DOI：10.11766/trxb201809100397

基于可见-近红外光谱与化学属性的土壤来源地判别*

张欣跃[1，2] 赵玉国[1，2†] 刘 峰[1] 曾 荣[3] 高 鸿[4] 林 卡[1] 张甘霖[1，2]

（1土壤与农业可持续发展国家重点实验室（中国科学院南京土壤研究所），南京 210008）

（2 中国科学院大学，北京 100049）

（3 南京信息工程大学地理科学学院，南京 210044）

（4 南京师范大学地理科学学院 南京师范大学虚拟地理环境教育部重点实验室，南京 210023）

摘 要 土壤是一种重要的法庭痕迹科学证据，可提供有价值的信息，在案件侦破和法庭审理中发挥关键作用。对于一个未知的土样，怎样确定其来源地，是一个值得研究的问题。分别从跨省和省内两种尺度，基于黑龙江、安徽和江苏三个省市的土壤可见-近红外波段光谱以及土壤化学数据，采用随机森林模型对土壤样本的来源地进行判别，比较了不同土壤测定数据集及其组合方案的判别效果，并分析了土壤化学属性和光谱数据在来源地判别中的相对重要性，以判别正确的样点占总样点数作为验证精度进行评价。结果发现：跨省尺度下，光谱主成分和化学数据结合建模判别验证精度最佳，为0.92；土壤光谱测量所需样品量少，当土壤物证材料量少，化学数据难以获取时，光谱主成分和吸收峰结合建模验证精度最高，为0.82。省内尺度下，依旧是光谱主成分和化学数据结合建模精度最佳，为0.83；在化学数据难以获取时，仅利用光谱主成分与吸收峰也取得了相当的精度（0.82），可见在省内尺度，可以利用光谱来替代化学数据进行建模判别。计算两种尺度下判别因子的重要性发现，跨省尺度下，影响模型判别的化学数据主要是土壤中的全钾、全磷，光谱数据主要是光谱第一主成分以及350～600 nm与1 800～2 100 nm波段的吸收峰；省内化学数据主要是全磷，光谱数据主要是第七主成分与1 300～1 600 nm以及2 100～2 200 nm波段的吸收峰。这表明，利用土壤可见-近红外光谱与化学数据可以有效地判别土壤的来源地。当模型的样点空间分布范围有差异时，可以考虑利用不同的判别因子建模和多个指标来评估判别结果。

关键词 土壤光谱；化学属性；来源地；随机森林

中图分类号 P934 **文献标识码** A

土壤受发生环境的影响，存在高度空间异质性[1-2]。直接表现为土壤化学属性的区域差异。这种土壤化学性质的差异已被应用于法庭土壤学中。一百多年前，Georg Popp作为第一个向法庭提交

* 国家重点研发计划（2017YFC0803807）、科技部基础性工作重点专项（2014FY110200）、公安部物证鉴定中心现场物证溯源技术国家工程实验室开放课题（2017NELKFKT03）、公安部物证鉴定中心协同创新工作项目（2016XTCX03）共同资助Supported by the National Key Research and Development Program of China（No.2017YFC0803807），Key Special Basic work of the Ministry of Science and Technology of China（No.2014FY110200），Open Research Projects of the National Engineering Laboratory of On-site Material Evidence Traceability Technology & Institute of Forensic Science of China（No.2017NELKFKT03），Collaborative Innovation Program of the Institute of Forensic Science of China（No.2016XTCX03）

† 通讯作者 Corresponding author，E-mail：ygzhao@issas.ac.cn

作者简介：张欣跃（1992—），女，河北保定人，硕士研究生，主要从事土壤光谱研究。E-mail：xyzhang @issas.ac.cn

收稿日期：2018-09-10；收到修改稿日期：2018-11-16；优先数字出版日期（www.cnki.net）：2019-01-28

土壤地质材料的科学家来帮助刑事案件侦查，随后法庭地质学逐渐在理论和实践中取得了较大进展[3]。土壤的成分组成可以用来预测土壤的来源位置[4]，Menchaca等[5]利用土壤的颜色、粒度等属性开发了南加州的表土变异性取证数据库，从而确定土壤样本证据的来源位置。利用土壤的基本化学属性数据进行土壤物证材料的空间溯源是法庭物证领域的重要方向[6-7]。

但在实际案例中，土壤物证材料往往是痕量、微量存在，不易获得上述应用中所需的化学分析指标，而土壤光谱能够综合反映部分土壤化学属性，且测量快速、便捷、所需样品量少，且无损[8]。现今，土壤光谱已经广泛地应用于土壤属性的预测、制图等方面，特别是在土壤有机质、全氮、Fe、水分、部分黏土矿物等具有光谱响应的属性上获得较高的预测精度[8-11]。研究表明土壤光谱用于分类很有前景[11,9]，Bellinaso等[11]建立了巴西几个农业区的光谱库，利用光谱主成分及每个剖面的光谱曲线图，来指示不同的土壤类型及剖面分类。土壤光谱也被用来指示土壤所在区域。吴豪翔和王人潮[12]均通过对南方山地丘陵的几种土壤光谱进行研究，选出10个特定波段的光谱来指示相应地区的土壤。徐彬彬等[13]通过对全国土壤光谱反射特性的研究，找出了我国土壤的分布规律，建立了土壤光谱与时空特征间的相关关系。但是，土壤光谱应用于物证溯源的研究尚少，基于土壤发生学的理解和土壤空间变异规律，借助土壤光谱这一综合性指标来指示土壤物证材料的来源地在法庭物证领域是否具有应用前景，以及土壤光谱的物证溯源能够达到怎样的空间精度，是一个具有实际价值的挑战性工作[14]。

本研究的目标是，基于土壤可见-近红外光谱与土壤化学属性数据，采用随机森林方法，分别从省域间尺度与省域内尺度上对土壤的来源地进行判别，判别未知土壤样点的来源，并探讨不同的判别因子对结果的影响。

1　材料与方法

1.1　研究区概况

本研究选取黑龙江、安徽和江苏三个省份以及安徽省的定远、宣城与蒙城三个县作为研究区。黑龙江省介于121°11′～135°05′E，43°26′～53°33′N之间，属温带大陆性季风气候，主要土壤类型为黑土、白浆土、草甸土、沼泽土等。安徽省介于114°54′～119°37′E，29°41′～34°38′N之间，属暖温带与亚热带的过渡地区，主要土壤类型有水稻土、红壤、砂姜黑土等。江苏省地跨116°18′～121°57′E，30°45′～35°20′N，同属于温带向亚热带的过渡性气候，主要土壤类型有：水稻土、潮土、棕壤、黄棕壤等。选择土壤类型相似的安徽省与江苏省，以及土壤差异较大的黑龙江省，比对模型方法在土壤类型相似与差异较大的省域间的应用精度。

1.2　土壤样本采集

为了尽可能避免样品数量对建模结果的影响，三省采集数量相当的样点：安徽100个，黑龙江98个，江苏89个（利用部分样点（77个）化学数据进行建模），样点分布如图1a，其位置的确定均采用综合地理单元法。按照中国土壤系统分类标准中土纲级别，对采集的样点分类，样点土壤类型见表1。为探索不同尺度下样点判别精度，本研究还选取了安徽的定远、宣城与蒙城三地各23、22、21个点，来进一步探究省域内样点的判别情况，样点分布如图1b，对应土壤类型见表2。在法庭土壤学中，获取的土壤物证多为表层土壤，本研究选取表层（依据土壤发生划分层次标准）土壤进行研究。

1.3　土壤光谱与化学数据的测定与处理

本研究中使用的土壤化学数据主要有：pH、有机质、全磷、全钾、速效磷、CEC。测定方法分别为：pH由水浸提法测定、有机质由重铬酸钾-硫酸消化法测定、全磷采用酸溶-钼锑抗比色法测定、全钾采用酸溶-火焰光度法测定、速效磷的测定根据不同的土壤性质选择不同的方法测定（中性和石灰性土壤采用碳酸氢钠浸提-钼锑抗比色法，酸性土壤采用氟化铵、盐酸浸提-钼锑抗比色法）、CEC采用乙酸铵-EDTA交换法测定[15]。

光谱测定在室内完成，将采集的表层土样经室内风干、去杂、研磨过60目筛，然后将过筛后的土样置于烘箱中，45℃条件下烘24 h。采集光谱前，将烘好的土样置于干燥器中待测。测量时取适量土样（约1.2～1.5g）于样品池，利用Cary 5000分光光度计采集光谱数据，采集的波段范围为350～2 500 nm。在可见光波段（350～700 nm）

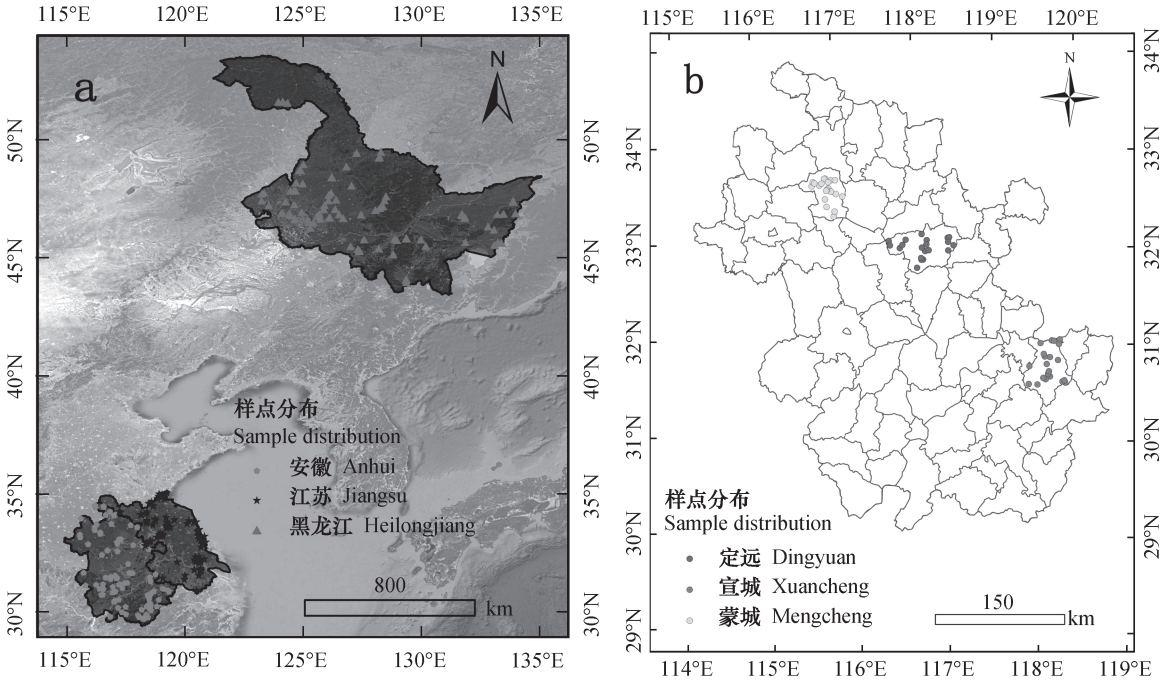

图1 研究区位置和土壤样点分布图

Fig. 1 Study area and distribution of soil sampling sites

表1 黑龙江、安徽和江苏三省样点土壤类型

Table 1 Soil types of the samples collected in Heilongjiang, Anhui and Jiangsu Provinces

黑龙江 Heilongjiang		安徽 Anhui		江苏 Jiangsu	
土壤类型 Soil type	样点数 Number	土壤类型 Soil type	样点数 Number	土壤类型 Soil type	样点数 Number
雏形土 Cambosols	23	雏形土 Cambosols	48	雏形土 Cambosols	33（27）
淋溶土 Argosols	6	淋溶土 Argosols	15	淋溶土 Argosols	12（11）
均腐土 Isohumosols	50	人为土 Anthrosols	34	人为土 Anthrosols	42（37）
火山灰土 Andosols	1	潜育土 Gleyosols	1	盐成土 Halosols	2（2）
潜育土 Gleyosols	2	新成土 Primosols	2		
新成土 Primosols	11				
盐成土 Halosols	5				

注：括号中样点个数为化学数据建模江苏省用到的部分样点数 Note: The digits in parentheses are numbers of the soil samples in Jiangsu Province used in the soil chemical property data based model

表2 定远、宣城和蒙城三县市样点土壤类型

Table 2 Soil type of the soil samples collected in Dingyuan, Xuancheng and Mengcheng Counties

定远 Dingyuan		宣城 Xuancheng		蒙城 Mengcheng	
土壤类型 Soil type	样点个数 Number of samples	土壤类型 Soil type	样点个数 Number of samples	土壤类型 Soil type	样点个数 Number of samples
雏形土 Cambosols	1	雏形土 Cambosols	1	雏形土 Cambosols	12
淋溶土 Argosols	3	人为土 Vertosols	21	变性土 Vertosols	9
变性土 Vertosols	3				
人为土 Anthrosols	16				

和近红外波段（700～2 500 nm）的光谱分辨率分别为<0.048 nm和<0.2 nm，光谱的重采样间隔为1 nm，共采集2 151个波段。

光谱数据的处理主要是吸收峰与主成分的提取。吸收峰特征参数是光谱经连续统去除后提取的，能反映对应波段相应物质含量，比如，Viscarra和Behrens[16]在2 300 nm附近存在一个与有机质相关的C-H的特征峰，Fabrizzi等[17]发现700～800 nm区域与土壤腐殖质以及色素密切相关。本研究提取连续统去除后的光谱吸收峰的部分相关特征参数来进行研究。具体步骤如下：（1）在The Unscrambler中对所有样点的光谱进行异常值剔除；（2）对光谱进行连续统去除，来提取分析土壤光谱吸收峰。连续统是一个逐点直线连接光谱上局部光谱反射极值点的凸壳，连续统去除法处理后的反射率等于在光谱吸收特征处的每个波段的反射率值除以连续统直线上相对应波段处的反射均值[18]。（3）计算连续统去除后前十个吸收峰（按深度排序）的部分特征参数：吸收峰的起始位置波长、终止位置波长、深度、宽度、面积、偏度、吸收峰最深处对应的波长共七个参数。

获取光谱主成分，以10 nm为采样间隔，对所有样点350 ～ 2 500 nm波段的光谱进行重采样，然后提取主成分。在提取出的主成分中，前7个主成分方差累计贡献率达99.9%，故取前七个主成分用于建模分析。土壤光谱数据的处理均在R语言计算环境中实现。

1.4　研究方法

本研究选择了随机森林方法来对土壤样点进行区分，从省域间（黑龙江、安徽与江苏）和省域内（安徽的定远、宣城与蒙城）两个尺度展开研究，分别利用化学数据、光谱数据作为判别因子，来判别未知样点的来源地。

判别因子从化学数据、光谱数据、化学与光谱组合三种数据组合中获取，其中光谱数据包括350～2 500 nm波段全波段的光谱反射率（下文简称全波段）、吸收峰的7个特征参数（下面简称吸收峰）、光谱的前7个主成分（下文简称主成分）、吸收峰与主成分的组合（下文简称吸收峰与主成分）。化学与光谱组合包括化学数据与全波段（下文简称化学与全波段）、化学数据与吸收峰（下文简称化学与吸收峰）、化学数据与主成分（下文简称化学与主成分）以及化学数据与吸收峰及主成分（下文简称化学、吸收峰与主成分）。通过不同数据的组合，探究最佳的模型判别因子。

分类器选择随机森林方法，随机森林模型（Random forest model）是一种对于大多数问题均有效的通用模型，可以处理分类和连续的特征[19]。RF利用bootsrap重抽样方法从原始样本中抽取多个样本，对每个bootsrap样本进行决策树建模，然后对多棵决策树的预测进行投票，得出最终预测结果，RF可以处理大量的数据，运算速度快，且不容易过拟合[20]。

对判别结果选择留一验证法进行验证，利用平衡精度（Balanced Accuracy）来衡量每一类别的判别精度，平衡精度是类别灵敏度（Sensitivity）与特效度（Specificity）的平均值，灵敏度衡量了分类器对正例的识别能力，例如安徽省判别正确样点占安徽总样点的比率；特效度衡量了分类器对负例的识别能力，例如，判别正确的非安徽样点占所有非安徽样点的比率。平衡精度则综合了正例与负例的判别精度。对于模型总精度根据精度以及Kappa 系数来衡量，其中精度为正确判别样点数占总样点数的比例，精度值越大说明样点判别结果越好。Kappa系数代表一致性的强弱程度，当Kappa系数小于0时，一致性程度极差；0～0.2之间一致性程度微弱；0.21～0.4之间一致性弱；0.41～0.6之间一致性中度；0.61～0.8之间一致性显著（或一致性高）；0.81～1.0之间一致性极佳[21-22]。

2　结果与讨论

2.1　基于土壤光谱建模判别

利用土壤光谱建模，对未知土壤样点的来源地进行判别。首先，比较三省可见-近红外波段光谱（图2），可以看出，三省光谱存在一定差异。以全波段、吸收峰、主成分、吸收峰与主成分四种判别因子数据方案分别建立随机森林模型，判别结果的混淆矩阵详见表3。

全波段、吸收峰、主成分三种不同光谱处理方法中，吸收峰建模整体精度最高，为0.81， Kappa系数为0.72，表现出较好的一致性，而且三省精度差别相对较小，安徽与江苏的判别精度也最高，利用吸收峰的相关参数能够对三省样点做出精度相对

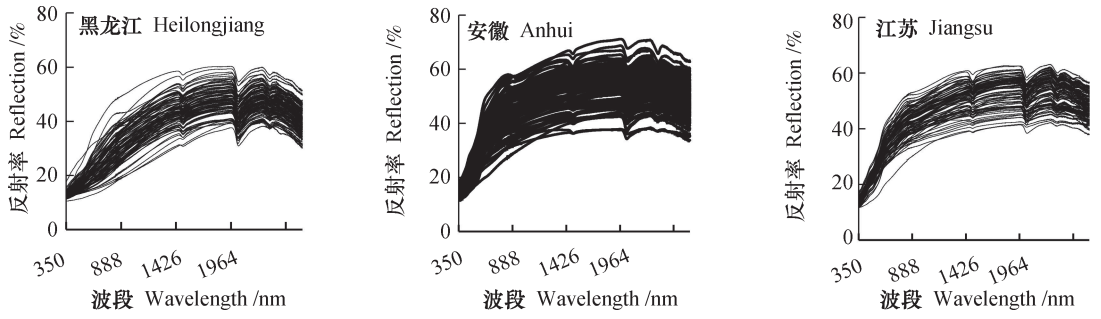

图2 黑龙江、安徽和江苏三省可见-近红外光谱

Fig. 2 Vis-NIR spectrum of Heilongjiang, Anhui and Jiangsu Provinces

表3 光谱数据建模判别结果混淆矩阵

Table 3 Confusion matrix of the performances of the spectral data based model

判别因子 Discriminating factor	真实来源地 Real originated location	判别结果 Discriminant result			评价指标 Evaluation indexes		
		安徽 Anhui	黑龙江 Heilongjiang	江苏 Jinagsu	平衡精度 Balanced Accuracy	精度 Accuracy	卡帕系数 Kappa
全波段 Full band	安徽 Anhui	65	3	32	0.73		
	黑龙江 Heilongjiang	3	89	6	0.93	0.71	0.57
	江苏 Jiangsu	32	6	51	0.69		
吸收峰 Absorption peak	安徽 Anhui	74	6	20	0.81		
	黑龙江 Heilongjiang	6	90	2	0.93	0.81	0.72
	江苏 Jiangsu	16	4	69	0.83		
主成分 Principal components	安徽 Anhui	67	6	27	0.78		
	黑龙江 Heilongjiang	3	94	1	0.95	0.79	0.68
	江苏 Jiangsu	17	7	65	0.79		
吸收峰与主成分 Absorption peak and Principal components	安徽 Anhui	76	4	20	0.82		
	黑龙江 Heilongjiang	6	90	2	0.94	0.82	0.72
	江苏 Jiangsu	17	4	68	0.83		

均一的判别。结合主成分与吸收峰两种数据方案，安徽与江苏的错判点有所降低，各省判别精度均有所提升，且整体精度达到了0.82，Kappa系数为0.72，说明主成分与吸收峰能够反映土壤不同的信息，结合建模能达到更好的判别结果。

三省中黑龙江的判别结果相对较好，错判点均少于10个，其中利用主成分判别结果最佳，错判点为4个，精度达到0.95。黑龙江省土壤类型主要是黑土、草甸土等，有机质含量高，而江苏、安徽省多为水稻土、红壤等富含铁元素，这说明，光谱主成分可以很好地代表不同类型土壤的光谱信息，从而对土壤来源地进行判别，这与Zeng等[23]的研究吻合，光谱主成分可以用于辅助土壤分类。所以利用光谱主成分来对三省样点进行判别，与安徽、江苏土壤类型相差较大的黑龙江省的样点得到了很好的判别，而安徽省与江苏省则相对较差。

通过计算基尼值平均降低量（Mean Decrease Gini，MDG），对判别因子的重要性进行分析。MDG通过基尼（Gini）指数计算每个变量对分类树每个节点上观测值的异质性的影响，从而比较变量的重要性，该值越大表示该变量的重要性越大[24]。模型判别因子的MDG（前10位）统计如表4所示。

根据表4，统计重要性高的吸收峰特征参数对应的波段，发现主要集中在500～800 nm、800～1 300 nm以及2 200～2 495 nm附近的波段。有研究表明大部分有机质光谱响应波段均集中在600～800 nm

表4　光谱判别因子基尼值平均降低量（MDG）统计

Table 4　Discriminant MDG statistics of spectral factors

全波段 Full band		吸收峰 Absorption peak		主成分 Principal components		吸收峰与主成分 Absorption peak and Principal components	
判别因子 Discriminant factor	MDG	判别因子 Discriminant factor	MDG	判别因子 Discriminant factor	MDG	判别因子 Discriminant factor	MDG
596 nm	2.27	depth.5	8.13	PC1	53.07	PC1	10.04
640 nm	2.16	skew.3	7.33	PC2	35.37	depth.5	8.74
629 nm	1.97	end_loc.7	7.25	PC3	27.70	depth.6	6.65
601 nm	1.96	start_loc	7.22	PC7	21.86	width	6.43
637 nm	1.65	width.1	6.95	PC6	19.85	depth.4	5.90
610 nm	1.65	depth.4	6.86	PC4	16.78	start_loc	5.78
589 nm	1.63	depth.6	6.7	PC5	14.99	area.1	5.46
620 nm	1.62	width	6.1			PC2	5.43
614 nm	1.33	area.1	5.86			width.1	5.41
603 nm	1.32	start_loc.1	5.45			skew.3	5.40

注：表中 start_loc、depth、skew、width、area、end_loc、depth_loc分别表示第一个吸收峰的起始位置、深度、偏度、宽度、面积、结束位置、峰最深处位置，相应的 start_loc.1等表示第二个吸收峰特征参数，以此类推。PC1、PC2等表示第一到第七个主成分 Note: The start_loc, depth, skew, width, area, end_loc, and depth_loc in the table denotes the starting position, depth, skewness, width, area, ending position and bottom of the first absorption peak, respectively, and the corresponding start_loc.1, etc. denotes respective characteristic parameters of the second absorption peak, and by parity of reasoning, PC1 ~ PC2 denotes and the first to seventh principal component

波段[18]；铁元素决定了760 ~ 1 300 nm波段内的光谱特性[18]；2 200 nm波段附近存在Al-OH黏土矿物的吸收带。黑龙江省多黑土、草甸土等富含有机质，而安徽省、江苏省多为水稻土、红壤等，土壤中铁元素含量较多。从而导致电磁波特性的差异，影响模型的判别，对土壤样点的来源地判别起到了主要作用。

以上分析可见，利用土壤可见-近红外光谱基本上可以对土壤样点来源地做出有效的判别。当土样化学数据难以获取时，可以通过光谱手段对其进行判别。

2.2　基于化学属性建模判别

土壤光谱对样点来源地进行判别能够达到较好的结果，当土壤样品量足够，可以获取其化学数据时，可以利用土壤的部分化学数据来辅助土壤样点的来源地判别。利用土壤化学属性建立随机森林模型，得到判别结果混淆矩阵（表5）。模型判别因子的MDG（前10位）统计如图3。

表5　化学数据建模判别结果混淆矩阵

Table 5　Confusion matrix of the predictions of the soil chemical property data-based model

		真实来源地 Real originated location		
		安徽 Anhui	黑龙江 Heilongjiang	江苏 Jiangsu
判别结果 Discriminant result	安徽 Anhui	91	3	12
	黑龙江 Heilongjiang	4	89	6
	江苏 Jiangsu	5	5	59
平衡精度 Balanced accuracy		0.92	0.93	0.86
精度 Accuracy			0.87	
卡帕系数Kappa			0.81	

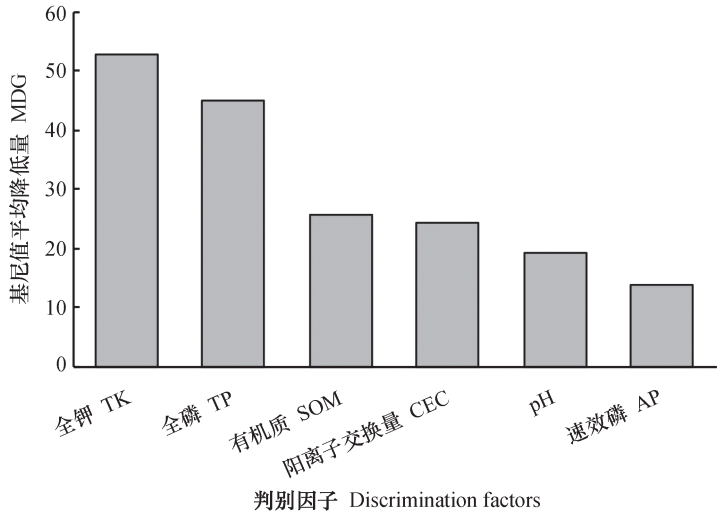

图3 化学判别因子基尼值平均降低量（MDG）统计

Fig. 3 Discriminant MDG statistics of chemical factors

表5判别结果显示，利用土壤化学数据对土壤样品来源地判别整体精度可达0.87，Kappa系数为0.81，其中黑龙江的验证结果最好，为0.93。在影响判别结果的因子中，全钾和全磷的重要性最大（图3），计算原始化学数据的均值、标准差等绘制正态分布图（图4），图中显示安徽的全钾与黑龙江省和江苏省的差异较大，而江苏省的全磷与安徽省和黑龙江省的也有较大差异，同样黑龙江省有机质含量与其他两省也存在一定差异。因而，在模型判别中，全钾、全磷、有机质重要性显著，与重要性计算结果一致。

2.3 化学数据与光谱数据结合建模判别

将化学数据与光谱数据组合建模进行来源地判别。分别利用化学与吸收峰、化学与主成分，以及三者组合建立随机森林模型，样点判别结果混淆矩阵详见表6。

表6混淆矩阵显示，利用化学与主成分相结合

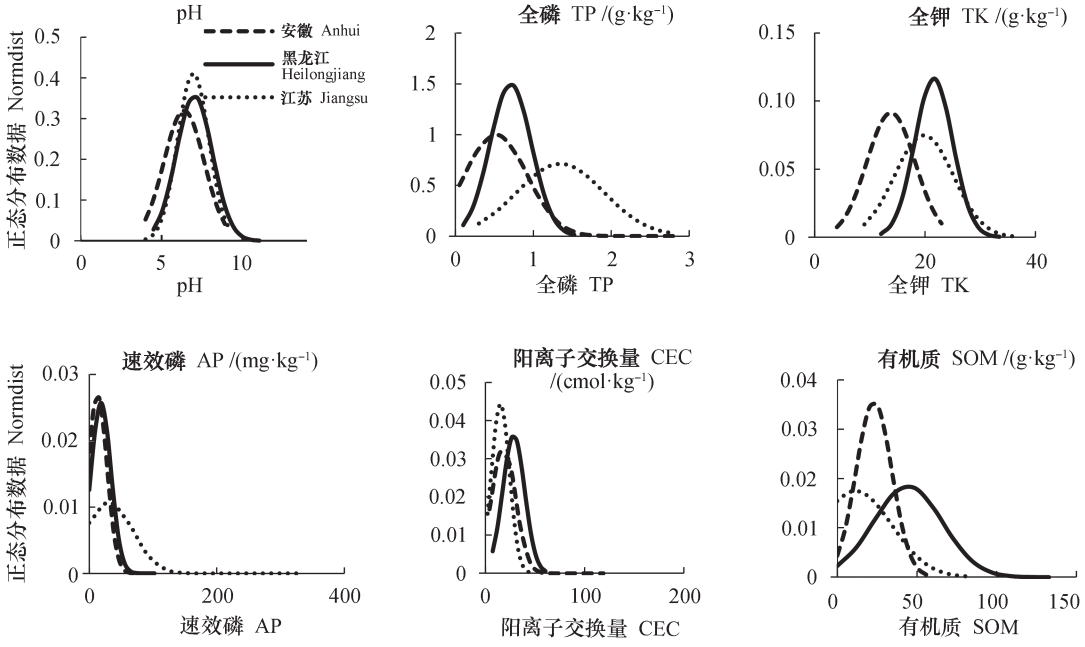

图4 三省化学数据正态分布图

Fig. 4 Normal distribution of soil chemical property data in three provinces

表6　化学与光谱数据建模判别结果混淆矩阵

Table 6　Confusion matrix of the performances of chemical property and spectral data based models

判别因子 Discriminant factor	真实来源地 Real originated location	判别结果 Discriminant result			评价指标 Evaluation indexes		
		安徽 Anhui	黑龙江 Heilongjiang	江苏 Jiangsu	平衡精度 Balanced accuracy	精度 Accuracy	卡帕系数 Kappa
化学与吸收峰 Chemical data and absorption peak	安徽 Anhui	89	3	8	0.90		
	黑龙江 Heilongjiang	3	93	1	0.96	0.89	0.83
	江苏 Jiangsu	12	3	62	0.88		
化学与主成分 Chemical data and Principal components	安徽 Anhui	93	2	5	0.94		
	黑龙江 Heilongjiang	1	96	0	0.98	0.92	0.88
	江苏 Jiangsu	9	4	64	0.90		
化学、吸收峰与主成分①	安徽 Anhui	92	1	7	0.92		
	黑龙江 Heilongjiang	3	94	0	0.97	0.91	0.87
	江苏 Jiangsu	10	3	64	0.90		

① Chemical data, absorption peak and Principal components

建模，黑龙江的错判样点只有1个，安徽与江苏错判点分别为7个与13个，两省间互判错误的点分别为5个与9个。三省各自的精度与三省整体精度均达到了0.9以上，整体精度为0.92，Kappa系数为0.88，表现出极佳的一致性。较之前单独利用化学数据、光谱数据判别结果有了进一步提升。综合三者建模后，判别精度变化不大。统计判别因子重要性（前十）见表7。

表7　判别因子MDG统计

Table 7　MDG statistics of discriminant factors

化学与吸收峰 Chemical data and absorption peak		化学与主成分 Chemical data and Principal components		化学、吸收峰与主成分 Chemical data，absorption peak and Principal components	
判别因子 Discriminant factor	Gini	判别因子 Discriminant factor	Gini	判别因子 Discriminant factor	Gini
TK	15.53	TK	32.92	TK	14.32
TP	10.25	TP	28.18	TP	10.37
width.1	6.15	PC.1	24.25	PC.1	7.81
start_loc	5.98	PC.2	17.98	start_loc	5.84
width	5.96	有机质	12.65	width	5.65
有机质	5.62	CEC	12.4	depth.6	5.58
skew	5.09	pH	10.22	depth.4	5.53
skew.3	5.04	PC.6	7.62	PC.2	5.06
depth.5	4.62	AP	7.55	skew.3	5.04
depth.4	4.41	PC.4	7.46	end_loc.7	4.53

重要的判别因子（表7）主要有全磷、全钾、光谱的第一主成分，其次还有光谱第一个与第二个吸收峰的相关参数、有机质等，其中吸收峰对应的为350～600 nm与1 800～2 200 nm附近的波段，有研究表明，氧化铁在300～580 nm可见光区可形成很强的铁谱带[18]，1 800～2 200 nm则受AL-OH黏土矿物吸收带的影响。化学与主成分判别结果较吸收峰与主成分要好，可能是因为吸收峰对应的矿物、铁元素等属性较化学数据判别效果差，使得结合化学与主成分的判别精度更佳。

基于以上分析，利用光谱前七个主成分与化学数据相结合对安徽省、黑龙江省与江苏省样点进行判别，验证精度能够达到0.92，说明在省域间大范围内可以对样点的来源地做出很好地判别。

2.4 省内不同县市样本的建模判别

在省级尺度可以对土壤来源进行很好地区分，省内县级尺度间的土壤变异更小，本研究尝试是否可以溯源到具体的县。在省内尺度对安徽省三个县的样本进行了来源地判别。判别结果的混淆矩阵及精度见表8。

表8 模型判别结果

Table 8 Performances of the Models

因子 Factors	真实来源地 Real originated location	判别结果 Discriminant result			精度 Accuracy		
		定远①	蒙城②	宣城③	平衡精度 Balanced accuracy	精度 Accuracy	卡帕系数 Kappa
全波段 Full band	定远①	7	8	8	0.48		
	蒙城②	6	14	1	0.72	0.48	0.23
	宣城③	9	2	11	0.65		
吸收峰 Absorption peak	定远①	12	4	7	0.67		
	蒙城②	4	16	1	0.84	0.70	0.55
	宣城③	4	0	18	0.82		
主成分 Principal components	定远①	13	6	4	0.68		
	蒙城②	7	14	0	0.77	0.71	0.57
	宣城③	2	0	20	0.91		
吸收峰与主成分 Absorption peak and Principal components	定远①	15	4	4	0.79		
	蒙城②	1	19	1	0.91	0.82	0.73
	宣城③	2	0	20	0.90		
化学 Chemical data	定远①	11	4	8	0.62		
	蒙城②	2	19	0	0.89	0.64	0.46
	宣城③	8	2	12	0.68		
化学与吸收峰 Chemical data and absorption peak	定远①	15	4	4	0.76		
	蒙城②	1	19	1	0.90	0.76	0.64
	宣城③	5	1	16	0.81		
化学与主成分 Chemical data and Principal components	定远①	17	3	3	0.81		
	蒙城②	3	18	0	0.90	0.83	0.75
	宣城③	2	0	20	0.92		
化学、吸收峰与主成分④	定远①	14	5	4	0.75		
	蒙城②	3	18	0	0.86	0.77	0.66
	宣城③	2	1	19	0.89		

①Dingyuan, ②Mengcheng, ③Xuancheng，④Chemical data，absorption peak and Principal components

由表8的混淆矩阵可以看出，在省域内建模，化学数据建模结果并未优于光谱数据，而且利用化学与主成分和利用主成分与吸收峰建模的判别精度只有0.1之差，说明在安徽省域范围内，可以利用光谱来取代化学数据建模。计算二者建模判别因子的MDG，在只用光谱数据的模型中，光谱第七主成分与第一个吸收峰特征参数重要性较高，第七主成分对应光谱的600 nm附近、1 900 nm附近与2 100 nm附近波段，吸收峰主要对应350～600 nm附近与1 800～2 100 nm附近波段，即受铁谱带、黏土矿物吸收带的影响；化学数据与主成分建模中，重要性较高的为光谱第七主成分与土壤全磷。

同样，利用原始化学数据均值、标准差等绘制化学属性正态分布图（图5），图中显示宣城与蒙城的全磷含量有差异，蒙城、定远和宣城的pH也存在明显差异，所以在省内判别时，全磷与pH的重要性比较大，与MDG计算结果一致。

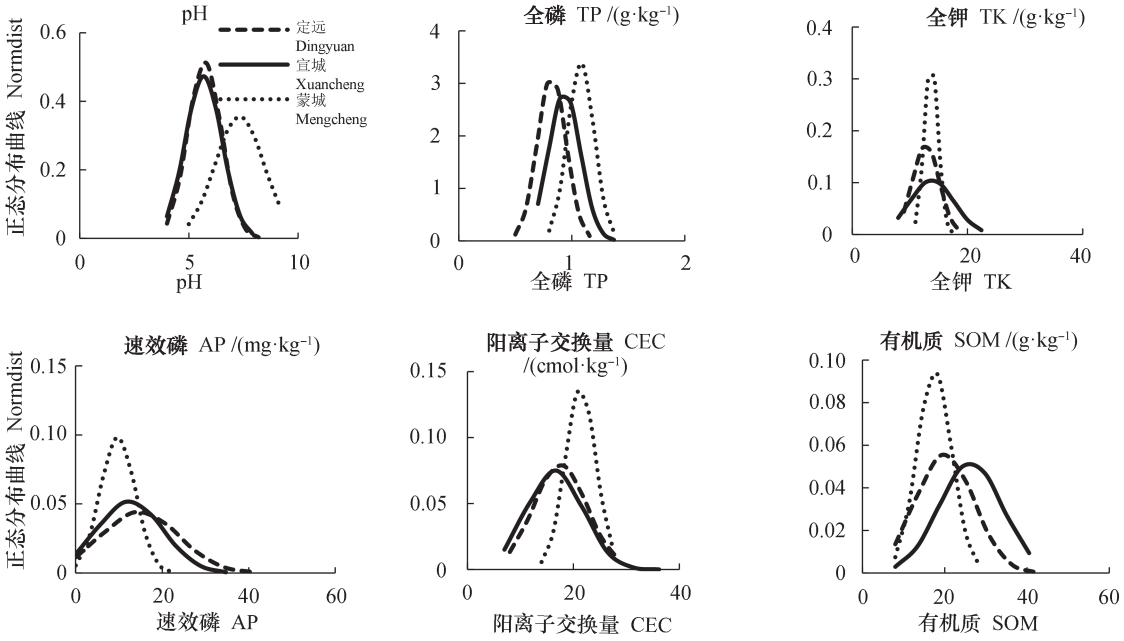

图5　三县化学数据正态分布图

Fig. 5　Normal distribution of soil chemical property data of three counties

3　结　论

基于土壤化学数据与可见光-近红外光谱数据，采用随机森林方法，对土壤来源地进行了判别与分析，主要结论如下：利用土壤可见-近红外光谱与化学数据可以对跨省及省内两种尺度下的土壤样本来源地进行有效判别。其中，跨省尺度下，光谱主成分和化学数据结合建模判别精度最佳，当土壤物证材料量少，化学数据难以获取时，可以结合利用光谱主成分和吸收峰建模；省内尺度下可以利用光谱来替代化学数据进行建模判别。两种尺度下判别因子有差异，可以根据尺度差异选取不同的判别因子进行建模。受本研究所采用的数据集所限，虽然跨省与省内尺度下土壤样本可以得到有效判别，但是能够反映土壤属性的指标还有很多，还需探究更多的判别因子，寻求模型在判别精度上的进一步提高。本研究是基于跨省和省内较大尺度展开的研究，关于对更进一步的局地土壤样本来源地的判别，值得进一步研究。

参 考 文 献

[1] 赵其国，史学正，等. 土壤资源概论. 北京：科学出版社，2007
Zhao Q G, Shi X Z, et al. Introduction to soil resources (In Chinese). Beijing: Science Press, 2007

[2] Nkheloane T, Olaleye A O, Mating R. Spatial

heterogeneity of soil chemical properties in contrasting wetland soils in two agro-ecological zones of Lesotho. Soil Research, 2012, 50(7): 579—587

[3] Ruffell A, McKinley J. Forensic geoscience: Applications of geology, geomorphology and geophysics to criminal investigations. Earth-Science Reviews, 2005, 69(3/4): 235—247

[4] Pirrie D, Dawson L, Graham G. Predictive geolocation: Forensic soil analysis for provenance determination. Episodes, 2017, 40(2): 141—147

[5] Menchaca P R, Graham R C, Younglove T. Developing and testing a soil property database for forensic applications in southern california. Journal of Forensic Sciences, 2018(2): 1—10

[6] Pirrie D, Rollinson G K, Andersen J C, et al. Soil forensics as a tool to test reported artefact find sites. Journal of Archaeological Science, 2014, 41(2): 461—473

[7] Uitdehaag S, Wiarda W, Donders T, et al. Forensic comparison of soil samples using nondestructive elemental analysis. Journal of Forensic Sciences, 2016, 62(4): 861—868

[8] Mcbratney A B, Minasny B, Rossel R V. Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. Geoderma, 2006, 136(1): 272—278

[9] Rossel R A V, Walvoort D J J, Mcbratney A B, et al. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma, 2006, 131(1/2): 59—75

[10] Castaldi F, Chabrillat S, Chartin C, et al. Estimation of soil organic carbon in arable soil in Belgium and Luxembourg with the LUCAS topsoil database: Estimation of SOC with the LUCAS topsoil database. European Journal of Soil Science, 2018, 69(4): 592—603

[11] Bellinaso H, Demattê J A M, Romeiro S A. Soil spectral library and its use in soil classification. Revista Brasileira De Ciência Do Solo, 2010, 34(3): 861—870

[12] 吴豪翔, 王人潮. 土壤光谱特征及其定量分析在土壤分类上的应用研究. 土壤学报, 1991, 28(2): 177—185

[13] 徐彬彬, 季耿善, 朱永豪. 中国陆地背景和土壤光谱反射特性的地理分区的初步研究. 环境遥感, 1991, 6(2): 142—151

[14] Barclay A D, Dawson L A, Donnelly L J, et al. Soils in forensic science: Underground meets underworld. Criminal and environmental soil forensics. Edinburgh: Springer, 2009

[15] 张甘霖, 龚子同. 土壤调查实验室分析方法. 北京: 科学出版社, 2012
Zhang G L, Gong Z T. Soil survey laboratory methods (In Chinese). Beijing: Science Press, 2012

[16] Viscarra R A, Behrens T. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 2010, 158: 46—54

[17] Fabrizzi K P, Moron A, Garcia F O. Soil carbon and nitrogen organic fractions in degraded vs. non-degraded mollisols in Argentina. Soil Science Society of America Journal, 2003, 67(6): 1831—1841

[18] 史舟. 土壤地面高光谱遥感原理与方法. 北京: 科学出版社, 2014
Shi Z. Hyperspectral remote sensing principle and method for soil surface (In Chinese). Beijing: Science Press, 2014

[19] Brett Lantz. 机器学习与R语言. 李洪成, 许金炜, 李舰. 译. 北京: 机械工业出版社, 2015
Lantz B. Machine learning with R. Li H C, Xu J W, Li J. trans. Beijing: China Machine Press, 2015

[20] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述. 统计与信息论坛, 2011, 26(3): 32—37
Fang K N, Wu J B, Zhu J P, et al. A review of technologies on random forests (In Chinese). Statistics & Information Forum, 2011, 26(3): 32—37

[21] 田苗, 王鹏新, 严泰来, 等. Kappa 系数的修正及在干旱预测精度及一致性评价中的应用. 农业工程学报, 2012, 28(24): 1—7
Tian M, Wang P X, Yan T L, et al. Adjustment of Kappa coefficient and its application in precision and agreement evaluation of drought forecasting models (In Chinese). Transactions of the Chinese Society of Agricultural Engineering, 2012, 28(24): 1—7

[22] Landis J R, Koch G G. The measurement of observer agreement for categorical data. Biometrics, 1977, 33: 159—174

[23] Zeng R, Zhang G L, Li D C, et al. How well can VNIR spectroscopy distinguish soil classes? Biosystems Engineering, 2016, 152(2016): 117—125

[24] 张晓羽. 黑龙江省森林植被空间分布及影响因子研究. 哈尔滨: 东北林业大学, 2016
Zhang X Y. Spatial distribution of forest vegetation and its impact factors in Heilongjiang Province (In Chinese). Haerbin: Northeast Forestry University, 2016

Identification of Sources of Soils Based on Vis-NIR Spectroscopy and Chemical Attributes

ZHANG Xinyue[1, 2]　ZHAO Yuguo[1, 2†]　LIU Feng[1]　ZENG Rong[3]　GAO Hong[4]　LIN Ka[1]　ZHANG Ganlin[1, 2]

（1 *State Key Laboratory of Soil and Sustainable Agriculture*，*Institute of Soil Science*，*Chinese Academy of Sciences*，*Nanjing* 210008，*China*）

（2 *University of Chinese Academy of Sciences*，*Beijing* 100049，*China*）

（3 *School of Geographical Sciences*，*Nanjing University of Information Science & Technology*，*Nanjing* 210044，*China*）

（4 *School of Geographic Sciences*，*Nanjing Normal University*，*Key Laboratory of Virtual Geographic Environment*，*Ministry of Education*，*Nanjing Normal University*，*Nanjing* 210023，*China*）

Abstract　【Objective】As an important kind of forensic evidence with valuable information, soil plays a key role in case detection and court trials. For an unknown soil sample, how to determine its source is an issue worth studying. 【Method】 In this paper, a stochastical forest model was adopted to identify sources of soil samples based on vis–NIR spectra and soil chemical properties of the soils in Heilongjiang Anhui and Jiangsu, on a trans-provincial and a provincial scale; comparison performed of the usages of different soil datasets and combination schemes in effect of the identification; analysis conducted of relative importances of soil chemical attributes and spectral data; and evaluation made of determination accuracy based on ratio of the number of the samples correctly determined to the total number of samples.【Result】 Results show that the model combining spectral principal component (PC) and chemical data is the best one in determining sources of soil samples on the cross-provincial scale, with accuracy being 0.92. As spectral measurement does not require many soil samples, in the case the amount of soil samples is limited and soil chemical data is hard to obtain, the spectral-PC-and-absorption-peak-combining model is the highest in accuracy, reaching 0.82. On the provincial scale, the combination of spectral PC and soil chemical property data is still the best one with accuracy being 0.83. When soil chemical property data are hard to obtain, the spectral-PC-and-absorption-peak-combining model can achieve considerable accuracy (0.82), which indicates that spectra can be used to replace soil chemical property data in modeling for determination of sources of soils on the provincial scale. To evaluate importance of discriminant factors on the two scales, it is found that the contents of total potassium (TK) and total phosphorus (TP), the first PC of spectra and spectral absorption peaks at 350 ~ 600 nm and 1 800 ~ 2 100 nm band are the most important indices in the model for determination on the trans-provincial scale. While the content of TP and the seventh PC of spectra and spectral absorption peaks at 350 ~ 600 nm and 1 800 ~ 2 100 nm band were in the model for determination on the province scale.【Conclusion】All the findings indicate that source of a soil sample can be accurately identified based on vis–NIR spectroscopy and soil chemical property data. When spatial distribution of sampling sites varies in range in the model, it is advisable to consider the use of different determination factors in modeling and multiple indices in evaluating accuracy of the determination.

Key words　Soil spectrum; Chemical data; Originated location; Random forests

（责任编辑：檀满枝）