

王 宇,汪 泓,肖玖军,等. 基于 MCC-GAPLS-PLSR 的辣椒叶绿素含量高光谱定量反演[J]. 江苏农业学报, 2024, 40(5): 865-873.

doi: 10.3969/j.issn.1000-4440.2024.05.011

## 基于 MCC-GAPLS-PLSR 的辣椒叶绿素含量高光谱定量反演

王 宇<sup>1</sup>, 汪 泓<sup>2</sup>, 肖玖军<sup>3,4,5</sup>, 邢 丹<sup>6</sup>, 李可相<sup>3,4</sup>, 张永亮<sup>2</sup>, 岳延滨<sup>7</sup>

(1.中国电建集团贵阳勘测设计研究院有限公司,贵州 贵阳 550081; 2.贵州大学矿业学院,贵州 贵阳 550025; 3.贵州科学院山地资源研究所,贵州 贵阳 550001; 4.贵州省土地绿色整治工程研究中心,贵州 贵阳 550001; 5.贵州大学资源与环境工程学院,贵州 贵阳 550025; 6.贵州省农业科学院辣椒研究所,贵州 贵阳 550009; 7.贵州省农业科学院农业科技信息研究所,贵州 贵阳 550009)

**摘要:** 为了准确监测辣椒生长,本研究对辣椒冠层光谱反射率进行对数处理、倒数处理、倒数的对数处理、连续统去除处理、一阶微分处理、二阶微分处理,并与 SPAD 值进行相关性分析,用最大相关系数法(MCC)选取相关性较好的特征波段生成特征波段数据集,再用遗传算法-偏最小二乘法(GAPLS)进行降维得到最优特征波段组合,采用偏最小二乘法(PLSR)、反向传播神经网络(BPNN)、随机森林(RF)和最小二乘支持向量机(LSSVM)4种机器学习算法构建辣椒叶绿素含量反演模型。结果表明,最优波段和对应处理分别为 700 nm(原始光谱)、699 nm(对数处理)、713 nm(连续统去除处理)、500 nm(二阶微分处理)、713 nm(二阶微分处理)。GAPLS 的降维效果较好,与降维前相比 PLSR 模型的精度提升率最高, $R^2$ 、RPD 分别提升了 82.22%、136.98%,RMSE 降低了 29.96%。4 种模型中,GAPLS 降维处理后的 PLSR 模型的精度最好, $R^2$ 、RMSE 和 RPD 分别为 0.82、1.94、4.55。本研究构建的 MCC-GAPLS-PLSR 模型具有较好的反演潜力,适用于研究区辣椒叶片叶绿素含量测定,推动辣椒高效种植。

**关键词:** 叶绿素含量; 辣椒; 高光谱; 光谱变换; 遗传算法-偏最小二乘法

**中图分类号:** S127;S641.3 **文献标识码:** A **文章编号:** 1000-4440(2024)05-0865-09

## Hyperspectral quantitative inversion of chlorophyll content in pepper based on MCC-GAPLS-PLSR

WANG Yu<sup>1</sup>, WANG Hong<sup>2</sup>, XIAO JiuJun<sup>3,4,5</sup>, XING Dan<sup>6</sup>, LI Kexiang<sup>3,4</sup>, ZHANG Yongliang<sup>2</sup>, YUE Yanbin<sup>7</sup>

(1.Guiyang Engineering Corporation Limited, Power China, Guiyang 550081, China; 2.College of Mining, Guizhou University, Guiyang 550025, China; 3.Institute of Mountain Resources, Guizhou Academy of Sciences, Guiyang 550001, China; 4.Engineering Research Center for Land Green Consolidation of Guizhou, Guiyang 550001, China; 5.College of Resources and Environmental Engineering, Guizhou University, Guiyang 550025, China; 6.Peper Research Institute, Guizhou Academy of Agricultural Sciences, Guiyang 550009, China; 7. Institute of Agricultural Science and Technology Information, Guizhou Academy of Agricultural Sciences, Guiyang 550009, China)

收稿日期:2023-05-31

**基金项目:**国家重点研发计划项目(2017YFD1100307);贵州省科技支撑计划项目[黔科合支撑(2020)1Y172、黔科合支撑(2021)一般173、黔科合支撑(2021)一般496];贵州省农业科学院青年基金项目[黔农科院青年科技基金(2021)22]

**作者简介:**王 宇(1998-),女,贵州安顺人,硕士研究生,主要从事农业摄影测量方面的研究。(E-mail)2029807592@qq.com

**通讯作者:**肖玖军,(E-mail)xiaojiujun0504@163.com

**Abstract:** In order to accurately monitor the growth of peppers, this study performed logarithmic treatment, reciprocal treatment, reciprocal logarithmic treatment, continuum removal treatment, first derivative treatment, second derivate treatment on the canopy spectral reflectance of

peppers, and conducted correlation analysis with SPAD values. The maximum correlation coefficient method (MCC) was used to select the feature bands with good correlation to generate a feature band dataset. And the genetic algorithm-partial least squares (GAPLS) was used to reduce the dimensionality to obtain the optimal feature band combination. Pepper chlorophyll content inversion model was constructed by using four machine learning algorithms: partial least squares regression (PLSR), backpropagation neural network (BPNN), random forest (RF) and least squares support vector machine (LSSVM). The results showed that the optimal wavelengths and corresponding treatments were 700 nm (original reflectivity), 699 nm (logarithmic treatment), 713 nm (continuum removal treatment), 500 nm (second derivate treatment), 713 nm (second derivate treatment). The dimensionality reduction effect by GAPLS was good. And compared with before dimension reduction, the accuracy improvement rate of PLSR model was the highest,  $R^2$  and  $RPD$  increased by 82.22% and 136.98% respectively, and  $RMSE$  decreased by 29.96%. Among the four models, PLSR model after dimensionality reduction by GAPLS had the best accuracy, with  $R^2$ ,  $RMSE$ , and  $RPD$  of 0.82, 1.94, and 4.55, respectively. The MCC-GAPLS-PLSR model constructed in this study has good inversion potential and is suitable for measuring the chlorophyll content of pepper leaves in the study area, thus promoting efficient cultivation of peppers.

**Key words:** chlorophyll content; pepper; hyperspectral; spectral transformation; genetic algorithm-partial least squares

叶绿素是植被进行光合作用的主要物质, 可以作为植被生长状况的监测指标<sup>[1]</sup>。监测叶绿素含量动态变化可帮助了解植物健康、营养状况和生产力<sup>[2]</sup>。前人常采用化学试验法对叶绿素含量进行测定, 该方法费时费力, 且在叶绿素提取过程中易造成色素损失, 破坏性取样也无法保证样品完整性<sup>[3]</sup>, 同时该方法仅能实现点状检测, 需达到一定样本量后才能借助一些空间插值法扩展为面状尺度, 离植物生长快速区域性监测有一定距离, 无法满足精准农业要求。SPAD (Soil and plant analyzer development) -502 叶绿素计被广泛用于叶绿素含量测定, SPAD 值能准确表征叶绿素相对含量, 可用于测量植被叶绿素含量<sup>[4-5]</sup>。由于高光谱遥感技术具有连续且丰富的光谱信息, 它可以更加精确地估算出植物的叶绿素含量, 并且高光谱遥感技术还具有快速、无损监测的优势, 可以在不损害植物的前提下, 对植物的长势进行监测<sup>[6]</sup>。

近年来, 针对高光谱反演 SPAD 值展开的研究取得了一定成果, 通常在对原始光谱进行预处理后利用相关模型对植被参数进行反演<sup>[7]</sup>。学者们在植被指数构建、光谱数据处理、特征波长选择等具体方面又各有所侧重。如 Zheng 等<sup>[2]</sup>研究原位高光谱测量反演山地草原叶绿素含量时发现植被指数 MSR (Modified simple ratio)、MCARI (Modified chlorophyll II absorption ratio index)、MCARI/OSAVI (Modified chlorophyll II ab-

sorption ratio index/Optimized soil-adjusted vegetation index) 和 TCARI (Transformed chlorophyll II absorption ratio index) 在估计冠层叶绿素含量方面具有较高准确性。Jiang 等<sup>[8]</sup>开发出三波段高光谱植被指数用于反演受不同程度病虫害威胁的红树林的叶片相对叶绿素含量。依尔夏提·阿不来提等<sup>[9]</sup>对棉花原始光谱进行连续统去除、立方根转换等处理, 发现转换后的光谱反射率与 SPAD 值相关性比原始光谱反射率高。李长春等<sup>[10]</sup>对光谱反射率进行分数阶微分处理, 结果表明处理后的光谱反射率与冬小麦叶绿素含量的相关系数绝对值有所提升。苑迎春等<sup>[11]</sup>提出在对颜色特征进行筛选时采用熵权法, 并利用机器学习法对小麦冠层叶绿素含量进行反演。张佳伟等<sup>[12]</sup>利用不同红边位置算法估测了玉米叶绿素含量。在模型构建方面, 线性模型法简单, 建模速度快, 如多元逐步线性回归 (MSLR)、主成分回归 (PCR)、偏最小二乘回归 (PLSR) 等, 其中 PLSR 应用最为广泛<sup>[13]</sup>。但线性模型无法理解多个变量之间的相互作用, 不能很好地拟合非线性数据。现阶段大量的非线性数据挖掘技术如人工神经网络 (ANN)、随机森林 (RF)、后向传播神经网络 (BPNN) 被广泛应用于遥感反演建模中。采用联合回归方法建立叶绿素含量混合反演模型的研究也有零星报道<sup>[14]</sup>。总体上多以单种光谱预处理后再经过特征波段选择来直接回归建模或构建植被指数间接回归建模。

目前作物叶绿素含量反演研究对象多为玉米、高粱、大麦等粮食作物,对经济作物研究较少,鲜见有对辣椒的高光谱反演研究。辣椒是重要的蔬菜作物,构建辣椒叶绿素含量反演模型对监测辣椒生长速度、智慧化种植、提升辣椒果实品质具有重要意义。本研究对辣椒冠层高光谱数据进行多种方法预处理,各预处理方法的特征波段提取采用最大相关系数法(MCC, Maximum correlation coefficient),组成特征波段数据集,再利用遗传算法-偏最小二乘法(GAPLS)对特征波段数据集进行变量选择,将选择的波段作为自变量输入偏最小二乘法回归(PLSR)、最小二乘支持向量机(LSSVM)、后向传播神经网络(BPNN)和随机森林(RF)4个机器学习模型中构建辣椒叶绿素含量反演模型,并对模型精度进行验证,优选最佳反演模型,探讨辣椒叶绿素含量高光谱反演方法,以期提高辣椒叶绿素含量反演模型的反演能力,推动辣椒高效、绿色种植。

## 1 材料与方法

### 1.1 研究区概况与试验设计

研究区位于贵州省遵义市辣椒研究基地( $107^{\circ}2'27''\text{E}$ ,  $27^{\circ}44'5''\text{N}$ ) (图1),属亚热带季风气候,年平均降水量1 200 mm,日照充足,其气候、土壤等自然条件十分适合辣椒栽培。

2021年4~10月在研究区对各辣椒品种进行不同施肥处理。选用贵州主栽辣椒的线椒和朝天椒,线椒品种为黔椒8号和红辣18号,朝天椒品种为辣研101号和红全球,共4个品种。每个辣椒品种都以 $0\text{ kg/hm}^2$ 、 $200\text{ kg/hm}^2$ 、 $350\text{ kg/hm}^2$ 、 $500\text{ kg/hm}^2$  4个施氮量进行处理,3次重复。共计48个试验区,每个试验区长6 m,宽5 m。分别于播种期和开花期施肥,播种期和开花期施肥量相等。在最能反映辣椒生长状况、决定辣椒产量与品质的盛花期、盛果期采集试验数据。

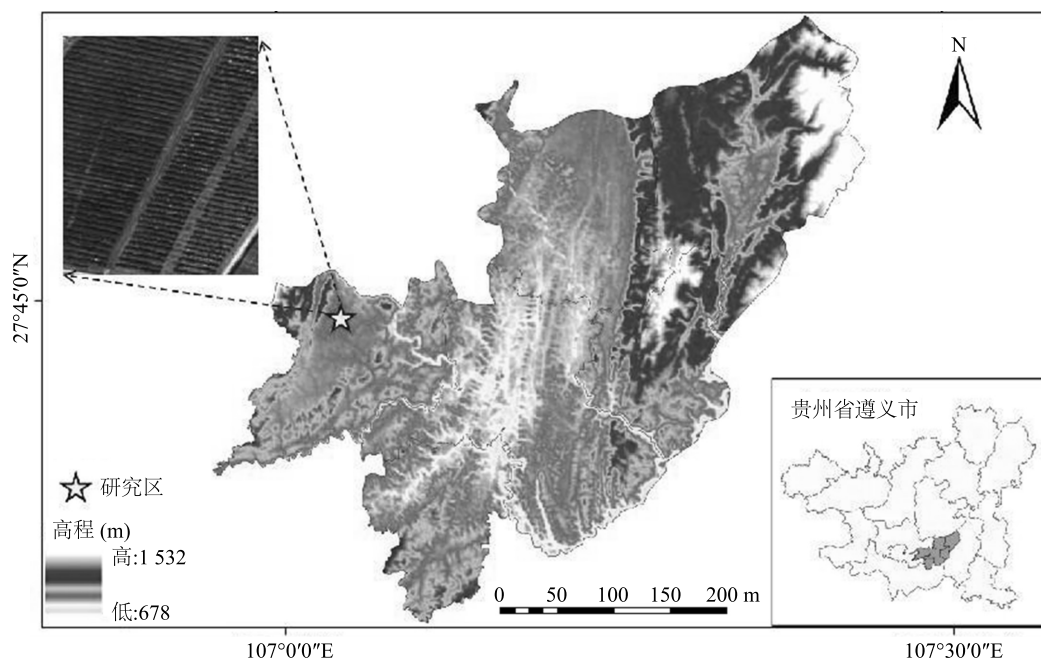


图1 研究区位置

Fig.1 Location of research area

### 1.2 光谱反射率和叶绿素含量测定方法

1.2.1 光谱反射率测定 利用美国 ASD 公司制造的 FieldSpec4 Standard-Res 便携式地物光谱仪,在350~2 500 nm 波段测量辣椒冠层的反射率,采样间隔为1 nm。在10:00~15:00对高光谱数据进行测量,测量当天为晴天,无风,无云。在测量过程中,传

感器探头垂直于地面,并且与辣椒植株顶部的高度距离为30 cm,视场角 $25^{\circ}$ 。每隔5株植株进行1次标准白板校正,每株植株重复测量5次,该植株最终光谱反射率为测量数据的平均值。在高光谱波段范围内,可见光和近红外波段为叶绿素的主要响应波段<sup>[15]</sup>,所以将研究光谱区域定为400~1 000 nm。



**1.2.2 叶绿素含量测定** 本研究用 SPAD 值表征辣椒叶绿素含量<sup>[16]</sup>。叶绿素含量采用 SPAD-502 Plus 叶绿素仪进行测定,与高光谱测量同步。每株辣椒测定 15~20 片辣椒冠层叶片叶绿素含量,测定时对每片叶片上、中、下部测定 6~10 次,取平均值作为辣椒叶绿素含量,测定时避开叶脉。

### 1.3 数据处理方法

**1.3.1 异常值处理** 用 3 倍标准差法检查 SPAD 数据集异常样本,剔除均值-3 倍标准差以及均值+3 倍标准差之外的数据,对剔除后的 SPAD 数据集再次用 3 倍标准差法检查,如此反复,直至无异常样本。用主成分分析法检测光谱数据,在主成分分布图中,95%置信椭圆之外的点被认为是异常样本,予以剔除,经多次反复检测,直至无异常样本。最后得到光谱数据和 SPAD 数据对,共 80 对。

**1.3.2 光谱预处理** 对光谱数据进行 Svitzky-Golay 平滑处理,对反射率进行降噪处理,以此曲线为基础再进行光谱变换。光谱变换方法为连续统去除、对数、倒数的对数、一阶微分、倒数和二阶微分处理,以不变换光谱为对照。Svitzky-Golay 卷积平滑利用 Origin 运算,光谱变换利用 Matlab 运算。

**1.3.3 特征变量选择** (1) 最大相关系数法 (MCC)。Pearson 相关性系数可用来度量变量  $X$  和变量  $Y$  之间的相关性,取值为 $[-1, +1]$ 。通过 MCC 可以一定程度上减少光谱数据的冗余信息,从而筛选特征波段<sup>[17]</sup>。将从各光谱变换中筛选出的特征变量进行组合,形成特征波段数据集。(2) 遗传算法-偏最小二乘法 (GAPLS)。特征波段对于构建模型至关重要<sup>[18]</sup>。偏最小二乘法擅于处理小样本数据,可降低特征维数,使低维数据更具解释性,从而获得更好的识别精度<sup>[19]</sup>。GAPLS 不仅具有遗传算法的全局优化搜索能力,还能利用偏最小二乘法对特征波段进行筛选<sup>[20]</sup>。其基本思路是将偏最小二乘法中交叉均方根  $RMSECV$  作为遗传算法中的适应度函数,用遗传算法进行变量筛选,提取特征参数。本研究控制参数为:规模 30,种群规模大小视样本个数决定,最大迭代次数 120,交叉概率 50%,变异概率 1%,收敛终止条件由迭代次数决定。对特征波段数据集执行 10 次 GAPLS 运算,根据贡献率高、 $RMSECV$  值小的原则提取特征波段,组成特征波段组合。

### 1.4 模型构建方法

为了保证模型构造与反演的稳定性与鲁棒性,本研究采用 Kennard-Stone 方法,选取 75%的数据进行建模,25%的数据进行验证。分别利用偏最小二乘回归 (PLSR)、反向传播神经网络 (BPNN)、随机森林 (RF) 和最小二乘支持向量机 (LSSVM) 4 种机器学习算法构建辣椒叶绿素含量反演模型,并对模型精度进行验证。(1) 偏最小二乘回归 (PLSR)。PLSR 不仅可以有效解决多重线性模型带来的问题,还可以解决诸如主成分分析等分析方法能解决的问题,进而提供更加深入、有用的信息<sup>[21-22]</sup>。(2) 随机森林 (RF)。RF 可以对数据进行分类和回归,它将问题划分成多棵树进行解决,从根源上提高模型的准确性,规避拟合现象,主要原理是集成学习思想<sup>[23]</sup>。其参数设置:决策树数为 200,方法为回归,每个树叶的最小观察数为 3。(3) 反向传播神经网络 (BPNN)。BPNN 属于非线性模型,具有很好的监督学习功能,被广泛应用到测绘遥感领域中<sup>[24]</sup>,其参数设置:迭代次数 1 000 次,误差  $1 \times 10^{-5}$ ,学习率 0.05,附加动量因子 0.9,最小性能梯度  $1 \times 10^{-5}$ ,最大失败次数 3。(4) 最小二乘支持向量机 (LSSVM)。LSSVM 将经典 SVM 中的不等式限制转化为等式限制,并通过求解一系列线性方程组来实现,从而提高了 SVM 运行效率,可以处理线性和非线性多元变量建模问题,还可以解决多元变量的复杂性问题<sup>[25]</sup>。

### 1.5 模型参数验证方法

用决定系数 ( $R^2$ )、均方根误差 ( $RMSE$ ) 和相对分析误差 ( $RPD$ ) 对高光谱信息和叶绿素含量进行评估。采用 Chang 等<sup>[26]</sup>提出的  $RPD$  评判等级,即当  $RPD < 1.4$  时,模型将不能对数据进行预测;当  $1.4 < RPD < 2.0$  时,模型可较好地对数据进行预测;而当  $RPD > 2.0$  时,模型可以非常好地对数据进行预测。当  $R^2$  越接近 1,  $RPD$  值越大,  $RMSE$  值越接近 0,模型预测结果越接近真实值。

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

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

$$RPD = SD / RMSE \quad (3)$$

式中,  $n$  为样本数,  $\hat{y}_i$  为反演值,  $y_i$  为实测值,  $\bar{y}$  为

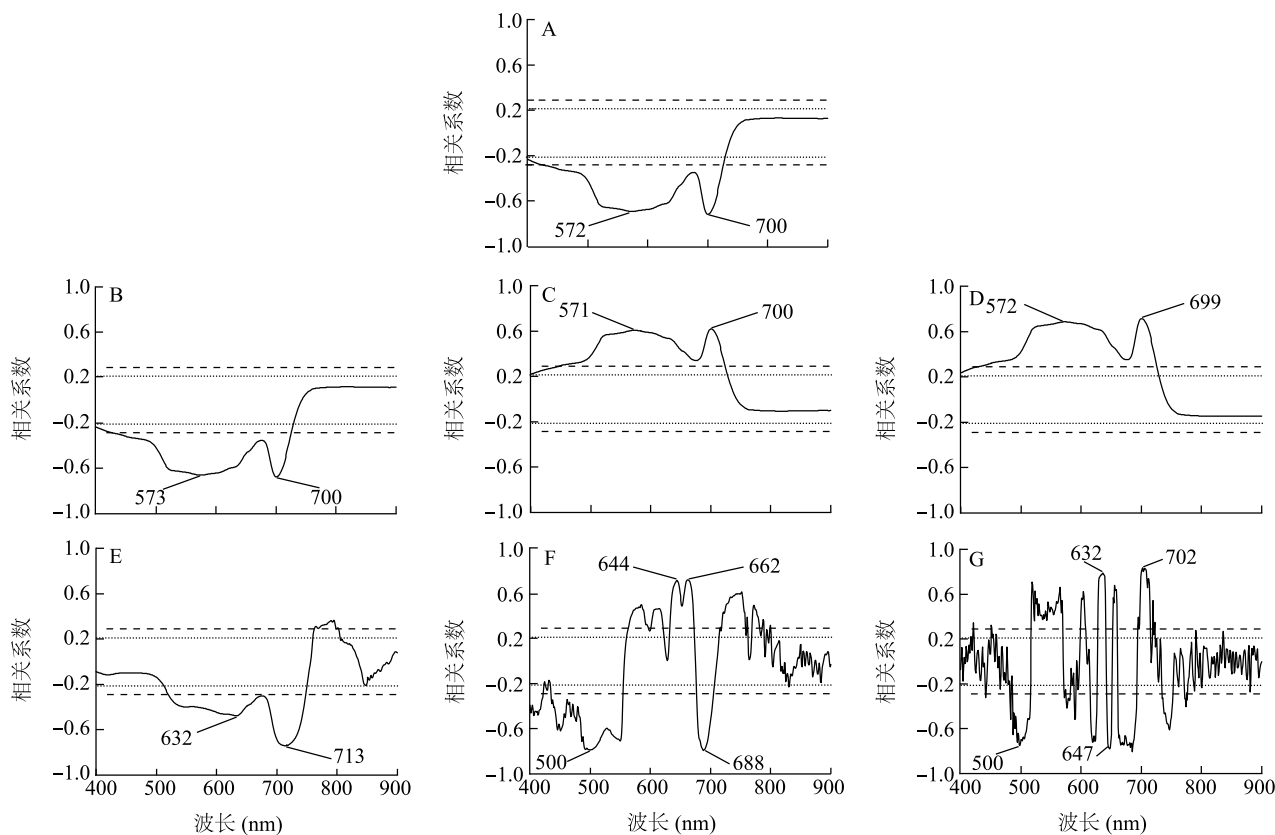
平均值,  $SD$  为  $y_i$  的标准差。

## 2 结果与分析

### 2.1 最大相关系数法生成特征波段数据集

相关系数被广泛用于度量变量间线性相关的强弱程度。将原始光谱的反射率和经处理后的反射率

与 SPAD 值进行 Pearson 相关性分析(图 2)。由图 2A 可见,辣椒原始光谱反射率与 SPAD 值在可见光 421~726 nm 波段呈极显著负相关( $P<0.01$ ),在近红外波段(780~1 000 nm)相关性不显著,相关性最高的波长是 572 nm 和 700 nm,相关系数分别为 -0.717 和 -0.691。



A:原始光谱;B:对数处理;C:倒数处理;D:倒数的对数处理;E:连续统去除处理;F:一阶微分处理;G:二阶微分处理。

图2 辣椒光谱反射率与 SPAD 值的皮尔逊相关性

Fig.2 Pearson correlation analysis between spectral reflectance and SPAD value of pepper

对数处理(LR)、倒数处理(IR)和倒数的对数处理(RL)因光照条件对光谱产生的乘性因素降低,可见光区光谱差异放大<sup>[27-28]</sup>。由图 2B~图 2D 可知,相对于原始光谱反射率,对数处理、倒数处理和倒数的对数处理仅改变了反射率与 SPAD 值相关系数大小及正负相关性,特征波段仅相差 1~2 nm,并没有突出其他特征波段。在 429~725 nm 波段,对数处理后的反射率与 SPAD 值呈极显著负相关( $P<0.01$ ),相关性最好的波长是 573 nm 和 700 nm,相关系数分别为 -0.658 和 -0.675。在 446~724 nm 波段,倒数处理后的反射率和 SPAD 值呈极显著正相

关( $P<0.01$ ),相关性最好的波长是 571 nm 和 700 nm,相关系数分别为 0.606 和 0.619。在 428~726 nm 波段,倒数的对数处理后的反射率和 SPAD 值呈极显著正相关( $P<0.01$ ),相关性最好的波长是 572 nm 和 699 nm,相关系数分别为 0.686 和 0.712。

连续统去除处理可削弱背景噪声,并放大弱吸收特征信息<sup>[29]</sup>。由图 2E 可知,在 520~748 nm 波段,连续统去除处理后的反射率与叶绿素含量呈显著负相关( $P<0.01$ ),相关性最好的波长是 632 nm 和 713 nm,相关系数分别为 -0.478 和 -0.741。

微分处理可提升光谱反射率与样品之间的敏感

度,其中,一阶微分处理主要解决的是线性杂声带来的影响问题,二阶微分处理主要解决的是土壤、其他植被等背景带来的影响问题,提高分析精度,这 2 种微分处理是近红外常用的光谱预处理方法<sup>[30]</sup>。从图 2F 可见,在 400~421 nm、435~554 nm、677~704 nm 波段一阶微分处理后的反射率和 SPAD 值呈极显著负相关( $P<0.01$ ),相关性最好的波长为 500 nm 和 688 nm,其对应的相关系数值分别为-0.789 和-0.791,在 563~621 nm、633~672 nm、716~758 nm 波段一阶微分处理后的反射率和 SPAD 值呈极显著正相关( $P<0.01$ ),相关性最好的波长为 644 nm 和 662 nm,其对应的相关系数值分别为 0.718 和 0.716。在 483~541 nm、573~580 nm、616~625 nm、643~651 nm、662~692 nm、736~752 nm 波段,二阶微分处理后的反射率和 SPAD 值呈极显著负相关( $P<0.01$ ),相关性最好的波段为 500 nm 和 647 nm,其对应的相关系数值分别为-0.721 和-0.832;在 518~568 nm、626~640 nm、654~659 nm、696~714 nm 波段,二阶微分处理后的反射率和 SPAD 值呈极显著正相关( $P<0.01$ ),相关性最好的波长为 632 nm 和 702 nm,其对应的相关系数值分别为 0.819 和 0.845。

经过 MCC 降维处理,7 种光谱数据由 4 027 个特征变量减少到 18 个,并由这 18 个特征变量组成特征波段数据集。

## 2.2 GAPLS 选择变量生成最优特征波段组合

用 GAPLS 算法对 MCC 生成的特征波段数据集进一步筛选,进而建立更加稳定、反演能力更强的模型。由表 1 和表 2 可见,当变量选择数为 1、2 时,交叉验证均方根误差( $RMSECV$ )在 0.2 以上,误差较大,当变量选择数>2, $RMSECV$ 降至 0.08 左右。根据贡献率高、 $RMSECV$  小的筛选原则选择特征波段,当变量选择数为 5,可获得最高贡献率(99.34%),最低  $RMSECV$  (0.081 0),故选出第 10、8、2、15、18 波段作为最优特征波段组合,最优波段和对应处理分别为 700 nm(原始光谱)、699 nm(对数处理)、713 nm(连续统去除处理)、500 nm(二阶微分处理)、713 nm(二阶微分处理)。

## 2.3 回归建模与评价

将 MCC 生成的特征波段数据集作为自变量输入到 4 种机器学习回归模型中,验证集精度评价结果如表 3 所示,根据评价指标  $R^2$  排序,4 种模型的精

度由高到低为 LSSVM>RF>BPNN>PLSR;根据评价指标  $RMSE$  排序,4 种模型的精度由高到低为 PLSR>LSSVM>RF>BPNN;根据评价指标  $RPD$  排序,4 种模型的精度由高到低为 PLSR>BPNN>RF>LSSVM。综合以上结果可以看出,线性模型优于非线性模型,三个非线性模型精度各异。

表 1 累积贡献率及交叉验证均方根误差

Table 1 Cumulative contribution rate and root mean square error of cross validation

选择变量数	累积贡献率(%)	交叉验证均方根误差
1	91.06	0.297 1
2	94.50	0.233 0
3	99.20	0.088 9
4	99.33	0.081 3
5	99.34	0.081 0
6	99.32	0.082 0
⋮	⋮	⋮

表 2 筛选出的波段与其对应被选用频次

Table 2 Selected bands and their corresponding selected frequencies

序号	波段	被选用频次
1	10	86
2	8	48
3	2	46
4	15	45
5	18	39
⋮	⋮	⋮

表 3 GAPLS 降维前的回归模型精度

Table 3 Regression model accuracy before dimensionality reduction by GAPLS

模型	$R^2$	$RMSE$	$RPD$
PLSR	0.45	2.77	1.92
RF	0.65	5.42	1.42
BPNN	0.62	6.36	1.62
LSSVM	0.67	4.96	1.28

PLSR:偏最小二乘法回归;LSSVM:最小二乘支持向量机;BPNN:后向传播神经网络;RF:随机森林。 $R^2$ :决定系数; $RMSE$ :均方根误差; $RPD$ :相对分析误差。

将经过 GAPLS 降维处理后得到的最优特征波段组合作为自变量也输入 4 种机器学习回归模型中,同样用 60 个校正样本进行模型构建,20 个验证

样本进行模型验证,其精度评价结果如表 4 所示。4 种回归模型  $R^2$  均大于 0.7,4 个模型反演值和实测值间均极显著相关( $P<0.01$ )。GAPLS 降维处理后的 PLSR 模型的  $RMSE$  最小,为 1.94,其余 3 种模型的  $RMSE$  在 4 以上。4 种模型的  $RPD$  均大于 1.4,均可对叶绿素含量进行粗略反演。GAPLS 降维处理后的 PLSR 模型的  $RPD$  最高,在 2 以上,达到了 4.55,表明该模型具有较好的反演能力。

根据  $R^2$  值越接近 1、 $RMSE$  值越小、 $RPD$  值越大模型越好的原则,通过对不同模型的反演值和实测值进行拟合(图 3),可以看出,4 种模型均能很好反演辣椒叶绿素含量。PLSR 模型反演效果最好,曲线斜率为 0.913 9,最接近 1,其次是 BPNN 模型,斜

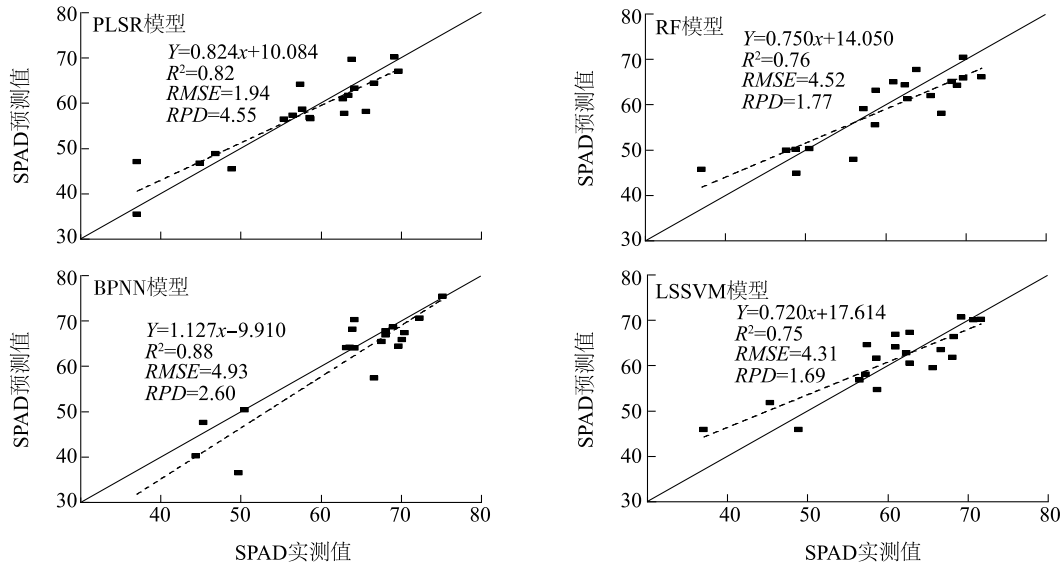
率达到 1.127 2。综上分析,GAPLS 降维处理后的 PLSR 模型最优,其  $R^2$  为 0.82, $RMSE$  为 1.94, $RPD$  为 4.55,对辣椒叶绿素含量反演效果最佳。

表 4 GAPLS 降维后的回归模型精度

Table 4 Regression model accuracy after dimensionality reduction by GAPLS

模型	$R^2$	$RMSE$	$RPD$
PLSR	0.82	1.94	4.55
RF	0.76	4.52	1.77
BPNN	0.88	4.93	2.60
LSSVM	0.75	4.31	1.69

PLSR:偏最小二乘法回归;LSSVM:最小二乘支持向量机;BPNN:后向传播神经网络;RF:随机森林。 $R^2$ :决定系数; $RMSE$ :均方根误差; $RPD$ :相对分析误差。



PLSR:偏最小二乘法回归;LSSVM:最小二乘支持向量机;BPNN:后向传播神经网络;RF:随机森林。 $R^2$ :决定系数; $RMSE$ :均方根误差; $RPD$ :相对分析误差。

图 3 4 种回归模型的拟合度

Fig.3 The fitting degree of four regression models

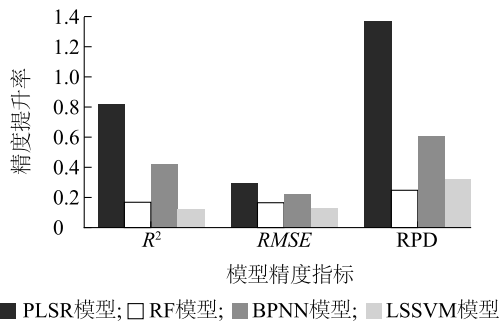
GAPLS 降维处理前后对比可见(图 4),GAPLS 降维处理后模型精度普遍比未经 GAPLS 降维处理升高,说明 GAPLS 处理达到了进一步降低 MCC 生成的特征波段数据集维数的效果。同时不同模型精度提升情况存在差异,精度提升率最高的是 GAPLS 降维处理后的 PLSR 模型,其次是 GAPLS 降维处理后的 BPNN 模型。

### 3 讨论

为了提高光谱数据与叶绿素含量的相关性,减

少噪声等环境的影响,本研究对辣椒冠层光谱反射率数据进行连续统去除等 6 种数学变换处理,并在此基础上用最大相关系数法筛选出 18 个特征波段,这些对辣椒叶绿素含量敏感的波段多位于 500 nm、570 nm、640 nm、700 nm 处,与大多数绿色植物特征波段<sup>[31]</sup>一致,方慧等<sup>[32]</sup>指出叶绿素响应光谱主要集中在红边位置(波段为 700 nm 左右)及绿峰位置(波段为 500~600 nm)。在 6 种数学变换处理中,微分处理后的反射率与 SPAD 值相关性最强,在 702 nm 处相关性系数达到最大值 0.845,这可能是由于





PLSR: 偏最小二乘法回归; LSSVM: 最小二乘支持向量机; BPNN: 后向传播神经网络; RF: 随机森林。  $R^2$ : 决定系数; RMSE: 均方根误差; RPD: 相对分析误差。

图4 GAPLS 降维处理后模型精度提升情况

Fig.4 Improvement of model accuracy after dimensionality reduction by GAPLS

辣椒没有封行现象,叶片空隙较大,辣椒栽培时的薄膜和土壤会产生较大背景噪声而影响光谱反射率,而微分处理能够较好地消除这些噪声。对数、倒数和倒数的对数处理后的特征波段与原始光谱的特征波段位置大致不变,这可能是3种变换的主要作用是降低光照影响因素,放大光谱差异,而本试验在测量辣椒光谱反射率过程中受光照条件因素影响不大。

特征光谱数据集经 GAPLS 算法降维处理后,4种机器学习回归模型的精度均有所提升,这主要是因为 GAPLS 算法改进了全局优化搜索能力,同时降低了特征维数,使低维数据更具解释性,从而获得更好的识别精度,不仅避免了光谱数据局部最优问题,还能解决光谱数据多重共线性问题,更利于搭建一个稳定、精度高的模型<sup>[33]</sup>。

PLSR、RF、LSSVM、BPNN 模型的  $R^2$ 、RMSE、RPD 均表现出较好的预测精度,故4个模型都具有较好的反演能力,其中 PLSR 模型的反演精度最好,远高于其他3种模型,可能是由于经过 GAPLS 处理后的最优特征波段数量减少,且已较大程度消除了共线性、数据噪声等问题<sup>[34]</sup>,故由2个变量构建的线性回归模型表现出较好的反演能力,在高维数据、高变量时表现更加突出,孔钰如等<sup>[35]</sup>研究结果进一步证实了该结论。而非线性模型中的 BPNN 模型也具有较好的反演效果,这可能是由于辣椒叶绿素含量与最优特征波段组合间除了线性关系之外,还有一定的非线性关系,值得后续进一步研究。

## 4 结论

本研究以西南地区盛花、盛果期辣椒冠层叶片为研究对象,对原始光谱反射率进行数学变换处理,并与辣椒 SPAD 值进行相关性分析,采用 MCC、GAPLS 算法对特征波段数据集进行降维,获取最优波段组合,从而建立辣椒叶绿素含量反演模型,并对各模型的精度进行比较,得到如下结论:(1)辣椒叶片叶绿素的敏感波段主要分布在红光区域和绿光区域,最优波段和对应处理分别为 700 nm(原始光谱)、699 nm(对数处理)、713 nm(连续统去除处理)、500 nm(二阶微分处理)、713 nm(二阶微分处理)。(2)GAPLS 对特征波段数据集的降维处理效果较为明显,各机器学习回归模型精度均呈现不同程度的提升,其中 PLSR 提升效果最突出, $R^2$ 、RPD 分别提升了 0.37、0.83, RMSE 降低了 2.63。(3)4种机器学习回归模型中,PLSR 模型的精度最好,由 MCC-GAPLS-PLSR 构建的辣椒叶绿素含量反演模型具有较好的反演潜力。

## 参考文献:

- [1] 王雪,吉海彦. 基于便携式光谱仪和反射、透射光谱技术的小麦叶片叶绿素定量分析研究[J]. 中国农学通报, 2011, 27(21): 39-43.
- [2] ZHENG F L, XU B, XIAO P F, et al. Estimation of chlorophyll content in mountain steppe using in situ hyperspectral measurements[J]. Spectroscopy Letters, 2021, 54(7): 495-506.
- [3] JANG S H, HWANG Y K, LEE H J, et al. Selecting significant wavelengths to predict chlorophyll content of grafted cucumber seedlings using hyperspectral images[J]. Korean Journal of Remote Sensing, 2018, 34(4): 681-692.
- [4] GUO Y H, CHEN S Z, LI X X, et al. Machine learning-based approaches for predicting SPAD values of maize using multi-spectral images[J]. Remote Sensing, 2022, 14(6): 1337.
- [5] 陈澜,常庆瑞,高一帆,等. 猕猴桃叶片叶绿素含量高光谱反演模型研究[J]. 西北农林科技大学学报(自然科学版), 2020, 48(6): 79-89, 98.
- [6] 徐若涵,杨再强,申梦吟,等. 苗期低温胁迫对“红颜”草莓叶绿素含量及冠层高光谱的影响[J]. 中国农业气象, 2022, 43(2): 148-158.
- [7] 李长春,施锦锦,马春艳,等. 基于小波变换和分数阶微分的冬小麦叶绿素含量反演[J]. 农业机械学报, 2021, 52(8): 172-182.
- [8] JIANG X P, ZHEN J N, MIAO J, et al. Newly-developed three-band hyperspectral vegetation index for estimating leaf relative chlorophyll content of mangrove under different severities of pest



- and disease[J]. *Ecological Indicators*, 2022, 140: 108978.
- [9] 依尔夏提·阿不来提,买买提·沙吾提,白灯莎·买买提艾力,等. 基于随机森林法的棉花叶片叶绿素含量反演[J]. *作物学报*, 2019, 45(1): 81-90.
- [10] 李长春,施锦锦,马春艳,等. 基于小波变换和分数阶微分的冬小麦叶绿素含量反演[J]. *农业机械学报*, 2021, 52(8): 172-182.
- [11] 苑迎春,周 毅,宋宇斐,等. 基于信息熵特征选择的小麦冠层叶绿素含量估测方法[J]. *农业机械学报*, 2022, 53(8): 186-195.
- [12] 张佳伟,王仲林,谭先明,等. 利用不同红边位置算法估测玉米叶绿素含量[J]. *浙江大学学报(农业与生命科学版)*, 2021, 47(4): 464-472.
- [13] 刘 涛,张 寰,王志业,等. 利用无人机多光谱反演小麦叶面积指数和叶绿素含量[J]. *农业工程学报*, 2021, 37(19): 65-72.
- [14] CANDIANI G, TAGLIABUE G, PANIGADA C, et al. Evaluation of hybrid models to estimate chlorophyll II and nitrogen content of maize crops in the framework of the future CHIME mission[J]. *Remote Sensing*, 2022, 14(8): 1792.
- [15] 王芳东,严志雁,赵小敏,等. 油茶叶片叶绿素含量高光谱估测的偏最小二乘模型参数选择[J]. *江西农业大学学报*, 2022, 44(1): 86-96.
- [16] NAUŠ J, PROKOPOVÁ J, ŘEBÍČEK J, et al. SPAD chlorophyll II meter reading can be pronouncedly affected by chloroplast movement[J]. *Photosynthesis Research*, 2010, 105(3): 265-271.
- [17] 付波霖,邓良超,张 丽,等. 联合星载高光谱影像和堆栈集成学习回归算法的红树林冠层叶绿素含量遥感反演[J]. *遥感学报*, 2022, 26(6): 1182-1205.
- [18] 李 胜,张培林,李 兵,等. 量子 GA-PLS 特征选择算法及其应用[J]. *量子电子学报*, 2014, 31(2): 194-201.
- [19] JIA W K, ZHAO D A, DING L. An optimized RBF neural network algorithm based on partial least squares and genetic algorithm for classification of small sample[J]. *Applied Soft Computing*, 2016, 48: 373-384.
- [20] SONG K S, LI L, LI S, et al. Hyperspectral retrieval of phycocyanin in potable water sources using genetic algorithm-partial least squares (GA-PLS) modeling[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2012, 18: 368-385.
- [21] 纪文君. 基于野外 vis-NIR 高光谱的土壤属性反演及田间水分影响去除研究[D]. 杭州:浙江大学, 2014.
- [22] 彭晓伟,张爱军,杨晓楠,等. 谷子叶绿素含量高光谱特征分析及其反演模型构建[J]. *干旱地区农业研究*, 2022, 40(2): 69-77.
- [23] ZHOU W, YANG H, XIE L J, et al. Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model[J]. *CATENA*, 2021, 202: 105222.
- [24] 郭云开,许 敏,张晓炯,等. 结合 PRO-4SAIL 和 BP 神经网络的叶绿素含量高光谱反演[J]. *测绘通报*, 2020(3): 21-24.
- [25] SHAO Y H, LI C N, HUANG L W, et al. Joint sample and feature selection via sparse primal and dual LSSVM[J]. *Knowledge-Based Systems*, 2019, 185: 104915.
- [26] CHANG C W, LAIRD D A, MAUSBACH M J, et al. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties[J]. *Soil Science Society of America Journal*, 2001, 65(2): 480-490.
- [27] SERRANO L, PEÑUELAS J, USTIN S L. Remote sensing of nitrogen and lignin in Mediterranean vegetation from AVIRIS data: Decomposing biochemical from structural signals[J]. *Remote Sensing of Environment*, 2002, 81(2): 355-364.
- [28] 李岚涛,汪善勤,任 涛,等. 基于高光谱的冬油菜叶片磷含量诊断模型[J]. *农业工程学报*, 2016, 32(14): 209-218.
- [29] 郭超凡,郭道宇. 基于可见光波段包络线去除的湿地植物叶片叶绿素反演[J]. *生态学报*, 2016, 36(20): 6538-6546.
- [30] 李敏夏,张林森,李丙智,等. 苹果叶片高光谱特性与叶绿素含量和 SPAD 值的关系[J]. *西北林学院学报*, 2010, 25(2): 35-39.
- [31] 罗 靖,杨再强,杨 立,等. 基于高光谱参数建立苗期高温条件下草莓叶片叶绿素含量反演模型[J]. *中国农业气象*, 2022, 43(10): 832-845.
- [32] 方 慧,宋海燕,曹 芳,等. 油菜叶片的光谱特征与叶绿素含量之间的关系研究[J]. *光谱学与光谱分析*, 2007, 27(9): 1731-1734.
- [33] 蔡建楠,刘海龙,姜 波,等. 基于 GA-PLS 算法的河网水体化学需氧量高光谱反演[J]. *灌溉排水学报*, 2020, 39(9): 126-131.
- [34] WANG C, FENG M C, YANG W D, et al. Extraction of sensitive bands for monitoring the winter wheat (*Triticum aestivum*) growth status and yields based on the spectral reflectance[J]. *PLoS One*, 2017, 12(1): e167679.
- [35] 孔钰如,王李娟,冯海宽,等. 无人机高光谱波段选择的叶面积指数反演[J]. *光谱学与光谱分析*, 2022, 42(3): 933-939.

(责任编辑:成纾寒)