**Pure Mathematics** 理论数学, 2024, 14(3), 260-280
Published Online March 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.143104

Hans汉斯

# 基于数学模型预测草原可持续发展能力研究

**赵牧艺**

上海理工大学理学院，上海

收稿日期：2024年1月29日；录用日期：2024年3月22日；发布日期：2024年3月29日

## 摘　要

草原作为重要陆地生物系统，一直以来都是重要研究课题之一。结合草原当地的自然环境和人文环境，建立合理模型来分析草原土地性质，将有助于草原可持续发展的研究。本文主要采取时间序列预测模型、多元线性回归模型和因子分析等多种数学模型，通过对草原土壤性质、植被生物量等预测及给出相应放牧策略，来实现草原可持续发展的目的。针对问题一，本文首先利用Pearson相关系数，对预处理过的数据进行特征选择排序，筛选出在不同放牧策略下对土壤湿度和植物生物量变化相关系数高于0.4的特征；最后利用LSTM时序模型建立不同放牧策略下草原土壤湿度和植被生物量的数学模型。针对问题二，本文利用ARIMA模型对土壤湿度相关性较高的特征建立模型进行预测，再利用LSTM模型对10 cm处土壤湿度进行预测；并在此基础上，完成40 cm等土壤湿度预测；由于LSTM模型得到的精度较低，最终选择ARIMA模型直接预测不同深度土壤湿度。针对问题三，本文首先建立时间与土壤化学性质的多元线性回归模型，预测不同放牧强度下土壤有机碳等值；随后通过绘制不同土壤化学性质与时间的散点图，发现其存在线性关系，选择线性回归模型构建不同放牧强度下各放牧小区五种土壤化学性质的预测模型进行预测。针对问题四，本文通过因子分析技术得到主因子以及沙漠化程度指数预测模型中各个特征因子的权重，利用线性回归求解调节参数从而得到指数预测模型，并计算出不同放牧强度下监测点的沙漠化程度指数值；通过建立多元线性模型构建土壤板结化与土壤湿度等之间的数学模型，计算出不同强度下监测点的土壤板结化程度；最后根据不同放牧强度下土壤板结化程度和沙漠化指数，选择使两者都最小的放牧强度作为最终放牧策略，本文选择轻牧的放牧策略。

## 关键词

土壤湿度，LSTM模型，ARIMA模型，因子分析，放牧策略

# Research on Predicting Grassland Sustainable Development Ability Based on Mathematical Model

**Muyi Zhao**

School of Science, University of Shanghai for Science and Technology, Shanghai

文章引用: 赵牧艺. 基于数学模型预测草原可持续发展能力研究[J]. 理论数学, 2024, 14(3): 260-280.
DOI: 10.12677/pm.2024.143104

Received: Jan. 29th, 2024; accepted: Mar. 22nd, 2024; published: Mar. 29th, 2024

## Abstract

**Grassland, as an important terrestrial biological system, has always been one of the important research topics. Combining the local natural and human environment of the grassland, establishing a reasonable model to analyze the land properties of the grassland will help the research on sustainable development of the grassland. This paper mainly adopts a variety of mathematical models such as time series prediction model, multiple linear regression model and factor analysis, and through predicting the soil properties and vegetation biomass of the grassland and giving corresponding grazing strategies, to achieve the purpose of sustainable development of the grassland. For the problem 1, this article first uses the Pearson correlation coefficient to select the pre-processing data to select the characteristics of the pre-grazing strategies under different grazing strategies, the characteristics of the correlation coefficient of soil humidity and plant bi quantity which is higher than 0.4; finally uses LSTM Time-order model to establish a mathematical model of soil humidity and vegetation biomass under different grazing strategies. For the problem 2, this article uses the ARIMA model to predict the features of high soil humidity correlation, then uses the LSTM model to predict the soil moisture at 10 cm, and on the basis, completes the soil moisture prediction at 40 cm; because the accuracy of the LSTM model is low, the ARIMA model is finally selected to directly predict different deep soil humidity. In response to the problem 3, this article first establishes a multi-linear regression model of time and soil chemistry, predicts the value of soil organic carbon equivalent under different grazing strength; then through drawing different soil chemical properties and time scattered dots, it is found that it exists in linear relationship, then a linear regression model is selected to construct predictive models for five soil chemical properties in each grazing plot under different grazing intensities for prediction. In response to the problem 4, this article obtains the main factor and desertification index prediction model through factor analysis technology and linear regression is used to solve the regulation parameters to obtain the index prediction model and to calculate the desertification index values for the monitoring sites at different grazing intensities; by establishing a mathematical model such as a diversified linear model to build a mathematical model between soil plates and soil humidity, the degrees of soil boards of the soil plate of monitoring points under different intensities are calculated. Finally, based on the degree of soil crusting and desertification index under different grazing intensities, the grazing intensity that minimised both is selected as the final grazing strategy, and in this paper, the grazing strategy of light grazing is selected.**

## Keywords

**Soil Humidity, LSTM Model, ARIMA Model, Factor Analysis, Grazing Strategy**



## 1. 前言

### 1.1. 研究背景

草原作为地球上分布最广的陆地植被类型，分成温带草原、高寒草原等多种类型，是陆地最重要的生态系统之一。草原素有孕育生命、沉淀历史的绿色花园之称，它不仅是数以万计的江河发源地和水源

涵养地，孕育着众多湖泊和冰川，还拥有着人类所需的水库、钱库和碳库等功能。林草兴则生态兴，草原对于一个国家的生态发展具有基础性和战略性作用，它有着生物多样性保育、营养元素循环、碳固持、调节气候及防水土流失等重要功能。作为世界上草原资源和种类最丰富的国度之一，中国的草原总面积约 4 亿公顷，占国土总面积的 41%，主要分布在完达山到青藏高原东麓的西北地区。

然而随着过度放牧等人为因素，我国的天然草原开始出现不同程度的土地退化和沙漠化，植物群落类型和土壤特性等发生变化，地上、地下生物量大幅度减少，极大地影响了整个生态系统的发展。解决草原生态环境问题迫在眉睫，为此 2003 年党中央、国务院颁发“退牧还草”决策，目的是加强草原保护与建设。十几年的显著成果可以看出这一重大措施使得草原逐渐从衰败走向复苏，草原植被开始恢复，草原环境得到改善，同时畜牧业也实现了可持续发展，保障了牧民的长远生计。

“放牧还草”并不意味着完全禁止放牧，通常需要因地制宜地去考虑放牧方式和放牧强度，以草定畜、严格控制载畜率等。除了部分地区禁牧以外，大部分草原区域实施选择划区放牧和生长季放牧等轮回制放牧法，高效利用天然草原，达到草畜平衡，实现了草原资源的可持续利用，建立起草原和畜牧业的共生平衡生态系统。适当放牧不仅可以改善土壤质量、提高生物多样性，还能保持草场相对稳定的生产率[1]。

### 1.2. 研究目的及意义

基于上述的研究背景，本文主要以草原土地性质和可持续发展为研究对象，通过对土壤湿度等土壤化学相关元素的数据建立多种数学模型，从而去预测不同深度及不同放牧强度下土壤化学元素的分布情况；因此我们可以更加方便了解草原土地状态的性质，为后续寻找最佳放牧政策提供良好的数据基础。

此外，为呼吁国家的“加强草原保护和建设”的重要决策，达到草原资源的可持续性发展，本文建立不同放牧强度下土壤板结化程度和沙漠化指数的数学模型，希望通过该定性数学模型，寻找到最佳的放牧策略，例如放羊数量等。从而建立草原和畜牧业的和谐生态系统，保证在草原可持续发展情况下实现经济效益最大化[2]。

## 2. 模型假设

假设 1：所有数据均真实可靠；

假设 2：在不同放牧强度下，样本小区的土壤湿度一致；

假设 3：除了提供的相关数据，无其余变量对锡林郭勒草原可持续发展产生影响。

## 3. 符号说明

文中符号说明如表 1。

**Table 1.** Symbol description
**表 1.** 符号说明

| 序号 | 符号 | 含义 |
|---|---|---|
| 1 | MSE | 均方误差 |
| 2 | RMSE | 均方根误差 |
| 3 | SM | 沙漠化程度指数 |
| 4 | $\eta$ | 调节系数 |
| 5 | $Q_i$ | 因子强度 |
| 6 | $W_{c_i}$ | 因子权重系数 |
| 7 | $S_{Q_i}$ | 因子对沙漠化程度的贡献度 |

续表

| | | |
|---|---|---|
| 8 | $p$ | 自回归阶数 |
| 9 | $q$ | 移动平均阶数 |
| 10 | $c_{up}$ | 特征因子影响沙漠化的上限 |
| 11 | $c_{down}$ | 特征因子影响沙漠化的下限 |
| 12 | AIC | 赤池信息准则 |
| 13 | BIC | 贝叶斯信息准则 |

注：其他符号在文中说明。

## 4. 问题一：模型的建立和求解

### 4.1. 问题一分析

问题 1 主要建立不同放牧策略下内蒙古锡林郭勒草原土壤湿度和植被生物数量的数学模型。现拟从以下三个步骤去处理问题一：

(1) 首先找出影响土地湿度和植被生物数量的主要特征变量。

(2) 分析各个变量之间的自相关性和对土壤湿度和植被生物数量的影响，采用皮尔逊相关分析对收集到的变量进行特征筛选，在不同的放牧强度下选择出对土壤湿度和植被生物数量具有显著影响的变量。

(3) 分别建立四种放牧强度下土壤湿度和植被生物数量的 LSTM 时间预测模型。

### 4.2. 数据预处理

经过查阅相关文献[3] [4]，本文归纳总结出草原土壤湿度和植被生物数量的主要影响因素，其中影响土壤湿度和植被生物数量的主要为降水、植物干重等约 30 多个变量因子。同时根据放牧与植物生长之间关系和土壤含水量－降水量－地表蒸发模型的两个简单模型，也在数据支撑验证下确定这些因素对土壤湿度和植被生物数量是存在一定影响关系。

基于问题一及所给数据，可以将所提及的放牧策略主要归结为放牧强度的不同，因此针对四种放牧强度(对照、轻度放牧强度、中度放牧强度和重度放牧强度)分别组建不同放牧强度下各个变量因子对土壤湿度和植被生物数量的影响模型。

为保证数据的有效性，本文首先进行初步统计特征分析并对所用数据进行预处理。数据的预处理主要分为以下三个板块：(1) 删除缺失率较高和含零值较高的变量；(2) 基于 KNN 算法对缺失率较少的变量进行数值插补；(3) 拉格朗日插值法。

### 4.3. 特征选择——皮尔逊 Pearson 相关系数

为减少变量数量、提高模型效率，使得模型泛华能力更强，避免过拟合，通常在数据预处理完以后会对变量进行特征选择，其原则是尽可能获取小的、稳定的、适应性强的特征子集，同时不显著降低分类精度、不显著影响分类分布。因此可以通过特征选择找到对土壤湿度和植被生物数量最具显著影响的变量，常用的降维方法为皮尔逊 Pearson 相关系数，其原理为：

英国数学家 Karl Pearson 提出了 Pearson 相关系数的统计指标，并将其广泛应用于变量间的线性相关程度定量分析。Pearson 相关系数的计算方式是将协方差除以两个变量的标准差，从而消除掉两个变量的量纲影响。计算公式为：

$$\rho_{x,y} = corr(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

该统计指标反映了两变量之间的线性相关程度，其值介于−1 和 1 之间。相关系数的绝对值越大，两者线性关系越强，相关系数越趋近于 1 或−1，相关度越强，相关系数越趋近于 0，相关度越弱。当一个变量增大，另一个变量也随之增加时，说明为正相关，相关系数大于 0，相反如果减小，则为负相关，相关系数小于 0。如果相关系数等于 0，则不存在线性相关关系。因此可以通过该方法来选择与土壤湿度相关性较强的特征变量。

由于问题一探讨的是不同放牧策略对土壤湿度和植被生物数量的影响，因此分为土壤湿度和植被生物数量两大部分去进行特征选择：

#### 4.3.1. 土壤湿度

查阅附件资料可以得出不同放牧强度下 SOC 土壤有机碳及鲜重等影响因子是不同的，但一些基本变量，例平均气温等因子是固定不变的，因此在做土壤湿度的特征选择时会分成两部分对特征变量做 Pearson 相关分析：

(1) 基本气候变量：首先对基本气候变量做 Pearson 相关分析，得到了基本气候变量与土壤湿度的相关系数，如表 2 所示：

**Table 2.** Correlation coefficient table of climate variables
**表 2.** 气候变量相关系数表

| 基本气候变量 | 相关系数 |
|---|---|
| 平均气温(℃) | 0.16 |
| 平均最高气温(℃) | 0.19 |
| ⋮ | ⋮ |
| 植被指数 | 0.79 |
| 径流量($m^3/s$) | 0.32 |
| 径流量($m^3$) | 0.32 |

结合相关系数表 2 可知，相关系数大于 0.4 的变量与土壤湿度具有一定程度的相关性，因此结合气候变量与土壤湿度的相关系数图可以推断：最低气温极值、降水量等 9 个变量对土壤湿度具有一定影响，但是由于降水量、单日最大降水量和降水天数两两之间的相关系数高达 80%，相关性较高，为避免变量间的耦合性带来的干扰，且降水量更能表现降水的特征，因此这三个降水相关的变量仅选择降水量，其余删除。

同理对剩余变量进行筛选，最终得出对土壤湿度有一定影响的基本气候变量为降水量、平均露点温度、平均最大持续风速共三个，如表 3 所示：

**Table 3.** Table of main climate variables
**表 3.** 主要气候变量表

| 变量(基本气候数据) | 相关系数 |
|---|---|
| 降水量(mm) | 0.67 |
| 平均露点温度(℃) | 0.59 |
| 平均最大持续风速(knots) | 0.41 |

(2) 土壤、植被相关变量：

由于这类变量随着放牧强度的不同而变化，因此在进行特征选择时分成四个不同放牧强度(对照、轻度放牧强度、中度放牧强度和重度放牧强度)，表 4 为不同放牧强度下土壤、植被相关数据与湿度的相关系数表：

**Table 4.** Soil and vegetation correlation coefficient table
**表 4.** 土壤、植被相关系数表

| | 鲜重(g) | 干重(g) | SOC 土壤有机碳 | SIC 土壤无机碳 | STC 土壤全碳 | 全氮 N | 土壤 C/N 比 |
|---|---|---|---|---|---|---|---|
| **对照组** | 0.12 | 0.06 | 0.53 | 0.22 | 0.47 | 0.50 | 0.46 |
| **轻度放牧** | 0.30 | 0.31 | 0.52 | 0.48 | 0.35 | 0.47 | 0.51 |
| **中度放牧** | 0.40 | 0.31 | 0.18 | 0.52 | 0.51 | 0.26 | 0.51 |
| **重度放牧** | 0.30 | 0.29 | 0.53 | 0.53 | 0.49 | 0.53 | 0.53 |

首先在表 4 里选取相关系数大于 0.4 的特征变量，分别得到不同放牧强度下土壤湿度的影响因素。其中在重度放牧强度的情况下，STC 土壤全碳与土壤湿度的相关系数达到 0.49，但与 SOC 土壤有机碳、全氮 N、土壤 C/N 比和 SOC 土壤无机碳的相关性过高，同样为避免变量间自相关性过强带来的干扰，因此选择可以删掉。

至此，整理出不同放牧强度下选择后的变量与土壤湿度的相关系数图，如图 1 所示：

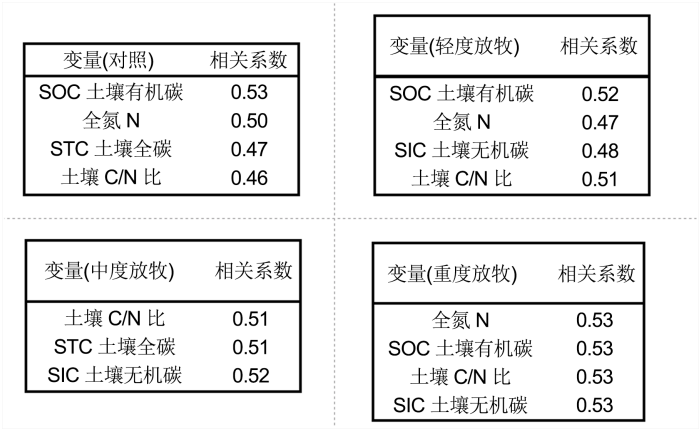


**Figure 1.** Soil moisture correlation coefficient diagram (corresponding from left to right and top to bottom, four grazing intensities: control, light grazing, moderate grazing and heavy grazing)
**图 1.** 土壤湿度相关系数图(从左到右从上到下分别对应对照、轻度放牧、中度放牧和重度放牧这四种放牧强度)

综合上述，最终得到了不同放牧强度下影响土壤湿度的不同特征变量，得到图 2：

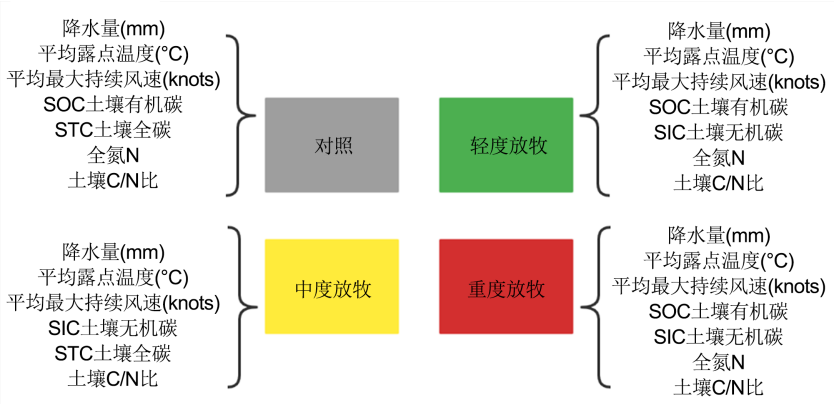


**Figure 2.** Characteristic variable diagram of soil moisture
**图 2.** 土壤湿度的特征变量图

#### 4.3.2. 植被生物数量

与土壤湿度模型不同，叶面积对植被生物数量可能存在一定的影响，因此在前面所有特征变量的基础上，本文在对植被生物数量模型进行特征选择时，增加了叶面积指数这一特征变量。对影响植被生物数量的三十多个特征变量做 Pearson 相关分析，得到与植被干重的相关系数表 5 (以轻度放牧强度为例)：

**Table 5.** Vegetation dry weight correlation coefficient table
**表 5.** 植被干重相关系数表

| 变量(轻度放牧) | 相关系数 |
|---|---|
| 鲜重(g) | 0.98 |
| 土壤 C/N 比 | 0.70 |
| ⋮ | ⋮ |
| 径流量($m^3$) | 0.13 |
| 湿度 | 0.31 |
| 平均能见度(km) | 0.60 |
| 全氮 N | 0.83 |

与土壤湿度模型的特征选择同理，结合表 5 选择相关系数大于 0.4 的变量，且在变量间相关性过高的变量中仅保留最能体现这一特征的变量，经过筛选后得到不同放牧强度下的显著影响因子对植被干重的相关系数如图 3。

| 变量(对照) | 相关系数 |
|---|---|
| 土壤 C/N 比 | 0.46 |
| SIC 土壤无机碳 | 0.43 |
| 平均露点温度(°C) | 0.41 |
| 平均最大持续风速(knots) | 0.48 |
| 平均风速 (knots) | 0.55 |

| 变量(轻度放牧) | 相关系数 |
|---|---|
| 土壤 C/N 比 | 0.70 |
| SIC 土壤无机碳 | 0.42 |
| 平均能见度(km) | 0.60 |
| SOC 土壤有机碳 | 0.63 |
| 全氮N | 0.83 |

| 变量(中度放牧) | 相关系数 |
|---|---|
| 平均露点温度(°C) | 0.52 |
| 植被指数(NDVI) | 0.49 |
| STC 土壤全碳 | 0.41 |
| 平均风速(knots) | 0.47 |
| 平均最大持续风速(knots) | 0.55 |
| SOC 土壤有机碳 | 0.64 |

| 变量(重度放牧) | 相关系数 |
|---|---|
| 土壤 C/N 比 | 0.67 |
| SIC 土壤无机碳 | 0.60 |
| 平均能见度(km) | 0.63 |
| 全氮 N | 0.70 |
| SOC 土壤有机碳 | 0.72 |

**Figure 3.** Correlation coefficient diagram of the number of vegetation organisms (corresponding from left to right and top to bottom, Four grazing intensities: control, light grazing, moderate grazing and heavy grazing)
**图 3.** 植被生物数量相关系数图(从左到右从上到下分别对应对照、轻度放牧、中度放牧和重度放牧这四种放牧强度)

至此，得到了不同放牧强度下影响植被生物数量的特征变量(图 4)：

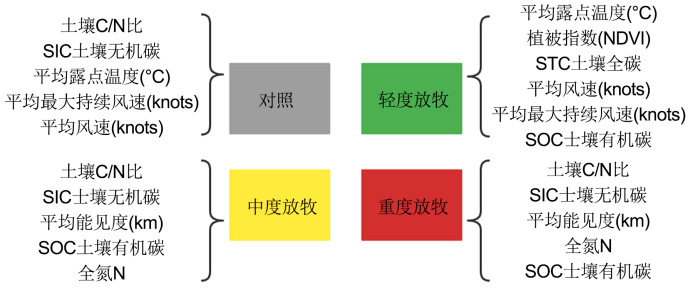


**Figure 4.** Characteristic variable diagram of the number of vegetation organisms
**图 4.** 植被生物数量的特征变量图

## 4.4. 基于 LSTM 时序预测模型的建立

### 4.4.1. LSTM 模型的原理

1997 年德国计算机科学家 Sepp Hochreiter 和 Schmidhuber 首次提出长短时间记忆(LSTM)模型，是一种时间循环神经网络，为解决 RNN 的长序列训练过程中梯度消失问题演进而来的，能够处理具有长期依赖关系的时间序列数据，并进行准确的预测。在实际应用中，LSTM 模型已经被广泛应用于各种时序预测任务[5]，如股票价格预测、交通流量预测、能源消耗预测等。LSTM 引进三个门来保护和控制细胞状态：遗忘门 + 输入门 + 输出门，通过这三个精心设计的门来新增或删除信息。

首先，遗忘门决定当前状态应该保留或丢弃多少上一时刻的信息。将先前时刻和当前时刻输入的信息同时输入到 sigmoid 函数，得到的输出值在 0~1 之间，越靠近 0 表示越应该遗弃，越接近 1 表明越应该保留。计算公式为：

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right)$$

其中 $W$，$b$ 分别为权重和偏置，$\sigma$ 为 sigmoid 函数，表示每个部分有多少量可以通过。其次，输入门过滤新信息，确定什么样的新信息需要输入到细胞状态里。同样将先前状态和当前状态的信息输入进 sigmoid 函数里，调整输出值来确定哪些信息应该更新，0 代表不重要，1 代表重要。同时将值输入到 tanh 函数，将数值压缩至−1 到 1 之间，最后将 sigmoid 输出值和 tanh 输出值进行相乘，决定哪些信息是重要的、需要保留的。计算公式为：

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right)$$

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

旧细胞通过遗忘门，待用信息通过输入门，两者相结合得到的结构来更新细胞状态。最后，输出门做出决策，确定当前时刻输出什么样的记忆信息，能够决定下一个状态(即隐藏状态)的值。前面两个门确定了隐藏状态需要携带的信息并当前输出，将新的单元状态和隐藏状态传递给下个时间步。计算公式为：

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

$\sigma$ 决定要输入的部分，结合 tanh 将细胞状态变换后相乘，输出结果信息。

LSTM 建立预测模型主要有以下几步：(1) 数据预处理：选取数据样本，进行数据标准化处理，对数据进行归一化处理有利于训练模型的损失(loss)迭代。将数据划分成训练集和测试集。(2) 预测模型：建立预测模型需要提前确定输入层、输出层的节点数和隐藏层的神经元等模型参数，将训练样本输入到模型中，进行深度学习和数据挖掘，训练得到预测模型。(3) 评价模型：为保证预测结果的准确性和预测误差波动的稳定性，需要对模型的预测值和真实值的拟合程度进行评价，通常采用平均绝对误差(MAE)等进行评估，误差越小，说明两者之间的离散程度越小，预测结果更可靠[6]。

### 4.4.2. 土壤湿度的模型建立

在进行数据预处理和特征选择两步的处理后得到了土壤湿度的影响变量数据，分别将这些数据划分成训练集和测试集(其中 80%为训练集，20%为测试集)，训练集的数据用于模型的学习训练，测试集的数据用于检验模型的准确性。

本文基于 Python 软件，求解出不同放牧强度对土壤湿度的 LSTM 预测模型，共四个。不同放牧强度

下 20%测试样本的测试结果如图 5 所示，可以看出 LSTM 模型预测的整体效果比较好，预测值和真实值的整体走向趋势一致，拟合程度较高。

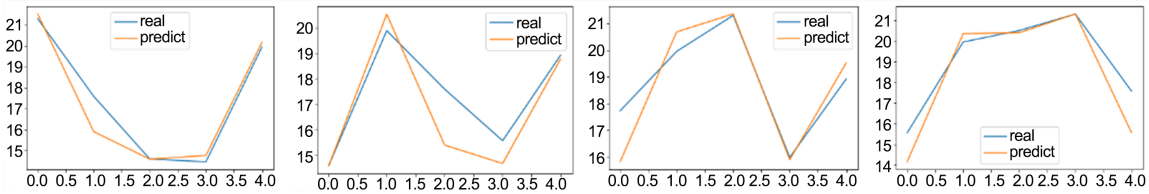


**Figure 5.** Measured values and true values of soil moisture (from left to right, top to bottom Corresponding to the four grazing intensities of control, light grazing, moderate grazing and heavy grazing respectively)
**图 5.** 土壤湿度的测值与真实值(从左到右从上到下分别对应对照、轻度放牧、中度放牧和重度放牧这四种放牧强度)

### 4.4.3. 土壤湿度的模型评价

为了更加准确地对模型性能进行客观评价，我们需要对模型的预测值和真实值的拟合程度进行评价，本次问题选取了常见的平均绝对误差(MAE)和均方根误差(RMSE)两个指标从不同的角度对预测模型进行判断，其有着简单易懂、可靠等优点。MAE 和 RMSE 的公式如下：

$$\mathrm{MAE}=\frac{1}{n}\sum_{t=1}^{n}\left|f_t-d_t\right|$$

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{t=1}^{n}\left(f_t-d_t\right)^2}{n}}$$

对建立好的 LSTM 模型计算不同放牧强度下的 MAE、RMSE 误差值，通常情况下 MAE、RMSE 误差值越小，预测值和真实值之间的差异就越小，模型的预测精度就越高：

**Table 6.** Soil moisture prediction error table
**表 6.** 土壤湿度预测误差表

| | 对照 | 轻度放牧 | 中度放牧 | 重度放牧 |
|---|---|---|---|---|
| **MAE** | 0.4906 | 0.7819 | 0.6803 | 0.7852 |
| **RSME** | 0.7819 | 1.1049 | 0.9581 | 1.1093 |

从表 6 可以看出这四个放牧强度下的土壤湿度 LSTM 预测模型的误差值均很小，预测效果较佳。

### 4.4.4. 植被生物数量的模型

与土壤湿度相同，将选择后的植被生物数量影响因素数据带入上述步骤中去，可以得到不同放牧强度下植被生物数量的 LSTM 预测模型图 6 和预测误差值表 7，从中可以看出植被生物数量 LSTM 时间预测模型的整体效果较佳，预测精度也较高。

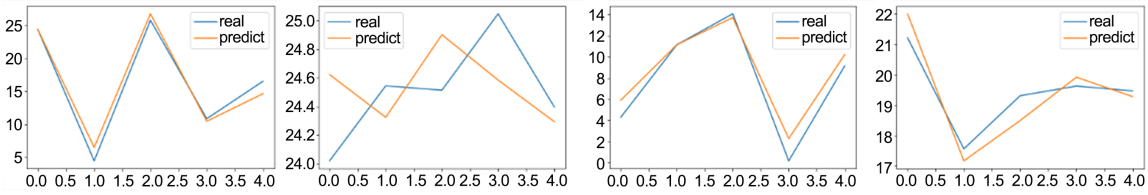


**Figure 6.** Predicted values and actual values of the number of vegetation organisms (from left to right, top to bottom corresponding to the four grazing intensities of control, light grazing, moderate grazing and heavy grazing respectively)
**图 6.** 植被生物数量的预测值与真实值(从左到右从上到下分别对应对照、轻度放牧、中度放牧和重度放牧这四种放牧强度)

**Table 7.** Prediction error table for the number of vegetation organisms
**表 7.** 植被生物数量预测误差表

| | 对照 | 轻度放牧 | 中度放牧 | 重度放牧 |
|---|---|---|---|---|
| MAE | 1.0602 | 0.3550 | 1.0329 | 0.4954 |
| RSME | 1.3280 | 0.3965 | 1.2929 | 0.5589 |

## 5. 问题二：模型的建立和求解

### 5.1. 问题二分析

问题二是问题一的延续，主要是对不同放牧强度下 2022~2023 年不同深度的土壤湿度进行预测。因此，针对问题二的预测，主要分为以下两步(图 7)：

(1) 首先根据往年数据对土壤湿度的三个影响因素(降水量、平均露点温度和平均最大持续风速)做 ARIMA 模型，预测得到 2022~2023 年这三个影响因素的值。(2) 第二步其实是问题一 LSTM 模型的延续，将第一步预测得到的值输入到 10 cm 土壤湿度的 LSTM 模型从而得到 2022~2023 年 10 cm 土壤湿度的预测值。对 40 cm 土壤湿度而言，不仅需要考虑降水量、平均露点温度和平均最大持续风速这三个影响因素，也要考虑 10 cm 土壤湿度对 40 cm 土壤湿度的影响，因此建立的是这四个因素对 40 cm 土壤湿度的 LSTM 模型从而预测得到结果。以此类推，最后预测得到 2022~2023 年不同深度土壤湿度值。

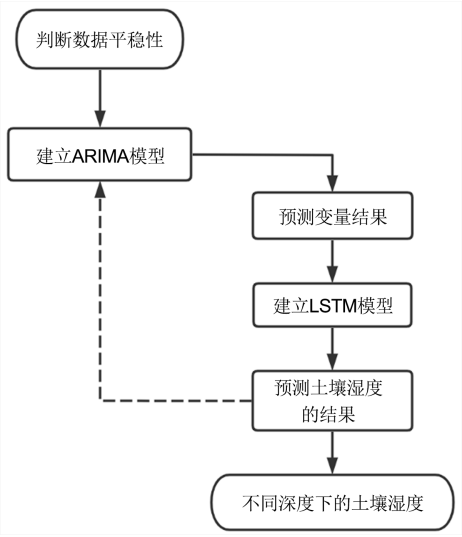


**Figure 7.** Flow chart of question 2
**图 7.** 问题二的流程图

### 5.2. ARIMA 模型

#### 5.2.1. ARIMA 模型原理

差分自回归滑动平均模型(简称 ARIMA 模型) [7]，是将自回归 AR 模型、移动平均 MA 模型和差分法相结合得到的一种时间序列预测方法，适用于解决经济学领域中众多实际问题，比如说非平稳时间序列。其原理是将非平稳时间序列转化为平稳序列，再将因变量的滞后值和随机误差项的滞后值等做回归。ARIMA($p$,$d$,$q$)模型的一般形式为：

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + \varepsilon_t + \theta_0 + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

其中 $X_t$ 为时间序列，$p$ 为自回归阶数，$q$ 为移动平均阶数，$\varphi_1,\varphi_2,\cdots,\varphi_p$ 为自回归模型系数，$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_q$ 为移动平均模型系数，$\varepsilon_t$ 为白噪声序列，$d$ 为差分阶数。

ARIMA 的建模过程分为以下：(1) 对数据绘制时间序列图，观察是否为平稳时间序列，若为非平稳时间序列，则不断进行差分直到转化为平稳数据序列。(2) 定阶过程：利用 AIC 信息和 BIC 信息进行判断，确定好模型的参数 $p,q$。(3) 用搭建好的模型去预测未来时间下的结果。

本文前后分别介绍了 LSTM 和 ARIMA 两个非常常见的时间序列预测模型[8]，并作为本次题目的关键模型，在不同情况下分别使用。其主要区别为：(1) 模型结构：ARIMA 是一种经典的统计模型，基于时间序列的自相关和移动平均性质建立模型。而 LSTM 是一种深度学习模型，它是一种循环神经网络的变体，具有记忆单元和门控机制，能够捕捉长期依赖关系。(2) 预测能力；由于 LSTM 具有记忆单元和门控机制，能够捕捉长期依赖关系，因此在处理具有较长时间依赖的序列数据时通常具有更好的预测能力。而 ARIMA 则适用于较短期的预测任务。

### 5.2.2. ARIMA 模型估计与建立

以平均最大持续风速为例，建立平均最大持续风速的 ARIMA 模型，剩下两个影响变量(降水量、平均露点温度)同理。

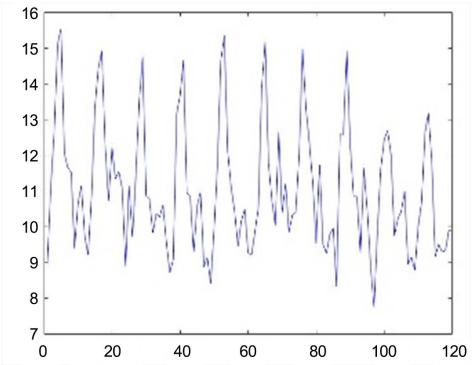


**Figure 8.** Time series graph of average maximum sustained wind speed
**图 8.** 平均最大持续风速的时间序列图

观察图 8 可知由平均最大持续风速的往年数据非平稳时间序列，并且具有较强的季节性，因此需要每次间隔 12 个月做差分来消除季节性，同时还需要进行 1 阶差分，将数据转化成平稳时间序列，再根据自相关 ACF 函数图 9 和偏自相关函数 PACF 图 10 判断出我们需要做 ARIMA 模型。

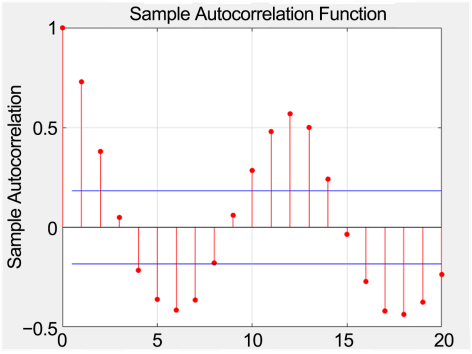


**Figure 9.** Autocorrelation ACF function graph
**图 9.** 自相关 ACF 函数图

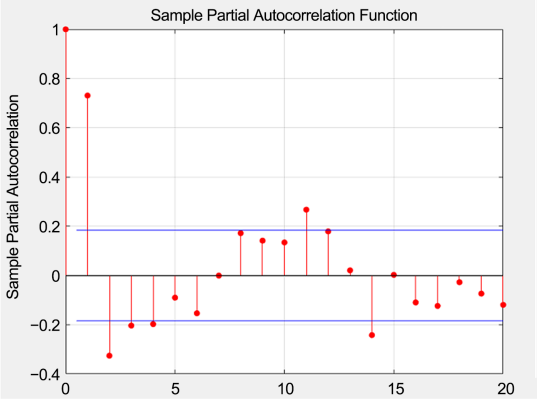


**Figure 10.** Partial autocorrelation PACF function graph
**图 10.** 偏自相关 PACF 函数图

为防止模型出现过拟合现象，需要用信息准则判断模型参数，从而选择更简单的模型。通常选用 AIC、BIC 信息两个准则：① 赤池信息准则：$\mathrm{AIC}=2K+2\ln(L)$；② 贝叶斯信息准则：$\mathrm{BIC}=K\ln(n)+2\ln(L)$，其中 $L$ 为似然函数，$K,n$ 分别为参数和取样个数。

一般情况下信息越小，模型参数越好，因此根据信息准则表去选取 AIC 值和 BIC 值均最小的自回归阶数 $p$ 和移动平均阶数 $q$，最终选择的 ARIMA 模型为 ARIMA(1,1,2) (如表 8)。

**Table 8.** ARIMA information table
**表 8.** ARIMA 信息表

| 平均最大持续风速 | | | |
|---|---|---|---|
| ARIMA(1,1,2) Model: | | | |
| Conditional Probability Distribution: Gaussian | | | |
| Parameter | Value | Standard Error | T Statistic |
| Constant | 0 | Fixed | Fixed |
| AR{1} | −0.92 | 0.07 | −12.62 |
| MA{1} | 0.0095 | 0.06 | 0.16 |
| MA{2} | −0.99 | 0.07 | −15.14 |
| Variance | 1.26 | 0.20 | 6.42 |

将平均最大持续风速的往年数据输入建立好的 ARIMA(1,1,2)模型中，就可以得到 2022~2023 年平均最大持续风速的值：如表 9 (仅展示部分数据)。

**Table 9.** Average maximum sustained wind speed prediction values
**表 9.** 平均最大持续风速预测值

| 月份 | 预测值(平均最大持续风速) |
|---|---|
| 23/08 | 9.09 |
| 23/09 | 9.23 |
| 23/10 | 8.95 |
| 23/11 | 9.80 |
| 23/12 | 9.54 |

### 5.3. 基于 LSTM 模型的预测

前面一步通过 ARIMA 模型求出平均最大持续风速、降水量等值。本文将平均最大持续风速、降水量和等往年数据作为训练集建立出 10 cm 土壤湿度的 LSTM 时间预测模型，再输入这三个因素 2022~2023 年的数据，最终预测得到 2022~2023 年 10 cm 土壤湿度的值。

由于 40 cm 土壤湿度不仅受到平均最大持续风速、降水量和平均露点温度的影响，也会受到 10 cm 土壤湿度的影响，因此需要建立平均最大持续风速、降水量、平均露点温度和 10 cm 土壤湿度这四个变量对 40 cm 土壤湿度的 LSTM 预测模型。同样 100 cm 土壤湿度也会受到 40 cm 土壤湿度的影响，以此类推，基于 LSTM 模型的预测，就可以得到 2022~2023 年不同深度下土壤湿度的预测值。

然而在模型训练过程中，本文发现，LSTM 模型对土壤湿度预测效果始终不佳，在对模型进行多次调参后，得到最优的 40 cm 土壤湿度训练结果，拟合效果不佳，得到的精度较低，如图 11 所示。因此需要思考 LSTM 模型是否为本次问题的最佳模型。经过分析后发现，造成这一现象的主要原因是在预测过程中先对特征变量进行预测，再用预测后的变量值对土壤湿度值进行预测，用预测值去预测未知值会增加模型的误差，从而使得模型拟合效果不佳。并且此次任务为较短期的预测任务，并在后续的数据分析中发现土壤湿度的数据特征和平均最大持续风速几乎趋于一致，符合 ARIMA 模型，因此优先选择 ARIMA 模型。

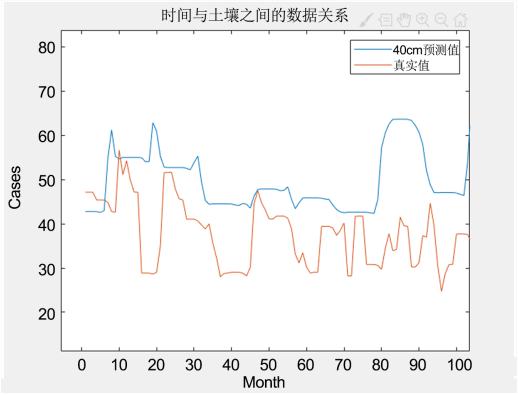


**Figure 11.** Model fitting diagram
**图 11.** 模型拟合图

本文通过数据分析发现，土壤湿度的数据特征和平均最大持续风速几乎趋于一致，符合 ARIMA 模型，模型的建立与求解基本相同，因此本文将不在此赘述，通过建立不同土壤湿度的 ARIMA 模型预测得到 2023 年不同湿度下的土壤湿度值(表 10) (仅展示部分数据)：

**Table 10.** Predicted values of soil moisture at different moisture levels
**表 10.** 不同湿度的土壤湿度预测值

| 年份 | 月份 | 10 cm 湿度 (kg/m$^2$) | 40 cm 湿度 (kg/m$^2$) | 100 cm 湿度 (kg/m$^2$) | 200 cm 湿度 (kg/m$^2$) |
|---|---|---|---|---|---|
| 2023 | 8 | 21.09 | 62.79 | 107.45 | 162.21 |
| | 9 | 21.75 | 66.01 | 109.13 | 161.91 |
| | 10 | 17.67 | 61.64 | 114.33 | 161.65 |
| | 11 | 14.65 | 59.02 | 115.23 | 161.55 |
| | 12 | 13.88 | 59.02 | 115.26 | 161.54 |

# 6. 问题三：模型的建立和求解

## 6.1. 问题三分析

本题以放牧强度、小区为一个单位，需要分别建立年份与 SOC 土壤有机碳等 5 个土壤化学相关因子的多元线性回归预测模型，并预测不同单位下 2022 年的土壤相关化学因子的值。因此分为以下两步：(1) 模型建立：一般情况下土壤性质变化受到时间的影响较大，且存在较强的线性关系，因此建立这 5 个土壤化学相关因子与年份之间的多元线性回归模型来进行求解。(2) 模型优化：对第一步建立好的模型进行评价时发现拟合效果并非最佳，因此考虑这 5 个土壤化学相关因子单独对年份做线性回归模型。

## 6.2. 多元线性回归模型

### 6.2.1. 多元线性回归模型原理

多元线性回归是多元统计分析里极其重要的分析方法之一，在实际中，一个元素的的变化常常受到多个变量因素的影响，为解释这一现象，通常采用多元线性回归模型，来描述一个因变量与其他多个自变量之间的相关关系。一般来说，多元线性回归模型的表达式为：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i \left(i = 1, 2, \cdots, n\right)$$

其中，$y_i$ 为因变量，有 $m$ 个自变量 $x_{im}\left(i = 1, 2, \cdots, n\right)$，$\beta_k\left(k = 1, 2, \cdots, m\right)$ 为回归系数，$\varepsilon_i$ 为随机误差，$\varepsilon_i \sim N\left(0, \sigma^2\right)$。写成矩阵形式：$\hat{Y} = X\hat{\beta} + \hat{\varepsilon}$。在服从正态分布的假设下，如果 $X^{\mathrm{T}}X$ 满秩，那么回归系数 $\beta$ 的最小二乘估计是：$\hat{\beta} = \left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}Y$。因此 $Y$ 的估计值 $\hat{Y} = X\hat{\beta}$，从而得到残差 $\varepsilon = Y - \hat{Y} = Y - X\hat{\beta}$，则随机误差方差 $\sigma^2$ 的最小二乘估计为：

$$\sigma^2 = \frac{\varepsilon^{\mathrm{T}}\varepsilon}{n - p - 1}$$

在求解多元线性回归模型时，我们是根据残差平方和达到最小来找到最合适的回归参数 $\beta_k\left(k = 1, 2, \cdots, m\right)$。

### 6.2.2. 数据处理

(1) 假设年份为因变量，5 个土壤化学相关变量为自变量，符号如表 11 所示：

**Table 11.** Soil chemistry related symbols table
**表 11.** 土壤化学相关符号表

| 变量 | 年份 | SOC 土壤有机碳 | SIC 土壤无机碳 | STC 土壤全碳 | 全氮 N | 土壤 C/N 比 |
|---|---|---|---|---|---|---|
| 符号 | $y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |

(2) 由于试验设计采取随机区分组形式，实验者每年在每个放牧小区将每个放牧强度都设置了 3 个重复，因此对每个放牧小区和放牧强度下的 5 个土壤化学相关变量都取平均值来替代。下面举例 G12、LGI 的平均值取法(表 12) (红色部分代表所求平均值)：

**Table 12.** Average table
**表 12.** 平均值表

| 年份 | 放牧小区 | 放牧强度 | SOC 土壤有机碳 | SIC 土壤无机碳 | STC 土壤全碳 | 全氮 N | 土壤 C/N 比 |
|---|---|---|---|---|---|---|---|
| 2012 | G12 | LGI | 13.90 | 8.13 | 22.04 | 1.93 | 11.40 |
| 2012 | G12 | LGI | 11.03 | 12.85 | 23.88 | 1.36 | 17.52 |
| | | | 12.47 | 10.49 | 22.96 | 1.65 | 14.46 |

### 6.2.3. 多元回归模型的求解

本题利用 Matlab 针对不同放牧强度及小区分别去求解年份与 5 个土壤化学相关变量的多元线性回归模型，由于有 12 种放牧小区，因此可以建立出 12 个多元线性回归模型。

基于模型计算得到多元线性回归方程为(以放牧小区 G17、放牧强度 NG 为例)：

$$y = 1997.96 + 0.84x_1 + 0.16x_2 - 0.14x_3 + 14.23x_4 + 3.24x_5$$

同理可得不同放牧强度下 12 个放牧小区的多元线性回归模型。为验证模型的拟合效果，利用*P*值和置信区间去检验每个多元线性回归模型的回归系数，发现 *P* 值几乎都远远大于 0.05，在统计意义上说明做的多元线性回归模型拟合效果并不好，为此下一步会对多元线性回归模型进行优化。表 13 为四种不同放牧强度下随机选取的一个放牧小区的多元线性回归系数的 *P* 值和置信区间表(仅展示对照组的数据)：

**Table 13.** Confidence table for control group
**表 13.** 对照组置信表

| | | 常量 | SOC | SIC | STC | N | C/N |
|---|---|---|---|---|---|---|---|
| *P* 值 | | 0.1350 | 0.1098 | 0.1055 | 0.0180 | 0.0080 | 0.1809 |
| 置信区间 | 下界 | 1880.9675 | −1.4285 | 0.1130 | −2.1133 | −28.3955 | 3.1133 |
| | 上界 | 2072.1356 | 3.1382 | 0.1339 | 1.7574 | 60.9216 | 4.7772 |

## 6.3. 多元线性回归模型的优化

我们观察每一个散点图可以发现在不同放牧强度和不同放牧小区的情况下，5 个土壤化学相关变量分别对年份具有一定的线性关系，因此可以考虑针对这 12 个放牧小区，分别建立 5 个化学土壤相关变量对年份的线性回归方程。

在不同放牧强度和不同放牧小区的情况下，利用 Matlab 软件分别建立 5 个土壤化学相关变量对年份的线性回归方程：$y = Kx + b$ (其中 $K$ 为回归系数，$b$ 为常数项)。以单个放牧小区和放牧强度为一个放牧点，总共可以得到 60 个线性回归模型。

以放牧小区 G12、放牧强度 LGI 为例，可以计算得到 5 个线性回归模型：$y = 2.02x_1 + 1986.33$，$y = -0.99x_2 + 2023.08$，$y = -1.39x_3 + 2046.46$，$y = 26.43x_4 + 1968.37$，$y = -1.60x_5 + 2036.14$。

同理对剩余的放牧点做线性回归方程，得到不同放牧小区和不同放牧程度下 5 个土壤化学相关变量分别对年份的线性回归系数表 14 (仅展示部分数据)：

**Table 14.** Regression coefficient table of soil chemistry related variables
**表 14.** 土壤化学相关变量的回归系数表

| 放牧强度 | 放牧小区 | 回归系数 | SOC 土壤有机碳 | SIC 土壤无机碳 | STC 土壤全碳 | 全 N | 土壤 C/N 比 |
|---|---|---|---|---|---|---|---|
| HGI | G9 | *K* | 1.63 | −0.89 | −1.46 | 8.34 | −0.87 |
| | | *b* | 1991.95 | 2022.57 | 2048.45 | 2000.86 | 2027.54 |
| | G13 | *K* | 1.47 | −0.96 | −1.99 | 11.07 | −1.02 |
| | | *b* | 1994.56 | 2022.60 | 2058.67 | 1996.23 | 2028.90 |
| | G20 | *K* | 1.37 | −1.00 | −0.74 | 22.96 | −1.04 |
| | | *b* | 1995.31 | 2024.26 | 2033.28 | 1972.70 | 2029.67 |

建立好模型后预测得到 12 个放牧小区 2022 年在不同放牧强度下土壤化学相关变量的值，最终结果如表 15 所示(仅展示部分数据)：

**Table 15.** Predicted values of soil chemistry related variables
**表 15.** 土壤化学相关变量的预测值

| 放牧强度 | 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 |
|---|---|---|---|---|---|---|
| HGI | G9 | 18.43 | 0.64 | 19.08 | 2.53 | 7.53 |
| | G13 | 18.63 | 0.62 | 19.26 | 2.33 | 8.27 |
| | G20 | 19.54 | 2.27 | 21.81 | 2.15 | 10.16 |

## 7. 问题四：模型的建立和求解

### 7.1. 问题四分析

问题四是要求在给定的沙漠化程度指数值和土壤板结化定性模型下，求出这两者值最小的放牧策略，因此需要分别对沙漠化程度指数和土壤板结化定性指数进行分析[6] [7]：

(1) 第一步主要是求出沙漠化程度指数预测模型表达式中的$\eta$值，表达式如下：

$$SM = \eta * \sum_{i=1}^{n} S_{Q_i} = \eta * \sum_{i=1}^{n} \left( Q_i * W_{c_i} \right)$$

其中$SM$为沙漠化程度指数，$\eta$为调节系数，$Q_i$为第$i$个指标因子的因子强度，$S_{Q_i}$为第$i$个因子对沙漠化程度的贡献度，$W_{c_i}$为第$i$个因子权重系数。为计算出$\eta$的值，首先需要计算出每个指标因子的因子强度$Q_i$和因子权重系数，从而求得每个指标对沙漠化程度的贡献值$S_{Q_i}$。最后对沙漠化程度$SM$和贡献值$S_{Q_i}$做线性拟合得到$\eta$的值。

(2) 第二步目的是求出不同放牧程度下土壤板结化程度值，为此需要先根据题目给出的定性数学模型$B = f(W, C, O)$建立模型求解，以土壤蒸发量作为土壤板结化的表征，建立土壤湿度、容重和土壤有机物等四个变量对土壤板结化的线性回归模型，再将不同放牧强度下的数值输入模型中去，得到四种放牧强度下土壤板结化程度值，绘制线性图去判断出在轻度放牧强度下的土壤板结化最小。

### 7.2. 沙漠化程度指数

#### 7.2.1. 因子强度

结合前人研究，现代的沙漠化过程主要受自然因素于人文因素的影响，其中自然因素又可以分为气象因素和地表因素两方面。本文首先选定气象因素中气温、风速、降水[9]；地表因素中的植被指数、地表水资源、地下水位，以及人文因素中的人口数量、牲畜数量[10]、社会经济水平共九个特征因子用以表征沙漠化程度。其中气象因素均以月平均数据表征，地表水资源和地下水位分别以地下 10 cm、地下 200 cm 土壤湿度表征。人文特征分别以统计年鉴中牧区人口数量，牲畜数量以及牧区人均经营性收入数据表征[11]。

由于其中有 4 个指标因子的数据存在过多缺失值，直接删除。至此，基于统计年鉴表和附件数据得到 5 个沙漠化相关指标因子的值，如表 16 所示(仅展示部分数据)：

**Table 16.** Desertification related indicator factor values
**表 16.** 沙漠化相关指标因子值

| 时间 | 植被指数(NDVI) | 平均气温(℃) | 降水量(mm) | 平均风速(knots) | 10 cm 湿度(kg/m$^2$) |
|---|---|---|---|---|---|
| 21/8 | 0.542 | 18.18 | 37.34 | 5.47 | 20.36 |
| 21/9 | 0.51 | 14.61 | 65.53 | 5.29 | 21.01 |
| 21/10 | 0.325 | 4.42 | 4.83 | 5.22 | 16.93 |
| 22/3 | 0.187 | −2.36 | 115.57 | 6.37 | 14.96 |

为判断这 9 个因子的因子强度($Q_i$取值范围为$[0,1]$)，主要分为以下两步：

① 单位标准化：使得这 5 个指标因子的统计年鉴数据单位与因子分级标准表中的单位相统一，比如统计年鉴表里的风速单位为 knots，需要统一为 m/s。② 因子强度分级：针对因子强度的分级界定，本文参考文献[12]量化特征因子的强度。具体量化公式如下：

$$Q(c)=\begin{cases}0 & c\le c_{down}\\ \dfrac{c-c_{down}}{c_{up}-c_{down}} & c_{down}\le c\le c_{up}\\ 1 & c\ge c_{up}\end{cases}$$

其中 $c$ 为特征因子的实测值，$c_{up},c_{down}$ 分别为特征因子影响沙漠化的上下限取值。

本文参考联合国关于影响沙漠化特征因子的分级标准和相关文献，给出如表 17：

**Table 17.** Factor upper and lower limit table
**表 17.** 因子上下限表

| 指标因子 | 上限 | 下限 |
|---|---|---|
| 植被指数(NDVI) | 0.5 | 0.1 |
| 平均气温(℃)) | 35 | 5 |
| 降水量(mm) | 100 | 10 |
| 平均风速(knots) | 2.61 | 1.43 |
| 10 cm 湿度(kg/m$^2$) | 21.47 | 1.22 |

经过数据处理后可以得到这 5 个指标因子的因子强度表如表 18 (仅展示部分数据)：

**Table 18.** Factor intensity table of desertification related indicators
**表 18.** 沙漠化相关指标的因子强度表

| 时间 | 植被指数(NDVI) | 平均气温(℃) | 降水量(mm) | 平均风速(knots) | 10 cm 湿度(kg/m$^2$) |
|---|---|---|---|---|---|
| 21/8 | 1.00 | 0.44 | 0.30 | 1.00 | 0.95 |
| 21/9 | 1.00 | 0.32 | 0.62 | 1.00 | 0.98 |
| 21/10 | 0.56 | 0.00 | 0.00 | 1.00 | 0.78 |
| 22/3 | 0.22 | 0.00 | 1.00 | 1.00 | 0.68 |

### 7.2.2. 因子权重系数和因子贡献值

得到沙漠化指标因子的因子强度后，需要对因子强度计算因子权重系数。常见的权重计算方法有因子分析、主成分分析和层次分析等，但因子分析主要利用信息浓缩原理，且其“旋转”功能使得因子具有更强的解释意义，因此在计算因子权重系数上选取因子分析法。因子分析法最早由英国数学家斯皮尔曼提出的一种基于降维思想的统计分析方法，它通过研究变量之间的相关系数矩阵，基于数据信息浓缩原理利用最大方差法对载荷因子进行旋转得到因子模型，计算因子得分从而得到权重系数，图 12 为因子分析模型：

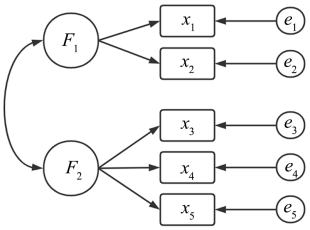


**Figure 12.** Factor analysis model
**图 12.** 因子分析模型

因子分析的基本步骤主要分为以下 4 步：

(1) 首先将数据标准化处理，为消除数据间的量纲影响。(2) 使用方差最大化旋转对处理好的数据做因子分析。(3) 计算每个主因子得分和方差贡献率 $e_j$，计算公式为：$F_j = \beta_{1j}x_1 + \beta_{2j}x_2 + \cdots + \beta_{nj}x_n$ $(j = 1,2,\cdots,m)$，其中 $F_i$ 为主成分，$x_1, x_2, \cdots, x_n$ 为各个指标因子，$\beta_{1j}, \beta_{2j}, \cdots, \beta_{nj}$ 为各个指标因子在每个主成分里的系数得分。(4) 求出各个指标因子的权重，计算公式为：$\omega_i = \frac{m\sum_{j=1}^{m}\beta_{ij} * e_j}{n\sum_{i=1}^{n}m\sum_{j=1}^{m}\beta_{ij} * e_j}$。

将这 5 个指标因子(植被指数、平均气温和降水量等)的因子强度做因子分析，可以得到因子权重系数表 19：

**Table 19.** Factor weight coefficient table of desertification related indicators
**表 19.** 沙漠化相关指标的因子权重系数表

| | 植被指数(NDVI) | 平均气温(℃) | 降水量(mm) | 平均风速(knots) | 10 cm 湿度(kg/m$^2$) |
|---|---|---|---|---|---|
| 权重 $W_{c_i}$ | 0.4864 | 0.2671 | 0.2340 | 0.0000 | 0.0125 |

得到因子强度和因子权重系数后，根据表达式中的 $S_{Q_i} = W_{c_i} * Q_i$，可以计算出因子对沙漠化程度的贡献值总和如表 20 (仅展示部分数据)：

**Table 20.** Contribution value table of desertification related indicator factors
**表 20.** 沙漠化相关指标因子的贡献值表

| 时间 | 21/5 | 21/6 | 21/7 | 21/8 | 21/9 | 21/10 | 22/3 |
|---|---|---|---|---|---|---|---|
| 因子贡献值 | 0.2526 | 0.5285 | 0.7512 | 0.6866 | 0.7286 | 0.2833 | 0.3483 |

### 7.2.3. 沙漠化程度指数

由于缺少锡林郭勒盟沙漠化程度指数的数据，本文参考相关文献[13]，以绿植覆盖率表征沙漠化程度指数。

**Table 21.** Standard table for classifying desertification degree and desertification degree index
**表 21.** 沙漠化程度及沙漠化程度指数划分标准表

| 划分内容 | 划分类型 | | | | |
|---|---|---|---|---|---|
| 沙漠化程度 | 非沙漠化 | 轻度沙漠化 | 中度沙漠化 | 重度沙漠化 | 极重度沙漠化 |
| 沙漠化程度指数 | [0,0.20] | (0.20,0.40] | (0.40,0.60] | (0.60,0.80] | (0.80,1.00] |

**Table 22.** Table of criteria for classification of desertification types
**表 22.** 沙漠化类型划分标准表

| 沙漠化类型 | 轻度沙漠化 | 中度沙漠化 | 重度沙漠化 | 严重沙漠化 |
|---|---|---|---|---|
| 绿植覆盖率(%) | 50~80 | 30~50 | 10~30 | 0~10 |

绿植覆盖率越高，沙漠化程度越小。对比表 21 和表 22，本次问题将沙漠化程度指数选定为 $SM = 1-$绿植覆盖率，得到如表 23 (仅展示部分数据)：

**Table 23.** Desertification related data table
**表 23.** 沙漠化相关数据表

| 时间 | 绿植覆盖率(%) | *SM* | $S_{Q_i}$ |
|---|---|---|---|
| 21/9 | 0.25 | 0.75 | 0.7286 |
| 21/10 | 0.09 | 0.91 | 0.2833 |
| 22/3 | 0.02 | 0.98 | 0.3483 |

已知 $SM$ 和 $S_{Q_i}$ 的值，结合公式 $SM=\eta*\sum_{i=1}^{n}S_{Q_i}$ 对调节系数 $\eta$ 进行线性拟合得到 $\eta=1.2722$。对拟合结果进行验证，发现 $P$ 值($P$ = 0.0049)远远小于 0.05，具有较高的统计学意义。求出 $\eta$ 的值后最终得到沙漠化程度指数预测模型表达式为：

$$SM=1.2722\sum_{i=1}^{n}S_{Q_i}=1.2722\sum_{i=1}^{n}\left(Q_i*W_{c_i}\right)$$

## 7.3. 土壤板结化程度

已知土壤板结化程度与土壤湿度、容重等有关，土壤板结化程度的定性数学模型为 $B=f\left(W,C,O\right)$，其中 $W$ 为土壤湿度，$C$ 为土壤容重，$O$ 为土壤有机物。土壤湿度越小，土壤容重越大，土壤有机物越少，则板结化程度越严重[14]。

### 7.3.1. 模型的建立

首先对数据进行处理(由于土壤容重为固定值 1.39，因此数据部分将不对土壤容重进行处理)：(1) 土壤板结化程度：由于土壤板板结化程度越严重会导致土壤蒸发量越低，因此在计算过程中本文均用土壤蒸发量代替土壤板结化程度。然后对土壤蒸发量进行[0,0.6]归一化处理，得到土壤蒸发量系数，则土壤板结化程度 = 1 − 土壤蒸发量系数。(2) 土壤有机物含量 = SOC 土壤有机碳 + SIC 土壤无机碳。

表 24 为土壤板结化程度的相关变量数值(仅展示部分数据)：

**Table 24.** Numerical table of variables related to the degree of soil compaction
**表 24.** 土壤板结化程度的相关变量数值表

| 时间 | 土壤蒸发量 ($W/m^2$) | 土壤板结化程度 | SOC 土壤有机碳 | SIC 土壤无机碳 | 全氮 N | 10 cm 湿度 ($kg/m^2$) |
|---|---|---|---|---|---|---|
| 21/10 | 9.78 | 0.1540 | 17.5565 | 1.6687 | 2.3819 | 16.93 |
| 21/11 | 1.39 | 0.0174 | 17.5565 | 1.6687 | 2.3819 | 13.91 |
| 21/12 | 0.87 | 0.0090 | 17.5565 | 1.6687 | 2.3819 | 13.14 |

(3) 皮尔逊相关系数分析：

为验证选出来的相关变量与土壤板结化程度的确有一定程度上的关系，因此得到皮尔逊相关系数表 25：

**Table 25.** Pearson correlation coefficient table
**表 25.** 皮尔逊相关系数表

| 变量 | 相关系数 |
|---|---|
| 全氮 N | 0.15 |
| 10 cm 湿度($kg/m^2$) | 0.81 |
| 土壤有机物 | 0.27 |

然后利用 Matlab 软件建立得到土壤容重、全氮 $x_1$、10 cm 土壤湿度 $x_2$、土壤有机物 $x_3$ 这 4 个变量对土壤板结化程度的多元线性回归模型：$y=1.1976+0.0181x_1-0.0437x_2-0.0094x_3$。

得到多元线性回归模型后，再对模型的回归系数做验证，发现$P$值几乎都小于 0.05，符合统计学意义上的相关性。同时得到模型的残差图 13，发现拟合效果较佳，因此后续求解不同放牧强度下的土壤板结化程度都使用这个多元线性回归模型。

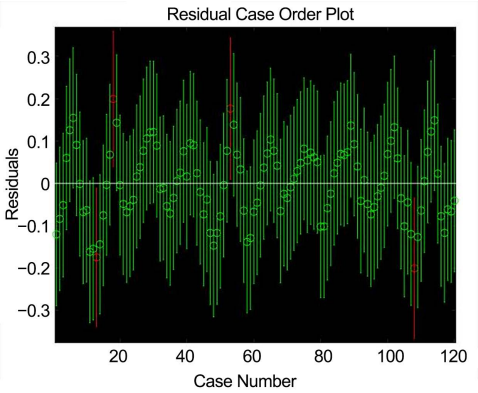


**Figure 13.** Residual plot
**图 13.** 残差图

### 7.3.2. 模型的求解

由于题目要求在土壤板结化的基础上给出放牧策略，因此针对这一问题，需要分别将不同放牧强度下的土壤容重、全氮 $x_1$、10 cm 土壤湿度 $x_2$、土壤有机物 $x_3$ 这 4 个变量的数据输入到土壤板结化程度的多元线性回归模型：$y=1.1976+0.0181x_1-0.0437x_2-0.0094x_3$，得到四种放牧强度下的土壤板结化程度表(表 26)和数据图(仅展示部分数据)：

**Table 26.** Soil compaction degree table
**表 26.** 土壤板结化程度表

| 时间 | 对照 | 轻度放牧 | 中度放牧 | 重度放牧 |
|---|---|---|---|---|
| 21/10 | 0.4616 | 0.4789 | 0.4955 | 0.4923 |
| 21/11 | 0.5407 | 0.5581 | 0.5747 | 0.5715 |
| 21/12 | 0.5609 | 0.5783 | 0.5948 | 0.5917 |

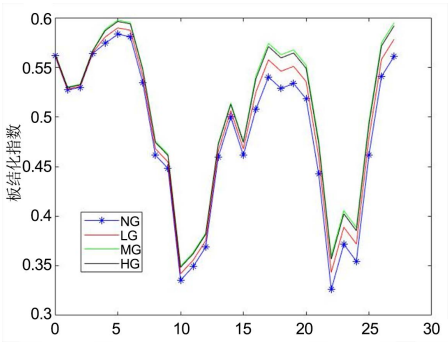


**Figure 14.** Map of soil compaction degree under different grazing intensities (NG is control, LG is light grazing, MG is moderate grazing, and HG is heavy grazing)
**图 14.** 不同放牧强度下土壤板结化程度图(NG 为对照，LG 为轻度放牧，MG 为中度放牧，HG 为重度放牧)

观察图 14 可以知道在轻度放牧强度的情况下的土壤板结化程度最小(由于 NG 为对照组故不予计入在内)，因此基于沙漠化程度指数值和土壤板结化定性模型就可以得出当放牧策略为轻度放牧时，沙漠化程度指数与板结化程度最小。

## 8. 模型评价

### 8.1. 模型优点

(1) 本文在解决每一问题时都对数据进行预处理，提高数据高效性；

(2) 参考联合国关于影响沙漠化特征因子额分级标准等文献，选取影响草原生态的关键特征，提高特征选择的有效性、降低实际挖掘所需有效影响因子的时间；

(3) 本文在分析解决问题时，充分考虑数据特点，并使用多个模型进行比较检验，选取最优解。

(4) 模型的计算成本低。

### 8.2. 模型缺点

(1) 由于时间紧张，在第二问中本文只依据 LSTM 模型对 10 mm、40 mm 处土壤湿度预测效果对其进行评价，具有一定片面性，同时未对其进行优化，后续可考虑组合模型对 LSTM 的初步预测结果中的残差进行训练拟合，从而提高模型的预测效果；

(2) ARIMA 模型在预测数据时仅考虑时序关系，忽略了实际生活中其他因素如新冠疫情等突发情况对草原发展变化的影响。

## 参考文献


[1] Katarzyna, P., Tembeck, M.J., Krzysztof, P., Gniewko, N. and Tomasz, W. (2022) Evaluation of Multiple Linear Regression and Machine Learning Approaches to Predict Soil Compaction and Shear Stress Based on Electrical Parameters. *Applied Sciences*, **12**, 8791. https://doi.org/10.3390/app12178791

[2] Woodward, S., Wake, G.C., *et al*. (1993) A Simple Model for Optimizing Rotational Grazing. *Agricultural Systems*, **41**, 123-155. https://doi.org/10.1016/0308-521X(93)90037-3

[3] 于世龙, 杨奉广. 某流域年均含沙量的非线性回归分析[J]. 吉林水利, 2022(8): 6-14.

[4] 宫海静, 王德利. 草地放牧系统优化模型的研究进展[J], 草业学报, 15(6): 1-8.

[5] 李阳, 杜睿山, 程永昌. 基于 ARIMA-LSTM 组合模型的原油产量时序预测研究[J]. 数学的实践与认识, 2022, 40-48.

[6] Xiao, X.X., Lv, W.C., Han, Y.C., Lu, F.K. and Liu, J.T. (2022) Prediction of CORS Water Vapor Values Based on the CEEMDAN and ARIMA-LSTM Combination Model. *Atmosphere*, **13**, 1453. https://doi.org/10.3390/atmos13091453

[7] Kong, L.Z., Li, G.S., Wajid, R., Shen, S.G., He, Q., Khosravi, M.R., Wang, R.L. and Qi, L.Y. (2022) Time-Aware Missing Healthcare Data Prediction Based on ARIMA Model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

[8] 杨洋, 程悦菲, 谯英, 刘炯. 基于时序动态分析的油井产量预测研究[J]. 西南石油大学学报(自然科学版), 2020, 42(6): 82-88.

[9] 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(5): 197-203.

[10] 王悦骅. 模拟降水对不同载畜率放牧荒漠草原植物多样性的影响[D]: [硕士学位论文]. 呼和浩特: 内蒙古农业大学, 2019.

[11] 许宏斌, 辛晓平, 宝音陶格涛, 等. 放牧对呼伦贝尔羊草草甸草原生物量分布的影响[J], 草地学报, 2020, 28(3): 768-774.

[12] 刘敦利. 基于栅格尺度的土地沙漠化预警模式研究[D]: [硕士学位论文]. 乌鲁木齐: 新疆大学, 2010.

[13] 张洁. 阿拉善左旗沙区生态产业化发展适宜性研究[D]: [硕士学位论文]. 呼和浩特: 内蒙古农业大学, 2021.

[14] 刘铠诚, 郭睿, 刘容川, 乔洪磊. 基于多元回归模型的钙质土固结特性优化模型[J]. 科技与创新, 2022(16): 67-70.