DOI: 10.5846/stxb202203180669

张子慧,吴世新,赵子飞,李向义,曾凡江,谢聪慧,侯冠宇,罗格平.基于机器学习算法的草地地上生物量估测——以祁连山草地为例.生态学报,2022,42(22):8953-8963.

Zhang Z H, Wu S X, Zhao Z F, Li X Y, Zeng F J, Xie C H, Hou G Y, Luo G P.Estimation of grassland biomass using machine learning methods: A case study of grassland in Qilian Mountains.Acta Ecologica Sinica,2022,42(22):8953-8963.

基于机器学习算法的草地地上生物量估测
——以祁连山草地为例

张子慧[1,5],吴世新[1,*],赵子飞[2,4,5],李向义[1,3],曾凡江[1,3],谢聪慧[1,5],侯冠宇[1,5],罗格平[1]

1 中国科学院新疆生态与地理研究所荒漠与绿洲生态国家重点实验室,乌鲁木齐 830011

2 中国科学院太空应用重点实验室,北京 100094

3 中国科学院新疆生态与地理研究所新疆荒漠植物根系生态与植被修复重点实验室,乌鲁木齐 830011

4 中国科学院空间应用工程与技术中心,北京 100094

5 中国科学院大学,北京 100049

摘要:草地地上生物量(Aboveground Biomass,AGB)是指导畜牧业生产管理的重要指标,是草畜平衡综合分析的基础。目前,有关祁连山草地 AGB 反演的研究较少,且多源数据间的尺度差异问题并未得到很好的解决。为了解祁连山草地 AGB 的空间分布状况,利用 Sentinel-2 多光谱数据、无人机(Unmanned Aerial Vehicle,UAV)数据以及 2021 年植被生长期实测草地 AGB 数据实现了空天地一体化监测,通过决策树回归(Decision Tree Regression,DTR)、随机森林回归(Random Forest Regression,RFR)、梯度提升决策回归树(Gradient Boosting Regression Tree,GBRT)以及极致梯度提升(eXtreme Gradient Boosting,XGBoost)共 4 种算法反演草地 AGB 的适用性分析,利用最优模型反演了祁连山草地的 AGB 空间分布状况。结果表明:研究区内多种植被指数所表现出的特性有所差异。祁连山地区 AGB 在空间分布上呈现出由西北向东南递增的趋势,平均 AGB 为 925.43kg/hm^2。6 种植被指数与实测 AGB 之间均表现为显著正相关,适合作为祁连山草地 AGB 遥感反演的指标; XGBoost 模型较其它模型具有最高的 R^2值(0.78)和精度(74.75%)、最低的均方根误差(RMSE,99.74 kg/hm^2)和平均绝对误差(MAE,71.60 kg/hm^2),模型反演效果最好;UAV 数据能够提供更加详细的空间细节特征,减小 Sentinel-2 数据和实地采样数据间的尺度差异;因此,基于 6 种植被指数与祁连山草地 AGB 间的相关性,构建 XGBoost 模型反演研究区草地 AGB 空间分布状况是具有实践意义的。研究结果将为指导祁连山草地畜牧业的发展和维护草地生态系统的平衡提供一定的参考价值与数据支撑。

关键词:地上生物量;空天地一体化;草地;回归模型;祁连山

Estimation of grassland biomass using machine learning methods: A case study of grassland in Qilian Mountains

ZHANG Zihui[1,5], WU Shixin[1,*], Zhao Zifei[2,4,5], LI Xiangyi[1,3], ZENG Fanjiang[1,3], XIE Conghui[1,5], HOU Guanyu[1,5], LUO Geping[1]

1 *State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi* 830011, *China*

2 *Key Laboratory of Space Utilization, Chinese Academy of Sciences, Beijing* 100094, *China*

3 *Xinjiang Key Laboratory of Desert Plant Roots Ecology and Vegetation Restoration, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi* 830011, *China*

4 *Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing* 100094, *China*

基金项目:第二次青藏高原综合科学考察研究(2019QZKK0302);中国科学院战略性先导科技专项(A 类)项目(XDA23100201)

收稿日期:2022-03-18; **采用日期**:2022-08-29

* 通讯作者 Corresponding author.E-mail: wushixin@ ms.xjb.ac.cn

5 *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*

Abstract: Aboveground biomass (AGB) is an important indicator to guide the management of livestock industry, and it is the basis of comprehensive analysis of the balance between grassland and livestock. To date, only few studies have studied the spatial distribution of grassland AGB in Qilian Mountains, and the scale differences of multi-sources data have not been well solved. Therefore, in order to understand the spatial distribution of AGB in Qilian Mountains, we used the space-air-ground integrated method based on Sentinel-2 multispectral data, Unmanned Aerial Vehicle (UAV) data and the measured AGB data during the growth period of vegetation in 2021. In addition, we analyzed the applicability of Decision Tree Regression (DTR), Random Forest Regression (RFR), Gradient Boosting Regression Tree (GBRT) and eXtreme Gradient Boosting (XGBoost) algorithms for AGB inversion. Finally, we mapped the spatial distribution of AGB in the study area using the optimal model among all the models. The results verified that the effectiveness of vegetation indices varied in study area. Generally, the results indicated that the spatial distribution had an increasing trend from northwest to southeast, with an average AGB density of 925.43 kg/hm^2. A significantly positive correlation was found between 6 vegetation indices and measured AGB, and both of the indices was suitable for inversion of grassland AGB in the Qilian Mountains. Moreover, compared with other models, the performance of XGBoost model was the best, with the highest R^2 of 0.78 and accuracy of 74.75%, the lowest Root Mean Squared Error (RMSE) of 99.74 kg/hm^{2}and Mean Absolute Deviation (MAE) of 71.60 kg/hm^2. In addition, UAV data provided spatial characteristics in detail, which reduced the scale difference between Sentinel-2 and the measured data. Therefore, on the basis of the correlation between 6 vegetation indices and AGB, it is of practical significance to construct the XGBoost model to invert the spatial distribution of AGB in grassland of Qilian Mountains. The results can provide a reference value and data support for guiding the development of livestock industry and maintaining the balance of grassland ecosystem.

Key Words: aboveground biomass; space-air-ground integrated; grassland; regression model; Qilian Mountains

草地资源是我国陆地上面积最大的生态系统类型,具有防风、固沙、保土、调节气候、净化空气、涵养水源等生态功能,是草地畜牧业发展最重要的物质基础,对于全球生态环境、碳循环和维系生态平衡起重要的作用[1]。其时空分布格局可以反映草地植被的碳汇潜力[2]。草地地上生物量(aboveground biomass,AGB)通常用草地地上部分植被的干重表示,是全球碳循环的重要组成部分,是指导畜牧业生产管理的重要指标[3]。估算草地 AGB 是草地资源空间格局动态研究的重要内容,也是草畜平衡综合分析的基础,祁连山作为我国西部地区的重要生态保护屏障、黄河流域的重要水源产流地,也是我国生物多样性保护的优先区域,区域土地覆盖以草地为主,其生态保护事关我国西部地区的生态安全及经济社会可持续发展。及时准确地获取祁连山草地产量数据、草地植被生长状况、AGB 大小及其空间分布,对区域复合生态系统功能的科学评估具有参考价值,对草地退化机理研究与治理、指导草地畜牧业生产,维护生态服务功能和区域可持续发展具有重要意义[4—5]。

目前,草地 AGB 反演面临着诸多挑战,数据源方面,遥感影像具有探测范围大、成本低、容易获取等优点,为通过间接测量的、非破坏性的手段反演草地 AGB 现状,遥感影像已成为生物量反演的主要数据源,目前使用较多的为 MODIS、Landsat-8 和 Sentinel-2 影像数据[6—9],其中 Sentinel-2 数据具有更高的空间分辨率(10m),可以更好地与实测 AGB 数据相匹配,是目前常用的性能最好的星载遥感数据[7—11]。虽然卫星影像可以进行大范围的监测,考虑到卫星影像与实地采集样方间的尺度差异,有些细节特征还是不能突出,因此许多学者引入 UAV 数据用于减少多源数据间的尺度效应[12—15]。特征提取方面,国内外学者构建了一系列与 AGB 密切相关的植被指数用于提高草地 AGB 遥感反演的精度[10—11,16—18],比如增强型植被指数(Enhanced Vegetation Index,EVI)能够改善归一化植被指数(Normalized Difference Vegetation Index,NDVI)在高植被覆盖区的饱和现象[16]、转换型土壤调节植被指数(Transformed Soil-adjusted Vegetation Index, TSAVI)对土壤背景值敏感[17]、比值植被指数(Ratio Vegetation Index, RVI)在植被覆盖度高的情况下对植被十分敏感[18]等,不同

植被指数能够反映不同的植被特征,根据研究区植被特征选取合适的植被指数用于 AGB 反演建模是极其重要的。反演算法方面,机器学习算法成为当前常用的构建模型的方法,许多学者利用数理方法构建草地 AGB 参数反演模型[19—24],且已被证实适用于草地 AGB 的反演中。目前常用的算法包括线性回归算法和非线性回归算法两大类,其中线性回归算法多以一元线性回归和多元线性回归为主[19—20],非线性回归算法多以支持向量机回归(Support Vector Regression, SVR)[21]和 DTR[22]、RFR[23]、GBRT[22]和 XGBoost[24]等基于决策树形成的回归算法为主,反演 AGB 过程中可能会遇到变量之间的关系并不是线性的,利用线性回归算法构建模型时受到限制,SVR 算法可以用于解决非线性回归问题,但容易出现过拟合现象且计算复杂度高、可解释性不好,基于决策树的算法不仅可以用于线性或非线性的回归问题中,还能够很好的避免过拟合现象,是目前应用最为广泛的反演算法。DTR 模型解释性好,可以使用图形化描述模型[22];RFR 是用训练数据随机计算出多颗决策树,实现简单,训练速度快,模型泛化能力强[23];GBRT 对异常值的鲁棒性非常强,在相对较少的调参时间下,预测的准确度较高[22];XGBoost 在代价函数中加入了正则项,能够控制模型的复杂度,可以防止过拟合现象还能够减少计算量[24]。因此,基于 4 种算法的独特优势,本文探讨了这 4 种算法反演祁连山草地 AGB 的适用性。以往的 AGB 反演研究取得了长足的进展,提出了许多新的方法和理论,综合以上内容依旧存在着三个重要问题,一是如何减少多源数据间的尺度差异;二是如何选取适合研究区内植被特点的植被指数;三是如何选择合适的反演算法用于构建草地 AGB 反演模型。

针对以上三个问题,本文以祁连山作为研究区,考虑到 UAV 数据具有实时性好、分辨率高、机动性高等优点[24—25]以及空天地一体化监测技术的发展趋势,本文将利用 Sentinel-2 数据、UAV 数据和地面实测数据实现祁连山草地 AGB 的空天地一体化监测,在实地考察的基础上,选取了 EVI、RVI、TSAVI 等在内的 6 种植被指数作为模型构建的输入特征,比较分析了 DTR、RFR、GBRT 和 XGBoost 4 种算法反演祁连山草地 AGB 空间分布状况的适用性,利用最优反演模型进行草地 AGB 空间分布制图,为祁连山草地生态系统的稳定发展、保护生物多样性提供参考依据。

1　研究区概况及数据处理

1.1　研究区概况

研究对象为祁连山的草地生态系统,位于青藏高原东北部(图 1,94°33′10.8″—100°39′7.2″E, 36°51′25.2″—39°54′37.2″N),是河西走廊地区重要的生态保护屏障和主要的水资源涵养地[26]。研究区内平均海拔约为 3800m,占地面积约为 128300km^2,年平均温度为 0.6℃,年平均降水量约为 400—700mm[27]。光照充足,冬季长而寒冷,夏季短而温和,干湿分明,常年受西风环流的影响,属于典型的高原大陆性气候[28]。水热条件随海拔的升高而发生变化,气候类型和植被类型表现为明显的垂直地带性差异[29]。主要植被类型有木本猪毛菜(*Salsola arbuscula* Pall.)、高山嵩草(*Kobresia Pygmaea* C. B. Clarke)、针茅(*Stipa capillata* L.)、鬼箭锦鸡儿(*Caragana Jubata* (Pall.) Poir.)、二裂委陵菜(*Potentilla bifurca* Linn.)等[30]。主要草地类型有荒漠草原、高寒荒漠草原、高寒苔草草原、高寒嵩草草甸、高寒针茅草原等[31—33]。

1.2　数据来源与处理

1.2.1　地面实测 AGB 数据

祁连山草地实测 AGB 数据获取时间为 2021 年 7 月 20 日—8 月 10 日,正值植被生长期。共布设 23 个能够代表祁连山草原植被、地形及土壤等特征的采样区(如图 1 所示),按一定方向在样地中间位置设 100m×100m 样线(图 2),记录每个采样区的基本信息,包括经度、纬度、海拔高度、盖度、株高等。为方便与 Sentinel-2 影像数据的分辨率(10m)相匹配,保证样方在采样区内覆盖主要草地类型且均匀分布,沿样线布置 5 个 1m×1m 样方,每个样方间隔距离为 20m(图 2),共计 115 个样方。收集样方内所有植物的地上部分,称其鲜重后在实验室内置于 105℃下杀青,采用 85℃烘干至恒重,获得每个样方的草干重,分别计算每个样方的生物量,获取 115 个样方的 AGB 实测数据。用于构建祁连山草地 AGB 反演的训练集和测试集。经实地考察,研究区

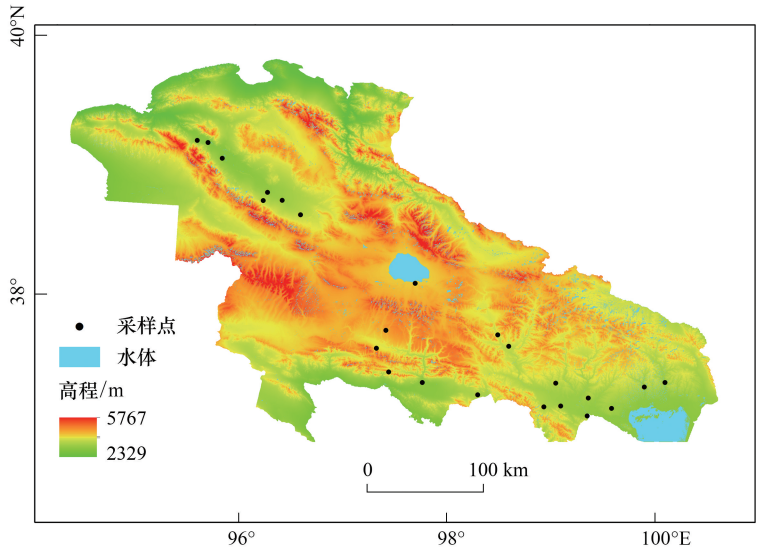

图 1 研究区示意图

Fig.1 Overview of Qilian Mountains in study area

群落由西北至东南方向具有明显差异,西北区域植被较稀疏,以荒漠草原为主,东南区域植被分布均匀且覆盖度均达 85%以上,以嵩草草原为主。

1.2.2 图像数据与处理

Sentinel-2 是高分辨率多光谱成像卫星,覆盖可见光、近红外等 13 个波段,其中包含 3 个红边波段(表 1),能够反映植被的反射率光谱特征,重采样后分辨率可达到 10m[6—9]。本研究中所需的 Sentinel-2 遥感产品通过欧洲航天局(European Space Agency, ESA)官网(https://scihub.copernicus.eu/dhus/#/home)下载云覆盖率低于 10%、与实地采样时间相近的 21 景影像,以及 2016 年植被生长期的影像。

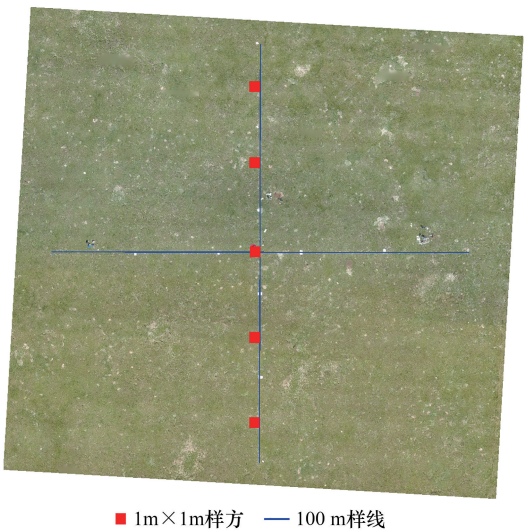

图 2 采样区示意图

Fig.2 Schematic diagram of sampling area

UAV 数据通过大疆 Phantom 4 Pro 获取,无人机按照 DJI GO 4 设定的飞行路线自动飞行至覆盖整个采样区,设置航向重叠率为 80%、旁向重叠率为 70%,飞行高度为 15m,并按一定的时间间隔(1.5s)垂直拍摄地表,共拍摄 23 个样地的 UAV 数据,空间分辨率为 0.4cm。基于 Sentinel-2 影像数据对 UAV 数据进行图像方向配准、尺度配准,提取两种影像数据特征向量,从而将 UAV 数据配准至 Sentinel-2 数据。UAV 数据作为联系 Sentinel-2 遥感影像和 1m×1m 实地样方的媒介,可以在一定程度上减少 Sentinel-2 影像数据和地面样方之间存在的尺度差异。研究区 DEM 数据来源于美国地质勘探局(United States Geological Survey,USGS)提供的 30m 分辨率 SRTM 数据(https://lpdaac.usgs.gov/products/srtmgl1v003/)。

利用 SNAP、ArcGIS 软件对 Sentinel-2 数据进行预处理,包括重采样、镶嵌、裁剪等(图 3)。使用 Pix4D Mapper 对 UAV 数据进行拼接、地理位置叠加、空三加密等操作,得到数字正射影像,最后导入 ENVI 图像处理软件中进行影像裁剪等处理(图 3)。

表 1　Sentinel-2 MSI 数据的参数信息

Table 1　The characteristic of Sentinel-2 MSI

波段名称 Band name	中心波长/nm Central wavelength	波段宽度/nm Bandwidth	空间分辨率/m Resolution
1-海岸/气溶胶波段 1-Coastal/Aerosol	443	20	60
2-蓝色波段 2-Blue	490	65	10
3-绿色波段 3-Green	560	35	10
4-红色波段 4-Red	665	30	10
5-红边波段 1 5- Vegetation red edge 1	705	15	20
6-红边波段 2 6- Vegetation red edge 2	740	15	20
7-红边波段 3 7- Vegetation red edge 3	783	20	20
8-近红外波段(宽) 8- NIR	842	115	10
8A-近红外波段(窄) 8A-Narrow NIR	865	20	20
9-水蒸气波段 9-Watervapour	945	20	60
10-短波红外波段 1 10-SWIR 1	1375	30	60
11-短波红外波段 2 11-SWIR 2	1610	90	20
12-短波红外波段 3 12-SWIR 3	2190	180	20

NIR:近红外波段 Near infrared;SWIR:短波红外波段 Shortwave infrared

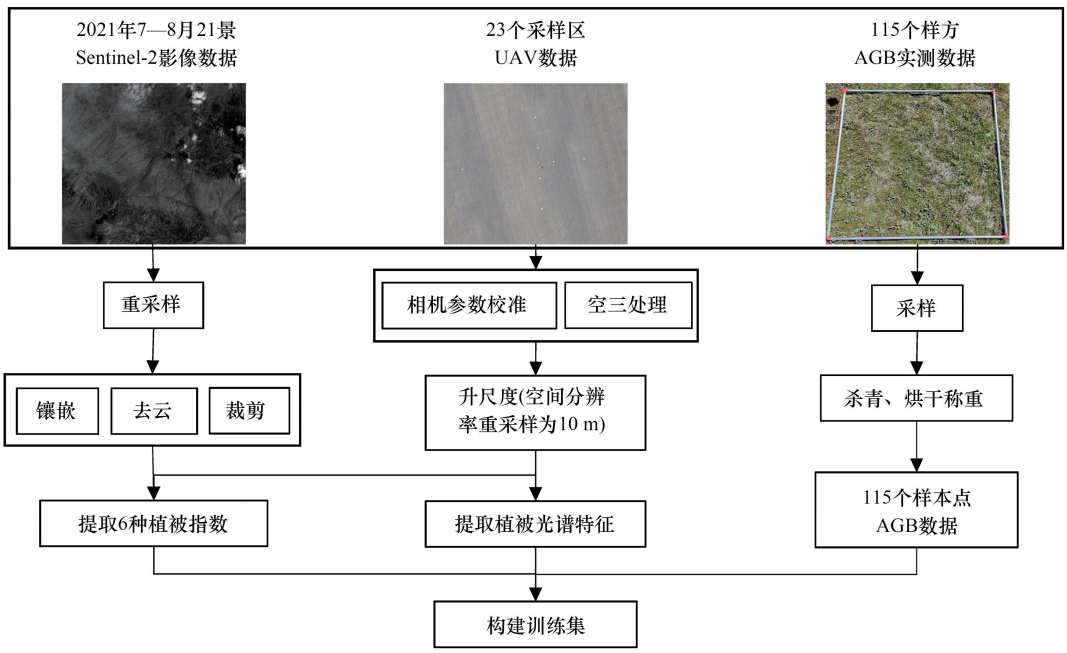

图 3　数据处理流程图

Fig.3　The process of multi-source data

1.2.3　植被指数选取

为了证实 Sentinel-2 数据在反演 AGB 方面的潜力,比较分析相关研究,在 UAV 数据辅助下,对处理后 Sentinel-2 影像数据提取了 EVI、RVI、红边归一化植被指数(Red Edge Normalized Difference Vegetation Index, $NDVI_{705}$)、TSAVI、红边简单比值(Red Edge Simple Ratio, SR_{re})和改进型简单比值(Modified Simple Ratio, MSR)6 种植被指数(表 2),用于反演祁连山 AGB 的空间分布。

2　研究方法

建立祁连山草地 AGB 反演模型,进行模型的适用性分析,技术路线图见图 4,图中 y、$y-y_1$、$y-y_1-y_2-\cdots$

$-y_{n-1}$为单颗决策树的相关 AGB 输入值;y_1、y_2、y_n、$y_1+\Omega(f_1)$、$y_2+\Omega(f_2)$、$y_1+\Omega(f_n)$分别为单颗决策树的 AGB 输出值;F(x)为算法的输出值。

表 2 选取的植被指数

Table 2 Selected vegetation indices

植被指数 Vegetation index	公式 Formula	参考文献 References
增强型植被指数 EVI	$2.5\times(R_{nir}-R_{red})/(R_{nir}+6R_{red}-7R_{blue}+1)$	[5]
比值型植被指数 RVI	R_{nir}/R_{red}	[11]
红边归一化植被指数 $NDVI_{705}$	$(R_{red-edge2}-R_{red-edge1})/(R_{red-edge1}+R_{red-edge2})$	[13]
转换型土壤调节植被指数 TSAVI	$0.5\times(R_{nir}-0.5R_{red}-0.2)/(0.5R_{nir}+0.5R_{red}+0.1)$	[13]
红边简单比值 SR_{re}	$R_{nir}/R_{red-edge1}$	[26]
改进型简单比值 MSR	$(R_{nir}/R_{red}-1)/\sqrt{R_{nir}/R_{red}+1}$	[26]

EVI:增强型植被指数 Enhanced vegetation index;RVI:比值型植被指数 Ratio vegetation index;NDVI705:红边归一化植被指数 Red edge normalized difference vegetation index;TSAVI:转换型土壤调节植被指数 Transformed soil-adjusted vegetation index;SRre:红边简单比值 Red edge simple ratio;MSR:改进型简单比值 Modified simple ratio

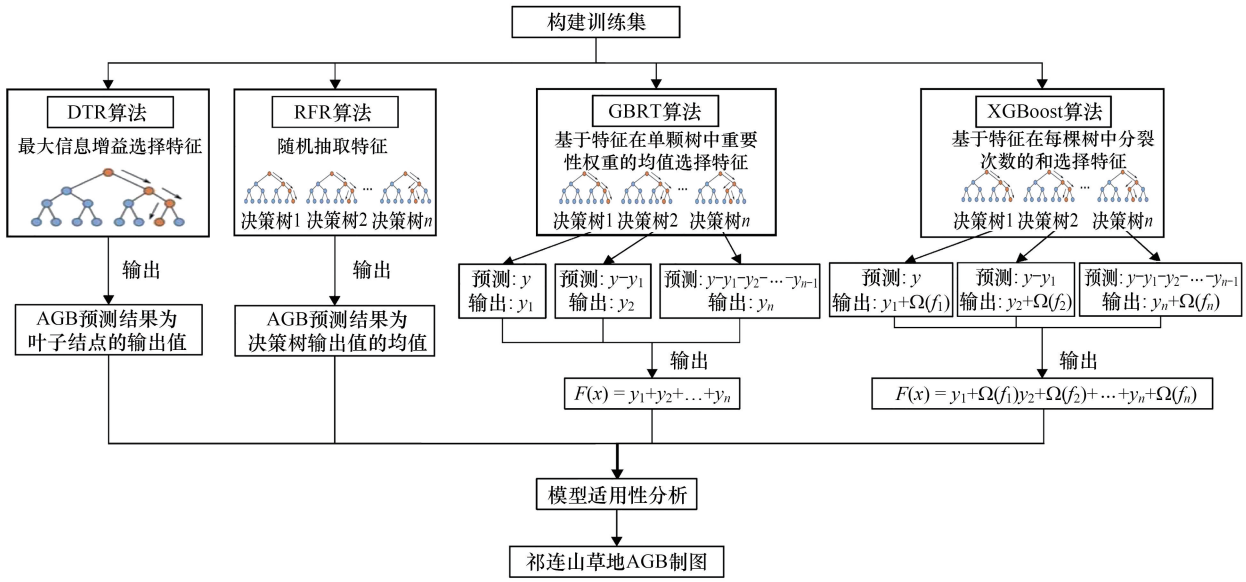

图 4 技术路线图

Fig.4 Technology roadmap

AGB: 草地地上生物量

本文利用了 DTR、RFR、CBRT 和 XGBoost 4 种算法构建祁连山草地 AGB 反演模型。这 4 种算法均基于分类与回归树(Classification and Regression Trees, CART)实现预测,CART 采用二分递归分割技术,通过构建一个二叉树将样本进行递归划分,使用样本的最小方差作为分裂节点的依据。

2.1 回归模型算法

DTR 在训练数据集所在的输入空间中,递归地将每个区域划分为两个子区域并决定每个子区域上的输出值,构建二叉决策树,根据样本中不同特征的信息纯度形成树的各个结点,输入样本数据流经整颗决策树,最终在叶子结点输出最终的结果[22]。RFR 作为一种集成学习方法,采用了 Bagging 的思想,在原始训练样本集中有放回地重复随机抽取新的训练样本集合,基于多颗决策树引入了随机属性选择对数据进行预测,输出所有决策树输出的平均值,可以避免单颗决策树的局限性,通过降低模型方差来提高性能[21,23]。GBRT 基于 CART 决策树及 Boosting 技术的集成学习算法,以串行的方式建立多颗决策回归树,是一种迭代的决策树算

法,能够抑制决策树的复杂性,降低单颗决策树的拟合能力,再通过梯度提升的方法集成多个决策树,使用一阶泰勒公式优化模型,通过降低残差来提高性能[22,34]。XGBoost 基于 GBRT 进行了改进,在传统的 Boosting 的基础上,使用了二阶泰勒公式优化模型,并在代价函数里加入了正则化项(公式(1)),能够降低模型的方差,用于控制模型的复杂度,使学习出来的模型更加简单,提高模型泛化能力[24,35]。

$$\Omega(ft) = \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{1}$$

式中:$\Omega(ft)$ 为第 t 颗树的正则化项;T 为第 t 棵树的叶子结点数;w_j 为叶子结点 j 的权重向量 $l2$ 范数;Υ 和 λ 为 XGBoost 算法的参数。

2.2　精度评定

交叉验证的基本思想是将原始数据进行分组,其中 80%的数据作为训练集,20%的数据作为测试集,首先用训练集对回归器进行训练,再利用验证集来测试训练得到的模型,作为评价回归器的性能指标,使用交叉验证是为了得到可靠稳定的模型,常见的交叉验证形式有 K 折交叉验证、留一交叉验证等。

为了分析 4 种算法反演祁连山草地 AGB 的适用性,本文利用 5 折交叉验证对模型进行评估,利用平均绝对误差(Mean Absolute Deviation,MAE,公式(2))、均方根误差(Root Mean Squared Error,RMSE,公式(3))和决定系数(R-squared,R^2,公式(4)) 3 个指标[21—24]用于模型精度评定。

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - f_i|}{n} \tag{2}$$

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - f_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{4}$$

式中:y_i 为实测 AGB;f_i 为预测 AGB;$\bar{y}$ 为实测 AGB 的均值;n 为样本个数。

3　结果与分析

3.1　变量相关系数矩阵分析

计算了植被指数与实测 AGB 间的 Pearson 相关系数(图 5),发现 AGB 与各植被指数均表现为显著正相关($P \leqslant 0.01$)。其中,EVI、TSAVI 和 AGB 的相关性相对较低;TSAVI 与其他植被指数的相关性也较低。造成该结果的原因是祁连山草地植被覆盖度较高,受土壤背景影响低,而 EVI 和 TSAVI 更适用于低覆盖区的植被研究中,能够降低土壤背景的影响,提高对植被的敏感性,主要与 EVI 和 TSAVI 本身所反映的植被特征有关。

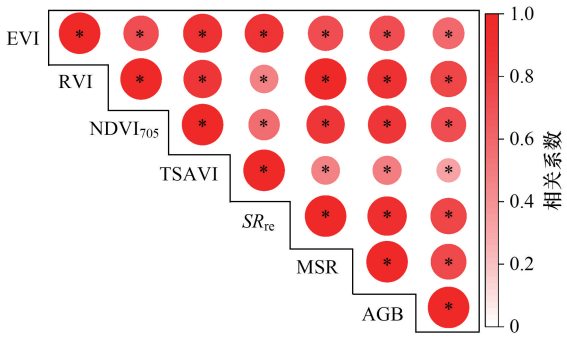

图 5　植被指数与 AGB 的相关系数矩阵

Fig.5　Correlation coefficient matrix between AGB and vegetation indices

* 显著相关,$P \leqslant 0.01$;EVI:增强型植被指数;RVI:比值型植被指数;NDVI705:红边归一化植被指数;TSAVI:转换型土壤调节植被指数;SRre:红边简单比值;MSR:改进型简单比值

3.2　模型的构建与因子重要性评估

本文评估了 4 种算法反演祁连山草地 AGB 的因子重要性(图 6),每种算法中各因子重要性表现也有其异同点。在 RFR 模型中,各因子间重要性差异较小,利用该模型反演 AGB 时考虑的因子较多,模型复杂度较高;

而 DTR 模型中,各因子间重要性差异较大,利用该模型反演 AGB 时考虑的因子较少,可能会出现影响因子考虑不全的结果。RVI 在 4 种算法中均占较高权重,对于反演祁连山 AGB 状况有着不可代替的作用;并且 RVI 对土壤背景变化的敏感性较弱,在植被覆盖度较高的情况下对植被信息反映敏感,表明祁连山草地的覆盖度较高; EVI 和 TSAVI 2 种植被指数的重要性均较低也能够表明研究区植被覆盖度较高的现状,因为这两个植被指数更适合对低覆盖度的植被监测,所表现出的结果与它们和 AGB 间的相关性也有关。在 XGBoost 算法中,降低了 EVI 的影响,说明该植被指数不适用于祁连山草地 AGB 的反演,不能有效地反映出研究区草地的植被特征;明显突出了 SR_{re} 的重要性,该指数与近红外与红边波段信息有关,这两个波段对植被信息反映比较敏感。

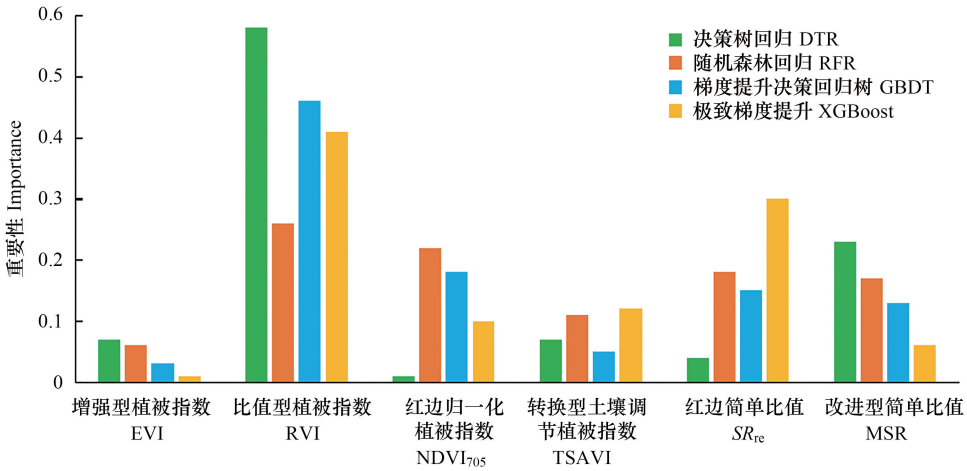

图 6 AGB 与各植被指数的因子重要性评估结果

Fig.6 The results of importance assessment between AGB and vegetation indices in study area

3.3 模型适用性评价

通过比较 4 种算法中反演祁连山草地 AGB 最优模型的 MAE、RMSE、R^2 和模型精度进行模型的适用性评价(表 3)。其中,XGBoost 模型具有最低的 MAE(71.60 kg/hm^2)和 RMSE(99.74 kg/hm^2),最高的 R^2(0.78)和精度(74.75%),反演效果最好,其次分别为 GBRT、RFR、DTR。因为 RFR、GBRT 和 XGBoost 3 种算法均是在 DTR 的基础上进行了改进,其中,RFR 采用的是 bagging 思想,而 GBRT 采用的是 boosting 思想,以串行的方式建立决策树;GBRT 在优化的时候只用到一阶导数信息,XGBoost 在代价函数中加入了正则项,与其他算法有其独特的先进性。XGBoost 算法中突出的 SR_{re} 植被指数,更适用于祁连山草地 AGB 的反演,对植被信息反映敏感,因此本文将选用 XGBoost 算法对祁连山草地 AGB 进行反演,用于进一步分析其空间分布状况。

表 3 4 种回归模型算法的适用性比较

Table 3 Applicability comparison of 4 regression model algorithms

回归模型算法 Regression model algorithms	平均绝对误差 MAE (kg/hm^2)	均方根误差 RMSE (kg/hm^2)	决定系数 R^2	精度 Accuracy/%
决策树回归 DTR	99.87	140.78	0.52	67.75%
随机森林回归 RFR	86.29	117.14	0.67	70.36%
梯度提升决策回归树 GBRT	81.45	113.53	0.69	72.59%
极致梯度提升 XGBoost	71.60	99.74	0.78	74.75%

DTR:决策树回归 Decision tree regression;RFR:随机森林回归 Random forest regression;GBRT:梯度提升决策回归树 Gradient boosting regression tree;XGBoost:极致梯度提升 eXtreme gradient boosting

3.4 地上生物量的空间分布

综合上述模型适用性的分析,本文利用 XGBoost 反演祁连山草地 AGB,总体上呈现出由西北向东南增加

的趋势(图 7),平均 AGB 为 925.43 kg/hm^2,AGB 分布主要受到水热条件的影响,靠近水源的区域 AGB 更大,水分条件差的区域 AGB 较低。东南区域靠近青海湖,属于高原大陆性气候,光照充足,具有良好的适宜植被生长的水热条件,同时在植被生长峰值期间受到西南季风的影响,带来的部分水汽适宜植被生长,因而 AGB 较高;西北区域水分条件较差,受季风影响弱,主要受到山间径流的影响,使得 AGB 沿水资源分布状况而发生变化,与东南区域相比,AGB 较低。

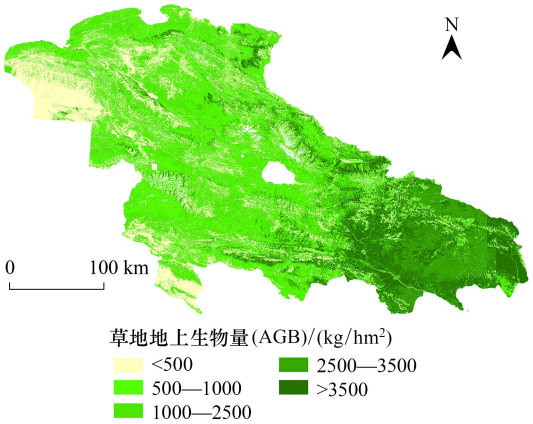

图 7　祁连山草地 AGB 空间分布图

Fig.7　Spatial distribution of AGB in grassland of Qilian Mountains

4　讨论

上述研究结果表明,在 UAV 数据的基础上,利用 Sentienl-2 影像数据提取出的植被指数,选用 XGBoost 模型反演祁连山草地 AGB 空间分布的方法具有很强的应用潜力,在交叉验证的基础上,模型精度也在可接受范围,与其他研究相比[23],本文针对祁连山草地 AGB 的空间制图结果是可信的,该研究能够提供一定的可信度与理论支撑,但依旧存在四类重要问题值得讨论分析。

4.1　降水量的年度波动对 AGB 反演的影响

在空间分布上受到降水量差异影响,草地 AGB 表现出空间分异性,由于研究区内降水量存在年际差异,大范围下草地 AGB 状况会存在一定的滞后性,以往研究结论也提出相关论证[36—37]。本研究及将在后续工作中继续探讨降水量的年波动是否会在很大程度上影响 XGBoost 模型反演祁连山草地 AGB 空间分布的效果,以及祁连山草地 AGB 呈现出怎样的动态变化模式。

4.2　草地地上生物量反演模型的不确定性与误差分析

本文在反演祁连山草地 AGB 过程中主要存在以下 3 各个方面的不确定性:首先是多源数据间的尺度效应[11,17],Sentinel-2 数据的空间分辨率为 10m,UAV 数据分辨率为 0.4cm,实地采集样方大小为 1m×1m,使用 Sentinel-2 数据和 UAV 数据协同反演祁连山草地 AGB 能够在一定程度上提高模型精度,但依旧会存在一些处理上以及算法上的误差。其次是在研究区内采样点布设较少,既对研究区 AGB 的反演会存在一定误差[20,38];也可能会造成过拟合现象,由于 XGBoost 算法中添加了正则化项来控制模型的复杂度,与其它 3 种算法相比,能够提高模型反演的精度,但不能完全消除反演时存在的高估和低估问题[19]。第三是本文中所使用的 Sentinel-2 数据由于受到云覆盖的限制,选择了与采样时间相近的影像用于研究,虽然时间不能完全匹配,但造成的这种不确定性对于祁连山草地 AGB 变化的影响是可以忽略不计的。

4.3　空天地一体化的耦合效应

利用 UAV 数据作为实现空天地一体化监测技术的中间媒介,将 Sentinel-2 影像数据和地面采样数据进行像元尺度的匹配,尤其是对于西北区域,建群种为木本猪毛菜,散布在该区域且自身生物量较重,导致在实地采样过程中会受到较大影响,而 Sentinel-2 数据的像元尺度是 10m,计算出的植被指数为 10m×10m 范围内的均值,本研究中通过提取 UAV 数据的特征波段,分析其与 Sentinel-2 影像数据间的差异,发现 UAV 数据能够观测到 Sentinel-2 数据所不能表现出的细节特征。因此,UAV 数据的引入能够降低匹配 Sentinel-2 数据和实地样方间的误差,该结论与孙世泽等[20]、刘沁茹等[39]的研究结果一致。UAV 数据可以减少尺度匹配过程中的误差,但这种误差是不能完全消除的,UAV 数据获取过程中会受到 GPS 误差的影响,以及对 UAV 数据处理时也会受到系统误差的影响。

4.4　XGBoost 模型反演草地 AGB 的适用性

本文所选用的 XGBoost 模型被证实适用于反演研究区草地 AGB 状况。在空间分布适用性上,XGBoost 模

型精度为 74.75%,MAE 为 71.60 kg/hm^2,RMSE 为 99.74 kg/hm^2,R^2为 0.78。除多等[19]基于 MODIS 数据,利用 GIS 空间数据处理技术和多元统计分析方法反演草地 AGB,R^2介于 0.1295—0.6929 之间。黄家兴等[40]基于 Sentinel-2 植被指数数据,利用二次曲线回归模型反演天祝县草地 AGB,R^2介于 0.4019—0.6983 之间,郭超凡等[41]基于 Sentinel-2 影像数据,利用随机森林回归模型反演草地 AGB,R^2为 0.74。与前人的研究结果相比,本研究所利用的 XGBoost 模型在反演祁连山草地 AGB 中具有一定的优势,反演效果较好,模型适用性较高。

5 结论

本文以 Sentinel-2 遥感影像计算的 6 种植被指数为特征,结合实拍 UAV 数据以及地面的 115 个样本点实测 AGB 数据,采用 DTR、RFR 、GBRT 和 XGBoost 4 种算法进行了祁连山草地地上生物量的遥感反演,以期为绿色推动祁连山草地畜牧业发展、维护生态平衡与生物多样性提供科学依据和理论支持。主要结论如下:

(1)6 种植被指数均与实测 AGB 表现为显著正相关,其中 RVI、$NDVI_{705}$、SR_{re}和 MSR 这 4 个植被指数与 AGB 的相关性更强,主要因为它们是由红色、红边和近红外波段对 AGB 较敏感的波段信息构成的;通过评估 6 种植被指数的因子重要性,RVI 和 SR_{re}在祁连山草地 AGB 反演中占据重要地位,RVI 本身对高覆盖区植被监测敏感,红边波段(SR_{re})的加入能够在一定程度上提高 AGB 反演模型的精度,减少植被在红色波段的饱和效应。

(2)通过对 4 种反演算法进行适用性评价,得出 XGBoost 算法反演祁连山草地 AGB 的效果最好,具有最强的适用性,MAE 为 71.60 kg/hm^2、RMSE 为 99.74 kg/hm^2,R^2为 0.78、模型精度为 74.75%,且适用于祁连山草地 AGB 的长时间序列反演。

(3)UAV 数据的应用能够减少 Sentinel-2 数据(10m)与实地采样(1m×1m)间尺度效应的影响,UAV 数据可以提供更详细的空间细节特征,便于 Sentinel-2 影像中提取出的植被指数与地面实测 AGB 数据相匹配。

(4)利用 XGBoost 算法对祁连山草地 AGB 进行反演制图,祁连山草地平均 AGB 为 925.43 kg/hm^2,其空间分布上表现为由西北向东南递增的趋势,主要原因是受到水资源分布的影响。

参考文献(References):

[1] 袁晓波, 牛得草, 吴淑娟, 蒲向东, 王龙, 滕家明, 傅华. 黄土高原典型草原地上生物量估测模型. 生态学报, 2016, 36(13): 4081-4090.

[2] Thurner M, Beer C, Santoro M, Carvalhais N, Wutzler T, Schepaschenko D, Shvidenko A, Kompter E, Ahrens B, Levick S R, Schmullius C. Carbon stock and density of northern boreal and temperate forests. Global Ecology and Biogeography, 2014, 23(3): 297-310.

[3] 沈海花, 朱言坤, 赵霞, 耿晓庆, 高树琴, 方精云. 中国草地资源的现状分析. 科学通报, 2016, 61(2): 139-154.

[4] 张福平, 王虎威, 朱艺文, 张枝枝, 李肖娟. 祁连县天然草地地上生物量及草畜平衡研究. 自然资源学报, 2017, 32(7): 1183-1192.

[5] Anaya J A, Chuvieco E, Palacios-Orueta A. Aboveground biomass assessment in Colombia: a remote sensing approach. Forest Ecology and Management, 2009, 257(4): 1237-1246.

[6] 田艳林, 刘贤赵, 毛德华, 王宗明, 李延峰, 高长春. 基于 MODIS 数据的松嫩平原西部芦苇湿地地上生物量遥感估算. 生态学报, 2016, 36(24): 8071-8080.

[7] Li C H, Zhou L Z, Xu W B. Estimating aboveground biomass using Sentinel-2 MSI data and ensemble algorithms for grassland in the Shengjin lake wetland, China. Remote Sensing, 2021, 13(8): 1595.

[8] Wang J, Xiao X M, Bajgain R, Starks P, Steiner J, Doughty R B, Chang Q. Estimating leaf area index and aboveground biomass of grazing pastures using Sentinel-1, Sentinel-2 and Landsat images. ISPRS Journal of Photogrammetry and Remote Sensing, 2019, 154: 189-201.

[9] Sibanda M, Mutanga O, Rouget M. Comparing the spectral settings of the new generation broad and narrow band sensors in estimating biomass of native grasses grown under different management practices. GIScience & Remote Sensing, 2016, 53(5): 614-633.

[10] 高明亮, 赵文吉, 宫兆宁, 赫晓慧. 基于环境卫星数据的黄河湿地植被生物量反演研究. 生态学报, 2013, 33(2): 542-553.

[11] Ren H R, Zhou G S. Estimating green biomass ratio with remote sensing in arid grasslands. Ecological Indicators, 2019, 98: 568-574.

[12] 车荧璞, 王庆, 李世林, 李保国, 马韫韬. 基于超分辨率重建和多模态数据融合的玉米表型性状监测. 农业工程学报, 2021, 37(20): 169-178.

[13] 刘杨, 冯海宽, 孙乾, 杨福芹, 杨贵军. 不同分辨率无人机数码影像的马铃薯地上生物量估算研究. 光谱学与光谱分析, 2021, 41(5):

1470-1476.

[14] 杨雪峰, 眷梅, 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. 农业工程学报, 2021, 37(1): 77-83.

[15] 刘艳慧, 蔡宗磊, 包妮沙, 刘善军. 基于无人机大样方草地植被覆盖度及生物量估算方法研究. 生态环境学报, 2018, 27(11): 2023-2032.

[16] 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. 遥感学报, 2017, 21(2): 318-328.

[17] Xu D W, Wang C, Chen J, Shen M G, Shen B B, Yan R R, Li Z W, Karnieli A, Chen J Q, Yan Y C, Wang X, Chen B R, Yin D M, Xin X P. The superiority of the normalized difference phenology index (NDPI) for estimating grassland aboveground fresh biomass. Remote Sensing of Environment, 2021, 264: 112578.

[18] 蒋馥根, 孙华, 李成杰, 马开森, 陈松, 龙江平, 任蓝翔. 联合 GF-6 和 Sentinel-2 红边波段的森林地上生物量反演. 生态学报, 2021, 41(20): 8222-8236.

[19] 除多, 德吉央宗, 姬秋梅, 唐红. 西藏高原典型草地地上生物量遥感估算. 国土资源遥感, 2013, 25(3): 43-50.

[20] 孙世泽, 汪传建, 尹小君, 王伟强, 刘伟, 张雅, 赵庆展. 无人机多光谱影像的天然草地生物量估算. 遥感学报, 2018, 22(5): 848-856.

[21] 邢晓语, 杨秀春, 徐斌, 金云翔, 郭剑, 陈昂, 杨东, 王平, 朱立博. 基于随机森林算法的草原地上生物量遥感估算方法研究.地球信息科学学报, 2021, 23(7): 1312-1324.

[22] Zhang Y Z, Ma J, Liang S L, Li X S, Liu J D. A stacking ensemble algorithm for improving the biases of forest aboveground biomass estimations from multiple remotely sensed datasets. GIScience & Remote Sensing, 2022, 59(1): 234-249.

[23] Zeng N, Ren X L, He H L, Zhang L, Zhao D, Ge R, Li P, Niu Z E. Estimating grassland aboveground biomass on the Tibetan Plateau using a random forest algorithm. Ecological Indicators, 2019, 102: 479-487.

[24] Li Y C, Li M Y, Li C, Liu Z Z. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. Scientific Reports, 2020, 10(1): 9952.

[25] 刘杨, 黄珏, 孙乾, 冯海宽, 杨贵军, 杨福芹. 利用无人机数码影像估算马铃薯地上生物量. 遥感学报, 2021, 25(9): 2004-2014.

[26] 钱大文, 曹广民, 杜岩功, 李茜, 郭小伟. 2000—2015 年祁连山南坡生态系统服务价值时空变化. 生态学报, 2020, 40(4): 1392-1404.

[27] 陈京华, 贾文雄, 赵珍, 张禹舜, 刘亚荣. 1982—2006 年祁连山植被覆盖的时空变化特征研究. 地球科学进展, 2015, 30(7): 834-845.

[28] 贾文雄, 赵珍, 俎佳星, 陈京华, 王洁, 丁丹. 祁连山不同植被类型的物候变化及其对气候的响应. 生态学报, 2016, 36(23): 7826-7840.

[29] 付建新, 曹广超, 郭文炯. 1998—2017 年祁连山南坡不同海拔、坡度和坡向生长季 NDVI 变化及其与气象因子的关系. 应用生态学报, 2020, 31(4): 1203-1212.

[30] 朱平, 陈仁升, 宋耀选, 韩春坛, 刘光琇, 陈拓, 张威. 祁连山中部 4 种典型植被类型土壤细菌群落结构差异. 生态学报, 2017, 37(10): 3505-3514.

[31] 高延锋, 仪律北, 张法伟, 马文婧, 李红琴, 王春雨, 罗方林, 杨永胜, 李英年. 祁连山东段 4 类草地生态系统 CO_2通量与叶面积指数的关系研究. 中国草地学报, 2022, 44(5): 1-8.

[32] 杨学亭, 樊军, 盖佳敏, 杜梦鸽, 金沐. 祁连山不同类型草地的土壤理化性质与植被特征. 应用生态学报, 2022, 33(4): 878-886.

[33] 赵文, 尹亚丽, 李世雄, 刘燕, 刘晶晶, 董怡玲, 苏世锋, 吉凌鹤. 祁连山不同类型草地土壤细菌群落特征研究. 草业学报, 2021, 30(12): 161-171.

[34] 刘悦. 基于梯度提升决策树与支持向量机融合模型的成矿预测研究[D]. 北京: 中国地质大学(北京), 2020.

[35] 张亦然, 刘廷玺, 童新, 段利民, 吴宇辰. 基于 XGBoost 算法的草甸地上生物量的高光谱遥感反演. 草业学报, 2021, 30(4): 1-12.

[36] 单楠. 京津风沙源区植被指数(NDVI)对气候变化响应研究[D]. 北京: 中国林业科学研究院, 2013.

[37] 刘成林, 樊任华, 武建军, 闫峰. 锡林郭勒草原植被生长对降水响应的滞后性研究. 干旱区地理, 2009, 32(4): 512-518.

[38] Morais T G, Teixeira R F M, Figueiredo M, Domingos T. The use of machine learning methods to estimate aboveground biomass of grasslands: a review. Ecological Indicators, 2021, 130: 108081.

[39] 刘沁茹, 孙睿. 森林生物量遥感降尺度研究. 生态学报, 2019, 39(11): 3967-3977.

[40] 黄家兴, 吴静, 李纯斌, 秦格霞, 钱娟冰, 李怀海. 基于 Sentinel-2 和 Landsat 8 数据的天祝县草地地上生物量遥感反演. 草地学报, 2021, 29(9): 2023-2030.

[41] 郭超凡, 陈泽威, 张志高. 基于最优模型选择的牧草地上生物量遥感估算研究. 草地学报, 2021, 29(5): 946-955.