划分的透明度与稳定性。在模型构建阶段，为每个输入特征指定光滑基函数（如自然样条），并通过惩罚项控制函数的自由度，利用广义交叉验证（GCV）或 AIC 准则自动选择最优平滑参数，以有效捕捉 ABLH 与各质量指标之间的平滑非线性关系，同时避免过拟合。模型训练完成后，在测试集上使用 MSE、RMSE、MAE 和 R^2 等指标评估预测性能，并对每个平滑项的响应曲线进行可视化展示，从而直观揭示不同 ABLH 取值区间内各空气质量指标的预期变化趋势，兼顾了预测精度与可解释性

对于随机森林算法，我们结合当地空气质量数据，首先对数据进行标准化处理，将不同量级和范围的特征值转换到相似的区间，以提高模型的训练效果和稳定性，再将预处理后的数据划分为训练集和测试集来训练随机森林模型，调整模型的参数，如决策树的数量、最大深度、最小样本分裂数等。通过交叉验证来选择最优的参数组合，以避免过拟合和提高模型的泛化能力，在训练过程中，随机森林会自动计算每个特征的重要性得分，这可以帮助我们了解 ABLH 以及其他因素对空气质量指标的相对影响程度。最后使用测试集评估模型的性能，采用均方误差（MSE）、均方根误差（RMSE）、平均绝对误差（MAE）等指标来衡量模型预测值与实际值之间的差异，计算决定系数R^2，评估模型对数据的拟合程度，R^2 越接近 1 表示模型拟合效果越好。通过模型训练得到的特征重要性得分，分析 ABLH 与其他因素相比，对各个空气质量指标（$PM_{2.5}$、PM_{10}、NO_2、AQI 等）的影响程度。

对于 XGBoost 算法，我们采用与随机森林算法相同的评估指标，如 MSE、RMSE、MAE 和 R^2 等，在测试集上评估模型的性能，确保模型具有良好的泛化能力，同时利用 XGBoost 的一些可视化工具或分析方法，绘制特征贡献图、进行局部可解释模型 - 不可知解释（LIME）分析等，来深入理解 ABLH 与空气质量指标之间的量化关系，包括不同 ABLH 取值下空气质量指标的预期变化情况。通过以上两种算法的应用，可以有效地分析 ABLH 与空气质量相关指标之间的量化关系,为进一步研究大气边界层高度对空气质量的影响提供有力的支持。同时，两种算法的结果可以相互验证和补充，提高分析的可靠性和准确性。

2.3 问题三的分析

问题的目标是根据天气预报数据来预测次日的 ABLH。天气预报数据中包含温度、湿度、风速、气压等多种因素，这些因素与 ABLH 之间存在一定的关系。可以利用第一问和第二问中提到的多种模型进行训练,通过比较选择出效果较好的基学习器，再进行融合以得到更优的最终模型。首先是基于第一问四种模型的

训练思路，使用线性回归模型、岭回归模型、套索回归模型和弹性网络回归模型分别对训练集数据进行训练。在训练过程中，通过调整模型的参数，如岭回归和套索回归中的正则化参数，来优化模型的性能再利用测试集对训练好的四个模型进行评估。比较四个模型的评估结果，选择出在测试集上表现最好的两个模型作为基学习器。再是基于第二问两种算法的训练思路，我们使用随机森林算法和 XGBoost 算法对训练集数据进行训练。在训练过程中，通过调整算法的参数，如随机森林中决策树的数量、最大深度，XGBoost 中的学习率、树的数量、最大深度等，来优化模型性能，同样利用测试集对训练好的两个模型进行评估，使用与前面相同的评估指标。比较两个模型的评估结果，选择出在测试集上表现更好的两个模型作为基学习器。最后我们使用 Stacking 融合，将前面选择的四个基学习器的输出作为新的特征，再使用一个新的模型（如逻辑回归、支持向量机等）进行训练，得到最终的融合模型。

我们认为多种模型结合具有一定的合理性，第一问中的四种回归模型和第二问中的两种机器学习算法都有各自的特点和优势。回归模型简单直观，能够直接建立变量之间的线性或近似线性关系；随机森林算法和 XGBoost 算法则能够处理复杂的非线性关系，并且具有较好的抗过拟合能力。通过结合不同类型的模型，可以充分利用它们各自的优点，从多个角度捕捉天气预报数据与次日 ABLH 之间的关系，提高模型的预测准确性。

三、模型假设

1、数据完整性假设：假设收集到的嘉兴地区历史数据（包括温度、湿度、风速、气压、ABLH 观测值以及空气质量相关指标等）是完整的，不存在大量缺失值或关键数据缺失的情况，且数据记录准确，误差在可接受范围内。
2、独立性假设：假设不同时间点的观测数据之间相互独立，即当前时刻的 ABLH 以及各相关因素不受过去时刻的直接影响，不考虑时间序列上的自相关性。
3、平稳性假设：假设在研究的时间范围内，嘉兴地区的气候条件和大气环境相对稳定，ABLH 与各影响因素之间的关系不随时间发生显著变化，即不存在明显的季节性、周期性或趋势性变化，或者已对数据进行了适当的处理以消除这些非平稳因素的影响。
4、线性可加性假设：对于第一问中建立 ABLH 与温度、湿度、风速、气压等因素的量化关系模型，假设 ABLH 可以表示为这些因素的线性组合，即各因素对 ABLH 的影响是线性可加的。即使考虑到可能存在的非线性关系，也假设可以通过适当的变换将其转化为线性关系来处理。
5、模型适用性假设：假设所选择的数学模型（线性回归模型、岭回归模型、套索回归模型、弹性网络回归模型、随机森林算法、XGBoost 算法等）能够合理地描述 ABLH 与各相关因素之间的关系，这些模型在该问题的背景下具有较好的拟合能力和泛化能力，能够准确地捕捉到数据中的规律并用于预测和分析。
6、局部性假设：对于根据天气预报数据预测次日 ABLH 的模型，假设当天的天气预报数据能够充分反映次日 ABLH 变化的主要影响因素，不考虑远处地区的天气状况或其他未纳入模型的因素对本地次日 ABLH 的显著影响，即认为 ABLH 的变化主要由本地的天气因素决定，具有一定的局部性。

7、忽略次要因素假设：假设除了考虑的温度、湿度、风速、气压等主要因素以及空气质量相关指标外，其他未纳入模型的因素对 ABLH 的影响可以忽略不计，或者这些次要因素的综合影响在模型中通过随机误差项来体现。

四、模型的建立与求解

4.1 问题一：建立数学模型，分析 ABLH 与温度、湿度、风速、气压等因素之间的量化关系，评估这些因素对 ABLH 的影响大小。

4.1.1 研究 ABLH 与影响因素量化关系模型的需求分析

在研究空气污染扩散、天气预报、气候变化等问题时，大气边界层高度(ABLH)是一个非常重要的指标。 在空气污染扩散理论中，BLH 决定污染物垂直扩散能力。边界层高度越高，污染物混合空间越大，近地面浓度越低，之夜间稳定边界层较低时易形成雾霾；在天气预报中，ABLH 变化反映大气稳定度，影响云的形成、降水过程和局地环流发展；在研究气候变化中，ABLH 参数化方案直接影响地表通量、湍流交换和辐射传输的计算精度。因此，对于我们深入研究 ABLH 是非常必要的。而 ABLH 与很多因素有关，比如温度、湿度、风速、气压等，这时候通过数学方法对 ABLH 进行量化的探索便比较合适。

4.1.2 ABLH 与因素的相关性分析

（1）六边形箱密度图（Hexbin Plot）分析

首先，为了更好地使用潜在因素比如温度、湿度、风速、气压等对 ABLH 进行量化，我们可以通过比较直观的方式初步探索 ABLH 与潜在因素之间的相关性。由于我们这次采用的数据规模极其庞大，采用传统的散点图会因为这个原因而无法从数据中传递有效消息。这时候我们便可以采用更适合大规模数据的六边形箱密度图（Hexbin Plot）了。

六边形箱密度图是一种用于可视化二维数据分布的图表，比较适用于在数据点过于密集而导致传统散点图难以区分单个点的情况。它的实现方式是将数据空间划分为一系列六边形单元（或“箱”），并根据每个单元中的数据点数量进行着色，从而提供了数据分布情况的直观表示。六边形箱密度图清晰地展示了数据密度分布、整体趋势与异常区域，相比于传统散点图，其优势在于避免了数据点重叠导致的“墨团效应”、以颜色梯度直观展示数据密度、保留数据的空间分布消息。

在 python 中，我们将 csv 文件数据导入，调用 hexbin 函数生成六边形箱密度图，通过调整 gridsize 控制六边形大小，bins 改变颜色分级策略，cmap 改变颜色映射方案，mincnt 控制最小显示数据点数，从而生成一系列美观、简洁的六边形箱密度图，如图 1 所示。

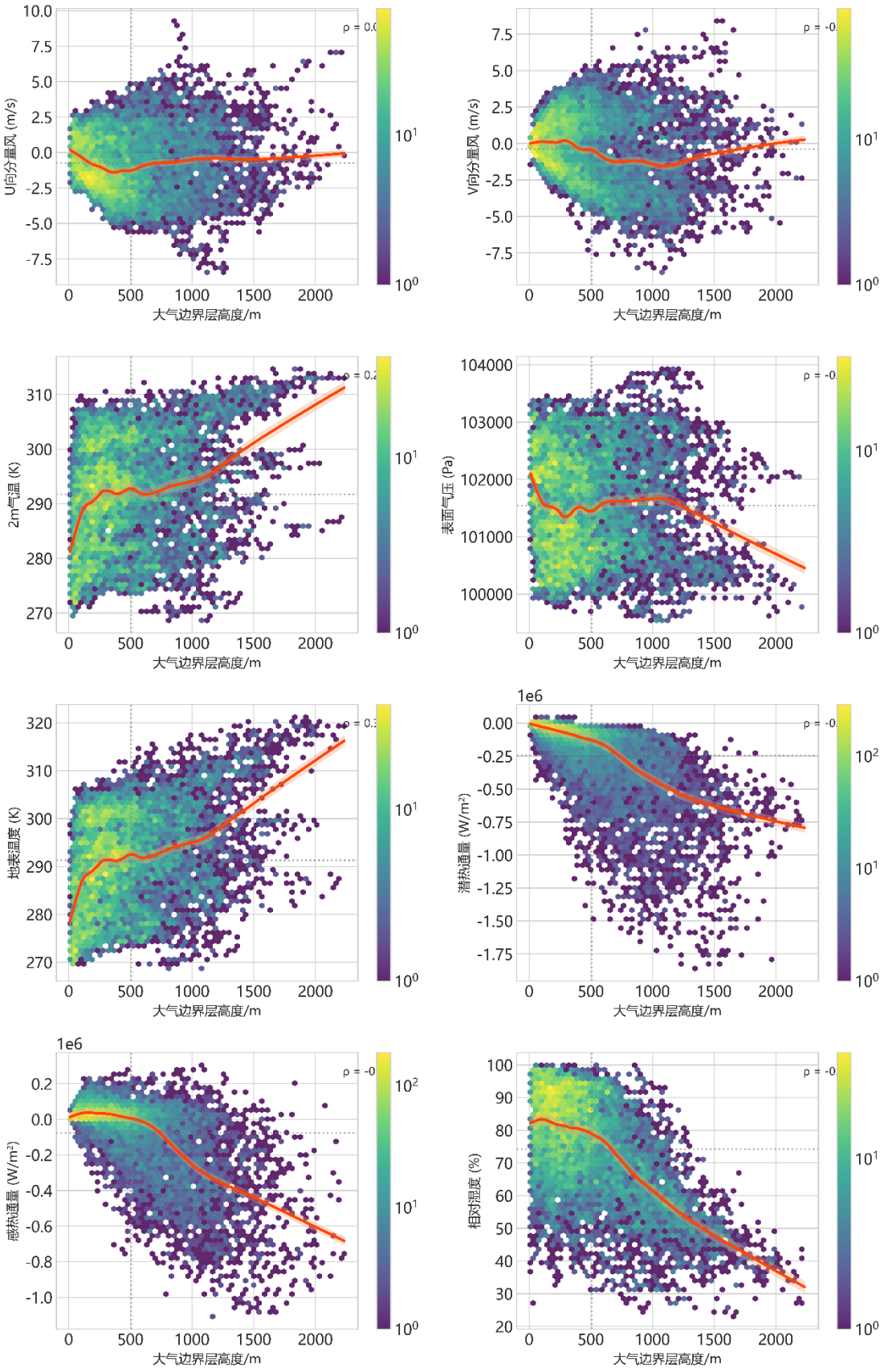

图 1 各因素与 ABLH 的六边形箱密度图

从图 1 可以看出 U 向分量风、V 向分量风与大气边界层高度 ABLH 之间没有明显的线性关系，数据点分布较为均匀分散。2m 气温、表面气压、感热通量、地表温度、相对湿度与大气边界层高度 ABLH 呈现出一定的负相关趋势，随着大气边界层高度的增加，2m 气温、相对湿度有下降的趋势，尤其是在较低的大气边界层高度时更为明显；表面压力显著降低，符合大气科学的基本原理;而地表温度的变化幅度较小。潜热通量与大气边界层高度 ABLH 之间存在一定的正相关关系，随着大气边界层高度的增加，潜热通量有增大的趋势，特别是在较高的大气边界层高度时更为明显。

（2）变量关系矩阵法分析

在使用六边形箱密度图探究完 ABLH 与可能影响因素的相关性后，我们也可以通过 SPLOM 来进行相关性分析这样也更有利于后续建立量化模型。

SPLOM，即散点图矩阵，是一种用于可视化多个变量之间关系的统计图表。它通过一个矩阵形式展示多变量数据集中的所有变量两两之间的关系。每个变量与其他变量的关系通过二维散点图表示，而每个变量自身的分布则通常通过对角线上的直方图或密度图来展示。为了在散点图中寻找潜在的关系，我们可以叠加 LOWESS 平滑曲线。

LOWESS 是一种非参数回归方法，适用于探索数据中的潜在趋势。它的核心原理是局部加权线性回归，其步骤如下：

① 局部加权线性回归

对于目标点x_i，选取邻域半径d，邻域内的点x_j满足$|x_j - x_i| < d$。

权重函数采用三次立方核（Tricube Kernel）：

$$w_j(x_i) = \begin{cases} \left(1 - \left|\frac{x_j - x_i}{d}\right|^3\right)^3, & if\,|x_j - x_i| < d \\ 0, & else \end{cases}$$

对于领域内每个点x_j，假设局部线性关系：

$$y_j = \beta_0 + \beta_1 x_j + \epsilon_j$$

通过加权最小二乘法求解系数β_0, β_1,目标是最小化：

$$\sum_j w_j(x_i) \cdot [y_j - (\beta_0 + \beta_1 x_j)]^2$$

解得：

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y$$

其中：

$X = [1, x_j]$为设计矩阵

$W = diag\left(w_j(x_i)\right)$为权重矩阵

$Y = [y_j]$为响应变量向量

目标点x_i的拟合值为：

$$\widehat{y_i} = \beta_0 + \beta_1 x_i$$

②稳健性迭代

为了降低异常值影响，引入迭代重加权：

1. 首次拟合后计算残差$r_j = y_j - \hat{y}_j$
2. 更新权重：

$$w_j^{robust}(x_i) = w_j(x_i) \cdot B\left(\frac{r_j}{6 \cdot MAD}\right)$$

其中：

$B(u)$为 Turkey 双权函数：

$$B(u) = \begin{cases} (1-u^2)^2, & |u| < 1 \\ 0, & else \end{cases}$$

MAD 为中位数绝对偏差

3. 用更新后的权重重新拟合，迭代至收敛，一般需要 3~5 次

LOWESS 的最终拟合曲线由所有局部回归结果组成：

$$\hat{y}(x) = \sum_{i=1}^{n} \hat{y}_i \cdot I\left(x \in \text{邻域}(x_i)\right)$$

其中$I(\cdot)$为指示函数，表示仅使用邻域内的拟合值。

LOWESS 无需预设全局函数形式就可以捕捉复杂趋势，还可以抵抗异常值干扰。这样，在 SPLOM 中，对角线上显示直方图，以展示单个变量的数据分布；非对角线上使用散点图，并叠加 LOWESS 平滑曲线，以揭示潜在的趋势和非线性关系。使用 python 进行编程绘图，得到如下图所示：

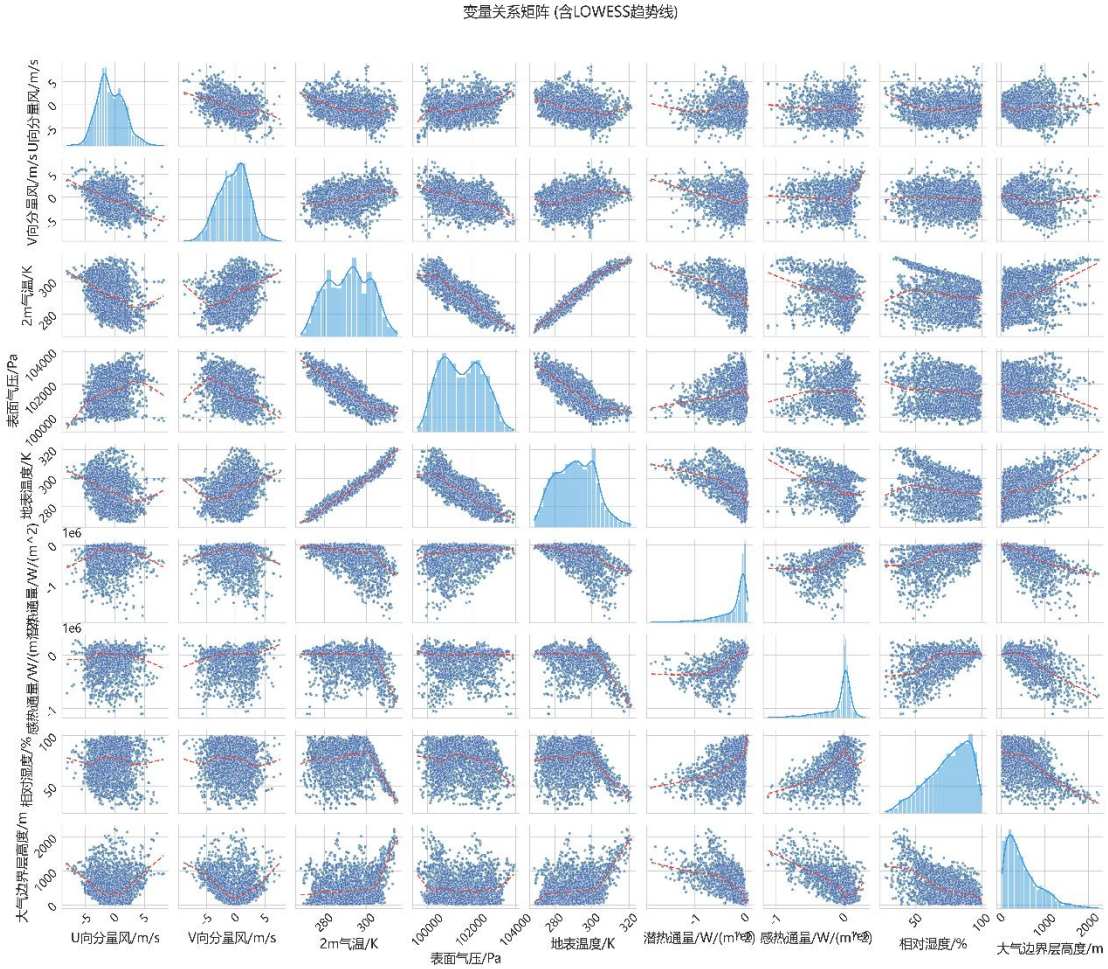

图 2 变量关系 SPLOM 矩阵（含 LOWESS 趋势线）

从矩阵对角线可以看出，U 向分量风、V 向分量风、2m 温度、地表温度的分布呈现出一定的正态性，整体较为稳定;潜热通量和感热通量的分布较为分散，在比较大的地方出现尖峰,显示出较大的变异性，这表明能量交换过程受多种因素影响，具有较高的复杂性；相对湿度的分布呈现出明显的偏斜，低湿度值较多，高湿度值较少；ABLH 的分布也呈现出明显的偏斜，低 ABLH 值较多，高 ABLH 值较少，毕竟符合地球的实际情况。

而从 LOWESS 平滑曲线可以大致看出，V 向分量风与 U 向分量风、表面气压与 2m 气温、表面气压与地表温度存在比较明显的负线性相关性，地表温度与 2m 气温存在比较明显的正线性相关性。

当然我们也可以使用 Pearson 相关系数矩阵来进行更加具体详细数字化的分析。Pearson 相关性分析是一种统计方法，用于衡量两个变量之间的线性关系强度和方向。我们通过建立 Pearson 相关性分析模型可以更加准确地分析与判断两个变量之间的关系。两个变量之间的皮尔逊相关系数定义为两个变量之间的协方差和标准差的商：

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - E[X])(Y - E[Y])]}{\sigma_X \sigma_Y}$$
$$= \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - [E(X)]^2}\sqrt{E(Y^2) - [E(Y)]^2}}$$

上式定义了总体相关系数，常用希腊小写字母ρ作为代表符号。估算样本的协方差和标准差，可得到皮尔逊相关系数。用矩阵的形式表现出来，得到图 3：

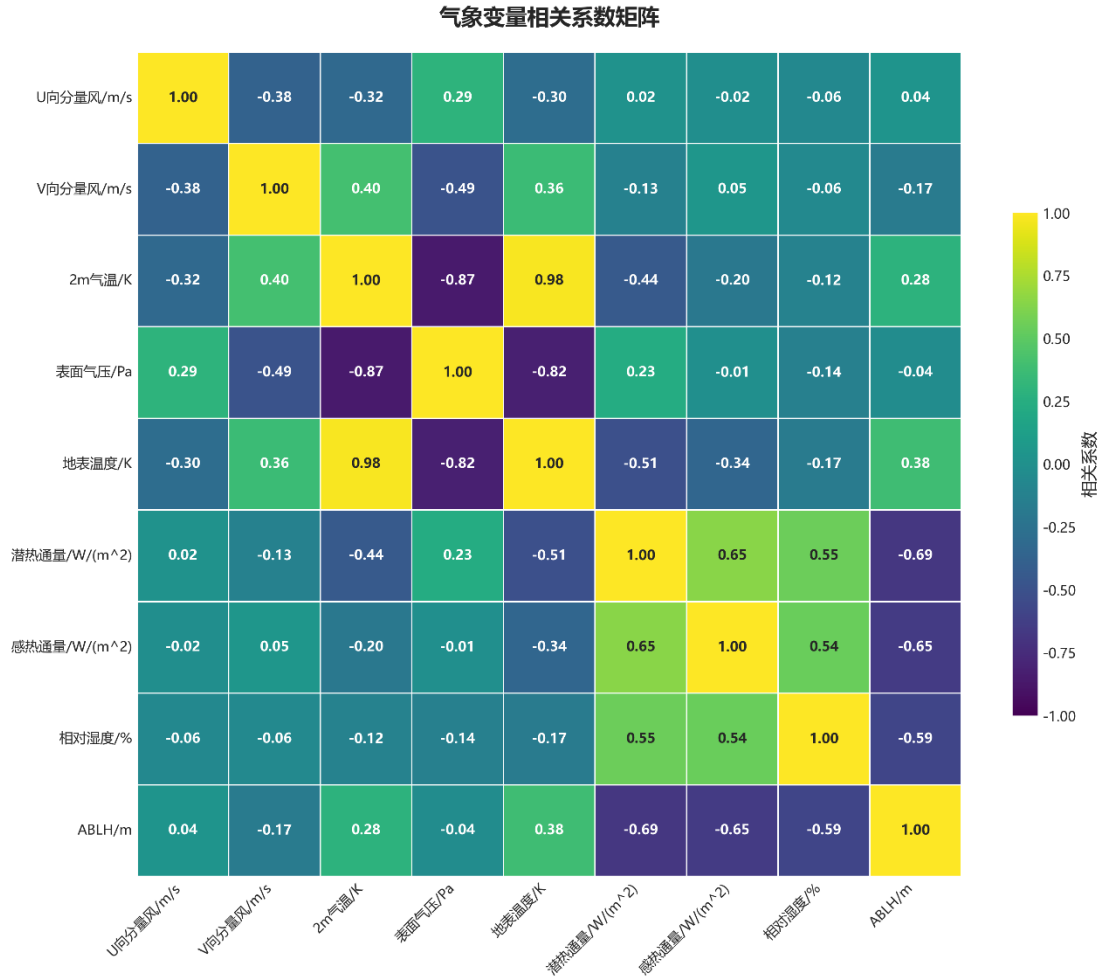

图 3 气象变量相关系数矩阵

以矩阵的形式表示可以通过颜色的不同比较清晰的看出相关系数的大小差异，由图 3 可知，表面气压与 2m 气温、地表温度具有较强的负相关性，ABLH 与潜热通量、感热通量、相对湿度具有中等程度的负相关性，地表温度与 2m 气温具有极强的正相关性，感热通量、相对湿度、潜热通量三者两两之间具有中等程度的正相关性，其他因素之间并不具备明显的相关性。

于是我们在建立三阶多项式扩展的多元线性回归模型时筛选删掉了 2m 气温、地表温度这两个指标。

4.1.3 建立 ABLH 与因素的多元回归分析模型

（1）共线性分析

共线性是指在多元回归分析中，自变量之间存在较高的相关性。这种情况可能会导致模型参数估计的不准确和不稳定，因为难以区分每个自变量对因变量的影响。在进行多元回归分析前我们需要进行共线性分析来判断自变量间的相关性，如果共线性的话我们便不可以轻易使用简单的多元线性回归模型，而应该使用更为复杂、更为合适的多元回归分析模型。

VIF 方差膨胀因子是共线性分析的重要指标。VIF 值越大，表示该自变量与其他自变量之间存在较强的线性依赖关系，即存在多重共线性。它的计算公式为：

$$VIF_i = \frac{1}{1 - R_i^2}$$

其中，

R_i^2是将第i个自变量作为因变量，对其他所有自变量进行回归分析得到的决定系数。

如果 VIF≥10,则表明自变量之间存在多重共线性。我们使用 python 编程计算出各影响因素的 VIF 如下表所示：

表 1 各因素共线性分析 VIF 表

因素	VIF
U 向分量风/m/s	1.378098
V 向分量风/m/s	1.417152
2m 气温/K	68983.372484
表面气压/Pa	855.707629
地表温度/K	68191.634399
潜热通量/W/(m^2)	3.797293
感热通量/W/(m^2)	5.183893
相对湿度/%	38.581440

从表 1 看出，2m 气温和地表温度的 VIF 远远大于 10，表明可能存在非常严重的共线性问题，对此我们虽然可以直接删除这些自变量，但是使用岭回归等其他方式缓解共线性带来的问题。

于是我们在建立三阶多项式扩展的多元线性回归模型时筛选删掉了 2m 气温、地表温度这两个指标。

（2）4 种多元线性回归分析模型的建立与求解

由 VIF 数据可以看出我们不能仅使用最简单的多元线性分析回归模型，否则会产生系数估计不稳定和过拟合风险的问题，所以我们打算采用正则化的方法予以解决。

在进行多元线性回归（OLS）时，对于模型的优劣可以通过损失函数来进行评估，它用来衡量模型预测值与真实值之间的差异。通过最小化损失函数，我们可以训练出一个尽可能准确的模型。

OLS 的损失函数如下：

$$\min_{\beta}\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

这也可以叫作均方误差（MSE）。

对 OLS 引入正则化也可以分为 3 种方法，分别为岭回归（Ridge）、Lasso 回归、Elastic Net 回归，这样可以降低模型复杂度，防止过拟合。

岭回归（Ridge）是在 OLS 的损失函数添加回归系数的平方和（L2 范数）作为惩罚项，它的损失函数如下：

$$\min_{\beta}\left[\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right]$$

岭回归可以改善多重共线性问题，提升模型稳定性，是以放弃无偏性、降低精度为代价解决病态矩阵问题的回归方法。

而 Lasso 回归在损失函数中添加回归系数的绝对值之和（L1 范数）作为惩罚项，它的损失函数如下：

$$\min_{\beta}\left[\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right]$$

Lasso 回归具有特征选择的能力，可以帮助选择最优的正则化参数λ，适用于高维数据。

如果同时引入 L1 范数和 L2 范数作为惩罚项，我们便可以得到 Elastic Net 回归，它的损失函数如下:

$$\min_{\beta}\left[\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\left((1-\alpha)\frac{1}{2}\sum_{j=1}^{p}\beta_j^2+\alpha\sum_{j=1}^{p}|\beta_j|\right)\right]$$

Elastic Net 回归平衡了岭回归和 Lasso 回归的特点。

我们使用 python 进行编程。经过计算机大量运算，得到 4 种线性多元回归分析模型。我们接下来要从这 4 个模型中选出拟合情况最好的模型作为最终结果来衡量 ABLH 与温度、湿度、风速、气压等因素之间的关系，我们可以使用R^2、MSE 和相对误差率来进行评估。

R^2分数，也被称为决定系数，是衡量回归模型拟合优度的一个统计量。它表示的是模型预测值与实际值之间相关程度的平方，具体来说，R^2分数反映了模型能够解释的变异量占总变异量的比例。它的计算公式如下：

$$R^2=1-\frac{SS_{res}}{SS_{tot}}$$

其中，SS_{res}是残差平方和，SS_{tot}是总平方和。

R^2值的范围从 0 到 1。一个 R^2值越接近于 1，表示模型对数据的拟合越好；越接近于 0，则表示模型的预测能力较差。

MSE 即均方误差，是衡量预测值与真实值之间差异的一个常用指标。它通过计算所有数据点上预测值与实际值之差的平方的平均值来量化模型预测的准确性。MSE 越低，表示模型的预测越准确。它的公式如下：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

其中，y_i是第i个样本的真实值，$\hat{y}_i$是第i个样本的预测值。

相对误差率（Relative Error Rate）是一种用于衡量预测值与真实值之间偏差程度的指标，尤其适用于不同量纲或数量级的数据比较。它表示的是误差相对于真实值的比例，通常以百分比形式表示，显然，相对误差率越小越好。它的公式如下：

$$Relative\ Error=\frac{|y-\hat{y}|}{|y|}$$

其中，y是真实值，$\hat{y}$是预测值。

我们使用 python 计算 4 种模型的R^2、MSE 和相对误差率，结果如表 3 所示：

表 2　4 种模型的R^2、MSE 和相对误差率表

Model	R^2	MSE	Relative Error(%)
LinearRegression	0.8379	26112.0899	32.31
Ridge	0.8357	26454.6096	32.52
Lasso	0.8030	31722.7791	35.61
ElasticNet	0.7130	46218.2698	42.98

由此我们发现 LinearRegression 模型R^2越大，MSE 和相对误差率越小，说明 LinearRegression 模型的拟合效果最好，最终我们选择 LinearRegression 模型对 ABLH 与温度、湿度、风速、气压等因素进行量化。生成的数学表达形式如下：

$$y = b + \begin{bmatrix} U & V & T & H & LH & SH \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_6 \end{bmatrix} + \mathbf{x}^\top \cdot \mathbf{C}_2 \cdot \mathbf{x} + \mathbf{x}^\top \cdot \mathbf{C}_3[\cdot] \cdot \mathbf{x} \cdot \mathbf{x}$$

$C1$：一次项向量
$C2$：二次项对称矩阵
$C33$：三次张量或稀疏列表形式

4.1.4 小结

在问题一中，我们探讨了大气边界层高度（ABLH）与温度、湿度、风速及气压等因素之间的量化关系，并评估了这些因素对 ABLH 的影响。首先，通过需求分析强调了 ABLH 在空气污染扩散、天气预报和气候变化研究中的重要性，以及建立数学模型进行量化探索的必要性。接着，采用了六边形箱密度图和变量关系矩阵法（SPLOM 加上 LOWESS 平滑曲线），分别从直观视觉化和统计分析的角度分析了各因素与 ABLH 之间的相关性。

进一步地，为了精确量化这种关系并构建预测模型，进行了共线性分析以识别自变量间的相关性，并针对多重共线性问题提出了使用岭回归等方法来解决。基于此，建立了包括普通最小二乘法回归（OLS）、岭回归（Ridge）、Lasso 回归以及 Elastic Net 回归在内的四种多元回归分析模型。通过对这四个模型的性能评估（使用R^2、MSE 和相对误差率作为评价指标），得出 LinearRegression 模型在拟合效果上表现最佳，因此被选定为最终模型用于量化 ABLH 与上述气象因素之间的关系。

4.2 问题二:结合当地空气质量数据，选择合理的数学模型分析 ABLH 与空气质量相关指标($PM_{2.5}$、PM_{10}、NO_2、AQI 等)的量化关系

4.2.1 研究 ABLH 与空气质量相关指标量化关系模型的需求分析

目前，由于改革开放初期粗放式经济发展模式对于环境的破坏，我国空气污染程度指标如 PM_{10}、NO_2等其他有害气体浓度较高。根据大气科学的理论，大气边界层高度 ABLH 会对空气中的悬浮物与气体分子造成比较大的影响，如果我们可以使用数学模型来量化 ABLH 与空气质量相关指标的关系，一方面我们可以还更好地对未来空气质量进行预测，及时的发布空气不良预警，保护中国人民的生命健康安全；另一方面我们也可以据此制定更加合理的排放计划，最小程度地改变大气质量，更好地促进可持续发展。

4.2.2 GAM 模型的建立

广义加性模型(GAM）是一种灵活且可解释的统计模型，它将广义线性模型（GLM）的非线性扩展与加性模型结合，能够捕捉特征与响应变量之间的复杂非线性关系，同时保持模型的可解释性。

在传统的线性回归模型中，我们假设响应变量y与预测变量$x_1, x_2, \ldots, x_p$之间存在线性关系：

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon$$

其中，ϵ为误差项，通常假设为正态分布。

然而，在很多实际情况下，这种线性关系并不总是成立的。GAM 通过引入平滑函数，使得每个预测变量可以以一种非线性的方式影响响应变量。GAM 的一般形式为：

$$g\big(E(y)\big) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

其中：

$g(\cdot)$为链接函数，它将响应变量的期望值$\mu = E(Y)$映射到实数域；

$E(y)$为响应变量的期望；

$f_j\big(x_j\big)$为特征x_j的平滑函数（如样条、核函数、多项式），用于捕捉该变量与响应变量之间的潜在非线性关系；

β_0为截距项。

对于给定的数据集$\big(y_i, x_{i1}, x_{i2}, \ldots, x_{ip}\big)$其中$i = 1,2, \ldots, n$,GAM 的目标是找到一组合适的平滑函数$f_j(\cdot)$,使得下式成立：

$$g\big(E(y_i)\big) = \beta_0 + \sum_{j=1}^{p} f_j\big(x_{ij}\big)$$

GAM 模型通过平滑函数自动适应数据中的非线性关系，提高了模型对复杂数据结构的拟合能力；同时保持了单个特征效应的直观解释，使得分析结果易于理解和传达给非技术受众。此外，GAM 可以灵活地处理各种类型的数据（如连续型、分类型响应变量），并通过正则化方法有效避免过拟合，适用于广泛的回归和分类问题。我们使用 python 进行 GAM 模型的建立，经过预处理、训练可以得到各因素与 ABLH 关系的 GAM 拟合曲线：

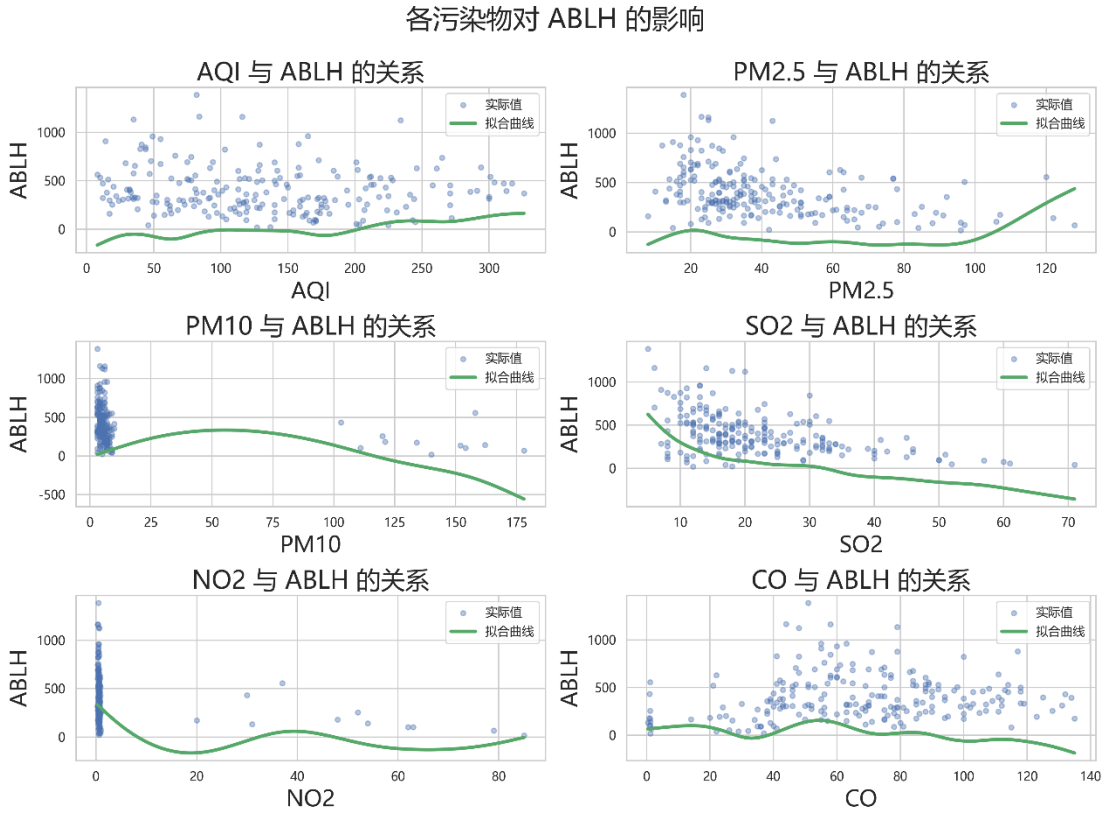

图 4 各因素与 ABLH 关系的 GAM 拟合曲线

我们当然也可以像分析随机森林与 XGBoost 一样使用R^2、RMSE 对 GAM

模型进行性能分析，得到表 7：

表 7　GAM 模型的R^2、RMSE 表

Model	R^2	RMSE
GAM	0.414	189.194

相较于基于决策树的随机森林和 XGBoost 模型

在 GAM 模型或者其他回归模型中，分析残差分布是评估模型拟合优度和检查模型假设是否满足的一个重要步骤。通过残差分析，我们可以发现数据中的模式、异常值以及模型可能存在的不足之处。

残差是指观测值与模型预测值之间的差异，对于第i个样本，残差表示为：

$$e_i = y_i - \hat{y}_i$$

其中，y_i为实际观测值，$\hat{y}_i$为预测值。

而将各个残差绘制在图上，可以比较清晰地看出残差的分布，如果残差的分布并不均匀，而是集中地分布在某一边，说明 GAM 模型拟合效果不好，我们通过 python 计算出每一组数据残差，然后绘制残差分布图：

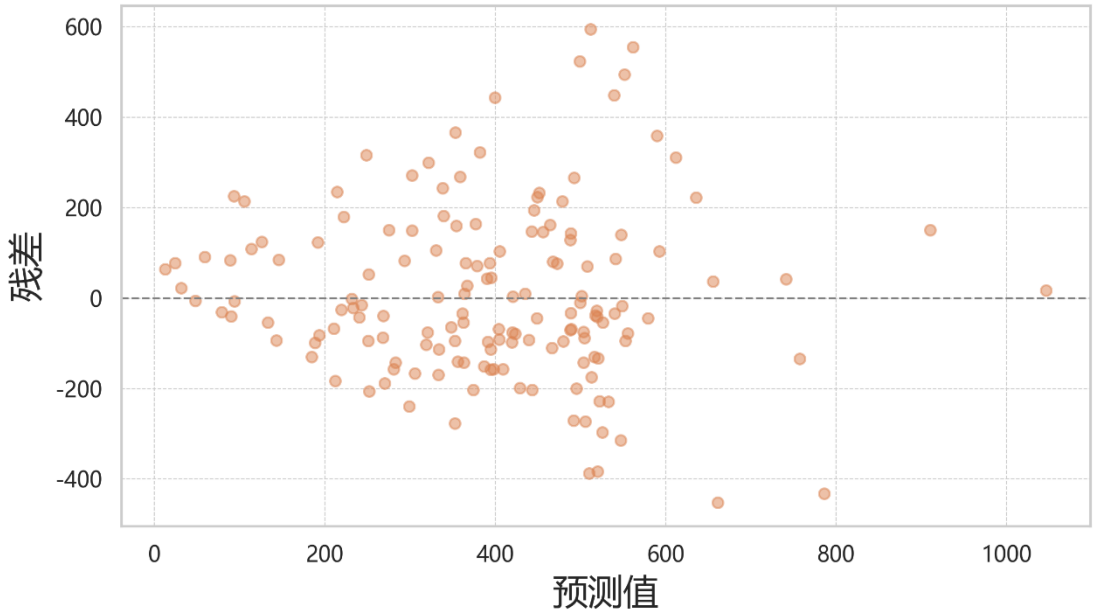

图 5 GAM 模型的残差分布图

残差分布图看出 GAM 模型产生的残差在 0 两边均匀分布，说明 GAM 模型的拟合效果较好，我们可以使用 GAM 模型进行 ABLH 与空气质量相关指标量化关系模型建立。

4.2.3 基于决策树的集成学习方法——随机森林与 XGBoost 回归模型的建立

在进行随机森林与 XGBoost 回归模型推演前，我们需要先了解一下决策树的概念。决策树算法是一种非参数的监督学习方法，用于分类和回归任务。它通过递归地选择最优特征并在该特征上进行数据集的最佳划分，逐步构建一个树形结构。每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，而每个叶节点代表一种类别或一个数值预测。在训练过程中，算法依据某种标准来选择最佳分裂点，以最大化不同类别的区分度或最小化预测误差。最终生成的模型易于理解和解释，能够清晰地展示决策规则。但是单一决策树容易出现过拟合问题，所以我们采用随机森林和 XGBoost 回归模型随机森林和 XGBoost 都是以决策树为基本构成模块的集成学习方法。

随机森林是一种基于集成学习的机器学习算法，通过构建多棵决策树并结合其预测结果以提升模型的泛化能力和鲁棒性。该算法采用 Bagging 框架，通过有放回 Bootstrap 抽样生成多样化的训练子集，并基于特征随机选择策略（如随机子空间法）为每棵树分配不同的特征子集进行节点分裂，从而有效降低模型方差并抑制过拟合。在回归任务中采用均值聚合输出连续值。其核心优势包括对高维数据的适应性、对噪声和异常值的强鲁棒性，以及天然支持并行化训练。此外，随机森林能够量化特征重要性，为模型可解释性提供依据，是处理复杂非线性关系的经典算法之一。图 4 生动形象地展示随机森林的原理。

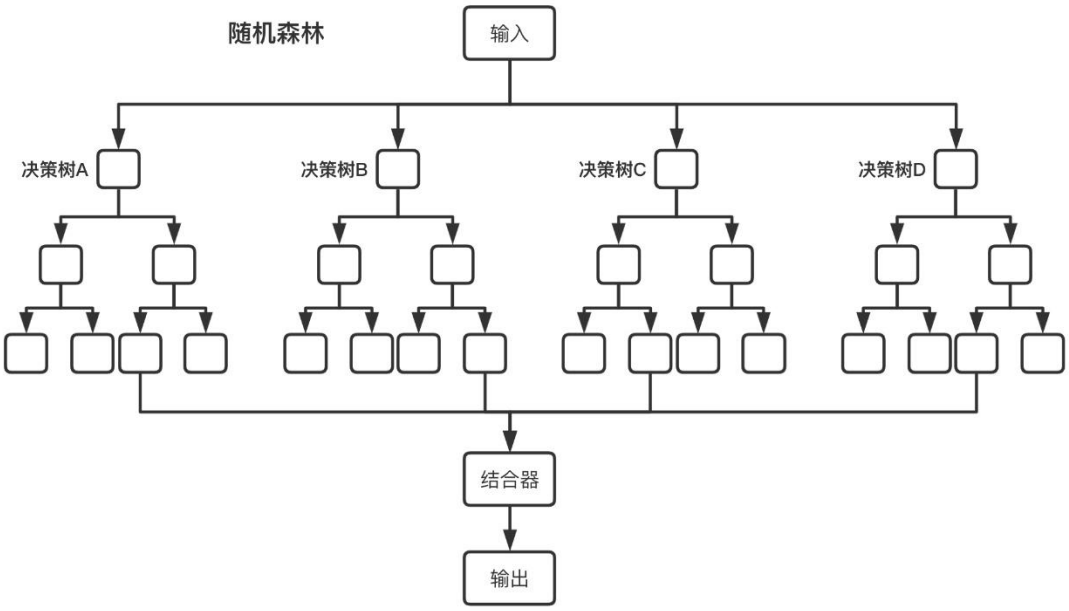

图 4 随机森林原理图

XGBoost 是一种高效且可扩展的梯度提升框架，通过集成多棵决策树以最小化损失函数，其核心创新在于引入二阶泰勒展开优化目标函数，并结合正则化项（L1/L2）控制模型复杂度，从而显著提升预测精度并抑制过拟合。它的目标函数由损失函数 L 和正则化项Ω组成：

$$Obj(\Theta) = \sum_{i=1}^{n} L(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k)$$

其中

$$\Omega(\mathrm{f}) = \lambda T + \frac{1}{2}\alpha||w||^2$$

T 为树的叶子节点数，w 为叶子结点的权重向量，λ、α为正则化系数。

该算法支持并行化计算和分布式训练，通过贪心算法生成树结构、加权残差逼近及特征分位点优化实现高效特征分裂，同时内置缺失值处理机制和特征重要性评估功能。XGBoost 在训练过程中通过列采样和行采样增强模型多样性，并支持自定义损失函数与评估指标，广泛应用于分类、回归等任务。

我们这一次使用了三种基于决策树的回归模型，分别为 RF_bootstrap 模型，RF 模型，XGB 模型。RF_bootstrap 和 RF 模型都是随机森林模型，其区别在于 RF_bootstrap 使用 Bootstrap 采样，RF 不使用 Bootstrap 采样，启用 Bootstrap 采样在每次构建树时对训练数据进行有放回抽样，增加模型多样性，降低方差；而使用原始数据训练每棵树，可能提高偏差但减少随机性。而 XGB

则是使用了 XGBoost 模型。使用 python 进行机器学习，对三个问题进行随机探索，分别得到各自模型的最佳参数：

表 4 RF_bootstrap 与 RF 模型最佳参数

参数	功能	RF_bootstrap	RF
n_estimators	控制森林中树的数量	50	1000
min_samples_split	节点分裂所需最小样本数	12	3
min_samples_leaf	叶节点最小样本数	1	8
max_sample	每棵树的样本采样比例	0.5	
max_features	分裂时的特征采样比例	0.7	Sqrt
max_depth	树的最大深度	15	5
bootstrap	启用 Bootstrap 采样	True	False

表 5 XGBoost 模型最佳参数

参数	功能	值
n_estimators	控制森林中树的数量	200
subsample	样本采样比例	0.4
reg_lambda	L2 正则化系数	0
reg_alpha	L1 正则化系数	10
max_depth	树的最大深度	12
learning_rate	学习率	0.01
gamma	分裂最小增益	0.1
colsample_bytree	特征采样比例	0.9

我们可以使用训练时间、R^2、RMSE 对 3 种模型进行性能分析。

训练时间体现了模型算法的复杂程度，消耗时间过长的模型不建议的，因为这个模型的时间复杂度太大需要的硬件资源会过高，而采用更加简单的模型可以方便高效进行拟合。

R^2在问题一已经使用过了，不必赘述。RMSE 为均方根误差，是衡量预测值与真实值之间差异的一个常用指标。它通过计算所有数据点上预测值与实际值之差的平方的平均值，然后取平方根来量化模型预测的准确性。RMSE 越低，表示模型的预测越准确。公式为：

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

其中，n为观测样本的数量，y_i是第i个样本的真实值，$\hat{y}_i$是第i个样本的预测值。

我们使用 python 计算 3 种模型的训练时间、R^2、RMSE，结果如表 6 所示：

表 6　3 种模型的训练时间、R^2、RMSE 表

Model	Time	R^2	RMSE
RF_bootstrap	0.053702	0.155982	16.95913
RF	0.717938	0.097133	17.54041
XGB	0.225946	0.139731	17.12162

由表 6 不难看出，在这三个模型中，RF_bootstrap 随机森林模型凭借其训练时间最短，R^2决定系数最大，RMSE 值最小成为拟合效果最好的模型。

表 7　4 种模型的对比表

Model	Time	R^2	RMSE
RF_bootstrap	0.053702	0.155982	16.95913
RF	0.717938	0.097133	17.54041
XGB	0.225946	0.139731	17.12162
GAM		0.414	189.194

在模型选择与性能对比中，广义加性模型（GAM）与决策树集成模型（随机森林、XGBoost）展现出显著的差异性。GAM 通过引入平滑函数（如样条或核函数）灵活刻画 ABLH 与 $PM_{2.5}$、NO_2等指标间的非线性关系，其加性结构可直观分解各变量的独立效应（如图 11 所示），为机理解释提供了清晰的量化框架，例如通过边际效应曲线识别 ABLH 阈值对污染物浓度的动态影响。相比之下，随机森林与 XGBoost 基于多棵决策树的集成策略，通过特征随机采样与模型聚合提升泛化能力，虽能高效处理高维数据与复杂交互作用（如表 6 中 RF_bootstrap 仅需 0.054 秒完成训练）。从性能指标看，GAM 的 R2R2（0.414）显著高于集成模型（RF_bootstrap: 0.156，XGB: 0.140），表明其对变量关系的解释能力更强；而集成模型在预测效率与工程适用性上更具优势，尤其是面对实时空气质量预警或海量数据场景时。残差分析进一步验证了 GAM 的稳健性（图 12 残差均匀分布），而集成模型可能因局部非线性效应残留未捕捉的系统偏差。综上，若研究目标聚焦于机理探索与政策制定（如减排阈值量化），GAM 凭借其透明性与科学解释力成为首选；若需快速响应或处理高维、缺失数据，则可优先采用随机森林或 XGBoost 提升预测效率。两类模型亦可协同应用——以 GAM 构建解释性框架，辅以集成模型验证预测一致性，从而实现空气质量分析的精准性与可操作性双重优化。

4.2.4 小结

在问题二中，我们旨在探讨大气边界层高度（ABLH）与空气质量相关指标之间的量化关系。首先，我们通过需求分析强调了研究这一关系的重要性。为了实现上述目标，我们尝试使用基于决策树的集成学习方法——随机森林和 XGBoost 回归模型，以及广义加性模型（GAM）来探索 ABLH 与空气质量指标之间的复杂非线性关系。尽管随机森林和 XGBoost 模型能够提供强大的预测能力，并且在处理高维数据方面具有显著优势，但考虑到这些模型的“黑箱”特性可能影响结果的解释性和透明度，我们还采用了 GAM 模型作为补充。GAM 模型不仅能够捕捉变量间的非线性关系，同时保持了较高的模型可解释性，这使得它成为一种理想的工具，尤其适合于需要理解各个因素如何单独影响响应变量的应用场景。

具体而言，在 RF_bootstrap 模型在 3 个决策树模型中凭借其较短的训练时间、较高的决定系数（R^2）以及较低的均方根误差（RMSE），显示出最佳的拟合效果。进行补充的 GAM 模型虽然其 RMSE 表现不如 RF_bootstrap 模型，但 R^2 值和残差分布图来看，GAM 模型具备良好的拟合效果。

综上所述，针对 ABLH 与空气质量相关指标的量化关系研究，我们推荐结合使用 RF_bootstrap 模型和 GAM 模型。RF_bootstrap 模型适用于需要高精度预测的情景，而 GAM 模型则更适合用于需要深入了解各因素影响机制的研究工

作。通过这种方式，不仅可以提高对未来空气质量状况的预测准确性，还能为制定有效的环境保护策略提供科学依据。

4.3 问题三：建立一个数学模型，根据天气预报数据来预测次日 ABLH，并说明该模型的合理性

4.3.1 预测次日 ABLH 的需求分析

前面的部分已经提到，大气边界层高度（ABLH）是评估大气情况的重要综合指标。ABLH 决定着污染物在垂直方向上的扩散能力，这对防止雾霾形成尤其重要；ABLH 的变化直接反映了大气的稳定程度，这对于研究天气现象有着重要意义。准确预测次日 ABLH 有助于更精准地进行空气质量预报，特别是对于 $PM_{2.5}$、PM_{10}等细颗粒物的浓度预测，能够帮助提前采取措施减少空气污染对公众健康的影响，还可以为短期天气预报提供支持，建立预测次日 ABLH 的模型是迫在眉睫的。

4.3.2 使用 Stacking 融合模型建立预测模型

在前面我们使用了多元回归分析模型 LinearRegression、Ridge、Lasso、ElasticNet 和基于决策树的回归模型 RF_bootstrap 模型，RF 模型，XGB 模型，它们分别具有各自的特点与优劣。如果我们将它们按照某一种方式组合起来，我们便可以在最大化所有模型的优点与最小化所有模型的缺点之间寻找到一个最佳的平衡点，这便是 Stacking 融合模型。

Stacking 融合模型是一种高阶集成学习方法，通过组合多个基学习器（的预测结果，再通过元学习器进行二次学习，从而提升整体模型的泛化能力。其核心思想是让元模型学习如何最优地组合基模型的预测结果。它的整体流程如下图所示：

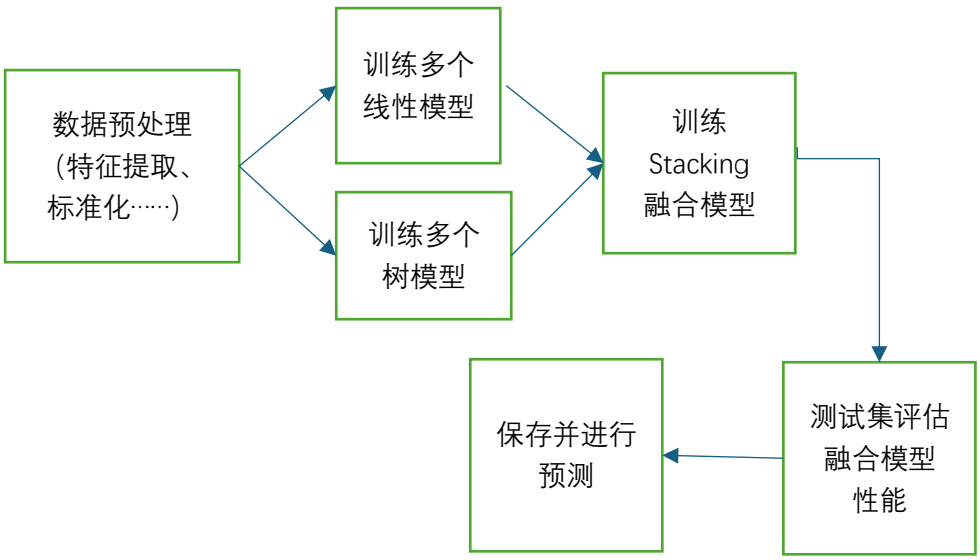

图 6 建立 Stacking 融合模型整体流程

我们使用 python 进行 Stacking 融合模型的建立。首先我们会进行数据的预处理，包括数据读取、特征提取、训练集/测试集划分、标准化这些操作。

接着我们训练多个线性模型,分别为 LinearRegression、Ridge、Lasso、ElasticNet。经过机器学习训练，我们可以得到 4 种多元回归分析模型。我们接下来继续使用R^2、MSE 和相对误差率来对 4 种多元线性回归分析模型进行评估。

我们使用 python 计算 4 种模型的R^2、MSE 和相对误差率，结果如表 9 所示：

表 8　4 种模型的R^2、MSE 和相对误差率表

Model	R^2	MSE	Relative Error(%)
LinearRegression	0.6448	57564.0112	38.85
Ridge	0.6462	57336.1083	38.77
Lasso	0.7296	43825.2614	33.90
ElasticNet	0.6350	59163.2656	39.33

4 种多元线性回归模型的R^2、MSE 和相对误差率均在可以接受的范围，于是我们对这 4 类线性模型均给予保留。

然后，我们训练多个树模型，大体上可以分为随机森林和 XGBoost，具体来说是 RF_bootstrap 模型、RF 模型、XGB 模型。使用 python 进行机器学习，对三个问题进行随机探索，分别得到各自模型的最佳参数：

表 9 RF_bootstrap 与 RF 模型最佳参数

参数	功能	RF_bootstrap	RF
n_estimators	控制森林中树的数量	800	500
min_samples_split	节点分裂所需最小样本数	2	5
min_samples_leaf	叶节点最小样本数	1	1
max_sample	每棵树的样本采样比例	0.5	
max_features	分裂时的特征采样比例	0.5	log2
max_depth	树的最大深度	None	None
bootstrap	启用 Bootstrap 采样	True	False

表 10 XGBoost 模型最佳参数

参数	功能	值
n_estimators	控制森林中树的数量	300
subsample	样本采样比例	0.2
reg_lambda	L2 正则化系数	10
reg_alpha	L1 正则化系数	1
max_depth	树的最大深度	4
learning_rate	学习率	0.05
gamma	分裂最小增益	0.2
colsample_bytree	特征采样比例	1.0

我们继续使用训练时间、R^2、RMSE 对 3 种模型进行性能分析，使用 python 计算 3 种模型的训练时间、R^2、RMSE，结果如表 12 所示：

表 11　3 种模型的训练时间、R^2、RMSE 表

Model	Time	R^2	RMSE
RF_bootstrap	9.54	0.780	188.067
RF	10.64	0.776	190.030
XGB	0.12	0.805	177.288

由此我们可以得出 XGB 的训练时间比另外 2 个树模型明显短很多，而R^2又是 3 种模型最高的，RMSE 是 3 种模型最小的，说明 XGB 模型是 3 种树模型中

表现最好的，但另外 2 个模型的综合性能表现也不错，也都给予保留。

最后，我们将 4 个线性模型与 3 个树模型融合起来，进行 Stacking 融合模型的训练与建立。在 python 中创建堆叠回归器 StackingRegressor：一级学习器列表 + 最终的线性回归元学习器 + 时间序列交叉验证，经过训练以后将生成的预测模型保存，以供随时提取使用进行预测。

我们可以使用R^2、MSE 对 Stacking 融合模型进行评估。我们使用 python 计算 Stacking 融合预测模型的R^2、MSE，结果如表 13 所示：

表 12　Stacking 融合预测模型的R^2、MSE 表

Model	R^2	MSE
Stacking Model	0.793446	33266.664335

从R^2、MSE 的值可以看出 Stacking 融合预测模型的综合性能比较良好。在使用 Stacking 融合技术建立起 ABLH 预测模型后，只需要给出前一天的天气预报详细数据，我们便可以比较轻松、准确地推测出次日的 ABLH，以做出相应的措施。

4.3.3 Stacking 融合模型的合理性

Stacking 融合模型的合理性主要体现在以下几个方面：

1. 理论基础与集成优势

Stacking 融合模型的核心理念在于集成不同模型的互补优势，通过多层次学习框架提升预测性能。就这个模型而言，线性模型（如 Ridge、Lasso）擅长从全局视角提取特征间的线性关系，可解释性强。而树模型（如 XGBoost、随机森林）则通过递归划分特征空间，能够捕捉复杂的非线性交互作用以及对异常值的鲁棒性。Stacking 融合模型对不同模型输出的加权组合，既保留了原模型的多样性，又通过正则化约束降低了单一模型的过拟合风险，从而在全局范围内实现泛化能力的提升。

2. 训练效果优异

实验结果表明，Stacking 模型在 ABLH 预测任务中展现出显著优势。从性能对比来看，Stacking 的 R^2 达到 0.793，MSE 为 33266.66，优于多数初始模型。尽管 XGBoost 单一模型的 R^2 略高（0.805），但 Stacking 通过融合多模型信息，在保持高精度的同时降低了预测结果的方差。此外，模型采用时间序列交叉验证（TimeSeriesSplit），严格按时间顺序划分训练集与验证集，避免未来信息泄露至历史数据中。这一策略有效契合气象数据的时序依赖性特点（如季节循环、昼夜变化），确保模型在真实场景中的可靠性。

3. 实际应用适配性

Stacking 模型在实际气象预测场景中表现出极强的适应性。首先，通过多源特征融合，模型能够整合温度、湿度、风速等实时气象因子及其滞后特征，从而捕捉 ABLH 的动态演变规律；其次，模型架构具有高度可扩展性，初始模型可灵活替换或增删，从而适应不同数据规模和预测需求。最后，在部署效率方面，尽管训练阶段耗时较长，但预测阶段仅需调用预训练的模型文件（如.pkl 格式），单次推理可在毫秒级完成，完全满足气象业务系统对实时性的要求。

4.3.4 小结

在问题三中，我们基于天气预报数据建立了 Stacking 融合模型以预测次日大气边界层高度（ABLH）。首先，通过需求分析明确了 ABLH 预测对空气质量预警和天气现象研究的重要性。随后，我们分别训练了 4 类线性回归模型和 3 类树模型，并通过性能评估筛选出最优基模型。最终，采用 Stacking 集成技术融合基

模型的预测结果，构建了高性能的 ABLH 预测模型。Stacking 融合模型的综合性能显著优于单一线性模型，与树模型接近甚至略好，但是它集齐所有初始模型的特点与优势，体现出 Stacking 融合模型的优势。最后，我们论证了 Stacking 融合模型的合理性，现在我们使用 Stacking 融合预测模型根据前一天的天气预报来获得相对准确的次日 ABLH 预测值。

五、模型的优缺点评价分析

5.1 优点

线性回归模型原理简单直观假设 ABLH 与各因素之间存在线性关系，直接通过最小二乘法估计参数，易于理解和解释，能清晰展示各因素对 ABLH 的影响方向和大致程度且计算成本低，求解过程相对简单，计算速度快，在处理大规模数据时效率较高，可快速得到模型结果和预测值。

岭回归模型和套索回归模型能够防止过拟合，通过在损失函数中加入正则化项，能够有效惩罚模型的复杂度，避免模型过度拟合训练数据，提高模型的泛化能力，使模型在面对新数据时具有更好的预测准确性。，套索回归模型还具有自动进行特征选择的能力，能够将一些对 ABLH 影响较小的因素的系数压缩为零，从而筛选出对 ABLH 影响较大的关键因素，有助于简化模型和理解问题。

弹性网络回归模型综合了岭回归和套索回归的优点，既通过 L2 正则化防止过拟合，又利用 L1 正则化进行特征选择，在不同的数据场景下可能表现出更好的稳定性和适应性。

随机森林算法能够很好地处理 ABLH 与各因素之间的非线性关系，通过构建多个决策树并进行集成学习，能够捕捉到数据中复杂的相互作用和规律，对复杂的实际问题具有较强的建模能力而且抗噪声能力强，随机森林通过随机抽样和特征选择构建多个决策树，减少了模型对个别数据点和特征的依赖，对数据中的噪声和异常值具有较好的鲁棒性，不易受到局部干扰而导致模型性能大幅下降。

XGBoost 算法则采用了梯度提升的思想，在迭代过程中不断优化模型，能够快速收敛到较优的解，同时具有很高的预测准确性，在处理大规模数据和复杂问题时表现出色，同时支持大规模分布式计算，能够处理海量数据，并且具有丰富的参数调整选项，可以根据具体问题进行灵活的配置和优化，以适应不同的数据集和建模需求。

5.2 缺点

线性回归模型：实际中 ABLH 与各因素之间的关系往往是非线性的，线性回归模型的假设过于简单，可能无法准确描述复杂的现实情况，导致模型拟合效果不佳，预测精度较低。 容易受到数据中异常值的影响，异常值可能会对回归系数的估计产生较大偏差，进而影响模型的整体性能和稳定性。

岭回归模型和套索回归模型：模型的性能对正则化参数非常敏感，需要通过交叉验证等方法仔细选择合适的超参数，这增加了模型调优的复杂性和计算成本。虽然套索回归具有特征选择功能，但随着正则化参数的变化，系数的变化可能不连续，导致对模型结果的解释相对困难，不如线性回归模型直观。

弹性网络回归模型：需要同时调整 L1 和 L2 正则化的参数，超参数的选择更加复杂，需要更多的经验和计算资源来进行优化，以找到最佳的模型配置。

随机森林算法模型复杂度高，构建了多个决策树，模型结构复杂，存储和计算成本较高。在处理大规模数据集时，训练和预测过程可能会占用较多的内存和时间，尤其是在实时性要求较高的场景下可能不太适用。

XGBoost 算法：对数据的质量和分布有一定要求，如果数据存在严重的不平衡或异常值，可能需要进行额外的数据预处理才能保证模型的性能，否则可能会导致模型过拟合或预测偏差较大。虽然具有丰富的参数可调整，但参数之间的相互作用复杂，调优过程需要一定的经验和技巧，否则很难找到最优的参数组合，以充分发挥模型的性能。

六、总结

本次数学建模围绕大气边界层高度（ABLH），综合多模型方法深入分析其与气象、空气质量因素关系并进行预测，在问题研究中，针对 ABLH 与温度、湿度、风速、气压等因素的量化关系，采用线性回归、岭回归、套索回归及弹性网络回归模型，通过数据预处理、模型训练与评估，依据回归系数及特征重要性评估各因素影响大小；在分析 ABLH 与空气质量指标（$PM_{2.5}$、PM_{10}、NO_2、AQI 等）关系时，运用随机森林和 XGBoost 算法，经数据处理、模型训练与评估，揭示两者量化关联；在预测次日 ABLH 方面，结合前两问的模型训练结果，从回归模型组和机器学习算法组中分别选出表现最优的两个基学习器，采用 Stacking 方式进行融合，构建最终预测模型。 研究过程中设定了数据完整性、因素独立性、关系线性可加性等多项假设，为模型构建提供前提。各模型优缺点显著，回归模型简单直观但对非线性关系处理欠佳；随机森林和 XGBoost 算法能处理复杂关系，但存在模型复杂、调参困难等问题。通过多模型融合，充分发挥不同模型优势，有效提升了模型的预测准确性与泛化能力。 此次数学建模为 ABLH 相关研究提供了可行的方法与思路，后续可进一步优化模型参数、拓展数据维度，为空气污染治理、天气预报及气候变化研究提供更有力的支持。

三、参考文献

[1]赵中行,付松琳,陈钧杰,等.基于深度森林的遥感融合反演大气边界层高度方法[J].光学学报,2025,45(06):261-272.
[2] 姜启源，大学数学实验，[M],清华大学出版社。
[3] 安德里亚斯·穆勒，莎拉·吉多，Python 机器学习基础教程，[M],人民邮电出版社。

四、附录