

基于离散小波的土壤全氮高光谱特征提取与反演

张娟娟^{1,2}, 牛 圳^{1,2}, 马新明^{1,2}, 王 健¹, 徐超越^{1,2},
时 雷^{1,2}, Bação Fernando³, 司海平^{1,2*}

1. 河南农业大学信息与管理科学学院, 河南 郑州 450002

2. 河南粮食作物协同创新中心, 河南 郑州 450002

3. Universidade Nova de Lisboa, NOVA Informantion Managment School, Lisboa, 1070-312, Portugal

摘 要 土壤全氮是重要的养分指标, 利用高光谱技术研究并构建砂姜黑土全氮含量高光谱估测模型, 为作物施肥及发展精确农业提供参考。尝试研究离散小波估测土壤全氮含量的可行性, 以河南省商水县不同小麦氮肥处理为试验区, 采集 100 份 0~20 cm 的砂姜黑土, 土壤样本风干并经研磨过筛等处理后, 在实验室暗室内采集光谱。利用含量梯度法, 将总样本(100 个砂姜黑土)划分为建模集 75 个和验证集 25 个。将原始光谱进行一阶导数变换, 并对一阶导数光谱分别进行相关分析和离散小波变换, 同时结合支持向量机和 K 邻近算法构建高光谱土壤全氮估测模型。系统分析了原始光谱和一阶导数光谱的单波段与土壤全氮的相关性, 结果表明, 经一阶导数变换后的光谱与土壤全氮有更好的相关性, 在 1 373 nm 处相关系数达到最高为 0.84。利用离散小波算法对一阶导数光谱进行最佳母小波和分解层次选择, 结果显示, 经 sym8 函数分解的小波系数能较好的重构土壤全氮光谱信息, 进一步基于分解层 L_1 — L_{11} 的低频系数分别建立支持向量回归和 K 邻近回归土壤全氮含量估测模型, 比较全部估测模型, 以分解层 L_5 的低频系数结合 K 邻近构建的模型最优, 建模决定系数为 0.90, 均方根偏差为 $0.09 \text{ g} \cdot \text{kg}^{-1}$, 相对分析误差为 3.78, 验证决定系数为 0.97, 均方根偏差为 $0.05 \text{ g} \cdot \text{kg}^{-1}$, 相对分析误差为 4.30。同时与全波段和经相关分析后挑选出的敏感波段作为输入构建的模型进行比较, K 邻近模型精度提高了 3.2%和 9%, 支持向量机模型精度提高了 6.7%和 11.6%。研究结果表明一阶导数变换与离散小波技术可有效减少噪声影响, 提高土壤全氮含量的估测精度, 又实现了光谱数据降维, 简化了模型复杂度, 为砂姜黑土全氮含量的精确估测提供参考。

关键词 砂姜黑土; 全氮; 高光谱; 离散小波; K 邻近算法

中图分类号: S151.9 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2023)10-3223-07

引 言

砂姜黑土是我国珍贵的土壤资源, 其发育于河湖相沉积物、河间洼地和岗丘间洼地环境, 经前期草甸潜育化过程和后期耕作熟化过程所形成的一种半水成土, 全国面积计有 370 万 hm^2 左右, 主要分布于淮北平原, 其特点是质地黏重, 结构松散^[1]。全氮作为砂姜黑土肥力的重要组成部分, 准确、快速地估测其含量对农田砂姜黑土生产力、粮食安全和农业可持续发展具有重要意义, 也是加快精确农业发展的重要手段^[2]。目前, 砂姜黑土养分的分析仍以传统的实验室

化验分析方法为主, 测试过程费时、费力。近年来, 高光谱分析技术以其快速、简便、不破坏等特点^[3], 为土壤养分的快速诊断提供新的思路和技术手段。

国内外学者利用高光谱技术反演土壤全氮已有较多研究。目前, 利用高光谱进行土壤养分估测多是首先对光谱进行预处理, 如对原始光谱反射率进行一阶导数, 二阶导数、对数、连续统去除和平滑等处理^[4], 然后结合不同的建模方法进行模型的构建和验证。预处理方法和建模方法的不同选择, 均显著影响模型的预测精度。如 Cheng 等^[5]通过估算土壤重金属含量表明, 将光谱数据进行一阶导数、二阶导数变换可以消除实验室光谱中的基线漂移和多次散射效应。随着

收稿日期: 2022-06-20, 修订日期: 2022-10-08

基金项目: 国家重点研发计划项目(2021YFD1700905), 河南省科技攻关项目(192102110012)和河南省现代农业(小麦)产业技术体系项目(S2016-01-G04)资助

作者简介: 张娟娟, 1979 年生, 河南农业大学信息与管理科学学院副教授 e-mail: zhangjuan_2003@126.com

* 通讯作者 e-mail: pingsss@126.com

非线性模型算法的出现,较多研究结合不同的机器学习方法开展了土壤养分含量的光谱模型构建,并取得了较好的精度。如孙小香等^[6]将全波段原始光谱作为输入变量,结合偏最小二乘、BP神经网络和支持向量机3种建模方法,构建全氮含量高光谱估测模型。任红艳等^[7]采用偏最小二乘回归方法建立了基于全波段高光谱反射率的全氮含量预测模型。胡贵贵等^[8]在主成分方法特征提取基础上,结合K邻近和支持向量机模型高效定量分析了土壤有机质含量。然而,上述研究多利用全波段进行建模,而土壤全氮因受到水分、粒径和微量元素等土壤组分的干扰^[9],全氮光谱吸收特征较为微弱。为凸显或提取土壤光谱中的全氮信息,降低数据的冗余性,筛选与研究对象相关的光谱特征和变量成为提高建模精度的一种有效方法^[10]。

离散小波是一种通过对数据在时域和频域上的分解实现更加精确的局部分析以及信号特征的分离,将光谱信号分解为不同的层,各层所包含的信息与土壤不同成分有关。低频信息决定土壤光谱基本形状的物质,高频信息反映光谱采集过程的各种噪声及土壤微量成分在原始光谱的特征^[11]。王延仓等^[12]利用小波变换和偏最小二乘算法构建土壤有机质含量估测模型,结果表明小波变换可以提高模型的估测能力。郭云开等^[13]运用小波变换获得小波系数,放大了土壤高光谱特征信息,而且降低了数据的维度,结合模型预测土壤重金属铬含量是可行的。土壤全氮含量较低时,在土壤光谱中的吸收特征较为细弱,噪声会对光谱信息造成很大的干

扰,因此,如何将土壤光谱数据内噪声信息剔除,并最大限度地保留原始土壤光谱信息,是进一步提高土壤全氮预测精度的关键。

因此,以砂姜黑土为研究对象,利用离散小波变换对土壤全氮进行光谱特征提取,将低频信息和高频信息的分离,并确定土壤全氮最佳母小波和分解层次选择,同时结合支持向量机和K邻近算法构建土壤全氮估测模型,以期同类研究提供技术参考。

1 实验部分

1.1 数据采集

1.1.1 土壤数据

供试土壤样本取自河南省商水县国营农场的小麦氮肥处理试验区,试验施氮量分别为 $0 \text{ kg} \cdot \text{hm}^{-2}$ (N0)、 $90 \text{ kg} \cdot \text{hm}^{-2}$ (N6)、 $180 \text{ kg} \cdot \text{hm}^{-2}$ (N12)、 $270 \text{ kg} \cdot \text{hm}^{-2}$ (N18)、 $360 \text{ kg} \cdot \text{hm}^{-2}$ (N24)。氮肥使用尿素,分别于播种期和返青期按5:5施入。如图1,研究区地势开阔,土地养分充足,是河南省粮食主产区。土壤类型为砂姜黑土。野外采样时,将每个采样点地表植物清除,在每个单元按照五点取样法的原则采集0~20 cm的表层土壤,将土样置于阴暗通风处风干。为避免土壤粒径带来的土壤光谱异向、噪音问题,对土样进行研磨、分别过筛0.9和0.15 mm处理,供光谱测试和化学分析使用。

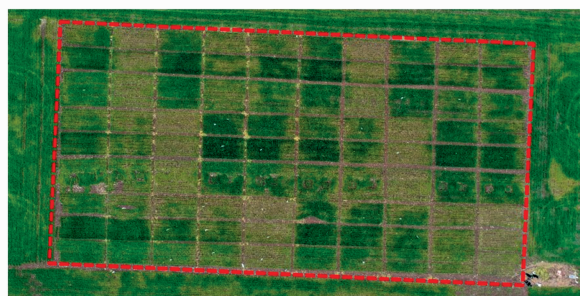
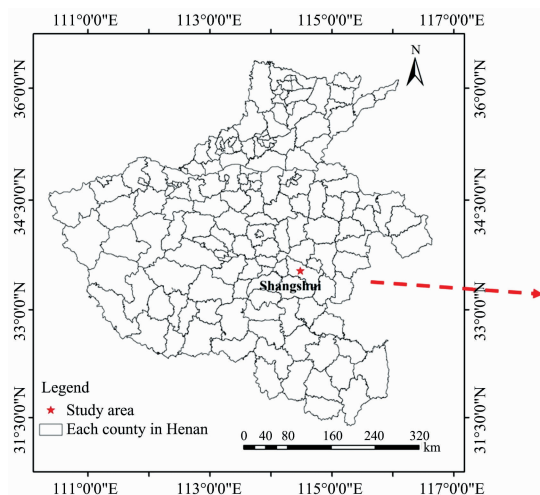


图1 研究区位置

Fig. 1 Location of study area

1.1.2 光谱数据

土壤光谱数据获取采用美国ASD公司生产的FieldSpec4地物光谱仪(光谱波段范围350~2500 nm)。测试在暗室进行,选择稳固的平台利用卤素光源和标准白板完成测量。把适量经处理的土壤样品倒入黑色盛样皿中,厚度为1.5 mm,用玻璃棒压实,使其表面尽可能平整。测量时为了减小土壤样品非目标因素的影响,测量过程中将样品旋转3次,每次采集10条共采集30条,计算30条曲线的平均值作

为样本的光谱反射率数据。删除350~399和2451~2500 nm信噪比低、噪声大的锯齿波段,共获得2051个波段数据。砂姜黑土耕层浅薄,较其他土壤类型光谱差异不明显。为扩大样本之间的光谱特征差异,对光谱曲线进行了一阶导数变换。

1.1.3 土壤全氮测定

土壤全氮测定采用凯式定氮法^[14],表1为土壤样本全氮含量统计特征。

表 1 土壤全氮含量统计 (g · kg⁻¹)

Table 1 Statistics of total nitrogen content in soil (g · kg⁻¹)

样本类型	样本数	最大值	最小值	均值	标准差	变异系数 /%
总体样本	100	1.71	0.80	1.20	0.27	22.01
建模集	75	1.71	0.80	1.22	0.26	21.64
验证集	25	1.59	0.87	1.16	0.27	22.80

1.2 分析方法

1.2.1 相关分析

相关分析(CA)是分别对原始光谱、一阶导数光谱的各个波段与全氮含量的相关计算,计算选定的土壤样本的全氮含量与每个波段的相关系数。在光谱波段范围的相关系数曲线中,选取相关系数曲线的波谷和波峰作为输入波段。相关系数的计算公式如式(1)

$$r_i = \frac{\text{Cov}(R, TN)}{\sqrt{D(R)} \sqrt{D(TN)}} = \frac{\sum_{n=1}^N (R_n - \bar{R}_i)(TN_n - \overline{TN})}{\sqrt{\sum_{n=1}^N (R_n - \bar{R}_i)^2} \sqrt{\sum_{n=1}^N (TN_n - \overline{TN})^2}} \quad (1)$$

式(1)中, r_i 为土壤全氮含量与光谱波段的单相关系数, n 为土壤样本的个数, i 为波段序号。 TN 为第 n 个土壤样品的全氮含量, \overline{TN} 为实测的 n 个土壤样品全氮含量的平均值, R_n 表示第 n 个土壤样本第 i 个波段的光谱反射率, \bar{R}_i 为 n 个土壤样本在 i 波段光谱反射率的平均值。

1.2.2 离散小波

离散小波变换的背景和原理可以被描述为有限长度信号和离散小波基的内积,其通过平移和缩放等运算功能可对光谱信号进行多尺度的细化分析^[15]。可利用低通与高通滤波器将信号分解为一系列高频和低频信号,高频信号(DC)为原始信息中的细微信号,低频信号(AC)为原始信息中的宏观信号^[16]。二者可从细微和宏观角度深入分析原始信息,同时还可以降低高光谱数据的维度。

1.3 建模方法

1.3.1 支持向量机

支持向量机(SVM)是由 Vapnik^[17]等提出的一种统计学理论,是用作分类和回归的机器学习方法,可以在有限数据下表现出良好的泛化能力和抗噪声能力^[18]。其目的是寻找数据之间的规律并依此来预测估计数据未来的变化趋势,在解决小样本、高位模式识别、大规模数据集的同时,可以有效地处理非线性问题。本研究,采用高斯核函数作为核函数,使用 GridSerachCV 函数发现最优参数,其中,惩罚系数代价=10 和 gamma=0.001。

1.3.2 K 邻近

K 邻近算法(KNN)是由 Cover 和 Hart 提出的,其原理是通过测量不同样本特征值之间的距离进行分类,同时该算法也可用于回归问题处理^[19]。邻近距离度量使用欧氏度量法,它定义于欧几里得空间中两点间直线距离,来衡量样本特征值之间的相似性。距离越远,即相似性越低。K 邻近算法采用 python 中的“sklearn.neighbors”包,交叉验证法用于确

定 K 值, $K=3$ 。

1.4 模型精度验证

采用决定系数(R^2)、均方根误差(RMSE)、相对分析误差(RPD)对估测模型进行评价,其计算公式分别为

$$R^2 = \frac{1 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (3)$$

$$\text{RPD} = \frac{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}{\text{RMSE}} \quad (4)$$

式中, y 和 \hat{y} 表示实测值和预测值, \bar{y} 表示实测值的平均值, n 表示样本数量。决定系数 R^2 越大,模型拟合越好。均方根误差 RMSE 越小,模型精度越高。RPD 可以反映模型的预测性能,当 $\text{RPD} < 1.4$ 时,模型的预测性能较差;当 $1.4 < \text{RPD} < 2.0$ 时,模型可用于定量预测;当 $2.0 < \text{RPD}$,可以实现更准确的定量预测。 R_c^2 、 R_v^2 、 RMSE_c 、 RMSE_v 、 RPD_c 、 RPD_v 分别代表建模和验证数据中的 R^2 、RMSE 和 RPD。

2 结果与讨论

2.1 土壤全氮含量与原始光谱及一阶导数光谱的相关性

已有研究表明,对光谱进行一阶导数变换能较好地消除背景信号或噪声、突出光谱曲线特征、去除或减弱其他因素的影响。将土壤全氮含量与原始光谱、一阶导数光谱分别进行相关性分析,见图 2。从图可以看出,原始光谱与土壤全氮的相关系数介于-0.48~0.59。经一阶导数变换后的光谱与土壤全氮的相关性更显著,峰值点清晰,相关系数介于-0.82~0.84 之间。综合比较可知,一阶导数光谱可以去除不同地背景噪声和基线漂移,解决重叠光谱特征,获取曲线中拐点和极值点,提高光谱与全氮的相关系数。其中,在 1 373 nm 处一阶导数光谱与全氮含量相关系数最高,为 0.84。选取一阶导数与全氮相关系数最高的 10 个峰值对应的波长作为全氮的特征波段,分别为 643、1 003、1 373、1 417、1 862、1 918、2 029、2 195、2 211 和 2 281 nm,用于

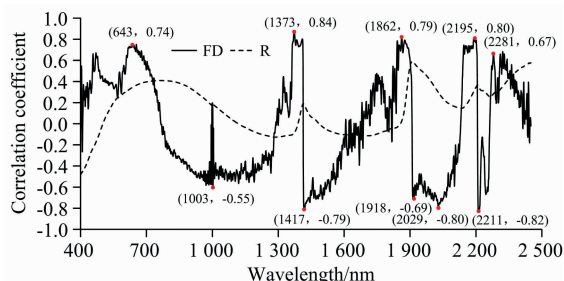


图 2 土壤全氮与原始光谱及一阶导数光谱的相关性
Fig. 2 Correlation between soil total nitrogen and original spectrum and first derivative spectrum

后续的建模分析。

2.2 基于离散小波的最佳母小波和分解层次选择

利用常见的 5 种母小波函数对一阶导数光谱进行分解,并用重构相似指标比较,分析不同母小波参数和分解层数的变化特征。随着分解水平增加,越来越多能够响应土壤的有用信息会被剔除,导致反射光谱的信息含量减少。DWT 系数的数量描述了数据压缩的程度,如表 2 所示,小波变换系

数的数量随着母小波和分解级别而变化。从 L_1 — L_{13} 趋于下降,下降稳定在 L_{11} 。5 个母小波中, sym8 压缩能力最强,而 coif5 最弱。例如,本研究中的波段总数为 2 051 个(400~2 450 nm),在分解级 11 后,母小波为 sym8 的 DWT 系数数量为 15,而 coif5 有 29。小波分解可以根据信号的长度和小波基长度,重复进行波长分解,直到达到最大尺度。

表 2 不同母小波和分解层数下的小波系数个数

Table 2 The number of wavelet coefficients under different mother wavelets and decomposition levels

函数	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	L_9	L_{10}	L_{11}	L_{12}	L_{13}
bior6.8	1 034	525	271	144	80	48	32	24	20	18	17	17	17
Db10	1 035	527	273	146	82	50	34	26	22	20	19	19	19
Rbio6.8	1 034	525	271	144	80	48	32	24	20	18	17	17	17
Sym8	1 033	524	269	142	78	46	30	22	18	16	15	15	15
Coif5	1 040	534	281	155	92	60	44	36	32	30	29	29	29

由于低频系数被认为是光谱全局信息的一个指标,分解级别 1—11 的每级变换光谱的低频系数被用来重构,以便发现低频系数如何描述反射光谱。反射光谱和重构信号之间的相关性如图 3 所示,相关系数从 L_4 一直下降到 L_{11} ,表明低频系数对光谱的解释和信号恢复能力从 L_4 到 L_{11} 逐渐下降;分解到 7 级后,相关性系数迅速下降到 0.6 以下, L_{11} 时一阶导数光谱相关系数为 0.4 左右;母小波 db10 比其他母小波函数更不稳定。考虑到数据压缩有效性、母小波的稳定性和保持光谱信息质量的能力,选择 L_1 — L_{11} 分解层的母小波 sym8 进行小波变换,用于分析与全氮含量的相关性。

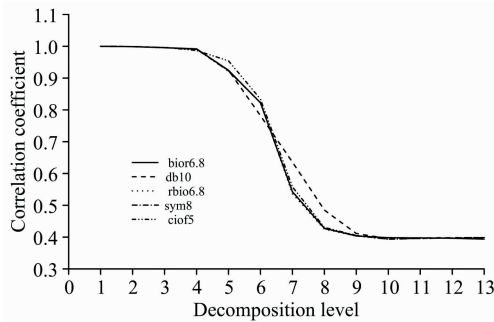


图 3 每个分解级不同母小波重构和一阶导数光谱之间的相关性

Fig. 3 Correlations between reconstructed signals and FD spectra for different mother wavelets at each decomposition level

2.3 土壤全氮含量估测模型建立

利用离散小波将土壤光谱数据分解为 11 个尺度的低频数据与高频数据,由于小波变换分析的高频系数代表光谱中的噪声或微小的吸收,图 3 显示,光谱信号和分解级 1—5 的 AC 重建信号之间的相关系数接近 1,这表明 L_1 — L_5 处的 DC 振幅非常小(接近 0),可以在信号信息内容没有重大损失的情况下去除,因此,利用各尺度低频数据作为输入结合机器学习方法构建模型。图 4 为不同分解级的低频系数结合 SVM 和 KNN 模型构建的土壤全氮含量建模和验证决定系数。

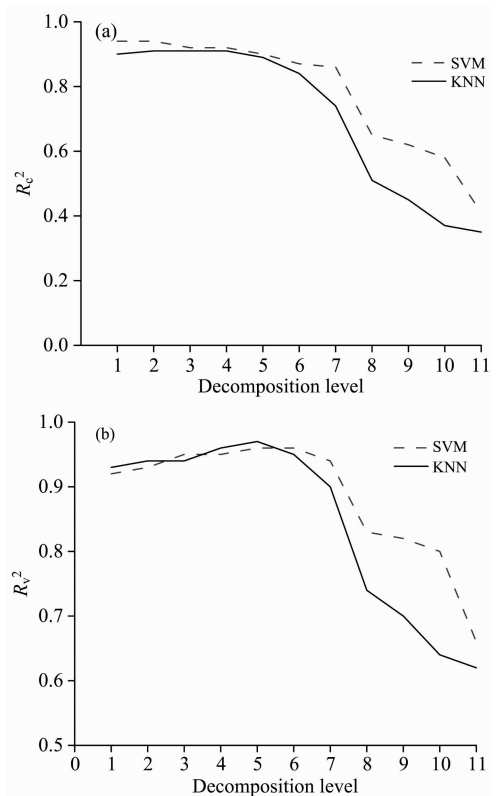


图 4 不同模型的决定系数与分解水平的关系

Fig. 4 Relationship between determination coefficient and decomposition level of different models

从图可以看出,随着分解程度的增加,各模型中的 R_c^2 的总体趋势减小, R_v^2 的总体趋势是先增加后减少,低频系数可以解释 90%~97% 的全氮含量在 L_1 — L_7 的变化。两种模型相比较,KNN 模型预测精度较好, R_v^2 在 L_4 、 L_5 和 L_6 分解级分别为 0.96、0.97 和 0.95。SVM 预测模型相比 KNN 模型精度较差,且在 7 级以后出现大幅度下滑,但在 L_4 、 L_5 和 L_6 分解级出现 0.95、0.96 和 0.96 较高的精度。表 3 显示了

不同分解级 R^2_V 、RMSE 和 RPD 的变化。图 5 展示了 L_5 分解级下 K 邻近模型实测值和预测值 1:1 散点图, 以分解级 L_5 构建的 KNN 模型表现最优。

表 3 低频系数不同分解级土壤全氮回归分析

Table 3 Regression analysis of soil total nitrogen with low frequency coefficient at different decomposition levels

分解级	SVM			KNN		
	R^2_V	RMSE _V	RPD _V	R^2_V	RMSE _V	RPD _V
AC ₁	0.92	0.08	3.03	0.93	0.07	4.18
AC ₂	0.93	0.07	3.28	0.94	0.06	4.63
AC ₃	0.95	0.06	3.74	0.94	0.06	4.08
AC ₄	0.95	0.05	3.76	0.96	0.06	4.09
AC ₅	0.96	0.05	3.78	0.97	0.05	4.30
AC ₆	0.96	0.06	2.97	0.95	0.06	3.98
AC ₇	0.94	0.09	2.21	0.90	0.09	2.38
AC ₈	0.83	0.15	1.95	0.74	0.17	2.06
AC ₉	0.82	0.16	1.85	0.70	0.18	1.68
AC ₁₀	0.80	0.17	1.70	0.64	0.20	1.58
AC ₁₁	0.66	0.19	1.62	0.62	0.21	1.54

0.90 以上, 全波段构建的模型验证决定系数均在 0.85 以上。其中, 以特征波段结合 KNN 建模及验证结果表现最好, 建模决定系数为 0.91, RMSE 为 $0.08 \text{ g} \cdot \text{kg}^{-1}$, RPD 为 3.30, 验证决定系数为 0.94, RMSE 为 $0.07 \text{ g} \cdot \text{kg}^{-1}$, RPD 为 3.72。

表 4 不同波段输入土壤全氮模型比较

Table 4 Comparison of soil total nitrogen models with different wave bands

模型	预处理	输入波段	建模			验证		
			R^2	RMSE	RPD	R^2_V	RMSE	RPD
SVM	相关分析	10	0.84	0.10	2.30	0.90	0.08	2.36
	全波段	2051	0.83	0.10	2.17	0.86	0.10	2.24
KNN	相关分析	10	0.91	0.08	3.30	0.94	0.07	3.72
	全波段	2051	0.86	0.09	2.61	0.89	0.08	2.95

2.4 不同建模结果比较

小波变换是信号处理的一种新型技术, 利用低通与高通滤波器将光谱数据分离为低频信息与高频信息。高频信息是光谱信号中包含的噪声和特殊信息, 低频信息是光谱信号全局行为的一种表达, 对应信号中的主趋势^[20]。已有研究表明土壤全氮的有益信息多集于低频信息内, 而高频信息内含噪声较多, 信噪比相对较低^[21]。本文采用离散小波处理土壤光谱, 并与传统降维方法进行对比建模。由表 3 和表 4 对比分析可知, 基于不同输入量构建的 SVM 模型中, 模型性能如下: 近似系数₁₋₇>相关分析>全波段>近似系数₈₋₁₁。基于近似系数的最优模型的预测精度比相关分析模型和全波段模型精度提高 6.7% 和 11.6%。基于不同输入量构建的 KNN 模型中, 模型性能如下: 近似系数₂₋₆>相关分析>近似系数₁>近似系数₇>全波段>近似系数₈₋₁₁。基于近似系数的最优模型的预测精度比相关分析模型和全波段模型精度提高 3.2% 和 9.0%, 表明离散小波可有效提升光谱对土壤全氮含量的敏感性, 压缩了光谱自变量并增加了各组分之间的区别, 进而提升模型的预测精度和稳定性, 这与其他学者的研究结果相似^[22-23]。

机器学习被应用到遥感数据的建模中, 但不同方法在构建模型时精度差别较大。Xu 等^[24]利用随机森林回归模型对滨海湿地的全氮含量估测并绘制空间分布地图, 模型决定系数为 0.65, 为大尺度全氮估测提供了技术指导。在本研究中, 使用小波变换提取的低频系数作为输入构建的模型要优于全波段和敏感波段作为输入构建的模型, 且整体上使用 K 邻近算法构建的模型精度要优于支持向量机, K 邻近算法可能是构建土壤全氮含量反演模型的可靠建模方法, 能获取更高的精度。本研究还存在不足之处, 如样本量较小, 模型的稳定性和预测精度需要进一步验证, 在后续研究会扩大取样范围, 获取更多的样本, 以进一步验证模型的可靠性。

3 结 论

以砂浆黑土为研究对象, 在系统分析土壤原始光谱和一

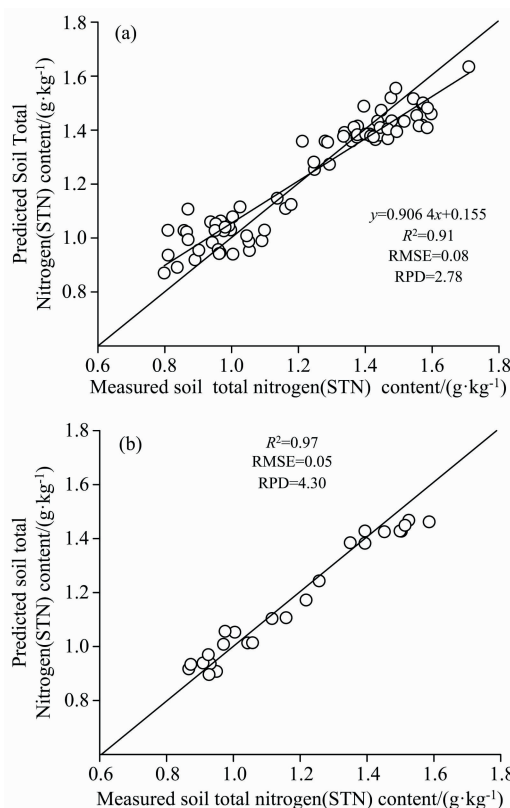


图 5 基于 L_5 近似系数的 K 邻近算法建模 (a) 及检验 (b)
Fig. 5 K proximity modeling (a) and testing (b) based on L_5 approximate coefficient

进一步将全波段和经过相关性分析后选择的 10 个特征波段分别作为输入, 结合支持向量机和 K 邻近建立土壤全氮模型, 见表 4。基于特征波段构建的模型验证决定系数均在

阶导数光谱的基础上,利用离散小波变换对一阶导数光谱进行小波分解和重构,提取特征参数,并结合支持向量机和 K 邻近算法建立机土壤全氮模型。结果表明:离散小波分析在保持原始光谱质量和降低光谱数据空间维度的基础上,可对土壤全氮光谱特征进行提取,利用 sym8 母小波对一阶导数

光谱进行压缩分解, L_5 层低频系数结合 K 邻近算法构建的土壤全氮模型表现最好,要优于全波段和以敏感波段作为输入构建的模型。研究结果可为利用高光谱遥感技术快速、实时、精确估测土壤全氮含量提供参考。

References

- [1] ZHANG Juan-juan, XI Lei, YANG Xiang-yang, et al(张娟娟, 席磊, 杨向阳, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2020, 36(17): 135.
- [2] Peng Y, Zhao L, Hu Y, et al. ISPRS International Journal of Geo-Information, 2019, 8(10): 437.
- [3] Guo L, Zhao C, Zhang H, et al. Geoderma, 2017, 285: 280.
- [4] Rinnan A, Van Den Berg F, Engelsen S B. TrAC-Trends in Analytical Chemistry, 2009, 28(10): 1201.
- [5] Cheng H, Shen R, Chen Y, et al. Geoderma, 2019, 336: 59.
- [6] SUN Xiao-xiang, ZHAO Xiao-min, XIE Wen(孙小香, 赵小敏, 谢文). Jiangsu Agricultural Sciences(江苏农业科学), 2018, 46(15): 287.
- [7] REN Hong-yan, SHI Xue-zheng, ZHUANG Da-fang, et al(任红艳, 史学正, 庄大方, 等). Remote Sensing Technology and Application(遥感技术与应用), 2012, 27(3): 372.
- [8] HU Gui-gui, YANG Fen-li, YANG Lian-an, et al(胡贵贵, 杨粉莉, 杨联安, 等). Arid Land Geography(干旱区地理), 2021, 44(4): 1114.
- [9] Hong Y, Chen Y, Yu L, et al. Remote Sensing, 2018, 10(3): 479.
- [10] Nijat Kasim, Rukeya Sawut, SHI Qing-dong, et al(尼加提·卡斯木, 茹克亚·萨吾提, 师庆东, 等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2018, 49(11): 155.
- [11] CHEN Hong-yan, ZHAO Geng-xing, LI Xi-can, et al(陈红艳, 赵庚星, 李希灿, 等). Chinese Journal of Applied Ecology(应用生态学报), 2011, 22(11): 2935.
- [12] WANG Yan-cang, YANG Xiu-feng, ZHAO Qi-chao, et al(王延仓, 杨秀峰, 赵起超, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(9): 2855.
- [13] GUO Yun-kai, XIE Xiao-feng, XIE Qiong, et al(郭云开, 谢晓峰, 谢琼, 等). Engineering of Surveying and Mapping(测绘工程), 2021, 30(3): 66.
- [14] Lin L, Wang Y, Teng J, et al. Sensors, 2015, 15(8): 17990.
- [15] Li F, Mistele B, Hu Y C, et al. Agricultural and Forest Meteorology, 2013, 180: 44.
- [16] Blackburn G A, Ferwerda J G. Remote Sensing of Environment, 2008, 112(4): 1614.
- [17] Vapnik V N. The Nature of Statistical Learning Theory, Springer, New York, 1995.
- [18] Ali I, Greifeneder F, Stamenkovic J, et al. Remote Sensing, 2015, 7(12): 16398.
- [19] Cover T, Hart P. IEEE Transactions on Information Theory, 1967, 13(1): 21.
- [20] Li F, Wang L, Liu J, et al. Remote Sensing, 2019, 11(11): 1331.
- [21] XIAO Yan, XIN Hong-bo, WANG Bin, et al(肖艳, 辛洪波, 王斌, 等). Remote Sensing for Natural Resources(国土资源遥感), 2021, 33(2): 33.
- [22] Ge Y, Thomasson J A. Transactions of the ASABE, 2006, 49(4): 1193.
- [23] Viscarra Rossel R A, Lark R M. European Journal of Soil Science, 2009, 60(3): 453.
- [24] Xu Y, Wang X, Bai J, et al. Ecological Indicators, 2020, 111: 106002.

Hyperspectral Feature Extraction and Estimation of Soil Total Nitrogen Based on Discrete Wavelet Transform

ZHANG Juan-juan^{1,2}, NIU Zhen^{1,2}, MA Xin-ming^{1,2}, WANG Jian¹, XU Chao-yue^{1,2}, SHI Lei^{1,2}, Bação Fernando³, SI Hai-ping^{1,2*}

1. Henan Agricultural University, College of Information and Management Science, Zhengzhou 450002, China

2. Henan Agricultural University, Collaborative Innovation Center of Henan Grain Crops, Zhengzhou 450002, China

3. Universidade Nova de Lisboa, NOVA Information Management School, Lisboa, 1070-312, Portugal

Abstract Soil total nitrogen is an important nutrient index. Hyperspectral technology is used to study and build a hyperspectral estimation model of total nitrogen content in Shajiang black soil, which provides a reference for crop fertilization and the development of precision agriculture. This paper attempts to study the feasibility of discrete wavelets to estimate soil total nitrogen content. Taking different wheat nitrogen fertilizer treatments in Shangshui County, Henan Province, as the experimental area, 100 samples of Shajiang black soil with a depth of 0~20 cm were collected. After the soil samples were air-dried in the dark and processed by grinding and screening, the spectra were collected in the darkroom of the laboratory. The total samples (100 sand ginger black soil) were divided into 75 modeling sets and 25 validation sets. The original spectrum was transformed by the first derivative, and the first derivative spectrum was analyzed by correlation analysis and discrete wavelet transform respectively. At the same time, the hyperspectral estimation model of soil total nitrogen content was constructed by combining the support vector machine and the k-nearest neighbor algorithm. The correlation between the single band of the original spectrum and the first derivative spectrum and soil total nitrogen were systematically analyzed. The results showed that after the first derivative transformation, the spectrum had a better correlation with soil total nitrogen, and the correlation coefficient reached 0.84 at 1 373 nm. The discrete wavelet algorithm selects the best mother wavelet and decomposition level of the first derivative spectrum. The results show that the wavelet coefficients decomposed by the Sym8 function can better reconstruct the spectral information of soil total nitrogen. Further, based on the low-frequency coefficients of decomposition layer L_1-L_{11} , the support vector regression and k-nearest neighbor regression estimation models of soil total nitrogen content were established respectively, and all the estimation models were compared. The model constructed by combining the low-frequency coefficients of decomposition layer L_5 with k-nearest neighbor is the best. The determination coefficient of modeling is 0.90, the root mean square deviation is $0.09 \text{ g} \cdot \text{kg}^{-1}$, and the relative analysis error is 3.78. The validation determination coefficient is 0.97, the root mean square deviation is $0.05 \text{ g} \cdot \text{kg}^{-1}$, and the relative analysis deviation is 4.30. At the same time, compared with the model constructed with the full band and the sensitive band selected after correlation analysis as input, the accuracy of the K-neighbor model is improved by 3.2% and 9%, and the accuracy of the support vector machine model is improved by 6.7% and 11.6%. The results show that the first derivative transform and discrete wavelet technology can effectively suppress the impact of noise, improve the estimation accuracy of soil total nitrogen content, reduce the dimension of spectral data, simplify the complexity of the model, and provide a reference for the accurate estimation of the total nitrogen content of Shajiang black soil.

Keywords Shajiang black soil; Total nitrogen; Hyperspectral; Discrete wavelet; K-neighbor

(Received Jun. 20, 2022; accepted Oct. 8, 2022)

* Corresponding author