基于高光谱数据预测土壤碱化程度最佳模型及其影响因素的研究①

王凯龙[1,3]，熊黑钢[2,3*]，张 芳[1,3]

(1 新疆大学资源与环境科学学院，乌鲁木齐 830046；2 北京联合大学应用文理学院，北京 100083；
3 教育部新疆绿洲生态重点实验室，乌鲁木齐 830046)

摘 要：为快速准确地估测土壤碱化程度，对实测波段范围为 400 ~ 900 nm 的土壤光谱数据进行了波段差、波段比、波段归一化 3 种预处理，采用偏最小二乘法(PLSR)建立了不同波段范围的土壤 pH 的预测模型，并利用测试集数据对模型进行精度检验。结果表明：采用归一化、波段比 2 种方式对原始光谱进行预处理，可有效地增强光谱与土壤 pH 的相关性，并抑制干扰信息，其中归一化最优。虽然可见光波段范围(400 ~ 750 nm)所建立的预测模型与全波段(400 ~ 900 nm)预测模型 R^2 相同，但其 $RMSE_P$ 比全波段减少了 0.059，RPD 提高了 0.2，说明该波段范围包括了反映土壤 pH 的大部分信息，是建立其预测模型的优势波段。因此，利用可见光波段的光谱数据，采用归一化预处理可以具有较好稳定性和预测能力地预测土壤 pH 的最佳模型(R^2 = 0.90，RMSECV = 0.104)。

关键词：实测光谱；土壤 pH；最佳波段

中图分类号：TP70；S15

盐碱和干旱一样，是长期困扰世界农业生产的重大问题之一。获取有关碱化土壤的性状、范围、面积、地理分布及碱化程度等方面的实时、可靠的信息，对治理碱化土壤，防止其进一步退化和农业可持续发展至关重要[1–2]。高光谱传感器能获取纳米级的地物连续光谱信息，其反映地物光谱细微特征，使得依据诊断性的光谱吸收特征来识别地物、进行遥感定量分析、研究地物的化学成分等成为可能。由于土壤的类型、地区及所处的环境等复杂性，其光谱特征不但受化学成分的影响，还会受到物理性状如颗粒大小、温度、湿度等多方面因素的影响，因此必须对原始光谱进行预处理，消除噪声，保留有用信息以便建立一个稳定、准确的土壤光谱模型。学者们通过多种方法对土壤光谱进行预处理，例如徐永明等[3]采用反射率 Reflectance、一阶导数 FDR、倒数之对数 log(1/R) 和波段深度 Depth 共 4 种光谱指标对土壤营养元素进行了估算。王淼等[4]采用 Savitzky-Golay 多项式对原始光谱进行了平滑、一阶微分、二阶微分处理，并利用 4 种窗口组合光谱对红壤有机质进行了研究，结果表明平滑与一阶微分是建立土壤有机质的最佳预处理形式。申艳军等[5]研究了多元散射校正 (MSC) 技术在光谱预处理中的应用，并利用其对黑土有机碳含量进行了估测。高洪智等[6]利用连续投影算法提取土壤总氮的近红外特征波长并对其进行了估算。周广柱等[7]提出了对数变换与小波变换相结合的降噪方法，采用空域相关算法对野外光谱进行了预处理。目前，提高建模精度的探讨主要集中在：①采用波段组合，即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，使其在突出主要信息的同时，使次要信息最小化。其可以消除多变量间的共线性影响、降低信噪比、消除背景干扰从而增强有用信息和抑制干扰信息[8–9]。②因土壤属性的不同，其对不同光谱波段的响应有所差异。因此通过土壤属性找出其对光谱响应的最佳波段，并建立最优模型。但通过波段组合，寻找最适合预测土壤 pH 的波段范围和模型的研究尚未见报道。本文以分布有大面积碱化土壤的新疆奇台绿洲为研究靶区，在选择出预测土壤 pH 最佳波段组合的基础上，结合偏最小二乘回归

①基金项目：国家自然科学基金项目(41171165、41261049)、北京联合大学人才强校计划人才资助项目(BPHR2012E01)和新疆大学博士启动基金项目(BS110124)资助。

* 通讯作者(xhg1956@sohu.com)

作者简介：王凯龙(1988—)，男，北京人，硕士研究生，主要研究方向为遥感应用及干旱区环境研究。E-mail: wkl1005@126.com

方法，建立了不同光谱范围的土壤 pH 预测模型，并分析了影响土壤 pH 模型的主要因素。旨在找出预测土壤 pH 的较为适合的预处理形式、最优波段范围以及最佳模型，为建立更为精准的土壤碱化高光谱定量监测提供了新的视角。

1　材料与方法

1.1　研究区概况

奇台县位于新疆维吾尔自治区东北部，天山山脉东段，博格达山北麓，准噶尔盆地东南缘(图 1)。地处 43°25′ ~ 45°29′N，89°13′ ~ 91°22′E。年平均气温为 5℃ 左右，年均降水 176 mm，年均蒸发势 2 141 mm。本文以奇台绿洲中部冲积平原区作为研究区域，范围为 43°56′56″~ 44°13′24″N，89°20′46″~ 90°3′43″E。碱化土壤属于龟裂碱土，碱化特征在新疆天山北坡、准噶尔盆地南缘区域具有一定的代表性。

1.2　土壤样本采集与处理

利用 GPS 定位技术在研究区内的典型荒地区域布点 40 个，采样点要求：①尽可能规则遍及所有荒地类型；②样点周围土壤性质相对成因一致，环境因子类似，异质性较小；③每个样点区域不小于 30 m × 30 m。采集土壤表层(0 ~ 20 cm)土样，采用梅花桩采样法进行采样，每个样区共设 5 个重复。采集的土壤样本在实验室内自然晾干，分散，过 1 mm 筛，按水土比 5∶1 配置土壤浸提液，pH 使用数字式酸度计测定，土壤有机质采用重铬酸钾容量法–外加热法测定。

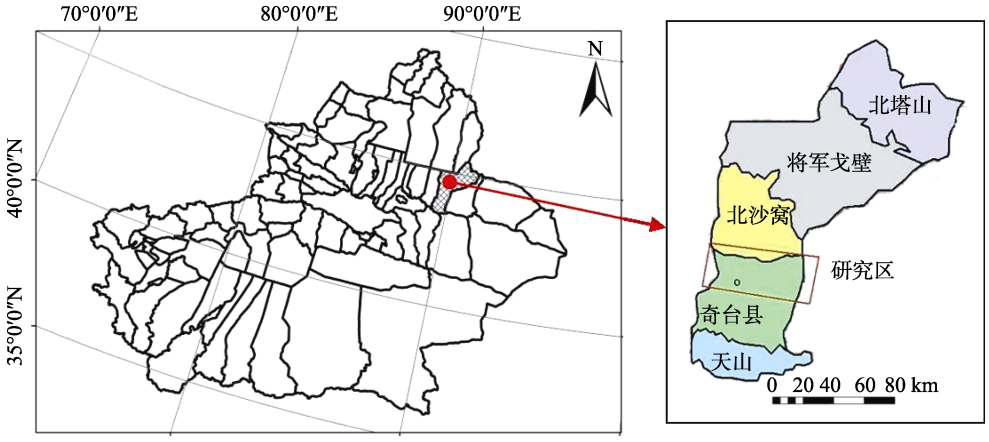

图 1　研究区示意图
Fig. 1　Sketch of study area

1.3　野外光谱测量与预处理

野外光谱测量与土壤取样同期同点进行，光谱测定使用 ASD FieldSpec HandHeld 光谱仪，测定范围 325 ~ 1 075 nm，并去除了噪声较大的 325 ~ 400 nm 和 900 ~ 1 075 nm 波段。测试时间为 10:00—14:00 (地方时)。测量期间天气状况良好，晴朗无云，风力较小，光谱仪采用垂直向下测量的方法，与多数传感器采集数据的方向一致。测点所处的地方地面平坦，能代表周围较大面积的特征。测量距离 15 cm，测量土壤光谱时避开植物影响，每次测定前严格按照操作规范，去除暗电流影响，进行标准白板定标。为保证光谱数据具有代表性，对同一种地物采取 5 次测量取算术平均值，得到该地物的反射光谱曲线。

首先采用 Savitzky-Golay 多项式对原始光谱进行了平滑，然后按照波段差值、归一化和波段比 3 种波段组合的方法对光谱进行预处理。具体计算公式如下：

波段差：$A = A_{i+1}-A_i$　　(1)

波段归一化：$A = (A_{i+1}-A_i)/(A_{i+1}+A_i)$　　(2)

波段比：$A = A_i/A_{i+1}$　　(3)

其中：A_i 和 A_{i+1} 分别为连续光谱波段的反射率值，i =(400, 401, …, 900 nm)，A 为预处理后的光谱。

1.4　偏最小二乘法(PLSR)

PLSR 继承了主成分分析和典型相关分析的提取主成分的思想，实现了数据结构的简化，解决了自变量之间多重相关的问题，同时又克服了主成分分析对自变量有较强解释能力但对因变量解释能力不够的缺点，适合用于光谱分析这种自变量较多的情况。使用 PLSR 建立定标模型时，主因子个数的确定直接影响到模型的预测能力，主因子个数太小，重建光谱拟合不够；反之，则将过度拟合。偏最小二乘模型利用 DPS 数据统计软件实现。

1.5　校正集和预测集的划分及验证评价参数

选用 KS (Kennard-Stone)[10]算法计算出各个样品吸光度值之间的欧氏距离，按照 3∶2 的比率划分为校正集和测试集(表 1)。本文选择主成分个数(number

of PLS factors)、校正集交互验证决定系数(cofficient of determination in cross validation , R^2_{CV})、预测集决定系数 (cofficient of determination, R^2_P)、校正集和预测集均方根误差($RMSE_{CV}$ 和 $RMSE_P$)、测定值标准差与标准预测误差的比值 RPD (ratio of standard deviation to standard error of prediction) 对模型精度进行评价。其中，校正集交互验证决定系数(R^2_{CV})越大，均方根误差($RMSE_{CV}$)越小，说明模型精度越好；预测集决定系数(R^2_P)和 RPD 越大，均方根误差($RMSE_P$)越小，表明预测模型越稳定，精度越高。另外，当 RPD>2 时，模型具有极好的预测能力；当 1.4<RPD<2 时，模型可对样品作粗略估测；而 RPD<1.4 则模型无法对样品进行预测[11]。

表 1 土壤样本 **pH** 统计特征
Table 1 Statistics of soil pH

成分	样本数	范围	均值	方差
校正集	24	7.77 ~ 10.58	9.33	0.59
预测集	16	7.90 ~ 10.32	9.16	0.56
全部样本	40	7.77 ~ 10.58	9.28	0.57

2 结果与分析

2.1 不同光谱组合与土壤 pH 的相关性分析

由于光谱测定是直接在野外进行的，受到外界干扰因素较多，因此原始光谱与土壤 pH 的相关系数较低(0.43 ~ 0.51)(图 2)。波段差与原始光谱相比，只有在可见光的少部分波段有较高的相关系数，大部分波段降低了其与土壤 pH 的相关性，显示出这种方法会引起更多的新的噪声。经过波段比和归一化这两种形式预处理后的光谱，相关性系数在可见光部分相对于原始光谱明显增加，最大值可达 –0.84，但在近红外波段则降低。这表明两种预处理方法可以在可见光波段有效增强与土壤 pH 有关的信息，同时也表现出可见光波段比近红外更为适合反映土壤 pH 信息，是建立预测土壤 pH 的优势波段。

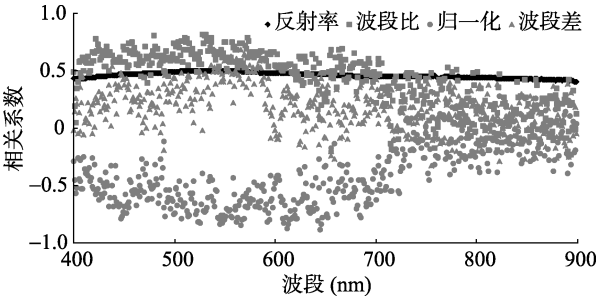

图 2 **pH** 与光谱变量之间的相关系数图
Fig. 2 Correlation coefficient between soil pH and spectrum reflectance

2.2 各光谱组合的土壤 pH 模型建立与检验

采用偏最小二乘法(PLSR)建立 4 种光谱指标与土壤 pH 的校正模型，并用预测集数据进行检验(表 2)。4 种模型除波段差外，其余 RPD 均大于 2，模型均为优等，说明采用 400 ~ 900 nm 光谱建立土壤碱化模型是可行的。从校正集模型来看，与原始光谱相比，经过归一化、波段比对光谱进行预处理后，降低了主成分的个数，模型预测精度明显提高，均方根误差($RMSE_{CV}$)有所降低。综合来看，归一化精度更高。预测集模型检验也同样说明了这一点 (表 2)。因此采用这种预处理形式可以增强 PLSR 模型的预测能力，是光谱在 400 ~ 900 nm 预测土壤 pH 的最佳模型。

表 2 不同光谱指标所建立的模型精度检验
Table 2 Accuracy tests for different models

光谱预处理形式	校正集模型建立			预测模型检验		
	判定系数 R^2	$RMSE_{CV}$	主成分个数	判定系数 R^2	$RMSE_P$	RPD
原始光谱	0.83	0.274	7	0.80	0.547	2.12
波段差	0.67	0.682	8	0.60	0.952	1.83
归一化	0.89	0.104	5	0.86	0.195	2.53
波段比	0.88	0.117	5	0.83	0.236	2.45

2.3 不同波段范围的模型对比

选择归一化预处理方式后的光谱数据分别建立可见光波段(400 ~ 750 nm)、近红外波段(750 ~ 900 nm)和全波段(400 ~ 900 nm)预测土壤 pH 的定量模型(表 3)。结果显示：从校正集模型来看，波段范围 400 ~ 750 nm 的模型效果最佳，400 ~ 900 nm 其次，而 750 ~ 900 nm 所建模型精度较差。预测集模型也说明了相似的情况：虽然可见光波段范围(400 ~ 750 nm)所建立的预测模型与全波段(400 ~ 900 nm)模型预测 R^2 相同，但其 $RMSE_P$ 比全波段减少了 0.069，RPD 提高了 0.12，说明该波段范围包括了反映土壤 pH 的大部分信息，是建立其预测模型的优势波段。同时，在可见光范围内的 490 ~ 590 nm 和 620 ~ 710 nm 对模型有较高的贡献率，明显大于近红外部分(图 3)，同样反映出该波段是提取土壤 pH 信息的最佳波段。

表 3　基于归一化预处理的 **PLSR** 建模和精度检验
Table 3　PLSR analysis based on normalization

波段范围 (nm)	模型建立			预测模型检验		
	判定系数 R^2	$RMSE_{CV}$	主成分个数	判定系数 R^2	$RMSE_P$	RPD
400 ~ 750	0.90	0.104	5	0.86	0.195	2.65
750 ~ 900	0.73	0.664	7	0.80	0.953	1.86
400 ~ 900	0.89	0.125	5	0.86	0.264	2.53

2.4　最佳模型对土壤碱化程度的响应

表 4 为最佳波段组合(归一化)和最佳波段范围(可见光)所建立的不同土壤 pH PLSR 模型。虽然不同土壤 pH 下，各模型主成分个数一致，但随着碱化程度的增强，模型判定系数 R^2 增加，$RMSE_{CV}$ 降低，即碱化程度越强，所建立的模型效果越好。从判定系数 R^2 来看，pH>9.5 的模型精度比 pH<8.5 的高 5 个百分点，而 $RMSE_{CV}$ 小 0.08(表 4)。这可能是因为碱化程度越强影响土壤光谱的因素越少，光谱中反映土壤 pH 的信息越“纯”，所建立的模型效果越好。

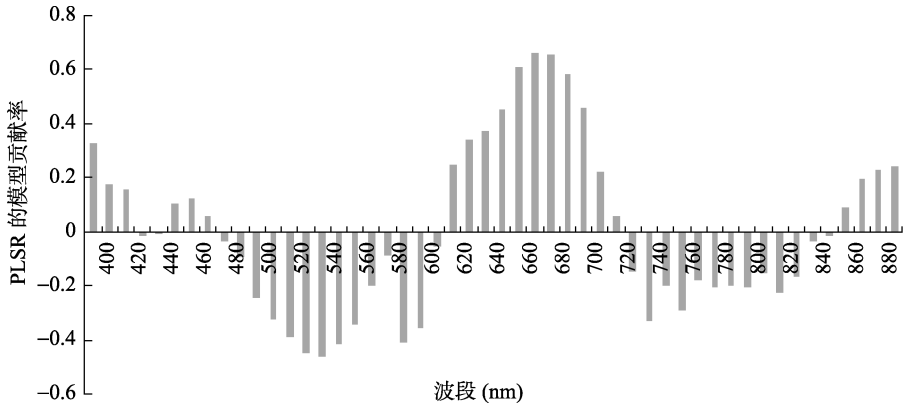

图 3　基于归一化预处理形式的 **PLSR** 模型贡献率
Fig. 3　Coefficients of PLSR model based on normalization

表 4　不同土壤 **pH** 下可见光波段的模型精度比较
Table 4　Comparison of PLSR models based on normalization under different pH ranges

PLSR 模型精度	pH<8.5	pH 8.5 ~ 9.5	pH>9.5
判定系数 R^2	0.86	0.88	0.91
$RMSE_{CV}$	0.168	0.143	0.088
主成分数	5	5	5

3　讨论

土壤表面及其内在因素的理化性质影响土壤的光谱特征。由于研究区土壤粒度、粗糙度等基本一致[12]，并且采样点无植被影响。因此本文认为土壤 pH 模型预测精度的差异主要是土壤碱化程度和其内部的有机质造成的，二者均通过影响土壤表面颜色来影响土壤光谱进而影响模型精度。

3.1　土壤有机质对模型精度的影响

土壤颜色的深浅在一定程度上可以反映有机质含量的多少。土壤有机质含量越高，颜色越暗。土壤有机质含量大于 2% 时，可见光波段是土壤有机质的敏感波段，也是预测土壤有机质的最佳波段范围，其对土壤光谱的影响占主导地位[13−16]。表 5 是不同土壤碱化程度下的土壤 pH 和有机质含量的 PLSR 预测模型精度及有机质含量。当土壤 pH<8.5 时，有机质含量为 38 g/kg，土壤 pH 模型精度最低，而有机质模型精度最高，说明此时有机质含量对土壤光谱的影响占主导地位，使得土壤 pH 的预测模型精度降低。随着土壤 pH 的增大，有机质含量逐渐降低(表 5)。当 pH>9.5 时，土壤有机质含量仅为 8.0 g/kg，土壤 pH 预测模型精度最高，并且土壤有机质模型精度最低。表明此时土壤碱化程度对光谱的影响占主导地位，所建立的土壤 pH 的定量模型受土壤有机质的影响达到最小。

表 5　土壤 **pH** 和有机质的预测模型精度及有机质含量
Table 5　Accuracy of prediction models of soil pH and organic matter based on pH values and contents of soil organic matter

项目		pH<8.5	pH 8.5 ~ 9.5	pH>9.5
土壤 pH 预测模型	PLSR 判定系数 R^2	0.84	0.85	0.88
	$RMSE_P$	0.183	0.154	0.121
土壤有机质含量预测模型	PLSR 判定系数 R^2	0.79	0.64	0.38
	$RMSE_P$	0.354	0.437	2.217
有机质含量(g/kg)		3.8	1.5	0.8

3.2　土壤颜色对土壤碱化程度的响应

土壤表面颜色是影响土壤光谱的重要因素之一。明度、色调是颜色的重要指标。利用中国标准色卡测量研究区土壤表层土样，其明度范围为 4 ~ 8。当 pH< 8.5 时，土壤明度主要为 5 和 6(33% 和 50%)；当 pH

在 8.5 ~ 9.5 范围内时，明度为 4 的样点消失，明度为 5 的样点减少，6 和 7 的样点数增加；当 pH>9.5 时，土壤明度以 7 为主(60.9%)，所有样点明度均大于 6。即随着土壤 pH 的增大，有机质减少，土壤明度也随之增强(图 3)。

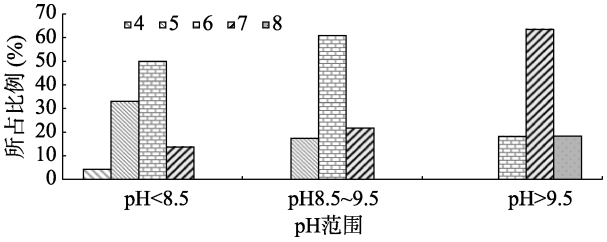

图 4 土壤明度分布

Fig. 4 Distribution of color value of soil samples

同时研究区的土壤色调以 10YR 为主，属色调黄红[17]，在可见光中的波段范围为 560 ~ 760 nm，而本文中的贡献率较高的波段 620 ~ 710 nm 刚好在此波段范围内。这也证明了可见光 400 ~ 750 nm 波段是预测碱化土壤的最佳波段。

4 结论

(1) 由于实测光谱测定是在野外进行的，受到外界干扰因素较多，采用波段组合对光谱进行预处理可有效增强光谱与土壤 pH 的相关性，其中归一化效果最好，最大相关系数可达 –0.84，是最优预处理方式。

(2) 不同波段范围所建立的模型中，400 ~ 750 nm 的模型效果最佳(R^2 = 0.90，$RMSE_{CV}$ = 0.104)，400 ~ 900 nm其次(R^2 = 0.89，$RMSE_{CV}$ = 0.125)，而 750 ~ 900 nm 所建模型精度最差(R^2 = 0.73，$RMSE_{CV}$ = 0.664)。因此利用归一化方法 + 可见光波段土壤光谱数据所建立的模型效果最好。

(3) 影响土壤 pH PLSR 模型预测精度的主要因素是土壤碱化程度和有机质。当 pH<8.5 时，土壤的有机质含量大于 38 g/kg，其对土壤光谱的影响占主导地位，所建立的土壤 pH 预测模型精度最低；当 pH>8.5 时，土壤碱化程度占主导地位，并且随着土壤碱性增强到 pH>9.5 时，模型精度提高了 4%。因此在推广到遥感影像监测土壤碱化程度时，应考虑二者对模型精度的影响。

(4) 本文采用的归一化预处理方法实际上是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，此类思想的信息增强与提取技术源于植被指数。而这种处理方法必然导致各波段数据之间可比性降低，并且增加一些噪声。由于条件限制，数据量较少，对于数据规律性的把握欠缺。因此在推广时，应注意这种数据处理方法所存在的风险性。

参考文献：

[1] 刘庆生, 刘高焕, 励惠国. 辽河三角洲土壤盐分与上覆植被野外光谱关系初探[J]. 中国农学通报, 2004, 20(4): 274–278

[2] 塔西甫拉提·特依拜, 张飞, 赵睿, 何祺胜. 新疆干旱区土地盐渍化信息提取及实证分析[J]. 土壤通报, 2007, 38(4): 235–341

[3] 徐永明, 蔺启忠, 王璐, 黄秀华. 基于高分辨率反射光谱的土壤营养元素估算模型[J]. 土壤学报, 2006, 43(5): 709–716

[4] 王淼, 解宪丽, 周睿, 王宝良，王昌昆，刘娅. 基于可见光–近红外漫反射光谱的红壤有机质预测及其最优波段选择[J]. 土壤学报, 2011, 48(5): 1 083–1 089

[5] 申艳, 张晓平, 梁爱珍, 范如芹, 杨学明. 多元散射校正和逐步回归法建立黑土有机碳近红外光谱定量模型[J]. 农业系统科学与综合研究, 2010, 26(2): 174–180

[6] 高洪智, 卢启鹏, 丁海泉, 彭忠琦. 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): 2 951–954

[7] 周广柱, 王翠珍, 杨锋杰 , 李寅明. 对数变换与小波变换用于野外采集植物波谱降噪[J]. 红外与毫米波学报, 2009, 28(4): 316–340

[8] 田庆久, 闵祥军. 植被指数研究进展[J]. 地球科学进展, 1998, 13(4): 327–333

[9] 张韬, 赵宇飞, 安慧君，陈秀兰. ETM + 影像提取伏沙地信息的最佳波段组合——以内蒙古锡林郭勒盟西乌旗为例[J]. 科技导报, 2011, 29(17): 29–32

[10] Volkan BA, van Es HM, Akbas F, Durak WD. Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey[J]. Journal of Arid Environments, 2010, 74: 229–238

[11] 杨海清. 基于光谱技术的土壤成分和植物生长信息快速获取建模和仪器研究(博士学位论文)[D]. 杭州: 浙江大学, 2012

[12] 张芳, 熊黑钢, 田源, 栾福明. 区域尺度地形因素对奇台绿洲土壤盐渍化空间分布的影响[J]. 环境科学研究, 2010(7): 731–738

[13] 程朋根, 吴剑, 李大军，何挺. 土壤有机质高光谱遥感和地统计定量预测[J]. 农业工程学报, 2009, 25(3): 142–147

[14] Galvao LS, Vitorello I. Role of organic matter in obliterating the effects of iron on spectral reflectance and colour of Brazilian tropical soils[J]. International Journal of Remote Sensing, 1998, 19(10): 1 969–1 979

[15] 何挺, 王静, 林宗坚, 程烨. 土壤有机质光谱特征研究[J]. 武汉大学学报, 2006, 31(11): 975–979

[16] 贺军亮, 蒋建军, 周生路, 徐军, 蔡海良. 土壤有机质含量的高光谱特性及其反演[J]. 中国农业科学, 2007, 40(3): 638–643

[17] 张芳, 熊黑钢, 卢文娟, 栾福明. 实测光谱对土壤碱化的响应特征[J]. 光谱学与光谱分析, 2011, 31(5): 1 245–1 249

Optimal Model of Soil pH and Influencing Factors By Using Hyperspectral Data

WANG Kai-long[1,3], XIONG Hei-gang[2,3*], ZHANG Fang[1,3]

(1 *College of Resources & Environment Science*, *Xinjiang University*, *Urumqi* 830046, *China*; 2 *Urban Department of College of Art&Science*, *Beijing Union University*, *Beijing* 100083, *China*; 3 *Key Laboratory of Oasis Ecology* (*Xinjiang University*), *Ministry of Education*, *Urumqi* 830046, *China*)

Abstract: Based on the monitored data of soil PH and measured VIS-NIR reflectance on given spots, the relationship between measured reflectance and soil PH was analyzed. Partial least squares regression (PISR) was used to build predicting model of pH value, band ratio, differential and normalization were calculated based on measured VIS-NIR reflectance within 400 – 900 nm. The results showed that band ratio and normalization can effectively enhance the correlations between spectral and soil pH, and interference information was suppressed. Accuracy of the model based on normalized got the best effect, R^2 was 0.90, which showed that the band contained most information of soil pH.

Key words: Field reflectance, Soil pH, Optimal band