2024 年第九届“数维杯”大学生数学建模挑战赛论文

题 目：应用统计检验与机器学习方法探讨生物质与煤炭的协同热解特性研究

摘 要

在解决能源和环境问题的大背景下，研究生物质和煤炭共热解技术，以提升能源利用效率并减轻环境负担，显得尤为重要。本研究旨在开发运用一系列数学模型，深入分析共热解过程中的关键因素，并探讨如何调整操作参数，以提高产品收率和能源转化效率。通过构建以及应用数学模型和数据分析，本文不仅深化了对共热解机制的理解，还为实际应用提供了理论支持和优化策略。

针对问题 1，本文分析了正己烷不溶物（INS）对热解产率的影响。先通过描述性统计和相关性分析方法明确 INS 与热解产物产率存在不同程度相关性。后续利用再线性回归模型证实了 INS 对焦油和焦渣产率在统计学上可认为有显著的正负影响，但对水产率影响不显著。此外，为了揭示了 INS 对热解产物产率的影响趋势，本文利用散点图、箱线图和趋势线图等多种可视化手段，直观展示了数据分布和相关性。

针对问题 2，本文深入研究了热解实验中正己烷不溶物（INS）和混合比例对热解产物的影响。基于已有的实验数据，通过构建线性回归模型，评估了这两个因素对热解产物产率的交互效应。结果显示，焦渣产率和水产率受到 INS 水平和混合比例的显著交互影响。具体而言，焦渣产率在 INS 存在且混合比例增加时显著下降，而水产率在高 INS 水平与特定混合比例结合时显著上升。为了更清晰地展示这些交互效应，本文采用等高线图等高级数据可视化技术，以深化对这些相互作用的理解。

针对问题 3，本文探讨了如何使用数学模型优化生物质和煤的混合比例，在共热解过程中提升产物利用率和能源转化效率。研究首先构建了一个多元线性回归模型，用于模拟不同混合比例对热解产物（如焦油、水和焦渣）的影响。基于已有的实验数据，本研究在实际优化过程中通过应用优化技术——模拟退火算法，来确定最优混合比例，目标是最大化焦油产率和最小化焦渣产率即提高焦油生产的整体效率。方法指出，基本使用纯煤进行热解时，成功实现焦油产率的最大化，产率可达到约 9.28%。

针对问题 4，本文深入探究了生物质和煤共热解过程中不同组合的产物收率实验值与理论计算值之间是否存在显著差异。通过配对样本 t 检验分析，发现特定混合比例下，实验值与理论值之间存在显著差异。表明理论模型可能需要针对这些条件下的产物收率预测进行进一步调整。通过进行子组分析每种组合在不同比例点的实验值与理论值差异，确定了差异最为显著的混合比例。该分析不仅有助于识别影响热解效率的关键因素，还为改进热解模型和优化操作参数提供了数据支持。

针对问题 5，本文深入研究了如何运用多种机器学习技术来预测热解产物的产率。采用了随机森林、梯度提升回归（XGBoost）两种模型，分别对焦油、水、焦渣以及正己烷可溶物的产率进行预测。在模型训练和测试中，随机森林展现了较高的预测精度，尤其是在预测水产率和焦渣产率方面，其决定系数分别高达 0.976 和 0.964。尽管 GBR 在某些产物的预测上表现不错，但随机森林模型在多数情况下更为优秀。我们还提供了可视化部分，展示了模型在测试集上的预测效果，并通过随机选取的五个样本的真实值与预测值对比，来证实模型的有效性。

关键词：相关性分析；线性回归；模拟退火算法；配对样本 t 检验； 随机森林

目录

一、问题重述

随着全球对可再生能源的需求日益增长，生物质和煤共热解技术受到关注。生物质，作为可再生能源，来源于植物和动物的有机物质，而煤则属于化石燃料。共热解过程中，生物质与煤在高温无氧条件下同时热解，生成气体、液体和固体产物，其中液体产物被称为热解油或生物油。探究生物质和煤共热解油的产率和品质机制，对提升能源利用效率、推动资源综合利用和保障能源安全具有深远意义。本文选取了多种生物质和中低阶煤的共热解实验数据，研究如何对共热解产物预测和优化，有助于提高生物质与煤共热解过程的效率和产物利用率。

现需通过数学建模具体完成以下问题：

- 评估正己烷不溶物(INS)对热解产率（焦油、水、焦渣）的显著性影响，并通过图像解释描述分析结果。
- 检验正己烷不溶物(INS)与混合比例的交互作用对热解产物产量的影响，并确定哪些产物对此交互作用最为敏感。
- 构建模型以优化共热解的混合比例，旨在提升产物利用率和能源转化效率。
- 比较共热解组合的产物收率实验值与理论值，找出显著差异，并分析这些差异在哪些混合比例中最明显。
- 利用实验数据建立模型，用于预测热解产物的产率。

二、问题分析

在问题 1 中，需要分析正己烷不溶物（INS）对热解产率的影响。这可以通过进行统计相关性分析，如方差分析或回归分析，来检查 INS 的存在或浓度对热解产物（焦油、水、焦渣）产率是否有显著影响。本文将构建一个包含 INS 作为自变量的模型，评估其对热解产率的统计显著性。此外，图像分析将用于直观表示描述这些影响。

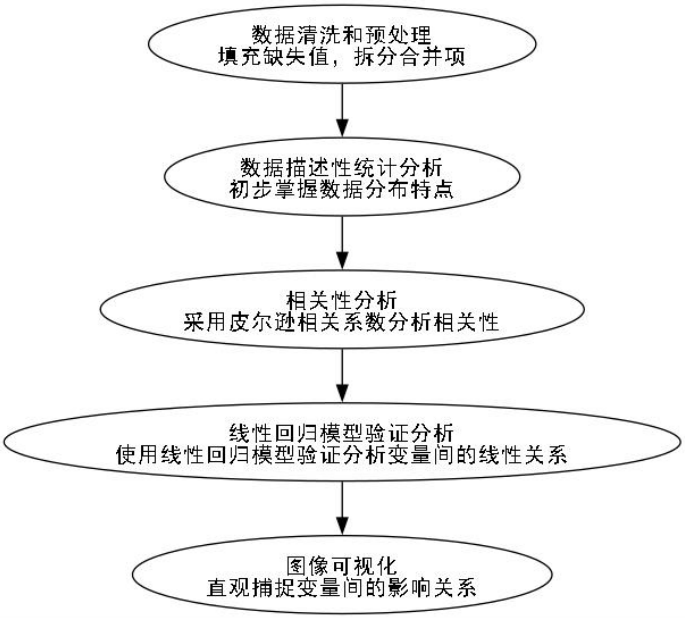


图 2-1：问题 1 分析解决流程

在问题 2 中，要求研究正己烷不溶物与生物质和煤的混合比例之间的潜在交互作用，以及这些作用如何影响热解产物的产量。本文通过建立一个包含交互项的多变量线性回归模型，基于此来评估和量化这些交互作用的统计显著性。关键是要识别在哪些热解产物中，混合比例和样品重量的交互影响最为突出。

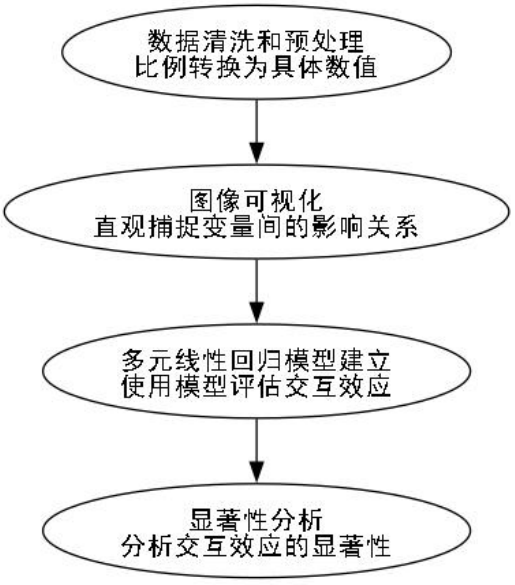


图 2-2：问题 2 分析解决流程

在问题 3 中，目标是构建一个数学模型，旨在优化共热解过程中生物质与煤的混合比例，从而提升目标产物的产率和能源转换效率。这涉及找出能够最大化特定热解产物（如焦油）产率的最优混合比例。实现这一目标可能需要采用优化算法，如遗传算法、蚁群算法、模拟退化算法等，来寻找最佳混合比例。

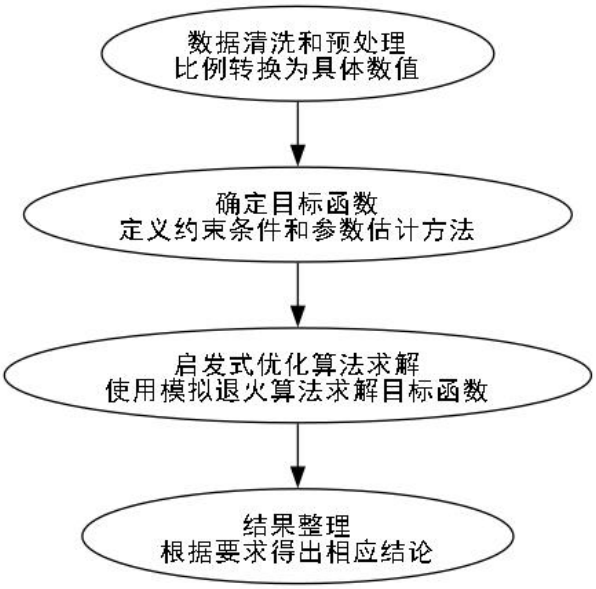


图 2-3：问题 3 分析解决流程

在问题 4 中，涉及到比较共热解实验中各种组合的产物收率实验值与理论值，以确定是否存在显著差异。通过进行配对样本 t 检验，可以评估实验值与理论值的一致

性，并分析这些差异在不同混合比例下的表现。这样的分析有助于了解理论模型的精确度和适用范围。

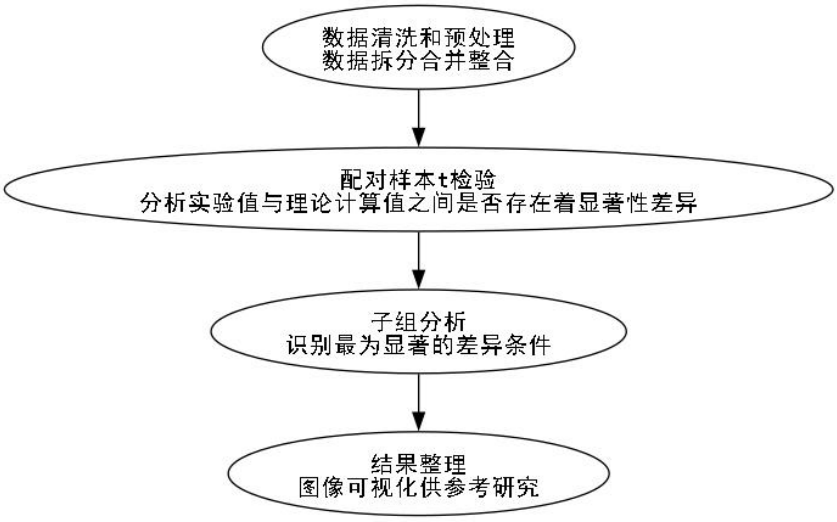


图 2-4：问题 4 分析求解流程

在问题 5 中，目标是利用已有的实验数据建立一个预测模型，以估计不同条件下热解产物的产率。关键步骤是选择合适的机器学习模型，并利用现有数据集训练这些模型。模型的选择将依赖于其预测的准确性以及对热解过程的解释能力。最终目标是提供一个精确且可靠的工具，以便高效预测和优化热解过程。

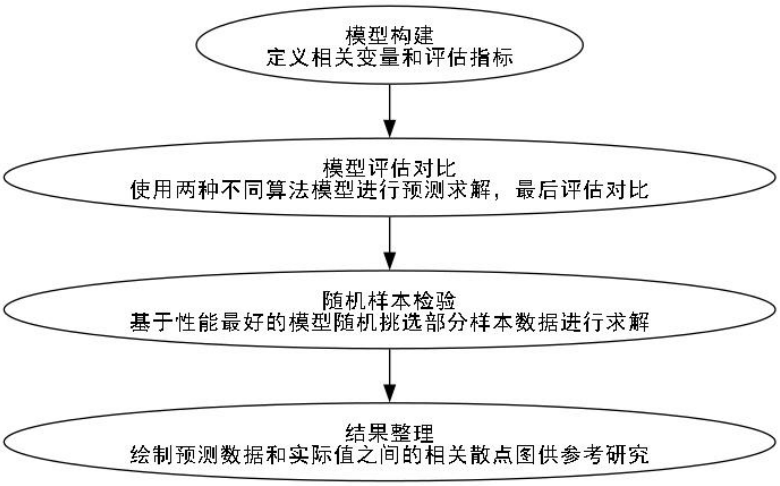


图 2-5：问题 5 分析求解流程

三、模型假设

- **均匀混合假设**：假设生物质和煤在共热解过程中能够完全混合，即它们的接触和反应是均匀的。
- **恒定反应速率假设**：假设共热解过程中，各组分的热解速率与温度的关系是恒定的，不受其他因素的影响。

- **化学平衡假设**：假设在共热解的高温条件下，所有的化学反应都能迅速达到平衡状态。
- **物质守恒假设**：假设在共热解过程中，反应前后物质的质量是守恒的，即反应物的质量等于产物的质量。
- **忽略副反应假设**：假设主要考虑的是生物质和煤的热解反应，忽略其他可能的副反应，如气体产物的二次反应。
- **理想热传导假设**：假设反应器内部的热传导是理想的，即温度在整个反应器内是均匀的。
- **恒定物理性质假设**：假设生物质和煤的物理性质（如热容、密度等）在共热解过程中保持不变。
- **正己烷不溶物(INS)的唯一性假设**：假设正己烷不溶物是影响热解产物产率的主要因素，其他因素如反应时间、温度等保持不变。
- **产率假设**：热解反应在一定条件下稳定进行，产率仅受原料组成和操作条件的影响。
- **线性关系假设**：假设影响热解产率的因素与热解产物产率之间存在线性关系或者能够通过线性模型进行近似拟合，便于模型的建立和分析。
- **实验数据准确性假设**：假设使用的实验数据是准确无误的，没有测量误差。

四、定义与符号说明

符号定义	符号说明
INS	表示热解过程中不溶于正己烷的物质
配比	代表生物质和煤的混合比例
Residue	代表焦渣产率，即热解过程中的残留固体物质
Tar	代表焦油产率，即热解过程中产生的焦油的量
Water	代表水产率，即热解过程中产生的水的量
HEX	代表正己烷可溶物产率
变量r_{xy}	皮尔逊相关性系数
R^2决定系数	用于衡量线性回归模型的拟合优度
P	用于统计检验中判断零假设是否成立的概率指标
MSE 均方误差	衡量模型预测与真实值的差异的评估指标

五、模型的建立与求解

5.1 问题 1 模型建立与求解

5.1.1 数据清洗与统计分析

在本小节中，针对题目中附件 1 的数据，如下图所示：

时间	试样	配比	样品g	焦油(Char)g	水(Water)mL	正己烷不溶物（INS)g	焦油产率	水产率	焦渣产率	正己烷可溶物产率
20131206	淮南煤(HN)	100	10.5737	1.3179	0.58		0.1246	0.0579	0.7599	
			10.2179	1.2832	0.59	0.4566	0.1256	0.0579	0.7587	0.0809
20140312	神木煤(SM)	100	10.3176	1.0082	0.94		0.0977	0.0914	0.7365	
			10.1371	1.0754	0.93		0.1061	0.0914	0.7281	
			9.2990	0.9803	0.85		0.1054	0.0914	0.7269	
			8.3511	0.8995	0.76	0.2990	0.1077	0.0914	0.7254	0.0719
20140105	内蒙褐煤(NM)	100	10.1726	0.4244	1.70		0.4244	0.1671	0.6244	
			10.1743	0.3613	1.70	0.0249	0.0355	0.1671	0.6288	0.0331
			10.1018	0.3670	1.69		0.3670	0.1671	0.6270	
20151015	黑山煤(HS)	100	8.7366	0.7328	0.79		0.0839	0.0904	0.7348	
			8.8229	0.7379	0.80	0.0775	0.0836	0.0904	0.7353	0.0749
20131123	棉杆(CS)	100	5.4439	0.9950	1.42		0.1828	0.2608	0.3139	
			5.5413	1.0075	1.45	0.4844	0.1818	0.2608	0.3140	0.0944
20151026	木屑(SD)	100	5.0725	1.2937	1.40		0.2550	0.2760	0.2805	
			5.1246	1.3254	1.41	0.8678	0.2586	0.2760	0.2779	0.0893
20131222	小球藻(GA)	100	5.3187	2.1076	0.65		0.3963	0.1222	0.2923	
			5.4069	2.1395	0.66	0.7096	0.3957	0.1222	0.2925	0.2645
20131212	稻壳(RH)	100	5.0970	0.9856	1.10		0.1934	0.2158	0.4086	
			5.3827	1.0548	1.16	0.4719	0.1960	0.2155	0.4001	0.1083
20131121	棉杆/淮南煤(CS/HN)	5/100	11.4190	1.4158	0.70		0.1240	0.0613	0.7362	
			12.0553	1.5529	0.74	0.4338	0.1288	0.0613	0.7355	0.0928
20131123		10/100	10.3167	1.3529	0.75		0.1311	0.0730	0.7168	
			10.2753	1.3481	0.75	0.4241	0.1312	0.0730	0.7229	0.0899
20131207		20/100	10.2329	1.2147	1.00		0.1187	0.0977	0.6860	
			10.3096	1.2000	1.00	0.3382	0.1164	0.0977	0.6900	0.0836
20131207		30/100	9.6363	1.1847	1.00		0.1229	0.1038	0.6759	
			9.9204	1.2475	1.03	0.3556	0.1258	0.1038	0.6772	0.0899
20131208		40/100	9.4570	1.2221	1.20		0.1292	0.1268	0.6317	
			9.4668	1.1912	1.20	0.3337	0.1258	0.1268	0.6295	0.0906
20131209		50/100	9.7279	1.2702	1.30		0.1306	0.1366	0.6090	
			9.9550	1.3127	1.33	0.4030	0.1319	0.1366	0.6087	0.0914

图 5-1：附件 1 的原数据部分展示

由图 5-1 可见，原数据中存在着部分缺失值，为了方便后续的数据处理以及模型建立分析研究，对原数据进行数据处理：将所有空白值补充为 0，同时将每种试样的不同实验数据的合并项拆分，最后得到经过处理后的数据，大致如下图所示：

时间	试样	配比	样品g	焦油(Char)g	水(Water)mL	正己烷不溶物（INS)g	焦油产率	水产率	焦渣产率	正己烷可溶物产率
20131206	淮南煤(HN)	100	10.5737	1.317900	0.580000	0.0000	0.124639	0.057944	0.759885	0.0000
20131206	淮南煤(HN)	100	10.2179	1.283200	0.590000	0.4566	0.125584	0.057900	0.758669	0.0809
20140312	神木煤(SM)	100	10.3176	1.008171	0.943029	0.0000	0.097714	0.091400	0.736547	0.0000
20140312	神木煤(SM)	100	10.1371	1.075369	0.926531	0.0000	0.106083	0.091400	0.728127	0.0000
20140312	神木煤(SM)	100	9.2990	0.980300	0.850000	0.0000	0.105420	0.091408	0.726852	0.0000
20140312	神木煤(SM)	100	8.3511	0.899509	0.763291	0.2990	0.107711	0.091400	0.725366	0.0719
20140105	内蒙褐煤(NI	100	10.1726	0.424401	1.700000	0.0000	0.424401	0.167116	0.624373	0.0000
20140105	内蒙褐煤(NI	100	10.1743	0.361274	1.700000	0.0249	0.035509	0.167100	0.628810	0.0331
20140105	内蒙褐煤(NI	100	10.1018	0.366989	1.688011	0.0000	0.366989	0.167100	0.626958	0.0000
20151015	黑山煤(HS)	100	8.7366	0.732800	0.790000	0.0000	0.083877	0.090400	0.734817	0.0000
20151015	黑山煤(HS)	100	8.8229	0.737910	0.800000	0.0775	0.083636	0.090400	0.735268	0.0749
20131123	棉杆(CS)	100	5.4439	0.995000	1.420000	0.0000	0.182773	0.260800	0.313911	0.0000
20131123	棉杆(CS)	100	5.5413	1.007500	1.450000	0.4844	0.181817	0.260800	0.314024	0.0944
20151026	木屑(SD)	100	5.0725	1.293700	1.400000	0.0000	0.255042	0.275998	0.280453	0.0000
20151026	木屑(SD)	100	5.1246	1.325410	1.414390	0.8678	0.258637	0.276000	0.277856	0.0893
20131222	小球藻(GA)	100	5.3187	2.107600	0.650000	0.0000	0.396262	0.122210	0.292289	0.0000
20131222	小球藻(GA)	100	5.4069	2.139500	0.660000	0.7096	0.395698	0.122200	0.292497	0.2645
20131212	稻壳(RH)	100	5.0970	0.985600	1.100000	0.0000	0.193369	0.215813	0.408613	0.0000
20131212	稻壳(RH)	100	5.3827	1.054800	1.160000	0.4719	0.195961	0.215505	0.400115	0.1083

图 5-2：处理后的部分热解统计数据

基于处理好的数据，进行数据描述性统计分析，详细结果如下图所示：

	时间	样品g	焦油(Char)g	水(Water)mL	正己烷不溶物（INS)g	焦油产率	水产率	焦渣产率	正己烷可溶物产率
count	1.350000e+02	135.000000	135.000000	135.000000	135.000000	135.000000	135.000000	135.000000	135.000000
mean	2.014103e+07	8.719148	1.038347	1.036556	0.142094	0.128227	0.122394	0.640920	0.046691
std	7.954360e+03	1.540324	0.316109	0.295224	0.195188	0.057044	0.040456	0.096857	0.087195
min	2.013022e+07	5.072500	0.361274	0.580000	0.000000	0.035509	0.057900	0.277856	0.000000
25%	2.013122e+07	7.587000	0.838238	0.850000	0.000000	0.106349	0.096012	0.610767	0.000000
50%	2.014032e+07	8.810600	0.999800	0.950000	0.000000	0.120000	0.111700	0.655395	0.000000
75%	2.015103e+07	9.974600	1.222482	1.159307	0.250650	0.132197	0.141400	0.704277	0.087650
max	2.015112e+07	12.055300	2.139500	1.909789	0.881100	0.424401	0.276000	0.759885	0.888900

图 5-3：数据描述性统计分析结果图

由图 5-3 可以得出，正己烷不溶物(INS)的平均克重约为 0.142g，标准差约为 0.195，这表明该数据分布具有一定程度上的波动性，后续分析需要考虑。同时也不难观察到，焦油产率、水产率和焦渣产率均呈现出不同水平的变异性。

为了便于本研究在后续的分析建模的过程中更好地掌握数据整体的特点的分布特征，绘制了如下的散点图来呈现焦油产率、水产率和焦渣产率这三者分别与正己烷不溶物(INS)质量的关系

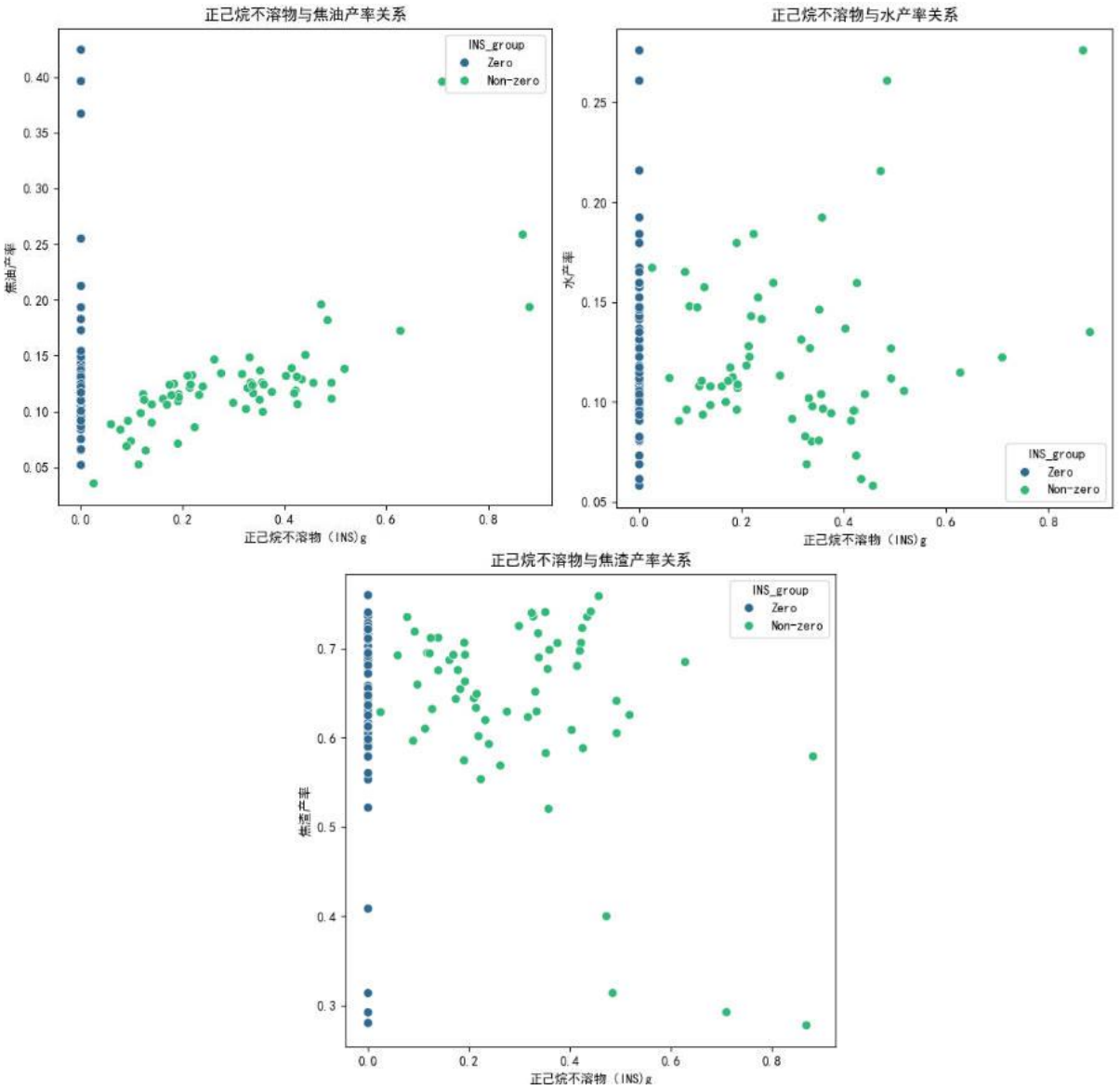


图 5-4：焦油产率、水产率和焦渣产率与正己烷不溶物(INS)质量的关系散点图

由图 5-4 得到以下分析结果：

- 正己烷不溶物(INS)质量与焦油产率之间的数据点分布整体呈现一个较为分散的状态，但是随着 INS 质量的增加，焦油产率整体上还是呈现着一个上升的趋势。
- 水产率与 INS 质量之间的数据点分布非常分散，关系较不明显。
- 焦渣产率与 INS 质量之间整体上近似趋于一个负相关的趋势，随着 INS 质量的增加，焦渣产率有一定水平的下降。

5.1.2 相关性分析

相关性分析是一种统计方法，用于衡量两个或多个变量之间的线性关系强度和方向。它是数据分析和数据科学中的一个重要工具，可以帮助我们理解变量之间的关联程度，从而为建立预测模型或做出决策提供依据。在本节中，利用基于皮尔逊相关系数[1]相关性分析方法分析焦油产率、水产率和焦渣产率这三者分别与正己烷不溶物(INS)质量的线性关系。建模过程大致如下：

1. 数据准备与相关变量定义：

首先，从实验结果中收集焦油产率、水产率、焦渣产率和正己烷不溶物(INS)质量的数据。这些数据应该涉及不同条件和比例下的共热解实验。定义以下变量：

X：正己烷不溶物(INS)的质量(g)

Y_1：焦油产率

Y_2：水产率

Y_3：焦渣产率

2. 计算相关性系数：

使用皮尔逊相关系数公式计算焦油产率、水产率、焦渣产率与正己烷不溶物(INS)质量之间的相关系数并以此来衡量其中的线性相关性。皮尔逊相关系数的计算公式如下：

$$r_{xy} = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

- n 是数据点的数量。
- $\sum \boldsymbol{xy}$是所有数据点x和x乘积的总和。
- $\sum \boldsymbol{xy}$ 和 $\sum \boldsymbol{y}$分别是变量x和y的总和。
- $\sum \boldsymbol{x^2}$和 $\sum \boldsymbol{y^2}$分别是变量x和y的平方和。

相关性系数分析原则如下：

- 相关性系数r_{xy} 的值介于 -1 和 1 之间。
- $r_{xy} = 1$ 表示完全正相关，即一个变量的增加与另一个变量的增加完全一致。接近于 1 表示强正相关。
- $r_{xy} = -1$ 表示完全负相关，即一个变量的增加与另一个变量的减少完全一致。接近于-1 表示强负相关
- $r_{xy} = 0$ 表示没有线性相关，但可能存在非线性关系。

经过上述公式的计算以及分析原则进行相关性分析，得到以下结果：

- 正己烷不溶物(INS)质量与焦油产率的相关性系数为 0.2091，表明二者之间存在着一定程度上的轻微的正相关关系。
- 正己烷不溶物(INS)质量与水产率的相关性系数为 0.0845，表明两者之间相关性关系较弱。
- 正己烷不溶物(INS)质量与焦渣产率的相关性系数为-0.1993，表明二者之间存在着一定程度上的轻微的负相关关系。

5.1.3 线性回归模型验证分析

经过 5.1.2 的相关性分析后，为了进一步验证热解过程中正己烷不溶物(INS)的质量对热解产率是否具有统计学定义上的显著性影响，本文基于此构建了一个线性回归模型进行线性回归分析[2]，进一步探究在控制好其他因素的前提下，正己烷不溶物(INS)的质量变化如何对焦油产率、水产率、焦渣产率产生影响。

线性回归是一种用于建立自变量（解释变量）和因变量（响应变量）之间关系的统计模型。线性回归的基本原理是通过寻找一条最佳拟合直线（在二维空间中）或超平面（在多维空间中），使得所有数据点到这条直线或超平面的垂直距离（即残差）之和最小。大致原理和建模流程如下：

1. 定义相关变量：

X：正己烷不溶物(INS)的质量比例

Y：热解产物产率（分为三个独立因变量，即焦油产率Y_{tar}，水产率Y_{water}，焦渣产率$Y_{residue}$）

2. **模型建立：**

构建 3 个线性回归模型如下：

$$Y_{tar} = aX + b + \epsilon$$
$$Y_{water} = mX + n + \epsilon$$
$$Y_{residue} = pX + q + \epsilon$$

其中，a, b, m, n, p, q 均为模型参数，需要通过数据进行拟合训练得到，ϵ代表模型本身的误差。

3. **参数拟合估计方法：**

计划采用最小二乘法[3]来拟合估计上述模型的各个参数，最小二乘法是一种常用的参数估计方法，用于找到线性回归模型中的最佳参数（斜率和截距），即本质目标是最小化误差ϵ的平方和，具体公式如下：

$$S(\beta) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

其中，$S(\beta)$ 是残差平方和，r_i是第i个观测值的残差，y_i是实际观测值，$\hat{y}_i$ 是模型预测值。

4. 模型效果验证分析：

- 使用R^2（决定系数）来具体评价模型的解释能力，决定系数（Coefficient of Determination），通常表示为 R^2，是一个统计量，用于衡量回归模型对因变量变异的解释程度。它告诉我们因变量的变异中有多少可以被自变量解释，从而评估模型的整体拟合优度[4] 。
- 同时使用 t 检验方法评估模型参数的显著性，t 检验是一种统计方法，用于确定两个样本均值之间的差异是否显著，或者一个样本均值与一个已知的总体均

值之间的差异是否显著。t 检验基于 t 分布，这是一种在样本大小较小或总体标准差未知时用于估计正态分布数据的分布

成功建立模型并应用求解后，线性回归结果数据如下表所示：

表 5-1：线性回归分析结果数据

	R^2	系数	P
焦油产率模型	0.044	0.0612	0.015
水产率模型	0.007	0.0175	0.370
焦渣产率模型	0.040	-0.0989	0.020

注：P 用于衡量假设检验的结果是否足够强烈地支持拒绝零假设，为 t 检验方法的结果

根据上表数据进行分析，**对于焦油产率模型**，$R^2 = 0.044$ 表示模型能解释焦油产率变异的 4.4%，这个水平较低，表明正己烷不溶物(INS)的质量变化对焦油产率的解释能力有限。系数=0.0612 表明正己烷不溶物(INS)的质量变化与焦油产率呈现正相关关系，即 INS 质量增加，焦油产率也会有上升的趋势。P=0.015，说明统计显著，这表明正己烷不溶物(INS)的质量变化对焦油产率存在显著性影响。

对于水产率模型，$R^2 = 0.007$，这个解释水平极低，表明正己烷不溶物(INS)的质量变化对水产率的影响非常小。系数=0.0175 表明正己烷不溶物(INS)的质量变化与水产率呈现正相关关系，即 INS 质量增加，水产率也有一定的上升的趋势。P=0.370，说明非统计显著，这表明正己烷不溶物(INS)的质量变化对水产率不存在显著性影响。

对于焦渣产率模型，$R^2 = 0.040$ 表示模型能解释焦渣产率变异的 4.0%，这个水平相对也较低，表明正己烷不溶物(INS)的质量变化对焦渣产率的解释能力有限。系数=-0.0989 表明正己烷不溶物(INS)的质量变化与焦渣产率呈现负相关关系，即 INS 质量增加，焦渣产率会有相反的下降的趋势。P=0.020，说明统计显著，这表明正己烷不溶物(INS)的质量变化对焦渣产率存在显著性影响。

总的来说，根据线性回归分析结果，正己烷不溶物(INS)的质量变化对焦油产率和焦渣产率有统计显著性的影响，而对水产率影响不显著。虽然由于R^2较低，正己烷不溶物(INS)的质量变化对和焦油产率和焦渣产率的解释能力有限，但显著的 P 值证明，INS 的变化对这些产率的改变有着一定的统计上的重要性。

为了进一步验证上述的结论，本文绘制相关的箱线图和趋势线散点图进行说明分析。

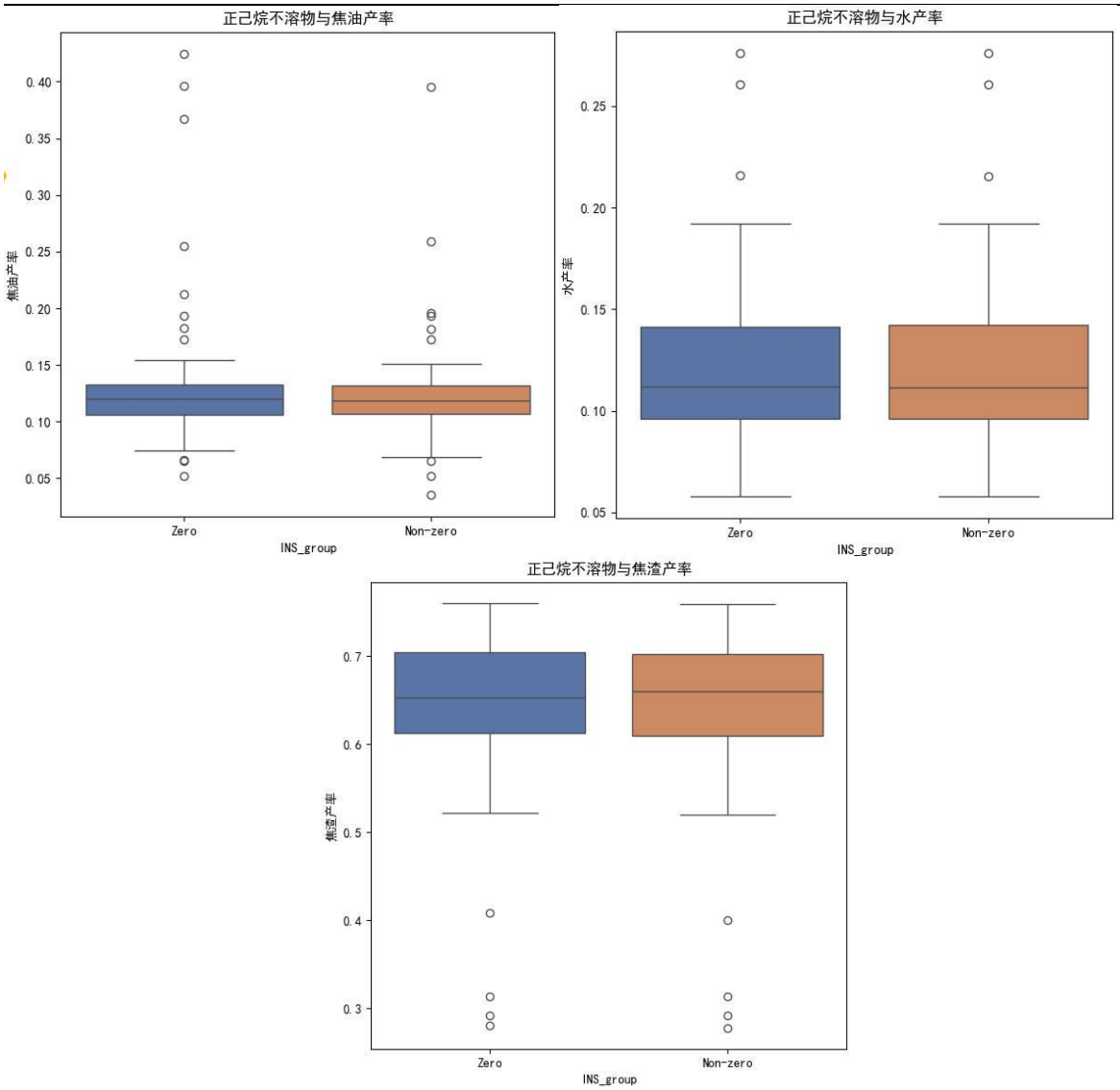


图 5-5：焦油产率、水产率和焦渣产率与正己烷不溶物(INS)质量的关系箱线图

由上图分析可知，对于焦油产率，在非零 INS 组中焦油产率的分布范围更广泛，表明 INS 的存在可能增加焦油产率，

而对于水产率，两组之间的差异不明显，这与上述的统计分析结论相吻合，即 INS 对水产率的影响不显著。

对于焦渣产率，非零 INS 组中焦渣产率的分布范围较低,中位数也略低，这上述与 INS 对焦渣产率有负相关关系的结论一致。

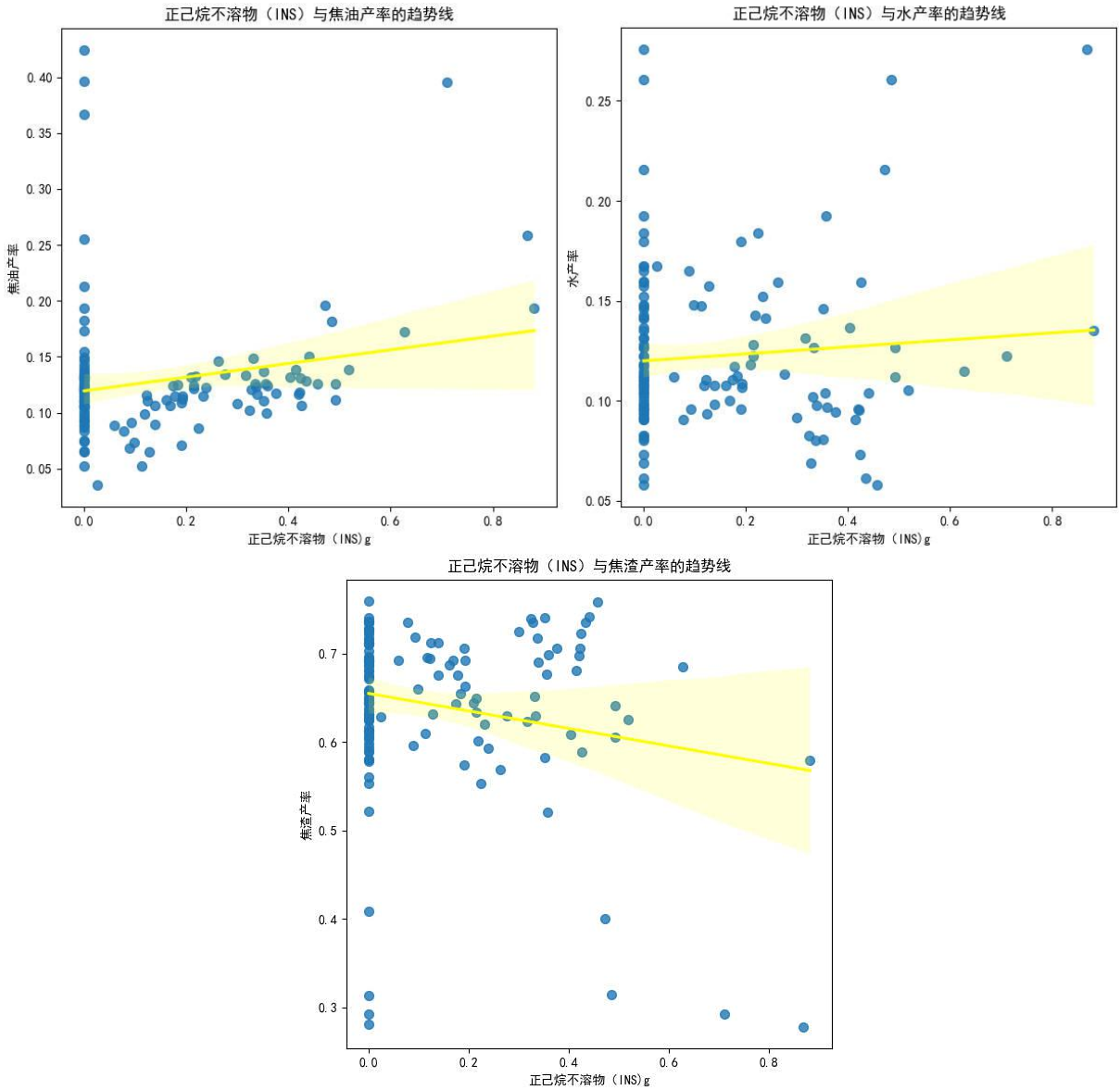


图 5-6：焦油产率、水产率和焦渣产率与正己烷不溶物(INS)质量的关系趋势线图

上述散点图和趋势线显示，针对焦油产率，随着 INS 含量的增加，焦油产率存在较小的上升趋势。这表明存在 INS 含量的提高促进了焦油的产生的可能性。

针对水产率，趋势线相对平缓，INS 对水产率的影响非常有限，这进一步验证了上述分析中的发现。

对于焦渣产率，随着 INS 含量增加，焦渣产率呈现下降趋势,这与上述回归分析的结论相符,显示出 INS 的含量提高可能对减少焦渣产率具有一定的作用。

5.2 问题 2 模型建立与求解

5.2.1 交互效应分析

在本节中，需分析正已烷不溶物和混合比例的交互效应对热解产物产量的影响。本文使用统计建模的方法来识别并量化这些交互效应，便于研究人员理解。

首先进行基本的数据处理，数据集需要包括热解实验中的各种相关变量，如正已烷不溶物的质量（INS_g），生物质与煤的混合比例以及各种热解产物的产率。基于题目所述，为了方便模型的拟合训练，将所有的配比转化成实际的比例值用于后续的分析，例如配比 100 表示一个成分在混合物中的比例为 100%，而 5/100 表示在 105 份样品中，生物质占 5 份，煤占 100 份。

为了评估 INS 与配比之间的交互效应，采取多元线性回归模型进行评估分析，模型建立公式如下：

$$Y = m * I + n * proportion + p * (I \times proportion) + q + \epsilon$$

其中：

- Y是热解产物产率，可以为焦油产率Y_{tar}、水产率Y_{water}或焦渣产率$Y_{residue}$
- I 是正已烷不溶物的质量
- $proportion$是生物质和煤的混合比例
- m, n, p, q都是模型参数
- ϵ代表模型本身的误差。

通过拟合求解上述模型，可以评估正已烷不溶物（INS）的质量变化和配比的主效应及其交互效应的统计显著性, 显著的交互效应 INS 的质量或可理解为存在程度会影响配比对热解物产率的效果。同时根据模型的输出，可以进一步得到哪些热解产物的产率受到 INS 存在程度和配比交立效应的显著影响。例如，如果n显著不为零，表明存在显著的交互效应。最后为了更直观清晰地展示交互效应，本文绘制了相关的散点图和热图以描述不同的 INS 质量和配比下热解产物产率的变化。

经过计算后得到结果，在此展示部分结果，如下图：

OLS Regression Results

Dep. Variable:	焦油产率	R-squared:	0.286
Model:	OLS	Adj. R-squared:	0.270
Method:	Least Squares	F-statistic:	17.49
Date:	Sun, 12 May 2024	Prob (F-statistic):	1.31e-09
Time:	06:36:00	Log-Likelihood:	218.31
No. Observations:	135	AIC:	-428.6
Df Residuals:	131	BIC:	-417.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0981	0.007	13.383	0.000	0.084	0.113
INS_g	0.0174	0.034	0.510	0.611	-0.050	0.085
配比	0.0835	0.017	4.965	0.000	0.050	0.117
INS_g:配比	0.0611	0.060	1.018	0.311	-0.058	0.180

Omnibus:	73.514	Durbin-Watson:	1.900
Prob(Omnibus):	0.000	Jarque-Bera (JB):	514.774
Skew:	1.749	Prob(JB):	1.65e-112
Kurtosis:	11.904	Cond. No.	16.9

图 5-6：计算结果部分展示

相关变量分布的散点图如下图所示：

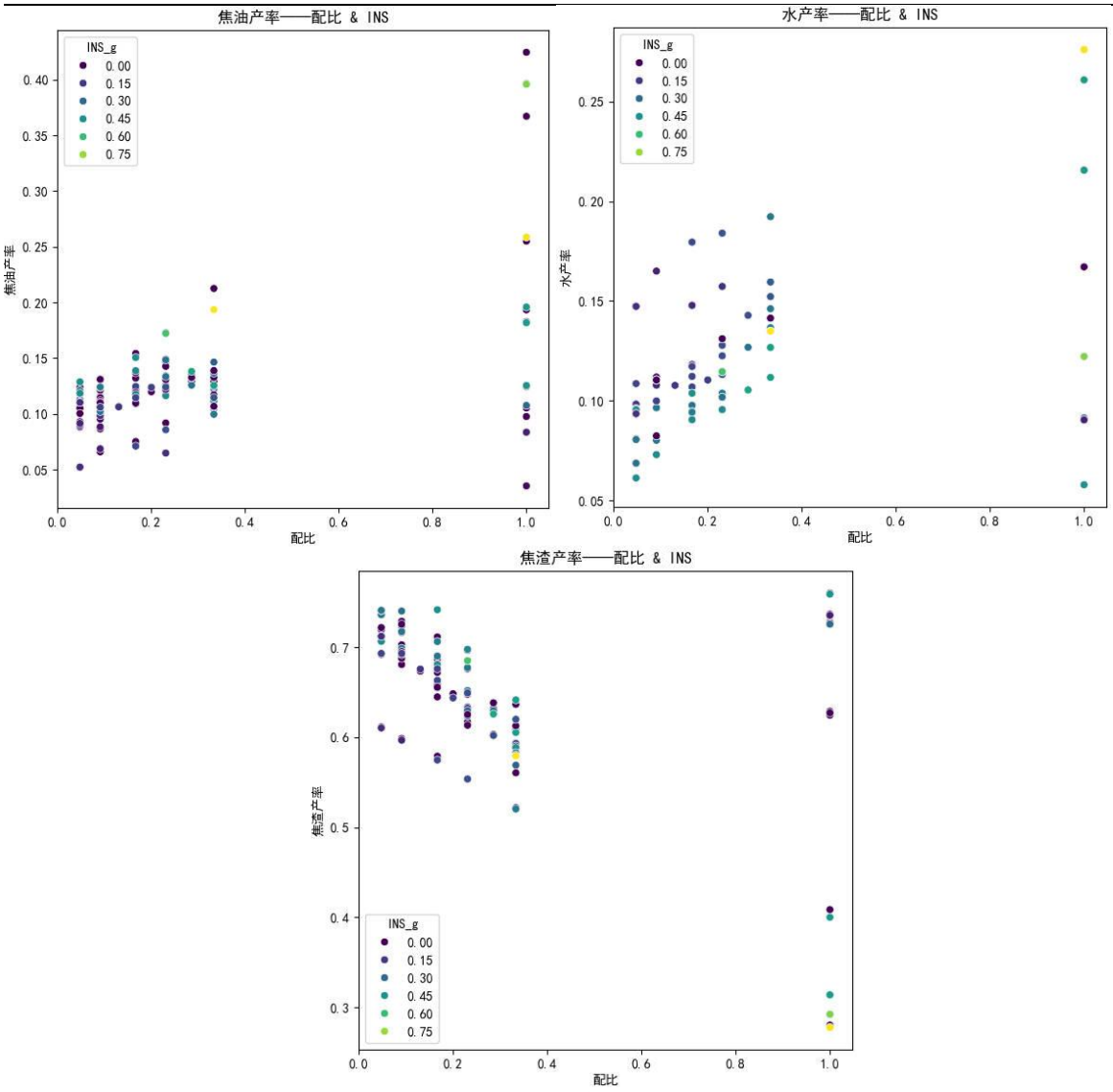
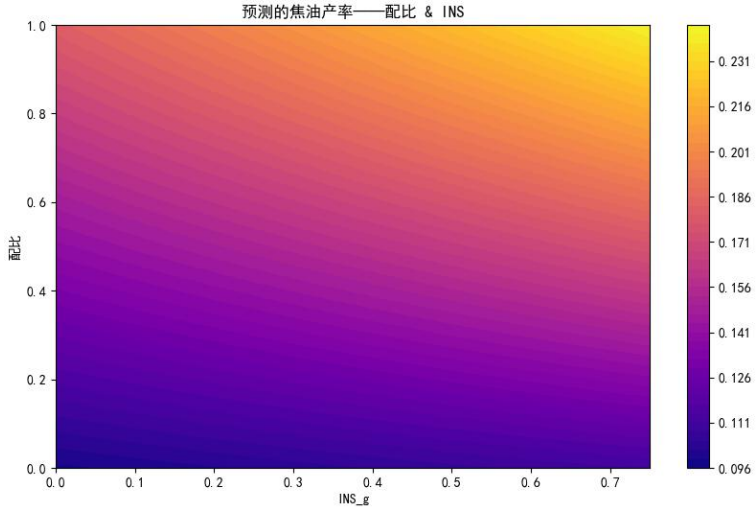


图 5-7：相关变量分布的散点图

同时，基于模型预测结果，绘制如下的等高线热图：



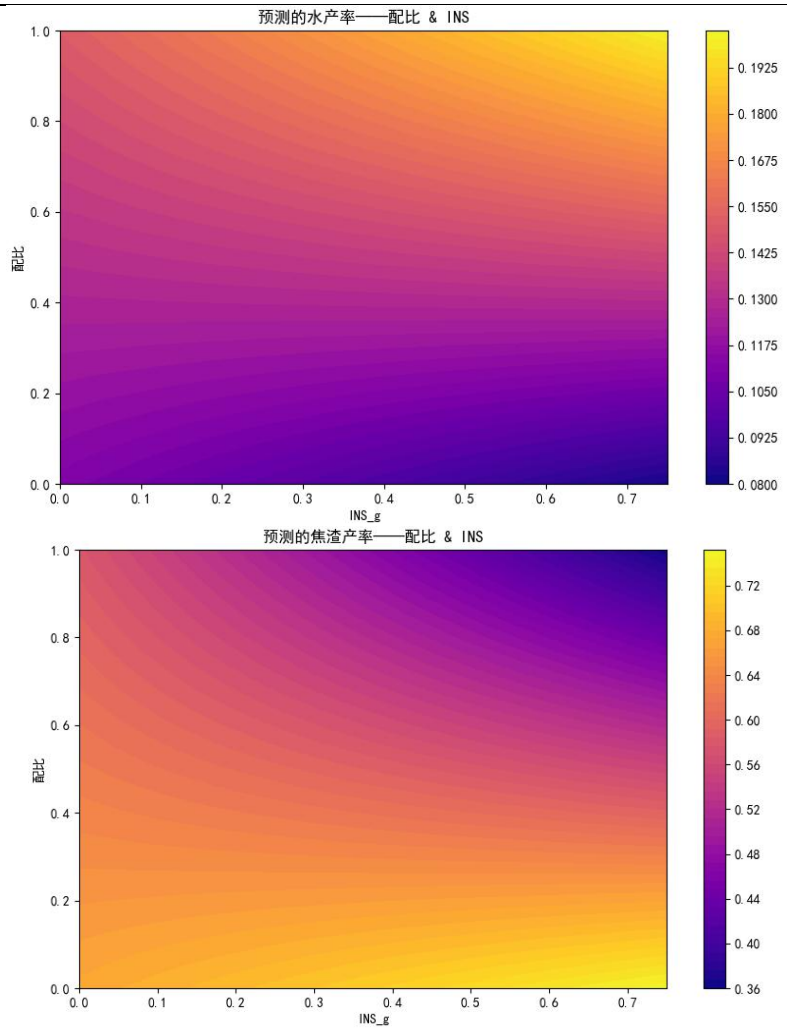


图 5-8：基于模型预测结果的等高线热图

基于上述散点图和等高线热图，可以得知，随着混合比例的增加，焦油产率和水产率呈现出上升趋势，特别是在低正己烷不溶物浓度区间内，这种上升趋势尤为显著。相反，焦渣产率随着混合比例的增加而下降，特别是在高正己烷不溶物浓度条件下，这种下降趋势更为明显。

5.2.2 显著性分析

基于 5.2.1 多元线性回归模型的求解结果，统计相关变量的系数和对应P值，记录在下表中：

表 5-2：多元线性回归分析数据

	$proportion$	P_1	$I \times proportion$	P_2
焦油产率	0.0174	0.001	0.0611	0.311
水产率	0.0389	0.002	0.1064	0.019
焦渣产率	-0.0969	0.001	-0.3849	0.0007

基于上表数据，可得到以下结论：

- **针对焦油产率**，P=0.311，正己烷不溶物（INS）与配比之间的交互效应不显著，但是混合比例对其有着显著正影响，即焦油产率随混合比例的增加而提高。
- **针对水产率**，P=0.019，$I \times proportion$的对应系数=0.1064，这表明正己烷不溶物（INS）与配比之间的交互效应对水产率存在着一定的显著的正向影响。具体而言，当正己烷不溶物（INS）的质量变化时，混合比例对水产率的影响也会有所变化，二者存在一定的协同作用。
- **针对焦渣产率**，$I \times proportion$的对应系数=-0.3849，P=0.001，正己烷不溶物（INS）与配比之间的交互效应存在显著负影响，即随着混合比例的增加，焦渣产率的下降会更加明显。同时，混合比例对焦渣产率也有显著负影响，即焦渣产率随混合比例的增加而下降。

由此不难得出，**交互效应最为明显的热解产物**为焦渣和水，前者体现为一个显著的负向影响，即在已有正己烷不溶物（INS）的情况下，随着混合比例的增加，焦渣产率的下降会更加显著。后者体现为一个显著的正向影响，即在已有正己烷不溶物（INS）的情况下，随着混合比例的增加，水产率的上升会更加显著。

相反，不显著的交互效应所对应的热解产物为焦油，虽然混合比例对焦油产率存在一定的显著影响，但正己烷不溶物（INS）与配比之间的交互效应对其的影响并不显著。

总而言之，在热解过程中，焦渣产率受到正己烷不溶物含量和混合比例的交互效应影响最为显著，这种交互效应在统计上表现出极高的显著性。这表明这两个因素的组合对焦渣产率的影响至关重要。相比之下，水产率虽然也受到这些因素交互作用的影响，但其在热解过程中的重要性略低于焦渣产率。这些发现对于优化热解过程的参数设置具有重要意义，有助于实现更高效的能源转化和提高产物的利用率。

5.3 问题 3 模型建立与求解

5.3.1 建立优化模型

基于题目要求，需要构建一个优化模型即一个目标函数，通过优化生物质和煤的混合比例，从而最大化热解产物中的焦油产率。其中，尽可能地尝试将生成焦油的质量接近于纯煤热解时产生的焦油，以便后续提高工作效率和节约处理成本。

下面是基于本问题的大致建模步骤：

1. **数据处理：**
   将配比数据的比例转化为具体数值，如 5/100 转换为 0.048，同时进一步清洗和准备数据用于后续的模型求解。
2. **确定变量关系：**
   Y：焦油产率
   X：混合比例
   I：正己烷不溶物（INS）的质量
   最后将目标函数形式化为$Y = F(I, X)$，其中F可看作是由数据驱动的回归模型
3. **具体模型建立：**
   本文采取多元线性回归模型来拟合焦油产率与正己烷不溶物（INS）的质量和混合比例的关系，具体模型构造如下：

$$Y = m * I + n * X + p * (I \times X) + q + \epsilon$$

其中：m, n, p, q都是模型参数，ϵ代表模型本身的误差。

4. **参数估计方法：**
   通过最小化预测值和模型预测之间的平方误差来估计参数
5. **定义目标函数：**
   最大化焦油产率，即求解：$MaxX\ s.t.Y = F(I,X)$，其中I应为一个固定的正己烷不溶物（INS）的质量，X即为自变量是一个可调整的混合比例，在 0—1 之间波动。

5.3.2 优化算法求解

在本小节中，为了求解 5.3.1 中提出的目标函数，本研究采用了模拟退火算法优化方法，模拟退火算法[5-8]（Simulated Annealing，SA）是一种通用概率算法，用来在一个大的搜寻空间内找寻问题的近似最优解，它是一种启发式算法模拟退火算法最早的思想是由 N.Metropolis 等于 1953 年提出。1983 年，S.Kirkpatrick 等成功地将退火思想引入组合优化领域。它是基于 Monte-Carlo 迭代求解策略的一种随机寻优算法，其出发点基于物理中固体物质的退火过程与一般组合优化问题之间的相似性。模拟退火算法从某一较高初温出发，伴随温度参数的不断下降，结合概率突跳特性，在解空间中随机寻找目标函数的全局最优解，即局部最优解能概率性地跳出，并最终趋于全局最优。该算法具有概率的全局优化性能，目前已在工程中得到了广泛应用，如 VLSI（超大规模集成电路）、生产调度、控制工程、机器学习、神经网络、信号处理等领域。[8]

模拟退火算法是通过赋予搜索过程一种时变且最终趋于零的概率突跳性，从而可有效避免陷入局部极小，并最终趋于全局最优的串行结构的优化算法。该算法还具有较强的鲁棒性、全局收敛性、隐含并行性及广泛的适应性，并且能处理不同类型的优化设计变量（离散的、连续的和混合型的），无须任何辅助信息，对目标函数和约束函数没有任何要求。它利用 Metropolis 算法，并适当地控制温度下降过程，在优化问题中具有很强的竞争力。

下面是结合模拟退火算法对本问题进行求解的详细步骤与分析：

1. **初始化：**随机创建一个初始解集，每个解代表一种参数方案。定义初始能量即初始的参数、标准偏差和方差、初始温度和退火速率，同时将问题转化为最小化-Y。
2. **模拟退火过程（邻域搜索）：**在当前温度下，在每一步迭代中随机选择一个或多个班次，并尝试改变参数。这个过程称为邻域搜索，生成的解称为邻域解。生成当前解的一个邻域解，即对当前解进行微小的随机扰动。
3. **接受准则：**计算当前解和邻域解的目标函数值。计算目标函数值的差值，并根据接受概率公式决定是否接受邻域解。如果邻域解的目标函数值优于当前解则接受这个新解作为当前解。如果邻域解的目标函数值劣于当前解,以一定的概率接受这个劣解，这个概率由当前温度和解的差异决定，遵循玻尔兹曼分布:

$$P(accept) = exp\left(-\frac{\Delta E}{k_B T}\right)$$

   其中，ΔE是新解和当前解的能量差即目标函数数值差，k_B是玻尔兹曼常数，T是当前温度。重复上述步骤，直到满足停止条件，如达到预定的迭代次数或目标函数值收敛。
4. **温度下降：**每次迭代后，根据退火速率降低温度。随着温度的降低，本质上劣解接受概率逐渐减小，算法逐渐收敛到最优解。
5. **结果分析：**输出最优解，即目标函数值最大的解，作为最大化焦油产率的混合比例。分析结果，评估其在实际操作中的可行性和效果。

6. **调整参数**：根据实验结果调整模拟退火算法的参数，如初始温度、退火速率等，以获得更好的优化效果。

针对本次问题，设置迭代次数为 1000 次，求解过程中适应度迭代图如下图所示：

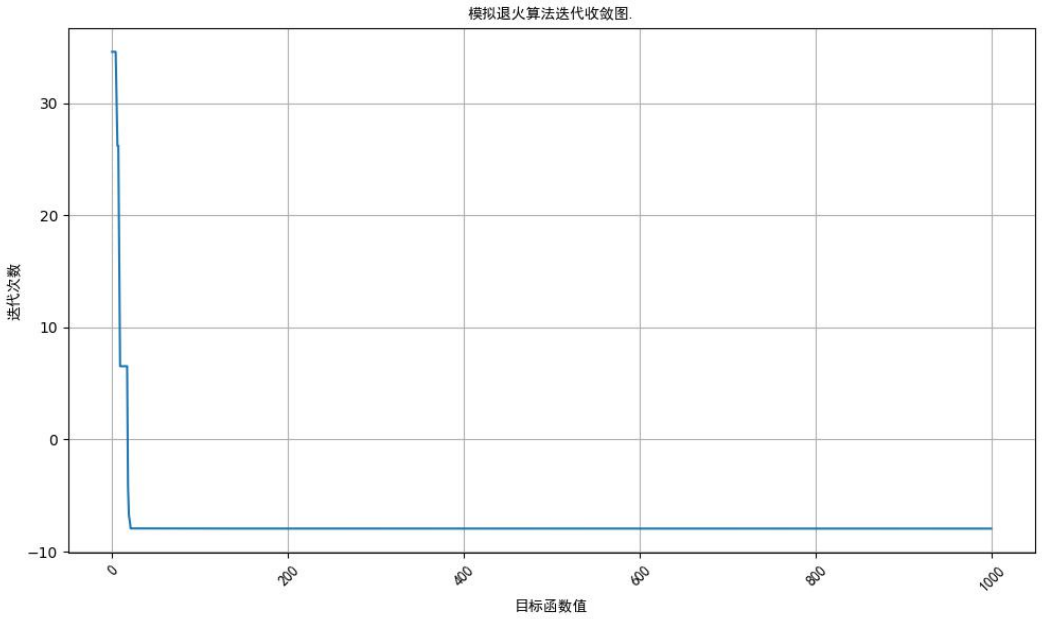


图 5-9：模拟退火算法优化迭代收敛图

从迭代图中得出，图中的纵坐标通常代表目标函数的值，而横坐标代表迭代次数或时间，可以看到，算法在第 40 次左右迭代时，已经达到了个相对稳定的状态，几乎没有进一步的变化，这表明算法已经接近收敛，找到了最优解或者是一个非常接近的解。说明模拟算法成功地对问题进行了优化，并且很可能找到了一个质量较高的解。这个平稳的收效表明了算法的稳定性和有效性。

求解结果为，最优配比：0.9999940391，预测的最大焦油产率：0.0927888，这说明若需要最大化焦油产率，几乎完全依赖于煤的热解，这与题目说明中焦油产率关于更倾向于煤焦油的质量的要求相符合在最优条件下，焦油的产率约为 9.28%。

经过启发式算法的优化分析，基本确定当混合比例接近纯煤状态时，焦油产率可以达到最高，并且焦油的质量与纯煤热解产生的焦油质量相近。这一结果与我们的目标一致，即在提高焦油产率的同时，利用煤的高纯度特性，并减少因生物质混合带来的杂质和处理成本。

据此分析，得出此结论：为了最大化焦油产率和质量，应当增加煤在混合物中的比例。这一结论为共热解过程的操作条件提供了理论依据和操作指导。

5.4 问题 4 模型建立与求解

5.4.1 数据预处理

为了便于后续的观察分析以及模型构建，基于题目附件二的数据，展示部分数据如下表：

表 5-3： 棉杆/淮南煤(CS/HN):热解产物收率实验数据与理论计算值(wt%, daf)

CS/HN	100/0	0/100	5/100	10/100	20/100	30/100	50/100
Tar	19.46	15.73	17.46	16.82	15.54	16.33	16.56

Calculated tar	—	—	15.97	16.17	16.51	16.78	17.21
HEX	10.74	11.33	12.58	13.02	11.16	11.67	11.48
Calculated HEX	—	—	11.34	11.27	11.22	11.17	11.1
Water	26.84	4.82	5.39	7.44	10.01	10.6	14.06
Calculated water	—	—	6.17	7.37	9.39	11.03	13.76
Char	29.21	70.77	67.01	65.06	61.66	60.55	53.46
Calculated char	—	—	68.79	66.99	63.84	61.18	56.92

表 5-4： 木屑/黑山煤(SD/HS):热解产物收率实验数据与理论计算值(wt%, daf)

SD/SM	100/0	0/100	5/100	10/100	20/100	30/100	50/100
Tar	27.29	11.78	12.77	12.99	14.86	15.78	15.87
Calculated tar	—	—	12.52	13.19	14.37	15.36	16.95
HEX	9.43	8.18	10.08	10.97	11.78	11.97	12.46
Calculated HEX	—	—	8.24	8.29	8.39	8.47	8.6
Water	23.8	6.1	7.95	8.06	8.76	10.07	13
Calculated water	—	—	6.94	7.71	9.05	10.18	12
Char	28.95	73.92	69.49	69.11	64.56	61.82	57.44
Calculated char	—	—	71.78	69.83	66.43	63.54	58.93

将所有数据整合合并处理为以下形式：

CS/HN	5/100	10/100	20/100	30/100	50/100
HEX	12.58	13.02	11.16	11.67	11.48
Calculated HEX	11.34	11.27	11.22	11.17	11.1
Water	5.39	7.44	10.01	10.6	14.06
Calculated water	6.17	7.37	9.39	11.03	13.76
Char	67.01	65.06	61.66	60.55	53.46
Calculated char	68.79	66.99	63.84	61.18	56.92
CS/SM	5/100	10/100	20/100	30/100	50/100
Tar	11.86	12.02	13.73	13.45	13.56
Calculated tar	12.17	12.52	13.13	13.64	14.45
HEX	9.13	9.67	11.13	10.55	9.89
Calculated HEX	8.3	8.42	8.63	8.8	9.07
Water	6.68	8.06	8.65	10.47	12.08
Calculated water	7.14	8.09	9.74	11.11	13.3
Char	71.7	70.74	65.52	62.72	58.6
Calculated char	71.79	69.86	66.47	63.6	59.02
CS/HS	5/100	10/100	20/100	30/100	50/100
Tar	10.18	9.97	10.15	9.26	9.35
Calculated tar	9.51	9.96	10.75	11.42	12.49
HEX	8.22	9.28	6.78	6.74	5.38
Calculated HEX	8.21	8.32	8.52	8.69	8.97
Water	6.99	8.58	9.73	11.19	14.5

图 5-10：数据整合处理后结果展示

5.4.2 配对样本t检验

在本小节中，采用配对样本 t 检验[9]来深入分析每种中间产物的实验值与理论计算值之间是否存在着显著性差异。配对样本 t 检验，也称为成对 t 检验或相关样本 t 检验，是一种用于比较两个相关样本均值差异的统计方法。这种检验适用于同一组受试者在不同条件或时间点的测量，或者是在两个相关样本中的相同个体在两个不同条件下的测量。下面是该方法的基本原理和相关说明：

1. **假设检验：**
   - 零假设（H0）：两个相关样本的均值相等。
   - 备择假设（H1）：两个相关样本的均值不相等。
2. **计算方法：**

   首先，计算每个配对样本的差值（d）。然后，计算差值的平均值（$d^{'}$）和标准差（s_d）。最后，计算 t 统计量：

   $$t = \frac{d^{'}}{s_d/\sqrt{n}}$$

   其中，n是配对样本的数量。
3. **p值和决策：**

   根据 t 统计量和自由度$(n-1)$，在 t 分布表中查找相应的临界值，或者计算p值。
   - 如果p值小于显著性水平（如 0.05），则拒绝零假设，认为两个样本的均值存在显著差异。
   - 如果p值大于显著性水平，则无法拒绝零假设，认为两个样本的均值没有显著差异。
4. **配对样本 t 检验的应用**
   - 比较同一组受试者在两种不同条件下的反应。
   - 比较同一组受试者在两个不同时间点的反应。
   - 比较同一产品在不同批次或不同处理条件下的性能。
5. **配对样本 t 检验的假设**
   - 数据是正态分布的。
   - 差值（d）是正态分布的。
   - 差值（d）的方差是恒定的，即各对之间的差异是相同的。

使用上述方法对数据进行检验，检验结果部分展示如下：

Combination	Product	T-Statistic	P-Value
CS/HN	Tar	0.030523	0.977112
CS/HN	HEX	2.354455	0.078134
CS/HN	Water	-0.175097	0.869509
CS/HN	Char	-4.411591	0.011586
CS/SM	Tar	-1.052632	0.351895
CS/SM	HEX	4.508776	0.010750
CS/SM	Water	-3.187304	0.033302
CS/SM	Char	-0.878522	0.429266
CS/HS	Tar	-1.485038	0.211710

图 5-11：配对样本 t 检验结果部分展示

(注：Tar 代表焦油，Water 代表水，Char 代表焦渣，HEX 代表正己烷可溶物，下同)

对于所有的检验结果，筛选出所有存在显著性差异即 P 值<0.05 的检验结果，如下图所示：

	Combination	Product	T-Statistic	P-Value
3	CS/HN	Char	-4.411591	0.011586
5	CS/SM	HEX	4.508776	0.010750
6	CS/SM	Water	-3.187304	0.033302
10	CS/HS	Water	3.873166	0.017945
11	CS/HS	Char	-5.024556	0.007362
13	SD/SM	HEX	8.511092	0.001045
15	SD/SM	Char	-6.231656	0.003378
17	SD/HS	n-hexane soluble(HEX)	3.661904	0.021544
26	GA/HN	Water	5.385132	0.005749
27	GA/HN	Char	-20.499814	0.000033
32	RH/HN	Tar	-8.983966	0.000850
33	RH/HN	HEX	-11.196172	0.000362
34	RH/HN	Water	6.913362	0.002297
37	RH/SM	HEX	5.563402	0.005112
39	RH/SM	Char	-3.374359	0.027931

图 5-12：存在显著性差异的数据

5.4.3 子组分析

在本小节中，根据 5.4.2 节中得出的所有显著性差异检验结果，会对实验值与理论计算值之间的差异进行子组分析，以识别在哪些混合比例下这些差异最为显著。需要计算每个显著差异产物的实验值与理论值在每个混合比例上的差异，并评估这些差异的显著程度。经过实际计算后得到结果部分展示如下：

	Combination	Product	Mixture	Actual	Calculated	Difference
0	CS/HN	Char	5/100	67.01	68.79	-1.78
1	CS/HN	Char	10/100	65.06	66.99	-1.93
2	CS/HN	Char	20/100	61.66	63.84	-2.18
3	CS/HN	Char	30/100	60.55	61.18	-0.63
4	CS/HN	Char	50/100	53.46	56.92	-3.46
...	...	...	...	...	...	...
70	RH/SM	Char	5/100	71.76	71.89	-0.13
71	RH/SM	Char	10/100	69.02	70.04	-1.02
72	RH/SM	Char	20/100	66.49	66.80	-0.31
73	RH/SM	Char	30/100	63.33	64.06	-0.73
74	RH/SM	Char	50/100	59.22	59.68	-0.46

图 5-13：子组分析后的检验数据汇总

这些数据揭示了实验测量值与理论预测值之间差异最大的混合比例。例如对于 CS/HN 组合的焦渣产物，不难发现其在 50/100 的比例上的差异最大，高达-3.46。

在实际操作中，往往关注这些最大差异，以便于对工艺进行优化和调整。为此，针对每个数据组合，直接指出在各个混合比例中差异最大的前三个，这样的信息更加便于理解和应用。通过这种方式，可以快速确定哪些比例需要特别注意，并可能需要进行调整。通过提取每个组合中差异最大的前三个的混合比例，并进行了可视化展示，使得实验值与理论值之间显著差异的观察更加直观明了。如下图所示：

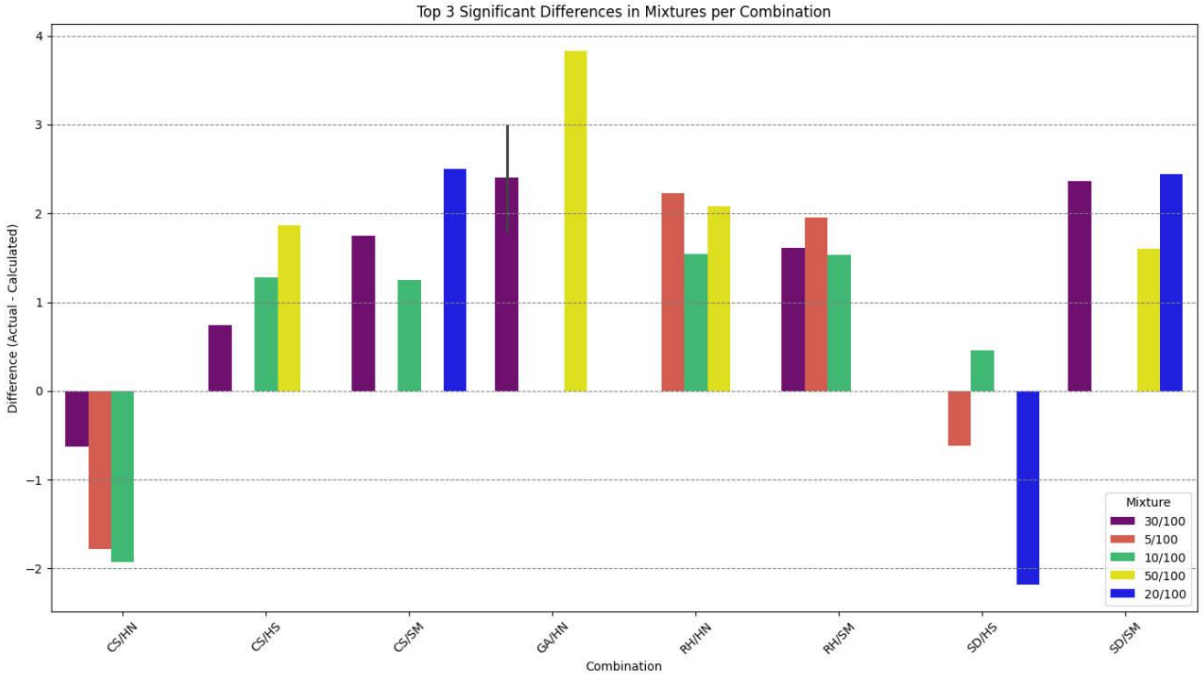


图 5-14：前三个差异最大的混合比例

5.5 问题 5 模型建立与求解

在本节中，针对问题 5 的需求，需要构建一个能够预测热解产物产率的数学模型，采用数据驱动的方法，结合统计学习理论来建立和优化机器学习算法模型。

5.5.1 模型建立与评估选择

首先，针对用于本次模型训练的数据，需要一定的数据理解和预处理。数据包含以下主要变量，需要理解数据中的每个变量及其对热解产物产率的潜在影响:

- 试样类型：不同的生物质和煤炭类型。
- 配比：生物质与煤的混合比例。
- 样品质量：实验中使用的样品总质量。
- 焦油(Char)g：实验中产生的焦油质量。
- 水(Water)mL：实验中产生的水量。
- 正已烷不溶物（INS）g：正已烷不溶物（INS）的质量

最后将分类数据转数值(如试样类型编码).

为了能够更好的预测热解产物产率，选择随机森林回归模型和 XGBoost 回归模型进行各自训练和预测，筛选出效果性能最好的模型算法。

对于模型的训练和验证，使用交叉验证——K 折交叉验证来评估模型的泛化能力和稳定性，同时利用随机搜索方法对模型超参数进行优化，以达到最好的预测效果。

最后针对模型的评估，选择使用均方误差 MSE（衡量模型精确度，越小越精确）和决定系数R^2（衡量模型对数据变异的解释能力，越接近于 1 越好）评估模型。

5.5.2 机器学习模型理论

随机森林（Random Forest）是一种集成学习方法，由 Leo Breiman 等人于 2001 年提出。[10]它通过构建多个决策树来预测目标变量，并利用这些决策树的结果进行投票或取平均值来提高预测的准确性。随机森林算法结合了多个树的预测能力，从而降低了过拟合的风险，提高了模型的稳定性和泛化能力。其基本原理如下：

1. **构建决策树**：随机森林由多个决策树组成，每个决策树用于预测目标变量。决策树是一种树形结构，它从根节点开始，每个节点包含一个特征和一个阈值，用于分割数据。
2. **随机选择特征**：在构建每个决策树时，随机森林算法从所有特征中随机选择一部分特征来分割数据。这种随机选择特征的方法可以增加模型的多样性，有助于防止过拟合。
3. **重复构建树**：随机森林算法重复构建多个决策树，每次构建时都从所有特征中随机选择特征。这样可以增加模型的多样性，提高模型的泛化能力。
4. **预测结果**：随机森林算法利用所有决策树对每个样本进行预测，然后根据投票或取平均值的方法来确定最终预测结果。对于分类问题，通常采用投票法；对于回归问题，通常采用取平均值的方法。

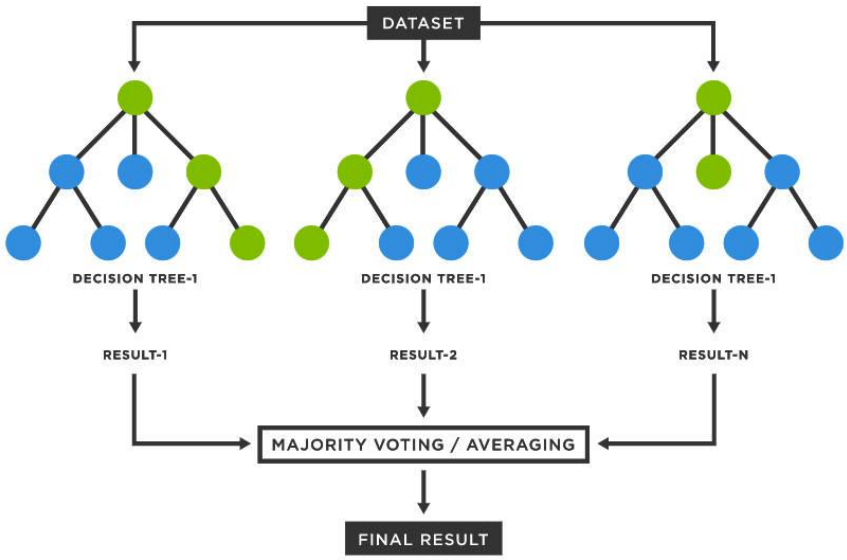


图 5-15：随机森林算法的结构图

XGBoost 是一个优化的分布式梯度增强库，旨在提供高效、灵活且可扩展的梯度提升框架。它是由 Tianqi Chen 在 2014 年开发的，受到他在卡内基梅隆大学（CMU）的实习经历和他在华盛顿大学的博士研究的影响,在研究机器学习算法的效率和可扩展性时，为了解决 Gradient Boosting 实现中的计算瓶颈和扩展性问题而创建的[11]。XGBoost 的核心算法是基于决策树的梯度提升框架。它通过不断地添加新的决策树来拟合上一次迭代的残差，最终将所有决策树的结果加权求和得到最终的预测结果。每棵决策树都是在一个新的子空间上学习，这样可以捕捉到数据中的复杂关系。XGBoost 通过在每次迭代时优化目标函数，不断添加新的决策树来构建一个强学习器。它的优势

在于能够自动处理缺失值，支持并行计算，提供了多种防止过拟合的技术，并且可以通过交叉验证来调整模型参数。下面是其计算流程以及相关公式的简单描述：

1. **初始化**：设定初始预测值 $F_0(x)$。在回归问题中，通常设为训练集标签的均值。
2. **迭代构建弱学习器**：（对于 $m = 1,2,\ldots,M$）：
   a. 计算残差：对于每个样本 ii，计算残差r_{mi}：
   $$r_{mi} = y_i - F_{m-1}(x_i)$$
   其中 y_i是真实标签，$F_{m-1}(x_i)$ 是上一轮的预测值。
   b. 训练弱学习器：使用残差r_{mi}作为标签，训练一个新的决策树 $h_m(x)$，用于拟合残差。
   c. 更新预测：$F_m(x) = F_{m-1}(x) + \eta h_m(x)$，其中 η是学习率。
3. **计算损失函数**：模型使用的是损失函数的泰勒二阶展开来优化模型。对于回归问题，常用的损失函数是平方损失
   $$L\big(y, F(x)\big) = \big(y - F(x)\big)^2$$
   其泰勒展开为：
   $$L\big(y, F(x)\big) = \big(y - F(x)\big)^2 = \big(y - F(x) - h(x)\big)^2 + 2\big(y - F(x)\big)h(x) + h(x)^2$$
   其中，第一项是常数项，第二项是关于$h(x)$的一次项，第三项是关于$h(x)$的二次项。
4. **定义目标函数**：XGBoost 的目标函数由损失函数和正则化项组成：
   $$\text{Obj} = \sum_i L\big(y_i, F(x_i)\big) + \sum_m \Omega(h_m)$$
   其中，$\Omega(h_m)$是正则化项，用于控制模型的复杂度。对于决策树，正则化项通常包括树的叶子节点数量和叶子节点的分数的平方和。
5. **优化目标函数**：对于每个决策树 $h_m(x)$，XGBoost 通过贪心算法选择最优的分割点，以最小化目标函数。具体的优化过程涉及到数值优化和梯度计算。
6. **输出最终模型**：最终的模型函数是所有弱学习器的加权和
   $$F(x) = \sum_{m=1}^{M} \eta h_m(x)$$

下面是模型的基本结构图：

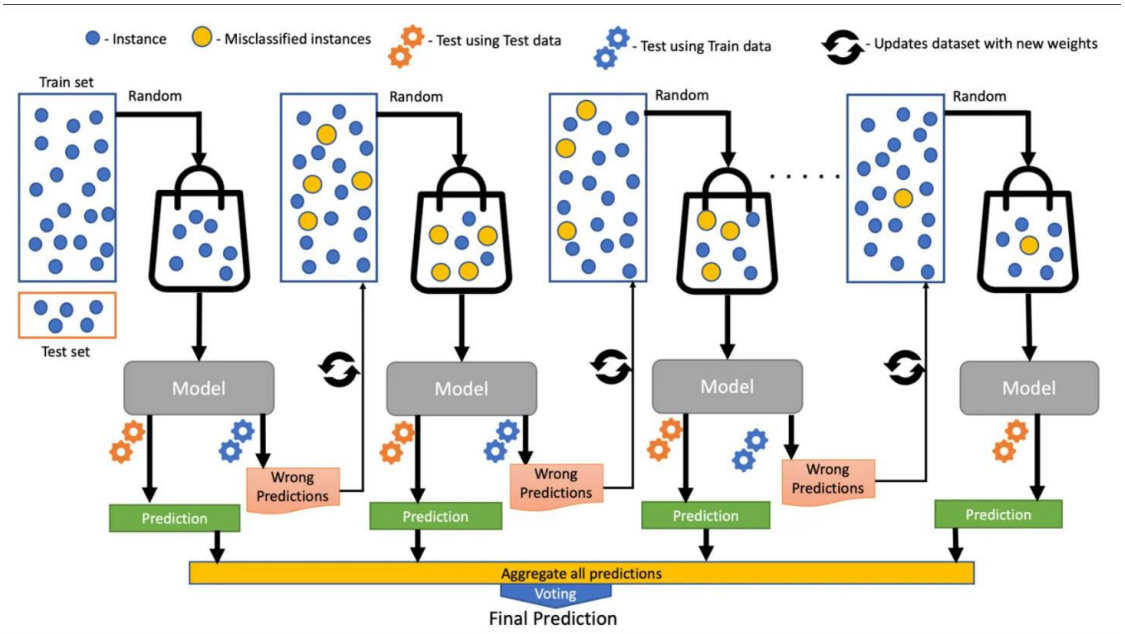


图 5-16：XGBoost 模型结构图

5.5.3 模型预测与对比选择

经过模型的实际运用预测后，得到的预测精度结果记录在下面的两个表格中：

表 5-5：随机森林模型预测精度结果

	MSE	R^2
焦油产率	0.0005	0.87
水产率	0.0001	0.98
焦渣产率	0.0005	0.96
正己烷可溶物产率	0.0030	0.19

表 5-6：XGBoost 模型预测精度结果

	MSE	R^2
焦油产率	0.0004	0.88
水产率	0.0001	0.97
焦渣产率	0.0003	0.98
正己烷可溶物产率	0.0110	-1.93

基于上述两表记录的数据，不难看出，随机森林模型预测精度结果整体表现更优，故后续选择随机森林模型进行预测。

在样本数据中，随机抽取 3 组数据，使用随机森林模型进行预测，展示预测结果，如下图所示：

```
样本 84:
焦油产率 - 真实值: 0.1213, 预测值: 0.1215
水产率 - 真实值: 0.0802, 预测值: 0.0865
焦渣产率 - 真实值: 0.7161, 预测值: 0.7208
正己烷可溶物产率 - 真实值: 0.0000, 预测值: 0.0000

样本 4:
焦油产率 - 真实值: 0.1054, 预测值: 0.1079
水产率 - 真实值: 0.0914, 预测值: 0.0955
焦渣产率 - 真实值: 0.7269, 预测值: 0.6981
正己烷可溶物产率 - 真实值: 0.0000, 预测值: 0.0000

样本 16:
焦油产率 - 真实值: 0.3957, 预测值: 0.2917
水产率 - 真实值: 0.1222, 预测值: 0.1397
焦渣产率 - 真实值: 0.2925, 预测值: 0.3033
正己烷可溶物产率 - 真实值: 0.2645, 预测值: 0.3156
```

图 5-17：随机 3 组数据预测结果展示

为了进一步展示模型的预测效果，绘制预测数据和实际值之间的相关散点图，红线即为预测的基准线越靠近红线说明预测结果越逼近真实结果，如下图所示：

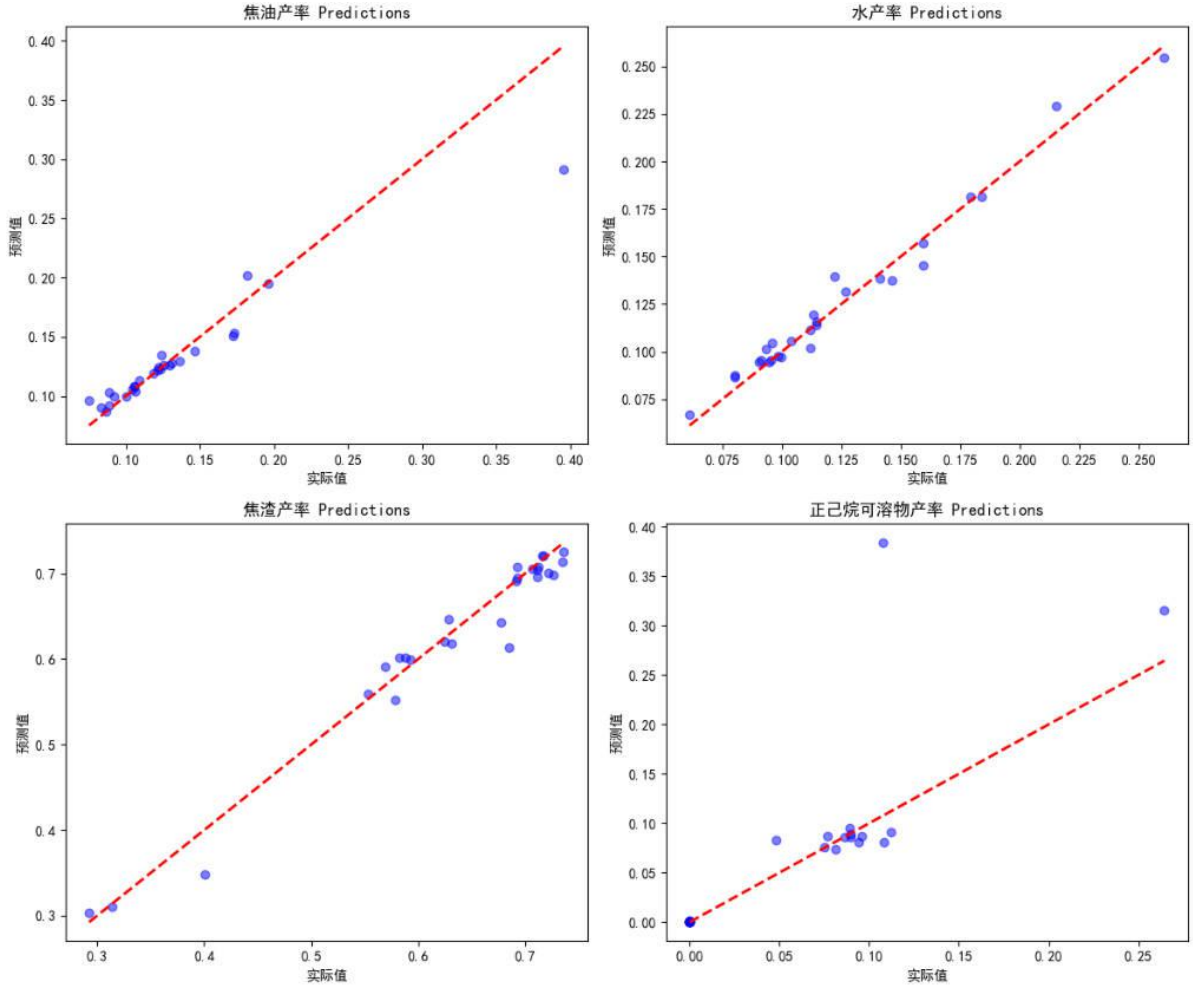


图 5-18：预测效果散点图展示

六、模型评价

6.1 模型优点

多元线性回归模型优点：

- 简单直观：模型结构简单，容易理解和解释，适合非专业人士把握模型的基本思路和原理。
- 可解释性强：可以明确地知道每个自变量对因变量的影响，提供清晰的参数估计。
- 可预测性强：在满足一定假设的情况下，模型具有良好的预测能力。
- 计算速度快：相对于其他复杂的模型，计算速度较快。

随机森林模型优点：

- 抗过拟合能力强：通过构建多个决策树，随机森林模型具有较强的抗过拟合能力。

- 预测能力强：随机森林模型可以处理高维数据，且预测准确率较高。
- 易于实现：相对于其他复杂的模型，随机森林模型容易实现。
- 不需要数据预处理：随机森林模型可以处理缺失值、异常值等，不需要过多的数据预处理。

6.2 模型缺点

多元线性回归模型缺点：

- 对异常值敏感：异常值可能会对模型产生较大的影响。
- 线性假设：模型假设因变量与自变量之间存在线性关系，可能会忽略变量间的非线性关系这可能不适用于所有数据。
- 过拟合风险：当模型过于复杂时，容易出现过拟合现象

随机森林模型缺点：

- 计算资源消耗大：随机森林模型需要构建多个决策树，因此计算资源消耗较大，尤其是在大规模数据集上进行训练。
- 参数调整复杂：随机森林模型中的参数较多，如树的个数、最大深度等，需要进行调整以获得最佳性能。
- 解释性差：随机森林模型的结果难以解释，因为它是基于多个决策树的投票或平均值，真题效果可能不如线性回归模型。

七、模型推广

在当今世界，对可再生能源和资源高效利用的需求日益增长。热解作为一种高效的生物质能转化技术，能够将有机废物转化为焦油、水和焦渣等多种有价值的产品。然而，热解过程的有效性和产物产率受到多种因素的影响，其中正己烷不溶物(INS)的添加量和混合比例是关键参数。

本研究通过一系列精心设计的实验，评估了正己烷不溶物(INS)对热解产率的影响，并揭示了它与混合比例之间的交互作用。通过高级统计分析，我们构建了一个预测模型，该模型能够准确预测不同条件下的热解产物产率。这一模型不仅帮助我们优化了共热解的混合比例，以提高产物利用率和能源转化效率，而且还允许我们比较实验值与理论值之间的差异，从而深入理解热解过程。

现在，可以将这一模型推广到更广泛的应用中。无论是生物质、塑料还是其他有机废物的热解，我们的模型都能够提供宝贵的指导和优化建议。通过调整模型参数以适应特定的原料和热解条件，本研究的模型能够帮助研究人员和工程师最大化能源转化效率，同时减少废物的产生。

此外，本模型还能够与其他热解参数（如热解温度、停留时间等）相结合的发展潜力，为复杂的热解过程提供全面的分析和优化策略。通过本研究建立的模型，在一定程度上不仅能够推动热解技术的发展，还能够为实现可持续能源和资源管理做出重要贡献。

八、参考文献

[1] P. Schober, C. Boer, L. A. J. A. Schwarte, and analgesia, "Correlation coefficients: appropriate use and interpretation," vol. 126, no. 5, pp. 1763-1768, 2018.
[2] G. A. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012.
[3] Å. J. H. o. n. a. Björck, "Least squares methods," vol. 1, pp. 465-652, 1990.
[4] D. J. J. P. b. Ozer, "Correlation and the coefficient of determination," vol. 97, no. 2, p. 307, 1985.
[5] P. J. Van Laarhoven, E. H. Aarts, P. J. van Laarhoven, and E. H. Aarts, *Simulated annealing*. Springer, 1987.
[6] 陈华根, 吴健生, 王家林, and 陈. J. 同. 自然科学版, "模拟退火算法机理研究," vol. 32, no. 6, pp. 802-805, 2004.
[7] 葛继科, 邱玉辉, 吴春明, and 蒲. J. 计算机应用研究, "遗传算法研究综述," vol. 25, no. 10, pp. 2911-2916, 2008.
[8] 岳琪 and 沈. J. 信息技术, "模拟退火算法在单目标规划问题中的应用," vol. 30, no. 5, pp. 27-28, 2006.
[9] W. G. J. B. Cochran, "The comparison of percentages in matched samples," vol. 37, no. 3/4, pp. 256-266, 1950.
[10] G. Biau and E. J. T. Scornet, "A random forest guided tour," vol. 25, pp. 197-227, 2016.
[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.

附录

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm

# 设置 matplotlib 绘图时可显示中文，不出现乱码
plt.rcParams['font.sans-serif'] = ['SimHei']  # 设置字体为黑体
plt.rcParams['axes.unicode_minus'] = False  # 使得坐标轴负号显示正常

# 加载数据
file_path = '附件一.xlsx'  # 修改为您本地的文件路径
df = pd.read_excel(file_path)

# 描述性统计和相关性分析并进行可视化
descriptive_analysis = df.describe()
correlation = df[['正己烷不溶物（INS)g', '焦油产率', '水产率', '焦渣产率']].corr()

# 输出描述性统计和相关性分析结果
print("描述性统计结果：\n", descriptive_analysis)
print("\n 相关性分析结果：\n", correlation)
```

```
# 线性回归模型建立和分析
X = sm.add_constant(df['正己烷不溶物（INS)g']) # 添加常数项用于截距
Y_tar = df['焦油产率']
Y_water = df['水产率']
Y_char = df['焦渣产率']

model_tar = sm.OLS(Y_tar, X).fit()
model_water = sm.OLS(Y_water, X).fit()
model_char = sm.OLS(Y_char, X).fit()

# 输出回归分析结果
print("焦油产率回归分析结果：\n", model_tar.summary())
print("\n 水产率回归分析结果：\n", model_water.summary())
print("\n 焦渣产率回归分析结果：\n", model_char.summary())

# 分组为零值和非零值
df['INS_group'] = df['正己烷不溶物（INS)g'].apply(lambda x: 'Zero' if x == 0 else 'Non
-zero')

# 自定义箱线图颜色
palette1 = 'deep'  # 可以设置为 'Set2', 'deep', 'muted' 等，或使用具体颜色的列表
palette2 = 'muted'  # 同上

# 绘制散点图和箱线图供后续研究
fig, axes = plt.subplots(2, 3, figsize=(18, 12))
# 箱线图
sns.boxplot(ax=axes[0, 0], hue='INS_group', y='焦油产率', x='INS_group', data=df, pale
tte=box_palette, legend=False)
axes[0, 0].set_title('正己烷不溶物与焦油产率')
sns.boxplot(ax=axes[0, 1], hue='INS_group', y='水产率', x='INS_group', data=df, palette
=box_palette, legend=False)
axes[0, 1].set_title('正己烷不溶物与水产率')
sns.boxplot(ax=axes[0, 2], hue='INS_group', y='焦渣产率', x='INS_group', data=df, pale
tte=box_palette, legend=False)
axes[0, 2].set_title('正己烷不溶物与焦渣产率')
# 散点图
sns.scatterplot(ax=axes[1, 0], x='正己烷不溶物（INS)g', y='焦油产率', hue='INS_grou
p', data=df, s=50,
                palette=scatter_palette)
axes[1, 0].set_title('正己烷不溶物与焦油产率关系')
sns.scatterplot(ax=axes[1, 1], x='正己烷不溶物（INS)g', y='水产率', hue='INS_group',
data=df, s=50,
                palette=scatter_palette)
axes[1, 1].set_title('正己烷不溶物与水产率关系')
```

```
    sns.scatterplot(ax=axes[1, 2], x='正己烷不溶物（INS)g', y='焦渣产率', hue='INS_grou
p', data=df, s=50,
                    palette=scatter_palette)
    axes[1, 2].set_title('正己烷不溶物与焦渣产率关系')
    plt.tight_layout()
    plt.show()

    # 自定义趋势线颜色
    trend_line_color = 'yellow'  # 可以设置为 'green', 'blue', 'red' 等

    # 绘制趋势线图
    fig, axes = plt.subplots(1, 3, figsize=(18, 6))
    sns.regplot(ax=axes[0], x='正己烷不溶物（INS)g', y='焦油产率', data=df, scatter_kws=
{'s': 50},
                line_kws={"color": trend_line_color})
    axes[0].set_title('正己烷不溶物（INS）与焦油产率的趋势线')
    sns.regplot(ax=axes[1], x='正己烷不溶物（INS)g', y='水产率', data=df, scatter_kws={'
s': 50},
                line_kws={"color": trend_line_color})
    axes[1].set_title('正己烷不溶物（INS）与水产率的趋势线')
    sns.regplot(ax=axes[2], x='正己烷不溶物（INS)g', y='焦渣产率', data=df, scatter_kws=
{'s': 50},
                line_kws={"color": trend_line_color})
    axes[2].set_title('正己烷不溶物（INS）与焦渣产率的趋势线')
    plt.tight_layout()
    plt.show()

    # 或者使用自定义的颜色映射
    fig, axes = plt.subplots(1, 3, figsize=(18, 6))
    sns.scatterplot(ax=axes[0], x='正己烷不溶物（INS)g', y='焦油产率', hue='INS_group',
data=df, s=50, palette='viridis')
    axes[0].set_title('正己烷不溶物与焦油产率关系')
    sns.scatterplot(ax=axes[1], x='正己烷不溶物（INS)g', y='水产率', hue='INS_group', da
ta=df, s=50, palette='viridis')
    axes[1].set_title('正己烷不溶物与水产率关系')
    sns.scatterplot(ax=axes[2], x='正己烷不溶物（INS)g', y='焦渣产率', hue='INS_group',
data=df, s=50, palette='viridis')
    axes[2].set_title('正己烷不溶物与焦渣产率关系')
    plt.tight_layout()
    plt.show()

    import pandas as pd
    from statsmodels.formula.api import ols
    import matplotlib.pyplot as plt
    import seaborn as sns
    from scipy.interpolate import griddata
    import numpy as np
```

```
import joblib
# 设置 matplotlib 绘图时可以显示中文
plt.rcParams['font.sans-serif'] = ['SimHei']  # 设置字体为黑体
plt.rcParams['axes.unicode_minus'] = False  # 使得坐标轴负号显示正常
# 加载数据
data_path = '附件一.xlsx'  # 请确保路径和文件名与实际情况相符
data = pd.read_excel(data_path)


# 配比数据处理，将配比数据转换为具体的数值形式
def convert_ratio(ratio):
    if isinstance(ratio, str) and '/' in ratio:
        num, denom = map(int, ratio.split('/'))
        return num / (num + denom)
    elif isinstance(ratio, str):
        return int(ratio) / 100
    else:
        return ratio / 100


data['配比'] = data['配比'].apply(convert_ratio)

# 重命名列，避免编码问题
data = data.rename(columns={'正己烷不溶物（INS)g': 'INS_g'})

# 构建模型，分析焦油产率、水产率和焦渣产率
model_tar = ols('焦油产率 ~ INS_g * 配比', data=data).fit()
model_water = ols('水产率 ~ INS_g * 配比', data=data).fit()
model_char = ols('焦渣产率 ~ INS_g * 配比', data=data).fit()

fig, axes = plt.subplots(1, 3, figsize=(18, 6))

# 焦油产率
sns.scatterplot(x='配比', y='焦油产率', hue='INS_g', data=data, ax=axes[0], palette='viri
dis')
axes[0].set_title('焦油产率——配比 & INS')
axes[0].set_xlabel('配比')
axes[0].set_ylabel('焦油产率')

# 水产率
sns.scatterplot(x='配比', y='水产率', hue='INS_g', data=data, ax=axes[1], palette='viridis
')
axes[1].set_title('水产率——配比 & INS')
axes[1].set_xlabel('配比')
axes[1].set_ylabel('水产率')

# 焦渣产率
```

```
    sns.scatterplot(x='配比', y='焦渣产率', hue='INS_g', data=data, ax=axes[2], palette='viridis')
    axes[2].set_title('焦渣产率——配比 & INS')
    axes[2].set_xlabel('配比')
    axes[2].set_ylabel('焦渣产率')

    plt.tight_layout()
    plt.show()

    # 输出模型结果
    print("焦油产率模型结果:\n", model_tar.summary())
    print("水产率模型结果:\n", model_water.summary())
    print("焦渣产率模型结果:\n", model_char.summary())

    # 生成预测数据网格
    pred_data = pd.DataFrame({
        '配比': np.repeat(np.linspace(0, 1, 50), 6),
        'INS_g': np.tile(np.linspace(0, 0.75, 6), 50)
    })

    # 预测焦油产率、水产率和焦渣产率
    pred_data['焦油产率预测'] = model_tar.predict(pred_data)
    pred_data['水产率预测'] = model_water.predict(pred_data)
    pred_data['焦渣产率预测'] = model_char.predict(pred_data)


    # 定义一个函数来绘制等高线图
    def plot_contour(data, x, y, z, title):
        # 创建网格数据
        xi = np.linspace(data[x].min(), data[x].max(), 100)
        yi = np.linspace(data[y].min(), data[y].max(), 100)
        zi = griddata((data[x], data[y]), data[z], (xi[None, :], yi[:, None]), method='cubic')

        # 绘制等高线图
        plt.figure(figsize=(10, 6))
        contour = plt.contourf(xi, yi, zi, levels=50, cmap='plasma')
        plt.colorbar(contour)
        plt.title(title)
        plt.xlabel(x)
        plt.ylabel(y)
        plt.show()


    # 假设 pred_data 中包含了以下列：INS_g, 配比, 焦油产率预测, 水产率预测, 焦渣产率预测
    # 调用函数绘制等高线图
```

```
plot_contour(pred_data, 'INS_g', '配比', '焦油产率预测', '预测的焦油产率——配比 & INS')
plot_contour(pred_data, 'INS_g', '配比', '水产率预测', '预测的水产率——配比 & INS')
plot_contour(pred_data, 'INS_g', '配比', '焦渣产率预测', '预测的焦渣产率——配比 & INS')



# 保存焦油产率模型
joblib.dump(model_tar, 'model_tar.joblib')

# 保存水产率模型
joblib.dump(model_water, 'model_water.joblib')

# 保存焦渣产率模型
joblib.dump(model_char, 'model_char.joblib')

import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.svm import SVR
from sklearn.multioutput import MultiOutputRegressor
from sklearn.metrics import mean_squared_error, r2_score
# 设置 matplotlib 绘图时可以显示中文
plt.rcParams['font.sans-serif'] = ['SimHei']  # 设置字体为黑体
plt.rcParams['axes.unicode_minus'] = False  # 使得坐标轴负号显示正常

# 载入数据
data = pd.read_excel('附件一.xlsx')  # 请确保文件路径正确

# 数据预处理
label_encoder = LabelEncoder()
data['试样编码'] = label_encoder.fit_transform(data['试样'])
data['配比'] = data['配比'].apply(
    lambda x: float(x.split('/')[0]) / float(x.split('/')[1]) if isinstance(x, str) else x)

# 选择因变量和自变量
X = data[['试样编码', '配比', '样品 g', '焦油(Char)g', '水(Water)mL', '正己烷不溶物（INS)g']]  # 自变量
y = data[['焦油产率', '水产率', '焦渣产率', '正己烷可溶物产率']]  # 因变量

# 数据分割
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
    # 创建并训练多输出随机森林模型
    multi_rf_model = MultiOutputRegressor(RandomForestRegressor(n_estimators=100, ra
ndom_state=42))
    multi_rf_model.fit(X_train, y_train)

    # 模型预测
    y_pred = multi_rf_model.predict(X_test)

    # 计算多输出的 MSE 和 R²
    mse_scores = [mean_squared_error(y_test.iloc[:, i], y_pred[:, i]) for i in range(y_test.sha
pe[1])]
    r2_scores = [r2_score(y_test.iloc[:, i], y_pred[:, i]) for i in range(y_test.shape[1])]

    # 可视化每个模型的测试集预测结果
    fig, axes = plt.subplots(2, 2, figsize=(12, 10))
    axes = axes.ravel()
    products = ['焦油产率', '水产率', '焦渣产率', '正己烷可溶物产率']

    for i, ax in enumerate(axes):
        ax.scatter(y_test.iloc[:, i], y_pred[:, i], alpha=0.5)
        ax.plot([y_test.iloc[:, i].min(), y_test.iloc[:, i].max()],
                [y_test.iloc[:, i].min(), y_test.iloc[:, i].max()], 'k--', lw=2)
        ax.set_title(f'{products[i]}\nMSE: {mse_scores[i]:.4f} | R2: {r2_scores[i]:.2f}')
        ax.set_xlabel('实际值')
        ax.set_ylabel('预测值')

    plt.tight_layout()
    plt.show()

    # 随机选择五个样本的索引
    sample = np.random.choice(X_test.index, size=5, replace=False)

    # 获取相对位置索引
    relative_indices = [np.where(y_test.index == idx)[0][0] for idx in sample]

    # 提取这些样本的真实值和预测值
    sample_true = y_test.loc[sample]
    sample_pred = y_pred[relative_indices]  # 使用修正后的相对位置索引

    # 创建数据框来展示结果
    results = []
    for i, idx in enumerate(sample):
        true_values = sample_true.loc[idx]
        pred_values = pd.Series(sample_pred[i], index=sample_true.columns)
        comparison = pd.DataFrame({'真实值': true_values, '预测值': pred_values})
        comparison = comparison.reset_index()
        comparison.rename(columns={'index': '产物'}, inplace=True)
        comparison.insert(0, '样本索引', idx)
```

```
        results.append(comparison)

    # 合并所有结果为一个 DataFrame
    final_results = pd.concat(results, axis=0)
    final_results.set_index(['样本索引', '产物'], inplace=True)

    # 展示表格
    print(final_results)

    # 载入数据
    data = pd.read_excel('附件一.xlsx') # 请确保文件路径正确

    # 数据预处理
    label_encoder = LabelEncoder()
    data['试样编码'] = label_encoder.fit_transform(data['试样'])
    data['配比'] = data['配比'].apply(
        lambda x: float(x.split('/')[0]) / float(x.split('/')[1]) if isinstance(x, str) else x)

    # 选择因变量和自变量
    X = data[['试样编码', '配比', '样品 g', '焦油(Char)g', '水(Water)mL', '正己烷不溶物（I
NS)g']] # 自变量
    y = data[['焦油产率', '水产率', '焦渣产率', '正己烷可溶物产率']] # 因变量

    # 数据分割
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    # 创建并训练多输出随机森林模型
    rf_model = MultiOutputRegressor(RandomForestRegressor(n_estimators=100, random_
state=42))
    rf_model.fit(X_train, y_train)

    # 创建并训练多输出支持向量回归模型
    svr_model = MultiOutputRegressor(SVR())
    svr_model.fit(X_train, y_train)

    # 创建并训练多输出梯度提升树模型
    gbr_model = MultiOutputRegressor(GradientBoostingRegressor(n_estimators=100, rand
om_state=42))
    gbr_model.fit(X_train, y_train)

    # 模型预测
    rf_pred = rf_model.predict(X_test)
    gbr_pred = gbr_model.predict(X_test)

    # 计算多输出的 MSE 和 R²
    models = {'Random Forest': rf_pred,  'Gradient Boosting': gbr_pred}
```

```
    mse_scores = {model: [mean_squared_error(y_test.iloc[:, i], pred[:, i]) for i in range(y_te
st.shape[1])] for model, pred
                  in models.items()}
    r2_scores = {model: [r2_score(y_test.iloc[:, i], pred[:, i]) for i in range(y_test.shape[1])]
for model, pred in
                 models.items()}

    # 输出 MSE 和 R² 结果
    for model in models:
        print(f"{model} Results:")
        for i, product in enumerate(products):
            print(f"{product} - MSE: {mse_scores[model][i]:.4f}, R²: {r2_scores[model][i]:.2f}
")
        print("\n")



    # 可视化随机森林模型的预测结果
    fig, axes = plt.subplots(2, 2, figsize=(12, 10))
    axes = axes.ravel()
    products = ['焦油产率', '水产率', '焦渣产率', '正己烷可溶物产率']

    for i, ax in enumerate(axes):
        ax.scatter(y_test.iloc[:, i], rf_pred[:, i], alpha=0.5)
        ax.plot([y_test.iloc[:, i].min(), y_test.iloc[:, i].max()],
                [y_test.iloc[:, i].min(), y_test.iloc[:, i].max()], 'k--', lw=2)
        ax.set_title(f'{products[i]} Predictions')
        ax.set_xlabel('实际值')
        ax.set_ylabel('预测值')

    plt.tight_layout()
    plt.show()

    # 展示随机森林模型预测的随机五个样本的真实值和预测值
    sample = np.random.choice(y_test.index, 5, replace=False)
    sample_true = y_test.loc[sample]
    sample_pred = pd.DataFrame(rf_pred, index=y_test.index).loc[sample]

    print("随机五个样本的真实值与预测值对比：")
    for i in sample:
        print(f"\n 样本 {i}:")
        for j, product in enumerate(products):
            print(
                f"{product} - 真实值: {sample_true.iloc[sample_true.index == i, j].values[0]:.4
f}, 预测值: {sample_pred.loc[i, j]:.4f}")

    import matplotlib.pyplot as plt
```

```
    # 设置颜色变量，可以根据需要修改这些颜色
    scatter_color = 'blue'  # 散点图颜色
    line_color = 'red'  # 线条颜色

    # 可视化随机森林模型的预测结果
    fig, axes = plt.subplots(2, 2, figsize=(12, 10))
    axes = axes.ravel()
    products = ['焦油产率', '水产率', '焦渣产率', '正己烷可溶物产率']

    for i, ax in enumerate(axes):
        ax.scatter(y_test.iloc[:, i], rf_pred[:, i], alpha=0.5, color=scatter_color)
        ax.plot([y_test.iloc[:, i].min(), y_test.iloc[:, i].max()],
                [y_test.iloc[:, i].min(), y_test.iloc[:, i].max()], color=line_color, linestyle='--', lw=
2)
        ax.set_title(f'{products[i]} Predictions')
        ax.set_xlabel('实际值')
        ax.set_ylabel('预测值')

    plt.tight_layout()
    plt.show()
```