参赛队号：

2024 年（第十届）全国大学生统计建模大赛

参 赛 作 品

参赛学校：东北大学秦皇岛分校

论文题目：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

参赛队员：李卓航 韩卓文 罗峥

指导老师：刘建波

目录

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

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灰色系统预测模型[1][2][3] 、时间序列预测模型[4]、最优化权值组合法[5]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。

二、 数据描述及数据预处理

（一）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。

（二）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或间接获得。

（三）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。

4.数据标准化

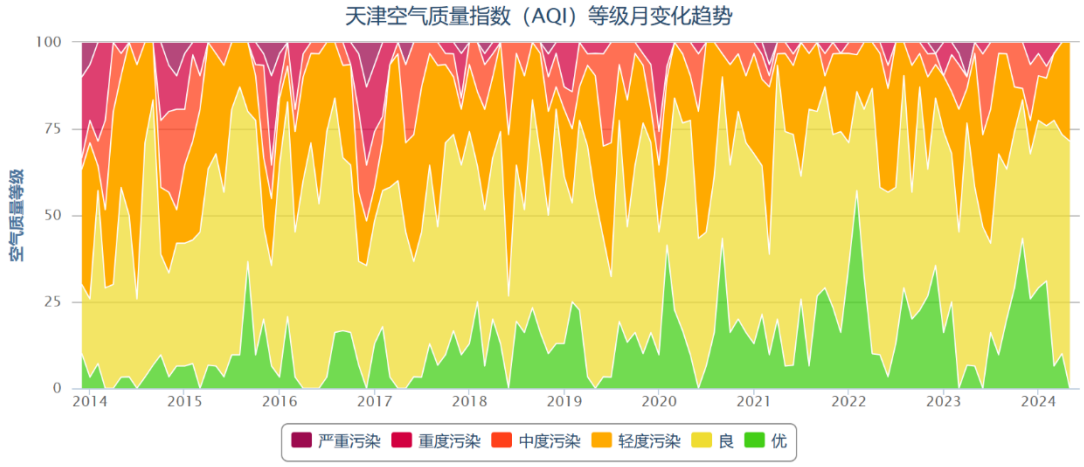
不同种类的的数据单位不统一将导致比较失去意义，由此我们需要借助数据标准化的方式是两者居于同一比较地位，相对于零均值标准化，我们倾向于采用 min-max 标准化对数据进行归一化处理，将表格中的数据经过变换转化成没有量纲的表达式，缩放到 0 和 1 之间；目的是使各个特征维度对目标函数的影响权重是一致的；改变了原始数据的一个分布，

Min-max 标准化公式如下:

$$x'_{ij} = \frac{x_{ij} - minx_{ij}}{maxx_{ij} - minx_{ij}}$$

5.数据可视化

对收集的数据进行可视化得到月变化趋势图，如图一：



图一 天津空气质量指数（AQI）等级月变化趋势图

三、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：

$$母序列: x_0 = [x_0(1), x_0(2), \cdots, x_0(n)]T$$

$$子序列: x_1 = [x_1(1), x_1(2), \cdots, x_1(n)]T$$

$$x_m = [x_m(1), x_m(2), \cdots, x_m(n)]T$$

为了考虑序列的整体变化趋势和局部变化特征，

$$a = \min_i \min_k |x_0(k) - x_i(k)|$$

$$b = \max_i \max_k |x_0(k) - x_i(k)|$$

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):

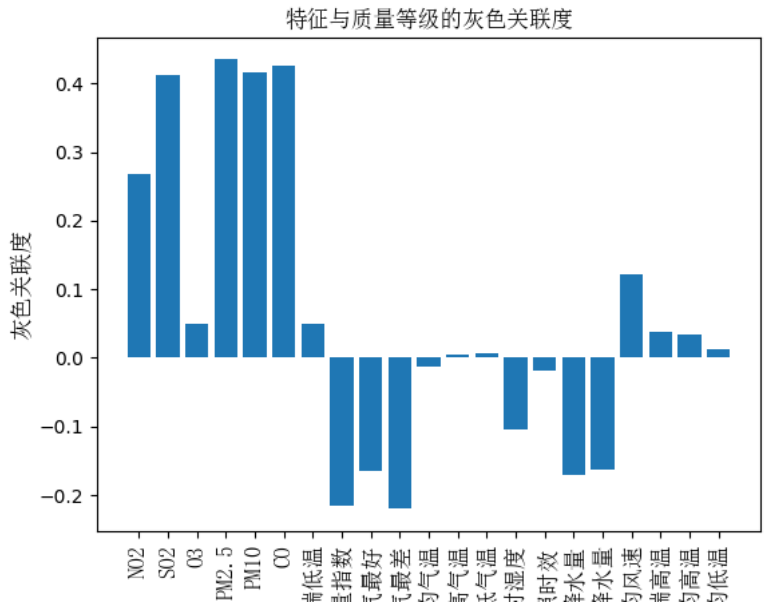
$$Y\big(x_0(k), x_i(k)\big) = \frac{a + \rho b}{|x_0(k) - x_i(k)| + \rho b}$$

代入分辨系数$\rho_{=0.5}$，求出关联系数$Y\big(x_0(k), x_i(k)\big)$的值

灰色关联度计算：

$$Y(x_0, x_i) = \frac{1}{n}\sum_{k=1}^{n} y\big(x_0(k), x_i(k)\big)$$

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。



图二 特征与质量等级的灰色关联度图

从结果可知，NO2，SO2，PM2.5，PM10，CO，降水量，平均风速，最高气温等特征与空气质量指数的关联度较高，意味着它们可能对于预测空气质量指数具有较强的影响力.

（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 R^2，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。

表 1 逐步回归分析部分结果

	非标准化系数		标准化系数	*t*	*p*	共线性诊断	
	B	标准差	Beta			VIF	容忍度
常数	-0.161	0.040	-	-4.050	0.000**	-	-
PM2.5	1.053	0.059	1.031	17.962	0.000**	1.508	0.663
03	0.625	0.083	0.797	7.569	0.000**	5.078	0.197
最高气温	-0.233	0.072	-0.340	-3.241	0.002**	5.030	0.199

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 $PM_{2.5}$，PM_{10}，CO，NO_2，SO_2，O_3,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 NO_2，SO_2，$PM_{2.5}$，PM_{10}，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。

四、基于随机森林创新模型的构建与评价

(一)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：

$$P(y \leqslant j) = \frac{e^{\alpha_j - \sum_{k=1}^{K} \beta_i x_{ij} + \varepsilon_i}}{1 + e^{\alpha_j - \sum_{k=1}^{K} \beta_k x_{ij} + \varepsilon_i}} \quad (1)$$

式中j为感知变化等级；i为自变量下标；$P(y \leqslant j)$代表移民对于水资源质量变化的感知的发生概率；X_{ij}为 10 个自变量；β_i为系数；α_j为截距；ε_i为误差。

$$O_i = \frac{P_i}{1 - P_i} \quad (2)$$

式中O_i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：

$$Y_i = \mathrm{Ln}\left[\frac{P_i}{1 - P_i}\right] \quad (3)$$

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，Y相对于X上的改变可以说是线性相关的:

$$\mathrm{Ln}\left(\frac{P_i}{1-P_i}\right) = b_0 + b_1 X_i = \log itY_i \quad (4)$$

则有：

$$P_i = \frac{e^{Y_i}}{1+e^{Y_i}} \quad (5)$$

由此可知,(5)式就是 logistic 回归模型的一般表达式。

2.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

H_0:有序 Logestic 回归满足平行性

H_A:有序 Logestic 回归不满足平行性

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。

3.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)

H_0:是否放入自变量两种情况时模型质量相等

H_A:是否放入自变量两种情况时模型质量不相等

表 2 有序 Logistic 回归模型似然比检验

模型	-2 倍对数似然值	卡方值	*df*	*p*	AIC 值	BIC 值
仅截距	188.770					
最终模型	136.352	52.418	8	0.000	156.352	184.715

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。

4.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

表 3 有序 Logistic 回归模型分析结果汇总

项	项	回归系数	标准误	z 值	Wald x^2	p 值	OR 值	OR 值 95% CI

因变量阈值	2.0	5.745	1.545	3.718	13.826	0.000	0.003	0.000~0.066
	3.0	10.388	1.937	5.364	28.776	0.000	0.000	0.000~0.001
自变量	NO2	-1.899	1.976	-0.961	0.924	0.337	0.150	0.003~7.198
	SO2	-0.415	2.460	-0.169	0.028	0.866	0.660	0.005~81.985
	PM2.5	7.787	3.701	2.104	4.426	0.035	2408.475	1.704~3405104.396
	PM10	1.518	2.648	0.573	0.328	0.567	4.562	0.025~818.955
	CO	2.860	3.245	0.881	0.777	0.378	17.459	0.030~10093.606
	降水量	0.169	1.365	0.124	0.015	0.901	1.185	0.082~17.191
	平均风速	1.087	1.255	0.866	0.750	0.386	2.965	0.253~34.694
	最高气温	2.554	1.218	2.097	4.398	0.036	12.861	1.182~139.980

并得到有序逻辑回归模型：

$$Logit(odds) = 5.745 - 1.899 * NO_2 - 0.415 * SO_2 + 7.787 * PM_{2.5} + 1.518 * PM_{10}$$

$$+ 2.860 * CO + 0.169 * 降水量 + 1.087 * 平均风速 + 2.554 * 最高气温$$

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。

建立模型之后，使用有序逻辑回归算法对已有的数据进行分类，可以判断在某一时刻该地区的空气质量等级， 通过模型预测准确率去判断模型拟合效果，由结果可知：

	precision	recall	f1-score
2	0.79	0.88	0.83
3	0.71	0.71	0.71
4	0.00	0.00	0.00
accuracy			0.77
macro avg	0.50	0.53	0.52
weighted avg	0.71	0.77	0.74

从结果可知模型对于类别 2 的性能表现较好。精确度为 0.79 表示在所有被模型预测为类别 2 的样本中，有 79%是真正属于类别 2 的。召回率为 0.88 表示

在所有真实属于类别 2 的样本中，有 88%被模型成功预测为类别 2。F1-score 为 0.83 是精确度和召回率的调和平均值，综合考虑了精确度和召回率的表现。对于类别 3，模型的性能较为一般。精确度为 0.71 表示在所有被模型预测为类别 3 的样本中，有 71%是真正属于类别 3 的。召回率为 0.71 表示在所有真实属于类别 3 的样本中，有 71%被模型成功预测为类别 3。

模型的整体准确率为 0.77，模型在所有样本上的正确分类比例为 77%。综合来看，该有序逻辑回归模型在类别 2 上表现良好，但在类别 3 和类别 4 上性能较为一般。所以我们选择其他的模型进行优化。

（二）随机森林、XGBoost、LightGBM、LSTM

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、XGBoost 和轻梯度提升机（Light Gradient Boosting Machine，简称轻 GBM）。同时，为了更好地对月份这一时间序列数据进行分析预测，我们使用 LSTM 这一长短期记忆递归神经网络进行时间序列上的分析和预测。

1.随机森林的原理

随机森林是一种使用多棵决策树对样本进行训练、分类和预测的方法。在数据分类过程中，可以通过每个变量的重要性来衡量其在分类中的地位。随机森林中的 "随机 "有两层含义：首先，样本选择是基于带撤回的抽样，这意味着每个样本都有可能被多次选中或不被选中。这种抽样方法可以有效地增加决策树之间的差异，进一步降低过度筛选的风险。其次，在构建决策树的过程中，随机森林并不是使用所有特征来构建每一棵决策树，而是从所有特征中随机抽取一部分特征来构建决策树。这种方法可以有效降低特征之间的相关性，提高模型的性能。在随机森林中，每棵决策树都是通过对数据进行迭代分区来构建的。在构建决策树的过程中，数据会根据特定的指标进行分割，直到达到预定的停止条件。为了避免过度分裂问题，随机森林还可以通过控制决策树的深度和节点停止分裂的最小样本数等参数来限制决策树的生长。

2.XGBoost 的原理

XGBoost 算法属于 Boosting 框架，其本质区别在于优化残差树所需的增益

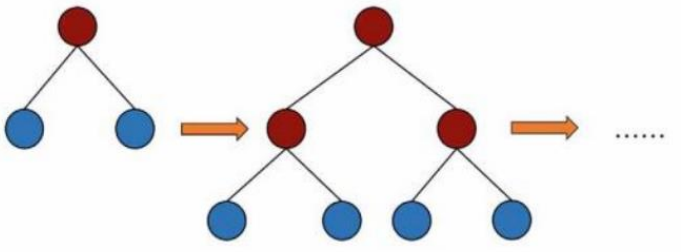
不同，XGBoost 使用的增益是分割前后的结构得分之差。XGBoost 的一个重要特点是引入了新的分割标准，在最佳分割点将分割损失降至最低。XGBoost 算法的核心思想分为三个步骤。首先，采用特征分割法不断增加树，每增加一棵树实质上都是学习一个新函数来修正上次预测的残差。其次，在完成训练并获得 k 棵树后，应预测样本的得分。第三，样本的预测值是每棵树对应分数相加的结果。XGBoost 模型迭代 m 次后，目标函数的定义如下式所示。

$$O^{(m)} = \sum^{n} l(y_i, \hat{y}_i^{(m-1)} + f_t(x_i)) + \Omega(f_i)$$

式中，l 和 Ω 分别为损失函数和正则项，yi 和 ybi 分别代表结果对应的真实值和模型的预测值。

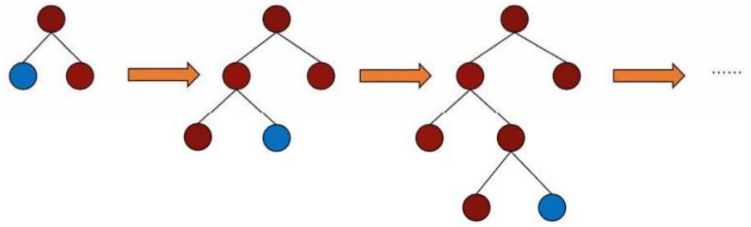
3.LightGBM 的原理

LightGBM 旨在解决 GBDT 在处理大规模数据时所面临的挑战，使 GBDT 能更好、更快地应用于工业实践。LightGBM 通过引入直方图算法和采用限制性分叶策略，克服了 XGBoost 算法内存消耗大、训练时间长等缺点。直方图算法利用直方图找出最佳分割点，处理连续变量，减少特征中的特征值数量，并减少叶节点分割时需要处理的特征值数量。其基本思想分为三个步骤。首先，将连续波动点特征值离散为 k 个整数，并构建宽度为 k 的直方图。然后，在遍历数据时，将离散值作为指数并累积到直方图中。在此基础上，进行遍历并找到最佳分割点。如下图所示，XGBoost 算法采用 Level-wise 作为增长策略。这种策略对数据进行一次遍历，可以同时分割同一层的叶子，有利于控制模型的复杂度，达到控制迭代的效果。但在实际应用中，大多数叶子的分割增益相对较小，因此没有必要对叶子进行搜索和分割，从而避免了不必要的计算。



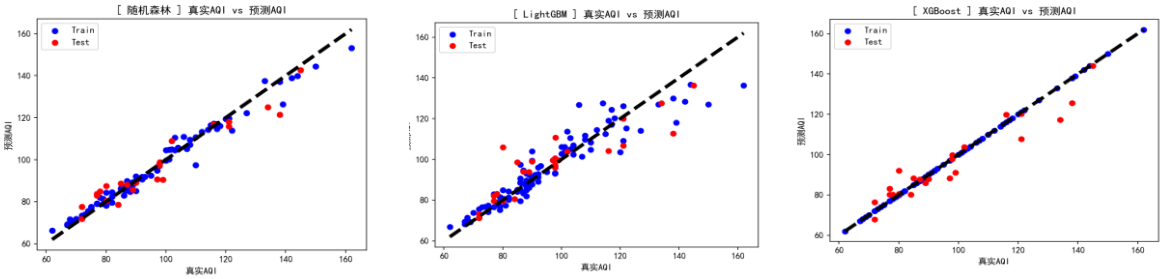
LightGBM 算法使用 ″ Leaf-wise ″作为生长策略。如图 13 所示，Leaf-wise 算法每次从当前的所有叶子中进行分割，找出分割增益最大的叶子，然后重复这一过程。与分层算法相比，分叶算法具有以下优点。在分割数量相同的情

况下，叶式算法能有效减少误差，提高算法的准确性。但叶式算法的缺点是会形成一棵更深的决策树，从而导致算法过度。因此，LightGBM 算法在 ″ Leaf-wise



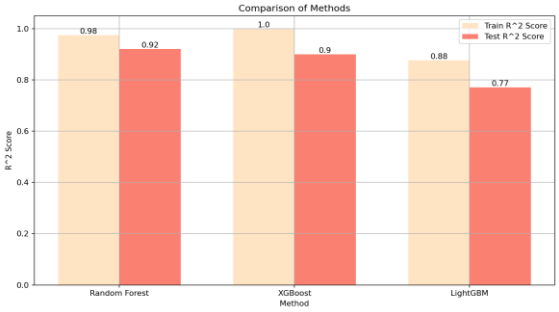
″算法的基础上增加了最大深度限制，以避免过密，提高计算效率。

使用三种算法进行分析预测，得到 R 方值并绘制拟合效果图：



图三 随机森林、XGBoost、LightGBM 拟合效果图

其中绘制出三个算法的 R 方值如下：



图四 随机森林、XGBoost、LightGBM R 方值可视化图

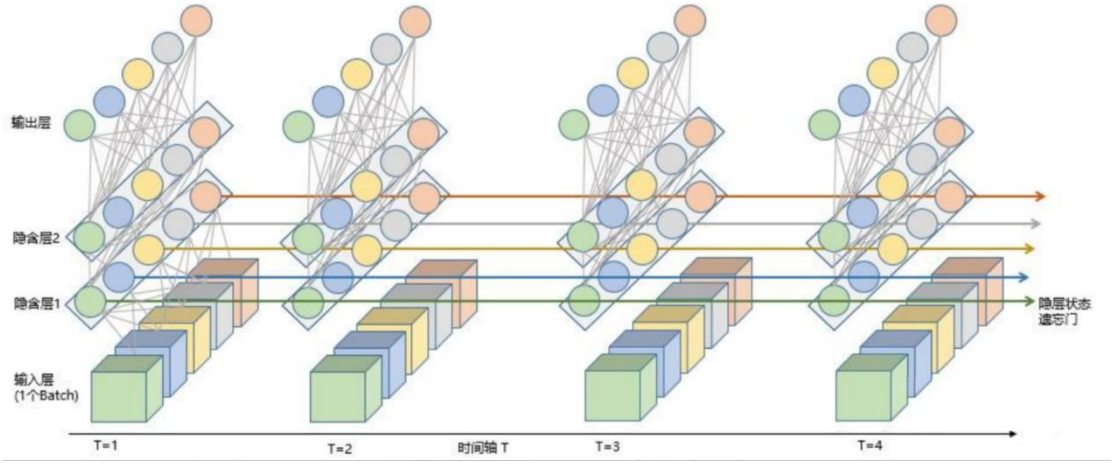
从图中可知，随机森林的训练和预测效果均较好，LightGBM 效果相对较差，XGBoost 对于训练集有很好的拟合度，但是对于预测集的效果不好。

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。

4. LSTM 的原理

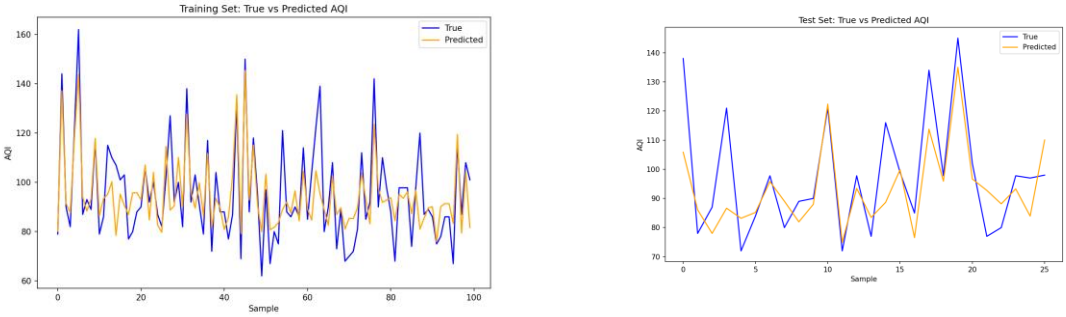
LSTM（Long Short-Term Memory）是一种循环神经网络（RNN）的变种。RNN 是一类用于处理序列数据的神经网络，它在每个时间步接受一个输入，并在下一个时间步产生一个输出，同时还会保存一些内部状态以处理序列信息。

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。



图五 LSTM 原理图

通过 LSTM 算法进行时间序列预测得到下图效果：



图六 LSTM 训练集和预测集拟合图

其中：

LSTM Train MSE Score: 202.85757446289062

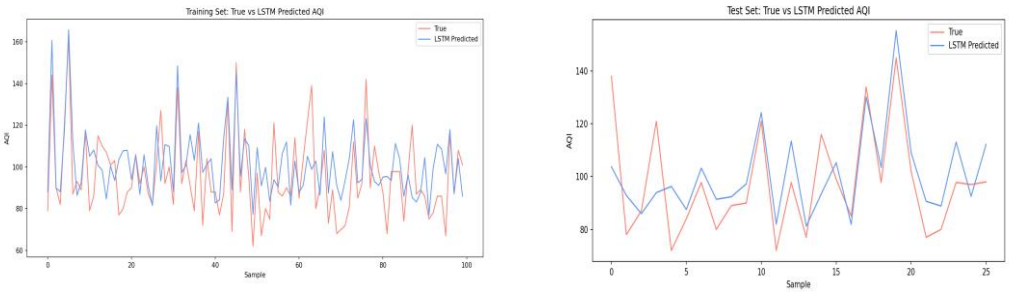
LSTM Test MSE Score: 198.930908203125

由图和数据可知：LSTM 能进行基本的拟合和预测，但是不能达到最优的效果，我们基于 LSTM 和随机森林进行下一步的模型优化。

（三）基于随机森林的 LSTM 时间序列分析

为了保留随机森林的良好预测结果和 LSTM 的时间预测特性，我们采取构建二者的集成算法，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，最后输出一个能进行时间预测的效果较好的模型。

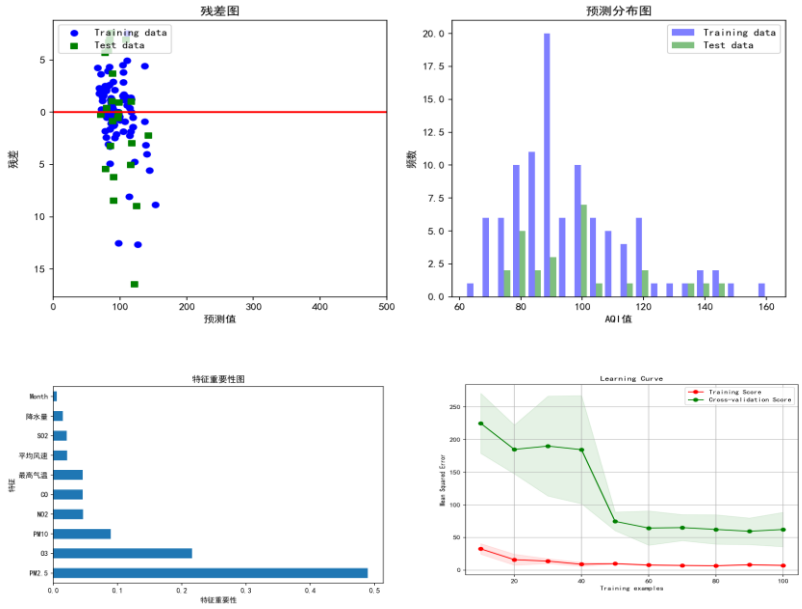
绘制出分析和预测的拟合图如下：



图七 基于随机森林的 LSTM 训练集和预测集拟合图

对最终模型进行评估：

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。

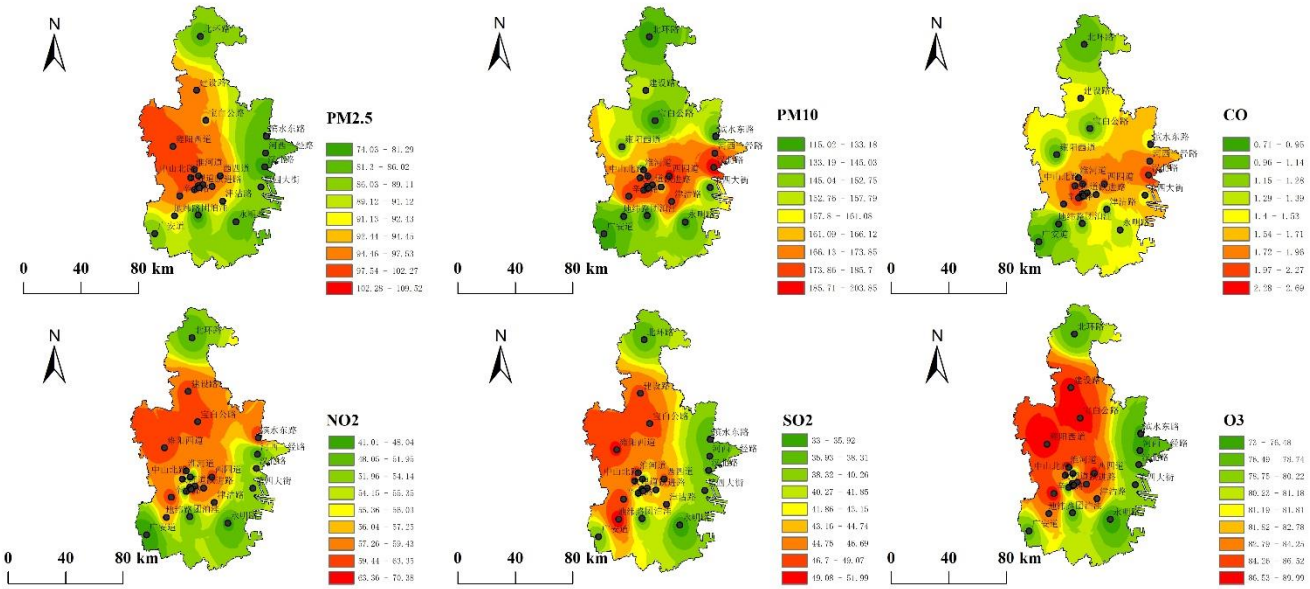


图八 模型评估可视化

至此构建和完整得到时间序列上的预测模型。

五、空间预测

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。因为 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 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 8.2%。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。



图九 空间插值分析图

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。

六、 结论与展望

（一）结论

在时间层面上,天津市天气污染有季节性变化趋势, 因为暖季节污染物反应速度较快：在暖季节，气温较高，大气中的化学反应速率加快，导致污染物的生成和转化速度加快。例如，光化学反应会促使 NO2 和挥发性有机化合物生成 O3，而暖季的高温和日照条件有利于这些反应的进行。另外天津市在暖季节通常有较高的风速。强风有助于将污染物稀释和扩散，减少空气中的污染物浓度。而在冷季节,风速较低,使得污染物在空气中停留时间更长,导致污染物的累积和积聚。最后冬季是供暖季节，天津市的能源消耗会大幅增加。燃煤、燃油和天然气等能源的燃烧会释放大量的污染物，如二氧化硫（SO2）、颗粒物（PM2.5 和 PM10）和一氧化碳（CO），从而导致冬季污染加重。

空间层面上,天津市部分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急需解决的问题。

（二）展望

大气污染是环境保护中很棘手的问题,针对天津市的大气污染问题，以下是一些具体的解决办法：

在交通工具层面上

1.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。

2.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。并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。

在公众层面上

1.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。并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。

2.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。并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。

在政府层面上

1.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。并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。

2.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。

参考文献

[1] 邓聚龙.灰色系统基本方法[M].武汉：华中理工大学出版社,1987.141-145.

[2] 王学萌,张继中,王荣.灰色系统分析及实用计算程序[M].武汉：华中科技大学出版社,2001.109-118

[3] 邓聚龙.灰色预测与决策[M].武汉：华中理工大学出版社,1987.96-120.

[4] 常学将,陈敏,王明生.时间序列分析[M].北京：高等教育出版社,1993.54-295.

[5] 唐小我.最优组合预测方法及其应用[J].数理统计与管理,1992,11(1)：31-35

附录

基于随机森林的 LSTM 时间序列算法代码：

```
import pandas as pd
import numpy as np
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestRegressor


# 读取数据
data = pd.read_excel("D:统计建模.xlsx")

# 将"Month"列转换为日期类型
data['Month'] = pd.to_datetime(data['Month'], format='%b-%y')

# 将"Month"列拆分为年和月，并添加这两列作为特征
data['Year'] = data['Month'].dt.year
data['Month'] = data['Month'].dt.month

# 确定特征和目标列
X = data.drop(["AQI"], axis=1)
y = data["AQI"]

# 划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# 创建随机森林模型
model_rf = RandomForestRegressor(n_estimators=100, random_state=42)

# 拟合随机森林模型
model_rf.fit(X_train, y_train)

# 使用随机森林模型对训练集和测试集进行预测
y_train_pred_rf = model_rf.predict(X_train)
y_test_pred_rf = model_rf.predict(X_test)

# 使用随机森林模型预测训练集和测试集的 AQI 值
y_train_pred_rf = model_rf.predict(X_train)
y_test_pred_rf = model_rf.predict(X_test)
```

```
# 将随机森林的预测结果添加到原始数据集中
X_train['RF_Prediction'] = y_train_pred_rf
X_test['RF_Prediction'] = y_test_pred_rf

# 数据缩放
scaler = MinMaxScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# 将数据重塑为 LSTM 所需的形状 (samples, time steps, features)
X_train_reshaped = np.reshape(X_train_scaled, (X_train_scaled.shape[0],
X_train_scaled.shape[1], 1))
X_test_reshaped = np.reshape(X_test_scaled, (X_test_scaled.shape[0],
X_test_scaled.shape[1], 1))

# LSTM 模型
model = Sequential()
model.add(LSTM(50, activation='relu',
input_shape=(X_train_reshaped.shape[1], X_train_reshaped.shape[2])))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')

# 训练 LSTM 模型
model.fit(X_train_reshaped, y_train, epochs=100, verbose=0)

# 使用 LSTM 模型进行预测
y_train_pred_lstm = model.predict(X_train_reshaped)
y_test_pred_lstm = model.predict(X_test_reshaped)
```

致谢

值此论文撰写完成之际，回望这将近一个月里的日日夜夜，首先我要由衷感谢我的队员。从论文的选题、数据收集、实证模拟到最后的论文撰写与修改，我们不断讨论、分析、相互协作，在每个深夜里挑灯夜读不知疲惫。正是有了彼此的鼓励和支持，有了集体的智慧，我们的论文才得以在规定时间内顺利完成。

同时，我们还要感谢刘建波老师，感谢他们一直以来对我们的关心和精心指导。在论文选题、梳理框架等环节，老师们都为我们付出的宝贵时间和极大心血，为我们在编写论文的大方向上提供了宝贵的意见。当我们在实证研究中遇到棘手的问题时，老师也耐心指导，让我们少走了很多弯路。在此衷心感谢各位老师对我们的无私帮助！

其次，感谢各位老师和同学们在学习上对我们的帮助，在统计学领域，使我们的专业素养得到了很大提升，同时也加深了对本专业的认知。

最后，我们要感谢含辛茹苦培养我们茁壮成长的父母，在钻研课题遇到困难的时候给予我们恰当的安慰和开导，并且在精神和物质上给予了我们稳固的支持。在这里，我们向我们的父母和家人表示最诚挚的谢意。

在整个论文的撰写过程中，有了队员们的默契合作、老师们的悉心指导和家人们的支持，我们才能够克服困难，完成了这篇论文。我们深知，每一份成果的背后都离不开众多人的辛勤付出和帮助。谨以此文表达我们最真挚的谢意，感谢每一位对我们论文完成所给予的支持和鼓励！