第 29卷第 1期
2010年 2月

红 外 与 毫 米 波 学 报
J. Infrared Millim. Waves

Vol 29, No. 1
February, 2010

文章编号: 1001 - 9014(2010)01 - 0032 - 06

用近红外光谱预测土壤碳含量的研究

沈掌泉[1], 王 珂[1], Xuewen Huang[2]
(1. 浙江大学 环境与资源学院,浙江 杭州 310029;
2. 密歇根州立大学 作物与土壤科学系,密歇根 东兰辛 48824;美国)

摘要:以田间行走式设备获取的近红外光谱数据为基础,利用最小二乘回归法(PLSR)建立了应用近红外光谱数据预测土壤碳含量的校正模型,与利用原始光谱数据建立的模型相比,应用经比值或归一化差值处理的光谱数据建立的校正模型可以提高预测精度.精度提高的原因可能是光谱数据经过波段算术组合处理后,能降低模型建立过程中产生过配的风险,使模型能包括更多的成分和信息.研究结果表明,利用偏最小二乘回归法,可以有效地建立田间近红外光谱与土壤碳含量之间的校正模型;同时,应用比值或归一化差值这些波段算术组合方法来处理近红外光谱数据,可以进一步提高模型的预测精度.因此,应用行走式设备获取的近红外光谱数据来快速测定田间土壤中碳的含量是可行的.

关 键 词:近红外光谱;土壤碳含量;行走式测定;波段算术组合;偏最小二乘回归法
中图分类号: S123; TH744.1 **文献标识码**: A

ESTIMATING THE CONTENT OF SOIL CARBON BY USING NEAR-INFRARED SPECTRA

SHEN Zhang-Quan[1], WANG Ke[1], Xuewen HUANG[2]
(1. College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310029, China;
2. Department of Crop and Soil Sciences, Michigan State University, East Lansing, MI 48824, USA)

Abstract: Partial least squares regression (PLSR) was employed to build predicting model of the content of soil carbon with on-the-go near-infrared reflectance spectroscopy (NIRS) measurements. The model based on band ratio or normalized difference of NIRS data can improve the prediction precision than the model with the original NIRS data. The reasons might be that the process of band arithmetic combination could reduce the risk of overfitting and it made the model include more useful components and information. The results show that the effective calibration model between field NIRS and the content of soil carbon can be set up by PLSR, and predicting precision can be improved while band arithmetic combination of ratio or normalized difference is performed on the NIRS data before modeling. Thus, it is feasible to estimate the content of soil carbon quickly in the field by on-the-go NIRS measurement.

Key words: near-infrared spectroscopy; soil carbon content; on-the-go measurement; band arithmetic combination; partial least squares regression (PLSR)

引言

近红外反射光谱分析(Near Infrared Reflectance Spectroscopy, NIRS)技术是近十年来发展最为迅速的高新分析技术之一,具有快速、简便的特点,已在农业及其他许多领域得到广泛应用,如在饲料成分、农产品品质分析方面已成为一种快速的例行分析方法[1~3].

近年来,随着NIRS技术应用领域的不断拓展,用于土壤成分分析和参数测定方面的研究日趋增多,并已出现了较多的报道[4];但目前的研究中,光谱数据的测定主要是在实验室条件下进行的,即使在田间测定的光谱数据,也无法快速获取;而且在田间自然状态下,对土壤光谱测定的影响因素更多、更复杂,同时由于土壤是一个复杂的混合体,因此对光谱数据的处理和分析提出了更高的要求.

收稿日期: 2009-04-16,**修回日期**: 2009-06-18 **Received date:** 2009-04-16, **revised date:** 2009-06-18
基金项目:国家科技支撑计划项目(2006BAD10A07);国家自然科学基金(40201021)
作者简介:沈掌泉(1969-),男,浙江桐乡人,副教授,博士,主要从事农业遥感、计算机应用及土壤空间变异等方面的研究, E-mail: zhqshen@zju.edu.cn.

土壤有机质由于自身特性使得它在作物养料的供给、土壤物理性质的改善、防止土壤侵蚀、实现土壤的可持续利用等方面发挥着重要的作用.而且,土壤被看作碳汇,土壤有机质的矿化、分解速度在很大程度上与全球变化有直接关系.因此,土壤中的有机质动态变化不但影响农业生态系统的可持续发展,也影响着大气圈、生物圈的可持续发展.大范围内如何有效、快速地获取土壤碳的含量与空间分布,对农业和全球变化等具有重要的价值和意义.

本研究以田间行走式近红外光谱测定设备所获取的土壤近红外光谱数据和经采样分析的土壤碳含量数据为基础,以偏最小二乘回归法作为建立校正模型的工具,分析和研究了将原始的和经过波段算术组合后的光谱数据用于建立田间土壤碳含量预测模型,并应用独立的测试数据集来检验、分析和比较预测精度上的差异.为田间行走式测定的近红外光谱数据的应用和快速获取田间土壤参数的研究提供依据.

1　材料和方法

1.1　研究区概况

研究区位于美国密歇根州 Kalamazoo县的 Carr农场,东西宽约 600m,南北长约 1000m,面积为 52公顷.土壤类型基本上为 Kalamazoo壤土,是美国北部玉米带代表性的冰渍土,由于土壤发育于冰渍物,加上试验区内的高程和坡度变化较大(高程在 290～303m之间,田面坡度在 0～14%之间),因此田块内土壤碳含量的变化相当大.根据采样分析数据,表层土壤的全碳含量在 0.551～2.637%之间,变异系数为 22.72%.该地块实行玉米/大豆轮作,并采用非灌溉的田间管理方式[5].

1.2　光谱测定与预处理

在 2004年 4月 19日,委托 Veris Technologies公司进行田间近红外光谱的测定[6].测定以南北条带的方式进行,共测定了 22个条带,条带之间的距离约为 25m,条带内点之间的距离约为 5m,共获取数据测定点约 3700个(见图 1).测定时,由拖拉机驱动的钢管插入土壤 10cm深,由固定于钢管内的钨丝灯照亮土壤,利用光纤把反射的光传输到光谱仪中进行测定和存储,测定深度保持在 7.5cm左右,测定的光谱波长范围为 920～1718nm,光谱分辨率为 6.3nm,共 128个波段,测定所获得的反射率通过倒数对数的方式转换为吸光率.在测定的同时利用 GPS获取测定点的位置信息并保存在计算机中.

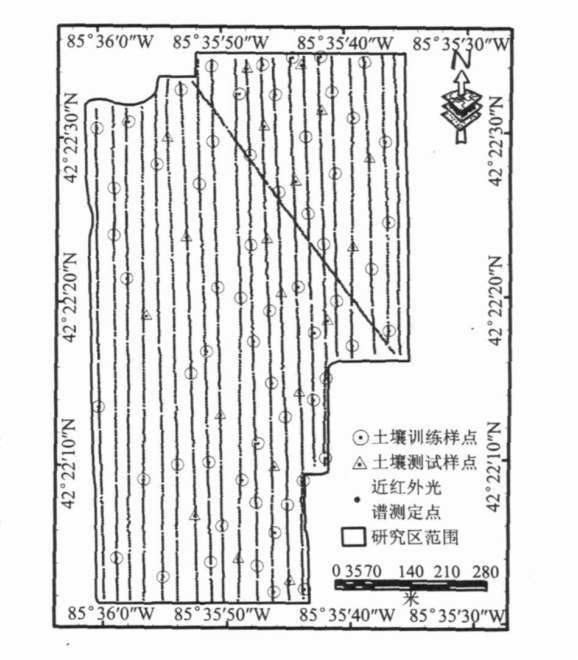

图 1　土壤训练、测试样点及近红外光谱测定点的分布图
Fig 1　Layout of soil train, test samples and near-infrared spectroscopy measurement points in study field

测定时田间土壤处于裸露和干燥状态.

以土壤采样点位置为中心,搜索 5m范围内的 NIRS测定点,如果测定点超过 1个,则取其各波段测定值的平均值作为该土壤样点红外光谱的值,反之则以该测定点的值作为此土壤采样点红外光谱的值.

1.3　土样采集与化学分析

在红外光谱数据测定后,以半随机的方式采集土壤样本 85个,也就是沿光谱测定的条带,随机地确定样本采集的位置,采样点位置用 GPS进行精确定位,采样深度为 10cm;土样在室温状态下风干后,清除掉植物残留物并过 100目筛,然后在密歇根州立大学作物与土壤科学系的实验室内通过 Carlo-Erba系列 2碳氮分析仪用干烧法测定土壤全碳含量[5].经统计分析剔除 1个异常点后,其余的 84个样点,在考虑到分布均匀的前提下,随机地分为独立的训练数据集和测试数据集,其所包含的样点数分别为 65和 19个(见图 1).

1.4　波段算术组合

近红外光谱分析技术属于弱光谱信号分析技术.近红外光谱的信息是分子内部振动的倍频与合频,包含键强度、化学组分、电负性和氢键等信息.当样品为固体时,受到散射、漫反射、反射光的偏振、样品的颗粒和尺寸等的影响,因此在发挥近红外光谱

的特点时,存在一系列分析的技术难点.它吸收强度较弱,测定不经过预处理的样品的光谱易受样品状态、测量条件等影响,尤其是在测定背景、样品成分复杂的情况下,导致光谱中谱峰重叠和不确定性较大.而且,作为信息源的近红外光谱中有效信息率低,对从复杂、重叠、变动的光谱中提取某个特定成分的微弱信息造成困难,需要应用有效的方法和技术来抑制噪声、增强有用的信息.

通常在建立红外光谱定量分析模型时,直接采用原始的或经过主成分分析、小波分析、相关分析、微分变换等处理的光谱数据来建立分析模型,而在遥感数据处理与信息提取中得到关注和广泛应用的植被指数,却并未在红外光谱数据处理中引起注意和研究.植被指数本质上是在综合考虑各有关光谱信号的基础上,把多波段的反射率作一定的数学变换,使其在突出感兴趣信息的同时,使非感兴趣的信息最小化,由于在遥感应用中最受关注的是植被的信息,因此把应用此类思想的信息增强与提取技术称为植被指数[7].其中提出来最早、应用也最广泛的植被指数是通过近红外和红波段之间的算术运算(波段间的加、减、乘或除运算)来得到的,包括差值、归一化差值、比值等.基于同样的考虑,本研究通过对不同红外光谱波段之间的吸光率进行算术运算,来达到增强有用信息和抑制干扰的目的.尽管目的是相同的,但由于与植被指数的含义存在差异,为了与之相区别,在本文中将其称为波段算术组合.

在本研究中,先将近红外光谱数据导入到 Matlab中,假设不同光谱波段的吸光率值分别为 $A_{\lambda 1}$和 $A_{\lambda 2}$,按照波段差、归一化差和波段比的计算公式:$A_{\lambda 1}-A_{\lambda 2}$、$(A_{\lambda 1}-A_{\lambda 2})/(A_{\lambda 1}+A_{\lambda 2})$和 $A_{\lambda 1}/A_{\lambda 2}$分别计算它们的值,然后进行进一步的分析和处理.

1.5 偏最小二乘回归法

偏最小二乘回归法(partial least squares regression, PLSR)是光谱多元定量校正最常用的一种方法,已被广泛应用于近红外、红外、拉曼、核磁和质谱等波谱定量模型的建立,几乎成为光谱分析中建立线性定量校正模型的通用方法[8].

偏最小二乘回归法由伍德、阿巴诺等人在 1983年提出,它是在普通多元回归的基础上揉合进主成分分析、典型相关分析的思想,很好地解决了自变量间多重共线性问题.偏最小二乘的分析原理为:偏最小二乘回归 = 主成分分析 + 典型相关分析 + 普通多元线性回归.偏最小二乘以最小二乘法为算法基础,在尽可能提取包含自变量更多信息的成分的基础上,保证了提取成分与因变量间最大的相关性,即偏爱与因变量有关的部分,所以称其为偏最小二乘回归[9].

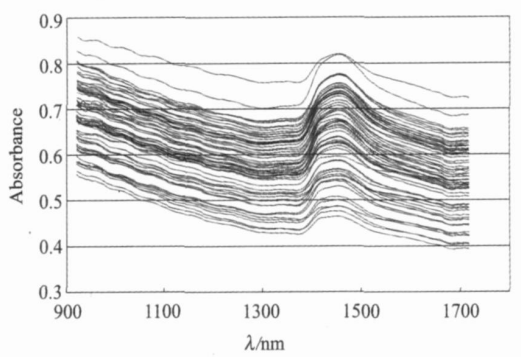

图2 训练数据集土壤近红外吸收光谱曲线

Fig. 2 The near-infrared absorbance spectra of soil samples in train dataset

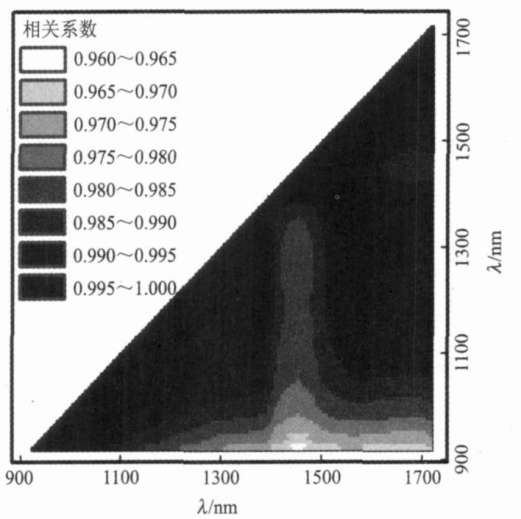

图3 训练数据集中土壤近红外光谱波段之间的相关性

Fig. 3 The correlative relationship between wavelengths of soil absorbance spectra in train dataset

偏最小二乘回归法的分析过程为:首先,应用主成分分析与典型相关分析的思想来提取成分,这不仅保证了提取的成分尽可能多地保留原始变量的信息且保持相互独立,而且使自变量与因变量间的相关性最大;然后,采用普通最小二乘法建立回归方程,由于成分间已不存在多重共线性,因此采用普通最小二乘估计所得结果稳定性较好.因此,偏最小二乘回归法集中了主成分分析、典型性相关分析及普通多元回归分析的优点.在分析过程中,如主成分分析那样,偏最小二乘回归法采用截尾的方式选择前几个重要的成分,因此需要确定模型所包含的成分的个数.一般可采用交叉验证法来确定保证模型较好的精度所需包含的成分的数量.

在本研究中,以 Nφrgaard等人开发的 iToolbox 工具箱中的 iPLS作为进行 PLSR分析和建模的工具[10],并在建模过程中,应用交叉验证的方法来确定模型需包含的成分的个数和防止模型过度拟合.

2　结果与讨论

2.1　田间土壤近红外光谱数据分析

图 2为研究区训练数据集中 65个样本的近红外吸收光谱曲线,由于土壤发育自冰渍物,田块内土壤差异较大,加上区内高程差异明显,使不同区域土壤水分条件差异也较大,导致有机质含量的差异也相当明显,因此样本之间的近红外光谱也存在明显的差异.

对训练数据集中光谱数据的 128个波段进行波段之间的相关性分析表明,不同波段之间均存在非常高的相关性 (见图 3),而这种波段之间密切的相关关系,导致一般的回归分析手段难以建立可靠的校正模型.

2.2　土壤碳含量与土壤近红外光谱之间的相关性

对土壤碳含量与近红外光谱之间的相关性进行分析表明,在测定的整个近红外光谱范围内,其相关性均不高,除 1399～1525nm范围内达到显著水平 (α=0.05)外,其余波段均未达到显著水平 (图 4).其原因可能与光谱测定直接在田间进行,田间土壤水分等变化大 (采样时土壤水分在 2.25～22.72%之间,变异系数达 24.59%),干扰因素多等有关.

而经过波段算术组合后的近红外光谱数据与土壤碳含量之间的相关性得到明显提高,从图 5可以

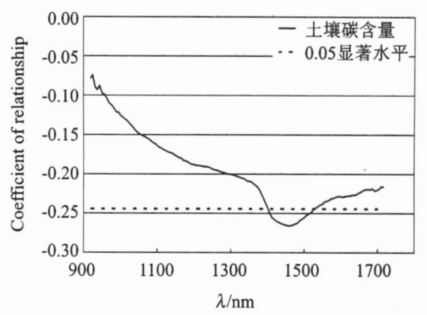

图 4　训练数据集的土壤碳含量与土壤近红外光谱之间的相关性

Fig 4　The correlative relationship between soil *C* and soil absorbance spectra in train dataset

发现,无论是差值、归一化差值还是比值,除少量变量未达到显著外,大部分均达到了极显著 (α=0.01)的相关性,相关系数最高的甚至接近 0.7,说明经过波段算术组合处理后,干扰信息被有效抑制,而与土壤碳含量有关的光谱信息得到了明显的增强,为应用近红外光谱来提取土壤碳含量提供了更好的基础.

2.3　偏最小二乘回归法的建模与预测结果分析

对原始的和经波段算术组合后的光谱数据,分

表 1　不同光谱数据处理方式下偏最小二乘回归法的结果

Table 1　Summary of results derived by PLSR from absorbance spectra with different processing methods

数据处理方法	RMSE(%) 交叉验证	训练数据集	测试数据集	包含的成分数
原始光谱数据	0.249	0.218	0.213	4
波段差	0.257	0.242	0.245	3
波段比	0.264	0.192	0.200	7
归一化差	0.267	0.197	0.201	7

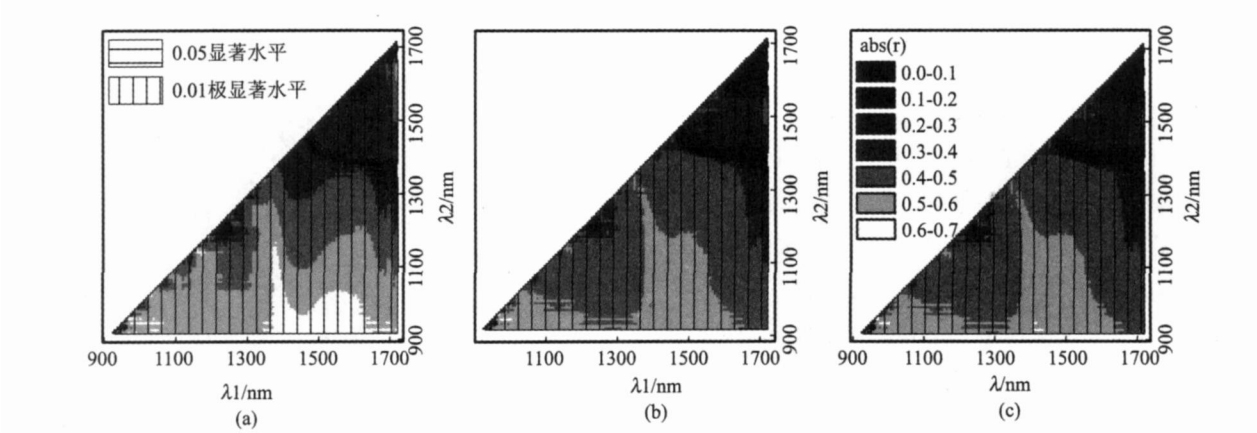

图 5　训练数据集中土壤样本的碳含量与经波段算术组合后近红外光谱数据之间的相关系数的绝对值　(a)是波段差　(b)是归一化差　(c)是波段比

Fig 5　The absolute correlativity between soil C and band arithmetic combinations of absorbance spectra in train dataset
(a) difference of bands　(b) normalized difference of bands　(c) ratio of bands

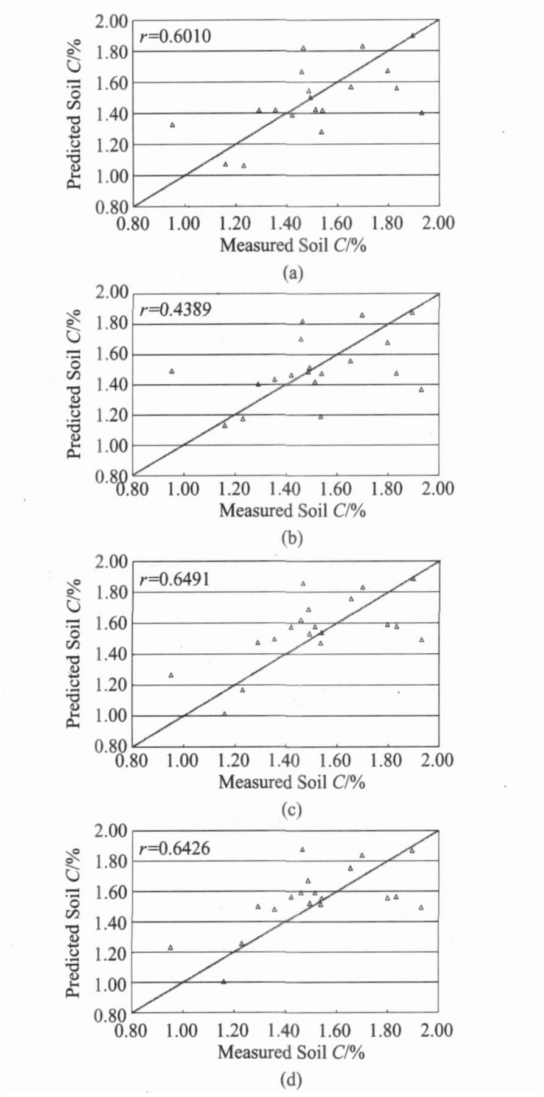

图 6 测试数据集中的测定值与预测值比较的散点图 (a)为原始光谱数据 (b)为波段差 (c)为波段比 (d)为归一化差

Fig 6 Plots of measured versus predicted soil *C* in test dataset by PLSR (a) original spectra (b) difference of bands (c) ratio of bands (d) normalized difference of bands

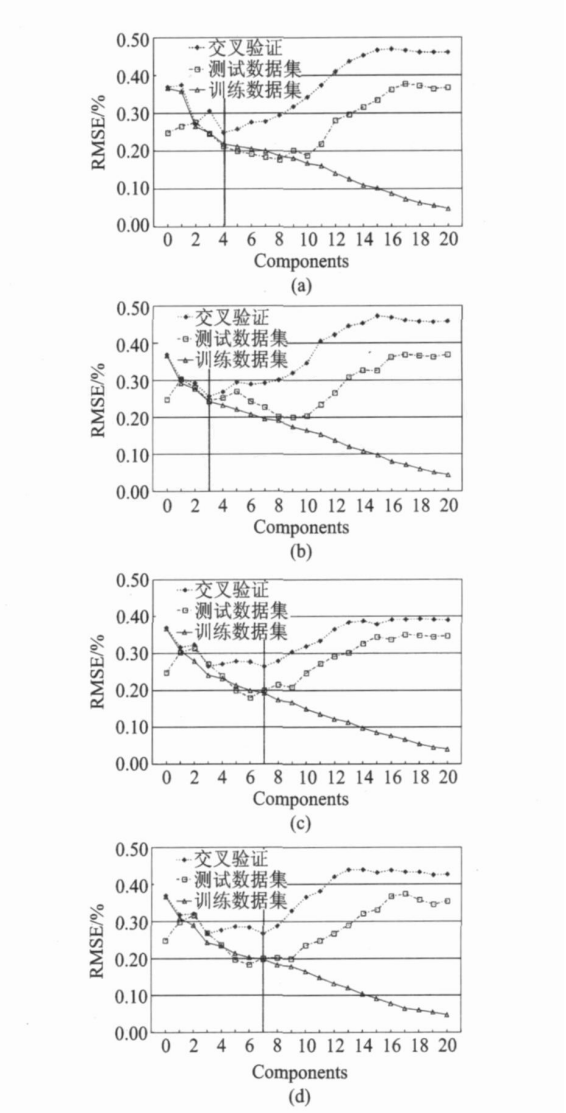

图 7 PLSR建模过程中 RMSE的变化及模型的选择 (a)为原始光谱数据 (b)为波段差 (c)为波段比 (d)为归一化差

Fig 7 Plots of RMSE changing in processing of PLSR and model selection (a) original spectra (b) difference of bands (c) ratio of bands (d) normalized difference of bands

别应用偏最小二乘回归法建立了校正模型,并对测试数据集进行预测.表 1的结果表明,与应用原始光谱数据所建立的校正模型相比,经波段比值和归一化差值处理后,尽管交叉验证的误差 RMSE (Root Mean Square Error)略有升高,但训练数据集和测试数据集的预测误差均有明显的降低,而差值处理的预测误差却反而提高了;从模型所包括的成分个数来看,也存在差异;从测试数据集中各样本测试值与预测值的散点图 (图 6)也说明了同样的情况,经比值和归一化差值处理的相关系数较高,其次为原始光谱的,而差值处理的最低.

在 PLSR回归模型建立过程中,随着模型包括的成分个数的增加,交叉验证、训练数据集和测试数据集的预测误差的变化 (图 7),可以发现,尽管训练数据集的预测误差在各模型建立过程中的表现比较相似,但交叉验证的误差的变化却有明显的差异,与原始光谱数据相比,经过比值和归一化差值处理的光谱数据,在建立 PLSR校正模型时,似乎能接纳更多的成分而不容易导致过配,而经差值处理的光谱数据却比原始光谱数据所能包括的成分数量还要少.

3 结论

在田间应用行走式设备测定的近红外光谱数

据,由于在田间条件下影响因素多,因此其与土壤碳含量之间的相关性较低,除小部分波长的相关性达到显著外,大部分无法达到显著,而且光谱波段之间的相关性非常高,全部在 0.96以上,给应用常规回归方法建立校正模型带来了困难.利用波段算术组合方式处理光谱数据后,其与土壤碳含量之间的相关性得到了明显的提高,除少量变量的相关性较低外,大部分经组合后的波长均达到了极显著水平,说明通过波段算术组合的处理方式可以抑制干扰、增强与土壤碳含量有关的信息.通过对原始的和经波段算术组合处理的光谱数据分别应用偏最小二乘回归法建立校正模型后,从独立的测试样本集的预测精度的比较可以发现,与原始光谱数据建立的模型相比,经过比值和归一化差值处理的光谱数据所建立的模型的预测精度有一定的改善,而差值处理的光谱数据所建立的模型的预测精度反而有所降低.究其原因,经比值和归一化差值处理后,似乎能延缓模型过配的风险而使模型能够包括更多的成分数,而差值处理却无法达到这样的效果.

本研究的结果表明,利用偏最小二乘回归法,可以有效地处理通过行走式设备获取的田间近红外光谱数据,并建立校正模型来测定田间土壤碳的含量.与直接利用原始光谱数据所建立的校正模型相比,应用比值和归一化差值这两种波段算术组合方法来处理光谱数据,可以进一步提高校正模型的预测精度.因此,利用行走式设备所测定的红外光谱数据来快速地获取田间土壤碳含量的空间分布是可行的.

REFERENCES

[1] BAO Yi-Dan, HE Yong, FANG Hui, *et al*. Spectral characterization and N content prediction of soil with different particle size and moisture content[J]. *Spectroscopy and Spectral Analysis* (鲍一丹,何勇,方慧,等.土壤的光谱特征及氮含量的预测研究.**光谱学与光谱分析**), 2007, **27**(1): 62—65.

[2] YI Qiu-Xiang, HUANG Jing-Feng, WANG Xiu-Zhen. Hyperspectral estimation models for crude fiber concentration of corn[J]. *J. Infrared Millim. Waves* (易秋香,黄敬峰,王秀珍.玉米粗纤维含量高光谱估算模型研究.**红外与毫米波学报**), 2007, **26**(5): 393—395.

[3] LIU Huan-Jun, ZHANG Bai, WANG Zong-Ming, *et al*. Soil saline-alkalization evaluation basing on spectral reflectance characteristics[J]. *J. Infrared Millim. Waves* (刘焕军,张柏,王宗明,等.基于反射光谱特征的土壤盐碱化评价.**红外与毫米波学报**), 2008, **27**(2): 138—142.

[4] COZZOLINO D, MORÓN A. The potential of near-infrared reflectance spectroscopy to analyse soil chemical and physical characteristics[J]. *Journal of Agricultural Science*, 2003, **140**: 65—71.

[5] HUANG X W, SENTHILKUMAR S, KRAVCHENKO A, *et al*. Total carbon mapping in glacial till soils using near-infrared spectroscopy, Landsat imagery and topographical information[J]. *Geoderma*, 2007, **141**: 34—42.

[6] CHRISTY C D, DRUMMOND P, LAIRD D A. An on-the-go spectral reflectance sensor for soil[C]. *American Society of Agricultural Engineers Meeting*, 2003, paper number: 031044.

[7] TIAN Qing-Jiu, MIN Xiang-Jun. Advances in study on vegetation indices[J]. *Advance in Earth Sciences* (田庆久,闵祥军.植被指数研究进展.**地球科学进展**), 1998, **13**(4): 327—333.

[8] FRANK I E, KALVAS J H, KOWALSKI B R. Partial least squares solutions for multicomponent analysis[J]. *Anal Chem.*, 1983, **55**(11): 1800—1804.

[9] WU Qiong, YUAN Zhong-Hu, WANG Xiao-Ning. Summary of Partial Least Squares Regression[J]. *Journal of Shenyang University* (吴琼,原忠虎,王晓宁.基于偏最小二乘回归分析综述.**沈阳大学学报**), 2007, **19**(2): 33—35.

[10] NØRGAARD L, SAUDLAND A, WAGNER J, *et al*. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy[J]. *Applied Spectroscopy*, 2000, **54**: 413—419.