张楠楠,张　晓,施明登,等. 基于 SPA 和 PLS 的南疆绿洲区土壤盐分的近红外光谱分析[J]. 江苏农业科学,2018,46(15):218 – 221.
doi:10.15889/j.issn.1002 – 1302.2018.15.057

基于 SPA 和 PLS 的南疆绿洲区土壤盐分的近红外光谱分析

张楠楠, 张　晓, 施明登, 范泽华, 王　涛, 白铁成
(塔里木大学信息工程学院/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:应用近红外光谱技术结合连续投影算法(SPA)对南疆绿洲区土壤盐分进行分析,对 92 份土壤样品进行光谱扫描,应用不同预处理方法,以相关系数(r)、交互验证预测均方差(RMSECV)、预测标准差(SEP)、准确率(Precision)作为模型评价指标。首先建立土壤盐分预测的全波段偏最小二乘法(PLS)模型,13 种预处理方法中卷积平滑(savitzky golay smoothing,简称 SG 平滑)、傅里叶变换、SG 平滑 + 傅里叶变换的模型最好,*SEP*、*RMSECV*、r、*Precision* 分别为 0.019 876、0.024 978、0.982 686、0.965 362。同时应用 SPA 提取特征波长作为 PLS 的输入变量,建立 SPA – PLS 模型,13 种预处理方法中傅里叶变换处理方式的模型较好,*SEP*、*RMSECV*、r、*Precision* 分别为 0.016 931、0.016 769、0.987 467、0.968 915。结果表明,经连续投影算法提取 6 个特征波长建立的模型,所用变量数仅占全波段变量数的 0.38%,但 SPA – PLS 模型优于全波段的 PLS 模型。说明应用连续投影算法和 PLS 建立土壤盐分模型是可行的,并能获得满意的预测精度,可为土壤盐分预测模型研究提供一种新的思路和方法。

关键词:光谱预处理;土壤盐分;连续投影算法;PLS 建模
中图分类号: S153.6;O657.33　　**文献标志码**: A　　**文章编号**:1002 – 1302(2018)15 – 0218 – 04

在干旱与半干旱地区,土壤盐渍化是一种常见的土地退化方式[1],已发展成为国内外土壤学研究的热点[2]。土壤盐渍化是威胁农业生产和生态系统稳定的一个重要因素[3],目前已成为全球性环境问题。新疆盐碱土总面积 848 万 hm^2,现有耕地中 31.1% 的面积受到了不同程度盐碱化危害[4]。新疆南疆绿洲区为土壤盐渍化大区,盐碱土具有面积大、类型多、积盐重、形成复杂的特点[5]。

近红外光谱(NIR)技术具有快速高效监测的特点,已经在农业及其他许多领域中得到广泛应用[6-8]。近年来,许多专家学者致力于变量的选择问题[9]和采用不同的光谱预处理方法使所建立的土壤模型更稳定和精确。如代希君等借助模糊 k – 均值聚类方法、归一化处理等方法,采用偏最小二乘回归法进行全局盐分预测,预测精度有所提高[10];张娟娟等选取全谱、合频、N—H 基团等组合的 8 个波段,采用多元散射校正等多种预处理方法组合进行土壤光谱样品处理,发现 4 000 ~ 5 500 cm^{-1} 波段的模型精度最好,其决定系数达到 0.90,说明模型具有极好的预测能力[11];黄帅等把原始光谱经微分变换后的 12 种高光谱指数与土壤含盐量进行相关性分析,筛选出对土壤含盐量变化敏感的特征光谱波段,构建了

收稿日期:2017 – 02 – 26
基金项目:国家自然科学基金(编号:61362026);塔里木大学校长基金(编号:TDZKQN201506);国家自然科学基金青年科学基金(编号:61501314);塔里木大学现代农业工程重点实验室项目(编号:TDNG20150503)。
作者简介:张楠楠(1987—),女,河南洛阳人,硕士,讲师,主要从事农业遥感及作物模型研究。E – mail:893209892@qq.com。
通信作者:白铁成,硕士,副教授,主要从事干旱区作物遥感研究。E – mail:baitiecheng1983@163.com。

[4]李英年. 海北高寒草甸生态系统定位站气候概述[J]. 资源生态环境网络研究动态,1998(3):30 – 33.
[5]刘玉杰,韩建国,杨　艳,等. 施肥对草地早熟禾草坪质量、剪草量及蒸散量的影响[J]. 中国草地,2003,25(4):50 – 55.
[6]彭　琴,董云社,齐玉春. 氮输入对陆地生态系统碳循环关键过程的影响[J]. 地球科学进展,2008,23(8):874 – 883.
[7]Beard J B. Turfgrass: science and culture [M]. New Jersey: Prentice – Hall Englewood Cliffs,1972:227 – 260.
[8]Theodose T A,Bowman W D. Nutrient availability,plant abundance, and species diversity in two alpine tundra communities[J]. Ecology, 1997,78(6):1861 – 1872.
[9]沈振西,陈佐忠,周兴民,等. 高施氮量对高寒矮嵩草甸主要类群和多样性及质量的影响[J]. 草地学报,2002,10(1):7 – 17.
[10]Synder G H, Burt E O, Davidson J M. Nitrogen leaching in bermdagrass turf[J]. Turfgrass Society,1981,107:313 – 314.
[11]刘高军,韩建国,魏臻武,等. 施氮对 1 年生黑麦草人工草地中硝态氮动态及氮素分配的影响[J]. 江苏农业科学,2010(5):307 – 310.
[12]Kowalenko C G, Yu S. Solution, exchangeable and clay – fixed ammonium in south coast British Columbia soils [J]. Canadian Journal of Soil Science,1996,76(4):473 – 483.
[13]廖继佩,林先贵,曹志洪,等. 土壤固定态铵的影响因素[J]. 土壤,2003,35(1):36 – 40.
[14]朱维琴,章永松,林咸永. 土壤矿物固定态铵研究进展[J]. 土壤与环境,2000,9(4):333 – 335.
[15]Scherer H W,Weimar S. Fixation and release of ammonium by clay minerals after slurry application [J]. European Journal of Agronomy,1994,3(1):23 – 28.

基于逐步多元线性回归和偏最小二乘回归模型,得出对数二阶微分变换形式模型的稳定性和预测精度最高[12];贾生尧等提出采用递归偏最小二乘法(recursive partial least squares regression,简称 RPLS)来提高模型的预测能力,并同偏最小二乘法(PLS)、局部加权 PLS、滑动窗口 PLS 对土壤速效磷与速效钾含量进行预测,结果表明,RPLS 模型取得了最优的预测结果,决定系数分别为 0.61、0.76[13];Lin 等利用平滑与多重散射校正联合的方法对光谱进行预处理,再利用 x-y 矩阵法(sample set partitioning based on joint x-y distance,简称 SPXY)算法挑选建模集样本,利用连续投影算法和遗传算法分别进行波长优选,得出 2 种算法均可减少参与建模的波长数且能提高模型的准确度,其中遗传算法的预测精度更高[14]。

本研究在总结前人研究的基础上,以南疆绿洲区为研究区,依据近红外光谱数据、土壤含盐量实测数据,通过多种处理方法对土壤光谱进行变换处理消除光谱噪声,运用连续投影算法(successive projections algorithm,简称 SPA)[15]选出特征波长,建立偏最小二乘法(partial least square,简称 PLS)和连续投影算法-偏最小二乘法(successive projections algorithm-partial least square,简称 SPA-PLS)预测模型,并将 2 种模型进行比较,以期为土壤盐分预测模型提供一种新的研究思路和方法。

1　材料与方法

1.1　研究区概况

本研究选取新疆维吾尔自治区南疆绿洲区土壤为试验对象,该地区最高气温为 35 ℃,最低气温为 -28 ℃。研究区太阳辐射强度平均每年为 0.56~0.61 MJ/cm^2。年均日照时数为 2 556.3~2 991.8 h,日照率为 58.69%。研究区雨雪稀少,地表蒸发强烈,年均降水量为 40.1~82.5 mm,年均蒸发量为 1 876.6~2 558.9 mm。

1.2　土壤采样

在南疆绿洲区所选的典型样点进行土样采集,取表层 0~10 cm 土壤,为保证所取土样样点的代表性,确定样方面积为 3 m×3 m,采用 5 点法采样,即在每个样方的 4 角和中心各取 1 个土样,混合均匀,取 500 g 土样放入密封袋中,并做好标记。室内阴干:将采回的各土样放到塑料布上摊开,并做好标记后依次排开,将较大的土块捏碎,以利于磨细;将石子、草渣等杂物捡出,以免杂物过多,防止在称质量时产生较大误差。研磨过筛,将阴干后的各土样倒入木盘中,用擀面杖或啤酒瓶研磨,并全部通过 1 mm 筛,分成 2 份,1 份用于土壤盐分测定,另 1 份用于近红外光谱测定。共取得 92 份土壤样本。

1.3　土壤盐分含量的测定

土壤含盐量的测定参照《土壤农化分析》中的电导法[16],采用标准曲线法计算土壤全盐含量。

1.4　光谱预处理

使用美国赛默飞世尔科技公司生产的 Antaris Ⅱ FT-NIR 型光谱仪,以仪器内部空气为背景,测量范围为 4 000~10 000 cm^{-1},采样点数为 1 557 个,每张光谱扫描次数为 32 次,分辨率为 8 cm^{-1},仪器使用 InGaAs 检测器,化学计量学分析软件为仪器自带的 TQ 软件。采集光谱前开机预热 0.5h,确保光源更稳定,采集样品时重复 3 次,取平均值作为土壤样品的原始光谱(图 1)。

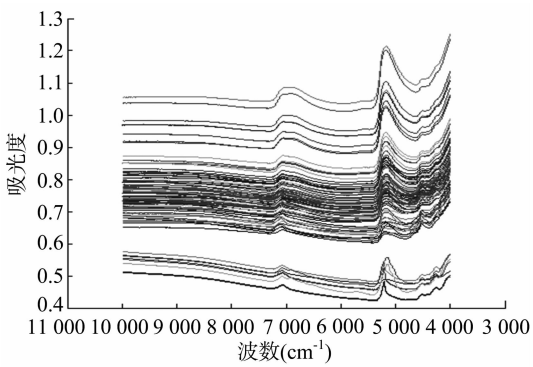


图1　土壤近红外原始光谱

应用 MATLAB 2010b 软件,采用多种处理方法对土壤光谱进行变换处理及相关分析。数据变换处理包括数据中心化(data centralized)、傅里叶变换(the fourier transform)、小波变换(wavelet transform)、归一化处理(the normalized processing)、一阶导数(savitzky golay first derivative)、二阶导数(savitzky golay second derivative)、多元散射校正(multiplicative scatter correction,简称 MSC)、卷积平滑(savitzky golay smoothing,简称 SG 平滑)。进行光谱预处理的目的在于比较分析不同光谱预处理方法对模型预测结果的影响,从而为后续提高预测模型精度打下基础。

1.5　连续投影算法

连续投影算法是一种新型变量选择方法,通过向量的投影分析,从光谱矩阵提取有效信息,并使光谱变量共线性最小[17],具体算法步骤参考文献[18]。

1.6　建模方法和模型验证指标

偏最小二乘法是一种多元数据统计分析方法,该方法是集主成分分析、普通多元线性回归和典型相关分析于一体的回归分析方法,解决了自变量多重共线性的问题[19],已经在光谱分析中得到了广泛应用。为了有效评价模型精度,本研究选取相关系数(r)、交互验证预测均方差(root mean standard error of cross validation,简称 RMSECV)、预测标准差(standard error of prediction,简称 SEP)、准确率(Precision)进行模型分析检验,其计算公式见表 1。其中,r 越接近 1,回归(或预测)结果越好;*RMSECV* 越小,说明该模型的预测能力越高;*SEP* 越小,则表示模型对外部样品的预测能力越高;对于同一批次的样本,*RMSECV* 和 *SEP* 越小,说明模型的精度越高,两者的值越接近,说明模型稳定性越好;Precision 用来验证模型的准确程度。

2　结果与分析

2.1　全局波长 PLS 模型

由表 2 可知,数据中心化和归一化处理经 SG 平滑后,各项指标均有小幅度改善;傅里叶变换、小波变换、SG 平滑、SG 平滑+傅里叶变换、SG 平滑+小波变换这 5 种处理方式的 4 项指标基本一样,是因为这 5 种处理算法都有平滑去噪的功能;一阶求导运用 SG 平滑处理后,*SEP* 变大,r 和 *Precision* 变小,*RMSECV* 有较大改善;二阶求导经 SG 平滑处理后,4 个指

表 1　偏最小二乘法模型的检验指标

验证指标	公式
相关系数(r)	$r=\sqrt{1-\frac{\sum_{i=1}^{n}(y_{实际}-y_{预测})^2}{\sum_{i=1}^{n}(y_{实际}-\bar{y}_{实际})^2}}$
交互验证预测均方差($RMSECV$)	$RMSECV=\sqrt{\frac{\sum_{i=1}^{n}(y_{实际}-y_{预测})^2}{n}}$
预测标准差(SEP)	$SEP=\sqrt{\frac{\sum_{1}^{n}(y_{实际}-y_{预测})^2}{n}}$
准确率($Precision$)	$Precision=1-\left\lvert\frac{y_{实际}-y_{预测}}{y_{预测}}\right\rvert$

注：$y_{实际}$、$\bar{y}_{实际}$、$y_{预测}$分别为实测值、实测值均值、预测值；n 为建模集(或验证集)的土壤样品数量。

标均有较明显改善。从 *SEP* 来看，傅里叶变换、小波变换、SG 平滑、SG 平滑 + 傅里叶变换、SG 平滑 + 小波变换的值较小，分别为 0.019 876、0.019 877、0.019 876、0.019 876、0.019 877；从 *RMSECV* 来看，SG 平滑 + 一阶求导的值最小，为 0.008 877，其次为傅里叶变换、小波变换、SG 平滑、SG 平滑 + 傅里叶变换、SG 平滑 + 小波变换，其值分别为 0.024 978、0.024 979、0.024 978、0.024 978、0.024 979；从 r 来看，SG 平滑、傅里叶变换、SG 平滑 + 傅里叶变换的值接近 1，为 0.982 686，其次是小波变换、SG 平滑 + 小波变换处理，为 0.982 685；但从 *Precision* 来看，傅里叶变换、小波变换、SG 平滑、SG 平滑 + 傅里叶变换、SG 平滑 + 小波变换的值最大，为 0.965 362。

综合来看，效果最好的是 SG 平滑、傅里叶变换、SG 平滑 + 傅里叶变换，由图 2 可知，模型的 *SEP*、*RMSECV* 相对较小，分别为 0.019 876、0.024 978，r 最接近 1，为 0.982 686，*Precision* 为 0.965 362。从全局波段来看，SG 平滑、傅里叶变换、SG 平滑 + 傅里叶变换适合土壤盐分含量的可见近红外光谱预处理。

表 2　PLS 模型对土壤中盐分含量的建模精度和预测能力

预处理方法	*SEP*	*RMSECV*	*r*	*Precision*
数据中心化	0.070 093	0.150 497	0.757 029	0.879 207
傅里叶变换	0.019 876	0.024 978	0.982 686	0.965 362
小波变换	0.019 877	0.024 979	0.982 685	0.965 362
SG 平滑	0.019 876	0.024 978	0.982 686	0.965 362
一阶求导	0.034 933	0.025 404	0.945 497	0.939 055
二阶求导	0.091 064	0.161 612	0.528 601	0.889 502
归一化处理	0.053 898	0.155 145	0.864 622	0.909 989
SG 平滑 + 数据中心化	0.070 087	0.150 485	0.757 075	0.879 217
SG 平滑 + 傅里叶变换	0.019 876	0.024 978	0.982 686	0.965 362
SG 平滑 + 小波变换	0.019 877	0.024 979	0.982 685	0.965 362
SG 平滑 + 一阶求导	0.045 164	0.008 877	0.907 057	0.923 056
SG 平滑 + 二阶求导	0.066 315	0.135 219	0.786 043	0.914 396
SG 平滑 + 归一化处理	0.053 801	0.154 636	0.865 147	0.910 187

2.2　SPA 选择特定波长后的 PLS 模型

2.2.1　基于 MSC + SPA 的优选波长　采用校正集 60 个样本的 1 557 个光谱变量建立的 PLS 全谱模型在建模过程中的光谱数据量很大，同时还会引入干扰变量，反而会降低模型的预测精度。在全谱范围内使用 MSC 进行光谱预处理，之后使用 SPXY 进行校正集样品划分处理，最后使用 SPA 算法进行光谱变量压缩。由图 3 可知，模型中包含的变量数为 6 时，其均方根误差(RMSE)最小，为 0.011 809。由图 4 可知，得到 6 个特征波长，波数分别为 4 393.047、4 285.053、4 971.587、3 999.64、7 293.461、5 210.717 cm^{-1}，其重要性依次减弱。

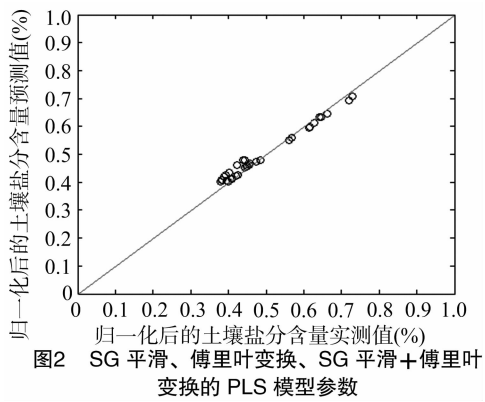


图2　SG 平滑、傅里叶变换、SG 平滑+傅里叶变换的 PLS 模型参数

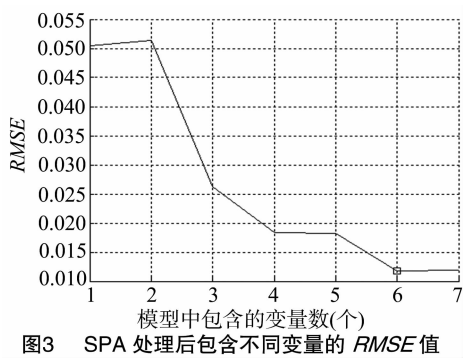
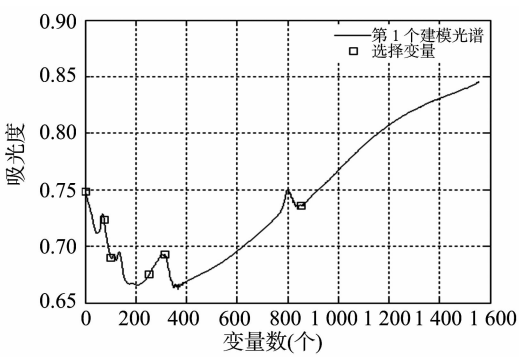


图3　SPA 处理后包含不同变量的 *RMSE* 值



图中 6 个点对应的波数从左到右依次为4 393.047、4 285.053、4 971.587、3 999.64、7 293.461、5 210.717 cm^{-1}

图4　SPA 选择的光谱特征波长

2.2.2　基于 SPA 特征波长的 PLS 模型　采用 13 种光谱预处理方法后建立 SPA - PLS 模型，由表 3 可知，数据中心化经 SG 平滑后，各项指标均变差；归一化处理经 SG 平滑后，各项指标均向好的方向变化；傅里叶变换、小波变换、SG 平滑、SG 平滑 + 傅里叶变换、SG 平滑 + 小波变换这 5 种处理方式的 4 项指标基本一样；一阶求导运用 SG 平滑处理后，4 项指标性能变差；二阶求导经 SG 平滑处理后，4 个指标性能变差，r 变化最剧烈。从 *SEP* 来看，傅里叶变换、小波变换、SG 平滑、SG 平滑 + 傅里叶变换、SG 平滑 + 小波变换的值均较小，分别为

0.016 931、0.019 144、0.017 688、0.017 688、0.019 262；从 *RMSECV* 来看，傅里叶变换的值最小，为 0.016 769，其次是 SG 平滑和 SG 平滑 + 傅里叶变换的值，均为 0.017 173；从 *r* 来看，傅里叶变换的值最接近 1，为 0.987 467；从 *Precision* 来看，傅里叶变换、小波变换、SG 平滑、SG 平滑 + 傅里叶变换、SG 平滑 + 小波变换的值较大，为 0.96 左右。

表 3　SPA – PLS 模型的预测结果

预处理方法	*SEP*	*RMSECV*	*r*	*Precision*
数据中心化	0.049 738	0.110 033	0.886 021	0.912 237
傅里叶变换	0.016 931	0.016 769	0.987 467	0.968 915
小波变换	0.019 144	0.020 123	0.983 949	0.964 959
SG 平滑	0.017 688	0.017 173	0.986 313	0.967 527
一阶求导	0.026 123	0.037 992	0.969 897	0.951 926
二阶求导	0.100 571	0.256 061	0.348 012	0.828 167
归一化处理	0.078 299	0.237 486	0.683 573	0.853 661
SG 平滑 + 数据中心化	0.066 022	0.157 805	0.788 183	0.884 479
SG 平滑 + 傅里叶变换	0.017 688	0.017 173	0.986 313	0.967 527
SG 平滑 + 小波变换	0.019 262	0.020 316	0.983 748	0.964 771
SG 平滑 + 一阶求导	0.033 315	0.076 562	0.950 556	0.941 548
SG 平滑 + 二阶求导	0.130 246	0.293 819	0.000 000	0.783 375
SG 平滑 + 归一化处理	0.071 329	0.218 335	0.746 929	0.866 047

综合来看，效果最好的是傅里叶变换，由图 5 可知，SPA – PLS 模型的 *SEP*、*RMSECV* 相对较小，分别为 0.016 931、0.016 769，*r* 最接近 1，为 0.987 467，*Precision* 为 0.968 915。从局部特征波段来看，傅里叶变换适合土壤盐分含量的可见近红外光谱预处理。

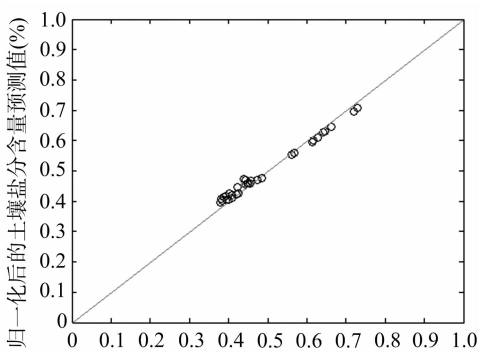


图5　傅里叶变换的 SPA–PLS 模型参数

3　结论与讨论

研究表明，利用可见近红外光谱技术、光谱预处理算法和连续投影算法检测土壤中的盐分是可行的。全波段建模过程中，经过 13 种光谱预处理后建立的 PLS 模型，效果最好的是 SG 平滑、傅里叶变换、SG 平滑 + 傅里叶变换，模型的 *SEP*、*RMSECV* 都较小，分别为 0.019 876、0.024 978，*r* 为 0.982 686，*Precision* 为 0.965 362。说明 SG 平滑、傅里叶变换、SG 平滑 + 傅里叶变换组合适合土壤盐分含量的可见近红外光谱预处理。

经 SPA 算法得到 6 个特征波长，将 6 个特征波长作为输入，经 13 种光谱预处理后建立的 PLS 模型中，效果最好的是傅里叶变换，模型的 *SEP*、*RMSECV* 相对较小，且比较接近，分别为 0.016 931、0.016 769，*r* 为 0.987 467，*Precision* 为 0.968 915。从局部特征波段来看，傅里叶变换适合土壤盐分含量的可见近红外光谱预处理。

比较全局波段和局部特征波段的模型，局部特征建模的精确度有所提高，而模型的运算量大大降低，并具有较好的稳定性。模型是否适合其他更广阔的区域有待进一步验证。

参考文献：

[1] Farifteha J, Farshada A, Georgeb R J. Assessing salt – affected soils using remote sensing, solute modeling, and geophysics [J]. Geoderma, 2006, 130(3/4): 191 – 206.

[2] 王永东，李生宇，徐新文，等. 塔里木沙漠公路防护林咸水灌溉土壤盐渍化状况研究[J]. 土壤学报，2012，49(5)：886 – 891.

[3] 王玉刚，肖笃宁，李　彦. 三工河流域中上游绿洲土壤盐化的时空动态[J]. 中国沙漠，2008，28(3)：478 – 484.

[4] 田长彦，周宏飞，刘国庆. 21 世纪新疆土壤盐渍化调控与农业持续发展研究建议[J]. 干旱区地理，2000，23(2)：177 – 181.

[5] 木合塔尔·吐尔洪，木尼热·阿布都克力木，西崎·泰，等. 新疆南部地区盐渍化土壤的分布及性质特征[J]. 环境科学与技术，2008，31(4)：22 – 26.

[6] 郑立华，李民赞，安晓飞，等. 基于近红外光谱和支持向量机的土壤参数预测[J]. 农业工程学报，2010，26(增刊 2)：81 – 87.

[7] 王　静，刘湘南，黄　方，等. 基于 ANN 技术和高光谱遥感的盐渍土盐分预测[J]. 农业工程学报，2009，25(12)：161 – 166.

[8] 石吉勇，邹小波，赵杰文，等. 近红外光谱技术快速无损诊断黄瓜植株氮、镁元素亏缺[J]. 农业工程学报，2011，27(8)：283 – 287.

[9] 袁越明，熊　伟，方勇华，等. 差分偏振 FTIR 光谱法探测水面溢油污染的模型分析[J]. 红外与激光工程，2011，40(3)：408 – 412.

[10] 代希君，彭　杰，张艳丽，等. 基于光谱分类的土壤盐分含量预测[J]. 土壤学报，2016，53(4)：909 – 918.

[11] 张娟娟，田永超，姚　霞，等. 基于近红外光谱的土壤全氮含量估算模型[J]. 农业工程学报，2012，28(12)：183 – 188.

[12] 黄　帅，丁建丽，李　相，等. 土壤盐渍化高光谱特征分析与建模[J]. 土壤通报，2016，47(5)：1042 – 1048.

[13] 贾生尧，杨祥龙，李　光，等. 近红外光谱技术结合递归偏最小二乘算法对土壤速效磷与速效钾含量测定研究[J]. 光谱学与光谱分析，2015，35(9)：2516 – 2520.

[14] Lin Z D, Wang Y B, Wang R J, et al. Improvements of Vis – NIRS model in the prediction of soil organic matter content using wavelength optimization [J]. Chinese Journal of Luminescence, 2016, 37(11): 1428 – 1435.

[15] Araújo M C U, Saldanha B T C, Galvão K H R, et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis [J]. Chemometrics and Intelligent Laboratory Systems, 2001, 57(2): 65 – 73.

[16] 鲍士旦. 土壤农化分析[M]. 3 版. 北京：中国农业出版社，2000.

[17] 陈　斌，孟祥龙，王　豪. 连续投影算法在近红外光谱校正模型优化中的应用[J]. 分析测试学报，2007，26(1)：66 – 69.

[18] 洪　涯，洪添胜，代　芬，等. 连续投影算法在砂糖橘总酸无损检测中的应用[J]. 农业工程学报，2010，26(增刊 2)：380 – 384.

[19] 陶劲松，杨亚帆，李远华. 基于 PLS 和 SVM 的纸张抗张强度建模比较[J]. 华南理工大学学报（自然科学版），2014，42(7)：132 – 137.