耦合机器学习和机载高光谱数据的土壤含水量估算

田美玲[1,2,3**]，葛翔宇[1,2,3]，丁建丽[1,2,3*]，王敬哲[1,2,3]，张振华[1,2,3]

[1]新疆大学资源与环境科学学院，新疆 乌鲁木齐 830046；

[2]新疆大学绿洲生态教育部重点实验室，新疆 乌鲁木齐 830046；

[3]新疆大学智慧城市与环境建模自治区普通高校重点实验室，新疆 乌鲁木齐 830046

摘要 准确估算土壤含水量(SMC)，对干旱区的精准农业、水资源管理具有重要意义。针对传统估算方法和野外测量耗时、费力的问题，通过无人机平台获取新疆阜康市冬小麦样地的高光谱影像数据，分别利用一阶导数、二阶导数、吸光度、吸光度一阶导数(FDA)、吸光度二阶导数对原始高光谱数据进行预处理；采用随机森林(RF)、梯度提升回归树(GBRT)和极端梯度提升(XGBoost)三种算法进行特征变量重要性遴选，基于地理加权回归(GWR)建立模型。结果表明：FDA 的预处理效果最佳，以 FDA-GBRT 为基础的模型效果最优，建模集与验证集的决定系数(R^2)分别为 0.890、0.891，四分位数间隔为 3.490；GBRT 算法相较于 RF 和 XGBoost 算法优势较为突出，多数模型建模集与验证集的 R^2 均大于 0.600；GWR 模型对 SMC 的预测建模有效，可为干旱区农业生态系统的管理与保护提供理论支撑。

关键词 土壤含水量；无人机；高光谱数据；机器学习；地理加权回归模型

中图分类号 O433 **文献标志码** A **doi**：10.3788/LOP57.093002

Coupled Machine Learning and Unmanned Aerial Vehicle Based Hyperspectral Data for Soil moisture Content Estimation

Tian Meiling[1,2,3**], Ge Xiangyu[1,2,3], Ding Jianli[1,2,3*], Wang Jingzhe[1,2,3], Zhang Zhenhua[1,2,3]

[1] *College of Resource and Environment Sciences, Xinjiang University, Urumqi, Xinjiang* 830046, *China*;

[2] *Key Laboratory of Oasis Ecology, Xinjiang University, Urumqi, Xinjiang* 830046, *China*;

[3] *Key Laboratory of Smart City and Environment Modelling of Higher Education Institute, Xinjiang University, Urumqi, Xinjiang* 830046, *China*

Abstract Accurate estimation of soil moisture content (SMC) is of great significance for precision agriculture and water resources management in arid areas. Traditional estimation methods and field measurements are time consuming and labor intensive. Therefore, we obtain hyperspectral image data of winter wheat plots in Fukang City, Xinjiang by unmanned aerial vehicle platform, and the original hyperspectral data are preprocessed through first derivative, second derivative, absorbance, first derivative of absorbance (FDA), and second derivative of absorbance. Random forest (RF), gradient boosted regression tree (GBRT), and extreme gradient boost (XGBoost) are used to select the importance of feature variables. A model is established based on geographical weighted regression (GWR). The results show that the pretreatment effect of FDA is the best. The model based on FDA-GBRT is optimal. The determination coefficient (R^2) of the modeling set and the verification set are 0.890 and 0.891, respectively, and the quartile interval reaches 3.490. Compared with RF and XGBoost algorithms, the advantages of the GBRT algorithm are more prominent. The R^2 of most of the model modeling set and the verification set are greater than 0.600. This indicates that the GWR model is effective in predictive modeling of SMC and can provide theoretical support for the management and protection of agro ecosystem in arid regions.

Key words soil moisture content; unmanned aerial vehicle; hyperspectral data; machine learning; geographical weighted regression model

收稿日期：2019-09-04；**修回日期**：2019-09-09；**录用日期**：2019-09-16

基金项目：国家自然科学基金(41771470，41961059)、新疆教育厅自然科学重点项目(XJEDU2018I008)

* **E-mail**：watarid@xju.edu.cn；** **E-mail**：tianmeiling_0911@163.com

OCIS codes 300.6320; 350.6980

1 引 言

土壤含水量(SMC)是评价地表作物生长发育的关键指标,也是制约干旱区绿洲作物生长的主要因素,同时会影响植被的生物物理和化学结构[1]。传统的 SMC 监测耗时、费力,且成本较大,遥感技术作为一种快速、简洁、无损的探测技术在 SMC 反演监测中被广泛应用[2]。近几年无人机(UAV)衍生的高光谱技术发展迅速,使大规模、高效率获取 SMC 信息成为可能。原始的 UAV 高光谱数据信息量大[3],且存在噪声和信息冗余,一定程度上增加了挖掘信息的难度。通常 SMC 与土壤的光谱间存在非线性、异方差性等复杂关系。为解决这些问题,引入预处理方案消除外部噪声,增强光谱特征,解译非线性关系,从而提高模型对特定目标的估算精度[4-5]。

机器学习为特征波段的选择提供了较为理想的方法,其中集成学习在数据挖掘中具有重要意义,可在一定程度上提高预测模型的估算精度[6]。随机森林(RF)算法在特征选择上应用广泛,苗松等[7]利用 RF 算法以哨兵 3A-OLCI 影像作为自变量进行重要性分析,捕获到对藻蓝蛋白影响最大的 3 个波段;王娜等[8]采用 RF 算法和单变量特征选择结合的方法提升了对遥感影像的分类精度;Zamani 等[9]利用 RF 算法对德黑兰市区 $PM_{2.5}$ 特征重要性进行预测,效果较好。近年来提出的梯度提升回归树(GBRT)和极端梯度提升(XGBoost)算法在相关应用中崭露头角,但其特征变量的优选鲜有报道。因此本文引入 RF、GBRT 和 XGBoost 三种集成算法进行敏感波段筛选,找出最优方案。

地理加权回归(GWR)是一种局部模型,通过不同空间子集受空间变化影响的自变量与因变量之间的关系构建模型[10-11],建模过程中融入了地理位置信息,广泛用于空间非平稳性领域,但在植被-土壤属性空间预测中报道较少。基于此,本文利用 UAV 遥感平台获取高光谱遥感影像,在光谱一阶导数(FDR)、光谱二阶导数(SDR)、吸光度(A)等 6 种预处理方案的基础上用 RF、GBRT 和 XGBoost 三种算法优选敏感波段,结合 GWR 模型构建该空间区域的 SMC 估算模型,获得研究区域内土壤墒情监测的最优模型,为 SMC 监测和农田灌溉管理提供了理论依据和技术支持。

2 材料与方法

2.1 土壤样本采集

研究区为新疆维吾尔自治区阜康市(87°51′15″E,44°21′14″N),该地区位于阜康绿洲北缘的古尔班通古特沙漠过渡带[12],年平均降水量不足200 mm,是典型的温带大陆性沙漠气候。采样区田块内的作物为冬小麦,且采样时间为返青期,植被覆盖度较高,土壤样品的采集与 UAV 空中作业同步进行,且范围相同。将采样区均匀分割为 70 个 0.5 m×0.5 m 的小区田,依据四点采样法对植株周围各点进行采集,采样深度为 0~10 cm,利用 GPS 获取各采样点的地理信息。土样经室内烘干法得到 70 个 SMC 数据,用联合 X-Y 距离算法(SPYX)对 SMC 数据进行建模集与验证集的划分,用于后期建模与验证。

2.2 UAV 高光谱数据获取

用六旋翼无人机(大疆创新科技有限公司 Matrice 600 Pro)搭载的高光谱传感器(Headwall 公司 Nano-Hyperspec)进行数据采集[13],范围为 400~1000 nm,高度为 100 m,焦距为 12 nm,空间分辨率为 4 cm。在 2018-04-17T15:00 时刻采集高光谱遥感数据,作业前对传感器进行暗电流矫正和白板矫正,并在晴朗无风、视野良好的情况下进行采集。测量前 5 d 内无降水、无人工干扰,保证数据的准确性。高光谱数据处理和校正分别在 Hyperspec Ⅲ 和 Headwall Spectral View 软件中完成。

2.3 数据预处理

利用 Savitzky-Golay(SG)方法对获取的高光谱图像进行平滑处理,以去除传感器自身带来的噪声[14]。光谱分析领域中一阶导数(FD)、二阶导数(SD)、吸光度(A)、连续统去除(CR)是有效的预处理方法,一定程度上可消除背景噪声,从而增强光谱特征。实验基于 IDL+ENVI5.3 平台,将 SG 滤波后的图像作为预处理的原始图像 R,通过光谱变换得到 FDR、SDR、CR、A、吸光度一阶导数(FDA)、吸光度二阶导数(SDA)的高光谱图像,同时计算每个采样单元光谱数据的平均值,为后续特征波段筛选、建模做准备,实验中的 SG 平滑在 Matlab R2016b 中实现。

2.4 建模分析及验证

考虑到 GBRT 和 XGBoost 算法性能比较相

似，且后者是在前者基础上改进的，均属于集成学习的预测模型；而 RF 与 GBRT、XGBoost 算法虽存在差异，但也具有很强的代表性。因此利用这三种集成算法筛选特征波段，根据重要性排序将前 20 个特征波段构建的 SMC 预测模型作为输入 GWR 模型的自变量。

RF 算法是一种基于决策树的集合学习算法，一定程度上可以平衡误差，相对简单。同时在参数优化、变量排序以及后续变量分析解释等方面优势明显，且能够充分利用样本数据[7-8,15]。实验设置决策树的数量 $n_{tree}=500$，节点数 $n_{try}=5$。

GBRT 算法是一种迭代的决策树算法，由多棵决策树组成，不断迭代直到决策树的个数达到预先给定的条件，在此过程中模拟各变量间的相互作用并根据变量的重要性进行排序，将所有树的结论作为最终输出结果[16]。GBRT 中也需要调整各项参数，为了满足对比条件，设置决策树的数量 $n_{tree}=500$。

XGBoost 算法是 2015 年提出的一种基于 GBDT 改进的算法，可有效构造增强树并运行、并行计算、近似建树及对稀疏数据进行有效处理[9]。与 GBRT 算法相比，该算法不再使用一阶导数信息，而基于二阶泰勒公式展开，能够提高输入特征变量重要性排序最优解的效率。实验设置迭代次数 $n_{round}=500$。

首先利用 RF、GBRT 和 XGBoost 算法在 R3.5.0平台内置的重要性函数计算各波段的函数值；其次根据函数值大小排列特征波段，函数值越大，表明该波段对 SMC 预测模型的影响程度越大。为了评估基于三种算法选取的特征变量建立的 GWR 模型的优劣程度，选取决定系数(R^2)、均方根误差(RMSE)和四分位数间隔(RPIQ)评价 21 种模型实测 SMC 与预测建模的效果和性能[17]。其中，R^2 值越大，模型的精度越高；RMSE 表示预测能力，大小与 R^2 成反比；RPIQ 作为一种预测指标已广泛应用于评估预测模型的准确性中，当RPIQ≥2.2 时，模型具有极佳预测能力；当 1.4≤RPIQ<2.2 时，模型预测能力比较均衡；当 RPIQ<1.4 时，模型可信度低。

3 结果与分析

3.1 SMC 统计分析

图 1 为 SMC 样本数据统计特征，采集的土壤全样本 SMC 范围为 12.230%～37.630%，平均 SMC 为 24.464%，标准差(S.D.)为 5.408。该冬小麦种植区域内土壤表层水分差异较大，受周围环境影响比较明显；建模集 SMC 的范围为 12.230%～37.630%，平均 SMC 为 25.017%，S.D.为 5.894；验证集 SMC 的范围为 14.950%～28.560%，平均 SMC 为 23.081%，S.D.为 3.714。通过 SPXY 算法在建模集和验证集中均保持了与全集 SMC 相似的统计分布结果，尽可能减小建模集和验证集中可能存在偏差的估计[12]。

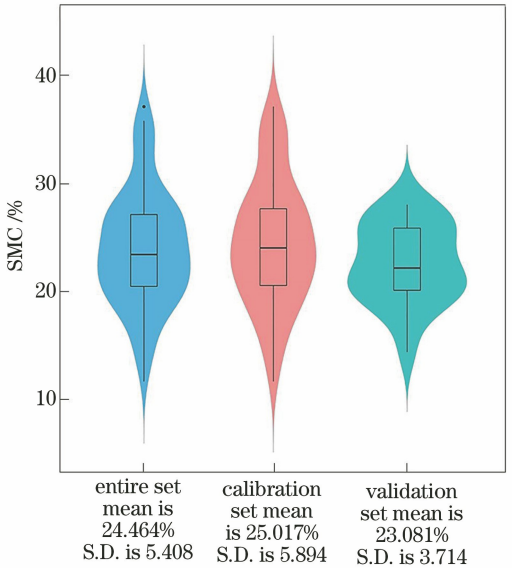

图 1 SMC 数据统计性特征

Fig. 1 Statistical characteristics of SMC

3.2 预处理后高光谱影像

预处理后得到 R、FDR、SDR、CR、A、FDA 和 SDA 共 7 种不同的高光谱影像及其光谱曲线分别如图 2、图 3 所示。由图 2 可以看出经过 SG 平滑后的 R、CR 和 A 的高光谱图像差异较小且均能反映地面真实情况。光谱微分是增加信噪比的主要手段之一，其中 FD、SD 可去除背景噪声，修复基线漂移[6]，图 2(c)、图 2(g)、图 2(d)、和图 2(h)是经过 FD 和 SD 后得到的高光谱图像，随阶数倒数的增加，反而掩盖了地面的真实情况，且噪声出现的频次也增加，值域缩小。图 3 中的曲线表示平均光谱曲线，围绕着平均光谱曲线的部分表示标准差范围。

3.3 集成学习变量优选

使用上述 3 种算法的重要性函数，分别以原始光谱波段和 6 种预处理后的波段作为重要性估计的输入变量，7 种光谱经 3 种算法筛选出重要性排序前 20 的波段如图 4 所示，图中 B001、B002、B003 等表示高光谱中对应的波段。可以发现当输入为 R 时，3 种方法筛选出的特征波段虽有差异，但也有重合。重要性排名靠前的均是波长为420～450 nm

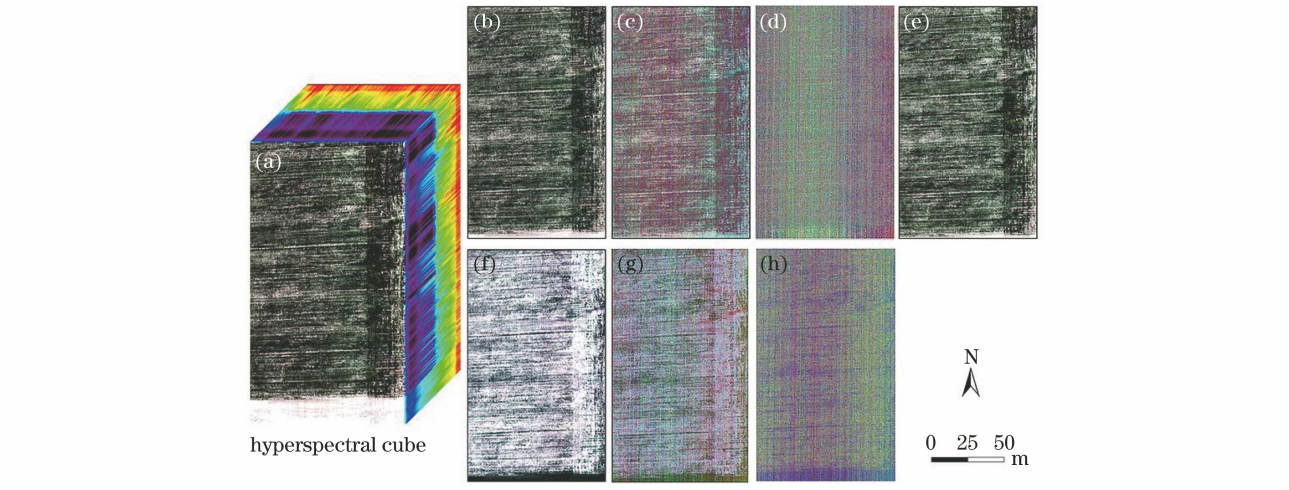

图 2 不同预处理后的高光谱影像。(a)三维图;(b) R;(c) FDR;(d) SDR;(e) CR;(f) A;(g) FDA;(h) SDA

Fig. 2 Hyperspectral images based on different pretreatments. (a) Three-dimensional image; (b) R; (c) FDR; (d) SDR; (e) CR; (f) A; (g) FDA; (h) SDA

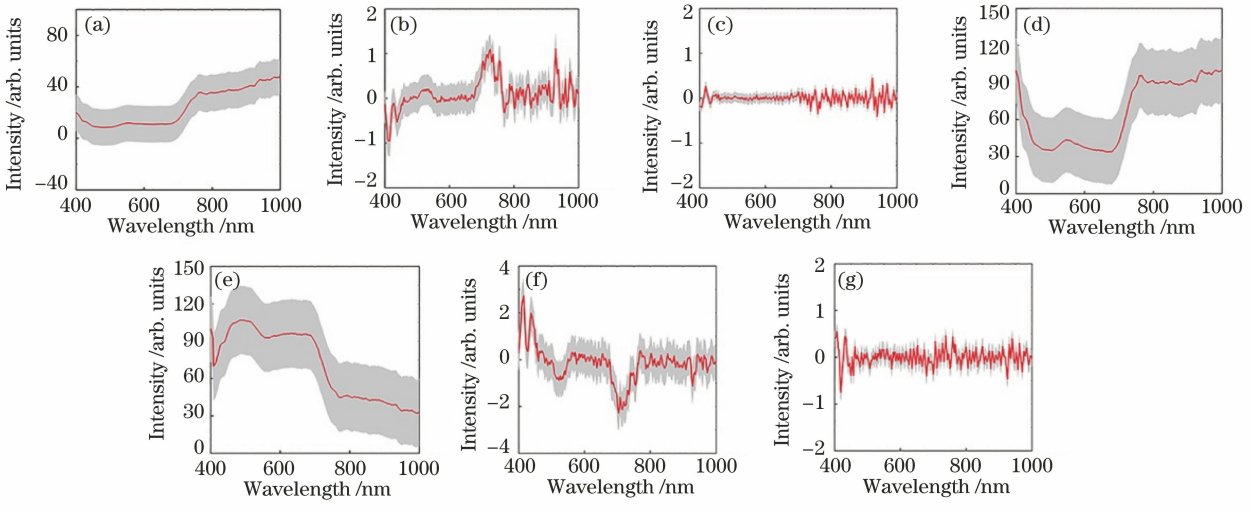

图 3 不同预处理后的光谱曲线。(a) R;(b) FDR;(c) SDR;(d) CR;(e) A;(f) FDA;(g) SDA

Fig. 3 Spectral curves based on different pretreatments. (a) R; (b) FDR; (c) SDR; (d) CR; (e) A; (f) FDA; (g) SDA

的蓝波段,其次是波长为 490～60 nm 的绿波段和 670～760 nm 的红边波段。以 RF 算法进行重要性筛选时未将波长为 892～1001 nm 的近红外波段纳入前 20,且红边波段占比较少,这表明不同算法在权衡特征变量重要性时考虑的因素有所差异。此外,对比 GBRT 算法以 R、FDR 和 SDR 为输入变量筛选出的特征波段可知,R 经 GBRT 算法(R-GBRT)筛选出的特征波段集中在蓝波段、红边波段和的近红外波段,而 FDR 经 GBRT 算法(FDR-GBRT)和 SDR 经 GBRT 算法(SDR-GBRT)筛选出的特征波段相较于 R-GBRT 增加了绿波段、红边波段和近红外波段的数量,这表明不同预处理方式对光谱波段信息筛选的影响不同,且筛选出的红边和近红外波段数量随预处理导数阶数的增加而增加。

3.4 GWR 建模分析与精度评价

GWR 是一种预测土壤属性的局部回归方法,在最小二乘法回归(OLS)模型上进行了改进,融入采样点的空间位置。当输入的自变量空间位置发生变化时,自变量的系数也随之改变。OLS 模型可表示为

$$y_i=\beta_0+\sum_{i=1}^{k}\beta_k x_{ik}+\varepsilon_i, \tag{1}$$

扩展后的 GWR 模型可表示为

$$y_i=\beta_0(u_i,v_i)+\sum_{i=1}^{k}\beta_k(u_i,v_i)x_{ik}+\varepsilon_i, \tag{2}$$

式中,y_i 为采样点的因变量,x_{ik} 为第 i 个采样点上对应的第 k 个变量的实测值,(u_i,v_i) 为采样点对应的空间坐标,$\beta_0(u_i,v_i)$ 为回归常数项,$\beta_k(u_i,v_i)$ 为

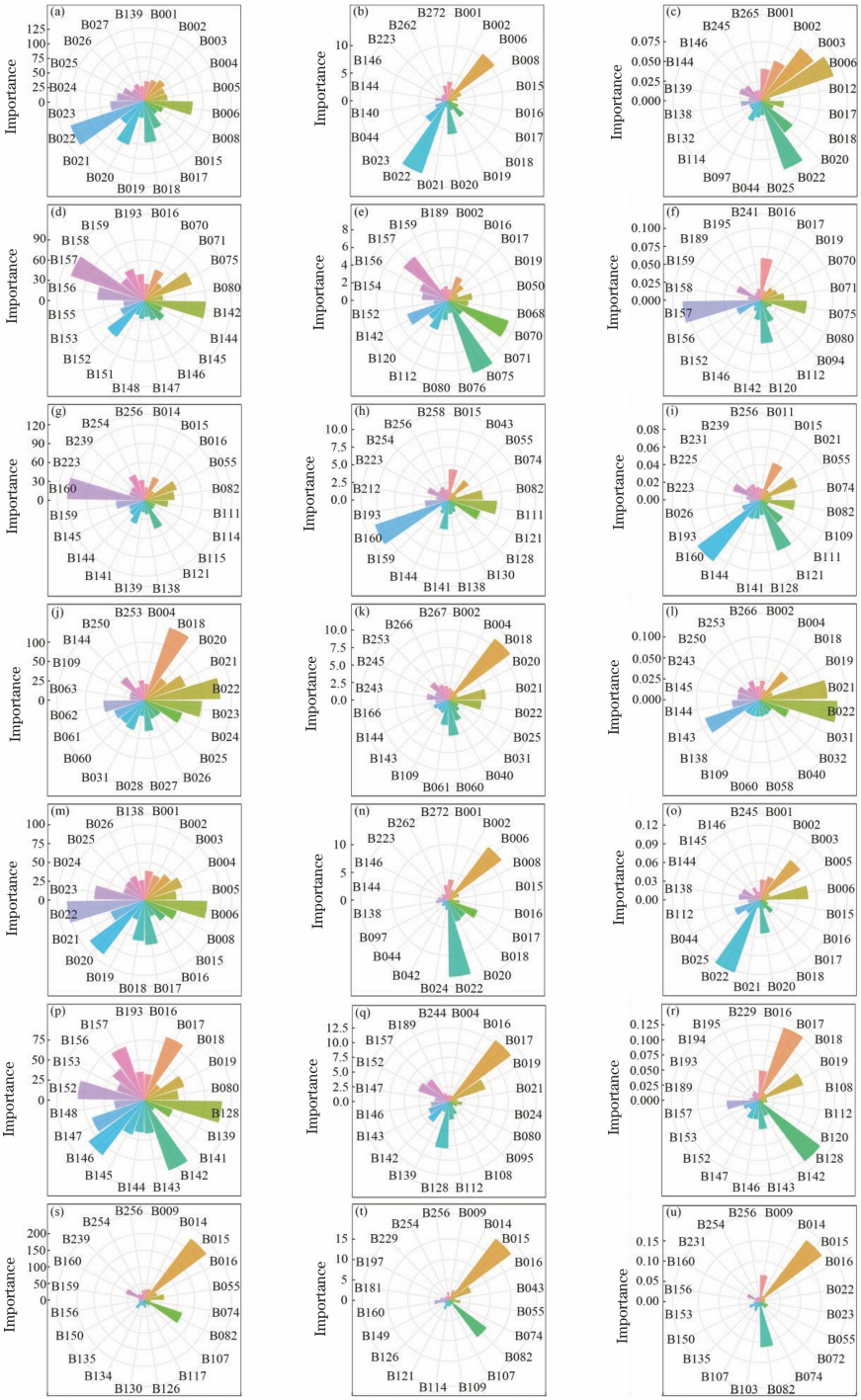

图 4 基于不同算法筛选的特征波段。(a)～(c) R 经 RF、GBRT、XGBoost 筛选后的特征波段;(d)～(f) FDR 经 RF、GBRT、XGBoost 筛选后的特征波段;(g)～(i) SDR 经 RF、GBRT、XGBoost 筛选后的特征波段;(j)～(l) CR 光谱经 RF、GBRT、XGBoost 筛选后的特征波段;(m)～(o) A 经 RF、GBRT、XGBoost 筛选后的特征波段;(p)～(r) FDA 经 RF、GBRT、XGBoost 筛选后的特征波段;(s)～(u) SDA 经 RF、GBRT、XGBoost 筛选后的特征波段

Fig. 4 Characteristic bands selected by different algorithms. (a)-(c) Characteristic bands of R after RF, GBRT, XGBoost screening; (d)-(f) characteristic bands of FDR after RF, GBRT, XGBoost screening; (g)-(i) characteristic bands of SDR after RF, GBRT, XGBoost screening; (J)-(l) characteristic bands of CR after RF, GBRT, XGBoost screening; (m)-(o) characteristic bands of RF, GBRT, XGBoost screening; (p)-(r) characteristic bands of FDA after RF, GBRT, XGBoost screening; (s)-(u) characteristic band of SDA after RF, GBRT, XGBoost screening

第 i 个采样点上对应的第 k 个回归参数，ε_i 为误差项。实验以 3 种方法优选后的 21 组最优 UAV 高光谱波段作为自变量，利用 GWR 模型对 SMC 进行回归建模，得到的结果如表 1 所示。

表 1 不同优选方法下最优特征变量与 SMC 的 GWR 模型

Table 1 GWR model of optimal variable SMC under different preferred methods

Independent variable	Modeling set		Validation set		
	R^2	RMSE /%	R^2	RMSE /%	RPIQ
R-RF	0.690	3.307	0.694	2.068	1.682
R-GBRT	0.700	3.214	0.698	2.019	1.890
R-XGBoost	0.653	3.440	0.657	2.230	1.410
FDR-RF	0.621	3.614	0.621	2.237	1.401
FDR-GBRT	0.800	2.624	0.801	1.654	3.007
FDR-XGBoost	0.771	2.802	0.772	1.764	2.943
SDR-RF	0.712	3.132	0.712	2.065	1.895
SDR-GBRT	0.744	2.973	0.745	1.90	2.724
SDR-XGBoost	0.690	3.268	0.692	2.563	1.688
CR-RF	0.726	3.062	0.724	1.932	2.212
CR-GBRT	0.681	3.312	0.680	2.202	1.436
CR-XGBoost	0.688	3.276	0.689	2.322	1.483
A-RF	0.694	3.239	0.692	2.076	1.724
A-GBRT	0.685	3.280	0.688	2.191	1.437
A-XGBoost	0.690	3.257	0.691	2.053	1.588
FDA-RF	0.842	2.434	0.843	1.454	3.115
FDA-GBRT	0.890	2.024	0.890	1.337	3.490
FDA-XGBoost	0.764	2.852	0.764	1.835	2.801
SDA-RF	0.599	3.727	0.598	2.317	1.362
SDA-GBRT	0.738	2.998	0.740	1.881	2.315
SDA-XGBoost	0.860	2.285	0.861	1.632	3.238

3 种不同集成算法学习的 GWR 模型估算值与实测值如图 5 所示，其中直线表示拟合线，阴影部分表示置信区间。对比不同变量的建模效果可知，以 FDA 为基础的 GBDT 算法筛选的波段建模效果最佳，其建模集的 R^2 为 0.890，RMSE 为 2.024%，验证集的 R^2 为 0.890，RMSE 为 1.337%。按照 RPIQ 预测分级标准，FDA-GBRT 模型预测能力最佳，RPIQ 为 3.490。其次为 SDA-XGBoost 和 FDR-GBRT，建模集和预测集的 $R^2 \geqslant 0.800$，且 RPIQ 大于 3.000。

从图 5 可以发现，以 R、FDR 和 SDR 为基础时，GBRT 算法的建模效果为最佳，且建模集和验证集的 R^2 最大，预测效果较好；以 CR 和 A 为基础时，RF 算法建模效果最佳，均具有较好的预测性能；以 SDA 为基础时，XGBoost 算法的建模效果最佳，预测精度较高。

当输入相同时，在 RF 算法中，建模效果从大到小依次为 FDA-RF、CR-RF 和 SDR-RF；在 GBRT 算法中，建模效果从大到小依次为 FDA-GBRT、FDR-GBRT 和 SDR-GBRT，且建模预测效果均优于 RF 算法；在 XGBoost 算法中，建模效果从大到小依次为 SDA-XGBoost、FDR-XGBoost 和 FDA-XGBoost。

综上所述，3 种算法中 GBRT 算法表现最优，且在 FDA 基础上筛选的敏感波段建模效果最佳。此外，以 SDA 为基础的 RF 算法所筛选的特征波段建模集和验证集 R^2 均未超过 0.600，其余模型建模集 R^2 均大于 0.680，其中 FDA-GBRT、SDA-XGBoost、FDA-RF 以及 FDR-GBRT 的建模集 R^2 均达到 0.800，这表明 GWR 模型在 SMC 预测建模中是有效的。

4 讨 论

特征波段分布如图 6 所示，可以发现基于 RF、GBRT 和 XGBoost 三种特征波段选择算法得到的敏感波段大多集中于蓝光波段和红边波段。原因是中心波长为 450 nm 的蓝光波段和中心波长为 670～760 nm 的红边波段为植被叶绿素的显著吸收波段；此外，植被水分显著吸收波段也集中在中心波长450 nm处[17-19]，原因是红边波段对植被的应力敏感性较强，对土壤背景影响较敏感。这与Ge等[12]

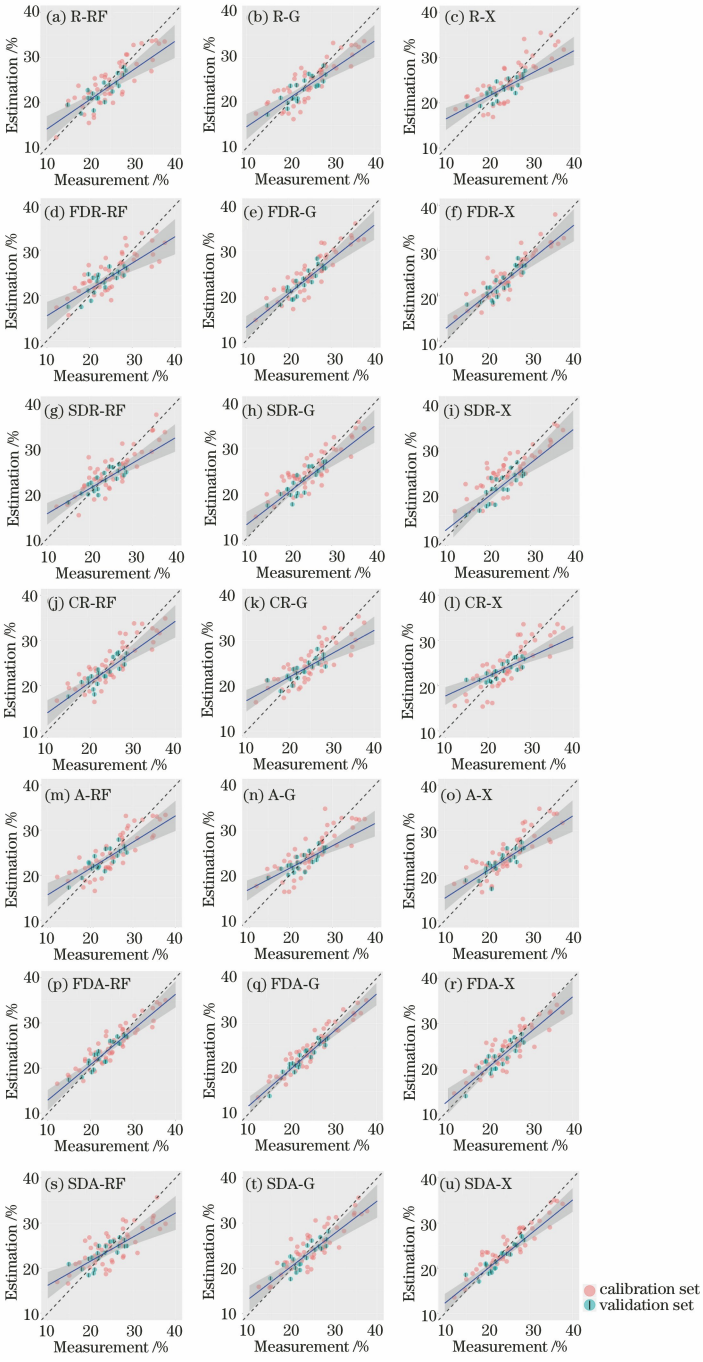

图 5 基于不同优选方法的 SMC 估测效果。(a)～(c) R 经 RF、GBRT、XGBoost 优选后的 SMC 估测效果;(d)～(f) FDR 经 RF、GBRT、XGBoost 优选后的 SMC 估测效果;(g)～(i) SDR 经 RF、GBRT、XGBoost 优选后的 SMC 估测效果;(j)～(l) CR经 RF、GBRT、XGBoost 优选后的 SMC 估测效果;(m)～(o) A 经 RF、GBRT、XGBoost 优选后的 SMC 估测效果;(p)～(r) FDA 经 RF、GBRT、XGBoost 优选后的 SMC 估测效果;(s)～(u) SDA 经 RF、GBRT、XGBoost 优选后的 SMC 估测效果

Fig. 5 SMC estimation results based on different preferred methods. (a)-(c) SMC estimation effect of R optimized by RF, GBRT and XGBoost; (d)-(f) SMC estimation effect of FDR optimized by RF, GBRT and XGBoost; (g)-(i) SMC estimation effect of SDR optimized by RF, GBRT and XGBoost; (j)-(l) SMC estimation effect of CR optimized by RF, GBRT and XGBoost; (m)-(o) SMC estimation effect of A optimized by RF, GBRT and XGBoost; (p)-(r) SMC estimation effect of FDA optimized by RF, GBRT and XGBoost; (s)-(u) SMC estimation effect of SDA optimized by RF, GBRT and XGBoost

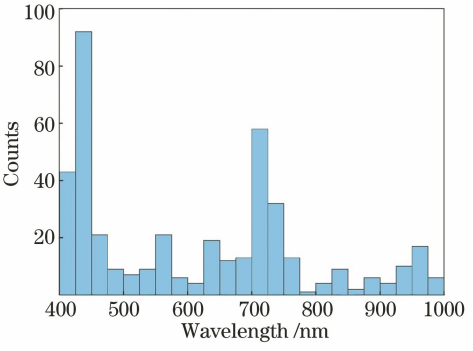

图 6 特征波段分布

Fig. 6 Distribution of characteristic bands

利用 RF 算法和极限学习机(ELM)算法结合最优光谱指标筛选的 SMC 敏感波段相近。在干旱区,作物受到不同程度的水分胁迫,导致植被生理特征发生变化,一定程度上造成了光谱差异。由此可见 SMC 的大小影响了地表植被叶绿素和水分含量,从植被冠层光谱信息中提取特征波段从而建立的预测模型在一定程度上能够反映 SMC 的变化情况。

将 UAV 获得的高光谱数据,进行 SMC 建模与预测,但 UAV 高光谱影像通常存在无效、冗余的信息。对原始光谱预处理,最大程度降低背景噪声对光谱数据的影响,增强高光谱特征对实测地物的敏感程度,进一步优化后期的特征变量筛选与建模。Wang 等[20]利用经过处理的 FD 光谱实现了盐渍土的预测与建模,且精度较高,因此实验采用 6 种预处理方式使特征波段尽可能被 3 种不同算法筛选。在 R、CR 和 A3 种预处理方式中,A 方式得到的光谱信息最佳,其中 FDA 建模效果最佳,RPIQ 达到 3.490,这表明利用 FDA 进行特征变量筛选与建模能在最大程度上挖掘冬小麦的高光谱信息并抑制土壤背景所带来的噪声影响。

对比 FDA-GBRT、FDA-RF 和 FDA-XGBoost 建模过程中使用的波段,FDA-GBRT 中红边波段占比 40%,蓝波段占比 25%,近红外波段占比 10%;FDA-RF 中,红边波段占比 70%,蓝波段占比 20%,近红外波段占比 5%;FDA-XGBoost 中,红边波段占比 40%,蓝波段占比 25%,近红外波段占比 25%。在同等预处理方式下 GBRT 算法在进行重要性排序时考虑了全波段响应特征,RF 算法只注重地表植被与光谱响应最为明显的红边波段,忽略了其他略微弱的波段,XGBoost 算法则较多考虑了近红外波段。比较 FDR-GBRT 与 FDA-GBRT 的建模效果,FDR-GBRT 建模过程中红边波段占比 35%,蓝波段占比 15%,近红外波段占比 5%。对比建模效果较好的特征波段发现,蓝波段、红边波段以及近红外波段占比相对均衡的模型建模效果更好。

GWR 模型考虑了变量间关系的空间非平稳性,故对 SMC 的预测效果依赖于 SMC 与各变量相关关系的空间非平稳性程度[11,21-22]。其中 GBRT 算法对 FDA 筛选的敏感波段 SMC 建模效果最佳,这表明该模型中 FDA-GBRT 自变量满足与 SMC 之间的空间非平稳性程度。对比不同输入变量的建模效果,多数模型建模集与预测集 R^2 均大于 0.600,RPIQ 均达到 1.400 以上,表明 GWR 模型在大尺度 SMC 建模预测中精度较高。

遥感影像只能反映部分信息,而植被光谱是多因素的综合表达[23],参量不完全是控制遥感信息的主导因子,虽然集成学习算法在特征变量筛选过程中的精度有所提高,但本身存在多参数或超参数的现象,在优选变量前需要进行复杂和系统的训练。实验利用有限的样本对 SMC 进行了较好的估算,在接下来的研究中,将着手大尺度 SMC 估算,实现空天地一体化的观测方法,为干旱区农业监测、生态系统管理提供技术支持[24]。

5 结 论

通过 UAV 获得的高光谱数据和实测的 SMC 数据,利用原始光谱及 6 种预处理后的高光谱数据,经三种算法筛选特征波段后,得到 21 种输入变量并通过 GWR 建立 SMC 预测模型。实验结果表明:在不同预处理方式下,以 A 为最佳,可释放出较好的光谱特性,其次为 CR;3 种特征波段重要性筛选的算法中,在 R、FDR、SDR 和 FDA 特征波段筛选中 GBRT 算法表现最优,RF 算法对 420~450 nm、670~760 nm 波段较为敏感,XGBoost 算法在 SDA 预处理方式下表现较好,综合分析得出,GBRT 算法在特征波段筛选中优势最大;多数模型建模集与验证集的 R^2 均大于 0.600,且预测能力较好,因此 GWR 模型在 SMC 预测建模中切实有效,具有大尺度、潜在高精度的特点。

参 考 文 献

[1] Park J, Baik J, Choi M. Satellite-based crop coefficient and evapotranspiration using surface soil moisture and vegetation indices in Northeast Asia[J]. Catena, 2017, 156: 305-314.

[2] Zhang Z T, Wang H F, Han W T, et al. Inversion of soil moisture content based on multispectral

remote sensing of UAVs[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(2): 173-181.
张智韬, 王海峰, 韩文霆, 等. 基于无人机多光谱遥感的土壤含水率反演研究[J]. 农业机械学报, 2018, 49(2): 173-181.

[3] Sankey T T, McVay J, Swetnam T L, et al. UAV hyperspectral and lidar data and their fusion for arid and semi-arid land vegetation monitoring[J]. Remote Sensing in Ecology & Conservation, 2018, 4(1): 20-33.

[4] Cheng H, Shen R L, Chen Y Y, et al. Estimating heavy metal concentrations in suburban soils with reflectance spectroscopy[J]. Geoderma, 2019, 336: 59-67.

[5] Ge X Y, Ding J L, Wang J Z, et al. Estimation of soil moisture content based on competitive adaptive reweighted sampling algorithm coupled with machine learning[J]. Acta Optica Sinica, 2018, 38(10): 1030001.
葛翔宇, 丁建丽, 王敬哲, 等. 基于竞争适应重加权采样算法耦合机器学习的土壤含水量估算[J]. 光学学报, 2018, 38(10): 1030001.

[6] Zhang Z P, Ding J L, Wang J Z. Spectral characteristics of oasis soil in arid area based on harmonic analysis algorithm[J]. Acta Optica Sinica, 2019, 39(2): 0228003.
张子鹏, 丁建丽, 王敬哲. 基于谐波分析算法的干旱区绿洲土壤光谱特性研究[J]. 光学学报, 2019, 39(2): 0228003.

[7] Miao S, Wang R, Li J C, et al. Retrieval algorithm of phycocyanin concentration in inland lakes from sentinel 3A-OLCI images[J]. Journal of Infrared and Millimeter Waves, 2018, 37(5): 621-630.
苗松, 王睿, 李建超, 等. 基于哨兵3A-OLCI影像的内陆湖泊藻蓝蛋白浓度反演算法研究[J]. 红外与毫米波学报, 2018, 37(5): 621-630.

[8] Wang N, Li Q Z, Du X, et al. Identification of main crops based on the univariate feature selection in Subei[J]. Journal of Remote Sensing, 2017, 21(4): 519-530.
王娜, 李强子, 杜鑫, 等. 单变量特征选择的苏北地区主要农作物遥感识别[J]. 遥感学报, 2017, 21(4): 519-530.

[9] Zamani Joharestani M, Cao C X, Ni X L, et al. $PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data[J]. Atmosphere, 2019, 10(7): 373.

[10] Song W Z, Jia H F, Huang J F, et al. A satellite-based geographically weighted regression model for regional $PM_{2.5}$ estimation over the Pearl River Delta region in China[J]. Remote Sensing of Environment, 2014, 154: 1-7.

[11] Yuan J, Zhang F, Ge X Y, et al. Leaf salt ion content estimation of halophyte plants based on geographically weighted regression model combined with hyperspectral data [J]. Transactions of the Chinese Society of Agricultural Engineering, 2019, 35(10): 115-124.
袁婕, 张飞, 葛翔宇, 等. 地理加权回归模型结合高光谱反演盐生植物叶片盐离子含量[J]. 农业工程学报, 2019, 35(10): 115-124.

[12] Ge X, Wang J, Ding J, et al. Combining UAV-based hyperspectral imagery and machine learning algorithms for soil moisture content monitoring[J]. PeerJ, 2019, 7: e6926.

[13] Wang J Z, Ding J L, Ma X K, et al. Detection of soil moisture content based on UAV-derived hyperspectral imagery and spectral index in oasis cropland[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(11): 164-172.
王敬哲, 丁建丽, 马轩凯, 等. 基于光谱指数的绿洲农田土壤含水率无人机高光谱检测[J]. 农业机械学报, 2018, 49(11): 164-172.

[14] Sun J, Cong S L, Mao H P, et al. CARS-ABC-SVR model for predicting leaf moisture of leaf-used lettuce based on hyperspectral [J]. Transactions of the Chinese Society of Agricultural Engineering, 2017, 33(5): 178-184.
孙俊, 丛孙丽, 毛罕平, 等. 基于高光谱的油麦菜叶片水分CARS-ABC-SVR预测模型[J]. 农业工程学报, 2017, 33(5): 178-184.

[15] Menze B H, Kelm B M, Masuch R, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data[J]. BMC Bioinformatics, 2009, 10(1): 213.

[16] Friedman J H . Greedy function approximation: a gradient boosting machine[J]. Annals of Statistics, 2001, 29(5):1189-1232.

[17] Im J, Jensen J R. Hyperspectral remote sensing of vegetation[J]. Geography Compass, 2008, 2(6): 1943-1961.

[18] Schoo R N, Ray S S, Manjunath K R. Hyperspectral remote sensing of agriculture[J]. Current Science, 2015, 108(5): 848-859.

[19] Bao Q L, Ding J L, Wang J Z. Prediction of soil moisture content by selecting spectral characteristics using random forest method [J]. Laser & Optoelectronics Progress, 2018, 55(11): 113002.
包青岭, 丁建丽, 王敬哲. 利用随机森林方法优选光谱特征预测土壤水分含量[J]. 激光与光电子学进展, 2018, 55(11): 113002.

[20] Wang J Z, Ding J L, Abulimiti A, et al. Quantitative estimation of soil salinity by means of different modeling methods and visible-near infrared (VIS-NIR) spectroscopy, Ebinur Lake Wetland, Northwest China[J]. PeerJ, 2018, 6: e4703.

[21] Jiang Z L, Yang Y S, Sha J M. Application of GWR model in hyperspectral prediction of soil heavy metals [J]. Acta Geographica Sinica, 2017, 72(3): 533-544.
江振蓝, 杨玉盛, 沙晋明. GWR 模型在土壤重金属高光谱预测中的应用[J]. 地理学报, 2017, 72(3): 533-544.

[22] Luo M, Guo L, Zhang H T, et al. Characterization of spatial distribution of soil organic carbon in China based on environmental variables[J]. Acta Pedologica Sinica, 2020(1): 48-69.
罗梅, 郭龙, 张海涛, 等. 基于环境变量的中国土壤有机碳空间分布特征[J]. 土壤学报, 2020(1): 48-69.

[23] Yue Y M, Wang K L, Zhang B, et al. Exploring the relationship between vegetation spectra and eco-geo-environmental conditions in Karst region, Southwest China [J]. Environmental Monitoring and Assessment, 2010, 160(1/2/3/4): 157-168.

[24] Shi Z, Xu D Y, Teng H F, et al. Soil information acquisition based on remote sensing and proximal soil sensing: current status and prospect[J]. Progress in Geography, 2018, 37(1): 79-92.
史舟, 徐冬云, 滕洪芬, 等. 土壤星地传感技术现状与发展趋势[J]. 地理科学进展, 2018, 37(1): 79-92.