DOI:10.11766/trxb201309220429

应用田间行走式红外光谱进行土壤碳含量估测研究*

沈掌泉[1,2†] 叶领宾[1,2] 单英杰[3]

(1 浙江大学农业遥感与信息技术应用研究所,杭州 310058)

(2 浙江省农业遥感与信息技术重点研究实验室,杭州 310058)

(3 浙江省土肥站,杭州 310020)

摘 要 对应用田间行走式设备获取的土壤红外光谱数据,通过特征变换和特征选择相结合,以提高所建立土壤碳校正模型的预测精度。首先应用独立成分分析(ICA)、主成分分析(PCA)和小波分析(WA)对土壤红外光谱数据进行特征变换,然后分别应用无信息变量消除法(UVE)、连续投影算法(SPA)、无信息变量消除结合连续投影算法(UVE-SPA)、基于遗传算法和偏最小二乘法的变量选择法(GA-PLS)来进行特征选择,基于所选择的特征建立了土壤碳校正模型。结果表明,通过 ICA 进行特征变换,然后进行特征选择,可以建立比直接对光谱数据进行波长选择精度更好的预测模型;而 WA 或 PCA 与特征选择方法结合,只能获得与对光谱数据直接进行波长选择相近的效果。因此,针对田间条件下通过行走式设备获得的光谱数据由于受复杂的环境条件下干扰多的情况,可以将 ICA 与特征选择方法结合起来对光谱数据进行特征变换和选择,以建立更可靠的土壤碳含量预测模型。

关键词 特征变换;特征选择;土壤碳;田间行走式测定;近红外光谱;偏最小二乘回归法

中图分类号 S123;TH744.1 **文献标识码** A

近红外光谱分析技术是近年来发展很快的一种分析方法,已经在农业及许多其他领域得到广泛应用[1-5]。精确农业需要详细、精确地了解土壤性质的空间分布和特性,对土壤性质获取的效率和精度提出了更高的要求;而光谱分析技术所具有的优势和特点,正好能满足精确农业的要求。因此,将光谱技术应用于土壤分析和数字化制图方面的研究,也受到了越来越多的关注和重视[5-8]。

由于近红外光谱主要是倍频和合频的吸收,光谱信息重叠严重。加上土壤是一个复杂的混合体,对土壤光谱的影响因素很多,因而在利用光谱测定来分析土壤性质时,大多为在实验室条件下对土壤样品进行风干、研磨等预处理后进行;近年来,逐步开展了田间条件下进行土壤性质测定的尝试,由于田间条件下土壤水分等变化更加明显,而且还受到杂质、测试环境变化等各种因素的影响,因此通过光谱测定来分析土壤性质,显得更加困难,因而通过对光谱数据的处理,剔除干扰,保留和增强与土壤性质相关的光谱信息,就显得尤为重要[6-8]。尽管通过波段比值、差值、差值归一化等手段可以改善光谱信息与土壤性质的相关性和提高建模精度[9-10],但这些手段无法充分利用丰富的光谱信息。

近红外光谱包含数百甚至上千个波段,在进行建模时,应用所有的波段来进行分析与建模,不但计算分析的复杂度大大增加,而且由于噪声和干扰的存在,反而影响了模型的精度。通过对光谱波段直接进行选择,可以提高所建模型的质量和预测精度[11-12],但通过光谱波段的选择,也必然会损失有用的光谱信息。近年来的研究发现,通过特征变换,可以降低光谱数据的维数、并将有用的信息集中到变换后的光谱特征中,同时降低了光谱变量之间的信息冗余。由于光谱信息是综合性的,经过特征变换后,有的特征与土壤性质有关,而其他特征则并不相关。因此,可以通过特征选择,识别并选择出与之相关的特征而剔除其余无关的特征,从而

* 国家科技支撑计划项目(2012BAH29B04)资助

† 通讯作者,E-mail:zhqshen@zju.edu.cn

作者简介:沈掌泉(1969—),男,浙江桐乡人,博士,副教授,主要从事农业遥感、计算机应用及土壤空间变异等研究。E-mail:zhqshen@zju.edu.cn

收稿日期:2013-09-22;收到修改稿日期:2014-02-28

提高所建校正模型的质量。特征选择在多变量校正分析与建模中的重要性得到越来越广泛的认同,已经成为光谱数据分析建模过程中的关键步骤[1-4,12-13]。

陈红艳等[6]先应用小波分析剔除土壤光谱数据噪声,然后建立了通过土壤光谱估测土壤有机质含量的模型;郑立华等通过小波分析剔除低频和高频的干扰信息,然后应用偏最小二乘法建立土壤有机质的预测模型[7];陈红艳等[8]在通过小波分析剔除低频和高频干扰信息后,应用相关分析选择光谱波段,然后建立土壤有机质的回归模型;Vohland等[14]应用遗传算法结合偏最小二乘法等建立基于可见-近红外光谱的土壤有机碳校正模型,发现对于独立的验证数据集而言,不同方法的预测精度差异不大。沈掌泉等[11]发现通过光谱波段选择可以提高土壤碳含量的预测精度。

通过田间行走式设备,可以快速、大范围获取土壤的光谱数据;但在自然的田间环境下,干扰因素众多,导致光谱信息处理和建模更加困难。随着田间行走式测定设备的逐步成熟和推广应用,可靠、有效地提取土壤信息,已成为一个迫切需要解决的问题。本文通过将特征变换与特征选择相结合,先应用光谱数据处理中常用的特征变换方法进行特征变换,然后应用常用的特征选择方法对光谱特征进行选择,来探索建立基于行走式设备测定的土壤近红外光谱的土壤碳含量测定的校正模型的有效方法,以研究大面积快速获取土壤信息的可行性。

1 材料与方法

1.1 研究区概况

研究区位于美国 Michigan 州 Kalamazoo 县的 Carr 农场,面积为 52 hm^2。土壤类型为 Kalamazoo 壤土;地块内地形变化较大,地势南高北低,高程相差达 10 m 以上(图 1),导致土壤水分的变化也较大,因此土壤碳含量的变化也较大。根据采样分析数据,表层土壤碳含量在 5.51 ~ 28.98 g kg^{-1},变异系数为 24.36%(表 1)。

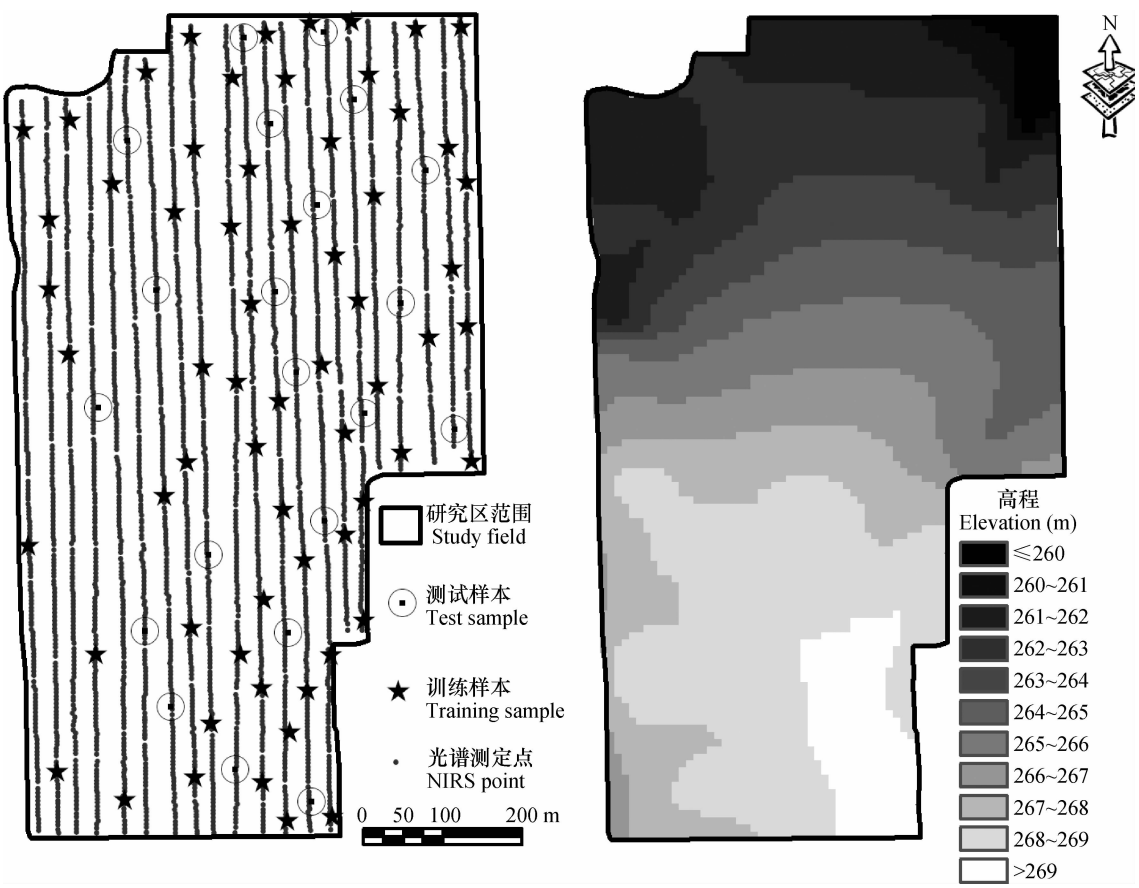


图 1 研究区土壤采样点、光谱数据测定点分布和高程图

Fig. 1 Distribution and contour map of the soil sampling sites and spectral data measuring sites in the study field

表 1　研究区土样的基本统计

Table 1　Statistics of soil samples in the study field

项目 Item	样本数 Number of samples	最小值 Minimum ($g\ kg^{-1}$)	最大值 Maximum ($g\ kg^{-1}$)	平均值 Mean ($g\ kg^{-1}$)	标准差 SD ($g\ kg^{-1}$)	变异系数 CV (%)
校正数据集 Calibration dataset	64	5.51	28.98	15.57	4.02	25.82
测试数据集 Test dataset	21	9.18	19.31	14.77	2.75	18.59
整个数据集 Whole dataset	85	5.51	28.98	15.38	3.75	24.36

1.2　光谱数据测定

由 Veris Technologies 公司的田间行走式光谱测定设备(VIS-NIR)进行光谱测量。测定以条带的方式进行,条带之间的距离为 25 m,测定点之间的距离为 5 m(图 1)。测定深度为 7.5 cm,光谱的波长范围为 920 ~ 1 718 nm,共 128 个波段,分辨率为 6.3 nm,测定的反射率应用倒数对数的方式转换为吸光率。测定时,田面处于裸露和干燥状态。以土壤采样点为中心,搜索出离其最近的光谱测定点,并以此光谱测定数据,作为该采样点土壤的光谱测定数据。

1.3　土壤样本采集与分析

沿光谱测定的条带,随机采集土壤样本 85 个。采样深度为 10 cm,土样经处理后,在 Carlo-Erba 2 碳氮分析仪中,用干烧法测定土壤的全碳含量。在考虑采样点空间均匀分布的情况下,将 85 个样点随机地分为独立的校正数据集和测试数据集,其基本统计特性见表 1。

1.4　光谱特征变换、选择及建模方法

本文用目前常用的主成分、独立成分和小波变换这 3 种方法对田间测定的光谱数据进行特征变换,然后对变换后的光谱特征应用无信息变量消除法、连续投影算法、无信息变量消除法结合连续投影算法及基于遗传算法和最小二乘法的变量选择法进行特征选择,最后基于选择出的特征,应用偏最小二乘回归法建立校正模型。各方法及本文中所采用的工具与参数如下:

主成分分析(Principle component analysis,PCA)的中心思想是降维,在损失很少信息的前提下,把多个指标转化为几个综合指标。PCA 被广泛应用于光谱数据分析,包括数据降维、特征提取与压缩、确定化学组分数、分类、聚类等。但 PCA 是一种全局性的分析方法,经变换后的主成分的物理意义一般也难于分析[13]。在本研究中,应用 SPSS 中相应的主成分分析方法来处理光谱数据,提取了累积贡献率达到 99.9% 的前 20 个主成分,以进行后续的分析和建模。

独立成分分析(Independent component analysis, ICA)起源于盲源信号分离,是近二十年发展起来的一种新的信号处理和数据分析方法,在信号处理、数据挖掘、特征提取等领域得到广泛的研究和应用。它分解出的各成分之间是相互独立的,具有很强的特征提取能力[15-16]。在本研究中,应用 FastICA 2.5 来进行光谱数据处理,通过分析后共获得 21 个独立成分。

小波分析(Wavelet analysis, WA)方法是在傅里叶分析的基础上发展起来的,它解决了傅里叶方法无法对局部时间信号的局部频谱特性进行分析的问题。在光谱数据处理方面,小波分析已经被应用于去噪、数据压缩和模型传递等,但主要应用于光谱数据的去噪。近来的一些研究表明,对光谱数据进行小波分析分解,然后以小波系数代替光谱数据进行选择并建模,可以取得很好的效果[6-8,17-19]。由于 Daubechies 小波函数在光谱数据去噪和特征提取方面比较有效[19],因此,本研究选择 2 阶的 Daubechies(Db2)小波函数对光谱数据进行分解,将分解后的小波系数作为光谱特征,用于进一步的处理和建模;小波分析通过 Matlab 的小波分析工具箱完成。

无信息变量消除法(Uninformative variable elimination, UVE)是通过人为地将一些随机变量加入到数据中,并以这些随机变量为参考,将那些在模型中不关键的变量加以剔除。通过 UVE 进行特征选择后建立的模型,能够更好地防止过配和提高预测精度。在本研究中,应用 CARS-PLS (Competitive adaptive reweighted sampling coupled with partial least

squares regression, Version 2.0) 工具箱中的 UVE-PLS 函数来进行运算[20]。

连续投影算法(Successive projection algorithm, SPA)是通过最小化多元线性回归中的共线性问题,在矢量空间中投影来获得最小共线性的变量子集,来进行变量的选择。尽管研究表明基于 SPA 特征选择所建立的模型比基于全波段的 PLSR 或 PCR 模型具有相当或更好的预测能力,但应用 SPA 进行特征选择也可能导致降低信噪比或在多变量分析建模中因变量过少的问题而影响模型的预测能力[1-3,13,20-22]。在本研究中,应用 GUI_SPA(Successive projections algorithm-graphical user interface)工具箱来进行有关 SPA 的运算。

Ye 等提出将 UVE 和 SPA 结合起来进行光谱特征选择,称为无信息变量消除法结合连续投影算法(UVE-SPA)[20]。与直接应用 SPA 相比,通过无信息变量的消除来加强光谱信息与预测值之间的联系,而且能使 SPA 更集中有效地进行变量选择,从而提高模型的预测能力[1-3,20-22]。

遗传算法(Genetic algorithm, GA)是一种通过模拟自然进化过程搜索最优解的方法。已被广泛地应用于组合优化、机器学习、信号处理、自适应控制和人工生命等领域。基于将特征选择问题作为优化来进行处理的想法,Leardi 等人提出并开发了 PLS-GA 工具箱(PLS-Genetic algorithm toolbox),将遗传算法和 PLS 结合起来,许多研究已证明是光谱数据特征选择的一种有效方法[23-25]。本研究应用 PLS-GA 工具箱进行运算,进化代数为 100,其他参数均采用工具箱的缺省值。

偏最小二乘回归法(Partial least squares regression,PLSR)是多元定量分析中一种常用的方法,被广泛应用于近红外、红外、拉曼、核磁和质谱等波谱定量分析模型的建立,已成为光谱分析中建立线性定量校正模型的通用方法。本研究在应用 PLSR 进行建模时,选择均值均匀化(Mean centering)为预处理方法,交叉检验的交叉数设为 3,采用顺序交替的方式来生成交叉检验的样本子集。所有 PLSR 建模,均由 Norgaard 等开发的 iToolbox[26] 中的 iPLS 工具来实现。

2 结果与讨论

2.1 光谱数据及变换后各光谱特征之间相关性分析

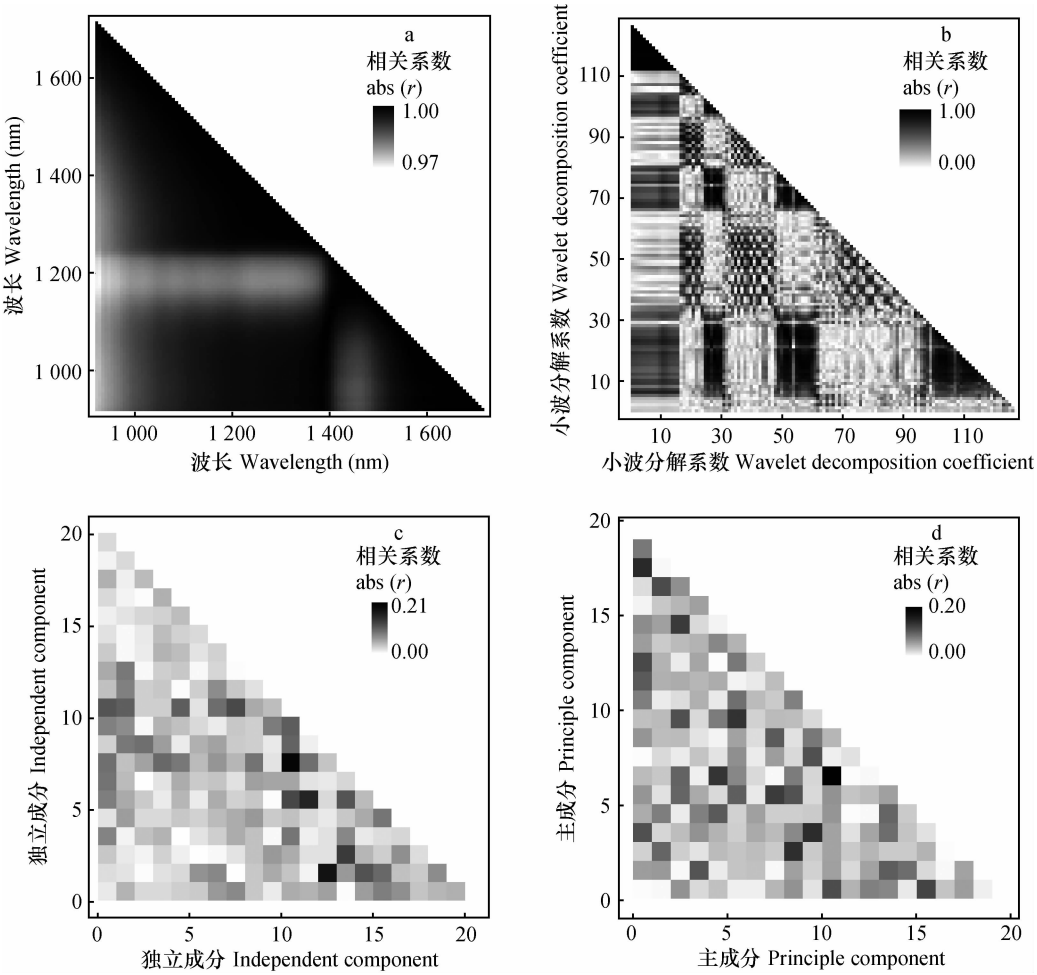
由于近红外光谱主要是倍频和合频的吸收,因此光谱信息重叠非常严重,这一点从图 2a 光谱波段之间的相关性可以明显地反映出来,从图 2 可以发现各光谱波段的相关性均在 0.97 以上,相关性非常高,说明各光谱波段之间信息的重叠程度很高。而在经过特征变换后,各特征之间的相关性,发生了明显的变化。经小波分解后,由于小波将光谱信号分解为低频和高频部分,因此小波系数之间的相关性呈现出明显的差异,低频部分相关性比较高,高频部分相关性比较低(图 2b)。而经独立成分变换和主成分变换后,各独立成分和主成分之间的相关性也比较低,均未达到显著水平(图 2c 和图 2d),说明经过 ICA 和 PCA 变换,光谱信息的重叠度明显减少。

2.2 光谱数据及变换后各光谱特征与土壤碳含量之间的相关性分析

图 3 为原始光谱变量及经特征变换后各光谱特征与土壤碳含量之间的相关性。从光谱变量与土壤碳含量之间的相关性来分析,可以发现,尽管原始光谱各变量与土壤碳含量之间的相关性存在明显差异,从 920 nm 开始,相关性逐渐提高,在 1 456 nm 达到 -0.209 8,然后逐渐降低,但均未达到显著水平;而经小波分解后,不同小波系数与土壤碳含量的相关性发生了较大的变化,部分系数达到了极显著水平,最高的相关系数达到了 0.579 1,说明通过小波变换,部分与土壤碳含量有关的光谱信息,得到了集聚与加强;通过 ICA 变换后的各独立成分,也表现出差异,其中第 2、7、12、16 独立成分达到了极显著水平;而通过 PCA 变换后的各主成分,同样也呈现出差异,其中第 2、4 主成分达到极显著相关,第 17 主成分也达到了显著相关。因此,通过特征变换,将与土壤碳含量相关的光谱信息,集中到部分光谱特征中,从而更加有利于选择与土壤碳含量相关的特征来建立有效的校正模型。

2.3 基于不同特征选择方法的模型的预测精度比较分析

表 2 为原始光谱变量及经特征变换后各光谱特征,在不同特征选择方法进行特征选择后,应用 PLSR 建立的校正模型对独立的测试数据集的预测精度。从中可以发现,在应用全部变量建立的模型的情况下,与基于原始光谱变量所建立的模型相比,经特征变换后所建立模型的预测精度,均非常接近。RMSE 在 2.08 ~ 2.15 g kg^{-1} 之间,决定系数在 0.436 4 ~ 0.468 7 之间,说明经过特征变换,尽管光谱信息的重叠与冗余得到消除,但由于光谱信息



a. 光谱数据;b. 小波分析;c. 独立成分分析;d. 主成分分析

a. Spectral;b. Wavelet;c. ICA;d. PCA

图 2　光谱数据及经特征变换后各光谱特征之间的相关性分析

Fig. 2　Correlation analyses of NIR spectra and spectral features after feature transformation

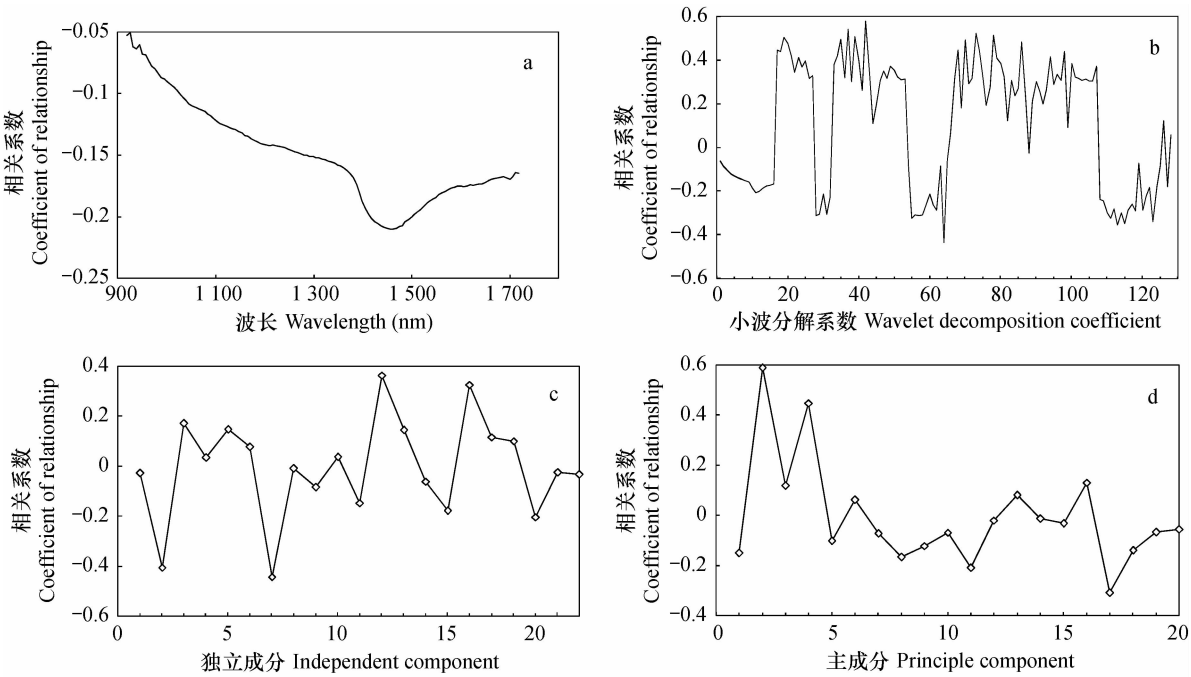
包含的是综合信息,既包含了与土壤碳含量相关的信息,同时也包含大量与之无关的信息。因此,应用所有特征来建立模型,仍然无法提高模型的质量。

对于原始光谱而言,应用 SPA 来选择光谱变量,效果并不理想,预测精度仅略有提高,而 UVE 能达到比较好的效果,但经选择后参与建模的变量数仍然较多。将 UVE 与 SPA 结合起来应用,能达到精简模型的同时保持模型的预测精度,测试数据集的决定系数达到 0.584 5,RMSE 达到 1.81 g kg^{-1}。UVE – SPA 的这一优点,在其他研究中也已得到证实[2-3,25-26]。

对于经过小波分析后获得的小波系数,通过 GA-PLS 和 SPA 进行特征选择后所建立的模型的预测精度较高,测试数据集的决定系数达到 0.577 9 和 0.597 5;而应用 UVE 和 UVE-SPA 的预测精度尽管高于应用全部小波系数建立模型的预测精度,但其效果明显比应用 GA-PLS、SPA 进行特征选择差;UVE-SPA 的效果,要略好于 UVE。

对于 ICA 而言,同样应用 GA-PLS 和 SPA 进行特征选择的效果最佳,测试数据集的决定系数达到 0.649 6 和 0.645 8,RMSE 达到 1.88 和 1.79 g kg^{-1},精度达到最高;UVE-SPA 明显好于 UVE,决定系数从 0.418 1 提高至 0.612 8,RMSE 从 2.22 提高至 1.90 g kg^{-1}。

对于 PCA 而言,GA-PLS 的效果并不理想,决定系数仅为 0.459 7,低于应用全部主成分的预测精度,说明未能有效地进行特征选择;SPA 的效果也不理想,决定系数仅达到 0.512 2;而 UVE 及 UVE-SPA 则效果较好,决定系数达到 0.594 7 和 0.553 5,RMSE 也达到 1.79 和 1.81 g kg^{-1},说明 UVE 对主成分的选择比较有效。



a 光谱数据； b 小波分析； c 独立成分分析； d 主成分分析

a. Spectral data; b. Wavelet analysis; c. ICA; d. PCA

图 3 土壤碳含量与光谱数据及特征变换后各光谱特征之间的相关性分析

Fig. 3 Correlation analysis of soil carbon with soil NIR spectra and spectral features after feature transformation

表 2 基于不同特征提取方法所建模的精度比较

Table 2 Comparison between PLSR models based on different feature extraction methods

特征变换方法 Data transformation	特征选择方法 Feature selection	变量数 Variable number	参与建模的成分数 Number of components included inthe model	RMSECV	RMSEP	R^2
光谱数据 Spectral data	全部变量 All variables	128	4	2.98	2.13	0.445 0
	GA-PLS	18	8	2.85	2.06	0.531 6
	SPA	9	7	2.74	2.14	0.477 6
	UVE	49	6	2.96	1.88	0.599 7
	UVE-SPA	5	5	3.24	1.81	0.584 5
小波分析 Wavelet analysis	全部变量 All variables	128	4	2.98	2.13	0.445 0
	GA-PLS	15	6	2.71	1.85	0.577 9
	SPA	4	4	3.11	1.84	0.597 5
	UVE	30	5	2.91	2.13	0.496 3
	UVE-SPA	14	6	2.97	2.05	0.518 4
独立成分分析 ICA	全部变量 All variables	21	3	2.87	2.15	0.436 4
	GA-PLS	8	8	2.64	1.88	0.649 6
	SPA	4	4	2.84	1.79	0.645 8
	UVE	7	6	2.79	2.22	0.418 1
	UVE-SPA	3	3	2.93	1.90	0.612 8
主成分分析 PCA	全部变量 All variables	20	3	2.95	2.08	0.468 7
	GA-PLS	10	2	2.52	2.09	0.459 7
	SPA	4	1	2.74	1.90	0.512 2
	UVE	3	1	2.82	1.79	0.594 7
	UVE-SPA	2	1	2.93	1.81	0.553 5

与直接对原始光谱数据进行变量选择所建立模型的预测精度相比,在经过小波分析和 PCA 的特征变换,然后应用适当的特征选择方法进行特征选择,所建立的模型,能达到与之相近的预测精度;而应用 ICA 进行特征变换,然后进行特征选择,能建立预测精度更高的校正模型,说明 ICA 在光谱信号分离与特征变换方面,具有更强的能力,同时也说明,选择适当的方法进行特征选择,同样是非常重要的,因此,特征变换和特征选择,需要联合应用,才能取得比较好的效果。

2.4　原始光谱及变换后各光谱特征之间相关性分析

图 4 为应用 GA-PLS、SPA、UVE 和 UVE-SPA 这 4 种特征选择方法对原始光谱和经小波分析、ICA、PCA 变换后的特征进行选择的结果。可以发现,对于原始光谱而言,UVE 选择的变量,包括了与土壤碳含量相关性最高的 1 387 ~ 1 637nm 的范围,但同时也包括了相关性最弱的 920 ~ 964 nm 区域,说明仅通过相关性的高低,无法完全揭示光谱信息与土壤碳含量之间的内在联系;而以此为基础再应用 SPA 进行选择,最终找出 926、932、964、1 406、1 481 nm 这 5 个特征变量。而 GA-PLS 与 SPA,只设别出了部分与 UVE-SPA 相同的特征波段,从而影响了其模型的预测精度。

对于小波分析而言,SPA 与 GA-PLS 选择的特征分布比较广泛,在低频和高频区域均有分布,而且 GA-PLS 选择的特征中,基本包含了 SPA 选择的特征,这就说明了两者所建模型的精度比较接近的原因,同时也说明 SPA 在特征选择方面更有效;而 UVE 选择的区域比较集中,这也限制了 SPA 在进一步选择中发挥作用。

对于 ICA,GA-PLS 选择的独立成分中,完全包括了 SPA 所选择的特征(第 2、7、12 和 19 独立成分);而 UVE 选择的独立成分,尽管也包括了 SPA 选择的特征,但同时也包括了 3 个无关的特征,因而影响了其所建模型的效果,而在继续应用 SPA 进行选择时,尽管剔除了这 3 个无关的特征,但同时也剔除了第 19 独立成分这个关键的特征,因此尽管所建模型的精度仍然较高,但与 GA-PLS、SPA 相比,有所降低。

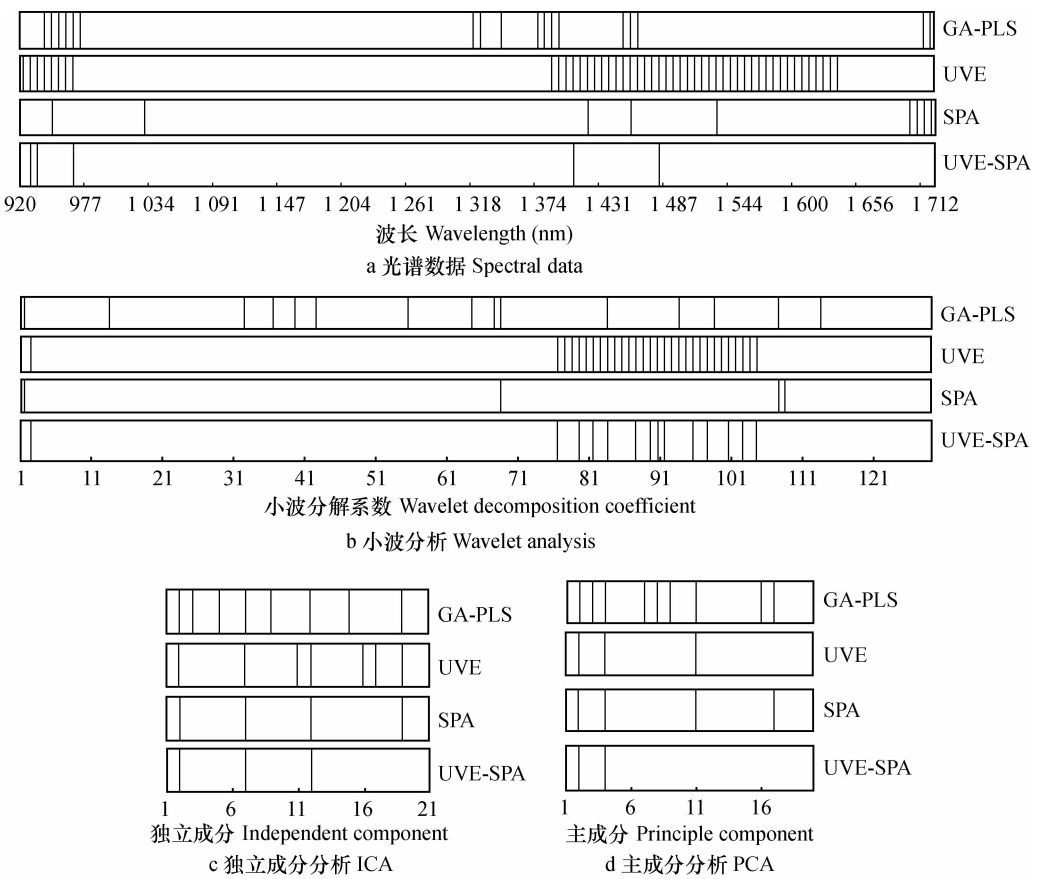


图 4　应用不同特征选择方法对 NIRS 和特征变换后特征进行特征选择的结果

Fig. 4　Feature selection of NIRS and features after transformation with different feature selection methods

对于 PCA 而言,UVE、SPA、UVE-SPA 均选择出了关键的特征(第 2、4、11 主成分),特别是 UVE 选择的效果较好,而 UVE-SPA 由于剔除了第 11 主成分,使模型的预测精度略有降低;SPA 尽管只较 UVE 多选了 1 个特征(第 17 主成分),但模型的预测精度下降比较明显。这可能对 PCA 而言,越后面的主成分,所包含的有用信息越少,反而对建模产生了负面影响;而 GA-PLS 选择的主成分数较多,所包括的无关信息也较多,影响了所建模型的质量。

对 3 种特征变换方法进行比较,可以发现,由于 ICA 具有很强的信号分解能力,因此,先应用 ICA 进行特征变换,然后进行特征选择,能取得较好的效果;而应用小波分析与 PCA 进行特征变换,然后进行特征选择,尽管也能取得一定的效果,但与直接对光谱信息进行特征选择所获得的效果相当,未能显示出优势。

而且对于不同的特征变换方法,特征选择方法的效果,也存在差异。对于小波分析、ICA,GA-PLS、SPA 表现出较好的效果,UVE-SPA 的效果,也好于 UVE,而 UVE 的效果较差,这可能与 UVE 选择的特征较多有关;但对于 PCA,UVE、UVE-SPA 的效果较好,而 GA-PLS、SPA 的效果则较差;这些差异,可能与经不同特征变换方法获得的光谱特征之间的内在关系有关。

3 结 论

土壤近红外光谱信息冗余度高,波段之间信息重叠严重,通过特征变换,可以将信息集中到少量特征中,而通过特征选择方法筛选出与土壤碳密切相关的特征来建立预测模型,可以提高模型的预测精度。本文研究发现,先通过 ICA 进行特征变换,然后应用 GA-PLS、SPA、UVE-SPA 进一步进行特征选择,可以建立较直接对原始光谱数据进行变量选择精度更高的土壤碳校正模型,其精度也达到较好的水平;而通过小波分析或 PCA 进行特征变换,再进行特征选择,仅能获得与直接对光谱数据进行变量选择精度相当的模型。说明 ICA 在光谱数据的特征变换时,较 PCA 和小波变换具有更好的信息集聚效果。因此,在应用行走式设备获得的近红外光谱数据来进行土壤碳含量的测定时,可以将 ICA 与 GA-PLS、SPA 或 UVE-SPA 结合起来进行特征变换和选择,然后应用 PLSR 建立可靠的校正模型,来进行大范围、快速、可靠的土壤碳含量信息获取与制图。

致 谢 对密歇根州立大学作物与土壤科学系 Xuewen Huang 博士和全球变化与对地观测研究中心 Jiaguo Qi 教授在研究数据获取和数据分析方面给予的帮助和支持,表示衷心的感谢。

参 考 文 献

[1] 吴瑞梅,赵杰文,陈全胜,等. 近红外光谱技术结合特征变量筛选快速检测绿茶滋味品质. 光谱学与光谱分析,2011,31(7):1782—1785. Wu R M, Zhao J W, Chen Q S, et al. Determination of taste quality of green tea using FT-NIR spectroscopy and variable selection methods (In Chinese). Spectroscopy and Spectral Analysis, 2011, 31(7): 1782—1785

[2] 吴迪,吴洪喜,蔡景波,等. 基于无信息变量消除法和连续投影算法的可见-近红外光谱技术白虾种分类方法研究. 红外与毫米波学报,2009,28(6):423—427. Wu D, Wu H X, Cai J B, et al. Classifying the species of exopalaemon by using visible and near infrared spectra with uninformative variable elimination and successive projections algorithm (In Chinese). Journal of Infrared Millimeter and Short Waves, 2009, 28(6): 423—427

[3] 黄凌霞,吴迪,金航峰,等. 基于变量选择的蚕茧茧层量可见-近红外光谱无损检测. 农业工程学报,2010, 26(2):231—236. Huang L X, Wu D, Jin H F, et al. Non-destructive detection of cocoon shell weight based on variable selection by visible and near infrared spectroscopy (In Chinese). Transactions of the Chinese Society of Agricultural Engineering, 2010, 26(2): 231—236

[4] Balabin R M, Smirnov S V. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. Analytica Chimica Acta, 2011, 692(1): 63—72

[5] Cecillon L, Cassagne N, Czarnes S, et al. Variable selection in near infrared spectra for the biological characterization of soil and earthworm casts. Soil Biology & Biochemistry, 2008, 40(7): 1975—1979

[6] 陈红艳,赵庚星,李希灿,等. 基于小波变换的土壤有机质含量高光谱估测. 应用生态学报,2011,22(11):2935—2942. Chen H Y, Zhao G X, Li X C, et al. Hyper-spectral estimation of soil organic matter content based on wavelet transformation (In Chinese). Chinese Journal of Applied Ecology, 2011, 22(11): 2935—2942

[7] 郑立华,李民赞,潘娈,等. 近红外光谱小波分析在土壤参数预测中的应用. 光谱学与光谱分析,2009,29(6):1549—1552. Zheng L H, Li M Z, Pan L, et al. Application of wavelet packet analysis in estimating soil parameters based on NIR spectra (In Chinese). Spectroscopy and Spectral Analysis, 2009, 29(6):1549—1552

[8] 陈红艳,赵庚星,李希灿,等. 小波分析用于土壤速效钾含量高光谱估测研究. 中国农业科学,2012,45(7):1425—1431. Chen H Y, Zhao G X, Li X C, et al. Application of wavelet

analysis for estimation of soil available potassium content with hyperspectral reflectance (In Chinese). Scientia Agricultura Sinica, 2012, 45(7): 1425—1431

[9] 沈掌泉,Qi J G,Huang X W,等. 田间行走式测定的红外光谱数据与土壤质地之间的相关性研究. 光谱学与光谱分析,2009,29(6):1526—1530. Shen Z Q, Qi J G, Huang X W, et al. Study on relationship between on-the-go near-infrared spectroscopy and soil texture (In Chinese). Spectroscopy and Spectral Analysis, 2009, 29(6): 1526—1530

[10] 沈掌泉,王珂,Huang X W. 用近红外光谱预测土壤碳含量的研究. 红外与毫米波学报,2010,29(1):32—37. Shen Z Q, Wang K, Huang X W. Estimating the content of soil carbon by using near-infrared spectra (In Chinese). Journal of Infrared Millimeter and Short Waves, 2010, 29(1): 32—37

[11] 沈掌泉,卢必慧,单英杰,等. 基于变量选择的偏最小二乘回归法和田间行走式近红外光谱进行土壤碳含量测定研究. 光谱学与光谱分析,2013,33(7):1775—1780. Shen Z Q, Lu B H, Shan Y J, et al. Study on soil carbon estimation by on-the-go near-infrared spectra and partial least squares regression with variable selection (In Chinese). Spectroscopy and Spectral Analysis, 2013, 33(7): 1775—1780

[12] Hoskuldsson A. Variable and subset selection in PLS regression. Chemometrics and Intelligent Laboratory Systems, 2001, 55(1): 23—38

[13] 吴迪,金春华,何勇. 基于连续投影算法的光谱主成分组合优化方法研究. 光谱学与光谱分析,2009,29(10):2734—2737. Wu D, Jin C H, He Y. Study on combinatorial optimization of spectral principal components using successive projections algorithm (In Chinese). Spectroscopy and Spectral Analysis, 2009, 29(10): 2734—2737

[14] Vohland M, Besold J, Hill J, et al. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. Geoderma, 2011, 166(1):198—205

[15] 张建明,林亚平,吴宏斌,等. 独立成分分析的研究进展. 系统仿真学报,2006,18(4):992—1001. Zhang J M, Lin Y P, Wu H B, et al. Advances of research in independent component analysis (In Chinese). Journal of System Simulation, 2006, 18(4): 992—1001

[16] 冯燕,何明一,宋江红,等. 基于独立成分分析的高光谱图像数据降维及压缩. 电子与信息学报,2007,29(12):2871—2875. Feng Y, He M Y, Song J H, et al. ICA-based dimensionality reduction and compression of hyperspectral images (In Chinese). Journal of Electronics & Information Technology, 2007, 29(12): 2871—2875

[17] 朱大洲,籍保平,史波林,等. 基于小波变换的苹果汁多光程近红外光谱信息提取研究. 红外与毫米波学报,2009,28(5):371—375. Zhu D Z, Ji B P, Shi B L, et al. Information extraction of multi-optical-path nir spectra for apple juice based on wavelet transformation (In Chinese). Journal of Infrared Millimeter and Short Waves, 2009, 28(5): 371—375

[18] 陈刚,陈小梅,李婷,等. 基于小波分解的光谱特征提取算法研究. 光谱学与光谱分析,2010,30(11):3027—3030. Chen G, Chen X M, Li T, et al. Research on spectral data feature extraction based on wavelet decomposition (In Chinese). Spectroscopy and Spectral Analysis, 2010, 30(11): 3027—3030

[19] Li H D, Liang Y Z, Xu Q S, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Analytica Chimica Acta, 2009, 648(1): 77—84

[20] Ye S F, Wang D, Min S G. Successive projections algorithm combined with uninformative variable elimination for spectral variable selection. Chemometrics and Intelligent Laboratory Systems, 2008, 91(2): 194—199

[21] Wu D, Chen X J, Zhu X G, et al. Uninformative variable elimination for improvement of successive projections algorithm on spectral multivariable selection with different calibration algorithms for the rapid and non-destructive determination of protein content in dried laver. Analytical Methods, 2011, 3(8): 1790—1796

[22] Araujo M C U, Saldanha T C B, Galvao R K H, et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics and Intelligent Laboratory Systems, 2001, 57(2):65—73

[23] Leardi R, Gonzalez A L. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. Chemometrics and Intelligent Laboratory Systems, 1998, 41(2): 195—207

[24] Leardi R. Application of genetic algorithm-PLS for feature selection in spectral data sets. Journal of Chemometrics, 2000, 14(5): 643—655

[25] Leardi R, Norgaard L. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. Journal of Chemometrics, 2004, 18(11): 486—497

[26] Norgaard L, Saudland A, Wagner J, et al. Interval partial least-squares regression (ipls): A comparative chemometric study with an example from near-infrared spectroscopy. Applied Spectroscopy, 2000, 54(3): 413—419

ESTIMATION OF SOIL CARBON USING A FIELD AMBULATORY INFRARED SPECTROSCOPY DEVICE

Shen Zhangquan[1,2†] Ye Lingbin[1,2] Shan Yingjie[3]

(1 *Institute of Agricultural Remote Sensing and Information Technology Application*, *Zhejiang University*, *Hangzhou* 310058, *China*)

(2 *Key Laboratory of Agricultural Remote Sensing and Information System*, *Zhejiang Province*, *Hangzhou* 310058, *China*)

(3 *Zhejiang Soil and Fertilizer Station*, *Hangzhou* 310020, *China*)

Abstract To improve accuracy of the prediction of soil carbon using a soil carbon calibration model, feature transformation and feature selection was done to soil infrared reflected spectral (NIRS) data obtained with a field ambulatory infrared spectroscopy device in situ. Firstly, feature transformation was done of the soil NIRS data through independent component analysis (ICA), principle component analysis (PCA) or wavelet analysis (WA), and then feature selection was through uninformative variable elimination (UVE), successive projection algorithm (SPA), uninformative variable elimination in combination with successive projection algorithm (UVE-SPA), and genetic algorithm with partial least squares regression (GA-PLS), separately. And in the end, a soil carbon calibration model was established. Results show that after the processing, a prediction model, better than subjecting the NIRS data to direct wave band selection in accuracy, can be built up, while the combination of the feature selection method with PCA or WA could only achieve some similar effects to those of subjecting NIRS data to direct wave band selection. Therefore, it is feasible to establish a more reliable soil carbon prediction model through feature transformation and selection with the feature selection method coupled with ICA of the NIRS data acquired with a field ambulatory device under complicated environmental condition.

Key words Feature transformation; Feature selection; Soil carbon; Field ambulatory measurement; Near-infrared spectra; Partial least square regression

(责任编辑:汪枞生)