doi:10.11733/j.issn.1007-0435.2021.08.019

基于随机森林的高寒草地地上生物量高光谱估算

高宏元[1]，侯蒙京[1]，葛 静[1]，包旭莹[1]，李元春[1]，刘 洁[1]，冯琦胜[1]，梁天刚[1*]，贺金生[1,2]，钱大文[3]

（1. 兰州大学草地农业科技学院，兰州大学草地农业生态系统国家重点实验室，兰州大学农业农村部草牧业创新重点实验室，兰州大学草地农业教育部工程研究中心，甘肃 兰州 730020；2. 北京大学城市与环境学院，北京 100871；3. 中国科学院西北高原生物研究所，青海 西宁 810008）

摘要：草地地上生物量（Aboveground biomass，AGB）是衡量草地生产力的关键因素，准确测定草地 AGB 具有重要意义。高光谱因具有时效性强、不破坏草地等特点被广泛用于草地生理生态指标的测定。本研究提取和计算了海北试验站高寒草地冠层的原始光谱（Original spectrum，OR）反射率、一阶微分光谱（First derivative spectrum，FD）反射率、光谱位置面积参数（Spectral parameters of spectral position and area，PA）和植被指数（Vegetation indices，VI）4 种不同类型的特征变量，使用连续投影算法（Successive projections algorithm，SPA）和递归特征消除算法（Recursive feature elimination，RFE）进行特征选择，采用随机森林算法（Random forest，RF）构建草地 AGB 估测模型。结果表明：在由 4 种特征变量分别构建的草地 AGB 估测模型中，基于 VI 的 RF 模型精度最高（测试集 $R^2=0.70$，$RMSE=557.87\ kg\cdot ha^{-1}$），实测 AGB 与估测 AGB 的线性 R^2 达到 0.72；不同类型特征变量组合构建的草地 AGB 估测模型中，PA+VI 组合的 RF 模型精度最高（$R^2=0.71$，$RMSE=548.97\ kg\cdot ha^{-1}$），实测 AGB 和估测 AGB 的线性 R^2 达到 0.73。

关键词：高寒草地；地上生物量；高光谱；随机森林；连续投影算法；递归特征消除

中图分类号：S11 **文献标识码**：A **文章编号**：1007-0435(2021)08-1757-12

Hyperspectral Estimation of Aboveground Biomass of Alpine Grassland based on Random Forest Algorithm

GAO Hong-yuan[1], HOU Meng-jing[1], GE Jing[1], BAO Xu-ying[1], LI Yuan-chun[1], LIU Jie[1], FENG Qi-sheng[1], LIANG Tian-gang[1*], HE Jin-sheng[1,2], QIAN Da-wen[3]

(1. College of Pastoral Agriculture Science and Technology, Lanzhou University; State Key Laboratory of Grassland Agro-ecosystem; Key Laboratory of Grassland Livestock Industry Innovation, Ministry of Agriculture and Rural Affairs; Engineering Research Center of Grassland Industry, Ministry of Education, Lanzhou, Gansu Province 730020, China; 2. College of Urban and Environmental Science, Peking University, Beijing 100871, China; 3. Northest Institute of Plateau Biology, Chinese Academy of Science, Xining, Qinghai Province 810008, China)

Abstract: Aboveground biomass (AGB) is a key indicator of grassland productivity and it is important to measure grassland AGB accurately in grassland resource survey. Hyperspectrum is an effective method to measure the physiological and ecological indexes of grassland without physical damage to the grassland. In this study, the original spectrum (OR), the first derivative spectrum (FD), spectral parameters of spectral position and area (PA) and the vegetation index (VI) of alpine grassland canopy were calculated near Haibei National Field Research Station for Alpine Grassland Ecosystem (Haibei Station). Based on the above variables, feature selection was performed with successive projections algorithm (SPA) and recursive feature elimination algorithm (RFE), and the model was constructed with random forest algorithm (RF). The results showed that among the grassland AGB estimation models constructed by different variables, the ac-

收稿日期：2021-04-13；**修回日期**：2021-05-19

基金项目：国家重点研发计划（2019YFC0507701）；国家自然科学基金（31672484，41805086，41801191）；中国工程院咨询研究项目（2021-HZ-5，2020-XZ-29）；兰州大学中央高校基本科研业务费专项资金（lzujbky-2021-kb13）；财政部和农业农村部：国家现代农业产业技术体系资助

作者简介：高宏元（1996-），男，汉族，甘肃静宁人，硕士研究生，主要从事草地遥感与地理信息系统，E-mail：gaohy2015@lzu.edu.cn；* 通信作者 Author for Correspondence，E-mail：tgliang@lzu.edu.cn

curacy of RF model based on VI was the highest ($R^2=0.70$, $RMSE=557.87$ kg·ha^{-1}), and R^2 of measured AGB and predicted AGB was 0.72. Among the grassland AGB estimation models constructed by different combination of variables, the accuracy of RF model of PA + VI combination was the highest ($R^2=0.71$, $RMSE=548.97$ kg·ha^{-1}), and R^2 of measured AGB and predicted AGB was 0.73.

Key words: Alpine grassland; Aboveground biomass; Hyperspectral; Random forest; Successive projections algorithm; Recursive feature elimination

草地地上生物量(Above ground biomass, AGB)是指某一时刻单位面积内的草地植物在地上部分的有机物的总量[1],它不仅是衡量草地群落生长状况与生产力水平的关键参数,还是表征群落能量流动和物质循环的重要指标之一[2]。在高寒地区,草地是家畜的主要食物来源与能量供给[3],草地AGB变化还能反映草地的退化与土壤的侵蚀程度[4]。因此,测定高寒草地AGB十分有必要。传统的草地AGB测定通常采用实地收获法,耗时费力,时效性较差,且对草地破坏性大。

高光谱遥感能在不直接接触目标物体的同时,获得其丰富的光谱信息[5]。对草地高光谱数据进行挖掘和分析,可以获得植被的物理化学组分和生理生态情况等指标,这使实时监测草地植被成为了可能,如纪童等[6]用高光谱数据实现了草坪草叶绿素含量的监测,高金龙等[7]利用高光谱数据构建了高寒天然草地氮、磷养分的估测模型,王磊[8]用高光谱反演了草地的叶面积指数,韩万强等[9]用高光谱数据识别了3种草地主要植物等。对于草地AGB的高光谱监测,目前也有不少相关研究。胥慧等[10]的研究表明基于光谱红谷吸收深度D和光谱绿峰反射高度H的高光谱特征参数(D−H)/(D+H)和草地生物量有较高的相关性;夏浪等[11]使用NDVI反演生物量并得到较高的模型精度;马维维[12]用高光谱卫星数据提取NDVI等5个植被指数用于反演草地生物量。这些研究大多直接以植被指数为建模特征,少有考虑原始光谱特征及其转换参数对模型的影响。张凯等[13]筛选了甘南地区草地冠层的6个光谱特征变量并分别进行AGB的线性、对数等模型的构建;安海波等[14]用不同植被指数对内蒙古天然和人工草地进行了AGB指数、对数等的非线性回归建模。上述研究构建的AGB反演模型大多为单因素或多因素参数统计模型,更注重数据的空间分布,这在研究样本较少且草地类型均一时是有效的,当样本较多且研究区情况复杂时模型的预测能力存在不稳定的问题[15]。因此,在进行草地AGB的高光谱研究时,建模特征的选择和模型算法是影响AGB估算模型的关键因素,在模型简化和稳定性方面有重要意义。

目前,用于高光谱研究的算法较多,连续投影算法(Successive projections algorithm, SPA)由于其良好的光谱冗余信息消除能力,常用于连续光谱数据的选择,杨晨波等人[16]研究表明SPA可以极大减小原始光谱数据的维度,从而简化模型;递归特征消除算法(Recursive feature elimination, RFE)是一种用于筛选最优特征子集的贪心算法,一般与机器学习结合使用,在降低数据维度的同时可找到精度最高的模型[17-18];随机森林(Random forest, RF)是一种决策树集成算法,具有不易过拟合和普适性广的优点[19],在植被生理生态等指标的研究中,常用于估测模型的构建[20-22]。综上,本研究基于SPA和RFE特征选择方法和RF模型,开展高寒地区草地AGB的高光谱研究,以期对高寒地区的放牧强度、草畜平衡和生态环境实时监测提供科学依据。

1 材料与方法

1.1 试验设计

本研究的试验区位于青海海北高寒草地生态国家野外科学观测研究站(简称海北站,37°37′ N, 101°19′ E)附近(图1),坐落于青藏高原东北隅,草地类型属于典型的高寒草地,年平均气温−1.7℃,年降水量范围为426~860 mm,植被类型是以金露梅(*Potentilla fruticosa*)为建群种的高寒灌丛草甸和以嵩草属(*Kobresia*)植物为建群种的高寒嵩草草甸[23]。试验区A,B是两块典型的高寒草地,每块试验地有15个小区(每个小区面积0.2 ha),设有禁牧(CK)、轻度放牧(Light, L)、中度放牧(Medium, M)和重度放牧(Heavy, H)4个放牧梯度,对应的放牧强度分别为0头·ha^{-1}(CK)、0.5头·ha^{-1}(L)、1.0头·ha^{-1}(M)和2.0头·ha^{-1}(H)。放牧家畜为牦牛,放牧方式为轮牧,时间为每年7—9月。放牧设置主要模拟了天然草地的复杂情况,能反映本研究构建的AGB模型的普遍适用性。

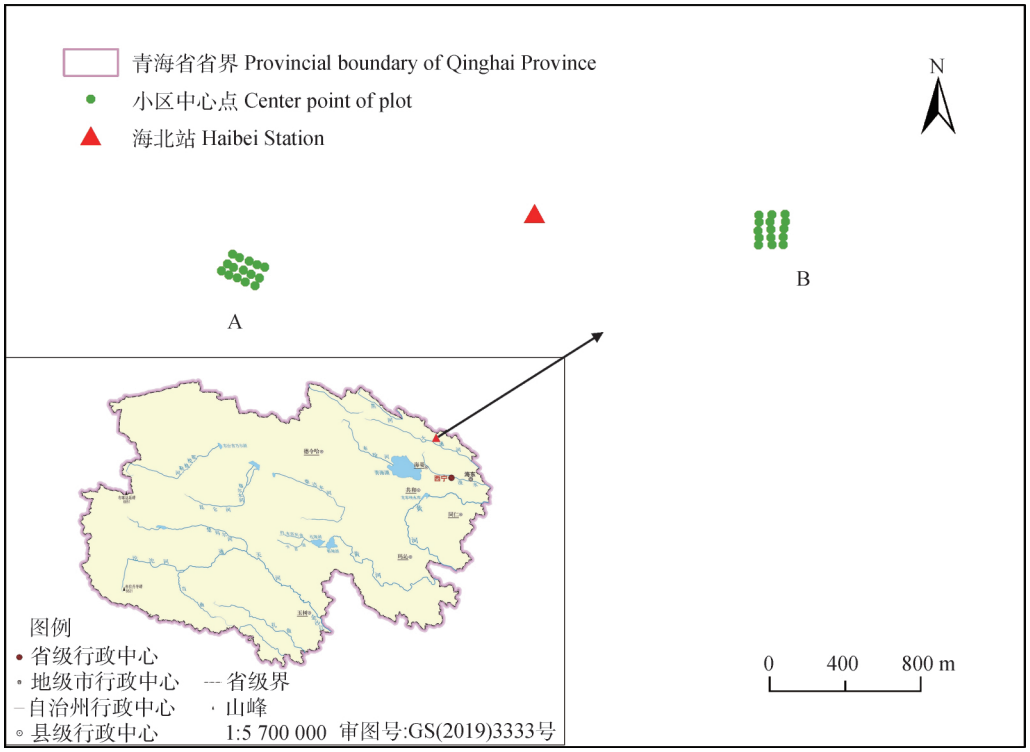

图 1　研究区位置图

Fig. 1　Location of the study area

1.2　数据获取与预处理

草地冠层光谱数据测量采用美国 ASD 公司的双光谱仪系统,该系统由移动端和固定端两台 ASD Field Spec 4 光谱仪组成,其波段范围为 350～2 500 nm。双光谱仪系统通过无线通讯,实时获得反射率数据,并且能自动完成波长交叉校准,可消除不同时间太阳辐射变化带来的误差,因此在多变的天气条件下也可以获得良好的特征光谱图[24]。测量草地冠层光谱要在晴朗少云的天气下进行,测量时探头垂直向下,距冠层高度大约 1 m,每个小区随机选取 3 个测量点,每个测量点测定 9 条光谱曲线,去掉异常光谱曲线,将其余光谱的平均值作为该小区的草地冠层反射率光谱。在小区的每个光谱测量位点放置样方框,齐地面刈割样方内的植物地上部分,带回实验室在 65℃烘箱中烘 48 h 至恒重,得到每个小区的草地 AGB。

由于光谱曲线有一定的噪声,用 Savitzky-Golay 平滑法[25]进行平滑去噪处理,得到原始光谱曲线,进而得到一阶微分光谱。此外,考虑到水分、氧气和仪器敏感性波动等因素的影响,剔除了 1 301～1 450 nm,1 801～2 050 nm 和 2 301～2 500 nm 波段的光谱曲线[26]。本研究在 2019 年 5 月、6 月、7 月、8 月、9 月和 2020 年 7 月、9 月总计开展了 7 次外业调查,在 2020 年的 7 月和 9 月,由于 B 放牧地封锁无法获得实测数据,因此共获得 180 个样本数据,去掉无效或错误数据,共有 179 个样本数据,具体的样本分布和描述性统计如表 1 所示。

表 1　AGB 数据描述性统计结果

Table 1　Descriptive statistical result of AGB data

月份 Month	5	6	7	8	9
样本个数 Number of samples	30	29	45	30	45
平均值 Mean/kg·ha^{-1}	714.41	1 504.16	2 688.99	2 897.95	3 139.91
最小值 Minimum/kg·ha^{-1}	187.20	1 105.20	1 528.53	1 520.00	1 398.93
最大值 Maximum/kg·ha^{-1}	1 220.00	2 172.40	3 934.53	3 739.73	5 237.07
标准误 Standard error	44.74	51.26	93.98	105.23	141.20

1.3　特征变量

本研究的特征变量有原始光谱(Original spectrum,OR)、一阶微分光谱 FD(First derivative spectrum,FD)、光谱位置面积参数(Spectral parameters of spectral position and area,PA)和植被指数(Vegetation indices,VI)4 类(表 2),这些不同类别的特征变量,将用于后续的特征选择和模型构建。

表 2 特征变量及其定义

Table 2 Variables and definition

变量类型 Type of variables	名称 Name	定义或公式 Definition or formula
原始光谱 Original spectrum(OR)	Ri	原始光谱波段反射率 Band reflectance of OR
一阶微分光谱 First derivative spectrum(FD)	R_i'	一阶微分光谱波段反射率 Band reflectance of FD
光谱位置及面积参数[10] Spectral parameters of spectral position and area(PA)	Db	蓝边(430～470 nm)内一阶微分光谱中的最大值 The maximum FD value in the blue edge (430～470 nm)
	Dr	红边(620～760 nm)内一阶微分光谱中的最大值 The maximum FD value in the red edge (620～760 nm)
	Dinr	近红外(780～1 300 nm)内一阶微分光谱中的最大值 The maximum FD value in the near-infrared region (780～1 300 nm)
	H	绿峰反射高度,计算公式为 $1-\{[R_{500}+6/17(R_{670}-R_{500})]/R_{560}\}$ The reflection height of green peak, calculated with $1-\{[R_{500}+6/17(R_{670}-R_{500})]/R_{560}\}$
	D	红谷吸收深度,计算公式为 $1-\{R_{670}/[R_{560}+11/20\ (R_{760}-R_{560})]\}$ The absorption depth of red valley, calculated with $1-\{R_{670}/[R_{560}+11/20\ (R_{760}-R_{560})]\}$
	SDb	蓝边波长范围内一阶微分波段反射率值的总和 The sum of the FD values in the blue edge
	SDr	红边波长范围内一阶微分波段反射率值的总和 The sum of the FD values in the red edge
	SDinr	近红外平台一阶微分波段反射率值的总和 The sum of the FD values in the near-infrared region
	VI_1	D/H
	VI_2	(D－H)/(D＋H)
	VI_3	SDr/SDb
	VI_4	SDinr/SDb
	VI_5	SDinr/SDr
	VI_6	(SDr-SDb)/(SDr＋SDb)
	VI_7	(SDinr-SDb)/(SDinr＋SDb)
	VI_8	(SDinr-SDr)/(SDinr＋SDr)
植被指数[27] Vegetation indices(VI)	CIred-edge	$(R_{800}/R_{720})-1$
	$NDVI_1$	$(R_{800}-R_{680})/(R_{800}+R_{680})$
	$NDVI_2$	$(R_{800}-R_{670})/(R_{800}+R_{670})$
	NDBleaf	$(R_{1540}-R_{2160})/(R_{1540}+R_{2160})$
	NDII	$(R_{850}-R_{1650})/(R_{850}+R_{1650})$
	NDMI	$(R_{1649}-R_{1722})/(R_{1649}+R_{1722})$
	GNDVI	$(R_{801}-R_{550})/(R_{801}+R_{550})$
	SAVI	$1.5\times(R_{800}-R_{670})/(R_{800}-R_{670}+0.5)$
	OSAVI	$(R_{800}-R_{670})/(R_{800}+R_{670}+0.16)$
	CARI	$(R_{700}-R_{670})-0.2\times(R_{700}+R_{670})$
	TCARI	$3\times(R_{700}-R_{670})-0.6\times(R_{700}-R_{550})(R_{700}/R_{670})$
	MCARI	$[(R_{700}-R_{670})-0.2\times(R_{700}-R_{550})]\times(R_{700}/R_{670})$
	HNDVI	$(R_{827}-R_{668})/(R_{827}+R_{668})$
	MTCI	$(R_{754}-R_{709})/(R_{709}-R_{681})$
	PRI	$(R_{531}-R_{570})/(R_{531}+R_{570})$
	SIPI	$(R_{800}-R_{450})/(R_{800}+R_{450})$
	PSNDa	$(R_{800}-R_{680})/(R_{800}+R_{680})$
	PSNDb	$(R_{800}-R_{635})/(R_{800}+R_{635})$
	PSSRa	R_{800}/R_{680}
	PSSRb	R_{800}/R_{635}
	VARIg	$(R_{560}-R_{670})/(R_{560}+R_{670}-R_{450})$
	VARIr	$(R_{700}-1.7\times R_{670}+0.7\times R_{450})/(R_{700}+2.3\times R_{670}-1.3\times R_{450})$
	SR	R_{744}/R_{667}
	TVI	$60\times(R_{800}-R_{550})-100\times(R_{670}-R_{550})$
	GRVI	R_{800}/R_{550}
	MSI	R_{1599}/R_{819}

注:i 表示 350～1 300 nm,1 451～1 800 nm 和 2 051～2 300 nm 范围内的任一波长

Note: i is any wavelength in the range of 350～1 300 nm, 1 451～1 800 nm and 2 051～2 300 nm

1.4　特征选择与建模

本研究先用连续投影算法 SPA 对原始光谱和一阶导光谱数据进行特征波段反射率的提取,再用递归特征消除算法 RFE 对特征波段反射率和其他特征变量进行特征选择,最后用随机森林 RF 算法构建草地 AGB 的反演模型,以上算法均由 Python 编程语言实现。

SPA 是一种使矢量空间共线性最小化的前向变量选择算法,它的优势在于可提取全波段的几个特征波段,能够消除光谱矩阵中冗余的信息[28],因此本研究用 SPA 提取 OR 和 FD 的光谱特征波段,从而降低数据维度,使模型更简单高效。

RFE 是一种寻求最优特征子集的贪心算法[29],基本思想是构建底层模型进行初始特征集的训练,并给每个特征赋予权重,然后去掉权重最小的特征,将其他特征组成新的特征子集,再进行训练,递归重复此过程直至达到最终所需的特征数目。在本研究中,RFE 构建 RF 底层模型时选取默认参数,选取训练结果决定系数 R^2 最大时对应的特征子集作为构建模型的特征组合。

RF 是多棵决策树构成的集成模型,模型的最终输出结果由森林中的每一棵决策树共同决定,在 RF 中以每棵决策树输出的均值为最终结果。RF 算法的具体过程[30]如下:在原始训练特征集中用 Bootstrap 抽样方法获得 n 个特征子集;对每个特征子集选择 m 个特征,并对每个训练特征子集构建决策树,得到 n 个决策树模型,建立起随机森林;计算每棵决策树的结果,将 n 棵决策树输出结果的均值作为最终结果。此外,RF 算法提供了特征重要性(总和为 1)的接口,方便比较建模特征的重要性。经过参数优选和多次训练,确定 RF 的主要参数决策树数量为 1 000。

1.5　模型评价

为了减少训练样本划分偶然性带来的结果误差,本研究采用 10 折交叉验证确保模型的稳定性。在 10 折交叉验证中,试验数据样本被划分为 10 份,轮流将其中 9 份作为训练集,1 份作为测试集,将测试集结果取平均值作为模型最终的评价结果。

本研究采用决定系数(Coefficient of determination,R^2)和均方根误差(Root mean square error,*RMSE*)评价 AGB 估测模型的精度,计算公式如下:

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

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

式中,$\hat{y}_i$ 为 AGB 的模型估测值,y_i 为 AGB 的实测值,$\bar{y}_i$ 为 AGB 实测值的均值,n 为样本数量。决定系数 R^2 反映了模型对实测数据的拟合程度,其范围在 0～1 之间,R^2 越大反映模型对因变量的解释程度越高。*RMSE* 反映了 AGB 实测值和模型估计值之间的离散程度,其值越小说明模型对因变量的预测效果越好。

2　结果与分析

2.1　原始光谱和一阶微分的 SPA 特征波段提取

从 SPA 特征波段提取的结果(图 2)来看,随着 OR 的特征波段数量的增加,*RMSE* 整体呈先减小再平缓最后增大的趋势(图 2a),而波段数量越多模型复杂度越高,因此需要选择适宜的特征波段数量。综合考量,OR 选取 9 个特征波段,此时特征波段较少且 *RMSE* 较小,其特征波段的反射率分布如图 2b 所示,具体特征波段为 R_{350},R_{371},R_{396},R_{749},R_{937},R_{985},R_{1130},R_{1294}和 R_{1603};同理,FD 提取 19 个特征波段(图 2c),分别为 R'_{542},R'_{659},R'_{783},R'_{824},R'_{1061},R'_{1070},R'_{1177},R'_{1201},R'_{1247},R'_{1451},R'_{1463},R'_{1541},R'_{1659},R'_{1690},R'_{1773},R'_{1779},R'_{1789},R'_{1798}和 R'_{2198}(图 2d)。与全波段相比,SPA 分别将 OR 和 FD 的波段数量从 1 551 减少到 9 和 19,波段数量分别减少了 99.4%和 98.8%,这极大地降低了数据维度,从而利于精简模型。

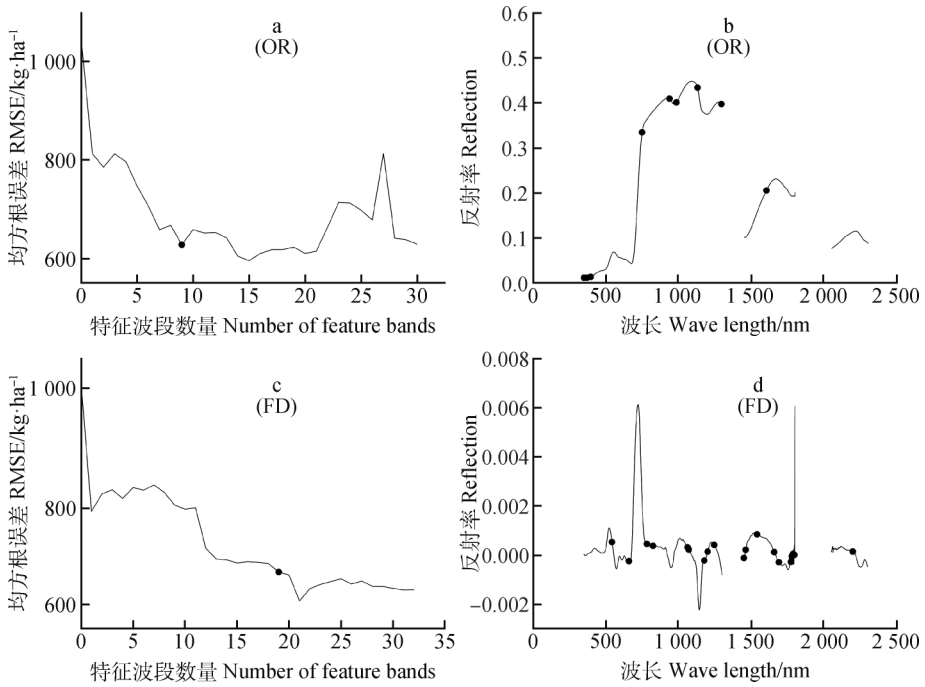

图 2 SPA 特征波段提取及波段分布

Fig. 2 SPA feature bands extraction and bands distribution

2.2 基于不同类型特征变量的 AGB 模型分析

对 OR 特征波段、FD 特征波段、光谱位置面积参数 PA 和植被指数 VI(表 2)4 类特征变量分别先进行 RFE 特征选择，将 RFE 选择出的特征子集作为 RF 的建模特征，再经过 10 折交叉验证，计算得到训练集和测试集的 *RMSE* 和 R^2(表 3)，并利用 RF 自带的特征重要性接口分析建模特征的重要性(图 3)。

从模型精度来看，基于植被指数 VI 构建的 RF 模型精度最高($R^2=0.70$，$RMSE=557.87$ kg·ha^{-1})，对应 7 个建模特征为 NDBleaf，OSAVI，TCARI，MTCI，PRI，SIPI，VARIg；光谱位置面积参数 PA 和一阶光谱特征波段 FD 的模型精度次之，分别为 $R^2=0.64$ ($RMSE=596.42$ kg·ha^{-1})和 $R^2=0.57$ ($RMSE=685.96$ kg·ha^{-1})；精度最差的是原始光谱 OR 特征波段构建的 RF 模型，测试集 R^2 仅为 0.52 ($RMSE=700.06$ kg·ha^{-1})。从 RF 模型构建的特征来看，经过 RFE 特征选择后，4 种模型的特征数量都有所降低，尤其是 VI 构建的 RF 模型，特征数量从 26 减少到 7，在较大降低数据维度的同时，模型精度也最高，其中由 R_{800} 和 R_{450} 构建的结构不敏感色素指数 SIPI 的重要性最高(0.47)，其次是由 R_{531} 和 R_{570} 构建的光化学反射系数 PRI (0.25)，这两个特征为模型贡献了 72%的重要性。

表 3 不同类别特征变量的 RF 估测模型结果

Table 3 Results of RF estimation model based on different class of features ($n=179$)

变量类型 Type of variables	初始特征数量 Number of initial features	RFE 特征数量 Feature numbers after RFE	RF 建模特征 RF modeling feature	训练集 Training set		测试集 Test set	
				RMSE/kg·ha^{-1}	R^2	*RMSE*/kg·ha^{-1}	R^2
OR	9	5	R_{371}，R_{396}，R_{749}，R_{1130}，R_{350}	269.34	0.94	700.06	0.52
FD	19	7	R'_{1541}，R'_{659}，R'_{542}，R'_{783}，R'_{1690}，R'_{824}，R'_{1779}	265.61	0.94	685.96	0.57
PA	16	8	H，Dr，Dinr，SDb，VI_1，VI_2，VI_6，VI_7	228.16	0.96	596.42	0.64
VI	26	7	NDBleaf，OSAVI，TCARI，MTCI，PRI，SIPI，VARIg	217.25	0.96	557.87	0.70

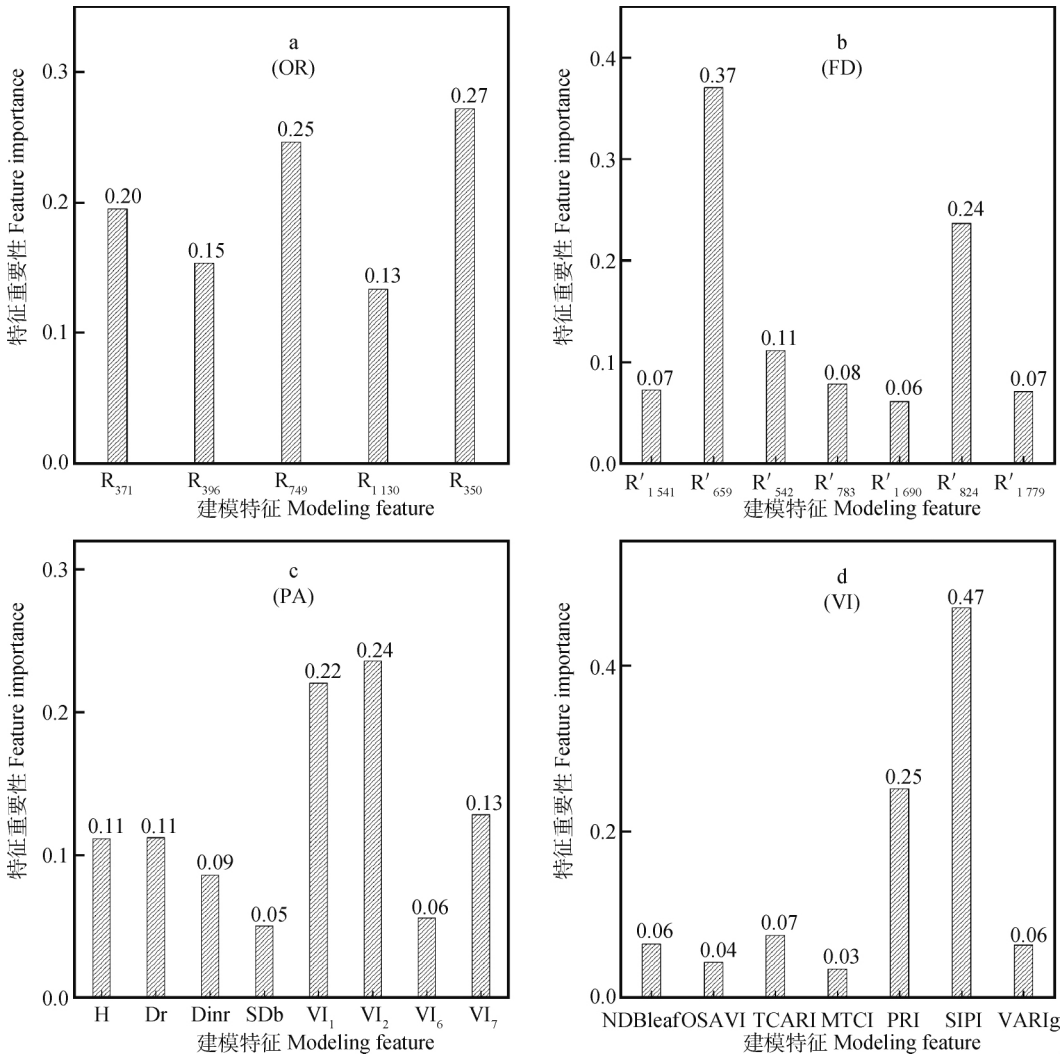

图 3　不同 RF 模型的特征重要性

Fig. 3　The importance of features in different RF models

对全集数据($n=179$)进行实测生物量与估测生物量的相关性分析,结果表明,在 4 类特征变量中,基于植被指数 VI 的 RF 地上生物量反演模型的估测效果最好,其线性决定系数 R^2 达到 0.72,其次是光谱位置面积参数 PA($R^2=0.68$),再次是一阶光谱特征波段 FD($R^2=0.59$)(图 4),模型估测效果最差的是原始光谱特征波段 OR,其实测 AGB 和估测 AGB 的决定系数 R^2 为 0.56,这和测试集的 10 折交叉验证评价结果相一致。

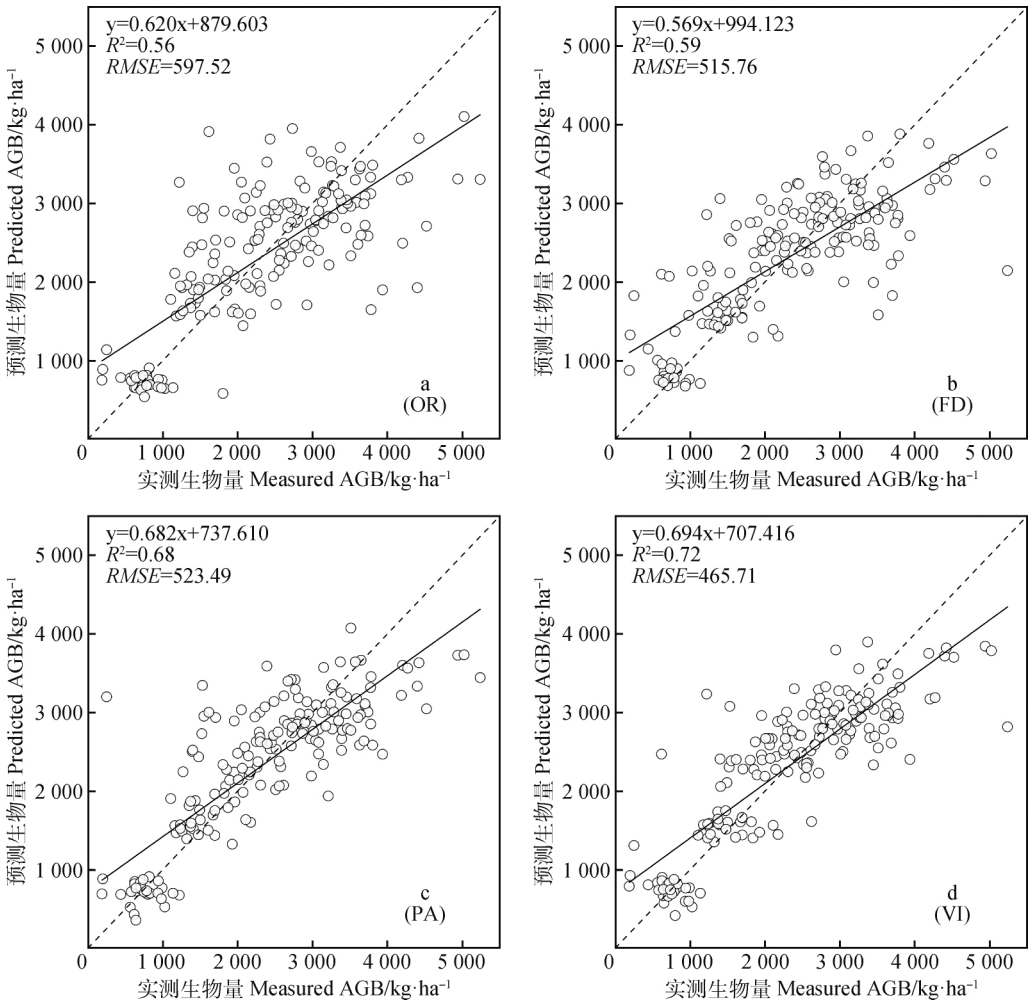

图 4 不同 RF 估测模型的实测 AGB 与估测 AGB 的相关性

Fig. 4 Correlation between measured AGB and predicted AGB in different RF estimation models

2.3 基于不同类型特征变量组合的 AGB 模型分析

进行不同类型特征变量组合时，考虑到 OR 特征波段和 FD 特征波段都属于单波段特征变量，因此可以将其作为整体，与光谱位置面积参数 PA 和植被指数 VI 任意组合，进而进行 RFE 特征选择和 RF 模型构建。表 4 结果显示，在 5 种组合中，PA＋VI 组合的模型精度最高，R^2 达到了 0.71，其对应的 *RMSE* 也最小，为 548.97 kg·ha^{-1}；精度次之的是所有变量构建的模型，拟合系数 R^2 = 0.70（*RMSE* = 562.93 kg·ha^{-1}）；OR＋FD 组合与 OR＋FD＋PA 组合的模型精度居中，R^2 分别为 0.69 和 0.67（*RMSE* 分别为 579.00 和 591.17 kg·ha^{-1}）；OR＋FD＋VI 的模型精度最差（R^2 = 0.64，*RMSE* = 620.05 kg·ha^{-1}）。从模型的建模特征及其重要性结果（图 5）来看，模型精度最高的 PA＋VI 组合，是所有组合中特征数量最少的，仅有 5 个建模特征 H，Dr，VI_2，PRI 和 SIPI，其中特征重要性最高的是 VI_2 和 PRI，分别为 0.37 和 0.30，二者为整个模型贡献了 67%的重要性；模型精度次之的全部变量组合，有 6 个建模特征，其中 PRI 和 VI_1 的重要性最高，分别为 0.30 和 0.26；OR＋FD 组合有 7 个建模特征，最重要的特征是 R_{350}，重要性为 0.24；OR＋FD＋PA 组合和 OR＋FD＋VI 组合的建模特征最多，均为 10 个，重要性最高的特征分别是 VI_2 和 PRI。

表 4　不同类别特征变量组合的 RF 估测模型结果

Table 4　Results of RF estimation model based on different combination of class of features (n=179)

变量类型 Type of variables	初始特征数量 Number of initial features	RFE 特征数量 Feature numbers after RFE	RF 建模特征 RF modeling feature	训练集 Training set		测试集 Test set	
				RMSE/ kg·ha^{-1}	R^2	*RMSE*/ kg·ha^{-1}	R^2
OR+FD	28	7	R_{371},R_{396},R_{749},R_{350},R'_{1247},R'_{824},R'_{1177}	226.45	0.96	579.00	0.69
OR+FD+PA	44	10	R_{371},R_{350},R'_{659},R'_{783},R'_{1773},R'_{824},H,Dr,VI_2,VI_7	228.79	0.96	591.17	0.67
OR+FD+VI	54	10	R_{371},R_{396},R_{350},R'_{542},R'_{783},R'_{1779},R'_{1798},MTCI,PRI,VARIg	233.28	0.95	620.05	0.64
PA+VI	42	5	H,Dr,VI_2,PRI,SIPI	204.85	0.96	548.97	0.71
OR+FD+PA+VI	70	6	R_{371},R_{350},H,Dr,VI1,PRI	214.25	0.96	562.93	0.70

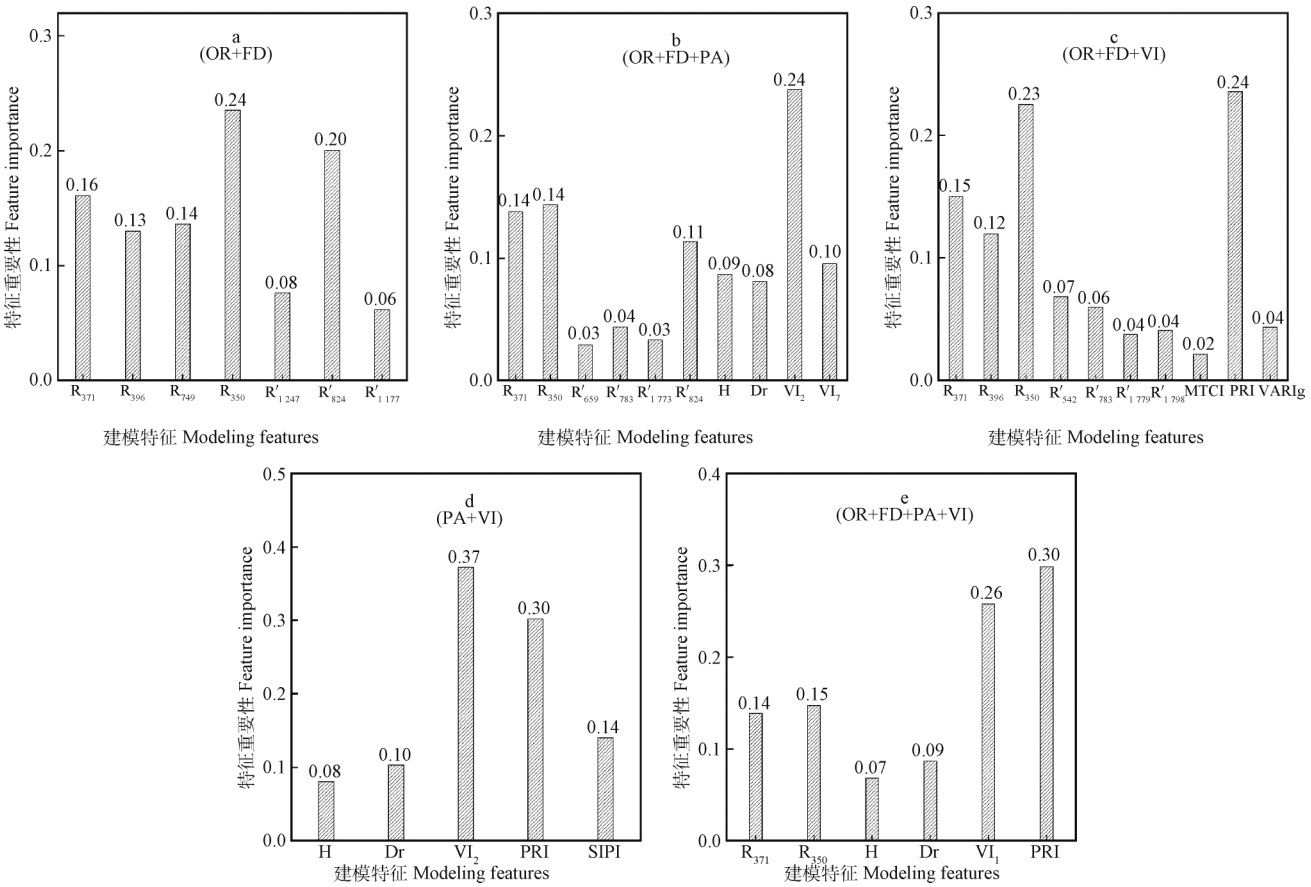

图 5　不同 RF 模型的特征重要性

Fig. 5　The importance of features in different RF models

利用全集数据(n=179),对不同类型特征变量组合 RF 模型的实测生物量与估测生物量进行相关性分析(图 6),可以得出,在不同的特征变量组合中,基于光谱位置面积参数 PA 和植被指数 VI 的组合 PA+VI 的模型 AGB 估测效果最好,其线性决定系数 R^2 达到 0.73,均方根误差 *RMSE* 为 476.93 kg·ha^{-1};估测效果次之的是 OR+FD+PA+VI 所有变量组合(R^2=0.72);OR+FD 组合和 OR+FD+PA 组合的 AGB 模型估测效果一致(R^2=0.68);OR+FD+VI 组合的 RF 模型的 AGB 估测效果最差,其实测 AGB 和估测 AGB 的决定系数 R^2 为 0.66。可见,不同类型特征变量组合模型的整体估测效果和其测试集 10 折交叉验证的评价结果基本一致。

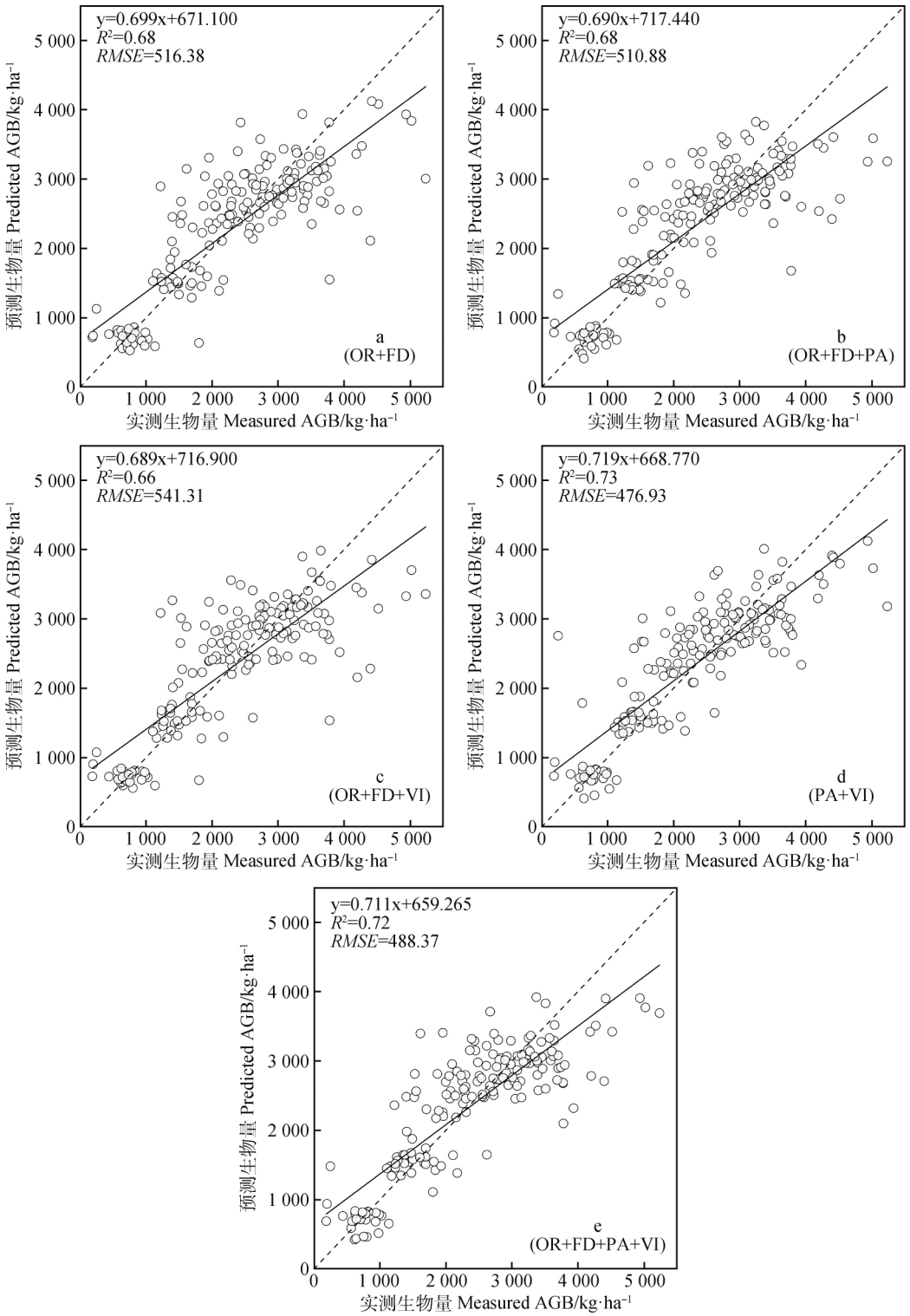

图 6 不同 RF 估测模型的实测 AGB 与估测 AGB 的相关性

Fig. 6 Correlation between measured AGB and predicted AGB in different RF estimation

3 讨论

与传统的统计方法相比，机器学习模型适用于较大规模数据和复杂场景，更关心模型的可用性与预测能力[15]。本文用 RF 构建的最优高寒草地 AGB 估测模型（PA＋VI 组合）的测试集精度为 R^2 ＝0.71，全集的预测精度为 R^2 ＝0.73，在不破坏草地的同时能够快速测定 AGB。

特征选择是高光谱数据模型构建前的一个重要步骤，主要起数据降维和模型简化的作用。本研究

先后用 SPA 和 RFE 算法进行草地 AGB 建模变量的筛选,其中 SPA 算法主要用于连续光谱的特征波段提取。本研究表明,SPA 将 OR 和 FD 的波段数量从 1 551 个分别减少到 9 个和 19 个(图 2),可极大地降低数据的维度,但在特征波段提取时,SPA 并未选择 *RMSE* 最小时的波段数量,而是综合考量波段数量和 *RMSE* 对模型的影响,选择了适宜数量的特征波段,这和王承克[31]、吴迪[32]等的研究结论相一致。此外,OR 的 SPA 特征波段主要集中在 740,950,1 100 和 1 300 nm 等附近,其中 740 nm 属于红边范围,950,1 100 和 1 300 nm 属于近红外范围。有研究表明红边包含了可以表征生物量、叶绿素等参量的光谱信息[33],而 920～1 120 nm 之间的近红外波段很有可能与叶片水分和干物质的吸收有关[34],1 300 nm 波段则和杨晨波等[16]对冬小麦的地上干生物量的敏感波段的提取结果相近。

不同类型特征变量的 RF 估测模型结果表明,PA 和 VI 的建模精度好于 OR,FD 特征波段的建模精度,这可能与 PA 和 VI 是由多个波段组合计算的指数,而 OR,FD 特征波段是单波段有关,在一定程度上说明了 PA 和 VI 等波段组合指数比 OR 等单波段包含的光谱信息丰富。在不同类型特征变量组合的 RF 估测模型结果中,精度最高的是 PA+VI 组合,其仅有 5 个建模特征,是所有组合中建模特征数量最少的组合,其中特征 VI_2 和 PRI 为整个模型提供了 67%的重要性。VI_2 是红谷吸收深度 D 和绿峰反射高度 H 的归一化指数(表 2),能较好地反映草地生物量等参量,如胥慧等[10]的研究表明(D—H)/(D+H)和草地生物量的相关系数达到 0.660。PRI 指光化学反射系数,是 R_{531} 和 R_{570} 的归一化指数[35],能反映植物光合作用过程中光能利用效率,可用于研究植被生产力[36],道日娜[37]研究发现,重度放牧时草地生物量最小,此时的 PRI 也最小,因此 PRI 也能较好的反映草地的 AGB。

然而,本研究也存在一定的局限性。例如,机器学习是个典型的数据驱动的问题,通常需要大量的数据[15],而本研究的样本数为 179 条,这对机器学习而言数据量较小,可能对模型的精度造成一定的影响。放牧干扰行为也可能影响到模型最终结果:苏日古嘎[38]研究发现,与未放牧草地相比,放牧会改变草地植被的群落物种多样性、优势度和植物种群的空间异质性,且不同放牧家畜对草地植被的影响程度不同;还有研究表明,由于长期围栏放牧家畜的践踏和选择性取食,导致植被环境斑块化[39],以上行为都会改变草地植被冠层的结构组成,使获取的高光谱反射率呈现不同的特征,进而对 RF 模型的精度造成一定的影响。此外,由于 ASD 光谱仪测定的是目标物"点"的光谱数据,是 350～2 500 nm 全光谱段(间隔 1 nm)的观测结果,由大量窄波段组成,所选最优模型变量也由其中部分特征波段组成,但目前可用的高时空分辨率的光学遥感卫星通常只有宽波段的观测结果,后续将进一步研究最优模型变量的特征波段和遥感影像波段的对应关系,进而实现大面积草地 AGB 的模型构建和反演。无人机高光谱成像技术具有获取数据快、操作灵活等特点,现被越来越多地应用到农作物等的监测,如陶惠林等[40]借助无人机高光谱遥感平台,获取了冬小麦各生育期的无人机影像,提取了植被指数和红边特征,最终构建了冬小麦产量模型。与 ASD 地物光谱仪相比,无人机高光谱成像技术的覆盖面积更广,因而可以构建较大尺度的生物量估算模型,这对高寒草地 AGB 的遥感估测具有参考意义。

4　结论

SPA 可有效降低高光谱数据维度,使原始光谱 OR 和一阶微分光谱 FD 的波段数量分别减少了 99.4%和 98.8%。在 4 种特征变量分别构建的草地 AGB 估测模型中,基于 VI 的 RF 模型精度最高(测试集 $R^2=0.70$,$RMSE=557.87\ kg\cdot ha^{-1}$),建模特征为 NDBleaf,OSAVI,TCARI,MTCI,PRI,SIPI,VARIg,其中特征 SIPI 和 PRI 共为模型贡献了 72%的重要性,实测 AGB 和估测 AGB 的线性决定系数 R^2 达到 0.72。不同类型特征变量组合构建的草地 AGB 估测模型中,PA+VI 组合的 RF 模型精度最高($R^2=0.71$,$RMSE=548.97\ kg\cdot ha^{-1}$),建模特征为 H,Dr,$VI_2$,PRI 和 SIPI,其中特征 VI_2 和 PRI 共为模型贡献了 67%的重要性,实测 AGB 和估测 AGB 的线性决定系数 R^2 达到 0.73。整体而言,相对 OR,FD 等特征变量,PA,VI 的特征变量组合构建的草地 AGB 的 RF 模型精度较高,建模效果更好。

参考文献

[1] 葛静,孟宝平,杨淑霞,等. 基于 ADC 和 MODIS 遥感数据的高寒草地地上生物量监测研究——以黄河源区为例[J]. 草业学报,2017,26(7):23-34

[2] 胡中民,樊江文,钟华平,等. 中国草地地下生物量研究进展

[J]. 生态学杂志,2005(9):1095-1101
[3] 孙义. 高寒草甸—藏羊放牧系统土草畜互作特征[D]. 兰州:兰州大学,2015:1
[4] 魏茂宏. 江河源区高寒草地放牧侵蚀能值模拟研究[D]. 兰州:兰州大学,2016:2
[5] 李莉. 高光谱遥感图像特征提取方法研究[D]. 无锡:江南大学,2020:1,4
[6] 纪童,王波,杨军银,等. 基于高光谱的草坪草叶绿素含量模拟估算[J]. 光谱学与光谱分析,2020,40(8):2571-2577
[7] 高金龙. 青藏高原东缘高寒天然草地牧草氮磷养分和生长状况的高光谱遥感研究[D]. 兰州:兰州大学,2020:26-58
[8] 王磊. 基于高光谱的草地冠层物种丰度估算与叶面积指数反演[D]. 南京:南京大学,2019:82-87
[9] 韩万强,靳瑰丽,岳永寰,等. 基于高光谱成像技术的伊犁绢蒿荒漠草地主要植物识别参数的筛选[J]. 草地学报,2020,28(4):1153-1163
[10] 胥慧,包玉海,包刚,等. 内蒙古典型草原干草生物量高光谱遥感估算研究[J]. 阴山学刊(自然科学版),2014,28(4):22-27
[11] 夏浪,张瑞瑞,陈立平,等. 基于无人机高光谱影像的地表植被生物量反演波段优选[J]. 电子测量技术,2018,41(9):87-90
[12] 马维维. 草地类型及其品质参数的遥感反演方法研究[D]. 上海:中国科学院研究生院(上海技术物理研究所),2015:74-78
[13] 张凯,郭铌,王润元,等. 甘南草地地上生物量的高光谱遥感估算研究[J]. 草业科学,2009,26(11):44-50
[14] 安海波,李斐,赵萌莉,等. 基于优化光谱指数的牧草生物量估算[J]. 光谱学与光谱分析,2015,35(11):3155-3160
[15] 胡林,刘婷婷,李欢,等. 机器学习及其在农业中应用研究的展望[J]. 农业图书情报,2019,31(10):12-22
[16] 杨晨波,冯美臣,孙慧,等. 不同灌水处理下冬小麦地上干生物量的高光谱监测[J]. 生态学杂志,2019,38(6):1767-1773
[17] 刘笑笑. 基于 RF-RFE 算法的森林生物量遥感特征选择方法研究[D]. 泰安:山东农业大学,2016:32-33
[18] 黄卫卫. 基于随机森林——递归特征消除的道路交通事故成因分析[J]. 电脑知识与技术,2018,14(14):240-243
[19] 李盛芳,贾敏智,董大明. 随机森林算法的水果糖分近红外光谱测量[J]. 光谱学与光谱分析,2018,38(6):1766-1771
[20] 岳继博,杨贵军,冯海宽. 基于随机森林算法的冬小麦生物量遥感估算模型对比[J]. 农业工程学报,2016,32(18):175-182
[21] 竟霞,白宗璠,高媛,等. 利用随机森林法协同 SIF 和反射率光谱监测小麦条锈病[J]. 农业工程学报,2019,35(13):154-161
[22] 张春兰,杨贵军,李贺丽,等. 基于随机森林算法的冬小麦叶面积指数遥感反演研究[J]. 中国农业科学,2018,51(5):855-867
[23] 李东,罗旭鹏,曹广民,等. 高寒草甸土壤异养呼吸对气候变化和氮沉降响应的模拟[J]. 草业学报,2015,24(7):1-11
[24] 北京理加联合科技有限公司. ASD FieldSpec Dual 软件[EB/OL]. http://www.li-ca.com/prod_view.aspx? TypeId=70&Id=180&FId=t3:70:3,2018-04-19/2021-04-10
[25] Savitzky A,Golay M. Smoothing and differentiation of data by simplified least squares procedures[J]. Analytical Chemistry,1964,36(8):1627-1639
[26] 谢莉莉,王福民,张垚,等. 基于多生育期光谱变量的水稻直链淀粉含量监测[J]. 农业工程学报,2020,36(8):165-173
[27] 田明璐. 西北地区冬小麦生长状况高光谱遥感监测研究[D]. 杨凌:西北农林科技大学,2017:22-23
[28] Hurt N E. Signal enhancement and the method of successive projections[J]. Acta Applicandae Mathematica,1991,23(2):145-162
[29] 吴辰文,梁靖涵,王伟,等. 基于递归特征消除方法的随机森林算法[J]. 统计与决策,2017(21):60-63
[30] 侯蒙京,殷建鹏,葛静,等. 基于随机森林的高寒湿地地区土地覆盖遥感分类方法[J]. 农业机械学报,2020,51(7):220-227
[31] 王承克,张泽翔,黄晓玮,等. 高光谱成像的豆腐形成过程中组分含量变化检测[J]. 光谱学与光谱分析,2020,40(11):3549-3555
[32] 吴迪,吴洪喜,蔡景波,等. 基于无信息变量消除法和连续投影算法的可见-近红外光谱技术白虾种分类方法研究[J]. 红外与毫米波学报,2009,28(6):423-427
[33] Hansen P M,Schjoerring J K. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression[J]. Remote Sensing of Environment,2003,86(4):542-553
[34] Lei Z,Zhou Z,Zhang G, *et al*. Monitoring the leaf water content and specific leaf weight of cotton (Gossypium hirsutum L) in saline soil using leaf spectral reflectance[J]. European Journal of Agronomy,2012,41:103-117
[35] Gamon J A,Penuelas J,Field C B. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency[J]. Remote Sensing of Environment,1992,41(1):35-44
[36] Hall F G,Hilker T,Coops N C. Data assimilation of photosynthetic light-use efficiency using multi-angular satellite data:I. Model formulation[J]. Remote Sensing of Environment,2012,121:301-308
[37] 道日娜. 放牧对克氏针茅(Stipa krylovii)草原植物性状、功能群与生产力的影响[D]. 呼和浩特:内蒙古大学,2016:20-23,39-42
[38] 苏日古嘎. 不同放牧家畜对内蒙古典型草原植物群落结构的影响[D]. 呼和浩特:内蒙古大学,2020:2-4,24-29
[39] 马景川,黄训兵,秦兴虎,等. 放牧干扰对典型草原植被光谱及蝗虫密度的影响[J]. 植物保护,2017,43(6):6-10,28
[40] 陶惠林,徐良骥,冯海宽,等. 基于无人机高光谱遥感数据的冬小麦产量估算[J]. 农业机械学报,2020,51(7):146-155

(责任编辑 闵芝智)