第42卷，第9期
2022年9月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 9, pp2768-2773
September, 2022

南疆农田土壤全氮含量的中红外光谱反演模型

白子金[1]，彭 杰[1*]，罗德芳[1]，蔡海辉[1]，纪文君[2]，史 舟[3]，柳维扬[1]，殷彩云[1]

1. 塔里木大学农学院，新疆 阿拉尔 843300
2. 中国农业大学土地科学与技术学院，北京 100083
3. 浙江大学环境与资源学院，浙江 杭州 310058

摘 要 快速准确监测农田土壤全氮含量，可显著提高土壤肥力诊断与评价工作的效率。传统测定土壤全氮的方法存在耗时费力、成本高、环境污染等缺点，而基于光谱学原理的土壤全氮定量方法克服了传统测量的劣势。中红外(MIR)光谱相较于可见光-近红外(VNIR)光谱而言，具有更多的波段数和信息量，如何利用中红外光谱监测土壤全氮含量是具有重要应用前景的研究课题。为了探索中红外光谱对土壤全氮监测的可行性，以新疆南疆地区采集的246个农田土样为研究对象，以室内测定的全氮含量和中红外光谱反射率数据为数据源，分析了不同全氮含量土样的中红外光谱特征差异，以主成分分析法(PCA)和连续投影算法(SPA)对光谱数据进行降维，然后采用偏最小二乘回归(PLSR)、支持向量机(SVM)、随机森林(RF)和反向传播神经网络(BPNN)四种建模方法分别构建基于全波段和降维数据的土壤全氮含量定量反演模型。研究结果表明：(1)土壤在中红外波段光谱反射率随全氮含量的增加而增加，在3 620，2 520，1 620和1 420 cm^{-1}附近存在明显的吸收谷；将中红外光谱数据进行最大值归一化处理后，可明显提高土壤光谱反射率与全氮含量的相关性。(2)对比两种数据降维方法，PCA和SPA分别使模型变量数减少了99.8%和97.5%，但以PCA提取的8个主成分为自变量建立的模型预测精度总体要高于SPA对应的模型，因此以PCA提取的主成分建模更适于土壤全氮模型的构建。(3)在建模集中，PLSR和SVM模型以全波段建模精度最高，但建模变量数多，建模效率较低，而RF和BPNN模型分别以PCA和SPA降维后的数据建立的模型在保持精度相当的前提下，可显著提高建模效率；在预测集中，基于PCA降维数据的BPNN模型预测能力最高，R^2和RMSE分别为0.78和0.12 $g \cdot kg^{-1}$，RPD和RPIQ值分别为2.33和3.54，模型具备较好的预测能力。研究结果可为农田土壤全氮含量快速估测提供一定的参考价值。

关键词 中红外光谱；土壤全氮；反演模型；光谱数据降维

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)09-2768-06

引 言

土壤全氮是衡量土壤肥力的重要指标之一，在植物生长的许多生物化学活动以及土壤合理施肥过程中都起着重要作用。如何准确、快速测定农田土壤全氮含量，对作物生长发育和合理施肥有着重要的意义。传统测定土壤全氮含量的方法，不仅存在耗时、耗力、成本高、环境污染等缺点，而且在测定过程中一些化学试剂容易对人体造成危害，显然不能满足生产中大面积快速监测土壤全氮含量的需求。近年来，基于光谱学原理的土壤全氮定量方法在实时、快速、非破坏、低成本、无污染等方面表现出了独特的优势。

土壤的光谱反射特性是土壤养分含量、土壤类型、土壤质地等光谱特征的综合响应，其中土壤全氮含量也是影响土壤光谱反射特性的重要因素之一。迄今，在土壤属性预测方面，光谱技术主要包括VNIR和MIR波段。国内外学者利用光谱研究土壤全氮的报道甚多，但多采用可见光-近红外波段，主要针对的问题是在光谱数据预处理及定量反演模型的构建等方面，对光谱特征参数提取、敏感波段筛选以及建模方法等进行了深入研究。卢艳丽等[1]利用可见光特征波段构建的光谱指数建立的土壤全氮预测模型R^2达到了0.82；Pudełko A等[2]基于可见光-近红外光谱，采用PCA，PLSR

收稿日期：2021-04-23，**修订日期**：2021-10-04

基金项目：国家重点研发计划项目(2018YFE01070006)，国家自然科学基金项目(41361048，41061031)资助

作者简介：白子金，1997年生，塔里木大学农学院硕士研究生 e-mail：bzjzky@163.com

*通讯作者 e-mail：pjzky@163.com

和 ANN 方法建立了土壤有机碳和全氮预测模型，发现利用 PCA-ANN 组合模型的建模预测精度最高；孙宇乐等[3]通过相关分析结合小波去噪筛选的相关性强的特征波段，构建的神经网络全氮预测模型 R^2 为 0.75。20 世纪末，随着漫反射傅里叶变换技术的完善，中红外漫反射光谱技术在土壤属性预测方面有了一定的进展。McCarty 等[4]分别使用 NIR 和 MIR 漫反射光谱建立了土壤有机碳和无机碳预测模型，结果表明 MIR 光谱的预测能力强于 NIR，其原因是 MIR 波段与土壤碳相关性波段更多。近年来，国内学者利用中红外漫反射技术在土壤属性预测方面开展了一些研究。陈颂超等[5]研究了 VNIR，MIR 和 VNIR-MIR 三种不同波段光谱对土壤有机质的预测能力，发现以 MIR 波段建模效果最佳；Gomez 等[6]使用法国光谱库中红外反射光谱数据，建立的模型能够较好地预测土壤有机碳和无机碳。迄今为止，关于土壤全氮含量的光谱预测研究主要集中于可见光-近红外波段，基于中红外波段的相关研究甚少，而中红外光谱相较于可见光-近红外波段具有更多的波段数量，数据量更大，如何有效进行数据降维，降低数据冗余，提高计算效率，是中红外光谱应用于土壤属性预测所面临的现实问题。

综上，以新疆南疆地区的温宿县、阿瓦提县、和田县和新和县为研究区，以农田土壤全氮含量为研究因子，基于室内测定的土样中红外光谱反射率数据和土壤全氮含量数据，对原始光谱进行预处理，利用 PCA 和 SPA 对光谱数据进行降维，运用 PLSR，SVM，RF 和 BPNN 四种方法建立土壤全氮含量预测模型。旨在探明不同全氮含量土样的中红外光谱特征差异，明确土壤全氮在中红外光谱的敏感波段，筛选一种高效率的中红外数据降维方法，构建一种高精度的土壤全氮含量的中红外光谱预测模型，为中红外光谱在土壤全氮含量预测的应用提供一定的理论依据与技术参考。

1　实验部分

1.1　研究区概况

新疆位于中国西北部，以天山为界分为北疆和南疆两个区域。采样区位于南疆阿克苏地区的阿瓦提县(39°31′—40°50′N，79°45′—81°05′E)、温宿县(40°52′—42°15′N，79°28′—81°30′E)、新和县(40°45′—41°45′N，80°55′—82°43′E)以及和田地区的和田县(34°22′—38°27′N，78°00′—80°30′E)。阿克苏地区位于天山南麓和塔里木盆地北缘，地势北高南低，气候干燥，年均降水量 68 mm，年蒸发量 1 200～1 500 mm，光热资源丰富，昼夜温差大，属暖温带干旱型气候。和田地区南抵昆仑山与藏北高原交界，北临塔里木盆地，地势北高南低，气候极其干燥，年均降水量 35 mm，年蒸发量 2 480 mm，光照充足，热量丰富，昼夜温差大，属暖温带极端干旱荒漠气候。两地主要土壤类型有棕漠土、灌淤土、水稻土和盐土，土壤贫瘠，盐渍化严重，主要种植棉花、玉米、小麦、果树等农作物。

1.2　土壤样品采集与化学测定

采用网格法从四个地区共采集 246 个表层(0～20 cm)土壤样品，土样采集地点、采集数量和土壤类型分别为阿瓦提县 60 个灌淤土土样，和田县 44 个水稻土土样，温宿县 97 个水稻土土样，新和县 45 个盐土土样。每个采样点采集大约 1 kg 的样品储存于自封袋中，贴标签记录编号。采样点的地理坐标由手持设备全球定位系统(GPS)记录，位置误差小于 5 m。样品带回室内经自然风干，去除杂草、砾石及动植物残骸等杂质，经研磨混匀后用四分法将其分成两份，一份过 2 mm 筛，用于室内光谱测定，一份过 0.25 mm 筛，用于室内全氮测定。

1.3　土壤光谱数据采集及预处理

MIR 光谱测量使用 Agilent Technologies(美国)公司生产的 Agilent 4300 手持式 FTIR 光谱仪，测量光谱范围为 4 000～650 cm^{-1}，采样间隔为 0.47 cm^{-1}，光谱分辨率为 4 cm^{-1}，每个频谱扫描 32 次。测量前将土壤样品在 45 ℃烘干 24 h，去除土壤样品中的水分，避免水分对光谱测定的影响，之后再进行 MIR 光谱测量。每次测量之前需进行白板校正，每个样本测量 10 次光谱，算数平均后得到该土样的实际反射率光谱数据。

将 MIR 光谱去除边缘噪声较大波段，保留 4 000～800 cm^{-1} 波段范围光谱数据。采用 Savitzky-Golay (SG) 平滑法对光谱数据先进行平滑处理，然后进行最大值归一化(maximum normalization，MAN)预处理。

1.4　光谱数据降维处理

采用主成分分析法(principal component analysis，PCA)和连续投影算法(successive projections algorithm，SPA)分别对 MIR 光谱数据进行降维。

PCA[7]是一种把多个变量转化为少数几个互相独立并且包含原来指标大部分信息变量的多元统计分析方法，是将光谱包含的大量信息，通过线性变换保留方差大、包含信息量多的组分，舍弃信息量少的组分，从而对数据进行降维；设置累计贡献率为 95%，根据累计贡献率计算最终降维度。PCA 在 The Unscrambler X10.5.1 中实现。

SPA 是一种使矢量空间共线性最小化的前向变量选取算法，在有效信息获取和去除众多波段之间共线性影响的研究中取得了较好的效果，可以极大减少数据量，有效提高运算效率和模型精度，缩减数据集建模时间，具有简便、快速等优点，SPA 以均方根误差(root mean square error，RMSE)为评价指标，以 RMSE 最小值下的波长个数确定敏感波段数[8]。连续投影算法在 MATLAB R2019a 中完成。

1.5　建模方法及模型评价指标

建模方法选用偏最小二乘回归(partial least squares regression，PLSR)、支持向量机(support vector machines，SVM)、随机森林(random forest，RF)和反向传播神经网络(back propagation neural network，BPNN)。PLSR 方法借鉴了主成分分析、典型相关分析和普通多元线性回归三种分析方法的优点，较好的解决了样本数少于变量数等的问题。SVM 是一种基于统计学习理论的结构化学习算法，用来研究有限样本预测的智能学习方法，在处理分类与回归问题时常常被用到，尤其是在对小样本非线性与高维模式分类处理时，更能展现出其最优的性能[9]。PLSR 和 SVM 模型构建在 The Unscrambler X10.5.1 完成。随机森林(RF)是一种较新

的数据挖掘模型[10]，具有运算速度快、稳定性高、数据适应能力强、在处理大数据集时预测精度高，且不易产生过拟合等优势，RF 模型构建在 R 语言中实现。BPNN 是一种多向层感知的前馈式神经网络[11]，具有模型结构简单、误差小、运行速度快等优势，由于其强大的学习能力已被广泛用于土壤光谱建模分析中，BPNN 模型构建在 MATLAB R2019a 中完成。

选取土壤全氮含量实测值与预测值的均方根误差 RMSE、决定系数(determination coefficient，R^2)、相对分析误差(relative percent deviation，RPD)以及样本观测值三四分位数 Q3 与一四分位数 Q1 之差与 RMSE 的比值(RPIQ)四个参数验证模型预测精度。R^2 表示预测值与实测值之间的拟合程度，R^2 越大，说明模型预测结果越出色；RMSE 表示样本的实测值与预测值的偏离程度，RMSE 越小，说明预测值越接近实测值；RPD 是标准差与均方根误差的比值，证明模型的预测能力，根据 Chang 等[12]对 RPD 的划分等级，当预测模型 RPD＞2 时，表示模型有较好的估测能力；当 1.4＜RPD＜2.0 时，表示模型可以对样品含量进行粗略估测；当 RPD＜1.4 时，表示模型预测能力很差，无法对样品含量进行估测；RPIQ 通常被用来更好的表示非正态变量，RPIQ 越高，模型精度越好。

2 结果与讨论

2.1 土壤全氮描述性统计

将全部数据根据全氮含量由低到高排序，采用固定间距抽样，按 2∶1 比例划分成建模集和预测集，其中 164 个样本用于建模，剩余 82 个样本用于模型预测，表 1 为 246 个土样土壤全氮含量的描述性统计结果。样本总体全氮含量范围为 0.07～1.66 g·kg^{-1}，其中建模集的样本全氮含量范围为 0.07～1.66 g·kg^{-1}，预测集的样本全氮含量范围为 0.17～1.50 g·kg^{-1}，建模集的全氮含量范围完全覆盖了预测集的含量范围，可确保预测集土样的全氮含量不会超出预测模型的量程。建模和预测集变异系数在 40%左右，属于中等变异。

表 1 土壤样本全氮含量统计结果

Table 1 Statistical results of total nitrogen content in soil samples

样本集	数目	最小值/(g·kg^{-1})	最大值/(g·kg^{-1})	平均值/(g·kg^{-1})	标准差/(g·kg^{-1})	变异系数/%
建模集	164	0.07	1.66	0.68	0.28	41
预测集	82	0.17	1.50	0.68	0.27	40
总体	246	0.07	1.66	0.68	0.28	41

2.2 土壤全氮中红外光谱特性描述

根据土样全氮含量高低，将全氮含量分为四个等级，分别为等级Ⅰ(＜0.5 g·kg^{-1})、等级Ⅱ(0.5≤TN＜1.0 g·kg^{-1})、等级Ⅲ(1.0≤TN＜1.5 g·kg^{-1})和等级Ⅳ(≥1.5 g·kg^{-1})。图 1 为每个等级的 MIR 光谱均值曲线。由图 1 可知，MIR 光谱反射曲线具有很多的反射峰和吸收谷，不同全氮含量土样光谱曲线形态基本一致，但在某些特征波段的吸收谷深度和反射峰高度存在一定差异，如 2 200 和 2 600 cm^{-1}附近

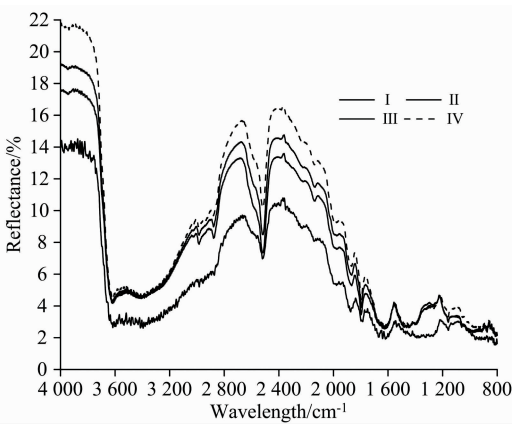

图 1 不同全氮含量土样的中红外光谱反射特性

Fig. 1 Mid-infrared spectral reflectance characteristics of soil samples with different total nitrogen contents

的反射峰高度呈现随含氮量增加而升高的趋势，2 500 cm^{-1}附近的吸收谷深度也呈现出相同的趋势。反射率变化范围在 0～22%之间，在全波段范围内均表现出土壤全氮含量越高反射率越高的趋势，在 3 620，2 520，1 620 和 1 420 cm^{-1}附近有明显的吸收特征。反射率在 3 900 cm^{-1}处达到最大，从 3 900～3 600 cm^{-1}反射率下降速度较快，且在此波段范围内反射率出现了交叉现象，可能是由于土壤反射率数值接近，平均之后差异较小；从 3 600～2 680 cm^{-1}，反射率呈增加的趋势，随后在 2 500 cm^{-1}处呈现出下降的趋势，在 2 500～2 400 cm^{-1}又呈增加趋势，随后至1 600 cm^{-1}缓慢下降，在 1 600～1 200 cm^{-1}再次增加，最后呈持续波动下降状态。

2.3 土壤全氮含量与中红外光谱相关性分析

将 MIR 原始光谱数据与最大值归一化处理后的光谱数据分别和土壤全氮含量做相关性分析，相关系数曲线见图 2。由图 2 可知，原始光谱曲线在全波段的相关性较差，相关系数曲线起伏较小，且多呈正相关态势，达到显著性水平的波段仅有 4 000～3 720，2 670～1 740 和 1 140～980 cm^{-1}，相关系数最高值在 2 516 cm^{-1}左右，相关系数仅为 0.31。将光谱反射率经最大值归一化处理后，相关系数曲线较原始光谱起伏较大，与土壤全氮含量的相关系数有了显著提高，相关系数在－0.67～0.67 之间，达到显著性相关的波段分布范围更广，主要集中在 3 800～2 650，2 600～2 470，2 130～2 020，1 740～1 170 和 970～800 cm^{-1}等波段内，全氮含量相关性在 2 732 cm^{-1}波段处达到最高，相关系数达到－0.67。上述分析说明对中红外光谱反射率值进行最大值归一

化可以有效地扩大一些细小的光谱特征，更能反映土壤全氮含量的变化特征。

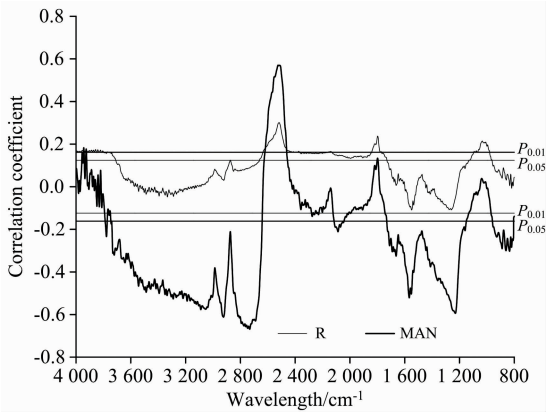

图 2　土壤全氮含量与中红外光谱的相关性分析

Fig. 2　Correlation analysis of soil total nitrogen content and mid-infrared spectral data

2.4　不同建模方法精度对比

利用 PLSR，SVM，RF 和 BPNN 四种建模方法结合中红外光谱数据以及经 PCA 和 SPA 降维后的数据分别和土壤全氮含量进行模型构建与验证(见表 2)。由表 2 可知，基于全波段光谱建模时，四种模型建模效果均较好(R^2＞0.77)，其中 PLSR，RF 和 BPNN 模型可以较好的估测土壤全氮含量(R^2＞0.75，RPD＞2.0，RPIQ＞3.24)，而 SVM 模型只能粗略估测土壤全氮含量(RPD＜2.0)。以 PCA 提取的主成分为自变量建模时，四种模型建模效果均非常出色(R^2＞0.75)，但只有 BPNN 和 RF 模型可以对土壤全氮含量进行较好的估测(RPD＞2.0，RPIQ＞3.03)，而 SVM 和 PLSR 模型只能对土壤全氮含量进行粗略估测(RPD＜2.0)。以 SPA 筛选的特征波段为自变量建立的模型较前两者效果较差(R^2＞0.69)，其中 BPNN 模型精度最高，可以较好估测全氮含量(RPD＝2.18，RPIQ＝3.41)，PLSR 和 RF 模型只能粗略估测全氮含量(RPD＜2.0)，而 SVM 模型基本无法对全氮含量进行估测(RPD＜1.4)。

表 2　土壤全氮含量不同模型精度对比

Table 2　Accuracy comparison of different models for soil total nitrogen content

建模因子	模型	建模集				预测集			
		R^2	RMSE	RPD	RPIQ	R^2	RMSE	RPD	RPIQ
全波段	PLSR	0.79	0.13	2.17	3.22	0.77	0.13	2.08	3.30
	SVM	0.82	0.12	2.26	3.36	0.73	0.15	1.86	2.88
	RF	0.77	0.14	2.12	2.92	0.75	0.13	2.09	3.24
	BPNN	0.81	0.12	2.31	3.43	0.75	0.12	2.12	3.28
PCA	PLSR	0.76	0.14	2.06	3.06	0.72	0.15	1.84	2.85
	SVM	0.81	0.12	2.24	3.33	0.75	0.14	1.95	3.01
	RF	0.75	0.12	2.14	3.37	0.74	0.15	2.05	3.03
	BPNN	0.84	0.11	2.49	3.70	0.78	0.12	2.33	3.54
SPA	PLSR	0.75	0.14	2.00	2.98	0.57	0.18	1.50	2.31
	SVM	0.69	0.16	1.69	2.52	0.51	0.20	1.39	2.15
	RF	0.74	0.14	2.02	2.96	0.71	0.16	1.77	2.68
	BPNN	0.89	0.10	2.75	4.10	0.76	0.13	2.18	3.41

从数据降维方法看，基于主成分建立的模型与全波段建立的模型相比，模型建模和预测精度均有所变化，PLSR 模型预测集 R^2，RPD 和 RPIQ 分别降低了 0.05，0.24 和 0.45，模型预测精度大幅降低；SVM 模型建模集 R^2 虽然下降了 0.01，但预测集 R^2，RPD 和 RPIQ 分别上升 0.02，0.09 和 0.13，模型精度略有提升；RF 模型预测精度变化较小；BPNN 模型建模和预测精度均有提高，其建模集 R^2，RPD 和 RPIQ 分别上升了 0.03，0.18 和 0.27，预测集 R^2，RPD 和 RPIQ 也分别上升了 0.03，0.21 和 0.26。基于连续投影算法提取波段建模与全波段建模相比，PLSR，SVM 和 RF 模型精度明显降低，其预测集 R^2 分别下降了 0.20，0.22 和 0.04，RPD 分别下降了 0.58，0.47 和 0.32，RPIQ 分别下降了 0.99，0.73 和 0.56；BPNN 模型精度略有提高，其建模集 R^2，RPD 和 RPIQ 分别上升了 0.08，0.44 和 0.67，预测集 R^2，RPD 和 RPIQ 分别上升了 0.01，0.06 和 0.13。对比降维算法和建模方法对建模结果的影响，两种降维算法均减少了建模数据量，改变了原始数据的结构，加快了建模效率，基于 PCA 建立的各种模型总体精度要高于基于 SPA 对应模型。对比 12 种模型精度发现，基于 PCA 建立的 BPNN 预测模型效果最佳(R^2＝0.78，RPD＝2.33，RPIQ＝3.54)，其在提高建模效率同时仍可保持与全波段建模的同等精度，可以很好地对土壤全氮含量进行预测，且不易受变量数的影响。

关于土壤全氮的高光谱反演已有较多研究，其中基于光谱波段选择、光谱预处理、敏感波段筛选、建模算法等化学计量方法逐步改善了模型的精度和稳定性。本工作采用中红外波段光谱数据对土壤全氮含量进行估测，由于中红外光谱波段数要远超于可见光-近红外波段数，其光谱数据载有土壤样品的结构和组成信息，在建模时有其独特的优势。基于 MIR 光谱建立的最优模型预测集 R^2 为 0.78，较杨梅花等[13]采集的 120 个样本使用 VNIR 光谱建立的 PCA-BPN 预测模型 R^2 高 0.26，说明中红外波段光谱较可见光-近红外建模具有一定的优势。光谱变量筛选可以剔除与待测目标无关的变

量，降低数据冗余，实现模型简化，提高建模效率，Yang[14]等研究表明基于田间较小范围的土壤光谱采用无信息变量消除(UVE)结合 SPA 选择特征波段建立的模型预测精度和全波段建立的模型精度相当；本工作选用的 PCA 和 SPA 筛选的变量建模对比发现，基于 PCA 建立的模型总体建模精度要优于 SPA，可能是由于 SPA 更适用于小样本试验[15]，而本研究的样本数和使用的中红外波段光谱数据量较大，导致 SPA 筛选变量建模效果较差。

在模型构建方面，土壤养分含量光谱估测模型可分为线性模型和非线性模型，合理选择建模方法是提高反演精度和效率的重要步骤。本工作建立了经典线性回归 PLSR 模型和非线性 SVM、RF 和 BPNN 模型。从模型预测能力看，基于不同建模变量的 BPNN 模型效果最佳，原因是光谱变量与土壤属性之间往往存在一定的非线性关系，BPNN 在处理非线性关系的样本数据时，具有很高的非线性和容错性，能够进行复杂的逻辑操作；PLSR 模型精度位于 BPNN 和 SVM 之间，PLSR 在进行全波段建模时，模型精度高于非线性的 SVM 模型，接近于 RF 和 BPNN 模型，在于 PLSR 可有效解决变量之间的多重共线性问题，但也存在建模效率低的问题；只有 SVM 模型建模总体效果较差，存在建模集的 R^2 较高而预测集 R^2 较低的问题，可能是由于 SVM 模型没有找到适当的核函数，导致数据没有进行适当分类，还可能在于 SVM 对大规模训练样本难以实施。目前，应用 MIR 技术进行土壤全氮分析较少，对于提高模型精度主要针对光谱预处理方法和建模方法的选择，很难提高建模效率，因此应引入一些无信息变量消除(UVE)、遗传算法(GA)等光谱特征变量选取方法，进一步优化和提高模型预测能力，探索一些新的光谱处理方法在土壤全氮分析中的应用。

3 结 论

筛选土壤全氮的光谱响应波段或对光谱数据降维是简化模型、提高模型预测能力和建模效率的关键技术。本研究以新疆南疆四县的农田表层土样为研究对象，测定土样的 MIR 波段光谱数据和全氮含量，分析了土壤在 MIR 波段的光谱响应曲线，采用 PCA 和 SPA 两种方法对光谱数据降维，并结合 PLSR，SVM，RF 及 BPNN 模型分别建立基于全波段和降维数据的土壤全氮含量预测模型，筛选最优模型，得出以下结论：

(1)土壤 MIR 反射率随全氮含量的增加而增加，在 3 620，2 520，1 620 和 1 420 cm^{-1}附近存在明显的吸收谷。将 MIR 光谱数据进行最大值归一化处理，可明显提高与土壤全氮含量的相关性，有利于提高模型精度。

(2)对比两种降维方法，基于 PCA 的 PLSR 模型精度较全波段有所下降，SVM 和 BPNN 模型精度有一定程度的提高，而 RF 模型没有明显的变化，但 PCA 使模型变量数减少了 99.8%，降低了模型复杂度，提高了建模效率。基于 SPA 的 PLSR，SVM 和 RF 模型较全波段精度均有不同程度的降低，而 BPNN 模型精度有明显的提高，预测集 R^2，RPD 和 RPIQ 分别提高了 0.01，0.06 和 0.13，SPA 使得模型变量数减少了 97.5%，减少了模型变量数，但模型总体精度较全波段建模有所下降。

(3)对比四种方法建模，BPNN 模型精度明显高于其余三种模型，其在处理多维特征数据和抗噪声能力方面具有独特的优势，基于 PCA 的 BPNN 模型精度和拟合度和稳定性最高，较全波段 BPNN 模型 R^2，RPD 和 RPIQ 分别提高到了 0.78，2.33 和 3.54，在减小建模复杂度的同时提高了建模效率，可为快速准确监测研究区农田土壤全氮含量提供新的思路。

References

[1] LU Yan-li, BAI You-lu, WANG Lei, et al(卢艳丽，白由路，王 磊，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2010, 26(1): 256.

[2] Pudełko A, Chodak M. Geoderma, 2020, 368: 114306.

[3] SUN Yu-le, QU Zhong-yi, LIU Quan-ming(孙宇乐，屈忠义，刘全明). Journal of Irrigation and Drainage(灌溉排水学报), 2020, 39(12): 120.

[4] McCarty G W, Reeves J B, Reeves V B, et al. Soil Science Society of America Journal, 2002, 66(2): 640.

[5] CHEN Song-chao, PENG Jie, JI Wen-jun, et al(陈颂超，彭 杰，纪文君，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(6): 1712.

[6] Gomez C, Chevallier T, Moulin P, et al. Geoderma, 2020, 375: 114469.

[7] Anowar F, Sadaoui S, Selim B. Computer Science Review, 2021, 40: 100378.

[8] Li J, Zhang H, Zhan B, et al. Infrared Physics & Technology, 2020, 104: 103154.

[9] 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.

[10] GUO Peng-tao, LI Mao-fen, LUO Wei, et al(郭澎涛，李茂芬，罗 微，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2015, 31(5): 194.

[11] QI Hai-jun, LI Shao-wen, KARNIELI Arnon, et al(齐海军，李绍稳，KARNIELI Arnon，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2018, 49(2): 166.

[12] Chang C W, Laird D A, Mausbach M J, et al. Soil Science Society of America Journal, 2001, 65(2): 480.
[13] YANG Mei-hua, ZHAO Xiao-min(杨梅花, 赵小敏). Scientia Agricultura Sinica(中国农业科学), 2014, 47(12): 2374.
[14] Yang H, Kuang B, Mouazen A M. European Journal of Soil Science, 2012, 63(3): 410.
[15] ZHANG Chao, LIU Yong-mei, SUN Ya-nan, et al(张　超, 刘咏梅, 孙亚楠, 等). Chinese Journal of Applied Ecology(应用生态学报), 2018, 29(9): 2835.

A Mid-Infrared Spectral Inversion Model for Total Nitrogen Content of Farmland Soil in Southern Xinjiang

BAI Zi-jin[1], PENG Jie[1*], LUO De-fang[1], CAI Hai-hui[1], JI Wen-jun[2], SHI Zhou[3], LIU Wei-yang[1], YIN Cai-yun[1]
1. College of Agriculture, Tarim University, Alar 843300, China
2. College of Land Science and Technology, China Agricultural University, Beijing 100083, China
3. College of Environment and Resources, Zhejiang University, Hangzhou 310058, China

Abstract Monitoring total nitrogen content rapidly and accurately in farmland soils can significantly improve the efficiency of soil fertility diagnosis and evaluation efficiency. Traditional methods for measuring total soil nitrogen have disadvantages such as time-consuming, high cost, and environmental pollution, while the quantitative method of total soil nitrogen based on spectroscopic principles overcomes the disadvantages of traditional measurements. Mid-infrared (MIR) spectroscopy has morebands and information than visible-near infrared (VNIR) spectroscopy, and how to use MIR spectroscopy to monitor soil total nitrogen content has not been systematically studied in China. In order to explore the feasibility of mid-infrared spectroscopy for soil total nitrogen monitoring, we selected 246 farmland soil samples in the southern Xinjiang region as the research objects and used total nitrogen content and mid-infrared spectral reflectance data measured in the laboratory to analyze the differences mid-infrared spectral characteristics of soil samples with different total nitrogen content. Firstly, the dimension of spectral data was reduced by the principal component analysis (PCA) and successive projections algorithm (SPA) and then used four modeling methods including the partial least squares regression (PLSR), support vector machine (SVM), random forest (RF) and back propagation neural network (BPNN) to construct the quantitative inversion model of soil total nitrogen content based on the full-band and dimension-reduced data, respectively. The results showed that: (1) the spectral reflectance of soil in the mid-infrared band increased with the increase of total nitrogen content with a significant absorption valley near 3 620, 2 520, 1 620 and 1 420 cm^{-1}. The correlation between soil spectral reflectance and total nitrogen content could improve significantly after the maximum normalization of mid-infrared spectral data. (2) Comparing the two data dimension reduction methods, PCA and SPA reduced the number of model variables by 99.8% and 97.5% respectively. However, the prediction accuracy of the model established with the eight principal components extracted by PCA as independent variables were generally higher than that of the corresponding model of SPA. Therefore, the modeling with the principal components extracted by PCA was more suitable for constructing the soil total nitrogen model. (3) In the modeling set, the PLSR and SVM models had the highest accuracy in full-band modeling, but the modeling efficiency was low due to alarge number of modeling variables. However, based on the RF and BPNN models, using the data after dimension reduction by PCA and SPA for modeling respectively, while maintaining the comparative accuracy, the modeling efficiency can be significantly improved. In the prediction set, the BPNN model based on PCA dimension-reduced data had the highest prediction ability, with R^2 and RMSE of 0.78 and 0.12 $g \cdot kg^{-1}$, RPD and RPIQ of 2.33 and 3.54, respectively, indicating that the model had great prediction ability. The study results can provide some reference values for the rapid estimation of total nitrogen content in farmland soil.

Keywords Mid-infrared spectrum; Soil total nitrogen; Inversion model; Dimension reduction of spectral data

(Received Apr. 23, 2021; accepted Oct. 4, 2021)

* Corresponding author