普通线性回归、Lasso与Ridge及其融合模型在径流预测中的应用与评估

|  |  |
| --- | --- |
| **小组成员：** | **包一宁 赵知微 林月磊** |
|  | **林谷 唐玉彬** |

[摘要 1](#_Toc210242971)

[第一章 问题背景及意义 2](#_Toc210242972)

[1.1 研究背景 2](#_Toc210242973)

[1.2 问题意义 2](#_Toc210242974)

[第二章 数据预处理 4](#_Toc210242975)

[2.1 数据校验 4](#_Toc210242976)

[2.1.1 缺失值校验 4](#_Toc210242977)

[2.1.2 异常值校验 4](#_Toc210242978)

[2.2 特征工程 6](#_Toc210242979)

[第三章 多模型集成径流预测方法构建 7](#_Toc210242980)

[3.1 模型选取与集成 7](#_Toc210242981)

[3.2 模型原理 7](#_Toc210242982)

[第四章 模型评估及结果分析 7](#_Toc210242983)

[4.1 模型评估 7](#_Toc210242984)

[4.2 结果分析 7](#_Toc210242985)

[第五章 模型讨论 7](#_Toc210242986)

[5.1 结果解释与成因分析 7](#_Toc210242987)

[5.2 模型局限性 7](#_Toc210242988)

[5.3 模型优化方向 7](#_Toc210242989)

[结论 7](#_Toc210242990)

[参考文献 7](#_Toc210242991)

摘要

1. 问题背景及意义
   1. 研究背景

水资源是维系生态平衡与支撑社会经济发展的关键自然要素，其可持续利用已成为全球关注的焦点。径流作为水文循环的核心组成部分，直接关系到区域水资源的分布、可利用量及灾害风险程度。随着全球气候变化与人类活动干扰的加剧，径流过程呈现出更强的非线性和时空异质性，使得传统水文模型在预测精度与适应性方面面临严峻挑战。在此背景下，数据驱动模型因其不依赖复杂的物理机制、仅通过历史数据构建预测关系的优势，逐渐成为径流预报的重要研究方向之一。

线性回归作为一类基础且广泛使用的数据驱动方法，因其模型结构简单、解释性强，被广泛应用于径流预测中。然而，普通线性回归模型在处理高维特征或存在多重共线性的数据时，容易出现过拟合问题，导致模型泛化能力下降。为提升预测稳定性与准确性，正则化技术被引入线性回归框架中，如Lasso回归通过L1正则化实现特征选择，Ridge回归通过L2正则化抑制参数膨胀，而Elastic Net则结合二者优势，在保持模型简洁性的同时增强鲁棒性。此外，通过集成学习与多模型耦合策略，进一步整合不同回归模型的优势，已成为提高径流预测性能的有效途径。

当前，尽管已有研究尝试将正则化线性回归应用于水文预报，但多数工作仍集中于单一模型的对比，缺乏系统性的多模型耦合与集成策略探讨，尤其在面对不同流域特性与水文情势时，模型的适应性与解释性仍有待深化。因此，开展基于线性回归的集成与正则化改进研究，对提升径流预测的准确性与可靠性具有重要理论与实用价值。

* 1. 问题意义

本研究围绕线性回归模型在径流预测中的优化与集成展开，重点探讨普通线性回归、Lasso、Ridge及其耦合模型的表现差异与适用条件，具有如下意义：

在理论层面，本研究通过系统比较不同正则化策略与模型融合方式，有助于揭示线性回归模型在水文序列预测中的泛化机制与抗干扰能力，推动数据驱动水文模型向更稳健、更可解释的方向发展。同时，模型耦合策略的研究为多算法协同建模提供新思路，丰富了水文预报的方法体系。

在实践层面，径流预测的准确性直接关系到水资源调度、洪水防控、干旱应对及水利工程运行等关键决策。本研究提出的正则化线性回归及其集成模型，能够有效提升预测精度与时效性，为区域水资源管理、灾害预警系统优化提供技术支撑，尤其在数据稀缺或异质性强的流域中展现出良好的应用潜力。此外，通过模型可解释性增强，有助于决策者理解预测结果背后的驱动因素，提升水资源管理的科学性与透明度。

综上所述，本研究不仅对线性回归模型在水文领域的应用进行了深化与拓展，也为应对复杂环境下的水资源挑战提供了方法论支持，具有较强的学术价值与现实意义。

1. 数据预处理
   1. 数据校验

本研究基于文件“qingshandataforregression.xlsx”中的降雨、蒸发和径流数据进行线性回归模型拟合。为确保数据质量并提升模型拟合的准确性和效率，在建模之前，必须对数据进行必要的预处理。这将有助于优化拟合过程，并保证最终模型的可靠性和良好性能。

* + 1. 缺失值校验

缺失值处理是数据预处理的首要步骤。本研究利用Python程序对文件中三个变量（降雨量、蒸发量、径流量）进行了校验。得到表 1：

表 1：降雨量、蒸发量和径流量缺失值检验结果表

|  |  |  |
| --- | --- | --- |
| 数据类型 | 缺失值数目 | 数据类型 |
| 降雨量 | 0 | dtype: int64 |
| 蒸发量 | 0 | dtype: int64 |
| 径流量 | 0 | dtype: int64 |

校验结果显示，三个变量的缺失值数目均为0。这表明数据集的完整性高，无需进行缺失值填充或删除处理，可以直接进入下一步的异常值校验。

* + 1. 异常值校验

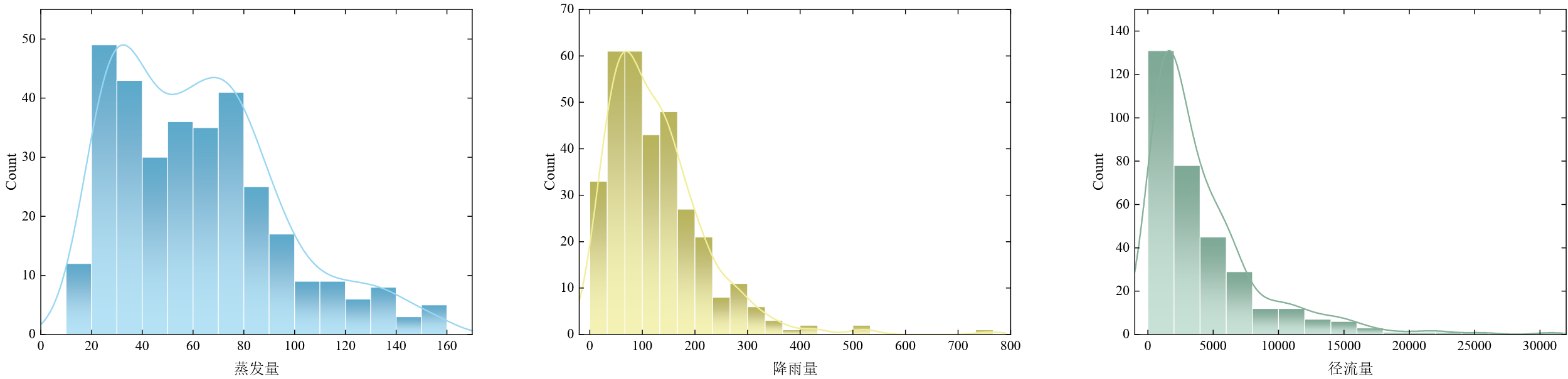
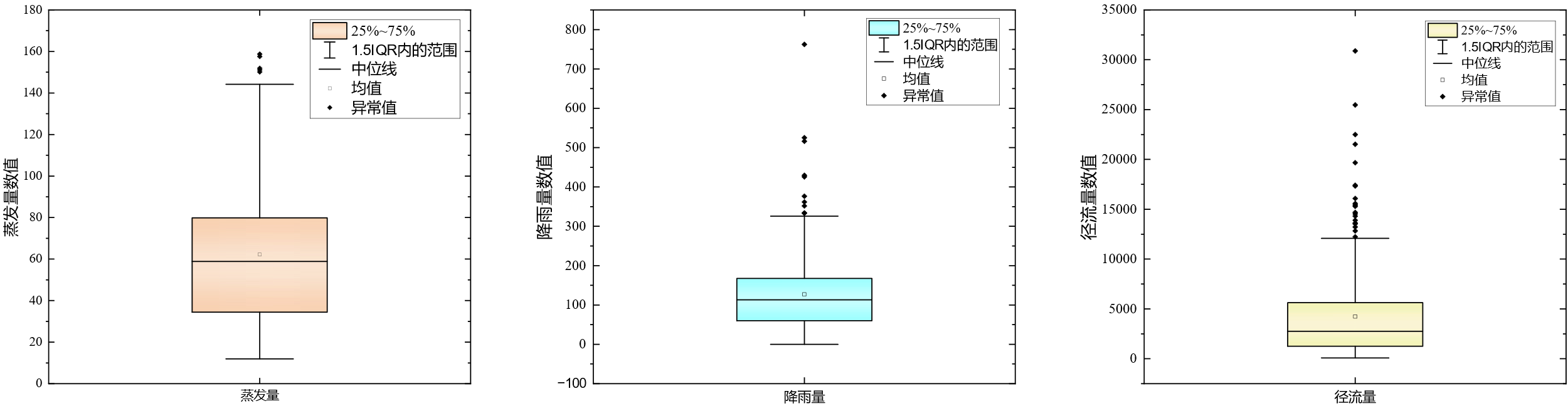
依据统计学原理，在进行异常值校验之前，先要对数据集进行偏度和正态性检测，即分布特征分析，通过Python得到图 1：

图 1：数据分布特征图

结果表明，三项数据均存在一定程度的右偏，其中降雨量与径流量数据右偏较为严重。

基于此判断，本研究后续采用四分位数间距（Interquartile Range, IQR）方法对径流预测模型的输入变量进行异常值检测。IQR方法基于数据的分位数特征，具有对极端值不敏感的优点，适用于非正态分布数据的异常值识别。检测标准为：

其中，和分别为第一四分位数和第三四分位数，为四分位数间距。

本研究利用IQR方法绘制箱线图对数据进行可视化以定位异常值，结果如图 2所示：

显然在不同数据列中均存在一定量的异常值，这也符合偏度预期，进一步分析得到表 2：

图 2：蒸发量、降雨量和径流量箱线图

表 2：异常值分析表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 变量 | 异常值数量 | 异常值比例(%) | 正常值范围 | 最大异常值 | 最小异常值 |
| 蒸发量(mm) | 5 | 1.52 | [-33.02, 147.57] | 158.70 | 150.10 |
| 降雨量(mm) | 10 | 3.05 | [-99.48, 326.93] | 762.50 | 333.10 |
| 径流量(m³/s) | 20 | 6.10 | [-5284.50, 12173.50] | 30890.00 | 12230.80 |

其中，**蒸发量变量**表现出较好的数据质量，仅有5个异常值（占比1.52%），均为上界异常值，数值范围为150.10-158.70 mm。这些异常值主要出现在极端高温或干旱条件下，符合气象学规律，属于自然现象的极端表现。**降雨量变量**检测出10个异常值（占比3.05%），最大异常值达762.50 mm，为均值的6.02倍。这些异常值反映了研究区域极端降水事件的存在，如暴雨、台风等极端天气现象。降雨量的变异系数为73.56%，表明其时空变异性较大。**径流量变量**异常值数量最多，共20个（占比6.10%），最大异常值为30890.00 m³/s，约为均值的7.31倍。径流量的变异系数高达103.71%，表明其具有显著的不均匀分布特征。这些异常值主要对应于洪水事件，是降雨量异常值的直接响应。

* 1. 特征工程

数据校验结果显示，径流量和降雨量均存在明显的右偏分布，特别是径流量的偏度值接近 2.315。同时，为保留水文系统极端事件的完整信息，本研究采取了保留异常值的策略。鉴于这些特性，必须对数据进行变换，以缓解异常值对模型拟合的不利影响，并使数据分布更接近线性回归模型所要求的正态分布假设。此外，特征变量之间的量纲和取值范围差异必须消除，以确保后续Lasso和Ridge等正则化模型的良好拟合。本研究采用双重策略进行数据变换：

（1）首先对于具有严重右偏特性的径流量和降雨量，采用对数变换进行处理。对数变换能有效将右偏分布拉向对称，显著降低偏度，使数据分布更接近正态。同时，它能够压缩极端高值的尺度，从而减轻保留的异常值对模型系数的过度影响，提升模型的稳健性。具体而言，即采用的形式。

表 3：对数变换后的数据偏度表

|  |  |
| --- | --- |
| 变量 | 偏度 |
| 蒸发量 | 0.7649 |
|  | -1.0561 |
|  | -0.0984 |

根据表 3可以发现经过对数变换后径流量偏度得到了有效减少，接近理想的对称状态，极大地优化了模型的拟合条件。虽然降雨量偏度从1.9855(高度右偏)到1.0561(中度左偏)显著改善了数据分布，但距离精确的正态分布仍存在一定偏差。不过，考虑到气象数据保留极端值信息的需求，-1.0561的偏度仍在可接受范围内。

（2）由于不同变量之间存在量纲区别，并且蒸发量仍保持中度右偏（0.7649），为了使其与经过变换的降雨量和径流量特征保持尺度一致性，并满足正则化模型的要求，需要对其进行Z-Scroe标准化。

Z-Score标准化，也称为标准化或Z标准化，它能将原始数据集中的数据点转换成标准分数，从而使转换后的数据集具有均值为0和标准差为1的特性。Z-Score的计算公式非常直观：

其中：Z是原始数据点x对应的Z-Score，x是数据集中的原始数据点（或称为观测值），是数据集的平均值，是数据集的标准差。

通过Python可以直接计算出标准化后的数据的均值和标准差均符合预期。至此，数据集已进行一定的特征工程，可进行数据划分为训练集和测试集以供后续模型拟合。

1. 多模型集成径流预测方法构建
   1. 模型选取与集成
   2. 模型原理
2. 模型评估及结果分析
   1. 模型评估
   2. 结果分析
3. 模型讨论
   1. 结果解释与成因分析
   2. 模型局限性
   3. 模型优化方向

结论

参考文献

刘昌明,陈志恺.中国水资源现状评价和供需发展趋势分析[M].北京:中国水利水电出版社,2001.

夏军,谈戈.全球变化与水文科学新的进展与挑战[J].水科学进展,2002,13(5):667-674.

程慧先.基于傅里叶变换、近似熵和线性回归的数据驱动径流预测模型及机理揭示[D].北京:中国水利水电科学研究院,2020.

黄瑾.岷江上游生态水遥感定量反演及径流预测模型研究[D].成都:成都理工大学,2017.

TIBshiraniR.Regressionshrinkageandselectionviathelasso[J].JournaloftheRoyalStatisticalSociety:SeriesB(Methodological),1996,58(1):267-288.

王浩,严登华,贾仰文,等.水文模型研究进展与展望[J].水科学进展,2010,21(4):451-457.

陈庆秋,薛联青,周祖昊.基于耦合模型的水资源系统模拟研究进展[J].水利学报,2005,36(10):1155-1161.

张建云,王宗志,刘克琳,等.变化环境下水文水资源系统模拟与调控研究[J].水利学进展,2014,25(5):735-742.