doi:10.6041/j.issn.1000-1298.2024.07.025

基于无人机高光谱遥感与机器学习的小麦品系产量估测研究

齐　浩[1]　吕亮杰[2]　孙海芳[1]　李　偲[1]　李甜甜[1]　侯　亮[1]

(1.河北省农林科学院农业信息与经济研究所,石家庄 050051;
2.河北省农林科学院粮油作物研究所,石家庄 050035)

摘要:为快速、准确地估测小麦产量,有效提高育种工作效率,本文以小麦品系为研究对象,收集小麦灌浆期无人机高光谱数据和产量数据。首先基于递归特征消除法筛选出特征波长作为模型输入变量,然后利用岭回归(Ridge regression,RR)、偏最小二乘回归(Partial least squares regression,PLS)、多元线性回归(Multiple linear regression,MLR)3 种线性算法和随机森林(Random forest,RF)、梯度提升回归(Gradient boosting regression,GBR)、极限梯度提升(eXtreme gradient boosting,XGB)、高斯过程回归(Gaussian process regression,GPR)、支持向量回归(Support vector regression,SVR)、K 最邻近算法(K-nearest neighbor,KNN)6 种非线性算法构建单一算法产量估测模型并进行精度比较,最后基于 Stacking 算法构建多模型集成组合,筛选最佳集成模型。结果表明,基于不同算法的产量估测模型精度差异显著,非线性模型优于线性模型,基于 GBR 的产量估测模型在单一模型中表现最优,训练集 R^2 为 0.72,RMSE 为 534.49 kg/hm^2,NRMSE 为 11.10%,测试集 R^2 为 0.60,RMSE 为 628.73 kg/hm^2,NRMSE 为 13.88 %。基于 Stacking 算法构建的集成模型性能与初级模型和次级模型的选择密切相关,以 KNN、RR、SVR 为初级模型组合,GBR 为次级模型的集成模型有效提高了估测精度,相比单一模型 GBR,训练集 R^2 提高 1.39%,测试集 R^2 提高 3.33%。本研究可为基于高光谱技术的小麦品系产量估测提供应用参考。

关键词:小麦品系;产量估测;无人机高光谱;遥感;机器学习;Stacking 算法

中图分类号:S127　　**文献标识码**:A　　**文章编号**:1000-1298(2024)07-0260-10　　**OSID**:

Yield Estimation of Wheat Lines Based on UAV Hyperspectral Remote Sensing and Machine Learning

QI Hao[1]　LÜ Liangjie[2]　SUN Haifang[1]　LI Si[1]　LI Tiantian[1]　HOU Liang[1]

(1. *Institute of Agricultural Information and Economics, Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang* 050051, *China*
2. *Institute of Cereal and Oil Crops, Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang* 050035, *China*)

Abstract: Rapid and accurate estimation of wheat yield can improve the efficiency of breeding. Yield data of wheat lines and hyperspectral data during grain filling period were collected. Firstly, the feature wavelengths were selected as model input variables by using recursive feature elimination method. Then three linear algorithms (ridge regression, partial least squares regression, multiple linear regression) and six nonlinear algorithms (random forest, gradient boosting regression, eXtreme gradient boosting, Gaussian process regression, support vector regression, K-nearest neighbor) were employed to establish single algorithm yield estimation models for precision comparison. Finally, the Stacking algorithm was adopted to develop multi-model ensemble combinations, aiming to identify the optimal ensemble model. The results showed that the accuracy of yield estimation models, based on different algorithms, varied significantly, and that the nonlinear models were better than the linear models. The yield estimation model based on GBR performed best in the single models, with R^2 of 0.72, RMSE of 534.49 kg/hm^2 and NRMSE of 11.10% in the training set, R^2 of 0.60, RMSE of 628.73 kg/hm^2, and NRMSE of 13.88% in the testing set. The performance of the ensemble models based on Stacking algorithm was closely

收稿日期:2024-03-22　修回日期:2024-05-13

基金项目:河北省现代农业产业技术体系小麦创新团队项目(21326318D)和河北省农林科学院基本科研业务费项目(2023090101)

作者简介:齐浩(1990—),男,助理研究员,主要从事农业信息化研究,E-mail:njsqihao@163.com

通信作者:侯亮(1980—),男,副研究员,主要从事农业信息化研究,E-mail:giantark@hotmail.com

related to the selection of primary and secondary models. The model with KNN, RR, SVR as primary models and GBR as the secondary model effectively improved the yield estimation accuracy. Compared with the single model GBR, the training set R^2 was increased by 1.39% and the testing set R^2 was increased by 3.33%. The research result can provide an application reference for yield estimation of wheat lines based on hyperspectral technology.

Key words: wheat lines; yield estimation; UAV hyperspectral; remote sensing; machine learning; Stacking algorithm

0　引言

小麦是中国三大粮食作物之一，2023 年我国小麦种植面积 2.363×10^7 hm^2，占全国粮食种植总面积的 19.86%[1]，培育高产优质小麦品种是保证我国粮食安全的关键。在小麦育种工作中，通常利用系谱法进行优良后代选育，杂种四代以后参加产量比较试验的系统称为品系[2]，在此试验过程中产量测量方法主要依靠人工取样，耗时耗力，效率低下。无人机遥感技术的快速发展，为小麦产量信息的快速、准确获取提供了新途径，对于提高育种效率和加快高产品种选育具有重要意义。

基于无人机的高光谱遥感具有探测范围广、分辨率高、光谱信息量大等特点，已成为预测作物性状的常用方法[3-6]。前人研究中多利用单一光谱指数或植被指数组合进行作物产量估测[7-11]，然而光谱指数仅使用了部分波段信息，这可能导致基于这些光谱指数的产量估测模型通常针对于特定品种[12]。高光谱反射率受植物形态和生理的影响，具有基因型特异性[13]，可作为一种高效的高通量表型获取手段，实现高产品系的准确筛选[14-16]。

近年来，机器学习算法被广泛应用于作物表型性状反演模型的构建[17-19]。传统的线性算法如岭回归、偏最小二乘回归、多元线性回归被应用于基于光谱指数的产量估测模型开发中[20-22]。随着更多光谱信息的加入，线性模型可能无法揭示输入变量和产量之间的复杂关系。非线性算法的开发利用可能改善估测结果，随机森林、梯度提升回归、高斯回归过程、支持向量回归、K 最邻近算法、极限梯度提升等算法被应用于小麦[23]、水稻[24]、大麦[25]、玉米[26-27]、大豆[28]等作物的产量估测。虽然前人开展了大量结合机器学习算法的作物产量估测研究，但大多针对单一或几个品种，对小麦品系的产量估测研究相对较少，且算法性能受到输入因子和应用场景的影响，因此，基于高光谱反射率评估不同机器学习算法在小麦育种计划中的产量估测能力是研究的目标之一。

为进一步提高作物产量估测精度，相关学者开展了集成模型的研究[29-34]。Stacking 作为一种多模型集成算法[35]，其通过次级模型对多个初级模型的输出值进行集成训练，从而将多个模型的解析能力进行整合。在大多数利用 Stacking 算法估测作物产量的研究中，线性模型 MLR 通常被用作次级模型，对多个初级模型进行集成，从而得到比单一模型更精准的估测结果[31-34]。也有研究表明，RF 模型作为次级模型时可得到较高的产量估测精度[14]。不同初级模型和次级模型必然会影响集成效果，因此，有必要探索不同的模型集成组合，以提升集成模型的估测能力。

本文以小麦品系为研究对象，将无人机高光谱反射率作为输入变量构建产量估测模型，评价不同机器学习算法在产量估测建模中的表现，利用 Stacking 算法构建集成模型，探究不同模型集成组合对估测精度的影响，以期为小麦品系产量的精确估算提供方法依据和应用参考。

1　材料与方法

1.1　研究区概况

试验在河北省农林科学院粮油作物研究所堤上试验站（37°56′32.63″N，114°42′53.13″E）进行（图 1），试验站地处河北省石家庄市藁城区，属暖温带季风区大陆性气候，地势平坦、排灌方便、肥力均匀、地力水平中上等。

1.2　产量数据获取

选取 2019—2021 年河北省农林科学院粮油作物研究所小麦研究室品系比较试验材料，其中，2019 年为黄淮北片水地组预试材料和河北省水地组区试材料，2020 年为黄淮北片水地组预试和区试材料，2021 年为中作联合体预试材料和黄淮北片水地组区试材料，育种材料播种在尺寸为 8 m × 1.5 m 的小区内，材料成熟后，对每个小区全部收获，晒干后称取小麦产量（单产），3 年共计采集小麦材料样本 957 个。

1.3　无人机影像获取及预处理

小麦灌浆期间，籽粒充实增重，叶片逐渐衰老，光合能力下降，营养器官如叶片、茎鞘和茎秆中的贮存性物质向穗部转移[36]。前人研究发现，在此时期进行遥感估产效果最好[10,30]。因此，获取了 2019—

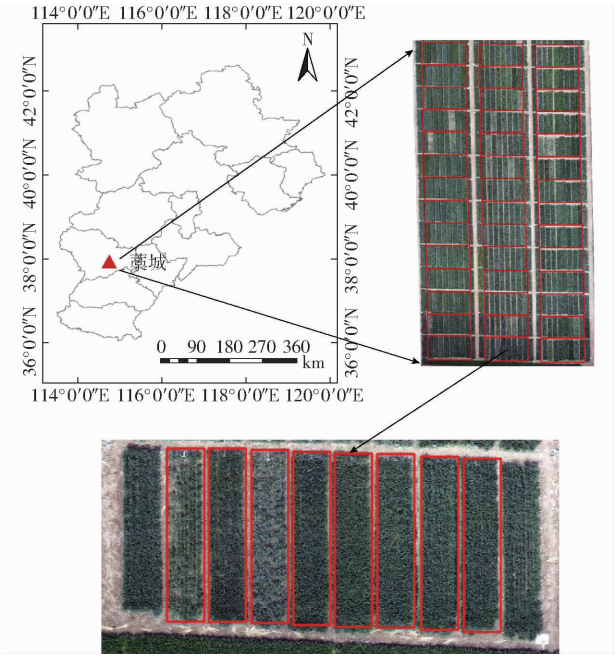

图 1 研究区概况

Fig. 1 Study area

2021 年小麦灌浆期高光谱数据,采集时间分别为 2019 年 5 月 10 日、2020 年 5 月 8 日、2021 年 5 月 12 日。数据获取采用大疆 S1000 型无人机(整机质量为 4.2 kg,载物质量 6 kg,续航时间 15 ~ 20 min),同步搭载 CubertUHD185 型机载高光谱成像光谱仪(德国,光谱范围 450 ~ 950 nm,光谱间隔 4 nm),飞行高度 100 m,飞行时天气晴朗、无云、无风,光谱数据稳定。

高光谱影像采集后,需对数据进行预处理,主要包括影像融合、影像拼接和反射率提取等步骤。利用软件 Cubert Cube - Pilot 对每幅高光谱影像进行融合,获得相应的全彩影像。利用软件 Agisoft PhotoScan 完成高光谱影像的拼接。拼接完毕的高光谱影像包含 125 个波段,空间分辨率约为 5 cm。在软件 ENVI 中提取影像的冠层光谱反射率,对每个育种小区划分感兴趣区域(Region of interest, ROI),ROI 尽可能覆盖小区内全部植被,提取感兴趣区域内所有像元光谱反射率计算平均值,即得到每个育种小区的冠层光谱反射率。

1.4 研究方法

1.4.1 遥感特征选择方法

在模型开发之前进行特征选择,以降低数据维度,提高模型泛化能力。递归特征消除法(Recursive feature elimination, RFE)是一种广泛应用的包裹式特征选择方法[37]。其实现过程为:①选择一种评估器,确定初始特征重要性/相关性。②剔除重要性/相关性最低的特征。③根据被移除变量的移除顺序对其排序,重复以上过程,直到确定所有输入特征的排名。在本研究中,选择随机森林算法作为评估器(RF - RFE)[38],计算各特征排名,进行特征变量选择。

1.4.2 模型构建

本研究基于 Python 3.8 进行数据分析,共选取 9 种机器学习算法,包括岭回归(RR)、偏最小二乘回归(PLS)、多元线性回归(MLR)3 种线性算法和随机森林(RF)、梯度提升回归(GBR)、极限梯度提升(XGB)、高斯过程回归(GPR)、支持向量回归(SVR)、K 最邻近算法(KNN)6 种非线性算法,建立高光谱反射率与小区实测产量的单算法回归模型,采用 Stacking 算法(图 2),对不同模型进行集成组合,以提高小麦产量估测精度。

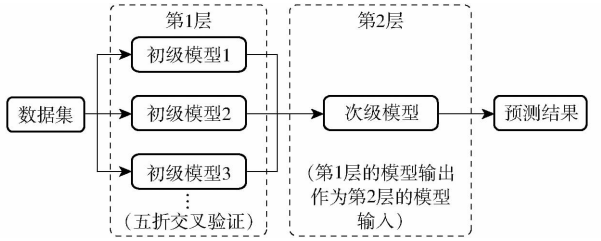

图 2 基于 Stacking 的集成学习算法原理图

Fig. 2 Schematic diagram of ensemble learning algorithm based on Stacking

1.4.3 精度验证

3 年共采集冬小麦样本 957 个,按照比例 4∶1 随机划分,765 个样本作为训练集用于模型训练,采用五折交叉验证策略进行参数优化,以提高模型泛化能力。192 个样本作为测试集用于模型精度验证。模型检验选用决定系数(Coefficient of determination, R^2)、均方根误差(Root mean square error, RMSE)和标准均方根误差(Normalized root mean square error, NRMSE)。R^2 越接近 1,则该模型拟合精度越高;RMSE 与 NRMSE 越小,表示估算值与实测值差异越小,模型估算能力越强。

2 结果与分析

2.1 不同产量下小区冠层光谱反射率

将小麦产量分为高、中、低 3 个等级,所有光谱数据按照对应的小麦产量进行划分,并取平均值得到 3 条平均反射率,如图 3 所示。在不同产量下光谱反射曲线的波动趋势总体保持一致,但在局部波段区域存在较大差异。在波段 450 ~ 680 nm 范围,3 条光谱曲线变化差异不大,反射率呈降低、升高再降低的趋势,分别形成 2 个吸收谷和 1 个反射峰,此波长范围受叶绿素影响,小麦灌浆期叶绿素含量下降,因此波峰和波谷较不明显。680 nm 后,曲线反射率开始急剧上升,不同产量的反射率曲线出现差异,较高产量对应

的光谱反射率曲线具有较大的斜率。780 nm 后曲线反射率上升趋近平缓，不同产量对应的光谱曲线出现明显分层，较高的产量具有较高的反射率数值。900 nm 之后，3 条光谱曲线均呈快速下降趋势。前人研究发现 UHD185 型高光谱仪在波长 454 nm 及 854～950 nm 处测量的光谱信息不稳定[39]，因此选用波段 458～850 nm 进行下一步分析。

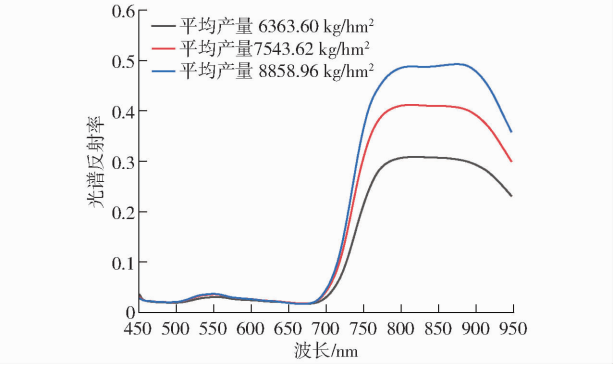

图 3　不同产量下对应的冠层光谱反射率

Fig. 3　Canopy spectral reflectance corresponding to different yields

2.2　光谱特征选择

首先采用 RF－RFE 策略对所有波长进行重要性排序，共计训练 250 次，如图 4 所示。排名靠前和靠后的波长均相对稳定，波长 458、470、566、726、786、818、830、834、838、842、846、850 nm 在超过 200 次的训练排名中始终保持在前 20 位，因此，为消除冗余特征，提高模型运算速率，将上述 12 个波长选为特征变量用于下一步建模分析。

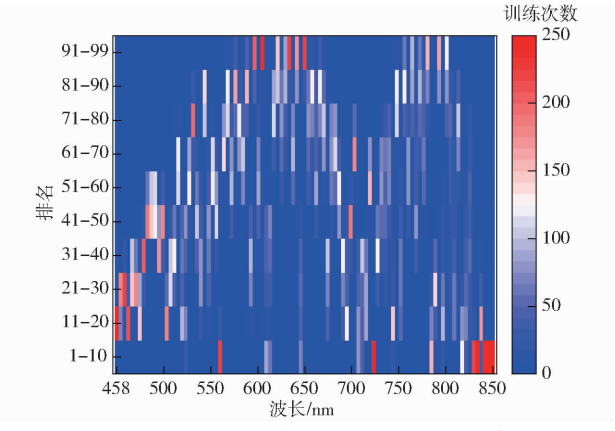

图 4　基于 RF－RFE 策略的特征重要性排序热图

Fig. 4　Heat map of feature importance ranking based on RF－RFE

2.3　基于单一算法的产量估测模型构建

以上述 12 个特征波长作为输入变量，以小麦产量作为响应变量，利用 RR、PLS、MLR 3 种线性算法以及 RF、GBR、XGB、GPR、SVR、KNN 6 种非线性算法进行模型构建。利用训练集进行建模，采用 5 折交叉验证进行调参，以提高模型泛化能力，结果如图 5 所示。训练集中，6 种非线性模型 R^2 为 0.64～0.72，RMSE 为 534.49～610.92 kg/hm^2，NRMSE 为 11.10%～12.69%，3 种线性模型建模效果差异不大，R^2 为 0.58，RMSE 为 661.04～663.10 kg/hm^2，NRMSE 为 13.73%～13.77%。测试集中，6 种非线性模型 R^2 为 0.55～0.60，RMSE 为 628.73～664.36 kg/hm^2，NRMSE 为 13.88%～14.67%，3 种线性模型 R^2 为 0.52，RMSE 为 685.94～691.35 kg/hm^2，NRMSE 为 15.14%～15.26%。可见，非线性模型估测精度远优于线性模型。

在 6 种非线性模型中，具有集成学习能力的 RF、GBR 和 XGB 模型表现较优，GBR 模型训练集和测试集的 R^2 分别为 0.72 和 0.60，RMSE 分别为 534.49 kg/hm^2 和 628.73 kg/hm^2，NRMSE 分别为 11.10% 和 13.88%。RF 模型训练集和测试集的 R^2 分别为 0.69 和 0.60，RMSE 分别为 572.43 kg/hm^2 和 630.01 kg/hm^2，NRMSE 分别为 11.89% 和 13.91%。XGB 模型训练集和测试集的 R^2 分别为 0.71 和 0.58，RMSE 分别为 545.34 kg/hm^2 和 641.56 kg/hm^2，NRMSE 分别为 11.33% 和 14.16%。GPR、KNN、SVR 模型在训练集和测试集上的表现各有优劣，其中，GPR 模型相比 KNN、SVR 在训练集上具有更好的估测精度（R^2 为 0.67、RMSE 为 584.45 kg/hm^2、NRMSE 为 12.14%），而 SVR 模型比 GPR、KNN 在测试集上更具稳定性（R^2 为 0.57、RMSE 为 650.07 kg/hm^2、NRMSE 为 14.35%）。综合比较模型在训练集与测试集的估测结果，GBR 模型表现最优，在小麦品系产量估测上有较大的应用潜力。

2.4　基于 Stacking 集成的产量估测模型构建

为进一步提高估测精度，本研究尝试对不同模型进行堆叠。Stacking 集成模型由初级模型组合和次级模型组成，以上述 9 种模型为基础，构建二模型组合（$C_9^2=36$ 组）、三模型组合（$C_9^3=84$ 组）、四模型组合（$C_9^4=126$ 组）、五模型组合（$C_9^5=126$ 组）、六模型组合（$C_9^6=84$ 组）、七模型组合（$C_9^7=36$ 组）、八模型组合（$C_9^8=9$ 组）以及所有模型组合（$C_9^9=1$ 组）共计 502 组分别作为初级模型组合，并以 RF、GBR、XGB、GPR、SVR、KNN、RR、PLS 和 MLR 分别作为次级模型进行迭代训练，不同次级模型下所有集成模型 R^2 变化分布如图 6 所示。以 RF、GBR、XGB 作为次级模型时，集成模型 R^2 变化幅度较小。基于次级模型 RF 训练集 R^2 处于 0.65～0.69，测试集 R^2 处于 0.57～0.62。基于次级模型 GBR 训练集 R^2 处于 0.67～0.73，测试集 R^2 处于 0.57～0.62。基于次级模型 XGB 训练集 R^2 处于 0.67～0.75，测试集 R^2 处于 0.55～0.60。以 KNN、RR、PLS、MLR 为次级模型

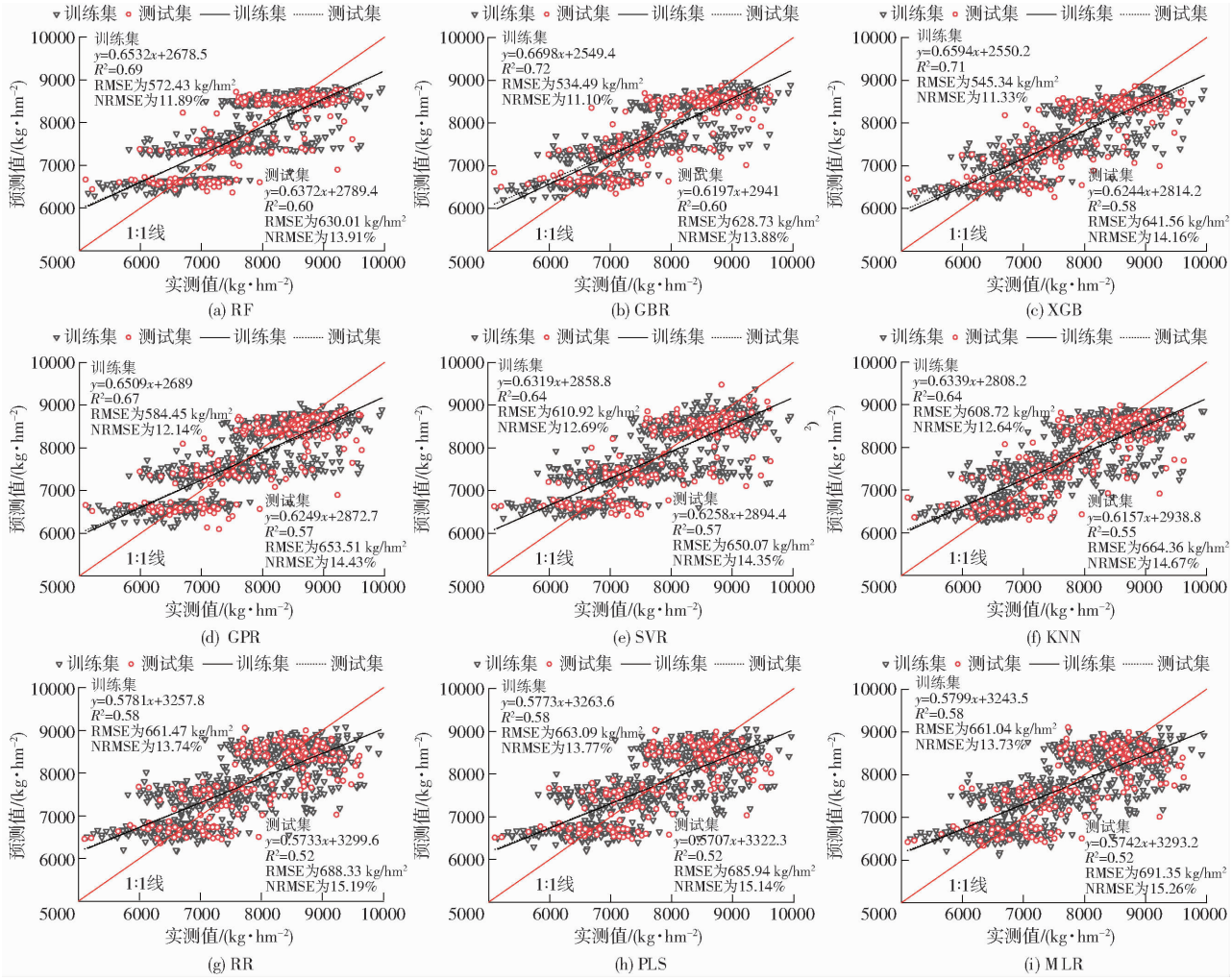

图 5 基于单一算法构建的产量估测模型估测结果

Fig. 5 Estimation results of yield estimation models based on single algorithms

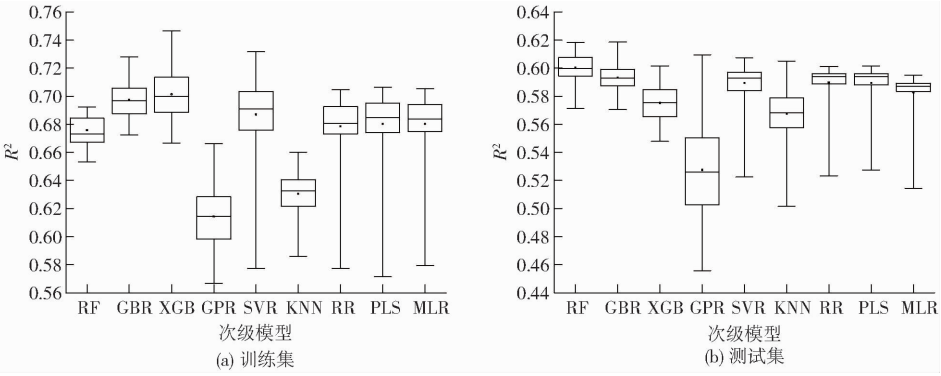

图 6 基于不同次级模型的所有集成模型 R^2 分布

Fig. 6 R^2 distribution of all ensemble models based on different secondary models

时,集成模型 R^2 变化幅度居中。基于次级模型 KNN 训练集 R^2 处于 0.59 ~ 0.66,测试集 R^2 处于 0.50 ~ 0.60。基于次级模型 RR 训练集 R^2 处于 0.58 ~ 0.70,测试集 R^2 处于 0.52 ~ 0.60。基于次级模型 PLS 训练集 R^2 处于 0.57 ~ 0.71,测试集 R^2 处于 0.53 ~ 0.60。基于次级模型 MLR 训练集 R^2 处于 0.58 ~ 0.71,测试集 R^2 处于 0.51 ~ 0.60。以 GPR、SVR 作为次级模型时,集成模型 R^2 波动较大,基于次级模型 GPR 训练集 R^2 处于 0.57 ~ 0.67,测试集 R^2 处于 0.46 ~ 0.61。以 SVR 为次级模型时训练集 R^2 处于 0.58 ~ 0.73,测试集 R^2 处于 0.52 ~ 0.61。综上可知,当选择 RF、GBR、XGB 作为次级模型时,受初级模型选择的影响较小,集成模型较为稳定。

综合考虑在训练集和测试集上的表现,筛选具有不同初级模型数量的最优集成模型(表 1),其 R^2 如图 7 所示。以 GBR、XGB 为次级模型时,最优集

表 1　不同初级模型数量下的最优集成模型

Tab. 1　Optimal ensemble models with different numbers of primary models

次级模型	初级模型组合（数量为 2）	初级模型组合（数量为 3）	初级模型组合（数量为 4）	初级模型组合（数量为 5）	初级模型组合（数量为 6）	初级模型组合（数量为 7）	初级模型组合（数量为 8）	初级模型组合（数量为 9）
RF	GBR/RR	GBR/RR/PLS	GBR/KNN/RR/PLS	GBR/KNN/RR/PLS/MLR	GBR/XGB/KNN/RR/PLS/MLR	GBR/XGB/SVR/KNN/RR/PLS/MLR	RF/GBR/XGB/SVR/KNN/RR/PLS/MLR	
GBR	KNN/RR	KNN/RR/SVR	KNN/SVR/MLR/RR	SVR/KNN/PLS/MLR/RR	XGB/SVR/KNN/PLS/MLR/RR	RF/XGB/SVR/KNN/PLS/MLR/RR	RF/GBR/XGB/SVR/KNN/RR/PLS/MLR	
XGB	KNN/MLR	KNN/RR/MLR	KNN/PLS/MLR/RR	SVR/KNN/PLS/MLR/RR	XGB/SVR/KNN/PLS/MLR/RR	RF/GBR/XGB/SVR/KNN/RR/PLS	RF/GBR/XGB/SVR/KNN/RR/PLS/MLR	
GPR	GBR/KNN	SVR/KNN/PLS	GBR/RR/PLS/MLR	SVR/KNN/RR/PLS/MLR	RF/SVR/KNN/RR/PLS/MLR	RF/XGB/SVR/KNN/RR/PLS/MLR	RF/GBR/XGB/SVR/KNN/RR/PLS/MLR	
SVR	GBR/XGB	GBR/XGB/GPR	GBR/GPR/SVR/PLS	GBR/XGB/KNN/PLS/MLR	GBR/XGB/GPR/KNN/PLS/MLR	GBR/XGB/GPR/KNN/RR/PLS/MLR	RF/GBR/XGB/SVR/KNN/RR/PLS/MLR	RF/GBR/XGB/GPR/SVR/KNN/RR/PLS/MLR
KNN	RF/PLS	GBR/PLS/MLR	RF/GBR/KNN/PLS	RF/SVR/KNN/RR/PLS	RF/SVR/KNN/RR/PLS/MLR	RF/XGB/SVR/KNN/RR/PLS/MLR	RF/GBR/XGB/SVR/KNN/RR/PLS/MLR	
RR	GBR/RR	GBR/XGB/RR	GBR/XGB/RR/MLR	GBR/XGB/RR/PLS/MLR	GBR/XGB/GPR/RR/PLS/MLR	GBR/XGB/GPR/SVR/RR/PLS/MLR	GBR/XGB/GPR/SVR/KNN/RR/PLS/MLR	
PLS	GBR/GPR	GBR/GPR/MLR	GBR/GPR/RR/PLS	GBR/GPR/RR/PLS/MLR	GBR/XGB/GPR/RR/PLS/MLR	GBR/XGB/GPR/SVR/RR/PLS/MLR	GBR/XGB/GPR/SVR/KNN/RR/PLS/MLR	
MLR	GBR/RR	GBR/XGB/RR	GBR/XGB/RR/PLS	GBR/XGB/GPR/RR/PLS	GBR/XGB/GPR/RR/PLS/MLR	GBR/XGB/GPR/SVR/RR/PLS/MLR	GBR/XGB/GPR/SVR/KNN/RR/PLS/MLR	

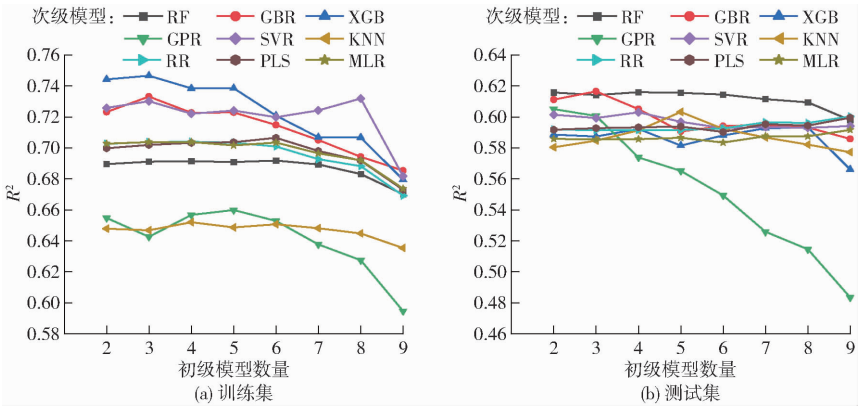

图 7　不同初级模型数量下最优集成模型的 R^2

Fig. 7　R^2 of optimal ensemble models with different number of primary models

成模型的初级模型数量为 3。以 RF 为次级模型时，最优集成模型的初级模型数量为 4，模型在训练集和测试集上均有较高的估测精度。以 GPR 为次级模型时，训练集上，初级模型数量为 5 的集成模型 R^2 最高，测试集上，R^2 随初级模型数量增加而急剧下降。以 SVR 为次级模型时，初级模型数量为 3 的集成模型在训练集和测试集上均具有较好的估测能力。以 KNN 为次级模型，初级模型数量为 5 时，集成模型具有较高的估测精度。以 RR、PLS、MLR 为次级模型，初级模型数量在 4 或 5 时集成模型具有较好的估测结果。此外发现，当初级模型数量为 9 时，大部分集成模型在训练集和测试集上的估测精度均有一定幅度降低。由此说明，对于不同的次级模型，最佳集成模型的初级模型数量不同，在一定范

围内增加初级模型数量能够提高集成模型性能。

选择不同次级模型下最优集成模型如图 8 所示。将集成模型与最优单一模型 GBR 对比，以 GBR 为次级模型时，最优初级模型组合为 KNN、RR、SVR，训练集 R^2 为 0.73，RMSE 为 532.22 kg/hm^2，NRMSE 为 11.05%，测试集 R^2 为 0.62，RMSE 为 618.78 kg/hm^2，NRMSE 为 13.66%，训练集 R^2 提高了 1.39%，测试集 R^2 提高了 3.33%。以 SVR、XGB 为次级模型时，最优初级模型组合分别为 GBR、XGB、GPR 和 KNN、RR、MLR，在训练集上的估测精度有所提升，R^2 分别为 0.73 和 0.75，在测试集上估测精度持平或下降，R^2 分别为 0.60 和 0.59。以 RF 为次级模型时，其最优初级模型组合为 GBR、KNN、RR、PLS，模型在测试集上估测精度有一定提升（R^2 为 0.62），但在训练集上并未体现集成模型的优势（R^2 为 0.69）。以 GPR、KNN 作为次级模型时，集成模型在测试集上的估测能力与单一模型 GBR 相似，但在训练集上的 R^2 分别降低 8.33% 和 9.72%。以 RR、PLS 和 MLR 作为次级模型时，集成模型在训练集和测试集上的估测精度均不如单一模型 GBR。综上而言，Stacking 集成方法能够提高小麦产量估测精度，但受到模型选择的影响，以 KNN、RR、SVR 作为初级模型组合，GBR 作为次级模型的集成模型兼具稳定性和精确性，具有比单一模型更好的估测能力。

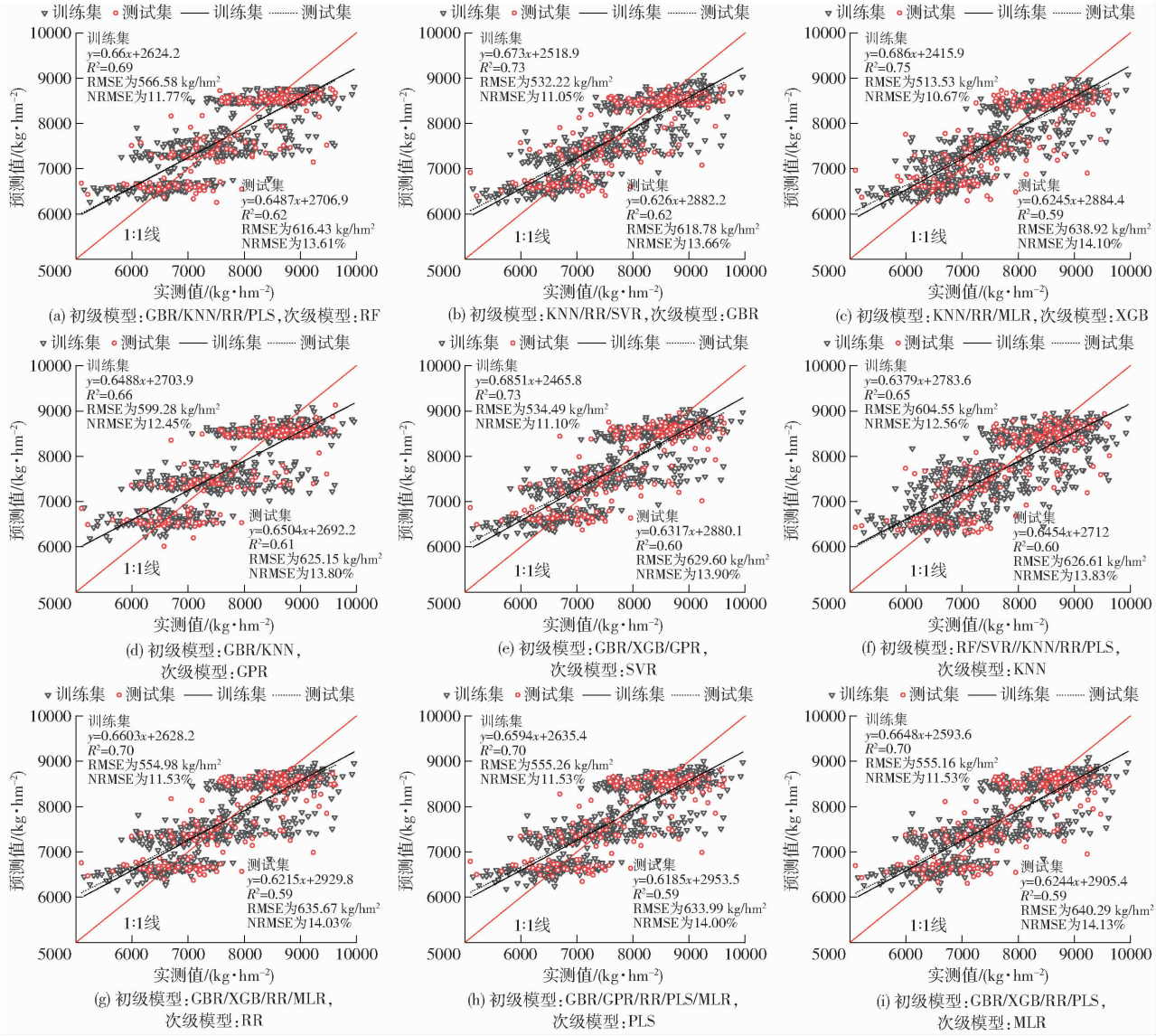

图 8 基于不同次级模型的最优集成模型估测结果

Fig. 8 Estimation results of optimal ensemble models based on different secondary models

3 讨论

利用无人机遥感平台进行小麦估产，能够帮助育种专家快速筛选小麦高产品种。高光谱遥感因其光谱信息量大、分辨率高、探测范围广等特点在作物生长监测中具有较大的优势，因此，本研究通过高光谱数据，探究了不同算法估算小麦品系产量的能力。

前人研究证实了利用递归特征消除法进行特征选择具有较好的应用效果[14,31,40]，本研究利用 RF－RFE 筛选出波长 458、470、566、726、786、818、830、834、838、842、846、850 nm 为特征波长，大幅度降低了特征波长数量，减少模型复杂程度和计算时间。其中波长 458、470 nm 处于蓝光范围，为叶绿素、类胡萝卜素和黄色素的吸收波段，因此反射率较低。波长 566 nm 处于黄绿光范围，是弱活性带之一，靠近波长 550 nm“绿峰”位置。波长 726 nm 处于红边位置，反射率随着波长的增加而急剧升高，曲线较陡[41]，红边是植被营养、长势、水分、叶面积等的指示性特征，相关研究表明，加入红边信息能够提高产量估测准确性[21]。波长 786、818、830、834、838、842、846、850 nm 位于近红外波段范围，受到绿色作物叶片结构的多次反射、散射影响，具有较强的反射特性，是作物表型参数与反射率关系较敏感的波段，对监测作物长势状况最有利[42]。上述波长与植被的各种理化参数紧密相关，本研究证实了通过 RFE 筛选出光谱反射率进行小麦产量估测具有一定可靠性。

不同的建模方法显著影响模型的估测精度。本研究利用 9 种机器学习算法构建小麦产量估测模型。经分析发现，基于非线性模型的产量估测精度明显高于线性模型。非线性模型能够解析光谱信息与产量的复杂关系，获得精度更高的估算结果，这与前人研究结果相似[43－44]。在非线性模型中，KNN 模型运算复杂度较低，训练时间较短，过度依赖样本点的欧氏距离[45]。SVR 及 GPR 模型的核心问题是核函数确定及相关参数选择[46－47]，构建的产量模型在一定程度上受到限制[43]。具有集成学习能力的 RF、GBR、XGB 模型能够处理具有高维特征的样本，其中 GBR 模型在基于单一算法构建的模型中具有最优的估测精度和稳定性。因此，在今后的研究中可以将 GBR 作为一种可靠的作物产量估算技术，提高 GBR 算法在作物表型监测中的应用价值。

Stacking 集成算法在作物表型评估中能够提升模型性能已得到广泛共识[14,31－34]。本研究共构建 4 518 种（502 组初级模型组合，9 组次级模型）模型集成组合进行小麦产量估测，结果表明，Stacking 集成算法能够提高产量估测精度，但受到初级模型和次级模型选择的影响。RF、GBR、XGB 模型具有强大的学习能力和抗噪能力，作为次级模型时，对于不同的初级模型组合均有一定的集成效果，而 SVR 和 GPR 模型作为次级模型时，因核函数的确定问题对其有一定的应用限制。此外，对于不同的次级模型，最佳集成模型的初级模型数量不同，在一定范围内增加初级模型数量能够提高集成模型估测能力，但过多的初级模型反而会使集成模型性能下降，可能原因是部分初级模型有一定相似性，过多的初级模型反而会造成一定冗余，降低了集成模型估测精度。因此，选取初级模型时要充分考虑模型之间的异质性[31]。在大多利用 Stacking 集成模型估测作物性状的研究中，线性模型通常被用作次级模型，但本研究发现以线性模型 PLS、RR、MLR 作为次级模型时，最优集成结果估测精度不如表现最佳的单一模型，而利用非线性模型作为次级模型时，集成模型具有比单一模型更高的估测精度。经分析发现，当初级模型组合为 KNN、RR、SVR，次级模型为 GBR 时，模型性能在训练集和测试集上均有一定的提升。因此，在平衡模型复杂性和准确性的基础上，选择适当的初级模型组合，并利用非线性次级模型进行集成，能够有效提升作物产量估测精度。

目前，基于高光谱与机器学习结合的小麦品系产量估算研究中仍存在待解决的问题。首先，本文仅涉及单点试验数据，如何提高估产模型在不同环境下的普适性有待进一步研究。此外，应当考虑冠层结构及纹理信息等因素对产量估算的影响，深入探讨多源遥感信息协同下产量模型的估测精度，在变量增多的同时，可考虑深度学习模型的开发与应用。本研究仅利用小麦灌浆期影像进行产量估算，同一影像下的不同品系生育进程不同必然导致估产结果存在差异，可考虑融合全生育期模型进行优化试验，以期达到更好的估产效果。

4　结论

（1）通过采集 3 年小麦品系高光谱影像数据，基于 RF－RFE 方法筛选出排名较高的波长，构建 9 种机器学习算法模型进行产量估测，并利用 Stacking 集成算法以提高产量估测的准确性。

（2）基于非线性算法的产量估测模型（训练集 R^2 为 0.64～0.72、测试集 R^2 为 0.55～0.60）优于基于线性算法的产量估测模型（训练集 R^2 为 0.58、测试集 R^2 为 0.52），其中 GBR 模型在训练集和测试集上的表现都较为稳定，训练集 R^2 为 0.72，RMSE 为 534.49 kg/hm^2，NRMSE 为 11.10%，测试集 R^2 为 0.60，RMSE 为 628.73 kg/hm^2，NRMSE 为 13.88%。

（3）Stacking 算法能够提升模型性能，但受到初级模型和次级模型的影响。以 KNN、RR、SVR 为初级模型组合，GBR 为次级模型的集成模型兼具准确性和泛化能力，训练集 R^2 为 0.73，RMSE 为 532.22 kg/hm^2，NRMSE 为 11.05%，测试集 R^2 为 0.62，RMSE 为 618.78 kg/hm^2，NRMSE 为 13.66%，

相比单一模型 GBR,训练集 R^2 提高 1.39%,测试集 R^2 提高 3.33%。

(4)本研究结果可为快速高效的小麦产量估测、高产品种筛选提供技术支持和辅助决策。

参 考 文 献

[1] 国家统计局. 中华人民共和国 2023 年国民经济和社会发展统计公报[R/OL]. 北京:国家统计局,2024. https://www.stats.gov.cn/sj/zxfb/202402/t20240228_1947915.html.

[2] 张天真. 作物育种学总论[M]. 四版. 北京:中国农业出版社,2022.

[3] 许童羽,白驹驰,郭忠辉,等. 基于无人机高光谱遥感的水稻氮营养诊断方法[J]. 农业机械学报,2023,54(2):189 - 197.
XU Tongyu, BAI Juchi, GUO Zhonghui, et al. Diagnosis method of rice nitrogen deficiency based on UAV hyperspectral remote sensing[J]. Transactions of the Chinese Society for Agricultural Machinery,2023,54(2):189 - 197. (in Chinese)

[4] YANG X, YANG R, YE Y, et al. Winter wheat SPAD estimation from UAV hyperspectral data using cluster-regression methods [J]. International Journal of Applied Earth Observation and Geoinformation, 2021, 105: 102618.

[5] WANG L, CHEN S, LI D, et al. Estimation of paddy rice nitrogen content and accumulation both at leaf and plant levels from UAV hyperspectral imagery[J]. Remote Sensing, 2021, 13(15): 2956.

[6] TAO H, FENG H, XU L, et al. Estimation of crop growth parameters using UAV-based hyperspectral remote sensing data[J]. Sensors, 2020, 20(5): 1296.

[7] 唐延林,王纪华,黄敬峰,等. 利用水稻成熟期冠层高光谱数据进行估产研究[J]. 作物学报,2004,30(8):780 - 785.
TANG Yanlin, WANG Jihua, HUANG Jingfeng, et al. Yield estimation by hyperserctral data of rice canopies in mature stages [J]. Acta Agronomica Sinica, 2004, 30(8): 780 - 785. (in Chinese)

[8] 张威,潘剑君,李勇,等. 基于冠层光谱特征的冬小麦产量估算研究[J]. 土壤通报,2015,46(1):169 - 176.
ZHANG Wei, PAN Jianjun, LI Yong, et al. Research on estimation of grain yield of winter wheat based on canopy spectrum characteristics[J]. Chinese Journal of Soil Science, 2015, 46(1): 169 - 176. (in Chinese)

[9] 赵晓庆,杨贵军,刘建刚,等. 基于无人机载高光谱空间尺度优化的大豆育种产量估算[J]. 农业工程学报,2017,33(1):110 - 116.
ZHAO Xiaoqing, YANG Guijun, LIU Jian'gang, et al. Estimation of soybean breeding yield based on optimization of spatial scale of UAV hyperspectral image[J]. Transactions of the CSAE, 2017, 33(1): 110 - 116. (in Chinese)

[10] 朱婉雪,李仕冀,张旭博,等. 基于无人机遥感植被指数优选的田块尺度冬小麦估产[J]. 农业工程学报,2018,34(11):78 - 86.
ZHU Wanxue, LI Shiji, ZHANG Xubo, et al. Estimation of winter wheat yield using optimal vegetation indices from unmanned aerial vehicle remote sensing[J]. Transactions of the CSAE, 2018, 34(11): 78 - 86. (in Chinese)

[11] WANG F, YAO X, XIE L, et al. Rice yield estimation based on vegetation index and florescence spectral information from UAV hyperspectral remote sensing[J]. Remote Sensing, 2021, 13(17): 3390.

[12] MONTESINOS-LÓPEZ O A, MONTESINOS-LÓPEZ A, CROSSA J, et al. Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data[J]. Plant Methods, 2017, 13: 1 - 23.

[13] SCHWEIGER A K, CAVENDER-BARES J, TOWNSEND P A, et al. Plant spectral diversity integrates functional and phylogenetic components of biodiversity and predicts ecosystem function[J]. Nature Ecology & Evolution, 2018, 2(6): 976 - 982.

[14] YOOSEFZADEH-NAJAFABADI M, EARL H J, TULPAN D, et al. Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean[J]. Frontiers in Plant Science, 2021, 11: 624273.

[15] FAN J, ZHOU J, WANG B, et al. Estimation of maize yield and flowering time using multi-temporal UAV-based hyperspectral data[J]. Remote Sensing, 2022, 14(13): 3052.

[16] WEBER V S, ARAUS J L, CAIRNS J E, et al. Prediction of grain yield using reflectance spectra of canopy and leaves in maize plants grown under different water regimes[J]. Field Crops Research, 2012, 128: 82 - 90.

[17] TAO H, FENG H, XU L, et al. Estimation of the yield and plant height of winter wheat using UAV-based hyperspectral images[J]. Sensors, 2020, 20(4): 1231.

[18] WANG T, GAO M, CAO C, et al. Winter wheat chlorophyll content retrieval based on machine learning using in situ hyperspectral data[J]. Computers and Electronics in Agriculture, 2022, 193: 106728.

[19] GRINBERG N F, ORHOBOR O I, KING R D. An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat[J]. Machine Learning, 2020, 109(2): 251 - 277.

[20] ADAK A, MURRAY S C, BOŽINOVIĆ S, et al. Temporal vegetation indices and plant height from remotely sensed imagery can predict grain yield and flowering time breeding value in maize via machine learning regression[J]. Remote Sensing, 2021, 13(11): 2141.

[21] 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.
TAO Huilin, XU Liangji, FENG Haikuan, et al. Winter wheat yield estimation based on UAV hyperspectral remote sensing data[J]. Transactions of the Chinese Society for Agricultural Machinery,2020,51(7):146 - 155. (in Chinese)

[22] 肖璐洁,杨武德,冯美臣,等. 基于高光谱植被指数的冬小麦估产模型[J]. 生态学杂志,2022,41(7):1433 - 1440.
XIAO Lujie, YANG Wude, FENG Meichen, et al. Development of winter wheat yield estimation models based on hyperspectral vegetation indices[J]. Chinese Journal of Ecology, 2022, 41(7): 1433 - 1440. (in Chinese)

[23] 王晶晶,李长硕,卓越,等. 基于多时相无人机遥感生育时期优选的冬小麦估产[J]. 农业机械学报,2022,53(9):197 - 206.

WANG Jingjing, LI Changshuo, ZHUO Yue, et al. Yield estimation of winter wheat based on optimization of growth stages by multi-temporal UAV remote sensing[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(9): 197 – 206. (in Chinese)

[24] ARUMUGAM P, CHEMURA A, SCHAUBERGER B, et al. Remote sensing based yield estimation of rice (*Oryza sativa* L.) using gradient boosted regression in India[J]. Remote Sensing, 2021, 13(12): 2379.

[25] SHARIFI A. Yield prediction with machine learning algorithms and satellite images[J]. Journal of the Science of Food and Agriculture, 2021, 101(3): 891 – 896.

[26] KARIMI Y, PRASHER S, MADANI A, et al. Application of support vector machine technology for the estimation of crop biophysical parameters using aerial hyperspectral observations[J]. Can. Biosyst. Eng., 2008, 50(7): 13 – 20.

[27] KUMAR C, MUBVUMBA P, HUANG Y, et al. Multi-stage corn yield prediction using high-resolution UAV multispectral data and machine learning models[J]. Agronomy, 2023, 13(5): 1277.

[28] HUBER F, YUSHCHENKO A, STRATMANN B, et al. Extreme Gradient Boosting for yield estimation compared with deep learning approaches[J]. Computers and Electronics in Agriculture, 2022, 202: 107346.

[29] FU P, MEACHAM-HENSOLD K, GUAN K, et al. Hyperspectral leaf reflectance as proxy for photosynthetic capacities: an ensemble approach based on multiple machine learning algorithms[J]. Frontiers in Plant Science, 2019, 10: 454448.

[30] LI Z, CHEN Z, CHENG Q, et al. UAV-based hyperspectral and ensemble machine learning for predicting yield in winter wheat[J]. Agronomy, 2022, 12(1): 202.

[31] FENG L, ZHANG Z, MA Y, et al. Alfalfa yield prediction using UAV-based hyperspectral imagery and ensemble learning [J]. Remote Sensing, 2020, 12(12): 2028.

[32] 费帅鹏,禹小龙,兰铭,等.基于高光谱遥感和集成学习方法的冬小麦产量估测研究[J].中国农业科学,2021,54(16):3417 – 3427.
FEI Shuaipeng, YU Xiaolong, LAN Ming, et al. Research on winter wheat yield estimation based on hyperspectral remote sensing and ensemble learning method[J]. Scientia Agricultura Sinica, 2021, 54(16): 3417 – 3427. (in Chinese)

[33] 甘甜,李雷,李红叶,等.基于多源遥感数据和机器学习算法的冬小麦产量预测研究[J].麦类作物学报,2022,42(11):1419 – 1428.
GAN Tian, LI Lei, LI Hongye, et al. Winter wheat yield prediction based on multi-source remote sensing data and machine learning algorithms[J]. Journal of Triticeae Crops, 2022, 42(11): 1419 – 1428. (in Chinese)

[34] FEI S, HASSAN M A, XIAO Y, et al. UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat[J]. Precision Agriculture, 2023, 24(1): 187 – 212.

[35] WOLPERT D H. Stacked generalization[J]. Neural Networks, 1992, 5(2): 241 – 259.

[36] 王伟青.小麦旗叶灌浆过程中光合作用及其籽粒灌浆之间的关系研究[D].青岛:中国海洋大学,2007.
WANG Weiqing. The associations of photosynthesis and grain filling during grain filling period in flag leaves of wheat species [D]. Qingdao: Ocean University of China, 2007. (in Chinese)

[37] CAI J, LUO J, WANG S, et al. Feature selection in machine learning: a new perspective[J]. Neurocomputing, 2018, 300: 70 – 79.

[38] GRANITTO P M, FURLANELLO C, BIASIOLI F, et al. Recursive feature elimination with random forest for PTR – MS analysis of agroindustrial products[J]. Chemometrics and Intelligent Laboratory Systems, 2006, 83(2): 83 – 90.

[39] 高林,杨贵军,于海洋,等.基于无人机高光谱遥感的冬小麦叶面积指数反演[J].农业工程学报,2016,32(22):113 – 120.
GAO Lin, YANG Guijun, YU Haiyang, et al. Retrieving winter wheat leaf area index based on unmanned aerial vehicle hyperspectral remote sensing[J]. Transactions of the CSAE, 2016, 32(22): 113 – 120. (in Chinese)

[40] SUN Q, GU X, CHEN L, et al. Hyperspectral estimation of maize (*Zea mays* L.) yield loss under lodging stress[J]. Field Crops Research, 2023, 302: 109042.

[41] 姚付启.冬小麦高光谱特征及其生理生态参数估算模型研究[D].杨凌:西北农林科技大学,2012.
YAO Fuqi. Hyperspectral characteristics and estimating models about physiological ecological parameters of winter wheat[D]. Yangling: Northwest A&F University, 2012. (in Chinese)

[42] 冯伟.基于高光谱遥感的小麦氮素营养及生长指标监测研究[D].南京:南京农业大学,2007.
FENG Wei. Monitoring nitrogen status and growth characters with canopy hyperspectal remote sensing in wheat[D]. Nanjing: Nanjing Agricultural University, 2007. (in Chinese)

[43] 张少华,段剑钊,贺利,等.基于无人机平台多模态数据融合的小麦产量估算研究[J].作物学报,2022,48(7):1746 – 1760.
ZHANG Shaohua, DUAN Jianzhao, HE Li, et al. Wheat yield estimation from UAV platform based on multi-modal remote sensing data fusion[J]. Acta Agronomica Sinica, 2022, 48(7): 1746 – 1760. (in Chinese)

[44] WANG Y, ZHANG Z, FENG L, et al. Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States[J]. Remote Sensing, 2020, 12(8): 1232.

[45] 宋成阳,耿洪伟,费帅鹏,等.基于多源数据的小麦品种产量估测研究[J].光谱学与光谱分析,2023,43(7):2210 – 2219.
SONG Chengyang, GENG Hongwei, FEI Shuaipeng, et al. Study on yield estimation of wheat varieties based on multi-source data[J]. Spectroscopy and Spectral Analysis, 2023, 43(7): 2210 – 2219. (in Chinese)

[46] GUNN S R. Support vector machines for classification and regression[J]. ISIS Technical Report, 1998, 14(1): 5 – 16.

[47] SCHULZ E, SPEEKENBRINK M, KRAUSE A. A tutorial on Gaussian process regression: modelling, exploring, and exploiting functions[J]. Journal of Mathematical Psychology, 2018, 85: 1 – 16.