Landsat 8 和机器学习估算蒙古高原草地地上生物量

赵 越，徐大伟，范凯凯，李淑贞，沈贝贝，邵长亮，王 旭※，辛晓平
（中国农业科学院农业资源与农业区划研究所/呼伦贝尔草原生态系统国家野外科学观测研究站，北京 100081）

摘 要：草地地上生物量（Above-Ground Biomass，AGB）是反映草地植被利用状况的重要参数，其精准监测对于草地科学管理与合理利用具有重要意义。近年来，遥感技术因其能快速、准确获取大尺度草地光谱信息，已经被广泛应用于草地地上生物量的估算中。该研究以中国内蒙古呼伦贝尔市与其毗邻的蒙古国东方省草原区为研究区，利用 Landsat 8 数据计算的 9 种植被指数、气象数据和地面调查数据，比较分析 6 种机器学习算法构建的回归模型性能，重新构建优化的随机森林回归模型。结果表明，以光谱、降水量、气温为特征的优化后的随机森林回归模型性能更稳定，预测值与实测值之间决定系数为 0.801，均方根误差为 43.709 g/m^2，相对均方根误差为 23.077%。研究区域地上生物量呈中部较低，东西两侧较高的空间分布特征，最高可达 357.2 g/m^2，最低为 33.01 g/m^2，与该区域降水量与草地利用方式的空间异质性密切相关。该研究表明，基于 Landsat 8 数据结合气象数据构建的机器学习模型在草地生物量遥感反演中有较大潜力，地上生物量反演结果可以为草地资源合理利用与评价提供参考。

关键词：遥感；反演；机器学习；地上生物量；蒙古高原
doi：10.11975/j.issn.1002-6819.2022.24.015
中图分类号：S127；TP79 **文献标志码**：A **文章编号**：1002-6819(2022)-24-0138-07

赵越，徐大伟，范凯凯，等. Landsat 8 和机器学习估算蒙古高原草地地上生物量[J]. 农业工程学报，2022，38(24)：138-144. doi：10.11975/j.issn.1002-6819.2022.24.015 http://www.tcsae.org
Zhao Yue, Xu Dawei, Fan Kaikai, et al. Estimating above-ground biomass in grassland using Landsat 8 and machine learning in Mongolian Plateau[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(24): 138-144. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2022.24.015 http://www.tcsae.org

0 引 言

草地是中国面积最大的陆地生态系统，不仅为畜牧业提供了重要资源，而且在防风固沙、水土保持和维护生态安全等方面扮演着重要角色[1]。根据第二次全国国土调查结果，中国草地面积 28 731.4 hm^2，占国土面积 40%以上[2]。草地地上生物量（Above-Ground Biomass，AGB）是评价草地生态系统健康和可持续利用状况的关键指标[3-4]，快速、准确获取草地地上生物量对于确定合理载畜量、科学管理草地资源至关重要。

蒙古高原是欧亚大陆草地的重要组成部分，蒙古国东方省和中国呼伦贝尔市毗邻，气候特征和草地类型相似，但由于人口、经济状况和草地利用方式不同，草地生物量分布存在较大的空间异质性，这为构建草地生物量遥感反演模型提供了理想的试验场景。

草地生物量传统调查方法主要是利用收获法实测单位面积绿色植物地上部分干质量，但该方法存在耗时耗力、时空尺度上局限性明显等问题[5]。相比之下，遥感技术具有监测空间尺度大、时效性强等优点，适用于大尺度草地生物量调查。近年来，随着遥感数据源不断完善，更多学者构建基于光谱反射率或计算植被指数与生物量间的统计关系模型、基于同化遥感数据的作物生长模型以及通过机器学习自动构建生物量预测关系模型等进行草地生物量反演和预测[5–9]。草地物种组成丰富，野外实测数据难以获取，基于机器学习算法的模型具有较强的鲁棒性的优点[10-12]，可以通过有限的数据量得到较好的地上生物量的估计结果。由于卫星遥感在成像过程中受到大气和土壤等因素的影响导致系统性误差[13]，现有草地生物量遥感反演机器学习模型在精度和效率上仍有提升空间，一些学者通过引入气象数据作为特征辅助建模[14]，可以显著提高草地生物量反演精度，从而获得更精确和稳定的反演结果。本研究基于 Landsat 8 遥感数据和地面调查数据，通过构建光谱指数与气象数据的机器学习模型，目的是：对该研究区草地生物量进行反演并对中国呼伦贝尔与蒙古国东方省进行对比，探讨草地生物量的高精度遥感反演方法。

1 研究区与数据

1.1 研究区概况

蒙古国东方省位于蒙古国东部地区，与中国内蒙古呼伦贝尔市接壤，地理位置介于 46°16'26"N～50°16'34"N，111°59'57"E～119°56'15"E 之间，海拔高度 560～1 300 m，地处北温带，气候干燥，年降水量 150～300 mm，气温在−27～21 ℃之间。呼伦贝尔市位于 48°6'14"N～48°22'57"N，119°29'43"E～119°44'54"E 之间，海拔高度

收稿日期：2022-08-31 修订日期：2022-11-17
基金项目：国家重点研发计划项目（2021YFD1300502）；国家科技基础资源调查专项（2019FY102000）；国家自然科学基金项目（32171567）
作者简介：赵越，研究方向为农业工程与信息技术。
Email：82101215505@caas.cn
※通信作者：王旭，博士，副研究员，研究方向为草地生态遥感。
Email：wangxu01@caas.cn

550～1 000 m，温带半湿润气候，年降水量 250～400 mm，气温在−18～−30 ℃之间。两地区处于相同纬度带，被肯特山脉和大兴安岭三面包围，气候类型均相似，但两地人口数量、密度和大型牲畜数量差异较大（表 1）。该区域草地类型主要包括草甸草原和典型草原等类型，草甸草原主要分布在大兴安岭西部和肯特山东部山麓，主要植物有贝加尔针茅（*Stipa baicalensis*）、羊草（*Leymus chinensis*）、线叶菊（*Filifolium sibiricum*）等；典型草原主要分布在研究区中部的克鲁伦河和乌尔逊河流域，主要植物有大针茅（*Stipa grandis*）、羊草（*Leymus chinensis*）、克氏针茅（*Stipa Krylovii*）、冷蒿（*Artemisisfrigida*）等[15]。土壤类型为栗钙土或暗栗钙土[16]。

表 1　人口和农业统计数据

Table 1　Population and agricultural statistics

区域 Region	人口数量 Population/万人	人口密度 Population density (人·km^{-2})	大型牲畜数量 Number of large livestock/万头			利用方式 Use pattern
			牛 Cow	马 Horse	骆驼 Camel	
蒙古国东方省 Dornod, Mongolia	8.098	0.7	243.14	278.81	0.606	自然放牧
中国呼伦贝尔市 HulunBuir, China	336.370	13.0	80.66	19.33	0.280	放牧/刈割

1.2　数据来源

1.2.1　野外调查与采样

2018 年、2020 年生长季期间分别在蒙古国东方省和中国呼伦贝尔市草原区开展了样带地面样方调查和生物量采样（图 1）。每个样点等距设置 3 个样方，每个样方间隔 15 m，样方大小为 1 m×1 m。调查样方内植物种类、盖度和高度等指标，并详细记录经纬度坐标、海拔、地形、利用方式、利用强度等信息。地上生物量采用收获法，将样方内植物所有绿色部分用剪刀齐地面剪取，烘箱 65 ℃烘干 24 h，称干质量。

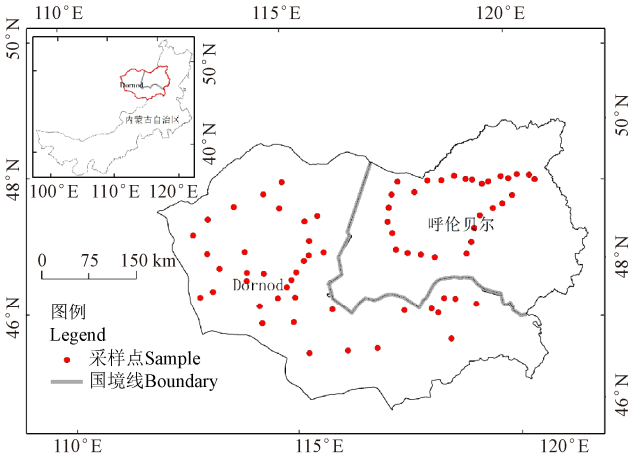


图 1　研究区和采样点分布

Fig.1　Study area and sample distribution

1.2.2　遥感数据

遥感技术作为一项 20 世纪 60 年代兴起的一种探测技术，广泛应用于生态研究领域[17-18]。美国的 Landsat 系列是运行时间最长的地球观测计划，空间分辨率为 15～60 m 不等，时间分辨率为 16 d。其中，Landsat 8 遥感数据空间分辨率较高，本研究采用 Google Earth Engine（GEE）平台提供的 2018 年和 2020 年 7—8 月分辨率 30 m 的 Landsat 8 影像来估算研究区内的地上生物量。为了保证数据的完整性，选择了所选时相内的同轨道数据筛选无云的部分，对不同轨道影像边界进行羽化并镶嵌，根据研究区范围裁剪影像。从经预处理的影像中提取各波段反射率（红，绿，蓝，近红外）作为冠层反射率，并计算植被指数（表 2）。

表 2　植被指数及公式

Table 2　Vegetation index and formula

植被指数 Vegetation index	公式 Formulation	引用 References
叶绿素指数 Chlorophyll Index_green (Cig)	CIg = (NIR / Green) − 1	[19]
增强型植被指数 Enhanced Vegetation Index (EVI)	EVI = 2.5 (NIR − Red) / (NIR + 6 Red − 7.5 Blue + 1)	[20]
全球环境监测指数 Global Environmental Monitoring Index (GEMI)	GEMI = eta · (1 − 0.25 eta) − (Red − 0.125) / (1 − Red) eta = (2×(NIR^2 − Red^2) + 1.5 NIR + 0.5Red) / (NIR + Red + 0.5)	[21]
绿光归一化差值植被指数 Green Normalized Difference Vegetation Index (GNDVI)	GNDVI = (NIR − Green) / (NIR + Green)	[22]
修正土壤调节植被指数 Modified Soil Adjusted Vegetation Index (MSAVI)	MSAVI = 1/2×(2(NIR+1) − sqrt((2 NIR+1) × 2 − 8 (NIR−Red)))	[23]
归一化差值植被指数 Normalized Difference Vegetation Index (NDVI)	NDVI = ((NIR − Red) / (NIR + Red))	[24]
调节土壤的植被指数 Soil Adjusted Vegetation Index (SAVI)	SAVI = (NIR − Red) / (NIR + Red + 0.5) × (1 + 0.5)	[25]
简单比值 Simple Ratio (SR)	SR = NIR / Red	[26]
可见光大气阻抗指数 Visible Atmospherically Resistant Index (VARI)	VARI = (Green − Red) / (Green + Red − Blue)	[27]

注：Blue：蓝色波段反射率；Green：绿色波段反射率；Red：红色波段反射率；NIR：近红外波段反射率。
Note: Blue: Blue band reflectance; Green: Green band reflectance; Red: Red band reflectance; NIR: Near infrared band reflectance.

1.2.3　气候数据

草地植被生长状况与气候条件密切相关，由于研究区地理位置位于山系包围的草原区，降水量和气温的变化显著影响草地生物量。气候数据是由 World Clim（https://www.worldclim.org/）提供的 GeoTiff 格式数据，包括当年生长季平均气温和生长季总降水量 1km 空间分辨率栅格数据。

1.3　研究方法

机器学习算法通过直接从数据拟合一个灵活的模型学习输入值（例如反射率）和输出值（例如地上生物量）之间的关系[28]。本研究通过 Python3.10 调用第三方包（Scikit-learn）构建机器学习模型，该包提供分类、回归、聚类、降维、调参、预处理等科学工具，本研究选用包括：决策树回归（Decision Tree Regressor）、K 临近回归

（K-Neighbors Regressor）、随机森林回归（Random Forest Regressor）、梯度增加回归（Gradient Boosting Regressor）、装袋回归（Bagging Regressor）、极端树（Extra Tree Regressor）在内的 6 种算法作为回归模型，探究不同机器学习算法预测地上生物量的性能，讨论适合作为反演研究区内地上生物量的机器学习算法。

根据野外调查样点的位置坐标，提取相应位置影像像元的植被指数、年均降水量（Prec）和年均气温（Tavg）。通过六折交叉验证将数据分为训练集和测试集。根据多源数据的训练集作为机器学习算法的输入，建立其与地上生物量的回归关系，并进行结果检验。基于机器学习算法得出的回归关系，反演研究区内地上生物量的空间分布。

反演结果精度评定选择相关系数（coefficient of determination，R^2），均方根误差（Root Mean Square Error，RMSE）和相对均方根误差（relative Root Mean Square Error，rRMSE）来评价模型性能，其公式为

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

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

$$\mathrm{rRMSE} = \frac{\mathrm{RMSE}}{\bar{y}} \times 100\% \quad (3)$$

式中 y_i 为实测值，$\hat{y}_i$ 为预测值，$\bar{y}$ 为实测值的平均值，n 为样本数量，RMSE 和 rRMSE 越小，表明误差越小，R^2 表示回归分析趋势线的预测值与对应的实测数据之间的拟合程度，当趋势线 R^2 趋近于 1 时，其可信度最高。

2 结果分析

2.1 基于多种机器学习算法的地上生物量回归分析

本文选择生长季总降水量，生长季平均气温和九种植被指数为特征建立机器学习回归模型，将 3 种数据组成 4 种组合，通过 3 种评价指标（R^2，RMSE 和 rRMSE）比较各种数据组合方式与生物量在多种机器学习模型中的性能（表 3）。

通过回归验证结果得知，6 种机器学习模型在采用单独的光谱数据作为输入的情况下，总体表现非常接近，且模型性能较差（RMSE 为 63.852～87.944 g/m^2，rRMSE 为 46.432%～33.712%，R^2 为 0.388～0.647）。其中随机森林（RMSE=63.852 g/m^2，rRMSE=33.712%，R^2=0.647）和装袋回归模型（RMSE=69.661 g/m^2，rRMSE=36.779%，R^2=0.637）得到的评价结果略高于其他模型。随着数据组合中特征数量的增加，所有回归方法的 RMSE 和 rRMSE 都逐渐减小，R^2 逐渐增大，表明对于不同回归模型而言，增加特征数量都能够有效地处理多数据输入特征的融合。当数据组合为光谱+降水和光谱+温度时，几种回归方法得到的评价结果均有提高，但评价指标的上升幅度不同，尤其是在光谱+温度作为数据源的情况下 K 近邻方法的误差最小（RMSE=49.566 g/m^2，rRMSE=26.170%）。当数据组合为光谱+降水+温度时，各回归方法得到了最好的模型评价结果（RMSE 为 51.702～70.683 g/m^2，rRMSE 为 27.297%～37.319%，R^2 为 0.664～0.749），随机森林算法效果最好（RMSE=51.702 g/m^2，rRMSE=27.297%，R^2=0.749）。

表 3 不同地上生物量预测模型统计验证

Table 3 Statistical validation of prediction models for different aboveground biomass

数据类型 Data type	评价指标 Metrics	决策树 DTR	K 近邻 KNN	随机森林 RFR	梯度增加 GBR	装袋回归 BR	极端树 ETR
光谱 Spectral	RMSE/(g·m^{-2})	75.058	87.944	63.852	70.955	69.661	68.182
	rRMSE/%	39.629	46.432	33.712	37.462	36.779	35.998
	R^2	0.511	0.388	0.647	0.487	0.637	0.561
光谱+降水 Spectral + Prec	RMSE/(g·m^{-2})	70.693	78.395	62.245	69.627	71.116	68.175
	rRMSE/%	37.324	41.390	32.864	36.761	37.547	35.995
	R^2	0.513	0.473	0.720	0.671	0.631	0.648
光谱+温度 Spectral + Tavg	RMSE/(g·m^{-2})	70.618	49.566	61.614	60.330	57.037	88.819
	rRMSE/%	37.285	26.170	32.530	31.852	30.114	46.894
	R^2	0.687	0.689	0.727	0.709	0.662	0.457
光谱+降水+温度 Spectral + Prec + Tavg	RMSE/(g·m^{-2})	59.271	70.683	51.702	64.827	64.134	53.379
	rRMSE/%	31.294	37.319	27.297	34.227	33.861	28.182
	R^2	0.707	0.664	0.749	0.665	0.695	0.685

注：Prec 为生长季总降水量，Tavg 为生长季平均气温。

Note: Prec is total precipitation during the growing season, Tavg is Average growing season temperature.

2.2 随机森林算法的特征选择和超参数优化

特征对目标变量预测的相对重要性可以通过（随机森林中树的决策节点的）特征使用的相对顺序（即深度）来进行评估。决策树顶部特征使用的相对顺序，对输入样本的选择做出决策。因此，可以使用接受每个特征对最终预测的贡献的样本比例，来评估该特征的相对重要性。多种机器学习算法中随机森林表现出优异的性能，并且对于不同数据组合形式作为输入特征，均有较高的精度和较低的误差。通过评估输入特征的相对重要性确定多源数据在模型中的权重（图 2）。输出结果显示降水量是重要性最高的随机森林算法输入特征，其权重大于 0.1 为最大值，远高于其他光谱数据。光谱数据中 VARI，MSAVI，GEMI 三种植被指数作为特征权重大于 0.09，高于其他植被指数。

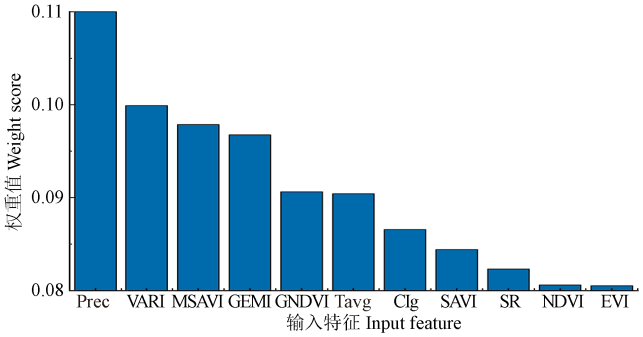
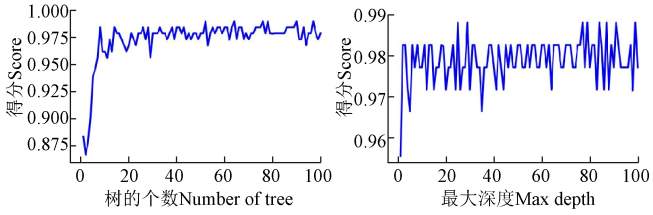


图 2 随机森林算法中各输入特征权重

Fig.2 Weight of each input feature in random forest algorithm

超参数是优化预测建模算法性能的模型配置参数，对模型内超参数进行优化能够使模型性能较好的同时使用较少的计算量。随机森林是基于 bagging 框架的决策树模型，其超参数包括：随机森林框架参数，如：树的个数（n_estimators）和随机森林决策树的参数，如：最大

深度（max_depth）。本研究对随机森林树的个数和最大深度进行超参数优化，优化后的结果通过交叉验证评分进行评价。通过遍历模型内不同树的个数和最大深度确定最佳值。n_estimators 值较小时模型出现明显的欠拟合。其数值增加至 n_estimators=52，模型性能显著提升并趋于稳定，继续增加 n_estimators 值对模型提升不明显（图 3a）。max_depth 值较大时模型有较高的复杂度，max_depth>25 时模型性能提升不再显著（图 3b）。



a. 树的个数优化　b. 最大深度优化
a. Number of estimators optimization　b. Max depth optimization

图 3　超参数优化

Fig.3　Hyperparameter optimization

为验证只选用权重较高的特征量的模型稳定性，优选降水量（Prec）、VARI、MSAVI、GEMI 四项在随机森林算法的输入中权重较高的数据作为输入特征，通过经超参数优化的随机森林算法进行回归分析（图 4）。模型评价结果显示，经筛选优化后的数据作为特征的随机森林回归模型，在减少数据量的同时模型性能得到提升（RMSE=43.709 g/m^2，rRMSE=23.077%，R^2=0.801）。并且筛选优化后的回归模型在误差和拟合程度，均高于未经筛选过的模型评价结果，这可能是由于权重较低的特征与地上生物量相关性较差导致的。

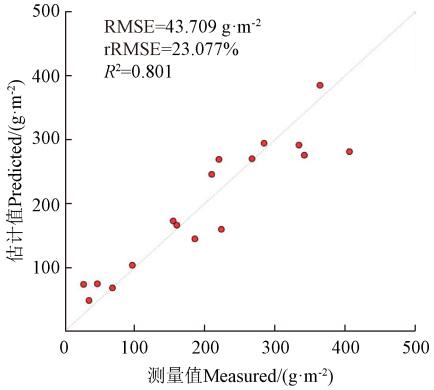


图 4　优化的随机森林算法地上生物量估计值与地面实测值

Fig.4　Aboveground biomass predicted and measured by optimized random forest algorithm

2.3　地上生物量反演

通过筛选优化后的随机森林回归算法作为估计模型，从空间上对研究区内地上生物量进行反演。从地上生物量空间分布图中易知，研究区内地上生物量呈现出明显的空间分布规律：蒙古国东方省西北地区和东部地区以及中国呼伦贝尔东部地区草地生物量较高，地上生物量最高可达 357.22 g/m^2，东方省中部和呼伦贝尔西部草地生物量较低，最低为 33.01 g/m^2，整个研究区分布格局来看，地上生物量呈中部较低，东西两侧较高的分布特征（图 5）。

东方省与内蒙古呼伦贝尔市处于同纬度带，地理特征和气候类型相似，通过对地上生物量分布情况的分析，两地在地上生物量空间格局上呈现出差异性（图 6），标准差越大表示该地区大部分生物量与其平均生物量差异较大，标准差越大地上生物量分布越分散，峰值越低，曲线越低平。即东方省地上生物量空间分布相对均一（标准差为 74.35 g/m^2），而呼伦贝尔草地地上生物量空间分布异质性较强（标准差为 107.43 g/m^2）。东方省人口数量、密度和牲畜数量均远低于呼伦贝尔市，且其利用方式以自然放牧为主，而呼伦贝尔以放牧和刈割利用为主，利用强度较大[29-30]（表 1）。这也表明，东方省草地生物量空间分布特征主要受气候主导，而呼伦贝尔草地生物量空间分布特征则受人类活动和气候共同影响。

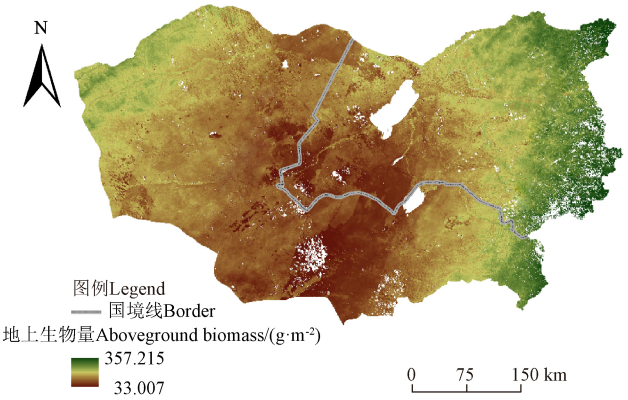
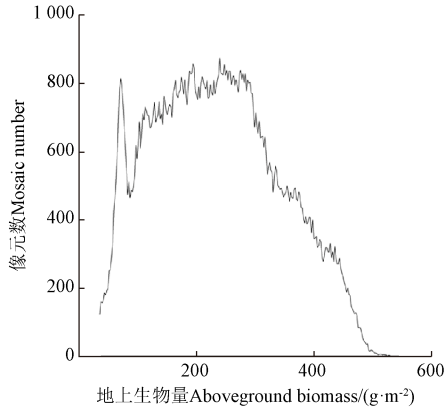
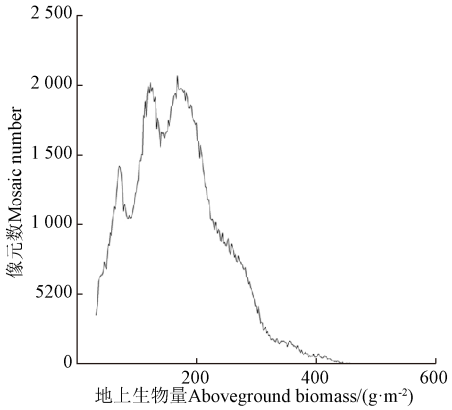


图 5　研究区地上生物量分布图

Fig.5　Distribution map of aboveground biomass in the study area



a. 东方省
a. Dornod



b. 呼伦贝尔
b. Hulubuir

图 6　地上生物量像元数分布统计

Fig.6　Statistics of aboveground biomass mosaic number distribution

3 讨 论

作物生长模型建模由于涵盖一定的生物学理论和计算机算法，取得良好的效果，常用在作物的生物量预测中。但由于作物生长模型生物量估算方法需要大量不同数据源，如水分和土壤养分等，并且需要大量生态学理论对模型进行解释，在解决生物量估算的快速应用上受到了很大制约[31]。而基于各波段反射率或计算植被指数与生物量间关系的机器学习模型，因其模拟精度高、受复杂环境影响小而被广泛使用[32-34]。本研究设计了不同的数据类型进行组合形式作为机器学习的特征输入，增加不同类型的特征对几种回归模型性能均有明显提升（表 3）。这是由于机器学习算法的结果精度和稳定性与训练数据量密切相关，对于传统机器学习算法而言，在一定数据量阈值范围内更多的数据作为特征的情况下，其性能按照幂律增长，一段时间后随着数据量继续增加，导致模型的鲁棒性开始降低，模型性能进入停滞不前的状态。以本研究中验证性能较好的随机森林算法为例，随着特征数量的增加，模型拟合程度（R^2）显著提高，误差明显减小，类似的结论在 Maimaitijing 等[35]的研究中也得到有效验证。

在机器学习模型中，对特征进行筛选优化有助于在减少数据量的同时提高特征总体质量，从而最大程度上发挥传统机器学习模型较好的鲁棒性[36]。将各特征在随机森林算法中的权重得分作为重要性衡量标准（图 2），结果表明降水量在模型中占较大权重，对回归结果有显著影响，这是由于研究区特殊的地形特征导致的，蒙古国东方省和中国呼伦贝尔市位于大兴安岭以西和肯特山脉以南之间的地势平坦区域，特殊的地形地势阻止了气流对降水的影响，降水受季风影响较弱，而受地形影响较强，形成了靠近山脉降水量较大而中心地带较少的分布格局。由于半干旱地区降水量显著地影响草地生物量和群落物种丰富度[37]，因而降水成为本研究生物量回归模型的重要特征量。

光谱数据中本研究选用 9 种植被指数作为模型特征，其中 GEMI、MSAVI、VARI 三种植被指数在模型特征的重要性评估中显著高于其他植被指数，而在估计生物量中广泛应用的 NDVI、EVI 等植被指数重要性最低（图 2）。优选的 3 种植被指数共同特点是受土壤背景和大气影响较小[21,23,27]。东方省中部降水量较小，人类活动干扰较少，出现草地斑块化和多年枯落物与当年生长植被混杂的情况，导致植被指数作为特征的估计效果较差。通过对特征进行筛选，优化后的机器学习算法得到了性能更加稳定的评价结果。因此，选用对裸土敏感的 MSAVI 等植被指数作为特征进行建模是有效的特征优化手段。

4 结 论

本研究利用 Landsat 8 数据提取的 9 种植被指数、生长季降水量和气温数据，评估了 5 种机器学习回归模型的性能，对评价结果最好的随机森林模型进行特征重要性分析，根据分析结果建立经特征选择优化后的回归模型，模型评价结果显著提高，基于机器学习算法的回归模型能够有效的估计地上生物量，增加特征数量显著的提高了模型性能。利用优化后的模型进行草地生物量反演，估算精度较好（RMSE = 43.709 g/m^2，rRMSE = 23.077%，R^2 = 0.801），研究区草地生物量呈中部地区较低、东西两侧较高的空间分布特征，最大地上生物量达到 357.22 g/m^2，最低为 33.01 g/m^2，这与该区域气候与草地利用方式的空间异质性密切相关。本研究表明，基于 Landsat 8 数据结合气象数据构建的机器学习模型在草地生物量遥感反演中有较大潜力，反演结果通过实测数据验证，地上生物量反演结果可以为草地资源合理利用与评价提供参考。

[参 考 文 献]

[1] 刘艳，聂磊，杨耘. 基于植被指数估算天山牧区不同利用类型草地总产草量[J]. 农业工程学报，2018，34(9)：182-188.
Liu Yan, Nie Lei, Yang Yun. Estimation of total yield of different grassland types in Tianshan pastoral area based on vegetation index[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(9): 182-188. (in Chinese with English abstract)

[2] 第三次全国国土调查主要数据公报[Z]. [2022-11-13]. http://www.gov.cn/xinwen/2021-08/26/content_5633490.htm.

[3] 孟宝平，陈思宇，崔霞，等. 基于多源遥感数据的高寒草地生物量反演模型精度：以夏河县桑科草原试验区为例[J]. 草业科学，2015，32(11)：1730-1739.
Meng Baoping, Chen Siyu, Cui Xia, et al. The accuracy of grassland vegetation biomass estimated model based on multi-source remote sensing data: As a case of experimental area in Sangke grassland in Xiahe County[J]. Pratacultural Science, 2015, 32(11): 1730-1739. (in Chinese with English abstract)

[4] Piao S L, Fang J Y, He J S, et al. Spatial distribution of grassland biomass in China[J]. Chinese Journal of Plant Ecology, 2004, 28(4): 491-498

[5] 王新云，郭艺歌，何杰. 基于多源遥感数据的草地生物量估算方法[J]. 农业工程学报，2014，30(11)：159-166.
Wang Xinyun, Guo Yige, He Jie. Estimation of above-ground biomass of grassland based on multi-source remote sensing data[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2014, 30(11): 159-166. (in Chinese with English abstract)

[6] Filho M G, Kuplich T M, Quadros F L F D. Estimating natural grassland biomass by vegetation indices using Sentinel 2 remote sensing data[J]. International Journal of Remote Sensing, 2020, 41(8): 2861-2876.

[7] Yang S X, Feng Q S, Liang T G, el al. Modeling grassland above-ground biomass based on artificial neural network and remote sensing in the Three-River Headwaters Region[J]. Remote Sensing of Environment, 2018, 204: 448-455.

[8] Clevers J G P W, Leeuwen H J C V. Combined use of optical and microwave remote sensing data for crop growth monitoring[J]. Remote Sensing of Environment, 1996, 56(1): 42-51.

[9] 吴门新，钱拴，侯英雨，等. 利用 NDVI 资料估算中国北方草原区牧草产量[J]. 农业工程学报，2009，25(Supp. 2)：149-155.
Wu Menxin, Qian Shuan, Hou Yingyu, et al. Estimation of

forage yield in Northern China based on NDVI data[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2009, 25(Supp. 2): 149-155. (in Chinese with English abstract)

[10] 严海军，卓越，李茂娜，等. 基于机器学习和无人机多光谱遥感的苜蓿产量预测[J]. 农业工程学报，2022，38(11)：64-71.
Yan Haijun, Zhuo Yue, Li Maona, et al. Alfalfa yield prediction using machine learning and UAV multispectral remote sensing[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(11): 64-71. (in Chinese with English abstract)

[11] 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 Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(10): 191-199. (in Chinese with English abstract).

[12] 陆军胜，陈绍民，黄文敏，等. 采用 SEPLS_ELM 模型估算夏玉米地上部生物量和叶面积指数[J]. 农业工程学报，2021，37(18)：128-135.
Lu Junsheng, Chen Shaomin, Huang Wenmin, et al. Estimation of aboveground biomass and leaf area index of summer maize using SEPLS_ELM model[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2021, 37(18): 128-135. (in Chinese with English abstract)

[13] 姚薇，李志军，姚珙，等. Landsat 卫星遥感影像的大气校正方法研究[J]. 大气科学学报，2011，34(2)：251-256.
Yao Wei, Li Zhijun, Yao Hong, et al. Atmospheric correction model for Landsat images[J]. Transactions of Atmosoheric Sciences, 2011, 34(2): 251-256. (in Chinese with English abstract)

[14] 张旭琛，朱华忠，钟华平，等. 新疆伊犁地区草地植被地上生物量遥感反演[J]. 草业学报，2015，24(6)：25-34.
Zhang Xuchen, Zhu Huazhong, Zhong Huaping, et al. Assessment of above-ground Biomass of Grassland using remote sensing, Yili, Xinjiang[J]. Acta Prataculturae Sinica, 2015, 24(6): 25-34. (in Chinese with English abstract)

[15] 王旭，闫瑞瑞，邓钰，等. 放牧对呼伦贝尔草甸草原土壤呼吸温度敏感性的影响[J]. 环境科学，2014，35(5)：1909-1914.
Wang Xu, Yan Ruirui, Deng Yu, et al. Effect of grazing on the temperature sensitivity of soil respiration in Hulunber meadow steppe[J]. Envirmental Science, 2014, 35(5): 1909-1914. (in Chinese with English abstract)

[16] 李淑贞，徐大伟，范凯凯，等. 基于无人机与卫星遥感的草原地上生物量反演研究[J]. 遥感技术与应用，2022，37(1)：272-278.
Li Shuzhen, Xu Dawei, Fan Kaikai, et al. Research of grassland aboveground biomass inversion based on UAV and satellite remoting sensing[J]. Remote Sensing Technology and Application, 2022, 37(1): 272-278. (in Chinese with English abstract)

[17] Weiss M, Jacob F, Duveiller G. Remote sensing for agricultural applications: A meta-review[J]. Remote Sensing of Environment, 2020, 236: 111402.

[18] 郑阳，吴炳方，张淼. Sentinel-2 数据的冬小麦地上干生物量估算及评价[J]. 遥感学报，2017，21(2)：318-328.
Zheng Yang, Wu Bingfang, Zhang Miao. Estimating the above ground biomass of winter wheat using the Sentinel-2 data[J]. Journal of Remote Sensing, 2017, 21(2): 318-328. (in Chinese with English abstract)

[19] Gitelson A A, Kaufman Y J, Merzlyak M N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS[J]. Remote Sensing of Environment, 1996, 58(3): 289-298.

[20] Huete A, Didan K, Miura T, et al. Overview of the radiometric and biophysical performance of the MODIS vegetation indices[J]. Remote Sensing of Environment, 2002, 83(1): 195-213.

[21] Pinty B, Verstraete M M. GEMI: A non-linear index to monitor global vegetation from satellites[J]. Vegetatio, 1992, 101(1): 15-20.

[22] Buschmann C, Nagel E. In vivo spectroscopy and internal optics of leaves as basis for remote sensing of vegetation[J]. International Journal of Remote Sensing, 1993, 14(4): 711-722.

[23] Qi J, Chehbouni A, Huete A R, et al. A modified soil adjusted vegetation index[J]. Remote Sensing of Environment, 1994, 48(2): 119-126.

[24] Rouse J W, Haas R H, Deering D W, et al. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation[R]. 1974.

[25] Huete A R. A soil-adjusted vegetation index (SAVI)[J]. Remote Sensing of Environment, 1988, 25(3): 295-309.

[26] Birth G S, Mcvey G R. Measuring the Color of Growing Turf with a Reflectance Spectrophotometer[J]. Agronomy Journal, 1968, 60(6): 640-643.

[27] Gitelson A A, Kaufman Y J, Stark R, et al. Novel algorithms for remote estimation of vegetation fraction[J]. Remote Sensing of Environment, 2002, 80(1): 76-87.

[28] Verrelst J, Munoz J, Alonso L, et al. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3[J]. Remote Sensing of Environment, 2012, 118: 127-139.

[29] 蒙古国畜牧业统计数据[EB/OL]. 蒙古国国家统计局，[2022-11-20]. https: //www. 1212. mn/mn.

[30] 内蒙古自治区盟市统计数据[EB/OL]. 内蒙古自治区统计局，[2022-11-01]. http: //tj. nmg. gov. cn/datashow/index. htm.

[31] 黄健熙，黄海，马鸿元，等. 遥感与作物生长模型数据同化应用综述[J]. 农业工程学报，2018，34(21)：144-156.
Huang Jianxi, Huang Hai, Ma Hongyuan, et al. Review on data assimilation of remote sensing and crop growth models[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(21): 144-156. (in Chinese with English abstract)

[32] 马国林，丁建丽，韩礼敬，等. 基于变量优选与机器学习的干旱区湿地土壤盐渍化数字制图[J]. 农业工程学报，2020，36(19)：124-131.
Ma Guolin, Ding Jianli, Han Lijing, et al. Digital mapping of soil salinization in arid area wetland based on variable optimized selection and machine learning[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2020, 36(19): 124-131. (in Chinese with English abstract)

[33] 周俊宏，王子芝，廖声熙，等. 基于 GF-1 影像的普达措国家公园森林地上生物量遥感估算[J]. 农业工程学报，2021, 37(4): 216-223.
Zhou Junhong, Wang Zizhi, Liao Shengxi, et al. Remote

sensing estimation of forest aboveground biomass in Potatso National Park using GF-1 images[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2021, 37(4): 216-223. (in Chinese with English abstract)

[34] 杨雪峰，咎梅，木尼热·买买提. 基于无人机和卫星遥感的胡杨林地上生物量估算[J]. 农业工程学报，2021，37(1)：77-83.
Yang Xuefeng, Zan Mei, Munire·Maimaiti. Estimation of above ground biomass of Populus euphratica forest using UAV and satellite remote sensing[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2021, 37(1): 77-83. (in Chinese with English abstract)

[35] Maimaitijing M, Sagan V, Sisike P, et al. Soybean yield prediction from UAV using multimodal data fusion and deep learning[J]. Remote Sensing of Environment, 2020, 237: 111599.

[36] 付波霖，孙军，李雨阳，等. 基于多光谱影像和机器学习算法的红树林树种 LAI 估算[J]. 农业工程学报，2022，38(7)：218-228.
Fu Bolin, Sun Jun, Li Yuyang, et al. Mangrove LAI estimation based on remote sensing images and machine learning algorithms[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(7): 218-228. (in Chinese with English abstract)

[37] 武建双，李晓佳，沈振西，等. 藏北高寒草地样带物种多样性沿降水梯度的分布格局[J]. 草业学报，2012，21(3)：17-25.
Wu Jianshuang, Li Xiaojia, Shen Zhenxi, et al. Species diversity pattern of alpine grasslands communities along a precipitation gradient across Northern Tibetan Plateau[J]. Acta Prataculturae Sinica, 2012, 21(3): 17-25. (in Chinese with English abstract)

Estimating above-ground biomass in grassland using Landsat 8 and machine learning in Mongolian plateau

Zhao Yue, Xu Dawei, Fan Kaikai, Li Shuzhen, Shen Beibei, Shao Changliang, Wang Xu※, Xin Xiaoping

(*Institute of Agricultural Resources and Regional Planning*, *Chinese Academy of Agricultural Sciences/National Field Scientific Observation and Research Station of Hulunbuir Grassland Ecosystem in Inner Mongolia*, *Beijing* 100081, *China*)

Abstract: Above-Ground Biomass (AGB) is one of the most important indicators to reflect the status of grassland use. Accurate and rapid monitoring is of great significance to scientific management and rational use. Alternatively, remote sensing technology has been widely used to estimate the AGB in recent years. However, the estimation errors can often be caused by the common phenomenon of "same spectrum, different species" in remote sensing. One of the potential solutions can be to use the spectral and meteorological data to invert the AGB grassland. In this study, a machine learning model was developed to characterize the spectral indices and meteorological data using Landsat 8 remote sensing and ground survey as data sources. A systematic investigation was implemented to explore the performance of regression models constructed by five machine learning algorithms. Specifically, the AGB of grassland was estimated to obtain the high accuracy inversion of remote sensing for the grassland biomass. Nine vegetation indices were selected to calculate in Hulunbuir of Inner Mongolia and Dornod of Mongolia in China. An optimal Random Forest (RF) regression model was then reconstructed by feature selection. The regression validation revealed that a similar overall performance was achieved in the six machine learning models. But the lower performance was found in the spectral data as the input only (Root Mean Square Error (RMSE): 63.852-87.944 g/m^2, relative Root Mean Square Error (rRMSE): 33.712%-46.432%, coefficient of determination (R^2): 0.388-0.647). Furthermore, the error of all regression decreased gradually, as the number of features increased in the data combination. The model fitting ability increased gradually as well, indicating that the increasing number of features in the different regression models was effectively handled through the fusion of multiple data inputs. The best evaluation was obtained from each regression model in the data combination of spectra + precipitation + temperature. The RF also obtained the best performance (RMSE=51.702 g/m^2, rRMSE=27.297%, and R^2=0.749). The weights of the multiple source data in the model were determined to assess the relative importance of the input data. The results showed that the precipitation was the most important input feature of the model, with a maximum weight of more than 0.1, much higher than the other spectral data. Three vegetation indices of VARI, MSAVI, and GEMI in the spectral data were weighted more than 0.09 as the features, which was higher than the rest. The more stable performance was achieved in the optimized RF regression model, with a correlation coefficient (R^2) of 0.801 between predicted and measured values, an RMSE of 43.709 g/m^2, and an rRMSE of 23.077%. The AGB spatial distribution in the study area was lower in the central area, but higher on the east and west sides, with a maximum of 357.2 g/m^2 and a minimum of 33.01 g/m^2. It was closely related to the spatial heterogeneity of climate and grassland use patterns.

Keywords: remote sensing; inversion; machine learning;aboveground biomass; Mongolian plateau