

基于田间原位土壤含水量估测的可见/近红外光谱建模方法

刘广霖¹ 郭焱¹ 劳彩莲^{2*} 徐兴阳³ 秦春丽³

(1. 中国农业大学 资源与环境学院,北京 100193;

2. 中国农业大学 信息与电气工程学院,北京 100083;

3. 云南省烟草公司昆明市公司,昆明 650051)

摘要 为实时、准确地获取原位土壤含水量信息,利用可见/近红外光谱技术,分别使用全局偏最小二乘(PLS)建模、局部 PLS 建模方法,对田间原位土壤含水量进行快速估测。结果表明:全局 PLS 模型中,其建模集的决定系数(R^2)、交叉验证均方根误差(RMSECV)分别为 0.943 和 1.750%,检验集的决定系数(R^2)、预测均方根误差(RMSEP)分别为 0.956 和 1.260%。局部 PLS 模型中,分别比较了选取定标子集的 2 种方法(欧氏距离法和马氏距离法),采用欧氏距离法和马氏距离法选取定标子集进行建模的 R^2 值分别为 0.974 和 0.979, RMSEP 值分别为 0.976% 和 0.943%。因此,将可见/近红外光谱技术应用到田间原位含水量测量是可行的,其中,使用局部建模方法的效果优于全局建模。

关键词 土壤含水量;可见/近红外;偏最小二乘;田间原位;全局建模;局部建模

中图分类号 S 152.7+1; O 657.3

文章编号 1007-4333(2016)08-0125-07

文献标志码 A

Estimation of soil water content *in situ* by using visible/near infrared spectrum modeling

LIU Guang-lin¹, GUO Yan¹, LAO Cai-lian^{2*}, XU Xing-yang³, QIN Chun-li³

(1. College of Resources and Environmental Sciences, China Agricultural University, Beijing 100193, China;

2. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China;

3. Yunnan Tobacco Company Kunming Branch, Kunming 650051, China)

Abstract In order to collect the data of soil water content *in situ* real-timely and accurately, this study used visible/near infrared (Vis/NIR) spectroscopy technology to estimate soil water content *in situ* rapidly, by global partial least squares (PLS) modeling and local PLS modeling. The results showed that, For global PLS modeling, the decision coefficient (R^2) and root mean square error of cross validation (RMSECV) of the modeling set were 0.943 and 1.750%, respectively. The decision coefficient (R^2) and root mean square error of prediction (RMSEP) set were 0.956 and 1.260%, respectively. For local PLS modeling, two methods (i. e., Euclidean distance method and Mahalanobis distance method) for subset selection were used. The R^2 of the two methods were 0.974 and 0.979, respectively. RMSEP were 0.976% and 0.943% respectively. This study suggests that it is feasible to measure soil water content *in situ* by using the Vis/NIR spectroscopy technology. And the result of local modeling was better than that of global modeling.

Keywords soil water content; Vis/NIR; partial least square; *in situ*; global modeling; local modeling

土壤含水量的快速测定一直是农业信息获取的难题^[1]。已建立的各种土壤含水量测量方法或者费时费力,或者在测量时需要考虑诸多因素,如土壤介

电特性、土壤电导率、电磁波等,给土壤含水量实时的测量带来不便。可见/近红外光谱技术是一种高效的现代分析技术,以实时、快速、精确、低成本、不

收稿日期:2015-06-12

基金项目:中国烟草总公司云南省公司资助项目(2013YN17)

第一作者:刘广霖,硕士研究生,E-mail:liu_liuwangwang@foxmail.com

通讯作者:劳彩莲,副教授,主要从事光谱分析技术研究,E-mail:wxjylh@126.com

消耗化学试剂、不污染环境以及易于实现原位和原位分析等诸多优点,已在农业、烟草、食品、化工、能源等行业中得到了广泛应用^[2]。其主要原理是通过含氢基团的倍频和合频进行吸收,由此得到样品中含氢基团的特征信息。土壤中水分含量越多,吸收就越强烈,通过光谱吸收或反射的强烈程度反映土壤中水分含量情况。其中,1 450、1 940、2 200 nm为其吸收峰,吸收或反射最强烈,也最敏感。因此,利用可见/近红外光谱的上述特性,将此技术引入到田间原位土壤水分测量上,实时、准确地获取土壤含水量信息,具有重要的科学意义和应用价值。

土壤由不同粒径的矿物颗粒、有机质、生物活体和水分等组成,是一个复杂而多相的系统。鉴于土壤的复杂性,目前对土壤水分的光谱研究基本停留在室内研究上。Bowers等^[3]发现在400~2 500 nm波段,主要吸收峰有1 450、1 940、2 200 nm,其中以1 450、1 940 nm吸收最为强烈。鉴于此,部分研究对此3个波长点下的吸光度建立多元线性回归(MLR)模型,均得到较为理想的效果^[4-5]。由于近红外光谱所包含的信息众多,容易造成信息上的重叠,为了消除多重共线性,很多研究把主成分分析(PCA)、全局偏最小二乘(PLS)等建模方法引入到土壤含水量研究中,并取得理想结果^[6-9]。随着近红外光谱技术快速发展,建模方法也日趋完善。为了使模型更加精确、稳定、可靠,部分研究提出了局部建模的思想^[10-11]。这种方法对模型的稳定性、预测性均有所提高,但伴随而来的是计算量的显著增大。

目前,土壤参数的光谱研究多采用野外布点取样,土样带回室内进行风干、过筛等处理后测量。土壤经过相应处理,消除了土壤质地、土壤结构、土壤紧实度等因素对土壤反射光谱的影响,因此,所得到的效果均较为理想。而田间原位土壤光谱测量将不可避免上述因素,因此,将可见/近红外光谱应用于土壤参数(尤其是土壤含水量)的田间原位测量上具有重要的应用和科研价值。目前,将可见/近红外光谱技术应用在田间原位土壤参数上的研究已开展,但多数模型在稳定性、预测精度上均不高^[8,12]。本研究对土壤进行田间原位光谱测量,旨在建立光谱含水量的数学模型,比较不同建模方法的优劣,探讨可见/近红外光谱技术在田间原位土壤含水量估测上的可行性。

1 材料与方法

大田试验于2014年在云南省寻甸县烟草科

技推广站(25°52'N,103°27'E)进行。试验地土壤类型为水稻土,质地为粘质壤土。种植作物为烟草,品种为K326。试验共设置A、B、C3个处理:A处理为自然降雨处理,不设置遮雨棚;B处理前期不进行遮雨棚设置,7月10日后设置遮雨棚,晴天时打开,晚上、阴雨天则关闭;C处理为极度干旱处理,遮雨棚从前期到烟草采收均未打开。各处理小区数分别为9、6、1个,随机区组设计。每个小区面积为36 m²,种植3垄烟株,株行距为0.5 m×1.2 m。

从烟草的团棵期开始进行田间原位土壤光谱测量并采集土壤样品,直至烟草采收完毕。测量周期定为每5天进行1次,但会根据天气变化做相应调整。每次测量均在1天内完成。试验从6月7日—8月17日,进行了14次测量,共322个样品。其中6月7日所做的为预试验。

1.1 主要仪器及分析软件

所使用的光谱仪器为美国ASD公司的FieldSpec3便携式分光辐射光谱仪。该仪器的波长范围为350~2 500 nm,其350~1 050 nm采样间隔为1.4 nm;1 050~2 500 nm间隔为2 nm。所使用软件有ViewSpecPro5.7.2光谱分析软件,Matlab R2012a数据分析软件,Excel 2013。

1.2 土壤光谱采集和含水量测定

在进行田间原位土壤光谱采集时,每个处理分别选取1个小区作为取样小区,进行随机取样,每个处理每次取5~10个样品不等,根据需要进行调整。使用小铁铲挖去5~15 cm垄上土壤,除去大石子、大根、枯叶等,用FieldSpec3高强度接触式探头直接接触土壤进行光谱反射率测定,每次测量前用标准白板对仪器进行校正。每个土样重复测定3次,每次采集2条光谱,共6条光谱。使用ViewSpecPro5.7.2软件进行有效性检查,剔除有异常的光谱,取平均值后将数据导出。

土壤光谱采集完毕后,立即取此点下的土壤约120 g,装袋密封,带回实验室采用烘干法进行土壤含水量测量^[13]。

1.3 土壤光谱建模方法及实现

偏最小二乘法是近年来广泛被应用的新型多元统计分析方法。它在建模过程中集中了主成分分析和线性回归分析的特点。在分析结果中,能够提供提供一个更加合理的模型。因此,被广泛应用在可见/近红外光谱分析中^[14]。本研究分别采用全局、

局部 PLS 方法进行建模。全局建模的思路是将样本划分为建模集和检验集,用建模集的样本进行建模,检验集样本进行预测,采用 10 折交叉验证,以决定系数、交叉均方根误差和预测均方根误差值为标准评价模型优劣。

单样本局部建模,其建模过程和全局建模大致相同,不同之处是全局建模只建立一个 PLS 模型,而局部模型则是对每一个检验样本均建立相应的 PLS 模型。其基本思想是:对建模集光谱数据进行主成分分析,采用前 2 个主成分构建主成分特征空间,计算检验集在该主成分空间的投影;计算检验集中的每个样本与建模集全部样本在主成分空间的欧式距离或马氏距离;通过限定距离,对检验集中的其中一个样本,选取局部建模的定标子集;建立此定标子集下的 PLS 模型,并对该样本进行预测;依次类推,完成检验集中每一个样本的预测;选出不同限定值下 RMSEP 值最小的局部 PLS 模型^[10,15]。欧氏距离、马氏距离的数学定义为:

$$D_e(x_i, x_j) = \left[\sum_{k=1}^n (x_{ik} - x_{jk})^2 \right]^{1/2} \quad (1)$$

$$D_m(x_i, x_j) = \left[(x_i - x_j)^T S^{-1} (x_i - x_j) \right]^{1/2} \quad (2)$$

式中: D_e 、 D_m 分别为欧氏距离、马氏距离; x_i 、 x_j 分别为第 i 个建模集样本和第 j 个检验集样本在主成分空间的投影得分; S 为建模集样本协方差; n 为主

成分空间维数,即主成分数,本研究 $n=2$ 。

由于外界及光谱仪器自身扰动影响,所获得的土壤光谱在 350~450 及 2 400~2 500 nm 波段处扰动较大,故将其去除,取 450~2 400 nm 作为建模的全波段。对经过吸光度预处理的相关波段(450~2 400、1 200~1 600 nm、1 600~2 000、2 000~2 300、1 200~2 300 nm)和土壤含水量间分别建立全局 PLS 模型。具体如下:使用 SPXY 法将样本划分为建模集(262 个)和检验集(60 个),分别建立全局 PLS 模型,使用 10 折交叉验证,以 R^2 、RMSECV 和 RMSEP 值为标准,选出最优的建模波段。

研究表明,土壤含水量与光谱反射率倒数的对数(即吸光度)相关^[4,6]。建立回归模型之前,本研究对土壤光谱进行吸光度预处理;使用最优波段进行建模;利用 2sigma 原则进行异常样本剔除(剔除异常样本 23 个);使用 SPXY 法划分检验集(239 个)和检验集(60 个);分别建立相应的全局、局部 PLS 模型。

2 结果及分析

2.1 土壤含水量

图 1 为测得的土样含水量频数分布直方图:322 个土壤样品含水量变幅较大,最小值 3.67%,最大值 40.45%,均值 21.97%。

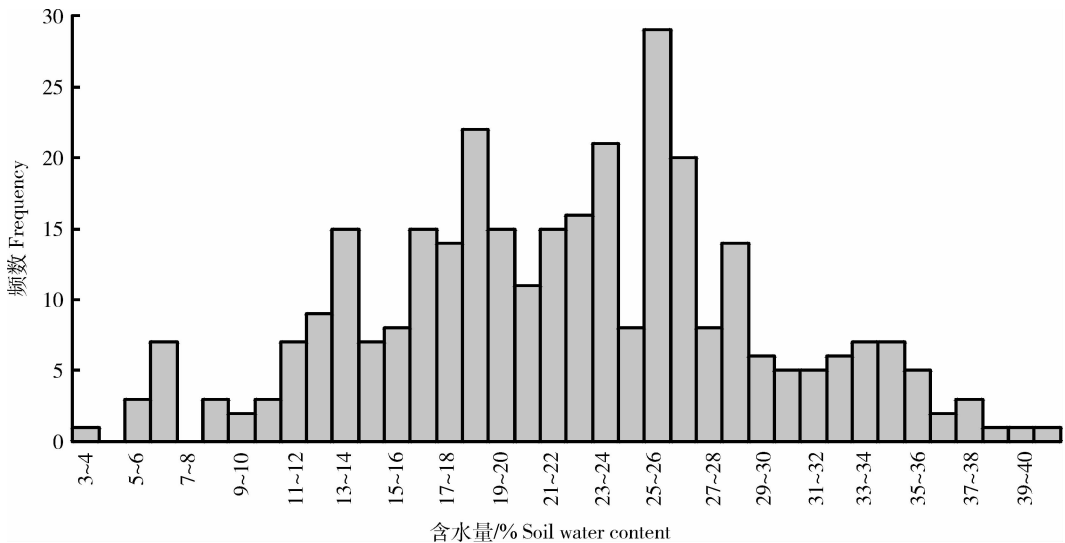


图 1 土样含水量分布直方图

Fig. 1 Soil water content distribution histogram

2.2 光谱预处理和特征波段选择

对土壤原始光谱反射率进行吸光度预处理

(图 2),选取各波段(450~2 400、1 200~1 600、1 600~2 000、2 000~2 300、1 200~2 300 nm)作为

建模波段,分别建立全局 PLS 模型,比较模型估测精度,以 1 200~2 300 nm 波段吸光度建模效果

最好(表 1),故以下分析均采用此波段作为建模波段。

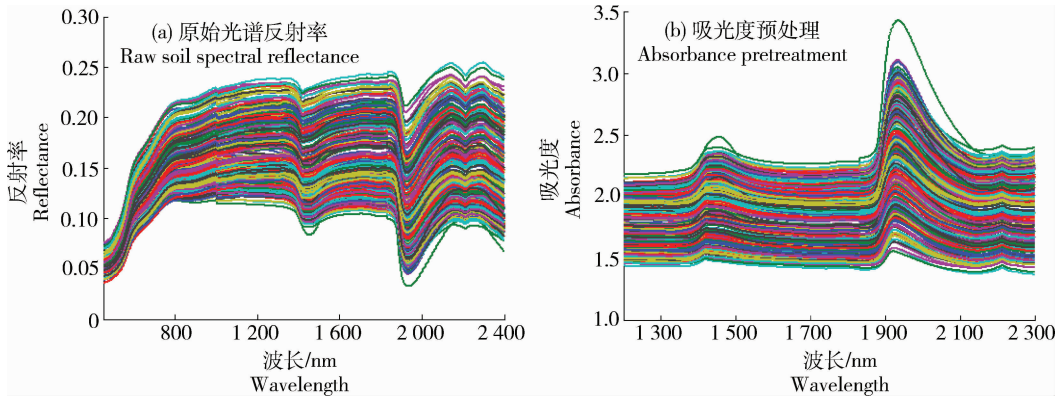


图 2 原始光谱反射率和吸光度预处理光谱图

Fig. 2 Raw soil spectral reflectance and absorbance pretreatment

表 1 不同波段建模精度

Table 1 The estimation accuracy of models in different wave bands

波段/nm Spectral band	建模集 Modeling set		检验集 Test set	
	R^2	RMSECV/%	R^2	RMSEP/%
450~2 400	0.848	2.945	0.936	1.624
1 200~2 300	0.860	2.833	0.945	1.427
1 200~1 600	0.851	2.930	0.930	1.650
1 600~2 000	0.850	2.952	0.925	1.586
2 000~2 300	0.844	2.957	0.928	1.791

注: R^2 为决定系数;RMSECV 为交叉均方根误差;RMSEP 为预测均方根误差。

Note: R^2 is decision coefficient; RMSECV is cross validation of root mean square error
RMSEP is prediction of root mean square error.

2.3 全局建模方法

使用全局 PLS 进行建模,得到建模集样本估

测拟合结果(图 3(a))和检验集样本估测拟合图(图 3(b))。

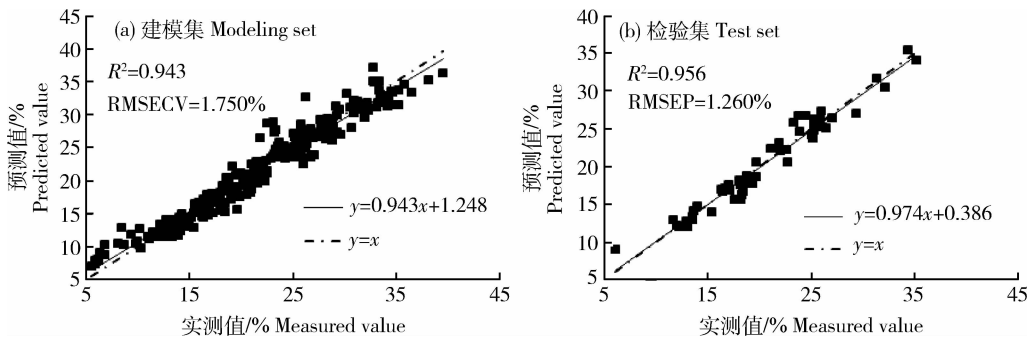


图 3 全局 PLS 模型预测值与实测值的比较

Fig. 3 Relationship between the measured and predicted values using global partial least squares regression method

2.4 局部建模方法

预测局部建模中,为了确定定标子集数与 RMSEP 的关系,本研究以定标子集数为限定距离,将定标子集数以 10 为间隔,从 10 依次取到 200。对每个定标子集,依次提取前 20 个主成分(定标子集数为 10、20 时,则依次提取前 8、15 个主成分),选出每个定标子集下 RMSEP 值最小的主成分数作为此定标子集下所提取的主成分数。图 4 示出采用欧氏、马氏距离方法选取不同定标子集数与 RMSEP 的关系。采用欧氏距离法,定标子集数为 20,提取的主成分数为 2 时, RMSEP 值最小;采用马氏距离法,定标子集为 70,提取主成分数为 4 时, RMSEP 值最小。此时,采用欧氏距离法和马氏距离法选取定标子集数,所提取的光谱主成分贡献率均达到 99.99% 以上。

表 2 为采用欧氏距离法和马氏距离法进行局部 PLS 建模下的最优结果,结果表明采用马氏距离法

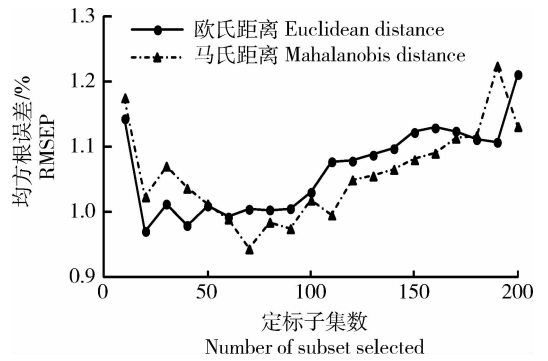


图 4 2 种距离方法下选取定标子集数与预测均方根误差的关系

Fig. 4 Relationship between the number of subset selected and RMSEP

的效果优于欧氏距离法,故本研究局部 PLS 建模以马氏距离法为标准选取定标子集数。图 5 示出最优结果下对检验集样本进行预测的拟合图。

表 2 基于局部 PLS 建模的不同距离方法下最优结果比较

Table 2 Comparison of optimal results based on different distance methods by local partial least squares regression

局部 PLS 建模 Local PLS modeling	定标子集数 Number of selected set	主成分数 Number of principal components	检验集 Test set	
			R ²	RMSEP/%
欧氏距离法 Euclidean distance method	20	2	0.976	0.970
马氏距离法 Mahalanobis distance method	70	4	0.976	0.943

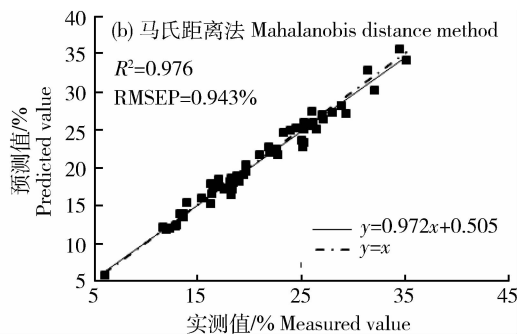
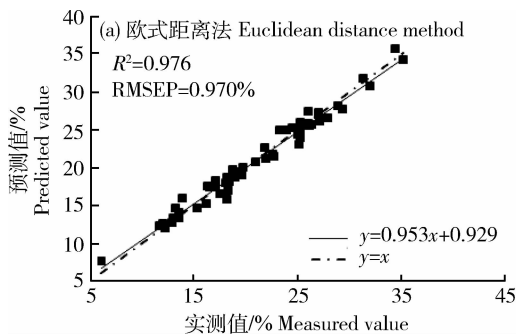


图 5 局部 PLS 模型预测值与实测值的比较

Fig. 5 Relationship between the measured and predicted values using local partial least squares regression method

欧氏距离法和马氏距离法均以聚类分析中的相似性作为判断的标准。为了直观地表示 2 种方法在样本相似性方面表达上的异同,图 6 示出采用欧氏距离法和马氏距离法选择定标子集数为 70 时的主

成分得分分布情况。可以看出,欧氏距离法所选择的定标子集近似于圆形分布,马氏距离则更近似于椭圆。这主要是马氏距离在计算时考虑到了不同方向上的权重,更能反映样本间的相似性。

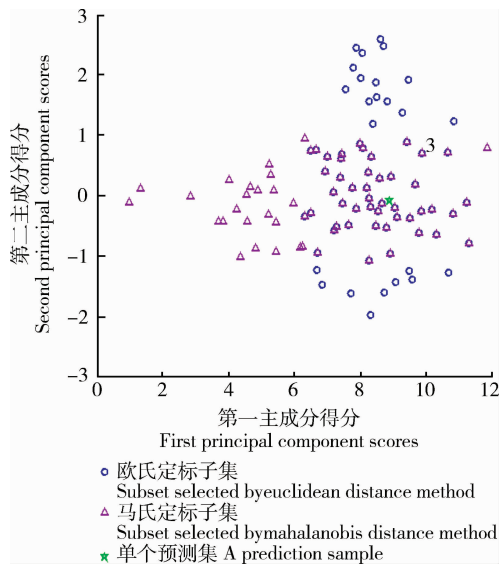
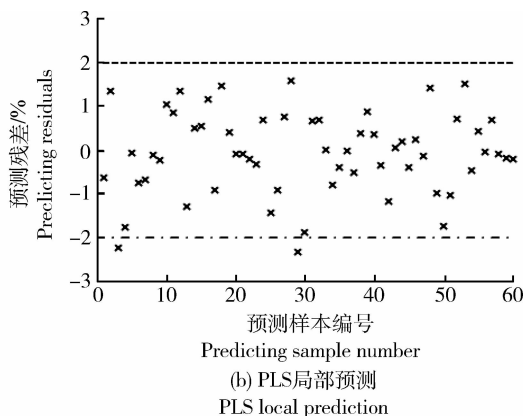
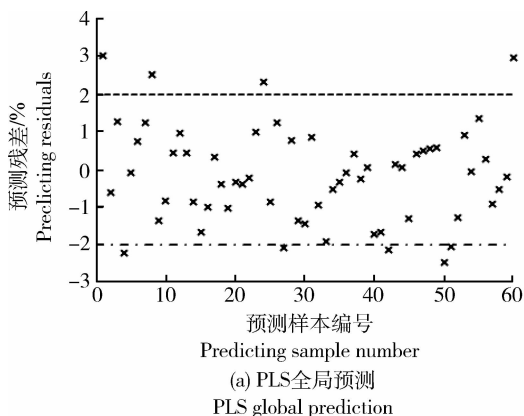


图6 欧氏距离法、马氏距离法选取的定标子集主成分得分分布

Fig. 6 Principal component score distribution of subset selected by Euclidean and Mahalanobis distance method



× 残差 Residual ---- $y=2$ - - - $y=-2$

图7 检验集预测残差图

Fig. 7 Predicting residuals

3 结论

本研究将可见/近红外光谱技术应用到田间原位含水量快速估测,采用全局和局部 PLS 法,建立了土壤含水量快速估测模型。从检验集预测结果看,在 1 200~2 300 nm 波段建模效果均较为理想,说明将可见/近红外光谱技术应用于土壤含水量的田间原位测量上是可行的。

比较了全局建模与局部建模的预测精度,局部 PLS 建模方法的预测精度优于全局 PLS 建模方法。其主要原因是局部建模样本与预测样本在主成分空

2.5 全局建模与局部建模方法预测结果比较

对全局和局部 PLS 回归模型建模效果进行比较。全局建模中,检验集的 R^2 、RMSEP 值分别为 0.956、1.260%(图 3(b));局部建模中(马氏距离法),检验集的 R^2 、RMSEP 值分别为 0.976、0.943%(图 5(b))。从检验集的预测结果上看,采用局部建模方法在估测精度上高于全局建模方法。

对全局和局部 PLS 回归模型的预测残差进行比较。全局 PLS 建模中,模型预测值和实测值之间的残差平方平均值为 1.587%,其中最大误差为 3.009%(图 7(a));局部 PLS 建模中(马氏距离法),模型预测值和实测值之间的残差平方平均值为 0.889%,其中最大误差为 2.344%(图 7(b))。从预测残差角度上看,采用局部建模在预测准确度上要优于全局建模。

间上具有相似性,预测效果更佳。在局部模型中,比较了选取定标子集数的方法,结果显示采用马氏距离法优于采用欧氏距离法。其主要原因是采用欧氏距离法计算距离时对各个方向同等考虑,而马氏距离法在计算距离时考虑到了不同方向上的加权,故采用马氏距离法预测效果优于采用欧氏距离法。

本试验研究对象为水稻土,所建立的模型是否适用于其他类型的土壤有待探讨。进一步可将可见/近红外技术引入到不同的土壤类型中,分别建立相应的可见/近红外光谱模型,最终建立不同土壤类型可见/近红外光谱数据库。

参 考 文 献

- [1] 罗锡文, 臧英, 周志艳. 精细农业中农情信息采集技术的研究进展[J]. 农业工程学报, 2006, 22(1): 167-173
Luo X W, Zhang Y, Zhou Z Y. Research progress in farming information acquisition technique for precision agriculture[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2006, 22(1): 167-173 (in Chinese)
- [2] 李民赞. 光谱分析技术及其应用[M]. 北京: 科学出版社, 2006: 78-79
Li M Z. *Spectral Analysis Technique and Application* [M]. Beijing: Science Press, 2006: 78-79 (in Chinese)
- [3] Bowers S A, Hanks R J. Reflection of radiant energy from soils [J]. *Soil Science*, 1965, 100(2): 130-138
- [4] Dalai R C and Henry R J. Simultaneous determination of moisture, organic carbon, and total nitrogen by near-Infrared reflectance spectrophotometry [J]. *Soil Science Society of America Journal*, 1986, 50(1): 120-123
- [5] 孙建英, 李民赞, 郑立华, 胡永光, 张喜杰. 基于近红外光谱的北方潮土土壤参数实时分析[J]. 光谱学与光谱分析, 2006, 26(3): 424-429
Sun J Y, Li M Z, Zheng L H, Hu Y G, Zhang X J. Real-time analysis of soil moisture, soil organic matter, and soil total nitrogen with NIR spectra [J]. *Spectroscopy and Spectral Analysis*, 2006, 26(3): 424-429 (in Chinese)
- [6] Mouazen A M, Kuang B, Baerdemaeker J D, Ramon H. Characterization of soil water content using measured visible and near infrared spectra[J]. *Soil Science Society of America Journal*, 2006, 70(4): 1295-1302
- [7] Mouazen A M, Kuang B, Baerdemaeker J D, Ramon H. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy[J]. *Geoderma*, 2010, 158(1-2): 23-31
- [8] Zhu Y D, David C, Chakraborty W S, Haggard B, Johnson S, Bakr F. Characterizing surface soil water with field portable diffuse reflectance spectroscopy[J]. *Journal of Hydrology*, 2010, 391(1): 133-140
- [9] 肖武, 李小昱, 李培武, 雷延武, 王为, 刘洁, 冯耀泽. 基于近红外光谱土壤水分检测模型的适应性[J]. 农业工程学报, 2009, 25(3): 33-36
Xiao W, Li X Y, Li P W, Lei T W, Wang W, Liu H, Feng Y Z. Adaptability of the model for soil moisture measurement based on near-infrared spectroscopy[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2009, 25(3): 33-36 (in Chinese)
- [10] 石雪, 蔡文生, 邵学广. 局部建模方法用于烟草样品的近红外光谱定量分析[J]. 光谱学与光谱分析, 2008, 28(11): 2561-2564
Shi X, Cai W S, Shao X G. A local regression method for near-infrared spectral quantitative analysis of tobacco samples[J]. *Spectroscopy and Spectral Analysis*, 2008, 28(11): 2561-2564 (in Chinese)
- [11] 吴茜, 杨宇虹, 徐照丽, 晋艳, 郭焱, 劳彩莲. 应用局部神经网络和可见/近红外光谱法估测土壤有效氮磷钾[J]. 光谱学与光谱分析, 2014, 34(8): 2102-2105
Wu Q, Yang Y H, Xu Z L, Jin Y, Guo Y, Lao C L. Applying local neural network and visible/near-infrared spectroscopy to estimating available nitrogen, phosphorus and potassium in soil [J]. *Spectroscopy and Spectral Analysis*, 2014, 34(8): 2012-2015 (in Chinese)
- [12] Cristine L S, Waiser T H, Brown D J, Hallmark C T. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy[J]. *Geoderma*, 2009, 151(3): 249-256
- [13] 鲍士旦. 土壤农化分析[M]. 北京: 中国农业出版社, 2010: 22-24
Bao S D. *Soil Agricultural Chemistry Analysis* [M]. Beijing: China Agriculture Press, 2010: 22-24 (in Chinese)
- [14] 高惠璇. 应用多元统计分析[M]. 北京: 北京大学出版社, 2005: 369-378
Gao H X. *Applied Multivariate Statistical Analysis* [M]. Beijing: Peking University Press, 2005: 369-378 (in Chinese)
- [15] 展晓日, 朱向荣, 史新元, 张卓勇, 乔延江. 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
Zhan X R, Zhu X R, Shi X Y, Zhang Z Y, Qiao Y J. Determination of hesperidin in tangerine leaf by near-infrared spectroscopy with SPXY algorithm for sample subset partitioning and Monte Carlo cross validation[J]. *Spectroscopy and Spectral Analysis*, 2009, 29(4): 964-968 (in Chinese)

责任编辑: 刘迎春