张树艳,王有武,白铁成,等. 基于近红外光谱和 SPA 算法的棉花叶面积指数定量分析[J]. 江苏农业科学,2018,46(6):174 – 177.
doi:10.15889/j.issn.1002 – 1302.2018.06.045

基于近红外光谱和 SPA 算法的棉花叶面积指数定量分析

张树艳[1,3], 王有武[2], 白铁成[1,3], 张　晓[1,3], 石鲁珍[1]
(1. 塔里木大学信息工程学院,新疆阿拉尔 843300; 2. 塔里木大学植物科学学院,新疆阿拉尔 843300;
3. 中国农业科学院农业信息研究所新疆南疆农业信息化研究中心,新疆阿拉尔 843300)

摘要:棉花叶面积指数是作物长势诊断和产量预测的重要参数。运用近红外光谱仪测定不同生育时期的棉花冠层光谱反射率,通过连续投影算法从 1 557 个近红外光谱波长中提取出 5 个有效特征波长,然后用最小二乘法对叶面积指数进行建模。将连续投影算法和最小二乘法(简称 SPA – PLS)模型与全光谱建立的 PLS 模型预测结果进行比较,预测相关系数(r)由 0.801 23 提高到 0.928 27,预测均方根误差($RMSEP$)由 0.501 22 降低到 0.294 7,建模均方根误差($RMSPCV$)由 0.425 33 降低到 0.294 20。结果表明,SPA – PLS 模型仅用占全波段 0.32% 的特征波长建模,不仅缩短了运算时间,而且模型精度、预测能力和稳定性均得到明显提高。

关键词:近红外光谱;偏最小二乘法;波长提取;连续投影算法;棉花;叶面积指数
中图分类号: O657.33;S126　　**文献标志码**: A　　**文章编号**:1002 – 1302(2018)06 – 0174 – 04

棉花是关系国计民生的重要物资,是仅次于粮食的第二大农作物,其产值占我国经济作物的 50% 以上,在国民经济发展中具有重要地位。新疆以其优越的光热资源条件成为我国最主要的棉花产区,棉花种植面积、单位面积产量和总产量一直居全国首位[1-2]。叶面积指数(leaf area index,简称 LAI)很好地反映了冠层结构是否合理、营养生长与生殖生长是否协调及其生育进程等信息,与生物量和作物产量密切相关,是群体特征的重要指标[3-4]。因此,棉花不同生育时期 LAI 的精确估测,对了解棉花长势、提高新疆棉花生产管理水平及遥感估产有着重要意义。

目前,利用高光谱获取 LAI 已经成为精准农业研究的热点问题之一[5-6]。植被冠层叶片特别是宽叶片在近红外光谱(简称 NIR)区域的高反射率和透射率可引起强烈的多重反射,NIR 光谱区(700 ~ 2 500 nm)主要是由含氢基团的倍频和组频吸收峰组成,吸收强度弱,灵敏度相对较低,吸收带较宽且重叠严重,近红外光谱通常包含数以千计的波长变量,光谱信息存在多重相关性等,如果采用全光谱数据建模,由于光谱含有大量冗余数据,必然会增加建模的工作量。因此,为了削弱以至于消除各种非目标因素对近红外光谱的影响,提高物系性质参数对光谱的分辨率和灵敏度,在利用光谱建立校正模型前,通常需对其进行波长选择[7],剔除不含有用信息的波长。另外,选择有较好代表性的校正集样本,可以提高预测模型的预测能力。

鉴于此,本研究以南疆棉花为研究对象,采用近红外光谱仪获得棉花冠层光谱,通过基于 X – Y 距离的样本集划分(sample set partitioning based on joint x – y distance,简称

收稿日期:2017 – 07 – 15
基金项目:国家自然科学基金应急管理项目(编号:61640413);兵团科技攻关与成果转化计划(编号:2015AC013);塔里木大学青年创新资金(编号:TDZKQN201508)。
作者简介:张树艳(1984—),女,甘肃白银人,硕士,讲师,主要从事农业信息化和光谱图像方面的研究。E – mail:zhangsy84@163.com。
通信作者:王有武,硕士,副教授,主要从事作物育种研究。E – mail:wangyw1975@126.com。

[5]雷　霖,代传龙,王厚军. 基于 Rough set 理论的无线传感器网络节点故障诊断[J]. 北京邮电大学学报,2007,30(4):69 – 73.
[6]冯志刚,王　祁,徐　涛,等. 基于小波包和支持向量机的传感器故障诊断方法[J]. 南京理工大学学报(自然科学版),2008,32(5):609 – 614.
[7]Jiang P. A new method for node fault detection in wireless sensor networks[J]. Sensors,2009,9(2):1282 – 1294.
[8]庄　夏,戴　敏,何元清. 基于人工免疫和模糊 K 均值的传感器节点故障诊断[J]. 计算机测量与控制,2013,21(3):611 – 613.
[9]Huang G B,Zhu Q Y,Siew C K,et al. Extreme learning machine:theory and applications[J]. Neurocomputing,2006,70(1/2/3):489 – 501.
[10]Pan C,Park D S,Yang Y,et al. Leukocyte image segmentation by visual attention and extreme learning machine[J]. Neural Computing and Applications,2012,21(6):1217 – 1227.
[11]Yang Y M,Wang Y N,Yuan X F. Bidirectional extreme learning machine for regression problem and its learning effectiveness[J]. IEEE Transactions on Neural Networks and Learning Systems,2012,23(9):1498 – 1505
[12]李智敏,陈祥光. 无线传感器节点模块级故障诊断方法的研究[J]. 仪器仪表学报,2013,34(12):2763 – 2769.
[13]谢迎新,陈祥光,余向明,等. 基于 VPRS 和 RBF 神经网络的 WSN 节点故障诊断[J]. 北京理工大学学报,2010,30(7):807 – 811.
[14]余成波,李　芮,何　强,等. 基于粒子群算法及高斯分布的 WSN 节点故障诊断[J]. 振动、测试与诊断,2013,33(1):149 – 152,172.

SPXY)法划分校正集样本和验证集样本,然后使用连续投影算法剔除光谱冗余信息,优选出棉花近红外特征波长,结合最小二乘法实现 LAI 的建模,比较连续投影算法和最小二乘法(简称 SPA－PLS)模型和 PLS 模型的预测精度和稳定性,以期为棉花叶面积指数的精确估测提供一种新的思路和方法。

1　材料与方法

1.1　试验地概况

本试验设于新疆阿拉尔市十团六连棉花试验区,地理坐标为 81°13′E,40°34′N,为典型的大陆性干旱荒漠气候,年均相对湿度为 51%,太阳辐射总量为年均 6 100 MJ/m^2 左右,生长季太阳辐射量为 1 300 MJ/m^2 左右,年均日照时数为 2 800～3 000 h,云雾天气较少,扬尘、浮尘、沙暴等天气较多。

1.2　试验设计

本试验于 2015 年实施,棉花品种为新陆中 67 号,小区面积为 300 m^2,种植密度为 24 万株/hm^2,行距为 40 cm＋20 cm 宽窄行,按当地高产栽培模式管理。选择晴朗无风沙天气,分别于棉花的蕾期(6 月 22 日)、初花期(7 月 3 日和 7 月 9 日)、盛花期(7 月 15 日)、初铃期(7 月 30 日)、盛铃期(8 月 9 日和 8 月 24 日)和吐絮期(9 月 10 日)进行数据采集,每次测定时间选择在当天 12:00—15:00(北京时间)。本试验区选取长势不同的 10 个采样点采样,共采集 80 个样本,试验区采集的数据包括冠层光谱和 *LAI*。

1.3　测定指标及方法

1.3.1　冠层测量　采用美国赛默飞世尔公司生产的 Antaris II FT－NIR 型光谱仪采集棉花冠层光谱,测量范围为 4 000～10 000 cm,扫描次数设置为 32 次,分辨率设置为 8 cm,采样点数为 1 557 点,使用的检测器为 InGaAs。在每个采样点采集 5 株棉花,立即摘叶,将叶片装入牛皮纸袋,标号封口,带回实验室进行近红外光谱测量。将近红外光谱仪开机预热约 30 min,用近红外光谱仪对棉花叶片进行扫描,获取近红外光谱图像,使用 EVNI 软件处理得到不同采样点棉花叶片的光谱反射值[8]。

1.3.2　棉花叶面积指数的获取　棉花冠层 LAI 的测量与光谱采集同步进行。每次测完冠层反射率光谱,将其不重叠地铺放在画有坐标网格的白色背景的纸上,用 500 万像素的数码相机拍照,相机取景以刚好框住所有叶片为宜,要求叶片上光线均匀,无阴影,同一采样点叶片重复拍摄 3 次,记录照片编号与采样点号。使用 LA－S 植物图像分析软件得到图片上的叶片面积,最后汇总计算得出不同采样点的棉花总叶面积[9]。*LAI* 的计算方法如下:

$$LAI = \text{单位面积棉花保苗数(株/m}^2) \times \frac{\text{样点叶面积(m}^2)}{\text{样点采样数(株)}}。$$

1.4　校正集和验证集样本选择

为了减小过拟合现象,使模型的预测能力增强,选择的校正集样本要具有较好的代表性。SPXY 方法是由 Galvão 等在 KS 法的基础上提出的[10],试验证明能够有效地用于 NIR 定量模型的建立。SPXY 在计算样品间距离时,将 x 变量和 y 变量同时考虑在内,标准化的 xy 的距离公式[11]为

$$d_{xy}(p,q) = \frac{d_x(p,q)}{\max[d_x(p,q)]} + \frac{d_x(p,q)}{\max[d_y(p,q)]};p,q \in [1,N]。$$

式中:$d_x(p,q)$为以棉花冠层光谱为参数计算出的样本间的距离;$d_y(p,q)$为以棉花 *LAI* 为参数计算出的样本间的距离。

采用 SPXY 方法将 80 个样本划分为 60 个校正集和 20 个验证集,分别用以建立 LAI 预测模型和验证所建模型的准确性。

1.5　SPA 提取有效波长

连续投影算法(successive projections algorithm,简称 SPA)最早由 Bregman 于 1965 年提出[12],是一种使矢量空间共线性最小化的前向变量选择算法,本研究用于剔除光谱冗余信息。设光谱矩阵为 $X^{n\times p}$,其中 n 为样本容量,p 为全谱波长数,要选出 m 个最优波长,选择步骤[13]如下:

步骤 1:第 1 次迭代之前($n=1$),将训练集光谱矩阵 $\boldsymbol{X}$ 的第 k 列赋值给 $x_{k(1)}$,$k \in (1,2,\cdots,p)$;

步骤 2:令 S 为所有未被选入的波长变量的集合,$S=\{k, 1\leqslant k\leqslant p, p \notin [k(1),k(2),\cdots,k(n)]\}$;

步骤 3:计算剩余列向量 x_k 与当前所选向量的投影;

步骤 4:记下投影值范数最大的波长的位置 $k(n+1)=\arg[\max(\|x_{k(n+1)}\|)]$,$n \in S$;

步骤 5:令 $n=n+1$,若 $n<p$,返回第 2 步进行下一波长矢量的选择;

步骤 6:分别使用各子集中的变量建立多元线性回归(简称 *MLR*)模型,选出均方根误差(简称 *RMSE*)最小的子集,然后进行逐步回归建模,在尽量不损失预测准确度的前提下,得到 1 个变量数较少的集合。该集合中的波长变量即为所选有效波长。

1.6　模型评价

校正模型性能评价参数[14]:相关系数(r)、建模均方根误差(简称 *RMSPCV*)和预测均方根误差(简称 *RMSEP*)。一个好的模型通常具有高的 r 值,低的 *RMSPCV* 和 *RMSEP*。计算公式如下:

$$r = \sqrt{1 - \frac{\sum_{i=1}^{m}(z_i - y_i)^2}{\sum_{i=1}^{m}(z_i - y_{i,av})^2}}; \quad (1)$$

$$RMSPCV = \sqrt{\frac{\sum_{i=1}^{m}(z_i - y_i)^2}{m-1}}; \quad (2)$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{m}(z_i - y_i)^2}{n}}。 \quad (3)$$

式中:m 为校正集的总样品数;n 为验证集的总样品数;z_i 为第 i 样品的测量值;y_i 为预测模型第 i 样品的预测值;$y_{i,av}$为预测模型所有样品的平均值。r 越接近 1,回归(或预测)结果越好;*RMSPCV* 越小,说明该模型的预测能力越高;*RMSEP* 越小,则表示模型对外部样品的预测能力越高;同一批次样本,*RMSPCV* 和 *RMSEP* 越小,说明模型的精度越高,两者值越接近说明模型稳定性越好。

2　结果与分析

2.1　光谱反射率特征

棉花各生育时期的冠层光谱如图 1 所示,在近红外波段,光谱反射率主要是受细胞结构的影响,在 910 nm 处反射率急剧升高,在 940 nm 波段附近出现反射率的峰值,直到

1 300 nm 的近红外反射率都维持较高水平，在 950 ~ 1 300 nm 处反射率、透射、吸收稳定，超过 1 300 nm，随波长的增加，吸收增加，反射减小，在 1 450 nm 处呈现吸收波谷，短波红外光谱区（1 300 ~ 2 600 nm）主要受叶片水分的影响，反射率升高。

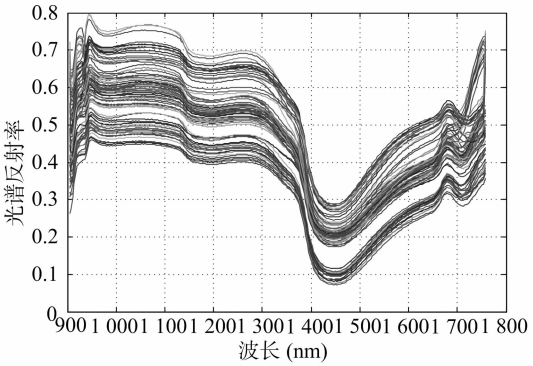


图1　棉花冠层反射率光谱

2.2　各生育时期 *LAI* 的变化

棉花 *LAI* 统计特征如下：样本数 80 个，*LAI* 平均值 2.99，*LAI* 中位数 4.08，*LAI* 标准差 0.79，*LAI* 最小值 1.64，*LAI* 最大值 4.29。图 2 为不同品种的棉花叶面积指数 *LAI* 在整个生育期内的变化曲线，苗期由于棉花未封垄，棉花冠层光谱受到土壤背景光谱的影响较大，所以叶面积的测定从蕾期开始。从蕾期到初花期，由于棉花枝叶数量的急剧增加，叶片面积不断增长，致使 *LAI* 不断升高，*LAI* 升高的速率从盛蕾期到花期再到盛花期较快，各品种的棉花 *LAI* 在铃期都达到了最大值。进入盛铃期后期，棉花叶片的光合作用已开始逐渐减弱，养分不断转移输送到棉铃上，植株下部的棉叶逐渐枯黄干落，*LAI* 在吐絮后急剧减小。

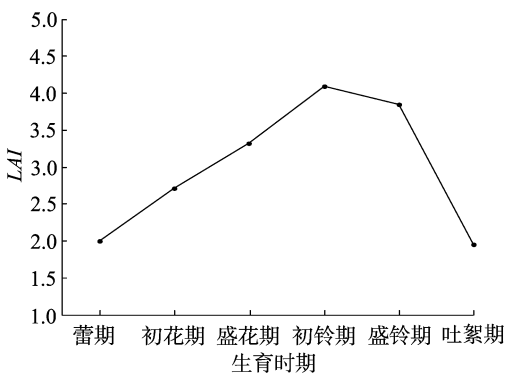


图2　不同生育时期棉花 *LAI* 的变化曲线

2.3　棉花叶面积指数建模

2.3.1　PLS 模型　原始光谱包含背景信息和除样品外的噪声信息，分别采用均值中心化、归一化、平滑去噪、一阶求导、多元散射校正（简称 MSC）5 种不同方法对光谱进行预处理。使用原光谱和预处理光谱分别对棉花 *LAI* 进行 PLS 模型建模，建模精度和预测能力如表 1 所示。可以看出，采用一阶求导光谱建立的 PLS 模型，其相关系数 *r* 最高，为 0.801 23，*RMSPCV* 和 *RMSEP* 最小，分别为 0.425 33 和 0.501 22，因此采用一阶求导预处理效果最佳。一阶导数光谱可以消除基线和其他背景干扰，分辨重叠峰，得到比原光谱更高的分辨率和更清晰的光谱轮廓变化[15]，后面在连续投影算法基础上建立 SPA - PLS 模型也以一阶求导光谱为基础进行，图 3 为经一阶求导处理后的光谱。

表 1　5 种预处理方法建立 PLS 模型结果综合比较

预处理方法	*RMSPCV*	*r*	*RMSEP*
原光谱	0.769 39	0.600 77	0.988 38
一阶求导	0.425 33	0.801 23	0.501 22
平滑去噪	1.265 05	0.619 65	1.168 96
归一化	0.443 25	0.739 15	0.653 88
数据中心化	0.520 06	0.741 73	0.646 15
MSC	0.640 23	0.719 15	0.704 38

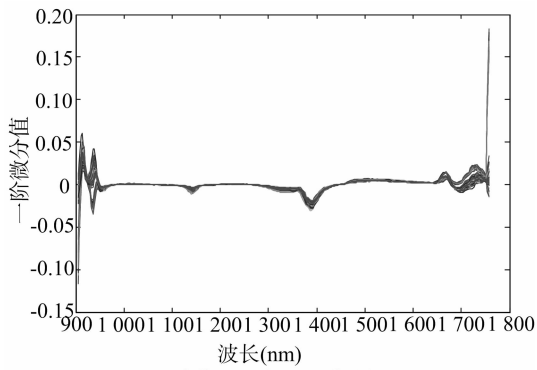
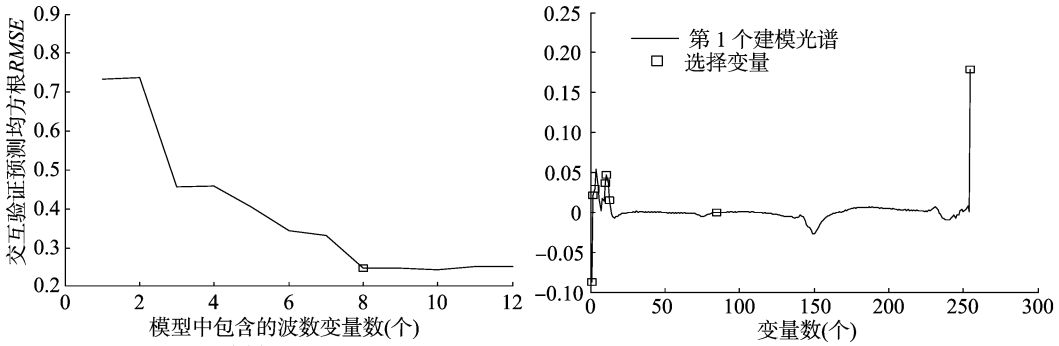


图3　棉花一阶求导光谱特征

2.3.2　SPA - PLS 模型　本研究的光谱波长区间为 900 ~ 1 700 nm，该区间含有大量冗长信息，故采用 SPA 方法在全光谱区间内提取有效波长。包含变量数的范围为 1 ~ 12 个，根据均方根误差确定选定的最终变量数，图 4 - a 为棉花近红外光谱中选用不同变量数的交互验证预测均方根 *RMSE*，当 *RMSE* 取最小值 0.248 68 时，对应的变量数是 8 个。用 SPA 算法对 ROI 区域的平均光谱进行筛选，结果如图 4 - b 所示，从全波段中提取出的 8 个特征波长分别为 904.830 02、1 754.28、936.030 03、932.900 02、911.06、1 172.61、907.950 01、942.289 98 nm。考虑到传感器边缘的光谱不能使用，故将边缘光谱剔除，留下 5 个有效特征波长（分别为 936.030 03、932.900 02、911.06、1 172.61、942.289 98 nm）作为叶面积指数的优选波长组合进行最小二乘法建模，SPA 算法选取的波长建立的 SPA - PLS 模型如图 5 所示。

3　讨论

经过 SPA 提取的波长建立的 SPA - PLS 模型与采用全谱建立的 PLS 模型结果进行对比，结果如表 2 所示，比较建模精度和预测能力可知，*RMSEP* 由 0.501 22 降低到 0.294 70，*RMSPCV* 由 0.425 33 降低到 0.294 20，*r* 由 0.801 23 提高到 0.928 27。试验结果表明，棉花的近红外光谱的谱峰重叠严重，冗余信息多，在全谱区包含大量与叶面积指数无关的信息，将全光谱的所有信息参与建模，使用 SPA 法剔除大量无用和冗余信息，从全光谱中优选出 5 个有效特征波长建立 SPA - PLS 模型，使用的变量数仅占全波段的 0.32%，然而 *RMSPCV* 和 *RMSEP* 却更小更接近，模型对外部样品的预测能力和模型稳定性也都得到了很大的提高，因此 SPA - PLS 模型的准确度和精度均优于 PLS 模型。



a. 不同变量的 *RMSE*　　b. 参加建模的有效波长分布

图4　棉花 *LAI* 预测模型最佳光谱变量总数和相应波长分布

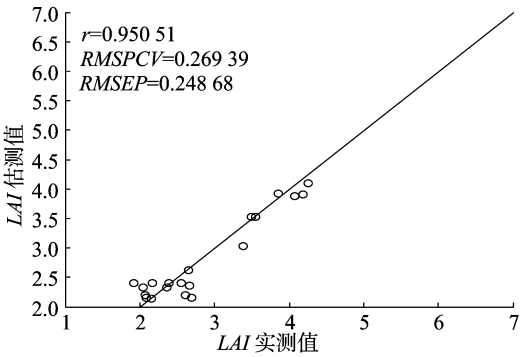


图5　SPA–PLS 模型中 *LAI* 实测值与估测值对比

4　结论

本研究运用近红外光谱仪获取棉花冠层光谱，通过一阶求导方法进行光谱预处理，分辨率和光谱轮廓比原光谱更高更清晰。采用 SPA 算法对 SPXY 法划分获得的 60 个棉花样本进行棉花 LAI 校正模型的建立及预测试验可以看出，校正样本集的选择和特征波段的选取都会影响模型的预测精度和稳定度。对全光谱使用 SPA 方法选取的有效特征波长基本上都分布在 930、1 100 nm 附近，建立的 SPA – PLS 模型效果明显好于使用全光谱建立的 PLS 模型。因此，利用 SPA 可以有效降低光谱矩阵的维数，不仅减小了参与建模的数据规模，而且降低了模型的计算量。

表 2　5 种预处理方法建立 SPA – PLS 模型结果综合比较

模型	波长(nm)	变量数(个)	*RMSPCV*	*r*	*RMSEP*
PLS	900 ~ 1 700	1 557	0.425 33	0.801 23	0.501 22
SPA – PLS	936.030 03、932.900 02、911.06、1 172.61、942.289 98	5	0.294 20	0.928 27	0.294 70

参考文献：

[1]黄乐珊，李　红，孙泽昭. 棉花产业在新疆区域经济中的地位[J]. 新疆农业科学，2006(6)：38 – 41.

[2]杨忠娜，唐继军，喻晓玲. 新疆棉花产业对国民经济的影响及对策研究[J]. 农业现代化研究，2013，34(3)：298 – 302.

[3]刘　轲，周清波，吴文斌，等. 基于多光谱与高光谱遥感数据的冬小麦叶面积指数反演比较[J]. 农业工程学报，2016，32(3)：155 – 162.

[4]谢巧云，黄文江，梁　栋，等. 最小二乘支持向量机方法对冬小麦叶面积指数反演的普适性研究[J]. 光谱学与光谱分析，2014，34(2)：489 – 493.

[5]Tang H，Brolly M，Zhao F，et al. Deriving and validating Leaf Area Index(LAI) at multiple spatial scales through lidar remote sensing：a case study in Sierra National Forest [J]. Remote Sensing of Environment，2014，143(5)：131 – 141.

[6]姚付启，蔡焕杰，王海江，等. 基于平稳小波变换的冬小麦覆盖度高光谱监测[J]. 农业机械学报，2012，43(3)：173 – 180.

[7]高洪智，卢启鹏，丁海泉，等. 基于连续投影算法的土壤总氮近红外特征波长的选取[J]. 光谱学与光谱分析，2009，29(11)：2951 – 2954.

[8]张怀志，曹卫星，周治国，等. 棉花适宜叶面积指数的动态知识模型[J]. 棉花学报，2013，03(09)：151 – 154.

[9]柏军华. 基于 LAI 的棉花产量近地遥感模型研究[D]. 石河子：石河子大学，2005：67 – 80.

[10]Kennard R W. Stone L A. computer aided design of experiments [J]. Technometrics，1969，11(1)：137 – 148.

[11]展晓日，朱向荣，史新元. SPXY 样本划分法及蒙特卡罗交叉验证结合近红外光谱用于橘叶中橙皮苷的含量测定[J]. 光谱学与光谱分析，2009，29(4)：964 – 968.

[12]Brègman L M. Finding the common point of convex sets by the method of successive projections [J]. Akademiia. Nauk SSSR Doklady，1965，162(3)：487.

[13]Galvão R H，Araújo M U，Fragoso W D，et al. Chemometrics and intelligent laboratory systems[Z]. 2008：83.

[14]刘姣娣，曹卫彬，马　蓉. 棉花叶面积指数的遥感估算模型研究[J]. 中国农业科学，2014，12(25)：4301 – 4306.

[15]陆婉珍，袁洪福，徐广通，等. 现代近红外光谱分析技术[M]. 北京：中国石化出版社，2000.