

乙醇耦合制备C4烯烃

摘 要

C4烯烃一种重要的化工原料,被广泛的应于化工产品以及医药中间体的生产,而通过以乙醇分子作为平台分子生产高附加值的C4烯烃是目前较为主流的一种制备方案。但在通过乙醇耦合制备C4烯烃的过程中,使用不同的催化剂组合和反应温度将会对C4烯烃选择性与收率产生很大影响。因此探究乙醇催化耦合制备C4烯烃的工艺条件具有重要的意义和价值。

针对问题一: 分别对 21 种不同催化剂组合下的反应结果进行相关性检验和显著性检验,计算不同皮尔逊相关系数之间的相对标准偏差 RSD,然后通过绘制催化剂组合一定时乙醇转化率、C4 烯烃的选择性与温度之间的折线图,探究温度与二者之间的关系;并基于控制变量的基本思想,在催化剂组合与温度一定的条件下绘制了不同产物选择性随反应时间变化的百分比堆积柱状图,最后根据相关文献的实验数据分析了反应时间对反应结果产生影响的可能原因。

针对问题二: 首先对实验数据进行了预处理,根据Co负载量、Co/SiO₂和HAP装料比等参数计算得到具体的HAP质量、SiO₂质量,将其作为影响反应结果的新指标,再对其进行标准差标准化。然后分别采用了熵权法评价模型和支持向量机回归模型(SVR)以及LIME局部可解释模型对问题中的不同反应条件所产生的乙醇转化率、C4烯烃选择性影响大小进行了定量分析。

针对问题三: 结合问题二中所建立的支持向量机回归模型,通过重新对催化剂组合与温度同C4烯烃收率进行拟合,然后采用多目标优化的方法进行最优解的计算,得到了在C4烯烃收率最大的情况下催化剂的组合与反应温度,同时结合实验已知数据进行了验证。

针对问题四: 基于问题三中所建立的C4烯烃收率模型与满足其最高时的反应条件,并根据原始实验所缺少数据从温度和催化剂组合两个不同的角度分别设计了5次实验,以此验证二者对C4烯烃收率产成的影响。

关键词: C4烯烃; 相关性检验; 熵权法; SVR; LIME模型; 多目标优化; 佐证性实验

1. 前言

1.1 问题背景

C4 烯烃作为重要的化工原料,被广泛的应用于化工产品及医药中间体的生产。传统的生产方法均采用化石能源为原料,但随着化石能源产量的短缺及对环境影响的加重,能源的供给逐渐趋向于多元化,开发新型的清洁能源显的愈加紧迫。乙醇分子可以通过生物质发酵制备,来源广泛、绿色清洁,以其为平台分子生产高附加值的 C4 烯烃,具有巨大的应用前景及经济效益,受到国内外广泛关注。

1.2 问题重述

➤ 任务一

在催化剂装配确定的条件下,研究反应温度与乙醇转化率和乙烯选择性的关系,并在温度和催化剂确定的条件下分析反应时间对反应结果的影响。

➤ 任务二

探讨不同的催化剂组合和温度对乙醇转化率以及 C4 烯烃选择性大小的影响。

➤ 任务三

确定最优的温度和催化剂组合,使得 C4 烯烃的收率最高。并在温度小于 350 度时再次确定最优解。

➤ 任务四

如果允许再增加 5 次实验,应如何设计,给出详细理由。

2. 模型准备

2.1 基本假设

以下假设均适用于本文中的所有模型:

1. 实验时操作均符合化学实验的基本规范,获得的数据均是准确,真实的。
2. 在实验中除了问题中的变量,其他环境条件均默认为恒定如压强,光照等。

2.2 名词解释

1. 选择性: 某一个产物在所有产物中的占比。
2. Co 负载量: Co 与 SiO₂ 的重量之比。例如,“Co 负载量为 1wt%”表示 Co 与 SiO₂ 的重量之比为 1:100,记作“1wt%Co/SiO₂”,依次类推。
3. HAP: 一种催化剂载体,中文名称羟基磷灰石。Co/SiO₂ 和 HAP 装料比: 指 Co/SiO₂ 和 HAP 的质量比。
4. 乙醇转化率: 单位时间内乙醇的单程转化率,其值为 $100\% \times (\text{乙醇进气量} - \text{乙醇剩余量}) / \text{乙醇进气量}$ 。

5. C4 烯烃收率：其值为乙醇转化率×C4 烯烃的选择性。

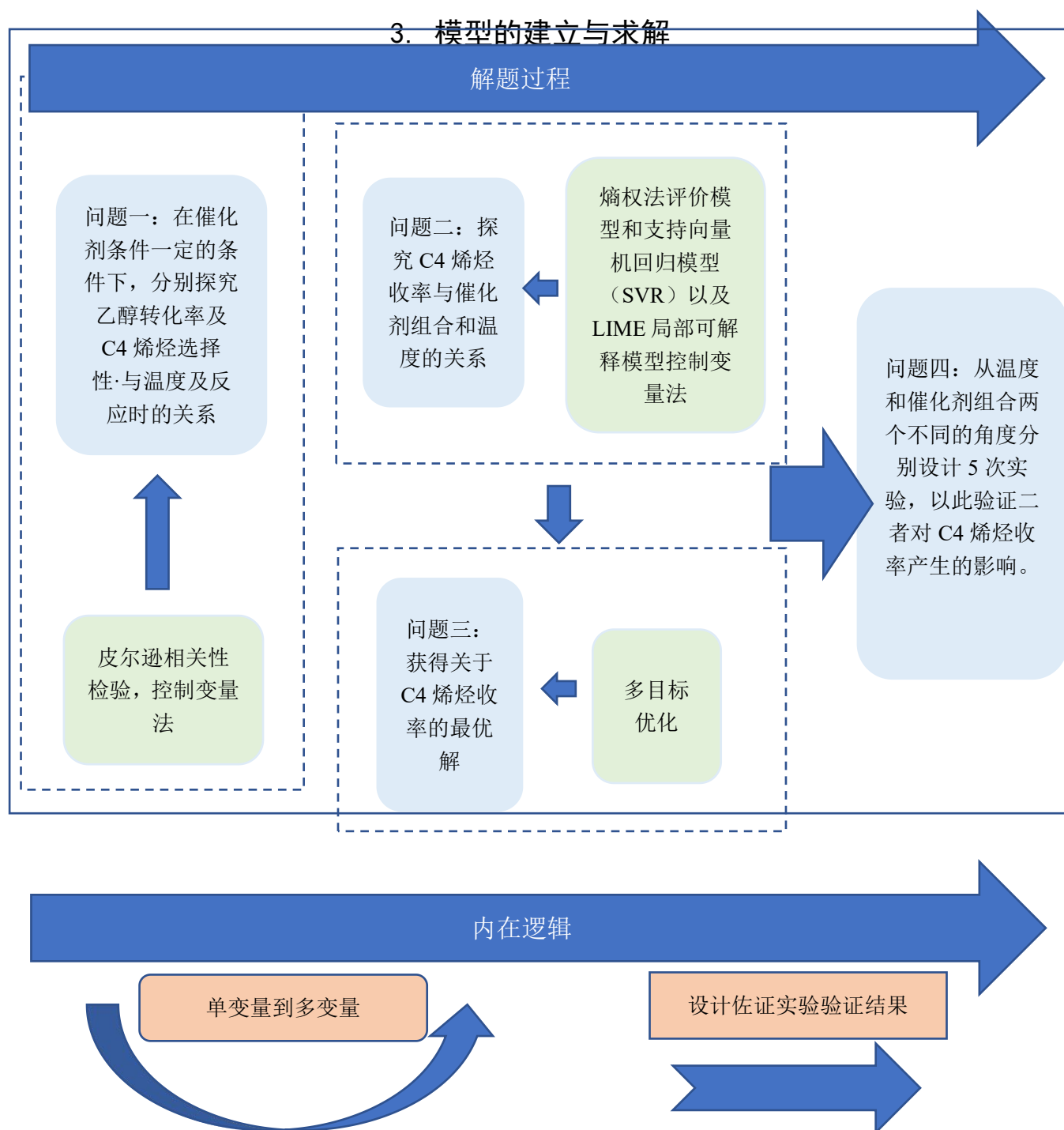


图 1.求解过程思维导图

3.1 问题一：反应结果与温度、时间相关性模型的建立与求解

3.1.1 问题一分析

对于任何化学反应而言，反应速率与温度均存在着一定的相关性，随着温度的升高，分子运动速度增大，分子间碰撞频率也会增加，反应速率因此加快。但对于不同的化学反应而言，温度对反应速率的影响也是不同的，在本题中由于温度对于生成 C_4 烯烃的反应速率与生成其它副产物的反应速率影响不同，因此也会造成在反应物乙醇通入速率一定的情况下不同温度会导致 C_4 烯烃的选择性和乙醇的转化率不同。而对于第二小问，在 350 度且给定的催化剂组合下随着时间的变化，乙醇的转化率与不同产物的选择性也发生了一定的改变，通过查阅相关文献我们找到了时间变化对其影响的主要因素——催化剂的活性。

针对本题，我们分别对不同催化剂组合下的反应结果进行皮尔逊相关性检验，探究温度与 C_4 烯烃的选择性和乙醇的转化率之间的关系；并根据相关文献的实验数据探究了在催化剂组合与温度一定的条件下反应时间对测试结果的影响。

3.1.2 温度对乙醇转化率、 C_4 烯烃的选择性影响

(1) 乙醇转化率与 C_4 烯烃选择性同温度变化趋势

根据题中所给数据，我们分别对不同催化剂组合时反应的乙醇转化率和 C_4 烯烃的选择性随温度的变化进行了定性分析。

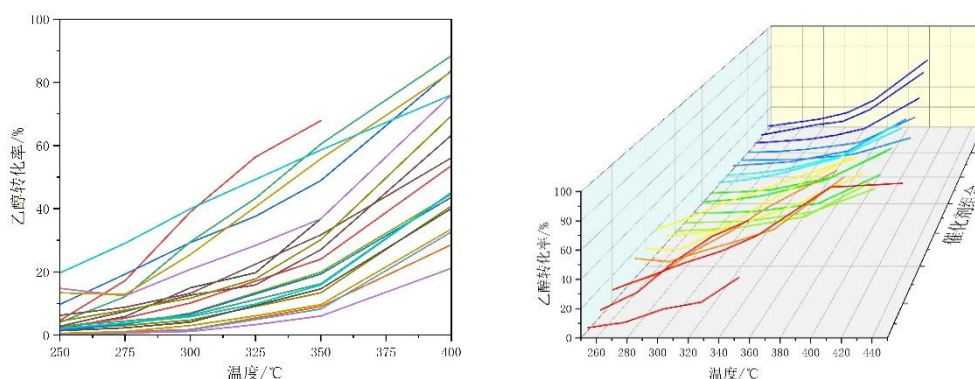


图 3-1 温度对乙醇转化率的影响

从图 3-1 中我们可以看出，随着温度的升高乙醇的转化率也在不断增大，且即使在不同催化剂组合的情况下这种增长趋势也十分类似，接近于线性增大。

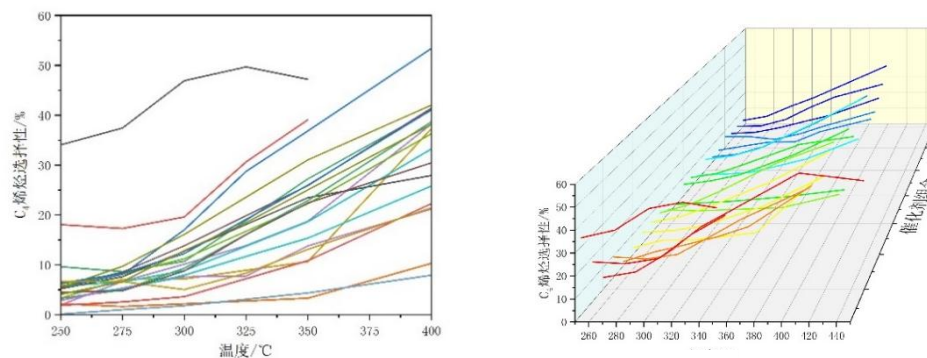


图 3-2 温度对 C_4 烯烃选择性的影响

从图 3 仍可以看出，随着温度的增大 C₄ 烯烃的选择性也在不断增大，即使在不同催化剂组合的情况下这种增长趋势也十分类似，均接近于线性增大。

(2) 皮尔逊相关性检验

皮尔逊相关系数也被称为皮尔森积矩相关系数，是一种线性相关系数，通常用来反应两个变量 X 和 Y 的线性相关程度，其值介于-1 到 1 之间，绝对值越大表明相关性越强，且大于零代表正相关，反之代表负相关。

假设有两个变量 X 和 Y，则其之间的皮尔逊相关系数可通过以下公式进行计算：

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X\sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X\sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

根据假设检验方法，在零假设成立的条件下，即假设两变量不存在相关性的前提下，计算出两变量不存在相关性的概率值（p 值），如果这个 p 值很小，说明两变量不存在相关性的概率很小，我们就可以拒绝原假设，接受备择假设。通常以 5%为阈值即显著性水平，如果 p<0.05,则说明可以拒绝原假设，接受备择假设，即两变量之间存在显著的线性关联。

因此，分别将不同催化剂组合下的温度与乙醇转化率、C₄ 烯烃选择性的数据代入上述计算公式便可得到不同催化剂组合下的皮尔逊相关系数与显著性检验的 p 值，如下表所示：

表 3-1.相关性检验与显著性检验

催化剂组合	皮尔逊相关系数 与p值	温度与乙醇转化率	p值	温度与C ₄ 烯烃选择性	p值
A1		0.96549	0.00766	0.88708	0.04477
A2		0.99499	0.00043	0.91429	0.02973
A3		0.98198	0.00008	0.95541	0.00079
A4		0.99751	0.00001	0.95776	0.00264
A5		0.93455	0.00629	0.96958	0.00137
A6		0.98359	0.00252	0.88541	0.04575
A7		0.99939	0.00002	0.96878	0.00678
A8		0.97715	0.00413	0.99158	0.00093
A9		0.92097	0.02635	0.99740	0.00016
A10		0.92286	0.02542	0.86146	0.06060
A11		0.90305	0.03570	0.98905	0.00137
A12		0.96306	0.00848	0.98316	0.00262
A13		0.93628	0.01912	0.98832	0.00151
A14		0.96390	0.00819	0.95915	0.00985
B1		0.96206	0.00882	0.98578	0.00203
B2		0.92927	0.02234	0.98478	0.00225
B3		0.89015	0.01744	0.97097	0.00125
B4		0.89929	0.01470	0.89500	0.01596
B5		0.91241	0.01117	0.97762	0.00075
B6		0.93991	0.00531	0.98212	0.00048
B7		0.93612	0.00599	0.99440	0.00005

从上表中可以看出显著性检验 p 值均小于 0.05，因此温度与乙醇转化率和 C₄ 烯烃的选择性具有显著的相关性。

从图 3-3 中可以看出不同催化剂组合时温度与乙醇转化率之间的皮尔逊相关系数的相对标准偏差 RSD=3.56%，说明不同催化剂组合对温度与乙醇转化率之间线性相关程度的影响不大，且由于皮尔逊相关系数较高，因此可以说明温度与

乙醇转化率之间具有较强的正线性相关性；温度与 C₄ 烯烃选择性之间的皮尔逊相关系数的相对标准偏差 RSD=4.364%,说明不同催化剂组合对温度与 C₄ 烯烃选择性之间的线性相关程度的影响不大，且由于皮尔逊相关系数较高，因此同样可以说明温度与 C₄ 烯烃选择性之间具有较强的正线性相关性。

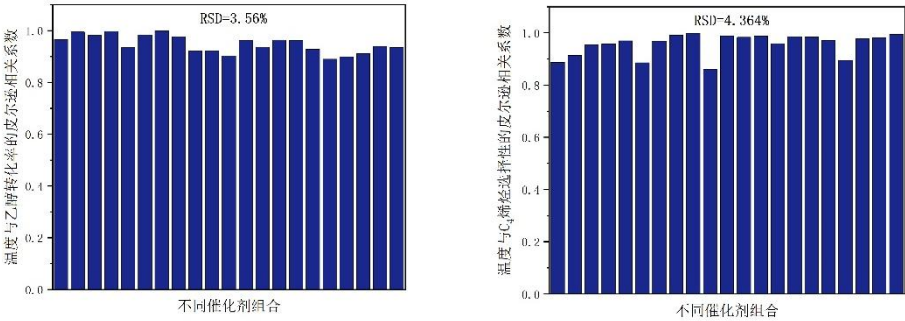


图 3-3 皮尔逊相关系数的相对标准偏差 RSD

3.1.3 反应时间对测试结果的影响

对于附件中 350 度时给定的催化剂组合在不同时间的测试结果进行分析。由于在温度和催化剂配比一定的情况下，只有反应时间这一单一变量对乙醇的转化率以及产物的选择性产生影响。因此我们在不同反应时间下，绘制了乙醇转化率以及不同产物的选择性关于反应时间的折线图以及百分比堆积柱状图。

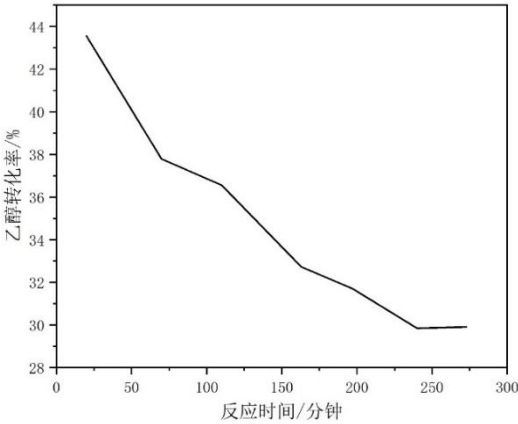


图 3-4 乙醇转化率随时间的变化

随着反应时间增加，由图 3-4 可知，在反应时间为 20min 时，乙醇转化率为 43.5%，反应时间为 70min 时，乙醇转化率为 37.8%，反应时间为 110min 时，乙醇转化率为 36.6%，反应时间为 163min 时，乙醇转化率为 32.7%，反应时间为 197min 时，乙醇转化率为 31.7%，反应时间为 240min 时，乙醇转化率为 29.9%，反应时间为 273min 时，乙醇转化率为 29.9%。由此可见随着反应时间的增加乙醇的转化率持续下降，并在反应时间超过 250 分钟时趋于稳定，即乙醇转化率稳定在 30%左右。

通过查阅相关文献及实验数据，对于乙醇转化率随着反应时间的增加呈明显的下降趋势，其主要原因是随着反应时间的增加，催化剂的活性受到了一定的影响[1]。如图 3-5 所示，Co 在 SiO₂ 及 HAP 表面处于一种吸附的状态，而随着反应时间的增加 Co/SiO₂ 表面 Co 颗粒的聚集和金属的流失便会导致催化剂与反应物的接触面积减小，催化剂的催化效果下降，即催化剂活性降低。

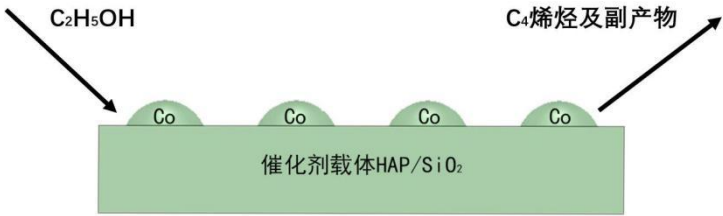


图 3-5 催化反应过程示意图

从图 3-6 和图 3-7 可以看出，对于生成的产物，反应时间由 20min 增加到 273min 时，乙烯的选择性的变化分别为 4.23%，4.28%，4.46%，4.63%，4.62%，4.76%，4.68%，由此可见时间的变化对乙烯选择性的影响几乎可以忽略；C₄ 烯烃的选择性的变化分别为 39.9%，38.55%，36.72%，39.53%，38.96%，40.32%，39.04%，由此可见我们所需要的产物乙醛的选择性在 40%上下波动，且随着时间增长，波动幅度不大；乙醛选择性的变化为 5.17%，5.6%，6.37%，7.82%，8.19%，8.42%，8.79%由此可见，乙醛在副产物中占比不大，且随着反应时间的增加，乙醛在副产物中的产量有所增加；碳数为 4-12 脂肪醇的选择性变化为 39.7%，37.36%，32.39%，31.29%，31.49%，32.36%，30.86%，显然 C₄-12 的脂肪醇在副产物中的占比最大，且随着反应时间的增加，脂肪醇的选择性略有下降。而甲基苯甲醛和甲基苯甲醇的选择性变化为 2.58%，4.28%，4.63%，4.8%，4.26%，4.48%，3.95%，由此可见该种副产物在总的副产物中占比不大且选择性具有一定不确定性。

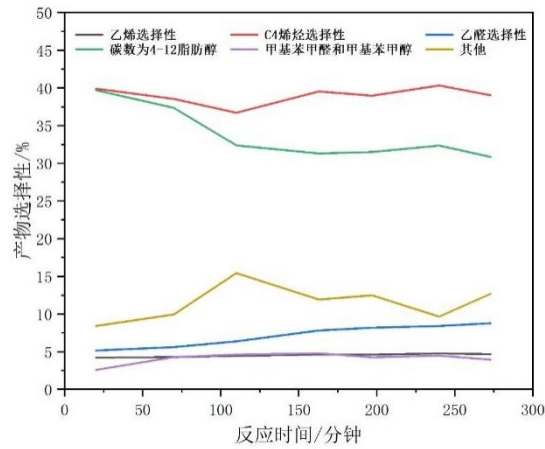


图 3-6 不同产物选择性随时间变化折线图

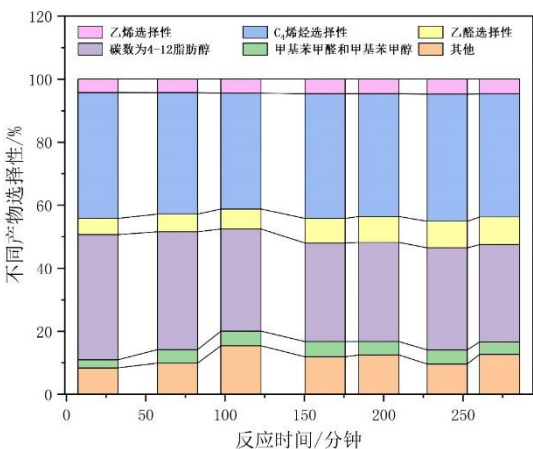
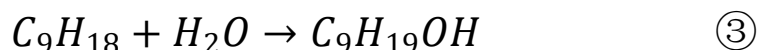
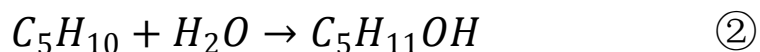
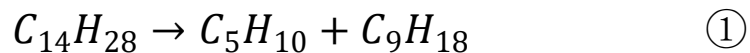


图 3-7 不同产物选择性随时间变化比例百分比堆积柱状图

根据相关有机反应原理可知,在反应过程中多碳的副产物会裂解成更小的烯烃,而这些烯烃又比较容易和水生成相对稳定的脂肪醇,所以 C4-12 的脂肪醇在副产物中的占比较高,生成 C4-12 脂肪醇可能发生的反应如下所示:



其中,式①为反应生成的副产物中的多碳烯烃裂解成含碳数更小的戊烯和己烯,式②和式③则为戊烯和己烯分别与水生成相应的脂肪醇。

3.2 问题二: 反应条件对乙醇转化率及 C4 烯烃选择性影响大小评估模型

3.2.1 问题二分析

通过上述对问题一的分析与解答可知,不同催化剂组合与温度均会对乙醇转化率和 C4 烯烃选择性的大小产生影响。并且在题中所给的实验数据中除 Co 负载量、Co/SiO₂ 和 HAP 装料比、乙醇通入速率外,装料方式 I 和装料方式 II 仍可能会对催化效果产生不同程度的影响,因此同样需考虑该因素。

与问题一的定性分析不同,针对问题二我们首先对实验数据进行了预处理,将 Co 负载量、Co/SiO₂ 和 HAP 装料比等参数处理为具体的 SiO₂、HAP 质量,紧接着对这些数据进行了标准化。然后分别采用了熵权法评价模型和支持向量机回归模型(SVR)以及 LIME 局部可解释模型对问题中的不同反应条件所产生的乙醇转化率、C4 烯烃选择性影响大小进行了定量分析。

3.2.2 数据预处理

(1) 指标映射

首先由于装料比这个特征包含两个信息,即 Co/SiO₂ 混合物与 HAP 各自的质量,以及两者质量大小的相对比值,仅考虑比值的大小并不能充分的利用数据,且数据中石英砂的本质仍为 SiO₂。因此,我们通过对 Co 负载量、Co/SiO₂ 和 HAP 装料比进行计算,得到了 SiO₂ 质量和 HAP 质量,以此作为新的影响催化反应的指标,具体计算方法为:

$$\text{Co 质量} = \left(\frac{\text{Co}}{\text{SiO}_2 \text{混合物的质量}} \right) * \text{Co 负载量} / (100 + \text{Co 负载量})$$

$$\text{SiO}_2 \text{质量} = \frac{\text{Co}}{\text{SiO}_2 \text{混合物质量}} - \text{Co 质量} + \text{石英砂质量}$$

(2) 标准化处理

为了避免某个特征过大或过小在训练模型中可能引起的作用不平衡,以及在核计算中所用到的内积计算或指数运算因数据不平衡而引起的计算困难,我们把这六个特征维度的数据进行了标准差标准化处理,其转化公式为:

$$x^* = \frac{x - \mu}{\sigma}$$

其中 μ 为所有样本数据的均值, σ 为所有样本数据的标准差。

3.2.3 熵权法评价模型的建立与求解

熵权法是根据指标变异性的的大小来确定客观权重[2],在计算过程中,指标的信息熵与该指标值的变异程度呈反比,当指标值的信息熵越小时,变异程度越大,提供的信息越多,在综合评价中所起的作用越大,反之亦然。

在本问中我们首先将不同催化剂组合的具体指标进行了标准化处理,并利用熵信息输出求取客观权重的方法求出每种指标的权重,然后根据权重大小判断不同反应指标对乙醇转化率和 C4 烯烃选择性影响的大小。

(1) 信息熵输出求取客观权重

对标准化决策矩阵 $Z = (z_{ij})_{n \times m}$, 令

$$p_{ij} = \frac{z_{ij}}{\sum_{i=1}^n z_{ij}}, i = 1, 2, \dots, n; j = 1, 2, \dots, m$$

根据相关信息论知识,可求得不同反应条件时的指标 G_i 的信息熵为

$$E_j = -(\ln n)^{-1} \sum_{i=1}^n p_{ij} \ln p_{ij}, j = 1, 2, \dots, m$$

其中人为规定当 $p_{ij} = 0$ 时, $p_{ij} \ln p_{ij} = 0$, 此时不同反应条件指标 G_i 的客观权重为

$$\mu_i = \frac{1 - E_j}{\sum_{k=1}^m (1 - E_k)}, j = 1, 2, \dots, m$$

将标准化的数据代入其中可得到以下结果:

表 3-2 不同反应条件指标熵权法所得权重表

反应条件指标	Co 负载量	SiO ₂ 质量	HAP 质量	乙醇浓度	温度	装料方式
--------	--------	---------------------	--------	------	----	------

权重	0.18895736	0.11342567	0.11081420	0.05419457	0.39312583	0.13948237
----	------------	------------	------------	------------	------------	------------

将所得权重大小进行排序：

温度>Co 负载量>装料方式>SiO₂质量>HAP 质量>乙醇浓度

由此可以看出采用熵权法即站在客观数据的角度分析，Co 负载量对反映结果的影响最大，其次是温度、SiO₂ 质量和 HAP 质量，乙醇浓度的影响最小。

3.2.5 支持向量机回归（SVR）模型的建立与分析

（1）支持向量机回归模型的建立

支持向量机(Support Vector Machines, SVM) 是一种快速可靠的线性分类器，在有限的数量下性能非常好。给定训练数据（监督学习），SVM 算法得到一个最优超平面，从而对训练数据进行分类。当支持向量机应用到回归预测等领域则称为支持向量回归，其基本原理如下：

给定训练样本 $x_i \in \mathbb{R}^p, i = 1, \dots, n, y \in \mathbb{R}^n$ ，引入 ε 不敏感函数作为损失函数：

$$L_{\varepsilon}(f(x_i) - y_i) = \begin{cases} 0, & |f(x_i) - y_i| < \varepsilon \\ |f(x_i) - y_i| - \varepsilon, & |f(x_i) - y_i| \geq \varepsilon \end{cases}$$

其中 ε 是一正数， ε 不敏感损失函数指：当观察值 y 与预测值 $f(x)$ 之差不超过事先给定的 ε 时，则认为在该点的预测值是无损失的。而回归问题就是对于训练样本集，能够找到一对参数 (w, b) , $f(x) = w \cdot \phi(x_i) + b$ ，和实际获得的目标之间有尽量小的偏差。

因此需要引入松弛因子 ξ, ξ^* ，根据统计学习理论的结构风险化准则，线性回归可转化为求解如下凸规划问题：

$$\begin{aligned} \min R(w, \xi_i, \xi_i^*) &= \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \\ \text{subject to } &\begin{cases} y_i - w \cdot x_i - b \leq \varepsilon + \xi_i \\ w \cdot x_i + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned}$$

然后引入拉格朗日函数，同时通过相应的鞍点条件简化得到：

$$\min \left\{ \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i \cdot x_j) + \sum_{i=1}^l \alpha_i(\varepsilon - y_i) + \sum_{i=1}^l \alpha_i^*(\varepsilon + y_i) \right\}$$

$$\text{subject to} \begin{cases} \sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases}$$

同时由于此次模型我们引入了多项式核函数 poly, $\phi(x) = (\gamma \langle x, x' \rangle + r)^d$

对上述函数方程求解可以得到w:

$$w = \sum_{i=1}^l (\alpha_i - \alpha_i^*) \cdot \phi(x)$$

由此可求得线性回归函数:

$$f(x) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) \phi(x_i)^T \phi(x_i) + b$$

其中

$$b = y_j - \sum_{i=1}^l (\alpha_i^* - \alpha_i) \phi(x_i)^T \phi(x_i) + \varepsilon, \quad j \in \{m | 0 < \alpha_m < C\}$$

再分别将预处理后的反应条件指标与乙醇转化率和 C4 烯烃选择性以及 C4 烯烃收率数据分别代入上述模型便可得到支持向量机的预测函数模型, 具体预测结果见支撑材料。

(2) 支持向量机回归模型的结果分析

对于乙醇转化率的拟合曲线如下所示:

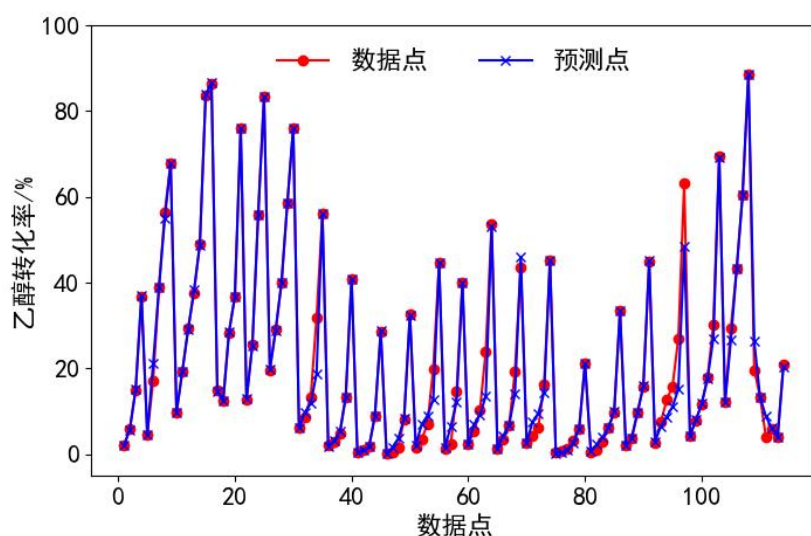
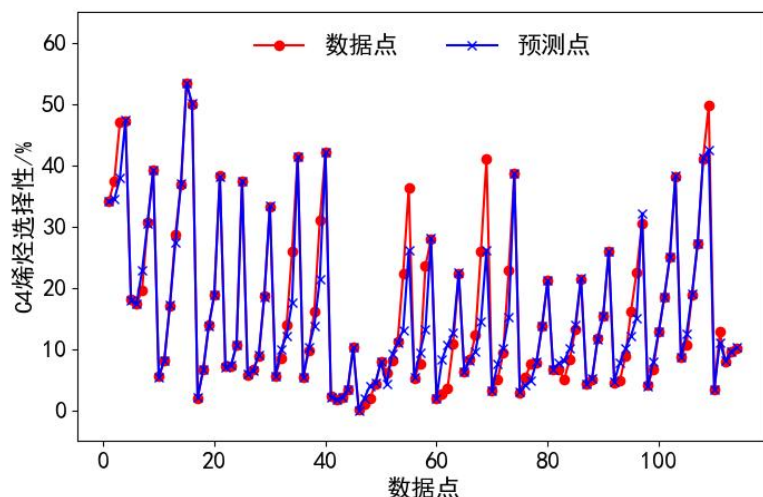


图 3-8 乙醇转化率拟合曲线

SVR 模型拟合之后的曲线 R^2 达到了 0.984，从图 3-8 也可以看出 SVR 模型对乙醇转化率的拟合效果较好，原始实验数据点与预测点重叠性较高，具有较高的可信度。

对于 C4 烯烃选择性的拟合曲线如下所示：



SVR 模型拟合之后的曲线 R^2 达到了 0.939，从图 3-8 也可以看出 SVR 模型对 C4 烯烃选择性的拟合效果虽然比乙醇转化率的拟合效果差一些，但仍可认为原始实验数据点与预测点重叠性较高，具有较高的可信度。

3.2.6 LIME 局部可解释模型的建立与分析

(1) LIME 模型的建立

LIME 是一种能够利用简单的模型（这种简单的模型包括但不限于线性模型、决策树等）局部近似目标黑盒模型的预测，但不深入其模型内部，主要通过对输入进行轻微的扰动，探测黑盒模型的输出发生的变化，最终根据这种变化在原始的输入点训练一个可解释性的模型。

LIME 的数学表达式如下所示：

$$\text{explanation}(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

其中 f 表示原始的模型，即需要解释的模型， g 表示简单模型， G 是简单模型的一个集合，如所有可能的线性模型， π_x 表示我们新数据集中的数据 x' 与原始数据的距离， $\Omega(g)$ 表示模型 g 的复杂程度。然后利用下式来衡量不同模型预测值之间的差：

$$L(f, g, w^y) = \sum_{i=1}^N w^y(z_i) (f(z_i) - g(z_i'))^2$$

最后我们针对想要解释的变量 x ，对数据集中的数据进行扰动得到新的数据，同时计算出支持向量机回归模型对这些新的数据的预测值。然后对这些新的取样点求出权重，即这些数据点与我们要解释的数据之间的距离。最后通过对所有数据点所取得的不同反应条件下的权重向量计算平均值便可得到对于拟合曲线而言不同反应条件指标对乙醇转化率和 C4 烯烃选择性的影响。

(2) LIME 模型的结果

在支持向量机回归模型中，我们分别对乙醇转化率 y_1 和 C4 烯烃选择性 y_2 进行了拟合计算，因此代入数据利用 LIME 模型对其进行计算后可以得到如下权重表：

表 3-3 不同反应条件 LIME 模型所得对乙醇转化率 y_1 权值表

反应条件指标	Co 负载量	SiO ₂ 质量	HAP 质量	乙醇浓度	温度	装料方式
平均权值	15.73010590	1.50143713	5.88115566	3.21012279	16.35613348	3.27020992

从上表可以看出对于乙醇转化率 y_1 而言，各个反应条件指标的平均权值由大到小排序为：

温度 > Co 负载量 > HAP 质量 > 装料方式 > 乙醇浓度 > SiO₂ 质量

表 3-4 不同反应条件指标 LIME 模型对 C4 烯烃选择性 y_2 权值表

反应条件指标	Co 负载量	SiO ₂ 质量	HAP 质量	乙醇浓度	温度	装料方式
平均权值	9.28970258	3.56058693	2.07270147	1.76974852	10.11063256	7.22876791

从上表可以看出对于 C4 烯烃选择性 y_2 而言，各个反应条件指标的平均权值由大到小排序为：

温度>Co 负载量>装料方式>SiO₂质量> HAP 质量>乙醇浓度

综上所述，利用 LIME 解释器可知不同的催化剂组合和温度对乙醇转化率、C4 烯烃选择性有着不同的影响，具体分析如下：

温度对乙醇转化率的影响最大，依次是 Co 负载量、HAP 质量、装料方式、乙醇浓度，SiO₂ 质量对乙醇转化率的影响最小；温度仍然对 C4 烯烃选择性的影响最大，依次是 Co 负载量、装料方式、SiO₂ 质量、HAP 质量，乙醇浓度对 C4 烯烃选择性的影响最小。

3.2.7 综合分析结果

在本题中，我们首先对装料方式是否会影响分别采用了熵权法和 LIME 模型对包括 HAP 质量、温度、乙醇浓度等变量在内对乙醇浓度和 C4 烯烃选择性的影响大小进行了分析。但由于熵权法仅仅是站在客观数据的角度，从数据分布离散性来分析不同反应条件的影响大小，并未考虑该条件与反映结果乙醇浓度和 C4 烯烃选择性的关系，且默认了乙醇浓度与 C4 烯烃选择性同自变量关系相同，因此所得到的结论是不够全面的。法二采用了 SVM 回归的方法对反应过程进行了拟合，拟合效果较好，然后采用 LIME 模型对其反应条件与反应结果之间的关系进行了权值量化计算，得到的结论可信度较强。即乙醇转化率和 C4 烯烃选择性的在很大程度上取决于反应温度的大小，其次才是催化剂的不同组合。

3.3 问题三： C4 烯烃收率最优解模型

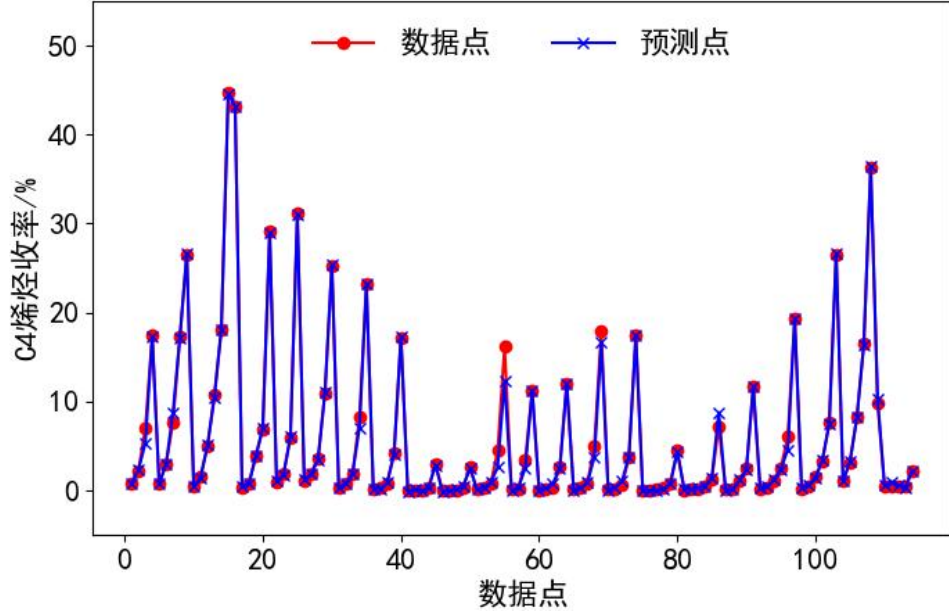
3.3.1 问题三分析

从前两问的结果可知在本实验过程中，不同催化剂组合与反应温度对乙醇转化率以及 C4 烯烃选择性大小有着不同的影响。因此选择合适的催化剂组合与反应温度对催化反应结果的优劣具有很大的意义。

针对本题，我们结合了问题二中所建立的支持向量机回归模型，通过重新对催化剂组合与温度同 C4 烯烃收率进行拟合，然后采用多目标优化的方法进行最优解的计算，找到了在 C4 烯烃收率最大的情况下催化剂的组合与反应温度，同时结合实验已知数据进行了验证，效果良好。

3.3.2 C4 烯烃收率回归模型的建立

与问题二中采用支持向量机回归方法相同，但训练的数据集将反应结果乙醇转化率与 C4 烯烃选择性进行了相乘，得到了代表反应结果的新指标——C4 烯烃收率，然后对其进行拟合，得到的拟合结果如下图所示：



其中 $R^2 = 0.9964196356232593$ ，拟合曲线与实际数据重合度较高，因此可以说明该模型的拟合结果具有一定的实际意义。

3.3.3 多目标优化模型的求解与分析

(1) 决策变量

我们选取的六个决策变量分别为，乙醇浓度 x_1 、温度 x_2 、Co 的负载量 x_3 、 SiO_2 质量 x_4 、HAP 质量 x_5 、装料方式 x_6 。

(2) 目标函数

乙醇转化率为单位时间内乙醇的单程转化率，其值为 $100\% \times (\text{乙醇进气量} - \text{乙醇剩余量}) / \text{乙醇进气量}$ ，C4 烯烃选择性为 C4 烯烃在所有产物中的占比。在问题二中，由支持向量机回归模型的拟合结果我们可以知道乙醇转化率 y_1 、C4 烯烃选择性 y_2 的计算方法为：

$$\begin{cases} y_1 = \sum_{i=1}^l (\alpha_{i1} - \alpha_{i1}^*)(x_1, x_{i1}) + b_1 \\ y_2 = \sum_{i=1}^l (\alpha_{i2} - \alpha_{i2}^*)(x_2, x_{i2}) + b_2 \end{cases}$$

以 C4 烯烃收率最大为优化目标，得到目标函数为：

$$\max y = y_1 * y_2 = \left(\sum_{i=1}^l (\alpha_{i1} - \alpha_{i1}^*)(x_1, x_{i1}) + b_1 \right) \left(\sum_{i=1}^l (\alpha_{i2} - \alpha_{i2}^*)(x_2, x_{i2}) + b_2 \right)$$

其中 α_{i1} 、 α_{i1}^* 、 α_{i2} 、 α_{i2}^* 为支持向量机回归模型根据五个决策变量计算得到

的参数。

(3) 约束条件

①反应温度 x_2 : 由题目中第一小问要求可知温度本身并无限制, 为了保证支持向量机回归模型预测的效果, 温度需在实际实验的范围内, 即

$$250^{\circ}\text{C} < x_2 < 450^{\circ}\text{C}$$

②反应温度 x_2 : 由题目中第二小问可知温度最大值需小于 350°C , 为了保证支持向量机回归模型预测的效果, 温度需在实际实验的范围内, 即

$$250^{\circ}\text{C} < x_2 < 350^{\circ}\text{C}$$

③其余约束条件: 所有催化剂组合需在已有实验的组合之内。

(4) 多目标优化模型的求解

结合实验已有数据, 使用遍历算法对于温度指标而言步长为 5°C 、对于其余催化剂组合的指标而言以提供的数据组合作为遍历的选择, 由此对 C4 烯烃收率进行遍历求解。在温度要求不同的条件下, 分别得到 C4 烯烃收率最大之时催化剂组合的类别与反应温度的大小, 其最优结果如下所示。

(1) 250°C – 350°C 温度拟合区间下
C4 烯烃收率的最优结果为: 54.144%
其实验条件为: 200mg 2wt%Co/SiO₂– 200mg HAP–乙醇浓度 1.68ml/min, 350°C , 装料方式 I

(2) 250°C – 450°C 温度拟合区间下
C4 烯烃收率的最优结果为: 54.264%
其实验条件为: 200mg 2wt%Co/SiO₂– 200mg HAP–乙醇浓度 1.68ml/min, 450°C (模拟条件下), 装料方式 I

(5) 可靠性检验

根据附件一中所提供的实验数据, 我们对不同催化剂组合以及温度指标进行了重新排序, 处理得到了在不同反应条件下 C4 烯烃收率的结果。然后将其进行降序排列, 得到了收率最高的几种可能情况, 如下图所示, 在 350°C 下, 可以看到我们位于第一名的实验条件: 200mg 2wt%Co/SiO₂– 200mg HAP–乙醇浓度 1.68ml/min– 350°C –装料方式 I, 与我们在 250°C – 350°C 温度拟合区间下预测的 C4 烯烃收率的结果的实验条件完美相符, 因此预测结果可靠性较强。

表 3-1. 真实数据中不同条件下 350 摄氏度以下 C4 烯烃收率降序排列前 5 个值

实验条件	C4 烯烃收率/%
200mg 2wt%Co/SiO ₂ – 200mg HAP–乙醇浓度 1.68ml/min, 350°C , 装料方式 I	26.54080464

200mg 1wt%Co/SiO ₂ - 200mg HAP-乙醇浓度 0.9ml/min, 350℃, 装料方式 I	18.03330095
200mg 1wt%Co/SiO ₂ - 200mg HAP-乙醇浓度 1.68ml/min, 350℃, 装料方式 I	17.37376011
200mg 0.5wt%Co/SiO ₂ - 200mg HAP-乙醇浓度 1.68ml/min, 350℃, 装料方式 I	16.47764673
50mg 1wt%Co/SiO ₂ - 50mg HAP-乙醇浓度 0.3ml/min, 350℃, 装料方式 I	10.9239714

3.4 问题四：佐证性实验的设计

化学领域是一门实验性科学领域, 相比于其他领域, 其可预测性偏低, 因而在创造性评判中往往需要实验数据进行辅助说明。[\[1\]赵凯, 韩强, 游学为, 等. 浅谈化学领域创造性答复中如何有效地补充实验数据\[J\]. 中国发明与专利, 2019, 016\(02\):33-36.](#) 因此在问题三中所预测到的满足 C4 烯烃收率尽可能大时催化剂组合与反应温度的条件存在一定的误差, 因此需要通过进一步的实验数据来对其进行佐证。

3.4.1 基于温度与催化剂组合的实验设计思路

从问题三中可以看出, 温度对 C4 烯烃收率有着较大的影响, 且在已有实验条件组合的情况下所预测的 C4 烯烃收率达到最大值时的反应条件为: 200mg 2wt%Co/SiO₂- 200mg HAP-乙醇浓度 1.68ml/min, 450℃ (模拟条件下), 装料方式 I, 但从下图可以看出温度在 400℃ 以后对催化效果的增强并不显著, 因此

为了探究此时温度的影响可以如**表 1** 设计方案一中的五组实验, 来探究温度在最优解附近的影响大小, 如果 C4 烯烃收率随温度的增长趋势不大, 则可以认为在 400℃ 左右温度对 C4 烯烃收率产生的影响达到了临界值, 即使继续提升温度也无法对 C4 烯烃的收率提高起到一定的作用。

同样的, 对于不同催化剂组合而言, 在第三问的最优解条件下也存在着缺少实验数据无法验证预测结果准确性的问题, 因此在温度为 450℃ 的条件下, 我们选取了 400℃ 时催化效果最好的五种催化剂组合进行实验, 探究催化剂组合在最优解附近的影响大小, 如果实验得到的催化剂组合为 **balabala**, 则验证了预测

C4 烯烃收率最大时的反应条件正确性。

3.4.2 具体实验设计方案

Balabalabala

4. 模型评价和改进

4.1 新冠疫情等级评估模型

在该模型中，我们采用了单指标风险评价法，系统聚类法以及层次分析法。以上方法的分类方法都具有一定的合理性，但是这些方法有各自的优缺点。

对于单指标风险评价法，优势是计算简单、容易操作但此方法主观性太强且没有考虑社会经济因素的影响，没有综合考虑影响疫情发展情况的各个因素。系统聚类方法将风险向量视为 4 维空间的一个点，然后基于各个地区特征的相似性将距离临近的点聚为一类，方法简单，易于操作，但该方法同等对待各个指标的重要性，结果的可解释性较弱。层次分析法综合多个指标的信息，可以对各个地区的风险进行量化评分，因此也可以对各个地区风险进行排序，便于领导层决策管理；其缺点是其判断矩阵较难获得，需要查询相关文献或请有关专家进行评判。

综上所述，针对问题一，层次分析法效果良好，能够较好的满足各城市地区疫情情况分类的要求。

4.2 Spearman 相关性检验

常用的 Pearson 相关检验适用于两变量均符合正态分布时的情况，因此在第二问中的疫情发展情况与城市数据指标的相关性探究中该方法并不适用。我们采用了另外一种相关性分析方法 Spearman 检验。该方法能够利用单调方程评价两个统计变量的相关性，且适用范围较广，能够较好的满足探城市疫情发展情况与已有数据的相关性检验。针对其他无法用数据衡量的城市指标，我们通过查询相关文献和资料也对其进行了一一分析和整理，并给出了相关性结论，较好的完成了任务二的要求。

4.3 温州疫情预测模型

4.3.1 ARIMA 时间序列分析模型

ARIMA 时间序列分析模型建立过程较为容易，只需要内生变量而不需要借助其他外生变量，在此处只需要借助原有的总确诊人数就可以进行时间序列分析预测但是要求时序数据是稳定的（stationary），或者是通过差分化(differencing)后是稳定的，并且本质上只能捕捉线性关系，而不能捕捉非线性关系，因为他只会捕捉到动态的信息，对于其它的一些客观信息譬如 GDP，人口密度，政策等信息是无法捕捉到的，因此会比较局限。

4.3.2 SEIR 模型

整体而言是非常适合用来模拟传染病类型的预测和分析，此处是新冠的预测和分析，并且综合考量了城市的政策情况，感染者接触情况，传染情况，疾病的致死率发病率，恢复率等等因素，根据之前的疫情数据，进行数值模拟来确定一系列相关的参数。

但 SEIR 只是传染病传播的基本模型，还有很多没有考虑的因素，如人口的出生与死亡，迁入和迁出等。还有可以考虑的更细致的因素，如人群流动速度、易感人群的年龄分布、不同人群对疾病的易感性，患病者的症状轻重，人口密度，医疗卫生程度，检验检疫手段，政府重视程度，人群心理因素等。这些因素都对暴露数、发病率、治愈率、传染期长度有着直接或间接的影响。但是 SEIR 整体而言算是一个比较优秀的模型了

SEIR 和 ARIMA 比较而言，这里更适合用 SEIR，因为 ARIMA 时间序列分析在这里很难只根据总确诊人数提取出有效信息，譬如发病率，致死率，人口密度，总人口数，经济状况，病人接触情况，因此很难预测出准确和有效的模型，但是 SEIR 模型这里的一系列参数，即潜伏期，传染率，病人接触情况，恢复率等可以根据官方的数据进行数值模拟，进行很好的估算，同时如果有政策改动，譬如临时封城等，SEIR 模型里面也会有较为直观的体现。因此这里更适合于用 SEIR 传染病动力学模型进行预测和评估。

4.4 模型扩展方向

(1) 疫情各个城市的分类模型在本文中主要利用了四项城市指标，但也可以通过进一步考虑城市政府的政策影响力等来优化改分类评估模型。

(2) 对于 ARIMA 模型可以考虑将一系列环境影响因素标量化，譬如政策改进 0-10 为封管程度，病人接触情况，传播情况等等也可以标量化，这样结果就不会那么单一仅仅根据总确诊人数来确定预测疫情的未来发展趋势；对于 SIER 模型可以考虑构建再多加一些权重因素，譬如人群流动速度、易感人群的年龄分布、不同人群对疾病的易感性，患病者的症状轻重，等等，如果再细致一点将这些特征引入，SIER 模型应该会表现更好。

5. 参考文献

- [1]贾闻达. Co/HAP 催化剂催化天然油脂及脂肪酸选择性加氢制备脂肪醇[D].中国科学技术大学,2018.
- [2]李翠云,叶劲松,万敏惠,程琴,何伟明,师洁. 基于熵权法优化 TGase 制备高乳化性蛋黄粉的反应条件[J]. 基因组学与应用生物学, 2020, 39(12):5526-5532.
- [3]GBD 2016 Healthcare Access and Quality Collaborators. Measuring performance on the Healthcare Access and Quality Index for 195 countries and territories and selected subnational locations: asystematic analysis from the Global Burden of Disease Study 2016. Lancet. 2018 Jun 2.
- [4]中华人民共和国中央人民政府, 抗击新冠肺炎疫情的中国行动
http://www.gov.cn/zhengce/2020-06/07/content_5517737.htm, 2021. 6. 5
- [5]特布鲁, 世界主要国家的教育水平,
<https://zhuanlan.zhihu.com/p/28398447>, 2021. 6. 5
- [6] 张斌,气温升高了, 新冠病毒会被遏制吗? 看看专家怎么说 | 特别关注,
https://www.sohu.com/a/391877717_100012122, 2021. 6. 5
- [7] 杨真真, 谢艳秋, 靳旭东, 庄桂敏. 基于 ARIMA 时间序列模型的传染病发展趋势预测——以 COVID-19 为例[J]. 中国科技信息, 2021 (Z1):70-72.
- [8] 曹盛力, 冯沛华, 时朋朋. 修正 SEIR 传染病动力学模型应用于湖北省 2019 冠状病毒病 (COVID-19) 疫情预测和评估[J]. 浙江大学学报(医学版), 2020, 49(02):178-184.

6. 附录

6.1 新冠疫情评分及等级（层次分析法）

地区	风险评分	评价结果
Arizona-Maricopa	-0.0142558	中等
Arizona-Pima	-0.242526644	轻微
Arizona-Pinal	-0.198235698	轻微
Arkansas-Jefferson	-0.187016326	轻微
California-Alameda	-0.053271894	中等
California-Contra Costa	-0.074058689	中等
California-Fresno	-0.172199376	轻微
California-Humboldt	-0.224110991	轻微
California-Imperial	-0.208455982	轻微
California-Los Angeles	0.589080144	严重
California-Madera	-0.185375844	轻微
California-Marin	-0.184300713	轻微
California-Orange	0.019833442	中等
California-Placer	-0.202237328	轻微
California-Riverside	-0.118324924	中等
California-Sacramento	-0.096701161	中等
California-San Benito	-0.198275999	轻微
California-San Diego	0.001413711	中等
California-San Francisco	-0.032812628	中等
California-San Luis Obispo	-0.329914664	轻微
California-San Mateo	-0.190295703	轻微
California-Santa Clara	-0.102864132	中等
California-Santa Cruz	-0.176672284	轻微
California-Shasta	-0.186657913	轻微
California-Solano	-0.180294378	轻微
California-Sonoma	-0.182148882	轻微
California-Ventura	-0.168281614	轻微
California-Yolo	-0.202915047	轻微
Chongqing	-0.088820436	中等
Colorado-Arapahoe	-0.155201107	轻微
Colorado-Denver	0.035721556	中等
Colorado-Douglas	-0.168777085	轻微
Colorado-Eagle	-0.132091129	中等
Colorado-El Paso	-0.126990576	中等
Colorado-Gunnison	-0.172023125	轻微
Colorado-Jefferson	-0.107374475	中等
Colorado-Larimer	-0.204081105	轻微
Colorado-Summit	-0.183564495	轻微
Connecticut-Fairfield	0.264577734	严重
Connecticut-Litchfield	-0.16367813	轻微
Florida-Alachua	-0.137251499	中等
Florida-Broward	0.214437608	严重
Florida-Charlotte	-0.177297009	轻微
Florida-Clay	-0.193222341	轻微
Florida-Collier	-0.161437908	轻微
Florida-Duval	-0.099053969	中等
Florida-Escambia	-0.177725958	轻微
Florida-Hillsborough	-0.061904872	中等
Florida-Lake	-0.17400545	轻微
Florida-Lee	-0.098530218	中等
Florida-Manatee	-0.165997492	轻微
Florida-Nassau	-0.18312866	轻微
Florida-Okaloosa	-0.158866673	轻微
Florida-Pasco	-0.163845456	轻微
Florida-Pinellas	-0.084722761	中等
Florida-Santa Rosa	-0.15702843	轻微
Florida-Sarasota	-0.176454891	轻微
Florida-Seminole	-0.142476874	中等
Florida-Sumter	-0.163940763	轻微
Florida-Volusia	-0.280327849	轻微
Georgia-Bartow	-0.138234552	中等
Georgia-Charlton	-0.193828147	轻微
Georgia-Cherokee	-0.12875632	中等
Georgia-Cobb	-0.066605724	中等
Georgia-Coweta	-0.207588177	轻微
Georgia-DeKalb	-0.039924827	中等
Georgia-Fayette	-0.158605726	轻微
Georgia-Floyd	-0.172744381	轻微
Georgia-Fulton	-0.034184061	中等
Georgia-Gordon	-0.172801167	轻微
Georgia-Gwinnett	-0.122873759	中等
Georgia-Lee	-0.156078353	轻微
Georgia-Lowndes	-0.229189211	轻微
Georgia-Polk	-0.174997543	轻微
Illinois-Cook	1.061550792	严重
Illinois-Kane	-0.143449204	中等
Illinois-McHenry	-0.195903811	轻微
Indiana-Adams	-0.213163632	轻微
Indiana-Boone	-0.196491075	轻微

Indiana-Hendricks	-0.173552578	轻微
Indiana-Howard	-0.199424606	轻微
Indiana-Johnson	-0.174659542	轻微
Indiana-Marion	0.089374799	中等
Indiana-Noble	-0.232680152	轻微
Iowa-Johnson	-0.168634662	轻微
Iowa-Pottawattamie	-0.198801992	轻微
Kansas-Johnson	-0.142423523	中等
Kentucky-Harrison	-0.250209287	轻微
Kentucky-Jefferson	-0.147742118	中等
Louisiana-Ascension	-0.139622708	中等
Louisiana-Caddo	-0.151470814	轻微
Louisiana-Jefferson	0.08012936	中等
Louisiana-Orleans	0.296409609	严重
Maryland-Baltimore-city	-0.142843275	中等
Maryland-Harford	-0.206794931	轻微
Maryland-Montgomery	-0.08062983	中等
Maryland-Prince George's	-0.084873629	中等
Massachusetts-Berkshire	-0.140821236	中等
Massachusetts-Essex	0.022049946	中等
Massachusetts-Middlesex	0.147870593	严重
Massachusetts-Norfolk	0.005793176	中等
Massachusetts-Suffolk	0.229027305	严重
Massachusetts-Worcester	-0.077777582	中等
Michigan-Oakland	0.219563732	严重
Michigan-Wayne	0.802633121	严重
Minnesota-Anoka	-0.167683464	轻微
Minnesota-Ramsey	-0.139422861	中等
Missouri-St. Louis	-0.043160994	中等
Nebraska-Douglas	-0.135027827	中等
Nebraska-Knox	-0.171763865	轻微
Nevada-Clark	-0.02094263	中等
Nevada-Washoe	-0.536662411	轻微
New Hampshire-Grafton	-0.195029579	轻微
New Hampshire-Rockingham	-0.150213048	轻微
New Jersey-Bergen	0.627994247	严重
New Jersey-Burlington	-0.120473632	中等
New Jersey-Camden	-0.128880159	中等
New Jersey-Hudson	0.385073908	严重
New Jersey-Monmouth	0.116987622	中等
New Jersey-Passaic	0.176417835	严重
New Jersey-Union	0.173170129	严重
New Mexico-Bernalillo	-0.141890437	中等
New Mexico-Santa Fe	-0.200388415	轻微
New Mexico-Socorro	-0.223064629	轻微
New York-Nassau	2.124899509	严重
New York-New York	12.55583512	严重
New York-Rockland	0.62139176	严重
New York-Saratoga	-0.131892424	中等
New York-Suffolk	1.594300932	严重
New York-Ulster	-0.109712581	中等
New York-Westchester	2.815403565	严重
North Carolina-Cabarrus	-0.202992673	轻微
North Carolina-Chatham	-0.156871712	轻微
North Carolina-Durham	-0.129743784	中等
North Carolina-Forsyth	-0.202449223	轻微
North Carolina-Johnston	-0.187949268	轻微
North Carolina-Mecklenburg	-0.046087027	中等
North Carolina-Onslow	-0.172563785	轻微
North Carolina-Wake	-0.165071568	轻微
Ohio-Cuyahoga	-0.050064294	中等
Ohio-Stark	-0.152020338	轻微
Oklahoma-Tulsa	-0.156603327	轻微
Oregon-Deschutes	-0.209275296	轻微
Oregon-Douglas	-0.169297699	轻微
Oregon-Jackson	-0.172973579	轻微
Oregon-Klamath	-0.186916672	轻微
Oregon-Marion	-0.155849674	轻微
Oregon-Multnomah	-0.156571305	轻微
Oregon-Polk	-0.193399881	轻微
Oregon-Umatilla	-0.224864519	轻微
Oregon-Washington	-0.124924508	中等
Pennsylvania-Bucks	-0.140966516	中等
Pennsylvania-Delaware	-0.059057094	中等
Pennsylvania-Monroe	-0.140194339	中等
Pennsylvania-Montgomery	-0.041966158	中等
Pennsylvania-Philadelphia	0.315962241	严重
Pennsylvania-Wayne	-0.243144995	轻微
Rhode Island-Providence	-0.092142748	中等
Shanghai	0.042051915	中等
South Carolina-Charleston	-0.182203903	轻微
South Carolina-Kershaw	-0.154257568	轻微
South Carolina-Lancaster	-0.192162587	轻微
South Carolina-Spartanburg	-0.167611942	轻微
South Dakota-Beadle	-0.175556675	轻微
South Dakota-Charles Mix	-0.188961852	轻微
South Dakota-Davison	-0.294327065	轻微
South Dakota-Minnehaha	-0.19266919	轻微
South Dakota-Pennington	-0.206573693	轻微
Tennessee-Davidson	-0.022705442	中等
Tennessee-Shelby	-0.054575196	中等
Tennessee-Sullivan	-0.167528089	轻微
Tennessee-Williamson	-0.133471353	中等
Texas-Collin	-0.215771713	轻微
Texas-Dallas	0.036129332	中等
Texas-Fort Bend	-0.121544414	中等
Texas-Gregg	-0.180550439	轻微
Texas-Harris	0.026912532	中等
Texas-Montgomery	-0.17521687	轻微
Texas-Tarrant	-0.13494275	中等
Utah-Davis	-0.134416677	中等
Utah-Weber	-0.170603617	轻微
Vermont-Bennington	-0.22745104	轻微
Virginia-Arlington	-0.006958029	中等
Virginia-Fairfax	-0.102710514	中等
Virginia-Henrico	-0.15154166	轻微
Virginia-Loudoun	-0.129791603	中等
Washington-Clark	-0.134803817	中等
Washington-Grant	-0.157546541	轻微
Washington-Island	-0.121856456	中等
Washington-Jefferson	-0.163462328	轻微
Washington-King	0.500824587	严重
Washington-Kitsap	-0.143110597	中等
Washington-Kittitas	-0.1644894	轻微
Washington-Pierce	-0.104037217	中等
Washington-Skagit	-0.268020029	轻微
Washington-Snohomish	0.149397799	严重
Washington-Whatcom	-0.131394433	中等
Wenzhou	-0.140217937	中等
Wisconsin-Dane	-0.156544213	轻微
Wisconsin-Pierce	-0.192015487	轻微

6.2 SEIR 模型 Python 代码

```
T = 50 #天数
N = 9300000 #总人数
E = [0]*T #潜伏者人数
I = [1]*T #感染者
S = [N-I[0]]*T #易感人数
R = [0]*T #康复人数
r = 12 #感染者接触易感者的人数
B = 0.053 #传染概率
r2 = 12 #潜伏者接触易感者的人数
B2 = 0.1 #潜伏者传染正常人的概率
a = 0.3 #潜伏者转化为感染者概率
y = 0.6 #康复概率
for i in range(T-1):
    if i+1>=10: #从第 9 天开始，有关部门采取措施，平均接触人数和平均感染率显著下降一直到 3 月 5 号左
        右接触封禁，即到 46 天解除封禁，
            r=3
            r2=3
            B = 0.001
            B2 = 0.001
    if i+1>=46:
        r= 0
        B= 0
        r2 = 0.000001
        B2 = 0.000001
        a= a* 0.3/2.0
        y = 0.6
    S[i+1] = S[i] - r*B*S[i]*I[i]/N - r2*B2*S[i]*E[i]/N
    E[i+1] = E[i] + r*B*S[i]*I[i]/N - a*E[i] + r2*B2*S[i]*E[i]/N
    I[i+1] = I[i] + a*E[i]-y*I[i]
    R[i+1] = R[i] + y*I[i]
import matplotlib.pyplot as plt
plt.xlabel('Date')
plt.ylabel('Number of latent patients')
#plt.plot(S)
plt.plot(E)
#plt.plot(I)
#plt.text(,0,' \leftarrow 正弦');
#plt.text(20,400,'number of E');
#plt.plot(R)
plt.show()
```


6.3 ARIMA 模型 Python 代码

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from datetime import datetime
from pandas import Series, DataFrame
import warnings
warnings.filterwarnings('ignore', 'statsmodels.tsa.arima_model.ARMA',
                        FutureWarning)
warnings.filterwarnings('ignore', 'statsmodels.tsa.arima_model.ARIMA',
                        FutureWarning)

data = pd.read_excel('testdf.xlsx', index_col= u'日期')
data.index = pd.DatetimeIndex(data.index).to_period()
data1 = pd.read_excel('testdf1.xlsx', index_col = u'日期')
forecastnum = 5
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False
data.plot()
data1.plot()
plt.show()
#以上是画时序图

from statsmodels.graphics.tsaplots import plot_acf
plot_acf(data).show()
#以上是自相关图

from statsmodels.tsa.stattools import adfuller as ADF
#print(u'原始序列的 ADF 检验结果为: ', ADF(data[u'确诊总人数']))
#原始序列的 ADF 检验结果为:  (-3.7667177304726867, 0.003266793578451666, 9, 19, {'1%': -3.8326031418574136,
'5%': -3.0312271701414204, '10%': -2.655519584487535}, 114.61375882813807)
#得出的结论是不平稳

D_data = data.diff().dropna()
D_data.columns = [u'确诊总人数差分']
#差分后的结果
D_data.plot()
plt.show()
#时序图
plot_acf(D_data).show()
plt.show()
#自相关图

from statsmodels.graphics.tsaplots import plot_pacf
#偏自相关图
#plot_pacf(D_data).show()
#平稳性检测
#print(u'差分序列的 ADF 检验结果为: ', ADF(D_data[u'确诊总人数差分']))
```

```

#差分序列的 ADF 检验结果为: (-1.2113874732141632, 0.6686780253791897, 9, 18, {'1%': -3.859073285322359, '5%':
-3.0420456927297668, '10%': -2.6609064197530863}, 123.53995327978535)
#from statsmodels.stats.diagnostic import acorr_ljungbox
#print(u'差分序列的白噪声检验结果为: ', acorr_ljungbox(D_data, lags=1))
#差分序列的白噪声检验结果为: (array([15.83641633]), array([6.90605033e-05]))
from statsmodels.tsa.arima_model import ARIMA
#定阶
#一般阶数不超过 length/10
pmax = int(len(data)/30)
#一般阶数不超过 length/10
qmax = int(len(data)/20)
#bic 矩阵
bic_matrix = []
for p in range(pmax+1):
    tmp = []
    for q in range(qmax+1):
        try:
            tmp.append(ARIMA(data, (p, 1, q)).fit().bic)
        except:
            tmp.append(None)
    bic_matrix.append(tmp)
#从中可以找出最小值
bic_matrix = pd.DataFrame(bic_matrix)
#先用 stack 展平, 然后用 idxmin 找出最小值位置。
p,q = bic_matrix.stack().idxmin()
print(u' BIC 最小的 p 值和 q 值为: %s、%s' %(p,q))
#取 BIC 信息量达到最小的模型阶数, 结果 p 为 0, q 为 1, 定阶完成。
#建立 ARIMA(0, 1, 1)模型
model = ARIMA(data, (p, 1, q)).fit(dis=0)
#作为期 20 天的预测, 返回预测结果、标准误差、置信区间。
print(model.forecast(20))

```