江苏农业学报(*Jiangsu J.of Agr.Sci.*),2022,38(3):706-713
http://jsnyxb.jaas.ac.cn

朱泊东,金　京,罗洪斌,等. 基于变量优选的机载激光雷达对林分平均高的反演[J].江苏农业学报,2022,38(3):706-713.
doi:10.3969/j.issn.1000-4440.2022.03.016

基于变量优选的机载激光雷达对林分平均高的反演

朱泊东[1]，　金　京[1]，　罗洪斌[1]，　龙　飞[1]，　李春干[2]，　岳彩荣[1]
(1.西南林业大学林学院,云南 昆明 650224；2.广西大学林学院,广西 南宁 530004)

摘要： 森林高度是反映森林数量和质量的重要指标,是森林经营管理的重要基础数据,准确获取森林高度信息一直是林业遥感研究的目标。本研究以广西高峰林场的105块地面实测样地数据和机载激光雷达(Light detection and ranging，LiDAR)数据为基础,从点云数据中提取35个特征变量,分别采用支持向量机-递归特征消除法(SVM-RFE)、轻量级梯度提升机(LightGBM)和主成分分析(PCA)法进行特征筛选,并结合参数模型(LR)和非参数模型(RFR、KNN)对林分平均高进行反演。研究结果表明,不同特征选择方法和估测模型的组合精度差异较大。其中,利用LightGBM进行特征筛选结合KNN回归反演效果最佳,建模的R^2和*RMSE*分别为0.83和1.64 m,验证的R^2和*RMSE*分别为0.81和1.56 m。此外,在SVM-RFE、LightGBM和PCA这3种特征筛选方法中LightGBM的效果最好,无论在RFR模型还是在KNN模型中均能得到较高的R^2,优于SVM-RFE和PCA。

关键词： 机载激光雷达；林分平均高；特征优选；LightGBM
中图分类号： TP79　　**文献标识码：** A　　**文章编号：** 1000-4440(2022)03-0706-08

Inversion of average forest stand height based on variable selection by airborne laser radar

ZHU Bo-dong[1]，　JIN Jing[1]，　LUO Hong-bin[1]，　LONG Fei[1]，　LI Chun-gan[2]，　YUE Cai-rong[1]
(1.*College of Forestry*, *Southwest Forestry University*, *Kunming* 650224, *China*; 2.*College of Forestry*, *Guangxi University*, *Nanning* 530004, *China*)

Abstract: Forest height is an important indicator of forest quantity and quality, and it is also an important essential parameter for forest management. Obtaining the height information of the forest accurately has always been the target of remote sensing study of the forestry. In this study, 35 characteristic variables were extracted from the point cloud data, based on the measured data from 105 sample ground plots and airborne light detection and ranging (LiDAR) data of Guangxi State-owned Gaofeng Forest Farm. Support vector machine-recursive feature elimination (SVM-RFE), light gradient boosting machine (LightGBM) and principal components analysis (PCA) were used to screen the characteristics, respectively. The average height of forest were inverted by combining parametric model (LR) and nonparametric models (RFR, KNN). The results showed that, there were large differences of accuracy between combinations of different feature selection methods and estimation models. Among them, the combination of LightGBM feature selection with KNN regression inversion showed the best effect. The R^2 and root mean square error (*RMSE*) of modeling were 0.83 and 1.64 m, respectively, and the R^2 and *RMSE* of verification were 0.81 and 1.56 m, respectively. In addition, among the three feature selection methods of SVM-RFE, LightGBM and PCA, LightGBM showed the best effect, which was high in R^2 both in the RFR model and in the KNN model, and was better than SVM-RFE and PCA.

Key words: airborne laser radar; average forest stand height; feature selection; light gradient boosting machine (LightGBM)

收稿日期: 2021-10-29
基金项目: 国家自然科学基金项目(42061072);云南省科技厅重大科技专项(202002AA00007-015);云南省教育厅项目(2018JS330)
作者简介: 朱泊东(1997-),男,云南曲靖人,硕士研究生,主要从事林业遥感研究。(E-mail)zhubodong@swfu.edu.cn
通讯作者: 岳彩荣,(E-mail)cryue@163.com

林分平均高度是生态系统模型重要的输入参数之一,是反映森林生长状况、计算森林蓄积量和生物量的重要因子,与碳循环研究高度相关[1]。森林树高一直是森林调查中最难准确获取的因子之一。传统森林资源调查主要是依靠人工地面调查,需要耗费大量的人力物力财力,且部分区域地形复杂难以进行精确的地面调查[2-3]。遥感技术的出现,给森林资源调查提供了一种全新的技术手段[4]。而对于林分尺度和单木尺度的森林垂直结构探测,激光雷达(Light detection and ranging, LiDAR)技术相较于其他遥感探测手段具有显著优势[5-6]。

在用 LiDAR 估测林分高度的研究中,通常采用从 LiDAR 数据中提取的大量特征中选取部分或全部特征,并在此基础上建立模型,估测林分平均高。穆喜云等[7]基于机载激光雷达数据与样地数据,构建不同树种的林分平均高与郁闭度反演模型,结果表明,不同树种模型的估测精度不同,混交林的林分平均高估测精度高于阔叶林和针叶林。彭涛等[8]分别设置圆形样地和方形样地,利用激光雷达特征变量中的高度分位数估测樟子松及落叶松的林分平均高,结果表明,利用累积高度百分位数中 99% 累积高度估测林分平均高时 2 种样地结果均最优。曾伟生等[9]研究发现,基于机载激光雷达对林分平均高等森林参数进行估测时,非线性模型优于线性模型。焦义涛等[10]基于激光点云数据计算植被点云高度阈值平均值,以此建立预测林分平均高的回归模型,结果表明,利用点云高度阈值的平均值进行林分平均高估测可靠性较高。赵勋等[11]利用随机森林、支持向量机以及组合模型对林分平均高进行估测,结果表明组合模型的泛化能力及预测精度最好。周蓉等[12]通过提取 32 个点云特征变量,并采用逐步回归法和 BP 神经网络算法对针叶林进行林分平均高估测,研究发现 BP 神经网络模型估测精度较高,优于逐步回归法,其 R^2 均在 87% 以上。沈剑波等[13]通过对比森林调查因子和不同神经网络训练算法对平均树高估测结果的影响,发现利用贝叶斯正则化 BP 神经网络能有效提升估测精度。Grace 等[14]研究发现,建模因子的增加并不意味着模型精度的提升,对建模因子进行筛选有助于提升模型精度。曹林等[15]采用主成分分析法、逐步回归法和贝叶斯模型平均法对建模因子进行筛选,结果表明,逐步回归法拟合模型效果最好。Paolo 等[16]比较了不同特征筛选方法(遗传算法特征筛选和随机森林算法特征筛选)对森林参数估测的影响,结果表明,将遗传算法(GAs)和赤池信息量准则(AIC)组合进行变量筛选是最有效的方法。郝红科[17]从归一化点云数据中提取 43 个点云变量,分别采用支持向量机回归、快速人工神经网络以及逐步回归建立林分平均高反演模型,研究结果表明,对建模变量进行筛选能有效提升模型预测精度。

综上所述,基于机载激光雷达的林分平均高反演中点云变量的特征筛选方法与估测模型的选择是最终决定反演精度的 2 个技术环节,针对不同的估测模型选择适当的特征筛选方法有助于提高估测精度,但还需进一步对估测模型与特征筛选方法的最优组合进行探讨和分析。为此,本研究以广西高峰林场为研究区域,利用支持向量机-递归特征消除法(SVM-RFE)、轻量级梯度提升机(LightGBM)和主成分分析(PCA)对 35 个点云特征变量进行建模因子筛选的基础上构建基于逐步回归、随机森林和 K 近邻法的林分平均高估测模型,对比基于不同特征筛选的模型估测结果,为提高机载激光雷达林分平均高的估测精度提供参考。

1 材料与方法

1.1 研究区概况

研究区位于广西最大的国有林场——国营高峰林场,研究区为一个呈东北-西南走向的近矩形区域,长 11.2 km,宽 4.2 km,面积约为4 770 hm^2,中心地理位置为 108°23′45″E,22°58′33″N[18]。高峰林场属于南亚热带季风气候区,日照充足,雨量充沛,其地貌主要为丘陵和山地,地形起伏较大,坡度大多为 20°~35°,最大坡度为 69.7°。

研究区内森林覆盖率达 90% 以上,约 95% 为人工林,主要为短周期经营的桉树人工林和长周期经营的杉木,主要树种包括尾叶桉(*Eucalyptus urohpylla* S.T. Blake)、巨尾桉(*Eucalyptus grandis*×*E. urohpylla*)、杉木(*Cunninghamia lanceolata*)、马尾松(*Pinus massoniana* Lamb.)、湿地松(*Pinus elliottii* Engelmann)、八角(*Lllicium verum* Hook. f.)等[18]。

1.2 地面样地数据

地面调查时间为 2016 年 5 月至 12 月,在研究区内设置 105 个30 m×30 m 的地面实测样地(包括 51 块针叶林样地和 54 块阔叶林样地),记录胸径大

于 5 cm 的每株树的树种、胸径、树高、冠幅等。林分平均高利用断面积加权法(公式 1)进行计算。样地调查信息如表 1 所示。

$$H=\frac{\sum_{i=1}^{n}h_i g_i}{\sum_{i=1}^{n}g_i} \quad (1)$$

式中,n 为林木株数,h_i表示第 i 株树的树高,g_i表示第 i 株树的胸高断面积。

表 1 样地数据统计结果

Table 1 Statistics of plot data

样地	类别	林分平均高(m)	蓄积量(m^3/hm^2)	断面积(m^2/hm^2)
针叶林(51 块样地)	范围	6.20~18.22	57.93~387.20	15.20~50.21
	平均值	13.35	17.59	25.60
	标准差	2.97	63.76	6.70
阔叶林(54 块样地)	范围	6.52~23.56	21.28~317.13	3.92~36.69
	平均值	13.74	139.23	19.25
	标准差	4.30	76.61	7.46

1.3 机载 LiDAR 数据

本研究以 R44 直升机为飞行平台,搭载奥地利 RIEGL 公司生产的 VUX1LR 激光雷达系统获取点云数据,数据获取时间为 2016 年 9 月。遥感平台飞行高度为1 000 m,激光器的波长为1 550 nm,激光发射角为 0.5 mrad,脉冲发射频率为 820 kHz,平均点云密度为 2.9 pts/m^2,飞行覆盖面积约为 55 km^2。该系统还搭载了3×10^7 像素的电荷耦合器件(CCD)相机以获取该研究区的航空影像,影像的空间分辨率为 0.2 m。

1.4 LiDAR 数据预处理

对激光雷达点云数据进行去噪后采用布料模拟滤波(Cloth simulation filter,CSF)法将点云数据分类为地面点和非地面点,将地面点通过不规则三角网(Triangulated irregular network,TIN)插值生成空间分辨率为 0.5 m 的数字高程模型(Digital elevation models,DEM),以此生成归一化点云数据。

1.5 特征变量提取

目前较多的研究者利用归一化点云数据提取高度变量、密度变量和强度变量并将其用于森林参数估测。本研究共提取了 35 个特征变量,其中包括 24 个高度变量、10 个密度变量和郁闭度,各点云特征变量描述如表 2 所示。

表 2 机载激光雷达数据中提取的特征变量指标

Table 2 Feature variable indices extracted from airborne laser radar data

点云变量符号	变量描述
$H_{min}/H_{max}/H_{mean}/H_{mad}$	表示每个统计单元内所有点的最低高度/最高高度/平均高度/高度中位数
H_{ccr}	冠层起伏率能表征冠层的表面粗糙程度,其计算公式为 $H_{ccr}=(H_{mean}-H_{min})/(H_{max}-H_{min})$
$H_{var}/H_{stdv}/H_{skew}/H_{kurt}/H_{cv}$	表示每个统计单元内所有点高度的方差/标准差/偏态/峭度/变异系数
$H_{10}/H_{20}/H_{25}/H_{30}/H_{40}/H_{50}/H_{60}/H_{70}/H_{75}/H_{80}/H_{90}/H_{95}/H_{99}$	点云累计分位数高度(3 个),是将归一化后的点云数据按高度排序,统计累计百分位数点云所在的高度
H_{iq}	表示高度四分位数的间距,计算公式为 $H_{iq}=H_{75}-H_{25}$
$D_0/D_1/D_2/D_3/D_4/D_5/D_6/D_7/D_8/D_9$	密度变量(0~9,共 10 个):按高度将点云分成相同高度的 10 层,每层表示其所对应的密度变量
CC	郁闭度

1.6 特征变量筛选

一个好的特征选择技术可使机器学习集中在最重要的特征上,能有效提升运算效率和模型精度,从而避免大量冗余特征的输入对机器学习的性能和结果造成影响。本研究使用支持向量机-递归特征消除法、LightGBM 算法以及主成分分析法对点云特征变量进行筛选,研究不同特征筛选方法对林分平均高估测结果的影响。

1.6.1 *支持向量机-递归特征消除法* 支持向量机-递归特征消除法(Support vector machine-recursive feature elimination,SVM-RFE),利用 SVM 的最大间隔原理,先训练样本再按特征贡献进行排序,每次迭代中去掉得分最低的特征,直至选出所需特征数[19]。

1.6.2 LightGBM LightGBM(Light gradient boosting machine)是一种基于决策树的快速、分布式、高性能的梯度下降树(GBDT)框架,是 GBDT 算法的改进[20]。与 GBDT 算法相比,LightGBM 算法具有训练效率更高、内存占用率更低、准确率更高、能够处理大规模数据等优点。

1.6.3 主成分分析法 主成分分析法(Principal component analysis,PCA)利用降维的思想通过正交变换将高维数据映射到低维空间中,在保留高维数据重要特征的同时又能去除不重要的特征或噪声。

1.7 林分平均高估测模型的建立

1.7.1 随机森林回归算法 随机森林回归算法(Random forest regression,RFR)是由 Breiman[21]提出的,利用 bootstrap 方法生成 m 个训练集,然后对于每个训练集构造一棵决策树,并在特征中随机抽取一部分特征作为最优解,应用于节点。选择重要性累计贡献率大于 0.85 的特征变量建模,并通过网格搜索法和十折交叉验证法确定决策树的数量、最小特征数及最小叶子节点。

1.7.2 K 近邻回归算法 K 近邻(K-nearest neighbor,KNN)算法常被用于大面积森林参数反演和制图。KNN 是通过测量不同样本在特征空间中的距离来进行分类的[22]。基本思路是:在某一样本的特征空间中如果有 K 个样本与之相似,那么这些样本大多数属于同一个类别,则该样本也属于这个类别。本研究采用欧氏距离作为激励度量方法。

1.7.3 线性回归方程 通过逐步回归法构建线性回归方程预测林分平均高,在回归建模过程中,每一步引入的新变量都要进行 F 检验并达到显著水平($P<0.05$)。

1.8 模型精度评价

采用决定系数(R^2)、均方根误差($RMSE$)和相对均方根误差($rRMSE$)作为指标评价回归模型的精度。R^2、$RMSE$ 和 $rRMSE$ 计算方法分别见公式(2)、公式(3)和公式(4):

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

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \tag{3}$$

$$rRMSE=\frac{RMSE}{\bar{y}_i}\times 100\% \tag{4}$$

式中,y_i为样地实测值;n 为样地个数;$\hat{y}_i$为预测值;$\bar{y}_i$为样本平均值。

2 结果与分析

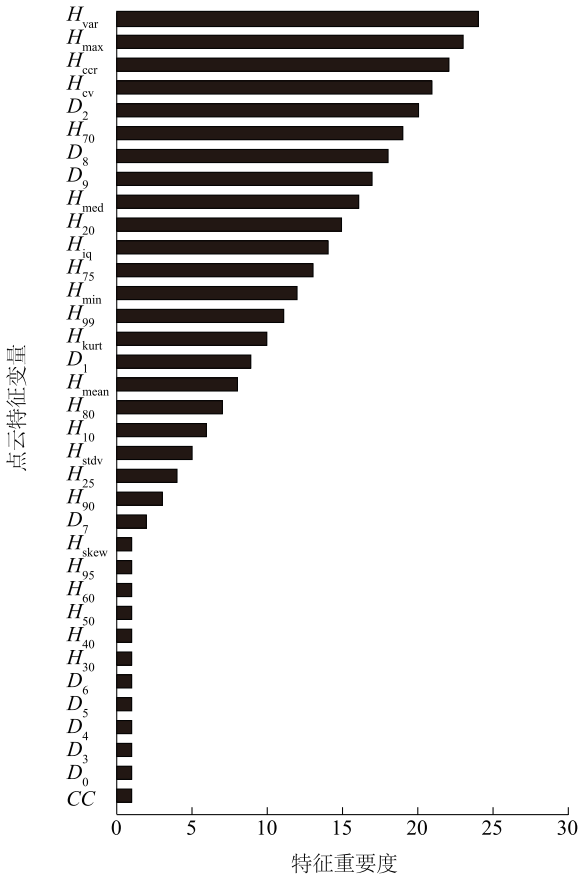
2.1 特征筛选及重要性评估

2.1.1 支持向量机-递归特征消除法特征筛选 利用 SVM-RFE 对所提取的 35 个特征进行特征重要性排序,排序结果如图 1 所示。在排序准则中的特征得分越小,其特征的重要性也就越高。因此,本研究选取特征的重要性为 1 的 12 个变量参与建模,分别为 CC、D_0、D_3、D_4、D_5、D_6、H_{30}、H_{40}、H_{50}、H_{60}、H_{95}、H_{skew},其中包括 5 个密度变量、6 个高度变量以及郁闭度。

2.1.2 LightGBM 特征筛选 基于 LightGBM 算法对所提取的 35 个特征进行特征重要性排序,排序结果如图 2 所示。在排序准则中重要度越大所对应特征的重要性也就越大,图 2 中前 16 个特征的重要度不为 0,认为其特征重要性高于重要度为 0 的其他特征,同时有 4 个特征重要度为 1 的变量也认为其特征重要性较低,故将其剔除。因此,本研究选取特征重要度大于 1 的 12 个特征变量参与建模,其中包括 7 个高度变量(H_{cv}、H_{max}、H_{90}、H_{iq}、H_{99}、H_{kurt}、H_{stdv})和 5 个密度变量(D_5、D_4、D_9、D_3、D_6)。

2.1.3 主成分分析法特征筛选 PCA 法依据贡献率高的前几个主成分与原始点云特征变量的相关关系筛选出每个主成分中相关性最高的变量参与建模,当累计贡献率大于 85%时就认为少数几个变量就足以解释原始数据的大部分信息。本研究结果显示,前 5 个主成分的累计贡献率达到了 89%,说明前 5 个主成分足以反映原始数据的大部分信息。利用主成分分析得到的各主成分与点云特征变量的相关关系如表 3 所示。

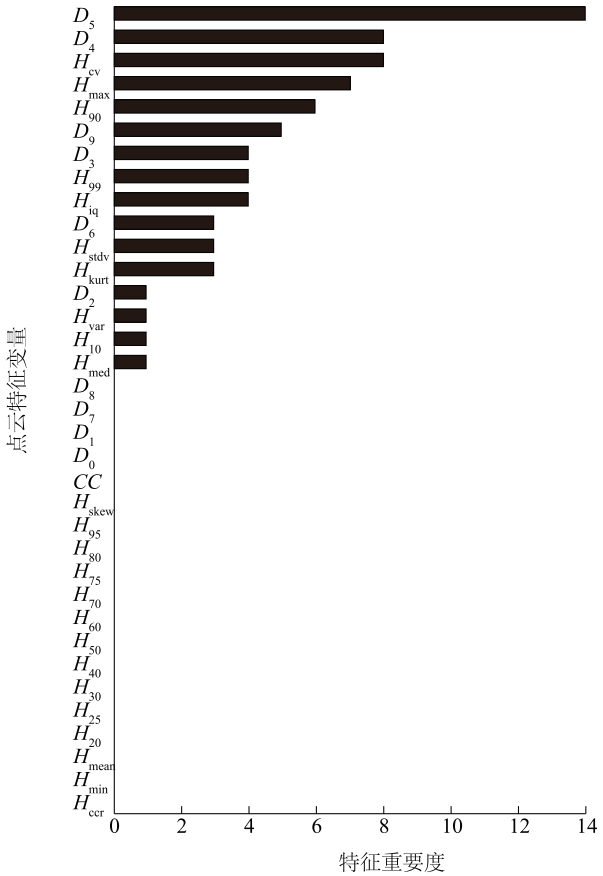
与第 1 主成分的相关性最高的变量为 H_{30} 和 H_{40},相关系数为 0.982;与第 2 主成分的相关性最高的是 H_{cv},相关系数为 0.943;与第 3 主成分的相关性最高的是 D_4,相关系数为 0.687;与第 4 主成分的相关性最高的是 D_9,相关系数为 0.463;与第 5 主成分的相关性最高的是 H_{min},相关系数为 0.834。故 H_{40}、H_{30}、H_{cv}、D_4、D_9和 H_{min}能够解释原始变量的绝大部分信息,将其作为回归模型的自变量。



H_{var}、H_{stdv}、H_{skew}、H_{min}、H_{med}、H_{mean}、H_{max}、H_{kurt}、H_{iq}、H_{cv}、H_{ccr}、H_{10}、H_{20}、H_{25}、H_{30}、H_{40}、H_{50}、H_{60}、H_{70}、H_{75}、H_{80}、H_{90}、H_{95}、H_{99}、D_0～D_9、CC 见表 2。

图 1 基于支持向量机-递归特征消除法(SVM-RFE)的点云变量特征重要性排序

Fig.1 Feature importance ranking of point cloud variables based on support vector machine-recursive feature elimination (SVM-RFE)



H_{var}、H_{stdv}、H_{skew}、H_{min}、H_{med}、H_{mean}、H_{max}、H_{kurt}、H_{iq}、H_{cv}、H_{ccr}、H_{10}、H_{20}、H_{25}、H_{30}、H_{40}、H_{50}、H_{60}、H_{70}、H_{75}、H_{80}、H_{90}、H_{95}、H_{99}、D_0～D_9、CC 见表 2。

图 2 基于 LightGBM 的点云变量特征重要性排序

Fig.2 Feature importance ranking of point cloud variables based on light gradient boosting machine (LightGBM)

2.2 林分平均高估测模型建立及精度评价

通过构建的 RFR、KNN 和 LR 3 种模型反演林分平均高,根据模型精度评价的指标(R^2、*RMSE* 和 *rRMSE*),对不同特征筛选方法所构建的 RFR、KNN 和 LR 模型进行精度评价,并做对比分析,其中建模精度与检验精度之差能有效反映不同估测模型间的泛化能力。具体模型评价指标对比见表 4。

随机森林算法所生成的决策树数目为 200,树节点分裂最小特征数为 3,最小叶节点为 2;KNN 模型以欧氏距离作为激励度量方法,当 K 值为 11 时,KNN 模型的误差最小;逐步回归方程为 $LorH=3.278+0.676H_{90}+8.661D_0$,式中 *LorH* 代表林分的胸高断面积加权平均高。

其中基于 LightGBM 进行特征筛选构建的 KNN 模型比较稳定,精度最高,为高峰林场林分平均高反演的最优模型,建模结果的 R^2 为 0.83,*RMSE* 为 1.64 m,*rRMSE* 为 12.04%;基于 PCA 进行特征筛选构建的 KNN 模型精度最低,R^2 为 0.53,*RMSE* 为 2.40 m,*rRMSE* 为 18.36%。在 3 种特征筛选方法中基于 LightGBM 进行特征筛选所构建的估测模型较为稳健,其泛化能力最强。而基于 SVM-RFE 特征筛选所构建的 RFR 和 KNN 模型,建模精度与验证精度相差很多,说明模型不够稳健,泛化能力差。利用 PCA 对建模因子进行降维后,其精度反而低于不降维的模型精度,且模型出现欠拟合。

除 PCA 方法外,进行特征筛选后所构建的估测模型精度均高于全部因子参与建模的模型精度,表

明通过对建模因子进行特征筛选能有效提升模型精　　度和运算效率。

表 3　各主成分与点云变量的相关关系

Table 3　Correlations between principle components and point cloud variables

点云变量	主成分 1	主成分 2	主成分 3	主成分 4	主成分 5	点云变量	主成分 1	主成分 2	主成分 3	主成分 4	主成分 5
H_{ccr}	0.818	−0.515	−0.093	0.055	−0.069	H_{skew}	−0.731	0.562	0.170	0.011	0.167
H_{cv}	−0.060	0.943	0.061	−0.157	−0.065	H_{stdv}	0.707	0.671	0.108	0.006	−0.116
H_{iq}	0.189	0.745	0.303	−0.297	0.186	H_{var}	0.616	0.708	0.053	0.092	−0.119
H_{kurt}	0.117	−0.527	−0.564	−0.152	0.241	CC	0.193	−0.258	0.511	0.306	0.009
H_{max}	0.908	0.259	0.248	−0.094	0.040	D_0	−0.483	0.649	−0.312	−0.212	0.137
H_{min}	0.140	−0.200	0.232	0.307	0.834	D_1	−0.584	0.668	−0.232	0.033	0.154
H_{mean}	0.981	−0.141	0.042	0.028	0.042	D_2	−0.639	0.460	0.073	0.261	−0.018
H_{med}	0.944	−0.218	−0.008	0.131	−0.031	D_3	−0.666	0.106	0.460	0.366	−0.164
H_{10}	0.872	−0.384	−0.111	0.061	0.045	D_4	−0.450	−0.342	0.687	0.126	−0.163
H_{20}	0.947	−0.202	−0.058	0.121	−0.035	D_5	−0.006	−0.626	0.655	−0.229	0.017
H_{25}	0.971	−0.080	−0.035	0.106	−0.047	D_6	0.404	−0.681	0.299	−0.417	0.074
H_{30}	0.982	0.003	−0.021	0.055	−0.037	D_7	0.716	−0.381	−0.333	−0.162	−0.015
H_{40}	0.982	0.085	0.017	0.016	−0.030	D_8	0.757	−0.002	−0.524	0.191	−0.142
H_{50}	0.976	0.161	0.062	0.003	−0.014	D_9	0.559	0.195	−0.307	0.463	0.072
H_{60}	0.966	0.208	0.092	−0.005	0.000	H_{90}	0.930	0.299	0.160	−0.057	0.045
H_{70}	0.958	0.250	0.111	−0.025	0.022	H_{95}	0.930	0.299	0.160	−0.057	0.045
H_{75}	0.947	0.280	0.112	−0.046	0.046	H_{99}	0.917	0.291	0.205	−0.060	0.038
H_{80}	0.943	0.285	0.130	−0.050	0.048						

H_{var}、H_{stdv}、H_{skew}、H_{min}、H_{med}、H_{mean}、H_{max}、H_{kurt}、H_{iq}、H_{cv}、H_{ccr}、H_{10}、H_{20}、H_{25}、H_{30}、H_{40}、H_{50}、H_{60}、H_{70}、H_{75}、H_{80}、H_{90}、H_{95}、H_{99}、D_0~D_9、CC 见表 2。

表 4　林分平均高模型建模精度和检验精度及其差值

Table 4　Modeling accuracy and testing accuracy of average forest stand height model and their difference

模型	特征筛选方法	建模精度			验证精度			建模精度与验证精度之差		
		R^2	*RMSE*(m)	*rRMSE*(%)	R^2	*RMSE*(m)	*rRMSE*(%)	R^2	*RMSE*(m)	*rRMSE*(%)
RFR	全部特征	0.78	1.90	14.07	0.71	2.04	15.05	0.07	0.14	−0.98
	SVM-RFE	0.81	1.76	13.04	0.68	2.04	15.00	0.13	0.28	−1.96
	LightGBM	0.80	1.80	13.33	0.72	1.96	14.48	0.08	0.16	−1.15
	PCA	0.73	2.05	15.14	0.73	1.51	11.14	0.01	0.54	4.00
KNN	全部特征	0.79	1.86	13.69	0.78	1.59	11.76	0.01	0.27	1.93
	SVM-RFE	0.79	1.81	13.35	0.73	1.75	12.95	0.06	0.06	0.40
	LightGBM	0.83	1.64	12.04	0.81	1.56	11.54	0.02	0.08	0.50
	PCA	0.53	2.40	18.36	0.57	1.93	14.84	0.04	0.47	3.52
LR	SR	0.69	2.22	16.33	0.57	2.70	20.21	0.12	0.48	−3.88

R^2:决定系数;*RMSE*:均方根误差;*rRMSE*:相对均方根误差。

2.3 林分平均高反演

利用基于 LightGBM 进行特征筛选所建立的 KNN 模型对高峰林场林分平均高进行反演。将反演的林分平均高分为 6 个等级，Ⅰ级为2.0~6.0 m，Ⅱ级为6.1~10.0 m，Ⅲ级为10.1~14.0 m，Ⅳ级为14.1~18.0 m，Ⅴ级为18.1~22.0 m，Ⅵ级为 22.0 m 以上。

高峰林场大部分区域林分平均高等级较低，林分平均高在Ⅰ、Ⅱ、Ⅲ级的相对较多，这是因为林场主要以木材生产作为经营项目，林分以人工纯林为主，林种结构单一，导致高度等级低的林分内部高度差异小，分布较为集中。

3 结论与讨论

本研究以中国南亚热带季风区域的人工林为研究对象，以机载激光雷达点云数据为数据源，在研究区内选取 105 个典型样地(2 种森林类型)作为验证数据，分别采用 PCA、支持向量机-递归特征筛选和 LightGBM 对提取的 35 个点云特征变量进行建模因子筛选，并结合 KNN、RFR 和 LR 对林分平均高进行反演，结果表明，对建模因子进行特征筛选能有效提升模型精度和计算效率，除 PCA 外其他 2 种基于特征筛选所建立的估测模型精度均高于所有特征参与建模的模型精度。基于 LightGBM 进行特征筛选所构建的估测模型均能得到较好的估测结果，其泛化能力优于其他特征筛选方法。其中，基于 LightGBM 进行特征筛选所构建的 KNN 模型建模精度和检验精度均最高。

以往的研究大多是比较某一种非参数模型与参数模型(如逐步回归)的精度差异，如曾伟生等[9]考虑到模型的实用性和可解释性，采用简单的线性与非线性模型构建估测模型，虽然研究结果表明，基于点云高度变量(中位数)与强度变量(75%分位数)所构建的二元非线性模型就能达到较为理想的估测结果，但也忽略了其他点云变量所包含的有用信息。此外，多数研究结果表明，非参数模型精度优于参数模型；周蓉等[12]和沈剑波等[13]比较了不同训练算法对 BP 神经网络所建立的林分平均高估测模型精度的影响并与逐步回归算法进行比较，但未考虑其他特征筛选方法和估测模型的差异对精度的影响。而比较多种特征筛选方法与多种模型的组合对模型精度影响的研究还较少，此外 LightGBM 算法虽然在其他领域表现出较强的预测能力以及数据挖掘能力，但在林分平均高建模及对点云变量筛选方面的研究还较少，本研究使用的不同特征筛选方法和多种模型结合的方法为林分平均高建模提供了参考。

目前，机器学习[23-25]被广泛应用于森林参数估测，通过对变量进行筛选可以有效降低自变量的维度，从而提高模型精度和运算效率。但用随机森林法和支持向量机法等机器学习模型进行森林参数估测时具有随机性，此外机器学习的拟合精度取决于训练样本和参数的设置，如果训练样本选取不合理或者参数设置不当，将会使机器学习模型过拟合或欠拟合。在今后的研究中将考虑用诸如遗传算法、粒子群算法等来对模型进行优化。对不同森林类型、不同树种及不同郁闭度下所提取的森林参数对模型估测精度的影响机理有待进一步研究。

参考文献：

[1] 董立新.林分平均高度卫星遥感新进展[J].遥感技术与应用，2016，31(5)：833-845.

[2] ZHAO F，GUO Q，KELLY M. Allometric equation choice impacts lidar-based forest biomass estimates：a case study from the Sierra National Forest，CA[J]. Agricultural and Forest Meteorology，2012，165：64-72.

[3] 刘亚男.基于多源遥感数据的森林地上生物量及净初级生产力估算研究[J].测绘学报，2020，49(12)：1641.

[4] GREGOIRE T G，NÆSSET E，MCROBERTS R E，et al. Statistical rigor in LiDAR-assisted estimation of aboveground forest biomass[J]. Remote Sensing of Environment，2016，173：98-108.

[5] PEDERSEN R Ø，BOLLANDSÅS O M，GOBAKKEN T，et al. Deriving individual tree competition indices from airborne laser scanning[J]. Forest Ecology and Management，2012，280：150-165.

[6] 刘鲁霞，庞 勇.机载激光雷达和地基激光雷达林业应用现状[J].世界林业研究，2014，27(1)：49-56.

[7] 穆喜云，张秋良，刘清旺，等. 基于机载 LiDAR 数据的林分平均高及郁闭度反演[J]. 东北林业大学学报，2015，43(9)：84-89.

[8] 彭 涛，邢艳秋，尤号田，等.LiDAR 采样形状及采样尺度对林分均高估测的影响[J].中南林业科技大学学报，2018，38(4)：27-32.

[9] 曾伟生，孙乡楠，王六如，等. 基于机载激光雷达数据估计林分蓄积量及平均高和断面积[J].林业资源管理，2020(2)：79-86.

[10] 焦义涛，邢艳秋，霍 达，等. 基于机载 LiDAR 点云估测林分的平均树高[J]. 西北林学院学报，2015，30(3)：170-174.

[11] 赵 勋，岳彩荣，李春干，等. 基于机载 LiDAR 数据估测林分平均高[J]. 林业科学研究，2020，33(4)：59-66.

[12] 周 蓉，赵天忠，吴发云.依据 BP 神经网络的机载 LiDAR 数据

估算林分平均高[J].东北林业大学学报,2021,49(9):60-66.
[13] 沈剑波,雷相东,李玉堂,等.基于 BP 神经网络的长白落叶松人工林林分平均高预测[J].南京林业大学学报(自然科学版),2018,42(2):147-154.
[14] GRACE Y,HE W Q,CARROLL R J. Feature screening with large scale and high dimensional survival data.[J]. Biometrics, 2021, 77(2):1-14.
[15] 曹 林,代劲松,徐建新,等. 基于机载小光斑 LiDAR 技术的亚热带森林参数信息优化提取[J]. 北京林业大学学报, 2014, 36(5): 13-21.
[16] PAOLO M, VIBRANS A C, MCROBERTS R E, et al. Methods for variable selection in LiDAR-assisted forest inventories[J]. Forestry,2017, 90: 112-124.
[17] 郝红科. 基于机载激光雷达的森林参数反演研究[D]. 杨凌:西北农林科技大学, 2019.
[18] 莫莉婕.国有林场森林资源可持续发展对策研究——以高峰林场为例[J].企业科技与发展,2021(4):222-224.
[19] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1): 389-422.
[20] KE G L, MENG Q, FINLEY T, et al. LightGBM: a highly efficient gradient boosting decision tree[J]. Advances in Neural Information Processing Systems, 2017, 30: 3146-3154.
[21] BREIMAN L. Random forests[J].Machine Learning,2001,45 (1):5-32.
[22] ACAR L. Some examples for the decentralized receding horizon control[C]. Tucson: IEEE, 1992: 1356-1359.
[23] 郝 玲,张 佩,史逸民,等. 基于机器学习的江苏省冬小麦气象产量客观区划及歉年预测[J].江苏农业科学,2021,49(12):162-168.
[24] 白 婷,丁建丽,王敬哲. 基于机器学习算法的土壤有机质质量比估算[J].排灌机械工程学报,2020,38(8):829-834.
[25] 赵献立,王志明. 机器学习算法在农业机器视觉系统中的应用[J].江苏农业科学,2020,48(12):226-231.

(责任编辑:陈海霞)