

基于GAC特征选择和AST回归算法的森林地上碳储量遥感反演

孙凯萍¹, 张加龙², 滕晨凯², 杨坤², 黄凯³, 雷启旺², 熊登亮⁴

1. 西南林业大学水土保持学院, 云南 昆明 650224;
2. 西南林业大学林学院(亚太林学院), 云南 昆明 650224;
3. 昆明理工大学国土资源工程学院, 云南 昆明 650093;
4. 云南省基础测绘技术中心, 云南 昆明 650034

摘要: 由于单一机器学习算法无法充分挖掘数据中的潜在信息, 本研究提出一种AST回归算法(AdaBoost-Stacking Tree-based, AST)以提高森林地上碳储量(Aboveground Carbon Stock, AGC)估测的稳定性和准确性。以云南省香格里拉地区的高山松为研究对象, 采用森林资源二类调查数据和Landsat 8 OLI影像, 构建融合遗传算法与CatBoost的特征选择方法(Genetic Algorithm and CatBoost, GAC), 并与递归特征消除法(RFE)分别进行遥感变量筛选, 采用Hyperopt超参数优化框架对各模型进行超参数调优, 将四种单一机器学习回归模型集成提升(AdaBoost、CatBoost、随机森林(RFR)、轻量级梯度提升(LightGBM))进行堆叠集成形成AST回归算法, 该算法通过基学习器的均值融合与元学习器自适应加权实现模型优化; 通过比较六种单一模型与AST集成模型的精度, 采用最优模型进行高山松碳储量反演及结果的不确定性制图。结果表明: 1) RFE筛选出9个变量, GAC筛选出7个变量, 其中GAC筛选出的7个变量对高山松AGC反演精度贡献更大; 2) 基于Hyperopt对各模型的超参数进行迭代调优, 发现GAC筛选出的最优特征子集与AST算法结合进行回归拟合时获得了最优的估测精度, 其决定系数 $R^2=0.885$, 均方根误差RMSE=8.321 t/hm², 预测精度P=86.4%; 3) 基于最优估测模型反演得到香格里拉市2016年高山松地上碳储量为770.953万t, 平均碳密度为40.015 t/hm²。综上, AST算法在多次交叉验证下展现出更高的稳定性与抗干扰能力, 为区域尺度森林碳储量反演提供了新的思路与技术支撑。

关键词: 碳储量, Hyperopt超参数调优, 机器学习, AST, 高山松, GAC, 遥感反演, 不确定性分析

中图分类号: S757

引用格式: 孙凯萍, 张加龙, 滕晨凯, 杨坤, 黄凯, 雷启旺, 熊登亮. XXXX. 基于GAC特征选择和AST回归算法的森林地上碳储量遥感反演. 遥感学报, XX(X): 1-24

SUN Kaiping, ZHANG Jialong, TENG Chenkai, YANG Kun, HUANG Kai, LEI QiWang, XIONG Dengliang. XXXX. Remote sensing inversion of forest aboveground carbon storage based on GAC feature selection and AST regression algorithm. National Remote Sensing Bulletin, DOI:10.11834/jrs.20265246]

1 引言

森林是陆地生态系统的重要组成部分, 在全球碳循环中发挥着至关重要的作用。森林碳储量作为衡量其碳汇功能的关键指标, 即体现了森林生态系统中的物质循环与能量传递过程, 也反映了植被与环境之间的复杂动态关系(Kauppi等, 1992; Pan等, 2011)。准确量化森林碳储量并监测其空间分布, 对于深化理解陆地碳循环机制、

评估区域碳汇潜力以及支撑碳中和政策具有重要科学意义(Huang等, 2024)。当前森林碳储量的估算主要依赖于实地调查和遥感反演两种方法(Meng等, 2024)。尽管实地测量具有较高精度, 但在大尺度区域应用中常面临成本高、效率低等问题(Li等, 2020)。相比之下, 遥感技术具备非接触、高时效性和覆盖广等优势, 已成为AGC监测的主要方法(陆灯盛等, 2025)。其中, Landsat等中等分辨率光学遥感影像因其长期连续性与公

收稿日期: 2025-08-06; 预印本: XXXX-XX-XX

基金项目: 国家自然科学基金(编号:32260390)资助, 云南省基础研究计划重点项目(编号:202501AS070047)。

第一作者简介: 孙凯萍, 研究方向为资源环境遥感。E-mail: sunkaiping0903@swfu.edu.cn

通信作者简介: 张加龙, 研究方向为林业遥感。E-mail: jialongzhang@swfu.edu.cn

开可获取性,在区域乃至全球尺度森林碳储量反演中得到广泛应用(滕晨凯等,2024)。然而,目前大多数反演模型仍受限于数据源差异和区域适应性,导致模型的泛化能力较差。因此,仍需对森林碳储量反演模型进行更深入的优化与改进。

在遥感反演建模过程中,特征选择是提高模型性能的关键步骤。中等分辨率遥感数据虽能捕捉植被结构与光谱特征,但波段冗余常导致模型复杂度增加、精度下降。已有研究表明,科学合理的特征选择方法有助于降低特征维度,提高模型预测的准确性。如王雁鹤等(2025)基于递归特征消除(RFE)与梯度提升树(GBRT)构建的AGC估测模型 R^2 达到0.84;黄天宝等(2024)利用Boruta算法筛选遥感变量并构建Stacking模型,取得 R^2 为0.73的估测结果。然而,传统特征选择方法仍易陷入局部最优,忽视变量之间的全局交互关系。遗传算法(GA)作为一种全局优化策略,具备强大的搜索能力与解空间并行探索能力,近年来在高维遥感建模中展现出良好表现(Cai等,2018),但在森林碳储量估算中的具体表现仍有待进一步验证。此外,超参数作为一种新兴工具,因其具有良好的可扩展性被广泛用于机器学习模型的细化中(Tahmouresi等,2024)。本研究拟将Hyperopt超参数优化框架引入到高山松碳储量机器学习反演模型中,在节省时间的同时获得最优预测性能。

除了变量的选择和超参数优化,建模算法也是影响森林AGC估算准确性的重要因素,目前森林AGC遥感反演模型主要分为参数模型和非参数模型。参数模型如多元线性回归,具有较强的可解释性,但难以捕捉遥感特征与森林AGC之间的非线性与多变量耦合关系(张子慧等,2022)。非参数模型如随机森林(RF)、支持向量机(SVM)、BP神经网络等,能够有效处理高维、非线性及异构数据,已在森林AGC估算中广泛应用(雷令婷等,2022)。然而,单一模型在应对噪声干扰、异质数据或边界样本时,仍易出现过拟合与泛化不足等问题。近年来,堆叠集成学习(Stacking)通过整合多个弱学习器,兼顾偏差与方差的基础上提升模型预测结果的稳健性,逐渐成为遥感大数据领域的研究热点(Li等,2024)。具体而言,由多个基模型和元模型的耦合形成“异构并行集成”结构(王可月等,2025),基模

型可捕捉数据集中难以识别的光谱、空间、纹理等多元特征模式,元模型则专注于将各个模型的输出合并以获得最终预测(Wolpert,1992)。Sun等(2024)基于袋装集成模型反演太湖总磷浓度,取得良好效果,但所用基模型为线性回归表现受限;Zhao等(2025)基于CatBoost和RFR等构建Stacking集成框架,显著提升滑坡敏感性评估精度,但所选模型结构相近,融合效果有限;Wang等(2025)将Stacking应用于土壤湿度估算,弥补了单模型的局限,但对基模型差异性缺乏系统分析。因此,现有集成模型在应用中仍存在不足:一是基学习器选择缺乏系统性,导致模型间互补性不足,削弱了集成效果;二是元学习器多依赖复杂的非线性模型,虽拟合能力强,但增加了过拟合风险和计算负担。针对这些问题,本研究提出了一种AST集成学习回归算法,在基学习器层引入多类型、结构差异显著的模型组合,通过差异性度量增强之间的互补性;采用基学习器跨折预测进行均值融合与元学习器自适应加权实现模型优化,以线性融合方式整合各基学习器输出,避免过度复杂化,使其具备较强的可解释性;此外,借助智能化Hyperopt超参数优化框架,对基学习器进行全局高效调优实现精度与效率的平衡。旨在解决单一机器学习算法在森林地上AGC估测中存在的信息挖掘不充分、稳定性差及精度偏低的问题,通过融合多种优势算法来提高模型的稳健性和准确性。

综上,以香格里拉市的高山松为研究对象,结合2016年森林资源二类调查数据与Landsat 8 OLI遥感数据,采用两种特征选择方法筛选特征子集,基于六种单一机器学习算法(AdaBoost、CatBoost、RFR、LightGBM、XGBoost、SVM)、AST集成算法和Hyperopt超参数优化来构建高山松AGC估测模型,利用最优模型进行高山松AGC空间分布及结果的不确定性制图。研究提出的新型堆叠模型及优化策略,为高山松AGC的高精度遥感估算提供创新的方法论和理论支撑,推动森林碳循环监测技术的发展。

2 研究区、材料与方法

2.1 研究区概况

研究区为云南省迪庆藏族自治州香格里拉市

(99°20′—100°19′ E, 26°52′—28°52′ N), 地处青藏高原横断山脉腹地, 位于滇、川、藏三省交界。海拔范围在 1503 m 至 5545 m 之间, 最大高差达 4042 m, 平均海拔 3459 m, 市域土地面积 11613 km²。高山松 (*Pinus Densata*) 为当地的优势树种, 能长时间地固定和存储大量的碳; 此外, 它在维持生物多样性等方面也发挥着巨大作用 (Huang 等, 2025)。研究区概况详见图 1。

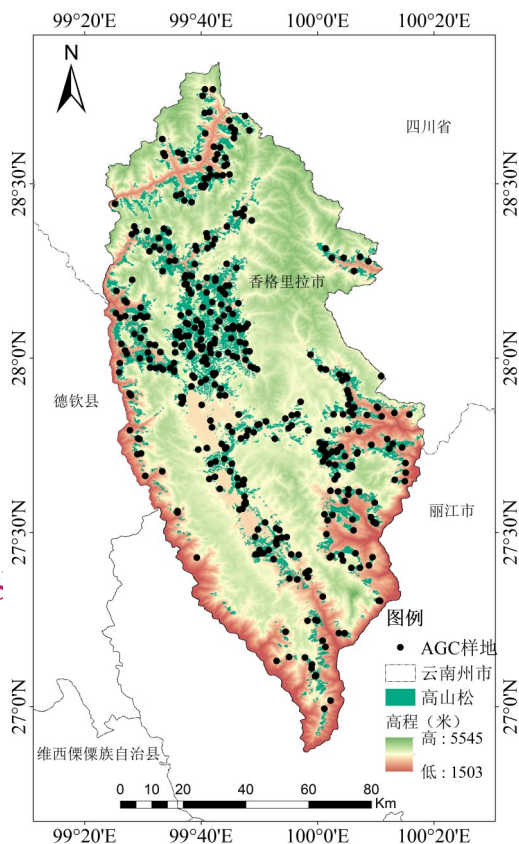


图 1 研究区概况图及样地分布图

Fig. 1 Overview map of the study area and distribution map of sample plots

该图基于审图号 GS(2024)0650 号的标准地图制作, 底图无修改。

2.2 数据来源与处理

2.2.1 样地数据

采用森林生物量膨胀系数 (forest biomass expansion factor, BEF) 方法, 按林分蓄积量—生

物量的转换公式完成每块小班生物量的计算, 进而得到高山松单位面积总生物量 (胥辉等, 2019), 具体计算公式如式 (1) 所示。参考韩雪莲等 (2023) 关于香格里拉地区高山松碳储量的研究, 采用 0.5 作为碳含量转换系数来计算单位面积的 AGC, 具体计算公式如式 (2) 所示。

样地数据为香格里拉市 2016 年森林资源二类调查小班数据, 首先筛选出平均胸径大于等于 5 cm 的小班 (卢腾飞等, 2020); 参考 Peng 等 (2025) 采用分层随机抽样的方法, 选取龄组做为分层变量, 在 95% 置信水平下, 抽样精度设定为 95%, 共选取 477 块小班样本。为进一步减少误差并提高数据准确性, 对样本进行目视解译, 剔除包含其他地类或存在明显异常的小班 (如采伐地、造林地、疏林地等) (戚玉娇等, 2015), 依据 Pauta 准则 (即 3 倍标准差原则) 去除极端异常值其范围为 (−14.46—110.87 t/hm²) (Yang 等, 2025; Liao 等, 2022; 周律等, 2020), 最终保留 398 块小班用于分析。由图 1 及表 1 可知, 所选样本小班分布得较为均匀, 且与剩余样本基本统计值相近, 具有代表性。

$$B = V \cdot SVD \cdot BEF \tag{1}$$

式中: B 为小班单位面积地上生物量 (t/hm²), V 为小班单位面积蓄积量 (m³/hm²), SVD 为基本木材密度 (0.413 t/m³), BEF 为森林生物量转换系数 (1.6509) (胥辉等, 2019)。

$$C = B \cdot Cc \tag{2}$$

式中: C 为高山松碳储量, B 为森林地上生物量, Cc 为碳转换系数。

2.2.2 遥感影像

研究使用的 3 景 2016 年 Landsat 8 OLI 遥感影像数据来自于美国地质调查局 (USGS, <http://glovis.usgs.gov/>, 引用日期: 2024 年 9 月 8 日), 影像信息如表 2 所示。采用 Xu 等 (2024) 的方法进行了辐射校正、大气校正、地形校正和镶嵌裁剪等预处理操作。

表 1 选取样本与剩余样本描述性统计

Table 1 Descriptive statistics of the selected sample and the remaining sample

总样本	样地数量/个	最小值/(t/hm ²)	最大值/(t/hm ²)	平均值/(t/hm ²)	标准差/(t/hm ²)
选取样本	398	2.16	108.48	41.54	18.42
剩余样本	7552	0.93	148.31	39.71	18.06

表2 遥感影像信息

Table 2 Remote sensing image information

影像类型	影像信息	空间分辨率(m)	获取日期	含水量/(%)
Landsat 8 OLI	LC81320412016348LGN00	30	2016-12-23	0.76
	LC81320402016348LGN01	30	2016-12-13	0.73
	LC81310412016325LGN00	30	2016-11-20	0.40

2.2.3 数字高程模型数据

数字高程模型 (DEM) 来自于地理空间数据云 (<http://www.gscloud.cn/>, 引用日期: 2024 年 9 月 5 日) ASTER GDEM 数据产品, 空间分辨率为 30 m, 该数据的水平精度和垂直精度分别为 30 m 和 20 m。

2.2.4 气象数据

气象数据来自国家青藏高原科学数据中心 (<https://data.tpdc.ac.cn/>, 引用日期: 2024 年 9 月 25 日), 采用 2016 年年平均气温、年降水量, 该数据的空间分辨率均为 1 km, 采用双线性插值法重采样至 30 m (Luo 等, 2024)。

2.3 研究方法

首先, 对遥感影像数据进行预处理; 随后, 提取遥感特征变量并采用 RFE、GAC 进行特征选择; 基于六种单一机器学习算法和 AST 回归算法

构建碳储量估测模型, 同时引入 Hyperopt 超参数优化框架, 并通过多项指标评价模型精度; 最后, 选择最优模型进行高山松地上 AGC 空间分布及不确定性制图。技术路线如图 2 所示。

2.4 遥感特征变量提取

研究参考 Yang 等 (2025) 的研究, 共提取了 256 个特征变量, 包括 251 个遥感特征因子 (41 个光谱特征、210 个纹理特征)、3 个地形因子、2 个气候因子。遥感特征因子在 ENVI 5.6 中提取, 其中, 纹理特征在 45° 方向 3×3、5×5、7×7、9×9 和 11×11 共 5 个窗口下提取, 在剔除相关性较弱的特征后, 分别保留 8 个、9 个、27 个、33 个、39 个, 共计 116 个。最终, 共选取 162 个特征变量用于森林碳储量估测模型的构建 (如表 3 所示), 包括 41 个光谱特征、116 个纹理特征、3 个地形因子和 2 个气候因子。

表3 遥感特征变量和因子信息

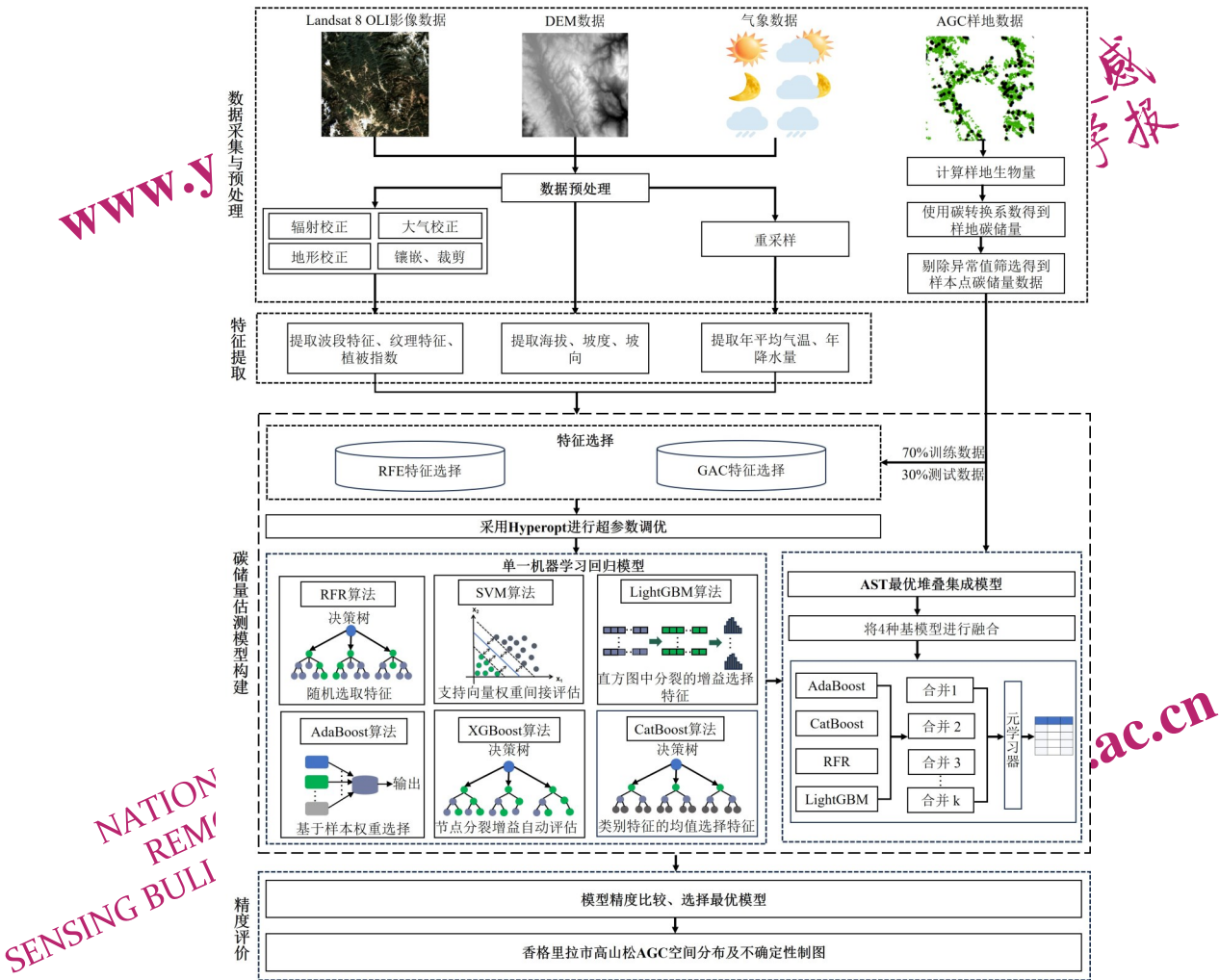


图2 技术路线图
Fig. 2 Technology roadmap

Table 3 Information on all characteristic variables and factor

特征类型	变量个数	特征名称	计算公式
原始单波段因子	6	B2-B7	—
		Albedo	$Albedo = B2 + B3 + B4 + B5 + B6 + B7$
		B24	$B24 = B2/B4$
		B53	$B53 = B5/B3$
		B65	$B65 = B6/B5$
		B73	$B73 = B7/B3$
		B74	$B74 = B7/B4$
		B76	$B76 = B7/B6$
		B245	$B245 = (B4 + B5) - (B2)$
		B345	$B345 = B3 \times B4/B5$
		B347	$B347 = B3 \times B4/B7$
		B547	$B547 = B4 \times B5/B7$
信息增强因子	3	主成分变换	—
	3	缨帽变换	—
植被指数	18	VIS_{234}	$B2 + B3 + B4$

续表

特征类型	变量个数	特征名称	计算公式
植被指数	3	大气阻抗植被指数 ARVI	$ARVI = \frac{B5 - 2B4 + B3}{B5 + 2B4 - B3}$
		绿色叶绿素指数 Clgree	$Clgree = \frac{B5}{B3} - 1$
		差值植被指数 DVI	$DVI = B5 - B4$
		增强植被指数 EVI	$EVI = \frac{2.5(B5 - B4)}{B5 + 6B4 - 7.5B2 + 1}$
		重归一植被指数 GRVI	$GRVI = \frac{B5}{B3}$
		ND43	$ND43 = \frac{B4 - B3}{B4 + B3}$
		ND67	$ND67 = \frac{B6 - B7}{B6 + B7}$
		ND563	$ND563 = \frac{B5 + B6 - B3}{B5 + B6 + B3}$
		归一化植被指数 NDVI	$NDVI = \frac{B5 - B4}{B5 + B4}$
		优化的土壤调节植被指数 OSAVI	$OSAVI = \frac{B5 - B4}{B5 + B4 + 0.16}$
		垂直植被指数 PVI	$PVI = 0.939B5 - 0.344B4 + 0.09$
		比值植被指数 RVI	$RVI = \frac{B5}{B4}$
		优化的土壤调节植被指数 SARVI	$SARVI = \frac{B5}{B4 + 0.084726/0.96916}$
		土壤调节植被指数 SAVI	$SAVI = \frac{1.25 \times (B5 - B4)}{B5 + B4 + 0.25}$
		有效叶面积指数 SLAVI	$SLAVI = \frac{B5}{B4 + B7}$
		中红外植被指数 VI3	$VI3 = \frac{B5 - B6}{B5 + B6}$
		可见光绿色植被指数 VIGreen	$VIGreen = \frac{B3 - B4}{B3 + B4}$
纹理特征	116	均值 (Mean, MEA)	$\sum_{i,j=0}^{N-1} i(P_{i,j})$
		同质性 (Homogeneity, HOM)	$\sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1 + i - j }$
		异质性 (Dissimilarity, DIS)	$\sum_{i,j=0}^{N-1} P_{i,j} i - j $
		二阶矩 (Angular Second Moment, ASM)	$\sum_{i,j=0}^{N-1} P_{i,j}^2$
		信息熵 (Entropy, ENT)	$\sum_{i,j=0}^{N-1} P_{i,j} (-\ln P_{i,j})$

续表

特征类型	变量个数	特征名称	计算公式
		相关性 (Correlation, COR)	$\left[\frac{\sum_{i,j=0}^{N-1} P_{i,j} (i - MEA_i)(j - MEA_j)}{\sqrt{VAR_i} \sqrt{VAR_j}} \right]$
		对比度 (Contrast, CON)	$\sum_{i,j=0}^{N-1} P_{i,j} (i - j)^2$
		方差 (Variance, VAR)	$\sum_{i,j=0}^{N-1} (i - MEA)^2$

注：B2, Blue, 蓝波段; B3, Green, 绿波段; B4, Red, 红波段; B5, NIR, 近红外波段; B6, SWIR1, 短波红外 1; B7, SWIR2, 短波红外 2。3 个主成分变换: PCA1, PCA2, PCA3 (PCA, Principal Component Analysis)。3 个缨帽变换: TCB, TCG, TCW (TC 代表缨帽变化, B 为亮度, G 代表绿色, W 代表湿度)。土壤调节植被指数 SAVI 引入一个土壤调节系数 L , 由于香格里拉地区植被覆盖度较高 (岳彩荣, 2012), L 取 0.25。 i, j : 灰度共生矩阵中第 i 行和 j 列的灰度值; $P_{i,j}$: 灰度值为 (i, j) 的概率; N : 影像的行列数; MEA : P 的均值; VAR : P 的方差。

2.5 特征选择方法

由于多光谱数据特征具有多重共线性, 易引发过拟合, 造成模型精度随特征数量增加而降低的情况。因此, 设置相关性阈值 (即相关性 > 0.9) 剔除高度自相关的冗余变量, 并在此基础上进行特征选择。

2.5.1 基于 GAC 算法的特征选择

遗传算法通过选择、交叉和变异、更新种群等操作来生成和优化问题的解决方案 (Samad 等, 2023)。本文尝试构建 GAC 特征选择算法操作如图 3 所示, 具体流程如下: (1) 种群初始化: 总共随机选择 60 个初始个体以形成初始种群。参考 Gordon 等 (2022) 在前期实验中对不同初始种群规模进行了对比, 结果表明在种群规模为 60 时, 算法的收敛速度与结果稳定性均表现最佳, 因此最终确定为 60。(2) 适应度评估: 使用经过 5 折交叉验证的 CatBoost 模型后的 RMSE 作为适应度评估函数, 为了提高算法运行的效率, 在 CatBoost 中使用随机子空间 (RSM) 策略和 GPU 加速。(3) 遗传操作: 选择算子为锦标赛选择 (规模=3); 交叉概率设定为 0.7; 个体以 0.2 的概率发生变异, 且在变异时每个基因以 0.2 的概率翻转。(4) 终止条件: 总迭代次数设为 80 代, 终止条件设定为达到最优特征集在 15 代内保持不变或者最大特征数为 10。当满足终止条件时, 算法停止搜索并返回当前的最优解。

2.5.2 递归特征消除

RFE 是一种包装器特征选择算法, 通过基模

型反复训练为各个特征分配不同权重逐一消除特征, 根据性能指标选择最佳特征子集 (Lin 等, 2023)。研究中使用随机森林作为基模型, 关键参数: `mtry`、`ntree` 和节点大小使用默认值, 从 1 开始, 采用逐一递增的策略增加至全部特征数, 通过 5 折交叉验证获得的最小 RMSE 来确定最优特征子集。

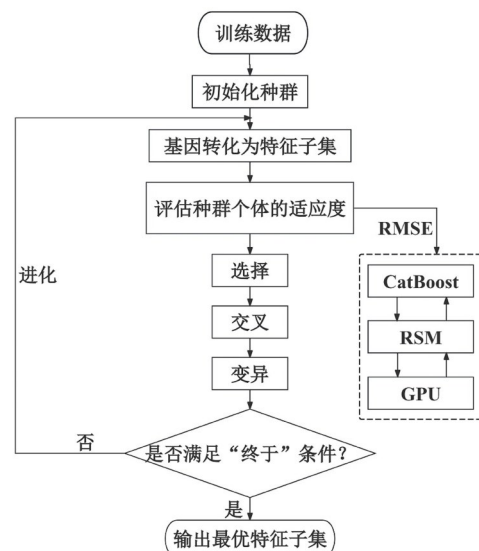


图 3 GAC 特征选择方法
Fig. 3 GAC feature selection method

2.6 超参数优化

研究运用超参数调优工具 (Hyperopt) 在建模初期进行调优, Hyperopt 使用贝叶斯优化策略, 利用 TPE (Tree-structured Parzen Estimator) 方法对超参数空间进行概率建模与连续更新 (崔慧莹, 2023), 使用 Hyperopt 提供的 `fmin` 函数来最小化目

标函数值, 并采用5折交叉验证以此确定每个模型的最优参数组合。

2.7 机器学习模型

本文采用六种机器学习回归模型反演高山松碳储量, 以上所有模型的构建基于Python 3.9语言的机器学习库“scikit-learn”实现。

2.7.1 自适应提升

AdaBoost属于提升类集成算法, 通过迭代调整样本权重以强化对难分类样本的学习, 逐步集成多个弱学习器形成最终预测结果 (Freund等, 1997), 具备抗噪能力和泛化性。

2.7.2 分类提升

CatBoost是以对称树为基础的GBDT框架擅长处理类别变量和有序特征 (Luo等, 2021), 结合独热编码等技术处理类别变量, 并通过迭代最小化损失函数以提升预测准确性与泛化能力。

2.7.3 随机森林

RFR通过集成多个决策树并采用投票或平均策略生成最终预测结果, 具有较强的鲁棒性与泛化性能 (Singh等, 2023), 在高维数据处理和过拟合抑制方面表现优越。

2.7.4 轻量级梯度提升

LightGBM作为改良版梯度提升算法, 通过基于直方图的特征分桶和高效的叶子生长策略, 在提升计算效率的同时实现高精度回归预测 (Ke等, 2017)。

2.7.5 极限梯度提升

XGBoost (eXtreme Gradient Boosting) 算法是梯度提升决策 (GBDT) 的改进。通过迭代集成学习和正则化提升计算效率与预测性能, 同时支持并行处理并能够自动处理缺失值, 广泛应用于分类、回归和排序任务 (Yao等, 2025)。

2.7.6 支持向量机

SVM (Support Vector Machine) 是一种基于统计学习理论的监督学习算法, 通过将低维非线性数据映射到高维空间进行线性回归, 从而实现原始空间中的非线性回归效果 (Yao等, 2025)。实验中采用径向基函数 (RBF) 作为核函数以刻画特

征与目标变量的复杂非线性关系。

2.8 堆叠集成学习回归模型 (AdaBoost-Stacking Tree-based, AST)

2.8.1 算法描述

AST算法的核心思想在于将多个不同的预测模型 (即“基学习器”) 组合成一个统一的模型, 通过K折交叉验证训练单一回归模型并输出预测结果, 再将这些输出作为新的特征输入至一个高层模型进行再训练。基于Zhang等 (2022) 的研究, 提出AST算法的基本假设为: 不同类型的基学习器在处理高维遥感数据时具有结构互补性, 能够从不同维度和角度捕捉地上AGC与遥感因子之间的映射关系。因此, AST算法在第一层引入四种性能互补的算法作为基学习器, 采用跨折预测策略, 通过对各折预测结果取平均来获得稳定的元特征; 第二层则采用多元线性回归作为元学习器, 对基学习器的输出进行自适应加权整合, 实现预测结果的最优融合。算法的结构和原理如图4所示。

2.8.2 算法具体步骤

步骤1: 设原始训练数据集为:

$$D = \{(x_i, y_i)\}_{i=1}^N \quad (3)$$

其中, $x_i \in R^d$ ($i=1, 2, \dots, d$) 表示第*i*个样本的遥感特征向量, R 表示实数集, R^d 表示*d*维实向量空间, 其中*d*是一个正整数表示这个向量空间的维度; $y_i \in R$ 表示第*i*个样本对应的地上AGC实测值, N 为样本总数。

步骤2: 基学习器训练与跨折预测, 本研究中共选取 $m=4$ 个不同的基学习器 $\{f_1, f_2, f_3, f_4\}$, 将训练集 D 划分为*K*个大小相等的子集, $D_1, D_2, \dots, D_K$, 对每个基学习器 f_m 循环进行*S*次训练与验证: 第*K*次, 使用 $D^{(k)} = D \setminus D_k$ 作为训练集, D_k 作为验证集; 得到每个样本 x_i 在第*m*个学习器上的预测值如式 (4)。在完成所有 $x_i \in D_k$ 折次后, 将每个样本在*S*个学习器上的预测值组合为新的元特征向量如式 (5):

$$\hat{y}_i^{(m)} = f_m^{(k)}(x_i) \quad (4)$$

$$z_i = [\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \hat{y}_i^{(3)}, \hat{y}_i^{(4)}]^T \in R^4 \quad (5)$$

式中, $\hat{y}_i^{(m)}$ 为第*m*个基学习器在交叉验证中对样本

x_i 的预测值, $f_m^{(k)}$ 为第 m 个学习器在第 k 折训练得到的模型, T 为转置符号表示列向量, z_i 为样本 i 的新特征向量由 4 个基模型的预测结果组成。

步骤 3: 在每折训练完成后, 对整个测试集进行预测, 交叉验证后对每列预测结果取平均, 融合得到 A1、A2、A3、A4 作为测试集的输入元特征。

步骤 4: 元学习器训练与自适应加权, 构建一

个多元线性回归模型 $g(z_i)$, 以 z_i 为输入特征, 拟合目标值 y_i 如式 (6):

$$\hat{y}_i = g(z_i) = w_0 + \sum_{m=1}^4 w_m \hat{y}_i^{(m)} = w_0 + w^T z_i \quad (6)$$

式中 $w_0, w_1, \dots, w_4$ 为回归系数, $\hat{y}_i$ 为最终模型对样本 i 的 AGC 预测结果, $w^T z_i$ 为矩阵乘法, 表示线性加权。

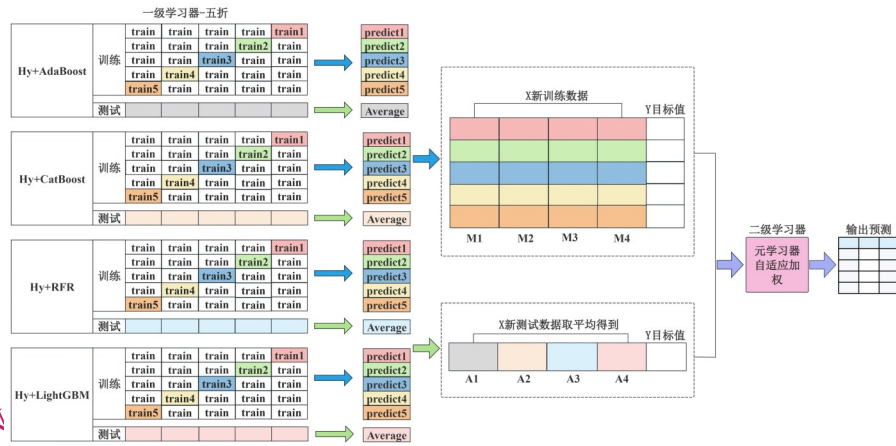


图 4 AST 堆叠集成模型结构

Fig. 4 AST stacked ensemble model structure

2.2 精度评估

本文通过随机选取 70% 的训练数据和 30% 的验证数据进行模型训练和验证, 采用决定系数 (R^2)、平均绝对误差 (MAE)、精度 (P)、相对均方根误差 (rRMSE) 和 RMSE 来评估模型在训练集上的拟合效果, 采用残差 (e_i) 来评价反演结果的不确定性。具体计算公式如下:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (8)$$

$$P = \frac{1}{n} \sum_{i=1}^n \left(1 - \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right| \right) \times 100\% \quad (9)$$

$$rRMSE = \frac{RMSE}{\bar{y}} \times 100\% \quad (10)$$

$$MAE = \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{n} \right| \quad (11)$$

$$e_i = \hat{y}_i - y_i \quad (12)$$

式中, n 是样本数量, y_i 为实测样地数据碳储量,

$\hat{y}_i$ 为模型预测值, $\bar{y}$ 为真实值均值。

3 结果与分析

3.1 特征选择结果

本研究在剔除高度自相关特征后, 剩余 62 个特征因子用于特征选择。包括 2 个光谱变量、10 个植被指数、38 个纹理特征因子、6 个波段组合和 3 个地形特征因子、温度及降水和 1 个主成分因子。如图 5 (a) 中在 RFE 特征选择时, 特征数量为 9 时 RMSE 最低, 因此选择特征数量为 9 个。与 RFE 算法相比, 如图 5 (b) 中 GAC 算法在第 15 代遗传进化后 RMSE 收敛, 从初始 62 个特征中优选出 7 个变量, 显示出更强的全局搜索能力和降维效果。

图 6 (a) 显示, 在 RFE 所选特征中, Pre 与 Temp 的贡献度最高, R11B6ENT、B347 次之, 其余特征贡献度相对接近。相关性分析如图 6 (b) 显示, Pre、ND67、R11B5ME、R9B5ME 和 PVI 与 AGC 呈极显著正相关, 而 Temp、B347、R7B6ME 与 R11B6ENT 则呈显著负相关, 其中 B347、

R11B6ENT呈极其显著负相关。此外，GAC方法结果进一步验证了气温与降水在特征选择中的主导地位，如图6(a)和(c)所示，两者的重要性排名均靠前。GAC方法的皮尔逊相关性热图，如图6(d)表明除Temp与AGC为弱负相关外，其余的特

征都与AGC呈极显著正相关，其中R11B6COR与R7B6COR表现出最强的正相关性。在特征贡献度图中如图6(c)，R7B6COR的贡献度较为突出，显著高于其余变量，而其余特征的贡献度则大致处于同一水平。

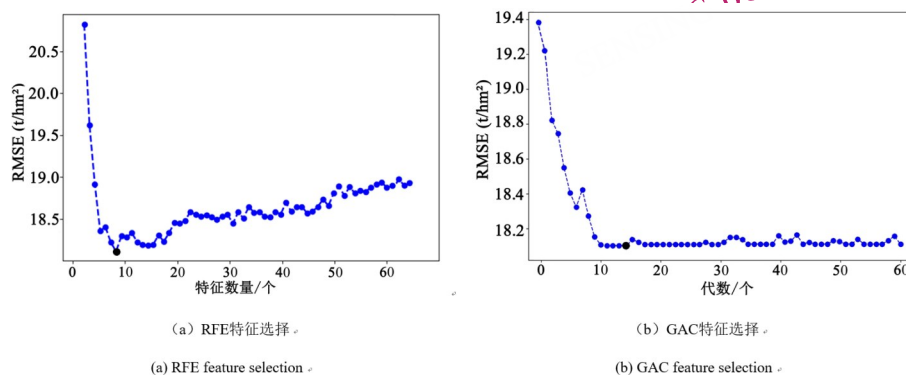
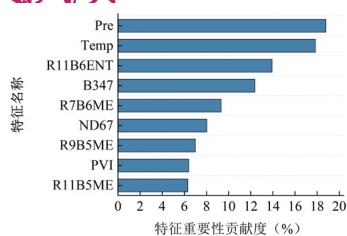


图5 两种特征选择方法的RMSE随特征数量变化趋势

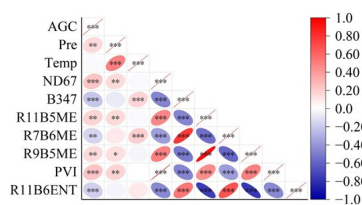
Fig. 5 Trend of RMSE of two feature selection methods with the number of features

在所选的结果中，表达式为“RXBYZZZ”。RX为窗口大小，BY为特定波段，ZZZ为特定纹理特征的缩写，降水为(Pre)，气温为(Temp)



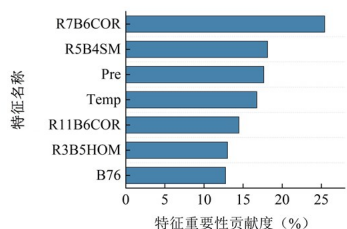
(a) RFE法特征贡献度。

(a) Characteristic contribution of the RFE method.



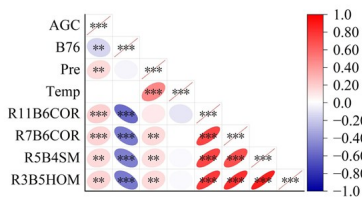
(b) RFE法特征皮尔逊相关性热图。

(b) Pearson correlation heat map of RFE features.



(c) GAC法特征贡献度。

(c) Characteristic contribution of the GAC method.



(d) GAC法特征皮尔逊相关性热图。

(d) Pearson correlation heat map of GAC features.

图6 两种特征选择方法的特征重要性排序和相关性

Fig. 6 Feature importance ranking and correlation of two feature selection methods

*表示显著相关($p < 0.05$), **表示极显著相关($p < 0.01$), ***表示极其显著相关($p < 0.001$)

3.2 超参数调优结果

在超参数调优过程中，Hyperopt调优框架通过TPE算法实现各模型关键超参数的搜索轨迹逐渐收敛，最终得到最小RMSE对应的最优参数组合。经“200”次迭代，得出不同最优特征子集与七种机

器学习组合的超参数调优结果及各模型参数及搜索范围(如表4)。

如图7所示，基于RFE方法筛选出的最优特征子集建模时，AdaBoost与LightGBM、XGBoost模型的性能最依赖于学习率(learning_rate)与最大深度(max_depth)。AdaBoost与LightGBM、XGBoost

的学习率分别收敛至 0.155 与 0.17、0.813, 最大深度稳定于 10、8.4、5, 表明学习率对模型更新的稳定性起关键调控作用, 而最大深度则有效控制了模型对特征交互的拟合复杂度。在 CatBoost 模型中, “bagging_temperature” 作为控制随机正则化强度的核心参数, 在调优过程中波动于 6 - 10 之间, 当取值为 9.312 且 learning_rate 为 0.17 时, 模型表

现出最佳稳定性。对于 RFR 模型, 最优参数配置为 n_estimators=256, depth=11, 其中 depth 对决策树泛化能力影响显著。而在 SVM 模型中核心参数 C 调节训练误差容忍度, epsilon 则影响支持向量选择与回归精度, 最优参数 epsilon 最终稳定于 0.156, C 为 10.02 时均方根误差达到最小。

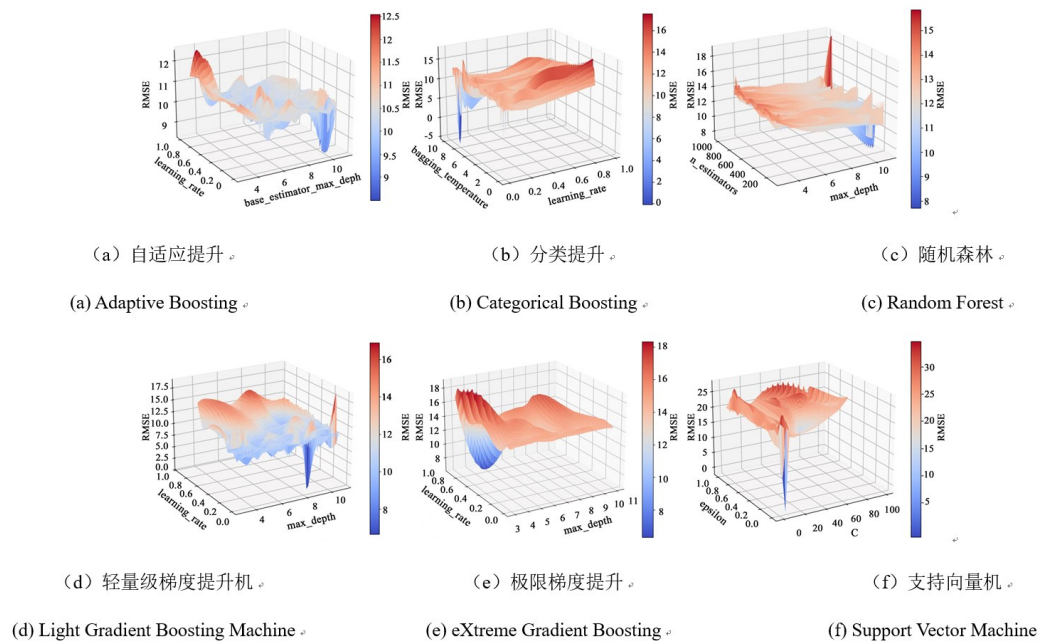


图 7 RFE-Hyperopt 各个模型超参数优化过程

Fig. 7 Hyperparameter optimization process of each model in RFE-Hyperopt

基于 GAC 方法筛选出的最优特征子集建模时如图 8 所示, RFR 模型当 n_estimators 在 800-1000 范围内波动, n_estimators 取 817 时, max_depth 为 7.5 时 RMSE 达到最小。对于 AdaBoost 模型的 learning_rate 收敛至 (0.2-0.8) 区间为 0.6, max_depth 稳定在 4-12 之间为 11.8 时均方根误差达到最小; CatBoost 模型的 bagging_temperature 在 4 - 10 之间收敛于 4.25, learning_rate 在 0.1 - 0.8 区间内最优为 0.181, 调优路径呈螺旋收缩态势;

LightGBM 模型的 learning_rate 在 (0.2 - 0.8) 区间内最终收敛至 0.614, max_depth 在 (4 - 10) 区间内稳定于 10, RMSE 分布呈现典型径向漏斗形态; XGBoost 模型的 learning_rate 在 (0 - 1) 区间内最终收敛至 0.104, max_depth 在 (4 - 10) 区间内稳定于 7, 而 SVM 模型最优参数 epsilon 在 (0-1) 区间内最终稳定于 0.892, C 为 6.49 时均方根误差达到最小。

表 4 各个模型超参数搜索范围及优化的结果

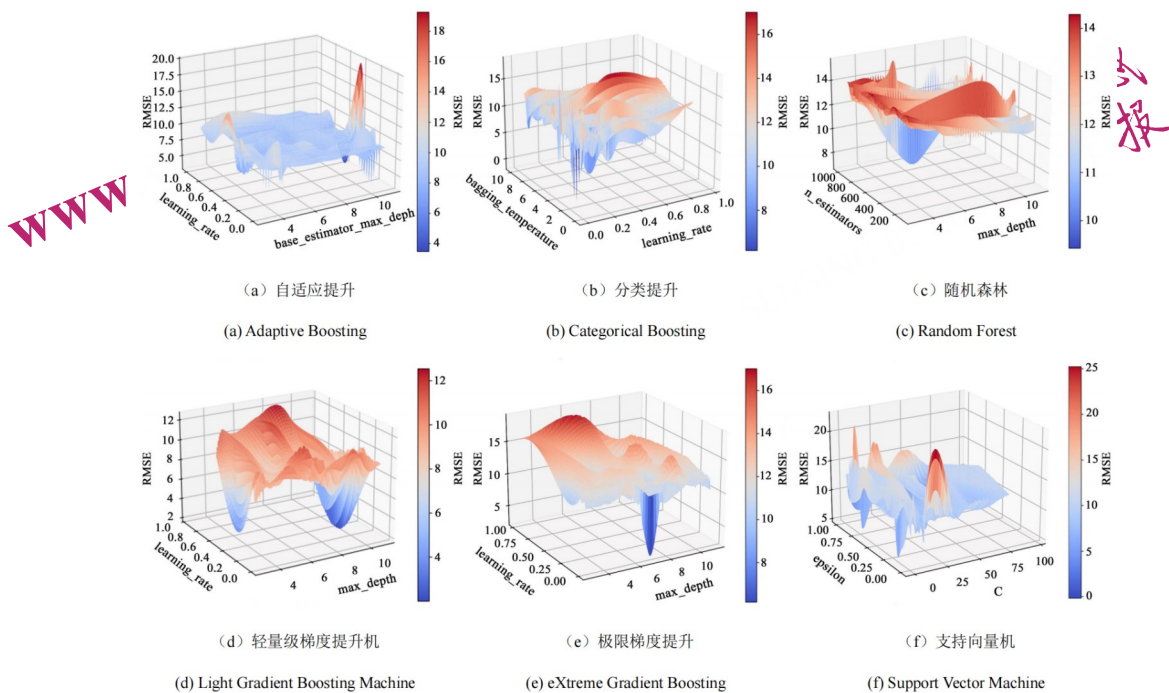


图 8 GAC-Hyperopt 各个模型超参数优化过程

Fig. 8 The hyperparameter optimization process of each GAC-Hyperopt model

Table 4 Search ranges for hyperparameters of each model and the results of optimization

模型	特征含义	超参数名称	搜索范围	RFE 最优特征子集	GAC 最优特征子集
XGBoost	每棵树的列采样比例	colsample_bytree	0.6—1	0.849	0.691
	最小分裂损失	gamma	0—5	2.779	2.766
	学习率	learning_rate	3—12	0.813	0.104
	树的最大深度	max_depth	3—10	5	7
	决策树的数量	n_estimators	100—1000	260	923
SVM	惩罚系数	C	0.1—1000	10.02	6.49
	多项式核阶数	degree	2—5	3	3
	ϵ 不敏感区间	epsilon	0.001—1	0.156	0.892
AdaBoost	每棵树的最大深度	base_estimator_max_depth	3—12	10	11.8
	学习率	learning_rate	0.0067—1	0.155	0.6
	决策树的数量	n_estimators	100—1000	313	654
CatBoost	采样温度	bagging_temperature	0—10	9.313	4.25
	树的最大深度	depth	3—12	4	4
	学习率	learning_rate	0.0067—1	0.17	0.181
	每颗树随机采样的样本比例	subsample	0.5—1	0.841	0.712
RFR	每棵树的最大深度	max_depth	3—12	11	7.5
	最大特征数	max_features	1—10	0.5	1
	叶子节点的最小样本数	min_samples_leaf	0—10	1	1
	节点再划分所需的最小样本数	min_samples_split	2—10	2	4
	决策树的数量	n_estimators	100—1000	256	817
LightGBM	学习率	learning_rate	0.0067—1	0.17	0.614
	树的最大深度	max_depth	3—12	8.4	10

续表

模型	表征含义	超参数名称	搜索范围	RFE 最优特征子集	GAC 最优特征子集
AST	决策树的数量	n_estimators	100—1000	466	905
	每棵树随机采样的样本比例	subsample	0.5—1	0.999	0.546
	L2 正则化系数	reg_lambda	0—1	4.521	0.394
	树的最大深度	cat_depth	3—12	6	7
	学习率	cat_learning_rate	0.0067—1	0.015	0.028
	每颗树随机采样的样本比例	cat_subsample	0.5—1	0.791	0.678
	采样温度	cat_bagging_temperature	0—10	8.251	5.35
	学习率	lgb_learning_rate	0.0067—1	0.016	0.015
	树的最大深度	lgb_max_depth	3—12	3	11
	决策树的数量	lgb_n_estimators	100—1000	440	377
	L2 正则化系数	lgb_reg_lambda	0—5	1.318	0.111
	每棵树随机采样的样本比例	lgb_subsample	0.5—1	0.713	0.897
	每棵树的最大深度	rf_max_depth	3—12	4	3
	节点再划分所需的最小样本数	rf_min_samples_split	2—10	3	2.3
	决策树的数量	rf_n_estimators	100—1000	554	643
	叶子节点的最小样本数	rf_min_samples_leaf	0—10	0.123	0.147
AdaBoost	最大特征数	rf_max_features	1—10	5	8
	学习率	ada_learning_rate	0.0067—1	0.131	0.0132
	决策树的数量	ada_n_estimators	100—1000	624	107
	每棵树的最大深度	ada_base_estimator__max_depth	3—12	8	5

3.3 基于不同机器学习回归算法的高山松估测模型精度对比

基于 RFE 筛选的 9 个最优特征子集建模时，四种机器学习回归模型的性能如图 9 所示，LightGBM 模型表现最佳，决定系数 R^2 达到 0.796，拟合精度优于其他模型。采用 GAC 方法筛选出的最优特征子集建模结果图 10 表明，AdaBoost 与 LightGBM、XGBoost、SVM 模型的预测性能显著提升， R^2 分别为 0.818 与 0.799、0.759、0.679。相较而言，CatBoost 与 RFR、SVM 模型在两种特征选择方法中表现略低，但在 RFE 方法下 CatBoost 的 R^2 提升 0.024，在 GAC 方法下 RFR 的 R^2 提升 0.074，SVM 的 R^2 提升 0.061，其中，SVM 模型在所有模型中的预测性能最低（如表 5 所示）。表明合适的特征选择对模型性能具有关键调控作用，可以减少预测偏差提高模型性能。

进一步比较六种单一模型与 AST 堆叠模型的估测结果如图 9、10 所示，使用 RFE 特征子集建模时 AST 模型 R^2 为 0.864，GAC 特征子集建模时 R^2 更高提升至 0.885，斜率更趋近于 1，拟合更优，

RMSE 为 8.321 t/hm²，rRMSE 为 17.672%，MAE 为 6.017 t/hm²，拟合优度 P 达 86.4%，结果表明，通过堆叠算法整合异质学习器的 AST 策略，具有更高的准确性和有效性。显著增强了模型对 AGC 非线性分布的响应能力与预测一致性，验证了集成学习与高质量特征选择在 AGC 反演中的协同增益效应。

3.4 高山松地上碳储量估测结果并与已有研究成果的对比

根据上述建模结果，采用 GAC 方法选择的遥感特征变量和 AST 堆叠回归算法来建立碳储量估测模型，并对香格里拉市高山松地上碳储量进行制图分析如图 11。结果显示，该区域 AGC 总量约为 770.953 万 t，碳密度为 40.015 t/hm²。空间分布格局表现为“西北高、东南低”，高值区集中于西北、中部及南部山地，低值区则主要位于东南部横断山区和三江并流流域，受水系密布和林地破碎化影响，碳密度相对偏低。

通过将已有文献中高山松生物量估算结果乘以碳转换系数并与研究对比发现（表 6），不同研

研究者 AGC 估算存在一定差异。首先，样地数据的差异可能导致 AGC 估算偏差。Xu 等（2024）虽与本研究使用相同的样地数据，但在样地选取过程中，Xu 等（2024）基于泊松圆盘随机抽取方法，而本研究则基于分层抽样方法。研究分层抽样样本均值（41.52 t/hm²）与总体均值（39.71 t/hm²）相对偏差为 4.6%，Xu 等（2024）泊松圆盘抽样相对偏差为 5.86%，分层抽样方法在碳储量估算中表现出更小的偏差，特别是在代表性较强的地类中。两种方法估算的碳储量分别为 770.953 万 t、787.279 万 t，差异为 16.33 万 t。分层抽样则根据不同地类的碳密度差异进行采样，能够更好地代表高碳密度区域，而泊松圆盘抽样方法是通过设置最小间距约束，使样点在空间上更加均匀分布，但这种均匀性会弱化高异质区域（如高碳密度林分）的样本权重，从而在碳密度反演中表现为轻微高估。因此不同抽样策略会导致各地类代表性差异，进而导致碳储量估算结果的偏差。此外，孙雪莲（2016）和 Huang 等（2025）则使用了不同来源的样地数据，样地条件与样本分布情况的差

异，会进一步造成估算结果在均值和离散度上的偏差。其次，遥感数据源差异也是影响 AGC 估算的重要因素。研究虽与 Xu 等（2024）、Huang 等（2025）、孙雪莲等（2016）均使用了 Landsat 8 OLI 光学影像，但孙雪莲等（2016）反演数据年份与本研究存在差异；而 Xu 等（2023）则采用 GEDI 激光雷达数据，且反演年份亦不相同。由于 GEDI 密度有限、空间采样不连续，且受轨道分布限制，区域尺度外推时可能引入系统性误差。最后，建模方法差异亦会造成估算结果差异。研究采用 AST 堆叠模型，而其他研究均使用 RFR 方法，虽然 RFR 方法被广泛用于森林碳储量的估测建模中，但因其存在“低值高估，高值低估”现象（Pang 等，2022），同样会导致已有研究结果与本研究的差异。这种偏差的根本原因在于训练数据中极端样本数量有限，使模型难以充分刻画碳储量全范围的变异性。同时，遥感特征因子与 AGC 之间的非线性关系复杂，也可能在一定程度上加剧这种偏差。

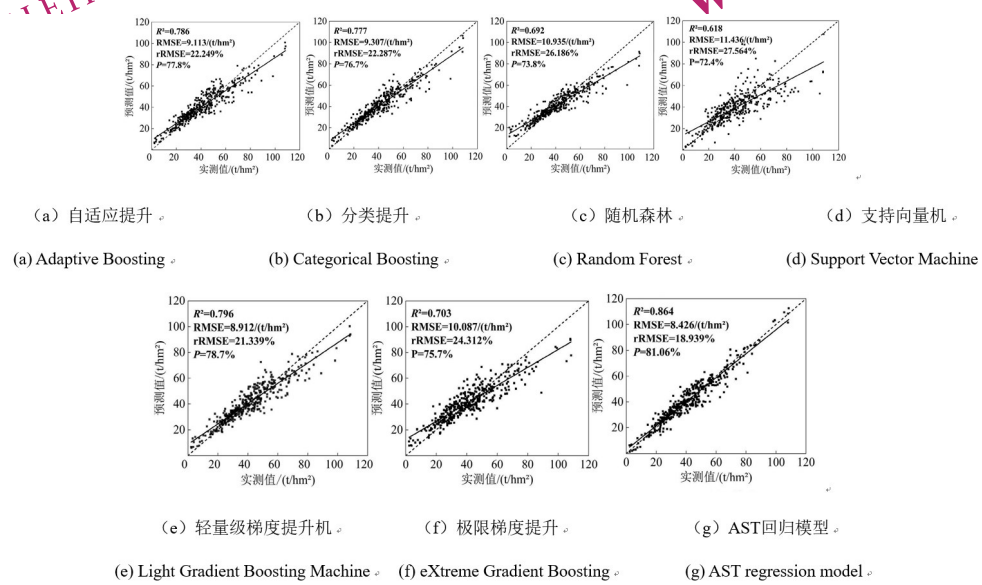


图9 基于RFE特征选择方法各模型的拟合精度散点图
Fig. 9 Scatter plot of the fitting accuracy of each model based on the RFE feature selection method

综上，尽管与不同研究者的估算结果存在一定差异，但均针对同一区域同一树种，差异并不显著；森林生态系统本身具有较高的稳定性，碳储量在短时间内的波动通常较为有限。此外，利用2016年香格里拉市森林资源二类调查数据计算得到高山松地上碳储量约为745万t，与本研究估

测结果相差约25万t，差异较小，因此本研究结果具有较好的可信度。

3.5 高山松地上碳储量反演结果的不确定性

为进一步评估 AST 模型对高山松碳储量估算的准确性，研究引入平均碳密度残差作为不确定性度量指标，通过比较预测与实测平均碳密度计

算残差,并在空间尺度上进行量化分析,得到如图 12 (a) 所示反演结果的不确定性分布。结果显示:研究区域 64.5% 的样地残差分布在 $0 \sim 2 \text{ t/hm}^2$ 的范围内, 28.9% 在 $2 \sim 10 \text{ t/hm}^2$ 之间, 6.6% 超过 10 t/hm^2 。Kolmogorov-Smirnov (KS) 检验不依赖数据分布的先验假设,能够在大量样本条件下有效评估残差与正态分布的一致性 (Massey 等, 1951)。通过“Kolmogorov-Smirnov”方法检验残差正态性

($P=0.5248 > 0.05$), 计算得到残差的偏度为 0.0417 , 峰度为 -0.2726 , 表明残差分布接近正态分布如图 12 (b)。因此,本研究中的模型误差可视为随机误差,可忽略系统性偏差。该区域大部分范围内空间不确定性较低,整体预测结果较为稳定,表明 AST 模型具有良好的稳健性,可为不同空间位置提供可靠的 AGC 估算结果。

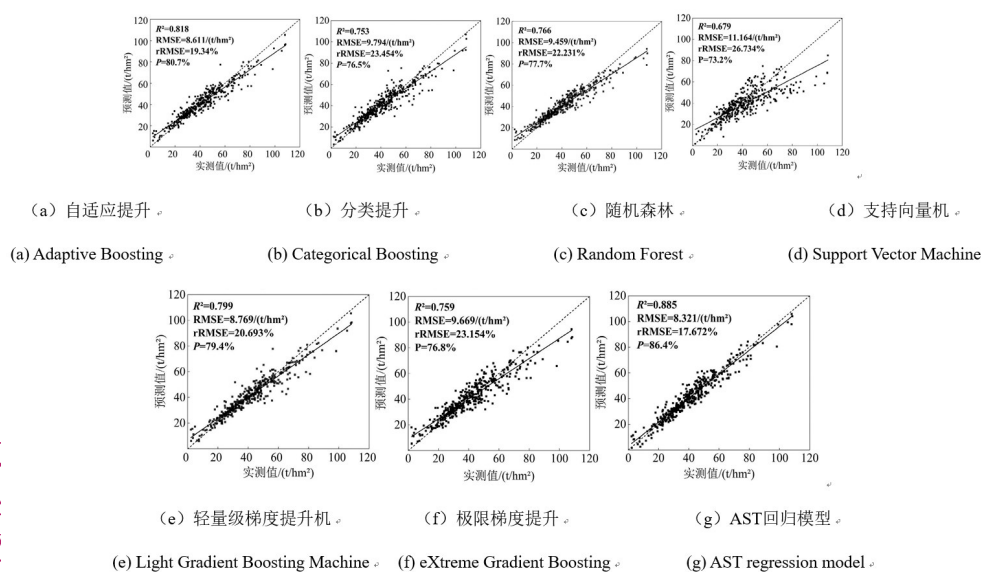


图 10 基于 GAC 特征选择方法各模型的拟合精度散点图

Fig. 10 Scatter plot of the fitting accuracy of each model based on the GAC feature selection method

表5 RFE/GA-Hyperopt 优化算法中各模型性能评估的结果

Table 5 Results of model performance evaluation in RFE/GA-Hyperopt optimization algorithms

特征选择方法	模型	R^2	RMSE/ (t/hm ²)	rRMSE/%	MAE/ (t/hm ²)	P/%
RFE	AdaBoost	0.786	9.113	22.249	7.224	77.8
	CatBoost	0.777	9.307	22.287	7.283	76.7
	RFR	0.692	10.935	26.186	8.332	73.8
	LightGBM	0.796	8.912	21.339	6.964	78.7
	XGBoost	0.703	10.087	24.312	7.936	75.7
	SVM	0.618	11.436	27.564	8.609	72.4
	AST	0.864	8.426	18.939	6.392	81.06
GAC	AdaBoost	0.818	8.611	19.34	6.686	80.7
	CatBoost	0.753	9.794	23.454	7.669	76.5
	RFR	0.766	9.459	22.231	7.742	77.7
	LightGBM	0.799	8.769	20.693	6.792	79.4
	XGBoost	0.759	9.669	23.154	7.532	76.8
	SVM	0.679	11.164	26.734	8.443	73.2
	AST	0.885	8.321	17.672	6.017	86.4

NATIONAL
REMOTE
SENSING BULLETIN | 遥感学报

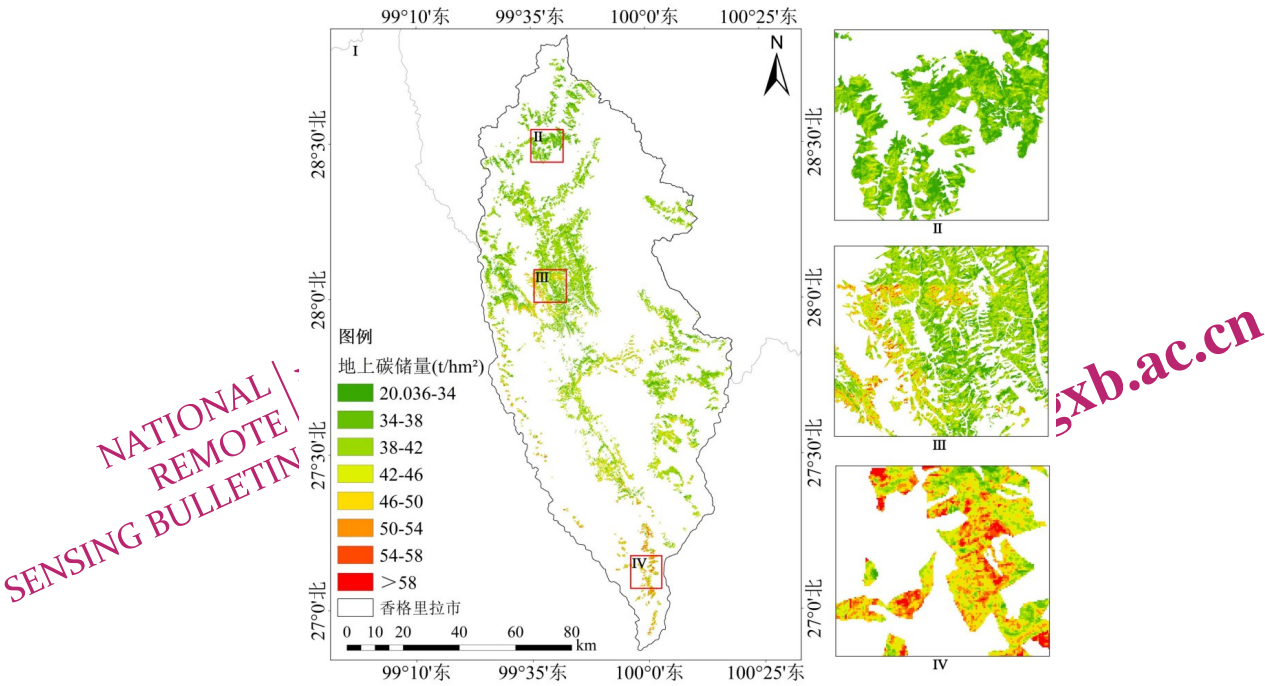
www.ygxb.ac.cn

www.ygxb.ac.cn

NATIONAL
REMOTE
SENSING BULLETIN | 遥感学报

表 6 不同研究对于香格里拉地区高山松碳储量估测结果
Table 6 Estimation of carbon storage in the Shangri-La region by different studies

来源	反演数据年份	样地数据	遥感数据源	估测模型	生物量/百万 t	碳储量(含转换后)/百万 t
Xu 等(2024)	2016	森林资源二类调查数据	Landsat 8 OLI	RFR	—	787.279
Xu 等(2023)	2019	森林资源连续清查数据	GED1	RFR	1260	631.260
孙雪莲等(2016)	2014	实测样地数据	Landsat 8 OLI	RFR	1171.961	587.152
Huang 等(2025)	2017	森林资源连续清查数据	Landsat 时间序列	RFR	—	704.000
本研究	2016	森林资源二类调查数据	Landsat 8 OLI	AST	—	770.953



该图基于审图号 GS(2024)0650 号的标准地图制作，底图无修改。

I 全区域高山松 AGC II 高山松 AGC 较低区域
III 高山松 AGC 中部区域 IV 高山松 AGC 较高区域

图 11 香格里拉市高山松碳储量空间分布图

Fig.11 Spatial distribution map of carbon storage of *Pinus densata* in Shangri-La City

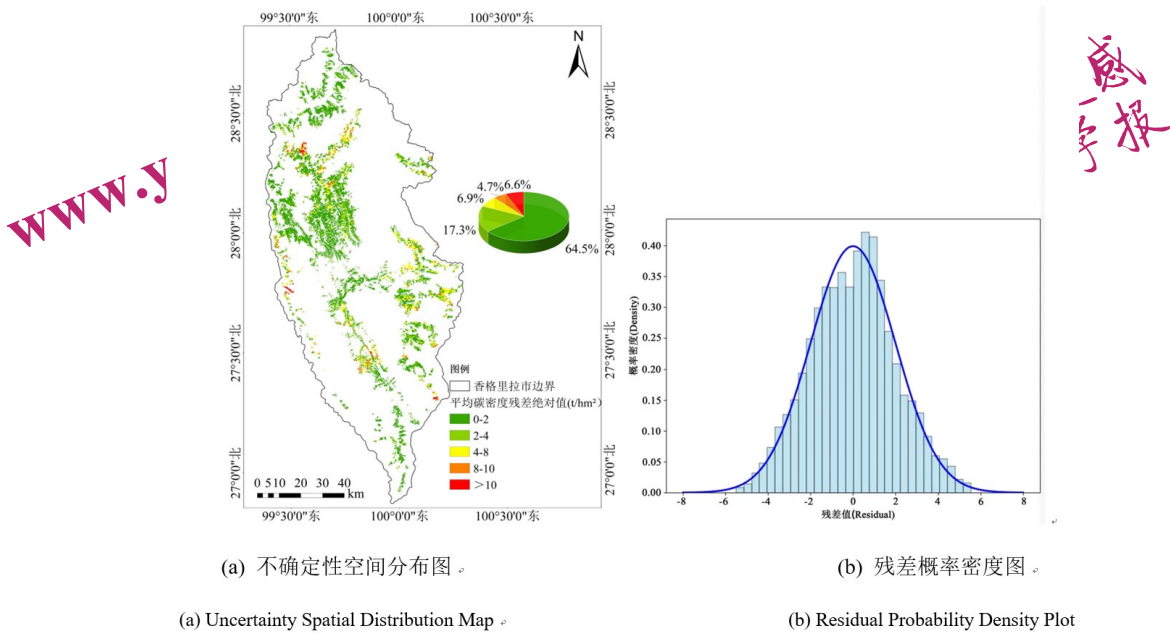


图 12 高山松碳储量反演结果不确定性分布及残差概率密度图

Figure 12 Uncertainty Distribution and Residual Probability Density Map of Carbon Stock Inversion Results for *Pinus densata*

4 讨论

4.1 特征选择的分析

在森林碳储量估测模型中，特征选择是影响模型精度与稳定性的关键因素。研究提出一种 GAC 特征选择方法，将遗传算法与 CatBoost 相结合，旨在提升模型性能的同时降低输入变量的维度。与 RFE 相比，GAC 能够选取更少且更有效的特征，减少了数据冗余，同时又保持了较高的计算效率和反演精度，有效规避了变量选择中不相关特征带来的干扰。此外，GAC 方法与 AdaBoost、LightGBM、RF、XGBoost、SVM 及 AST 等模型结合时，均表现出显著的拟合性能，验证了该方法在不同算法框架下的通用性与稳健性；而 CatBoost 与 RFE 特征子集结合时表现更优，进一步表明合理的特征子集对于提升模型的泛化能力与预测精度具有重要作用。当特征空间较大或样本量增加的情况下，模型训练时间可能会相对较长，且该方法对参数设置较为敏感，若缺乏合理的参数调优，可能会导致算法收敛速度减慢，上述问题有待后续进一步研究以此来优化计算效率与参数调控策略。Hyperopt 超参数调优的结果表明，该方法有效通过不断尝试不同参数组合来寻最优地模型参数配置，在降低计算成本时获得了最优的模型性能，与之前的研究一致（崔慧莹，2023）。不同

于以往仅针对单一模型的调优策略。研究还对堆叠模型进行复杂的超参数调优，通过合理限制最大迭代次数，在保证搜索效率的同时使所获得的超参数能够接近全局最优，避免过拟合与冗余计算。在变量选择结果方面，两种特征选择方法都选择了平均气温和平均降水量，说明降水和温度是影响高山松碳储量的主要因素，这与 Luo 等（2024）的研究成果相吻合。在未来的研究中，可进一步引入土壤数据、年均蒸发量等环境因子，采用其它超参数调优方法以提升模型性能。

4.2 估测模型的不确定性分析

本研究中，经超参数优化的 LightGBM、AdaBoost 作为单个学习器，其预测效果优于 RFR、CatBoost。具体而言，LightGBM 凭借直方图算法增强了对复杂生态因子的识别能力，但对异常值较为敏感；AdaBoost 通过样本加权机制提升了对异质区域的拟合稳定性，但易受噪声影响导致过拟合；RFR 与 CatBoost 相比，预测性能存在略微差异，但 CatBoost 在处理高维非线性关系时表现更稳定。此外引入流行的 XGBoost、SVM 模型进行对比分析，在 GAC 方法中 XGBoost 预测效果优于 SVM、CatBoost，RFE 方法中则优于 RFR、SVM，表明 XGBoost 依托梯度提升策略在平衡偏差与方差方面表现良好，能够较为稳定地刻画碳储量空间分布特征；而 SVM 受核函数类型及参数敏感性影响，

其对复杂生态变量间非线性关系的拟合能力相对不足, 预测效果最差。

总体而言, AST集成模型在所有AGC反演模型中表现出较高的预测精度, 该算法能够有效整合异构学习器的优势, 在降低随机偏差的同时提升了非参数模型的估测精度与可靠性。为进一步量化模型不确定性, 研究将不同方法建模后得到的高山松地上碳储量进行了比较(表7)。结果表明, AST模型的估测结果最接近真实值, 尤其与GAC特征子集结合时, 方差与标准差均为最小, 显示出更低的离散程度, 说明模型具有较高的稳定性且不易过拟合。尽管如此, AST模型仍存在一定偏差, 如图12(a)所示, 约6.6%的样地残差值超过10 t/hm², 其不确定性可能来源于多个方面: 首先, 基学习器的选取直接影响模型结果的

稳定性和精度; 其次, 高山松碳储量的空间分布特征及局地空间自相关性会对模型性能与结果产生影响(黄雅君等, 2023); 此外, 特征选择的稳定性与光学遥感数据的局限性; 特征选择在不同训练集或折次中波动, 使模型响应不一致。同时光学遥感数据易受天气条件与云层遮挡干扰, 造成植被信息缺失或失真, 此外集成学习在融合多模型输出时倾向于平滑异质区域的差异, 致使在森林结构变异显著的区域表现不佳(Reichstein等, 2019)。未来可引入空间自相关结构, 以此表征碳储量的空间非平稳性; 融合多源遥感数据以弥补光学遥感在植被结构感知方面的不足, 构建多特征协同的估算模型; 通过增强样本代表性、提高数据质量, 从而进一步提升模型的泛化能力与估测精度。

表7 RFE/GAC特征选择在不同模型对高山松地上碳储量的估测结果对比

Table 7 Comparison of RFE/GAC feature selection results for estimating carbon storage above ground in mountain pine using different models

特征选择方法	模型	碳储量值/(t/hm ²)	方差	标准差
RFE	高山松地上碳储量(真实值)	2.16—108.48	339	18.42
	AdaBoost	10.97—93.21	369.06	19.21
	CatBoost	3.45—106.44	370.94	19.26
	RFR	11.79—91.32	400.64	20.0
	LightGBM	5.95—100.29	363.54	19.07
	XGBoost	7.64—100.42	388.25	19.77
	SVM	3.69—107.69	432.64	23.92
	AST	1.39—112.65	349.68	18.69
GAC	AdaBoost	5.84—105.36	356.66	18.89
	CatBoost	3.71—107.02	380.18	19.50
	RFR	8.42—94.13	376.25	19.40
	LightGBM	5.99—105.46	360.72	18.99
	XGBoost	5.35—104.51	376.58	19.48
	SVM	2.9—84.91	413.73	20.91
	AST	1.54—105.74	345.65	18.59

4.3 纹理特征提取方向参数对AST估测模型的影响分析

相较于光谱特征, 纹理特征的方向参数能够显著影响森林冠层空间异质性的捕捉能力(Zhou等, 2017)。因此, 基于本文提取45°方向纹理的基础上, 提取其余三个方向0°、90°、135°的纹理特征并计算四个方向的均值, 再进一步构建单一方向(0°、90°、135°、均值)的纹理特征组合,

系统评估不同方向纹理对AST模型碳储量估测性能的影响。根据文章中方法采用相关性分析与共线性诊断对所有特征进行冗余剔除, 并在此基础上使用RFE与GAC对均值纹理及各单一方向纹理分别获取其最优特征组合。结果如表8所示, 不同方向纹理特征组合在AST模型中的建模精度呈显著差异。45°方向纹理的组合依然表现最佳, 其R²和RMSE均优于其他方向组合; 0°、90°及135°方

向纹理组合精度明显低于45°方向，而四个方向均值纹理组合的性能也未超过45°单方向。这一现象表明，纹理方向与森林冠层空间结构的一致性是影响模型精度的关键因素。从机理上看，研究区高山松林冠在西南季风与坡向的共同作用下呈斜向排列（杜燕等，2022），使冠层边界、枝叶纹理及林分聚集度在45°方向上最为显著。因此，45°方向纹理更能捕捉与碳储量紧密相关的结构信息，从而提升模型的解释力与预测精度。这一结果与廖易等（2023）、Yang等（2025）的研究一致，他们验证了默认45°方向及一个像元间隔所获得的纹理在森林参数建模中表现更佳。相反，0°、90°、

135°方向纹理所反映的空间模式与冠层实际结构不一致，其纹理对关键冠层结构的敏感性较弱，导致其建模贡献有限。四个方向均值纹理虽能整合多方向信息，但同时会稀释优势方向的结构信号并削弱纹理对影像空间结构模式的表征能力（Zhang等，2017）。此外，不同方向遥感、纹理特征组合的特征数量对AST模型性能显著影响（表8），尽管RFE与GAC已对变量进行了筛选，但冗余或弱相关性特征的增加仍会削弱关键方向纹理的信息贡献，从而降低特征集整体信噪比进而影响AST模型的预测精度。

表8 AST模型基于RFE/GAC特征选择的不同方向纹理特征组合建模精度对比

Table 8 Comparison of modeling accuracy of the AST model for different directional texture feature combinations selected by RFE/GAC feature selection methods

特征选择方法	特征变量组合	R ²	RMSE/ (t/hm ²)	MAE/ (t/hm ²)	rRMSE/%
RFE	0° SLAVI、B53、R11B7CON、R9B7CON、R5B3DIS、R7B7CON、R5B6ENT、R7B6CON、R7B7ME、R7B3HOM、B2	0.763	9.912	7.815	22.932
	45° Pre、Temp、B347、DN67、PVI、R11B6ENT、R7B6ME、R9B5ME、R11B5ME	0.864	8.426	6.392	18.939
	90° MSR、R11B3SEM、R7B6ENT、B345、R11B7ENT、Pre	0.819	8.881	7.323	20.957
	135° R11B4HOM、NDVI、R9B4HOM、MSR、R7B4HOM、SARVI、RVI、R11B7HOM、R11B3ME、Temp	0.752	10.045	7.882	23.703
	均值 B24、B74、ND67、Temp、DEM、2016_slope、R11B4HOM、R3B7HOM、R11B4ENT、R3B6ME、R11B3HOM、AR11B4SEM	0.838	8.762	6.841	19.552
GAC	0° R7B4HOM、R11B4HOM、R9B4HOM、Pre、R9B4SEM、R7B6ME	0.774	8.294	7.207	21.104
	45° Pre、Temp、B76、R7B6COR、R5B4SM、R11B6COR、R3B5HOM	0.885	8.321	6.017	17.672
	90° R11B4HOM、NDVI、R9B4HOM、MSR、R11B3HOM、SARVI、RVI、R11B4SEM、SLAVI、R11B7HOM	0.769	8.943	7.490	22.390
	135° SLAVI、Pre、B547、B345、R5B7HOM、R11B3SEM、R7B6ENT、R11B6DIS、	0.849	8.619	6.547	19.277
	均值 NDVI、RDVI、ND67、Temp、R11B2ME、R9B2ME、R11B4HOM、R7B2ME、R5B4ME	0.854	8.452	6.32	18.983

5 结论

本研究基于2016年森林资源二类调查数据和Landsat 8 OLI影像，提出一种GAC特征选择方法与RFE进行特征优选，基于六种单一机器学习算法及AST堆叠集成模型构建了森林碳储量估测模型，并结合Hyperopt超参数框架进行调优，对香格里拉市高山松碳储量进行估测。得到以下结论：（1）GAC特征选择方法能有效提升碳储量估测精度。与RFE方法相比，除CatBoost算法之外，各机器学习算法与GAC筛选的特征子集相结合构建的碳储量估测模型效果更优。（2）Hyperopt超参数

优化结合机器学习应用于森林碳储量估算中，提高了模型性能和计算效率。不同特征子集对应的最优超参数存在差异，其中AdaBoost和LightGBM回归模型对fmin函数选择与学习率调优较为敏感。（3）对比了单一机器学习回归模型，堆叠集成学习回归模型在森林碳储量估算中具有明显的优势。基于GAC选择的特征子集所构建的AST集成学习模型在高山松碳储量估测模型中精度最高（R²为0.885，RMSE为8.321 t/hm²），为森林碳储量的最佳优化反演模型。（4）纹理特征的方向性显著影响AST模型森林碳储量的估测精度，在0°、45°、90°方向、均值中，45°对角线方向被进一步证实为

碳储量估算的最佳方向。

参考文献(References)

- Cai J, Luo J, Wang S L and Yang S. 2018. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300: 70-79 [DOI: 10.1016/j.neucom.2017.11.077]
- Cui H Y. 2023. Improvement of extreme random forest algorithm and study on second-hand housing price prediction. Liaoning: Dalian Jiaotong University). (崔慧莹. 2023. 基于极端随机森林算法的改进与二手房价预测的研究. 辽宁: 大连交通大学)
- Du Y and Bao W K. 2022 Research progress on nature geographic distribution, growth law and ecological adaptability of *Pinus densata*. *Journal of Sichuan Forestry Science and Technology*, 43(6): 1-10 (杜燕, 包维楷. 2022. 高山松自然地理分布、生长与生态适应性研究进展. 四川林业科技, 43(6): 1-10 [DOI: 10.1217/202208220001])
- Freund Y and Schapire R E. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1): 119 - 139 [DOI: 10.1006/jcss.1997.1504]
- Gordon N, Kambhampati C and Mabeed A. 2022. Addressing Optimisation Challenges for Datasets with Many Variables, Using Genetic Algorithms to Implement Feature Selection. *AI Computer Science and Robotics Technology*, 1: 1 - 21 [DOI: 10.5772/acrt.01]
- Huang k, Teng C K, Zhang J L, Bao R, Liao Y, He Y R, Qiu B and Xu W R. 2025. A New Spatiotemporal Filtering Method to Reconstruct Landsat Time-Series for Improving Estimation Accuracy of Forest Aboveground Carbon Stock. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18: 6503 - 19 [DOI: 10.1109/JSTARS.2025.3539395]
- Huang L, Huang Z H, Zhou W L, Wu S M, Li X J, Mao F J, Song M X, Zhao Y Y, Lv L J, Yu J C and Du H Q. 2024. Landsat-based spatiotemporal estimation of subtropical forest aboveground carbon storage using machine learning algorithms with hyperparameter tuning. *Frontiers in Plant Science*, 15. [DOI: 10.3389/fpls.2024.1421567]
- Huang T B, Ou G L, Wu Y, Xu X W, Wang Z H, Lin R X and Xu C. 2024. Multi-source remote sensing estimation of forest biomass based on machine learning algorithms. *Journal of Northwest Forestry University*, 39(01): 10 - 18 (黄天宝, 欧光龙, 吴勇, 徐熊伟, 王振会, 蔺如喜, 徐灿. 2024. 基于机器学习算法的森林生物量多源遥感估测. 西北林学院学报, 39(01): 10 - 18 [DOI: 10.3969/j.issn.1001-7461.2024.01.02])
- Huang Y J, Zhou W and Ma M G. 2023. Characterize the spatial heterogeneity of fragmented land-surface and its relationship with spatial scale in southwest China. *National Remote Sensing Bulletin*, 27(3): 802-809 (黄雅君, 周伟, 马明国. 2023. 西南地区破碎地表空间异质性刻画与空间尺度关系初探. 遥感学报, 27(3): 802-809 [DOI: 10.11834/jrs.20232134])
- Han X L, Zhang J L, Liu L, Liao Y, Tang J H and Han D Y. 2023. Sampling estimation of *Pinus densata* carbon storage based on remote sensing feature variables. *Journal of Southwest Forestry University*, 43(6): 117-124 (韩雪莲, 张加龙, 刘灵, 廖易, 唐金灏, 韩东阳. 2023. 基于遥感特征变量的高山松碳储量抽样估算. 西南林业大学学报(自然科学), 43(6): 117-124 [DOI: 10.11929/j.swfu.202208047])
- Kauppi P E, Mielikäinen K and Kuusela K. 1992. Biomass and Carbon Budget of European Forests, 1971 to 1990. *Science*, 256(5053): 70-74. [DOI: 10.1126/science.256.5053.70]
- Ke G L, Meng Q, Finley T, Wang T F, Chen W, Ma W D, Ye Q W and Liu T-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30: 3149 - 3157 [DOI: 10.5555/3294996.3295074]
- Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N and Prabhat. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743): 195 - 204 [DOI: 10.1038/s41586-019-0912-1]
- Lei L T, Gao J P, Zhang X L and Gao X L. 2022. Forest volume estimation based on Sentinel-1 and Sentinel-2A data. *Journal of Yunnan University (Natural Science Edition)*, 44(06): 1174 - 82 (雷令婷, 高金萍, 张晓丽, 高显连. 2022. 基于 Sentinel-1 和 Sentinel-2A 数据的森林蓄积量估算. 云南大学学报(自然科学版), 44(06): 1174 - 82 [DOI: 10.7540/j.ynu.20210612])
- Li X H, Chen J J, Chen Z Z, Lan Y P, Ling M, Huang Q Y, Li H C, Han X W and Yi S H. 2024. Explainable machine learning-based fractional vegetation cover inversion and performance optimization - A case study of an alpine grassland on the Qinghai-Tibet Plateau. *Ecological Informatics*, 82: 102768 [DOI: 10.1016/j.ecoinf.2024.102768]
- Li Y C, Li M Y, Li C and Liu Z Z. 2020. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. *Scientific Reports*, 10(1): 9952 [DOI: 10.1038/s41598-020-67024-3]
- Liao Y, Zhang J L, Bao R and Xu D F. 2023. A model for estimating the dynamic change of aboveground biomass in Alpine pine based on Landsat was constructed. *Journal of Southwest Forestry University (Natural Science)*, 43(01): 117 - 25 (廖易, 张加龙, 鲍瑞, 许冬凡. 2023. 基于 Landsat 的高山松地上生物量动态变化估测模型构建. 西南林业大学学报(自然科学), 43(01): 117 - 25 [DOI: 10.11929/j.swfu.202111058])
- Lin X Q, Ren C, Li Y, Yue W T, Liang J Y and Yin X C. 2023. Eucalyptus Plantation Area Extraction Based on SLPSO-BEE Feature Selection and Multi-Temporal Sentinel-1/2 Data. *Forests*, 14(9): 1864 [DOI: 10.3390/f14091864]
- Lu D S, Jiang X D, Li Y H, Wang R Q and Li G Y. 2025. Forest Biomass Estimation with LiDAR Data. *National Remote Sensing Bulletin*, 29(0): 1 - 30 (陆灯盛, 蒋先蝶, 李云鹤, 王若琦, 李桂英. 2025. 基于激光雷达数据的森林生物量估测研究. 遥感学报, 29(0): 1 - 30 [DOI: 10.11834/jrs.20255022])
- Lu T F, Xu H and Ou G L. 2020. Remote sensing estimation of aboveground biomass in Yunnan pine forest in Qujing City based on mixed-effects model. *Journal of Southwest Forestry University (Natural Science)*, 40(01): 104 - 15 (卢腾飞, 胥辉, 欧光龙.

2020. 基于混合效应模型的曲靖市云南松林地上生物量遥感估测. 西南林业大学学报(自然科学), 40(01): 104 - 15 [DOI:10.11929/j.swfu.201907044]
- Luo K, Feng Y F, Liao Y, Zhang J L, Qiu B, Yang K, Teng C K and Yin T Y. 2024. Developing a Method to Estimate Above-Ground Carbon Stock of Forest Tree Species *Pinus densata* Using Remote Sensing and Climatic Data. *Forests*, 15(11): 2023 [DOI:10.3390/f15112023]
- Luo M, Wang Y F, Xie Y H, Zhou L, Qiao J J, Qiu S Y and Sun Y J. 2021. Combination of Feature Selection and CatBoost for Prediction: The First Application to the Estimation of Aboveground Biomass. *Forests*, 12(2): 216 [DOI:10.3390/f12020216]
- Liao Y, Zhang J L, Bao R, Xu D F and Han D Y. 2022. Modelling the Dynamics of Carbon Storages for *Pinus densata* Using Landsat Images in Shangri-La Considering Topographic Factors. *Remote Sensing*, 14(24), 6244 [DOI:10.3390/rs14246244]
- Meng Ge, Zhao D, Xu C, Chen J H, Li X W, Zheng Z J and Zeng Y. 2024. Forest aboveground biomass estimation combining ICESat-2 and GEDI spaceborne LiDAR data. *National Remote Sensing Bulletin*, 28(6): 1632-47 [DOI:10.11834/jrs.20222120]
- Massey F J. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253): 68 - 78 [DOI:10.1080/01621459.1951.10500769]
- Pan Y, Birdsey R A, Pang J Y, Houghton R, Kauppi P E, Kurz W A, Phillips O L, Shvidenko A, Lewis S L, Canadell J G, Ciais P, Jackson R B, Padala S W, McGuire A D, Piao S L, Rautiainen A, Sitch S and Hayes D. 2011. A Large and Persistent Carbon Sink in the World's Forests. *Science*, 333(6045): 988 - 93 [DOI:10.1126/science.1201609]
- Pang A, Chang M and Chen Y. 2022. Evaluation of Random Forests (RF) for Regional and Local-Scale Wheat Yield Prediction in Southeast Australia. *Sensors*, 2022, 22(3): 717 [DOI: 10.3390/s22030717]
- Peng M, Xu M R, Zhang, J L, Qiu B., Teng, C K, and Chen C Q. 2025. Mapping forest aboveground carbon stock of combined stratified sampling and RFRK model with mean annual temperature and precipitation. *Scientific Reports*, 15:17410 [DOI:10.1038/s41598-025-02338-8]
- Qi Y J and Li F R. 2015. Remote sensing estimation of forest aboveground carbon storage in the Greater Khingan Mountains based on KNN method. *Forestry Science*, 51(5): 46 - 55 (戚玉娇, 李凤日. 2015. 基于KNN方法的大兴安岭地区森林地上碳储量遥感估算. 林业科学, 51(5): 46 - 55 [DOI:10.11707/j.1001-7488.20150506])
- Samad A, Ahmad I, Kano M and Caliskan H. 2023. Prediction and optimization of exergetic efficiency of reactive units of a petroleum refinery under uncertainty through artificial neural network-based surrogate modeling. *Process Safety and Environmental Protection*, 177: 1403 - 14 [DOI:10.1016/j.psep.2023.07.046]
- Singh A, Kushwaha S K P, Nandy S, Padalia H, Ghosh S, Srivastava A and Kumari N. 2023. Aboveground Forest Biomass Estimation by the Integration of TLS and ALOS PALSAR Data Using Machine Learning. *Remote Sensing*, 15(4): 1143 [DOI: 10.3390/rs15041143]
- Sun X L. 2016. A remote sensing model for biomass estimation of Shangri-La alpine pine forest based on Landsat8-OLI. Yunnan: Southwest Forestry University (孙雪莲. 2016. 基于Landsat8-OLI的香格里拉高山松林生物量遥感估测模型研究. 云南: 西南林业大学)
- Sun Z H, Guo L, Tao X, Li Y, Zhan Y, Li S L and Zhao Y. 2024. Water Quality Inversion Framework for Taihu Lake Based on Multilayer Denoising Autoencoder and Ensemble Learning. *Remote Sensing*, 16(24) [DOI:10.3390/rs16244793]
- Teng C K, Xiao Y Y, Zhang J L, He Y R and Chen C Q. 2024. Estimation of *Pinus densata* carbon storage based on Landsat time series data and ATC filtering algorithm. *National Remote Sensing Bulletin*, 28(11): 2927 - 42 (滕晨凯, 肖月瑶, 张加龙, 和云润, 陈朝情. 2024. 基于Landsat时间序列数据和ATC滤波算法的高山松碳储量估测. 遥感学报, 28(11): 2927 - 42 [DOI:10.11834/jrs.20244006])
- Tahmouresi M S, Niksokhan M H and Ehsani A H. 2024. Enhancing spatial resolution of satellite soil moisture data through stacking ensemble learning techniques. *Scientific Reports*, 14:25454 [DOI: 10.1038/s41598-024-77050-0]
- Wang K Y, Wang Y F, Chen X, Zheng J P, Li J and Sun Y J. 2025. Remote sensing estimation of forest carbon storage in Jiangxi Province based on ensemble learning algorithms and Optuna tuning. *Ecological Journal*, 45: 685-700 (王可月, 王轶夫, 陈馨, 郑峻鹏, 李杰, 孙玉军. 2025. 基于集成学习算法和Optuna调优的江西省森林碳储量遥感估测. 生态学报, 45: 685-700 [DOI:10.20103/j.stxb.202403150536])
- Wang L G and Gao Y. 2025. Estimating and Downscaling ESA-CCI Soil Moisture Using Multi-Source Remote Sensing Images and Stacking-Based Ensemble Learning Algorithms in the Shandian River Basin, China. *Remote Sensing*, 17(4): 716 [DOI:10.3390/rs17040716]
- Wang Y H and Xing Y M. 2025. Optimized Modeling of Alpine Forest Carbon Storage in the Three-River Source Region via SAR and Multispectral Synergistic Inversion. *Chinese Environmental Science*, 1 - 12 [2025-06-11] (王雁鹤, 邢英梅. 2025. SAR与多光谱协同反演的三江源高寒森林碳储量优化. 中国环境科学, 1 - 12 [2025-06-11]) <https://link.cnki.net/urlid/11.2201.x.20250610.1816.014>
- Wolpert D H. 1992. Stacked generalization. *Neural Networks*, 5(2): 241-259. [DOI:10.1016/S0893-6003(05)80023-1]
- Xu H, Zhang Z Y and Ou C L. 2019. Study on the estimation and distribution of forest biomass and carbon storage in Yunnan Province. Kunming: Yunnan Science and Technology Press: 1-195 (胥辉, 张子翼, 欧光龙. 2019. 云南省森林生物量的碳储量估算及分布研究. 昆明: 云南科技出版社: 1-195)
- Xu L, Shu Q, T Fu H Y, Zhou W W, Luo S L, Gao Y Q, Yu J G, Guo C H, Yang Z D, Xiao J N and Wang S W. 2023. Estimation of Quercus Biomass in Shangri-La Based on GEDI Spaceborne Lidar Data. *Forests*, 14(5): 876 [DOI:10.3390/f14050876]

- Xu M R, Han X L, Zhang J L, Huang K, Peng M, Qiu B and Yang K. 2024. Integrating Ward's Clustering Stratification and Spatially Correlated Poisson Disk Sampling to Enhance the Accuracy of Forest Aboveground Carbon Stock Estimation. *Forests*, 15(12): 2111 [DOI:10.3390/f15122111]
- Yang K, Luo K, Zhang J L, Qiu B, Wang F P, Xiao Q L, Cao L, He Y R and Yang J. 2025. Estimating forest aboveground carbon sink based on landsat time series and its response to climate change. *Scientific Reports*, 15(1): 589 [DOI: 10.1038/s41598-024-84258-7]
- Yao Y R, Lu Y M. 2025. Heterogeneous Ensemble Landslide Susceptibility Assessment Method Considering Spatial Heterogeneity. *Remote Sensing*, 17(21): 3639 [DOI: 10.3390/rs17213639]
- Yue C R. 2012. Forset Biomass Estimation in Shangri-La County Based on Remote Sensing. Beijing: Beijing Forestry University (岳彩荣. 2012. 香格里拉县森林生物量遥感估测研究. 北京: 北京林业大学)
- Zhang Z H, Wu S X, Zhao Z F, Li X L, Zeng F J, Xie C H, Hou G Y and Luo G P. 2022. Estimation of aboveground biomass of grassland based on machine learning algorithm: A case study of grassland in Qilian Mountains. *Ecological Journal*, 42(22): 8953 - 63 (张子慧, 吴世新, 赵子飞, 李向文, 曹凡江, 谢聪慧, 侯冠宇, 罗格平. 2022. 基于机器学习算法的草地地上生物量估测——以祁连山草地为例. 生态学报, 42(22): 8953 - 63 [DOI:10.5846/stxb202203180669])
- Zhao Z F, Li Z Y, Lv P H, Zhao F and Niu L. 2025. The Study on Landslide Hazards Based on Multi-Source Data and GMLCM Approach. *Remote Sensing*, 17(9): 1634 [DOI:10.3390/rs17091634]
- Zhou L, Ou G L, Wang L F and Xu H. 2020. Remote sensing estimation of biomass and determination of light saturation point of Simao pine forest based on spatial regression model. *Forestry Science*, 56(03): 38 - 47 (周律, 欧光龙, 王俊峰, 胥辉. 2020. 基于空间回归模型的思茅松林生物量遥感估测及光饱和点确定. 林业科学, 56(03): 38 - 47 [DOI:10.11707/j.1001-7488.20200305])
- Zhang Y Z, Ma J, Liang S L, Li X S and Liu J D. 2022. A stacking ensemble algorithm for improving the biases of forest aboveground biomass estimations from multiple remotely sensed datasets. *GI-Science & Remote Sensing*, 59(1): 234 - 249 [DOI: 10.1080/15481603.2021.2023842]
- Zhang X, Cui J T, Wang W S and Lin C. 2017. A Study for Texture Feature Extraction of High-Resolution Satellite Images Based on a Direction Measure and Gray Level Co-Occurrence Matrix Fusion Algorithm. *Sensors*, 17(7): 1474. [DOI:10.3390/s17071474]
- Zhou J J, Yan Guo R Y, Sun M G, Di Tu T G, Wang S, Zhai J Y and Zhao Zhong. 2017. The Effects of GLCM parameters on LAI estimation using texture values from Quickbird Satellite Imagery. *Scientific Reports*, 7(1): 7366 [DOI:10.1038/s41598-017-07951-w]

Remote sensing inversion of forest aboveground carbon storage based on GAC feature selection and AST regression algorithm

SUN Kaiping¹, ZHANG Jialong², TENG Chenkai², YANG Kun², HUANG Kai³, LEI QiWang², XIONG Dengliang⁴

1.School of Soil and Water Conservation, Southwest Forestry University, Kunming 650224, Yunnan, China;

2.College of Forestry, Southwest Forestry University (Asia-Pacific Forestry College), Kunming, Yunnan 650224;

3.Faculty of Land and Resources Engineering, Kunming University of Science and Technology, Kunming, Yunnan 650093;

4.Yunnan Provincial Basic Surveying and Mapping Technology Center, Kunming, Yunnan 650034, China

Abstract: Objective Accurately estimating the aboveground carbon stock of forests is crucial for forest management and the sustainable development of forest ecosystems. Since single machine learning models still face issues such as weak generalization, low estimation accuracy, and high uncertainty when estimating forest carbon stocks, this study aims to explore new hybrid models to improve the efficiency of model building and prediction accuracy, thereby enhancing the accuracy of aboveground forest carbon stock estimation. Methods Taking Forest Resources Class II Survey data and Landsat 8 OLI imagery as data sources, a feature selection method integrating Genetic Algorithm (GA) and CatBoost (Genetic Algorithm and CatBoost, GAC) was proposed, combining GA's global optimization capability in feature subset exploration with CatBoost's prominent strengths in mining nonlinear feature relationships and quantitatively evaluating feature importance; on this basis, GAC was systematically compared with the traditional Recursive Feature Elimination (RFE) method to screen remote sensing feature variables that effectively improve model accuracy, while the Hyperopt hyperparameter optimization algorithm was adopted to iteratively search the hyperparameter space of each machine learning regression model to obtain the optimal parameter combinations. Subsequently, a stacked ensemble AST regression algorithm based on base learner mean fusion and meta-learner-based adaptive weighting was constructed, which selects four optimized single machine learning models—Adaptive Boosting (AdaBoost), CatBoost, Random Forest Regression (RFR), and Light Gradient Boosting Machine (LightGBM)—as base learners to fully leverage the unique advantages of each model and significantly enhance the overall performance of the algorithm. Finally, remote sensing estimation models for carbon stock were established based on six single machine learning regression models (AdaBoost, CatBoost, RFR, LightGBM, Support Vector Machine

(SVM), and Extreme Gradient Boosting (XGBoost)) as well as the AST ensemble model, and after comprehensively comparing the estimation accuracies of all candidate models, the optimal one was selected to conduct high-precision inversion mapping of *Pinus densata* carbon stock in Shangri-La City. Results: 1) RFE selected 9 variables, while GAC selected 7 variables, among which the 7 variables selected by GAC contributed more to the accuracy of *Pinus densata* AGC inversion; 2) Based on Hyperopt, the hyperparameters of each model were iteratively optimized, and it was found that the optimal feature subset selected by GAC, when combined with the AST algorithm for regression fitting, achieved the best estimation accuracy, with a coefficient of determination $R^2 = 0.885$, a root mean square error $RMSE = 8.321 \text{ t/hm}^2$, and a prediction accuracy $P = 86.4\%$; 3) Based on the optimal estimation model, the aboveground carbon stock of *Pinus densata* in Shangri-La City in 2016 was estimated to be 7.70953 million t, with an average carbon density of 40.015 t/hm^2 ; 4) The directionality of texture features has a significant impact on the estimation accuracy of forest carbon storage by the AST model, and the 45° diagonal direction is the optimal direction for carbon storage estimation under this model. Conclusion The AST algorithm exhibits higher stability and anti-interference ability under multiple cross-validations, which effectively improves the estimation accuracy of nonparametric models and thereby reduces model uncertainty. This method can provide an effective reference for the dynamic monitoring of forest resources in other high-altitude areas.

Key words: Carbon stocks, Hyperopt hyperparameter tuning, machine learning, AST, *Pinus densata*, GAC, remote sensing inversion, uncertainty analysis

Supported by Supported by the National Natural Science Foundation of China (NO.32260390) and Key Project of Yunnan Fundamental Research Projects (NO.202501AS070047).

NATIONAL
REMOTE
SENSING BULLETIN | 遥感学报

www.ygxb.ac.cn

www.ygxb.ac.cn

NATIONAL
REMOTE
SENSING BULLETIN | 遥感学报