DOI: 10.11766/trxb202505280243 **CSTR: 32215.14.trxb trxb202505280243**

沈芸，刘峰，李德成，郑光辉，潘启凤，曾荣. 基于土壤属性相似性聚类与成土环境推测的土壤空间溯源比较研究[J]. 土壤学报，2026，

SHEN Yun, LIU Feng, LI Decheng, ZHENG Guanghui, PAN Qifeng, ZENG Rong. Comparative Study on Soil Spatial Provenance Based on Soil Property Similarity Clustering and Pedogenic Environment Inference[J]. Acta Pedologica Sinica, 2026,

基于土壤属性相似性聚类与成土环境推测的土壤空间溯源比较研究*

沈 芸[1]，刘 峰[2]，李德成[2]，郑光辉[1]，潘启凤[1]，曾 荣[1†]

（1.南京信息工程大学地理科学学院，南京 210044；2.土壤与农业可持续发展全国重点实验室（中国科学院南京土壤研究所），南京 211135）

摘 要：土壤空间溯源在法医土壤学和司法鉴定领域具有重要应用价值。本研究基于安徽省 265 个表层土壤样本，比较两种土壤溯源策略：一是基于土壤属性相似性匹配与空间聚类分析，利用土壤光谱与理化属性相似度筛选相似样本，结合 DBSCAN 算法确定未知样品可能的来源区域；二是基于成土环境要素逆向推断，采用随机森林模型预测母质、土地利用、地形、气候和植被等环境要素，结合环境要素空间分布图进行来源地推断。通过模拟溯源分析，评估不同溯源策略的精度与适用性。结果表明，相似性匹配策略在空间邻近性强、数据库完善的条件下具有较高的定位精度，而成土环境推测策略在数据库有限或空间异质性强的区域溯源中表现出更优的空间限定能力。两种策略各具优势，融合应用有望进一步提升土壤空间溯源的准确性与空间解析分辨率。

关键词：法庭土壤学；相似性匹配；机器学习；土壤空间溯源

中图分类号：P934　　**文献标志码**：A

Comparative Study on Soil Spatial Provenance Based on Soil Property Similarity Clustering and Pedogenic Environment Inference

SHEN Yun[1], LIU Feng[2], LI Decheng[2], ZHENG Guanghui[1], PAN Qifeng[1], ZENG Rong[1†]

(1. *School of Geographic Sciences, Nanjing University of Information Engineering, Nanjing* 210044, *China*; 2. *State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing* 211135, *China*)

Abstract: 【Objective】Soil spatial traceability has significant application value in forensic soil science and judicial identification.【Method】This study, based on 265 surface soil samples from Anhui Province, compares two soil provenance strategies: (1) a similarity-matching and spatial clustering approach, which filters similar samples using spectral and physicochemical property similarities and applies the DBSCAN algorithm to determine the potential source area of unknown samples; and (2) an inverse inference approach based on pedogenic environmental factors, which employs a random forest model to predict environmental variables such as soil parent material, land use, topography, climate, and vegetation, and infers provenance by integrating spatial distribution maps. By simulating

* 国家自然科学基金项目（42107322）、中国科学院重点部署项目（KGFZD-135-19-10）和中国烟草总公司重点研发项目（110202402016）共同资助 Supported by the National Natural Science Foundation of China (No.42107322), the Key Deployment Projects of Chinese Academy of Sciences (No.KGFZD-135-19-10), and the Key Research and Development Project of China National Tobacco Corporation (No.110202402016)

† 通讯作者 Corresponding author，E-mail：rzeng@nuist.edu.cn

作者简介：沈 芸（2001-），女，安徽六安人，硕士研究生，主要从事土壤遥感研究。E-mail:shenyun1209@163.com

收稿日期:2025-05-28；收到修改稿日期:2025-12-22；网络首发日期（www.cnki.net）:

provenance analysis, the accuracy and applicability of the two strategies were evaluated. 【Result】The results indicate that the similarity-matching strategy achieves higher localization accuracy under conditions of strong spatial proximity and well-established databases, while the inverse pedogenic environment inference strategy demonstrates superior spatial constraint capabilities in regions with limited databases or strong spatial heterogeneity. 【Conclusion】Each strategy has its own advantages, and their integrated application holds promise for further improving the accuracy and resolution of soil spatial provenance analysis.

Key words: Forensic soil science; Similarity matching; Machine learning; Soil spatial traceability

土壤作为地球表层的重要组成部分，在不同区域表现出显著的空间异质性[1]。这种空间差异性赋予土壤独特的地域指示性，使其成为分析和追溯空间来源的重要依据[2]。土壤空间溯源是指通过分析土壤的理化、生物或光谱等属性特征，推断其可能来源区域的空间解析过程，其核心目标在于实现土壤样本与其地理来源之间的空间匹配与定位。近年来，土壤溯源技术在司法鉴定领域的应用日益广泛，为解决涉土刑侦案件提供了科学依据和技术支持[3]。

传统的土壤溯源策略依赖土样间的属性相似性比对，通过分析未知样点与数据库中已知样点的光谱、理化等特征的相似程度，筛选出若干最相似的参考样点，并据此推断其可能的来源区域。常用的相似性度量方法包括欧氏距离、马氏距离及余弦相似度等[4]，部分研究结合行政区划[5]、格网划分[6]或空间插值[7]等方式进行空间限定。然而目前常用的空间限定方法易导致预测范围过大、空间解析精度不足的问题。而 Maestre 和 Cortina [8]研究结果表明属性相似的样本在地理空间上亦可能呈现一定程度的空间聚集性，为传统溯源提供了新的思路。本研究强调相似点群的空间集聚特征可作为识别来源区域的重要依据。通过解析未知样点与参考样点在属性空间中的相似关系，结合空间聚类方法识别属性相似样点在地理空间中的聚集区域，推断未知样点的潜在来源地，进一步缩小未知样点的来源地预测范围。

基于成土环境要素推测的策略，源于对成土环境要素学说的逆向运用。土壤的形成过程受气候、地形、生物、母质和时间等成土环境要素的共同影响，这些要素深刻影响着土壤的物理、化学等特性。因此，理论上若能准确刻画土壤属性与成土环境要素的映射关系，便可由已知的土壤属性信息反向推测其形成背景。这一构想随着机器学习技术的兴起得以实现。通过机器学习模型建立土壤属性与成土环境要素之间的统计关联，可直接推测未知样本的成土环境特征。徐佳等[9-10]建立了基于成土环境要素推测和基于土壤数据库空间分析两种溯源模型，分别对未知土壤样品的空间来源范围进行了溯源研究，并对比了随机森林和人工神经网络两种机器学习模型预测成土环境要素的精度。İnci 等[11]基于元素浓度和可见-近红外光谱数据，实现了 4 种母质类型的分类预测。该策略通常结合成土环境要素空间分布图展开溯源分析，以提高预测精度、并缩短调查时间和降低成本。如 Aitkenhead 等[12]开发了一套集成全国土壤数据库与神经网络算法的系统，借助地形、母质等环境协变量，生成预测未知样本来源地的概率分布图，缩小样点来源的搜索区域。因此，融合多个成土环境要素的空间分布特征，利用多重成土环境要素分布图有望进一步缩小未知样点的来源地预测范围。

然而，目前单一溯源策略在土壤空间溯源的应用仍存在局限。相似性匹配策略受限于已有数据库的完备程度及空间解析尺度，而基于成土环境要素推测的策略则依赖成土环境信息的完整性与准确性，且常基于单一或少量环境要素，难以充分应对复杂多变的地理环境。鉴于此，本研究基于安徽省土壤数据，筛选典型样本开展模拟溯源，对基于土壤属性相似性聚类与基于成土环境推测的两种空间溯源策略进行比较研究，采用未知样点实际空间位置是否落于预测空间范围内作为评价指标，比对与评估二者的精度与差异，探索整合潜力，以期缩小未知样点来源地的预测范围，为司法鉴定提供更精准的土壤溯源方法。

1 材料与方法

1.1 研究区概况与样品采集

安徽省（29°41′-34°38′N、114°54′-119°37′E ）位于我国长江三角洲地区，总面积约 14 万 km^2。地处暖温带与亚热带过渡地区，地势西南高、东北低，地形地貌南北迥异。母质组成复杂，河流冲积物和湖积物分布最广，淮北平原中南部集中分布黄土性古河湖相沉积物，其余母质交错分布在南部地区。安徽省作为农业大省，耕地是最主要的土地利用类型，旱地主要分布在北部，水田集中于中南部。

安徽土壤数据库的土壤样品来源于全国土系调查、土壤碳库调查及宣城小样区土壤调查[13]。数据覆盖典型的土壤类型，包含样点坐标、土壤理化属性、可见-近红外光谱等完备记录，能够较好地覆盖区域环境特征。由于法医土壤学主要涉及土壤表层数据[14]，本研究仅利用表层土壤作为基础数据集，共计 265 个样点（图 1）。

本研究为溯源模拟试验，为确保模型训练与模拟验证具有代表性与科学性，挑选 10 个理化性质差异显著且空间分布具有疏密差异的土壤样点作为模拟未知样点，为溯源模拟提供多样化的测试条件。其余 255 个土壤样点构成参考集，在基于土壤属性相似性的聚类方法中，用于与未知样点进行属性比对与空间聚类分析；在基于成土环境要素逆向推测的方法中，构建机器学习预测模型，推断模拟未知样点的各环境变量值。

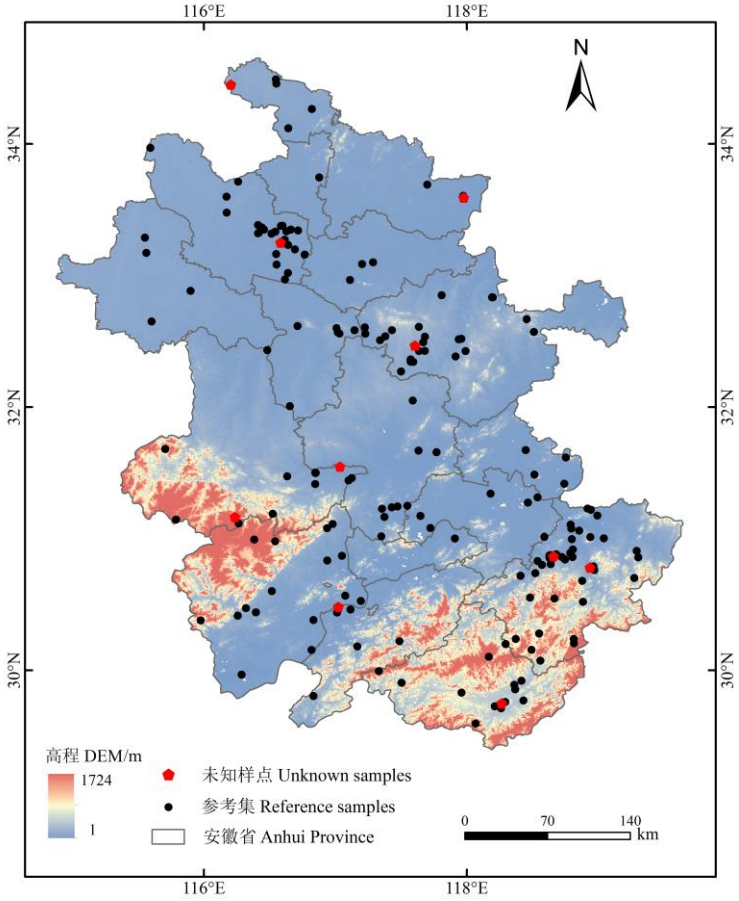

图 1 安徽省土壤溯源数据库样点分布图

Fig.1 Distribution of sample points of the soil traceability database in Anhui Province

1.2 土壤数据测定与预处理

数据库中已测定的理化属性包括土壤有机质含量（SOM）、酸碱度（pH）、阳离子交换量（CEC）以及颗粒组成（砂粒（SAND）、粉粒（SILT）、黏粒（CLAY）），测定过程遵循标准化方法[15]。土壤颗粒分级标准采用 USDA 制：黏粒（<0.002 mm）、粉粒（0.05～0.002 mm）、砂粒（2～0.05 mm）；土壤有机质含量采用经典的重铬酸钾氧化—分光光度法测定；pH 采用水浸提法测定；阳离子交换量采用醋酸铵-EDTA 交换法（NH_4OAc-EDTA）测定。

利用 Cary 5000 分光光度计采集土壤光谱数据，采集的波段范围为 350～2500 nm。在可见光波段（350～700 nm）和近红外波段（700～2500 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。并提取光谱特征吸收峰、光谱主成分、弓曲差（DOA）等特征信息：利用小波变换、连续统去除等光谱预处理方法，增强光谱吸收特征，提取典型吸收峰的特征参数[5,9]，如吸收峰位置、深度、偏度、宽度、面积等；提取方差累积贡献率达 99%的前 7 个主成分用于溯源分析；基于原始光谱数据计算弓曲差，即 550 nm 和 650 nm 处的平均光谱反射率与 600 nm 处光谱反射率之间的差异[16]。

1.3 环境变量数据

综合选取具有代表性的环境变量组合：母质（PM）、土地利用类型（LU）、地形（高程（DEM）、坡度（Slope）、坡向（Aspect））、气候（年均温（Tem）、年均降水量（MAP）、地表温度（LST））、归一化植被指数（NDVI）。并收集土地利用图、母质图等主要成土环境要素图辅助溯源。

母质数据来自全国土系调查[17]；土地利用数据来自国家地球系统科学数据共享平台[18]；地形数据采用 30 m 高程数据（2009 年），来自美国地质调查局（https://earthexplorer.usgs.gov），坡度和坡向数据利用 ArcGIS10.8 软件从高程数据中提取；气候数据采集的是年均温、年均降水量和地表温度，其中年均温和年均降水量为多年气温、降水的平均值（2001—2010 年），来自国家青藏高原科学数据中心平台（https://data.tpdc.ac.cn）；地表温度数据（LST）是在美国国家航空航天局（NASA）官网（https://ladsweb.modaps.eosdis.nasa.gov）下载的 MODIS 影像数据（获取时间为 2010 年 6 月），并对其进行拼接、投影转换、重采样等处理后得到；生物数据采集的是归一化植被指数，该指数也是基于 MODIS 影像数据获取的。其中母质和土地利用类型是离散数据，而地形、气候、归一化植被指数是连续数据。

将所有环境变量数据统一为 WGS1984 坐标系（World Geodetic System-1984 Coordinate System），且均采用通用横轴墨卡托投影（Universal Transverse Mercator Grid System, UTM），并通过栅格重采样将所有环境变量统一分辨率为 1 km。

1.4 基于相似性匹配算法的溯源方法

该策略基于具有相似特性的土壤样点可能来自相似地理空间的假设，认为未知样点与已建立的数据库中光谱和理化属性相似的土壤样点来自研究区域内相邻或邻近的区域。

首先，采用 Zeng 等[4]研究选择的光谱角（SAM）来衡量土壤光谱相似性，使用欧氏距离（ED）计算土壤理化属性相似性。通过比较模拟未知样点与参考集样点的光谱特征和理化属性的相似性，分别筛选出对应的最相似的 15 个参考样点。其次，基于每一个模拟未知样点筛选得到的最相似样本点群，采用密度聚类算法（DBSCAN）对其进行空间聚类分析，识别出空间上若干个高密度的聚类簇（cluster）。借助 ArcGIS 软件中 “最小边界几何”工具，对每一个聚类簇确定其最小外接圆，并以圆心为中心，半径外扩 500 m 建立缓冲区。若干缓冲区被视为未知样点的潜在来源区域，实现土壤来源地预测的空间定位与边界划定。

DBSCAN（Density-Based Spatial Clustering of Applications with Noise）是一种典型的基于密度的聚类算法，基本原理是以邻域半径（ε, epsilon）和邻域内最小样本数（MinPts, min

points）为参数，对任一最相似样本点p，在样本空间中以距离 ε 构建邻域。若该邻域内样本点数量不少于 *MinPts*，则将该点p及其邻域内的样本归为同一簇（cluster）。通过局部样本密度差异识别聚类结构，描述样本分布的紧密性[19]。该方法具有适应不同密度分布、识别不规则形状簇以及有效检测离群点等优势。同时结合 ArcGIS 等可视化工具可以更加直观和高效地处理数据[20]。本研究利用 Python 的 *sklearn* 库实现聚类，针对不同邻域半径（5～40 km，间隔 5 km 递增），分别设定最小样本数（2、3），以优化聚类效果，探索最小样本数与邻域半径对聚类结果的影响。

1.5 基于成土环境要素推测的溯源方法

该策略是对成土环境要素学说的“逆”运用，即土壤是多种环境因素综合作用的产物，利用土壤属性可以映射特定的成土环境要素。机器学习算法作为一种数据探索工具，通过分析参考集土壤属性与成土环境因素的关系构建预测模型，推测未知样本的成土环境，并借助成土环境要素分布图预测未知样点的来源地。

首先，利用随机森林算法（RF）构建成土环境要素推测模型。算法的核心思想是通过在样本和特征层面引入随机性构建多棵决策树，并汇总多个决策树的预测结果，从而提高模型的泛化能力和稳定性[21]。研究针对参考集中 255 个土壤样点，调用 R 中 *randomForest* 包进行建模，调整两个重要参数：*mtry*（每个节点随机选择的特征数）和 *ntree*（决策树的数量）。为了避免随机划分数据可能产生的偏差，采用留一交叉验证法评估判别结果。计算均方根误差（RMSE）、决定系数（R^2）和一致性相关系数（CCC）来评价连续数据的模型预测准确度，RMSE 越小，R^2、CCC 越大，预测准确度越高；计算正确预测的样本数量占样本总数的比值来评价类别数据的模型预测准确度，比值越大，预测准确度越高。

其次，确定各成土环境要素预测值范围，依据未知样点实际空间位置是否落于预测值范围内判定预测结果。母质、土地利用类型等类别变量，预测结果即为值域；而对于地形、气候及归一化植被指数等连续变量，预测结果是具体的数值。预先验证所选取的未知样点的环境组合并未超出参考集的分布范围，采用徐佳等[9]的方法通过计算随机森林模型对参考集样点的各成土环境要素的预测值和真实值的残差确定各成土环境要素阈值 ε，进而划定值域区间。以未知样点 34-001 的年均降水量模拟为例，模型计算参考集样点关于年均降水量的预测值和真实值的残差 ε 为 107.31，并预测该未知样品的 MAP 值为 957.9 mm，因此 34-001 的年均降水量预测值域为[850.6,1065]。具体公式如下：

$$\varepsilon = \frac{\sum|\mathrm{P}_{si}-\mathrm{V}_{si}|}{N} \tag{1}$$

$$R_{si} = [P_{si} - \varepsilon, P_{si} + \varepsilon] \tag{2}$$

式中，ε 为某一成土环境要素的阈值，P_{si}为参考集中成土环境要素的模型预测值，V_{si}为参考集中成土环境要素的实际值，N 为建模土壤样本个数，R_{si}为成土环境要素预测值范围。

最后，基于 GIS 空间分析技术在环境变量图层中筛选符合未知样点预测值域条件的栅格单元，利用图层中某栅格的值和该未知样点的某成土环境要素预测值的相似度，计算未知样点在各环境要素图层中落入相应栅格的概率，生成各成土环境要素预测概率图，并将筛选得到的单一环境要素概率图层进行叠加分析，综合确定未知土壤样品的空间来源概率分布。细节参见徐佳[10]，算法如下：

$$P_{sg} = e^{-|V_{sg}-P_{si}|} \tag{3}$$

$$P_{us} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} P_{sg}(\mathrm{i}) \tag{4}$$

式中，P_{sg}为未知样点位于某一环境变量图层中某栅格的概率，V_{sg}为该环境变量图层中某栅

格的值，P_{si}为该未知样点的某成土环境要素的预测值，P_{us}为预测未知土壤样品的空间来源位置的概率。

2 结果与讨论

2.1 未知样点的特征统计

图 2 为模拟未知样点在可见-近红外光谱区域的反射光谱曲线，呈现典型的土壤光谱特征。随着波长的增加， 700 nm 处土壤光谱反射率迅速上升；近红外波段光谱反射率升高，1400 nm、1900 nm 和 2200 nm 处出现明显吸收峰，对应土壤中羟基、矿物组分等的吸收特征[22]。不同样点之间光谱曲线的反射幅度与吸收深度存在一定差异，样点 34-061 整体反射率最低，近红外波段反射率约为 30%～40%；样点 34-032 整体反射率最高，近红外波段反射率为 55%～65%，为后续土壤来源地判别提供差异化光谱信息基础。

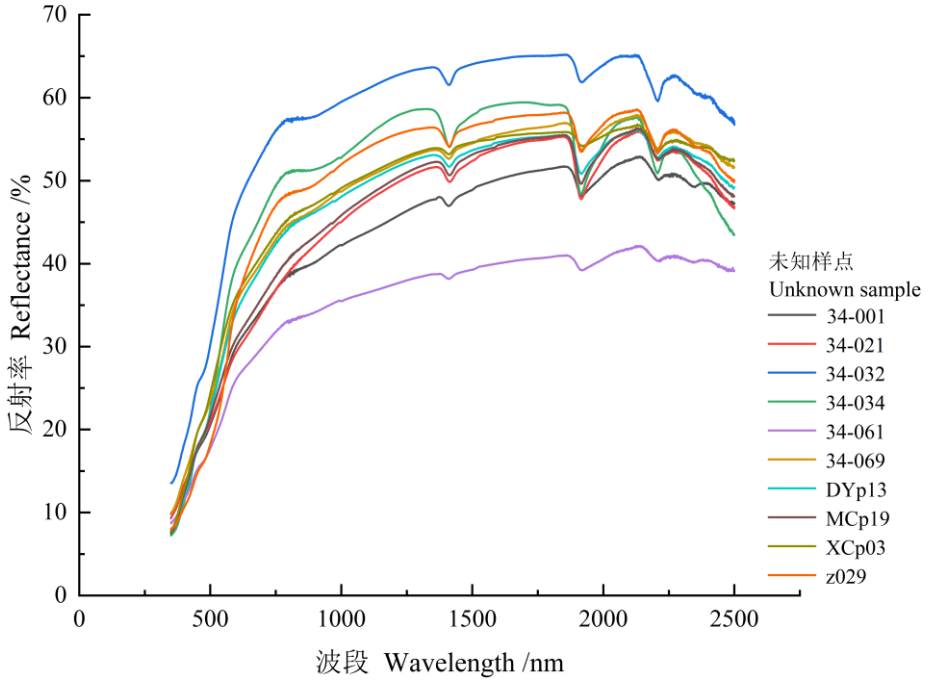

图 2 未知样点光谱特征统计

Fig. 2 Spectral characteristics statistics of unknown sample points

未知样点的理化属性特征统计如表 1 所示，不同未知样点间理化属性存在差异。样点间颗粒组成差异较大，粉粒、砂粒值域较宽，粉粒变异系数高达 59.79%；土壤 pH 均值为 6.706，整体偏酸性；有机质均值达 18.51 g·kg^{-1}，含量较高；阳离子交换量范围为 54.98～271.3 mmol·kg^{-1}，属于中等变异程度（10%≤CV≤100%）。

表 1 未知样点理化属性特征统计

Table 1 Characteristics of physical and chemical properties of unknown sample points

属性 Properties	黏粒 Clay/%	粉粒 Silt/%	砂粒 Sand/%	pH	有机质 SOM/（g·kg^{-1}）	阳离子交换量 CEC/（mmol·kg^{-1}）
最大值 Maximum	37.05	58.36	75.64	8.300	24.52	271.3
最小值 Minimum	9.437	4.204	15.85	4.860	7.63	54.98
均值 Mean	20.32	25.75	53.34	6.706	18.51	161.6
标准差	8.556	15.39	18.33	1.251	5.255	87.00

Standard deviation 变异系数/% Coefficient of variable	42.10	59.79	34.36	18.65	28.39	53.82

统计模拟未知样点环境变量特征，其中母质主要包括河流冲积物、异源母质、基岩风化残积物与坡积物、黄土 4 类；土地利用类型以旱地和水田为主。7 种连续型环境变量的统计特征见表 2，高程和坡度具有显著的空间变异性，表明未知样点覆盖山区、平原等复杂地形；坡向、归一化植被指数和年均降水量呈中等程度变异性，地表温度和年均温的变异程度较弱。环境变量的统计结果说明未知样点的环境因素差异性较强，为后续溯源模拟提供了丰富且具代表性的环境条件。

表 2 未知样点连续环境变量描述性统计

Table 2 Descriptive statistics of continuous environmental variables in unknown sample points

环境变量 Environmental variable	最大值 Maximum	最小值 Minimum	均值 Mean	标准差 Standard deviation	变异系数/% Coefficient of variable
坡度 Slope /°	13.15	0.031	1.512	4.092	270.5
高程 Dem /m	753.6	10.40	114.1	227.5	199.4
坡向 Aspect /°	306.2	64.90	217.5	82.50	37.93
归一化植被指数 NDVI	0.806	0.367	0.534	0.148	27.66
年均降水量 MAP /mm	1632	869.7	1270	293.2	23.08
地表温度 LST /℃	33.70	24.58	30.60	2.721	8.894
年均温 Tem /℃	17.57	13.68	16.33	1.133	6.939

2.2 基于相似性聚类的空间溯源

研究对比未知样点与参考集样点的土壤属性相似性，基于光谱和理化属性的相似度分别筛选出最相似的 15 个参考样点。结果显示，两种算法在 7 个未知样点中筛选出相同的参考样点，其中 34-061 和 34-001 分别重复匹配到 4 个和 3 个参考样点。经冗余剔除后，每个未知样点对应 26～30 个特征点群，构成空间聚类分析的基础。

采用 DBSCAN 聚类方法确定相似点群，对模拟未知样点的来源范围进行框定。设置不同的邻域半径（5 km 至 40 km，间隔 5 km 递增）及最小样本数（即每簇至少包含的样本数量），优化聚类效果。对预测结果产生显著影响的参数设置如表 3 所示，邻域半径和最小样本数均对未知样点来源地追溯精度产生影响。

固定最小样本数阈值，适当扩大邻域半径可提高预测成功率。当最小样本数设定为 2（即每簇至少包含 2 个样本）时，预测精度在聚类邻域半径 5 km 时为 0.5，在邻域半径 40 km 时提高至 0.9。在邻域半径为 30 km 时，预测精度达到 0.7，与张欣跃等[5]基于光谱和化学属性在安徽省进行县域尺度判别的预测精度相当，证实了基于相似性匹配的点群聚类分析方法在土壤溯源中的可行性，并有效缩小潜在来源地范围。样点 34-001 在聚类邻域半径扩展至 30 km 时被成功预测，样点 34-032 和 34-061 在聚类邻域半径扩展至 40 km 时实现预测。该现象主要与未知样点周围参考样点分布不均有关，设置不同邻域半径筛选簇，得到的溯源范围内包含的参考样点数量存在差异。因此，聚类分析中邻域半径应结合参考样点的空间分布密度合理设定，避免因过度压缩预测范围而导致样本量不足，影响溯源精度。

固定邻域半径，最小样本数阈值降低，预测准确率提高。最小样本数设置为 3 的整体预测精度低于最小样本数设置为 2，在 5 km 和 10 km 聚类邻域半径下差异尤为显著。因此，

合理选择邻域半径与最小样本数阈值对于准确识别未知样点的潜在来源区域至关重要，尤其在参考样点分布不均的情况下，更需谨慎调整相关参数。

表 3 基于相似性匹配策略的 10 个模拟未知样点预测结果对比

Table 3 Prediction results of the similarity matching strategy for 10 simulated unknown samples under different parameters

最小样本数 Min points	邻域半径 Epsilon/ km	溯源准确地模拟未知样点 Samples with correct source prediction	预测精度 Accuracy
2	5	34-034、DYp13、MCp19、XCp03、Z029	0.5
	10	34-034、DYp13、MCp19、XCp03、Z029、**34-069**	0.6
	30	34-034、DYp13、MCp19、XCp03、Z029、34-069、**34-001**	0.7
	40	34-034、DYp13、MCp19、XCp03、Z029、34-069、34-001、**34-032、34-061**	0.9
3	5	MCp19、Z029	0.2
	10	MCp19、Z029、**DYp13、XCp03**	0.4
	30	MCp19、Z029、DYp13、XCp03	0.4
	40	MCp19、Z029、DYp13、XCp03、**34-001、34-032、34-034、34-061**	0.8

注：表中加粗显示的模拟样点，表示在相较前一邻域尺度下新增的来源判别正确的样点。Note：Bolded simulated samples represent additional correctly predicted points compared to the previous spatial scale.

参考样点筛选得越集中，聚类效果越显著，来源地范围限定得越精确。以最小样本数阈值设定为 2、邻域半径为 10 km 的聚类分析溯源结果为例（图 3），该尺度下，共有 6 个模拟样点被成功预测，分别为 34-034、34-069、DYp13、MCp19、XCp03 和 z029。其中，模拟样点 MCp19 的预测范围缩减至总研究区域的 10.43%，显著提高了溯源的空间分辨率。

样点 34-021 在本研究设置的各尺度下均未能成功溯源。图 3b 显示与该模拟样点空间位置最接近的参考样点 34-022 已被筛选，二者在光谱整体形态及典型吸收峰[22]（如 1 400 nm、1 900 nm 处）表现出较高的一致性，且土壤有机质（SOM）含量差异约为 2 $g \cdot kg^{-1}$，粉粒和砂粒含量差异在 10 %以内（图 4）。然而，由于 DBSCAN 算法基于点群密度进行聚类，34-021 周围参考样本稀疏，受限于邻域半径内未满足最小样本数要求，导致聚类失败。因此，参考样本数量充足且分布均匀是基于密度与相似性聚类的土壤溯源方法成功应用的基础条件。建立覆盖面广、样本分布合理的高质量参考数据库，对提高未知样品来源的判别精度具有重要意义。

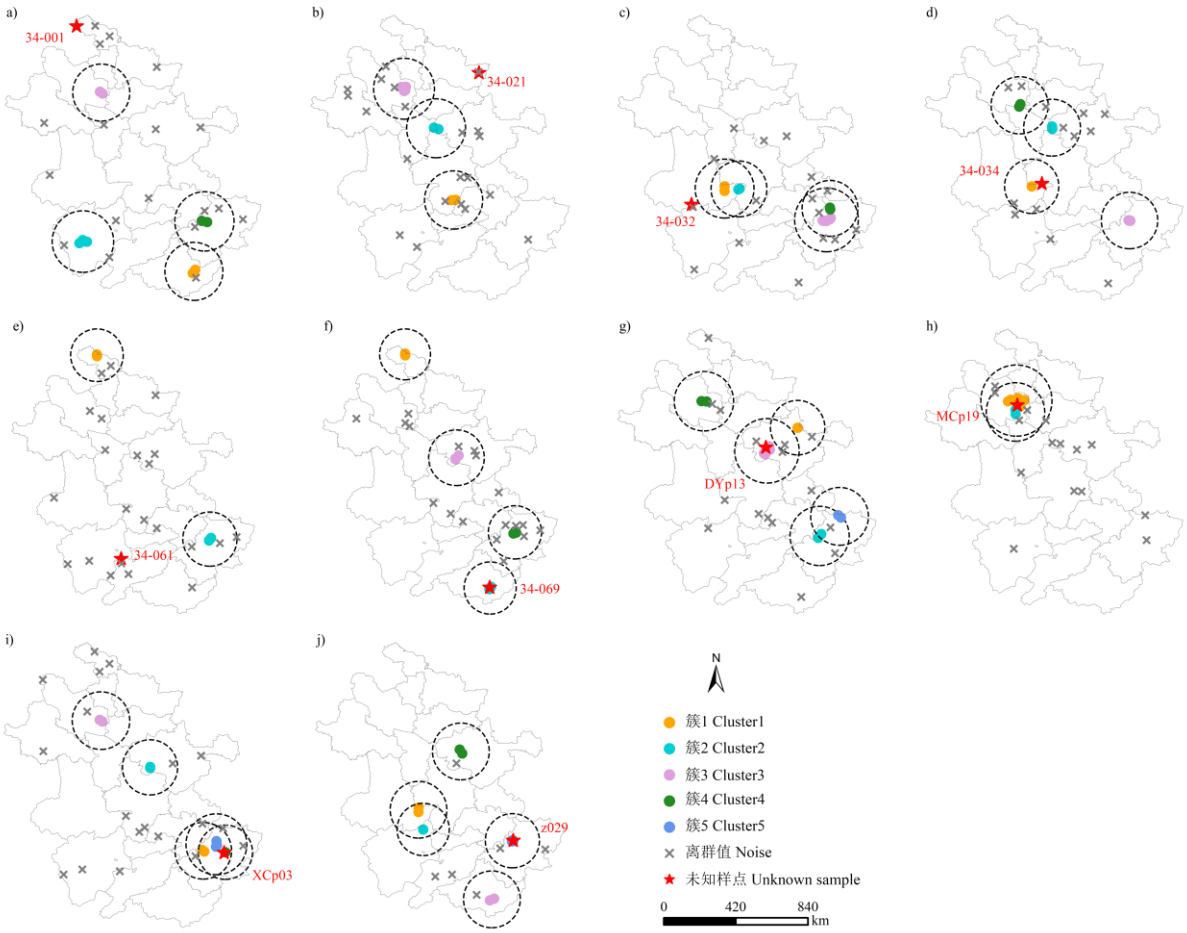

图 3 模拟样点基于最相似点群聚类方法的空间预测结果（聚类邻域半径：10 km；最小样本数：2）

Fig. 3 Spatial prediction of simulated samples using the most similar point-cluster method (ε = 10 km, MinPts = 2)

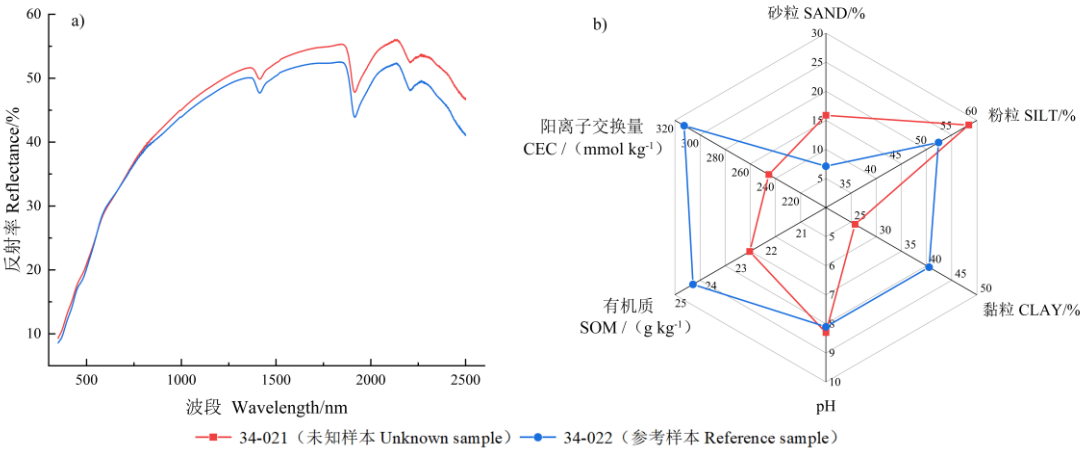

图 4 未知样点 34-021 和距离最近的相似样本 34-022 的光谱（a）和理化属性（b）特征

Fig. 4 Spectral (a) and physicochemical properties (b) characteristics of unknown sample 34-021 and the nearest similar sample 34-022

DBSCAN 聚类方法受最小样本数和邻域半径设置的限制，在参考样点分布稀疏时实用性受限。尝试将相似点群空间聚类分析结果和行政区划框定溯源范围相结合，以模拟样点 34-001 为例，采用相似点群空间聚类分析方法时，DBSCAN 聚类分析的邻域半径设为 40 km

时，可实现较为准确的预测结果，预测范围约占安徽省总面积的三分之一（图 5b）。而基于属性相似性的两种算法共同筛选出的 3 个参考样点均集中于宿州市域范围内（图 5a）。二者结合可有效缩减溯源范围至宿州市域面积的 1/2。因此，未来研究可探索多种方法协同应用，辅助相似性点群聚类分析，提升土壤来源地判别的准确性和精度。

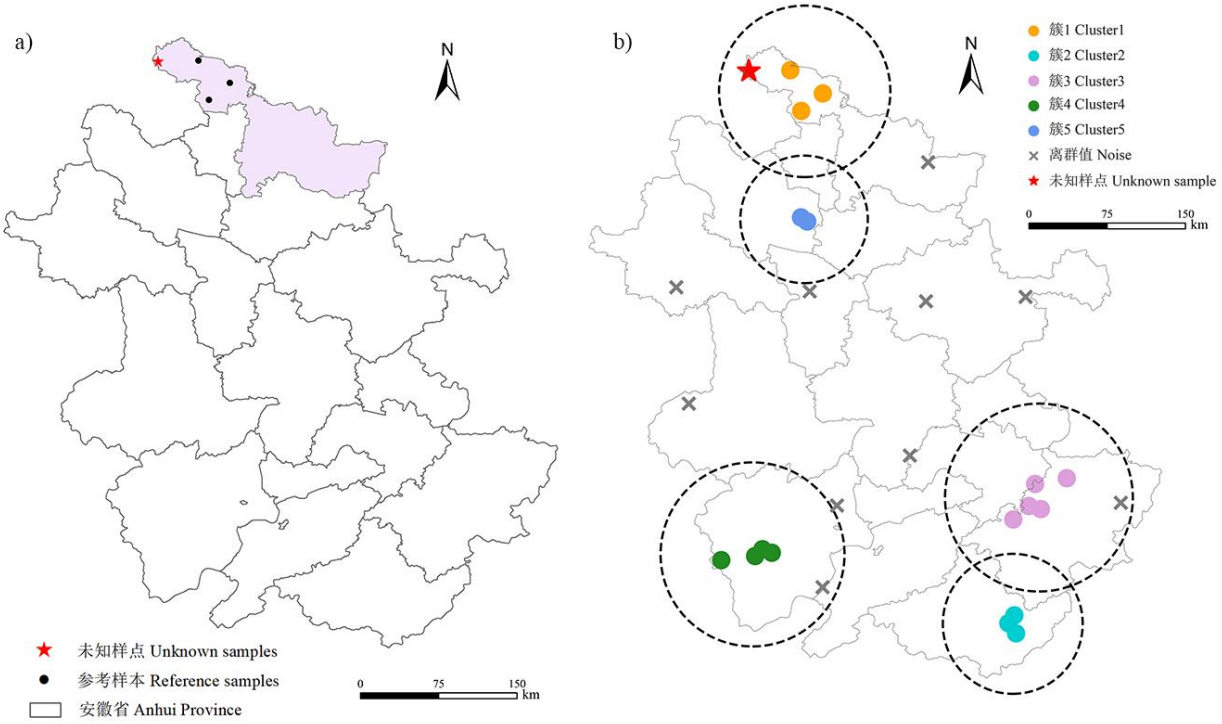

图 5 不同方法确定未知样点 34-001 来源地范围：SAM 与 ED 共同筛选参考样本限定市域范围（a）及 DBSCAN 相似性点群聚类成功预测范围（b）

Fig. 5 Different methods for determining the range of unknown sample 34-001 source locations: SAM and ED jointly screened reference samples to limit the municipal range (a) and the predicted range of DBSCAN similarity point group clustering (b)

2.3 基于成土环境要素推测的空间溯源

2.3.1 成土环境推测模型的精度评价　基于参考集 255 个土壤样点提取的光谱主成分、特征吸收峰参数等光谱数据和理化属性数据，采用随机森林算法构建成土环境要素推测模型，比较 9 种环境变量的预测结果，如表 4 所示。RF 模型对年均降水量（MAP）、地表温度（LST）、土地利用类型（LU）和母质（PM）等环境要素的预测精度较高。安徽省位于暖温带与亚热带过渡区，受季风气候影响显著，800 mm 等降水量线贯穿研究区，南北降水分布不均，同时地表温度梯度变化显著。研究区内母质类型多样，土地利用方式复杂，旱地与水田广泛分布，均对土壤理化属性产生较为强烈的影响。因此，MAP、LST、LU 和 PM 等环境因子与土壤属性之间存在较强的相关性，RF 模型能够较好地捕捉其内在映射关系，从而实现较高的预测精度。

相较之下，模型对高程（DEM）、坡度（Slope）和坡向（Aspect）等地形要素的预测精度相对较低，与徐佳等[9]关于地形对土壤属性影响显著的研究结果不同。主要原因在于不同环境因子对土壤属性作用的空间尺度存在差异[23]。徐佳等[9]以我国东部四省两市为研究区，涵盖平原、丘陵与山地等多样地貌类型，地形异质性强（海拔范围为−12～2 661 m）。而本研究采样点主要位于平原与丘陵地带，地形起伏有限（图 1），DEM、Slope 和 Aspect 等地

形因子的变异性较小，难以显著表征土壤属性的空间差异，导致模型预测能力受限。

表 4 RF 模型对各环境变量的预测准确度对比

Table 4 Comparison of prediction accuracy of RF model for each environmental variable

环境变量 Environmental variable	决定系数 R^2	均方根误差 RMSE	一致性相关系数 CCC	预测精度 Accuracy
年均降水量 MAP/mm	0.721	115.4	0.854	/
地表温度 LST /℃	0.520	1.458	0.723	/
归一化植被指数 NDVI	0.368	0.103	0.607	/
年均温 Tem /℃	0.090	0.687	0.335	/
高程 Dem /m	0.057	155.3	0.321	/
坡度 Slope /°	0.057	1.818	0.362	/
坡向 Aspect /°	0.045	78.23	0.242	/
母质 PM	/	/	/	0.667
土地利用类型 LU	/	/	/	0.604

2.3.2 未知样点空间溯源模拟结果　综合比较模型对 9 种成土环境要素的预测结果，筛选年均降水量、地表温度、土地利用类型、母质 4 种预测精度较高的成土环境要素进行未知样点空间溯源模拟研究。年均降水量和地表温度为连续数据，通过上述随机森林模型输出参考集的预测值和真实值的残差，计算预测阈值 ε（107.3，2.584），确定各因子预测值域范围。基于预测值域检索环境要素栅格，获得未知土壤样品在各环境要素图层上的空间来源预测结果（表 5）。

年均降水量与地表温度的预测成功率较高，分别在 9 个未知样点的模拟预测中正确框定来源位置。两类环境变量主要来源于高时效性、高分辨率的卫星遥感产品，具有良好的空间连续性与观测精度，在预测未知样点的真实位置时表现出较强的成土环境推断能力。但在实际溯源应用中，该方法对潜在来源区域的空间约束能力有限。以样本 34-034 和 34-069 为例，年均降水量与地表温度的匹配区域分别覆盖研究区总面积的 50%以上。结合研究区内成土环境要素的空间分布特征可知，两类环境因子的空间分异程度较低，反映出连续型成土环境要素在土壤来源区域筛选中的局限性，具体表现为预测准确性较高但空间指向性较弱，且易受区域环境背景特征影响。而土地利用和母质两类成土环境因子的预测准确率相对较低，大部分样本未能有效溯源。由于数据多依赖于人工调查和分类编制，受人为划分误差、数据更新频率低及分类标准不统一等因素影响，在实际溯源应用中判别效果受限。因此，年均降水量与地表温度在土壤来源地推断中适合作为提升初步预测准确率的主导因子，而土地利用和母质则在高精度溯源阶段，通过进一步空间约束发挥关键作用。合理结合连续型与分类型成土环境因子特征，有望在提高未知样点来源地判别准确率的同时，兼顾溯源区域的有效收敛。

表 5 正确预测未知样点的单一成土环境因子预测值域在研究区的空间覆盖率统计

Table 5 Spatial coverage statistics of the domain of predicted values of single soil-forming environmental factors in the study area for correctly predicted unknown sample sites

未知样点 Unknown sample	年均降水量 MAP/%	地表温度/% LST	土地利用类型/% LU	母质 /% PM
34-001	28.23	50.57	/	/
34-021	41.68	65.90	/	/

34-032	95.14	/	/	/
34-034	96.01	71.51	5.530	/
34-061	87.24	33.74	/	/
34-069	87.03	63.16	/	/
DYp13	27.10	74.23	/	/
MCp19	/	16.65	/	18.72
XCp03	82.47	81.48	25.87	/
z029	58.81	40.94	/	53.10

结合单一成土环境要素预测分析结果，选取年均降水量、地表温度、土地利用和母质 4 种成土环境要素，权重均设定为 0.25，进行加权叠加分析，确定模拟未知样点的空间来源概率分布（图 6）。不同未知样点基于成土环境要素加权叠加的空间预测概率分布存在显著差异。共计 8 个未知样点被成功预测，但其真实位置预测概率普遍不高（均小于 50%）。其中 34-034、XCp03 和 z029 样本的真实位置预测概率超过 30%，而 34-001、34-021、34-061、34-069 和 DYp13 等样本预测概率均低于 25%。结果表明 4 种成土环境要素在一定程度上可表征土壤属性特征，且多重有效环境因子叠加有助于提高空间推断的准确性，但整体空间映射能力存在局限。例如，34-032 样本仅在年均降水量单因子预测中溯源成功（表 5），叠加分析未能实现有效预测。此外，不同样点来源地预测的空间限定效果存在差异。34-001 和 DYp13 样本的预测范围相对集中，潜在来源区域显著收敛；34-021、34-061 和 34-069 样本虽然预测准确，但概率区域覆盖较广，缺乏集中趋势，实际溯源成本增加。而未知样点 MCp19 的预测结果虽限定在狭长地带，显著缩小了潜在区域，但预测未成功。观察表 5 发现，采用地表温度或母质作为单一成土环境因子预测时可有效溯源该样本，且将预测范围缩小至研究区总面积的 20%以下。因此，多环境因子组合在特定情形下未必优于单一成土环境因子的预测效果。

综上，多种成土环境要素的加权叠加分析在部分未知样点中能够有效缩小潜在来源区，提高溯源判别的指向性。但在成土环境要素预测准确率较低的情况下，预测空间范围会扩大，指向性下降。因此，在实际应用中需提高成土环境要素预测模型的准确率，根据样本特性动态优化成土环境要素的筛选及权重配置策略，以进一步提升土壤溯源的精度与应用可行性。

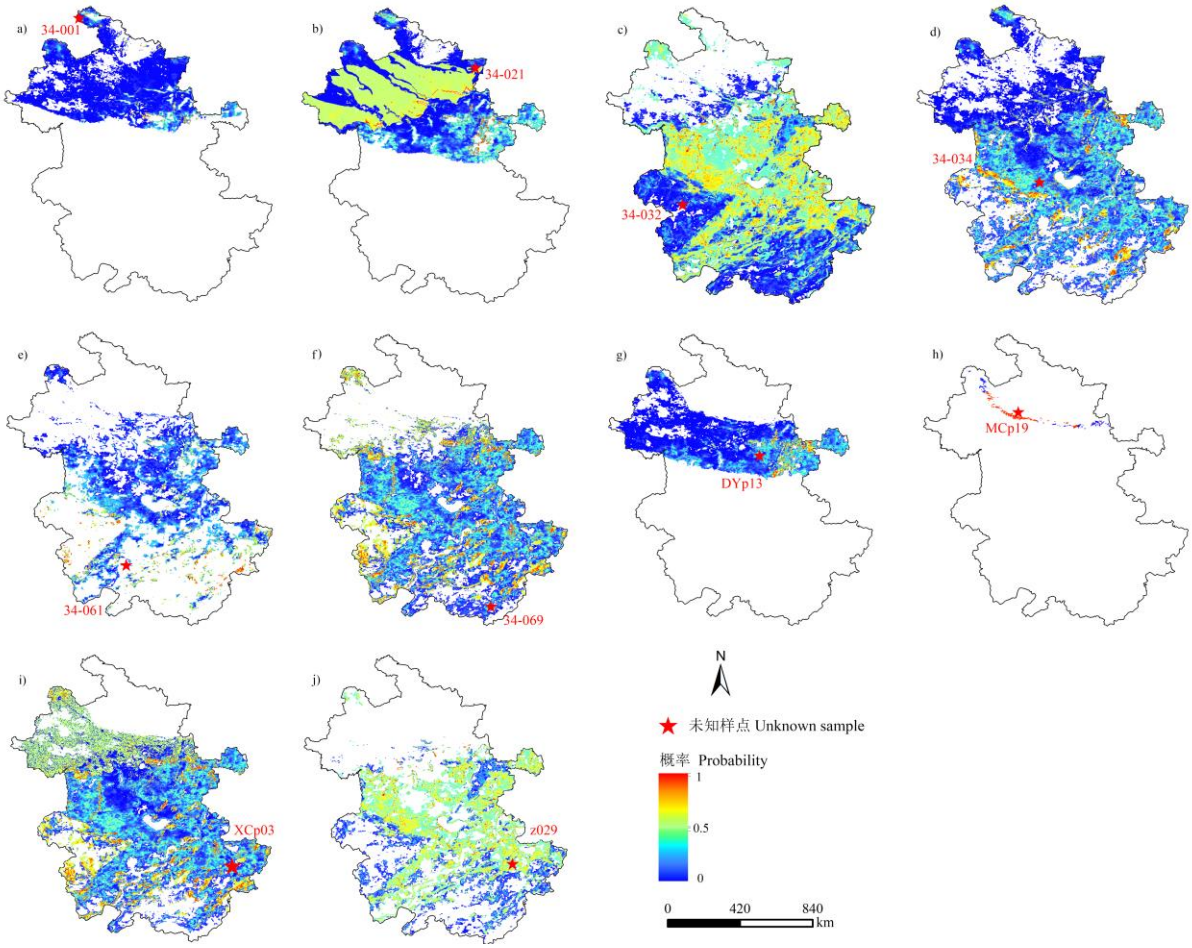

图 6 基于成土环境要素推测的未知样点空间来源概率分布图

Fig. 6 Probability distribution of spatial origin of unknown sample points based on inferred soil-forming environmental factors

2.4 空间溯源策略比较分析

基于属性相似性匹配与成土环境要素推测的两种土壤空间溯源策略原理存在显著差异，相似性匹配策略是将未知样点的光谱或理化属性与已建数据库中的特定样品进行比对，利用若干最相似样点的空间聚类特征，结合限定的邻域半径，推断未知样点的潜在来源区域；而成土环境推测策略则是预先通过对已有数据库的样本集分析，建立成土环境预测模型，直接推测未知样点可能的成土环境来源，进而在多重环境因子空间分布图中筛选符合条件的区域。

从溯源结果来看，两种溯源策略均可用于未知样品的溯源分析。相似性匹配策略在空间定位方面表现出明显的优势，采用优化的邻域半径和最小样本阈值的 DBSCAN 算法，从属性空间相似性推理到地理空间上的有效溯源。例如，未知样点 MCp19 成功聚类出密集点群，预测的溯源范围缩减至总研究区的 1/10，且包含真实采样位置。但受限于参考样本数量及空间分布密度，若参考集覆盖不足或样本聚集不均，可能导致局部溯源失败。未知样点 34-021 在参考样本稀缺区域溯源过程中出现较大偏差，研究设置的各尺度下均未能成功溯源。

成土环境要素推测策略的本质是经验统计意义上的间接推断，利用机器学习模型根据土壤性质和成土环境要素之间的映射关系，较少依赖于参考样本的空间接近性，有效缩小某些未知样点的可追溯范围。如未知样点 DYp13 结合年均降水量和地表温度等成土环境要素筛选，预测区域集中于研究区中部，覆盖实际来源地附近区域。该策略从统计层面利用机器学

习模型进行“反推”，反映的是条件概率而非决定性因果关系，其有效性依赖于模型训练所用数据集的完备性、代表性和空间均衡性。缺乏充足、有代表性的样本数据可能导致模型泛化能力受限，类别型成土环境要素图的更新滞后等问题亦可能降低溯源可靠性。受安徽省环境背景特征影响，在大多数连续型环境因子主导下，潜在来源区域通常覆盖研究区较大比例空间，限定效果相对较弱。仅当土地利用、母质等空间异质性强的分类型环境因子成功匹配时，空间收敛性才会显著提升。例如，未知样点 34-032 在年均降水量、地表温度等连续因子的筛选下，推测结果覆盖研究区超过 90%的面积，溯源范围较大，且土地利用、母质分类型因子预测失败，导致概率分布与实际来源地存在明显偏移。

两种溯源策略在不同应用情境下各具优势，相似性匹配策略适用于参考样本数量充足的精确溯源；而成土环境推测策略则适用于参考样本稀缺或需快速初筛潜在来源区域的情形。为探究两种策略联合应用的潜力，本研究尝试叠加两种策略的预测范围，分析交集区域能否进一步缩小溯源范围并覆盖未知样品的真实来源地。以未知样品 DYp13 为例，相似性匹配策略推定的溯源面积约为 42486.65 km^2，占研究区总面积的约 1/3（图 3g）；成土环境推测策略推定的溯源面积占研究区总面积的 22.89%（图 6g）；两种策略叠加后的综合预测区域面积约占总研究区的 10.81%（图 7a），有效缩小了溯源范围，降低了实际溯源成本。而未知样品 34-021 在单一相似性匹配策略中未能成功定位来源地（图 3b），两种策略叠加后预测结果与真实来源地亦存在明显偏移，导致溯源失败（图 7b）。基于上述结果，未来研究可进一步结合两种策略的优势，发展多源信息融合的综合判别框架，以提升土壤样品来源地识别的精度与空间约束能力。

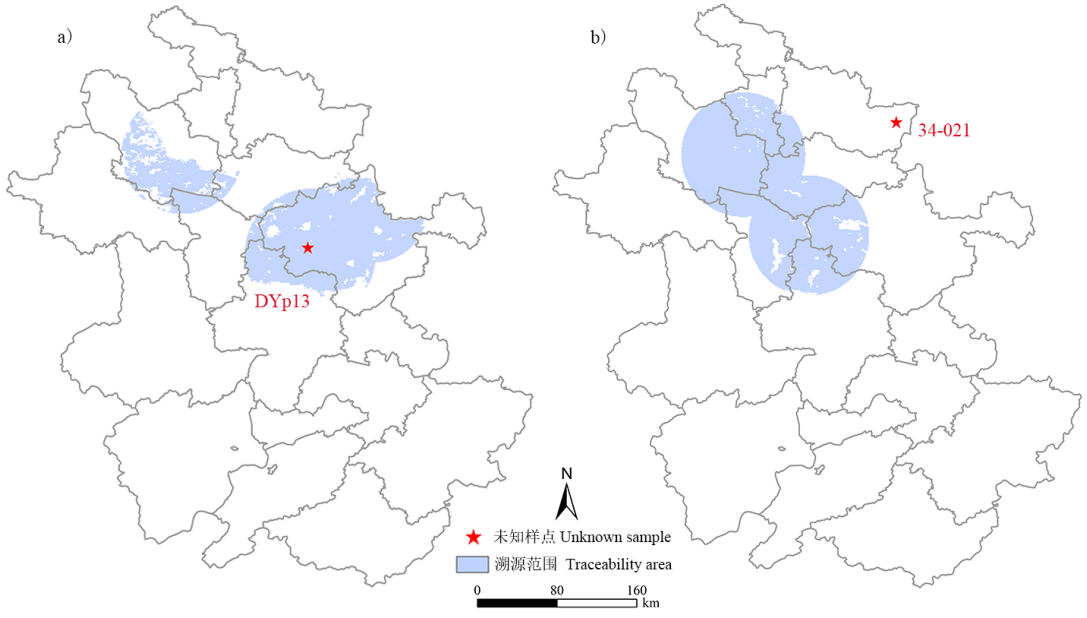

图 7 两种策略结合预测未知样点 DYp13（a）和未知样点 34-021（b）来源地

Fig. 7 Combination of both strategies to predict the origin of unknown sample DYp13 (a) and unknown sample 34-021 (b)

3 结 论

本研究围绕土壤属性相似性聚类与成土环境要素推断两种溯源策略，比较了模拟未知样点空间来源地的判别方法。结果表明，属性相似性匹配结合聚类分析能够有效缩小潜在来源区域，合理设置邻域半径与最小样本数阈值对于提高溯源准确性具有重要作用；基于随机森

林预测的成土环境要素，在空间分异显著的条件下，能够进一步增强溯源推断的稳定性和精度。综合对比两种策略，属性相似性聚类在具备充分参考样本时表现出较高的空间聚焦能力，而成土环境推测在参考样本稀疏或环境异质性强的情况下展现出良好的适用性，二者的互补集成有望为土壤空间来源地判别提供更加可靠的技术支撑。该研究基于安徽省及其选定的土壤属性与成土环境要素，对相关研究具有一定参考价值。研究结果强调构建高质量、空间平衡的土壤数据库，以及选择符合局部样本分布和环境异质性的溯源策略的重要性。未来的工作应侧重于综合考虑土壤属性测量误差与环境数据产品差异，结合样本特征动态优化成土环境要素的筛选与权重分配，通过多重环境变量建模实现潜在来源区域的快速初筛，并结合空间相似性分析进一步收缩未知样品来源地预测范围的溯源方法，提高司法鉴定领域土壤来源识别的应用效率。

参考文献（References）

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

[2] Cengiz S, Cengiz Karaca A, Çakır İ, et al. SEM–EDS analysis and discrimination of forensic soil[J]. Forensic Science International, 2004, 141(1): 33-37.

[3] de Caritat P, Simpson T, Woods B. Predictive soil provenancing (PSP): An innovative forensic soil provenance analysis tool[J]. Journal of Forensic Sciences, 2019, 64(5): 1359-1369.

[4] Zeng R, Rossiter D G, Zhao Y G, et al. The choice of spectral similarity algorithms influences suspected soil sample provenance[J]. Forensic Science International, 2023, 347: 111688.

[5] Zhang X Y, Zhao Y G, Liu F, et al. Identification of sources of soils based on vis-NIR spectroscopy and chemical attributes[J]. Acta Pedologica Sinica, 2019, 56(5): 1060-1071.[张欣跃, 赵玉国, 刘峰, 等. 基于可见-近红外光谱与化学属性的土壤来源地判别[J]. 土壤学报, 2019, 56(5): 1060-1071.]

[6] Haddadchi A, Hicks M, Olley J M, et al. Grid-based sediment tracing approach to determine sediment sources[J]. Land Degradation and Development, 2019, 30(17): 2088-2106.

[7] Aberle M G, Troitzsch U, Robertson J, et al. Conjunctive use of mineralogy and elemental composition for empirical forensic provenancing of topsoil from Canberra, Australia[J]. Forensic Chemistry, 2023, 36: 100524.

[8] Maestre F T, Cortina J. Spatial patterns of surface soil properties and vegetation in a Mediterranean semi-arid steppe[J]. Plant and Soil, 2002, 241(2): 279-291.

[9] Xu J, Liu F, Wu H Y, et al. Predicting of key environmental factors from soil properties based on artificial neural network and random forest learning model[J]. Chinese Journal of Soil Science, 2021, 52(2): 269-278. [徐佳, 刘峰, 吴华勇, 等. 基于人工神经网络和随机森林学习模型从土壤属性推测关键成土环境要素的研究[J]. 土壤通报, 2021, 52(2): 269-278.]

[10] Xu J. Study on soil spatial traceability method based on soil-environment relationship[D]. Beijing: University of Chinese Academy of Sciences, 2021. [徐佳. 基于土壤—环境关系的土壤空间溯源方法研究[D]. 北京: 中国科学院大学, 2021.]

[11] İnci Y, Bilgili A V, Gündoğan R, et al. Machine learning-based classification of soil parent materials using elemental concentration and vis-NIR data[J]. Sensors, 2024, 24(16): 5126.

[12] Aitkenhead M J, Coull M C, Dawson L A. Predicting sample source location from soil analysis using neural networks[J]. Environmental Forensics, 2014, 15(3): 281-292.

[13] Zeng R, Zhao Y G, Li D C, et al. Selection of “local” models for prediction of soil organic matter using a regional soil vis-NIR spectral library[J]. Soil Science, 2016, 181(1): 13-19.

[14]Fitzpatrick R W. Nature, Distribution, and origin of soil materials in the forensic comparison of soils //Ubelaker D H, Birkby W H. Soil Analysis in Forensic Taphonomy[M]. Boca Raton, USA: CRC Press, 2008: 1-28.

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

[16] Jiao C X, Zheng G H, Xie X L, et al. Prediction of soil organic matter using visible-short near-infrared imaging spectroscopy[J]. Spectroscopy and Spectral Analysis, 2020, 40(10): 3277-3281. [焦彩霞, 郑光辉, 解宪丽, 等. 可见-短近红外成像光谱数据的土壤有机质含量估算[J]. 光谱学与光谱分析, 2020, 40(10): 3277-3281.]

[17]Li D C, Zhang G L, Wang H. Soil Series of China·Anhui Volume[M].. Beijing: Science Press, 2017.[李德成, 张甘霖, 王华. 中国土系志·安徽卷[M]. 北京：科学出版社, 2017.]

[18]Pan X Z. Anhui Province 1:100,000 land use dataset (2010) [DS]. National Earth System Science Data Sharing Platform-Yangtze River Delta Data Center, 2016.[潘贤章. 安徽省 1:10 万土地利用数据集(2010 年)[DS]. 国家地球系统科学数据共享平台—长江三角洲数据中心, 2016.]

[19] Zhou Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016. [周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.]

[20] Tu X Q, Fu C, Huang A, et al. DBSCAN spatial clustering analysis of urban “production–living–ecological” space based on POI data: A case study of central urban Wuhan, China[J]. International Journal of Environmental Research and Public Health, 2022, 19(9): 5153.

[21] Aleksander G P, Yifan T, Fuming Z. Predicting service life of polyethylene pipes under crack expansion using “random forest”Method[J]. International Journal of Engineering, 2023, 36(12): 2243-2252.

[22] Viscarra Rossel R A, Behrens T, Ben-Dor E, et al. A global spectral library to characterize the world’s soil[J]. Earth-Science Reviews, 2016, 155: 198-230.

[23] Zhang G L, Shi Z, Zhu A X, et al. Progress and perspective of studies on soils in space and time[J]. Acta Pedologica Sinica, 2020, 57(5): 1060-1070. [张甘霖, 史舟, 朱阿兴, 等. 土壤时空变化研究的进展与未来[J]. 土壤学报, 2020, 57(5): 1060-1070.]

（**责任编辑：**檀满枝）