doi:10.6041/j.issn.1000-1298.2025.10.040

基于 MCMC 联合多模型框架的大兴安岭地区森林蓄积量估算

孙乡楠[1,2]　庞　勇[1]　曾伟生[2]　陈新云[2]　谢运鸿[3]　王藏姣[1]

(1. 中国林业科学研究院资源信息研究所, 北京 100091; 2. 国家林业和草原局林草调查规划院, 北京 100714;
3. 北京林业大学森林资源和环境管理国家林业和草原局重点开放性实验室, 北京 100083)

摘要: 为精准估算大兴安岭地区森林蓄积量,本研究选取多源遥感 Sentinel-1 和 Sentinel-2 影像,结合地理因子和气候因子,利用 Boruta 方法筛选自变量,采用随机森林(RF)、梯度提升机(GBM)、轻量梯度提升机(LGBM)、极端梯度提升(XGB)、自适应增强(ADB)、极端随机树(ETR)、K-最近邻(KNN)和支持向量机(SVM)8 种机器学习模型进行样地尺度森林蓄积量回归建模,并通过马尔科夫链蒙特卡洛(Markov chain Monte Carlo,MCMC)算法联合多模型框架算法组合多模型和优化组合权重,对测试数据集进行预测,生成联合多模型预测结果。结果表明,单模型预测中,极端随机树模型表现最佳,其十折交叉验证的 Pearson 相关系数 r、决定系数 R^2、均方根误差(RMSE)和平均相对误差(MRE)均值分别为 0.752、0.584、1.884 m^3 和 29.3%。联合模型中,五模型(LGBM + XGB + ADB + ETR + SVM)和六模型(RF + LGBM + XGB + ADB + ETR + SVM)的联合 R^2 最高(0.62),RMSE 最低,双模型的联合 R^2 提升最大(0.037)。根据 R^2 和 RMSE 评判,联合模型整体表现优于单一模型,显示出更高的预测准确性和稳定性。利用多源数据和 MCMC 联合多模型框架能有效提高大兴安岭地区森林蓄积量估算精度,为森林资源管理提供科学数据支持,并可推广应用于其他森林生态参数预测与模拟。

关键词: 森林蓄积量; 多源遥感; 机器学习; 马尔科夫链蒙特卡洛算法

中图分类号: S771.8; S757.2　　**文献标识码**: A　　**文章编号**: 1000-1298(2025)10-0458-12　　**OSID**:

Estimation of Forest Stock Volume in Greater Khingan Mountains Based on MCMC Joint Multi-model Framework

SUN Xiangnan[1,2]　PANG Yong[1]　ZENG Weisheng[2]　CHEN Xinyun[2]　XIE Yunhong[3]　WANG Cangjiao[1]

(1. *Institute of Resource Information, Chinese Academy of Forestry, Beijing* 100091, *China*
2. *Institute of Forest and Grassland Inventory and Planning, National Forestry and Grassland Administration, Beijing* 100714, *China*
3. *National Forestry and Grassland Administration Key Laboratory of Forest Resources and Environmental Management, Beijing Forestry University, Beijing* 100083, *China*)

Abstract: Aiming to accurately estimate forest stock in the Greater Khingan Range region, selecting Sentinel-1 and Sentinel-2 images, combined with geographical and climatic factors, and the Boruta method was used to screen independent variables. Totally eight machine learning models, including random forest (RF), gradient boosting machine (GBM), lightweight gradient boosting machine (LGBM), eXtreme gradient boosting (XGB), adaptive boosting (ADB), extremely randomized trees (ETR), K-nearest neighbors (KNN), and support vector machine (SVM), were used to regress forest stock at the sample scale. The Markov chain Monte Carlo (MCMC) joint multi model framework algorithm was used to combine multiple models and optimize the combination weights for testing, and predict the dataset and generate joint multi model prediction results. The results showed that the ETR performed the best in single model prediction, with Pearson correlation coefficient r, determination coefficient R^2, root mean square error (RMSE), and mean relative error (MRE) mean values of 0.752, 0.584, 1.884 m^3, and 29.3%, respectively, obtained through ten-fold cross-validation. In the joint

收稿日期: 2024-06-21　修回日期: 2024-09-11

基金项目: 国家重点研发计划项目(2023YFE0105100)

作者简介: 孙乡楠(1990—),男,博士生,主要从事森林经理学和林业遥感研究,E-mail: sunxn@caf.ac.cn

通信作者: 庞勇(1976—),男,研究员,博士,主要从事林业遥感机理模型和激光雷达信号处理研究,E-mail: pangy@ifrit.ac.cn

model, the joint R^2 of the five models (LGBM + XGB + ADB + ETR + SVM) and the six models (RF + LGBM + XGB + ADB + ETR + SVM) was the highest (0.62), the RMSE was the lowest, and the joint R^2 improvement of the two models was the highest (0.037). According to the judgment of R^2 and RMSE, the overall performance of the combined model was better than that of the single model, showing higher prediction accuracy and stability. The use of multi-source data and MCMC combined with a multi model framework can effectively improve the estimation accuracy of forest stock in the Greater Khingan Range region, provide scientific data support for forest resource management, and can be widely applied to the prediction and simulation of other forest ecological parameters.

Key words: forest stock volume; multi-source remote sensing; machine learning; Markov chain Monte Carlo algorithm

0　引言

森林是地球上最大的生态系统,其蓄积量作为量化其数量的重要指标之一[1],对森林经营和管理具有重要意义[2]。传统的森林生物量测量方法准确性较高,但主要依赖于现场观察,不仅成本高、耗时长,而且在山地中,尤其是在原始森林中,会对环境造成潜在的破坏。然而,随着遥感(Remote sensing,RS)技术的发展,该情况得到了显著改善[3]。借助卫星传感器,RS 技术在全球范围内收集到具有不同空间、光谱和时间分辨率的数据(例如 Landsat-8、Sentinel-1、Sentinel-2 等)。这些数据已成为一种具有成本效益的工具,广泛用于估算森林参数,包括森林蓄积量、生物量和碳储量[4]。此外,加入额外的因子,如地理因子和气候因子可以一定幅度地增加估算参数精度[5]。

遥感技术的广泛应用大大促进了森林参数估算的进步,但在复杂的森林环境中,建立精确模型仍存在挑战。为了应对这些挑战,预测回归方法被广泛应用,这些方法可以概括为 6 类参数模型[6]:线性模型和广义线性模型、最小二乘法、投影法、LASSO 回归和岭回归、贝叶斯模型、分位数回归,这些模型通过明确的数学表达式建立遥感变量与森林结构参数间的关系。线性模型和广义线性模型具有计算简单、解释性强的优点,但在处理复杂的非线性关系时表现不足。LASSO 回归和岭回归通过正则化方法减少过拟合,但在高维度数据下可能难以捕捉复杂的模式[7]。贝叶斯模型通过引入先验分布,能结合先验知识进行推断,但计算复杂度较高;8 类非参数模型:广义加性模型、高斯过程、最近邻法[8]、回归树和规则[9]、随机森林、支持向量回归、神经网络和深度学习[10],这些模型不依赖于预设的数学形式,能灵活地捕捉数据中的非线性关系;还有一类是集成方法(可包含参数和非参数模型):装袋法和提升法。尽管参数模型可以方便地建立遥感变量与森林结构参数的回归方程,但在复杂森林环境中,这种关系往往是非线性的。相比之下,非参数模型(如随机森林算法)和集成方法能识别出遥感变量与森林结构参数间的非线性关系,并忽略遥感变量间的协方差[11]。此外,作为非参数模型的一种,机器学习模型相较于传统经验模型,不仅能克服特征因子之间的共线性问题,还具有更好的泛化能力,能处理类别特征,尤其适用于大数据量和多特征的数据。因此,机器学习方法在遥感数据分析和森林参数估算中的作用日益重要[12-13]。

与单一模型相比,集成模型通过整合多个基础模型,提供了更优越和全面的预测能力[14]。根据学习器的构建和结合方式,集成学习可以分为基于同质学习器的集成(袋装法、增强法、堆叠泛化)、基于异质学习器的集成(投票法、混合法)和基于贝叶斯方法的集成(贝叶斯平均法、马尔科夫链蒙特卡洛[15])。在森林变量预测研究中,集成模型相对单一模型表现更出色。例如,YU 等[16]在预测塞罕坝森林冠层高度时发现,多层感知机、支持向量机、最近邻法和随机森林模型均存在不同程度的高估和低估,而堆叠集成模型显著减少了这些误差,达到了最佳拟合效果。文献[13,17-18]均通过堆叠集成的方法估测森林蓄积量,相比单一模型均呈现更高的精度和更低的均方根误差。大部分研究以堆叠式集成学习为主,基于贝叶斯方法的集成学习更少运用于森林变量预测。

在现代统计学和计算科学中,马尔科夫链蒙特卡洛(Markov chain Monte Carlo, MCMC)方法因其在处理复杂概率模型中的独特优势,已成为一种广泛应用的工具。MCMC 方法通过构建一个满足马尔科夫性质的随机过程,利用其长期平稳分布逼近复杂的多维积分和概率分布。其核心思想是利用依赖性(马尔科夫性质)和随机性(蒙特卡洛模拟)实现高效的样本生成,从而对目标分布进行估计和推断[19]。MCMC 联合估测相比于其他集成学习,可以更稳健地处理数据中的不确定性和复杂性,增强法的集成学习可能会对噪声和异常值敏感;MCMC 方

法可以直接从后验分布采样,提供更全面的模型参数估计,而堆叠泛化式集成学习可能受限于基模型多样性和元模型能力;有些梯度的优化方法可能在梯度信息可以计算的情况下有效,但 MCMC 方法不依赖于梯度信息,适用于梯度难以计算或模型包含非参数化组件的情况;贝叶斯平均法的集成学习通过结合多个模型的预测提高集成模型稳定性和准确性,但可能忽略模型预测的不确定性,MCMC 方法却可以提供每个模型预测的不确定性估计[20]。将 MCMC 方法与多模型预测框架相结合,能进一步提升预测因变量的估算精度,现有的林业遥感研究中,很少有学者利用 MCMC 将多模型联合估测因变量。

为精准反演大兴安岭地区森林蓄积量,以第九期国家森林资源清查(National forest inventory,NFI)中黑龙江省大兴安岭地区样地数据作为验证,利用 Sentinel-1 和 Sentinel-2 影像,结合地理因子和气候因子作为多源数据,选取 8 种表现优异的机器学习模型采取 MCMC 联合架构估算样地尺度的森林蓄积量,为森林资源管理和保护提供更为科学的数据支持,为森林资源监测领域提供一种高效科学的指导方法。

1 材料与方法

1.1 研究区概况

黑龙江省大兴安岭地区是中国北部重要生态屏障,地理坐标为 50°05′~53°34′N,121°41′~127°10′E,总面积约 84 000 km^2。平均海拔 897 m,地形以低山丘陵为主,最高峰摩天岭海拔 1 712 m。气候属寒温带季风,年均温度低于 0℃,冬季严寒,夏季短暂湿润。年降水量约 450 mm,集中于夏季。地形复杂,南北气候差异大,北部冬季极寒,南部相对温和。植被以兴安落叶松(*Larix gmelinii*)和樟子松(*Pinus sylvestris* var. mongolica)为主,森林覆盖率高,生物多样性丰富,是中国唯一的天然针叶林区域,同时大兴安岭是研究森林资源和生态系统变化的关键区域,图 1 为大兴安岭地区样地分布以及横纵向样地森林蓄积量分布。

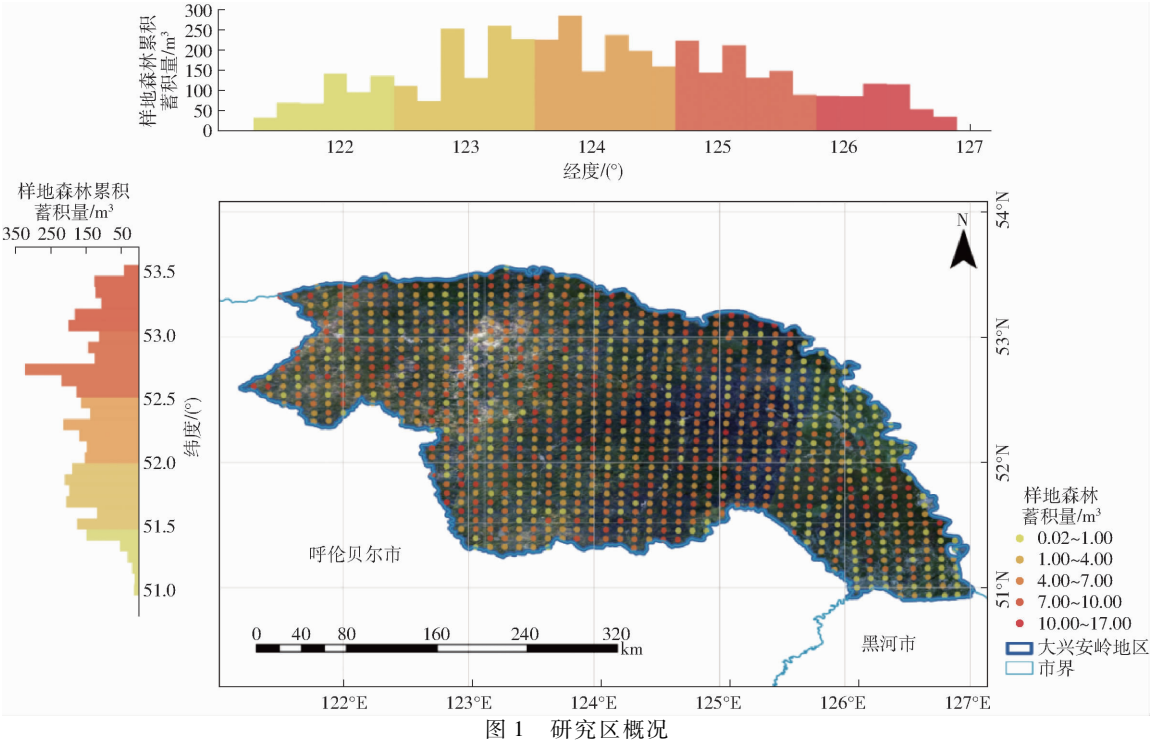


图 1 研究区概况

Fig. 1 Map of study area

1.2 数据源及其处理

1.2.1 国家森林资源清查样地数据

样地数据选自第九期国家森林资源清查数据,以大兴安岭地区为研究地,样地类型只保留乔木林地,以样地森林蓄积量为回归因变量,根据分位数离散法剔除掉 2% 以下和 98% 以上的样地,样地森林蓄积量依据不同树种的一元材积方程计算,参照 LY/T 1353—1999《立木材积表》[21]。纳入建模的样地数量为 887 块,样地为 10 m×60 m 的矩形样地,主要树种包括兴安落叶松、白桦(*Betula platyphylla*)、黑桦(*Betula davurica*)、红松(*Pinus koraiensis*)、樟子松、山杨(*Populus davidiana*)、紫椴(*Tilia amurensis*)、臭冷杉(*Abies nephrolepis*)、青扦(*Picea koraiensis*)、黄波罗(*Juglans mandshurica*)、胡桃楸(*Pterocarya stenoptera*)、榆树(*Ulmus pumila*)。表 1 为样地调查参数的统计特征。

表 1　大兴安岭地区样地调查特征统计

Tab.1　Statistics of plot survey characteristics in Greater Khingan Mountains

统计特征	林分年龄/a	林分平均胸径/cm	林分平均高度/m	郁闭度	森林蓄积量/m^3
均值	61	12.41	12.56	0.51	4.91
最大值	185	52.40	33.00	0.90	16.87
最小值	7	5.00	5.60	0.20	0.02
标准差	33.57	5.30	4.05	0.15	3.09
变异系数	0.55	0.43	0.32	0.29	0.63

1.2.2　遥感影像获取和光谱因子构建

为与第九期 NFI 数据相互对应，基于 Google Earth Engine（GEE）平台，选取 2018 年 6—9 月期间的 Sentinel－1 和 Sentinel－2 影像。首先，对影像进行辐射定标，将原始的数字数值转换为物理单位（如反射率或辐亮度），确保数据的物理意义。其次，使用 FLAASH（Fast line-of-sight atmospheric analysis of hypercubes）模型进行大气校正，消除大气成分（如水汽、气溶胶）对光谱数据的影响。然后，进行几何校正，消除影像获取过程中由于传感器运动、地形起伏等因素引起的几何畸变，确保影像的地理坐标与实际位置一致。最后，在 GEE 平台上对校正后影像进行去云和中值合成处理，以覆盖研究区无云无缝的高质量影像。光谱因子包括：Sentinel－2 影像的 10 个波段，包括蓝（B2）、绿（B3）、红（B4）、红边 1（B5）、红边 2（B6）、红边 3（B7）、近红外（B8）、非近红外（B8A）、短波红外 1（B11）和短波红外 2（B12）；草地叶绿素指数（Green chlorophyll index，gCI）；绿归一化植被指数（Green normalized difference vegetation index，greenNDVI）；改进的土壤调整植被指数（Modified soil-adjusted vegetation index，MSAVI）；改进的三角植被指数 2（Modified triangular vegetation index 2，MTVI2）；归一化植被指数（Normalized difference vegetation index，NDVI）；垂直植被指数（Perpendicular vegetation index，PVI）；红边 1 叶绿素指数（Red-edge1 chlorophyll index，r1CI）；红边 1 归一化差异植被指数（Red-edge1 normalized difference vegetation index，r1NDVI）；红边 1 三角植被指数核心（Red-edge1 triangular vegetation index core，r1RTVIcore）；红边 1 简单比值（Red-edge1 simple ratio，r1SR）；红边 2 叶绿素指数（Red-edge2 chlorophyll index，r2CI）；红边 2 归一化差异植被指数（Red-edge2 normalized difference vegetation index，r2NDVI）；红边 2 三角植被指数核心（Red-edge2 triangular vegetation index core，r2RTVIcore）；红边 2 简单比值（Red-edge2 simple ratio，r2SR）；红边 3 叶绿素指数（Red-edge3 chlorophyll index，r3CI）；归一化水体指数（Normalized difference water index，NDWI）；红边归一化植被指数（Red normalized difference vegetation index，REDNDVI）；红边位置指数（Red edge position index，REP）；红边反射指数 2（Red-edge reflectance index 2，RRI2）；红边 3 归一化差异植被指数（Red-edge3 normalized difference vegetation Index，r3NDVI）；红边 3 三角植被指数核心（Red-edge3 triangular vegetation index core，r3RTVIcore）；红边 3 简单比值（Red-edge3 simple ratio，r3SR）；土壤调节植被指数（Soil-adjusted vegetation index，SAVI）；比值植被指数（Simple ratio index，SR）；转换的土壤调整植被指数（Transformed soil-adjusted vegetation index，TSAVI）；可见光大气阻抗指数（Visible atmospherically resistant index，VARI）；叶绿素红边（Chlorophyll red-edge，CR）；叶绿素红边激发（Chlorophyll red-edge excitation，CRE）；差异色素比植被指数（Differential pigment specific vegetation index，DPSVIm）；倒红边叶绿素指数（Inverted red-edge chlorophyll index，IRECI）；地表水指数（Land surface water index，LSWI）；改良叶绿素吸收的反射率指数 2（Modified chlorophyll absorption in reflectance index 2，MCARI2）；改进的红边简单比值（Modified simple ratio red-edge，MSRre）；陆地叶绿素指数（MERIS terrestrial chlorophyll index，MTCI）；归一化差异指数（Normalized difference index，NDI）；比值植被指数（Ratio vegetation index，RVI）；三角植被指数（Triangular vegetation index，TVI）；垂直垂直极化（Vertical-vertical polarization，VV）；垂直水平极化（Vertical-horizontal polarization，VH）。基于预处理后的影像数据，经过一系列光谱因子计算，以支持后续的回归分析。

1.2.3　地理和气候因子获取

地理因子通常包括样地经纬度坐标（Longitude、Latitude）、坡向（aspect）、坡度（slope_deg）、坡位（slope_pos）、海拔（altitude），其中坡向、坡度、坡位利用 STRM DEM 数据提取。气候因子通过 Climate AP 软件获取，分辨率为 1 km，通过克里金插值与哨兵数据 30 m 分辨率一致进行重新提取气候因子，包含 2018 年 4 个春夏秋冬季节（简称 sp、sm、at、wt）的 13 个因子，分别为最高气温（Tmax，℃）、最低气温（Tmin，℃）、平均气温（Tave，℃）、降雨量（PPT，mm）、气温低于 0℃的温度和（DD_0，℃）、生长积温（DD5，℃）、气温低于 18℃的温度和（DD_18，℃）、气温高于 18℃的温度和（DD18，℃）、无霜期时间

(NFFD, d)、降雪量(PAS, mm)、哈格里夫斯蒸发参考量(Eref, mm/d)、哈格里夫斯水汽亏损量(CMD, mm/d)、相对湿度(RH, %)。

1.2.4 基于 Boruta 方法的自变量筛选

在建模前,对 52 个气候因子、6 个地理因子和 49 个光谱因子进行皮尔森相关性分析(图 2,特征

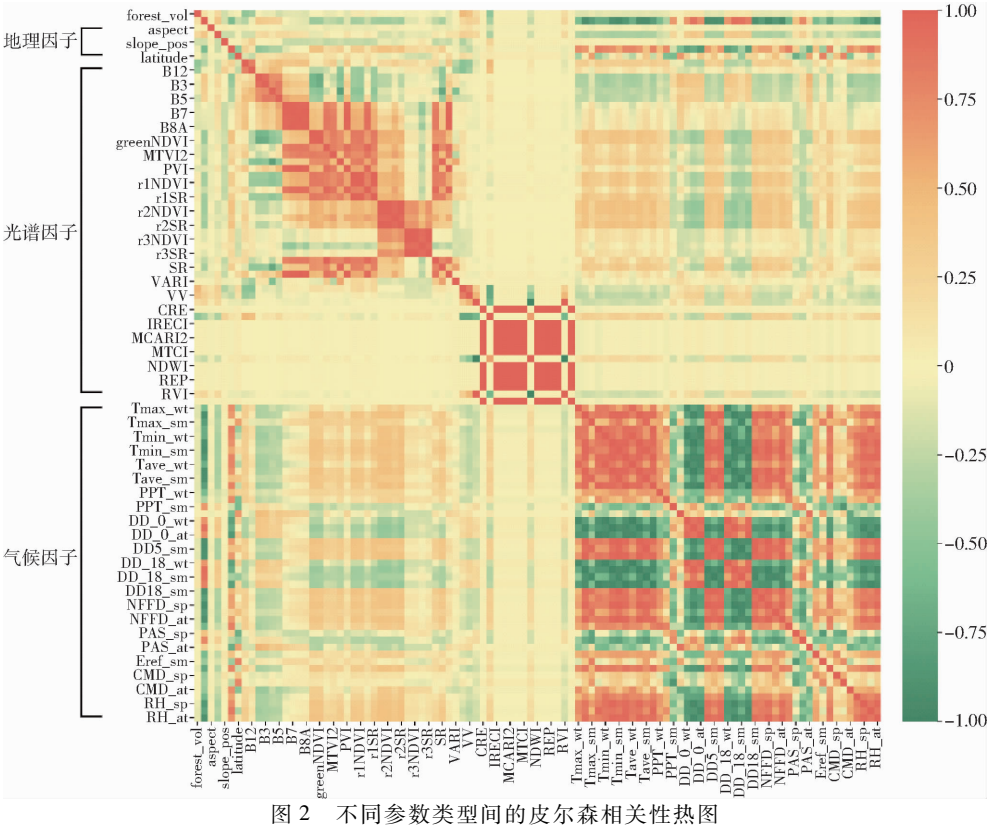


图 2 不同参数类型间的皮尔森相关性热图

Fig. 2 Pearson correlation heatmap between different parameter types

变量数量过多,只显示部分名称,其中 forest_vol 为森林蓄积量)。气候因子间有较强的相关性,其他因子间呈现相对较弱的相关性。图 3 为 Boruta 特征重要性累计分布图,选取 90% 的特征重要性累计值,累计特征(58 个)能解释大多数特征,筛选的特征中包括光谱因子 38 个、地理因子 5 个、气候因子 15 个。

1.2.5 机器学习回归模型

选择 8 种主流的机器学习回归算法:随机森林(Random forest, RF)、梯度提升机(Gradient boosting machine, GBM)、轻量梯度提升机(Lightweight gradient boosting machine, LGBM)、极端梯度提升(eXtreme gradient boosting, XGB)、自适应增强(Adaptive boosting, ADB)、极端随机树(Extremely randomized trees, ETR)、K-最近邻(K-nearest neighbors, KNN)、支持向量机(Support vector machine, SVM)[22]进行样地尺度森林蓄积量回归建模,比较各模型的拟合效果。这些模型存在互补性,基于树的模型(如 RF、GBM、LGBM、XGB、ETR)通常在处理具有复杂非线性关系的数据时表现良好,而 SVM 在处理高维数据时具有优势;KNN 作为一种基于实例的学习算法,可以提供对局部模式的敏感性;

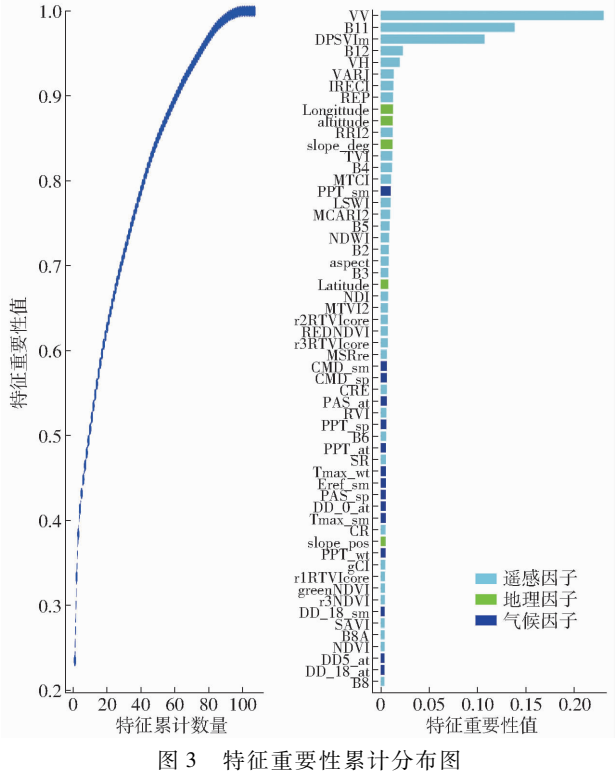


图 3 特征重要性累计分布图

Fig. 3 Cumulative distribution of feature importance

ADB 则在处理噪声数据时表现出鲁棒性。通过结合这些算法,利用它们各自的优势,提高模型的泛化

能力和预测准确性。表 2 为各模型的超参数，各模型超参数使用高斯过程建模获得目标函数的性能，然后根据建模结果选择下一个要评估的参数组合。通过迭代评估新的参数组合逐渐优化目标函数，以找到最佳参数组合。

表 2　各模型超参数

Tab. 2　Hyperparameters for each model

模型	超参数	描述
RF	n_estimators = (10,200)	决策树数量
	max_depth = (1,50)	最大深度
	min_samples_split = (2,20)	最小样本分裂数
	min_samples_leaf = (1,20)	叶子节点最小样本数
	max_samples = (0.1,1)	最大样本比例
GBM	n_estimators = (50,500)	基学习数量
	learning_rate = (0.01,0.3)	学习率
	max_depth = (3,10)	最大深度
	subsample = (0.5,1)	子样本比例
	min_samples_split = (2,20)	最小样本分裂数
	min_samples_leaf = (1,20)	叶子节点最小样本数
LGBM	n_estimators = (50,500)	基学习数量
	learning_rate = (0.01,0.3)	学习率
	max_depth = (3,10)	最大深度
	subsample = (0.5,1)	子样本比例
	colsample_bytree = (0.5,1)	列采样比例
	reg_alpha = (0,1)	L1 正则化系数
	reg_lambda = (0,1)	L2 正则化系数
	min_child_weight = (1,10)	最小子节点权重
XGB	n_estimators = (50,500)	基学习数量
	learning_rate = (0.01,0.3)	学习率
	max_depth = (3,10)	最大深度
	subsample = (0.5,1)	子样本比例
	colsample_bytree = (0.5,1)	列采样比例
	gamma = (0,1)	最小分裂损失
ADB	n_estimators = (50,500)	基学习数量
	learning_rate = (0.01,0.3)	学习率
	max_depth = (3,10)	最大深度
	min_samples_split = (1,50)	最小样本分裂数
	min_samples_leaf = (1,50)	叶子节点最小样本数
	min_weight_fraction_leaf = (0,0.5)	叶子节点最小权重比例
	max_features = (0.1,1)	最大特征数比例
ETR	n_estimators = (50,500)	基学习器数量
	max_depth = (3,10)	最大深度
	min_samples_split = (2,20)	最小样本分裂数
	min_samples_leaf = (1,20)	叶子节点最小样本数
	max_features = (0.1,1)	最大特征数比例
KNN	n_neighbors = (1,50)	邻居数
	uniform = (1,10)	距离权重
	leaf_size = (10,100)	叶子节点尺寸
SVM	C = (10^{-6},10)	正则化参数
	epsilon = (10^{-6},10)	ε-不敏感损失函数
	gamma = (10^{-6},10)	核函数系数
	tol = (10^{-3},10)	优化容忍度

1.2.6　马尔科夫链蒙特卡洛算法

集成策略核心为 MCMC 算法，该算法是一类基于马尔科夫链的随机采样算法，用于从复杂的概率分布中生成样本。一个马尔科夫链是一个随机过程，其中未来状态仅依赖于当前状态，而不依赖于过去状态。马尔科夫链具有状态空间和转移概率矩阵，转移概率决定从一个状态转移到另一个状态的概率。蒙特卡洛模拟是一种通过随机采样计算数学问题的方法，它常用于近似复杂的积分和期望值。MCMC 算法的核心思想是构建一个马尔科夫链，使其状态分布随着迭代趋近于目标分布。通过足够长时间的模拟，马尔科夫链的状态分布将收敛到目标分布，从而可以从中进行采样。本研究使用的 MCMC 算法基于 Metropolis - Hastings，通过以下步骤生成采样：①从当前状态生成候选状态。②根据候选状态与当前状态的目标分布概率比率，决定是否接受候选状态。③若接受，则移动到候选状态；否则保持在当前状态。以下是基于 MCMC 和多模型框架的森林蓄积量估测的算法说明。

算法 1：基于 MCMC 和多模型框架的森林蓄积量估测

输入：多源数据，初始模型参数

输出：森林蓄积量值估算

初始化 MCMC 参数：

- 设定总采样步数(T) = 20 000
- 设定烧入期(B) = 10 000
- 初始化模型权重的 Dirichlet 分布

For each model M_i($i \in \{1,2,\cdots,N\}$), do

　初始化模型 M_i参数

For t = 1 to T, do

　For each model M_i, do

　　利用当前参数从模型 M_i中生成样本

　　计算样本的似然概率

　　利用 MCMC 步骤更新模型参数

　If $t > B$, do

　　基于 Dirichlet 分布存储模型权重

采用加权平均的方式组合模型预测：

　For each model M_i, do

　　利用存储的权重计算加权预测

综合结果估算森林蓄积量

返回估测森林蓄积量

end

根据算法 1，确定输入和输出，设定采样步数和调优步数，初始的采样结果可能不可靠，通常会丢弃一定数量的初始样本(称为“烧入期”)。各模型的权重采用狄利克雷分布(Dirichlet distribution)设定，

之后进行 MCMC 采样得到每个模型权重后进行联合估测森林蓄积量。

1.2.7 精度评价指标

利用皮尔逊相关系数 r、决定系数 R^2、均方根误差(Root mean squared error,RMSE)、平均相对误差(Mean relative error,MRE)和相对均方根误差(Relative root mean squared error,rRMSE)评价不同回归模型准确性。

2 结果与分析

2.1 单模型精度评价

图 4a 为 8 个模型在十折交叉验证中的 R^2 分布。ETR、GBM、LGBM、XGB 表现出较高的中位数和较小的离散程度,表明其预测结果较为稳定。图 4b 为 8 个模型 R^2 分布,进一步验证了 ETR、GBM、LGBM 和 XGB 模型的优越性。表 3 为各模型十折交叉验证评价指标均值,综合表 3 和图 4a,ETR 具有最高中位数、均值(0.584)和最小箱体长度,显示出较小离散程度。GBM、LGBM 和 XGB 箱体长度也相对较短,但略长于 ETR。从图 4b 可以看出,ETR 小提琴形状最窄,表示其 R^2 分布较集中,离散程度最小。GBM、LGBM 和 XGB 小提琴形状稍宽,但总体分布也较为集中。综上所述,ETR 模型在森林蓄积量预测中表现最佳,具有较高的预测准确性和稳定性,而 KNN 和 SVM 模型则表现较差。

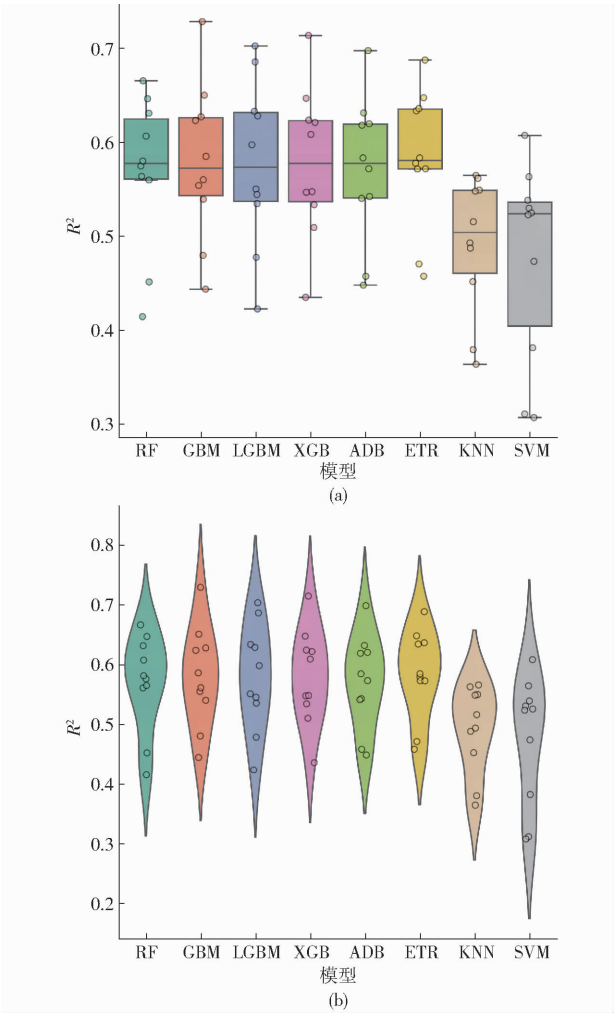


图 4 8 个模型的十折交叉验证 R^2 箱型图和小提琴图

Fig. 4 Boxplot and violin plot of R^2 values from ten-fold cross-validation for eight models

表 3 各模型十折交叉验证评价指标均值

Tab. 3 Mean evaluation metrics of ten-fold cross-validation for each model

模型	r	R^2	RMSE/m^3	MRE/%
RF	0.740	0.569	1.916	33.9
GBM	0.746	0.579	1.893	31.3
LGBM	0.745	0.577	1.895	30.3
XGB	0.745	0.578	1.894	30.1
ETR	0.752	0.584	1.884	29.3
ADB	0.745	0.571	1.912	32.6
KNN	0.688	0.491	2.091	35.8
SVM	0.697	0.476	2.121	37.8

将数据以比例 8∶2随机划分训练集和测试集,经过 8 个模型的训练和测试得到测试集中森林蓄积量的拟合散点图(图 5,图中阴影为 99% 置信区间)。由图 5 可知,ETR 模型在所有模型中表现最佳,具有最高 R^2 和最低误差。KNN 和 SVM 模型表现最差,预测误差较大,但 SVM 预测散点分布较为集中,其 r 仅次于 ADB 模型。几乎所有模型在蓄积量调查值约为 7.5 m^{3}后出现高值低估现象,SVM 模型出现高值低估的现象还更早一些,在蓄积量调查值约为 5.5 m^{3}时便呈现该现象。

2.2 MCMC 联合多模型精度评价

基于比例 8∶2随机划分数据集中的测试集进行构建 MCMC 联合多模型估测森林蓄积量架构,分别试验双模型、三模型、四模型、五模型、六模型、七模型和八模型的联合估测效果,寻找最合适的模型组合,分析不同模型组合的差异和影响因素。

2.2.1 不同模型数量联合预测

不同数量模型组合的精度评价中八模型只有一个组合,其联合 R^2 为 0.616,ΔR^2 为 0.014,RMSE 为 2.034 m^3,ΔRMSE 为 −0.036 m^3。如表 4 所示,联合模型中联合 R^2 最大(0.62)的为五模型和六模型,联合 R^2 提升最大(0.037)的为双模型,RMSE 最小(2.023 m^3)的为五模型和六模型,RMSE 下降量最大(0.089 m^3)的为三模型。

2.2.2 MCMC 联合预测结果

图 6 为不同数量模型组合的联合模型 R^2 及其提升值变化趋势,不同数量模型进行排列组合并联合测试拟合精度,双模型至八模型组合数量分别为

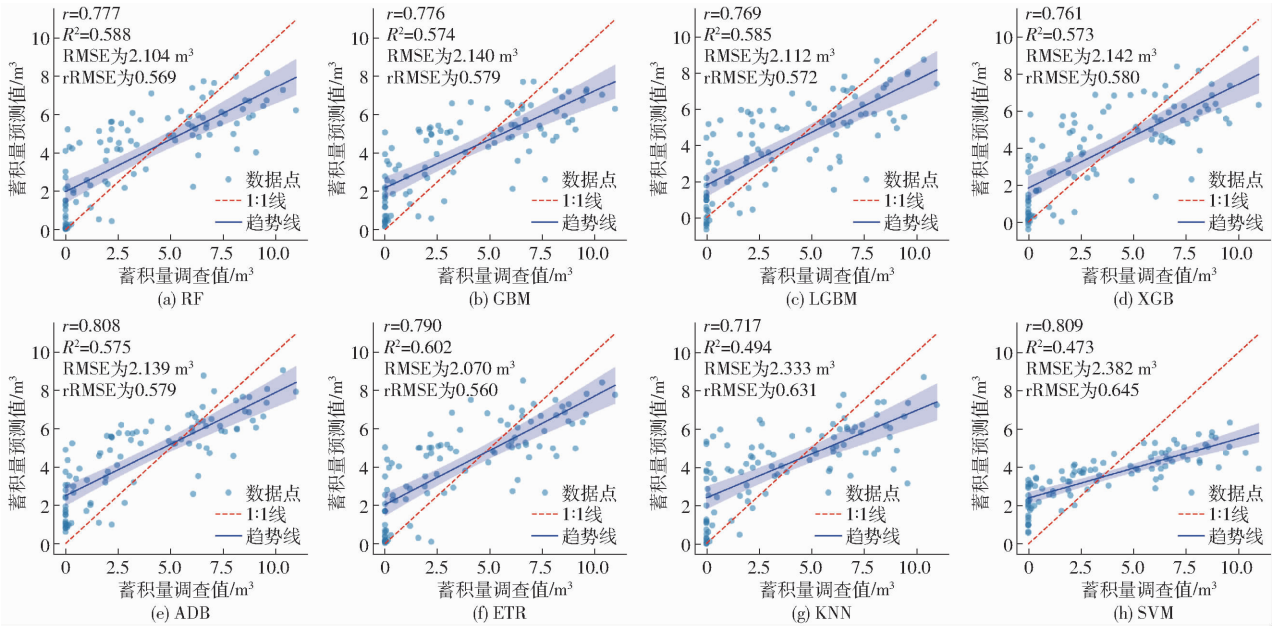


图 5　各模型森林蓄积量预测散点图

Fig. 5　Scatter plots of forest stock volume predictions for each model

表 4　不同数量模型组合精度评价

Tab. 4　Accuracy evaluation of combinations with different numbers of models

参数	双模型				三模型			
	R^2	ΔR^2	RMSE/m³	ΔRMSE/m³	R^2	ΔR^2	RMSE/m³	ΔRMSE/m³
均值	0.596	0.013	2.085	−0.032	0.605	0.015	2.061	−0.038
最大值	0.610	0.037	2.246	0.011	0.617	0.035	2.124	0.009
最小值	0.531	−0.004	2.048	−0.087	0.581	−0.003	2.031	−0.089
参数	四模型				五模型			
	R^2	ΔR^2	RMSE/m³	ΔRMSE/m³	R^2	ΔR^2	RMSE/m³	ΔRMSE/m³
均值	0.609	0.016	2.050	−0.041	0.612	0.015	2.044	−0.040
最大值	0.618	0.032	2.092	0.005	0.620	0.030	2.068	−0.003
最小值	0.593	−0.002	2.028	−0.084	0.603	0.001	2.023	−0.079
参数	六模型				七模型			
	R^2	ΔR^2	RMSE/m³	ΔRMSE/m³	R^2	ΔR^2	RMSE/m³	ΔRMSE/m³
均值	0.613	0.015	2.039	−0.039	0.615	0.015	2.036	−0.038
最大值	0.620	0.029	2.056	−0.013	0.618	0.026	2.044	−0.026
最小值	0.607	0.005	2.023	−0.075	0.612	0.010	2.026	−0.067

注：ΔR^2 为联合模型 R^2 减去单个模型中最高的 R^2，ΔRMSE 为联合模型 RMSE 减去单个模型中最低的 RMSE。

28、56、70、56、28、8、1。由图 6 可知，只有较少的组合出现联合 R^2 零提升或者负提升的现象，所有组合基本上在不同的程度上实现了正提升，最大的联合 R^2 提升值可达 0.037，说明 MCMC 多模型联合框架的有效性。

表 5 为最优 R^2 联合模型加权公式及其评价指标。由表 5 可知，SVM 在联合模型中权重占比相较其他模型均为最低，与其独立模型精度有一定关联性，KNN 在联合模型中权重占比也处于较低排名，说明模型间的权重与独立模型精度可能存在一定程度的正向相关关系。其中 ETR 在双模型、三模型、四模型、五模型、六模型和八模型联合中多次占最高权重，ETR 独立模型在十折交叉验证中也获得最高的拟合精度，再次证明模型的权重与独立模型拟合精度存在一定的正向关系。此外，SVM 凭借其独立模型拟合散点图中的散点收敛性，除了双模型中未成为最优组合，在三模型、四模型、五模型、六模型和七模型中均出现在最优组合之列，表明最优联合模型不止对独立模型的拟合评价指标敏感，还对拟合散点收敛性有着一定程度上的偏向性。

图 7 为联合模型散点图。对比图 5 和图 7 可知，联合模型在对接近 0 m³ 的样地尺度森林蓄积量较为收敛，该估计值较大偏差出现在 4 m³ 附近，而独立模型中的该估计值较大偏差基本出现在 5 m³ 附近，极小值估计的情况下联合模型拟合效果表现更优越。随着联合模型数量的增加，散点基本呈现逐

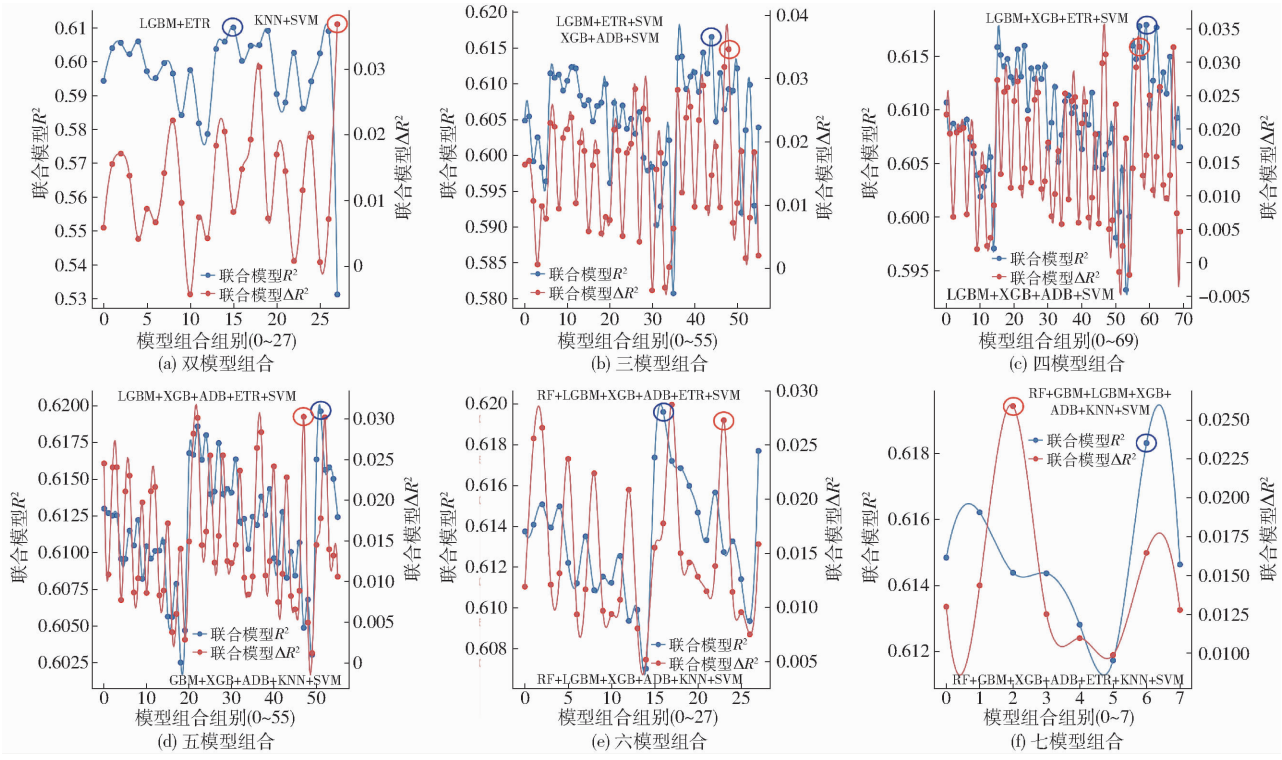


图 6 不同数量模型组合的联合模型 R^2 及其提升值变化曲线

Fig. 6 Trend chart of combined model R^2 and its improvement values for different numbers of model combinations

表 5 最优 R^2 联合模型加权公式及其评价指标

Tab. 5 Weighted formulas and evaluation metrics of optimal R^2 combined models

模型	加权公式	R^2	RMSE/m^3
双模型	$y_2=0.446x_{LGBM}+0.554x_{ETR}$	0.610	2.048
三模型	$y_3=0.366x_{LGBM}+0.404x_{ETR}+0.23x_{SVM}$	0.617	2.031
四模型	$y_4=0.274x_{LGBM}+0.241x_{XGB}+0.293x_{ETR}+0.192x_{SVM}$	0.618	2.028
五模型	$y_5=0.216x_{LGBM}+0.2x_{XGB}+0.2x_{ADB}+0.217x_{ETR}+0.166x_{SVM}$	0.620	2.023
六模型	$y_6=0.166x_{RF}+0.177x_{LGBM}+0.169x_{XGB}+0.169x_{ADB}+0.177x_{ETR}+0.143x_{SVM}$	0.620	2.023
七模型	$y_7=0.148x_{RF}+0.157x_{GBM}+0.15x_{LGBM}+0.149x_{XGB}+0.153x_{ADB}+0.119x_{KNN}+0.124x_{SVM}$	0.618	2.026
八模型	$y_8=0.13x_{RF}+0.115x_{GBM}+0.139x_{LGBM}+0.133x_{XGB}+0.128x_{ADB}+0.134x_{ETR}+0.11x_{KNN}+0.112x_{SVM}$	0.616	2.034

注：x_{RF}、x_{GBM}、x_{LGBM}、x_{XGB}、x_{ADB}、x_{ETR}、x_{KNN}、x_{SVM}分别表示 8 个模型的预测结果。

步集中于拟合线的趋势，说明联合模型数量的增加在一定程度上有利于整体的拟合效果。根据联合模型的评价指标，表现最优的为五模型，证明并不是联合模型数量越多整体拟合精度越高。

3 讨论

3.1 MCMC 联合架构有效性及联合模型方式

根据图 7 结果可知，MCMC 联合架构在多个方面显著优于传统的独立模型，展现出更高的准确性、计算效率和处理复杂性的能力。首先，联合模型能更好地捕捉变量间的复杂关系，提高预测准确性，比起独立模型，其有较高的 r(0.791 ~ 0.812) 和 R^2(0.610 ~ 0.620)，表明其预测结果与实际数据相关性强，拟合度较好；其次，联合模型通过同时考虑多个变量的联合分布，显著降低了预测误差，与独立模型相比，MCMC 联合模型有较低的 RMSE 和 rRMSE，说明其在控制误差方面表现良好，能提供更可靠的预测结果；再次，联合模型在处理高维数据时表现较好，通过利用变量间的相关性，提高了采样效率，降低了计算复杂度，这使得联合模型在大数据背景下具有更强的适应能力；最后，联合模型在处理缺失数据时，能通过对联合分布的建模，直接进行有效的采样和推断，减少因缺失数据带来的误差。

总之，MCMC 联合架构在多方面超越了传统的独立模型。在未来的研究中，还可以通过引入 MCMC 的改进算法[23]和更为科学有效的集成算法，以及考虑贝叶斯平均[3]等集成学习思路，进一步提高回归模型的预测准确性和鲁棒性。

3.2 多模态数据估测

利用遥感数据进行森林蓄积量估测的研究中，

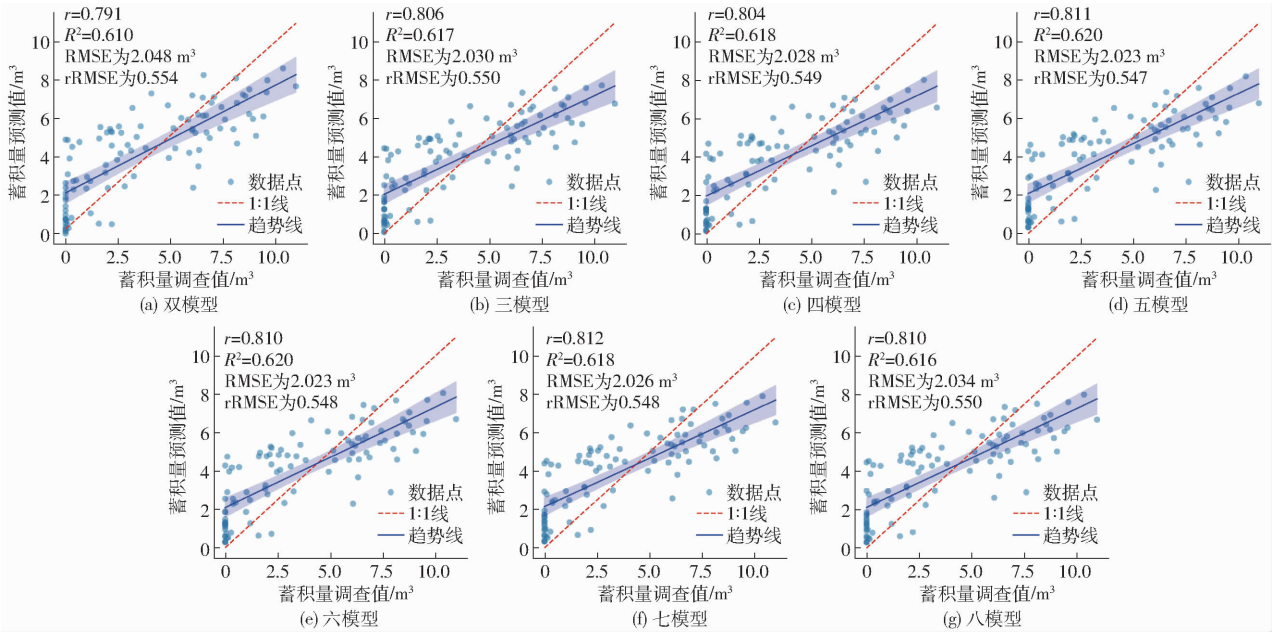


图 7　MCMC 最优联合模型预测散点图

Fig. 7　Scatter plots of MCMC optimal combined model predictions

经常出现“高值低估”的现象，本研究虽然使用联合模型取得较好的估测效果，但也无法避免这种现象，而且并没有比独立模型在更晚时出现这种现象。所以，引入更多的数据源或者遥感因子，或许可以更好地发挥机器学习的非线性拟合性能、降低森林参数估测饱和现象。随着星载遥感技术的发展，越来越多数据可以提供更丰富的信息，从而提高模型预测能力[24]，如 GEDI[25] 和 ICESat－2[26] 等数据。这些开源数据不仅可以提供关于森林结构的高分辨率数据，以便更准确地预测森林生物量和碳储量；还可以提供相应的气候模型数据，用来预测气候变化对森林生态系统的影响。在未来的研究中，可以协同多源遥感数据，利用更丰富的遥感信息，更全面地预测森林生态系统状态和变化，为森林管理和生态保护提供科学依据。

3.3　回归方法优化

总体而言，本研究选取的 8 种机器学习回归算法均取得了较好的效果，为了进一步优化回归方法，可以通过采用更为合适的特征选择方法、引入深度学习中的深度神经网络（Deep neural network，DNN）[27] 以及引入 LIME（Local interpretable model-agnostic explanations）和 SHAP（SHapley additive exPlanations）[28] 等可解释性分析工具，进一步提升回归方法的性能和解释性。一方面，选取合适的特征选择是提高模型性能的关键步骤之一，可以考虑使用基于模型的特征选择方法，如基于树模型的特征重要性评分，以提高特征选择的准确性；另一方面，DNN 能通过多层结构捕捉数据中的复杂模式和非线性关系，且与传统机器学习方法相比，DNN 在特征提取和表示学习方面具有优势；此外，LIME 和 SHAP 作为主流的模型解释性工具，可以看出哪些特征对模型的预测结果有显著影响。这些优化措施有助于构建更加强大和可靠的预测模型，为森林资源监测提供更加有力的支持。

4　结论

（1）RF、GBM、LGBM、XGB、ADB、ETR、KNN、SVM 8 个机器学习模型算法在对样地尺度蓄积量预测中，ETR 的十折交叉评价指标最优，r、R^2、RMSE、MRE 均值分别为 0.752、0.584、1.884 m^3、29.3%。

（2）不同数量的模型联合预测结果中，五模型和六模型组合的联合 R^2 最高（0.62），联合 RMSE 最低，双模型的联合 R^2 提升最大（0.037）。

（3）独立模型的权重与精度在联合模型中有一定关联，相比各独立模型，联合模型更能有效地实现最佳拟合。

（4）基于 Sentinel－1、Sentinel－2、地理因子和气候因子的多源数据弥补了单一数据的信息量不足，MCMC 联合多模型比单一模型表现出更强的模型鲁棒性、更优越的拟合效果以及更高的预测精度。研究结果克服了单一数据源、单一模型预测不稳定等问题，促进了森林蓄积量预测的精确度和可靠性。为大范围林业调查、森林资源管理提供了有效的技术支持，推动了精准林业的发展。

参 考 文 献

[1] 尤静妮. 基于高分遥感纹理信息的森林蓄积量估测研究[D]. 西安:西安科技大学, 2017.
YOU Jingni. Forest volume estimation based on high-resolution remote sensing texture information [D]. Xi'an: Xi'an University of Science and Technology, 2017. (in Chinese)

[2] ZHANG D, NI H. Inversion of forest biomass based on multi-source remote sensing images[J]. Sensors, 2023, 23(23): 9313.

[3] 林文科,陆亚刚,蒋先蝶,等. 协同多源遥感数据的北亚热带森林蓄积量贝叶斯分层估测[J]. 遥感学报, 2022, 26(3): 468 - 479.
LIN Wenke, LU Yagang, JIANG Xiandie, et al. Modeling forest growing stock volume in a north subtropical region using the hierarchical Bayesian approach based on multi-source data[J]. Journal of Remote Sensing, 2022, 26(3): 468 - 479. (in Chinese)

[4] 张少伟,惠刚盈,韩宗涛,等. 基于光学多光谱与 SAR 遥感特征快速优化的大区域森林地上生物量估测[J]. 遥感技术与应用, 2019, 34(5): 925 - 938.
ZHANG Shaowei, HUI Gangying, HAN Zongtao, et al. Estimation of forest aboveground biomass in large area based on fast optimization of optical multi-spectral and SAR remote sensing features[J]. Remote Sensing Technology and Application, 2019, 34(5): 925 - 938. (in Chinese)

[5] NI W, YU T, PANG Y, et al. Seasonal effects on aboveground biomass estimation in mountainous deciduous forests using ZY - 3 stereoscopic imagery[J]. Remote Sensing of Environment, 2023, 289: 113520.

[6] FERNANDEZ M, SIRSAT M S, CERNADAS E, et al. An extensive experimental survey of regression methods[J]. Neural Networks, 2019, 111: 11 - 34.

[7] 王宇璇,屈忠义,白燕英,等. 基于 Sentinel - 2 影像的黄河南岸典型改良示范区土壤含盐量反演模型[J]. 农业机械学报, 2024, 55(4): 290 - 299, 439.
WANG Yuxuan, QU Zhongyi, BAI Yanying, et al. Soil salt inversion of typical improvement demonstration area of south bank of Yellow River based on Sentinel - 2 images[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(4): 290 - 299, 439. (in Chinese)

[8] 陈尔学,李增元,武红敢,等. 基于 k - NN 和 Landsat 数据的小面积统计单元森林蓄积估测方法[J]. 林业科学研究, 2008, 21(6): 745 - 750.
CHEN Erxue, LI Zengyuan, WU Honggan, et al. Forest volume estimation method for small areas based on k - NN and Landsat data[J]. Forest Research, 2008, 21(6): 745 - 750. (in Chinese)

[9] 周小成,黄婷婷,李媛,等. 结合遥感林龄因子的亚热带森林蓄积量估算方法[J]. 林业科学, 2023, 59(4): 88 - 99.
ZHOU Xiaocheng, HUANG Tingting, LI Yuan, et al. A method for estimating subtropical forest stock by combining remotely sensed forest age factors[J]. Scientia Silvae Sinicae, 2023, 59(4): 88 - 99. (in Chinese)

[10] 王鹏新,杜江莉,张悦,等. 基于遥感多参数和 CNN - Transformer 的冬小麦单产估测[J]. 农业机械学报, 2024, 55(3): 173 - 182.
WANG Pengxin, DU Jiangli, ZHANG Yue, et al. Yield estimation of winter wheat based on multiple remotely sensed parameters and CNN - Transformer[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(3): 173 - 182. (in Chinese)

[11] 边瑞. 基于多源遥感数据的祁连山国家公园森林生物量估算研究[D]. 兰州:兰州大学, 2021.
BIAN Rui. Estimation of forest biomass in Qilian Mountain National Park based on multi-source remote sensing data[D]. Lanzhou: Lanzhou University, 2021. (in Chinese)

[12] ZHANG C, DENKA S, COOPER H, et al. Quantification of sawgrass marsh aboveground biomass in the coastal Everglades using object-based ensemble analysis and Landsat data[J]. Remote Sensing of Environment, 2018, 204: 366 - 379.

[13] 付永浩,李伟坡,张爱军. 面向对象和集成学习算法在森林蓄积量遥感估测中的应用[J]. 东北林业大学学报, 2024, 52(6): 64 - 69, 78.
FU Yonghao, LI Weipo, ZHANG Aijun. Application of object-oriented and integrated learning algorithms in remote sensing estimation of forest stock volume[J]. Journal of Northeast Forestry University, 2024, 52(6): 64 - 69, 78. (in Chinese)

[14] 李坤,吴达胜,方陆明,等. 运用 Sentinel - 2 遥感影像数据估测森林蓄积量[J]. 东北林业大学学报, 2021, 49(5): 59 - 66.
LI Kun, WU Dasheng, FANG Luming, et al. Forest volume stock with Sentinel - 2 remote sensing image[J]. Journal of Northeast Forestry University, 2021, 49(5): 59 - 66. (in Chinese)

[15] ZHAO J, LIU D, CAO Y, et al. An integrated remote sensing and model approach for assessing forest carbon fluxes in China [J]. Science of the Total Environment, 2022, 811: 152480.

[16] YU J, NIE S, LIU W, et al. Mapping global mangrove canopy height by integrating Ice, Cloud, and Land elevation Satellite - 2 photon-counting LiDAR data with multi-source images[J]. Science of the Total Environment, 2024, 939: 173487.

[17] 崔博文. 基于机器学习的森林蓄积量估测研究[D]. 长沙:中南林业科技大学, 2022.
CUI Bowen. Estimation of forest stock volume based on machine learning[D]. Changsha: Central South University of Forestry and Technology, 2022. (in Chinese)

[18] 黄宇玲. 基于特征选择和 Stacking 集成学习算法的森林蓄积量估测研究[D]. 杭州:浙江农林大学, 2019.
HUANG Yuling. Estimation of forest stock volume based on feature selection and Stacking integrated learning algorithm[D]. Hangzhou: Zhejiang A&F University. 2019. (in Chinese)

[19] ALI M, DEO R C, DOWNS N J, et al. Cotton yield prediction with Markov chain Monte Carlo-based simulation model integrated with genetic programing algorithm: a new hybrid copula-driven approach[J]. Agricultural and Forest Meteorology, 2018, 263: 428 – 448.

[20] PAN J, DURAND M T, VANDER J B J, et al. Application of a Markov chain Monte Carlo algorithm for snow water equivalent retrieval from passive microwave measurements[J]. Remote Sensing of Environment, 2017, 192: 150 – 165.

[21] LY/T 1353—1999 立木材积表[S]. 1999.

[22] 王鹏新,王颖,田惠仁,等. 基于 LightGBM 的冬小麦产量估测与可解释性研究[J]. 农业机械学报, 2023, 54(12): 197 – 206.
WANG Pengxin, WANG Ying, TIAN Huiren, et al. Interpretability on yield estimation of winter wheat based on LightGBM [J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(12): 197 – 206. (in Chinese)

[23] 马祥跃,杜晓婷,采青,等. 深度学习框架测试研究综述[J]. 软件学报, 2024, 35(8):1 – 33.
MA Xiangyue, DU Xiaoting, CAI Qing, et al. Survey on testing of deep learning frameworks[J]. Journal of Software, 2024, 35(8):1 – 33. (in Chinese)

[24] TRAVERS S H, COOPS N C, MULVERHILL C, et al. Mapping vegetation height and identifying the northern forest limit across Canada using ICESat – 2, Landsat time series and topographic data[J]. Remote Sensing of Environment, 2024, 305: 114097.

[25] 黄佳鹏,李国元,刘诏. 星载激光雷达估测森林结构参数研究现状与发展分析[J]. 农业机械学报, 2024, 55(6):18 – 33.
HUANG Jiapeng, LI Guoyuan, LIU Zhao. Current status and development trend in estimating forest structural parameters with spaceborne LiDAR[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(6):18 – 33. (in Chinese)

[26] 宋涵玥,舒清态,席磊,等. 基于星载 ICESat – 2/ATLAS 数据的森林地上生物量估测[J]. 农业工程学报, 2022, 38(10): 191 – 199.
SONG Hanyue, SHU Qingtai, XI Lei, et al. Remote sensing estimation of forest above-ground biomass based on spaceborne LiDAR ICESat – 2/ATLAS data[J]. Transactions of the CSAE, 2022, 38(10): 191 – 199. (in Chinese)

[27] 韩瑞. 基于变量优选及深度学习的森林蓄积量估测[D]. 杭州:浙江农林大学, 2020.
HAN Rui. Estimation of forest volume based on variable optimization and deep learning[D]. Hangzhou: Zhejiang A&F University, 2020. (in Chinese)

[28] 蒋星驰,温苏雅勒图,李俊瑶,等. 基于 SHAP 值方法的乌拉特梭梭(*Haloxylon ammodendron*)林保护区土壤表层碳氮储量及其影响因素研究[J]. 中国沙漠, 2023, 43(2): 170 – 183.
JIANG Xingchi, WEN Suyaletu, LI Junyao, et al. Study on storage of soil surface carbon and nitrogen and its influencing factors in Urad National Nature Reserve of *Haloxylon ammodendron* based on SHAP values[J]. Journal of Desert Research, 2023, 43(2): 170 – 183. (in Chinese)