油菜光谱的函数特征分析及叶绿素诊断建模

许 健

(湖南农业大学 信息与智能科学技术学院, 湖南 长沙 410128)

摘 要: 作物冠层光谱的特征及其变化是各种作物信息提取的基础, 是作物信息学的核心研究内容之一. 函数型数据分析将观测到的一条光谱看作一个整体, 从函数的角度描述光谱曲线变化特征. 为探索油菜冠层光谱与叶绿素含量的关系, 以 24 个移栽种植小区和 24 个直播种植小区的高油酸油菜在苗期、抽薹期和荚果期三个不同生长期的样本为研究对象, 利用函数型主成分分析方法分析不同生长期的油菜冠层光谱的典型变化特征, 发现不同样本在光谱的可见光、近红外、中红外区间都显示出明显差异. 但是, 利用最大信息系数方法做进一步分析发现, 原始的光吸收信息与叶绿素含量之间的关系微弱, 而不同光谱区间吸光度的比值却显示出与叶绿素含量之间有较强的非线性关系. 鉴于这种复杂的非线性关系, 最终利用增强回归树, 基于不同光谱区间吸光度的比值, 建立油菜叶片叶绿素含量的预测模型, 交互验证的均方根误差为 5.5%, 接近试验测量误差水平.

关键词: 油菜; 光谱; 最大信息系数; 增强回归树; 函数型数据分析

中图分类号: O213.9 **文献标识码**: A **文章编号**: 1672-5298(2023)03-0016-06

Functional Feature Analysis of Rape Spectrum and Chlorophyll Diagnostic Modeling

XU Jian

(College of Information and Intelligence, Hunan Agriculture University, Changsha 410128, China)

Abstract: The characteristics and changes of crop canopy spectrum are the basis of crop information extraction and one of the core research contents of crop informatics. Functional data analysis regards the observed spectrum as a whole, and describes the characteristics of spectral curve from the perspective of function. In order to explore the relationship between canopy spectrum and chlorophyll content of rape, we took the samples of high oleic acid rape from 24 transplant planting plots and 24 direct seeding planting plots at three different growth stages: seedling stage, bolting stage and pod stage. First of all, the functional principal component method was used to analyze the typical characteristics of rape canopy spectrum in different growth periods. It was found that there are significant differences among different samples in the visible, near infrared and mid infrared regions of the spectrum. However, further analysis using the maximum information coefficient method found that the relationship between the original light absorption information and the chlorophyll content is weak, while the absorbance ratio in different spectral intervals shows a strong nonlinear relationship with the chlorophyll content. In view of this complex nonlinear relationship, finally, based on the absorbance ratio of different spectral intervals, a BRT model was established for the prediction of the chlorophyll content of rape leaves. The root mean square error of cross validation is 5.5%, which is close to the error level of test measurement.

Key words: rape; spectrum; maximum information coefficient; boosted regression trees; functional data analysis

0 引言

自 20 世纪 90 年代起, 精确农业从概念走向实践并开始商业化运营, 作物生长状况实时监测是精确农业众多技术环节中的重要一环. 作物信息科学的重点内容是如何利用作物的信息对其进行无损营养诊断, 光谱分析便是一个有效可行的途径. 对于油菜而言, 冠层光谱特征是描述其营养状况的重要指标. 随着观测技术的进步, 光谱分辨率不断提高, 高光谱技术开始广泛用于作物营养状况实时监测. 高光谱技术中的光谱分辨率大大提高, 光谱波段之间存在很强的自相关现象[1,2]. 此时, 光谱适宜于看成是一条连续的曲线, 而曲线的形状变化特征则蕴含了作物信息. 实际上, 许多植被指数都可以看成是对某个特殊的光谱特征的描述, 比如, 比值植被指数、归一化植被指数、土壤调节植被指数等都重点利用了光谱中的红边特征信息.

收稿日期: 2022-07-19

基金项目: 湖南省教育厅科学研究项目(19C0939)

作者简介: 许 健, 男, 博士, 讲师. 主要研究方向: 函数型数据分析

基于叶片或冠层反射光谱的分析技术是最近几十年来广为流行的植物营养状况无损监测手段之一. 绿色植物组织的各种色素以及细胞结构特点导致了其反射光谱中会呈现一些独特的形状特征, 比如红边特征[3,4]. 大量研究证实了红边特征与植物的众多生理物理性质有着密切关系, 比如叶面积指数、生物量、叶绿素密度等[5,6]. 从最初的比值型植被指数以及归一化植被指数开始, 各种改进的光谱植被指数不断被构造出来以提取反射光谱中的信息, 进而用于植物的生物物理性质研究[7~9]. 早期的地球资源卫星上搭载的是分辨率较低的多光谱传感器, 随着传感器技术的进步, 高光谱技术的应用逐渐普及, 研究者已经意识到低分辨率的多光谱技术不足以准确探测植被的生物化学性质, 波段宽度在10 nm以下的高光谱则有着广阔的应用前景[9]. 但是, 基于多光谱时代的植被指数构建方式, 使用高光谱数据建模的效果不一定比使用多光谱数据的效果好[5]. 高光谱所携带的信息量显然要比多光谱大得多, 出现这一困境说明高光谱分析方法还需要进一步开发.

针对高光谱数据, 有许多研究工作延续了多光谱数据的思路, 如首先挑选出少数几个最优波长, 再对经典的基于多光谱的植被指数进行优化或另外构建新的指数[10~13]. 多光谱分析技术习惯于将各个波段当作相互独立的变量来对待, 这样的假设在处理多光谱数据时是合适的, 比如红光波段和近红外波段看起来很像是两个“相互独立”的信息源, 叶绿素在红光波段和近红外波段的光吸收情况也确实存在显著的差异, 但是类似的看法在处理高光谱数据时是不恰当的. 高分辨率情况下, 相邻的波长变量之间的“实体意义”很难区分清楚, 光谱波长之间存在很强的自相关现象, 哪个波长最优本身是有一定模糊性的. 此时, 光谱更应当看成是一条连续的曲线, 而曲线的形状变化特征则反映了作物的信息[14]. 目前, 已经有一些研究工作探讨了从连续曲线的视角来研究光谱数据[15,16].

函数型数据分析(Functional Data Analysis)方法于 20 世纪 80 年代开始在统计学领域提出. 以光谱数据为例, 该方法最大的特点是将整条光谱曲线看作一个整体, 从函数角度研究光谱曲线的变化特征. 通过各个波长的“离散”观测值恢复出背后的函数, 即连续光谱曲线, 进而从整体上研究曲线的典型变化特征, 不再仅仅从最优的几个波长提取光谱信息. 从函数的角度研究光谱特征拓宽了光谱信息提取的思路, 如果选择最优波段构建光谱指数可以看成是一种“离散”的思路, 那么基于函数的方法则代表一种“连续”的思路, 避免了因高度相关导致的最优波段挑选的模糊性, 以及由此造成的信息丢失.

1 理论与方法

1.1 函数型数据分析

用 $\{(t_j, y_j), j = 1, 2, \cdots, n\}$ 表示一条光谱观测记录, 其中 t_j 表示波长, y_j 为相应波段位置上的光谱观测值, n 是光谱中所使用的波长个数. 假设所观测到的离散光谱观测值实际上来自一条连续的光谱曲线, 用函数 $x(t)$ 表示, 则观测值与真实光谱函数的关系为

$$y_j = x(t_j) + \varepsilon_j. \tag{1}$$

即在 t_j 位置上观测到的光谱值 y_j 实际上是由真实的光谱值 $x(t_j)$ 加上一个观测误差 ε_j 而得到. 函数型数据分析方法首先要根据观测到的光谱估计出真实的光谱曲线函数 $x(t)$. 一般的, 一条光谱观测值就可以估计出一个函数, 为了提高估计精度, 也可以同时使用实验中在某个观测点上的若干重复光谱观测值来估计同一个函数. 误差项 ε_j 的方差在所有波长 t_j 上相同或不同, 要根据实际情况决定, 具体的估计方法参考文[17].

函数型主成分分析(Functional PCA)[18,19]是探索函数曲线典型变化特征的一个非常有用的方法. 在函数型 PCA 中, 一个数据样本即为一个函数, 计算出的主成分也是函数, 画出来是一条曲线, 表示的是在一组函数(即若干条曲线)中出现最多的曲线变化特征, 因此函数型 PCA 可以用来识别光谱或导数光谱中的典型变化特征.

1.2 增强回归树

增强回归树(Boosted Regression Trees, BRT)是 Schapire[20]于 2002 年提出的一种机器学习方法，它融合了两类算法——分类回归树与 boosting，通过 boosting 整合大量的分类回归树模型来改进单独一棵树的性能.与 bagging 方法所使用的直接对大量普通模型取平均不同，boosting 是逐步序列式进行的，在建立下一棵树时会以更高的权重考虑在之前已经建立的树上表现不够好的那些样本，强调在这部分样本上改进预测性能，通过这种方式，boosting 不仅能够减小预测方差，还能减小预测偏差. 有关细节参见文[21,22].

1.3 最大信息系数

最大信息系数(Maximal Information Coefficient, MIC)是由 D. Reshef 等提出的一种用于测量两个随机变量之间相依关系的方法，Speed 将其称为 21 世纪的相关系数[23]. MIC 方法具备两个非常吸引人的特点：广义性(generality)和等价性(equitability). 广义性指的是它能够度量变量之间可能存在的各种关系，而不仅仅局限于线性或者某几个明确的函数关系. 而等价性指的是 MIC 的取值大小只与变量之间的关系强度有关，不依赖于具体的关系类型，只要关系强度类似，MIC 取值大小就会差不多. MIC 的取值范围在 0 到 1 之间，取值越靠近于 1 说明两个变量之间的关系越密切，若 MIC = 0，则说明两个变量相互独立.

2 试验与数据

以湖南农业大学浏阳试验基地的 48 个小区的中熟高油酸油菜为研究对象，48 个小区中 24 个为移栽种植方式，另 24 个为直播种植方式. 选择晴朗、无风无云的三个日期：2014 年 12 月 2 日(苗期)，2015 年 1 月 22 日(抽薹期)，2015 年 4 月 16 日(荚果期)，于每个日期的 10:00 至 13:00，利用美国 ASD 公司(Analytical Spectral Device)生产的 FieldSpec® 3 高分辨便携式地物波谱仪进行采样. 波谱仪的有效波段范围为 350 ~ 2500 nm，其中，350 ~ 1000 nm 范围采样间隔为 1.4 nm，光谱分辨率为 3 nm；1000 ~ 2500 nm 范围采样间隔为 2 nm，光谱分辨为 10 nm. 测量时距离冠层垂直高度约 0.7 m，采用标准白板校正，因此所采集光谱为无量纲的相对反射率.

48 个种植小区苗期、抽薹期和荚果期所采集的光谱如图 1 所示. 抽薹期光谱相较于苗期和荚果期，在波长大于 1000 nm 的中红外区域，光谱反射峰明显低于苗期和荚果期. 为了获取叶绿素含量信息，在每个种植小区选取长势平均的 5 个植株的第三片展开叶，利用日本 MINOLI.A 公司生产的 SPAD 502 叶绿素仪，测定 5 个点的叶片 SPAD 值. 具体来说，每个叶片从叶尖到叶枕分成 3 段，在每段的二分之一处进行测定，各段重复 5 次，取平均值，作为该点位的叶片 SPAD 值. 如图 2 所示，与冠层光谱情况类似，抽薹期的叶片 SPAD 分布比其他两个生长期明显要高，平均值达到了 55；苗期和荚果期相比，苗期的 SPAD 值分布相对集中，而荚果期的则更加分散.

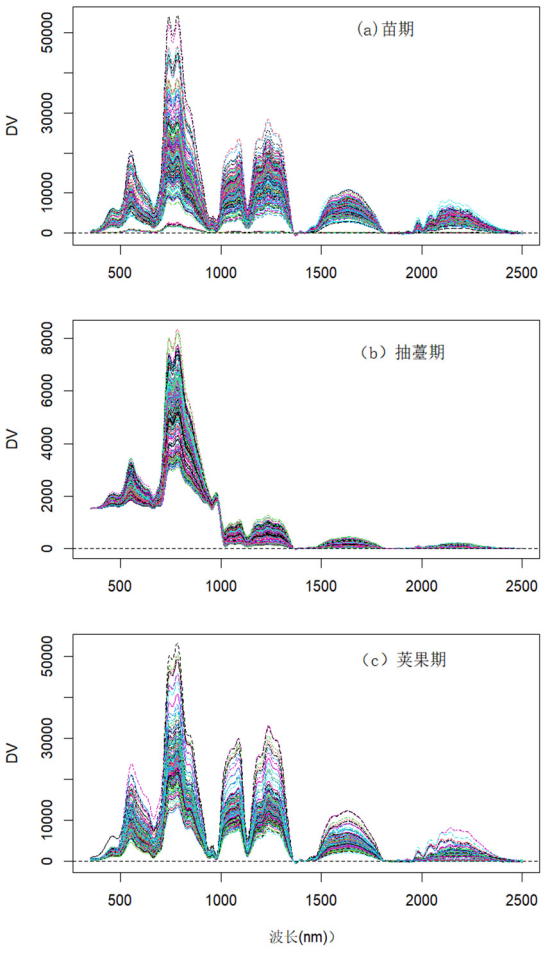

图 1 48 个种植小区所采集的光谱

3 结果与讨论

3.1 光谱特征与函数主成分

图 3 为三个生长期数据的第一、第二主成分函数, 其中第一主成分解释了数据中 82.9%的波动信息, 而第二主成分解释了 5.1%的信息. 图 3 中的实线是光谱平均值, 而“+”和“−”表示的是给均值加上或减去一定量的主成分之后的效应. 容易观察到, 不同样本光谱间的差异主要在各反射峰区域出现, 即“+”和“−”所标示的曲线分离得最开. 比起第一主成分, 第二主成分更加强调 1500 ~ 2500 nm 范围内反射峰的差异信息. 但是, 如图 4 所示, 第一主成分得分值的大小与叶片 SPAD 值之间无明显规律性, 说明光谱在这些反射峰上的波动信息并不能直接用于叶绿素含量建模. 偏最小二乘(PLS)方法是一种使用非常广泛的适用于解释变量远多于样本个数情况的线性建模方法. 图 5 为 5 个主成分的 PLS 模型的交互验证预测情况, 5 个主成分解释了解释变量 99.61%和响应变量 68.63%的信息, 交互验证均方根预测误差(RMSEP)为 4.609, 达到样本 SPAD 平均值的 10.3%. 从图 5 中容易看出, PLS 模型给出的预测值与样本测量值之间几乎没有呈现出线性关系.

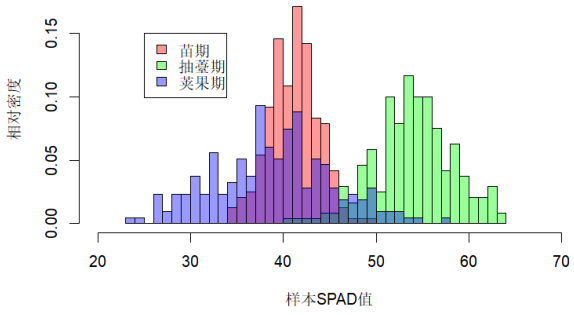

图 2　苗期、抽薹期、荚果期叶片 SPAD 测量值频数分布

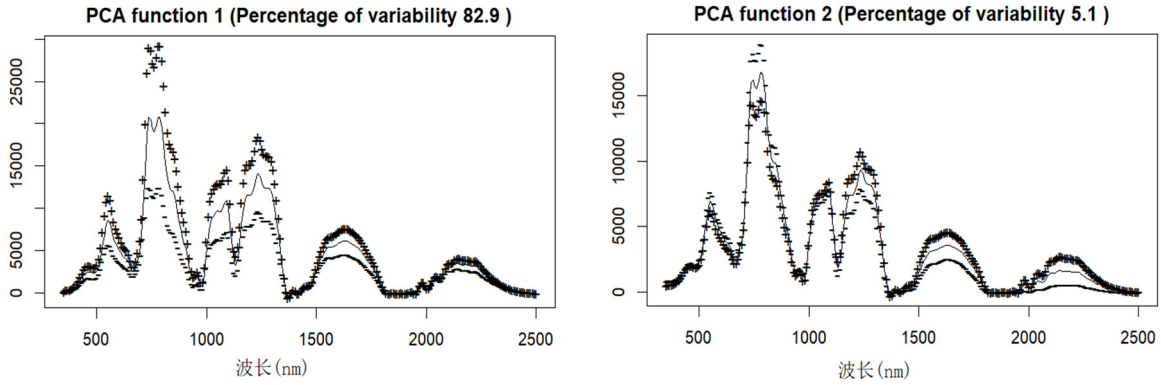

图 3　第一主成分与第二主成分函数

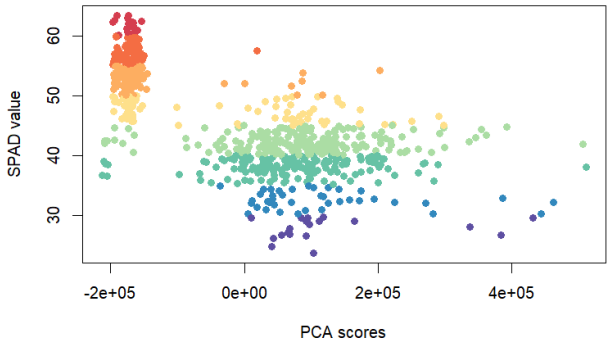

图 4　样本第一主成分得分值与样本 SPAD 值散点图

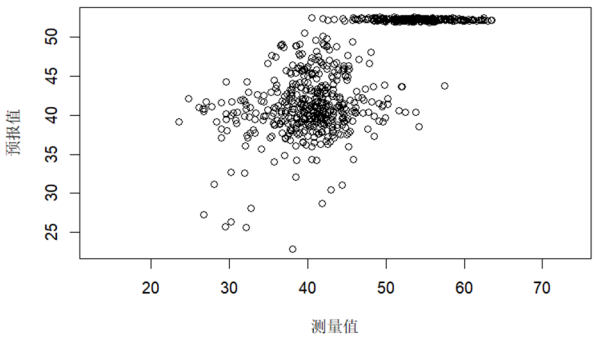

图 5　偏最小二乘模型预测值与样本测量值散点图

3.2 最大信息系数比值反射率特征选择

3.2.1 基于最大信息系数的比值植被指数构造

构造光谱参数通常能够消除光谱中的背景噪声, 提高光谱信息的利用效率[24], 而比值植被指数 RVI

因为构造简单、计算方便, 在植被光谱分析中应用最为广泛. 文[25]根据直线方程、幂函数、指数函数的决定系数大小在众多的比值参数中寻找重要的部分, 这种处理方式的局限性很明显, 因为如果一个重要的比值指数的取值规律与叶绿素含量之间的关系并不是线性函数、幂函数、指数函数中的任何一种, 那么在搜寻过程中就会产生遗漏.

由于并不能事先确定某个植被指数的取值规律与叶片叶绿素含量之间的具体关系是什么, 故本研究使用普适性更强的最大信息系数(MIC)来挖掘重要的比值参数. MIC 的计算开销较大, 如果对原始光谱中每一个离散波段的比值进行计算, 那么 350～2500 nm 范围内需要处理的比值参数多达 4626801 个. 因此, 本研究利用 B 样条基函数, 分段提取光谱曲线中的信息, 考虑到离散光谱波段本身存在高相关性, 这种处理方式既不会造成明显的信息损失, 同时又能大大降低计算开销.

如图 6 所示, 通过 100 个 B 样条基系数概括光谱信息, 计算 100 个基系数比值与对应样本 SPAD 值之间的 MIC 系数. 图 6 中的小矩形色块是对应光谱区段所对应比值的 MIC 值, MIC 值越趋向 1, 颜色越偏向于深红, 表示对应光谱区段比值与叶片 SPAD 值之间存在某种关系; MIC 值越趋向 0, 颜色越偏向于淡黄, 表示对应光谱区段比值与叶片 SPAD 值之间几乎没显示出任何关系. 400～1000 nm 范围与1000～1800 nm 范围的光谱比值与叶片 SPAD 值之间有密切联系, 此外, 350～500 nm 范围与 500～900 nm 范围的光谱比值, 以及1380～1500 nm 范围与1500～1800 nm 范围的光谱比值, 与叶片 SPAD 值之间关系密切. 图 7 为通过 MIC 找到的两个典型的比值指数取值与对应样本 SPAD 值的散点图, 其中(a)横坐标: 440~463 nm 区间与 531~553 nm 区间光谱反射率比值; (b)横坐标: 1187~1210 nm 区间与 1798~1821 nm 区间光谱反射率比值. 显然, 样本 SPAD 值确实随着光谱区间比值变化而变化, 但是其变化规律却无法用某个已知函数准确表达.

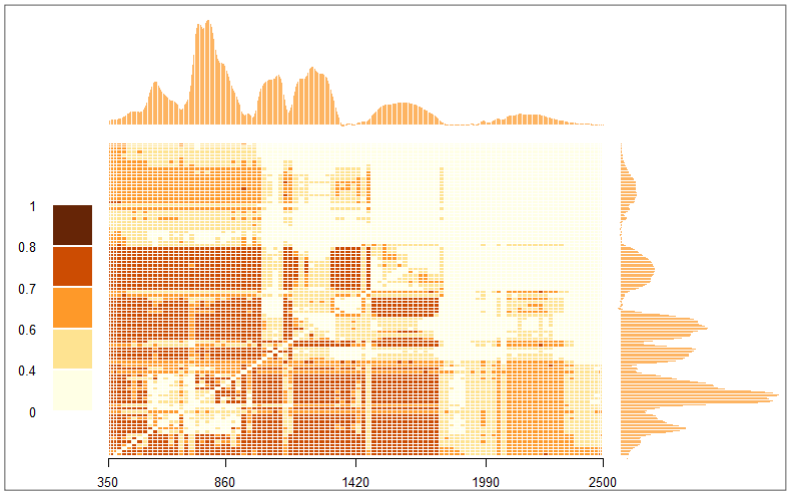

图 6 基于 B 样条基函数的比值参数与叶片 SPAD 值的 MIC 强度

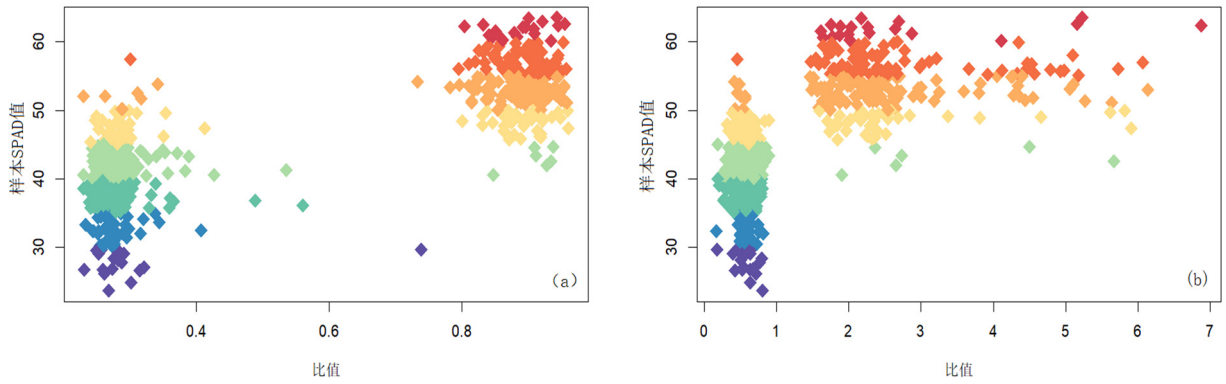

图 7 两个典型的比值指数取值与对应样本 SPAD 值的散点图

3.2.2 基于增强回归树的叶片 SPAD 值预测建模

设定树复杂度(树的节点数)为 5, 学习率(learning rate)为 0.005, 同时为了降低计算时间开销, 根据 MIC 值大小, 从 100 个 B 样条基系数两两之间的比值中挑选最大的 2686 个, 作为解释变量用于叶片 SPAD

值的增强回归树建模. 所挑选的 2686 个比值主要集中在 400～1000 nm 范围与 1000～1800 nm 范围的光谱比值, 350～500 nm 范围与 500～900 nm 范围的光谱比值, 以及 1380～1500 nm 范围与 1500～1800 nm 范围的光谱比值. 最终共建立回归树 1350 棵, 图 8 为最终的增强回归树模型交互验证预测情况, 均方根误差为 2.48, 为 SPAD 测量平均值的 5.5%, 已经接近实验数据的测量误差. 对比于 3.1 节中所采用的偏最小二乘方法, 预测精确度提升十分明显.

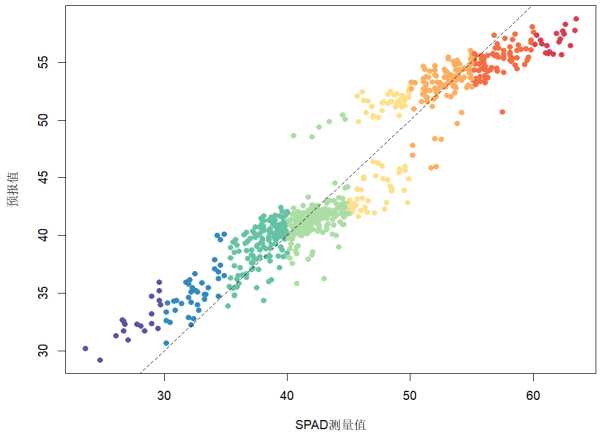

图 8　增强回归树模型交互验证预测情况

4 结束语

为研究油菜冠层光谱的函数特征与叶片叶绿素含量之间的关系, 本文以苗期、抽薹期和荚果期三个生长期的油菜为研究对象, 利用 B 样条基函数提取局部光谱片段的信息, 构建比值植被指数, 再利用增强回归树建立油菜叶片 SPAD 值定量预测模型. 油菜冠层光谱样品间的差异主要出现在各反射峰位置, 但是这种差异对叶片 SPAD 值的影响并非简单线性的, 利用增强回归树模型则能够有效处理二者间的复杂关系, 建立有效的 SPAD 预测模型.

参考文献:

[1] 张国云, 欧阳慧婷, 涂　兵, 等. 空间一致核协同优化的高光谱异常检测方法[J]. 湖南理工学院学报(自然科学版), 2022, 35(3):10−16+43.
[2] 易嘉闻, 李　希, 欧阳尔, 等. 基于自编码的高光谱图像波段加权分类网络研究[J]. 湖南理工学院学报(自然科学版), 2021, 34(1): 34−39.
[3] Horler D N H, Dockray M, Barber J. The red edge of plant leaf reflectance[J]. International Journal of Remote Sensing, 1983, 4(2): 273−288.
[4] Jordan C F. Derivation of leaf-area index from quality of light on the forest floor[J]. Ecology, 1969, 50(4): 663−666.
[5] Broge N H, Leblanc E. Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density[J]. Remote Sensing of Environment, 2000, 76: 156−172.
[6] Vogelmann J E, Rock B N, Moss D M. Red edge spectral measurements from sugar maple leaves[J]. International Journal of Remote Sensing, 1993, 14(8): 1563−1575.
[7] Xue J, Su B. Significant remote sensing vegetation indices: a review of developments and applications[J]. Journal of Sensors, 2017, 2017: 1353691.1−1353691.17.
[8] Blackburn G A. Spectral indices for estimating photosynthetic pigment concentrations: a test using senescent tree leaves[J]. International Journal of Remote Sensing, 1998, 19(4): 657−675.
[9] Blackburn G A. Hyperspectral remote sensing of plant pigments[J]. Journal of Experimental Botany, 2007, 58(4): 855−867.
[10] Chappelle E W, Kim M S. Ratio analysis of reflectance spectra (RARS): an algorithm for the remote estimation of the concentrations of chlorophyll A, chlorophyll B, and carotenoids in soybean leaves[J]. Remote Sensing of Environment, 1992, 39: 239−247.
[11] Elvidge C D, Chen Z. Comparison of broad-band and narrow-band red and near-infrared vegetation indices[J]. Remote Sensing of Environment, 1995, 54: 38−48.
[12] Huang J, Wang F, Wang X, et al. Relationship between narrow band normalized deference vegetation index and rice agronomic variables[J]. Communications in Soil Science and Plant Analysis, 2004, 35(19-20): 2689−2708.
[13] Thenkabail P S, Smith R B, Pauw E D. Hyperspectral vegetation indices and their relationships with agricultural crop characteristics[J]. Remote Sensing of Environment, 2000, 71: 158−182.
[14] Tsai F, Philpot W. Derivative analysis of hyperspectral data[J]. Remote Sensing of Environment, 1998, 66: 41−51.
[15] Talsky G, Mayring L, Kreuzer H. High-resolution, higher-order UV/VIS derivative spectrophotometry[J]. Angewandte Chemie International Edition in English, 1978, 17(11): 785−799.
[16] Demetriades-Shah T H, Steven M D, Clark J A. High resolution derivative spectra in remote sensing[J]. Remote Sensing of Environment, 1990, 33: 55−64.
[17] Wang J L, Chiou J M, Müller H G. Functional data analysis[J]. Annual Review of Statistics and Its Application, 2016, 3: 257−295.
[18] Dong J J, Wang L, Gill J, et al. Functional principal component analysis of glomerular filtration rate curves after kidney transplant[J]. Statistical Methods in Medical Research, 2018, 27(12): 3785−3796.
[19] 宁　贺. 函数型主成分分析的研究及其应用[D]. 长春: 吉林大学, 2021.
[20] Denison D D, Hansen M H, Holmes C C, et al. Nonlinear Estimation and Classification[M]. New York: Springer, 2002.
[21] Elith J, Graham C H, Anderson R P, et al. Novel methods improve prediction of species' distributions from occurrence data[J]. Ecography, 2006, 29(2): 129−151.
[22] Leathwick J R, Elith J, Francis M P, et al. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees[J]. Marine Ecology Progress Series, 2006, 321: 267−281.
[23] Speed T. A correlation for the 21st century[J]. Science, 2011, 334: 1502−1503.
[24] 王福民, 黄敬峰, 唐延林, 等. 采用不同光谱波段宽度的归一化植被指数估算水稻叶面积指数[J]. 应用生态学报, 2007, 18(11): 2444−2450.
[25] 姚　霞, 朱　艳, 冯　伟, 等. 监测小麦叶片氮积累量的新高光谱特征波段及比值植被指数[J]. 光谱学与光谱分析, 2009, 29(8): 2191−2195.