参赛队号：（参赛队无须填写，参赛队号由大赛官网自动生成）

2024年（第十届）全国大学生统计建模大赛

参 赛 作 品

|  |  |
| --- | --- |
| 参赛学校： | XXXXXX大学 |
| 论文题目： | XX相对贫困治理成效的统计测度研究 |
| 参赛队员： | XXX XXX XXX |
| 指导老师： | XXX XXX XXX |

目录

[摘要 3](#_Toc166049701)

[一、绪论 5](#_Toc166049702)

[（一）研究背景及意义 5](#_Toc166049703)

[（二）研究现状 5](#_Toc166049704)

[二、模型构建思路和创新 5](#_Toc166049705)

[（一）多元有序逻辑回归模型构建思路 5](#_Toc166049706)

[（二）空气质量的分析和预测的模型创新 5](#_Toc166049707)

[三、数据描述及数据预处理 5](#_Toc166049708)

[（一）研究区域概况 5](#_Toc166049709)

[（二）变量描述 5](#_Toc166049710)

[（三）数据来源 5](#_Toc166049711)

[（四）数据预处理 5](#_Toc166049712)

[四、多元有序逻辑回归模型的构建 5](#_Toc166049713)

[（一）多元有序逻辑回归模型的基本原理 5](#_Toc166049714)

[（二）多元有序逻辑回归模型的建模步骤 5](#_Toc166049715)

[**1.相关性分析** 5](#_Toc166049716)

[**2.灰色关联度分析进行特征选择** 5](#_Toc166049717)

[**3.逐步回归剔除共线性和没有相关性的变量** 5](#_Toc166049718)

[**4.利用逻辑回归对数据进行分类** 5](#_Toc166049719)

[**5.使用机器学习的逻辑回归模型** 5](#_Toc166049720)

[（三）多元有序逻辑回归模型的评价 5](#_Toc166049721)

[五、基于随机森林和时间序列相结合的创新模型的构建与评价 5](#_Toc166049722)

[（一）时间序列分解 5](#_Toc166049723)

[（二）参数优化 5](#_Toc166049724)

[（三）模型构建与评价 5](#_Toc166049725)

[六、结论与展望 5](#_Toc166049726)

[（一）结论 6](#_Toc166049727)

[（二）展望 6](#_Toc166049728)

XX相对贫困治理成效的统计测度研究

摘要

本研究以天津市为研究对象，通过基于多元有序逻辑回归和随机森林优化算法的大气质量的时空分析和预测，旨在深入了解大气污染的时空特征，并提供准确的大气质量预测。本研究采集了天津市的大气污染监测数据和相关气象数据，并结合时空特征，建立了多元有序逻辑回归模型和随机森林优化模型。首先，我们利用多元有序逻辑回归模型对天津市的大气质量进行分析。该模型能够识别和分析影响大气质量的关键因素，并将不同污染级别进行分类。通过对大量历史观测数据的分析，我们发现气象因素、工业排放和交通状况等因素对大气质量具有显著影响。多元逻辑回归模型能够通过这些因素，对大气质量水平进行准确的分类和预测。其次，我们构建了基于随机森林的优化模型，能够从多个角度综合考虑各种影响因素，并进行大气质量的时空预测。通过对历史观测数据的训练和验证，我们评估了模型的性能和准确度，并进行了模型的优化和调整。结果显示，优化模型在时空预测方面表现出很好的性能，能够准确预测未来不同地区的大气质量状况。综合研究结果表明，多元有序逻辑回归模型和随机森林优化模型能够有效地分析和预测天津市的大气质量。通过综合应用这些方法，我们能够更全面地了解大气污染的时空变化，并为决策者提供科学支持和管理建议。这对于天津市及其他城市的大气质量管理和预警具有重要意义。同时，本研究还存在一些局限性，例如数据的可用性和质量等方面的限制。未来的研究可以进一步优化模型，引入更多的数据和特征，以提高大气质量预测的准确性和可靠性。此外，可以考虑将其他因素，如人口密度和土地利用等纳入模型，以更全面地分析大气质量的影响。因素总之，本研究通过基于多元有序逻辑回归和随机森林优化算法，对天津市的大气质量进行了时空分析和预测。研究结果为大气质量管理和决策提供了重要的参考，同时也为类似研究提供了方法和思路。

关键词：多元有序逻辑回归；随机森林；大气质量；时空分析；预测

表格与插图清单

表1.XXXXXXXXXXXX

表2.XXXXXXXXXXXX

表3.XXXXXXXXXXXX

……

图1.XXXXXXXXXXXX

图2.XXXXXXXXXXXX

图3.XXXXXXXXXXXX

……

XX相对贫困治理成效的统计测度研究

一、绪论

（一）研究背景及意义

天津市作为中国北方的重要城市，面临着严重的大气污染问题。随着城市化进程和工业化的快速发展，大量的工业排放、机动车尾气、燃煤和扬尘等因素导致了大气污染的严重积累。根据《天津市生态环境质量报告（2020）》的数据显示，天津市PM2.5年平均浓度超过国家空气质量标准限值，空气质量状况较差。这对居民的健康和环境的可持续发展带来了严重威胁。

此外，天津市大气污染研究的意义还体现在可持续发展方面。大气污染不仅威胁人类健康，还对生态环境和社会经济发展产生负面影响。空气质量的下降会影响旅游业、经济发展和外来投资等方面，阻碍城市的可持续发展。通过深入研究天津市大气污染的形成机制和影响因素，可以为制定绿色发展战略和环境保护政策提供科学依据。

最后，天津市大气污染的研究对于全国范围内的大气污染治理具有示范和引领作用。天津作为人口众多、经济发达的城市，其大气污染治理经验和技术应用可供其他城市借鉴和推广。通过研究天津市大气污染问题，可以为其他地区提供重要的参考和指导，促进全国大气污染治理工作的进展。

（二）研究现状

近年来，生态环境部门以及众多学者基于现阶段公开数据对大气质量预测模型进行了大量的研究，但大多模型都面临将时空规律分开进行研究、影响因素考虑片面化、预测精度较低等问题。经查阅相关研究资料后，我们发现如今的大气质量预测研究模型主要分为机理性大气质量预测模型和非机理性大气质量预测模型。前者能够考虑更多的物理和化学机制，但需要较多的输入数据和计算资源。后者通过数据分析和统计建模，能够快速预测大气质量，但对于复杂的污染物传输和化学反应过程的理解相对较少。后者适用范围广，并且建模和预测过程相对简单，不需要大量的专业知识和复杂的参数设置。这使得非专业人员也能够使用这些模型进行大气质量预测，从而在实践中提供快速、有效的决策支持。常用的非机理性大气质量预测模型和方法有时间序列预测模型、灰色系统预测模型、最优化权值组合法等，这些模型和方法已经在大气质量预测领域得到广泛应用。

1. 时间序列模型

时间序列方法在大气质量预测中的原理是基于历史观测数据的时间模式和趋势来进行预测。它假设过去的数据模式会在未来一段时间内保持一定的稳定性和重复性，从而可以利用这些模式和趋势来推断未来的大气质量情况。

二、数据描述及数据预处理

1. 研究区域概况

天津市位于中国的北方沿海地区，东临渤海，北濒黄海。它紧邻北京市，距离北京市约120公里。天津市地处华北平原，地势相对平坦。市区内有海河流经，形成了独特的城市景观。

根据2020年的统计数据，天津市的人口约为15.6万人。天津市是中国人口密度较高的城市之一，人口集中在市区和周边地区。由于地理位置的优势和经济发展的吸引力，天津市吸引了大量的人口流入。

天津市是中国重要的工业中心之一，拥有发达的制造业和重工业。天津市的化工产业较为发达，涵盖了石油化工、化学制品生产等领域。化工生产会排放大量的有害气体和颗粒物，如挥发性有机物、硫化物、氮氧化物等，对大气环境造成污染和风险。天津市的燃煤电厂是主要的能源供应来源之一。燃煤电厂排放二氧化硫、氮氧化物、颗粒物等大气污染物，对空气质量产生显著影响，尤其是在燃煤过程中未经有效净化处理的情况下。汽车制造业和交通运输业是天津市的重要工业部门。汽车尾气排放是大气污染的重要来源，包括颗粒物、氮氧化物和挥发性有机物。交通拥堵也会导致排放物积累，进一步影响空气质量。

天津市政府在工业发展和环境保护方面采取了一系列措施来减少工业项目对大气环境的影响。这包括加强排放控制和净化设施的建设、推广清洁生产技术、限制高污染和高能耗产业的发展等。

1. 变量描述
2. 数据来源

我们收集数据文件夹中的数据为从由天津市统计局提供的每一年的统计年鉴中的天津市各月份气象资料以及天津市气象局提供的各区气象资料直接获得。

1. 数据预处理

**1.数据的获取与整理**

为了使大气质量的预测更具现实意义，在选取数据时选择以月为单位进行，共提取了 2013 年 12 月——2024 年4 月各数据指标。利用 Python 表格操作对数据作预处理，删除无用列数据。

**2.缺失值处理**

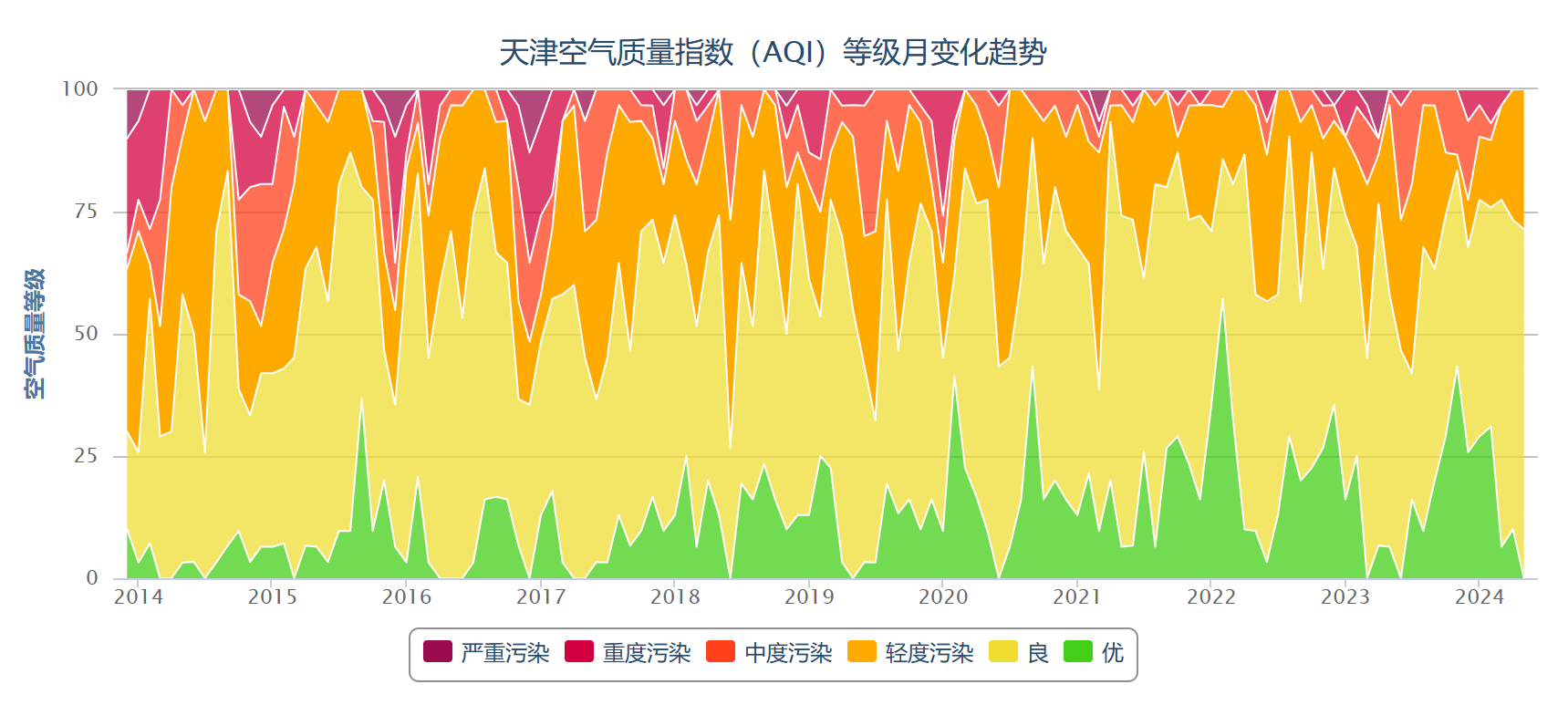
本文对数值型数据使用均值进行填补,对分类型数据使用众数补全。

**3.异常值处理**

本文使用Z-score方法来识别和处理异常值，(原理)

**4.数据标准化**

不同种类的的数据单位不统一将导致比较失去意义，由此我们需要借助数据标准化的方式是两者居于同一比较地位，相对于零均值标准化，我们倾向于采用 min-max 标准化对数据进行归一化处理，将表格中的数据经过变换转化成没有量纲的表达式,缩放到0和1之间；目的是使各个特征维度对目标函数的影响权重是一致的；改变了原始数据的一个分布，(min-max公式)



Min-max标准化公式如下:

公式（1）

四、相关性分析

（一）灰色关联度分析

我们将AQI的值作为重要的评判指标,也就是因变量。自变量作为参考序列以反映空气质量等级的评判标准，对因变量有影响的各指标即 21 个自变量作为比较序列，分别用 X0 和 Xi（i=1, 2, …, 21）

由于分析序列过分冗长，在此不做具体展示，对比子母序列并结合下列计算(公式)，

为了考虑序列的整体变化趋势和局部变化特征，

通过上述公式，我们可以计算出两级最小差值（a）和两级最大差值（b）。通过综合考虑最大值和最小值，我们能够更全面地描述序列的特征和变化趋势，并进而准确计算关联度。使用这种方法，可以在一定程度上平衡整体趋势和局部特征之间的关系，从而提高灰色关联模型的准确性和可靠性。

(因此，我们将其代入最终的关联系数计算公式)：

代入分辨系数，求出关联系数的值

灰色关联度计算

通过除以样本量 n，可以对关联系数进行归一化。由于元素组成和粒径所反应的 每个数据点上的关联系数取值范围可能存在差异，并且样本量的大小也可能会影响计 算结果的尺度。为了使得不同样本量和不同数据范围的序列能够进行比较和对比，我 们需要对关联系数进行归一化处理，将其限制在[0,1]的范围内。

（二）逐步回归分析

逐步回归分析方法的是自动从21种可供选择的变量中选取最重要的变量，首先将自变量逐个引入，引入的条件是其偏回归平方和经检验后是显著的。同时每引入一个新的自变量后，要对旧的自变量逐个检验，剔除偏回归平方和不显著的自变量。这样一直边引入边剔除，直到既无新变量引入也无旧变量删除为止。它的实质是建立“最优”的多元回归方程。本文使用的是向前法建立回归分析的预测或者解释模型。首先分析模型拟合情况R2,以及可对VIF值或者容忍度值。容忍度=1/VIF值进行分析判断，VIF>5一般说明特征间有共线性。

图

分析X的显著性,如果显著,则说明X对Y有影响关系,接着具体分析影响关系方向；将PM2.5, PM10, CO, NO2, SO2, O3,平均气温, 最高气温, 最低气温, 平均相对湿度, 日照时效, 降水量, 空气最差, 空气最好, 一日最大降水量, 平均风速, 平均空气质量指数, 平均高温, 平均低温, 极端高温, 极端低温作为自变量，而将AQI作为因变量进行逐步回归分析，经过模型自动识别，最终余下NO2, SO2, PM2.5, PM10, CO,降水量, 平均风速, 最高气温作为自一共8种特征,最后得到模型的R方值为0.790，意味着这几种特征可以解释AQI的79.0%变化原因。并且模型通过F检验(F=120.716，p=0.000<0.05)，说明模型有效。

灰色关联度分析和逐步回归分析,我们决定将NO2, SO2, PM2.5, PM10, CO, 降水量, 平均风速, 最高气温这8种特征作为自变量.

1. 基于随机森林创新模型的构建与评价
2. 多元有序逻辑回归

通过前面的相关性分析,我们决定将NO2, SO2, PM2.5, PM10, CO, 降水量, 平均风速, 最高气温这8种特征作为自变量，而将质量等级作为因变量进行有序logistic回归分析，并且使用Logit连接函数来预测当地某一时刻的天气状况属于哪一种质量等级。

**1.连接函数选择**

因为选取的8种变量的分布较为均匀,所以我们选用Logit函数作为连接函数。

原理

式中为感知变化等级；为自变量下标；代表移民对于水资源质量变化的感知的发生概率；为10个自变量；为系数；为截距；为误差。

式中为比数,指相对于不发生的可能性而言,发生的可能性。可以得到有意义的解释,并且除去上限的限制。

步骤二：取(2)式的对数：

公式（3）

可知(3)式无上下限,并且以0.5为中点对称,概率上很小的改变会引起更大的变化,使得数据更易观测。

步骤三：在没有了上限和下限的情况下, 相对于上的改变可以说是线性相关的:

公式（4）

则有：

公式（5）

由此可知,(5)式就是logistic回归模型的一般表达式。

1. **平行性检验**

平行性是有序Logit回归的前提条件，如果不满足平行性就无法使用该模型,因此进行平行性检验

从结果可知p值大于0.05,说明模型接受原假设，即模型满足平行性检验,接受原假设。

1. **似然比检验**

首先对模型整体有效性进行分析(模型似然比检验)

从上表可知：chi=52.418,p=0.000<0.05此处拒绝原假设,且说明本次构建模型时，放入的自变量具有有效性，本次模型构建有意义。

1. **模型建立和参数假设**

参数估计原理:最大数似然估计

参数估计方法:梯度下降算法

得到有序逻辑回归模型

**5.基于机器学习的逻辑回归**

最后使用有序逻辑回归算法进行预测, 通过模型预测准确率去判断模型拟合质量，由结果可知：研究模型的整体预测准确率为77%，模型拟合情况可以接受。

1. 随机森林、XGBoost、LightGBM

在数据分析的初始阶段，决策树对于探索多个因素与目标变量（即AQI值）之间的关系非常有用且易于解释。为了进一步确保结果更加准确可靠，我们首先介绍三种基于决策树的集合方法，包括随机森林、轻梯度提升机（Light Gradient Boosting Machine，简称轻GBM）和XGBoost。随机森林是一种袋算法，每棵树都是独立构建的，最后的预测结果由每棵树的预测结果投票决定。轻型GBM是一种梯度提升算法，树是按顺序建立的，每棵后续树都会纠正前一棵树的错误，以改善结果。XGBoost（eXtreme Gradient Boosting）是另一种集成学习算法，它在决策树基础上进行了扩展和优化。与传统的梯度提升算法相比，XGBoost通过引入正则化项和特定的分裂点查找算法，提高了模型的泛化能力和训练速度。

这三种模型有着共同的建树底层逻辑，并应用了不同的算法，通过比较模型的训练情况，我们可以根据测试集上准确率相对较高的算法，大致预测AQI的值。为了在一开始就测试三种方法的性能，我们将每个月的数据提前分成两组不同的数据集：训练数据集和评估数据集。其中70%的数据分配给训练集，其余30%分配给评估集。

1. **随机森林的基本原理**

随机森林(Random Forest)是一种基于集成学习思想的算法，通过构建多个决策树来实现分类、回归等任务。其核心原理涉及决策树的构建、Bagging集成、特征随机性以及预测过程。

①**集成学习**

随机森林作为一种集成学习思想的算法，通过训练学习出多个估计器，当需要预测时通过结合器将多个估计器的结果整合起来当作最后的结果输出，如图所示。

②决策树的构建

决策树是随机森林中的基本组成单元。在构建每棵决策树时，通过递归地将数据集分割成更小的子集。

分割时选择的特征和分割点是通过某种标准来确定的，例如信息增益、基尼系数等。最终得到的树形结构包含根节点、内部节点和叶节点，其中叶节点表示最终的分类或回归结果。

③Bagging集成

随机森林使用了Bagging（Bootstrap Aggregating）集成方法，通过随机抽样构建多个决策树，然后对它们的预测结果进行组合。

Bagging算法是一种集成学习算法，其全称为自助聚集算法（Bootstrap aggregating），由 Bootstrap 与 Aggregating 两部分组成，下图展示了Bagging 算法使用自助取样（Bootstrapping4）生成多个子数据的示例。

算法的具体步骤为：假设有一个大小为 N 的训练数据集，每次从该数据集中有放回的取选出大小为 M 的子数据集，一共选K次，根据这K个子数据集，训练学习出K个模型。当要预测的时候，使用这K个模型进行预测，再通过取平均值或者多数分类的方式，得到最后的预测结果。

随机抽样时采用的是有放回抽样，即从原始数据集中随机选择一部分样本形成训练集，使得每棵决策树所使用的训练集都有一定的差异性。

④特征随机性

在构建每棵决策树的过程中，不仅仅是对样本进行随机抽样，还对特征进行了随机选择。

随机选择一部分特征用于节点的分裂，而不是基于所有特征进行选择。这有助于减少模型的方差，增加模型的多样性，提高整体模型的性能。

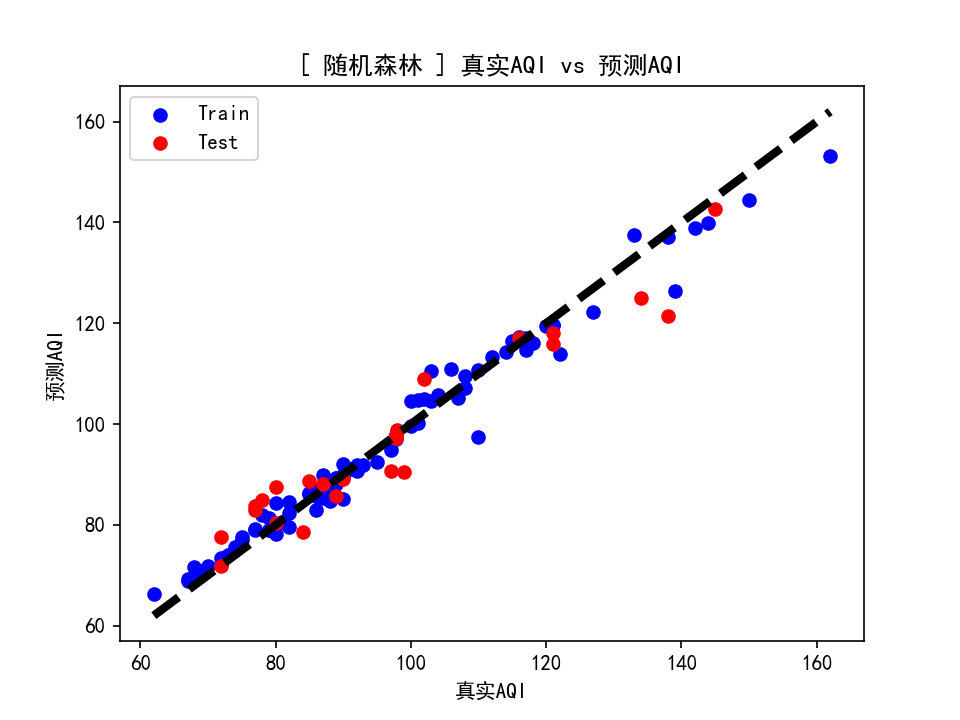
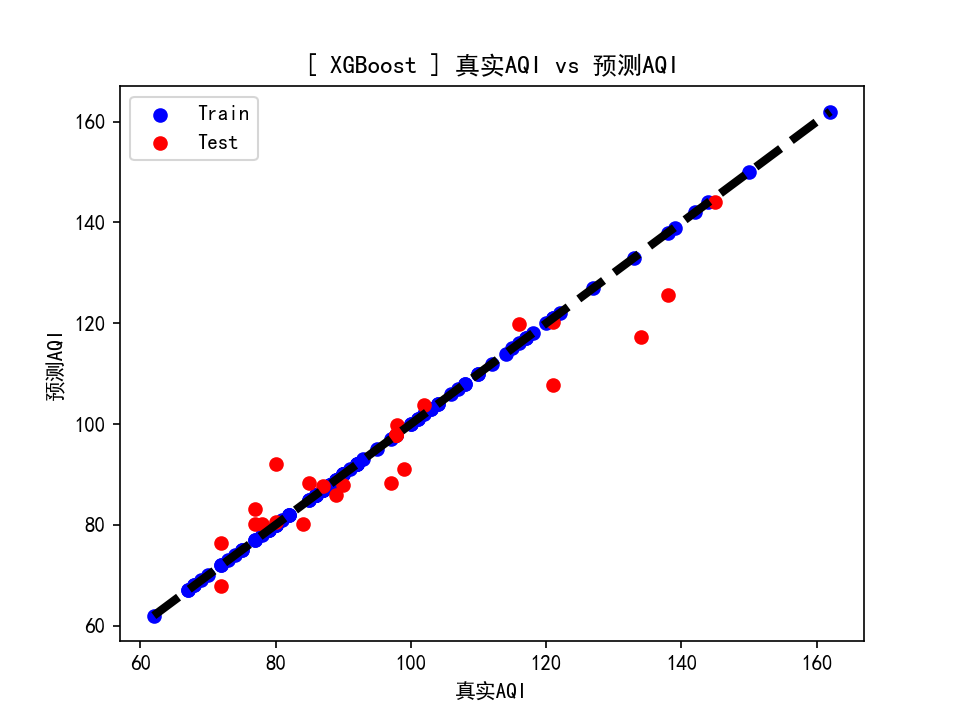
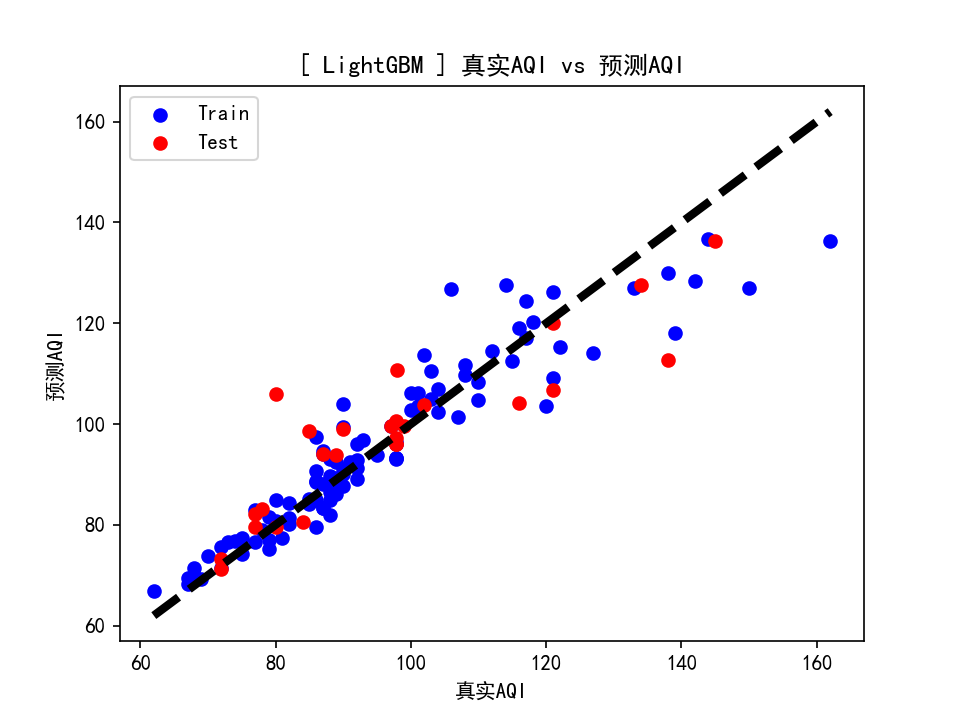
⑤预测过程

当需要对新数据进行预测时，随机森林中的每棵决策树都会输出一个预测结果。如下图所示，将多个决策树结合在一起，每次数据集是随机有放回的选出，同时随机选出部分特征作为输入。

对于分类任务，采用投票的方式，即最终的分类结果是所有决策树中出现最多次数的类别；对于回归任务，采用取平均值的方式得到最终的预测值。

通过以上步骤，随机森林能够有效地利用多个决策树的集体智慧，从而提高模型的性能和泛化能力。

三种算法的拟合效果如图所示：



其中：

放个表：

随机森林：

Train R^2 Score: 0.9757477919112507

Test R^2 Score: 0.9208221928727248

LightGBM:

Train R^2 Score: 0.8766610200363871

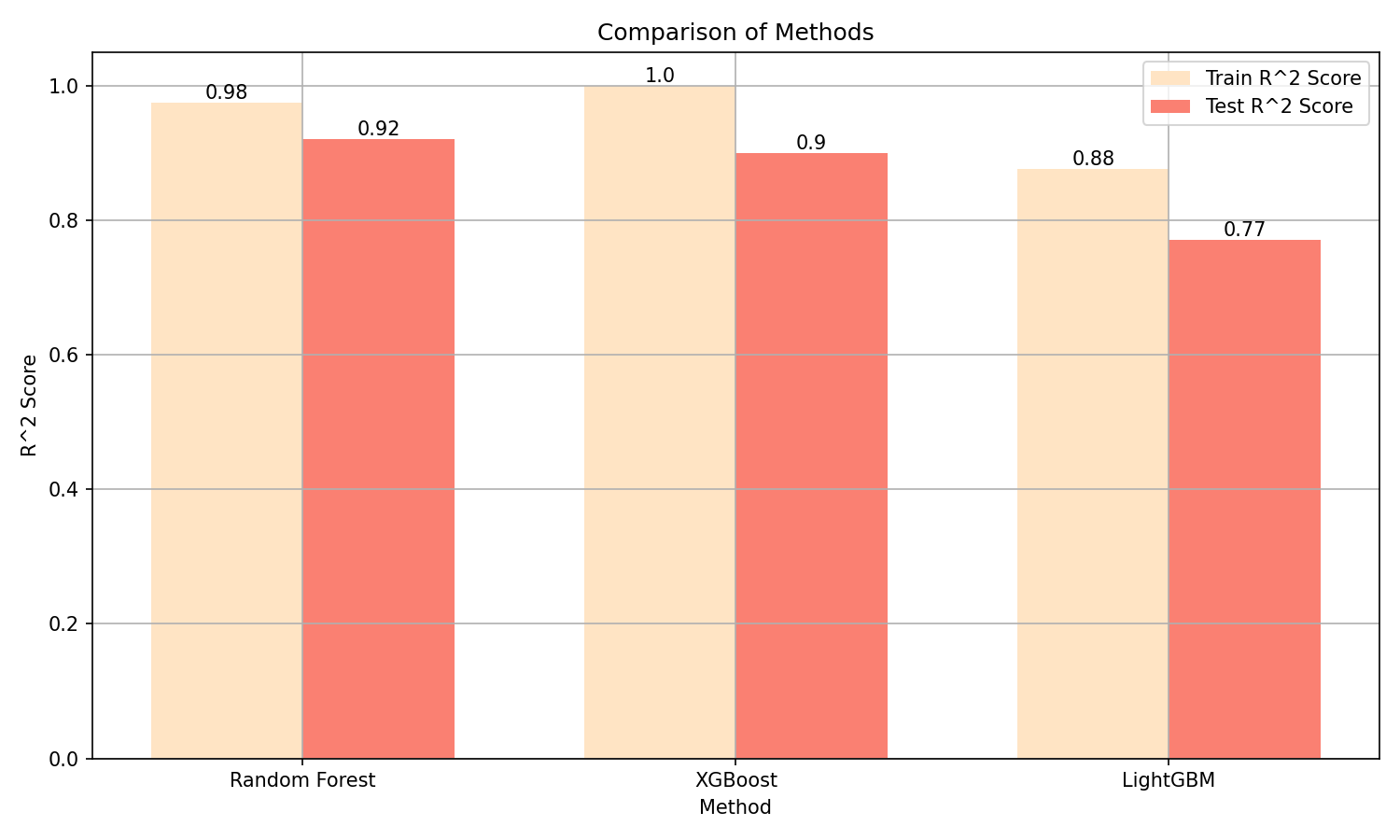
Test R^2 Score: 0.7713688256175901

XGBoost：

Train R^2 Score: 0.9999999971717937

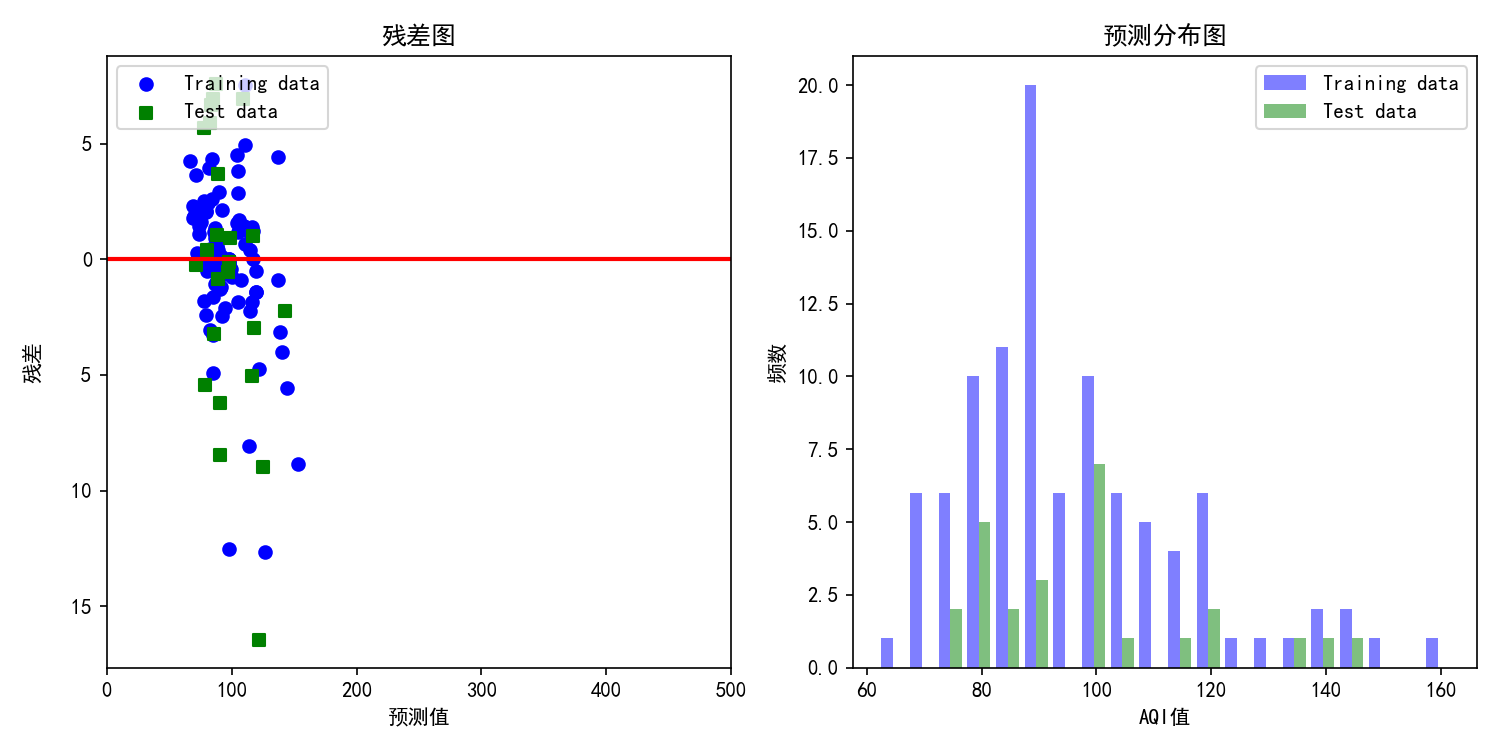
Test R^2 Score: 0.8999337510900779

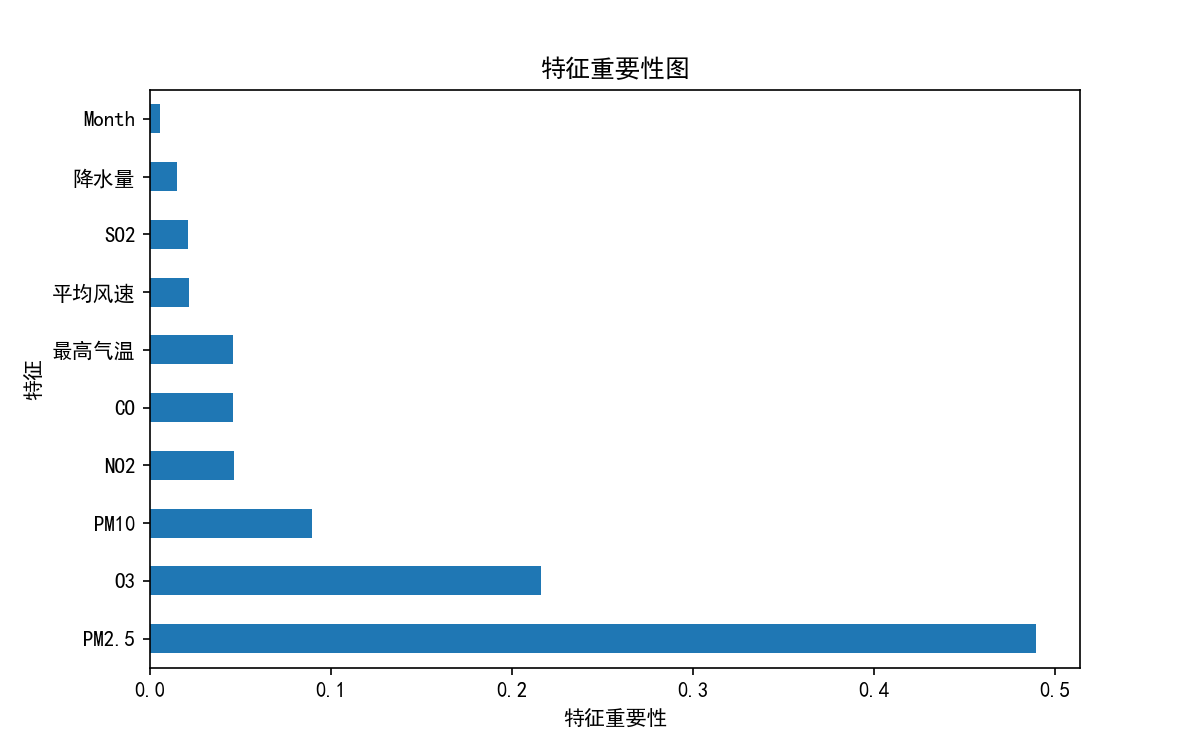
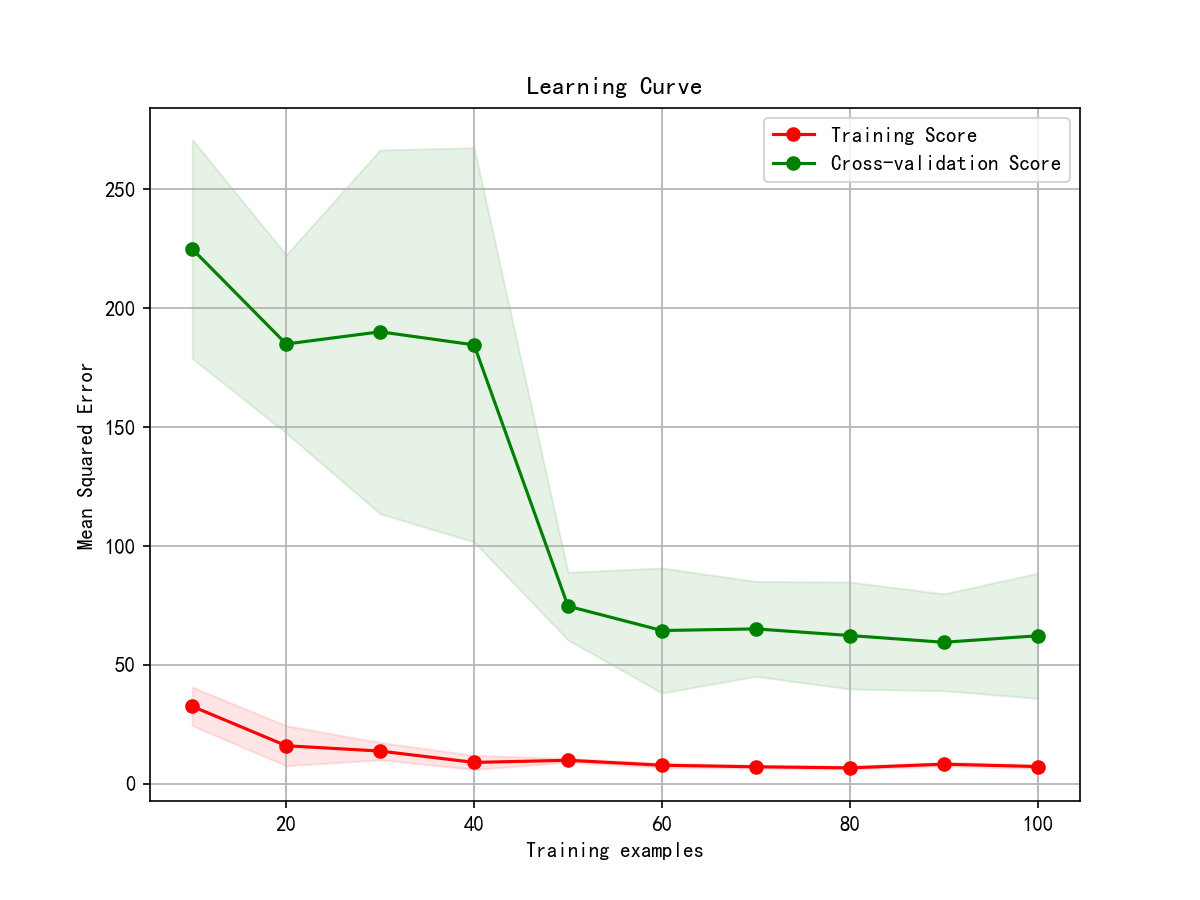
绘制出三个算法的



其中，随机森林的训练和预测效果均较好，LightGBM效果相对较差，XGBoost对于训练集有很好的拟合度，但是对于预测集的效果不好。

基于随机森林良好和稳定的分析预测效果，我们选择其进行进一步的优化算法。



v

（对其他可视化的分析）

学习曲线：Training Score逐渐减小且趋于稳定，可见随着训练集大小的增加，模型在训练集上的性能逐渐提高，但随着数据量的增加，提高的幅度逐渐减小，最终达到一个稳定的水平。这表示模型能够很好地拟合训练数据。

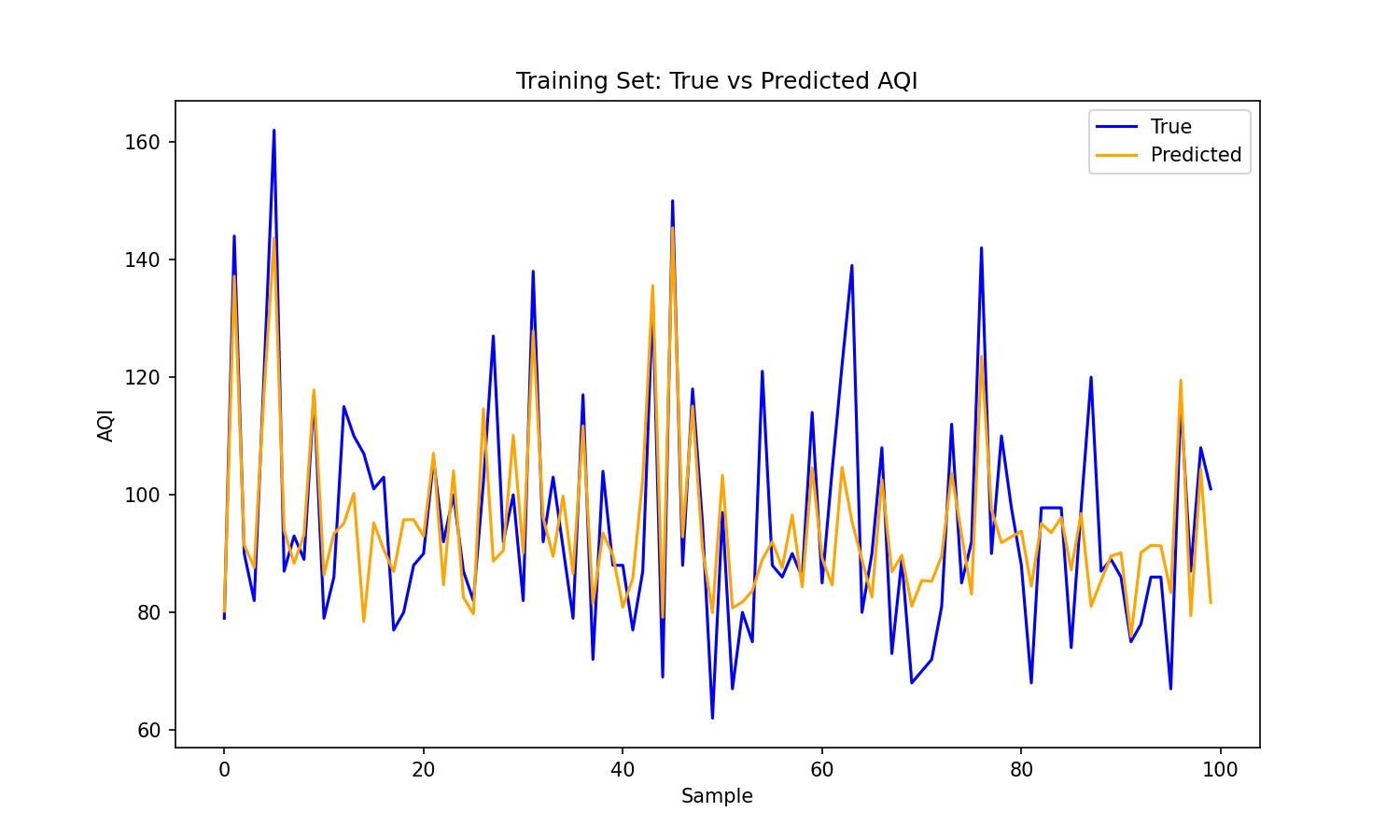
但是Cross-validation Score并没有逐渐增大且趋于稳定，可见随着训练集大小的增加，模型在交叉验证集上的性能并未逐渐提高，反而在不断下降。

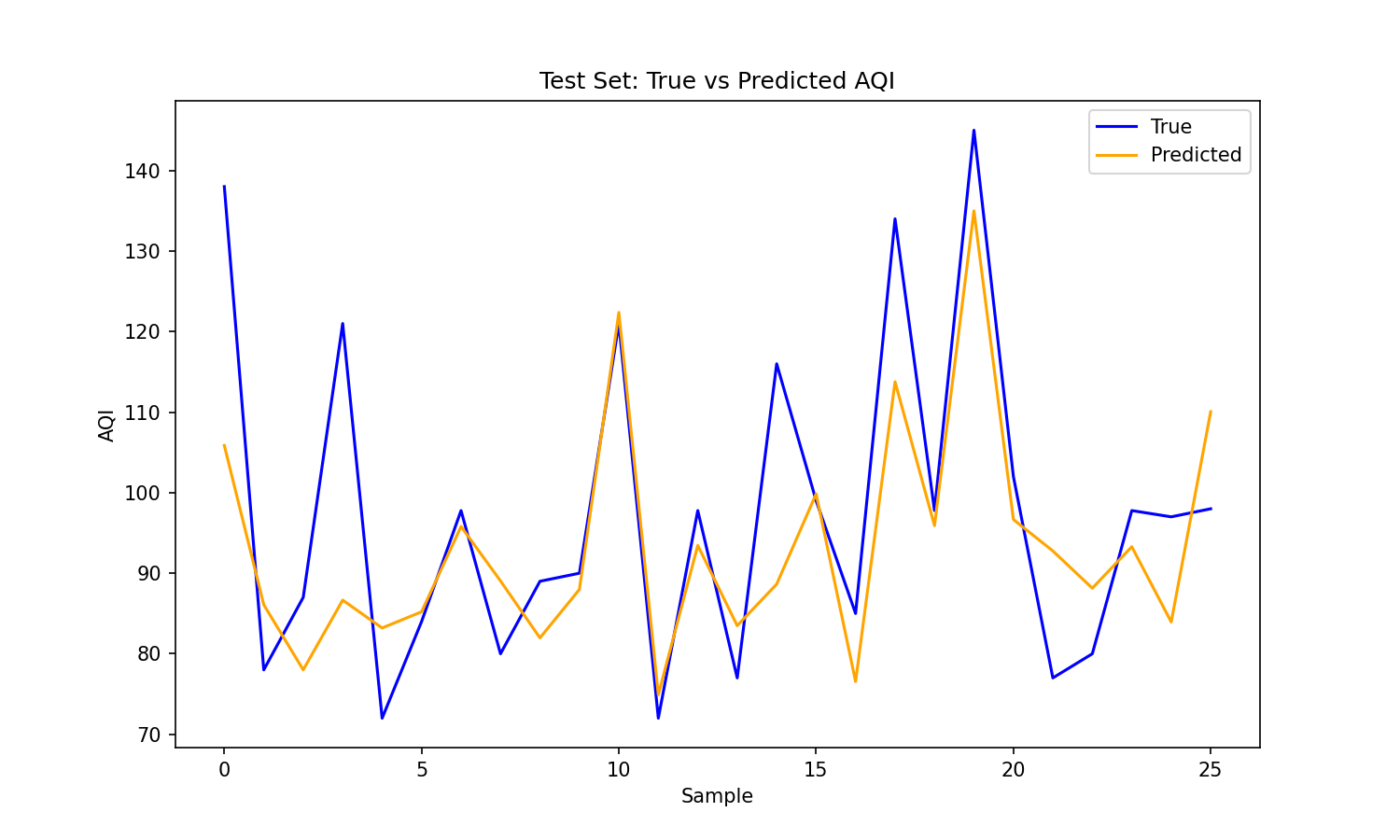
1. LSTM

LSTM（Long Short-Term Memory）是一种循环神经网络（RNN）的变种。RNN是一类用于处理序列数据的神经网络，它在每个时间步接受一个输入，并在下一个时间步产生一个输出，同时还会保存一些内部状态以处理序列信息。

LSTM是RNN的一种特殊类型，旨在解决传统RNN在处理长期依赖性时容易出现的梯度消失或梯度爆炸问题。它通过引入门控机制（如输入门、遗忘门、输出门）来控制信息的流动，有效地捕捉和利用长期依赖关系。

（放LSTM原理图 神经网络）

直接通过LSTM算法进行时间序列预测得到下图效果：

 其中：

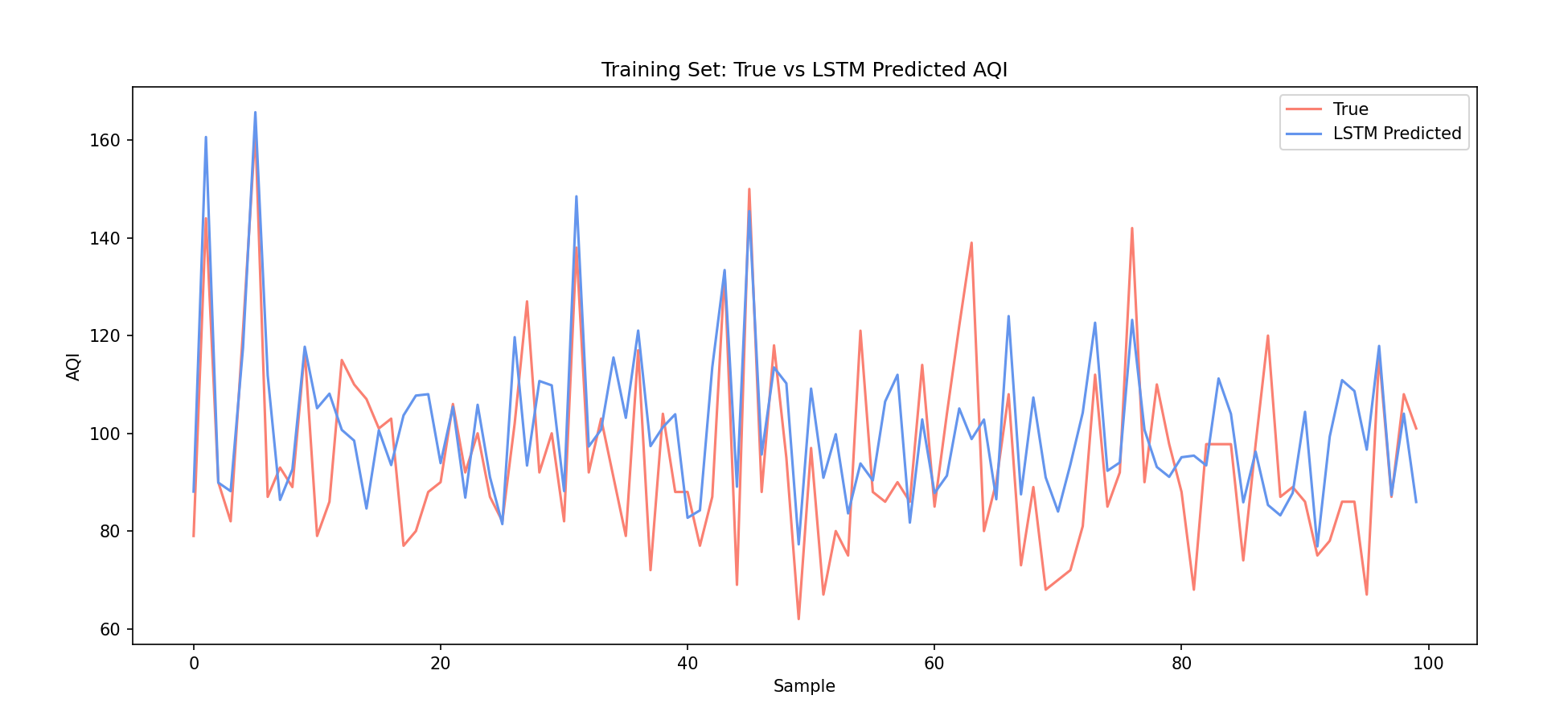
LSTM Train MSE Score: 202.85757446289062

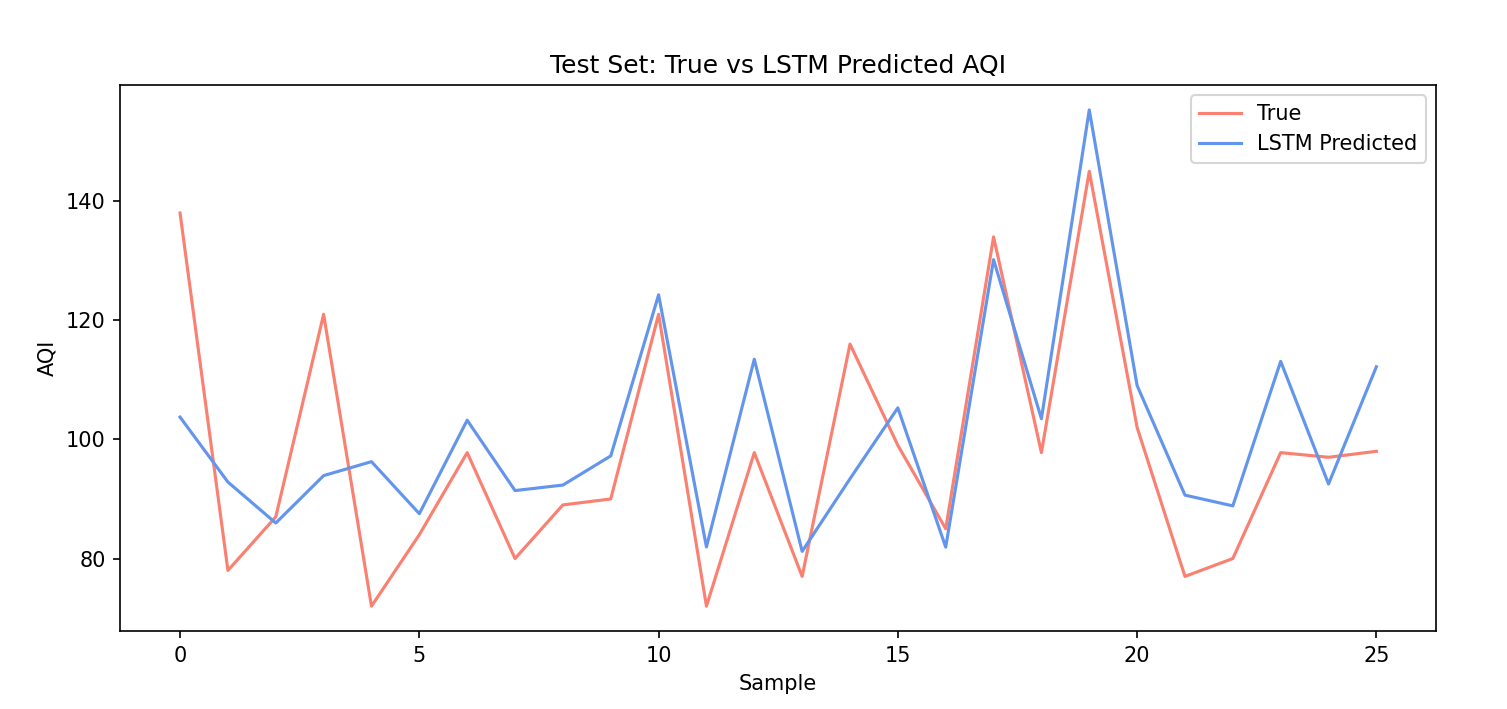
LSTM Test MSE Score: 198.930908203125

LSTM能进行基本的拟合和预测，但是不能达到最优的效果。

1. 基于随机森林的LSTM时间序列分析

首先使用随机森林模型对数据进行训练和预测，然后将随机森林的预测结果作为特征加入到LSTM模型中





Random Forest Train R^2 Score: 0.977653914328056

Random Forest Test R^2 Score: 0.9115190859292949

LSTM Train MSE Score: 150.00856018066406

LSTM Test MSE Score: 181.08969116210938

六、结论与展望

1. 结论

本文以天津市23个监测点的经纬度和2023年的6种大气污染物年浓度均值为基础数据，进行Kriging插值分析。由于监测点位不多，Kriging插值结果与站点监测值存在一定误差，故需要对插值结果进行可信度验证。Kriging 插值分析是一种基于地理空间数据的插值方法，它可以通过已知数据点的空间分布和属性值来估计未知位置的属性值. 本文选取了80%的监测点数据作为训练集来进行空间插值，剩余20%的监测点数据中山北路和西四道等四个监测点数据作为验证集来进行可信度验证。将模型预测值与站点实际监测值进行对比，发现平均绝对偏差为8%。由此可知，使用Kriging插值法对天津市大气污染物浓度的插值预测偏差较小，可信度高。

通过kriging插值得到的分析结果显示，天津市中心城区的空气污染物浓度分布图中，PM2.5和PM10的浓度呈现"西高东低"的趋势。而SO2和CO的分布格局变化较大，但年均浓度最高的地区仍然是天津市中心。重度污染区逐渐向城市郊区转移。NO2的空间分布呈现"城镇高、乡村低"的格局，同时随着时间推移，城市区域的污染有明显缓解的趋势，滨海新区核心区成为污染重点区域。O3的污染主要集中在西部和北部地区，并且随着时间的推移，O3的污染程度逐年加剧。

由图可知天津市各种大气污染物浓度分布有明显差异，天津市污染物浓度分布态势与本市的社会经济要素有关，城市中心人口密集，其污染主 要与机动车尾气和居民生活源排放有关，而新区和郊区的污染主要与工业污染和机动车尾气有关目前，天津市的大气污染得到了有效的缓解，尽管NO2的质量浓度波动较大，但2019年的年均质量浓度与2015年相比变化不大。机动车尾气排放仍然是天津市大气污染的重要贡献源，而NO2等污染物经光化学反应又能促进O3的生成。因此，控制机动车辆的污染排放是天津市当前大气污染治理中亟待解决的问题

（二）展望

针对天津市的大气污染问题，以下是一些具体的解决办法：

控制机动车辆尾气排放：推广清洁能源汽车，如电动汽车或混合动力汽车，并提供相应的充电基础设施。

强化车辆尾气排放标准，限制高排放车辆的进入和使用，并加强尾气排放监管和执法。

鼓励公共交通工具的使用，提供高效、便捷的公共交通网络，以减少私人汽车的使用量。

加强工业污染治理：严格执行和监督工业企业的污染排放标准，加强对污染物排放的监测和处罚力度。

鼓励工业企业采用清洁生产技术和设备，减少污染物的排放。

加强工业园区的环境管理，建设和改造污水处理设施，防止工业废水的直接排放。

提倡节能减排和可持续发展：推广能源高效利用技术，鼓励企业和居民使用节能设备和产品。

加强建筑节能标准，鼓励绿色建筑和能源管理措施的采用。

鼓励可再生能源的利用，如太阳能和风能等，以减少对传统能源的依赖。

加强环境监测和数据共享：建立健全的大气污染监测网络，覆盖城市中心、新区和郊区等不同区域，实时监测和报告污染物浓度。

加强数据共享和信息公开，提高公众对大气污染状况的认识，促进公众参与和监督。

加强跨部门合作和政策协调：建立跨部门的大气污染治理机制，加强政府各部门之间的协调合作，形成合力。

制定综合性的大气污染治理政策，包括法规、经济手段和激励措施，并加强政策的执行和监督。

这些解决办法需要政府、企业和公众的共同努力，通过改变行为习惯、技术创新和政策支持，逐步改善天津市的大气质量，减少污染物的排放，实现可持续发展和生态环境的保护。同时，需要持续监测和评估措施的效果，并不断优化和调整治理策略。