


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	电子科技大学
参赛队号	22106140093
队员姓名	1.杨秉鸿
	2.丁谦
	3.曹英杰

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　　锡林郭勒草原放牧与可持续性发展研究

摘　　要：

自 2003 年国家开始实施退牧还草政策以来，草原生态得到了有效保护，促进了可持续发展。锡林郭勒草原是内蒙古草原的天然草场之一，是华北地区重要的生态屏障，对其因地制宜优化放牧方案有着重大的战略意义。本文研究如下：

针对问题一，通过现有文献和粗略模型，探究不同放牧策略下锡林郭勒草原土壤物理性质（土壤湿度）和植被生物量的运作机理，分别给出了放牧强度、土壤湿度、放牧强度和植被生物量的微分方程符号表达式，并对符号一一解释。表达式包含了放牧强度、降水、土壤特征和植被等因素，为后面的问题提供了理论依据。

针对问题二，为了筛选出更具显著代表性的因子，建立了**加权集成特征筛选模型**。首先使用相关的**方差信息**对冗余特征进行过滤，接着使用**斯皮尔曼相关系数**对冗余变量进行过滤筛选出具有独立性和代表性的变量；最后采用加权集成式的特征筛选模型（包含斯皮尔曼相关系数、**距离相关系数、随机森林和弹性网络法**），选取全局重要性**占 75%的代表性变量**。在未来土壤湿度预测阶段，使用基于时间序列的 **SARIMA、SARIMAX 模型**进行预测，并使用 **MSE、MAE、RMSE** 和 **R2 指标**对模型进行验证，证明了我们提出的模型具有较高的鲁棒性和准确性。

针对问题三，我们通过相关**文献调查**研究获取与土壤化学性质的**可能相关变量**，使用问题二提出的**时间序列模型**获取 2022 年的土壤湿度数据弥补附件数据的空缺，并结合**数据可视化分析、最小二乘法**探索特征变量和土壤化学性质之间的关系。最后**建立数学微分方程**、并使用**决策树模型**适用于小样本数据的特点对未来的土壤化学性质进行预测。

针对问题四，首先通过锡林郭勒盟整体的数据修正给定的**沙漠化程度指数预测模型**参数，之后通过监测点的经纬度确定其在地图上的坐标，然后根据监测点与周边城市的**空间相关性**，估算出了监测点缺失的气象数据，接着**检索附件数据**得到人文因素数据。得到以上数据后，代入**修正的预测模型**得到不同放牧强度下的监测点沙漠化程度指数值，并给出了沙漠化程度的**定量分析**，其次通过**层次分析法**定量分析了土地板结化程度。最后，联合沙漠化程度和板结化指数通过**联合粒子群优化算法与遗传算法**求解最优放牧策略模型。

针对问题五，结合问题四中给出的沙漠化程度和板结化程度预测模型，在保证草原可持续发展的基础上，实现最大的经济效益。于是，进行**约束优化问题建模**，简化为在遏制沙漠化和土地板结化的情况下，寻找不同降雨量对应的最大的放牧强度以达到可持续发展下的经济效益最大化的问题。采取**粒子群优化算法**与**遗传算法**来搜索最优策略，为每个牧户设计了不同降雨量条件下的可持续发展最优策略。

针对问题六，对各示范牧户的当前放牧策略和各种环境因素的值进行了**仿真设置**，使用**图例演示**土地变化状态。根据前面问题得到的放牧强度与植物生长以及土壤化学成分含量变化的预测模型对 2023 年 9 月牧户采取不同放牧策略下的土地状态进行了预测，预测

结果也体现了各示范牧户的放牧策略与我们在问题四提出的放牧策略的优势。

关键词：锡林郭勒草原，放牧强度，土壤性质，沙漠化程度，板结化程度，可持续发展

一、问题重述

1.1 问题背景

草原生态是我国生态环境的重要组成部分，在防风固沙、保护水土、生物多样性维持等方面起到了重要作用，同时也是牧民群众的主要经济来源[1]。近些年，由于全球气候变化以及不合理的人类活动，草原正面临着退化的风险[2]。早在 2003 年，国家就开始实施“退牧还草”的政策以保护草原生态。2021 年，《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》及《内蒙古自治区“十四五”生态环境保护规划》中进一步指出，发展需要坚持尊重自然、顺应自然、保护自然，实施可持续发展战略，促进人与自然和谐共生。

锡林郭勒草原处于内蒙古自治区中部地区，属于温带草原，是内蒙古四大草原之一，是华北地区重要的生态屏障，是距首都北京最近的草原牧区，也是全国唯一被联合国教科文组织纳入国际生物圈监测体系的锡林郭勒国家级草原自然保护区。锡林郭勒草原的植被分布如图 1-1 所示，总共有四种植被类型，草甸草原集中分布于锡盟东北部，是森林向草原的过渡地段，以高平原、低山丘陵与宽谷平原地形为主，是水草丰美的牧场；典型草原主要分布于锡盟中部，分为东部和西部两个部分，是锡林郭勒草原的主体，地形以平原和低山丘陵为主，地表水资源比较丰富，牧草质量好；荒漠草原分布于锡盟西部，植被属旱生类型，植物群落主要由旱生丛生小禾草组成，并混生小半灌木与葱属植物，适宜饲养羊和骆驼；沙地植被主要分布于锡盟的西部和中南部地区，植被是发育在纯沙性母质土壤上的植物群落的组合，沙生系列植物为沙地植被的主体，伴有大量榆、柳、桦等灌木、半灌木林。锡林郭勒草原依托其丰富的生物多样性，有着巨大的生态调节能力、研究空间和经济价值。

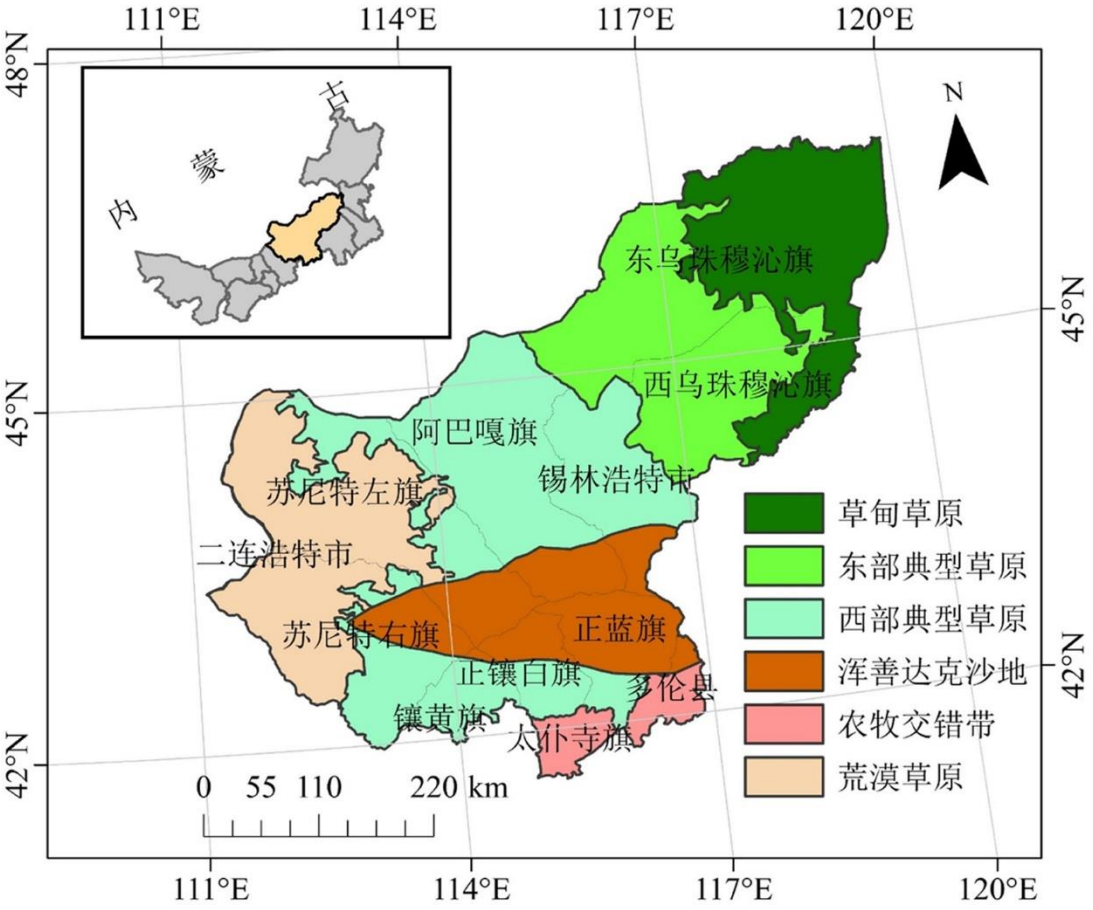


图 1-1 锡林郭勒草原的植被图[3]

一方面，放牧强度与放牧策略与草原植被保护、土壤质量以及生物多样性息息相关，也会影响牧民的经济收入水平[4]。另一方面，土壤水分可以溶解化学物质、稳定土壤温度、运输矿物养分，是植物生长的必备物质，也是连接气候变化和植被覆盖动态的关键因子[5]。因此，对锡林郭勒草原进行降水量监测、土壤分析并进一步完善“退牧还草”策略，因地制宜优化放牧方案有着重大的战略意义。

1.2 需解决的问题

基于以上背景，本文需要研究完成以下问题：

对于问题一，通过对锡林郭勒草原进行机理分析，需要找到放牧策略与锡林郭勒草原土壤湿度以及植被生物量之间的变化规律，从而建立放牧策略（自变量）分别与土壤湿度和植被生物量（因变量）的数学关系式；

对于问题二，附件 3-11 给出了相关数据，我们需要基于这些已有的数据来建立不同深度土壤湿度对于时间的数学模型，从而推算出 2022 年与 2023 年的不同深度土壤湿度；

对于问题三，通过对锡林郭勒草原进行机理分析，我们需要找到放牧策略与锡林郭勒草原土壤化学性质的变化规律，从而建立放牧策略和时间（自变量）与土壤各个化学性质（因变量）的数学模型，利用附件 14 中的数据求得模型中的未知参数；

对于问题四，利用附件和自己收集的数据，确定沙漠化程度指数模型的影响因子和权重系数，从而根据监测点的数据计算出不同放牧强度下监测点的沙漠化程度指数值。并且确定 B=f(W,C,O)的定量数学表达式，并结合问题 3 的数学模型，求解令沙漠化程度指数和板结化程度最小的目标函数，得到相应的放牧策略；

对于问题五，利用前面问题所得的模型，当降水量作为自变量分别取 300mm、600mm、900mm、1200mm 时，保证在实验草场内能够可持续发展的情况下，即前面问题所求得模型的因变量在满足值域在一定范围的约束下，求解目标函数得到最大的放牧强度；

对于问题六，利用前面问题所得的模型和附件的数据，分别在附件 13 的示范牧户的放牧策略和问题四中得到的放牧策略下，预测示范区 2023 年 9 月土地状态，并进行结果图示或动态演示。

二、总体技术路线图

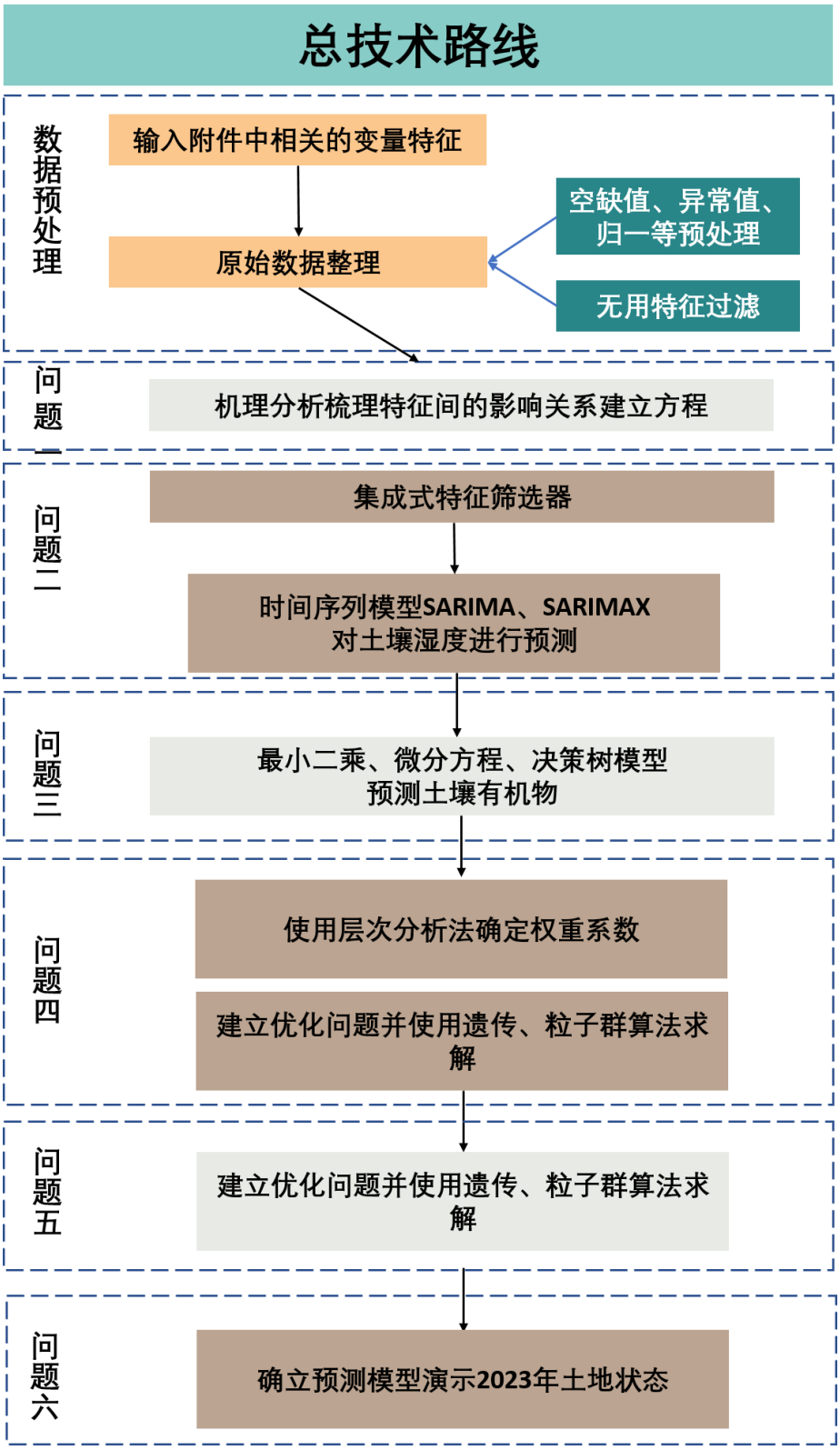


图 2-1 总技术路线图

三、数据预处理

3.1 特征可视化分析

首先剔除无效变量。如图 3-1，我们绘制了题示的相关附件数据的分布直方图，发现如**海拔高度、平均海平面气、最大能见度**等特征变量几乎是恒定的，因此删除这部分对于土壤湿度无影响的**全局不变的特征变量**。同时如积**雪深度、平均最大瞬时风速、最大瞬时风速极值**只包含部分数据、缺失率极高，因此删除**缺失率高于 40%**的无效特征。

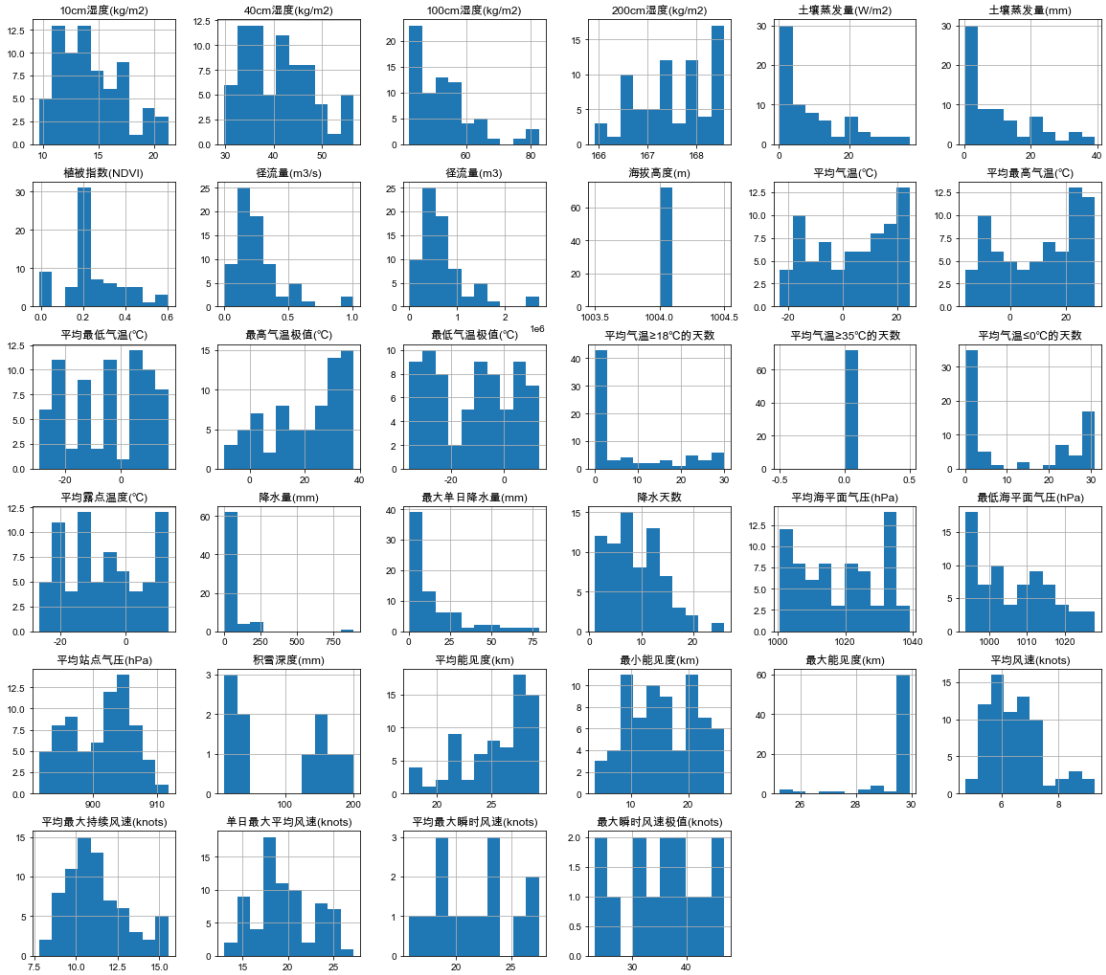


图 3-1 初始化特征分布

3.2 脏数据、空缺值、异常值的处理

首先对于空缺值我们采用**三次样条插值**进行处理。三次样条插值就是把已知数据分割成若干段，每段构造一个三次函数，并且保证分段函数的衔接处具有 0 阶连续，一阶导数连续，二阶导数连续的性质（也就是光滑衔接）。

在离散数据中，存在一些特征数据有空缺值的情况，对空缺值进行插值即根据已知点去计算未知点的数值。一般预处理会采用平均值或者众数代替特征的空缺值，但在问题二中，数据特征是与时间序列相关联的，因此特征值具有一定的趋势，因此需要找到一个插值函数使得函数在未知点上的预测值更准确、更真实、更合理。常用插值方法有最近邻域插值、线性插值、多项式插值等。其中邻近域插值直接采用相邻数据点的值作为空缺值，线性插值法使用左右相邻的点的数据并求取平均值，这两种方法插值函数都过于简化，最近邻域插值法适合于方差小，整体波动小的特征值，线性插值法更适用于整体随时间序列

呈线性变化的特征，但是我们对特征与时间之间关系进行二维图展示发现，最近邻域插值法和线性插值法都过于简化特征趋势，有较大的误差。因此多项式插值法更加趋于精确。但是多项式插值法也存在弊端，多项式插值随着阶数升高，插值的精度也越来越高越来越合理，但是多项式插值存在两个问题：1）随着多项式的阶数的增长，多项式拟合的计算量会越来越大；2）随着多项式的阶数的增长，插值的精度不会与之成正比，相反的是，函数的曲线会出现剧烈的震荡，即龙格现象，如图 3-2 所示：

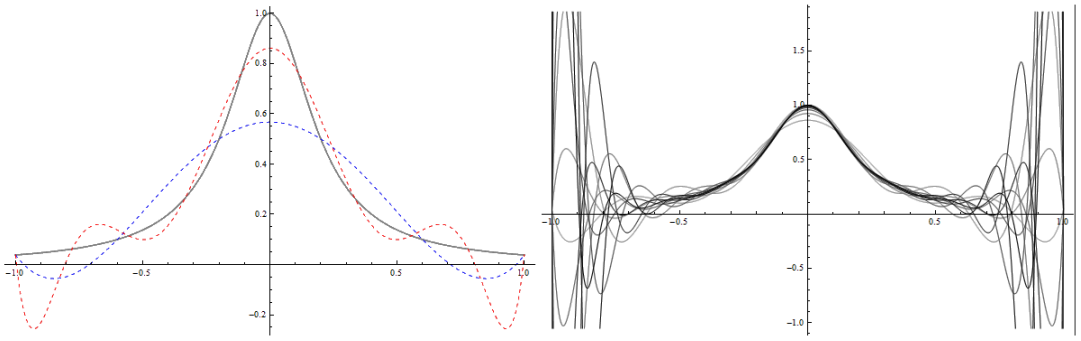


图 3-2 传统多项式插值的震荡情况

随着阶数增长在给定数据点之外曲线会越来越震荡，也容易产生过拟合现象，因此需要一种插值方法既能穿过已知点又能巧妙地避免龙格现象。因此我们采用三次样条插值方法把已知的特征数据分割成若干段，每个段都对应一个函数即插值函数，最终组合在一起得到一个插值函数的序列。若插值函数与插值函数之间存在彼此衔接不够流畅，此时使用 x 的三次方形式构造每一段插值函数，这样使得多个插值函数之间衔接的更加光滑。即每一段三次函数都用如下的方程来定义：

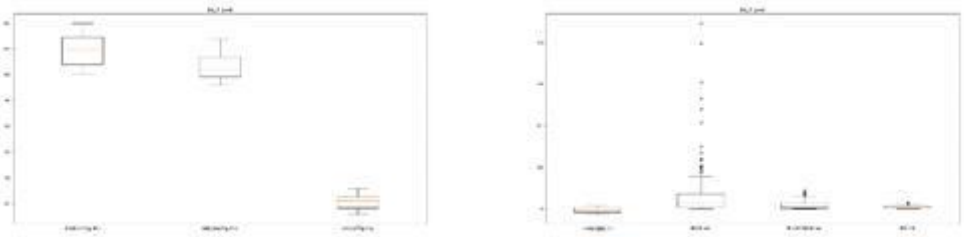
$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, i = 0,1, \dots, n-1$$

那么 n 个区间对应的三次函数s(x)其对应的数学表达式如下：

$$S(x) = \begin{cases} S_0(x) = a_0 + b_0(x - x_0) + c_0(x - x_0)^2 + d_0(x - x_0)^3 & if\ x_0 \le x \le x_1 \\ S_1(x) = a_1 + b_1(x - x_1) + c_1(x - x_1)^2 + d_1(x - x_1)^3 & if\ x_1 \le x \le x_2 \\ \dots \\ S_n(x) = a_n + b_n(x - x_n) + c_n(x - x_n)^2 + d_n(x - x_n)^3 & if\ x_n \le x \le x_n \end{cases}$$

每个子方程函数含有 a, b, c, d 四个未知参数，因此至少需要四个数据点，n 段函数则需要 4n 个数据点分段函数进行求解从而求解得到对应的空缺值。

第二，对于特征中的异常值我们使用箱线图进行处理。箱线图主要适用于反映连续型的数据分布的中心位置和分散的范围，箱型图包含丰富的数学统计变量，能够巧妙地反应不同特征之间各个层次水平的差异，还能够有效的揭示各个特征分布的离散程度、异常值检测和特征之间的分布差异。



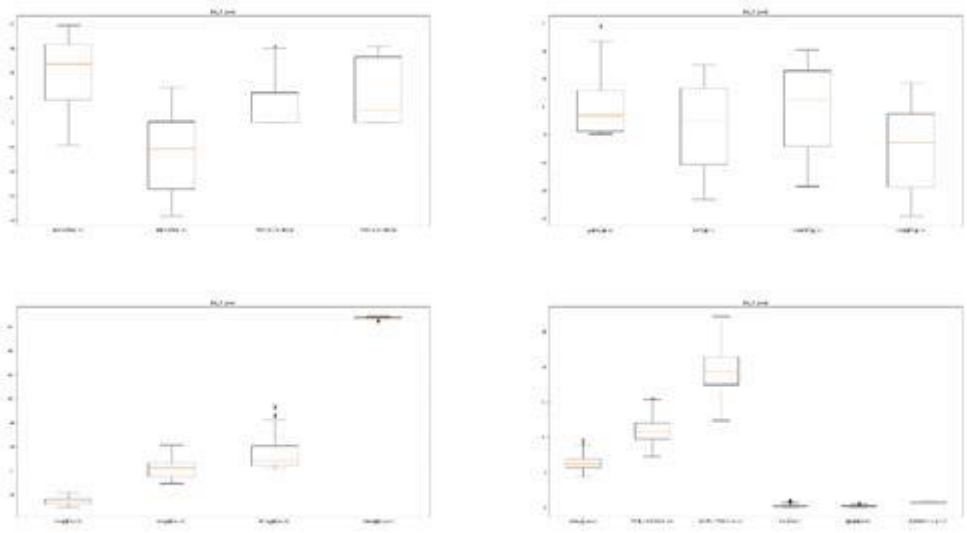


图 3-3 可视化特征箱型线清洗异常值

如图所示，我们对附件中的特征进行箱线图可视化，有效且直观地展示了离群的异常值，其中圆圈标注的地方就是相应特征的异常值。因此我们将异常值剔除并替换为三次样条插值得到的结果，使得脏数据得到了更为合理的替换。

3.3 数据归一化处理

最后我们对数据进行归一化处理以消除数据的纲量影响，其归一化公式如下所示：

$$x = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

四、问题一的求解

4.1 问题分析

根据问题一要求，需要从机理分析的角度，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的数学模型。鉴于锡林郭勒盟自 1998 年以来大力推广划区轮牧的策略并且渐成规模[6]，本文只考虑划区轮牧的放牧方式。

由于生物特征的多样性、土壤性质的复杂性以及数据的缺失，很难用精确的数学公式来表述放牧策略与土壤性质以及植被生物量之间的关系，现有的大部分研究通常通过数据分析的角度来分析两者的相关性来得到大致的结论[4]。

基于上述分析，本文通过总结现有研究工作，探究土壤和植被生物量的运作机理，分别给出了：

1) 放牧强度和土壤湿度的符号关系式；
2) 放牧强度和植被生物量的符号关系式。

其技术路线如图 4-1 所示：

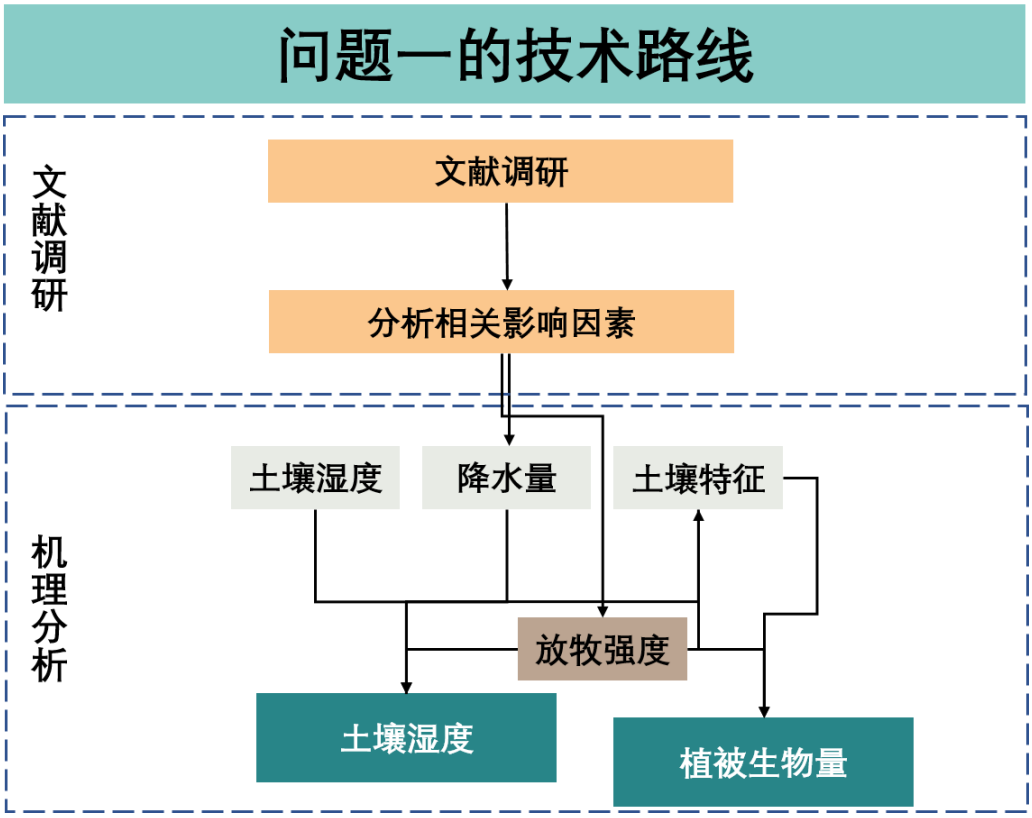


图 4-1 问题一的技术路线图

4.2 放牧强度和土壤湿度之间的关系

土壤的水分含量受到各种各样的因素影响，如土壤深度、降水量、土壤纹理、土壤孔隙度以及植被覆盖率等。有研究表明，前 20cm 深度的土壤湿度受到牧压强度的影响较大，是导致草原退化的重要因素[4]。综合考量以上因素，并结合垂直土柱情况下的水分迁移公式[8]，牧压强度和土壤湿度之间的符号关系式可以表示如下：

$$\frac{\partial \theta}{\partial t} = \underbrace{\frac{\partial}{\partial z}\left[K(\psi)\left(\frac{\partial \psi}{\partial z}+1\right)\right]}_{\text{土壤深度项}} - \underbrace{I\frac{\partial \rho}{\partial z}}_{\text{放牧强度项}} + \underbrace{A(\psi)}_{\text{降水项}} - \underbrace{B(\psi)-C(\psi)-D(\psi)}_{\text{土壤特征项}},$$

式中各个参数符号解释如表 4-1 所示。

表 4-1 各符号所表示含义及其单位

符号	含义	单位
$\boldsymbol{t}$	时间	天(d)
$\boldsymbol{\theta}$	t时间的土壤湿度	-
$\boldsymbol{z}$	距地表的垂直深度	cm
$\boldsymbol{\psi}$	土壤的毛管势	cm
$\boldsymbol{K}$	非饱和导水率	$cm*d^{-1}$
$\boldsymbol{\rho}$	牧压强度对不同深度土壤的影响系数	$cm*unit^{-1}*d*hm^{2}$
$\boldsymbol{I}$	放牧强度	$unit*d^{-1}*hm^{-2}$
$\boldsymbol{A}$	降水率	d^{-1}
$\boldsymbol{B}$	植物根系对土壤水分的吸收率	d^{-1}
$\boldsymbol{C}$	饱和土壤的排水率	d^{-1}
$\boldsymbol{D}$	土壤大孔隙水交换率	d^{-1}

现对符号关系式中每一项作详细阐述。

土壤深度项：土壤深度是影响土壤水含量的重要因素，从地表到地下 360cm 可以分为四个特征层，多变层（地表-地下 20cm）容易受气象条件和农业技术措施的影响，变异系数较大；贮水层（地下 20cm-80cm）是土壤水库的深水地带，是主要的供水层，对作物生长及其产量有着重要的意义;缓变层(地下 80cm-280cm)受气象条件和农业活动影响较小，所以变异系数较小,同时对作物生长的水分供应有着较大影响;均稳层(地下 280cm-360cm)作为最深的特征层，受外界条件影响最小，变异系数也最小，土壤水分处于比较稳定的状态，含水量较高。在表达式中，本文用非饱和导水率K刻画土壤湿度与深度之间的关系，其表达式如下[8]：

$$K(\psi)=\begin{cases}K_s[\Theta(\psi)]^{\beta}\left[\dfrac{1-\left(1-[\Theta(\psi)]^{\frac{1}{m}}\right)^{m}}{1-\left(1-[\Theta(\psi_b)]^{\frac{1}{m}}\right)^{m}}\right]^{2}, \psi\le\psi_b\\ K_s, \psi>\psi_b\end{cases},$$

$$\Theta(\psi)=\frac{\theta(\psi)-\theta_r}{\theta_s-\theta_r}=\begin{cases}(1+|\alpha\psi|^{n})^{-m}, \psi\le\psi_b\\ 1, \psi>\psi_b\end{cases},$$

$$\alpha=\frac{\zeta}{|\psi_b|},$$

$$n=\lambda+1,$$

$$m=1-\frac{1}{n},$$

式中各个参数符号解释如表 4-2 所示。

表 4-2 各符号所表示含义及其单位

符号	含义	单位
$\boldsymbol{\Theta}$	毛管势为ψ的有效饱和度	-
$\boldsymbol{K_s}$	饱和导水率	$cm*d^{-1}$
$\boldsymbol{\beta}$	经验孔隙连通参数	-
$\boldsymbol{\psi_b}$	进气势头	cm
$\boldsymbol{\theta_r}$	土壤残留湿度	-

$\boldsymbol{\theta_s}$	土壤饱和湿度	-
$\boldsymbol{\lambda}$	孔径分布指数	-
$\boldsymbol{\zeta}$	土壤滞后系数	-

放牧强度项：适当的放牧可以降低表层土壤湿度、PH，一定程度增加土壤容重，但是过度的放牧会导致表层土壤水分流失严重，造成草原退化。由于浅层土壤水土流失受到放牧强度的影响比较大，本文用放牧强度对不同深度土壤的影响系数ρ来刻画该现象。

降水项：降水可以改善水循环，是影响土壤水分的主要气象因素。锡林郭勒草原地理坐标介于东经 110°50′~119°58′，北纬 41°30′~46°45′之间，年均降水量 340mm。有研究统计无降水日占 75.4%，降水日中，小降水事件占 86.7%，大降水事件发生频率低，占 4.3%，但是对总降水量有着近一半的贡献[5]。因此，土壤湿度关系式中需要考虑降水量。

土壤特征项：土壤湿度还与土壤自身的特征相关，比如该土壤的植被量、土壤的饱和度以及孔隙度。每单位面积土壤的植被覆盖率越高，那么该单位面积土壤的植物根系对土壤水分的吸收率B就越高。对于饱和土壤来说，排水率C和大孔隙水交换率D都是正的，即造成水流失；对于非饱和土壤，这两个参数为负数，即造成水流入。

4.3 放牧强度和植被生物量之间的关系

Logistic 方程作为一个经典的生态学方程，可以描述资源种群变化这种非线性阻滞增长过程。根据种群生态学理论，在无放牧以及其他人为干扰的情况下，草原的植被生物量随时间变化的规律符合 Logistic 曲线[9]：

$$W = \frac{K}{1 + e^{a-Gt}}$$

式中，W为植被生物量，K为环境容纳量（最大植被生物量），G为植被内禀增长率。那么通过对时间的一阶导，我们可以得到植被生物量的增长速率：

$$\frac{dW}{dt} = GW(1 - \frac{W}{K})$$

考虑放牧强度因素的影响，不同的放牧强度指草原上实际放牧不同数量的家畜，相应的植被摄入量随着放牧强度的上升而增加。在放牧过程中，家畜的采食会影响植被生物量的增长速率[7]，结合该影响因素，植被生物量的增长速率重新表示为：

$$\frac{dW}{dt} = GW\left(1 - \frac{W}{K}\right) - aIW$$

式中，I为放牧强度。

许多研究结果表明，植被生物量的形成和积累不仅受群落中各种群生育节律的支配，而且受外界干扰和环境因子的制约。在放牧的过程中，通过适当的放牧，家畜在采食时的践踏能够促进枯落物分解，充分进入土壤，改善土壤的质量，促进植物的生长，而过度的放牧反而会破坏植被结构，增大土壤裸露面积，促进土壤表面的蒸发量，降低土壤质量，不利于植物的生长。考虑到放牧强度对草原土壤理化特征的影响，利用文献[10]提出的多因素植物生长模型，将植被生物量的增长速率重新表示为：

$$\frac{dW}{dt} = \left(G + \sum_{i=1}^{N} b_i x_i(W, I)\right) W\left(1 - \frac{W}{K(x1, x2, \dots, xi)}\right) - aIW$$

式中，环境容纳量$K(x1, x2, \dots, xi)$是由环境影响因素x_i共同决定的变量。$x_i(W, I)$是受放牧强度影响的环境影响因素（如土壤有机质含量、土壤湿度等）关于植被生物量和放牧强度的函数，N表示环境影响因子的个数，对于任意x_i影响因子，都满足植物生长模型，即$\frac{dW}{dx_i} = r_i W(1 - \frac{W}{K})$。当获取到各变量的实际数据时，利用最小二乘法进行多元回归计算可以

得到所需的未知参数b_i和a。

五、问题二的求解

5.1 问题分析

分析问题二可以得到，需要对土壤湿度数据、土壤蒸发数据、降水数据、径流量等与土壤湿度相关的特征数据进行预处理并建立相关模型进行分析预测。观察数据可知，要想使得训练的预测模型保持鲁棒性，就需要对特征进行筛选工作，从而使得影响模型结果的变量是与湿度密切相关的特征。因此对于问题二我们主要需要解决两个关键问题：第一，需要将**不包含高度相关性的重复特征**、**独立性较好**且与土壤湿度**密切相关联**的特征筛选出来，从而得到相应的因变量数据。第二，由于需要预测的对象是不包含相应因变量的模型，且与时间相关，因此**需要建立基于时间序列预测的模型**对未来关键时间点的土壤湿度进行预测。

对于第一个关键问题，首先我们分析题目提示的附件以及其他相关附件发现其中存在很多**冗余的特征**，因此需要对数据进行数据清洗。数据清洗阶段我们先计算每个特征数据的缺失率，如果样本特征**缺失率超过 40%**认为是无效特征，同时认为在时间变化情况下**全局不变的特征**为无效特征，并进行**类内类间相关性分析**从而去除掉高度相关的重复特征，至此完成剔除冗余特征的工作。其次，对保留的特征进行预处理工作，使用**三次样条插值方法**对缺失值进行插值处理，使用**箱线图**对特征异常值进行检测。最后特征筛选不同于特征提取，特征筛选希望降维后的特征是具有可解释性并保留原始特征空间的性质，使得一定程度上特征的噪声降到最低从而促进模型拟合。而特征提取相关的 PCA 降维等方法将原始空间的特征进行了空间变化，因此并没有保留降维后的特征的语义信息，因此在我们的特征筛选阶段，我们使用**集成使用**基于线性关系的 **Spearman 相关系数**、基于非线性的**距离相关系数**和基于机器学习的**随机森林模型**三种方法，对特征与土壤湿度的相关性进行排序筛选，从而分别筛选出几个重要性显著的特征作为模型的因变量，保留了筛选后的特征的语义信息。

对于第二个关键问题，首先我们选用**适用于趋势和季节性成分的时间序列模型 SARIMA 模型**对自变量**降水、蒸发量、植被指数（NDVI）、低层植被（LAIL）**进行自回归预测，并将降水、蒸发量、植被指数（NDVI）、低层植被（LAIL）作为预测 10cm 土壤湿度的外协变量，使用 **SARIMAX 模型得到**相应时间点的 10cm 土壤湿度。进一步我们使用 **Spearman 相关系数**对不同深度的土壤湿度分析发现，不同深度的土壤湿度之间也是高度相关的，因此将得到的 **10cm 土壤湿度、降水量、蒸发量、径流量和植被指数**作为预测 40cm 土壤湿度的协变量，继续使用 **SARIMAX 模型**进行训练预测。同上依次得到 10cm、40cm、100cm、200cm 的土壤湿度。其中，100cm 深度土壤湿度的预测使用的协变量为 40cm 的土壤湿度，200cm 深度土壤湿度的预测使用的协变量为 40cm 和 100cm 的土壤湿度。

最后我们使用**均方误差（MSE）、平均绝对（MAE）、均方根误差（RMSE）**和**决定系数（R2）**对我们的模型进行了评估，实验结果显示我们的模型对于未来土壤湿度的预测具有良好的鲁棒性和准确度。

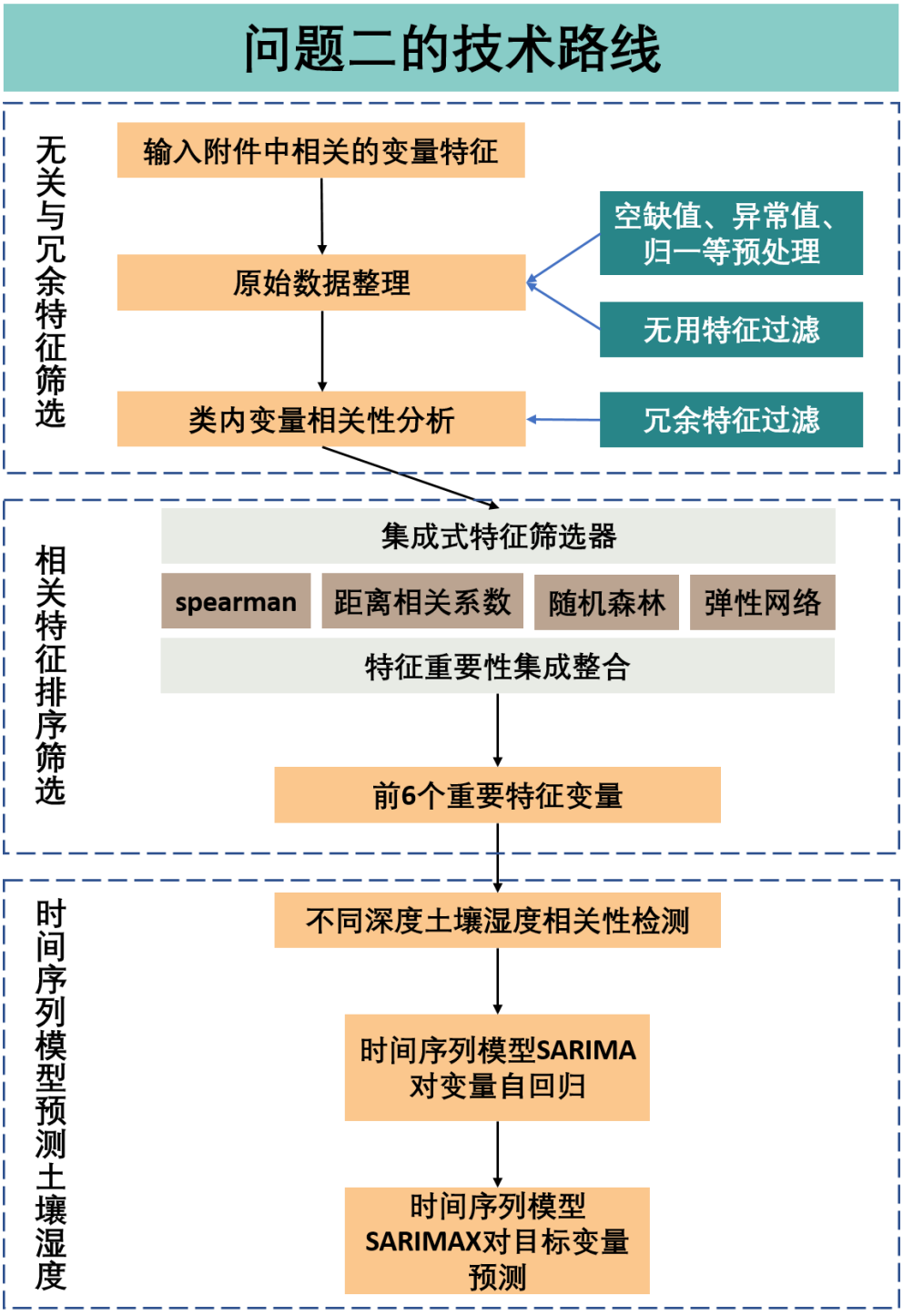


图 5-1 问题二的技术路线

5.2 冗余特征筛选模型的建立

第二，我们使用 Spearman 相关系数对与土壤湿度相关的变量进行分析，选取相关度高的特征作为模型的因变量进行训练。针对相关特征进行可视化分析后发现，各个特征之间具有一定的相关性，因此需要计量各个特征之间的相关性，并将高度相关的特征进行筛选剔除，使得筛选出的特征具有高度的独立性。在相关性分析中，常用的方法包括皮尔逊 Pearson、斯皮尔曼 Spearman、肯德尔 Kendall 三种相关系数矩阵法。下面将对这三种方法进行分析和比较，从而选择出更适合我们的建模问题的方法。

1）Pearson 相关系数：主要用于度量 X、Y 这两个变量之间的相关性的方法。Pearson 相关系数是用协方差除以两个变量的标准差从而得到的，虽然协方差能反映两个随机变量的相关程度，但是协方差值的大小并不能很好地度量两个随机变量的关联程度，为了更好的度量两个随机变量的相关程度，引入了 Pearson 相关系数，其在协方差的基础上除以了两个随机变量的标准差。Pearson 是一个介于-1 和 1 之间的值，当两个变量的线性关系增强时，相关系数趋于 1 或-1；当一个变量增大，另一个变量也增大时，表明它们之间是正相关的，相关系数大于 0；如果一个变量增大，另一个变量却减小，表明它们之间是负相关的，相关系数小于 0；如果相关系数等于 0，表明它们之间不存在线性相关关系。其具体的公式表示如下：

$$\mathrm{cov(X,Y)} = \frac{\sum_{n}^{i=1}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y\text{）}}{\sigma_X \sigma_Y} = \frac{E[(X-\mu x)(Y-\mu y)]}{\sigma_X \sigma_Y}$$

2）Spearman 相关系数：是秩相关系数的一种，也叫做斯皮尔曼秩相关系数。其计算方式和 Pearson 相关系数非常的类似，但要将两个变量转化为有序数即需要对两个变量成对取值并排序取秩，根据变量在数据内的位置进行计算。区别于 pearson 方法，pearson 更适用于变量服从正态分布的场景，spearson 假设这两组变量数据并不需要服从正态分布，因此其适用性更为广泛。其计算公式表征如下：

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

3）Kendall 相关系数：是秩相关系数的一种，也叫做肯德尔秩相关系数。不过区别于上述两种相关系数度量方法，Kendall 相关系数是一种衡量有序分类型数据的序数相关性方法，取值同样在-1-1 之间，相关系数为 1 为极度相关，反之-1 为极度不相关。Kendall 相关系数使用了“对数”这一方式来判断相关程度的强弱，其公式表征如下：

$$\rho_{X,Y} = \frac{n_m - n_n}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

其中n_m和n_c分别表征为变量 X 和变量 Y 中一致性的元素的数量和不一致的元素的数量。和前两种方法相对比 kendall 方法更适合相对有序的变量。

由于本题目中的相关变量并未完全符合正态分布且不满足相对有序的条件，因此我们选择 spearman 相关序列对变量特征之间的相关性进行了度量，得到相应的特征变量之间的相关系数矩阵：

$$Matrix_\rho = \begin{pmatrix} \rho_{11}\ \rho_{12}\ \cdots\ \rho_{1n} \\ \rho_{21}\ \rho_{22}\ \cdots\ \rho_{2n} \\ \cdots \\ \rho_{n1}\ \rho_{n2}\ \cdots\ \rho_{nn} \end{pmatrix}$$

其中ρ_{xy}代表不同特征变量之间的相关系数大小，n 表示特征变量的总数量。根据计算得到的相关系数矩阵，找出对应的ρ_{xy}大于相应阈值的变量对，认为这一对变量是具有强相关性的变量对，因此保留代表性更强的变量，从而可以有效的剔除类内相关性强的冗余变量。如土壤蒸发量（W/mm）和土壤蒸发量（mm）无论从相关系数表征上还是直观字面意义上都具有相互耦合的内在属性，因此我们二者之间只保留一项，作为后续时间序列模型

依赖的变量特征，以保持筛选出来的特征因子之间是具有强代表性和独立性特征的变量，至此满足了冗余特征筛选的要求。

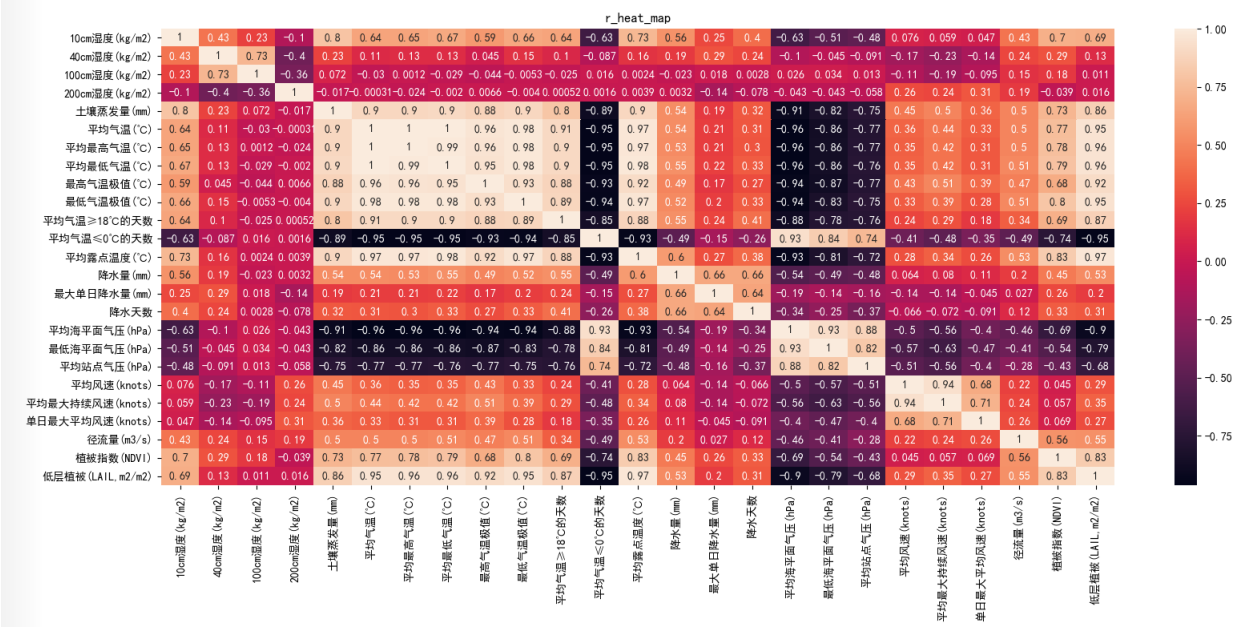


图 5-2 冗余特征的 spearman 相关性分析

5.3 特征筛选模型的建立

我们需要对这些独立性高、有一定变化且经过异常值、空缺值检测的变量特征进行筛选，从而选出与土壤湿度高度相关的特征变量，作为模型训练的基础变量。在特征筛选阶段现有的多种方法可用于特征筛选，如基于 Spearman 相关系数、距离相关系数度量、基于随机森林的特征筛选模型和基于弹性网络的筛选等等方法，我们选则集成多种经典特征筛选的方法，从而使得决策更加的科学理性。

使用 Spearman 相关系数主要用于研究两个变量之间的线性相关程度，同样从 spearman 相关系数矩阵中可以得到特征因子变量和目标变量之间的线性关联程度作为特征筛选的参考标准之一。但同时考虑到特征因子变量和目标变量之间还存在非线性关联的情况，因此需要进一步采用距离相关系数、基于随机森林的特征筛选模型和基于弹性网络的筛选模型探索特征因子和目标变量之间的非线性关系。其技术流程图如下所示。

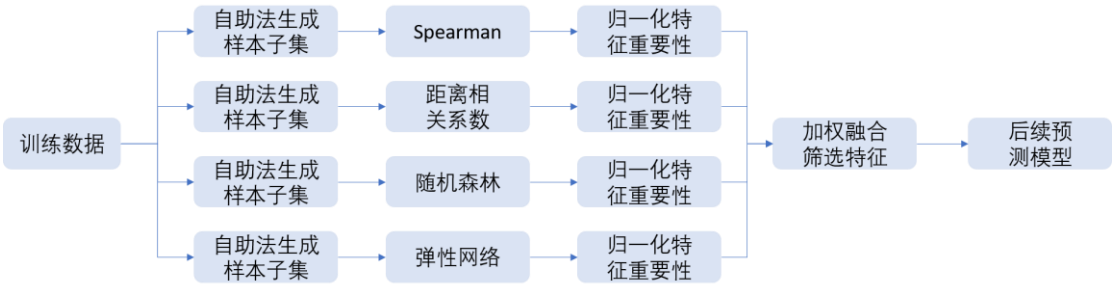


图 5-3 集成特征筛选技术路线

第一，距离相关系数克服了传统的 pearson、spearman 等相关系数度量方法只能探索线性关系的缺点，距离相关系数能够衡量变量特征和目标变量之间的非线性相关性程度。在一些特定的场景下，pearson 相关系数结果为 0 只能将两个变量判定为线性无关，但也可能存在非线性相关性的情况。在距离相关系数为 0 的情况下，能够将两个变量确定为相互独

立的变量，因此在特征筛选阶段更具有科学性。任意两个随机变量 X，Y 的距离相关系数其公式表征如下所示：

$$R^2(x,y)=\frac{h^2(x,y)}{\sqrt{h^2(x,x)h^2(y,y)}}$$

其中：

$$h^2(x,y)=\frac{1}{n^2}\sum_{i,j=1}^{n}X_{i,j}\,Y_{i,j}$$

$$X_{i,j}=\left\|x_i-x_j\right\|_2-\frac{1}{n}\sum_{k=1}^{n}\left\|x_k-x_j\right\|_2-\frac{1}{n}\sum_{i=1}^{n}\left\|x_i-x_j\right\|_2+\frac{1}{n^2}\sum_{k,i=1}^{n}\left\|x_k-x_j\right\|_2$$

$$Y_{i,j}=\left\|y_i-y_j\right\|_2-\frac{1}{n}\sum_{k=1}^{n}\left\|y_k-y_j\right\|_2-\frac{1}{n}\sum_{i=1}^{n}\left\|y_i-y_j\right\|_2+\frac{1}{n^2}\sum_{k,i=1}^{n}\left\|y_k-y_j\right\|_2$$

在距离相关系数度量中，距离相关系数的取值范围为 0-1 之间，其值越大则意味着两个变量之间的相关性越强，当距离相关系数为 0 时其表示两个变量之间是相互独立的关系即无关的。

第二，随机森林（Random forest，简称 RF）作为机器学习的一种代表性算法之一，常用来解决特征筛选的问题，作为 2000 年后新兴的一种算法在许多分类回归的研究问题中都得到了很好的准确率，且随机森林模型可以用来评估各个特征对于目标变量的重要性，因此也可以用来进行特征筛选工作，也是一种用来探究特征因子与目标变量之间的非线性相关性的常用方法。因此我们选用随机森林作为集成特征筛选模型中的子模型，来判断各个特征对于土壤湿度影响的比重。

随机森林是以决策树模型为基决策器，进行多个决策树集成后得到的一个组合的决策器。随机森林最终的结果是由每一个集成的决策树进行投票决定的,其主要技术路线如下图所示：

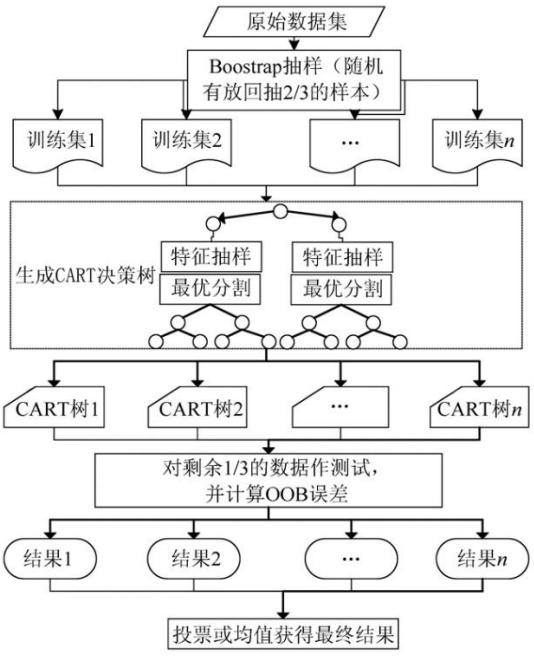


图 5-4 随机森林训练流程

其中决策树的随机变量序列主要是由包（Bagging）思想决定的，又称又称 Bootstrap aggregation，是基本的集成技术之一。Bagging 主要基于统计学中 Bootstraping 自助采样法而来，Bagging 思想指的是从对应的训练样本集中有放回随机抽取 M 个与原来样本数据集同样大小的样本集$T_1, T_2, \dots, T_m$，每个训练集对应一个决策树模型，其中 m 的数量分别取 100，500，1000 调参得到。则基于自主采样法随机森林的构建过程为：

1）从原始训练数据集中使用自主采样法随机有放回采样 m 个数据，共进行 n 次（共 n 棵子决策树），从而生成 n 个子训练集；

2）对于生成的 n 个训练集分别进行训练得到 n 个对应的决策树模型；

3）对于每一个子决策树，我们使用基尼指数选择最好的即对于信息增益影响最大的特征进行分裂，其中基尼指数的公式为

$$\mathrm{Gini(D)} = \sum_{k=1}^{|y|} \sum_{k'!=k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} p_k^2$$

$$\mathrm{Gini\_index(D,a)} = \sum_{v=1}^{V} \frac{|D^v|}{D} Gini(D^v)$$

其中$\mathrm{Gini(D)}$表示数据的基尼系数，其反映了一个数据集的纯度，其值越大数据集纯度越高，反之则反。$\mathrm{Gini_index(D,a)}$表示候选属性的基尼系数，每一次分裂选择使得划分后基尼指数最小的属性作为当前选择划分属性，其先后分类的顺序也揭示了特征对于目标函数的相关性。

4）每棵树都选择这样的分裂方式，直到所有的数据都归属于同一类别，这个过程不需要进行剪枝；

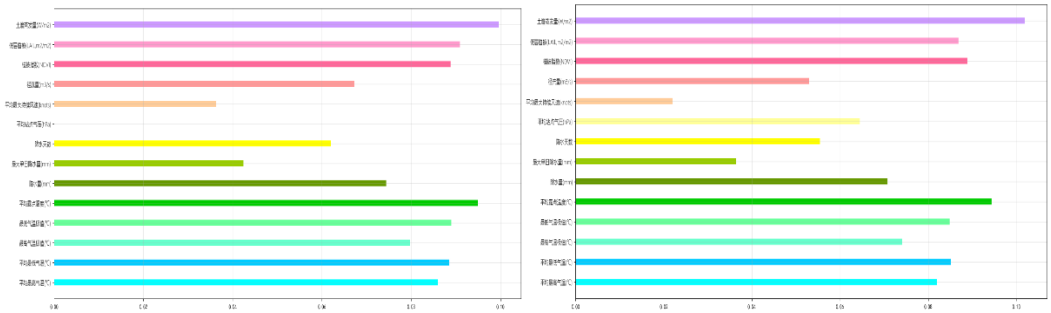
5）将多棵子决策树组合成随机森林，由于该问题是一个回归预测问题因此对多个决策树的预测值取均值得到最终的结果。

第三，弹性网络方法所得到的模型就像纯粹的 Lasso 回归一样稀疏，但同时具有与岭回归提供的一样的正则化能力，是综合使用 L1， L2 正则范数作为先验正则项训练的线性回归模型。这种组合允许拟合到一个只有少量参数是非零稀疏的模型，就像 Lasso 一样，但是它仍然保持了一些类似于 Ridge 的正则性质。其拟合的目标函数为：

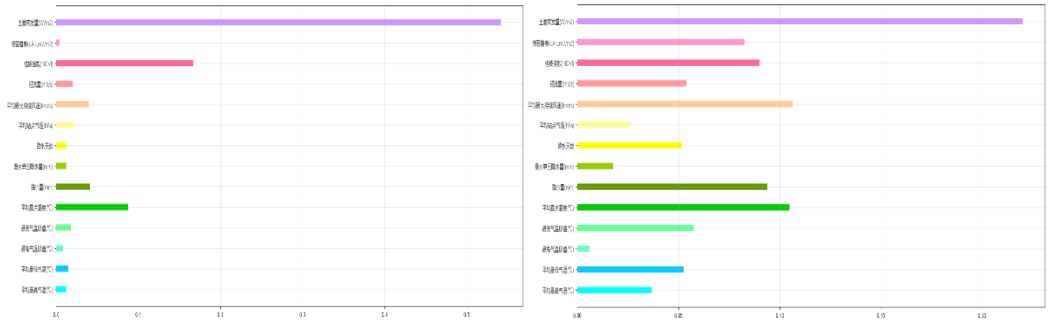
$$min_w \frac{1}{2n_{samples}} \|x_w - y\|_2^2 + \alpha\rho\|w\|_1 + \frac{\alpha(1-\rho)}{2}\|w\|_2^2$$

其中尾部的$\alpha\rho\|w\|_1 + \frac{\alpha(1-\rho)}{2}\|w\|_2^2$即就是综合了 L1L2 正则范数的弹性网路惩罚项，观察公式可以得到当系数参数$\rho = 1$时，弹性网络此时转换为 Lasso 回归，当$\rho = 0$时，弹性网络转化为岭回归，巧妙地则 Lasso 回归和岭回归之间及进行了折中的选择，保留了 Lasso 回归和岭回归的一些特性。使得拟合的目标函数更为灵活。

四种模型对与特征的排序分别如图所示：



（a）spearman 相关系数　　（b）距离相关系数



（c）随机森林相关系数　　（d）弹性网络相关系数

图 5-5 四种重要性分析法对应的特征重要性

5.4 征筛选模型的结果和模型的集成

我们使用 python 编程语言建立四种特征筛选模型，并在提供的数据集上进行实验，并画出了四种方式下的每个特征的重要性直方图，发现四种模型具有一部分的重叠特征，因此侧面反映我们的特征选择模型具有一定程度的统计代表意义。

根据四种方法我们建立集成特征筛选模型，使得得到的特征因子综合了线性相关性、非线性相关性、高耦合性等内容，且集成四种特征筛选模型得到的模型更加灵活、过拟合风险更低。

具体而言，我们将四种模型对不同特征的相关性分别进行归一化处理，即表征每一种关联性分析模型下每一个特征占所有特征重要性之和的比值：

$$y = \frac{y_0 - y_{min}}{y_{max} - y_{min}}$$

使用归一化后的特征系数作为四种模型的系数，并相应的对特征值重要性系数进行加权集成，得到最终对于土壤湿度重要性高的特征变量：

$$Y = \sum_{i=1}^{n} \alpha_i y_i$$

其中α_i为各个模型在集成融合模型种所占的权重，y_i为各个模型归一化后的重要性程度，最后集成模型筛选出的特征其重要性如下表所示：

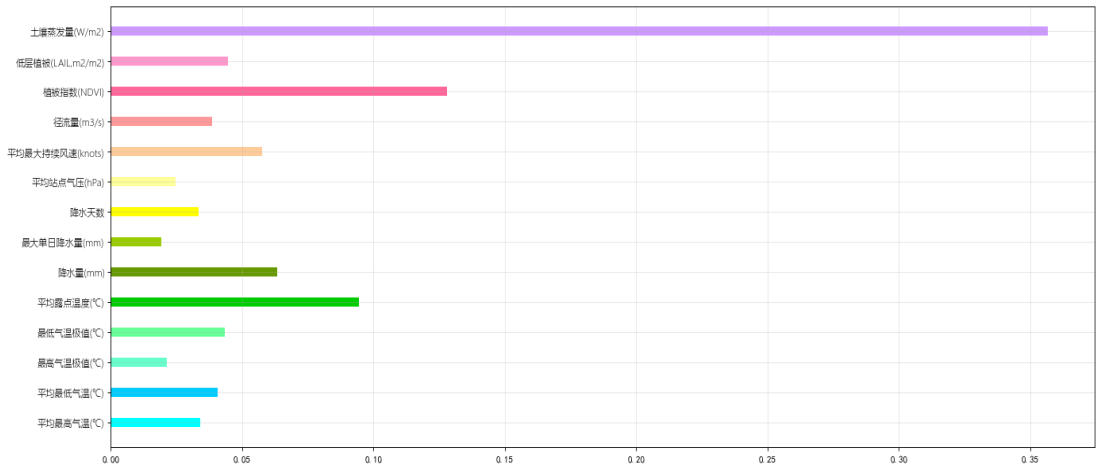


图 5-6 集成特征筛选模型得到的重要性系数

由上图可知，我们初步选取排名前六的特征因子作为后续自回归模型的自变量，因为前六名特征的重要系数占全局重要系数的 75%，具有高度的代表性和相关性，具体重要性系数如表所示：

表 5-1 集成特征筛选模型选出前六名重要性特征

名称	重要系数	名称	重要系数
土壤蒸发量(W/m2)	0.35679312390390994	植被指数(NDVI)	0.12807912656525067
平均露点温度(℃)	0.09438438057772598	降水量(mm)	0.06322453875447637
平均最大持续风速(knots)	0.057650813383692136	低层植被(LAIL,m2/m2)	0.044688705956638416

进一步分析发现，前六名特征中，平均露点温度和平均最大持续风速与土壤蒸发量互相耦合，为了避免时间序列预测时的多重共线性问题，故最终只取土壤蒸发量、降水量、植被指数、低层植被四个特征量作为外生变量对土壤 10cm 深处的湿度进行预测。

同理可以得到与土壤 40cm 深处湿度相关性相对较高的五个变量：10cm 深湿度、土壤蒸发量、降水量、植被指数、径流量；与土壤 100cm 深处湿度相关性相对较高的一个变量：40cm 深湿度；与土壤 200cm 深处湿度相关性相对较高的二个变量：40cm 深湿度、100cm 深湿度。

5.5 基于时间序列预测模型的建立

5.5.1 基于变量特征自回归的 SARIMA 模型的建立

得到显著影响土壤湿度的重要特征后，我们综合调研基于时间序列预测的模型，最终采用 SARIMA、SARIMAX 模型对土壤湿度进行预测。我们主要参考了 11 种经典的时间序列预测方法，分别为自回归、移动平均线、自回归平均线、自回归综合移动平均线、季节性自回归整合移动平均线 (Seasonal Autoregressive Integrated Moving-Average, **SARIMA**)、具有外生回归量的季节性自回归整合移动平均线 (Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors, **SARIMAX**)、向量自回归、向量自回归移动平均、具有外源回归量的向量自回归移动平均值、简单指数平滑 (Simple Exponential Smoothing, SES)以及霍尔特·温特的指数平滑。

经过机理分析可以得到，与土壤湿度相关性较高的特征变量如蒸发量、降水量都是具有季节性因素高度影响的特征，且具有强的周期性特征，因此我们选择使用 SARIMA 对于问题表格中对应时间节点的特征因变量进行季节性自回归预测，从而得到对应时间点的因

变量特征。SARIMA 是季节性整合自回归移动平均模型，将季节差分与 ARIMA 模型相结合的 SARIMA 模型用于具有周期性特征的时间序列数据建模。应用于包含趋势和季节性的单变量数据时，SARIMA 由趋势和季节要素组成的序列构成。SARIMA$(\mathrm{p},\mathrm{d},\mathrm{q})(P,D,Q)_S$主要分为两个部分，前一部分为对应的非季节模型，其参数分别为 p，d，q，其次是季节性模型将周期性因素考虑进来，更适合与周期性变量自回归预测相关的变量，其中季节性模型的参数对应位 P、D、Q，其中 S 是季节性周期的周期长度。其中 p 代表的是去世的自回归阶数，d 代表趋势差分阶数，q 代表的是趋势的移动平均阶数，季节性模型中参数 P 代表的是季节性自回归阶数，D 代表的是季节性差分阶数，Q 代表的是季节性移动平均阶数。主要建模步骤如下所示：

1）观察时间序列的周期，从而确定周期长度 S；

2）通过对非季节差分和周期为 S 的季节差分，从而确定 d 值。目的是消除时间需恶劣的趋势性、周期性和季节性，从而能够有效的保证时间序列的平稳性；

3）通过自相关系数 ACF 和偏自相关系数 PACF 确定非季节性模型 p、q 值，其中 ACF 和 PACF 的计算方式如下：

$$\rho_k = \frac{Cov(X_{t-k}, X_t)}{\gamma_0} = \frac{\gamma_k}{\gamma_0}\left(无偏ACF\right)$$

$$\rho_k = \frac{Cov(X_{t-k}, X_t)}{\gamma_0} = \frac{(N-k)\gamma_k}{\gamma_0}\left(有偏PACF\right)$$

4）通过季节分解的序列图、自相关系数 ACF 和 PACF 确定季节模型的 P、D、Q 值；

5）将选定的 pdq 和 PDQ 的可能值代入 SARIMA 模型；

6）根据池信息准则 AIC 最小值和 Q 检验来选取最优模型，其中 AIC 的计算表达式如下所示：

$$\mathrm{AIC} = \mathrm{nln}\left(\frac{\mathrm{RSS}}{\mathrm{n}}\right) + 2\mathrm{k}$$

其中 RSS 为剩余平方和的计算方式为：

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

5.5.2 基于目标变量回归的 SARIMAX 模型的建立

经过 SARIMA 模型的自回归预测模型得到对应预测时间点降水量、蒸发量等相关性高的特征因子。将降水量、蒸发量等相关特征作为协变量并结合 SARIMAX 模型对 10cm 土壤湿度进行预测。区别于 SARIMA 模型，SARIMAX 不再仅依赖于单变量进行自回归预测，还增加了对已知外部变量进入回归模型的支持，从而增加了时间序列预测模型的先验信息。同时经过对 10cm 土壤湿度、40cm 土壤湿度、100cm 土壤湿度、200cm 土壤湿度进行相关性分析可知，不同深度的土壤湿度之间存在高度的相关性。从机理角度分析也可以推测，土壤湿度在不同深度之间存在一定的渐变关系，因此在进行下一层土壤湿度的预测时，将前几层的土壤湿度加入协变量因子中建立相应的 SARIMAX 模型依次对 40cm 土壤湿度、100cm 土壤湿度、200cm 土壤湿度进行回归预测。

5.5.3 模型的评价指标

我们选择 MSE、MAE、RMSE 和 R2 四项指标进行模型的评估。

MSE（Mean Square Error）是指真实值与预测值的差值的平方然后求和平均。通过平方的形式便于求导，所以常被用作线性回归的损失函数。其计算公式表示如下：

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

MAE（Mean Absolute Error）是绝对误差的平均值。可以更好地反映预测值误差的实际情况。其计算公式表示如下：

$$\mathrm{MAE(X,h)} = \frac{1}{m}\sum_{i=1}^{m}|h(x_i) - y_i|$$

RMSE（Root Mean Square Error）是均方根误差，主要常被用来作为衡量模型预测结果的指标，其计算方法如下：

$$\mathrm{RMSE(X,h)} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2}$$

R2 是决定系数，主要用来反应因变量的全局变化通过模型回归关系被自变量所能够解释的比例，其计算公式如下：

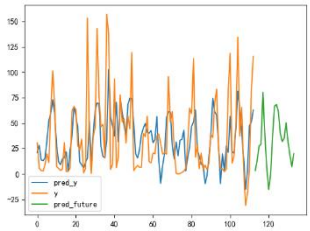
$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - y^2)^2}$$

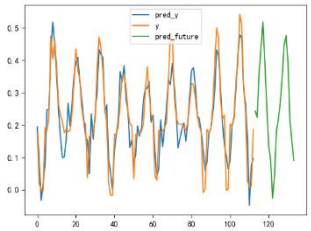
进一步可以简化为：

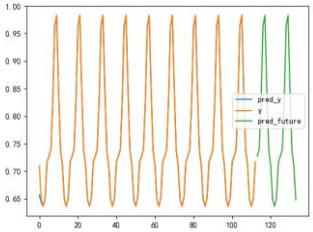
$$R^2 = 1 - \frac{\frac{\sum_i^n (y_i - \hat{y}_i)^2}{n}}{\frac{\sum_i^n (y_i - y^2)^2}{n}} = 1 - \frac{RMSE}{Var}$$

由该公式可以看出对于 R2 之变可以通俗地理解为使用均值作为误差基准，看预测误差是否大于或者小于均值基准误差，综合 RMSE 和 R2 两项指标能更好的评估模型的有效性。

5.6 结果展示







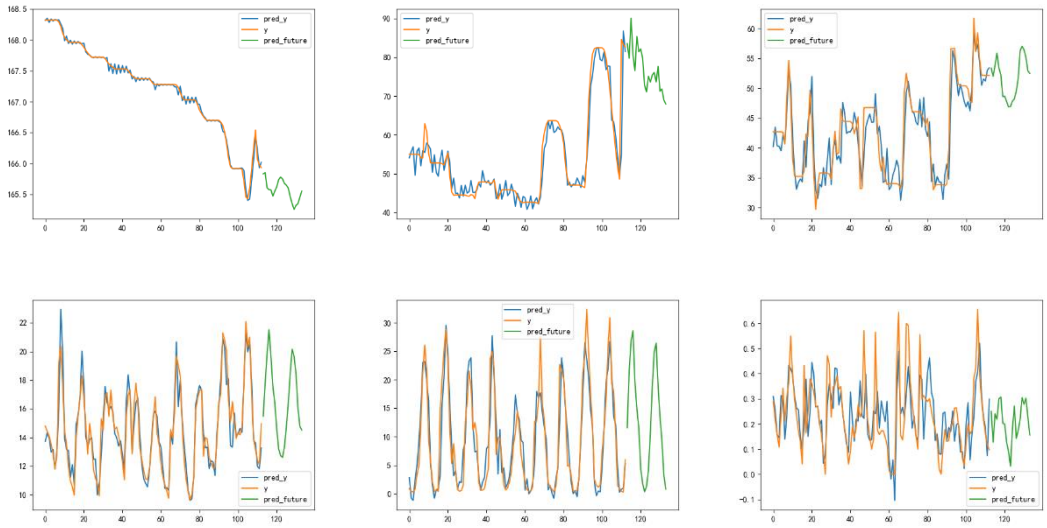


图 5-7 外生变量及土壤湿度变量预测结果图

图 5-7 中，橙色线为真实值、蓝色线为回归拟合值、绿色线为预测值。第一排的第一幅图表示对降水量的自回归时间序列预测的结果，第一排第二幅图表示对植被指数（NDVI）的自回归时间序列预测的结果，第一排第三幅图表示对低层植被（LAIL）的自回归时间序列预测的结果；第二排第一幅图表示对 200cm 深度土壤湿度的预测结果，第二排第二幅图表示对 100cm 深度土壤湿度的预测结果，第二排第三幅图表示对 40cm 深度土壤湿度的预测结果；第三排第一幅图表示对 10cm 深度土壤湿度的预测结果，第三排第二幅图表示对蒸发量的自回归时间序列预测结果，第三排第三幅图表示对径流量的自回归时间序列预测结果。

各个预测结果对应的评价指标如下表所示：

表 5-2 自回归预测评价指标表

预测变量\评价指标	MAE	MSE	RMSE	R2
降水量	22.68	996.03	31.56	-0.83
蒸发量	2.53	12.15	3.49	0.82
径流量	0.08	0.01	0.11	0.17
植被指数	0.05	0.00	0.06	0.77
低层植被	0.00	2.52	0.01	1.00
10cm 土壤湿度	0.74	0.99	0.99	0.88
40cm 土壤湿度	2.35	9.03	3.00	0.80
100cm 土壤湿度	2.11	13.88	3.73	0.89
200cm 土壤湿度	0.04	0.00	0.06	0.99

根据图 5-7 的预测曲线及表 5-2 的预测评价指标可以看出，除了降水量外，其它变量的预测误差都很小，预测效果都很不错。对于降水量，误差大的主要原因一方面是降水量的基础值过大，另一方面可能是影响降水的因素过多，数据中的外生变量并没有考虑全面导致。

根据上述预测结果，得到对于问题二的表格预测如下所示：

表 5-3 土壤不同深度湿度预测结果表

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	15.48	53.35	83.53	165.83

	05	17.76	51.99	79.76	165.85
	06	19.82	53.86	90.03	165.61
	07	21.53	55.88	80.56	165.58
	08	19.70	53.36	76.52	165.58
	09	17.57	52.25	85.36	165.47
	10	16.42	48.60	81.42	165.55
	11	14.63	48.65	82.13	165.62
	12	13.20	47.76	79.75	165.74
2023	01	12.73	46.87	72.79	165.78
	02	12.61	46.94	71.05	165.75
	03	13.31	47.84	75.09	165.68
	04	14.74	48.20	73.61	165.65
	05	16.41	49.51	75.40	165.60
	06	18.44	51.65	76.11	165.46
	07	20.17	56.27	73.63	165.35
	08	19.65	57.02	77.66	165.26
	09	18.46	56.44	71.26	165.32
	10	16.09	55.18	71.82	165.35
	11	14.78	52.90	68.98	165.46
	12	14.51	52.48	67.97	165.55

六、问题三的求解

6.1 问题分析

首先，通过对草原土壤化学性质**相关文献的综合调研**，发现影响土壤化学性质的主要因素包括**气候（温度等）、生物量、土壤湿度**关键因素。因此我们将这三个因素也作为预测土壤化学性质的关键自变量。

根据问题一和问题二的分析可以得到不同的放牧策略（放牧方式和放牧强度）会对锡林郭勒草原土壤物理性质（主要是土壤湿度）和植被生物量产生影响，问题二中更进一步建立了放牧强度和土壤湿度之间的基于时间序列预测的模型，因此结合问题一和问题二的放牧强度对其他因素的影响，我们可以**得到 2022 年土壤湿度的数据**。

对于 2022 年土壤化学性质的预测部分，首先利用并延续问题二的思路使用**时间序列模型**对未来时间点的化学含量进行预测，但由于给定已知的相关数据量较少，因此进一步根据放牧强度和土壤化学性质之间的关系进行直观的**可视化分析**，发现放牧强度和土壤化学性质之间存在非线性关系，因此分析其内在机理，建立了相应的**单变量微分方程**和**多变量微分方程**。其中单变量微分方程根据附录 14 提供的数据进行了初步拟合，多变量微分方程由于数据有限，未进行参数求解。由于可获取的数据有限，单变量微分方程的拟合的结果只能大致进行趋势分析，无法进行准确地预测，因此，对于 2022 年个小区土壤化学成分含量的预测部分采用了更适用于小样本预测的**决策树模型**进行回归预测。

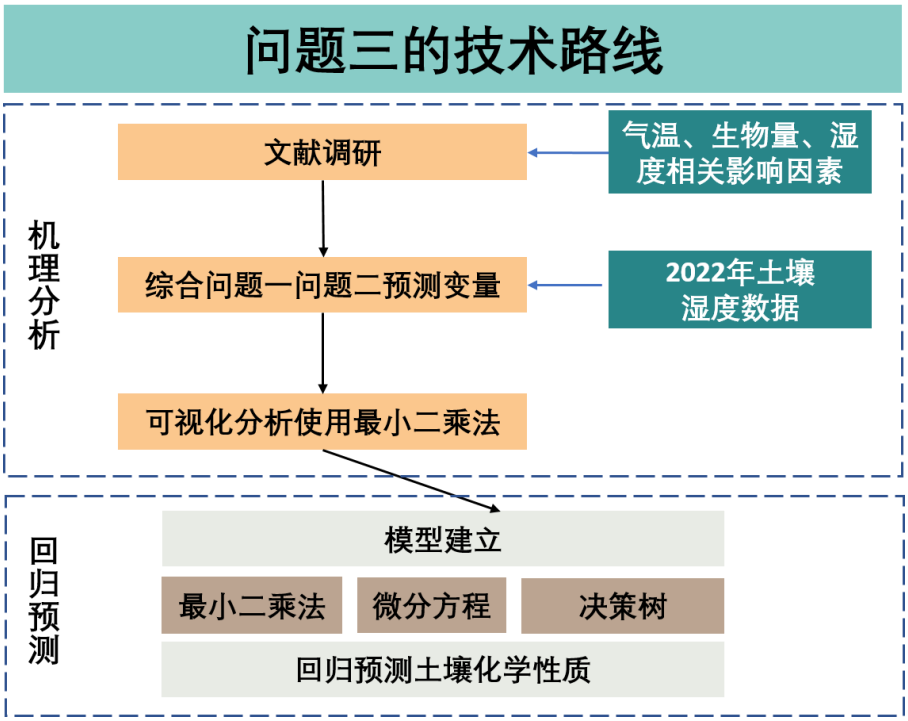


图 6-1 问题三的技术路线

6.2 综合文献进行机理分析

土壤的化学性质，如有机碳（SOC）、总氮（STN）是植物生长的必要养分，同时也在全球氮碳循环中起着重要作用。氮元素是条件陆地生态系统生产量、结构和功能的关键性元素，可以限制群落初级和次级生产量。放牧是人类活动影响草原土壤的重要干扰因素，过度放牧是影响草原生态系统土壤有机碳含量最主要的因素，一方面牲畜的进食减少了植物中的碳向土壤的归还量，另一方面过度放牧可以加速土壤的呼吸作用，导致有机碳的缺失[13]。

同时，也有研究表明，在土壤有机碳估算模型中，地形湿度指数和年均温具有重要影响，在土壤的全氮模型中，植被指数和地形湿度指数是重要因素[14]。地形湿度指数可以捕获土壤水分分布的能力，植被指数则描绘了土壤的植物生长情况，对氮碳循环有着重要指示。而年均温是重要的气候因素，气候指标会影响初级生产力、土壤中化学成分的输入，影响生物活性、凋落物的积累和分解速度，从而影响土壤有机碳和全氮空间分布。

综上，我们根据论文获取影响锡林郭勒草原土壤化学性质的主要影响因素除了放牧强度，还包括气候（温度等）、植被覆盖（生物量）、土壤湿度。

6.3 建立数学模型

6.3.1 特征可视化分析

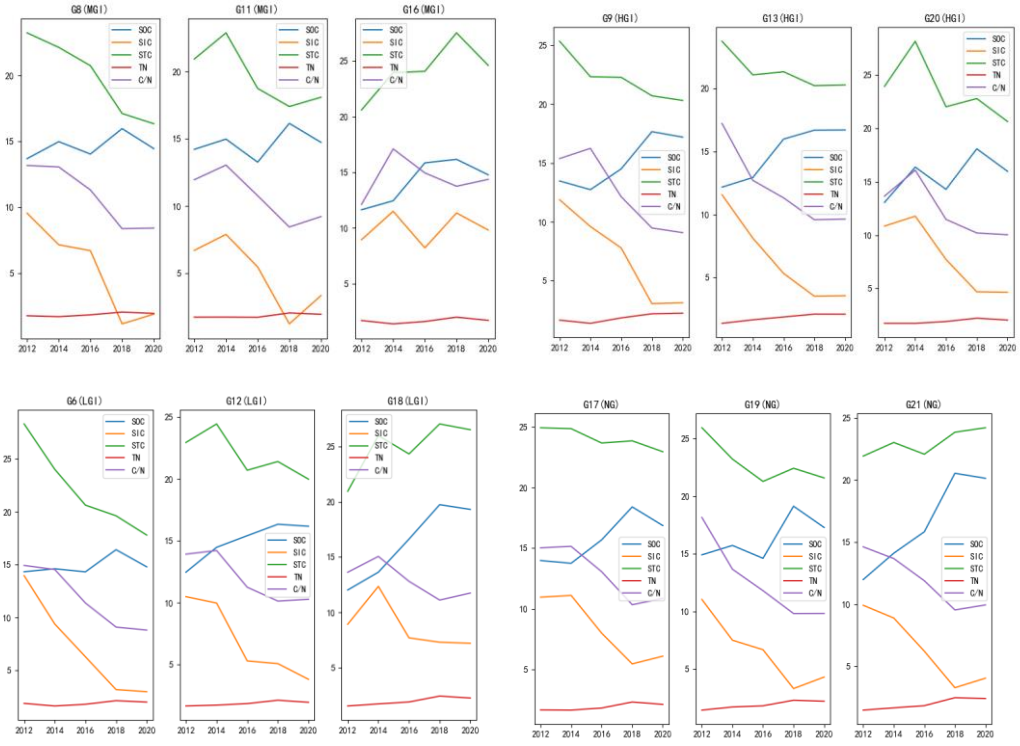


图 6-2 土壤化学性质在不同强度下的变化趋势

为了更好的对问题进行分析，我们将相关特征和土壤化学性质之间的关系进行

了可视化分析进行观察，综合可视化结果以及论文给出的导向，我们建立假设即土壤化学湿度与其他相关的因子之间成非线性关系。

例如极端天气条件（温度过高或过低）对土壤化学性质呈现负面影响，而在中间适度温度下，对土壤化学性质呈现正面促进影响，因此排除特征因子对于土壤化学性质线性关系的可能。因此后续我们将对因变量进行处理并使用数学模型进行求解。

6.3.2 因变量的选取

根据前面结合论文对土壤化学性质相关变量的机理性分析我们筛选出附件中对应的温度、10cm、40cm 湿度（根据采样土壤的位置选择），这三项数据分别取附件对应的年平均值作为最终结果。在求取年平均值之前对特征中的异常值进行预处理，同前面问题中的做法，我们使用箱型线剔除异常值并使用三次样条插值来补上剔除的异常值。

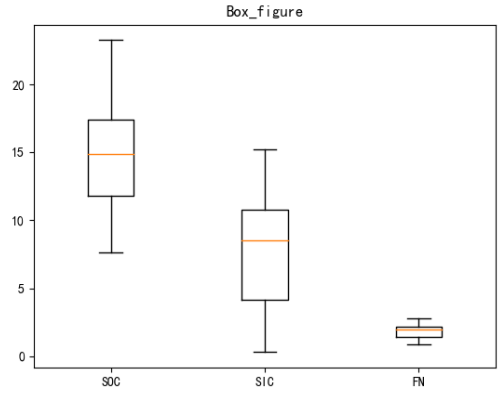
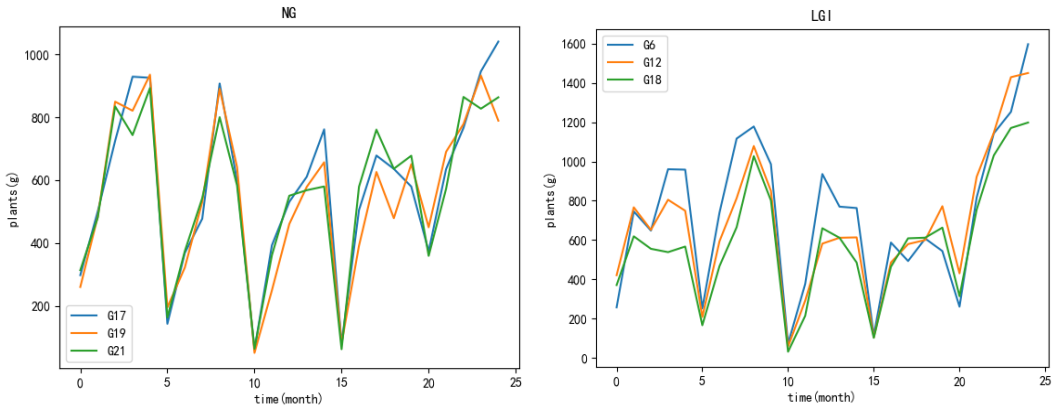


图 6-3 与土壤化学性质相关特征的箱型线图

其中植被覆盖即生物量未选择主要是因为在数据中同一放牧强度的不同小区生物量基本相同（如图 6-4 所示），因此无法从所给数据中衡量生物量对土壤化学性质的影响。



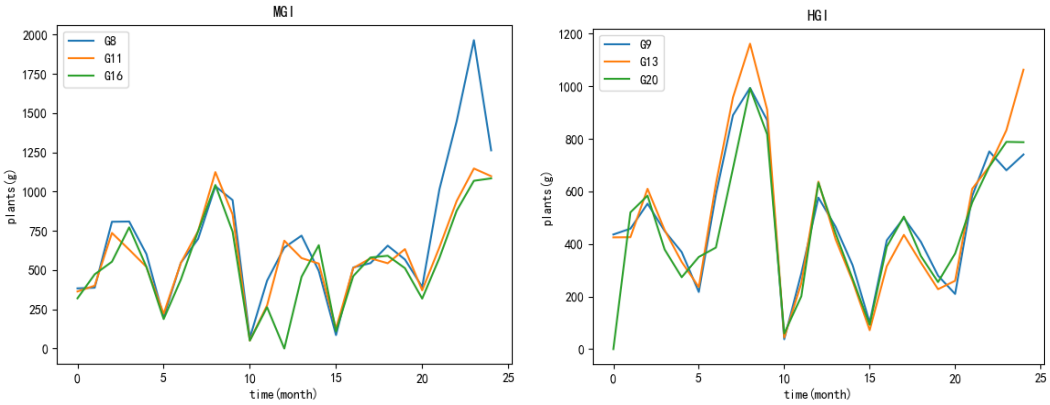


图 6-4 四种放牧强度和生物量的趋势

6.3.3 建立相应的数学模型

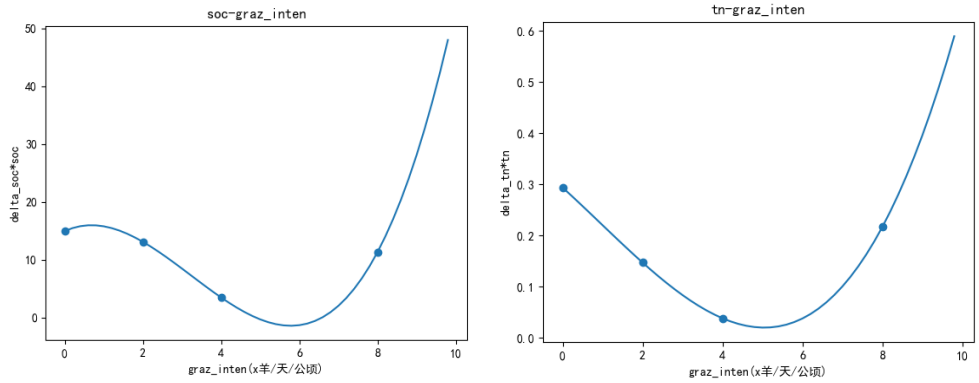
首先使用最小二乘法拟合放牧强度和土壤化学性质，最小二乘法是一种常用的数学优化的技术。主要通过最小误差的平方和寻找与训练数据最佳匹配的函数，并使得这些数据与实际数据点之间的误差最小化。

其基本思想是，另拟合的目标函数为：

$$f(x) = a_1\varphi_1(x) + a_2\varphi_2(x) + \cdots + a_m\varphi_m(x)$$

其中每一组对应的$\varphi_m(x)$是一组线性无关的函数，对应的a_m是线性无关函数的系数。

通过观察所给数据的散点图，大致分析出土壤化学含量的变化量与放牧强度呈现高次非线性关系，同时由于土壤化学成分含量存在一定的饱和值，故当土壤化学成分含量高而接近饱和时，放牧强度对化学成分含量的变化影响会变小，因此建立下述的微分方程关系，并使用最小二乘法拟合数据得到曲线如下所示:



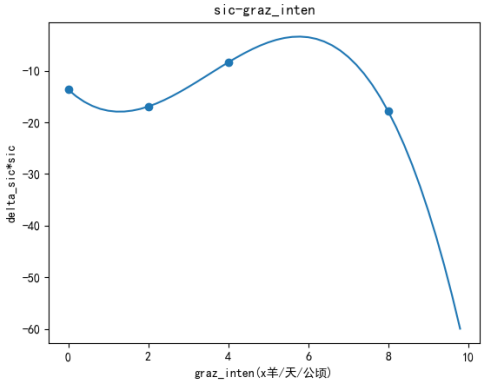


图 6-5 放牧强度和土壤化学性质的关系

其拟合的结果如下公式所示，其中 x 为放牧强度：

$$\frac{dy_{SIC}}{dt} = \frac{-0.32333783x^3 + 3.42005388x^2 - 7.18917494x - 13.65762912}{y_{SIC}}$$

$$\frac{dy_{SOC}}{dt} = \frac{0.26260973x^3 - 2.5478305x^2 + 3.12184967x + 14.9201808}{y_{SOC}}$$

$$\frac{dy_{TN}}{dt} = \frac{0.00147669x^3 - 0.00416676x^2 - 0.07081778x + 0.29414477}{y_{TN}}$$

进一步结合土壤温度、湿度相关因素以及其他因素，其他因素例如由于不同小区的各类数据较少，故将各类小区各自特有因素的影响考虑到拟合方程的常数项中，并对每个小区的 SOC、SIC、TN 指标分别建立多元加权系数预测模型（注：其他两指标 STC、C/N 可由前 3 个指标得出）。因此写出下列方程：

$$\frac{dy_{SIC}}{d_t} = \frac{\sum_{i=1}^{n}(\mu_{3i}x^3{}_i + \mu_{2i}x^2{}_i + \mu_{1i}\mathrm{x}_i) + \mathrm{a}}{y_{SIC}}, \mu_i\text{为}x_i\text{自变量的权重指标，a 为常量}$$

$$\frac{dy_{SOC}}{dt} = \frac{\sum_{i=1}^{n}(\alpha_{3i}x^3{}_i + \alpha_{2i}x^2{}_i + \alpha_{1i}\mathrm{x}_i) + \mathrm{b}}{y_{SOC}}, \alpha_i\text{为}x_i\text{自变量的权重指标，b 为常量}$$

$$\frac{dy_{TN}}{dt} = \frac{\sum_{i=1}^{n}(\beta_{3i}x^3{}_i + \beta_{2i}x^2{}_i + \beta_{1i}\mathrm{x}_i) + \mathrm{c}}{y_{TN}}, \beta_i\text{为}x_i\text{自变量的权重指标，c 为常量}$$

由于获取到的样本数量有限，因此根据上述机理分析得到的上述微分公式并未进行参数拟合。

6.3.4 适用于小样本学习量的决策树模型

考虑到用于 6.3.3 节单变量微分方程参数拟合的数据量过小，其预测效果可能不佳，故本问题对于 2022 年土壤化学成分含量的预测采用了适用于小样本情况下的决策树回归模型进行预测。

在决策树中，每一次进行节点的分裂主要是依据属性对于决策树的信息增益。信息增益是建议在信息熵上的一个概念，主要用来衡量一个属性来划分数据样本的能力。

信息熵主要用于描述事物的不确定性程度，假设存在样本集合 X 共有 N 个对应的类

别，设定第 m 个类别样本所占全局样本比例为p_m，则对应的样本数据集 X 的信息熵为：

$$\mathrm{H(X)} = -\sum_{k=1}^{N} p_m log_2 p_m$$

信息熵主要用于描述在事件发生之前可能对信息量的期望，主要是对于信息不确定性的一种描述方式，当样本对应的信息熵越大其样本的不确定就越大，样本的种类较为繁杂，反之则说明样本的纯度较高。

信息增益是基于信息熵的一种概念，主要用来描述一个样本属性用来区别样本类别的能力，信息增益越大则样本属性更容易区分样本类型，对应的决策树也就更加简洁。信息增益的计算公式如下：

$$\mathrm{IG(Y|X)} = \mathrm{H(Y)} - \mathrm{H(Y|X)} \geq 0$$

在决策树的结构中主要组成由决策节点、叶子节点、决策树的深度，是一种基于信息增益建立的一种常用模型，是直观的运用概率学相关知识分析问题的一种图解法，由于决策树的分支最终形成的图形像树的枝干故称为决策树。其中：

1）决策节点：代表通过需要通过条件的判断从而进行树干分支的节点。比如将样本进行分类或回归时，需要将样本的对应属性值和决策节点的值进行比较从而判断选择分支的方向。

2）叶子节点：即没有子节点的节点，也代表最终的决策结果。

3）决策树的最大深度：对应决策树的最大层的次数。从第一层开始根节点层数定为 0，从下面起每一层子节点层数+1。

其算法流程对应如下：

```
If so return 类标签：
Else
    寻找划分数据集的最好特征
    划分数据集
    创建分支节点
        for 每个划分的子集
            调用函数 createBranch()并增加返回结果到分支节点中
        return 分支节点
```

图 6-6 决策树算法流程

使用决策树模型对土壤化学性质进行预测，结果在下个小结展示。

6.4 结果与分析

我们根据问题二中预测出的 2022 年 10cm、40cm 湿度，并通过 SARIMA 预测 2022 年平均温度，得到所有自变量，带入预测模型进行预测。因为数据样本量较少因此使用决策树进行小样本数量的回归问题分析，并对 2022 年不同区域小区相应的土壤化学性质进行预测，其预测结果如表 6-1 所示。

表 6-1 2022 年不同区域小区相应的土壤化学性质预测结果

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	16.86	6.09	22.95	2.12	10.85
	G19	17.30	4.30	21.60	2.20	9.83
	G21	20.13	4.06	24.20	2.43	9.95
LGI	G6	14.79	3.00	17.79	2.01	8.82
	G12	16.19	3.80	19.99	1.94	10.28
	G18	19.31	7.21	26.51	2.26	11.75
MGI	G8	14.44	1.89	16.33	1.94	8.41
	G11	14.74	3.35	18.10	1.96	9.23
	G16	14.78	9.82	24.60	1.71	14.36
HGI	G9	17.20	3.12	20.32	2.24	9.08
	G13	16.69	3.56	20.26	2.10	9.63
	G20	15.97	4.66	20.63	2.05	10.04

七、问题四的求解

7.1 问题分析

根据问题四要求，需要给出沙漠化程度指数的预测模型，并利用该模型计算监测点的沙漠化程度，其次对监测点的土地板结化进行定义，最后给出使沙漠化程度和土地板结化程度最小的放牧策略模型。

由于监测点部分数据的缺失，本文先通过层次分析法得到因子权重系数，结合附件 2 中所给全盟数据得到对应的因子强度，之后计算整体沙漠化程度预测模型参数。利用监测点的经纬度可以得到监测点的近似气象数据，再结合附件 13 给出的数据，可以用整体沙漠化程度预测模型近似预测监测点在不同放牧强度下的沙漠化程度。

虽然目前并没有精确的土地板结化定量表达式，但是本文通过分析土壤湿度、容重以及有机物含量等要素对土地板结化的影响，并结合实际数据，给出了合理的土地板结化定义。

基于以上分析，本文给出了：

1) 锡林郭勒盟整体沙漠化程度预测模型
2) 监测点沙漠化程度
3) 监测点土地板结化定量分析
4) 最优放牧策略模型

其技术路线如下图所示：

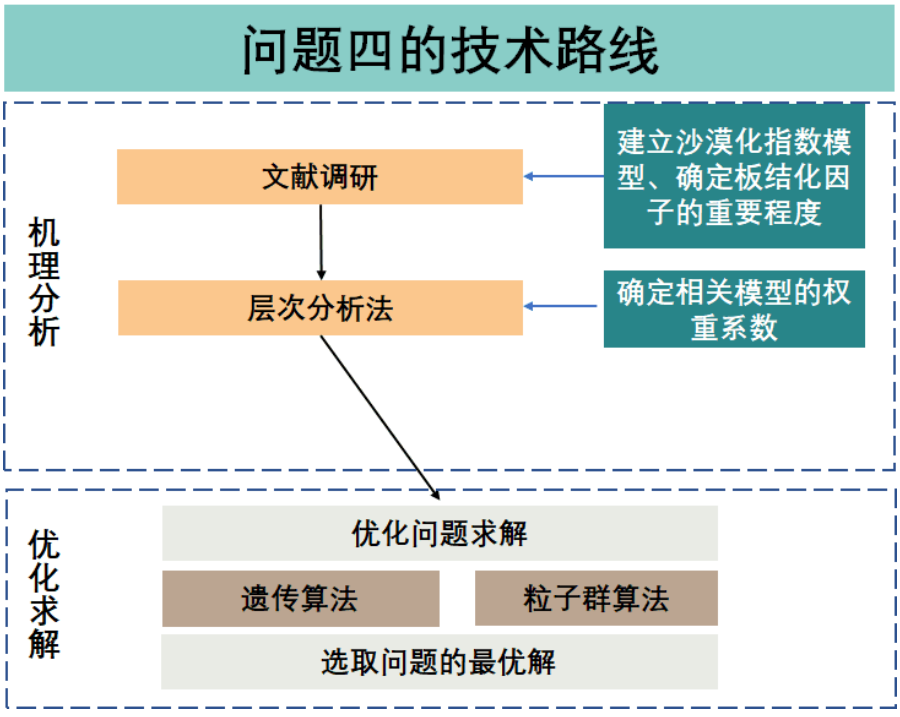


图 7-1 问题四的技术路线图

7.2 锡林郭勒盟整体沙漠化程度

一个地区的沙漠化程度受到各种各样的因素影响，通常地，可以将影响因素分为三类：气象因素、地表因素、人文因素[11]。气象因素主要包括风速、降水和气温，风是沙漠形成的动力，一定等级的风可以吹起沙物质；降水与沙漠化的发生和发展密切相关，降水量的减少可以导致植被生长受限；气温也是沙漠化发展的原因之一，高温会导致强烈的蒸发，

低温则会导致植物生长受限，植物的光合作用只有在最适宜的温度下才能发挥到最大限度。地表因素主要包含植被盖度、地表水资源以及地下水位，植被是土地的保护层，可以抑制地表沙粒的流动；地表水资源可以为生态环境提供可持续发展的前提，有利于人们生活和植物生长；地下水位的加深会使地表土壤退化，影响植物生长。人文因素主要包含人口数量、牲畜数量和社会经济水平，农牧地区的人数直接决定了人类经济活动对土地资源、植被资源的利用强度；牲畜数量是影响草场生态平衡的主要指标，过多的牲畜数量，会导致草场利用强度超过其自身承载力，造成退化；社会经济水平可以反映沙漠化程度，一般条件下，沙漠化灾害的风险等级随社会经济发展水平呈指数上升。

7.2.1 数据收集与处理

对于气象数据，鉴于锡林浩特市是锡林郭勒盟的盟府，位于锡盟中部，因此可以用该地区的年均风速、降水量和气温来表示锡林郭勒盟整体的气象数据平均值。

对于地表数据，由于草原的面积远远大于森林和耕地的面积，植被盖度的计算主要考虑草原，计算公式如下：

$$\text{植被盖度} = \frac{\text{草原面积}}{\text{行政区划面积}} \times 100\%$$

地表水资源可以根据历年年鉴中的表 1-1 直接计算，计算公式如下：

$$\text{地表水资源量} = \frac{\text{水资源总量}}{\text{行政区划面积}}$$

而地下水位由于数据的缺失，替换为水域及水利设施用地，该数据可以反映该地区的水资源利用程度和水资源丰富度。

对于人文数据，人口密度可以反映人口数量的变化，计算如下：

$$\text{人口密度} = \frac{\text{人口数量}}{\text{行政区划面积}}$$

同样的，牲畜数量通过牲畜密度间接反映，需要注意的是，不同的牲畜种类的食量不同，可以通过折算系数化为羊单位，计算公式如下：

$$\text{牲畜密度} = \frac{6 \times \text{成年大牲畜数量} + 3 \times \text{大牲畜幼崽数量} + \text{羊数量}}{\text{行政区划面积}}$$

在沙漠化灾害研究中常用人均纯收入这一指标来反映灾害危险度中社会经济水平方面的内容[11]，可以根据下面的公式进行计算：

$$\text{人均纯收入} = \text{人均总收入} - \text{人均总支出}$$

总结附件 2 中 2016-2021 年锡林郭勒盟的数据，如表 7-1 所示。

表 7-1 锡林郭勒盟各因素

年份 / 因素	2016	2017	2018	2019	2020	2021
风速（m/s）	3.1	3.2	3.2	3.3	3.2	3.2
降水（mm）	412.8	309	168.7	277.6	293.5	389.7
气温（^oC）	3.9	3.5	4.7	4.3	4.4	3.3
植被盖度（%）	95.3	95.3	88.6	95.4	95.4	95.4
地表水资源（$10^4m^3/km^2$）	1.724	1.639	1.641	1.641	1.641	1.601

水域及水利设施用地（km^2）	4866	4866	4866	4866	738	738
人口密度（$person/km^2$）	5.146	5.167	5.191	5.206	5.224	5.473
牲畜密度（$unit/km^2$）	116.39	124.02	111.11	105.62	109.52	113.57
人均纯收入（$yuan$）	404	614	297	1404	3700	8735

因子强度的确定需要上限与下限，一般文献需要根据实地考察确定经验范围，本文由于缺乏实地考察数据，采取均值与标准差的方式确定上下限，根据正态分布落在$[\mu-\sigma,\mu+\sigma]$的概率为 65.26%，可以得到上下界的计算公式如下：

$$\text{下限} = \text{均值} - \text{标准差}$$

$$\text{上限} = \text{均值} + \text{标准差}$$

计算得到的上下界如表 7-2 所示。

表 7-2 因子的下限和上限

因子	下限	上限	因子	下限	上限
风速（m/s）	3.1368	3.2632	水域及水利设施用地（km^2）	1358	5622
降水（mm）	221.15	395.95	人口密度（$person/km^2$）	5.1144	5.3546
气温（^oC）	3.4711	4.5623	牲畜密度（$unit/km^2$）	107.01	119.74
植被盖度（%）	0.9147	0.9699	人均纯收入（$yuan$）	-770	5822
地表水资源（$10^4m^3/km^2$）	1.6073	1.6884			

7.2.2 模型参数计算

本文将因素分为两种，正比关系和反比关系的因素。对于正比关系的因素，其值越大，沙漠化程度越重。风速、人口密度、牲畜密度以及人均纯收入属于正比关系因素，其因子强度计算公式如下：

$$Q=\begin{cases}0，\text{值}<\text{下限}\\ \dfrac{\text{值}-\text{下限}}{\text{上限}-\text{下限}}，\text{下限}\le\text{值}<\text{上限}\\ 1，\text{值}\ge\text{上限}\end{cases}$$

对于反比关系的因素，其值越大，沙漠化程度越轻。降水、气温、植被盖度、地表水资源和水域及水利设施用地属于反比关系因素，其因子强度计算公式如下：

$$Q=\begin{cases}1，\text{值}<\text{下限}\\ \dfrac{\text{上限}-\text{值}}{\text{上限}-\text{下限}}，\text{下限}\le\text{值}<\text{上限}\\ 0，\text{值}\ge\text{上限}\end{cases}$$

因子权重系数的确定是一个复杂的过程，由于缺少锡林郭勒盟长期的观察数据，本文依照文献[1]给出的层次分析法（AHP，Analysis Hierarchy Process）来确定各个因子权重系数。层次分析法中判断矩阵的构造需要长期的实地考察以及经验数据，本文未能收集到锡林郭勒盟的对应数据，故直接采用文献[1]中给出的判断矩阵，并在之后的小节对其进行修正。

在得到判断矩阵之后，就可以进行权重的计算，先对判断矩阵进行特征值分解，找出最大特征值λ_{max}以及其对应的特征向量W，对特征向量进行标准化后就可以得到该层对于上一层的权重。

不过并不是所有判断矩阵都是可行，在采用该特征矩阵之前，需要对其进行一致性检验，判断标准是$CR \leq 0.1$，具体计算公式如下：

$$CI = \frac{\lambda_{max} - n}{n - 1},$$

$$CR = \frac{CI}{RI},$$

式中，n为判断矩阵维度，$RI = 0.58$为随机一致性指标。判断矩阵的构建与一致性检验如表 7-3-7-6 所示。

表 7-3 $A - B$层判断矩阵、权重及一致性检验

A	B_1	B_2	B_3	权重W	一致性检验
B_1	1	1	1	0.3275	$\lambda_{max} = 3.05$
B_2	1	1	2	0.4126	$CI = 0.025$
B_3	1	1/2	1	0.2599	$CR = 0.0431$

表 7-4 $B_1 - C$层判断矩阵、权重及一致性检验

B_1	C_1	C_2	C_3	权重W	一致性检验
C_1	1	2	3	0.5503	$\lambda_{max} = 3.02$
C_2	1/2	1	1	0.2404	$CI = 0.01$
C_3	1/3	1	1	0.2093	$CR = 0.0172$

表 7-5 $B_2 - C$层判断矩阵、权重及一致性检验

B_2	C_4	C_5	C_6	权重W	一致性检验
C_4	1	2	2	0.4934	$\lambda_{max} = 3.05$
C_5	1/2	1	1/2	0.1958	$CI = 0.025$
C_6	1/2	2	1	0.3108	$CR = 0.0431$

表 7-6 $B_3 - C$层判断矩阵、权重及一致性检验

B_3	C_7	C_8	C_9	权重W	一致性检验
C_7	1	1/2	1/2	0.1958	$\lambda_{max} = 3.05$
C_8	2	1	2	0.4934	$CI = 0.025$
C_9	2	1/2	1	0.3108	$CR = 0.0431$

对上述层次进行因子权重合成，可以得到表 7-7 所示权重系数。

表 7-7 沙漠化程度中因子的权重系数

	权重系数W_A	影响因素B_i	权重系数W_{B_i}	指标因子C_i	权重系数W_{C_i}
沙漠化程度指标体系	1	气象因素B_1	0.3275	风速	0.1802
				降水量	0.0787
				气温	0.0685
		地表因素B_2	0.4126	植被盖度	0.2036
				地表水资源	0.0808
				水域及水利设施用地	0.1282
		人文因素B_3	0.2599	人口密度	0.0509
				牲畜密度	0.1282
				人均纯收入	0.0808

虽然层次分析流程中$CR \leq 0.1$均成立，该预测模型对于塔里木盆地的沙漠化预警具有良好的效果，但是将此模型应用至锡林郭勒盟的沙漠化程度的计算仍旧需要对模型参数进行修正。

7.2.3 沙漠化程度计算以及模型参数修正

根据前述预测模型参数，可以计算得出各个年份锡林郭勒盟的沙漠化程度，公式如下：

$$SM = \eta \sum_{i=1}^{9} W_{C_i} Q_i,$$

式中，$\eta = 0.5$为修正系数，用来描述塔里木盆地与锡林郭勒盟之间的差异，鉴于塔里木盆地的沙漠占地较多，而锡林郭勒盟草原面积占地较广，沙漠化程度较小，所以η的取值小于 1。

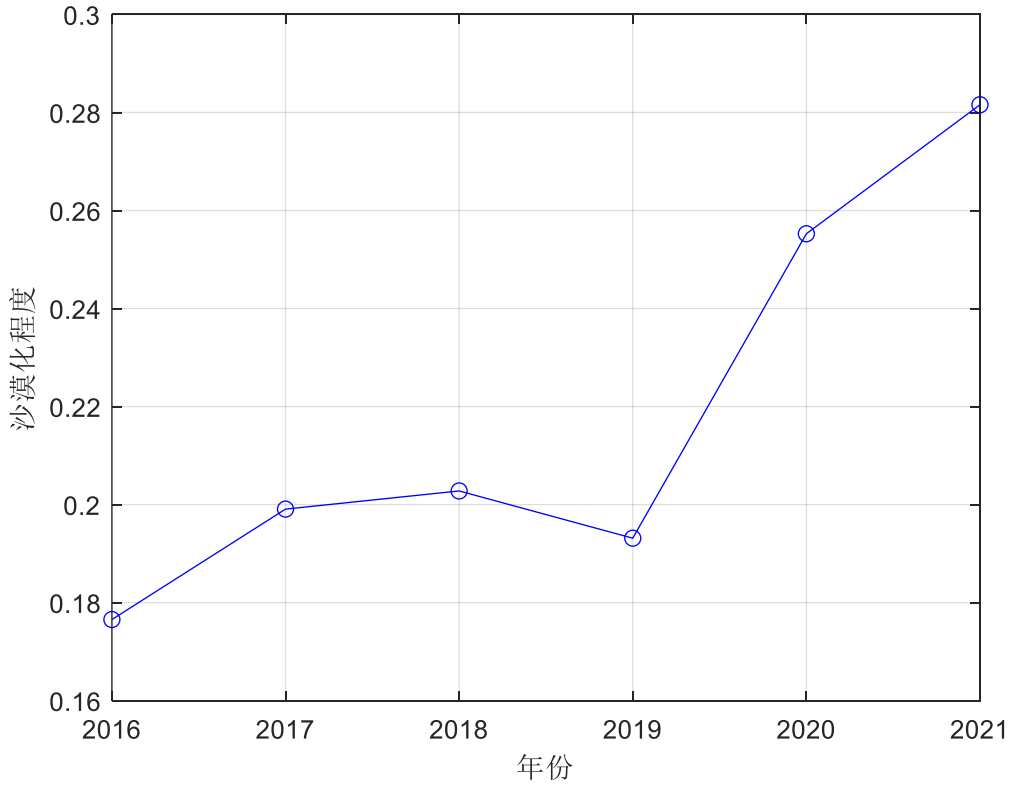


图 7-2 锡林郭勒盟年份-沙漠化程度示意图

从图 7-2 可以看出，2016 年锡林郭勒盟的沙漠化程度最低，之后随着年份小幅增长，一直到 2018 年，之后 2019 年略有下降，之后一直到 2021 年飞速上涨。究其原因，可以解释为锡林郭勒盟 2016 年至 2019 年之间生产方式和生活水平相似，但是 2020 年之后，人口的涌入以及牲畜数量的上升导致了经济飞速发展，此类人类活动加大了沙漠化程度。

7.3 不同放牧强度下监测点沙漠化程度

对于监测点来讲，其数据与锡林郭勒盟的整体数据稍有不同，但是依旧在其附近，因此本文采用上一小节得出的整体预测模型对监测点的沙漠化程度进行近似预测。

7.3.1 数据收集与处理

对于气象因素，根据监测点的经度 115°47′和纬度 43°44′可以得到其地图上的精确位置，如图 7-2 所示，红圈所在位置为监测点，可以看出与之相近的区城市或者区域为锡林浩特、阿巴嘎旗、苏尼特左旗、镶黄旗、正向白旗和正蓝旗，因此，本文用这六个城市或者区域的平均值来得到监测点的气候因素。

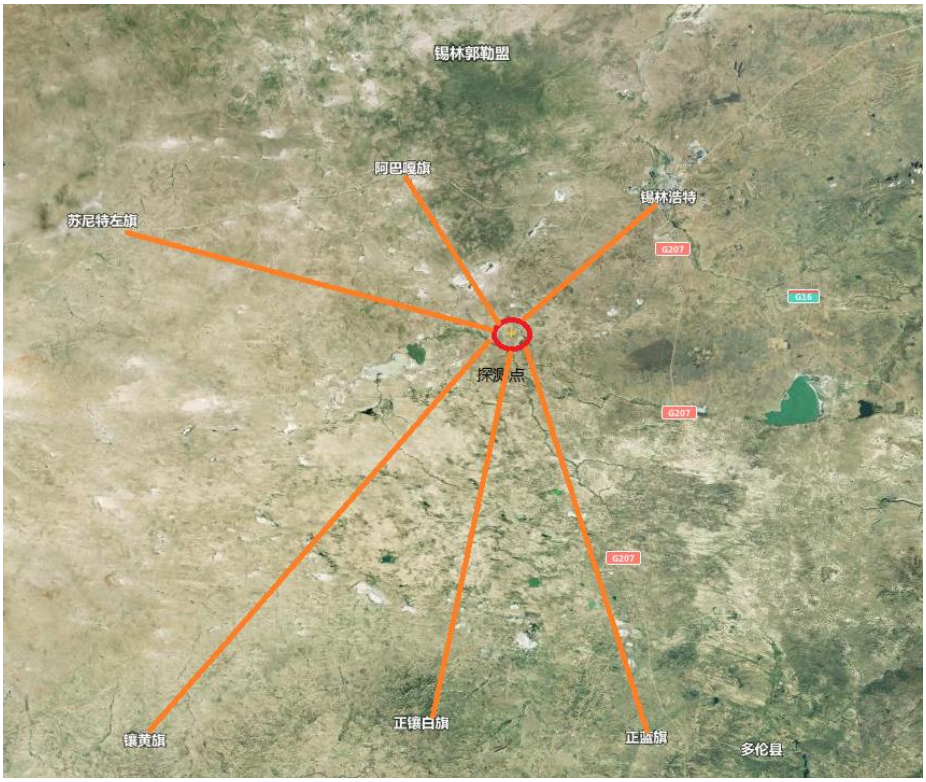


图 7-3 监测点经纬度位置示意图

对于地表因素，本文用锡林郭勒盟的整体数值来代替，会造成有限的可接受的误差。

对于人文因素，本文根据附件 13 所给数据以牧户为单位进行计算。值得注意的是，由于数量样本较少，且具有特殊性，与全盟的平均数据差异较大，所以全盟人文因素的上下限不再适用，为了解决该问题，本文将牧户的人文因素根据均值和方差等尺度映射到全盟人文因素的尺度下。具体变换方法可以如下表示：

$$\bar{x} = \frac{\mu_1}{\mu_2} x$$

其中$\bar{x}$表示尺度变换后的值，x表示尺度变换之前的值，μ_1表示全盟人文因素均值，μ_2表示牧户人文因素均值。

通过上述数据收集和处理，得到的各牧户每年的因素如表 7-8 所示。

表 7-8 各牧户每年的因素

因素 \ 年份			2018	2019	2020
气象因素	风速（m/s）		3.33	3.63	3.42
	降水（mm）		206.37	322.85	285.15
	气温（^oC）		4.78	4.32	4.5
地表因素	植被盖度（%）		88	95.4	95.4
	地表水资源（$10^4m^3/km^2$）		1.641	1.641	1.641
	水域及水利设施用地（km^2）		4866	4866	738
人文因素	人口密度（$person/km^2$）	牧户 1	0.83	0.83	0.83
		牧户 2	1	1	1
		牧户 3	0.5	0.5	0.5
		牧户 4	0.47	0.47	0.47
		牧户 1	275.83	272.22	256.39

	牲畜密度（$unit/km^2$）	牧户 2	230	210	200
		牧户 3	601.75	492	492
		牧户 4	245.9	245.9	255.7
	人均纯收入（$10^4 yuan$）	牧户 1	10.56	15.54	12.86
		牧户 2	21.95	24.195	31.175
		牧户 3	38.925	42.3375	44.8375
		牧户 4	3.33	3.33	3

7.3.2 沙漠化程度计算

将表 7-8 中数据代入沙漠化程度预测模型进行计算，可以得到不同牧户和不同年份的沙漠化程度，如图 7-4 所示。

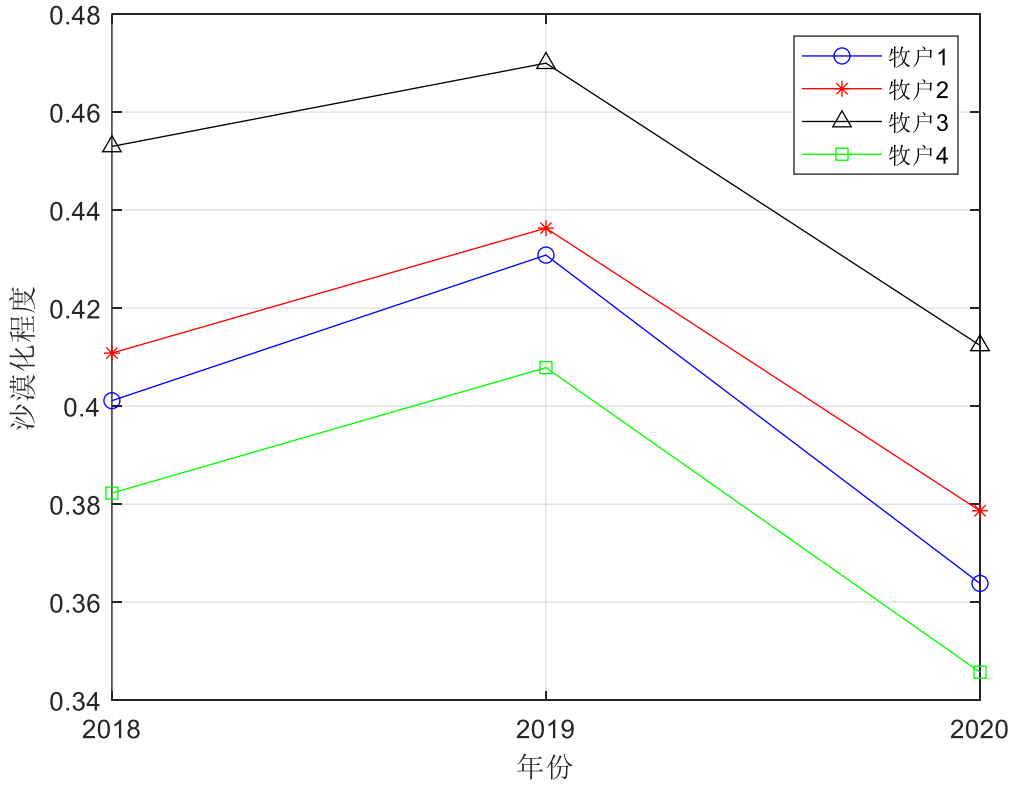


图 7-4 不同年份不同牧户沙漠化程度示意图

可以看出，2018 年到 2020 年，该地区的沙漠化程度呈现先上升后下降的趋势，归结于更加科学的放牧方式。牧户 3 的放牧强度最高且收入最高，所以其沙漠化程度也最高，与之相对，牧户 4 的放牧强度较低，收入最少，其沙漠化程度也最低。牧户 2 和牧户 3 的沙漠化程度相近，介于牧户 3 和牧户 4 之间。

为了更清楚地展示不同放牧强度对于沙漠化程度的影响，将同一年内不同放牧强度作为横坐标，沙漠化程度作为纵坐标进行画图。

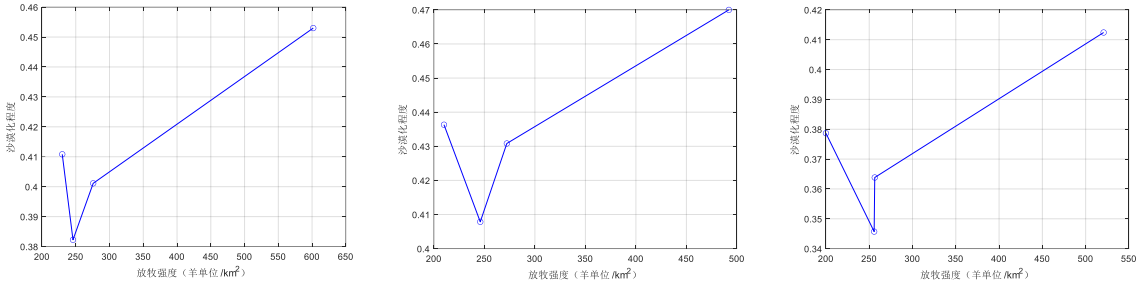


图 7-5 2018-2020 年放牧强度-沙漠化程度示意图

图 7-5 给出了 2018-2020 年沙漠化程度随着放牧强度的变化示意图，可以看到每一年的趋势都相似，沙漠化程度随着放牧强度先下降，再上升。其原因可以归结为适当的放牧，通过牲畜的践踏，造成枯落物分解，可以改善土地质量，提高草原生物的多样性，从而抑制沙漠化。一旦放牧强度过大，超过草原承受能力，植被的生长周期跟不上牲畜的食用，就会造成草原退化，加剧沙漠化。

7.3.3 沙漠化程度定量分析

为了计算最优放牧策略，需要对不同放牧强度下的沙漠化程度进行定量分析。如果改变放牧强度，受影响的因素有植被盖度，收入以及牲畜密度。放牧强度越大，牲畜密度越高，收入越高，而植被盖度先升高再降低。为了便于描述，将放牧强度I归一化到 0 和 1 之间，牲畜密度的强度因子Q_8与归一化的放牧强度相等，即：

$$Q_8 = I$$

植被盖度的强度因子Q_4可以用下式表示：

$$Q_4 = 2 \times (0.5 - I)^2 - 0.2$$

人均纯收入的强度因子Q_9可以用下式表示：

$$Q_9 = \frac{1}{\gamma(1 - I) + 0.1},$$

式中γ为不同牧户的放牧变现能力系数。

7.4 监测点土地板结化定量分析

土壤板结是指土壤表层因结构不良，有机质含量不足，在灌水或者降雨等外因作用下结构破坏、土料分散，而干燥后受内聚力作用使土面变硬的现象，是农业上经常碰到的问题。土地板结化与土壤有机物、土壤湿度和土壤的容重有关，当土壤湿度下降、有机物含量减少、土壤容重增大，土地的孔隙度减少，板结化程度就会加重。土壤板结化的危害有很多，例如延缓有机质的分解，土壤的理化性质逐渐恶化，地力逐渐衰退，影响植物的正常发育。为了预防土壤板结化，对其进行定量分析十分重要。

7.4.1 数据收集与处理

土壤湿度数据和有机物含量分别参照附件 3 和附件 14，但是土地容重并没有现成的数据，需要进行推算。本文根据公开数据库的土壤质地分布确定监测点的土壤质地构成，推算出大致的土壤容重。

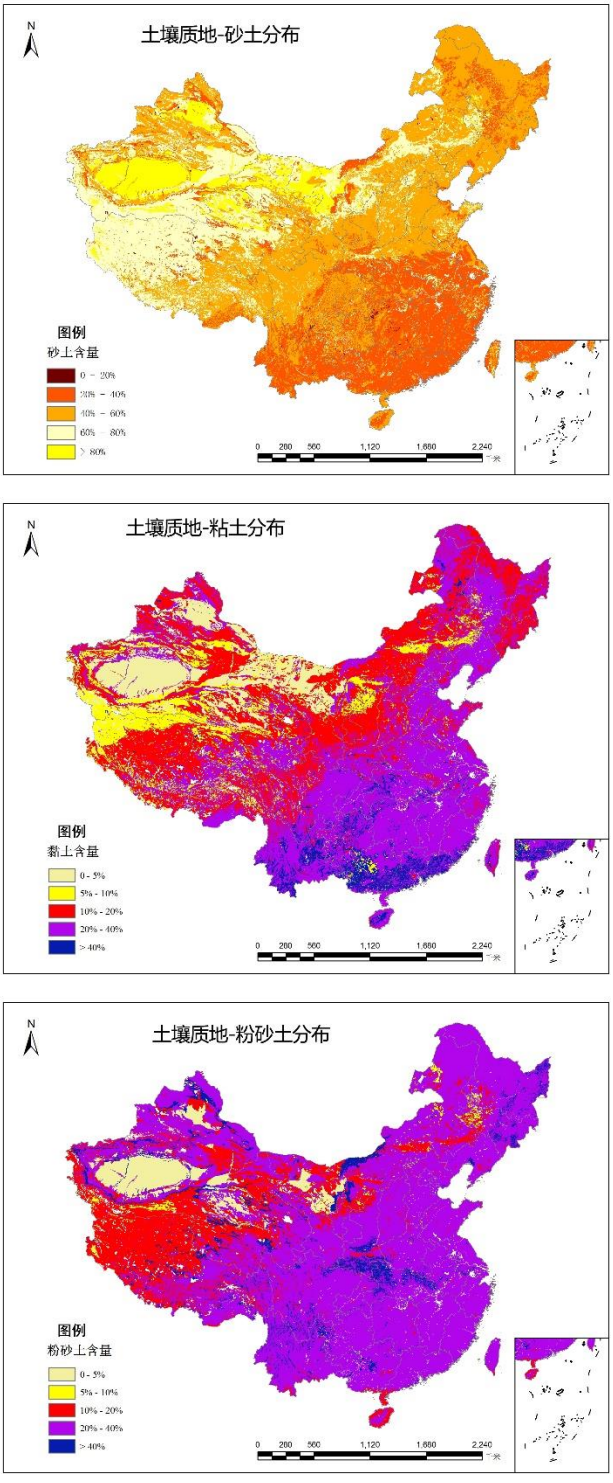


图 7-6 中国土壤质地分布数据[12]

根据图 7-6，可以得到检测点的土质大致由 50%的砂土、20%的粘土和 30%的粉砂土构成，根据表 7-9 所给的数据可以计算得到监测点的大致土壤容重，如下：

$$C = 1.5 \times 50\% + 1.2 \times 20\% + 1.4 \times 30\% = 1.41 g/cm^3$$

表 7-9 各个土质的经典容重

土质	容重（g/cm^3）
砂土	1.5
粘土	1.2

粉砂土	1.4

7.4.2 板结化程度定量分析

土壤板结化的三个影响因素中，土壤容重是最重要的因素，其反映了土壤的质地情况，而土壤的质地是造成土壤板结化的本质原因。土壤板结多出现在土壤质地黏重、黏土中的黏粒含量较多的土壤中，因为其质地比较黏厚，土壤中毛细管孔隙含量十分稀少，从而土壤本身的通气、透水、增温等性能都比较差。这样的土壤团粒结构很容易遭到破坏，造成土壤表层结皮。土壤有机质含量是影响土壤板结化的次要因素，土壤肥力和土壤团粒结构的一个重要指标，有机质含量决定着土壤理化性质的表现，有机质含量偏低，土壤理化性质就会变差，抑制微生物的活性，从而影响了土壤团粒结构的形成，造成土壤的酸碱性不平衡，导致土壤板结问题发生。表层土壤（0-20cm）湿度是影响土壤板结化最小的因素，它决定了植物的水分供应状况。表层土壤湿度降低，土壤中的水分供应不足，植物根部无法吸收足够的水分，影响植物的正常生长，同时也影响微生物的活动，致使土壤有机物含量降低，从而促进土壤板结化。综上所述，给出土壤板结化与土壤容重、土壤有机物和土壤湿度之间的符号关系表达式：

$$B = a_1 Q(C)^{k_1} + a_2(1 - Q(O)^{k_2}) + a_3(1 - Q(W)^{k_3})$$

式中，a_i与k_i(i=1,2,3)分别为权重系数和幂次，取$k_1 = 3, k_2 = k_3 = 1$；O 为土壤有机物含量，通常可以通过 1.724 倍的土壤有机碳含量来衡量；W 通常可用土壤 10cm 深度的湿度来表示。函数 Q 为影响因子强度的计算公式，详情见 7.2.2，三个影响因子的下限与上限如表 7-10 所示：

表 7-10 因子的下限和上限

因子	下限	上限
土壤容重 C(kg/dm^3)	1.0	2.0
土壤有机物含量 O(g/kg)	10	70
土壤湿度 W(kg/m^2)	10	60

本文依照层次分析法（AHP）来确定各个因子的权重系数，依据上述土壤容重、土壤有机质含量以及土壤湿度对于土壤板结化的重要程度，构造判断矩阵并进行一致性检验，具体如表 7-11 所示：

表 7-11 判断矩阵、权重及一致性检验

指标	C	O	W	权重系数	一致性检验
C	1	3	4	0.6232	$\lambda_{max} = 3.0183$
O	1/3	1	2	0.2395	$CI = 0.0091$
W	1/4	1/2	1	0.1373	$CR = 0.0176$

最终，我们可得定量的土壤板结化定义：

$$B = 0.6232(C-1)^3 - \frac{0.2395(O-10)}{60} - \frac{0.1373(W-10)}{50} + 0.3768$$

7.5 最优放牧策略模型求解

得到沙漠化程度和板结化程度的定量公式后，就可以对放牧策略进行优化，抑制草原的退化。

7.5.1 问题提出

为了使沙漠化程度和板结化程度最小，需要优化放牧强度I，得出最优的放牧策略，于

是可以提出以下问题：

$$\min_{I} SM + B$$

$$s.t.\ 0 \leq I \leq 10$$

7.5.2 遗传算法介绍

遗传算法（GA，Genetic Algorithm）作为普遍使用的智能优化算法，是从自然进化原则“自然选择，适者生存”中启发而来，它把优化问题模拟成为了自然界生物进化过程，从种群的初始状态开始搜索，伴随着种群的迭代，通过选择、交叉、变异等一系列操作对染色体不断进行最优筛选，从而得到最优的染色体，即为求解问题的最优解。遗传算法有着自适应、随机和高度并行的特点，本文利用遗传算法求解最佳放牧策略，其算法流程如下：

- Step1：在可行域内随机生成一组染色体作为初始种群P_0，每一染色体对应一个可行解；
- Step2：计算种群P_i中的每个个体的适应度值（Fit）；
- Step3：判定是否满足收敛条件，若满足则停止计算，输出结果，否则 i=i+1，执行下一步；
- Step4：依据 Step2 中求得的适应度值，按照一定的规则或方法，选择出一些优良个体进化到下一代；
- Step5：从当前种群中随机选出成对个体，以预先设定的交叉概率P_d在选定的交叉位置交换两个个体的染色体，由此产生新的个体并将其并入到当前种群中。
- Step6：对于种群中所有个体，以预先设定的变异概率P_m将个体中的某些基因改为其他基因，得到下一代种群P_{i+1}，然后转回步骤 2。

7.5.3 粒子群优化算法介绍

粒子群优化算法（PSO, Particle Swarm optimization）可以求解优化问题，具有收敛速度较快、原理通俗易懂、设置参数较少、算法结构简单、易于工程实现等优点，本文利用粒子群优化算法求解最佳放牧策略。

粒子群优化算法起源于复杂适应系统，主要研究具有适应性的主体在与环境和其他主体交互中学习，从而改变自身行为或者结构的现象。粒子群的思想源自鸟群，可以将整个优化过程视作鸟群寻找食物的过程，食物对应问题的最优解，在空中飞行寻找食物的鸟对应基本搜索单元，即粒子。于是，粒子群算法可以理解为一群粒子在可行域中寻找最优解的过程。粒子具有四个属性：当前位置、自身最佳位置、群体最佳位置和当前速度。当前位置表示本次迭代后粒子所处的位置，用适应值来表示；自身最佳位置是粒子以往最接近解的位置；群体最佳位置是所有粒子中最接近解的位置；当前速度是粒子每次迭代的步长，根据前三个属性进行调节。粒子群算法的流程可以表述如下：

- Step1：在可行域内随机生成一定数量的粒子，粒子的初始属性为随机值，初始化群体最佳位置为 0；
- Step2：将群体最佳位置更新为所有粒子位置中的最佳位置；
- Step3：开始迭代，每个粒子根据自身当前位置和速度更新迭代后的新位置；
- Step4：每个粒子根据当前的自身位置和群体最佳位置更新下次迭代的速度；
- Step5：每个粒子判断是否满足终止条件结束迭代，若满足则返回最优解；若不满足，则转 Step2。

7.5.4 最优放牧策略

本题实验数据采用 2020 年 G17 小区的有机物含量和附录三提供的 2022 年 10cm 土壤湿度作为板结化公式的参数，使用示范牧区的人文因素、地理因素等作为沙漠化程度指数

的参数。通过 PSO 算法和遗传算法分别对 7.5.1 定义的问题进行求解可以得到如下图所示结果：

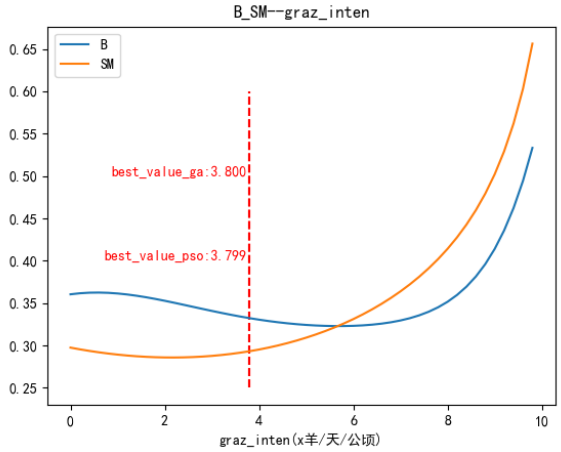


图 7-7 B/SM 指标与放牧强度的关系图

由图中可以看出，两种算法都得到了相近的最优解，提高最优解的可靠性。其中使沙漠化程度指数与土壤板结化程度最小的放牧策略为：放牧强度选取为 3.80 羊/天/公顷，其对应的沙漠化程度指数为：0.29，对应的土壤板结化程度为：0.33。

八、问题五的求解

8.1 问题分析

根据问题五所述要求，需要在给定降水量的情况下，给出最大放牧强度的阈值，在实现经济效应最大化的同时，保证草原的可持续发展。

本文利用沙漠化程度和土地板结化程度作为草原可持续发展的衡量指标，结合问题四得出的预测模型，用附件 13 所给牧户数据，分析不同降水量情况下 4 个牧户的沙漠化程度和土地板结化程度指标。并进一步利用粒子群优化算法来搜索每个牧户在不同年降水下的最大放牧强度。

其技术路线图如下所示：

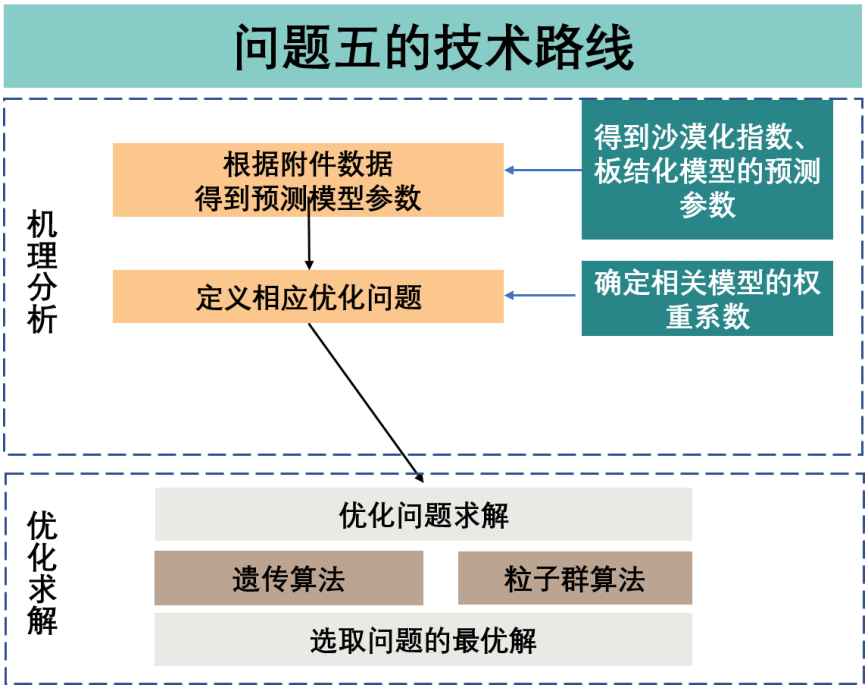


图 8-1 问题五的技术路线图

8.2 数据计算及问题提出

本文假定除降水量以外的气象因素、地表因素不变，再根据附件 3 算出各个牧户对应的人文因素，从而得到沙漠化程度曲线。对于板结化程度，可以根据降水量与湿度的关系进行估测。

适当的放牧可以促进草本分解，提升土壤养分，减少沙漠化和板结化程度，但过度放牧往往会加剧草原退化。在休牧和过度放牧之间找到最优放牧强度点是可以促进草原经济可持续发展的重要举措。为了得到可持续发展下的最大放牧强度I，构造以下问题：

$$\max I$$

$$s.t.\ SM \leq 0.29$$

$$B \leq 0.35$$

条件约束是实现可持续发展条件下，沙漠化程度和板结化程度必须满足的上界约束条件。

8.3 优化算法求解最佳放牧强度

表 8-1 遗传算法求解结果

降水量（mm） 牧户	300	600	900	1200
1	0.91	1.02	7.77	8.07
2	3.87	0.56	6.02	6.40
3	2.22	0.37	5.53	5.91
4	3.56	1.09	2.75	8.65

表 8-2 粒子群算法求解结果

降水量（mm） 牧户	300	600	900	1200
1	6.40	3.72	3.13	7.96
2	6.86	4.69	4.19	7.95
3	7.74	5.97	5.48	7.98
4	8.04	6.36	5.87	8.64

表 8-1 与表 8-2 分别为遗传算法和粒子群优化算法对 8.2 提出的优化问题进行求解的结果，由表中数据可以看出，在相同可持续发展的指标约束下，粒子群优化算法求解出的解能带来更大的经济效益。故最终解选用粒子群优化算法的结果，不同降水量下的各牧户的放牧强度与 SM 和 B 的关系及放牧数量阈值如下图所示：

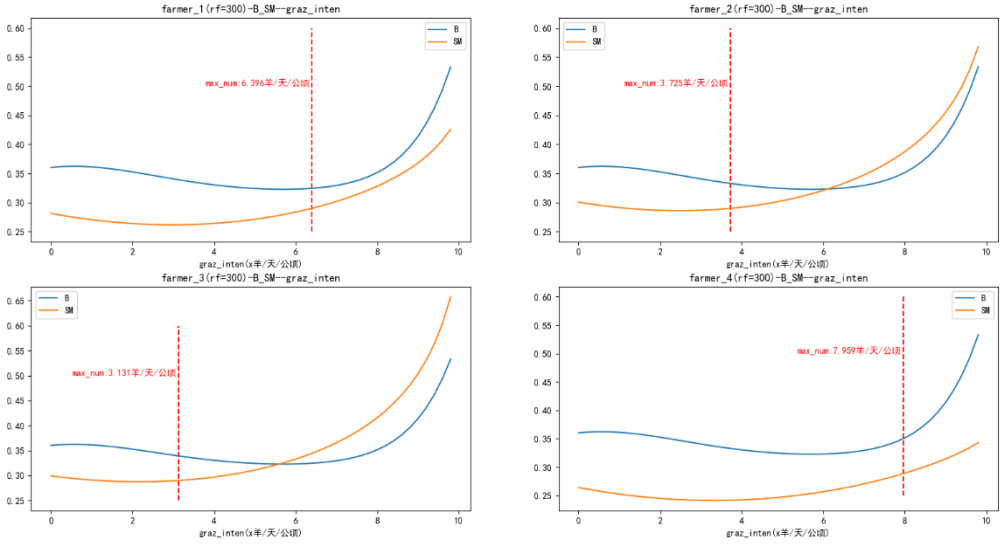


图 8-2 降水量 300mm 情况下 4 个牧户的最优放牧强度示意图

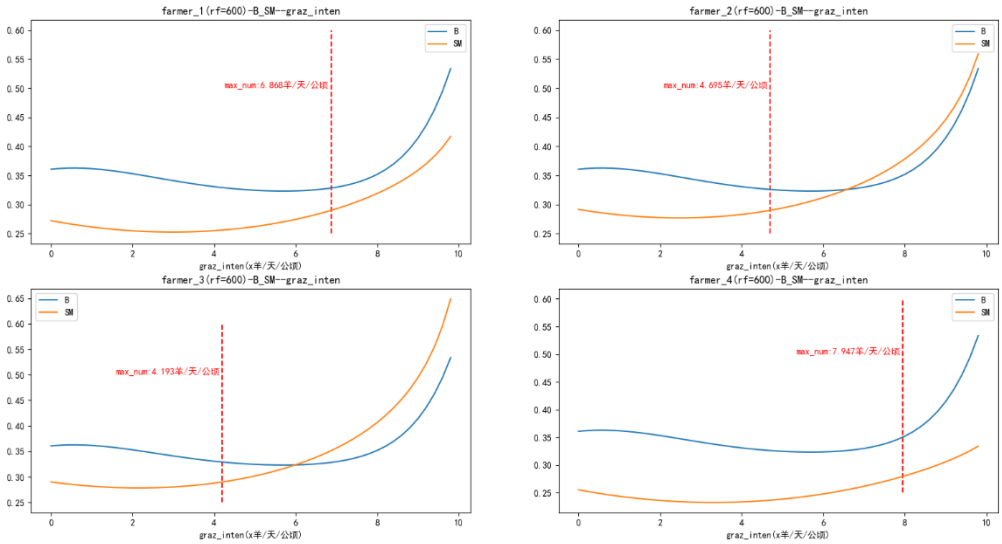


图 8-3 降水量 600mm 情况下 4 个牧户的最优放牧强度示意图

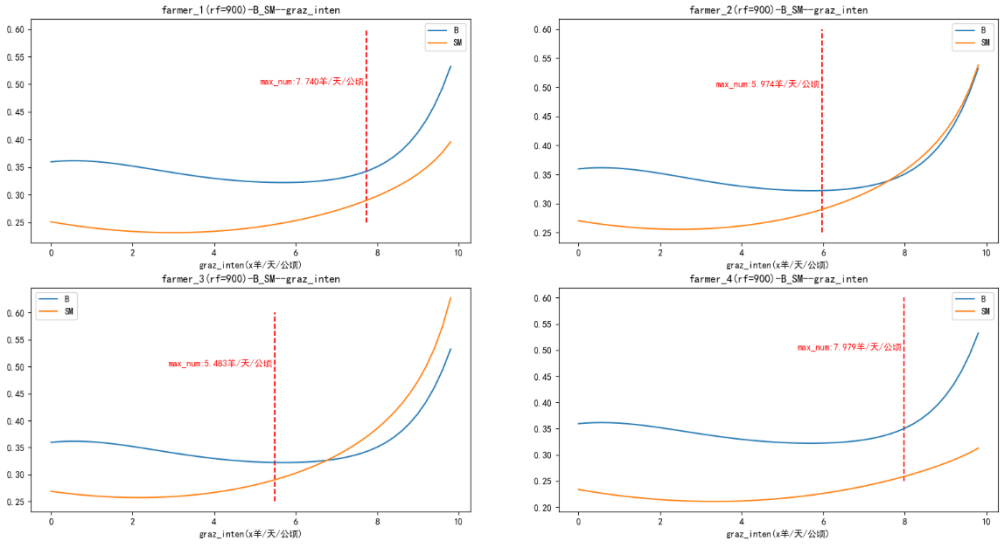


图 8-4 降水量 900mm 情况下 4 个牧户的最优放牧强度示意图

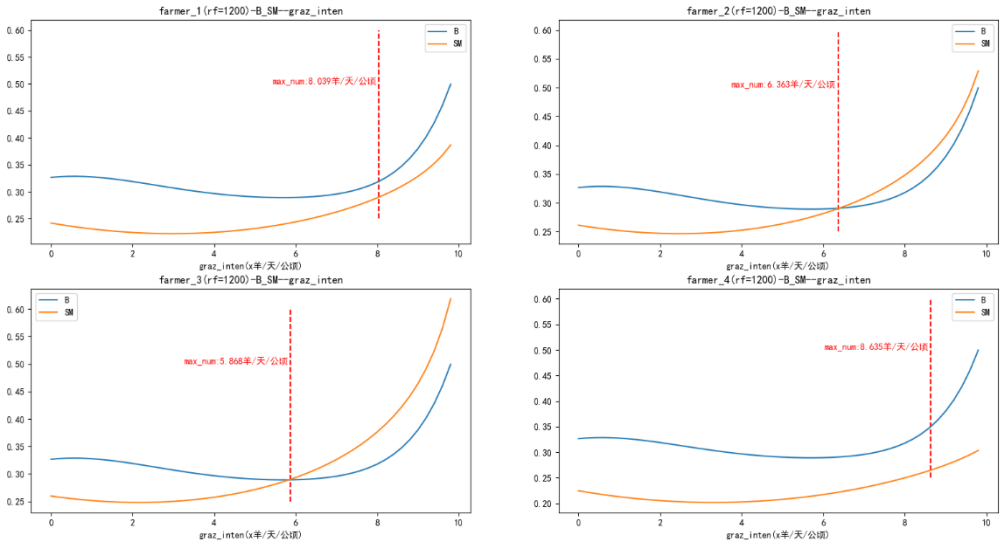


图 8-5 降水量 1200mm 情况下 4 个牧户的最优放牧强度示意图

比较图 8-2 到图 8-5，可以看出沙漠化程度和板结化程度都随着放牧强度的增加先下降再上升，但是其最低点通常在不同位置。同时降雨量的增加可以加大土地湿度，提升植被多样性，从而降低沙漠化程度和板结化程度，以支持更大的放牧强度。

在不同的降雨量下，最优放牧强度大小如下排列：牧户 4＞牧户 1＞牧户 2＞牧户 3，这是因为牧户 3 的放牧强度过高，在 6 羊/天/公顷左右，该强度已经对草原造成了沉重的压力，需要减少放牧强度来保护草原，而对应的，虽然牧户 4 的放牧强度在 2 羊/天/公顷左右，与牧户 1 和牧户 2 差不多，但是其经济收入过低，属于未能完全开发草原的情况，故其最优放牧强度最大。

九、问题六的求解

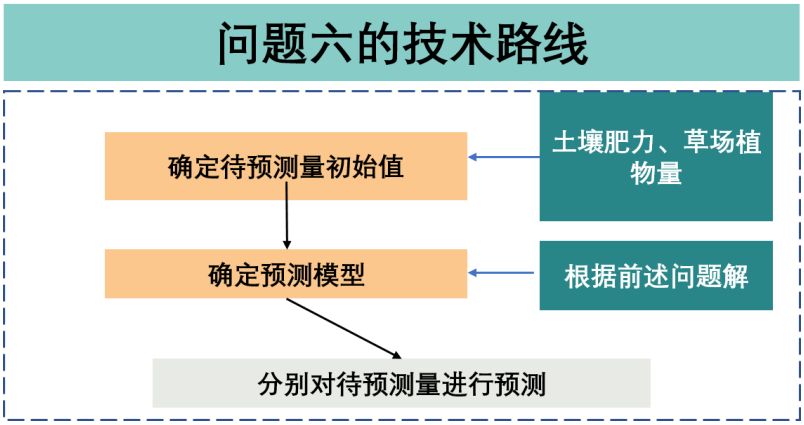


图 9-1 问题六的技术路线图

9.1 示范牧户的当前放牧策略

由附录 13 给出的数据中可以得到牧户 1、2、3、4 这四个牧户 2018 年到 2020 年的放牧强度如下表：

表 9-1 2018-2020 示范牧户的放牧强度表（单位：羊/km^2）

牧户\年份	2018	2019	2020	均值
1	275.83	272.22	256.39	268.16
2	230	210	200	213.33
3	601.75	492	520.75	538.17
4	245.9	245.9	255.7	249.17

对每个牧户三年的放牧强度取均值并除以 100，然后换算为羊/天/公顷为单位来作为 2020-2023 年 9 月的放牧策略，即牧户 1：2.68 羊/天/公顷；牧户 2：2.13 羊/天/公顷；牧户 3：5.38 羊/天/公顷；牧户 4：2.49 羊/天/公顷。

9.2 示范牧户当前草场的植物量与土壤肥力（主要是 SOC 和 TN 的含量）

由附录 12 中的不同年份平均生物量表得到下表数据：

表 9-2 2018-2020 示范牧户草场不同年份的生物量（单位：g/m^2）

牧户\年份	2018	2019	2020
1	53.96	65.45	124.2
2	124.74	108.59	92.81
3	73.92	49.78	100.16
4	107.2	96.82	61.79

这里将表中的 2020 年的生物量作为初始值来进行预测 2023 年 9 月的各牧户草场生物量。对于不同牧户草场的土壤肥力数据，我们根据表 14 中的 2020 年 12 个小区的土壤肥力来进行估算。具体估算方法为：由于牧户 1、牧户 2、牧户 4 的放牧强度接近轻牧（2 羊/天/公顷），我们选取 G6、G12、G18 这三个连续多年采取轻牧措施的小区，并对其土壤肥力取均值来作为牧户 1、2、4 的 2020 年的土壤肥力。对于牧户 3 我们选取 G8、G11、G16 这三个连续多年采取中牧措施的小区并对其土壤肥力取均值来作为牧户 3 的 2020 年土壤肥力的估计值。最终结果如下：

表 9-3 2020 年示范牧户草场土壤肥力估算结果表（单位：g/kg）

牧户\年份	SOC	TN
1	16.76	2.07
2	16.76	2.07
3	14.65	1.87
4	16.76	2.07

9.3 示范牧户 2023 年 9 月土地状态预测

9.3.1 草场生物量的预测

对于草场生物量的预测，由于该四个牧户所在地的其它因素缺失，我们只能根据放牧与植物生长之间的关系来对草场 2023 年 9 月的生物量进行预测，预测采用 Woodward 等建立的如下模型[7]：

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

式中，w 为植被生物量，S 为单位面积的载畜率。

9.3.2 草场肥力的预测

采用问题三中得到的 SOC 与放牧强度的拟合模型，TN 与放牧强度的拟合模型进行预测。

9.4 预测结果展示

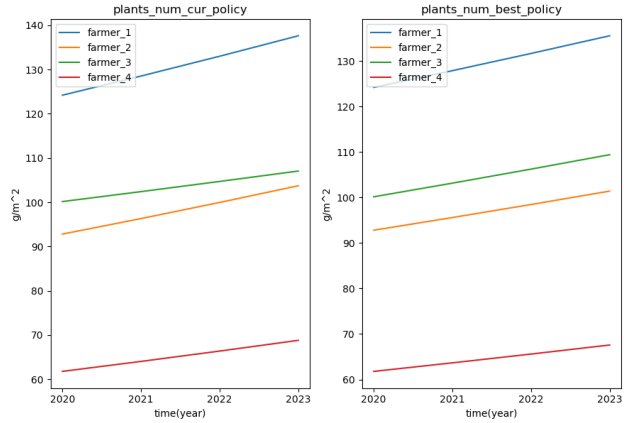


图 9-2 四个牧户在不同策略下的草场植物量随时间的变化.（左）四个牧户保持当前策略（右）四个牧户采取问题四求出的策略

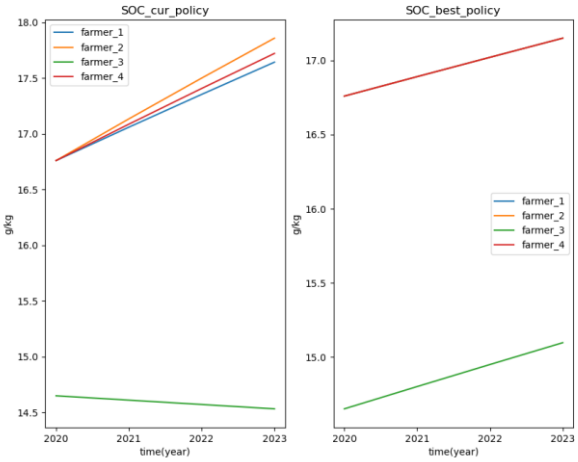


图 9-3 四个牧户在不同策略下的草场 SOC 随时间的变化.（左）四个牧户保持当前策略（右）四个牧户采取问题四求出的策略

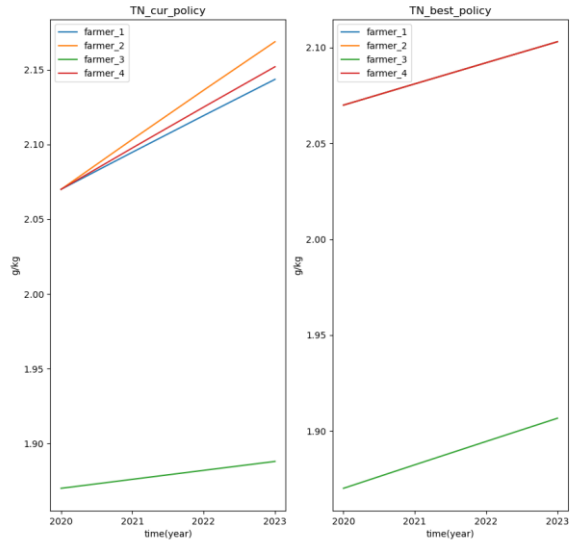


图 9-4 四个牧户在不同策略下的草场 TN 随时间的变化.（左）四个牧户保持当前策略（右）四个牧户采取问题四求出的策略

由上面展示的预测结果可以看出，各示范牧户当前采取的放牧策略与我们问题四求解出的策略都有利于草场植物量的增长与草场土地肥力的增长。

十、模型评价与改进

针对问题一，我们通过总结现有研究工作，探究土壤和植被生物量的运作机理，充分考虑了放牧强度对土壤湿度以及植被生物量的影响中涉及的因素，受到现有模型的启发，建立了合理且较为完善的数学模型，适用性广泛。但是由于现有数据不足，无法给出定量的数学表达式，后期如果对于模型中的变量有收集到充分的实验数据，相信求解出来的模型会非常适用于当地的环境。另一方面，土壤湿度和植被生物量之间并不是独立的关系，可能存在统一的表达式将两个因素结合起来。

针对问题二，我们创新的采用基于多种特征筛选的集成特征筛选模型，使得模型自变量特征筛选更加灵活和准确。但由于比赛时间限制未来及和其他模型进行对比。

针对问题三，我们结合文献中对于土壤化学物质影响因素的研究，创新式结合实际问题设计数学微分模型使用最小二乘进行回归求解，并使用决策树模型对未来时间的土壤有机物进行预测，但由于给定样本的数量有限，因此不能求解出更高阶的微分方程以更好地拟合问题，期待能够得到更多的数据来支持模型的先进性、创新型。

针对问题四，本文所利用的沙漠化程度预测模型是适用于塔里木盆地的模型，虽然我们利用修正系数进行了模型的拟合，但是因子权重系数并没有修正。锡林郭勒草原的地形、气候等因素与塔里木盆地有很大的差异，需要更加详细的分析以及实地考察，确定各个因子权重系数。

针对问题五，本文直接将降水量和放牧强度作为自变量，沙漠化程度作为因变量进行预测，并没有考虑到降水量对地面植被、人类活动的影响，会存在一定的误差。同时在预测土壤板结化程度时，并未考虑土壤容重的变化，需要更加深刻的分析。

针对问题六，对各示范牧户的土地状态进行预测时，由于获取到的相关数据有限，导致放牧强度对土地状态的影响比较单一，未考虑放牧强度与土地状态之间的一些中间变量，需要获取更加多的数据来进行更加复杂的预测。

十一、参考文献

[1] 才旦.草原生态保护与畜牧经济可持续发展研究[J].农家参谋,2021(23):126-127.
[2] 冯刚,尚维轩,丁勇.利用文献计量学分析草原退化对草地生态系统的影响[J].内蒙古大学学报(自然科学版),2022,53(04)
[3] 陈章,李磊,赵晋灵,王建柱,殷贺,李昂.2003-2020 年间锡林郭勒草原凋落物生物量指数变化趋势的研究[J].内蒙古大学学报(自然科学版),2022,53(03):290-298.
[4] 张伟华,关世英,李跃进.不同牧压强度对草原土壤水分、养分及其地上生物量的影响[J].干旱区资源与环境,2000(04):62-65.
[5] 张存厚,杨丽萍,越昆,刘朋涛,张德龙.锡林郭勒典型草原土壤水分对降水过程的响应[J].干旱区资源与环境,2022,36(08):133-139.
[6] 张军,亢志杰,海山,格日勒图.锡林郭勒盟农业综合开发划区轮牧现状与前景分析[J].内蒙古草业,2011,23(04):8-11.
[7] Woodward, Simon JR, Graeme C. Wake, and David G. McCall. Optimal grazing of a multi-paddock system using a discrete time model[J]. Agricultural Systems 48.2 (1995): 119-139.
[8] Wang, Xixi, Ruizhong Gao, and Xiaomin Yang. Responses of soil moisture to climate variability and livestock grazing in a semiarid Eurasian steppe[J]. Science of The Total Environment 781 (2021): 146705.
[9] 张彩琴,张军,李茜若.草地植被生物量动态研究视角与研究方法评述[J].生态学杂志,2015,34(04):1143-1151.DOI:10.13292/j.1000-4890.20150311.020.
[10]张彩琴. 内蒙古典型草原生长季内植物生长动态的数学模型与计算机模拟研究[D].内蒙古大学,2007.
[11]刘敦利. 基于栅格尺度的土地沙漠化预警模式研究[D].新疆大学,2010.
[12]中国土壤质地空间分布数据. https://www.resdc.cn/data.aspx?DATAID=260
[13]刘楠,张英俊.放牧对典型草原土壤有机碳及全氮的影响[J].草业科学,2010,27(04):11-14.
[14]杨珺婷,李晓松.应用哨兵 2 号卫星遥感影像数据和机器学习算法对锡林郭勒草原土壤表层有机碳及全氮的估算[J]. 东北林业大学学报,2022,50(01):64-71.DOI:10.13759/j.cnki.dlxb.2022.01.022.