



中国研究生创新实践系列大赛

“华为杯”第十七届中国研究生

数学建模竞赛

学　　校	重庆大学
参赛队号	20106110058
队员姓名	1. 郑维伟
	2. 钟　健
	3. 黄小芹

中国研究生创新实践系列大赛

“华为杯”第十七届中国研究生

数学建模竞赛

题 目 辛烷值损失较大的影响因素与优化措施

摘 要：

近年来，随着中国汽车保有量的快速增长，中国对于车用汽油需求量的逐年增加，作为成品汽油主要来源的催化裂化汽油，必须对其进行精制处理，有效降低汽油产品中辛烷值损失，以满足对汽油质量的要求，同时有效降低环境污染，并可带来可观的经济效益。在此背景下，本文以某石化企业催化裂化汽油精制脱硫装置 4 年来的历史观察数据为样本，通过数据预处理、提取影响辛烷值损失的核心影响因素、对次要操作变量进行主成分分析降维、多元非线性回归模型验证主要变量提取合理性、多种单一及复合机器学习预测算法、DEA 效率评价以及混合整数规划模型优化算法等步骤，探讨了该企业汽油产品辛烷值损失的影响因素及其主要操作变量可行的优化方案，并最终对相关研究结果进行了图示化展示。

针对问题一，依据相关数据处理原则，本研究对附件三中 285 和 313 号样本的操作变量数据进行了**删除空值位点**、**缺省值替换**、**第 5 和 95 百分位截尾处理**以及**“混频”数据集成**等预处理，并将其统一整合添加到附件一对应样本号中，供后续研究调用。

针对问题二，首先，通过文献研究法确定了包含**反应温度**、**反应压力**、**氢油比**、**质量空速**等 **10 个**影响辛烷值损失的核心变量，对其进行原则性保护；其次，再次依据相关数据处理原则，对附件一中相关数据进行预处理，确保待降维数据集基本信息正常；再次，在**保护核心变量**和**剔除异常指标**的基础上，运用**主成分分析**对余下操作变量数据集进行降维处理，并提取信息**累积贡献率达 80.1467%**的**前 15 个主成分**作为余下操作变量的精炼指标，最终提取出包含原料辛烷值在内的共计 **29 个建模变量**；最后，通过逐步建立**多元非线性回归模型**，对比了回归结果中各核心变量系数及其显著性与已有文献基于自然实验得到的相关影响情况总体一致，侧面验证了所提取主要建模变量的合理性与有效性。

针对问题三，基于数据挖掘技术，建立了基于**随机森林算法**、**线性核（Linear Kernel）的支持向量机**以及**岭回归**的机器学习算法，并在此基础上设计了**基于 Stacking 策略**将以上三种算法融合的**复合预测算法**对辛烷值损失进行预测，相应样本均方误差分别为 **0.033522**、**0.033478**、**0.0352379** 以及 **0.035339**，表明其均具有较好的拟合效果，且均较为稳定。

针对问题四，首先，通过引入“**辛烷值损失率**”以及“**产品硫去除率**”两个概念，找到了同时处于“辛烷值损失率”负向排序以及“产品硫去除率”正向排序前 20 的 **194 号样本**，并将其作为历史样本数据中处于“绝对占优”的理想样本，作为其他样本操作变量进

行优化操作的“标杆”；进一步基于投入产出视角引入 **DEA-BCC 效率评价模型**对样本进行多指标综合评价，进而由 21 个 DEA 有效样本得到其变量均值构造出“**理想样本**”作为各样本参数调整的“标杆”（该“理想样本”产品硫含量仅为 3.285μg/g），并依据各主要操作变量的正负属性，图示了各样本中主要操作变量需进行改进的方向及幅度；最后，构造了基于**混合整数规划模型**的优化算法，并结合问题三的预测算法，得到了当主要操作变量反应温度、质量空速、反应压力、氢油比、原料汽油硫含量的最优取值分别为（**400℃，7h-1，2 Mpa，0.2，590**）时，325 个样本中仅有 6 个辛烷值损失降幅小于 30%，47.69%的样本损失降幅大于 60%，且有 **81.23%**的样本可在满足产品硫含量不大于 5$\mu g/g$的条件下实现辛烷值损失降幅大于 30%。

针对问题五，以各操作变量取值范围和单次可调整幅度为约束，首先，本文根据问题四中 DEA-BCC 模型综合评价所得到的 21 个 DEA 有效样本均值所构造的“**理想样本**”为改进方向对 133 号样本进行 **10 次分阶段优化**，并将每阶段操作变量的优化参数值带入问题三所构建的预测模型中，最终得到 133 号样本的产品硫含量可由最初的 3.598 $\mu g/g$降低到**3.327**$\mu g/g$，辛烷值损失由最初的 1.31 降低到 **0.937** 个单位，即降低了约 **31.61%**；其次，利用基于**多元线性回归模型**，分**单次**和**批量**两种优化调整策略对操作变量分别进行 **44** 和 **8** 次优化操作，最终 133 号样本的产品辛烷值均可由 89.1 提高到 **89.7** 个单位，产品硫含量则可分别由 3.63 降低到 **3.51** 以及由 3.68 降到 **3.38**$\mu g/g$。

此外，本文基于**中介效应模型**，进一步研究了产品硫含量去除通过烯烃最终影响到产品辛烷值损失的作用传导机制，且验证了**烯烃**在此过程中存在**完全中介效应**。最后，对本文相关建模过程及实证结果进行了评价，并提出了未来改进和推广方向，使之更加具有普适性、应用性和现实意义。

关键词：主成分分析；多元非线性回归；复合机器学习；DEA-BCC **模型；中介效应分析**

目录

一、研究背景与问题提出

1.1 研究背景

近年来，随着中国汽车市场的迅速发展，汽油作为小型车辆的主要燃料，导致了中国对原油需求的持续增长。据美国能源信息署统计信息显示，2019 年中国的原油进口增长便已达到日均 1010 万桶，相较于 2018 年而言实现了日均增加 90 万桶的历史性新高。而据油气行业统计数据显示，自 2017 年中国超过美国成为世界上最大的原油进口国以来，近两年中国的石油对外依存度已超过 70%。与此同时，伴随着汽车数量的持续增长，中国每年因大量的汽油燃烧而导致温室气体排放、空气污染物（如氮氧化物、一氧化碳以及或颗粒物质等）的逐年增加，在建设“美丽中国”的时代背景下，给中国环境治理带来巨大压力。

因此，自 2017 年伊始，中国便开始实行了供应符合相关标准的含硫量小于 10 μg/g 的汽油。其中，催化裂化（FCC）汽油作为中国车用汽油的主要调组分和来源，提供了中国近 70%的商品汽油。值得注意的是，在实际生产过程中，催化裂化汽油的含硫量却往往远高于国家的相关标准，如中石化对未进行脱硫处理的催化裂化（FCC）汽油进行硫含量测量发现，其含硫量高达 580-942$\mu g/g$[1]。可见，对催化裂化汽油进行精制处理，以降低汽油中的硫含量是一个亟待解决的现实问题。

然而，在中国当前该领域的现有技术条件和工业操作过程中，极易出现在降低汽油硫含量的同时，导致汽油产品中辛烷值的较大程度损失。而辛烷值作为反映汽油燃烧性能的一项最为重要的衡量指标，如果能在进行汽油脱硫过程中有效地降低其损失值，那么其不仅能够极大程度地提高对汽油的利用效率，同时还可以有效地降低对环境的污染，并最终带来极为可观的经济效益。

在此背景下，本研究将着重对某石化企业的催化裂化汽油精制脱硫装置运行 4 年所获得的历史观测数据展开研究，力图通过数据挖掘技术，对其大量历史数据所集中反映出的汽油产品辛烷值损失值明显高于同类装置的最小损失值（即该某石化企业汽油产品辛烷值平均损失 1.37 个单位，而行业均值仅有 0.6 个单位）的核心影响因素展开探讨，并在此基础上运用合适的优化模型与算法，为该企业在脱硫过程中有效地降低汽油产品辛烷值损失提出可能的优化操作方案。

1.2 问题重述

问题一：数据预处理

由于数据采集过程中，原料、产品和催化剂的数据主要来源于 LIMS 实验数据库（仅对应一个测度值），而主要操作变量的采集频次则为 3 分钟或 6 分钟/次，导致在进行实证研究之前，需要对相关数据进行“混频”数据处理；同时考虑到原始数据中可能存在部分异常值。因此，需参考某石化企业近 4 年的历史数据，参照附件二所提供的“样本确定方法”，对所给定的 285 号和 313 号样本进行数据预处理，并将处理后的数据添加到附件一所对应的样本号之中。

问题二：寻找主要变量

根据附件所提供的 325 个样本数据，选择合适的降维方法从 367 个变量中筛选出具有代表性、独立性，并且可以较好地测度辛烷值（RON）损失的主要变量，而为了工程应用方便，还需尽量保证降维后的主要变量在 30 个以下，并尽可能包含原料的辛烷值（RON）变量，同时还需详细说明建模主要变量的筛选过程及其合理性。

问题三：建立预测模型

以问题二所筛选出的主要建模变量为基础，运用合适的数理模型以及机器学习算法等数据挖掘技术建立辛烷值（RON）的损失预测模型，并进行模型验证。

问题四：主要变量操作方案的优化

在保证产品硫含量不大于$5\mu g/g$的前提下，建立合理的模型对 325 个样本数据中的操作变量进行优化分析与改进，并给出辛烷值（RON）损失降幅大于 30%的样本所对应的主要变量优化后的操作条件。同时需考虑到优化过程中原料、待生吸附剂以及再生吸附剂的性质保持不变的基本情况。

问题五：模型的可视化展示

在保持样本的原料性质、待生吸附剂以及再生吸附剂的性质不变的前提下，着重针对 133 号，通过前期的数理模型与优化算法，以各主要操作变量每次允许调整幅度为条件，图示其在主要操作变量的优化调整过程中，所对应的汽油辛烷值以及硫含量的变化轨迹。

二、基本假设

为确保对汽油辛烷值损失的预测和优化方案更加具有针对性以及合理性，本文提出如下基本假设：

（1）对原始变量进行 5%和 95%的截尾处理后的样本数据能够较好地符合附件二中相关工艺要求与操作经验所限幅的操作范围。

（2）在同一转置中各个位点各项统计信息相似的情况下，通过方差最小原则筛选出其中一个代表性指标具有合理性和可行性。

（3）基于已有文献研究基础所筛选出影响辛烷值损失的核心指标具有重要的经验和现实意义。

（4）产品硫含量去除程度和产品辛烷值损失情况均受到其他某些共同因素影响，导致两者之间可能存在统计意义上的相关性，但并不存在直接的因果关系，即两者之间可能存在某些中介传导机制。

三、符号说明

编号	符号	符号说明
1	R	相关系数矩阵
2	λ_j	特征值
3	pca_i	第i个主成分
4	b_j	主成分y_j的信息贡献率
5	X_i	影响辛烷值损失的主要指标（解释变量）
6	θ	DEA 效率值
7	M	中介变量

四、问题分析与研究方案

4.1 问题一：数据预处理

主要依据附件二所提供的“样本确定方法”，首先对附件三中的 285 和 313 号样本进行基本统计特征分析，以初步判断其是否存在缺省值、异常值以及变量取值范围是否符合实际等情况。在此基础上，针对样本所出现的不同问题，给出相应的预处理方案。然后，借助 Stata 数据处理软件对以上样本数据进行预处理，并根据附件二“样本确定方法”将其操作变量进行高频数据转换为低频数据，并最终将整合后的样本观察数据添加到附件一所对应的样本号之中。

4.2 问题二：寻找建模主要变量

首先，通过文献研究法，以初步把握导致辛烷值损失的核心影响因素，对于这类变量所对应的指标进行原则性预保留，以避免后期进行数据降维过程中造成有效信息损失。其次，为避免在后续模型优化过程中对操作变量进行优化调整时，由于缺少对该变量基本信息的了解和把握，故在筛选建模的主要变量前，有必要剔除到对应在附件四中中文名称以及相关单位或取值范围信息缺失的变量。再次，为确保对变量进行降维时，因样本原始数据中某些变量缺失或遗漏数据影响降维效果，还需结合附件二中“样本确定方法”及其他合理的数据处理原则，对附件一中原始数据进行预处理。最后，在通过上述步骤初步筛选得到影响辛烷值损失的核心变量的情况下，为尽量保留原有统计数据信息，对余下的操作变量指标进行降维处理，以得到其他用于解释辛烷值损失的数据集，进而筛选得到最终的建模主要变量。此外，为验证所筛选出的变量是否合理，本节还将参考相关文献研究结果，建立辛烷值损失预测的多元非线性回归模型，通过各变量影响方向与已有研究结论的对比，初步验证模型主要变量筛选的合理性。

4.3 问题三：建立预测模型

以问题二所得的主要建模变量数据集为基础，从数据挖掘技术方法出发，通过构建合理的数理模型，运用机器学习算法建立辛烷值（RON）的损失预测模型。同时，考虑到应用单一的方法进行预测时可能存在片面性和难以确保准确性等弊端，因此，有必要建立含支持向量机、随机森林以及岭回归等在内的多种模型与算法。此外，为进一步提高辛烷值损失预测模型的精度，本文还基于集成学习的思想，考虑建立多种机器学习算法融合的辛烷值损失预测复合模型。最后，为确保模型预测的可靠性，还将对所采用的各种预测模型与算法进行误差对比与评价，进而确定相对误差最小的辛烷值损失预测模型，为后续研究做准备。

4.4 问题四：主要变量操作方案的优化

为了研究在产品硫含量在不大于 $5\mu g/g$ 条件下，以辛烷值损失降幅超过 30%（即尽可能低）为主要目标时，各观察样本所对应的操作变量的优化操作情况。本文首先考虑引入“辛烷值损失率”以及“产品硫去除率”两个概念，并分别以两个指标作为评价对象对样本数据进行评价排序，以试图发现历史样本数据中处于“绝对占优”的理想样本，并考虑将其对应的操作变量取值情况作为最优取值参考。其次，考虑到一般而言，实际过程中鲜有两个指标排序均靠前的观察样本，故考虑沿袭以上思路，通过构建合适的综合评价模型，以找到综合相关评价指标下综合得分靠前的样本数据，并考虑将其作为“理想样本”或“标杆”，探讨得分靠后样本的操作变量向其靠近“学习”以及优化调整的操作条件，并可结合问题三所构建的预测模型，模拟当操作变量参数变动后，各样本相应辛烷值损失的预测情

况。最后，鉴于历史样本中的“理想样本”在一定程度上限制了各样本可能具有的更大优化和改进空间，因此，有必要从优化算法思路出发，构建合理的优化算法，并结合问题三所建立的预测模型，基于数据优化层面，从另一思路出发，再次考察降低辛烷值损失以及产品硫含量不大于 $5\mu g/g$ 条件下各操作变量的优化操作。

4.5 问题五：模型的可视化展示

在问题四的研究基础上，本节主要针对 133 号样本，以各操作变量的单次调整幅度为限制，并通过多次运用问题三所得到的辛烷值的预测模型（另外还需对样本硫含量进行预测），以图示在对其操作变量进行逐步优化操作过程中，所对应的汽油辛烷值和硫含量的变化轨迹。在以上基本思路下，本文还将具体从以下两种思路出发制定具体的操作变量优化策略：其一，考虑以问题四中根据历史样本所确定的“理想样本”所对应的操作变量参数作为优化参考；其二，从问题四中所建立的优化算法思路出发，探讨在操作变量取值范围以及单次调整幅度的约束条件下，133 号样本所存在的理论上最优操作变量优化操作策略。

此外，在对以上问题进行整体分析的基础上，为便于后续建模过程中思路的清晰化，本文使用 Visio 绘制了最终的思维导图（见图 1）。

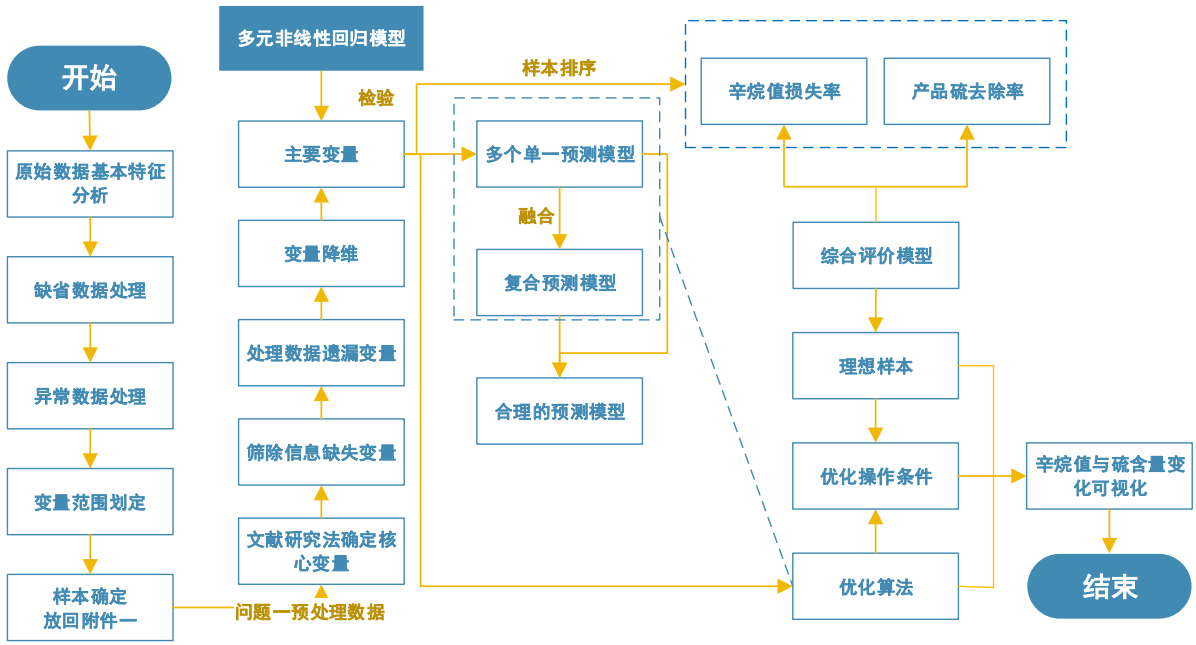

图 1　思路导图

五、模型的建立与求解

5.1 问题一：数据预处理

5.1.1 原始数据初步分析

对附件二以及附件三展开分析得知，附件三所提供的原始数据由某石化企业的催化裂化汽油精制脱硫装置在 2017/7/17 6:03:00 - 2017/7/17 8:00:00 以及 2017/5/15 6:03:00 - 2017/5/15 8:00:00 时间段采集所得（数据采集频次为 3 分钟/次），相应的样本编号分别为 285 和 313，其中包含了诸如反映原料、产品、待生吸附剂以及再生吸附剂基本性质在内的观测数据，同时也涵盖了如氢油比（S-ZORB.CAL_H2.PV）、反应过滤器压差（S-ZORB.PDI_2102.PV 和 S-ZORB.SIS_PDT_2103B.PV）、还原器压力（S-ZORB.PT_2801.PV）、还原器流化氢气流量（S-ZORB.FC_2801.PV）、反应器上部温度（S-ZORB.TE_2103.PV）等在内的 354 个操作变量数据，两个样本数据共计 28,320 条。

实际上，结合附件二进一步了解发现，附件三中的原始数据中大部分变量数据正常，但每个样本编号下的数据均可能存在部分位点异常现象，例如部分变量可能仅含有部分时间段的数据、部分变量的数据全部为空值或部分数据为空值。通过对附件三中的数据进行观察，发现部分变量的数据确实存在全部为空值或部分为空值的现象，如新氢进装置流量（S-ZORB.FT_1501.PV）、进料调节阀旁路流量（S-ZORB.FC_1104.DACA）等变量在两个样本中的数据均全部为空值；D-106 热氮气流量（S-ZORB.FC_2432.DACA）、3.0 步骤 FIC2432.SP（S-ZORB.FC_2432.PIDA.SP）等变量在 313 号样本中出现部分为空值的现象。对此，本文将其统一归为 I 类数据缺失问题。值得注意的是，结合“样本确定方法”对照分析，本文还发现除 I 类数据缺失问题外，附件三中原始数据还普遍存在 II 类超出操作范围（即根据工艺要求与操作经验，原始数据中某变量的观测值大幅超出了正常的操作范围）和 III 类数据异常（即超出拉依达准则“3σ准则”以外的数据观测值）等问题。

5.1.2 数据处理基本原则

根据附件二“样本确定方法”及相关数据情况，总结提炼出如下数据处理基本原则：

（1）对于只含有部分时间点的位点，若其残缺数据较多（本文定义为超过样本量 30%），则认为即使进行了缺省值处理，也极易存在较大程度的误差，故将其删除。

（2）某位点所对应的 325 个样本全部为空值，将其删除（实际上已包含于上一原则）。

（3）若某位点仅存在部分数据缺失（本文定义为低于样本量 50%），则采用其前后两小时该变量的均值数据代替。

（4）对原始数据采用 5%和 95%截尾处理，以作为其工艺要求和操作经验所得出的变量现实可操作范围，并以此为准则剔除未在此范围的样本。

（5）附件二要求根据拉依达准则（3σ准则）去除可能存在的数据异常值，事实上，由于正负 3 个标准差内实际包含了各统计数据约 99.73 百分位上的数据，由于在上一原则中已经采取了 95 百分位上数据的结尾处理，故这一原则实际处理过程中已确保得到满足。

采用附件二中“样本确定”的基本原则，即以辛烷值数据测定的时间点为基准时间，取其前 2 个小时的操作变量数据的平均值作为对应辛烷值的操作变量数据。

5.1.3 附件三数据预处理

根据 5.5.1 节对原始数据的初步分析发现，其大体存在三类问题，故结合 5.1.2 节数据处理基本原则，本文针对不同类型问题给出了相应较为合理的处理方式（见表 1）。

表 1　问题数据处理方案

问题类型	具体表述	处理方式
I 类数据缺失问题	1.大部分为空值，无法补充	依据规则（1）删除位点
	2.位点部分数据为空值	依据规则（3）插补数据
II 类超出操作范围	变量的数据不在给定范围内	依据规则（4）删除样本
III 类数据异常	在 5%和 95%结尾前含有较大误差值	依据规则（4）删除样本

按照以上问题数据处理思路，本文针对附件三中相关数据进行预处理的具体过程如下：

Step 1. 依据数据缺失比例替换或剔除数据

在应用 Stata 分别对 285 和 313 号样本原始数据进行描述性统计特征分析的基础上，结合相应的数据处理方式，最终对两个样本进行缺省值处理情况（见表 2）。其中，285 号样本替换了 11 个位点下的部分数据，313 号样本替换了 8 个位点下的部分数据。此外，还对 313 号样本的 5 个位点数据进行了缺省值均值替换（见表 3）。

表 2　附件三原始数据情况

样本号	变量处理数量	位号	中文名称
285	11	S-ZORB.FT_1501.PV	新氢进装置流量
		S-ZORB.FT_1002.PV	1#催化汽油进装置流量
		S-ZORB.FC_1202.PV	D121 顶去放火炬流量
		S-ZORB.FT_1501.TOTAL	新氢进装置流量
		S-ZORB.FT_5102.PV	-
		S-ZORB.FT_2901.DACA	D-109 松动风流量
		S-ZORB.FC_1104.DACA	进料调节阀旁路流量
		S-ZORB.FT_2803.DACA	紧急氢气去 D-102 流量
		S-ZORB.FT_1502.DACA	补充氢压缩机出口返回管流量
		S-ZORB.TEX_3103A.DACA	EH-102 加热元件/A 束温度
		S-ZORB.FT_5102.DACA.PV	D-201 含硫污水排量
313	8	S-ZORB.FT_1501.PV	新氢进装置流量
		S-ZORB.FT_1002.PV	1#催化汽油进装置流量
		S-ZORB.FT_1501.TOTAL	新氢进装置流量
		S-ZORB.FT_2901.DACA	D-109 松动风流量
		S-ZORB.FC_1104.DACA	进料调节阀旁路流量
		S-ZORB.FT_2803.DACA	紧急氢气去 D-102 流量
		S-ZORB.FT_1502.DACA	补充氢压缩机出口返回管流量
		S-ZORB.TEX_3103A.DACA	EH-102 加热元件/A 束温度

表 3　313 号样本局部数据调整情况

样本号	位号	中文名称	缺失个数	均值数据填补
313	S-ZORB.FT_1204.PV	-	2	49.70743
	S-ZORB.FC_2432.DACA	D-106 热氮气流量	16	55.02505
	S-ZORB.FT_2431.DACA	-	6	255.4798
	S-ZORB.FC_2432.PIDA.SP	3.0 步骤 FIC2432.SP	16	55.97 741
	S-ZORB.FT_1204.DACA.PV	D-121 含硫污水排量	2	49.70743

Step 2. 数据截尾处理

如前文所述，本文采用数据 5%和 95%分位数作为最大最小值的限幅方法，剔除部分不在此范围的样本。值得注意的是，经过以上数据处理，实质上已经剔除掉了根据拉依达准则（3σ准则）所可能存在的异常值。Stata 指令最终的处理结果显示，285 和 313 号样本均删除了 21 个样本观察数据（即 21 行）。

Step 3. “混频”数据集成处理

“混频”数据集成，即根据低频数据的周期对高频数据做平均或累加处理。依据数据基本原则（6），本文分别取 285 和 313 号前 2 个小时操作变量数据的平均值作为对应辛烷值的操作变量数据，并在提取完成之后，将操作变量数据预处理后的 285 和 313 号样本，连同附件三中其各自“原料”、“产品”、“待生吸附剂”和“再生吸附剂”工作表下数据一起合并后追加到附件一之中（Stata 数据预处理代码见附件 1），具体结果可参见附件：“附件一：325 个样本数据(附件三已追加).xlsx”。

5.2 问题二：寻找建模主要变量

5.2.1 基于文献研究法与数据处理原则的核心变量筛选

为初步把握导致辛烷值损失的核心影响因素，在查阅相关文献的基础上，本文对影响辛烷值损失的主要因素以及这些因素在装置实际运转过程中对去除硫含量的影响情况等进行了整理提炼，并最终绘制了相关文献研究结论的归纳分析表（见表 4）。实际上，从表 4 中本文不仅能够初步了解到影响辛烷值损失的潜在因素，而且这些基于大量现实生产实践所总结出的影响因素也可以较好地为本文提供影响辛烷值损失的核心指标。此外，以上研究还能提升本文在汽油去除硫含量的过程中是如何影响到辛烷值损失的具体作用机理的理解（如烯烃在脱硫装置运行过程中究竟是如何既影响到产品硫含量，又影响产品辛烷值的？）。总体而言，在对相关已有文献进行整理归纳的基础上，本文初步找到了 11 个对辛烷值损失可能具有较大影响的因素，如反应温度、反应压力、原料汽油硫含量、氢油比等因素（见表 4）。

表 4　相关文献研究结论归纳分析

影响因素	选取指标名称	影响方向	原理概述	脱硫率影响	参考文献
反应温度	R-101 温度	+	反应温度升高，增大烯烃的饱和程度，产品辛烷值的损失增大	—	赵小燕等（2015）[2]
	反应器温度			—	田勇震等（2019）[3]
				选择性加氢脱硫反应器 R-201 反应温度升高，反应速度增大，有利于脱硫反应进行	赵小燕等（2015）[2]
		-	提高反应温度可以抑制烯烃饱和反应，产品辛烷值损失值减小	产品脱硫率随着温度升高先增大后减小，脱硫率最高的温度点在 427℃左右效果最好	马强和赵昌明（2020）[4]
		未涉及	—	提高反应器出口温度到 426℃时，脱硫反应速率也会相应增加，超过 426℃时，脱硫率将会下降	周欢等（2019）[5]

反应压力	反应系统压力/反应器顶部压力	**+**	增加反应压力将导致精制汽油中的烯烃占比减少，辛烷值损失增加	增大反应压力也会增大脱硫率，产品硫含量会降低	马强和赵昌（2020）[4]、周欢等（2019）[5]
原料汽油硫含量	原料汽油硫含量	**U**	当催化裂化汽油硫含量较高或过低时，汽油辛烷值损失大	—	简建超等（2017）[6]
氢油比	氢油比	**+**	增大氢油使烯烃加氢饱和形成烷烃的速率增加，产品辛烷值损失增大	根据孙同根操作法，氢油比在 0.18（mol/mol）情况下，脱硫和辛烷值损失效果最好	马强和赵昌明（2020）[4]
				—	周欢等（2019）[5]
质量空速	反应器质量空速	**-**	增大质量空速减缓脱硫反应以及烯烃加氢反应，使辛烷值损失降低	脱硫效率也会减小，产品硫含量升高	马强和赵昌明（2020）[4]
				精制汽油硫含量升高	周欢等（2019）[5]
待生吸附剂的持硫率、持碳率	焦炭, wt%、S, wt%	**-**	待生、再生吸附剂的持硫率、持碳率越低，烯烃加氢反应程度越大，辛烷值损失增大	—	马强和赵昌明（2020）[4]
再生剂的持硫率				—	周欢等（2019）[5] 于善宝等（2018）[7]
稳定汽油中 C4 组分的流失	待定	**+**	降低稳定汽油中的 C4 组分流失，可以提高辛烷值	—	马强和赵昌明（2020）[4]
剂油比	待定	**+**	增大剂油比时汽油辛烷值增大	—	柳文等（2017）[8]
轻汽油与重汽油切割比	待定	**-**	分割精度低时，轻组分被加氢饱和，造成重汽油辛烷值损失增大	—	赵小燕等（2015）[2]
	待定	**U**	轻汽油与重汽油切割比过高或过低，均导致辛烷值损失	—	田勇震等（2019）[3]

注：“—”表示该研究未涉及此内容。

在总结已有文献，初步确定影响辛烷值损失的核心影响因素的基础上，本文结合附件四“354 个操作变量信息”，将以上指标进行一一匹配，由于附件四中可能存在未涵盖相关指标对应位点以及部分位点并未标注中文名称等情况，秉持谨慎性原则，本文最终成功匹配到如反应温度、质量空速、反应压力、氢油比等在内的 5 个核心操作变量影响因素（见表 5）[1]。同时，由于存在单个指标对应多个位点的情况，为避免信息重复，导致在后续建模过程中浪费样本自由度的情况，本文根据同一指标下各个位点变量的方差最小原则确定最终的衡量指标（实际上，从 Stata 显示的相关变量统计特征结果来看，处于同一指标下的各个位点的各项统计信息及其相似，故本文认为通过方差最小的原则所遴选出最终的衡量指标具有合理性和可行性）。进一步地，本文认为基于已有文献研究基础所筛选出影响辛烷值损失的核心操作变量指标具有重要的经验和现实意义，因此，在后续数据降维处理中，将对以上核心指标采取保护性原则，即不再统一纳入到待降维样本数据集中。

[1] 实际上，通过文献整理分析，本研究还确定了待生吸附剂持硫和持碳、再生吸附剂持硫和持碳 4 个影响辛烷值损失大小的核心因素。

表 5　辛烷值损失核心影响因素指标确定

影响因素	对应变量（组）位号	中文名称	最终确定指标
反应温度	S-ZORB.TE_2004.DACA	R-101 床层下部温度	R-101 床层下部温度
	S-ZORB.TE_2003.DACA	R-101 床层下部温度	
	S-ZORB.TE_2002.DACA	R-101 床层中部温度	
	S-ZORB.TE_2001.DACA	R-101 床层中部温度	
	S-ZORB.TE_2104.DACA	R-101 顶反应产物出口管温度	
质量空速	S-ZORB.CAL.SPEED.PV	反应器质量空速	反应器质量空速
反应压力	S-ZORB.PT_2101.PV	反应器顶部压力	反应系统压力
	S-ZORB.PC_1202.PV	反应系统压力	
氢油比	S-ZORB.CAL_H2.PV	氢油比	氢油比
原料汽油硫含量	S-ZORB.AT_1001.DACA	原料汽油硫含量	原料汽油硫含量

在依据文献筛选出核心指标并对其进行保护的基础上，接下来可对余下 354 个操作变量进行降维处理。首先，在对附件一与附件四中各操作变量的位点及中文名称进行匹配后发现，原始数据集中存在部分变量的中文名称或单位缺失的情况，由于缺少这些位点的基本含义信息，若继续纳入此类变量到后续研究中，将会有较大程度对其基本概念界定不准，以及给后续优化操作方案设计带来干扰等缺陷，因此，秉持谨慎性原则，本文将在进行降维处理之前先剔除掉此类指标。

Step4: 依数据处理原则对上述步骤处理后的数据进行进一步处理。

在对各操作变量进行以上基本处理后（即对核心指标采取稳健性的保护措施、剔除信息不完整变量），为确保待降维数据集基本统计信息无误，本文再次依据附 5.1.2 节数据处理基本原则，对附件一中余下的操作变量进行数据预处理。其中，除对缺省值低于 30%的变量以均值进行替换外，还根据缺省值大于 30%的原则剔除掉了部分操作变量（见表 6）。

表 6　缺省值超过 30%操作变量基本情况

位点	中文名称	空值数（个）
S-ZORB.FC_2301.PV	D105 流化氢气流量	145
S-ZORB.FT_1501.PV	新氢进装置流量	288
S-ZORB.FT_1004.PV	3#催化汽油进装置流量	126
S-ZORB.FT_9101.PV	污油出装置	134
S-ZORB.FT_1002.PV	1#催化汽油进装置流量	137
S-ZORB.FC_1202.PV	D121 顶去放火炬流量	219
S-ZORB.FC_3103.PV	再生冷氮气流量	214
S-ZORB.FT_1501.TOTAL	新氢进装置流量	123
S-ZORB.FT_2803.DACA	紧急氢气去 D-102 流量	297
S-ZORB.TEX_3103A.DACA	EH-102 加热元件/A 束温度	214

5.2.2 基于主成分分析的数据降维处理

经过 5.2.1 节对附件一中操作变量进行筛选和剔除后，可以对余下的操作变量数据集进行降维处理。事实上，当前对对变量进行的降维方法有很多，其中比较常用的主要有层次分析法（AHP）、熵值法、聚类分析、因子分析（Factor Analysis，FA）以及主成分分析（Principal Component Analysis，PCA）等。从以上方法各自的优势与不足来看：层次分析法主要依据专家打分的原则以生成最终综合评价指标中各指标的权重，而这往往带有个人

主观偏好，因此存在较明显的缺陷。熵值法主要根据各项指标的变异程度，利用信息熵计算出各指标的权重，虽然其权重确定方式相较于 AHP 而言更加客观，但由于不能从其权重中直观反映出原始变量的基本性质，因此也不适用于本研究。而类分析虽然可以依据数据本身所具有的特征对数据进行分组归类，且通常能够得到各个具有相似特征样本数据集，但该方法在处理较大样本量的数据集时，其聚类效果往往差强人意。此外，因子分析作为一种数据简化方法，可以依据众多变量内部之间的依赖关系（相关系数），并用少数几个假想变量（因子）作为原始数据的替代指标，通过因子分析所提取的因子通常可以找到比较明显的现实含义，但由于其特征因子在拟合原始变量的过程中存在不可避免的方差损失，故其仅能够在一定程度上反映出原始变量的真实信息。而对于主成分分析而言，其主要通过对原始变量集进行数学上的线性变换，转换为一组不相关的新变量（各新变量直接属于正交关系），在能够实现对多变量进行有效降维的同时，还能够最大程度地保留原有数据的信息。综上，本文认为主成分分析更加适用于本研究需要，故接下来将采用主成分分析法对余下操作变量进行降维处理。

参考司守奎和孙玺菁（2011）[9]603-604 对主成分分析法的基本介绍，该方法基本思路如下：假定进行主成分分析的指标变量有m个：$x_1, x_2, \dots, x_m$，共有n个评价对象，第i个评价对象的第j个指标的取值为a_{ij}。

首先，对原始数据进行标准化处理，将各指标值a_{ij}转换成标准化指标$\widetilde{a_{ij}}$：

$$\widetilde{a_{ij}} = \frac{a_{ij} - \mu_j}{s_j},\ (i = 1,2,\dots,n; j = 1,2,\dots,m) \tag{1}$$

其中，$\mu_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$，$s_j = \frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \mu_j)^2\ (j = 1,2,\dots,m)$，即$\mu_j$、$s_j$为第$j$个指标的样本均值和样本标准差。

其次，计算变量间的相关系数矩阵：

$$R = (r_{ij})_{m\times n} = \left(\frac{\sum_{k=1}^{n} \widetilde{a_{ki}} \cdot \widetilde{a_{kj}}}{n-1}\right)_{m\times n} \tag{2}$$

其中，r_{ij}是第i个指标与第j个指标的相关系数，$r_{ii} = 1$，$r_{ij} = r_{ji}$。

再次，构造主成分：通过计算特征值$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$，以及对应的特征向量$u_1, u_2, \dots, u_m$，其中$u_j = (u_{1j}, u_{2j}, \dots, u_{mj})^T$，最终可由特征向量组成$m$个新的指标变量：

$$\begin{cases} y_1 = u_{11}\widetilde{x_1} + u_{21}\widetilde{x_2} + \cdots + u_{m1}\widetilde{x_m} \\ y_2 = u_{12}\widetilde{x_1} + u_{22}\widetilde{x_2} + \cdots + u_{m2}\widetilde{x_m} \\ \cdots \\ y_m = u_{1m}\widetilde{x_1} + u_{2m}\widetilde{x_2} + \cdots + u_{mm}\widetilde{x_m} \end{cases} \tag{3}$$

其中，y_1为第 1 主成分，y_2为第 2 主成分，以此类推。

然后，根据特征值$\lambda_j (j = 1,2,\dots,m)$计算得到各主成分的信息贡献率以及累积贡献率：

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{m}\lambda_k},\ (i, j = 1,2,\dots,m) \tag{4}$$

$$\alpha_p = \frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{m}\lambda_k} \tag{5}$$

其中，当α_p接近于 1（如$\alpha_p = 0.80, 0.85, 0.90$或$0.95$时），可以认为其所对应的前$p$个主成分较好地保留了原始数据信息，据此可将其作为原来m个变量的替代指标进行后续分析。

采用 MATLAB 对数据基本筛选和处理后余下的操作变量数据集进行主成分分析降维处理（参见附录 3），表 7 汇报了各主成分对应的信息贡献率和累计贡献率。可见，标准化变量的前 4 个主成分的累计贡献率才超过 50%；然而，式（5）所对累计贡献率的选取准则要求其至少应不低于 80%，而本研究结果显示，当提取前 15 个主成分时，才刚好达到这一阈值；同时，从单个主成分的贡献率来看，自第 18 个主成分的信息贡献率已不到 1%。此外，考虑到后期工程应用的方便，最好确保在进行数据降维后的总变量不超过 30。由于本文前期已通过文献研究法提取了部分已有研究表明影响辛烷值损失的核心影响因素，以及在后续建模过程中还需考虑各样本“原料”、“产品”、“待生吸附剂”以及“再生吸附剂”中的主要变量。因此，结合主成分分析结果已经实际情况，本研究最终选择提取余下操作变量降维后的前 15 个主成分，并结合 5.2.1 节得到的 5 个核心指标以及其他汽油产品处理过程中涉及到的“原料”、“产品”、“待生吸附剂”以及“再生吸附剂”中指标作为后续研究的主要建模变量。

表 7　主成分分析结果

序号	特征值	差值	贡献率	累积贡献率
1	53.90	29.26212114	28.07	28.0731
2	24.64	12.6199932	12.83	40.9056
3	12.02	0.720261703	6.26	47.1651
4	11.30	2.495879993	5.88	53.0495
5	8.80	0.722997345	4.58	57.6339
6	8.08	2.169419249	4.21	61.8418
7	5.91	0.40509072	3.08	64.9198
8	5.50	1.067127488	2.87	67.7868
9	4.44	0.347206642	2.31	70.0980
10	4.09	0.383925246	2.13	72.2283
11	3.71	0.192818297	1.93	74.1587
12	3.51	0.398897536	1.83	75.9887
13	3.11	0.482802712	1.62	77.6109
14	2.63	0.394959499	1.37	78.9817
15	2.24	0.249024361	1.17	80.1467
16	1.99	0.038109849	1.04	81.1821
17	1.95	0.130040354	1.02	82.1976
18	1.82	0.268555429	0.95	83.1453
⋮	⋮	⋮	⋮	⋮
21	1.43	0.099065328	0.75	85.4530
⋮	⋮	⋮	⋮	⋮
29	0.94	0.013708299	0.49	90.0379
⋮	⋮	⋮	⋮	⋮
43	0.47	0.033246487	0.25	95.0767
⋮	⋮	⋮	⋮	⋮
191	0	0	0	100

5.2.3 基于多元非线性回归模型的辛烷值损失影响因素分析

为了验证本研究基于已有文献对核心指标保留以及操作变量降维过程的合理性，通过构建如下多元非线性回归模型，以逐步加入相关变量的方式，对文献研究法保留的关键变量和主成分分析法提取得主要变量的进行具体说明。

$$\begin{cases}\left(RON_{原}-RON_{产}\right)=\beta_0+\beta_1\left(S_{原}-S_{产}\right)+\beta_i X_i+\beta_j X_j^2+\beta_k PCA_k+\varepsilon \\ X=(olef, carb_d, carb_z, sulf_d, sulf_z, temp, speed, press, h_2\_o, sulf_c) \\ PCA=(pca1, pca2, \ldots, pca15) \\ \varepsilon\sim(0,\sigma^2)\end{cases} \tag{6}$$

其中，$\left(RON_{原}-RON_{产}\right)$表示辛烷损失值；$\left(S_{原}-S_{产}\right)$表示硫含量去除值；$olef$为烯烃；$carb_d$和$sulf_d$分别代表待生吸附剂持碳率、持硫率；$carb_z$和$sulf_z$分别代表再生吸附剂持碳率、持硫率；$temp$为反应温度，结合相关文献，最终以 R-101 床层下部温度作为替代指标；$speed$表示质量空速，使用反应器质量空速来衡量；$press$为反应压力，使用反应系统压力作为衡量指标；h_2_o表示氢油比；$sulf_c$表示原料汽油硫含量。

简建超等（2017）[6]研究发现，原料汽油硫含量存在最优值，过高或过低均会对辛烷值造成额外。对此，本研究通过引入该变量的二次项，以验证其基本研究结论，同时也可作为衡量本研究主要建模变量提取优劣的一个参考。表 8 汇报了逐步加入变量过程中多元回归模型的相关结果，整体来看，所提取出的主要变量对于辛烷值损失的拟合度接近 90%，总体表明所寻找到的主要建模变量具有合理性。

此外，将本研究实证回归结果与基于自然实验文献所得到的基本结论进行对比后发现：第一，已有文献表明，反应压力、氢油比的增加均会减少处理过程中辛烷值损失，而反应温度、质量空速、再生吸附剂的持硫率、持碳率等会对辛烷值损失产生负影响。这与本文的研究结论基本一致；第二，汽油进行处理的过程中，烯烃也会对辛烷值损失产生负影响，本文的结果也证明了这一观点；第三，对于简建超等（2017）[6]认为原料汽油硫含量属于一个适度指标的结论，本模型中该变量的二次项回归系数为正但不显著，表明在理论上原料汽油硫含量在一定程度上的确存在一个最优点，即取值过大或过小均会对辛烷值损失造成负面影响；第四，对比模型（3）、（5）、（7）后发现，逐步加入原料汽油硫含量二次项、前 15 个主成分后的模型相较于模型（3）而言并没有在各项统计特征上取得明显的改善。如模型（5）中原料汽油硫含量二次项系数并不显著、模型（7）中仅有少数几个主成分通过最低 10%的显著性水平检验，从侧面表明本文所提取的主要变量具有一定合理性。

可见，本文基于已有文献提取得到的核心解释变量已可在较大程度上满足本研究后续对于模型预测等方面的要求。因此，在后续研究过程中，本文将遵循优先尝试如模型（3）所示核心变量进行建模，如果拟合精度和效果存在不足，再尝试将原料汽油硫含量的平方项以及各个主成分纳入模型中。

表 8　多元非线性回归模型结果

被解释变量 RON 损失	(1)	(2)	(3)	(4)	(5)	(6)	(7)
去除硫含量	-0.035***		-0.022***		-0.022***		-0.011
	(-4.40)		(-2.72)		(-2.68)		(-1.26)
烯烃		-0.002	0.001	-0.002	0.000	-0.005	-0.004
		(-0.57)	(0.18)	(-0.63)	(0.13)	(-1.18)	(-0.89)
		-0.038***	-0.035***	-0.038***	-0.036***	-0.023*	-0.023*

待生吸附剂持碳		(-3.07)	(-2.88)	(-3.11)	(-2.89)	(-1.78)	(-1.78)
待生吸附剂持硫		0.025**	0.025**	0.025**	0.025**	0.019	0.018
		(2.09)	(2.11)	(2.11)	(2.12)	(1.43)	(1.35)
再生吸附剂持碳		0.073***	0.071***	0.074***	0.072***	0.062***	0.061***
		(4.17)	(4.12)	(4.20)	(4.12)	(3.21)	(3.15)
再生吸附剂持硫		-0.041***	-0.039***	-0.041***	-0.039***	-0.036***	-0.035***
		(-3.02)	(-2.93)	(-3.01)	(-2.93)	(-2.71)	(-2.63)
反应温度		0.010*	0.010*	0.010*	0.010*	0.027	0.023
		(1.92)	(1.93)	(1.91)	(1.93)	(0.87)	(0.76)
质量空速		-0.020	-0.020	-0.018	-0.019	-0.016	-0.016
		(-1.02)	(-1.04)	(-0.89)	(-0.96)	(-0.28)	(-0.27)
反应压力		0.507	0.552	0.539	0.571	2.587*	2.531*
		(1.26)	(1.38)	(1.32)	(1.41)	(1.87)	(1.83)
氢油比		2.719***	2.739***	2.750***	2.758***	-0.880	-0.576
		(3.44)	(3.50)	(3.47)	(3.51)	(-0.66)	(-0.43)
原料汽油硫含量		-0.001***	-0.000**	-0.001	-0.001	-0.001	-0.001
		(-2.90)	(-2.50)	(-1.57)	(-1.23)	(-1.27)	(-1.11)
原料汽油硫含量^2				0.001	0.000	0.001	0.000
				(0.56)	(0.34)	(0.57)	(0.44)
pca1						-0.017**	-0.015**
						(-2.32)	(-2.05)
pca2						0.005	0.004
						(0.58)	(0.51)
pca3						-0.006	-0.006
						(-0.65)	(-0.60)
pca4						-0.017	-0.017
						(-1.32)	(-1.34)
pca5						0.027**	0.025**
						(2.39)	(2.18)
pca6						0.020	0.021
						(0.93)	(0.98)
pca7						0.014**	0.012*
						(2.38)	(1.90)
pca8						0.017	0.017
						(1.44)	(1.40)
pca9						-0.003	-0.003
						(-0.36)	(-0.38)
pca10						0.005	0.005
						(0.72)	(0.75)
pca11						0.007	0.007
						(0.74)	(0.81)
pca12						0.006	0.006
						(0.84)	(0.78)
pca13						-0.007	-0.005
						(-0.78)	(-0.64)
pca14						0.012	0.012
						(1.27)	(1.23)
pca15						0.004	0.005
						(0.44)	(0.48)
_cons	1.396***	-4.489*	-4.593*	-4.536*	-4.621*	-15.137	-13.666
	(40.83)	(-1.80)	(-1.86)	(-1.82)	(-1.87)	(-1.23)	(-1.11)
N	325	325	325	325	325	325	325
R-Square	0.857	0.895	0.813	0.896	0.814	0.890	0.894
BIC	-52.95	-52.38	-54.21	-46.92	-48.55	-0.65	3.39

注：1. 括号内为 t 统计量；2. AIC 信息准则虽然为模型选择提供了有效的规则，但存在当样本容量很大时，即在 AIC 准则中拟合误差提供的信息就要受到样本容量的放大，而参数个数的惩罚因子却和样本

容量没关系（一直是 2）的明显缺陷，故当样本容量很大时，使用 AIC 准则选择的模型不收敛与真实模型，因为它通常比真实模型所含的未知参数个数要多。因此，本文主要汇报了 Schwartz 在 1978 年根据 Bayes 理论提出的 BIC（Bayesian InformationCriterion）贝叶斯信息准则，也称为 SBC 准则，据此可弥补 AIC 的明显不足。其定义为：$BIC = ln(模型中参数的个数n) - 2ln(模型的极大似然函数值)$。

5.3 问题三：辛烷值损失预测模型的建立与求解

以问题二所得的主要建模变量数据集为基础，从数据挖掘技术方法出发，通过构建合理的数理模型，运用机器学习算法建立辛烷值（RON）的损失预测模型。同时，考虑到应用单一的方法进行预测时可能存在片面性和难以确保准确性等弊端，因此，有必要建立含支持向量机、随机森林以及岭回归等在内的多种模型与算法。此外，为进一步提高辛烷值损失预测模型的精度，本文还基于集成学习的思想，考虑建立多种机器学习算法融合的辛烷值损失预测复合模型。最后，为确保模型预测的可靠性，还将对所采用的各种预测模型与算法进行误差对比与评价，进而确定相对误差最小的辛烷值损失预测模型，为后续研究做准备。

尽管本文已在 5.2.3 节尝试构建了多元回归模型对辛烷值损失情况进行预测，且各项统计特征表明模型拟合效果较高（如 R 平方普遍达到 80%以上），但由于基于回归的方式实质上只是寻找所有观察样本中各变量的一个“平均”情况，有可能并不能较好地反映出装置实际运转过程中各观察样本因各变量取值不同而潜在存在的差异性（即个体效应）。同时，应用单一方法进行预测时还有极大可能存在片面性和难以确保准确性等弊端。因此，为更加有效地对该企业历史数据中辛烷值的损失情况展开预测，本文进一步从数据挖掘技术出发，采用支持向量机回归、随机森林、岭回归以及集成学习思想下的 Stacking 策略对以上方法进行融合的多种机器学习算法，建立相应的辛烷值损失预测模型（如图 2 所示）。并通过最终结果间的对比，对以上算法的拟合情况进行相互间的检验与评价。

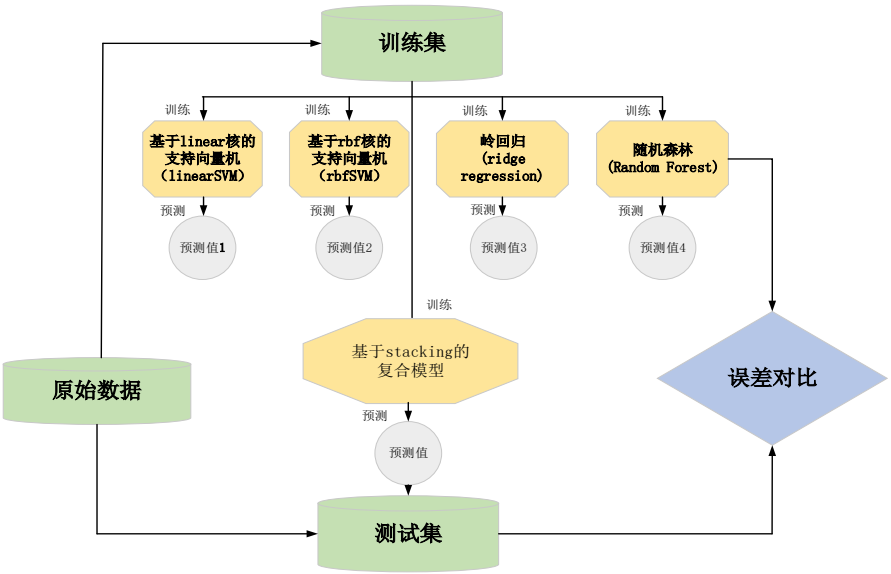

图 2　构建辛烷值损失预测模型思路图

5.3.1 数据预处理

本文使用部分历史数据作为训练集来训练模型，然后用测试集上的误差作为最终模型在应对现实场景中的泛化误差，将各模型误差进行对比。通常需要在开始构建模型之前把数据集进行划分，防止数据窥探偏误。

训练集：本问题中原始数据为前期基于文献研究法得到的核心变量数据、根据数据处理原则和主成分分析最终得到的主要变量数据，在原始数据中提取除测试集以外的 275 样本数据作为训练集。

测试集：选取原始数据，样本数据获取时间最近的前 50 个样本数据最为测试集。

5.3.2 基于机器学习算法的预测模型构建

模型一：随机森林算法

随机森林作为一种以决策树为基分类器的集成算法，通过组合多棵独立的决策树后根据投票或取均值的方式得到最终预测结果的机器学习方法，往往比单棵树具有更高的准确率和更强的稳定性。如其能够处理很高维度的数据，并且不用做特征选择；在创建随机森林的时候，模型泛化能力强；对于不平衡的数据集可以平衡误差；且即便存在较大部分数据特征遗失的情况下，也可以较好地维持准确度。随机森林算法的基本思路如下[10]：

Step1. 给定包含 N 个样本的数据集，经过 n 次有放回的随机抽样操作，得到 T 个含 n 个训练样本的采样集。

Step2. 对每个采样集，从所有属性中随机选择 k 个属性，选择最佳分割属性作为节点建立 CART 模型，最终建立拥有 T 个 CART 模型的随机森林，其中$k = \log_2 d$，（d 为样本所有属性）。

Step3. 将模型用于测试集，对于测试每个样本会有 T 个预测值，对分类任务使用简单投票法确定该样本最终预测值，进而对回归任务使用简单平均法确定最终预测值。

本文使用随机森林预测模型对本文的训练集进行训练后，得到一组预测数据，将本文预测出的辛烷值损失与实际值进行对比可以得到图 3 的随机森林预测结果。结果中绿色实心点代表实际值，红色实心点为预测值。从图中可以看出基于随机森林算法预测出来的结果随着测试次数的增加波动较为平缓，而实际值的波动较大，但其中有部分预测结果与实际值差距较小，如当测试次数为 40 次之左右时，两个值较为接近，说明随机森林模型预测结果具有一定的可行性。

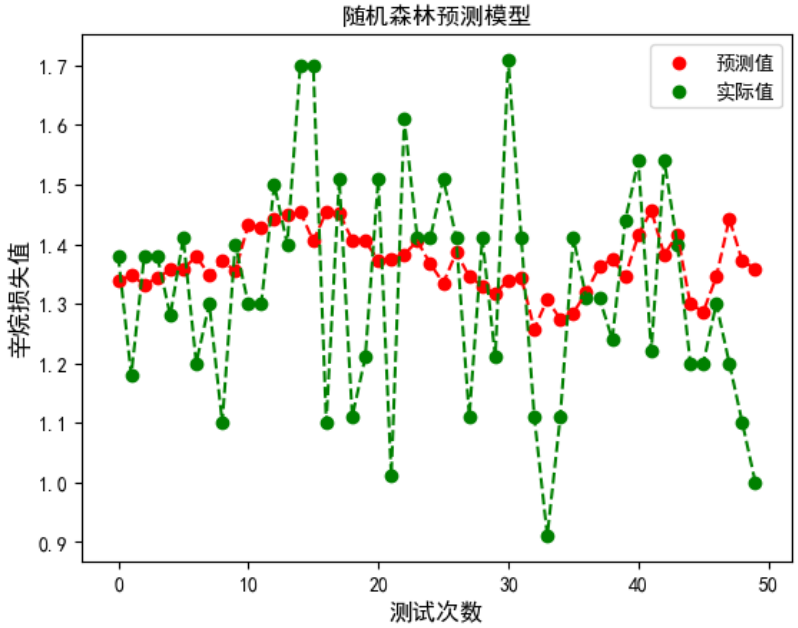

图 3　随机森林预测结果

模型二：基于线性核（Linear Kernel）的支持向量机算法

线性核（Linear Kernel）主要用于线性可分的情况，特征空间到输入空间的维度是一样的，在原始空间中寻找最优线性分类器，最大的优势为参数少速度快。对于线性可分数据，其分类效果很理想。基于线性核的支持向量机参数设置如下[11]：

- 惩罚系数 C：在模型准确率与模型复杂度之间取得一个平衡。当 C 较大时，支持向量机的决策间隔会较小；而当 C 较小时，则会牺牲一定准确度。
- 距离误差 epsilon：训练集中的样本需满足特定的约束条件。
- 样本权重 class weight：指定样本各类别的权重，防止训练集某些类别的样本过多。
- 正则化 penalty：可以选择‘l1’即 L1 正则化或者‘12’即 L2 正则化。
- 布尔变量 dual：控制是否使用对偶形式来优化算法。
- 损失函数 loss：具有特定的向量机回归的损失度量标准。

同时设定目标函数的原始形式如下：

$$min\frac{1}{2}\left|\left|w\right|\right|_2^2, s.t\ \ |y_i - w\cdot\phi(x_i) - b| \le \epsilon(i = 1,2,\dots m) \tag{7}$$

式（7）需要对每个样本(x_i, y_i)加入松弛变量$\xi_i \ge 0$，对两个不等式两边都需要松弛变量，定义为$\xi_i^{\vee}$，$\xi_i^{\wedge}$。在加入松弛变量后，本文回归模型的损失函数度量转变为：

$$min\frac{1}{2}\left|\left|w\right|\right|_2^2 + \mathrm{C}\sum_{i=1}^{m}(\xi_i^{\vee} + \xi_i^{\wedge}) \tag{8}$$

$$\mathrm{s.t.} -\epsilon - \xi_i^{\vee} \le y_i - w\cdot\phi(x_i) - b \le \epsilon + \xi_i^{\wedge}$$

$$\xi_i^{\vee} \ge 0, \xi_i^{\wedge} \ge 0, (i = 1,2,\dots m)$$

将训练集导入到基于线性核（Linear Kernel）的支持向量机预测模型中，最终得到其拟合效果图（见图 4）。可见，应用基于线性核（Linear Kernel）的支持向量机对该企业历史数据进行预测时，同样发现预测值波动趋势较实际值更为平缓，但仍然有部分值基本重合，且重合度较高，特别是当测试次数在 20 次上下时，因此可以认为基于线性核（Linear Kernel）的支持向量机预测模型预测结果的准确性较高，预测模型具有可信度。

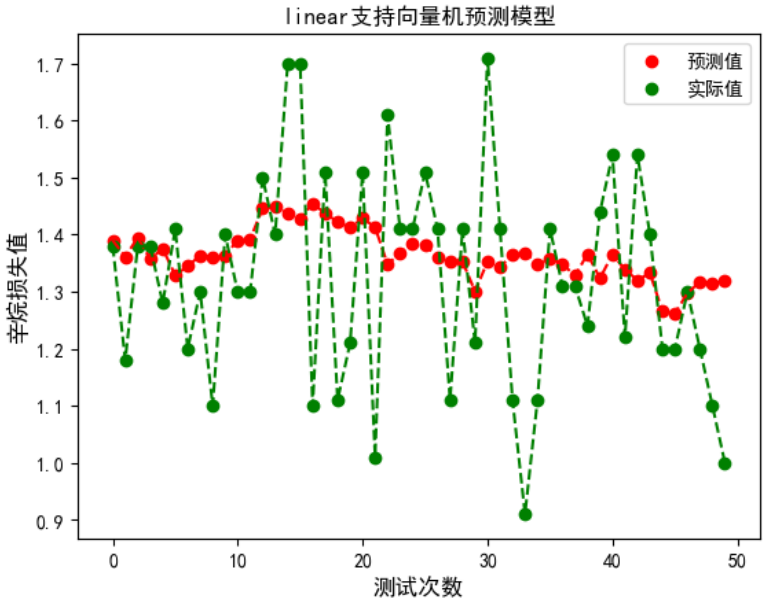

图 4　Liner 支持向量机预测结果

模型三：岭回归算法

岭回归实质上是一种改良的最小二乘估计法，即通过放弃最小二乘法的无偏性、以损失部分信息、降低精度为代价获得回归系数更加符合实际、更加可靠的回归方法。实质上，由于岭回归在变量增加了一个小的平方偏差因子（正则项），这种平方偏差因子向模型中引入少量偏差，但大大降低了方差。此外，其还具有可以解决特征数量比样本量多的情况、可以对一个模型增加偏差的同时减少方差等优点。

岭回归中核心参数——岭参数 k 的一般选择原则为使各回归系数的岭估计基本稳定、岭估计的符号合理、回归系数没有不合乎经济意义的绝对值以及残差平方和增加程度不大等；而对于扰动因子 λ 的选择则一般通过观察选取喇叭口附近的值。具体来看，岭回归的具体分析步骤主要有 2 步：

Step1：结合岭迹图寻找最佳 K 值。
Step2：输入 K 值进行回归建模。

图 5 汇报了基于岭回归的产品辛烷值损失预测结果，结果表明预测值和实际值得重合度较高，且波动的方向较为一致，这说明岭回归预测结果误差较小，运用次预测模型对辛烷值损失进行预测是可行的。

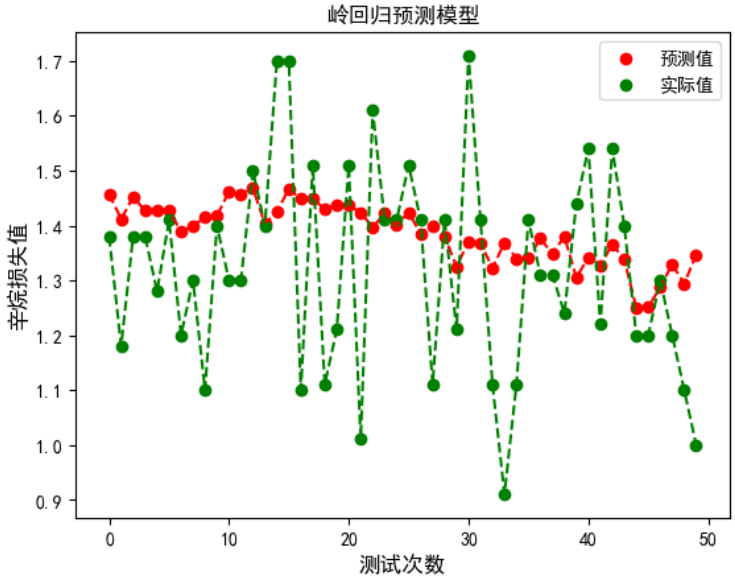

图 5　岭回归预测结果

模型四：复合预测模型算法

基于集成学习的思路，通过合并前文所构建的单一的随机森林算法、基于线性核（Linear Kernel）的支持向量机算法以及岭回归算法来提高机器学习性能。其中，集成学习的基本方法主要有 Bagging（主要用于减少方差）、Boosting（可以减少模型偏差）以及 Stacking（主要用于提升预测结果）。从理论上看，采用复合预测模型算法构建的预测模型提出能相较单一模型得到更好的拟合结果。

由于本研究主要对产品辛烷值的损失情况进行预测，因此选择 Stacking 融合方法对本文构建的 4 个单一模型进行合并。Stacking 算法的基本原理如下[12]：

Step1：对于 Model1，将训练集 D 分为 k 份，对于每一份，用剩余数据集训练模型，然后预测出这一份的结果；

Step2：重复上面步骤，直到每一份都预测出来，得到次级模型的训练集；

Step3：得到 k 份测试集，平均后得到次级模型的测试集；

Step4：对于 Model2、Model3…重复以上情况，得到 M 维数据；

Step5：选定次级模型，进行训练预测，一般最后一层用的是 LR。

基于 Starcking 策略复合模型预测结果如图 6 所示，可见，采用复合预测模型算法对辛烷值损失的拟合预测值在波动方向上和实际值具有更高程度的同步性。映证了相较于单一预测模型而言，采用复合预测模型可更好地实现对历史数据中辛烷值损失情况的预测。

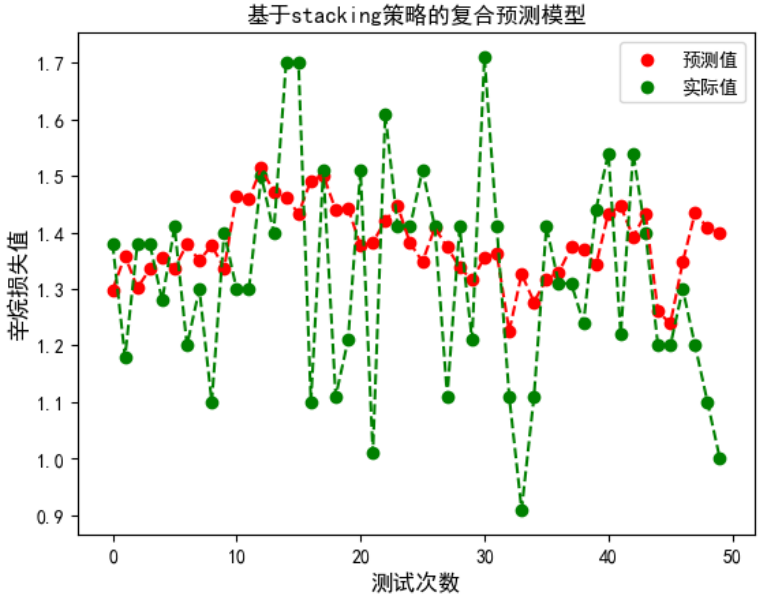

图 6　基于 Starcking 策略复合模型预测结果

5.3.3 预测模型对比与评价

尽管 5.3.2 节已经分别对各预测模型的拟合效果进行了简要评价，但却没有更科学合理地对其进行横向对比，亦即无法准确判断出各个预测模型算法的优劣性；而反映样本拟合效果的样本均方误差则可以用来较好地评价一个模型的拟合效果优劣。从各个预测模型算法的样本均方误差结果来看，随机森林预测模型、基于线性核（Linear Kernel）的支持向量机预测模型、岭回归预测模型以及基于 Starcking 策略的复合预测模型的样本均方误差依次为 0.033522、0.033478、0.0352379 和 0.035339。可见，各模型的样本均方误差较为接近。相较而言，基于线性核（Linear Kernel）的支持向量机预测模型样本均方误差最小。

此外，本文还分别汇报了各个预测模型对产品辛烷值损失预测的测试误差情况。可以看出，4 个模型的预测误差随测试次数的波动都表现得较为一致（图 7 左）。而从样本累积误测来看（图 7 右），随着测试次数的增加，以上四种预测模型算法的结果依然较为接近。可见，本文所构建的基于机器学习算法的多种预测模型之间并不严格存在绝对的优劣，且预测结果相对一致，也间接映证了所构建预测模型的稳健性。

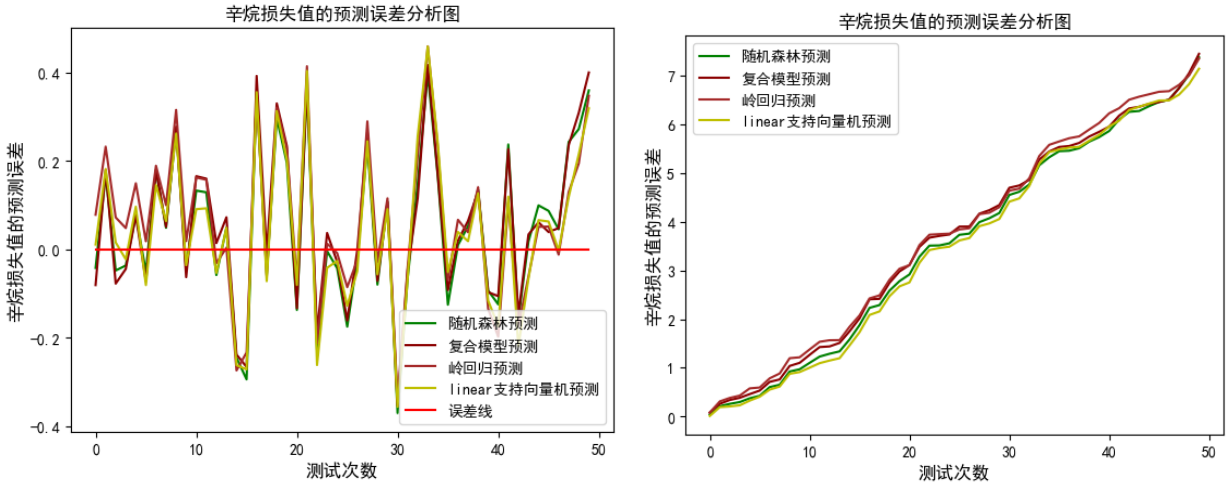

图 7 各预测模型算法的均方误差及累计均方误差对比图

5.4 问题四：主要变量操作方案的优化

5.4.1 两种评价指标引入与“理想样本”选择

由于本文的主要目的是优化操作使产品硫含量在不大于$5\mu g/g$条件下使辛烷值损失尽可能降低，优化操作的主要观测指标是辛烷值损失值和产品硫去除量，因此，本文提出“辛烷值损失率”以及“产品硫去除率”两个概念作为选择“理想样本”的初步尝试。其中，辛烷值损失率=（原料辛烷值-产品辛烷值）/原料辛烷值，产品硫去除率=（原料硫含量-产品硫含量）/原料硫含量。

本文首先基于“产品硫去除率”对 325 个历史观察样本进行降序排序；然后，同样对 325 个历史样本以“辛烷值损失率”进行升序排列。然而，通过对采用以上两个指标分别进行排序后的样本进行对比发现（表 9）：两者之间在前 20 个的样本中仅有 194 号样本均被涵盖在内（序号分别为 17 和 8）；而其他的样本均没有同时落入到其中。这与本研究最初设想通过以上方法找到历史样本数据中处于“绝对占优”的理想样本，并考虑将其对应的操作变量取值情况作为最优取值参考的思路变得可行性不高。因为即便可以将 194 号样本作为最初思路中的“绝对占优”理想样本对待，但由于其在两种排序方案中的排名并非很靠前，若纯粹使用该样本所对应的操作变量参数取值设定作为其他样本的潜在改进方向，则很容易导致结果出现较大偏差。可见，基于以上两个单一指标分别进行观测以发现“理想样本”的方案不具备现实可取性和可操作性。因此，采用一种可以综合以上评价要素的更加客观的评价模型对历史数据中各观察样本进行有效性评价显得尤为重要。

表 9 产品硫去除率与辛烷值损失率排序比较

序号	产品硫去除率降序排列					辛烷值损失率升序排列				
	样本号	产品硫去除率	辛烷值损失率	产品硫含量	RON损失	样本号	产品硫去除率	辛烷值损失率	产品硫含量	RON损失
1	238	99.092%	1.329%	3.20	1.20	142	97.914%	0.234%	3.20	1.20
2	315	99.053%	1.272%	3.30	1.15	185	97.293%	0.494%	5.10	1.15
3	318	99.053%	1.394%	3.30	1.25	179	98.142%	0.599%	4.30	1.25
4	240	99.039%	1.436%	3.30	1.30	167	98.742%	0.681%	3.20	1.30
5	308	99.034%	1.389%	3.20	1.25	292	94.843%	0.734%	6.10	1.25

6	310	99.034%	1.421%	3.20	1.28		242	98.382%	0.777%	4.80	1.28
7	264	99.014%	1.605%	3.20	1.44		195	97.336%	0.785%	6.50	1.44
8	210	98.986%	1.122%	3.60	1.01		194	98.959%	0.787%	3.20	1.01
9	158	98.978%	1.256%	3.20	1.12		265	97.597%	0.825%	7.80	1.12
10	156	98.978%	1.470%	3.20	1.32		139	98.144%	0.916%	3.20	1.32
11	157	98.978%	1.571%	3.20	1.42		270	97.853%	0.935%	5.40	1.42
12	159	98.976%	1.365%	3.20	1.22		129	98.710%	1.004%	3.20	1.22
13	207	98.970%	1.244%	3.20	1.11		220	98.167%	1.018%	4.80	1.11
14	208	98.970%	1.577%	3.20	1.41		34	97.778%	1.033%	3.20	1.41
15	155	98.960%	1.594%	3.20	1.42		172	98.624%	1.034%	3.20	1.42
16	154	98.960%	1.603%	3.20	1.42		171	98.624%	1.055%	3.20	1.42
17	194	98.959%	0.787%	3.20	0.71		190	98.084%	1.074%	3.20	0.71
18	192	98.959%	1.720%	3.20	1.51		87	97.516%	1.082%	7.70	1.51
19	112	98.936%	1.540%	3.20	1.40		90	98.745%	1.083%	3.20	1.40
20	113	98.903%	1.326%	3.30	1.20		170	98.624%	1.105%	3.20	1.20
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24	201	98.892%	1.568%	3.20	1.41		210	98.986%	1.122%	3.60	1.41
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
76	205	98.713%	1.572%	4.00	1.41		207	98.970%	1.244%	3.20	1.41
77	129	98.710%	1.004%	3.20	0.90		150	96.786%	1.247%	3.20	0.90
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
108	172	98.624%	1.034%	3.20	0.92		305	97.684%	1.301%	5.90	0.92
109	171	98.624%	1.055%	3.20	0.94		89	98.839%	1.304%	3.20	0.94
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
183	3	98.192%	1.521%	3.20	1.38		310	99.034%	1.421%	3.20	1.38
184	220	98.167%	1.018%	4.80	0.92		95	98.884%	1.421%	3.20	0.92
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
226	142	97.914%	0.234%	3.20	0.20		156	98.978%	1.470%	3.20	0.20
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
244	34	97.778%	1.033%	3.20	0.91		189	98.084%	1.535%	3.20	0.91
245	54	97.778%	1.332%	3.20	1.20		112	98.936%	1.540%	3.20	1.20
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
261	87	97.516%	1.082%	7.70	0.98		29	98.447%	1.572%	3.20	0.98
262	143	97.477%	1.702%	3.20	1.51		208	98.970%	1.577%	3.20	1.51
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
270	267	97.406%	1.387%	7.80	1.24		155	98.960%	1.594%	3.20	1.24
271	15	97.355%	1.936%	3.20	1.70		154	98.960%	1.603%	3.20	1.70
272	49	97.348%	1.220%	7.00	1.10		264	99.014%	1.605%	3.20	1.10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
321	292	94.843%	0.734%	6.10	0.65		76	97.333%	2.029%	3.20	0.65
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
325	287	87.654%	1.209%	10.60	1.08		153	97.472%	2.080%	3.20	1.08

注：上表中以红色字体标注了历史数据中“产品硫含量”大于$5\mu g/g$的情况；以黄色和绿色色块标注了分别在两种排序情况下处于排序前 20 名的样本在另一种排序下的对应情况。限于篇幅，更加详细的对照情况请参照“附件一：325 个样本数据(附件三已追加)(数据预处理).xIsx”中“产品硫去除率与辛烷值损失率排序比较”工作表。

5.4.2 客观综合评价模型的对比选择与建模

由于数据来源均为某石化企业的催化裂化汽油精制脱硫装置运行 4 年的真实历史数据，即所有样本数据均可在实际操作中得到实现。对此，本文可通过更加客观合理的综合评价模型，对各观察样本的有效性进行排序，并以排序靠前的历史观察样本对应的主要操作变量基本情况作为主要变量操作方案优化的切入视角，符合现实可操作性地提出相应的改进方案。因此，从以上基本思路出发，本文可以本企业的历史数据中综合评价得分较高的样本（如前 5 名或前 10 名）作为标杆，通过观察其各项操作变量的基本取值范围以及统计特征，以此作为综合评价得分靠后样本的基本改进方向。

实际上，结合相关理论以及本文实际，一种基于熵权法改进的 TOPSIS 理想点法以及常用于工程操作及管理效率评价的 DEA（数据包络分析）法均比较适用于对该企业 4 年以来观察样本的有效性评价。然而，通过仔细对比，本文认为，尽管相较于传统 TOPSIS 而言，本文所提出的通过熵权法进行相应指标权重的客观生成改进后的熵权 TOPSIS 模型在很大程度上已经能够实现本研究对 325 个观察数据进行排序的目的，同时也能够满足以得分靠前样本为标杆，对得分较低样本进行主要操作变量优化的的现实要求。但是，相较于熵权 TOPSIS 法而言，基于线性优化模型的 DEA 效率评价模型不仅可以基于研究中对各投入产出变量做出的合理假定（如选择不同的投入（Input Oriented）和产出（Output Oriented）为导向，以及基于催化裂化汽油精制脱硫装置运行中各操作变量对最终辛烷值影响的规模报酬不变（CRS）或可变（VRS）假定），选用合适的 DEA 具体形式，实现对历史数据的评优排序要求，而且能够结合以上假定，测度出各观察样本处于规模效率不变（-）、递增（irs）以及递减（drs）的具体阶段，从而为本研究后续对主要操作变量的优化改进提供更加丰富且可取的信息。此外，DEA 模型还具有测度结果不受统计数据量纲影响，即在评价之前无需进行数据标准化处理的良好优点。

综上，本文最终采用以投入为导向下、规模报酬可变的 DEA-BCC 模型对该企业历史观察样本进行客观评价[2]，其基本思路如下：

假定样本数据由n个决策单元(DMU)，每个决策单元都有m种“输入”和s种“输出”，则$x_{ij}(\ x_{ij}>0)$表示第j个决策单元的第i种输入量，且$i=1,2,\cdots,m;\ j=1,2,\cdots,n$。$y_{rj}(\ y_{rj}>0)$表示第$j$个决策单元的第$r$种输出量，且$r=1,2,\cdots,s$。据此得到各个决策单元的各种输入与输出指标的向量表示：

$$X_j=(x_{1j},x_{2j},\cdots,x_{mj})^T, Y_j=(y_{1j},y_{2j},\cdots,y_{mj})^T \tag{9}$$

相较于 CCR 模型关于固定规模报酬不变的基本假定，本文认为不同的规模报酬同样具有较为明显的效率差异。对此，借鉴 Banker 等（1984）[13]对 CCR 模型加以扩展改进得到的 BCC 模型，即在 CCR 模型的基本约束条件上，增加一个凸性假设条件:

$$\sum_{j=1}^{n}\lambda_j=1 \tag{10}$$

[2] 本文同时汇报了熵权 TOPSIS 模型的相关模型介绍（附录 6:“基于熵权法改进的 TOPSIS 理想点法模型”）、MATLAB 源代码（附录 7：“基于熵权法改进的 TOPSIS 模型的 MATLAB 源代码（Q4_SZF_Topsis.m）”）以及评价结果（附件一：325 个样本数据(附件三已追加)(数据预处理).xlsx”中“熵权 TOPSIS 结果”工作表），感兴趣的读者可自行查阅。

由此建立基于生产可能集下的 DEA-BCC 模型(P_{BCC})，进而在引入投入松弛变量S^-和产出松弛变量S^+后，上述模型可进一步转化为如下对偶问题(D_{BCC})：

$$min\ \ \theta$$

$$s.t.\ \begin{cases} \sum_{j=1}^{n} X_j\lambda_j + S^- = \theta X_0 \\ \sum_{j=1}^{n} Y_j\lambda_j - S^+ = Y_0 \\ \sum_{j=1}^{n} \lambda_j = 1 \\ \lambda_j \geq 0 \\ j = 1,2,\cdots,n \\ S^-, S^+ \geq 0 \end{cases} \tag{11}$$

事实上，若线性规划问题(D_{BCC})存在最优解λ_0、S^{-0}、S^{+0}和θ^0，则：（1）当$\theta = 1$时，决策单元为 DEA 有效；（2）当$\theta <= 1$时，为 DEA 无效。

鉴于催化裂化汽油精制脱硫装置在现实生产实践中，原料、待生吸附剂、再生吸附剂的原有性质总是保持不变的；与此同时，考虑到使用 DEA 模型进行效率评价时，为确保结果可靠，决策单元数量应至少不低于总投入产出指标的 2 倍，且当投入产出指标过多时，极有可能导致有效决策单元过多的情况（即多个决策单元效率值为 1），不符合本文的研究目的和预期。因此，结合相关文献以及本文前期研究结果，在对历史观察样本进行效率评价时，将不考虑优化过程中原料、待生吸附剂、再生吸附剂等性质不会发生变化的指标，进而综合指标数量限制以及最大程度反映评价多维度的原则，最终选择投入产出测度指标如表 10 所示。

表 10　DEA 效率评价模型的投入产出指标体系

指标类型	具体指标	衡量指标	单位	指标属性
投入指标	反应温度	S-ZORB.TE_2003.DACA	℃	-
	质量空速	S-ZORB.CAL.SPEED.PV	h-1	+
	反应压力	S-ZORB.PC_1202.PV	MPa	+
	氢油比	S-ZORB.CAL_H2.PV	%	-
	原料汽油硫含量	S-ZORB.AT_1001.DACA	mg/kg	U
产出指标	产品硫去除率	产品硫去除率	%	+
	辛烷值损失率	辛烷值损失率	%	-
	产品硫含量	产品硫含量	$\mu g/g$	-

值得注意的是，在进行效率评价之前，需对投入产出指标中的负向指标进行取倒数处理，以反应指标对最终效率值大小的真实影响情况。此外，在运用 DEAP 软件测度 DEA-BCC 的效率结果时，还需对其相应的向导文件（Q4_DEA_BCC.ins）进行如下设置：

```
Q4_DEA_BCC.dta          DATA FILE NAME                      //数据源文件
Q4_DEA_BCC.out          OUTPUT FILE NAME                    //结果存放文件
325                     NUMBER OF FIRMS                     //决策单元个数
1                       NUMBER OF TIME PERIODS              //时期数
3                       NUMBER OF OUTPUTS                   //产出变量个数
5                       NUMBER OF INPUTS                    //投入变量个数
0                       0=INPUT AND 1=OUTPUT ORIENTATED     //投入 or 产出导向
```

```
1                    0=CRS AND 1=VRS                                    //规模报酬可变 或 不变
0                    0=DEA(MULTI-STAGE), 1=COST-DEA, 2=MALMQUIST-DEA, 3=DEA(1-STAGE), 4=DEA(2-
STAGE)                                                                  //具体算法
```

通过以上数据处理和模型设定，本文最终得到该企业 4 年以来的 325 个历史观察样本的 DEA-BCC 效率评价结果（见表 5.4.3）。实证结果表明：首先，DEA-BCC 效率评价显示达到 DEA 有效的 21 个样本中，有多达 15 个观察样本分别处于 5.4.1 节中按产品硫去除率与辛烷值损失率单个指标排序的前 20 位之中（样本号依次为 142、152、154、155、189、192、194、210、223、224、238、308、310、315 和 318），这在很大程度上从侧面验证了本研究所采用的 DEA-BCC 效率评价结果具有合理性和可靠性。其次，从 21 个 DEA 有效的样本来看，其实际产品硫含量均介于[3.20,4.10]区间，且带有明显的右偏性质（众数<中位数<平均数），说明只要未来的生产过程中的各核心操作变量达到了本企业历史数据中 DEA 有效样本所对应的阈值，则对于问题中“保证产品硫含量不大于 $5\mu g/g$”的前提将几乎可以完美实现。再次，若以单个 DEA 有效的样本作为排名靠后样本的改进方向，将可能存在较大程度的评价指标偏误，如产品硫去除率区间为$[96.24\%, 99.09\%]$、辛烷值损失率区间为$[0.23\%, 1.72\%]$，因此，为进一步降低“理想条件”确定过程中受单个样本异常值的影响，一种可取的方式是对以上 21 个 DEA 有效的样本取其各变量的均值作为最终排序靠后历史样本改进的目标“理想样本”（如表 11 中“理想样本”行所示）。最后，以本研究所得到的 21 个 DEA 有效样本均值所构造的“理想样本”，进而可按 325 个观察样本的效率得分情况，分别对应“理想样本”的主要操作变量特征进行优化。从理论上看，依据“优化过程中原料、待生吸附剂、再生吸附剂的性质保持不变”的基本假定，对除此以外的其他核心解释变量与“理想样本”进行横向比较，最终可得到每个样本主要操作变量的基本优化方向（如图 8 所示）。

表 11　投入导向与可变成本（VRS）下 DEA-BCC 效率评价结果

效率值排序	样本号	可变成本效率	规模效率	所处阶段	产品硫去除率	辛烷值损失率	产品硫含量	反应温度	质量空速	反应压力	氢油比	原料汽油硫含量
1	11	1.000	1.000	-	96.24%	1.44%	3.20	415.87	2.98	2.26	0.29	86.11
2	12	1.000	1.000	-	97.26%	1.44%	3.20	415.85	2.97	2.26	0.29	92.21
3	73	1.000	1.000	-	98.25%	1.57%	3.20	420.70	3.32	2.18	0.36	216.79
4	142	1.000	1.000	-	97.91%	0.23%	3.20	413.88	4.49	2.25	0.30	393.68
5	152	1.000	1.000	-	97.47%	1.50%	3.20	419.39	4.71	2.30	0.32	92.81
6	154	1.000	1.000	-	98.96%	1.60%	3.20	426.09	3.05	2.22	0.35	173.47
7	155	1.000	0.977	drs	98.96%	1.59%	3.20	422.23	3.97	2.27	0.32	138.95
8	189	1.000	1.000	-	98.08%	1.53%	3.20	421.16	5.39	2.19	0.28	58.90
9	192	1.000	0.991	drs	98.96%	1.72%	3.20	421.06	4.77	2.25	0.25	90.20
10	194	1.000	0.976	drs	98.96%	0.79%	3.20	422.87	4.45	2.28	0.26	240.49
11	210	1.000	0.988	drs	98.99%	1.12%	3.60	423.47	4.39	2.26	0.27	122.25
12	223	1.000	1.000	-	98.78%	1.46%	3.20	423.10	4.09	2.32	0.25	1.65
13	224	1.000	1.000	-	97.89%	1.13%	4.10	422.16	4.02	2.32	0.25	1.58
14	238	1.000	0.961	drs	99.09%	1.33%	3.20	418.62	5.56	2.36	0.25	280.78
15	308	1.000	0.972	drs	99.03%	1.39%	3.20	420.62	5.85	2.28	0.26	312.12
16	310	1.000	0.983	drs	99.03%	1.42%	3.20	421.29	5.59	2.25	0.27	640.51
17	315	1.000	1.000	-	99.05%	1.27%	3.30	420.91	6.58	2.20	0.25	240.03
18	318	1.000	1.000	-	99.05%	1.39%	3.30	420.60	6.68	2.20	0.24	241.44

19	321	1.000	1.000	-	98.82%	1.29%	3.20	420.64	6.69	2.20	0.24	225.02
20	322	1.000	1.000	-	98.82%	1.40%	3.20	420.88	6.48	2.20	0.24	224.01
21	323	1.000	0.999	drs	98.82%	1.51%	3.20	421.17	6.44	2.20	0.24	213.34
理想样本	—	1.000	0.992	/	98.48%	1.33%	3.28	420.57	4.80	2.25	0.28	193.65
22	13	0.999	0.966	irs	93.88%	1.68%	6.30	415.68	3.00	2.26	0.29	91.48
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
98	132	0.980	1.000	-	98.71%	1.69%	3.20	417.25	4.42	2.25	0.31	274.16
99	133	0.980	1.000	-	98.71%	1.47%	3.20	417.91	4.42	2.25	0.30	270.38
100	134	0.980	1.000	-	98.71%	1.47%	3.20	415.38	4.41	2.25	0.31	255.31
101	141	0.980	1.000	-	97.91%	1.14%	3.20	419.27	4.55	2.25	0.30	227.92
102	145	0.980	0.986	irs	97.24%	1.36%	3.50	413.23	4.45	2.25	0.30	180.17
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
198	174	0.964	1.000	-	98.62%	1.35%	3.20	418.04	5.48	2.30	0.25	211.59
199	176	0.964	1.000	-	98.62%	1.27%	3.20	419.73	5.49	2.30	0.25	247.53
200	178	0.964	0.989	irs	97.80%	1.27%	5.10	419.31	5.45	2.30	0.25	253.66
201	180	0.964	1.000	-	98.62%	1.49%	3.20	419.90	5.44	2.30	0.24	263.95
202	181	0.964	0.996	irs	98.53%	1.38%	3.40	416.70	5.41	2.30	0.24	262.77
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
299	252	0.959	1.000	-	98.48%	1.45%	3.20	421.58	5.77	2.36	0.25	281.54
300	253	0.959	0.992	irs	98.15%	1.46%	5.10	420.89	5.56	2.34	0.24	289.90
301	256	0.959	0.990	irs	97.93%	1.42%	5.70	420.07	5.59	2.34	0.25	302.03
302	261	0.959	0.994	irs	98.38%	1.46%	4.90	418.76	5.96	2.36	0.24	299.24
303	266	0.959	0.989	irs	97.84%	1.38%	6.50	418.38	6.07	2.34	0.27	348.39
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
320	40	0.958	1.000	-	98.64%	1.61%	3.20	418.33	3.71	2.36	0.28	255.50
321	41	0.958	1.000	-	98.64%	1.73%	3.20	419.68	3.73	2.36	0.29	208.86
322	201	0.958	1.000	-	98.89%	1.57%	3.20	421.90	4.51	2.33	0.25	222.99
323	204	0.958	1.000	-	98.89%	1.57%	3.20	420.44	4.59	2.35	0.23	243.65
324	205	0.958	0.998	irs	98.71%	1.57%	4.00	421.41	4.69	2.35	0.23	236.36
325	222	0.958	0.995	irs	98.47%	1.57%	4.00	422.61	4.13	2.32	0.25	230.29

注：以红色字体标注了历史数据中“产品硫含量”大于$5\mu g/g$的情况；限于篇幅，更详细的模型评价测度结果见附件“Q4_DEAP”文件夹下“BCC.out”文档。

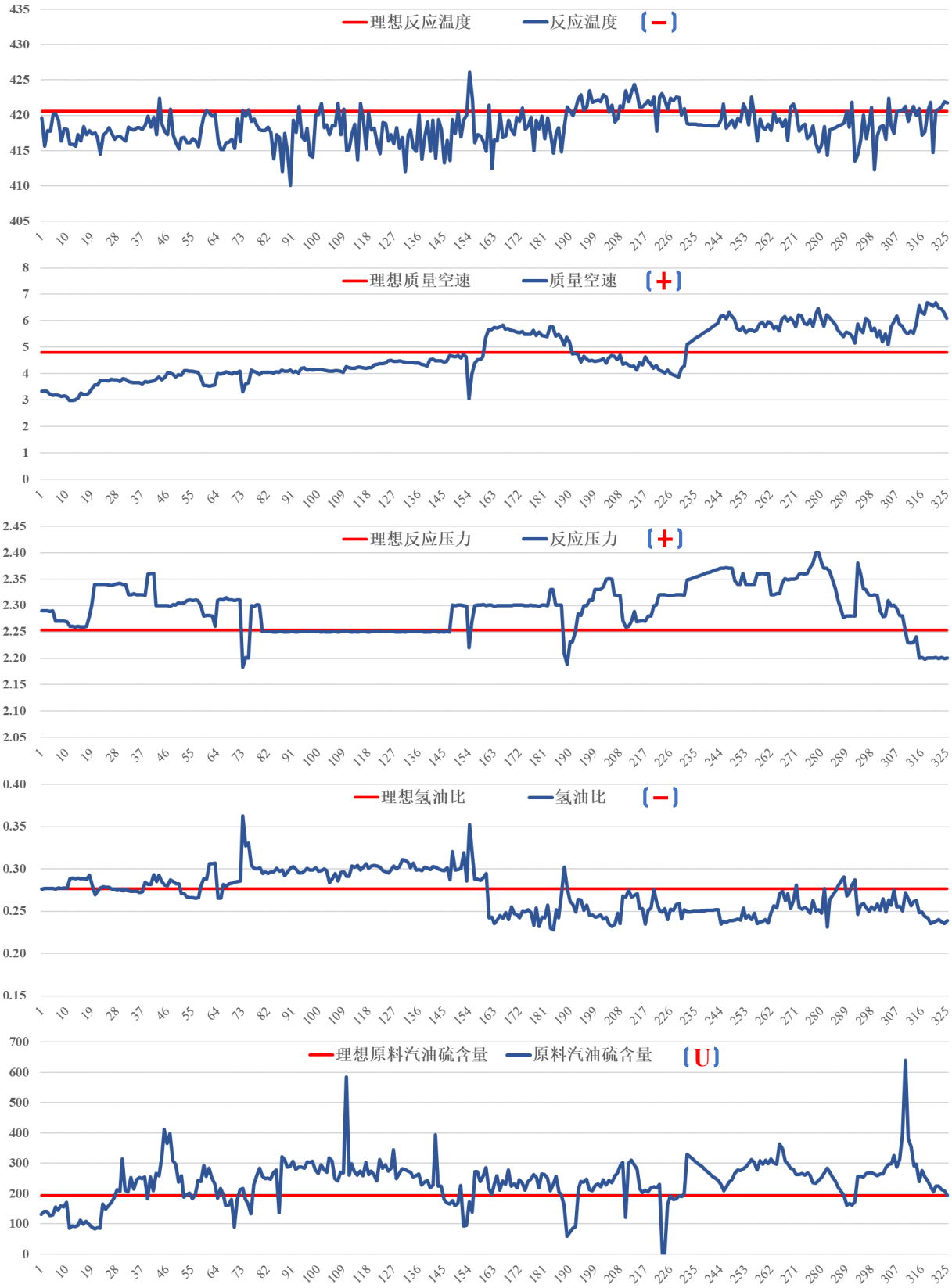

图 8　325 个历史观察样本主要操作变量的基本优化方向

注：图例右侧括号内为根据相关文献及 5.2.5 节多元非线性回归结果整理的该主要操作变量属性（“+”表正向指标、“-”表负向指标、“U”表适度指标）。其中，对于正向指标，则处于“理想样本”阈值以下的样本需尽可能调高相应参数值，反之亦然；而对于负向指标可反向类比正向指标的优化思路与方向。

5.4.3 基于辛烷值预测模型的优化操作

本文 5.4.1 与 5.4.2 节分别通过引入和构造“辛烷值损失率”和“产品硫去除率”两个评价指标以及基于 DEA 效率评价的方式寻找可以作为其他样本改进方向的“理想样本”，进而提出其他样本相关操作变量参数优化可以“理想样本”作为“标杆”的优化操作思路，由于“理想样本”很大程度上来源于该企业的历史观察数据，故确保了其对应各操作变量参数的实际可行性。

但是，由于以上研究思路总体基于在该企业自身历史数据中进行排序寻优，考虑到该企业存在产品辛烷值损失均值远高于同装置 0.6 个单位的实际情况（1.37 个单位），故仅从该企业自身历史数据出发，虽然可以为其提供各样本操作变量的优化操作指导，但也在较大程度上限制了该企业辛烷值损失降低幅度，即可能存在一个辛烷值下降幅度的下限（且该下限有可能同样高于同装置的平均水平）。

因此，本节首先考虑从 5.2.3 节所提出的多元线性回归方法出发，探寻出主要操作变量与辛烷损失值及硫去除值以及主要操作变量与辛烷值损失之间的关系，并通过数据挖掘技术，应用 5.3 节所构建的拟合预测模型，通过测试样本验证其相互关系的合理性。其次，依据主要操作变量与辛烷损失值和硫去除值之间的约束条件，构建以辛烷值损失最小为目标、以产品硫含量不大于 $5\mu g/g$以及各操作变量对应取值范围为约束条件的混合整数规划模型；并通过 Gurobi 求解器求解该模型，以获取主要操作变量调整参数的最优值。再次，在保证原料、待生吸附剂和再生吸附剂性质不变的情况下，将其余的主要操作变量设定为求解得到的最优值，再次应用 5.3 节所构建的拟合预测模型测算 325 个样本在操作变量调整至理论最优值的条件下对应的产品辛烷值损失和产品硫含量情况。最后，符合题目要求地将调整优化后各样本辛烷值损失降低超过 30%的情况进行相应的统计分析，并最终对结果进行客观合理的评价。

Step1：多元线性回归挖掘主要操作变量与辛烷损失值的关系。

为了进一步全面的分析主要操作变量与辛烷值损失之间的关系，利用机器学习与传统的回归方法，来挖掘出主要操作变量与辛烷损失值之间的相关关系。鉴于，问题三中已使用机器学习方法来挖掘主要操作变量与辛烷损失值之间的相关关系，此部分只重点描述传统的回归方法，并对比各方法在挖掘主要操作变量与辛烷值损失之间的相关关系上的准确度。以 RON 损失作为被解释变量Y_1，解释变量特征提取：原料硫含量（X_1）、原料辛烷值 RON（X_2）、烯烃（X_3）、待生吸附剂持碳率（X_4）、待生吸附剂持硫率（X_4）、再生吸附剂持碳率（X_6）、再生吸附剂持硫率（X_7）、R-101 床层下部温度（X_8）、反应器质量空速（X_9）、反应系统压力（X_{10}）、氢油比（X_{11}）、原料汽油硫含量（X_{12}）。

训练集与测试集划分和问题三一致。即原始数据为前期基于文献研究法得到的核心变量数据、根据数据处理原则和主成分分析最终得到的主要变量数据，在原始数据中提取除测试集以外的 275 样本数据作为训练集。选取原始数据，样本数据获取时间最近的前 50 个样本数据最为测试集。

由 5.3 节本文已经得出随机森林预测模型均方差为 0.033522，基于线性核（Linear Kernel）的支持向量机预测模型为 0.033478，岭回归预测模型为 0.0352379，基于 Starcking 策略复合模型为 0.035339，本节中多元线性回归拟合的均方差为 0.0407，相对来说误差较大。

Step2：机器学习方法与多元线性回归挖掘主要操作变量与硫去除值的关系

本小节将利用机器学习与传统的回归方法，来挖掘主要操作变量与辛烷损失值之间的相关关系。本节将详细描述机器学习方法与传统回归方法，并对比各方法在挖掘主要操作变量与硫去除值的相关关系上的准确度。其中传统回归方法的被解释变量硫去除（Y_2），解释变量、训练集和测试集的选取与上一节相同。

图 9 分别给出了每个机器学习方法预测模型的预测值与真实值的比较。上图直观地表明了四种预测模型都具有较高的预测准确度，与真实值的重合度较高。除随机森林预测模型外，其余预测模型的在预测硫去除上都具有理想的可行度。

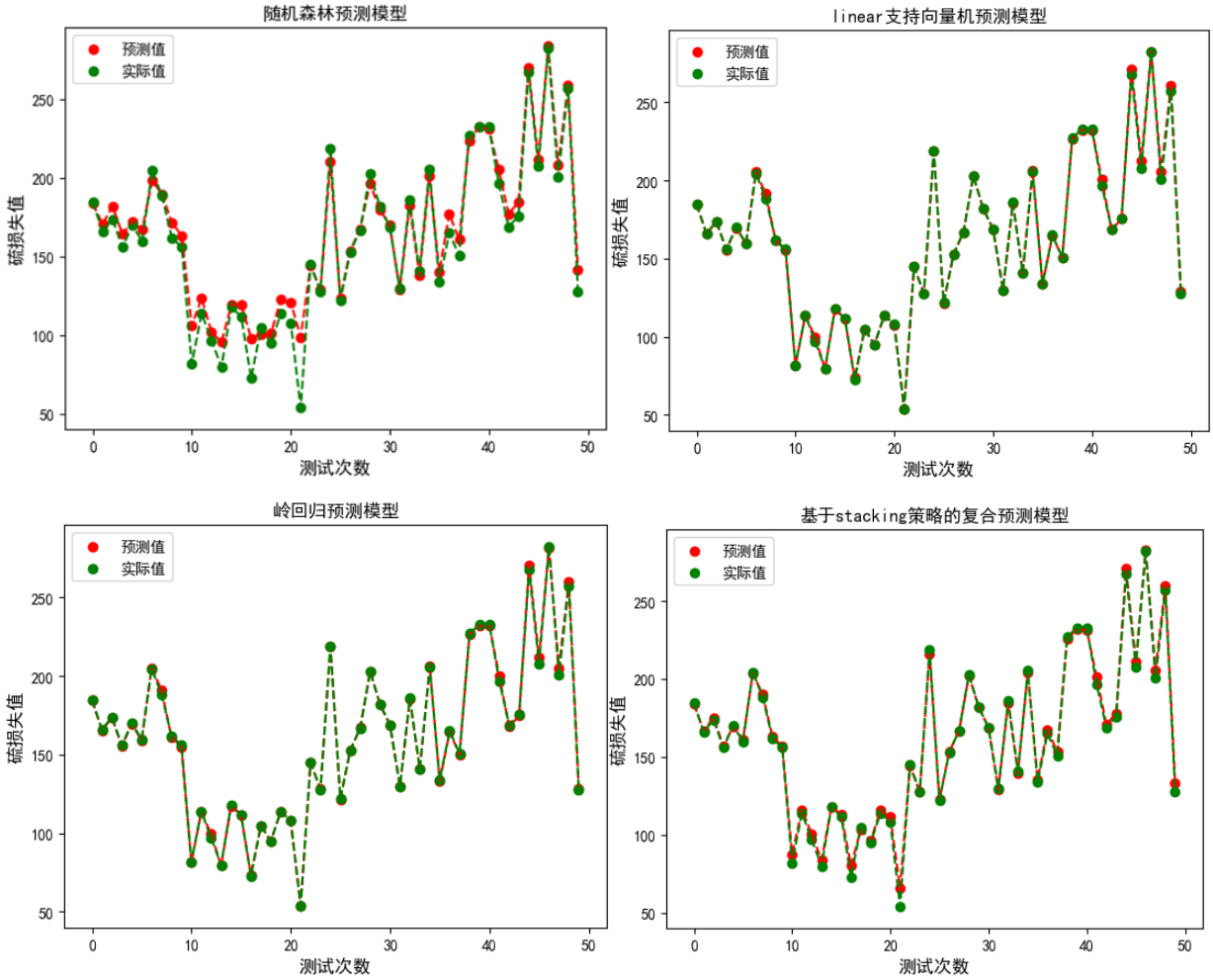

图 9　基于机器学习方法的预测模型

图 9 左上图为硫去除预测误差随测试次数的变动，随机森林模型的预测结果的误差随测试次数的上下波动幅度较大，在测试次数低于 20 次时最为明显，而基于线性核（Linear Kernel）的支持向量机、岭回归预测模型的误差一直在一个较小的范围内波动。图 10 右图是预测模型的累积误差，图中仍然显示出随机森林模型在对辛烷值损失进行预测累积误差较大，基于线性核（Linear Kernel）的支持向量机、岭回归预测模型的累积误差一直处于一个相对低的水平。本文根据较小误差评价基于线性核（Linear Kernel）的支持向量机、岭回归预测模型对预测结果较为准确，这与本文的辛烷值损失预测模型的选择基本一致。

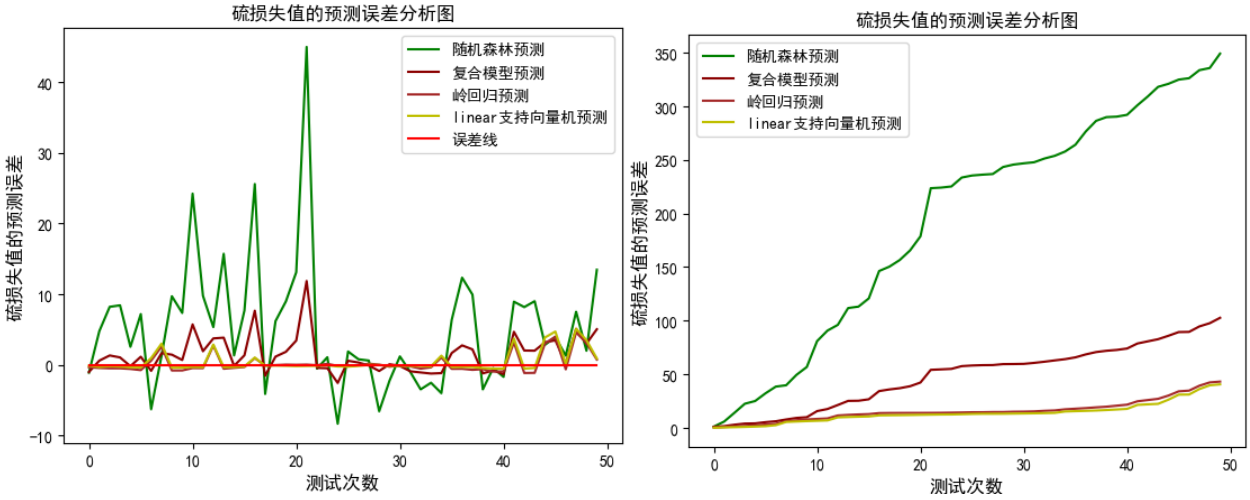

图 10　基于机器学习方法的各模型误差对比

本文通过多元线性回归模型进行损失预测的均方差为 1.5956，随机森林预测模型均方差为 107.3236，基于线性核（Linear Kernel）的支持向量机预测模型为 2.2809，岭回归预测模型为 1.8759，基于 Stacking 策略复合模型为 8.8234。多元线性回归拟合的均方差最方差最小，一定程度上可以说明多元线性回归模型可以对硫去除进行预测。

Step3：构建辛烷值操作变量的优化模型

考虑到机器学习方法在挖掘主要操作变量与辛烷损失和硫去除相关关系的过程中，由于其非线性的特征，在构建优化模型时极大的提高了模型求解的时间复杂性。本节选择多元线性回归方法获取主要操作变量与辛烷损失和硫去除之间的相关关系，来构建辛烷值操作变量的优化模型。在上一节中，分析了多元线性回归方法在描述主要操作变量与辛烷损失和硫去除之间相关关系，相比其它方法而言，其预测的准确性也是相当高的。表 12 给出了优化模型的决策变量与其具体信息。

表 12　决策变量

决策变量	决策变量名称	取值范围		最小变动单位	变量类型
		最小值	最大值		
X1	原料硫含量	57	392		连续变量
X2	原料辛烷值 RON	85.3	91.7		连续变量
X3	烯烃 v	14.6	34.67		连续变量
X4	待生吸附剂持碳率	1.01	12.15		连续变量
X5	待生吸附剂持硫率	2.94	14.31		连续变量
X6	再生吸附剂持碳率	1.43	13.34		连续变量
X7	再生吸附剂持硫率	0.25	8.92		连续变量
X8	R-101 床层下部温度	400	450	1	整数变量
X9	反应器质量空速	2.95	7	0.5	整数变量
X10	反应系统压力	2	2.45	0.1	连续变量
X11	氢油比	0.2	0.37	0.01	连续变量
X12	原料汽油硫含量	1.5	645	10	整数变量

注：上表中黄色色块标注的数据表示从样本中获取的最小值/最大值，黄色色块标注的数据表示从附件 4 获取的最小值/最大值与最小变动单位。

通过提取 Stata 软件所求解出的相关系数，最终构建目标函数如下：

$$
\begin{aligned}
min \quad & -0.0002X_1 + 0.0198X_2 - 0.0027X_3 - 0.0378X_4 \\
& +0.0250X_5 + 0.0768X_6 + 0.0421X_7 + 0.0098X_8 - 0.0136X_9 \\
& +0.5043X_{10} + 2.848329X_{11} - 0.0005X_{12} - 6.21
\end{aligned}
$$

$$
s.t. \quad \begin{cases}
S_{原} - S_{损失} \leq 5 \\
57 \leq X_1 \leq 392 \\
85.3 \leq X_2 \leq 91.7 \\
14.6 \leq X_3 \leq 34.67 \\
1.01 \leq X_4 \leq 12.15 \\
2.94 \leq X_5 \leq 14.31 \\
1.43 \leq X_6 \leq 13.34 \\
0.25 \leq X_7 \leq 8.92 \\
400 \leq X_8 \leq 450 \\
3 \leq X_9 \leq 7 \\
2 \leq X_{10} \leq 2.45 \\
0.2 \leq X_{11} \leq 0.37 \\
2 \leq X_{12} \leq 645
\end{cases} \tag{12}
$$

其中，约束条件$S_{原} - S_{损失} \leq 5$，即$X_1 - (0.9994X_1 + 0.0293X_2 - 0.1002X_3 - 0.1212X_4 + 0.0053X_5 + 0.0812X_6 - 0.0751X_7 + 0.0020X_8 + 0.0228X_9 - 2.1278X_{10} - 0.8242X_{11} - 0.0032X_{12} + 1.5959) \leq 5$，且$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_{10}, X_{11}$为连续变量，$X_8, X_9, X_{12}$为整数。使用 Gurobi 求解器求解该模型，解得最优解向量为(392.00,85.30,14.61,12.15,2.94,1.43,8.92,400.00,7.00,2.00,0.20,595.00)。

由于题目中要求了每个决策变量的最小变动单位，原料汽油硫含量（X_{12}）的最小变动单位是 10，所以本文将原料汽油硫含量（X_{12}）的最优解 595 进行调整，其近似最优解的值为 590，模型近似最优解为(392.00,85.30,14.61,12.15,2.94,1.43,8.92,400.00,7.00,2.00,0.20,590.00)。

Step4：结果分析

根据 Step3 所得到的各决策变量的最优值，本文将其定义为最优工艺参数。根据这一最优值对每个样本进行调整，反应温度调整为 400℃，质量空速调整为 7h^{-1}，反应压力调整为 2 MPa，氢油比调整为 0.2，原料汽油硫含量调整为 590。部分样本的具体优化操作条件在表 13 中给出。

表 13　最优操作条件表

样本号	原料性质			待生吸附剂性质		再生吸附剂性质		工艺参数（最优工艺参数）				
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
1	188.00	90.60	24.40	2.32	7.30	1.84	5.98	400.00	7.00	2.00	0.20	590
2	169.00	90.50	26.40	2.37	7.34	0.55	4.38	400.00	7.00	2.00	0.20	590
3	177.00	90.70	26.31	2.43	7.27	1.89	5.82	400.00	7.00	2.00	0.20	590
4	159.00	90.40	26.10	3.08	7.35	0.98	4.67	400.00	7.00	2.00	0.20	590
5	173.00	89.60	26.67	2.45	6.58	0.83	4.52	400.00	7.00	2.00	0.20	590
…	…	…	…	…	…	…	…	…	…	…	…	…
323	271.43	89.40	31.30	5.72	10.96	3.34	7.41	400.00	7.00	2.00	0.20	590
324	266.00	89.40	33.78	4.33	9.71	3.13	6.97	400.00	7.00	2.00	0.20	590
325	266.00	89.90	33.78	8.33	11.16	4.82	8.72	400.00	7.00	2.00	0.20	590

注：完整表见“附件一：325 个样本数据(附件三已追加)(数据预处理).xIsx”工作表“最优操作条件”。

按照最优工艺参数进行调整后，提取出本文比较关心的指标对优化操作进行评价分析，验证所找出的最优工艺参数可以在实际操作中降低辛烷值损失的同时，保证较低的产品硫含量，进而给出了在最优操作条件下部分样本的辛烷值损失及优化后的产品硫含量情况（见表 14）。

表 14　最优操作条件下各样本的辛烷值损失统计表

样本号	产品硫含量	产品辛烷损失	优化产品硫含量	优化辛烷损失值	辛烷损失降幅
1	3.2	1.38	4.24	0.55	60%
2	3.2	1.18	4.42	0.52	56%
3	3.2	1.38	4.42	0.56	59%
4	3.2	1.38	4.46	0.51	63%
5	3.2	1.28	4.48	0.49	62%
……	……	……	……	……	……
323	3.2	1.35	5.39	0.52	62%
324	3.6	1.28	5.46	0.53	58%
325	11.8	1.25	5.92	0.48	61%

注：完整表见“附件一：325 个样本数据(附件三已追加)(数据预处理).xIsx”工作表“最优操作条件下各样本辛烷值损失”。

从图 11 中可以看出优化后辛烷值损失降幅小于 30%的样本个数很少，仅占总体样本的 1.85%（6 个），超过 98%的样本（319 个）辛烷值损失降幅大于 30%，且有 47.69%的样本损失降幅大于 60%，最高降幅达 81%。在优化后辛烷值损失降幅大于 30%的 319 个样本里，有 264 个样本符合产品硫含量不大于$5\mu g/g$，结果表明最优工艺参数能使 81.23%的样本在符合产品硫含量不大于 $5\mu g/g$条件下辛烷值损失降幅大于 30%，说明了最优工艺参数的具有合理性。

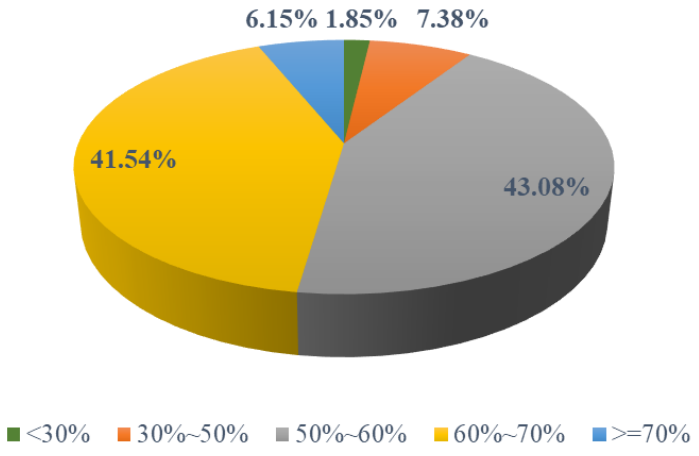

图 11　辛烷值损失降幅情况

5.5 问题五：模型的可视化展示

5.5.1 基于 DEA-BCC 效率评价模型的操作变量优化图示

根据本研究 5.4.2 节的相关研究思路与结果，可通过绘制出 133 号样本与“理想样本”在主要操作变量上的取值差异，并结合各个主要操作变量的取值范围及其每次变动 Δ 值情况，进行改进方案可行性评估（见表 15）。

表 15　133 号样本向“理想样本”改进可行性分析

操作变量	属性	具体指标	133 号样本	理想样本	差值	Δ 值	调整方向	距理想样本次数	可行性
反应温度	-	S-ZORB.TE_2003.DACA	417.91	420.5696	2.656438	1	调低	0	Yes
质量空速	+	S-ZORB.CAL.SPEED.PV	4.42	4.801541	0.377698	0.5	调高	1	Yes
反应压力	+	S-ZORB.PC_1202.PV	2.25	2.252921	0.002399	0.1	调高	1	Yes
氢油比	-	S-ZORB.CAL_H2.PV	0.30	0.276844	0.024526	0.01	调低	3	Yes
原料汽油硫含量	U	S-ZORB.AT_1001.DACA	270.38	193.6496	76.72983	10	调低	8	Yes

由表 15 的可行性报告可知，133 号样本向“理想样本”具有潜在的改进方向与空间，从而可根据各操作变量相应的取值范围、与“理想样本”的差异情况以及各操作变量的基本属性（正向指标越高越好、负向指标越低越好、适度指标趋向于理想样本值），对 133 号样本主要操作变量进行分阶段的动态优化，并运用 5.3 节的预测模型分别对产品辛烷值损失值以及去除硫含量进行模拟预测，最终调整可视化方案如图 12 所示。模拟预测结果表明：当以 5.4.2 节中所提出的 21 个 DEA 有效样本均值所构造“理想样本”作为改进方向“标杆”时，133 号样本在向其所对应的主要操作变量调整过程中，可以实现其进一步提高“去除硫含量”和降低“辛烷值损失”的双重功效。具体来看，在经过 10 次分阶段动态优化调整后，理论上，133 号样本的去除硫含量值可以从最初的 244.402$\mu g/g$提高到 244.673$\mu g/g$，即相应的产品硫含量从最初的 3.598$\mu g/g$降低到 3.327$\mu g/g$；而其辛烷值 RON 损失由最初的 1.31 降低到 0.937 个单位，即相较于 133 号样本最初而言降低了约 31.61%。

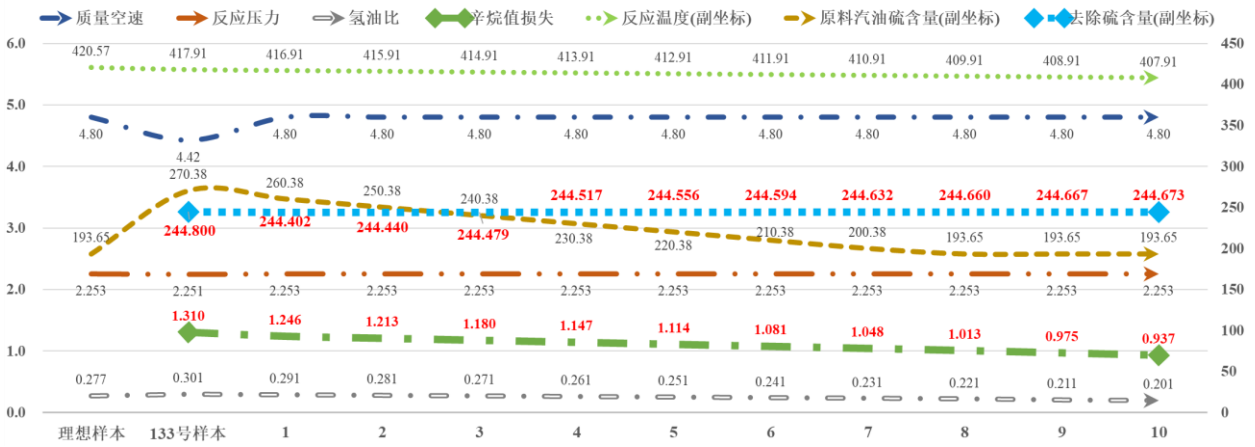

图 12　133 号样本主要操作变量向“理想样本”优化示意

5.5.2 基于多元线性回归模型的操作变量优化图示

根据本研究的 5.4.3 节中已经获得了主要操作变量的最优值，工业装置为了平稳生产，优化后的主要操作变量只能逐步调整到位，现以 133 号样本（原料性质、待生吸附剂和再生吸附剂的性质数据保持不变，其余操作变量设为最优值）为例，并以图形展示这些主要操作变量在优化调整过程中对应的汽油辛烷值和硫含量的变化轨迹。

133 号样本原始 R-101 床层下部温度、反应器质量空速、反应系统压力、氢油比、原料汽油硫含量分别为 417.91℃、4.42 h-1、2.25 MPa、0.3、270.38，依据最优工艺参数对以上操作进行调整，表 16 为 133 号样本的原始数据及优化数据的详细情况。优化前产品硫含量、产品辛烷损失分别为 3.2 和 1.31，进行优化后，产品硫含量虽然由一定程度上增加，但其产品辛烷损失下降为 0.48，降幅为 63.36%，符合题目给出的条件。

表 16　133 号样本参数特征

133 号样本特征	原料性质			待生吸附剂性质		再生吸附剂性质		工艺参数（最优工艺参数）				
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
原始参数值	248.00	89.40	20.60	2.53	8.57	1.30	6.69	417.91	4.42	2.25	0.30	270.38
优化后参数值	248.00	89.40	20.60	2.53	8.57	1.30	6.69	400.00	7.00	2.00	0.2	590

然后根据附件四操作变量信息给出参数调整要求，R-101 床层下部温度最小调整单位为 1℃，反应器质量空速的最小调整 0.5 h^{-1}，反应器顶部压力的最小调整 0.1 MPa，氢油比最小调整 0.01，原料汽油硫含量最小调整 10，表 17 给出了 133 号样本参数调整方案。

表 17　133 号样本参数调整方案表

调整参数	参数含义	调整前的参数值	最优参数值	最小调整单位	调整次数
X_8	反应温度	417.91	400.00	1	17
X_9	质量空速	4.42	7.00	0.5	3
X_{10}	反应压力	2.25	2.00	0.1	2
X_{11}	氢油比	0.30	0.2	0.01	10
X_{12}	原料汽油硫含量	270.3	590	10	12

需要观测参数调整方案对 133 号样本产品辛烷值和硫含量，因此需要分别建立辛烷值观测函数、硫含量观测函数：

133 号样本汽油的辛烷值观测函数为：$RON_{产} = RON_{原} - RON_{产}$，即：

$$\begin{aligned} RON_{产} = {} & 89.4 + 0.0002X_1 - 0.0198X_2 + 0.0027X_3 + 0.0378X_4 \\ & -0.0250X_5 - 0.0768X_6 - 0.0421X_7 - 0.0098X_8 + 0.0136X_9 \\ & -0.5043X_{10} - 2.848329X_{11} + 0.0005X_{12} \end{aligned}$$

133 号汽油的硫含量观测函数：$S_{产} = S_{原} - S_{去}$，即：

$$\begin{aligned} S_{产} = {} & 248 - 0.9994X_1 - 0.0293X_2 + 0.1002X_3 + 0.1212X_4 \\ & -0.0053X_5 - 0.0812X_6 + 0.0751X_7 - 0.0020X_8 - 0.0228X_9 \\ & +2.1278X_{10} + 0.8242X_{11} + 0.0032X_{12} \end{aligned}$$

对操作变量按照需要的操作次数从多到少进行调整，顺序为（反应温度→原料汽油硫含量→质量空速→反应压力→氢油比）。

按照操作方案和操作顺序进行调整的结果如表 18，一共进行了 44 次调整。首先，进行反应温度调整，每调整一个单位，辛烷值和硫含量都会有一定程度的增加；其次，进一步增加原料硫含量，辛烷值上升，硫含量增加速度；再次，增加质量空速会使辛烷值增加的同时降低硫含量；然后，降低反应器顶部压力时，产品辛烷值会有轻微的减少，但硫含量会有一定程度的增加；最后，通过降低氢油比也可以在使辛烷值增加的同时降低硫含量。

表 18　参数调整时汽油中硫含量及辛烷值的变化趋势表

调整参数	次数	调整幅度	辛烷值	硫含量
X_8 R-101 床层下部温度	1	1	89.12879	3.639871
	2	1	89.1386	3.641836
	3	1	89.14841	3.643802
	4	1	89.15823	3.645767
	5	1	89.16804	3.647732
	6	1	89.17785	3.649698
	7	1	89.18767	3.651663
	8	1	89.19748	3.653628
	9	1	89.20729	3.655594
	10	1	89.2171	3.657559
	11	1	89.22692	3.659524
	12	1	89.23673	3.661489
	13	1	89.24654	3.663455
	14	1	89.25636	3.66542
	15	1	89.26617	3.667385
	16	1	89.27598	3.669351
	17	1	89.28579	3.671316
X_{12} 原料汽油硫含量	18	10	89.29115	3.703468
	19	10	89.2965	3.73562
	20	10	89.30185	3.767772
	21	10	89.3072	3.799924
	22	10	89.31255	3.832076
	23	10	89.31791	3.864228
	24	10	89.32326	3.89638
	25	10	89.32861	3.928532
	26	10	89.33396	3.960684
	27	10	89.33931	3.992836
	28	10	89.34467	4.024988
	29	10	89.35002	4.05714
X_9 反应器质量空速	30	0.5	89.35681	4.045738
	31	0.5	89.36361	4.034336
	32	0.5	89.3704	4.022935
X_{10} 反应器顶部压力	33	-0.1	89.42083	3.810159
	34	-0.1	89.47127	3.597384
X_{11} 氢油比	35	-0.01	89.49975	3.589142
	36	-0.01	89.52823	3.580901
	37	-0.01	89.55672	3.572659
	38	-0.01	89.5852	3.564417
	39	-0.01	89.61368	3.556176
	40	-0.01	89.64217	3.547934
	41	-0.01	89.67065	3.539692
	42	-0.01	89.69913	3.531451
	43	-0.01	89.72762	3.523209
	44	-0.01	89.7561	3.514968

为更加直观地观察出辛烷值和硫含量的变化情况，分别描绘了在操作变量优化调整过程中汽油辛烷值和硫含量的变化轨迹（见图 13）。在优化操作的过程中，汽油中的辛烷值一直在增加，当操作次数进行到第 30 次以后时，辛烷值增加的速度明显加快，同时，几乎同时段产品硫含量也急剧下降。可以发现硫含量也有一段时间在快速增加，但此时，辛烷值也在缓慢的增加。133 号样本的在调整过程中，硫含量先增加，后减少，但最大值也没有超过 5μg/g。

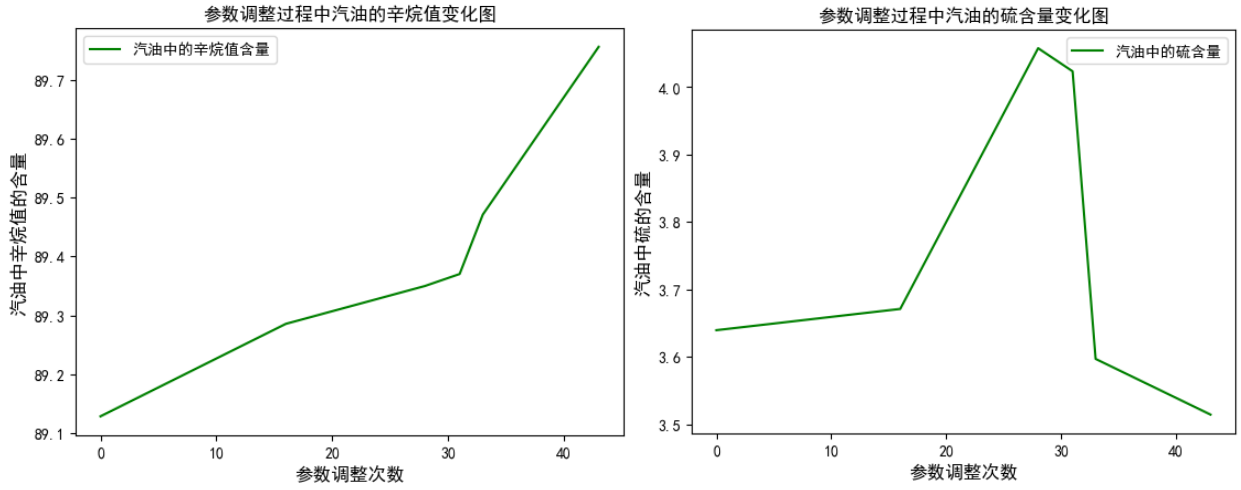

图 13　汽油辛烷值和硫含量的可视化

5.5.3 基于参数调整批次优化的操作变量优化图示

在 5.5.2 中本文分成 44 个批次调整参数，每次调整一个参数，每个参数每次变化一个单位，当一个参数调整到上限后再进行下一个参数的调整。但是这种操作不太符合实际操作要求，与此同时，每次只进行一个参数的调整，会造成其产品硫含量的波动比较大，从而造成产品质量下降。因此，本节构建了整数规划模型，用以降低在参数调整过程中汽油硫含量的波动，同时优化参数的调整批次。

在参数调整批次优化模型中，以降低汽油硫含量的波动为目标。在调整参数的过程中，汽油硫含量受到调整参数的影响可表示为：

$$\text{汽油硫含量的变动} = \text{参数调整幅度}(X_{ij}) \times \text{参数对汽油硫含量的影响系数}$$

X_{ij}表示在第i批次中，对参数j的调整幅度值，各主要操作变量每次调整的幅度值为 Δ 的倍数。

因此，该目标函数可表示为：

$$\underset{X_{ij}}{Min} \sum_{i=1}^{k} (-0.0020X_{i8} - 0.0228X_{i9} + 2.1278X_{i10} + 0.8242X_{i11} + 0.0032X_{i12})^2$$

$$s.t.\begin{cases} 15 \le \sum_{i=1}^{k} X_{i8} \le 17 \\ 2 \le \sum_{i=1}^{k} X_{i9} \le 3 \\ 1 \le \sum_{i=1}^{k} X_{i10} \le 2 \\ 8 \le \sum_{i=1}^{k} X_{i11} \le 10 \\ 10 \le \sum_{i=1}^{k} X_{i12} \le 12 \\ X_{ij} \in N^{+}, i = (1,2,\dots,k), j = (8,9,\dots,12) \end{cases} \tag{13}$$

通过 Gurobi 求解该优化模型，结果表明当变量调整的批次为 8 次时，汽油硫含量的波动幅度最小。表 19 中给出了具体每一个批次中需要调整的参数，以及参数所对应的调整幅度。如，在第一批次中，将会对参数X_8，X_{11}，X_{12}进行调整，具体对X_8（R-101 床层下部温度）的调整幅度为 3 个 Δ 单位，同时，对X_{11}（氢油比）调整 3 个 Δ 单位，X_{12}（原料汽油硫含量）调整 2 个 Δ 单位。

表 19 批次优化模型的最优解

参数 / 批次	X_8	X_9	X_{10}	X_{11}	X_{12}
1	3	0	0	3	2
2	3	0	0	2	0
3	3	0	1	0	3
4	2	0	0	2	0
5	0	0	0	1	0
6	0	1	1	0	3
7	2	1	0	1	0
8	2	1	0	1	0

在每次参数调整的批次中，通过硫含量及辛烷值的估算模型，获得汽油的硫含量和辛烷值，并将该数据展现在图 14 中。每一次的优化操作都会使辛烷值含量上升的同时降低硫含量。

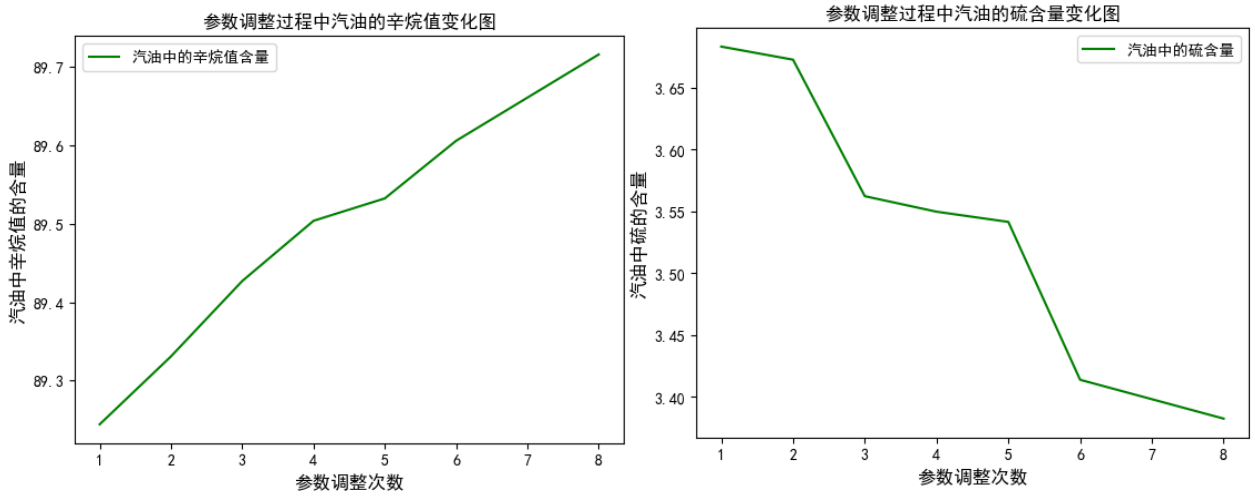

图 14　参数调整过程中汽油硫含量和辛烷值的变化

六、进一步讨论：基于中介效应模型的传导机制检验

参考 Hayes（2009）[14]以及温忠麟和叶宝娟（2014）[15]等人的研究思路，采用 Baron 和 Kenny（1986）[16]的逐步法构建如下中介效应模型，并根据前期文献研究法提取合适的中介因子，进而探讨产品硫含量去除对辛烷值损失的影响作用机制，最终尝试打开去除产品硫含量的同时如何导致辛烷值损失的“黑箱”。

考虑将烯烃作为中介变量，RON 损失为被解释变量，去除硫含量为解释变量，检验去除硫含量的过程中，是否会通过烯烃变量对 RON 损失产生中介效应。其基本模型示意图如图 15 所示，其中，假定M为中介变量，c为解释变量X对被解释变量Y的总效应，a为X对中介变量M的效应，b为当控制X的影响后，M对Y的效应，c'为当控制M的影响后，X对Y的效应。$e_i(i=1,2,3)$是回归残差。其中$c=c'+ab$。

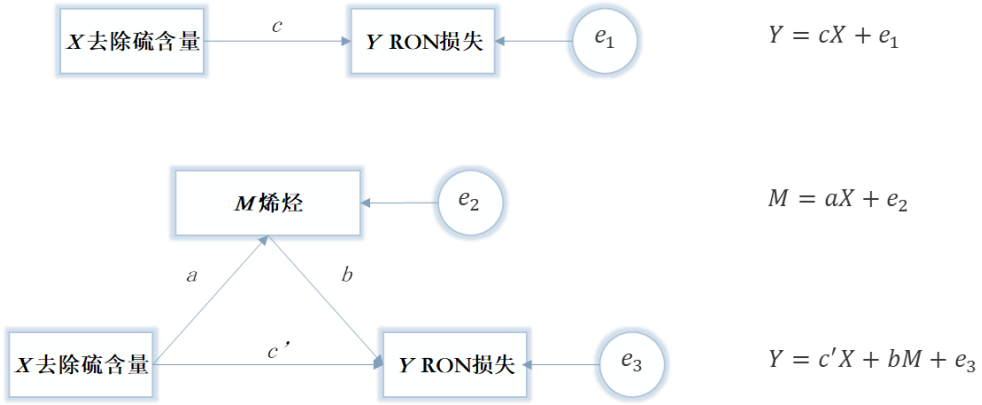

图 15　中介效应模型示意图

依据 Baron and Kenny（1986）[16]的逐步法思想，对于中介效应的探讨需逐步按照以下步骤进行检验：

Step1：检验$H_0: c=0$；
Step2：检验$H_0: a=0$；
Step3：检验$H_0: b=0$；

运用 Stata 进行中介效应检验结果如表 20 所示，解释变量去除硫含量对被解释变量 RON 损失的总效用为负，且显著；去除硫含量对被解释变量烯烃的效应为正同样也时显著的；当控制去除硫含量的影响后，烯烃对 RON 损失具有显著的负效应，说明中介效应显著；而当控制中变量烯烃的影响后，去除硫含量对 RON 损失的负效应变得不显著了，表明此过程的确是一个完全中介过程，亦即去除硫含量（X）的确通过烯烃（M）对 RON 损失（Y）产生了中介传导机制。

表 20 中介效应检验模型结果

	(1)	(2)	(3)
被解释变量	**RON 损失**	**烯烃**	**RON 损失**
去除硫含量	-0.001***	0.029***	-0.000
	(-2.73)	(7.53)	(-1.54)
烯烃			-0.007**
			(-2.58)
_cons	1.372***	18.791***	1.503***
	(30.77)	(20.59)	(22.35)
N	325	325	325
R-Square	0.023	0.149	0.042
BIC	-41.43	1920.59	-42.28

注：1. 括号内为 t 统计量；2. *，**和***分别表示通过了 10%，5%和 1%显著性水平检验。

七、模型评价与推广

7.1 优点

（1）在对变量进行降维过程中，本文创新性使用了文献研究法对核心变量进行提取，以避免后期进行数据降维过程中造成有效信息损失，也为后续优化操作提供理论基础与经验支撑。

（2）进行模型预测时，为尽可能使模型具有合理性，构建了四个模型进行预测并比较各模型的误差，择优确定了最终的预测模型算法。

（3）为从多个角度寻找操作优化条件，本文首先引入了“辛烷值损失率”以及“产品硫去除率”两个概念，并对这找个指标进行排序以寻找“理想样本”；同时采用了更加客观的 DEA-BCC 综合评价模型，以此测度得到了 21 个 DEA 有效样本，并利用其成功构造了“理想样本”；同时本文附录还报告了基于熵值 TOPISS 法选择“理想样本”另一思路。

（4）从两个方向展示了操作变量在逐步优化操作过程中 133 号样本汽油辛烷值和硫含量的变化轨迹。其一，以 “理想样本”所对应的操作变量参数作为优化参考对样本操作条件进行调整；其二，优化算法思路出发，探讨在操作变量取值范围以及单次调整幅度的约束条件下，进行单次单变量调整和单次多变量调整方案 。

（5）运用中介效应模型对本研究进行进一步的讨论，发现当控制解释变量去除硫含量的影响后，中介变量烯烃对被解释变量 RON 损失具有显著的负中介效应；而当控制中变量烯烃的影响后，去除硫含量对 RON 损失的负效应变得不显著，此过程是一个完全中介过程。

7.2 不足

（1）在数据预处理中采用了 5%和 95%百分位对数据进行截尾处理，可能会导致数据信息存在一定误差，可以尝试对其进行缩尾处理，并具体比较两者差异。

（2）主成分分析模型在对变量降维过程中仅提取了前 80.1467%贡献率的信息，导致损失了原始数据部分信息损失；且在进行后续操作优化时，由于无法准确对应到主成分降维前所涉及的非核心操作变量原有信息，限制了优化操作方案。

（3）以机器学习算法来构建的预测模型，其精度与特征向量的提取有者密切的关系，同时算法参数的设置对模型精度也有较大的影响，因此在以机器学习算法来构建辛烷值预测模型的精度，还有进一步提升的空间。

7.3 未来改进与推广

（1）可考虑采用其他方法对数据进行进一步的处理以减小误差，如采用截尾法处理异常数据，或者异常值不进行删除而是替换为合理区间的端点值等。

（2）在测度多元非线性回归模型时，采用了传统 OLS 估计法，未来可以尝试通过极大似然估计、系统矩估计等不同的运算内核来确保回归系数的一致性和无偏性。

（3）可以考虑将本文所得到的基本研究发现，如影响辛烷值损失核心变量提取、辛烷值损失预测模型、降低辛烷值损失的优化操作条件等推广到同行业其他公司进行实践参考与验证。

参考文献

[1] 王慧，张睿，刘海燕，孟祥海，催化裂化汽油脱硫精制技术研究进展，化工进展，39(6)：2354-2362，2020。

[2] 赵小燕，李成文，朱玉新，优化操作降低汽油加氢装置重汽油辛烷值损失，化工管理，(14)：164，2015。

[3] 田勇震，杨忠义，马健波，降低汽油加氢装置辛烷值损失的优化措施，石化技术与应用，37(5)：345-348，2019。

[4] 马强，赵昌明，降低S-Zorb装置汽油辛烷值损失的优化操作，当代化工研究，(15)：43-45，2020。

[5] 周欢，齐万松，李宏勋，S Zorb装置辛烷值损失大原因的分析与措施，云南化工，46(9)：90-91，2019。

[6] 简建超，孙浩，冯海春，炼油厂国V排放标准汽油生产方案中S Zorb装置的优化，石油炼制与化工，48(6)：61-64，2017。

[7] 于善宝，郭宏，孙同根，陈刚，加强工艺管理降低S Zorb装置精制汽油辛烷值损失，石油炼制与化工，49(10)：15-19，2018。

[8] 柳文，赵欣，邢东，隋志国，徐惠丽，Mip技术在石蜡基原料催化裂化装置上的应用，石油炼制与化工，48(6)：83-87，2017。

[9] 司守奎，孙玺菁，数学建模算法与应用，北京：国防工业出版社，1-256，2011。

[10] 阿童，随机森林原理，https://www.jianshu.com/p/d4ed4a0c540f，2020-09-20。

[11] 键盘流，Scikit-Learn(Sklearn)支持向量机(Svm)算法类库介绍，http://blog.sina.com.cn/s/blog_62970c250102xg0g.html，2020-09-20。

[12] 周先森爱吃素，机器学习-Stacking方法的原理及实现，https://blog.csdn.net/zhouchen1998/article/details/89253879，2020-09-20。

[13] Banker R. D.，Charnes A.，Cooper W. W.，Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis，Management Science，30(9)：1078-1092，1984。

[14] Hayes A. F.，Beyond Baron and Kenny: Statistical Mediation Analysis in the New Millennium，Communication Monographs，76(4)：408-420，2009。

[15] 温忠麟，叶宝娟，中介效应分析:方法和模型发展，心理科学进展，22(5)：731-745，2014。

[16] Baron R. M.，Kenny D. A.，The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations，J Pers Soc Psychol，51(6)：1173-1182，1986。

附录

附录 1：问题一：对附件三原始数据处理的 Stata 源代码（Q1_数据处理.do）

```
// 问题一：数据处理
* 2020-09-17

* 指定全局路径
clear
global path "E:\ArchiveforStudy\数学建模\2020 年 B 题--汽油辛烷值建模"
cd "E:\ArchiveforStudy\数学建模\2020 年 B 题--汽油辛烷值建模"

* ================================================================
* ==================== 285 号样本原始数据处理 ====================
* ================================================================
import excel "$path\附件三：285 号和 313 号样本原始数据.xlsx", firstrow sheet("操作变量 (2)") cellrange(a2:mq42) clear
   gen 样本号 = _n
   order 样本号 采样日期
* 查看数据基本特征
   summarize
* for 循环查找并根据原则 3 替换数据存在 0 值（空值）情况
foreach var of varlist SZORBCAL_H2PV-SZORBPC_1001APV{
   quietly summarize `var'
   replace `var' = r(mean) if `var'==0
}
* 采用数据 5%和 95%分位数作为最大最小的限幅方法，剔除部分不在此范围的样本
foreach var of varlist SZORBCAL_H2PV-SZORBPC_1001APV{
   egen `var'_group=cut(`var'), group(20)
   drop if `var'_group==0|`var'_group==20
}
* 取前 2 个小时操作变量数据的平均值作为对应辛烷值的操作变量数据
foreach var of varlist SZORBCAL_H2PV-SZORBPC_1001APV{
   quietly summarize `var'
   replace `var'= r(mean)
}
   keep if _n==1
   replace 样本号=285
   replace 采样日期="2017/7/17 8:00:00"
save "$path\附件三：285 号样本原始数据处理完成.dta", replace

* ================================================================
* ==================== 313 号样本原始数据处理 ====================
* ================================================================
import excel "$path\附件三：285 号和 313 号样本原始数据.xlsx", firstrow sheet("操作变量 (2)") cellrange(a44:mq84) clear
   gen 样本号 = _n
   order 样本号 采样日期
* 查看数据基本特征
   summarize
* for 循环查找并根据原则 3 替换数据存在 0 值（空值）情况
foreach var of varlist SZORBCAL_H2PV-SZORBPC_1001APV{
   quietly summarize `var'
   replace `var'= r(mean) if `var'==0
}
* 采用数据 5%和 95%分位数作为最大最小的限幅方法，剔除部分不在此范围的样本
foreach var of varlist SZORBCAL_H2PV-SZORBPC_1001APV{
   egen `var'_group=cut(`var'), group(20)
   drop if `var'_group==0|`var'_group==20
}
* 取前 2 个小时操作变量数据的平均值作为对应辛烷值的操作变量数据
foreach var of varlist SZORBCAL_H2PV-SZORBPC_1001APV{
   quietly summarize `var'
   replace `var'= r(mean)
```

```
}
   keep if _n==1
   replace 样本号=313
   replace 采样日期="2017/5/15 8:00:00"
save "$path\附件三：313 号样本原始数据处理完成.dta", replace

* ============================================================
* ============== 附件三 285 和 313 号样本处理数据合并 ==============
* ============================================================
* 提取并转换 EXCEL 格式为 Stata 格式数据（便于后续数据合并）
import excel "$path\附件三：285 号和 313 号样本原始数据.xlsx", firstrow sheet("原料") clear
   save "$path\附件三：285 号和 313 号样本原始数据-原料", replace
import excel "$path\附件三：285 号和 313 号样本原始数据.xlsx", firstrow sheet("产品") clear
   drop F-FN
   save "$path\附件三：285 号和 313 号样本原始数据-产品", replace
import excel "$path\附件三：285 号和 313 号样本原始数据.xlsx", firstrow sheet("待生吸附剂") clear
   save "$path\附件三：285 号和 313 号样本原始数据-待生吸附剂", replace
   import excel "$path\附件三：285 号和 313 号样本原始数据.xlsx", firstrow sheet("再生吸附剂") clear
save "$path\附件三：285 号和 313 号样本原始数据-再生吸附剂", replace

* 合并 285 和 313 号样本处理数据
use "$path\附件三：285 号样本原始数据处理完成.dta", replace
   append using "$path\附件三：313 号样本原始数据处理完成.dta"
   merge 1:1 样本号 using "$path\附件三：285 号和 313 号样本原始数据-原料.dta", keepusing()
      drop _merge
   merge 1:1 样本号 using "$path\附件三：285 号和 313 号样本原始数据-产品.dta", keepusing()
      drop _merge
   merge 1:1 样本号 using "$path\附件三：285 号和 313 号样本原始数据-待生吸附剂.dta", keepusing()
      drop _merge
   merge 1:1 样本号 using "$path\附件三：285 号和 313 号样本原始数据-再生吸附剂.dta", keepusing()
      drop _merge
save "$path\附件三：285 和 313 号样本原始数据处理完成.dta", replace

* 285 和 313 号样本处理数据与附件 1 合并
import excel "$path\附件一：325 个样本数据.xlsx", firstrow sheet("Sheet1 (2)") cellrange(a3:nf328) clear
   drop if 样本号==285|样本号==313
   append using "$path\附件三：285 和 313 号样本原始数据处理完成.dta"
   sort 样本号
   drop 采样点名称 样品名称 SZORBCAL_H2PV_group-SZORBPT_2101PV_group
   replace RON 损失= 辛烷值 RON - 辛烷值 RON_产品 if 样本号==285|样本号==313
   gen 去除硫含量 = 硫含量 μgg - 硫含量 μgg_产品
   drop *group
   rename (硫含量 μgg 辛烷值 RON 硫含量 μgg_产品 辛烷值 RON_产品 RON 损失不是变量 焦炭 wt Swt 焦炭 wt_再生 Swt_再生) (原料硫含量 原料辛烷值 RON 产品硫含量 产品辛烷值 RON RON 损失 待生吸附剂_焦炭 待生吸附剂_S 再生吸附剂_焦炭 再生吸附剂_S)
   order 样本号-RON 损失 去除硫含量
export excel "附件一：325 个样本数据(附件三已追加).xlsx", firstrow(variables) replace
```

附录 2：问题二：寻找建模主要变量的 Stata 源代码（Q2_寻找建模主要变量.do）

```
// 问题二：寻找建模主要变量
* 2020-09-17

* 指定全局路径
clear
global path "E:\ArchiveforStudy\数学建模\2020 年 B 题--汽油辛烷值建模"
cd "E:\ArchiveforStudy\数学建模\2020 年 B 题--汽油辛烷值建模"

* ============================================================
* ========= 附件三数据处理追加到附件一后数据预处理 ===========
```

```
* ============================================================
import excel "$path\附件一：325 个样本数据(附件三已追加).xlsx", firstrow clear
// 文献研究法 提取经验指标
* 对于多个指标的情况，采用方差最小的原则确定核心解释变量
* 反应温度
  summarize SZORBTE_2004DACA SZORBTE_2003DACA SZORBTE_2002DACA SZORBTE_2001DACA SZORBTE_2104DACA SZORBTE_2005PV SZORBTC_1606PV SZORBTE_2103PV
  * 根据赵小燕等（2015）文献选择 R-101 床层温度 SZORBTE_2003DACA
* 质量空速 SZORBCALSPEEDPV
* 反应压力
  summarize SZORBPT_2101PV SZORBPC_1202PV
  * 选择 SZORBPC_1202PV
* 氢油比 SZORBCAL_H2PV
* 原料汽油硫含量 SZORBAT_1001DACA
  global literature_var "SZORBTE_2003DACA SZORBCALSPEEDPV SZORBPC_1202PV SZORBCAL_H2PV SZORBAT_1001DACA"

// 附件数据异常情况
* 附件一中数据变量无单位无中文名称情况
  global novarname "SZORBFT_9102PV SZORBPT_1501PV SZORBFT_1002TOTAL SZORBFT_1004TOTAL SZORBFT_9001TOTAL SZORBFT_5104TOTAL SZORBFT_5201TOTAL SZORBFT_5101TOTAL SZORBFT_9101TOTAL SZORBFT_1003TOTAL SZORBFT_3301TOTAL SZORBFT_9201TOTAL SZORBFT_9202TOTAL SZORBFT_9301TOTAL SZORBFT_9302TOTAL SZORBFT_9401TOTAL SZORBFT_9402TOTAL SZORBFT_9403TOTAL SZORBFT_1202TOTAL SZORBFT_1204PV SZORBFT_5102PV SZORBFT_1204TOTAL SZORBFT_5102TOTAL SZORBFT_2431DACA"
* 附件一中数据变量无单位有中文名称情况
  global nounit "SZORBPC_2401PIDAOP SZORBFC_2432PIDASP SZORBPC_2401PIDASP SZORBFT_1504DACAPV SZORBFT_1504TOTALIZERAPV SZORBFT_1503DACAPV SZORBFT_1503TOTALIZERAPV SZORBPC_2401BPIDASP SZORBPC_2401BPIDAOP SZORBLT_1002DACA SZORBPC_1001APV SZORBFC_2432DACA SZORBFT_2433DACA SZORBZT_2533DACA SZORBPDT_2503DACA SZORBPT_2502DACA SZORBPT_2501DACA SZORBFT_2502DACA SZORBTE_2902DACA SZORBFT_2901DACA SZORBLT_2901DACA SZORBPT_2905DACA SZORBTC_2702DACA SZORBFC_2702DACA SZORBFT_3201DACA SZORBFT_2701DACA SZORBFT_3001DACA SZORBFT_3304DACA SZORBFT_5102DACAPV SZORBLC_5102DACA SZORBLT_9001DACA SZORBPC_9002DACA SZORBTC_2201PV SZORBTC_2201OP SZORBTXE_2203ADACA SZORBSIS_TEX_3103BPV SZORBTE_1605DACA SZORBPT_1601DACA SZORBPT_6008DACA SZORBTC_1607DACA SZORBPT_1604DACA SZORBTE_7102DACA SZORBPT_7103DACA SZORBPT_7103BDACA SZORBPT_7107BDACA SZORBPT_7503DACA SZORBPT_7505DACA SZORBPT_7503BDACA SZORBPT_7505BDACA SZORBPT_7508DACA SZORBPT_7510DACA SZORBPT_7508BDACA SZORBPT_7510BDACA SZORBFT_1301DACA SZORBPT_2106DACA SZORBPT_2901DACA SZORBPDI_2903DACA SZORBPDT_1004DACA SZORBPDT_3002DACA SZORBPDT_2906DACA SZORBPDT_3502DACA SZORBPDT_3503DACA SZORBPDT_2409DACA SZORBPDT_1002DACA SZORBPDT_1003DACA SZORBPDT_2001DACA SZORBDT_2107DACA SZORBDT_2001DACA SZORBLC_2601DACA SZORBZT_2634DACA SZORBPDT_2606DACA SZORBPT_2607DACA SZORBPDT_2605DACA SZORBPT_2603DACA SZORBAT0002DACAPV SZORBAT0003DACAPV SZORBAT0004DACAPV SZORBAT0005DACAPV SZORBAT0006DACAPV SZORBAT0007DACAPV SZORBAT0008DACAPV SZORBAT0009DACAPV SZORBAT0010DACAPV SZORBAT0011DACAPV SZORBAT0012DACAPV SZORBAT0013DACAPV SZORBLI_2107DACA SZORBFT_3702DACA SZORBFT_3701DACA SZORBPC_2401BDACA SZORBPC_2401DACA SZORBBS_AT_2401PV SZORBBS_AT_2402PV SZORBBS_LT_2401PV SZORBFT_1502DACA SZORBPC_2902DACA SZORBPDI_2105DACA SZORBPDI_2301DACA SZORBCALLEVELPV SZORBPC_6001PV SZORBPT_6003DACA SZORBTE_2104DACAPV SZORBPDI_2801DACA SZORBLT_9101DACA SZORBFT_9102TOTAL SZORBFT_1006TOTALIZERAPV SZORBFT_1006DACAPV SZORBPT_6002PV SZORBRXL_0001AUXCALCAPV SZORBPC_3501DACA SZORBFC_1104DACA SZORBPT_5201DACA SZORBFC_5203DACA SZORBPDT_3602DACA SZORBPDT_3601DACA SZORBFC_5103DACA SZORBFT_3501DACA SZORBLI_2104DACA SZORBTE_6008DACAPV SZORBPDI_2501DACA SZORBPDC_2702DACA SZORBPDT_2703BDACA SZORBPDT_2704DACA"
* 确保文献核心解释保护变量前提下，剔除无单位无中文名称数据
drop $novarname $nounit

// 按照附件二数据处理原则替换和剔除数据
* 删除观测值 50%以上均为 0 值的变量名
mvdecode 原料硫含量-SZORBFT_5204TOTALIZERAPV, mv(0=.)
/* SZORBFC_23~V: 145 missing values generated
     SZ~FT_9301PV: 4 missing values generated
     SZORBFT_15~V: 288 missing values generated
     SZORBF~104PV: 126 missing values generated
```

```
    SZORBFT_91~V: 134 missing values generated
    SZORBF~402PV: 1 missing value generated
    SZORBF~002PV: 137 missing values generated
    SZORBF~003PV: 4 missing values generated
    SZORBF~004PV: 19 missing values generated
    SZ~FC_1202PV: 219 missing values generated
    SZORBF~103PV: 214 missing values generated
    SZORBFT_15~L: 123 missing values generated
    SZO~3303DACA: 3 missing values generated
    SZO~2303DACA: 10 missing values generated
    SZO~2302DACA: 3 missing values generated
    S~T_2002DACA: 22 missing values generated
    SZORBFT_28~A: 297 missing values generated
    SZORBTEX_3~A: 214 missing values generated
    S~5204DACAPV: 84 missing values generated */
drop SZORBFC_23~V SZORBFT_15~V SZORBF~104PV SZORBFT_91~V SZORBF~002PV SZ~FC_1202PV
SZORBF~103PV SZORBFT_15~L SZORBFT_28~A SZORBTEX_3~A
mvencode 原料硫含量-SZORBFT_5204TOTALIZERAPV, mv(.=0)

* for 循环查找并根据原则 3 替换数据存在 0 值（空值）情况
foreach var of varlist 原料硫含量-SZORBFT_5204TOTALIZERAPV{
   quietly summarize `var'
   replace `var'= r(mean) if `var'==0
}

* 生成 辛烷值损失率 和 产品硫去除率 两个核心评价指标
gen 辛烷值损失率 =RON 损失 / 原料辛烷值 RON
gen 产品硫去除率 = 去除硫含量 / 原料硫含量
order 样本号 采样日期 产品硫去除率 辛烷值损失率 去除硫含量 原料硫含量 产品硫含量 RON 损失 原料辛烷值
RON 产品辛烷值 RON
order 样本号-再生吸附剂_S $literature_var

export excel "$path\附件一：325 个样本数据(附件三已追加)(数据预处理).xlsx", firstrow(variables) replace
```

附录 3：问题二：基于主成分分析降维的 MATLAB 源代码（Q2_PCA.m）

```
%==========================================================================
%                          读取附件一中数据，进行主成分分析
%==========================================================================
%   2020-09-17

clc, clear;
[X,textdata] = xlsread('附件一：325 个样本数据(附件三已追加)(数据预处理).xlsx','Sheet1');
data = X(:,25:end);
XZ = zscore(data);                                  % 数据标准化

% 调用 pca 函数根据标准化后原始样本观测数据作主成分分析，返回主成分表达式的系数矩阵 COEFF，
% 主成分得分数据 SCORE，样本相关系数矩阵的特征值向量 latent 和每个观测的霍特林 T2 统计量
[COEFF,SCORE,latent,tsquare] = pca(XZ);

% 为了直观，定义元胞数组 result1，用来存放特征值、贡献率和累积贡献率等数据
explained = 100*latent/sum(latent);                 % 计算贡献率
[m, n] = size(data);                                % 求 data 的行数和列数
result1 = cell(n+1, 4);                             % 定义一个 n+1 行，4 列的元胞数组
result1(1,:) = {'特征值', '差值', '贡献率', '累积贡献率'};
result1(2:end,1) = num2cell(latent);                % 存放特征值
result1(2:end-1,2) = num2cell(-diff(latent));       % 存放特征值之间的差值
result1(2:end,3:4) = num2cell([explained, cumsum(explained)]); %存放(累积)贡献率

% 为了直观，定义元胞数组 result2，用来存放前 15 个主成分（累计贡献率：80.1467%）的系数数据
varname = textdata(1,25:end)';                      % 提取变量名数据
```

```
result2 = cell(n+1, 16);                                    % 定义一个 n+1 行，8 列的元胞数组
result2(1,:) = {'标准化变量', '特征向量 t1', '特征向量 t2', '特征向量 t3', '特征向量 t4', '特征向量 t5', '特征向量 t6', '特征向量 t7', '特征向量 t8', '特征向量 t9', '特征向量 t10', '特征向量 t11', '特征向量 t12', '特征向量 t13', '特征向量 t14', '特征向量 t15'};
                                                            % result2 的第一行
result2(2:end, 1) = varname;                                % result2 的第一列
result2(2:end, 2:end) = num2cell(COEFF(:,1:15));            % 存放前 15 个主成分表达式的系数数据
[s1, id] = sortrows(result2(2:end, 2:end),1);               % 将主成分得分数据按第一主成分得分从小到大排序

% 计算生成的主成分指标
w = result2(2:end,2:end);
w2 = cell2mat(w);
pca_score = XZ*w2;

% matlab 自带 xlswrite 命令，格式 xlswrite('excel 文件名',数据变量名，第几个工作表，'单元格'）
xlswrite('附件一：325 个样本数据(附件三已追加)(数据预处理).xlsx',result1,'主成分贡献率','B1');
xlswrite('附件一：325 个样本数据(附件三已追加)(数据预处理).xlsx',pca_score,'主成分提取','Y2');
```

附录 4：问题二：构建多元非线性回归模型的 Stata 源代码（Q2_多元非线性回归模型.do）

```
// 问题二：寻找建模主要变量
* 构建多元非线性回归模型
* 2020-09-17

clear
global path "E:\ArchiveforStudy\数学建模\2020 年 B 题--汽油辛烷值建模"
cd "E:\ArchiveforStudy\数学建模\2020 年 B 题--汽油辛烷值建模"

* ============================================================
* ============ 主成分分析降维后多元非线性回归模型分析 ============
* ============================================================
/* 模型变量说明：
   被解释变量： RON 损失 =(原料辛烷值 RON - 产品辛烷值 RON)
   核心解释变量（基于文献研究法所提取）： R-101 床层温度、反应器质量空速、反应器顶部压力、氢油比、原料汽油硫含量（该项含平方）
   前 15 主成分特征解释变量（累计贡献率：80.1467%）：pca1-pca15
 */
import excel "$path\附件一：325 个样本数据(附件三已追加)(数据预处理).xlsx", sheet("主成分提取") firstrow clear
* 逐步构建回归模型，并将结果导出到 Word
putdocx begin
putdocx save "多元非线性回归模型结果.docx", replace
reg RON 损失 去除硫含量
   est store m1
reg RON 损失 去除硫含量 产品硫含量
   est store m2
reg RON 损失 去除硫含量 产品硫含量 烯烃 v 待生吸附剂_焦炭-氢油比 原料汽油硫含量
   est store m3
reg RON 损失 去除硫含量 产品硫含量 烯烃 v 待生吸附剂_焦炭-氢油比 c.原料汽油硫含量##c.原料汽油硫含量
   est store m4
reg RON 损失 去除硫含量 产品硫含量 烯烃 v 待生吸附剂_焦炭-氢油比 c.原料汽油硫含量##c.原料汽油硫含量 pca1-pca15
   est store m5
reg2docx m1 m2 m3 m4 m5 using "多元非线性回归模型结果.docx", ///
            r2(%6.3f) b(%9.3f) t(%7.2f) bic(%7.2f) ///
            title("表 1: 多元非线性回归模型结果") append

 keep 样本号 采样日期 产品硫去除率 辛烷值损失率 去除硫含量 产品硫含量 RON 损失 原料辛烷值 RON 产品辛烷值 RON 烯烃 v 待生吸附剂_焦炭-pca15
export excel "$path\附件一：325 个样本数据(附件三已追加)(数据预处理).xlsx", sheet("后续分析数据基础") firstrow(variables)
```

附录 5：问题三：多种数据挖掘分析算法的 Python 源代码（Q3_辛烷值预测.py）

```
# -*- coding: utf-8 -*-
"""
Created on Sat Sep 19 14:24:53 2020

"""
import xlrd
import numpy as np
from sklearn.svm import SVR
import matplotlib.pyplot as plt
from sklearn.linear_model import Ridge
from sklearn.metrics import mean_squared_error
from mlxtend.regressor import StackingRegressor
from sklearn.ensemble import RandomForestRegressor


#============================================================================
#**************************    数据清洗    ************************
#============================================================================
def ReadData(path,table):
    wb = xlrd.open_workbook(path)          # 打开 Excel 文件
    sheet = wb.sheet_by_name(table)        # 通过 excel 表格名称(rank)获取工作表
    original_data = []
    for a in range(sheet.nrows):           # 循环读取表格内容（每次读取一行数据）
        cells = sheet.row_values(a)        # 每列数据赋值给 cells
        original_data.append(cells)
    return original_data

def constructFeatureSet(original_data,y):
    temp = np.array(original_data)
    temp = np.delete(temp,0,axis=0)        # 删除第一行文字
    temp = np.delete(temp,0,axis=1)        # 删除第一列
    temp = np.delete(temp,0,axis=1)        # 删除第一列
    temp = temp.astype(np.float64)         # 转换数据格式为 float64
    if y == 1:
        temp = np.delete(temp,1,axis=1)    # 删除第二列

    else:
        temp = np.delete(temp,0,axis=1)    # 删除第一列
    return temp

def generateTTdata(feature_set,num_train):
    y_data = feature_set[:,0]
    x_data = np.delete(feature_set,0,axis=1)
    x_test,x_train = np.split(x_data,[num_train])
    y_test,y_train = np.split(y_data,[num_train])
    return x_train,x_test,y_train,y_test

#============================================================================
#**************************    绘图分析预测误差的函数    ************************
#============================================================================
def plotLinearError(pred,y_test,name,y):
    if y==1:
        object_temp = '辛烷损失值'
    else:
        object_temp = '硫去除值'
    #线性图分析各算法的预测误差
    plt.figure()
    plt.rcParams['font.sans-serif'] = ['SimHei']    # 解决中文显示
    plt.rcParams['axes.unicode_minus'] = False   # 解决符号无法显示
    plt.title(name)
    plt.ylabel(object_temp,size=12,position=(-1,0.5))
    plt.xlabel('测试次数',size=12)
```

```
    #绘图
    plt.scatter(np.arange(len(y_test)), pred,c='r',label='预测值')
    plt.plot(np.arange(len(y_test)), pred,linestyle='--',c='r')
    plt.scatter(np.arange(len(y_test)),y_test,c='g',label='实际值')
    plt.plot(np.arange(len(y_test)), y_test,linestyle='--',c='g')
    plt.legend(loc="best",fontsize=10)
    plt.show()

def plotSumError(errorValue,y_test,y):
    if y==1:
        object_temp = '辛烷损失值的预测误差'
    else:
        object_temp = '硫去除值的预测误差'
    if y==1:
        title_temp = '辛烷损失值的预测误差分析图'
    else:
        title_temp = '硫去除值的预测误差分析图'
    #绘制各模型误差的汇总图
    plt.figure()
    plt.rcParams['font.sans-serif'] = ['SimHei']          # 解决中文显示
    plt.rcParams['axes.unicode_minus'] = False          # 解决符号无法显示
    plt.title(title_temp)
    plt.ylabel(object_temp,size=12,position=(-1,0.5))
    plt.xlabel('测试次数',size=12)
    #绘图
    y = [0 for i in range (len(y_test))]
    plt.plot(np.arange(len(y_test)), errorValue[:,2],c='g',label='随机森林预测')
    plt.plot(np.arange(len(y_test)), errorValue[:,3],c='darkred',label='复合模型预测')
    plt.plot(np.arange(len(y_test)), errorValue[:,1],c='brown',label='岭回归预测')
    plt.plot(np.arange(len(y_test)), errorValue[:,0],c='y',label='linear 支持向量机预测')
    plt.plot(np.arange(len(y_test)), y,c='red',label='误差线')
    plt.legend(loc="best",fontsize=10)
    plt.show()

def plotAccumulateError(errorValue,y_test,y):
    if y==1:
        object_temp = '辛烷损失值的预测误差'
    else:
        object_temp = '硫去除值的预测误差'
    if y==1:
        title_temp = '辛烷损失值的预测误差分析图'
    else:
        title_temp = '硫去除值的预测误差分析图'
    #计算累计误差值
    accumulat_error = []
    temp_error = [0 for i in range(len(errorValue[0]))]

    for i in range(len(errorValue)):
        new_temp_error = []
        for j in range(len(errorValue[0])):
            temp = temp_error[j] + abs(errorValue[i][j])
            new_temp_error.append(temp)
        temp_error = new_temp_error.copy()
        accumulat_error.append(temp_error)
    accumulat_error = np.array(accumulat_error)
    #绘制各模型累计误差的汇总图
    plt.figure()
    plt.rcParams['font.sans-serif'] = ['SimHei']          # 解决中文显示
    plt.rcParams['axes.unicode_minus'] = False          # 解决符号无法显示
    plt.title(title_temp)
    plt.ylabel(object_temp,size=12,position=(-1,0.5))
    plt.xlabel('测试次数',size=12)
    #绘图
```

```
        plt.plot(np.arange(len(y_test)), accumulat_error[:,2],c='g',label='随机森林预测')
        plt.plot(np.arange(len(y_test)), accumulat_error[:,3],c='darkred',label='复合模型预测')
        plt.plot(np.arange(len(y_test)), accumulat_error[:,1],c='brown',label='岭回归预测')
        plt.plot(np.arange(len(y_test)), accumulat_error[:,0],c='y',label='linear 支持向量机预测')
        plt.legend(loc="best",fontsize=10)
        plt.show()


'''==========================================================================
############################## 1 运行主程序 ###################################
=========================================================================='''
#*******************************  读入数据  **********************************
table = '后续分析数据基础'
path = r'C:\Users\54074\Desktop\2020 数学建模比赛\4_算法代码\1_辛烷值损失预测模型\附件一：325 个样本数据(附件三已追加)(数据预处理).xlsx'
original_data = ReadData(path,table)

#*******************************  构造特征集  ********************************
y = 1#'RON 损失'
feature_set = constructFeatureSet(original_data,y)

#*****************************  生成训练集数据和测试集  **********************
num_train = 50 #训练集样本
x_train,x_test,y_train,y_test = generateTTdata(feature_set,num_train)

'''==========================================================================
############################## 2 构建预测模型 ##################################
=========================================================================='''
#*****************************  1.linear_SVR 预测模型  ***********************
svr_lin = SVR(kernel='linear', gamma='auto')
#*****************************  2.岭回归预测模型  ****************************
ridge = Ridge(random_state=2019)
#*****************************  3.随机森岭预测模型  **************************
rfr = RandomForestRegressor(n_estimators=200, max_features=0.7)

'''==========================================================================
############################## 3 训练辛烷损失预测模型 ##########################
=========================================================================='''
#*******************************  训练模型  **********************************
models = [svr_lin,ridge,rfr]
name = ['linear 支持向量机预测模型','岭回归预测模型','随机森林预测模型','基于 stacking 策略的复合预测模型']
#预测值与真实值的对比图
print('base model')
count = 0
errorValue = np.zeros(len(y_test))
for model in models:
        model.fit(x_train, y_train)
        pred = model.predict(x_test)
        temp = pred - y_test
        errorValue =np.c_[errorValue,temp]
        if model is ridge:
                object_coefficient = model.coef_
        plotLinearError(pred,y_test,name[count],y)
        print(name[count],"loss is {}".format(mean_squared_error(y_test, pred)))
        count += 1
#*************************  构建 stacking 预测模型  ***************************
sclf = StackingRegressor(regressors=models, meta_regressor=ridge)
#*************************  训练 stacking 预测模型  ***************************
sclf.fit(x_train, y_train)
pred = sclf.predict(x_test)
temp = pred - y_test
errorValue =np.c_[errorValue,temp]
errorValue = np.delete(errorValue,0,axis=1)
```

```
print('stacking model')
print("loss is {}".format(mean_squared_error(y_test, pred)))

# ************************  绘图分析预测误差  ********************************
plotLinearError(pred,y_test,name[3],y)
plotSumError(errorValue,y_test,y)
plotAccumulateError(errorValue,y_test,y)

'''==========================================================================
############################## 4 构建硫去除预测模型 ########################
=========================================================================='''
#*******************************  构造特征集  ******************************
y = 2#'硫去除'
feature_set = constructFeatureSet(original_data,y)

#*****************************  生成训练集数据和测试集  ***********************
num_train = 50 #训练集样本
x_train,x_test,y_train,y_test = generateTTdata(feature_set,num_train)

#*****************************  1.linear_SVR 预测模型  ************************
svr_lin = SVR(kernel='linear', gamma='auto')
#*******************************  2.岭回归预测模型  ****************************
ridge = Ridge(random_state=2019)
#*******************************  3.随机森岭预测模型  **************************
rfr = RandomForestRegressor(n_estimators=200, max_features=0.7)

'''==========================================================================
############################## 5 训练硫去除预测模型 ########################
=========================================================================='''
models = [svr_lin,ridge,rfr]
name = ['linear 支持向量机预测模型','岭回归预测模型','随机森林预测模型','基于 stacking 策略的复合预测模型']
#预测值与真实值的对比图
print('base model')
count = 0
errorValue = np.zeros(len(y_test))
for model in models:
    model.fit(x_train, y_train)
    pred = model.predict(x_test)
    temp = pred - y_test
    errorValue =np.c_[errorValue,temp]
    if model is svr_lin:
        variable_coefficient = model.coef_
    plotLinearError(pred,y_test,name[count],y)
    print(name[count],"loss is {}".format(mean_squared_error(y_test, pred)))
    count += 1
#************************  构建 stacking 预测模型  ***************************
sclf = StackingRegressor(regressors=models, meta_regressor=ridge)
#************************  训练 stacking 预测模型  ***************************
sclf.fit(x_train, y_train)
pred = sclf.predict(x_test)
temp = pred - y_test
errorValue =np.c_[errorValue,temp]
errorValue = np.delete(errorValue,0,axis=1)
print('stacking model')
print("loss is {}".format(mean_squared_error(y_test, pred)))
#************************  绘图分析预测误差  ********************************
plotLinearError(pred,y_test,name[3],y)
plotSumError(errorValue,y_test,y)
plotAccumulateError(errorValue,y_test,y)
```

附录 6：问题四：基于熵权法改进的 TOPSIS 理想点法模型

假设原始数据由 m 个评价对象和 n 个评价指标组成，相应的判断矩阵为$A=(a_{ij})_{m\times n}$，$(i=1,2,\dots,m;j=1,2,\dots,n)$。

首先，对原始数据中低优指标做取倒数变换转化为高优指标，确保指标同趋势化，并进行如下 0-1 标准化处理。

其次，计算信息熵并定义指标j的权重：

$$\omega_i=\frac{(1-e_i)}{\sum_{i=1}^{m}(1-e_i)} \tag{A1}$$

其中，$e_i=-\frac{1}{\ln m}\sum_{i=1}^{m}p_{ij}\ln p_{ij}$，$p_{ij}=\frac{b_{ij}}{\sum_{i=1}^{m}b_{ij}}$。

再次，对原始决策矩阵进行向量规范化，得到加权规范化决策矩阵：

$$B=\left(\omega_i\frac{a'_{ij}}{\sqrt{\sum_{i=1}^{m}a'^{2}_{ij}}}\right)_{m\times n} \tag{A2}$$

然后，根据规范化决策矩阵B得到正负理想解：

$$\begin{cases}C^{+}=(b_{i1}^{+},b_{i2}^{+},\cdots,b_{in}^{+})\\C^{-}=(b_{i1}^{-},b_{i2}^{-},\cdots,b_{in}^{-})\end{cases} \tag{A3}$$

紧接着，计算各评价对象到正负理想解的距离：

$$\begin{cases}D^{+}=\sqrt{\sum_{j=1}^{n}\left(b_{ij}^{+}-b_{ij}\right)^2}\\D^{-}=\sqrt{\sum_{j=1}^{n}\left(b_{ij}^{-}-b_{ij}\right)^2}\end{cases} \tag{A4}$$

最后，得到第 i 个观察值的综合评价指数：

$$C_i=\frac{D^{-}}{D^{+}+D^{-}} \tag{A5}$$

其中，$C_i\in[0,1]$，其值越趋近于 1，表明该评价对象越接近最优水平。

附录 7：问题四：基于熵权法改进的 TOPSIS 模型的 MATLAB 源代码（Q4_SZF_Topsis.m）

```
%==========================================================================
%       提取附件一第工作表 4“后续分析数据基础”数据，进行熵权 TOPSIS 评价
%==========================================================================
%    2020-09-19
```

```
clc, clear;
[X,textdata] = xlsread('附件一：325 个样本数据(附件三已追加)(数据预处理).xlsx','后续分析数据基础');
% 评价指标：产品硫去除率 3 辛烷值损失率 4- 待生吸附剂_焦炭 11 待生吸附剂_S12- 再生吸附剂_焦炭 13- 再生吸附剂_S14 反应温度 15- 质量空速 16 反应压力 17- 氢油比 18-
data = X(:,[3,4,11:18]);

% 结合相关文献与多元回归模型拟合结果，负向指标取倒数转换为正向指标，其中：
%    辛烷值损失率 为负向指标，取倒转正
%    待生吸附剂持硫、再生吸附剂持碳、反应温度、反应压力、氢油比 越低越好
%    产品硫去除率 越高越好
XZ = data;
XZ(:,[2,4,5,7,9,10]) = ones(325,6)./data(:,[2,4,5,7,9,10]);
XZ=zscore(XZ);                                  % 数据标准化
[obs,vars]=size(XZ);                            % obs 为总观察期数；n 为评价指标个数

%% ==================== 熵权法确定权重 w ==========================
for i=1:obs
    for j=1:vars
        p(i,j)=XZ(i,j)/sum(XZ(:,j));            % 计算第 j 个指标下，第 i 个对象占该指标比重
    end
end
k=1/log(obs);
for j=1:vars
    e(j)=-k*sum(p(:,j).*log(p(:,j)));     % 计算第 j 个指标的熵值 e(j)
end
d=ones(1,vars)-e;                               % 计算信息熵冗余度
w=d./sum(d);                                    % 求权值 w
s=w*XZ';
sq_score=s';                                    % 熵值法的综合得分

%% ============ 基于熵权法改进的 TOPSISI 评价模型测度 ===============
B=XZ.*repmat(w,obs,1);                          % 求加权矩阵，权重来源于 熵值法
Cstar=max(B);                                   % 求正理想解
C0=min(B);                                      % 求负理想解
for i=1:obs
    Dstar(i)=norm(XZ(i,:)-Cstar);               % 求到正理想解的距离
    D0(i)=norm(XZ(i,:)-C0);                     % 求到负理想的距离
end
C=D0./(Dstar+D0);                               % 求出综合得分
sqtopsis_score=C';                              % 得到评价结果

%% 为了直观，定义元胞数组 result1，用来存放熵权法得分、名次排序和 sqtopsis 得分排序
result1 = cell(obs+1,12);                         %定义一个 obs+1 行，12 列的元胞数组
result1(1,:) = {'样本号','熵权 TOPSIS 得分','产品硫去除率','辛烷值损失率','待生吸附剂_焦炭','待生吸附剂_S','再生吸附剂_焦炭','再生吸附剂_S','反应温度','质量空速','反应压力','氢油比'};
result1(2:end,1) = num2cell(X(:,1));            % 样本号
result1(2:end,2) = num2cell(sqtopsis_score);    % 熵权 TOPSIS 得分
result1(2:end,3:end) = num2cell(data);          % 存放原始变量

xlswrite('附件一：325 个样本数据(附件三已追加)(数据预处理).xlsx',result1,'熵权 TOPSIS 结果','A1');
```

附录 8：问题四：辛烷值工艺优化模型的 Python 源代码（Q4_辛烷值工艺优化模型.py）

```
# -*- coding: utf-8 -*-
"""
Created on Sat Sep 19 21:18:42 2020
"""
import re
import gurobipy as gp
from gurobipy import GRB
```

```
def readData(path):
    data = open(path,'r').readlines()
    variable_coefficient = []
    object_coefficient = []
    data_temp = []
    for i in range(len(data)):                          # 清洗:数据中的空格符，换行符
        location = re.findall(r'[\-\d][\.\d]*',data[i])
        data_temp.append(location)
    for i in data_temp[14]:
        object_coefficient.append(float(i))
    for i in data_temp[17]:
        variable_coefficient.append(float(i))
    return variable_coefficient,object_coefficient
'''==========================================================================
############################  构建工艺优化模型 ##########################
=========================================================================='''
#*****************************  读入数据  ***********************************
path = r'C:\Users\54074\Desktop\2020 数学建模比赛\3_写作资料\模型误差数据.txt'
variable_coefficient,object_coefficient =   readData(path)
#*****************************  创建变量名  *********************************
variables_name =[str(i) for i in range(len(variable_coefficient)-1)]
#*****************************  创建模型  ***********************************
m = gp.Model("process_optimize")
#*****************************  创建决策变量  *******************************
x1 = m.addVar(lb =57 ,ub =392 ,vtype = GRB.CONTINUOUS,name = variables_name[0])
x2 = m.addVar(lb=85.3,ub=91.7,vtype = GRB.CONTINUOUS,name = variables_name[1])
x3 = m.addVar(lb=14.6,ub=34.67,vtype = GRB.CONTINUOUS,name = variables_name[2])
x4 = m.addVar(lb=1.01,ub=12.15,vtype = GRB.CONTINUOUS,name = variables_name[3])
x5 = m.addVar(lb=2.94,ub=14.31,vtype = GRB.CONTINUOUS,name = variables_name[4])
x6 = m.addVar(lb=1.43,ub=13.34,vtype = GRB.CONTINUOUS,name = variables_name[5])
x7 = m.addVar(lb=0.25,ub=8.92,vtype = GRB.CONTINUOUS,name = variables_name[6])
x8 = m.addVar(lb=400,ub=450,vtype = GRB.INTEGER,name = variables_name[7])
x9 = m.addVar(lb=3,ub=7,vtype = GRB.INTEGER,name = variables_name[8])
x10 = m.addVar(lb=2,ub=2.45,vtype = GRB.CONTINUOUS,name = variables_name[9])
x11 = m.addVar(lb=0.2,ub=0.37,vtype = GRB.CONTINUOUS,name = variables_name[10])
x12 = m.addVar(lb=2,ub=645,vtype = GRB.INTEGER,name = variables_name[11])
#*****************************  创建目标函数  *******************************
obj = gp.LinExpr()
obj += x1*object_coefficient[0]
obj += x2*object_coefficient[1]
obj += x3*object_coefficient[2]
obj += x4*object_coefficient[3]
obj += x5*object_coefficient[4]
obj += x6*object_coefficient[5]
obj += x7*object_coefficient[6]
obj += x8*object_coefficient[7]
obj += x9*object_coefficient[8]
obj += x10*object_coefficient[9]
obj += x11*object_coefficient[10]
obj += x12*object_coefficient[11]
obj += object_coefficient[12]
m.setObjective(obj, GRB.MINIMIZE)
#*****************************  增加约束条件  *******************************
constr1 = gp.LinExpr()
constr1 += x1*(variable_coefficient[0]-1)
constr1 += x2*variable_coefficient[1]
constr1 += x3*variable_coefficient[2]
constr1 += x4*variable_coefficient[3]
constr1 += x5*variable_coefficient[4]
constr1 += x6*variable_coefficient[5]
constr1 += x7*variable_coefficient[6]
constr1 += x8*variable_coefficient[7]
constr1 += x9*variable_coefficient[8]
constr1 += x10*variable_coefficient[9]
```

```
constr1 += x11*variable_coefficient[10]
constr1 += x12*variable_coefficient[11]
constr1 += variable_coefficient[12]
m.addConstr(-1*constr1,GRB.LESS_EQUAL,5, "c0")
#============================================================================
#*****************************  运行优化模型  *********************************
#============================================================================
m.write('model.lp')                          # 导出模型
try:
    m.optimize()
except gp.GurobiError:
    print("Optimize failed due to non-convexity")
# Solve bilinear model
m.params.NonConvex = 2
m.optimize()
#*******************************  输出结果  ***********************************
m.printAttr('x')
m.optimize()
m.printAttr('x')
```

附录 9：问题五：133 号样本操作变量调整可视化的 Python 源代码（Q5_工艺调整可视化.py）

```
# -*- coding: utf-8 -*-
"""
Created on Sun Sep 20 14:03:13 2020
"""
import matplotlib.pyplot as plt
#============================================================================
#*****************************  导入基本数据  *********************************
#============================================================================
par_optimazation = [248.00,89.40,20.60,2.53,8.57,1.30,6.69,400.00,7.00,2.00,0.2,590] # 优化后的参数
par_original = [248.00,89.40,20.60,2.53,8.57,1.30,6.69,417.91,4.42,2.25,0.30,270.38]  # 原始的参数
par_unit = [-1,10,0.5,-0.1,-0.01]                    # 可优化参数的最小变化的单位
par_name = ['X8(R-101 床层下部温度)','X9(反应器质量空速)','X10(反应系统压力)','X11(氢油比)','X12(氢油比)']
#**************************  导入测算函数的数据  ******************************
coef_xw = [0.0001802,-0.019784,0.0026811,0.0378414,-0.025014,-0.0767658,0.0420859,\
           -0.0098129,0.0135879,-0.5043296,-2.848329,0.0005352,96.61]
coef_l = [-0.9994184,-0.0292569,0.100233,0.1212143,-0.0053392,-0.0811863,0.0751495,\
          -0.0019653,-0.0228035,2.127755,0.8241614,0.0032152,246.4044]

#============================================================================
#*****************************  计算汽油的辛烷值和硫含量  *********************
#============================================================================
number_adjust = [17,12,3,2,10]
temp_li = [7,10,6,6,6]

value_xw = []
value_l = []

for i in range(len(number_adjust)):
    for j in range(number_adjust[i]):
        par_original[i+temp_li[i]] += par_unit[i]
        temp_value_xw = 0
        temp_value_l = 0
        #计算参数调整后的辛烷值和硫含量值
        for k in range(len(par_optimazation)):
            temp_value_xw += par_original[k]*coef_xw[k]
            temp_value_l += par_original[k]*coef_l[k]
        temp_value_xw += coef_xw[-1]
        temp_value_l += coef_l[-1]
        value_xw.append(temp_value_xw)
        value_l.append(temp_value_l)
```

```
#==========================================================================================
#**********************  汽油辛烷值和硫含量变化过程的可视化  ******************
#==========================================================================================
object_temp = '汽油中辛烷值的含量'
title_temp = '参数调整过程中汽油的辛烷值变化图'
plt.figure()
plt.rcParams['font.sans-serif'] = ['SimHei']        # 解决中文显示
plt.rcParams['axes.unicode_minus'] = False      # 解决符号无法显示
plt.title(title_temp)
plt.ylabel(object_temp,size=12,position=(-1,0.5))
plt.xlabel('参数调整次数',size=12)
#绘图
plt.plot([i for i in range(len(value_xw))], value_xw,c='green',label='汽油中的辛烷值含量')
plt.legend()
plt.show()
#***************************  硫含量变化可视化  ******************************
object_temp = '汽油中硫的含量'
title_temp = '参数调整过程中汽油的硫含量变化图'
plt.figure()
plt.rcParams['font.sans-serif'] = ['SimHei']        # 解决中文显示
plt.rcParams['axes.unicode_minus'] = False      # 解决符号无法显示
plt.title(title_temp)
plt.ylabel(object_temp,size=12,position=(-1,0.5))
plt.xlabel('参数调整次数',size=12)
#绘图
plt.plot([i for i in range(len(value_l))], value_l,c='green',label='汽油中的硫含量')
plt.legend()
plt.show()
```

附录 10: 问题五：133 号样本操作变量调整可视化的 Python 源代码（Q5_批次优化模型.py）

```
# -*- coding: utf-8 -*-
"""
Created on Sun Sep 20 22:26:12 2020
"""
import matplotlib.pyplot as plt
import gurobipy as gp
from gurobipy import GRB

#==========================================================================================
#*****************************  导入基本数据  ********************************
#==========================================================================================
par_optimazation = [248.00,89.40,20.60,2.53,8.57,1.30,6.69,400.00,7.00,2.00,0.2,590]  # 优化后的参数
par_original = [248.00,89.40,20.60,2.53,8.57,1.30,6.69,417.91,4.42,2.25,0.30,270.38]   # 原始的参数
par_unit = [-1,0.5,-0.1,-0.01,10]                       # 可优化参数的最小变化的单位
par_name = ['X8(R-101 床层下部温度)','X9(反应器质量空速)','X10(反应系统压力)','X11(氢油比)','X12(氢油比)']
#***************************  导入测算函数的数据  ******************************
coef_xw = [0.0001802,-0.019784,0.0026811,0.0378414,-0.025014,-0.0767658,0.0420859,\
                    -0.0098129,0.0135879,-0.5043296,-2.848329,0.0005352,96.61]
coef_l = [-0.9994184,-0.0292569,0.100233,0.1212143,-0.0053392,-0.0811863,0.0751495,\
                    -0.0019653,-0.0228035,2.127755,0.8241614,0.0032152,246.4044]

object_coefficient1 = [-0.0019653,-0.0228035,2.127755,0.8241614,0.0032152]
object_coefficient2 = [-1,0.5,-0.1,-0.01,10]

#*****************************  创建变量名  *********************************
variables_name =[str(i) for i in range(1,16)]
#*****************************  创建模型  ***********************************
m = gp.Model("process_optimize")
#*****************************  创建决策变量  *******************************
```

```
k=8                                                    # 调整参数的批次
n =5
v_n = [i for i in range(k*5)]
vars = m.addVars(v_n,lb = 0,ub = 3, vtype=GRB.INTEGER, name='c')
#*****************************   创建目标函数   ********************************
count = 0                                              # 计数的变量
temp_object = []                                       # 构建 tuplist 存储表达式
temp_object = gp.tuplelist(temp_object)

for i in range(k):                                     # 第 k 个批次的值
    constr = []                                        # 初始化一个列表
    constr = gp.LinExpr()                              # 初始化一个表达式
    for j in range(n):                                 # 依次循环每一个特征
        constr +=vars[j+count]*object_coefficient1[j]*object_coefficient2[j]
    constr +=246.4044
    temp_object.append(constr)                         # 获得 k 个表达式
    count+=5

obj = gp.QuadExpr()                                    # 创建一个新的表达式
for i in range(len(temp_object)):                      # 创建目标函数
    obj += temp_object[i]*temp_object[i]
m.setObjective(obj, GRB.MINIMIZE)
#*****************************   增加约束条件   ********************************
p=0
temp = []
temp = gp.tuplelist(temp)
for i in range(0+p,n*k-1,5):
    temp.append(vars[i])
constr1 = gp.LinExpr()
for i in temp:
    constr1 += i
m.addConstr(constr1,GRB.LESS_EQUAL,17, "c2")
m.addConstr(constr1,GRB.GREATER_EQUAL,15, "c3")
p +=1


temp = []
temp = gp.tuplelist(temp)
for i in range(0+p,n*k-1,5):
    temp.append(vars[i])

constr2 = gp.LinExpr()
for i in temp:
    constr2 += i
m.addConstr(constr2,GRB.LESS_EQUAL,3, "c2")
m.addConstr(constr2,GRB.GREATER_EQUAL,2, "c3")
p+=1


temp = []
temp = gp.tuplelist(temp)
for i in range(0+p,n*k-1,5):
    temp.append(vars[i])

constr3 = gp.LinExpr()
for i in temp:
    constr3 += i
m.addConstr(constr3,GRB.LESS_EQUAL,2, "c4")
m.addConstr(constr3,GRB.GREATER_EQUAL,1, "c5")
p+=1

temp = []
temp = gp.tuplelist(temp)
for i in range(0+p,n*k-1,5):
    temp.append(vars[i])
```

```
constr4 = gp.LinExpr()
for i in temp:
    constr4 += i
m.addConstr(constr4,GRB.LESS_EQUAL,10, "c6")
m.addConstr(constr4,GRB.GREATER_EQUAL,8, "c7")
p+=1

temp = []
temp = gp.tuplelist(temp)
for i in range(0+p,n*k-1,5):
    temp.append(vars[i])

constr5 = gp.LinExpr()
for i in temp:
    constr5 += i
m.addConstr(constr5,GRB.LESS_EQUAL,12, "c8")
m.addConstr(constr5,GRB.GREATER_EQUAL,8, "c9")

#============================================================================
#*****************************  运行优化模型  *********************************
#============================================================================
m.write('oder_optimization.lp')# 导出模型
try:
    m.optimize()
except gp.GurobiError:
    print("Optimize failed due to non-convexity")
# Solve bilinear model
m.params.NonConvex = 2
m.optimize()
#*******************************  输出结果  ***********************************
m.printAttr('x')
m.optimize()
m.printAttr('x')
#*******************************  保存结果  ***********************************
decision_variables = m.getVars()
adjust_times = []
temp = []
count = 0
for i in range(len(decision_variables)):
    temp.append(decision_variables[i].x)
    if len(temp)==5:
        adjust_times.append(temp)
        temp = []

# # #============================================================================
# # #*****************************  计算汽油的辛烷值和硫含量  *********************
# # #============================================================================
value_xw = []
value_l = []
for i in range(len(adjust_times)):
    par_original[7] += par_unit[0]*adjust_times[i][0]
    par_original[8] += par_unit[1]*adjust_times[i][1]
    par_original[9] += par_unit[2]*adjust_times[i][2]
    par_original[10] += par_unit[3]*adjust_times[i][3]
    par_original[11] += par_unit[4]*adjust_times[i][4]

    temp_value_xw = 0
    temp_value_l = 0
    #计算参数调整后的辛烷值和硫含量值
    for k in range(len(par_optimazation)):
        temp_value_xw += par_original[k]*coef_xw[k]
        temp_value_l += par_original[k]*coef_l[k]
    temp_value_xw += coef_xw[-1]
    temp_value_l += coef_l[-1]
```

```
        value_xw.append(temp_value_xw)
        value_l.append(temp_value_l)

# #=============================================================================
# #**********************  汽油辛烷值和硫含量变化过程的可视化  ******************
# #=============================================================================
object_temp = '汽油中辛烷值的含量'
title_temp = '参数调整过程中汽油的辛烷值变化图'
plt.figure()
plt.rcParams['font.sans-serif'] = ['SimHei'] #解决中文显示
plt.rcParams['axes.unicode_minus'] = False #解决符号无法显示
plt.title(title_temp)
plt.ylabel(object_temp,size=12,position=(-1,0.5))
plt.xlabel('参数调整次数',size=12)
#绘图
plt.plot([int(i) for i in range(1,len(value_xw)+1)], value_xw,c='green',label='汽油中的辛烷值含量')
plt.legend()
plt.show()
#**************************  硫含量变化可视化  ******************************
object_temp = '汽油中硫的含量'
title_temp = '参数调整过程中汽油的硫含量变化图'
plt.figure()
plt.rcParams['font.sans-serif'] = ['SimHei'] #解决中文显示
plt.rcParams['axes.unicode_minus'] = False #解决符号无法显示
plt.title(title_temp)
plt.ylabel(object_temp,size=12,position=(-1,0.5))
plt.xlabel('参数调整次数',size=12)
#绘图
plt.plot([int(i) for i in range(1,len(value_l)+1)], value_l,c='green',label='汽油中的硫含量')
plt.legend()
plt.show()
```

附录 11：中介效应模型影响机制检验的 Stata 源代码（test_中介效应模型传导机制分析.do）

```
// 进一步讨论：中介效应检验
* 2020-09-20

clear
global path "E:\ArchiveforStudy\数学建模\2020 年 B 题--汽油辛烷值建模"
cd "E:\ArchiveforStudy\数学建模\2020 年 B 题--汽油辛烷值建模"

* ============================================================
* ================== 进一步讨论：中介效应检验 ================
* ============================================================
import excel "$path\附件一：325 个样本数据(附件三已追加)(数据预处理).xlsx", sheet("主成分提取") firstrow clear

* 逐步构建中介效应检验回归模型，并将结果导出到 Word
putdocx begin
putdocx save "中介效应检验模型结果.docx", replace
* 模型（1）检验
reg RON 损失 去除硫含量
   est store m1
* 模型（2）检验
reg 烯烃 v 去除硫含量
   est store m2
* 模型（3）检验
reg RON 损失 烯烃 v 去除硫含量
   est store m3
reg2docx m1 m2 m3 using "中介效应检验模型结果.docx",   ///
            r2(%6.3f) b(%9.3f) t(%7.2f) bic(%7.2f) ///
            title("表 1: 中介效应检验模型结果") append
```