


# 中国研究生创新实践系列大赛

# 中国光谷·“华为杯”第十九届中国研究生数学建模竞赛



| | |
|---|---|
| 学　　校 | 重庆大学 |
| 参赛队号 | 22106110005 |
| 队员姓名 | 1. 靳晓东 |
| | 2. 王逸飞 |
| | 3. 张海峰 |

# 中国研究生创新实践系列大赛

# 中国光谷·“华为杯”第十九届中国研究生数学建模竞赛

题　目　基于机理分析与组合模型的草原放牧策略研究

## 摘　　要：

内蒙古锡林郭勒草原是我国重要的畜牧业生产基地，在维持草原可持续发展的前提下，充分发掘草原经济资源具有重要意义。本文以草原可持续发展为约束，从机理角度分析、量化放牧策略对草原的影响，综合考虑了**并行集成学习随机森林低方差、串行集成学习 LightGBM 低偏差**的预测土壤性质的优势，并通过算例验证模型的准确性。

针对问题一，从土壤性质连续变化的机理角度分析，针对缺少单一变量下土壤湿度、植被生物量与对应放牧策略的数据，本文使用**水分平衡方程**将土壤含水量微分方程转化为**土壤贮水量微分方程**，采用 **K-Means** 方法对文献算例提供的围封区域与放牧区域土壤贮水量的差值聚类，量化不同放牧策略下对土壤含水量的影响。由于羊对植被生物量的影响具有随机性，本文构造以羊单位日食量 1.8 千克为期望、方差为标准差的正态分布随机生成函数，模拟不同羊对植被生物量的影响，基于载畜率与植被生物量的简单模型，引入放牧时长因子、强度因子等系数，综合表征放牧策略对植被生物量的影响，通过算例论证了模型的有效性。

针对问题二，采用 **Pearson**、**LASSO**、**Isomap** 和 **XGBoost** 方法进行特征筛选，排序投票选出 4 个包含 10 个变量的候选集，对候选集取交集，最终确定土壤蒸发量、降水天数、平均能见度、平均最大持续风速和平均露点温度作为 4 种深度土壤湿度的自变量。对于统计学习模型，单一预测模型的预测结果难以把偏差和方差同时满足最优，于是本文选择**串行-并行相结合的组合预测模型**，通过与构建的 **SVM**、**神经网络**、**GradientBoosting** 等**十一类**算法比较，以 5 折交叉验证的方式采用 **MSE**、**MedAE**、**可决系数**等多个指标评估，**随机森林-LightGBM** 组合预测模型效果最好，并接着对其参数进行了调优，调优后在土壤 10cm 湿度测试集上的平均预测精度 MSE 为 **0.6479**。使用灰色预测模型按统一月份输出 2022-2023 年的筛选出的特征变量后，本文使用组合预测模型预测不同深度土壤的湿度，见表 5.9。

针对问题三，从**土壤碳循环守恒**的机理角度分析，考虑植被枯萎率、牧群排泄物等增碳、减碳途径，构建土壤有机碳循环方程。基于皮尔逊系数挖掘各元素之间相关性，结果表明全 N 与有机碳高度强相关，相关文献证实二者符合线性关系，于是本文采用一次线性回归方程预测全 N 含量。由于无机碳与有机碳之间呈现负强相关，本文则采用三次回归方程预测无机碳，最后可推算得土壤全碳、土壤 C/N 比，见表。

针对问题四，本文选择 6 种指标反映土壤板结化程度，为了避免评价过程中主观

因素的影响，本文采取**熵值法-Topsis 综合评价法**客观衡量各指标对板结化的影响，使用 K-Means 方法将评价结果分为 3 类，量化土壤板结化等级。基于沙漠化程度指数预测模型，将放牧方式和放牧强度任意组合，利用 **K-Means** 方法进行聚类，得到不同放牧强度下沙漠化程度指数值。以沙漠化程度、板结化程度最小构建多目标优化模型，采用**基于动态线性标定的精英遗传算法**求解。

针对问题五，可持续发展体现在维持草原的化学因素和物理因素，题目中指出：土壤全氮含量反映了土壤状态，间接决定了牧羊阈值；且草地的植被直接决定放牧的强度，而植被的截流量能最好反映植被的生长能力。因此本文构建**基于土壤全氮含量和基于植被截断流量的约束方程**，使用熵值法-Topsis 方法客观评价各类指标，以经济效益最大为目标，构建单目标优化模型，并**基于重升温策略的模拟退火算法**求解。

针对问题六，基于以上模型，本文模拟了 4 位示范牧户的放牧策略，结果表明牧户 3 处于最佳的轻度放牧强度，土壤全 N 含量上升最快，值得其他牧户借鉴。通过模拟问题四给出的放牧策略，发现草原在较长时间内沙漠化、板结化仍处于非、轻状态。

**关键词：随机森林-LightGBM 组合预测；熵值法-Topsis 综合评价法；精英遗传算法；模拟退火算法；K-Means；XGBoost**

# 目录

# 1. 问题重述

## 1.1 问题背景

草原作为世界上分布最广的重要的陆地植被类型之一，其分布范围十分广阔。中国拥有 3.55 亿公顷的草原，占全球草原总面积的 6%~8%，位居全球第二。草原在保持生物多样性，涵养水土，净化空气，固碳，调节水土流失，防治沙尘暴等方面发挥着重要作用。自 2003 年党中央、国务院实施“退牧还草”以来，草原生态环境得到了较好的保护和改善，人民生活水平得到了明显改善。“退牧还草”并非是禁牧，而是在某些地区实行季休牧。因此，放牧政策的制定对促进区域经济发展、防止荒漠化、确保人民生活水平的提高起着至关重要的作用。

在草原上放牧，一般要综合考虑两个因素：放牧方式和放牧强度。过度放牧，常造成草地植被结构破坏，暴露土地面积扩大，使土体内水分的相对移动受阻，土壤积盐和脱盐平衡发生紊乱，盐分在地表沉积，使土壤盐碱化加剧。从而导致草场退化，土地荒漠化，破坏生态系统的平衡。

而适当的放牧能提高草原土壤质量，一方面能增加土壤中有机质和氮和钾含量，从而促进植物生理代谢，增强抗逆性，并促进植物对氮素营养的吸收和利用，减少土壤的板结。另一方面放牧能够降低表层土壤湿度、PH，一定程度增加土壤容重，促进土壤中凋落物的分解速率、微生物数量及活性以及有机碳和养分的积累

通过合理的放牧，可以改善草地土壤的品质，使土壤中的有机质、N、K 含量得到明显的提高，进而促进作物的生理代谢，提高作物的抗逆性，有利于作物吸收和利用 N 养分，降低土壤的板结。而放牧则可以减少土壤表层的湿度、 pH 值，并在一定程度上提高土壤的容重、加速土壤中的腐殖质分解、微生物数量活力、土壤有机质和营养物质的积累。

本试题针对草原的放牧策略进行建模研究，给出了包括内蒙古锡林郭勒草原概况在内的 15 个附件。本文在此基础上探讨了不同放牧策略下对锡林郭勒草原土壤物理性质和化学性质的影响，此外还需确定不同放牧强度下监测点的沙漠化程度指数值及



## 1.2 问题解决

基于上述研究背景，本题目共提供了“锡林格勒草原概况”等 11 项基础数据以及“内蒙古自治区锡林郭勒盟不同牧户生态畜牧业模式群落样方调查数据集”等 4 项监测点数据。基于 15 个附件内容，拟解决以下关键问题：

问题一：构建微分方程

从机理分析的角度探讨不同放牧策略（主要为放牧方式和放牧强度）对锡林郭勒草原土壤物理性质（主要为土壤湿度）和植被生物量影响的数学模型。

问题二：数据降维与回归预测建模

根据附件 3 土壤湿度数据、附件 4 土壤蒸发数据以及附件 8 中降水等数据，建立定量预测模型，在放牧策略不变的情况下对 2022 年、2023 年不同深度土壤湿度进行预测，并将结果填入表中。

问题三：碳循环平衡方程和线性回归建模

从机理分析的角度探讨锡林郭勒草原不同放牧方式、放牧强度对土壤化学特性

的影响。并请结合附录 14 的数据，对锡林郭勒草原（12 个牧区）的土壤同期有机碳、无机碳、全 N、土壤 C/N 比等值进行定量预测，并将结果填入表中。

问题四：沙漠预测模型

采用沙漠化程度指数预测模型及附加资料，计算出各放牧强度下不同监测点的荒漠化程度指数值，并提出了一个量化的土壤板结化定义，并以此为基础，结合问题 3，提出一个使得荒漠化程度和板结化程度最低的放牧策略。

问题五：

锡林郭勒草原 10 年的降雨量（包括降雪）一般为 300 毫米至 1200 毫米，请在给定的降水量（300 毫米、600 毫米、900 毫米和 1200 毫米）的条件下，在保持草原可持续发展情况下对实验草场内（附件 14、15）放牧羊的数量进行求解，找到最大阈值。

问题六：

在保留附件 13 中的示范放牧战略不变和问题 4 中得到的放牧计划两种情况下，以图表或动态演示的形式分别对各示范区 2023 年 9 月的土壤状况进行预测（例如土壤肥力变化、土壤湿度、植被覆盖。

## 2. 模型假设

1.不同放牧策略（放牧方式和放牧强度）对锡林郭勒草原土壤湿度、植被生物量、化学性质等因素影响呈现正态分布。

2.土壤含水量、植被生物量的变化为连续变化，均在空间上分布均匀。

3.忽略极端天气因素、环境污染对草原植被的影响。

4.针对不同等级对应的范围可能存在的重叠部分，本文用重叠部分数据的平均值作为分割点。



## 3. 符号说明

| 符号 | 符号含义 |
|---|---|
| W | 土壤水分总贮存量 |
| h | 土层厚度 |
| β | 土壤重量含水率 |
| $D_F$ | 时长因子 |
| $I_F$ | 强度因子 |
| C | 土壤容重 |
| O | 有机物含量 |
| a | 羊的单价 |
| Qi | 因子强度 |
| E | 地表蒸散发率 |
| α | 土壤植被覆盖率 |
| S | 牧羊数量 |

# 4. 问题一：模型的建立与求解

## 4.1 问题分析

问题一要求本文从机理分析角度建立不同放牧策略对锡林郭勒草原土壤物理性质和植被生物量影响的数学模型。拟从以下三个步骤解决问题一：（1）从**宏观角度**（土壤贮水量）和**微观角度**（土壤湿度）建立了**土壤水分动态模型**（2）将放牧方式进行不同的**函数化处理**，放牧强度则采用 K-Means 计算期望和方差，并计算每种放牧方式和强度下对土壤含水量影响的**高斯分布**（3）将影响植被生物量的因素加权组合成环境影响因子，牧群对植被生物量的波动性影响通过建立微分方程得到不同放牧策略对植被生物量的影响机理。

问题一的总体思路如图 4.1 所示

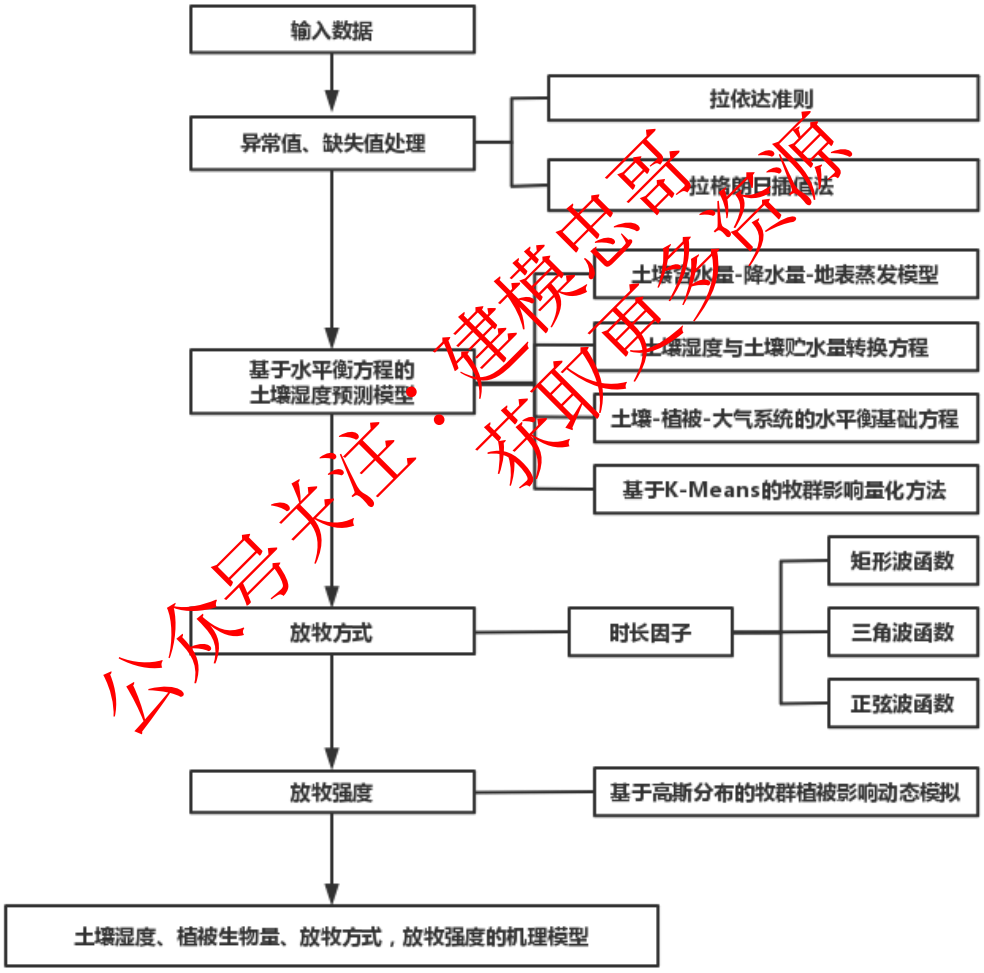


图 4.1 问题一的解题思路

## 4.2 数据预处理

对于缺失值主要有删除和插补处理，由于本题各类数据中样本点较少，因此对于缺失数据采用拉格朗日插值法进行补全处理。

$$l_i(x)=\prod_{j=0,j\neq i}^{k}\frac{x-x_j}{x_i-x_j}=\frac{x-x_0}{x_i-x_0}\cdots\frac{x-x_k}{x_i-x_k} \tag{1}$$

### 4.3 土壤物理性质机理分析

1.基于水分平衡方程转化的土壤湿度预测模型

（1）土壤含水量-降水量-地表蒸发模型

土壤湿度和土壤含水量的本质相同，只不过表现方式不同，后者通常是指 100 g 干燥的土壤中含有水分的重量，也就是所谓的土壤含水率。通过对土壤水分的测量，可以了解农作物对水分的需求，为农业生产提供有力的依据。

题目中虽然给出了土壤含水量-降水量-地表蒸发模型：

$$\frac{d\beta}{dt}=P-E(a) \tag{2}$$

式中，$P$ 为该牧区供水率；$E$ 为地表蒸散发率；$\beta$ 为土壤含水量；$\alpha$ 为土壤植被覆盖率可表达为 $\alpha^{*}G(w)$，$w$ 为成草数量，$G(w)=(1-e^{-\varepsilon_g w/w^{*}})$ 为草原的盖度，内蒙古草原盖度在 0.25~0.8 之间，$\alpha^{*}$ 为最大增长率，依赖于牧区草地除成草量外的环境条件（如光照、气温、土壤养分等），$\alpha^{*}$ 为有量纲系数，其余系数则是无量纲量。

公式中相关数据获取存在一定难度，如无法准确测量土壤植被覆盖率，最大增长率 $\alpha^{*}$ 的影响条件光照，气温，土壤等条件在现有技术条件下难以准确测量。因此，本文从土壤的微观特性和宏观特性出发，对土壤的物理性质进行综合分析。

（2）土壤湿度和土壤贮水量的衡量

**土壤湿度体现出土壤的微观特性，土壤贮水量则表现了宏观特性**，土壤贮水量和土壤湿度之间的关系如下[1]

$$W=p\times h\times\beta\times 10 \tag{3}$$

其中 $W$ 为土壤水分总贮存量（mm）；p 为地段实测土壤容重（g/cm$^3$）；h 为土层厚度（cm）；$\beta$ 为土壤重量含水率（%）。

土壤水分贮存量是指某一层土壤的蓄水总量，依据土壤湿度可以求得一定厚度土壤总的贮存水量，这两者之间存在着相互制约的关系。但是，土壤水分贮存量是由气象、土壤、植被、人为活动等因素共同作用的结果，具有极强的非线性特征，水分在土壤—农作物—大气间连续循环，通过降水、渗透、吸收、呼吸作用等形式周而复始地循环，过程机理复杂。在没有人为干预的条件下，可以使用下列公式来描述土壤-植物-大气-水平衡的基本方程式：

$$\Delta W=W_{t+1}-W_t=P+G_u+R_{in}-(Et_a+G_d+R_{out}+IC_{store}) \tag{4}$$

其中，$\Delta W$ 为土壤贮水变化量，$W_{t+1}$ 和 $W_t$ 分别为一段时间内的始末土壤含水量，$P$ 为降水量，$G_u$ 和 $G_d$ 分别为毛管上升量和土壤水渗透量，$Et_a$ 为实际蒸腾量，$R_{in}$ 和 $R_{out}$ 分别为入径出径流量，$IC_{store}$ 为植被截流量。由于锡林郭勒草原地势比较平坦，降水量和降水强度较少，水分循环以垂直方向的水量交换为主，绝大部分降水被蓄积在土壤中，尽管在遇到较大降水时会产生局地径流，但仍在整个草原区域内，其出入径流可视为相等，降水产生的径流量一般情况下可不考虑。

本文主要研究内蒙古锡林郭勒草原，其位于内蒙古高原锡林河流，地理坐标介于东经 110°50′~119°58′,北纬 41°30′~46°45′之间，年均降水量 340mm。地处其中心地带的锡林浩特国家气候观象台野外试验研究基地，位置处于东经 116° 19′50″北纬 43° 07′58″，年降水量 286．6mm。锡林浩特国家气象观测站位于内蒙古锡林郭勒大草原中部，其年平均降雨量处于平稳状态，极具代表性。因此，本文以内蒙古锡林郭勒草原地区的地下水状况，近似表征内蒙古锡林郭勒草原的地下水情况。

本文研究区地下水埋藏较深，多在三、四十米以下。现有文献已证明当地下水埋深大于 4m 后，土壤中毛管上升水对 2m 土壤水分循环的作用可忽略不计[2]。地下水毛管上升量对根系层的补给量也可忽略。若土壤水分测定结果为初始值 $W_t = W_0$，则可参考相关文献[3]将（3）式可简化为

$$W_{t+1} = W_t = P + Wo - (Et_a + G_d + IC_{store}) \tag{5}$$

植物根系在土壤中具有一定的锚固性，从而提高了土壤的粘附性和抗蚀能力，从而降低了水土流失。同时，在一定的海拔高度，可以有效的抑制雨水对地面的直接冲击，并在极小的高度形成二次降水，从而有效地避免了土壤受到溅蚀。植物根的锚固性和植物茎、叶片的水文效应都会对储水能力产生一定的影响，因此需要着重考虑[4]。

植被截流量与降水量、植被覆盖度、叶面积指数（LAI）等密切相关。植被覆盖率是植物群落覆盖地表状况的一个综合量化指标，能够直观反映地表植被的丰度。降水量小、植被覆盖度高、LAI 大时植被截流量大，其表达式为：

$$IC_{store} = c_p \cdot IC_{\max} \cdot \left[1 - \exp\left(-k \cdot R_{cum} / IC_{\max}\right)\right] \tag{6}$$

上式中，$IC_{store}$ 为植被截流量（mm）；$c_p$ 为植被覆盖率；$IC_{\max}$ 特定植被的最大截流量（mm）；$k$ 为植被密度校正因子，与LAI有关；$R_{cum}$ 为累积降雨量（mm）。

$IC_{\max}$ 可以通过 LAI 来估算：

$$IC_{\max} = 0.935 + 0.498 \cdot LAI - 0.00575 \cdot LAI^2 \tag{7}$$

（3）土壤水分动态模型

由于土壤的含水量属于模拟量，在大自然中出现土壤含水量突变的情况很少，所以本文认为土壤的含水量为连续变化，因此当 $t = t_0$ 时，对土壤含水量-降水量-地表蒸发模型（1）式进行积分可得：

$$\beta_{t_0} = \int_0^{t_0} (P - E(\alpha))dt \tag{8}$$

当 $t = t_1$ 时，可得：

$$\beta_{t_1} = \int_{t_0}^{t_1} (P - E(\alpha))dt + \beta_{t_0} \tag{9}$$

其中供水率在短时间内变化很小，本文认为其在微小的时间段内相同，随着时间跨度的延长，供水率应当进行改变，当 $t = t_n$ 时可得：

$$\beta_{t_n} = \sum_{i=1}^{i=n} \int_{t_{i-1}}^{t_i} (P - E(\alpha))dt \tag{10}$$

基于上式可得土壤贮水量为：

$$W_{t_n} = p \times h \times \beta_{t_n} \times 10 = p \times h \times 10 \times \sum_{i=1}^{i=n} \int_{t_{i-1}}^{t_i} (P - E(\alpha))dt \tag{11}$$

此时土壤水分动态模型为：

$$W_{t_{n+1}} = P + p \times h \times 10 \times \sum_{i=1}^{i=n} \int_{t_{i-1}}^{t_i} (P - E(\alpha))dt - (Et_a + G_d + IC_{store}) \tag{12}$$

（4）放牧策略对土壤含水量的影响

从公式（11）可知，土壤含水量的变化土壤贮水量的变化来表达，在放牧过程中，由于家畜的密度太高，会造成草地植被的破坏，暴露的土地面积增加，从而加速土壤的蒸发，从而使土中的水的相对运动受到负面的影响。放牧模式与放牧强度对土壤含水量的影响最大，其中放牧模式为：全年连续放牧、禁牧、选择划区轮牧、轻度放牧、生长季休牧，对放牧时间有较大的影响，因此本文选择放牧时长因子表征：

表 4.1 放牧时长因子表征

| 放牧方式 | 放牧时长因子 |
|---|---|
| 全年连续放牧 | 1 |
| 禁牧 | 0 |
| 选择划区轮牧 | 矩形波函数 |
| 轻度放牧 | 三角波函数 |
| 生长季休牧 | 阶跃函数 |

文献[5]指出，锡林郭勒草原的生长季大约为每年的 4 月至 10 月，阶跃函数起始时间与终止时间的设置，依照草原生长季时间位于全年时间的起始时间和种植时间。所有时长因子的最高值为 1。

放牧强度可以分为四种，分别为：对照、轻度放牧强度、中度放牧强度、重度放牧强度，但是这四者之间的区别，在定义上比较模糊，有 2 种区分方式，如下表所示：

表 4.2 不同放牧标准

| | 标准一 | 标准二 |
|---|---|---|
| 对照 | 0 羊/天/公顷 | 0 羊/天/公顷 |
| 轻度放牧强度 | 2 羊/天/公顷 | 1-2 羊/天/公顷 |
| 中度放牧强度 | 4 羊/天/公顷 | 3 -4 羊/天/公顷 |
| 重度放牧强度 | 8 羊/天/公顷 | 5-8 羊天/公顷 |

文献[6]的算例中给出了锡林浩特国家气候观象台野外试验研究基地围封地区的贮水量和放牧地区的贮水量在某次实验中的数据，二者的差值即为不同放牧强度下贮水量的变化，由于处于自然放牧状态的牧群，包含了轻、中、重度放牧强度，所以进行非监督学习聚类，即可得到不同放牧强度对贮水量的影响，经过转化后即为土壤含水量，本文选择 K-Means 算法进行聚类。

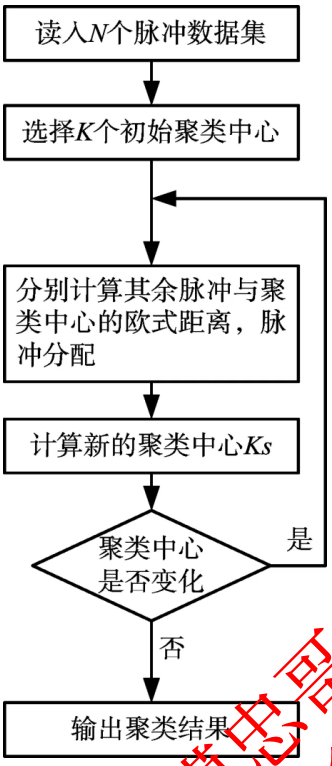


图 4.2 K-Means 算法流程图

因此本文针对贮水量差值进行 K-Means 聚类并确定质心数量为 3，为了准确模拟不同土壤含水量的大小，本文对每类点的平均值作为中心，计算每类点的方差、期望，采用期望和方差参数化描述每种强度下对土壤含水量影响，最终放牧方式、放牧强度对土壤含水量的影响如下：

$$\frac{d\beta}{dt}=\frac{P-(Et_a+G_d+IC_{store})}{p\times h\times 10}-D_f\times I_f\times W \tag{13}$$

其中 $D_f$ 表示时长因子，$I_f$ 表示强度因子，$W$ 表示相关经验权重。

2、基于牧群动态模拟的植被生物量预测模型

自然条件下，降水、温度、湿度、植被种类等是影响植被水分平衡的重要因子，而植物的生长不仅要考虑水分，还必须考虑土壤湿度、 PH 值、土壤中的养分等其他因素。对于放牧与植物生长之间的关系，Woodward 等建立了如下一个简单模型：

$$\frac{dw}{dt}=0.049w(1-\frac{w}{4000})-0.0047Sw \tag{14}$$

式中， $w$ 为植被生物量， $S$ 为单位面积的载畜率。

该模型只考虑放牧影响，简单地反映了载畜率对植被生物量的作用，没有考虑其他因素的影响，只是从某一个侧面刻画某一个因素对于植被生长的影响。但由于植被生物量处于连续变化的过程，因此对上式进行积分：

该模型仅考虑了放牧效应，仅考虑了载畜率对植物生物量的影响，而忽略了其它因素，仅从一个角度刻画了某一因子对植被生长的影响。但是，因为植物的生物量是一个不断变化的过程，因此对原式进行积分。其中 $W_0$ 表示植被生物量的初始值，

在$t=t_1$时植被生物量为：

$$W_{t_1}=\int_{t_0}^{t_1}(0.049w_{t_0}(1-\frac{w_{t_0}}{4000})-0.0047Sw_{t_0}) \tag{15}$$

当$t=t_n$时：

$$W_{t_n}=\sum_{i=1}^{i=n}\int_{t_{i-1}}^{t_i}(0.049w_{t_i}(1-\frac{w_{t_i}}{4000})-0.0047Sw_{t_i})dt \tag{16}$$

公式（13）只对原有简单模型进行积分，**未考虑到植被生物量受到多种环境因素的影响，以及牧群对植被生物量影响的随机性等问题**，后续将对上述两种影响因素进行分析

（1）植被生物量受哪些因素的影响

根据文献，本文选择土壤 PH 值、平均气温、土壤含水量、土壤全碳作为影响植被生物量的主要外界因素，每种因素通过加权的方式表征其对植被生物量的影响，多种因素组合成环境影响因子。

（2）每只羊对植被生物量的影响

标准[7]指出，对于“羊单位日食量”的定义为：1 只羊单位家畜每天所需从草地摄取含水量 14%的标准干草为 1.8 千克，对于牛、马、骆驼等大牲畜，采取折算系数 6.0，对于大牲畜幼崽采取折算系数 3.0，针对每只羊进食量会出现波动的情况，本文选择均值为 1.8，符合标准方差的正态分布来表征每只单位羊对植被生物量的波动性影响。此时植被生物量和环境因素、放牧测量之间的关系为：

$$W_{t_n}=\sum_{i=1}^{i=n}\int_{t_{i-1}}^{t_i}(0.049w_{t_i}(1-\frac{w_{t_i}}{4000})-0.0047D_f\times Sw_{t_i}+f(\beta_{t_i},PH_{t_i},T_{t_i},STC_{t_i})+D_f\times S\times N_s)dt \tag{17}$$

其中$D_f$表示时长因子，$N_s$表示均值为 1.8，符合方差为 1 的正态分布，通过每次随机生成每只羊食草量，模拟每只羊每天食量的变化。

3、模型验证

为了充分验证本文模型的合理性，利用文献[7]算例的数据，在相同土壤含水量初始条件下，采用全年放牧的方式，对该草原采用不同强度的放牧，进行为期 6 个月的迭代计算后，土壤含水量的变化如下所示：

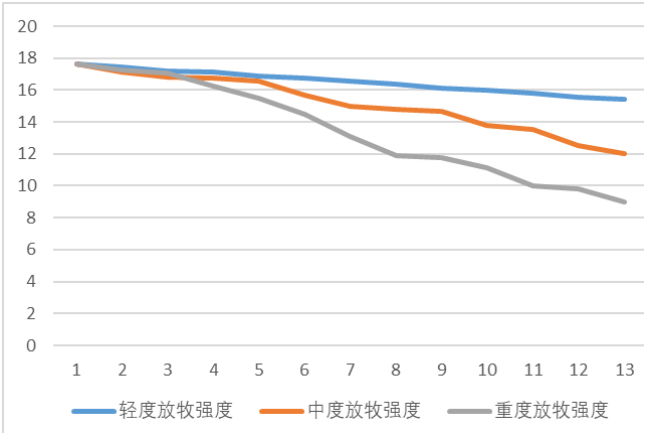


图 4.3 土壤含水量变化图

从上图中可以发现，随着时间的推移，处于重度放牧状态下的土壤含水量出现快速的下降，其中出现了斜率快速增大的点，客观反映了土壤含水量降低后，土壤重植被量减少，造成大片土壤裸漏，加快了土壤水分的蒸发；中度牧羊强度在初期和轻度牧羊强度效果类似，而处于轻度放牧状态的土壤含水量保持稳定，以上实验结果说明本模型能够客观反映土壤中含水量的变化情况。

为了验证本文模型的真实性和准确性，本文使用附件 15 中 G6 放牧小区多种植物的干重之和作为主要研究对象，结果如下图所示：

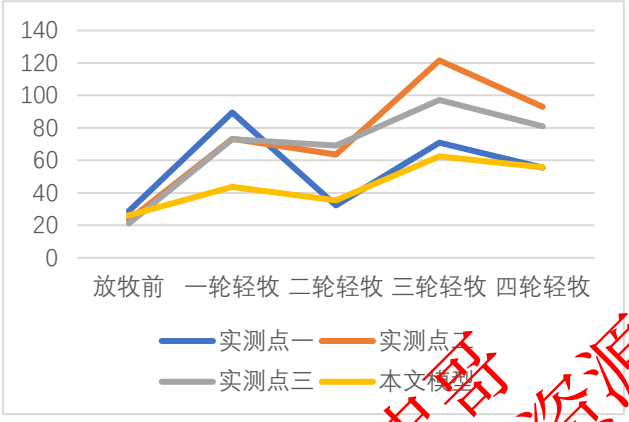


图 4.4 轻牧状态下植被生物量变化

从上图可以发现，本文模型的模拟效果与其他对比实测点的变化大致相似，但是对每轮轻牧后，植被生物量的生长量变化模拟不够精确，这主要原因是影响植被生物量生长的因素有很多，例如在放牧后，羊群的粪便会增加土壤中化学物质的含量，间接的增加植被的生长速度，而这一因素在本模型中没有体现，所以本模型的预测结果与实测点相比较为缓慢。

# 5. 问题二：模型的建立与求解

## 5.1 问题分析

问题二要求依据土壤蒸发数据以及降水等数据对未来 2 年土壤不同深度湿度进行定量预测，本文从以下三个步骤解决问题二：（1）对附件 8 中数据异常值采用**拉依达(PauTa)准则**进行剔除，采用**割线法**对缺失值进行补全；（2）对附件 8 中原始气候变量进行筛选，采用 **Pearson、LASSO、Isomap** 和 **XGBoost** 方法先剔除冗余的强相关性变量，再按特征重要性排序，然后采用投票方式选出 4 组分别对土壤 4 个不同深度湿度具有显著影响的变量候选集，最后将 4 个候选集取交集后得到最终筛选出的特征变量。（3）对筛选出的变量采用**灰色预测模型**进行预测，得出预测结果作为后续步骤的代入数据集；（4）建立 4 个**随机森林+LightGBM** 的组合预测模型分别对土壤 4 个不同深度湿度进行预测。

问题二的总体思路如图 5.1 所示

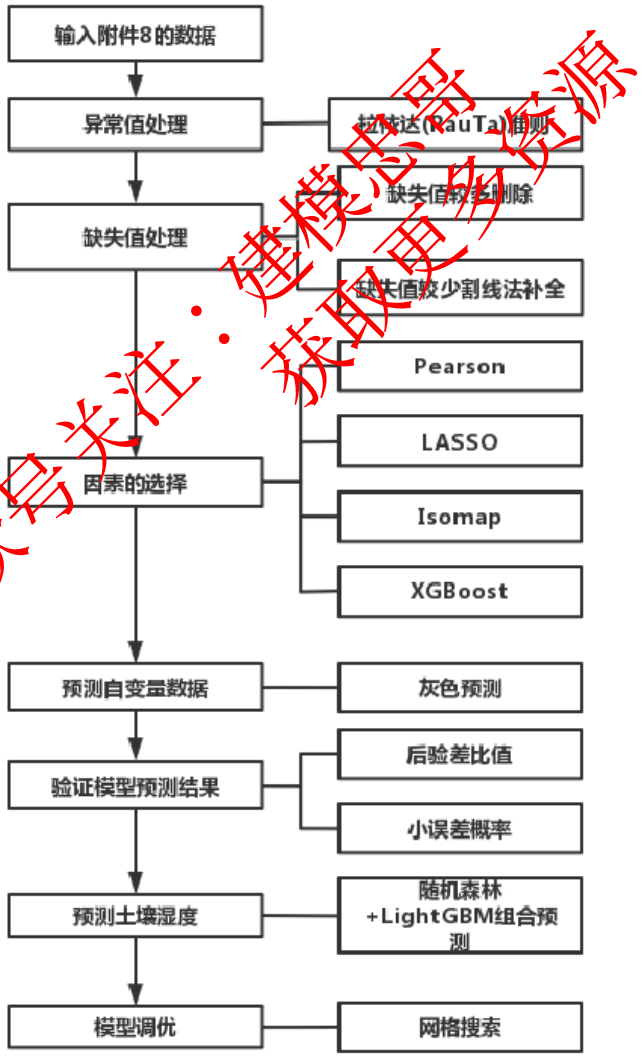


图 5.1 问题二的解题思路

## 5.2 数据预处理

1.数据异常值处理

对“附件 8、锡林郭勒盟气候”中数据采用**拉依达(PauTa)准则**将超过 3s 的数据进行剔除。

$$v_b = \left| x_b - \bar{x} \right| > 3\sigma, 1 \leqslant b \leqslant n \tag{18}$$

其中

$$\sigma = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{19}$$

同时剔除掉按月份取值均为固定值的气候因素：平均气温≥18℃的天数、平均气温≥35℃的天数、平均气温≤0℃的天数 3 项因素。

2.缺失值处理

将同一个月份的几年间缺失值多于 3 个的气候因素进行剔除。

剩下缺失值较少的数据采取**割线法**进行补全，迭代式如下：

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \tag{20}$$

剔除积雪深度、平均最大瞬时风速(knots)、最大瞬时风速极值(knots)3 项因素。保留的变量如表 5.1 所示。

表 5.1 保留的 18 个变量

| | | |
|---|---|---|
| 平均气温(℃) | 平均最高气温(℃) | 平均最低气温(℃) |
| 最高气温极值(℃) | 最低气温极值(℃) | 平均露点温度(℃) |
| 降水量(mm) | 最大单日降水量(mm) | 降水天数 |
| 平均海平面气压(hPa) | 最低海平面气压(hPa) | 平均能见度(km) |
| 平均站点气压(hPa) | 最大能见度(km) | 最小能见度(km) |
| 单日最大平均风速(knots) | 平均最大持续风速(knots) | 平均风速(knots) |

## 5.3 变量筛选

通过对变量进行筛选，找出对土壤湿度影响最显著的变量。通过 4 个方法，可以对所有变量进行近似表达。各变量间的作用可以用相关性来表达，且相关性越大，其作用愈明显。由于传统的特征筛选法很难精确地反映各个变量间的相关性，而且各个方法的原理及衡量结果也不尽相同，所以本文采用了四种包括线性和非线性特征筛选法的综合选取，并将四种方法中得到的变量相关性排序结果进行了综合，从中得到最好的特征变量，并给出了如下的特征筛选模型：

假定在方法 1 至 4 中，所选的变量集为 B1、B2、B3、B4，各个变量集包含 10 个变量，且根据变量的相关性程度进行排序。选取最佳变量的指标为（1）变量出现频数（愈多愈好）（2）变量排序（愈往前愈好）。由于各项方法的评估结果都存在差异，所以本文选取了人工的投票模型选择变量。最后将 4 个 B5 集合中所得出的变量集取交集得出影响不同深度土壤湿度的“公有变量”为最终结果。

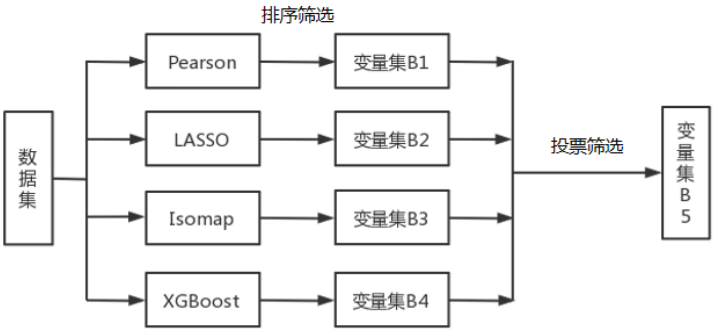


图 5.2 变量筛选流程图

（1） Pearson 相关系数分析（画热力图）

皮尔逊相关系数输出范围为-1—+1，绝对值越大其相关性越强。

$$\rho_{XY}=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{n}\frac{(X_i-E(X))}{\sigma_X}\frac{(Y_i-E(Y))}{\sigma_Y}}{n} \tag{21}$$

Pearson 相关系数分析得到的变量集 B1 中的变量作相关性分析，部分变量间的相关系数如图 5.3 所示为：

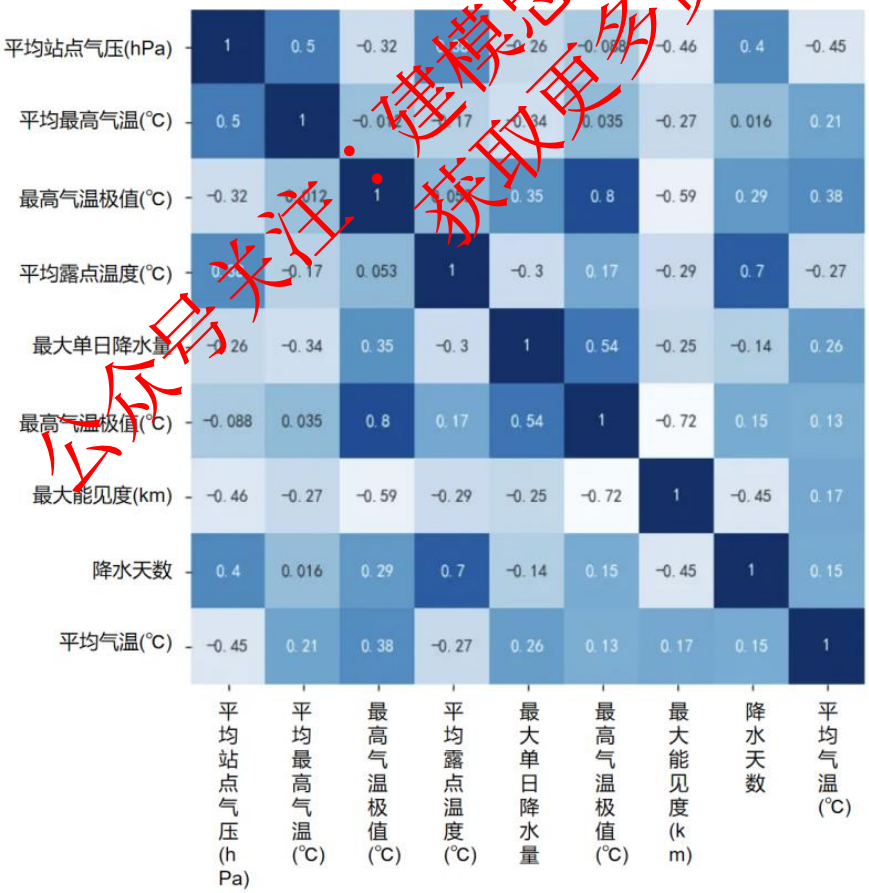


图 5.3 变量集 B1 部分变量相关系数热力图

（2） Lasso 回归

Lasso 回归的特点是建立一维连续因变量、多维连续因变量、非负次数因变量、二元离散因变量、多元离散因变量。lasso 可以处理因变量是否连续或离散，并通过

筛选变量的方式以减少模型的复杂性。

（3）Isomap

Isomap 相比于 MDS 其优势在于它采用了“测地距离”技术，而非欧几里德距离，从而可以有效地减少数据的丢失，同时能够显示出更多的高维数据。

（4）XGBoost

XGBoost 是 boosting 算法的一种实现方式，能够有效减少模型的误差。基本思路为不断生成新的决策树，每棵树都是基于上一颗树和目标值的差值来进行学习，从而降低模型的偏差。

$$Obj^{t}=\sum_{i=1}^{n}[l(yi,y_i^{t-1})+g_i f_t(x)+1/2g_i f_t^{\ 2}(x)]+\Omega(f_t)+constant \tag{22}$$

（5）排序，投票，交集

通过对 4 种不同方法筛选出的变量集进行比较，我们可以看出，不同的方法可以得到不同的特征集，但也有部分变量在每个变量集中重复出现，比如土壤蒸发量(mm)和降水天数。将每个变量集中的变量按照其是否出现分别赋值 1（出现）和 0（未出现），选择那些在多个变量集中值为 1 的变量进入候选集，对土壤不同深度湿度对应的 4 个候选集取交集，最终交集包含的筛选出的变量如表 5.4 所示。

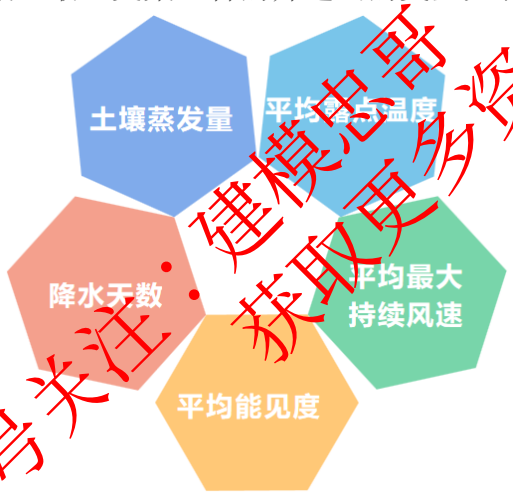


图 5.4 最终交集包含的变量



## 5.4 灰色预测模型

通过上述筛选后的数据能够对 2022、2023 年不同深度的土壤湿度进行更好的预测。根据题目已知条件、需要先对 2022 年、2023 年气候因素及土地蒸发量进行预测。由于相同的时间段（每年相同的月份）内气候因素存在一定的内部关联，因此可采用连续 10 年的数据值预测第 11 年及第 12 年的数据。

由于题目中所给数据量较少，对于小样本学习来说，若采用 MLP 神经网络模型训练学习，容易产生过拟合，导致模型泛化能力差，因此本文采用**适用于小样本时间序列预测的灰色预测模型**对气候因素及土地蒸发量进行短期预测。首先利用某一月 2012—2019 年数据对 2020-2021 年数据预测，验证模型验证误差可行后，再利用 2012—2021 年数据预测 2022-2023 年数据。

1.模型可行性验证

（1）给定观测数据列

$$x^{(0)}=\{x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(N)\} \tag{23}$$

（2）经一次累加得

$$x^{(1)}=\{x^{(1)}(1),x^{(1)}(2),\cdots,x^{(1)}(N)\} \tag{24}$$

（3）设 $x^{(1)}$ 满足一阶常微分方程，其中 a 是常数，u 称为发展灰数,为内生控制灰数，是对系统的常定输入

$$\frac{dx^{(1)}}{dt}+ax^{(1)}=u \tag{25}$$

（4）在此条件下对方程进行求解

$$x^{(1)}(k+1)=[x^{(1)}(1)-\frac{u}{a}]e^{-ak}+\frac{u}{a}. \tag{26}$$

（5）对 $x^{(0)}$ 的均值 $S_1$ 和残差 D 的均值 $S_2$ 分别求得后，计算后验差比值 C 和小误差概率 P。

$$C=\frac{S_2}{S_1} \tag{27}$$

$$P=P\left\{\left|E(k)-\overline{E}\right|<0.6745S_1\right\} \tag{28}$$

下图展示了在 2012—2021 年间 1 月份的平均能见度(km)的预测结果如下图 5.5 和表 5.2 所示；7 月份的平均露点温度(℃)的预测结果如下图 5.6 和表 5.3 所示。

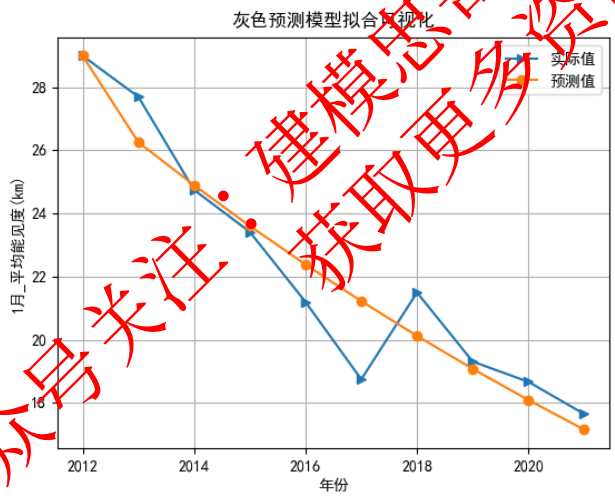


图 5.5 1 月份平均能见度(km)预测结果

表 5.2 1 月份的平均能见度(km)精度检验

| 数据是否通过光滑检验 | 数据是否通过级比检验 | 后验差比值 | 小误差概率 |
|---|---|---|---|
| 是 | 是 | 0.296 | 1 |

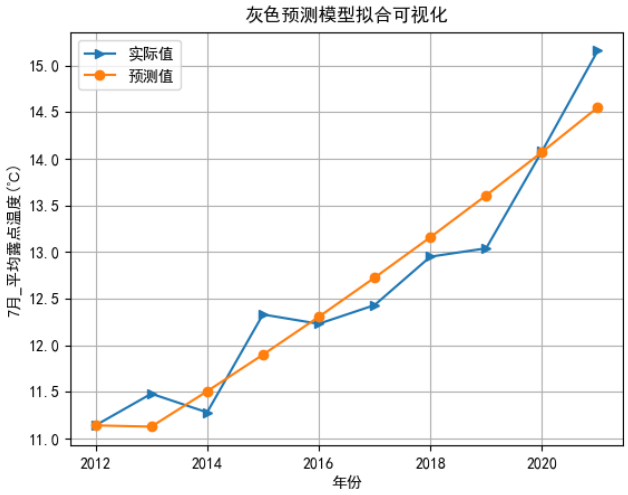


图 5.6 7 月份的平均露点温度(℃)

表 5.3 7 月份的平均露点温度(℃)

| 数据是否通过光滑检验 | 数据是否通过级比检验 | 后验差比值 | 小误差概率 |
|---|---|---|---|
| 是 | 是 | 0.287 | 1 |

从以上图表可以看出，两个气候因素 16 年间的数据均通过了光滑检验和比级检验。1 月份平均能见度(km)的预测结果计算结果 C≈0.296，P=1；7 月份的平均露点温度(℃)的预测结果计算结果 C≈0.287，P=1，将计算结果与后验差预测精度等级对照表 5.4 进行对比，得出该气候因素的灰色预测模型预测精度等级为“好”。

表 5.4 预测精度等级对照表

| 预测精度等级 | P | C |
|---|---|---|
| 好 | ＞0.95 | ＜0.35 |
| 合格 | ＞0.80 | ＜0.45 |
| 勉强 | ＞0.70 | ＜0.50 |
| 不合格 | ≤0.7 | ≥0.65 |

## 5.5 随机森林+LightGBM 组合预测

1.回归预测方法调研

针对问题二，需要对筛选出的变量进行回归预测分析。目前主流的回归分析方法有回归算法、正则化方法、决策树学习、集成算法。对这些回归方法的特点和优缺点进行调研，得到结果如下表 5.5 所示。

表 5.5 回归分析方法比较

| | 方法描述 | 优缺点 |
|---|---|---|
| 回归算法 | 通过测量误差，探讨了线 | 模型快速，适合数据量 |

| | | |
|---|---|---|
| | 性回归、逻辑回归等因素与自变量的关系。 | 小，关系简单，对非线性数据的拟合效果差 |
| 正则化方法 | 一般是一种基于复杂度的回归方法的扩展。例如，利亚特自动选择和选择操作（LASSO）、岭回归等 | 虽然可以避免过度拟合，改善模型的推广效果，但也会产生不足的拟合。 |
| 决策树学习 | 依据数据属性，采用树型结构，构建了基于属性的决策模型。例如分类和回归树，C4.5，随机森林等 | 该方法具有较高的复杂性和高度非线性，易于说明，但存在过度拟合、行速度缓慢、存储量大等缺点。 |
| 集成算法 | 通过几个比较薄弱的学习模式，分别对同一样品进行单独的训练，把预测的结果综合在一起，从而实现整体的预测。例如，Boosting, Bagging, AdaBoost，梯度推进器，GBM 等. | 最早的几种预测方法都是采用综合算法。比其它单一模式所预测的精度更高 |

对于预测模型的在实际应用过程中往往会产生泛化误差，从下式推导可得出泛化误差主要包括了方差和偏差，方差和偏差的含义如下图所示：

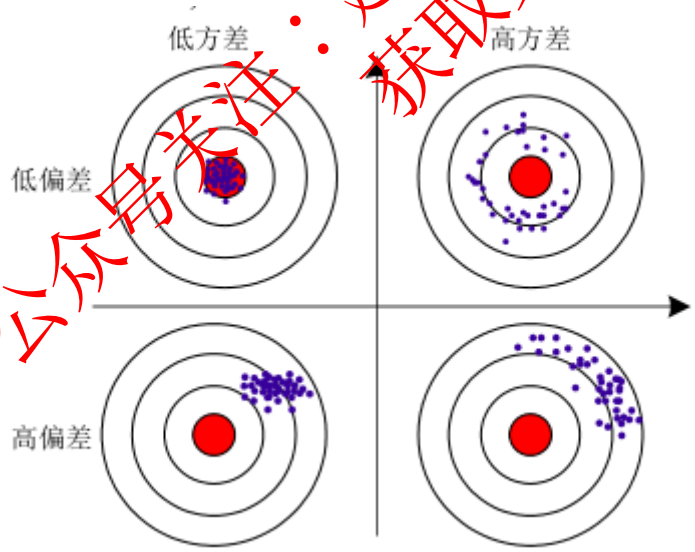


图 5.7　统计学习类算法偏差与方差的分布情况

除了数据分布中存在的噪声，泛化误差里主要存在着方差和偏差，下式中$E$指泛化误差，$f$是由训练集$D$学得的模型，$x$是测试样本，bias 指偏差，var 指方差，$\varepsilon^2$是噪声。

$$E(f;D)=\mathrm{E}_D[(f(x;D)-y_D)^2]$$
$$=\mathrm{E}_D[(f(x;D)-\bar{f}(x)+\bar{f}(x)-y_D)^2]$$
$$=\mathrm{E}_D[(f(x;D)-\bar{f}(x))^2]+\mathrm{E}_D[(\bar{f}(x)-y_D)^2]+\mathrm{E}_D[2(f(x;D)-\bar{f}(x))(\bar{f}(x)-y_D)]$$
$$=\mathrm{E}_D[(f(x;D)-\bar{f}(x))^2]+\mathrm{E}_D[(\bar{f}(x)-y_D)^2]$$
$$=\mathrm{E}_D[(f(x;D)-\bar{f}(x))^2]+\mathrm{E}_D[(\bar{f}(x)+y-y-y_D)^2]$$
$$=\mathrm{E}_D[(f(x;D)-\bar{f}(x))^2]+\mathrm{E}_D[(\bar{f}(x)-y)^2]+\mathrm{E}_D[(y-y_D)^2]+2\mathrm{E}_D[(\bar{f}(x)-y)(y-y_D)]$$
$$=\mathrm{E}_D[(f(x;D)-\bar{f}(x))^2]+(\bar{f}(x)-y)^2+\mathrm{E}_D[(y-y_D)^2]$$

综上可得，泛化误差可表示为：

$$E(f;D)=\mathrm{bias}^2(x)+\mathrm{var}(x)+\varepsilon^2 \tag{29}$$

相对于单一预测模型，由于模型原理的限制，预测结果往往只能侧重减少一项误差，这在实际应用过程中带来了较大的误差，而通过将多个预测模型的预测结果加权处理，能够实现减小偏差的同时减小方差，目前组合预测方法是主流的预测方法。

2.基于随机森林的并行集成学习方式

随机森林是最典型的并行集成学习方法，它可以分成两个阶段，即自助采样和投票组合，其算法结构如图 5.8 所示。随机森林采用自助采样方法样本，由此生成一个随机样本，假定在该数据集中存在 m 个样本，每一次随机放回 1 个样本，重复 m 次，获得包含 m 个样本的新变量集，随机有放回的样本运算，以使每一样本均可选取，且新变量集可能出现重复样本。在 n 次的自我抽样中，可以获得 n 个样本，每一个变量集都包含 m 个样品，在此基础上，训练 n 个预测模型，再用投票和策略将 n 个预测模型合并。

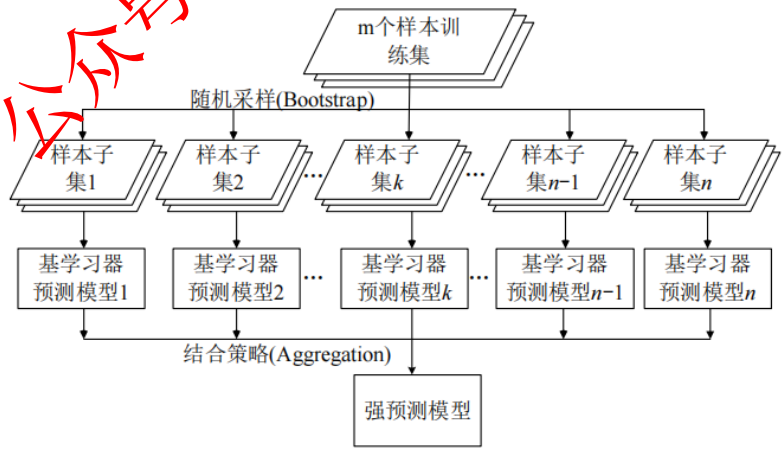


图 5.8　基于随机森林的并行集成学习机理，

3.基于 LightGBM 的串行集成学习方式

LightGBM 串行集成学习算法机理如下，对于已知数据集

$D=\{(x_i,y_i)\}(|D|=n,x_i\in R^m,y_i\in R)$ 树的集成模型由下式表示：

$$y=\sum_{k=1}^{K} f_k(\mathrm{x_i}),f_k\in F \tag{30}$$

式中 $F=\{f(x)=w_{q(x)}\}(q:R^m\to T, w\in R^T)$ 是树的集合空间；xi 为第 i 个数据点的特征向量；q 为每一棵树的结构映射到样本所对应的叶子的索引；T 为树上叶子的数量，每一棵树 fk 对应一个独立的树结构 q 和叶子的权重 w。基于 LightGBM 的串行集成学习机理如图 5.9 所示。

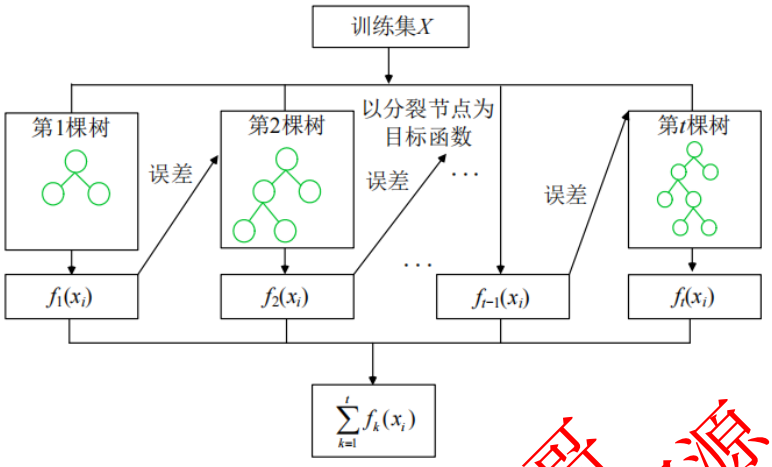


图 5.9 基于 LightGBM 的串行集成学习机理

4. 基于串行——并行组合预测模型分析

通过对统计习类算法的误差分析，可以看出，训练后的模型在保证最大程度上符合实际情况的前提下，保证了较小的方差，减小了数据情况对模型的影响。通常，很难使偏差与方差都能得到最佳的结果。在给定的学习任务条件下，在初始阶段，模型的拟合能力很差，且训练数据的干扰不能引起模型的显著改变，这时，偏差是最主要的影响因素，模型处于欠拟合状态。模型的拟合性能随训练水平的提高而增强，其偏差值逐渐减小，拟合性能逐渐达到最优状态。在此基础上，通过不断地训练模型，逐步学习到数据的干扰原理，并逐步由方差控制模型精度，学习数据本身的非全局性质，从而产生了过拟合。

LightGBM 是一种基于树状结构的串行集成学习算法，其整合方法是由基学习器的连续迭代叠加而成。通过反复迭代，得到的模型会根据上一次迭代的结果进行修正，从而使损失函数的顺序最小化，从而逐步减小误差。但是，采用了序列化、适应性优化的方法，使得各个子模型间存在着很强的关联度。串行集成的学习方法对模型的变异和过度拟合的危险没有明显的减少。与 LightGBM 的串行整合训练模式不同， Bagging 的并行集成学习算法是将原始训练数据中的任意一组进行抽样，然后对每个基学习器进行单独的训练，并进行平均值运算。尽管每个基学习器的输入都遵循同样的数据分布情况，但每一次所选取的数据都是独立的。该方法能有效地减少方差，避免模型过拟合的影响，提高了模型的泛化性。表 5.6 显示了串行和并行集成的相关机制。

表 5.6 不同算法机理分析

| 算法名称 | 训练方式 | 基学习器 | 基学习器相关性 | 主要针对误差类型 |
|---|---|---|---|---|
| 随机森林 | 并行集成 | 强预测模型 | 弱相关性 | 减小方差 |
| LightGBM | 串行集成 | 弱预测模型 | 强相关性 | 减小偏差 |

基于上述分析，本文以自变量的灰色预测值作为输入数据，并采用**随机森林**

**+LightGBM 组合预测**的方法对土壤不同深度湿度进行预测。随机森林+LightGBM 的组合预测模型综合考量了串行集成学习算法与并行集成学习算法的特点，既可以最大限度地利用 2 种模型的优势，又可以减小模型的方差和偏差，从而使模型更加稳定，具有更好的泛化能力。

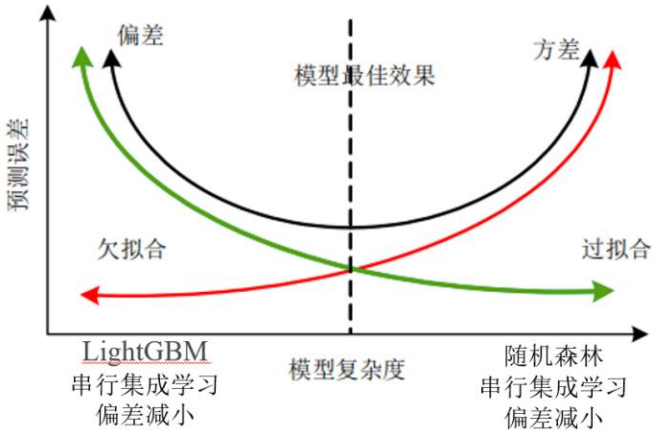


图 5.10 组合预测效果分析

## 5.6 对比分析与模型验证

本文对 2012-2021 年间 10 年的数据共 120 条按 8：2 划分训练集和测试集，统一采用 5 折交叉验证的方式对包括 **MLP、DecisionTree、AdaBoost** 等 **11** 个回归预测模型进行验证，检验其平均精度。采用平均绝对误差(MAE)、均方误差(MSE)、中位绝对误差(MedAE)和 R2 (R-Square)可决系数作为 4 个评价指标来确定最优回归预测模型。

以 10cm 湿度(kg/m2)的模型为例（其他深度的土壤湿度预测实验对比效果类似），各个模型交叉验证的最终实验结果如图 5.11 所示。设置合适的参数对机器学习方法十分重要，为了保证实验结果的客观性和可对比性，这里用到的模型均采用默认参数。

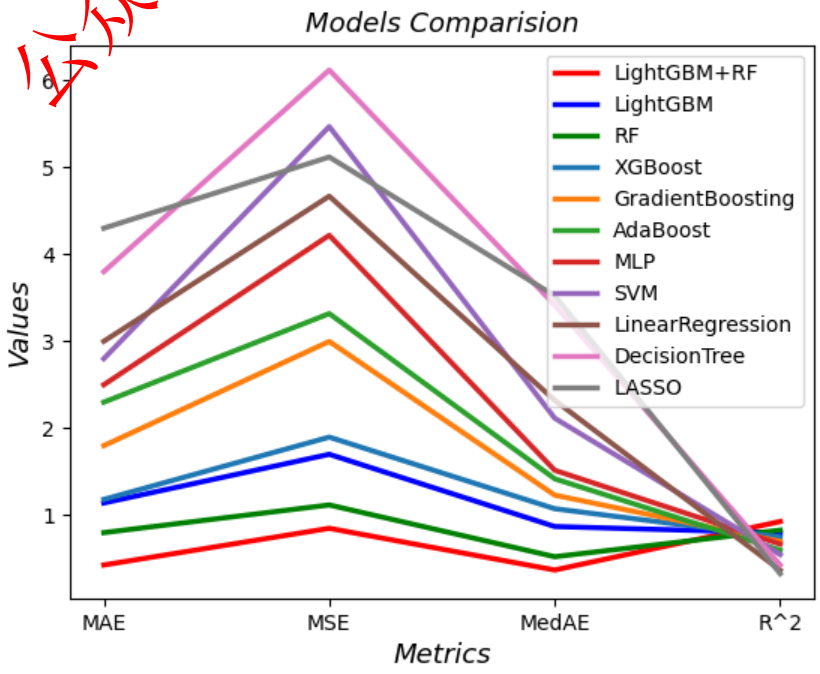

图 5.11 **11** 个回归预测模型对比分析可视化

从实验结果可以看出，LightGBM 和随机森林 RF 的回归预测效果优于其他回归器。LightGBM+RF 在四个指标上的表现均最好，因此本文后续选择采用**随机森林+LightGBM** 组合预测模型进行土壤湿度预测。

## 5.7 参数调优与求解

为了进一步提高 RF+LightGBM 回归预测模型的表现，对模型参数的调整十分必要，本文将采用 GridSearchCV (网格搜索)算法对 LightGBM 模型中 max_depth、learning_rate 等参数进行调优，对 RF 模型中 n_estimators、 max_features 等参数进行调优。所有调优都是依据 sklearn 包中 GridSearchCV（）函数的默认 scoring 评价指标进行 5 折交叉验证的平均结果。

LightGBM 调优：

LightGBM 初始参数设置如表 5.8 所示。

表 5.8 回归模型初始参数设置

| 参数名 | 含义 | 初始参数设置 |
|---|---|---|
| num_ iterations | 模型迭代次数 | 200 |
| learning_rate | 模型每次迭代产生的模型的权重 | 0.1 |
| min_ data_ in_leaf | 叶子节点可能具有的最小记录数 | 40 |
| bagging_fraction | 模型每次迭代所使用的数据比例 | 0.6 |
| max_depth | 树的最大深度，主要是为了防止模型过拟合 | 7 |
| feature_fraction | 模型每次迭代时使用的特征比例 | 0.8 |
| max_bin | 表示将特征存入的 bin 的最大数量 | 35 |

本文对对模型结果影响比较大的两个参数 max_depth 和 learning_rate 进行了较大跨度的的排列组合，调优结果如图 5.12 所示，从图中可以看出，理想的参数值分别为 **6** 和 **0.14**，随后将模型中对应的参数设置为最优解。

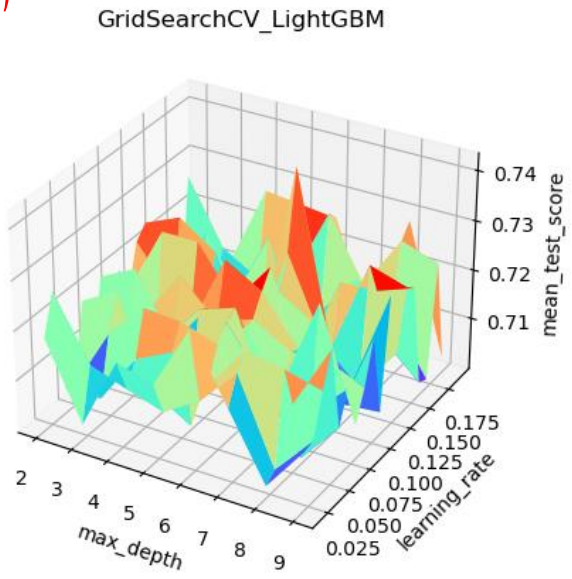


图 5.12 LightGBM 的 max_depth 和 learning_rate 参数调优结果

随机森林调优：

同样，本文也对对随机森林结果影响比较大的两个参数 n_estimators 和 max_features 进行了较大跨度的的排列组合，调优结果如图 5.13 所示，从图中可以看出，理想的参数值分别为 **80** 和 **3**，随后将模型中对应的参数设置为最优解。

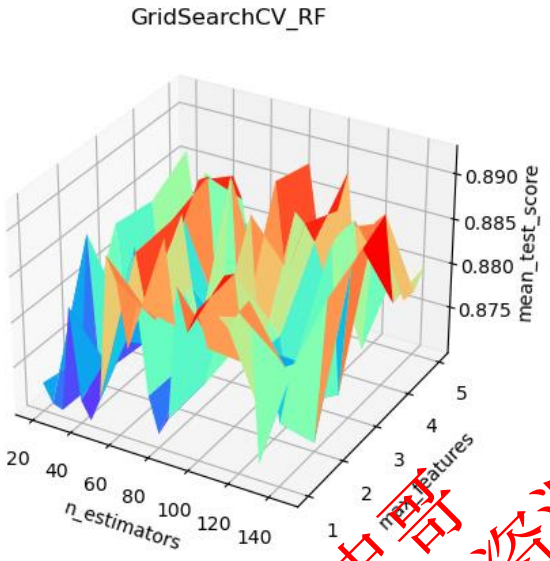


图 5.13 随机森林的 n_estimators 和 max_features 参数调优结果

将分别调优后的参数代入随机森林+LightGBM 的组合预测模型再次进行 5 折交叉验证，精度结果如图 5.14 所示。

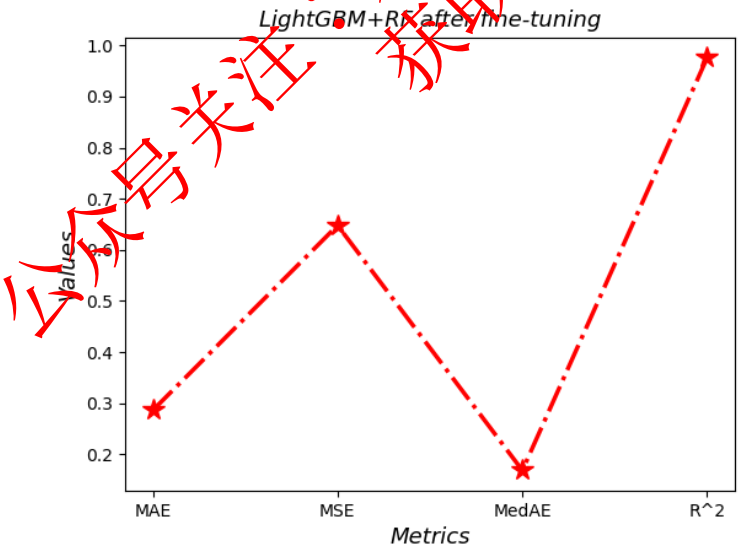


图 5.14 调参后模型验证结果

从上图可以看出，调优后的 LightGBM+RF 组合预测模型，精度效果相比较达到了最优值，具体如表 5.8 所示：

表 5.8 调优后的 LightGBM+RF 交叉验证精度

| MAE | MSE | MedAE | R^2 |
|---|---|---|---|
| 0.286 | 0.6479 | 0.1683 | 0.9759 |

将灰色模型预测出的 2022-2023 年的 5 个特征变量的值分别代入 4 个土壤不同深

度湿度(kg/m2)的**随机森林+LightGBM 组合预测模型**，即可得到 10cm 湿度(kg/m2)、40cm 湿度(kg/m2)、100cm 湿度(kg/m2)和 200cm 湿度(kg/m2)在 2022-2023 年的最终预测结果，如表 5.9 所示。

表 5.9 2022-2023 年不同深度土壤湿度**预测结果**

| 年份 | 月份 | 10cm 湿度 (kg/m2) | 40cm 湿度 (kg/m2) | 100cm 湿度 (kg/m2) | 200cm 湿度 (kg/m2) |
|---|---|---|---|---|---|
| 2022 | 04 | 13.4 | 38.9 | 93.46 | 164.48 |
| | 05 | 14.85 | 37.49 | 90.45 | 164.48 |
| | 06 | 18.21 | 42.95 | 87.22 | 164.45 |
| | 07 | 17.05 | 44.82 | 89.25 | 164.22 |
| | 08 | 20.03 | 54.43 | 95.32 | 164.03 |
| | 09 | 19.51 | 51.5 | 99.26 | 163.68 |
| | 10 | 15.43 | 44.74 | 100.28 | 163.42 |
| | 11 | 12.83 | 45.63 | 101.22 | 162.18 |
| | 12 | 11.98 | 46.22 | 102.33 | 162.03 |
| 2023 | 01 | 13.62 | 42.07 | 102.39 | 162.03 |
| | 02 | 12.02 | 45.61 | 102.24 | 162.03 |
| | 03 | 12.74 | 43.01 | 102.36 | 162.03 |
| | 04 | 13.12 | 37.91 | 101.98 | 162.03 |
| | 05 | 15.32 | 37.73 | 98.27 | 161.55 |
| | 06 | 16.52 | 39.38 | 96.17 | 161.22 |
| | 07 | 19.79 | 54.65 | 95.28 | 161.03 |
| | 08 | 19.79 | 54.65 | 104.17 | 160.63 |
| | 09 | 19.88 | 51.73 | 107.97 | 159.03 |
| | 10 | 15.82 | 44.26 | 114.87 | 158.93 |
| | 11 | 12.84 | 46.4 | 114.47 | 158.53 |
| | 12 | 12.5 | 46.93 | 114.97 | 158.48 |

# 6. 问题三：模型的建立与求解

## 6.1 问题分析

问题三要求本文从机理分析角度建立不同放牧策略对锡林郭勒草原土壤化学性质影响的数学模型。并对土壤中碳氮数据进行预测，拟从以下三个步骤解决问题三：（1）在化学循环的基础上，通过建立 SOC、SIC 的微分方程，计算出 2022 年 SOC、SIC 的预测量。（2）以全 N 为因变量，土壤无机碳或土壤有机碳作自变量进行回归分析，计算出全 N 的预测量

问题三的总体思路如图 6.1 所示

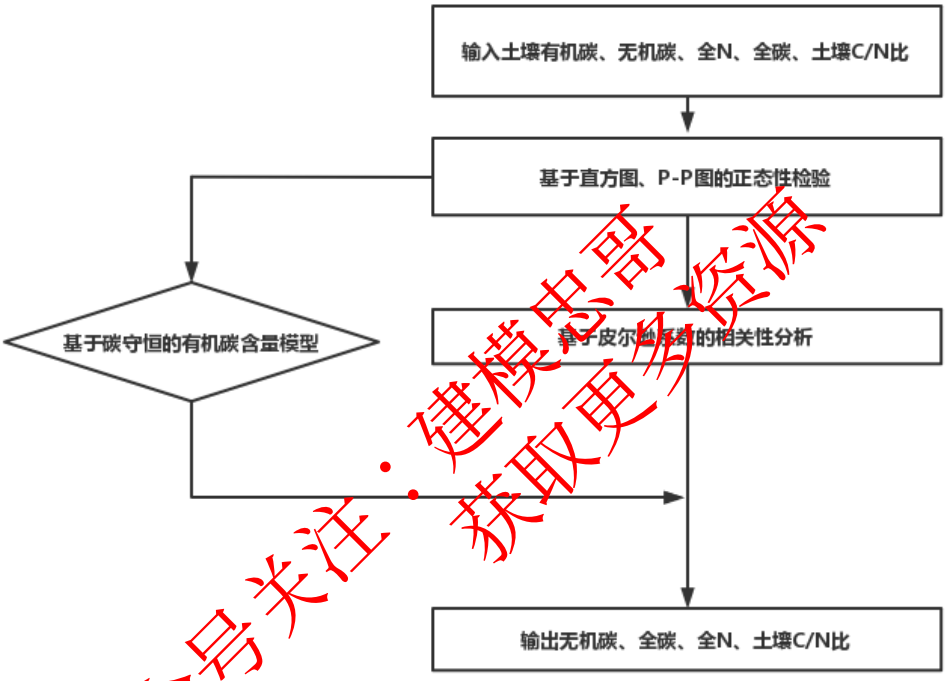


图 6.1 问题三的解题思路

## 6.2 基于碳守恒的有机碳含量模型土壤化学性质参数预测

根据题目分析可知土壤中的化学性质满足以下动态


图 6.2 土壤中化学循环

通过图 6.1 可知一段时间内土壤中的各化学性质满足质量守恒，具体分析如下：

（1）SOC 测量

对于 SOC 土壤中的有机碳，令土壤有机碳的含量为 X，则将 t 时刻土壤有机碳的含量记为 $X_t$，则存在

$$X_t=X_{t-1}+X_{现t}-X_{消t} \tag{31}$$

其中 $X_{现t}$ 表示从 t-1 时刻到 t 时刻土壤中有机碳的增长量，其中 $X_{消t}$ 表示从 t-1 时刻到 t 时刻土壤中有机碳的消耗量量，其中 $X_0$ 表示土壤中有机碳的初始量。

对于 $X_{消t}$，土壤中有机碳的主要作用是用于植被生长，而部分有机碳会受水土流失而随之丢失。本文在研究过程中，△t 的数值取值较小，在实际生活中因水土流失带来的土壤中有机碳的消耗可忽略不计，因此土壤中的有机碳的消耗量与植被数量相关。令植被吸收有机碳的系数为 $h_1$ 则

$$X_{消t}=hX_{tn}=h_1\sum_{i=2}^{i=n}\int_{t_{i-1}}^{t_i}v_i dt+W_0 \tag{32}$$

对于 $X_{消t}$ 而言，根据资料可知，主要由以下部分组成植被自然枯萎转化为土壤中的无机碳 $X_1$ 和放牧时深处排出的排泄物转化为土壤中的无机碳 $X_2$ 组成。令枯萎的植被吸收有机碳系数为 $h_2$，则：

$$X_1=W_{tn}*D_{tn}*h_2=W_{tn}*\beta_{tn}(e^{\varepsilon W_{tn}/W_{t-1}})h_2 \tag{33}$$

$D_{tn}$ 表示植被的枯萎率，对于放牧过程中。身处的日食量为点几吧。代谢比例为 u 排泄物中无机。碳的吸收系数为 h3，则

$$X_2=\mathrm{h}X_{tn}=\sum_{i=2}^{i=n}\int_{t_{i-1}}^{t_i}1.8uh_3dt \tag{34}$$

由式 36，37 可知

$$X_{现t}=X_1+X_2 \tag{35}$$

最终可得：

$$X_t=X_{t\text{-}1}+W_{tn}*\beta_{tn}(e^{\varepsilon W_{tn}/W_{t-1}})h_2+\sum_{i=2}^{i=n}\int_{t_{i-1}}^{t_i}1.8uh_3dt-\mathrm{h}_1\sum_{i=2}^{i=n}\int_{t_{i-1}}^{t_i}v_idt+W_0 \tag{36}$$

（2）SIC：无机碳测量

在自然状态下，土壤中各类化学元素呈现一定的相关性[8]，各类元素之间相关性的大小取决于土壤的性质，无机碳主要受土壤表面数量的影响；当植被数量增大时有利于防止土壤软化，无机盐的成分降低，反之则增高。为了挖掘锡林郭勒草原土壤中各类化学元素之间的关系，本文首先针对土壤有机碳进行正态性检验，如下图所示：

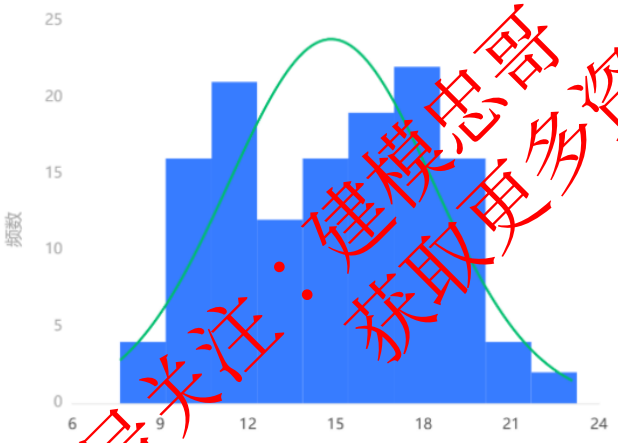


图 6.3 SOC 正态性检验直方图

土壤有机碳样本数量小于 5000，因此采用 S-W 检验，显著性 P 值为 0.015，其峰度-0.87 绝对值小于 10，并且偏度 0.032 绝对值小于 3，可以结合正态分布直方图、P-P 图进一步分析，其正态性检验 P-P 图如下所示：

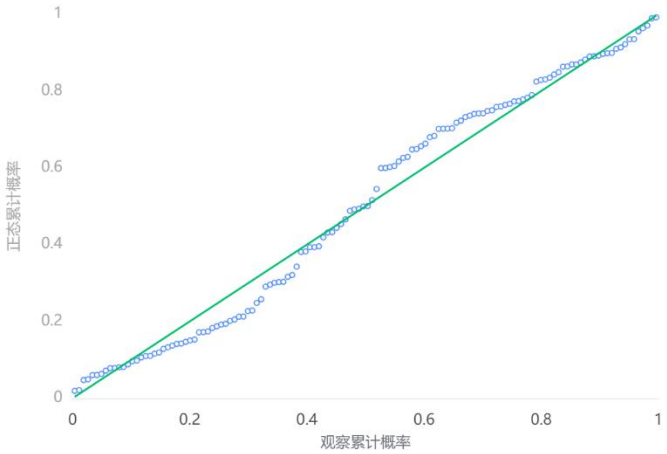

图 6.4 正态性检验 P-P 图

从上图可以发现，土壤有机碳计算观测的累计概率与正态累计概率的拟合情况较好，所以本文认为其近似服从正态分布，对于土壤无机碳、土壤全碳、全氮、土壤 C/N 比均做类似的处理。因为所有数据都近似服从正态分布，所以选用皮尔逊系数挖掘各个元素之间的相关性，其相关性如下图所示：


图 6.5 元素相关性图

从上图可以发现，土壤有机碳和无机碳之间存在负相关关系，呈现中等程度的相关性，因此对于土壤中无机碳的确定，即可根据土壤中有机碳含量确定，无机碳和有机碳之间的关系如下式：

$$SIC = 0.009 \times SOC^3 - 0.417 \times SOC^2 + 5.174 \times SOC - 8.153 \tag{37}$$

当已知土壤有机碳和土壤无机碳的含量后，最二者求和即可得到土壤中全碳的含量，即：

$$STC = SOC + SIC \tag{38}$$

（3）全氮含量的测定

文献[9]指出：土壤有机质和氮素的消长，主要决定于生物积累和分解作用的相对强弱以及气候、植被、耕作制度诸因素，特别是水热条件，对土壤有机质和氮素含量有显著的影响，有机质含量和土壤全氮之间的符合一次函数线性关系。本文研究主要研究内蒙古锡林郭勒草原，是十分具有代表性的温带草原，水热条件相对稳定，则可推断土壤全氮和土壤有机质之间的关系也呈现一次函数线性关系。

文献[10]指出，我国目前沿用的“van bemmelen 因数”，常设置为 1.724，即土壤中有机质 SOM 含量可以用土壤中的有机碳比例乘以有机碳百分数而求得，转换公式如下：

$$SOM = 1.724 \times SOC \tag{39}$$

从热力图中也可发现，土壤全氮含量与土壤有机碳之间的相关系数为 0.933，二者高度相关，充分应证了本文的推断，因此本文直接将全氮含量与土壤有机碳之间的含量设置为一次线性函数关系，即：

$$N = 0.117 \times SOC + 0.08 \tag{40}$$

全氮含量与有机碳之间的关系如下图所示，当已知土壤有机碳、无机碳和全氮含量以后，即可推算求得土壤 C/N 比。

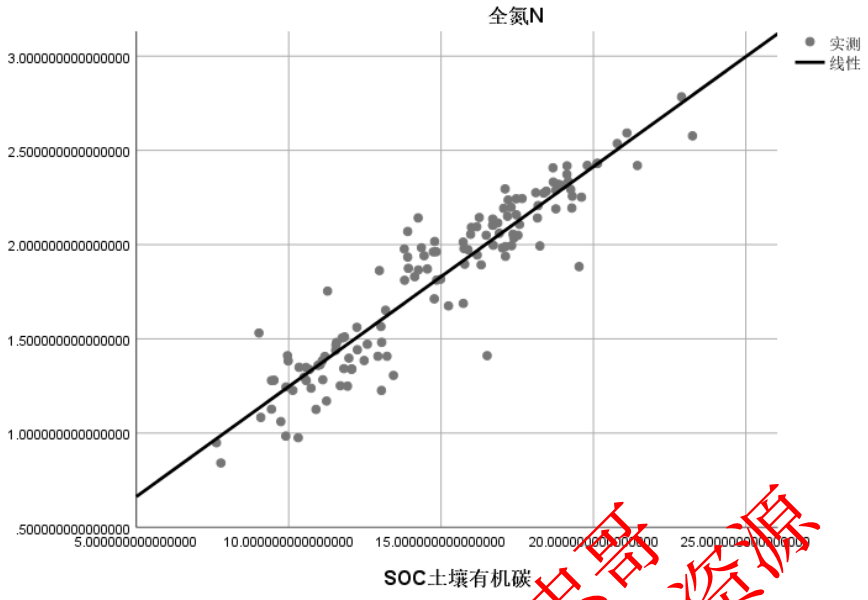


图 6.6 土壤 C/N 比

本文基于各个放牧小区不同的放牧强度下有机碳的预测结果，通过相关性分析求得无机碳和全氮含量，之后推算求得土壤全碳和土壤 C/N 比，具体结果如下表所示：

表 6.1 土壤化学性质预测图

| 放牧强度 | Plot 放牧小区 | SOC 土壤有机碳 | SIC 土壤无机碳 | STC 土壤全碳 | 全 N | 土壤 C/N 比 |
|---|---|---|---|---|---|---|
| NG | G17 | 14.07 | 7.162112 | 21.23211 | 1.72619 | 12.29999 |
| | G19 | 15.01 | 5.994389 | 21.00439 | 1.83617 | 11.43924 |
| | G21 | 15.11 | 5.868157 | 20.97816 | 1.84787 | 11.35262 |
| LGI | G6 | 14.52 | 6.60852 | 21.12852 | 1.77884 | 11.8777 |
| | G12 | 13.28 | 8.094623 | 21.37462 | 1.63376 | 13.08309 |
| | G18 | 14.26 | 6.929922 | 21.18992 | 1.74842 | 12.11947 |
| MGI | G8 | 14.09 | 7.137791 | 21.22779 | 1.72853 | 12.28083 |
| | G11 | 14.37 | 6.794417 | 21.16442 | 1.76129 | 12.01643 |
| | G16 | 14.94 | 6.082615 | 21.02261 | 1.82798 | 11.50046 |
| HGI | G9 | 13.05 | 8.353586 | 21.40359 | 1.60685 | 13.32021 |
| | G13 | 14.32 | 6.856101 | 21.1761 | 1.75544 | 12.06313 |
| | G20 | 14.01 | 7.234896 | 21.2449 | 1.71917 | 12.35765 |

# 7. 问题四：模型的建立与求解

## 7.1 问题分析

问题三要求本文根据题目中给出沙漠化程度指数预测模型及附件提供数据测算不同放牧强度下监测点的沙漠化程度指数值，并对土壤板结化定义定量给出。拟从以下三个步骤解决问题四：（1）采用**因子强度法、改进的主成分分析法**建立沙漠化程度指数评价模型。（2）对土壤板结化采用**熵权＋topsis** 进行综合评价，将评价结果使用 **K-means** 聚类得出不同土壤的板结化程度（3）基于改进的遗传算法求解多目标优化模型

问题四的总体思路如图 7.1 所示

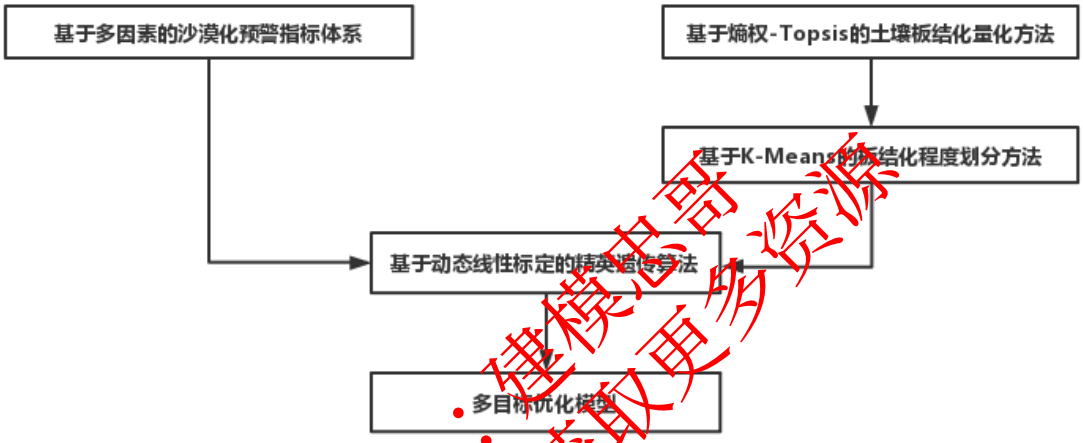


图 7.1 问题四的解题思路

## 7.2 沙漠化程度指数评价模型

本文针对研究区的沙漠化影响因素从多方面作了较为详细分析，并提取出能够代表其与沙漠化相关关系的指标因子，即气象因素、地表因素以及人文因素。气象因素包含的指标：平均风速(knots)、平均气温(℃)、降水量(mm)数据均由附件 8 直接获得。地表因素：植被覆盖度、土壤贮水量。其中植被覆盖度的缺失数据由拉格朗日插值法补全。相比于地表水资源量，本文采用的土壤贮水量能够从宏观角度更好的体现出土壤的水资源变化，且在问题一中已通过建立微分方程得出该项指标具体值。人为因素包括：人口密度、人均家庭经营纯收入、牲畜密度。其中人口密度及人均家庭经营纯收入数据由《中国统计年鉴》2013-2021 年直接搜集获取，牲畜密度的计算公式本文在问题一中已计算得出：时长因子×强度因子。从研究区的经济收入结构来看，研究区居民为兵团，其人均收入中除家庭经营收入以外，还包含了工资、财产性、转移性收入等。这些对沙漠化的影响极小，可忽略不计。沙漠化预警指标体系如图 7.2 所示：

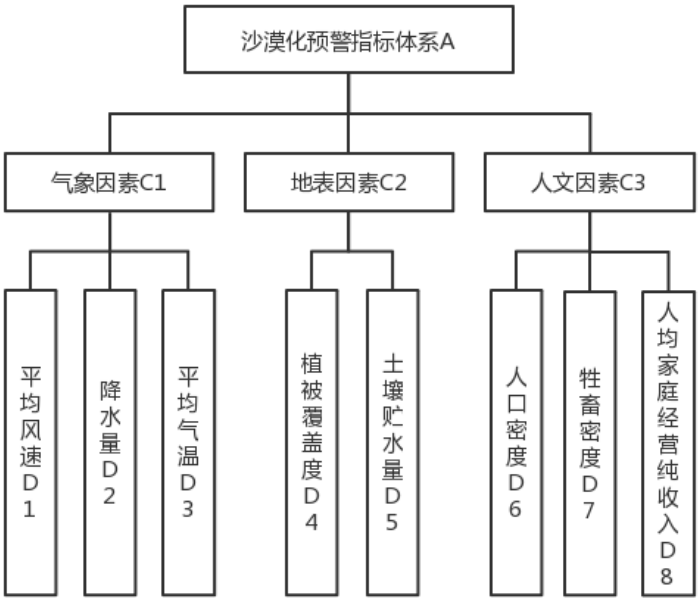


图 7.2 沙漠化预警指标体系

本文在原有模型的基础之上，完善并建立了本文的沙漠化预测模型表达式：

$$SM=\eta\cdot\sum_{i=1}^{n}S_{Q_i}=\eta\cdot(\sum_{i=1}^{n}(Q_i\cdot W_i)+D_f\times I) \tag{41}$$

其中 Qi 为因子强度，为避免人为主观因素的过多参与，本文对预测模型中的因子强度不再进行人为的分级，而是以因子对沙漠化影响的上限和下限为基准量化其因子强度。设因子 D 的实测值为 $X_i$~Xn,当 $D_i$ 的值为 $X_i$; ($X_1\leqslant X_i\leqslant X_n$)时，开始对沙漠化起作用；当 $D_i$ 的值为 $X_j$；($X_i\leqslant X_i\leqslant X_j\leqslant X_n$)时，对沙漠化起决定性作用，所以 $X_i$ 和 $X_j$ 分别为 $C_i$ 的下限和上限。当 $D_i\leqslant X_i$ 时，因子 $D_i$ 的因子强度 $Q_i$ 均取 0；当 $X_i<C_i<Xj$ 时，$Q_i=(C_i-X_i)/(X_j-X_i)$ $(0<Q_i<1)$；当 $C_i\geqslant X_j$ 时，$Q_i=1$。所以 $Q_i$ 的取值为[0，1]。

因子权重系数（$W_i$）由主成分分析法计算得出，并通过 CR 值取代 CI 来进行一致性检验，在 CR<0.1 的情况下，判断矩阵的一致性是可以接受的；如果没有，就必须修改判断矩阵。最后计算最底层各元素相对于最高层的权重值，结果如表 7.1 所示

表 7.1 预警指标体系因子权重系数

| | 权重系数 | 影响因素 | 权重系数 | 影响因子 | 权重系数 |
|---|---|---|---|---|---|
| 沙漠化预警指标体系 | 1 | 气象因素 | 0.3599 | 平均风速 D1 | 0.1802 |
| | | | | 降水量 D2 | 0.0787 |
| | | | | 平均气温 D3 | 0.1010 |
| | | 地表因素 | 0.4126 | 植被覆盖度 D4 | 0.2458 |
| | | | | 土壤贮水量 D5 | 0.1668 |
| | | 人文因素 | 0.2275 | 人口密度 D6 | 0.0882 |
| | | | | 牲畜密度 D7 | 0.0509 |

| | | |
|---|---|---|
| | 人均家庭经营纯收入 D8 | 0.0884 |

调节系数（η）调节系数的意义在于对预测模型的输出结果进行整体调整，从而得到满意的 SM 值，对其进行分级划分后使得沙漠化程度尽可能符合实际情况。本文利用 ArcGIS 软件实现对 1990 和 2000 年两个时期的数据进行沙漠化程度指数（SM）计算，SM 取值范围为 0～1。设定沙漠化程度及沙漠化程度指数划分标准表中定义临界值为 X，1990 年和 2000 年计算的临界值分别为 $X_1$ 和 $X_2$，调整临界值尽可能使得对应沙漠化程度与监测分布符合。基于 ENVI 和 ArcGIS，对 $X_1$、$X_2$ 进行尝试取值后，对沙漠化程度分布计算值与监测值进行栅格差值计算，对于任意像元的差值结果为两种：沙漠化程度相同时为 0，否则不为 0；统计两种像元的数量，差值为 0 的像元数占总像元数的百分比即为符合率。

当 $X_1$、$X_2$ 分别取表 7.2 中对应临界值时，两个时期中每一类型沙漠化土地的像元符合率均能够达到 90%，符合本研究的精度要求

通过对其进行分类，可以使荒漠化程度尽量接近于现实。采用 ArcGIS 软件，对 1990-2000 期间的荒漠化程度指标进行了分析，得到的 SM 指标在 0~1 之间。确定了荒漠化程度和荒漠化程度指标的指标，确定了 1990 和 2000 年的 X1、X2，通过调节阈值，尽量使相应的荒漠化与监测的分布相一致。在 ENVI 和 ArcGIS 的基础上，对 X1、X2 进行了试验，并用网格差对实测数据进行了分析，得出了两种不同的结果：当荒漠化程度一致时，结果为 0，将两个像元的数目进行统计，其差异为 0 的像元与全部像元的比例为相似性。

当 X1 和 X2 分别采用表 7.2 所示的相应阈值时，各阶段的像元符合率可 90%，满足了本文的精度要求。

表 7.2 沙漠化程度划分临界值

| X | 临界值 | | | | η 取值 |
|---|---|---|---|---|---|
| X（定义值） | 0.2 | 0.4 | 0.6 | 0.8 | |
| X1（1990 年） | 0.22 | 0.41 | 0.62 | 0.78 | $\eta_1$=1.0554x-0.0356，$R_2$ = 0.9974 |
| X2（2000 年） | 0.21 | 0.4 | 0.63 | 0.81 | $\eta_2$=0.9832x-0.0039，$R_2$ = 0.9979 |

建立 X1,X2 和 X 之间的数学关系，从而使它们的数值接近 X，从而获得相应的线性趋势线 η1，η2，也就是相应的调整因子 η。1990 年和 2000 年的输出结果 SM 值，基于 η1，η2 的数值，对其进行了全面的转换和标准等级分割（X 的定义值），获得了它们的荒漠化程度栅格分布图（附图 5-1 至 5-3)，并对数据进行了统计和计算，得出了两个阶段中的各种荒漠化类型的符合率都在 90%以上。在这些指标中，1990、2000 年时的 91.56%、90.88%、95.66%、95.66%。因此，本预测模式的调整因子 η 被设置为：η=0.9832x-0.0039。依据上文的模型，可以模拟出不同放牧强度下，沙漠化程度指数值，对于重合的区间范围，本文对相关数据求平均值，然后将平均值作为分割点。

| 放牧强度 | 沙漠化程度指数值 |
|---|---|
| 对照 | [0,0.12]] |
| 轻度放牧强度 | (0.12,0.18] |
| 中度放牧强度 | (0.18,0.26] |
| 重度放牧强度 | (0.26,0.38] |

### 7.3 熵权-Topsis 评价模型

土壤板结化与土壤有机物、土壤湿度和土壤的容重有关，目前还没有明确的定量表达式，其数学模型可定性描述为如下：

$$B = f(W,C,O) \tag{42}$$

土壤湿度$W$越少，容重$C$越大，有机物含量$O$越低，土壤板结化程度$B$越严重。其中土壤湿度$W$包含了放牧方式对土壤板结化的影响，有机物含量$O$用附件 14 中有机碳来表征。土壤容重则由土壤性质决定，变化不大。土壤容重的计算公式如下：

$$pb = Ms/Vt = Ms/(Vs+Vw+Va) \tag{43}$$

其中$Ms$为一定重量土壤烘干后的重量，$(Vs+Vw+Va)$为相同容积下水的重量。

由于土壤容重取决于土壤质地、粒间孔隙、有面质含量等因素[11]，属于土壤的基本物理性质，主要取决于土壤的类型，对于附件 7 中提供的容重为常数，难以表征整个草原上土壤容重的变化，本文以附件 7 中提供的常数为期望，以方差为标准差的正太分布生成不同的容重指数，选取集中的中间部分表征草原不同地区的容重。

（1）评价指标的选取

土壤板结化主要原因为：土壤质地粘重、耕作层浅、有机物质投入少、塑料废弃物污染、长期单一施用化肥、镇压、翻耕等农业生产措施造成土壤结构破坏、有害物质积累、暴雨造成水土流失。因此在选取土壤湿度、容重以及有机物含量作为土壤板结化的影响因素。容重题目中所给值为固定值，因此本文以附件 7 中提供的容重为期望值，方差为标准差构造不同地区土壤的容重。由于有机物含量的数据难以获取，本文以土壤有机碳、无机碳、全氮、土壤 C/N 比、4 项指标对土壤有机质共同表征。

表 7.3 土壤板结化影响指标

| 类型 | 指标 |
|---|---|
| 土壤湿度 | W |
| 容重 | C |
| 土壤有机碳 | M1 |
| 无机碳 | M2 |
| 全氮 | M3 |
| 土壤 C/N 比 | M4 |

（2）熵值法＋Topsis 评价

1）指标的权重计算

首先对原始数据进行无量纲化。为提高数据的可比性，采用正、反两种方法对离差进行规范化，得到的指数在[0,1]范围内，具体的公式是：

正向指标：

$$x'_{ij} = \frac{x_{ij} - x_{ijmin}}{x_{ijmax} - x_{ijmin}} \tag{44}$$

逆向指标：

$$x'_{ij} = \frac{x_{ijmax} - x_{ij}}{x_{ijmax} - x_{ijmin}} \tag{45}$$

其中，$x'_{ij}$指第 i 个评价样本的第 j 个评价指标的标准化数据，$x_{ijmax}$和$x_{ijmin}$为同一年份第 i 个系统第 j 项指标的最大值和最小值。标准化后的指标矩阵如下：

$$P_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} \tag{46}$$

计算指标信息熵：

$$e_j = -\frac{1}{\ln(n)} \times \sum_{i=1}^{n} (P_{ij} \times \ln(P_{ij})) \tag{47}$$

计算指标权重：

$$d_j = 1 - e_j \tag{48}$$

计算差异系数：

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j} \tag{49}$$

2）指标的得分计算

向量标准化：将原始数据归一化，以消除量纲

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \tag{50}$$

构造加权矩阵

$$Z^* = \begin{bmatrix} z_{11} \cdot w_1 & \cdots & z_{1n} \cdot w_n \\ \vdots & \ddots & \vdots \\ z_{m1} \cdot w_1 & \cdots & z_{mn} \cdot w_n \end{bmatrix} \tag{51}$$

寻找最优、最劣方案

$$\begin{cases} z_{ij}^{*+} = \max\limits_{m.n}(z_1^{*+}, z_2^{*+}, \ldots, z_n^{*+}) \\ z_{ij}^{*-} = \min\limits_{m.n}(z_1^{*-}, z_2^{*-}, \ldots, z_n^{*-}) \end{cases} \tag{52}$$

最优、最劣距离

$$D_i^+ = \sqrt{\sum_j (z_{ij}^* - z_j^{*+})^2} \tag{53}$$

$$D_i^- = \sqrt{\sum_j (z_{ij}^* - z_j^{*-})^2} \tag{54}$$

得分计算

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{55}$$

最终得到各项指标的权重得分如下表所示

表 7.4 土壤板结化影响指标

| 类型 | 指标 | 权重得分 |
|---|---|---|
| 土壤湿度 | W | 0.394 |
| 容重 | C | 0.103 |
| 土壤有机碳 | M1 | 0.797 |
| 无机碳 | M2 | 0.625 |
| 全氮 | M3 | 0.742 |
| 土壤 C/N 比 | M4 | 0.228 |

$$B = 0.394W + 0.103C + 0.797M_1 + 0.625M_2 + 0.742M_3 + 0.228M_4 \tag{56}$$

将归一化的指标代入到上升进行计算，从样本点中随机选择 3 个点作为初始簇中心，采用 K-means 算法对所有点聚类。经过转化以后，最终聚类结果如图 7.6 及表 7.5 所示。

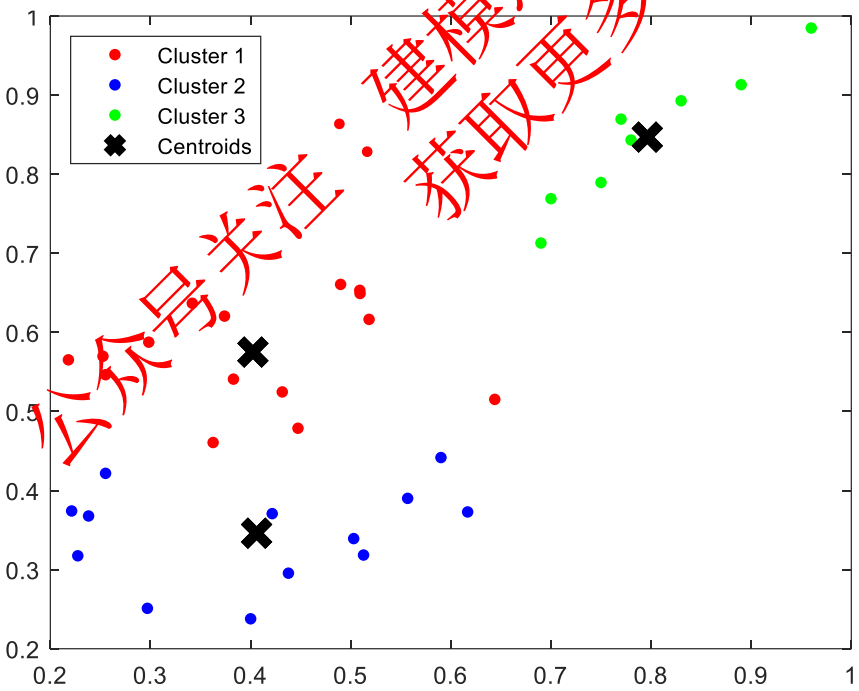




图 7.6 土壤板结化聚类结果

表 7.5 板结化程度分类

| 土壤板结化程度 | 划分类型 |
|---|---|
| 非、轻度板结化 | [0,0.78] |
| 中度板结化 | (0.78,0.83] |

| | |
|---|---|
| 重度板结化 | (0.83,1.00] |

### 7.4 多目标优化

为了保证锡林郭勒草原在减少沙尘暴和恶劣天气的发生方面的优势，本文目标函数包含两个：沙漠化程度最小和土壤板结化程度最小，如下：

$$\begin{cases} \min \quad SM=\eta\cdot\sum_{i=1}^{n}S_{Q_i}=\eta\cdot(\sum_{i=1}^{n}(Q_i\cdot W_{c_i})+D_f\times I_f) \\ \min \quad B=0.394W+0.103C+0.797M_1+0.625M_2+0.742M_3+0.228M_4 \end{cases} \tag{57}$$

上式中时长因子矩形波采用占空比为 50%的矩形波表示选择性轮牧的情况，三角波函数的峰值为 1，二者震荡周期相同，由于时长因子中包含了 0-1 变量和连续变量，极大的增加了优化难度和计算量，本文在优化过程中，对时长因子设置为范围在 $[0,1]$ 的变量，最后算得时长因子在一定时间内的积分面积，选择三角波函数、矩形波函数与之误差最小的作为最终的放牧类型。不同的放牧策略直接影响了土壤湿度、土壤中的有机碳等多种因素的含量，依据问题 1 和问题 2 中涉及的公式进行推算，由于影响土壤有机碳的因素过多，为了减小误差，本文限制问题 3 中得到的有机碳含量与自然状态下有机碳含量 $M_5$ 共同决定，依据二者的平均值推断土壤无机碳、全氮等多种因素含量，相关变量的限制如下：

$$\begin{aligned} s.t.\ & D_f\in(0,1] \\ & I_f\in(0,14.4] \\ & M_5\in[7.63,23.25] \\ & M_z=\frac{M_1+M_5}{2} \\ & M_2=0.009\times M_1-0.417\times M_1^2+5.174\times M_1-8.153 \\ & M_3=0.117\times M_1+0.08 \end{aligned}$$

### 7.5 基于动态线性标定的精英遗传算法求解

遗传算法是基于自然界生物遗传变异、物竞天择适者生存的原理而提出的。对于某类特定问题，其 $X=[x_1,x_2,...,x_n]^T$ 称为决策变量，而 $f(X)$ 为目标函数，目标是在某可行解中找到满足约束条件的最优解，这类问题被称为优化问题，而遗传算法正适用于此类问题，但不同于其他方法，遗传算法的优点是可以进行全局优化搜索最优解，其效率得到极大提高。**精英遗传算法**是对原算法的经典改进方法之一，具体步骤就是在选择个体时，优先选出适应度值最大的前几个个体即“精英个体”，再在精英个体的基础上进行后续的交叉和变异操作。

GA 算法乃至所有的智能优化方法都有很高的灵活性，无论是在参数的选取，操作方法的选择以及改进方法的实现上都能得以体现。以 GA 为代表的智能优化算法大都有在解域上从广域搜索能力向局域搜索能力过渡的趋势，使得算法在初期尽可能地搜索潜在的最优解，后期逐渐收敛于最优解。基于这一思想，本文采取了**动态线性标定**这一改进策略来调节算法的选择压力，相关原理如下：

函数表达式： $f^{'}=a^{k}f+b^{k}$， $k$ 为迭代指标

最大化问题 $a^{k}=1$， $b^{k}=-f_{\min}{}^{k}+\xi^{k}$

函数表达式： $f^{'}=f-f_{\min}{}^{k}+\xi^{k}$

$\xi^{k}$ 的取值： $\xi^{0}=M$ ， $\xi^{k}=\xi^{k-1}\cdot r$ ， $r\in[0.9,\ 0.999]$

（调节 $M$ 和 $r$，从而来调节 $\xi^{k}$ ）

当采取动态线性标定对解的适应值进行标定后，在算法初期，种群间个体的适值差异有了一定的缩小，这样就使得当前劣的个体也有机会被选择进行后续的交叉和变异操作，从而生成可能的潜在最优解，增强了算法在初期的广域搜索能力；而随着迭代的进行， 逐渐减小，种群中个体间差异回升，选择压力加大，GA 算法的局域搜索能力开始显著，最终找到最优解。本题采取**基于动态线性标定的精英 GA 算法**对模型进行求解，从最终结果可以看出，在连续放牧、轻度放牧强度的条件下，沙漠化程度、板结化程度最小，土壤沙漠化程度和板结化程度寻优过程如下图 7.7，7.8 所示。

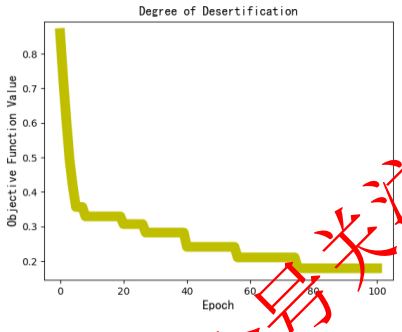


图 7.7 沙漠化程度迭代曲线

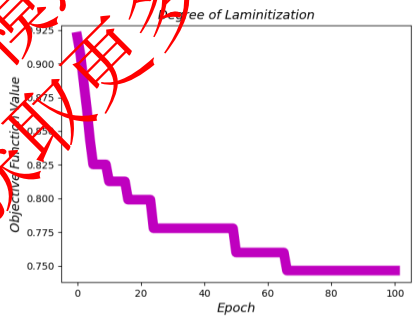


图 7.8 板结化程度迭代曲线

# 8. 问题五：模型的建立与求解

## 8.1 问题分析

问题五要求本文根据题目中给出沙漠化程度指数预测模型及附件提供数据测算不同放牧强度下监测点的沙漠化程度指数值，并对土壤板结化定义定量给出。拟从以下三个步骤解决问题四：（1）采用**因子强度法、改进的主成分分析法**建立沙漠化程度指数评价模型。（2）对土壤板结化采用**熵权＋topsis** 进行综合评价，将评价结果使用 **K-means** 聚类得出不同土壤的板结化程度（3）采用基于重升温策略的模拟退火算法（对算法引入随机性，避免算法陷入局部最优）对不同降水量下羊群数量最大阈值进行求解。

问题五的总体思路如图 8.1 所示：

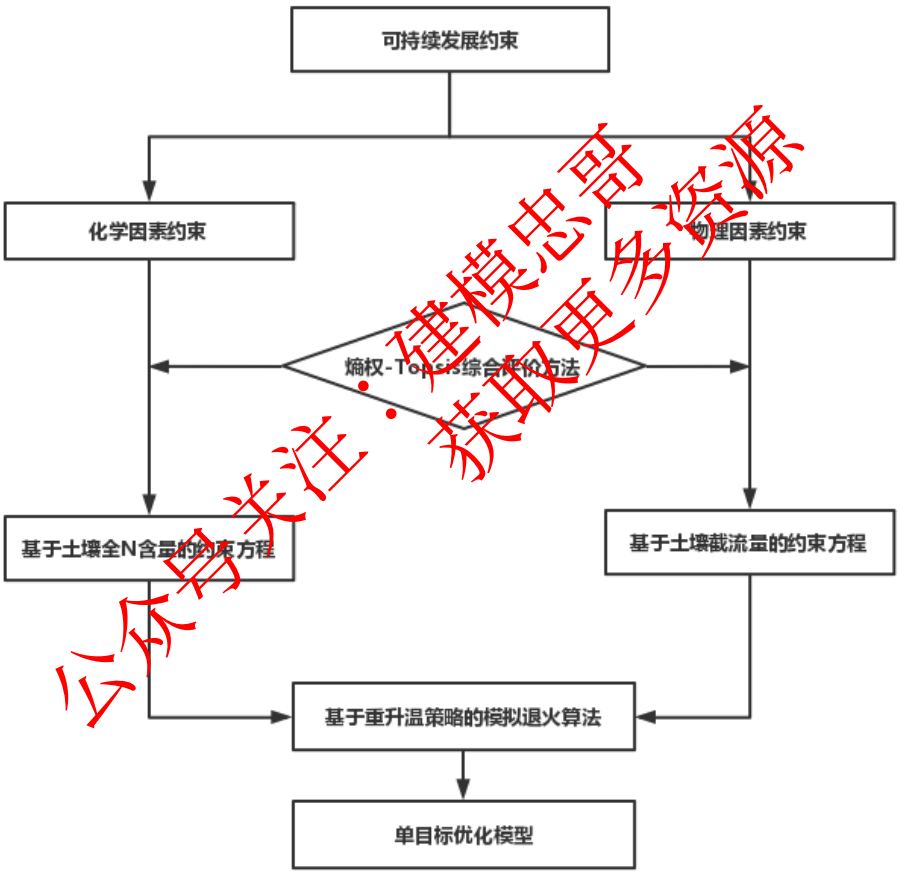


图 8.1 问题五解题思路

## 8.2 基于重升温策略的模拟退火算法求解羊群数量阈值

1.目标函数，令羊的单价为 a，数量为 S，则通过牧羊所取得的中利润为 P

$$P = a \times S \tag{58}$$

其中：a 值根据现有市场价格进行取值[12]，为 1000 元/只

2.约束条件

为了防止草原被破坏，同时也为了实现经济效益最大，需要探索土壤各类化学因素、植被等因素的承载极限，既能维持草原可持续发展，也能实现牧民、牧场经济效益最大，本文研究的可持续发展的限制条件主要为化学因素和植被因素。

（1）化学因素的约束

一方面，在放牧过程中，往往由于牲畜密度过大，可能导致草原植被结构破坏，土壤裸露面积增大，促进了土壤表面的蒸发，土体内水分相对运动受到不利影响，破坏了土壤积盐与脱盐平衡，增加了盐分在土壤表面的积累，土壤盐碱化程度加重，造成了土壤无机物含量浓度发生变化。

另一方面，放牧时由于由于家畜的采食践踏造成枯落物分解，充分进入土壤，从而提高土壤有机质和氮和钾含量，减少土壤的板结，促进了土壤有机物含量的增加。

关于放牧对化学因素的影响，题目中指出：土壤全氮含量随着放牧强度的增加而降低。有研究表明：高寒草甸的土壤全氮含量沿着放牧梯度呈下降趋势。因此，为了保证土壤达到合适的状态，找到放牧羊（标准羊）数量的阈值是问题的关键，可见，全氮含量与降水量关系不大，本文选择土壤含氮量作为衡量土壤化学因素对放牧羊数量制约的主要指标。

文献[13]中指出不同类别土壤中全氮含量不同，耕地土壤全氮量大约为每千克 0.4-3.8g，，平均全氮含量时每千克 1.3g，在自然界中植被未受损的土壤全氮含量一般为每千克 0.4-7.5g，平均每千克 2.9g，本文的研究对象大多处于天然状态，耕种面积较小，因此本文选择[0.4,7.5]作为约束范围。

由于附件 15 中放牧小区的初始条件类似，由于轻度、中度、重度放牧强度对于土壤含氮量的影响具有随机性，所以相同情况下做了三组对照实验，同时由于每年土壤含氮量会有明显的差异性，基于此考虑，为了使模型充分计及牧群的随机影响，本文选择以年为单位，因为忽略了放牧方式，所有羊的时长因子为 1，以 0、2、4、8 反映不同的放牧强度，但是由于本题不限制变量结果的类型，即可以为小数，因此本文采用随机数生成的方式，设置 4 个期望为 0、2、4、8 且方差为标准差的正态分布随机数为变量，采用三次多项式对捕捉放牧强度 GI 和土壤全氮量之间的关系，如下表所示：

表 8.1 各年份函数方程

| 年份 | 公式 |
|---|---|
| 2012 | $N_1 = -0.032GI^3 + 0.197GI + 1.529$ |
| 2014 | $N_2 = -0.041GI^3 + 0.066GI^2 + 1.674$ |
| 2016 | $N_3 = -0.032GI^3 + 0.058GI^2 + 1.824$ |
| 2018 | $N_4 = 0.006GI^3 - 1.112GI + 2.058$ |
| 2020 | $N_5 = 0.017GI^3 - 0.193GI + 2.248$ |

由于数据量很少，每个方程都具有片面性和局限性，因此使用上文中使用的熵权-Topsis 综合评价方法，对每个方程附加权重系数 m，如下式所示：

$$N = m_1N_1 + m_2N_2 + m_3N_3 + m_4N_4 \tag{59}$$

（2）植被因素的约束

植被因素对羊数量的约束关系主要体现在羊食用植被、践踏植被等方面，对于植被生物量是减少的作用。

值得注意的是，题目中指出：草地的植被直接决定放牧的强度，而植被的截流量能最好反映植被的生长能力，依照递推关系，放牧强度与植被的截流量存在正相关关系。植被截流量与降水量、植被覆盖度、叶面积指数（LAI）等密切相关，如式（6）中 $R_{cum}$ 为累积降雨量（mm）表达了降雨量对植被生物量生长的影响，因此本文选择植被截流量作为主要指标，表征植被因素对羊数量的约束。

由于羊的标准食量是 1.8 千克的标准干草，所以本文选择附件 14 中植物干重作为研究对象，对所有类型的植物的干重求和，以年为单位，利用三次多项式捕捉降雨量、放牧强度对植被截流量的影响，与全氮含量的处理方式类似，利用上文的熵权-Topsis 综合评价方法赋予每个方程权重，对于缺少的数据，本文选择历史数据中相似时间、相似经纬度的数据作为补充。

（3）单目标优化模型与求解

综上所述，为了在可持续发展的限制条件下，实现经济效益的最大化，本文建立了针对标准羊数量 S 的单目标优化模型，如下所示：

$$\max \mathrm{P} = a \times S$$

$$s.t.$$

$$\begin{cases} \mathrm{BIO} = a_0 \cdot c_p \cdot IC_{\max} \cdot \left[1 - \exp\left(-k \, R_{cum} / IC_{\max}\right)\right] \\ \mathrm{BIO}_0 + \sum_{i=1}^{n} (BIO - D_f \times X_f) \geq \mathrm{BIO}_g \\ N \in [0.4, 7.5] \\ S > 0 \\ a_0 \in [0.5, 3] \end{cases}$$

上式中 BIO 表示表示一定时间内，在规定的时间段内、不同降雨量情况下植被生物量的生长量；$a_0$ 表示植被截流量和植被生物量之间的转换系数，本文设置其范围为[0.5,3]，为了保证解能够实现可持续发展，植被生物量在经过多次不断消耗和增加后，采用**基于重升温策略的模拟退火算法（对算法引入随机性，避免算法陷入局部最优）**对不同降水量下羊群数量最大阈值进行求解。SA 算法基本步骤如下，以降水量为 300nm 的情况为例，退火寻优过程如图 8.2 所示。最终的求解结果如表 8.2 所示。

给定初温 $t=t_0$，随机产生初始状态 $s=s_0$，令 k=0 ;
Repeat
　Repeat
　　产生新状态 $s_j$=Genete(s) ;
　　if min{1,exp[-(C($s_j$)-C(s))$t_k$]}≥randrom[0,1] s=$s_j$;
　Until 抽样稳定准则满足;
　退温 $t_{k+1}$=update($t_k$)并令 k=k+1;
Until 算法终止准则满足;
输出算法搜索结果。

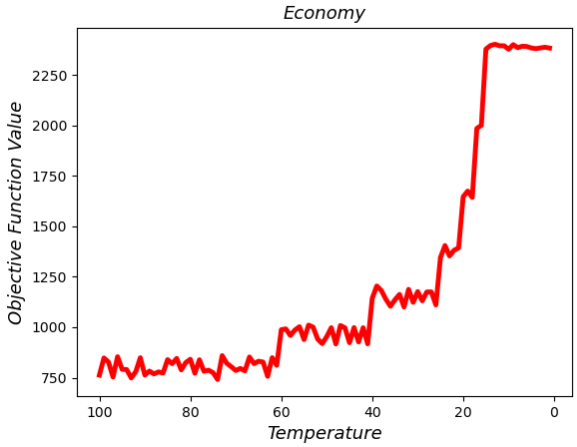


图 8.2 退火寻优过程可视化

表 8.2 不同降水量下羊群数量最大阈值

| 降水量 | 最大阈值（羊/天/公顷 ） |
|---|---|
| 300mm | 2.2 |
| 600mm | 3.1 |
| 900mm | 5.8 |
| 1200mm | 7.3 |

由于温带草原降水量在 400mm 以下，在附近的 300mm 降水量和 600mm 降水量时，最大阈值的变化为 1 左右，但是当降雨量达到 900mm 和 1200mm 时，最大阈值的变化接近为 2，说明保持在这样的降水状态下，草原的性质发生了巨大的变化。

# 9. 问题六：模型的建立与求解

## 9.1 问题分析

问题六的主要任务是先利用附件 13 不同示范牧户牲畜数量调查数据集中的已知信息量化 4 位牧户的不同放牧强度，结合附件 12 给出的 4 个示范地不同年份的植被生物量信息，然后利用前面问题 1 和问题 3 建立的不同放牧策略对锡林郭勒草原土壤的物理性质、化学性质和植被生物量影响的机理模型，构建模型预测出 4 位示范牧户的草场的土地状态，包括土壤肥力变化、土壤湿度、植被覆盖等。

由于土壤肥力受土壤物理性质和化学性质的影响，本文基于前面的机理模型建模分别预测 2023 年 9 月 4 个示范地的 100cm 湿度(kg/m2)，SOC 土壤有机碳、SIC 土壤无机碳、全氮 N、植被指数(NDVI)和径流量(m3/s)6 个指标，并可视化了 4 个示范地 2020 年-2023 年四年间 9 月份在不同指标上的对比。

## 9.2 示范牧户放牧强度的量化求解

对附件 13 中给出的 4 位示范牧户放羊压力按**羊/天/公顷**统一量纲后，取 2018-2020 三年间的平均值，按照对照（NG， 0 羊/天/公顷 ）、轻度放牧强度（LGI， 2 羊/天/公顷 ）、中度放牧强度（MGI，4 羊/天/公顷 ）和重度放牧强度（HGI，8 羊/天/公顷 ）的评价标准划分不同放牧强度等级，得到如表 9.1 结果：

表 9.1 4 位示范牧户放牧强度等级对比

| | 放牧压力 | 放牧强度 |
|---|---|---|
| 牧户 1 | 2.6815 | 介于 LGI~MGI 之间 |
| 牧户 2 | 2.1333 | 介于 LGI~MGI 之间 |
| 牧户 3 | 5.3817 | 介于 MGI~HGI 之间 |
| 牧户 4 | 2.4917 | 介于 LGI~MGI 之间 |

从表中可以看出，牧户 3 的放牧强度约高出其他牧户一个放牧强度等级。再对附件 12 给出的 4 个示范地的植被生物量按 2018-2020 三年取平均值，得到如表 9.2 结果。

表 9.2　4 个示范区平均植被生物量对比

| | 平均植被生物量 |
|---|---|
| 示范区 1 | 81.2 |
| 示范区 2 | 108.71 |
| 示范区 3 | 74.62 |
| 示范区 4 | 88.6 |

从表中可以看出，示范地 3 的平均值被生物量少于其他示范地，示范地 2 的植被量最多。由于在本题中假设保持示范牧户放牧策略不变，4 位牧户的放牧方式均为**生长季休牧**的，可将以上整理出的因子代入构建的模型用于各个土地状态指标的预测。

## 9.3 模型构建与预测结果对比分析

基于前面如下建立的预测土壤物理性质（土壤湿度）和化学性质的机理模型，构建分别预测 4 个示范地的 100cm 湿度(kg/m2)，SOC 土壤有机碳、SIC 土壤无机碳、全氮 N、植被指数(NDVI)和径流量(m3/s)共 6 个土壤状态指标，针对每个状态指标，分别给出了 4 个示范地在 2020 年 9 月、2021 年 9 月、2022 年 9 月和 2023 年 9 月时期的预测指标值，以便于对比分析。

$$IC_{store} = c_p \cdot IC_{\max} \cdot \left[1 - \exp\left(-k \cdot R_{cum} / IC_{\max}\right)\right]$$

$$W_{t_{n+1}} = P + p \times h \times 10 \times \sum_{i=1}^{i=n} \int_{t_{i-1}}^{t_i} (P - E(\alpha))dt - (Et_a + G_d + IC_{store})$$

$$\frac{d\beta}{dt} = \frac{P - (Et_a + G_d + IC_{store})}{p \times h \times 10} - D_f \times I_f \times W$$

$$X_t = X_{t-1} + W_{tn} * \beta_{tn}(e^{\varepsilon W_{tn}/W_{t-1}})h_2 + \sum_{i=2}^{i=n} \int_{t_{i-1}}^{t_i} 1.8uh_3 dt - \mathrm{h}_1 \sum_{i=2}^{i=n} \int_{t_{i-1}}^{t_i} v_i dt + W_0$$

$$SOM = 1.724 \times SOC$$

$$N = 0.117 \times SOC + 0.08$$

土壤 100cm 湿度(kg/m2)，SOC 土壤有机碳、SIC 土壤无机碳、全氮 N、植被指数(NDVI)和径流量(m3/s)的预测结果如图 9.1-9.6 所示：

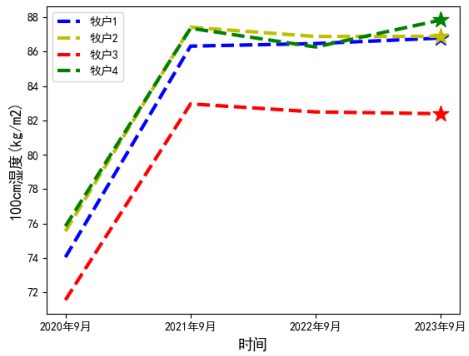


图 9.1 100cm 土壤湿度

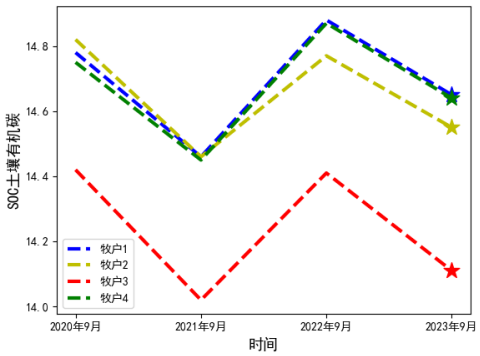


图 9.2 SOC 土壤有机碳

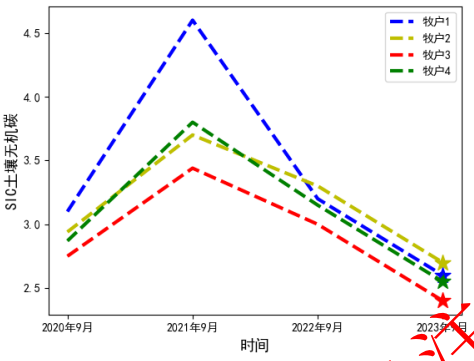


图 9.3 SIC 土壤无机碳

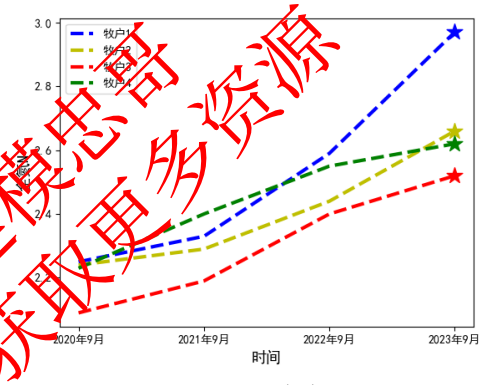


图 9.4 全氮 N

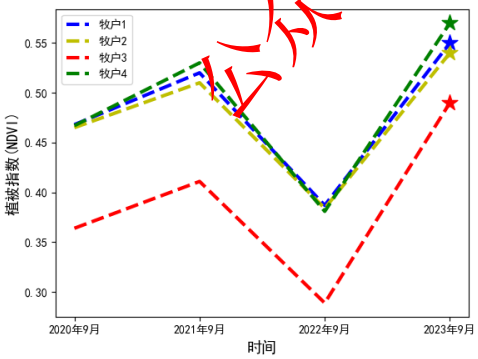


图 9.5 植被指数

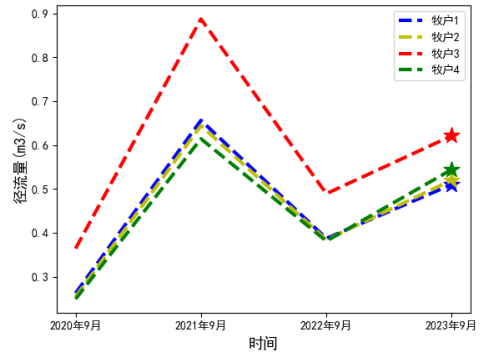


图 9.6 径流量

从以上图中可以对比分析出，牧户 3 的示范地主要由于相较于其他示范地较大的放牧强度（介于 MGI~HGI 之间），直接影响了其植被指数，较小的植被指数也间接影响了其土壤物理性质的湿度较小。同时，由于放牧强度大，示范地 3 的 SOC 土壤有机碳、SIC 土壤无机碳和全氮 N 的土壤化学性质指标也较小。

植被的减少削减了草场上植物对降水的拦截作用，增大了雨滴对地表的溅蚀和地表径流的形成，同时也增大了水土流失量。从上图 9.6 的对比中也可以看出，更大放牧强度的牧户 3 的示范地的径流量也高于其他 3 个示范地。

### 9.4 2023 年 9 月示范区土地状态预测结果展示

下表 9.3 给出了 4 个示范区在 2023 年 9 月的 6 个土地状态指标预测结果。

表 9.3　4 个示范区 2023 年 9 月土地状态预测结果汇总

| | 100cm 湿度(kg/m2) | SOC 土壤有机碳 | SIC 土壤无机碳 | 全氮 N | 植被指数(NDVI) | 径流量(m3/s) |
|---|---|---|---|---|---|---|
| 示范区 1 | 86.79 | 14.65 | 2.61 | 2.97 | 0.55 | 0.51 |
| 示范区 2 | 86.89 | 14.55 | 2.72 | 2.66 | 0.54 | 0.52 |
| 示范区 3 | 82.39 | 14.11 | 2.41 | 2.52 | 0.49 | 0.62 |
| 示范区 4 | 87.84 | 14.64 | 2.55 | 2.62 | 0.57 | 0.54 |

# 10. 模型的评价、改进与推广

## 10.1 模型的优点

1. 模型具有现实物理意义，具有很强的可解释性。
2. 模型采用熵值+Topsis 的模型进行评价，消除了主观因素可能对模型结果造成的影响。
3. 问题二中，土壤不同深度的湿度预测模型的选择与建立是在**超过 10 个**回归预测模型的 5 折交叉验证精度对比基础上的，具有相当的说服力。

## 10.2 模型的改进

缺少对极端天气气候、垃圾污染等因素对草原放牧策略的选择。

## 10.3 模型的推广

在现有模型的基础上，增加对极端气候的评价因子使得模型的应用场景更广泛。

# 11. 参考文献

[1]李保国，龚元石，左强. 农田土壤水的动态模型及应用[M]. 科学出版社, 2000.

[2]侯琼,王英舜,杨泽龙,师桂花.基于水分平衡原理的内蒙古典型草原土壤水动态模型研究[J].干旱地区农业研究,2011,29(05):197-203.

[3]陈皓锐,黄介生,伍靖伟,杨金忠.冬小麦根层土壤水量平衡的系统动力学模型[J].农业工程学报,2010,26(10):21-28.

[4]王悦骅,王占海,沈婷婷,王忠武.浅谈降水变化对植物多样性和生产力的影响[J].草原与草业,2022,34(02):15-19.

[5]赵冰茹,刘闯,王晶杰,陈文波.锡林郭勒草地 MODIS 植被指数时空变化研究[J].中国草地,2004(01):2-9.

[6] 侯琼,王英舜,杨泽龙,等.基于水分平衡原理的内蒙古典型草原土壤水动态模型研究[J].干旱地区农业研究,2011,29(05):197-203.

[7]NY/T 635-2015 天然草地合理载畜量的计算标准.

[8]曾水泉.温州市土壤化学元素含量特点[J].中山大学学报(自然科学版),1996(S1):215-220.

[9]张启新,李洁.土壤有机质与全氮相关关系分析[J].谷,2010(16):122+162.

[10]林俊,杨红,王春峰.长江口沉积物有机碳及溶解有机质结构对外源磷输入的响应[J/OL].上海海洋大学学报:1-10[2022-10-09].

[11]李猛,李海瑜,高明.保护性耕作对黑土不同土层土壤固氮菌丰度和群落结构的影响[J].土壤与作物,2022,11(03):273-281.

[12]梁亚俊,邢娜,贺俊平.羊驼简介以及经济价值[C]//.2013 中国驼业进展,2013:147-150.

[13]史锟,陈进.几种人造湿地植物和土壤含氮量分析[J].大连交通大学学报,2010,31(06):75-78.

## 12. 附录

**附录 1**

灰色预测及精度检验代码 Pyhton

```
from decimal import *
class GM11():
    def __init__(self):
        self.f = None

    def isUsable(self, X0):
        '''判断是否通过光滑检验'''
        X1 = X0.cumsum()
        rho = [X0[i] / X1[i - 1] for i in range(1, len(X0))]
        rho_ratio = [rho[i + 1] / rho[i] for i in range(len(rho) - 1)]
        print("rho:", rho)
        print("rho_ratio:", rho_ratio)
        flag = True
        for i in range(2, len(rho) - 1):
            if rho[i] > 0.5 or rho[i + 1] / rho[i] >= 1:
                flag = False
        if rho[-1] > 0.5:
            flag = False
        if flag:
            print("数据通过光滑校验")
        else:
            print("该数据未通过光滑校验")

        '''判断是否通过级比检验'''
        lambds = [X0[i - 1] / X0[i] for i in range(1, len(X0))]
        X_min = np.e ** (-2 / (len(X0) + 1))
        X_max = np.e ** (2 / (len(X0) + 1))
        for lambd in lambds:
            if lambd < X_min or lambd > X_max:
                print('该数据未通过级比检验')
                return
        print('该数据通过级比检验')

    def train(self, X0):
        X1 = X0.cumsum()
        Z = (np.array([-0.5 * (X1[k - 1] + X1[k]) for k in range(1,
len(X1))])).reshape(len(X1) - 1, 1)
        # 数据矩阵 A、B
        A = (X0[1:]).reshape(len(Z), 1)
        B = np.hstack((Z, np.ones(len(Z)).reshape(len(Z), 1)))
```

```
            # 求灰参数
            a, u = np.linalg.inv(np.matmul(B.T, B)).dot(B.T).dot(A)
            u = Decimal(u[0])
            a = Decimal(a[0])
            print("灰参数 a：", a, "，灰参数 u：", u)
            self.f = lambda k: (Decimal(X0[0]) - u / a) * np.exp(-a * k) + u / a

        def predict(self, k):
            X1_hat = [float(self.f(k)) for k in range(k)]
            X0_hat = np.diff(X1_hat)
            X0_hat = np.hstack((X1_hat[0], X0_hat))
            return X0_hat

        def evaluate(self, X0_hat, X0):
            '''
            根据后验差比及小误差概率判断预测结果
            :param X0_hat: 预测结果
            :return:
            '''
            S1 = np.std(X0, ddof=1)  # 原始数据样本标准差
            S2 = np.std(X0 - X0_hat, ddof=1)  # 残差数据样本标准差
            C = S2 / S1  # 后验差比
            Pe = np.mean(X0 - X0_hat)
            temp = np.abs((X0 - X0_hat - Pe)) < 0.6745 * S1
            p = np.count_nonzero(temp) / len(X0)  # 计算小误差概率
            print("原数据样本标准差：", S1)
            print("残差样本标准差：", S2)
            print("后验差比：", C)
            print("小误差概率 p：", p)


    if __name__ == '__main__':
        import matplotlib.pyplot as plt
        import numpy as np
        import pandas as pd

        plt.rcParams['font.sans-serif'] = ['SimHei']  # 步骤一（替换 sans-serif 字体）
        plt.rcParams['axes.unicode_minus'] = False  # 步骤二（解决坐标轴负数的
负号显示问题）

        # 原始数据 X
        df = pd.read_excel(r'D:\MODEL\model2\graypic.xlsx')
        df.head()
        # X1 = df[['土壤蒸发量(mm)', '平均露点温度(℃)', '降水天数', '平均能见度
```

```
(km)', '平均最大持续风速(knots)']]
            #    X    =    Decimal(np.squeeze(np.array(df[['    土    壤    蒸    发    量
(mm)']]))).quantize(Decimal("0.00"))

            input = '露点温度'
            X = np.squeeze(np.array(df[[input]])).astype('float')
            # 训练集
            # X_train = X[:int(len(X) * 0.7)]
            X_train = X
            # 测试集
            # X_test = X[int(len(X) * 0.7):]

            model = GM11()
            model.isUsable(X_train)  # 判断模型可行性
            model.train(X_train)  # 训练
            Y_pred = model.predict(len(X)+2)  # 预测
            Y_train_pred = Y_pred[:len(X_train)]
            Y_test_pred = Y_pred[len(X_train):]

            # 记录
            if input == '降水天数':
                A = np.round(np.resize(Y_test_pred, (2, 1)), 0).astype('int')
            else:
                A = np.round(np.resize(Y_test_pred, (2, 1)), 2)

            score_test = model.evaluate(Y_train_pred, X_train)  # 评估

            # 可视化
            fig = plt.figure()
            plt.grid()
            plt.plot(np.arange(2012, 2022, 1), X_train, '->')
            plt.plot(np.arange(2012, 2022, 1), Y_train_pred, '-o')
            plt.legend(['实际值', '预测值'])
            plt.title('灰色预测模型拟合可视化')
            plt.xlabel('年份')
            plt.ylabel('7 月_平均露点温度(℃)')
            plt.show()
            fig.savefig('7 月平均露点温度(℃)拟合曲线.png')
```

**附录 2**

LightGBM 及调优代码 Python

```
import pandas as pd
import numpy as np

import lightgbm as lgb

import matplotlib.pyplot as plt
from IPython.core.display import display

from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV

from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
# from sklearn.metrics import mean_squared_log_error
from sklearn.metrics import median_absolute_error
from sklearn.metrics import r2_score

from sklearn.metrics import accuracy_score
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score



# 对于 xlsx 的文件
df = pd.read_excel(r'D:\MODEL\model2\huiguiall.xlsx')
df.head()

X = df[['土壤蒸发量(mm)', '降水天数', '平均能见度(km)', '平均最大持续风速(knots)', '平均露点温度(℃)']]
# X = df.drop(columns=['10cm 湿度(kg/m2)', '40cm 湿度(kg/m2)', '100cm 湿度(kg/m2)', '200cm 湿度(kg/m2)', '月份', '年份'])
# Y = df[['10cm 湿度(kg/m2)', '40cm 湿度(kg/m2)', '100cm 湿度(kg/m2)', '200cm 湿度(kg/m2)']]
Y = df[['10cm 湿度(kg/m2)']]

X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=123)
# 划分训练集和测试集

# model = lgb.LGBMRegressor(max_depth=6, learning_rate=0.14)  # 回归
model = lgb.LGBMRegressor()  # 回归
# model = lgb.LGBMClassifier()  # 分类
model.fit(X_train, y_train)
```

```
y_pred = model.predict(X_test)

## 写入，方便比较
# a = pd.DataFrame()  # 创建一个空的 DataFrame
# a['预测值'] = list(y_pred)
# a['实际值'] = list(y_test)

# 模型误差（回归）
print('平均绝对误差 MAE：', mean_absolute_error(y_test, y_pred))
print('均方误差 MSE：', mean_squared_error(y_test, y_pred))
# print('均方误差对数 MSLE ：', mean_squared_log_error(y_pred, y_test))
print('中位绝对误差：', median_absolute_error(y_test, y_pred))
print('可决系数 R^2 ：', r2_score(y_test, y_pred))

# 模型准确度评分（分类）
# score = accuracy_score(y_test， y_pred)
# model.score(X_test, y_test)

# 模型 ROC 曲线（分类）
# y_pred_proba = model.predict_proba(X_test)
# fpr, tpr, thres = roc_curve(y_test, y_pred_proba[:, 1])
# plt.plot(fpr, tpr)
# plt.show()

# 模型 ROC 曲线的 AUC 值（分类）
# score = roc_auc_score(y_test.values, y_pred_proba[:, 1])

# 特征重要性排序
features = X.columns  # 获取特征名称
importances = model.feature_importances_  # 获取特征重要性
# 通过二维表格形式显示
importances_df = pd.DataFrame()
importances_df['特征名称'] = features
importances_df['特征重要性'] = importances
importances_df.sort_values('特征重要性', ascending=False)

# 模型参数调优
# parameters = {'num_leaves': [10, 15, 31], 'n_estimators': [10, 20, 30], 'learning_rate':
[0.05, 0.1, 0.2]}
parameters = {'max_depth': [2, 3, 4, 6, 8, 10, 12, 15, 17, 20], 'learning_rate': [0.02, 0.04,
0.06, 0.08, 0.1, 0.2, 0.12, 0.14, 0.16, 0.24, 0.26]}
grid_search = GridSearchCV(model, parameters, cv=5)  # 5 折交叉验证
grid_search.fit(X_train, y_train)  # 传入数据
```

```
print('最优参数：  ', grid_search.best_params_)
print('最优得分：  ', grid_search.best_score_)
# print('不同参数情况下交叉验证的结果：  ', grid_search.cv_results_)
#  结果：{'learning_rate': 0.1, 'n_estimators': 20, 'num_leaves': 15}



"""画出拟合曲线"""
fig = plt.figure()
plt.rcParams['font.sans-serif'] = 'default'   #  中文黑体为'SimHei'
plt.rcParams['axes.unicode_minus'] = False     #  中文时，正常显示负号
font_dict = dict(fontsize=13,
                 color='k',
                 family='default',                 #  中文黑体为'SimHei'
                 weight='light',
                 style='italic',
                 )
plt.title('Curve Fitting', fontdict=font_dict)
plt.xlabel('Date', fontdict=font_dict)
plt.ylabel('Qualified Rate', fontdict=font_dict)
plt.scatter(X_test, y_test, label='Label', linestyle='-', color='r', marker='*', linewidth=2.5)
plt.scatter(X_test,  y_pred,  label='Prediction',  linestyle='--',  color='b',  marker='*',
linewidth=2.5)
plt.legend()
plt.show()
fig.savefig('拟合曲线.png')
```

**附录 3**

随机森林及调优代码 Python

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
from IPython.core.display import display

from sklearn.ensemble import RandomForestRegressor

from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV

from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
# from sklearn.metrics import mean_squared_log_error
from sklearn.metrics import median_absolute_error
from sklearn.metrics import r2_score

from sklearn.metrics import accuracy_score
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score



# 对于 xlsx 的文件
df = pd.read_excel(r'D:\MODEL\model2\huiguiall.xlsx')
df.head()

X = df[['土壤蒸发量(mm)', '降水天数', '平均能见度(km)', '平均最大持续风速(knots)', '平均露点温度(℃)']]
# X = df.drop(columns=['10cm 湿度(kg/m2)', '40cm 湿度(kg/m2)', '100cm 湿度(kg/m2)', '200cm 湿度(kg/m2)', '月份', '年份'])
# Y = df[['10cm 湿度(kg/m2)', '40cm 湿度(kg/m2)', '100cm 湿度(kg/m2)', '200cm 湿度(kg/m2)']]
Y = df[['10cm 湿度(kg/m2)']]

X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=84)
# 划分训练集和测试集



# model = RandomForestClassifier()  # 分类
# model = RandomForestRegressor(n_estimators=80, max_features=3)  # 回归
model = RandomForestRegressor()
```

```
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

# '''记录'''
# dff = pd.read_excel(r'D:\MODEL\model2\huiguijieguo.xlsx')
# dff.head()
#
# X_testt = dff[['土壤蒸发量(mm)', '降水天数', '平均能见度(km)', '平均最大持续风速(knots)', '平均露点温度(℃)']]
# y_predd = np.round(np.resize(model.predict(X_testt), (24, 1)), 2)


## 写入，方便比较
# a = pd.DataFrame()  # 创建一个空的 DataFrame
# a['预测值'] = list(y_pred)
# a['实际值'] = list(y_test)

# 模型误差（回归）
print('平均绝对误差 MAE：  ', mean_absolute_error(y_test, y_pred))
print('均方误差 MSE：  ', mean_squared_error(y_test, y_pred))
# print('均方误差对数 MSLE ：  ', mean_squared_log_error(y_pred, y_test))
print('中位绝对误差：  ', median_absolute_error(y_test, y_pred))
print('可决系数 R^2 ：  ', r2_score(y_test, y_pred))

# 模型准确度评分（分类）
# score = accuracy_score(y_test，  y_pred)
# model.score(X_test, y_test)

# 模型 ROC 曲线（分类）
# y_pred_proba = model.predict_proba(X_test)
# fpr, tpr, thres = roc_curve(y_test, y_pred_proba[:, 1])
# plt.plot(fpr, tpr)
# plt.show()

# 模型 ROC 曲线的 AUC 值（分类）
# score = roc_auc_score(y_test.values, y_pred_proba[:, 1])

# 特征重要性排序
features = X.columns  # 获取特征名称
importances = model.feature_importances_  # 获取特征重要性
# 通过二维表格形式显示
importances_df = pd.DataFrame()
importances_df['特征名称'] = features
```

```
importances_df['特征重要性'] = importances
importances_df = importances_df.sort_values('特征重要性', ascending=False)

# 模型参数调优
parameters = {'max_features': [4, 3, 2, 1], 'n_estimators': [20, 100, 50, 80, 120, 150, 60, 70, 115, 130]}
grid_search = GridSearchCV(model, parameters, cv=5)  # 5 折交叉验证
grid_search.fit(X_train, y_train)  # 传入数据
print('最优参数： ', grid_search.best_params_)
print('最优得分： ', grid_search.best_score_)
# print('不同参数情况下交叉验证的结果： ', grid_search.cv_results_)
# 结果：{'learning_rate': 0.1, 'n_estimators': 20, 'num_leaves': 15}



"""画出拟合曲线"""
fig = plt.figure()
plt.rcParams['font.sans-serif'] = 'default'  # 中文黑体为'SimHei'
plt.rcParams['axes.unicode_minus'] = False  # 中文时，正常显示负号
font_dict = dict(fontsize=13,
                 color='k',
                 family='default',  # 中文黑体为'SimHei'
                 weight='light',
                 style='italic',
                 )
plt.title('Curve Fitting', fontdict=font_dict)
plt.xlabel('Date', fontdict=font_dict)
plt.ylabel('Qualified Rate', fontdict=font_dict)
plt.scatter(X_test, y_test, label='Label', linestyle='-', color='r', marker='*', linewidth=2.5)
plt.scatter(X_test, y_pred, label='Prediction', linestyle='--', color='b', marker='*', linewidth=2.5)
plt.legend()
plt.show()
fig.savefig('拟合曲线.png')
```