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

徐 佳[1,2], 刘 峰[1*], 吴华勇[1], 宋效东[1], 赵玉国[1,2], 张甘霖[1,2,3]

（1. 中国科学院南京土壤研究所土壤与农业可持续发展国家重点实验室，江苏 南京 210008；2. 中国科学院大学，北京 100049；
3. 中国科学院南京地理与湖泊研究所流域地理学重点实验室，江苏 南京 210008）

摘 要：土壤与其发生环境密切相关。如何利用土壤属性准确地推测环境要素的信息，是法庭土壤学的重要研究问题。本文以我国东部 4 省 2 市（北京、天津、河北、山东、安徽和江苏）为研究区，基于 746 个土壤表层样本的理化性质和光谱数据构建特征，使用人工神经网络和随机森林两种机器学习模型对海拔高度、年均温、年均降雨量和地表温度四个关键环境要素进行预测，并对两种模型的预测准确度进行了对比分析。结果显示：两个模型对四个目标环境变量的预测准确度 R^2 在 0.39 ~ 0.61 之间；与神经网络模型相比，随机森林模型能够解释的环境变量的空间变异分别提高了 9.9%、16.5%、10.3%、10.9%；同时发现，对海拔高度和降雨的预测效果要优于其他环境要素。这表明，利用机器学习的方法可以有效地从土壤属性反推其成土环境条件的信息，这为法庭土壤物证研究学中未知土壤样本的来源地范围识别提供了技术参考。

关 键 词：法庭土壤学；神经网络；随机森林；环境要素；土壤属性

中图分类号：S159.2　　**文献标识码：**A　　**文章编号：**0564−3945(2021)02−0269−10

DOI: 10.19336/j.cnki.trtb.2020090601

徐 佳, 刘 峰, 吴华勇, 宋效东, 赵玉国, 张甘霖. 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

XU Jia, LIU Feng, WU Hua-yong, SONG Xiao-dong, ZHAO Yu-guo, ZHANG Gan-lin. 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

当前土壤地理学研究的主要方向之一是如何从成土环境信息推测土壤类型和属性信息，使用的方法有多元线性回归[1−3]、地统计学[4−5]、模糊聚类[6]、人工神经网络[7]、随机森林[8−9] 等。例如，Liu 等[10] 利用气候、母质、地形等成土环境因素信息对全国土壤质地进行了预测性制图；Liang 等[11] 利用气候、植被、地形等成土因素绘制了全国土壤有机碳的空间分布图；黄魏等[12] 利用成土母质、地形等环境因子对土壤类型进行了推理制图；Ramcharan 等[13] 利用地形、气候、植被等环境因子预测了美国土壤有机碳、容重、pH 等属性和土壤类型；Viscarra Rossel 等[14] 根据气候、生物、地形等环境因素对澳大利亚土壤质地进行了预测性制图。

然而，通过土壤属性推测环境要素信息的研究鲜有报道。法庭土壤物证研究迫切需要挖掘土壤样本的来源地范围信息[15−16]。2007 ~ 2017 年十年间，仅地方疑难案件送交公安部相关单位的涉土检测和勘验案件数量逐年增加。2016 年，由公安部物证鉴定中心牵头，包括中国科学院南京土壤研究所在内的 9 家单位联合成立了现场物证溯源国家工程实验室，土壤物证的检测溯源是其重要研究方面之一。由于土壤样本往往数量少、空间分布不均、单独依靠有限数量的样本难以进行有效的溯源，而环境数据较易获取，且空间全覆盖，因此土壤样本的成土环境要素更易与地理空间位置建立直接的联系，可以从土壤属性逆推其成土环境要素，再推测地理空间位置，可为土壤物证的来源地溯源提供技术支撑，有助于缩小侦查范围，提高效率，减少时间和经济成本，这对于法庭土壤学研究具有重要意义。

土壤景观关系理论为从土壤属性逆推成土环境

收稿日期：2020−09−06；**修订日期：**2021−02−22

基金项目：国家重点研发计划（2018YFC1800104、2017YFC0803807）及国家自然科学基金项目（42071072）资助

作者简介：徐 佳（1997−），女，安徽省安庆市人，硕士，主要从事数字土壤制图研究。E-mail: xujia@issas.ac.cn

*通讯作者：E-mail: fliu@issas.ac.cn

条件提供了基础。土壤是气候、生物、地形、母质、时间等环境因素长期综合作用的产物[17]。不同的环境条件，发育的土壤往往也是不同的。土壤与环境之间存在复杂的关系，尤其对于大面积区域，各成土因素的综合作用影响着土壤属性的空间分布，因而土壤属性的分布情况又能在一定程度上反映环境条件的差异。由此，土壤属性信息有潜力用于推断样本的发育环境条件的信息。

本文旨在对这一潜力进行检验，探讨从定量土壤属性信息（常规实验室土壤理化分析和室内土壤光谱测定数据）推测关键成土环境要素信息的技术方法，以我国东部自然条件较为复杂的大面积区域为研究区，比较不同机器学习方法（随机森林和人工神经网络）的建模与推测效果。一方面，机器学习方法相较于一般的统计学方法通常具有更好的模型预测效果[18]。另一方面，土壤属性和光谱特征与成土环境之间具有密切的发生学关系。土壤光谱具有分辨率高、测量快速、无损等优点，能够综合反映土壤属性信息，如土壤质地、pH、阳离子交换量、有机质、全氮、全碳、矿物组成、含水量、粒径以及土壤颜色等，已广泛应用于土壤科学研究中[19−21]，其所含的丰富的土壤信息，能够很好地指示各环境要素对土壤的作用，进而能够对土壤的成土环境特征进行推测。如 600 ~ 800 nm 波段是土壤有机质含量的光谱响应波段[22]，可利用该波段范围内的反射特性来进一步推出对土壤有机质含量影响较大的植被信息[9]。

1 材料与方法

1.1 研究区概况

研究区主要位于我国东部地区，如图 1 所示，包括京津冀、山东省、安徽省和江苏省，地理坐标范围在 113°27′ ~ 122°42.3′ E，29°41′ ~ 42°40′ N 之间，属于东部沿海地区，人口密度大，涉及土壤物证的案件较多。研究区内环境因素在地理空间上分布的差异较大，地貌类型多样，有平原、丘陵和山地等类型，海拔−12 ~ 2661 m，跨中温带、南温带和北亚热带气候区，气温由南向北逐渐降低，降水量由南向北逐渐减少，成土母质以细、粉砂质河流冲积物为主，土地利用类型 55% 是耕地，主要是水田和旱地。由此形成的土壤类型丰富复杂，空间异质性较强，其中北部河北省地处华北腹地，内嵌北京、天津两市，主要土壤类型有褐土、潮土、棕壤和栗钙土；中部山东省的土壤类型以潮土、褐土和棕壤为主；南部安徽省和江苏省主要分布有潮土、水稻土、砂姜黑土、棕壤、黄褐土、黄棕壤、滨海盐土、红壤等土类[23]。

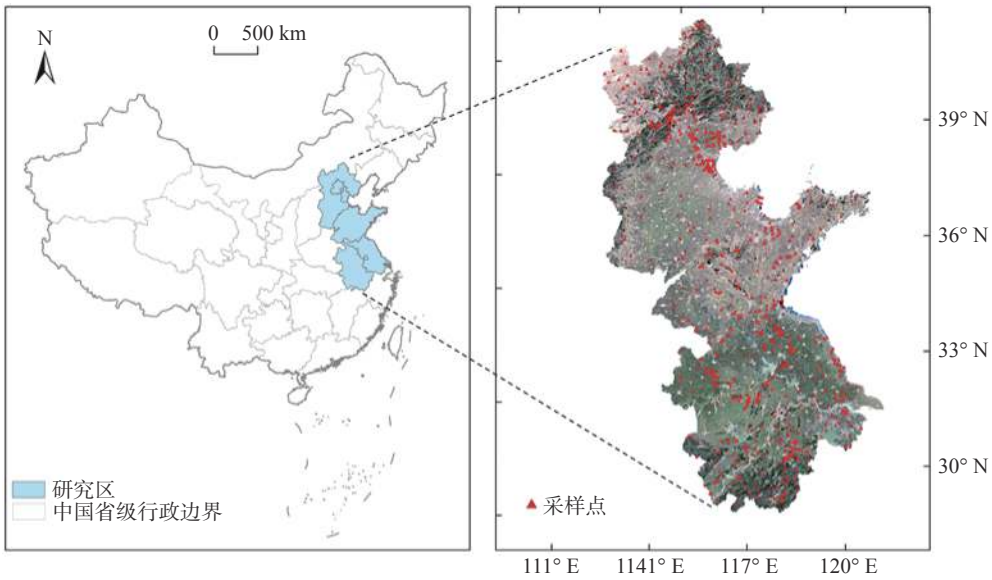


图 1 研究区地理位置及样点分布

Fig.1 Location of the study area and the distribution of sampling sites

1.2 土壤样本数据

土壤样本数据来源于科技部基础专项“我国土系调查与《中国土系志》编制”项目，包括各省主要土壤景观的典型土壤剖面。本研究区内有 746 个土

壤剖面样点，其中河北省 127 个样点，北京市 84 个样点，天津市 70 个样点，山东省 141 个样点，安徽省 185 个样点，江苏省 139 个样点，分别取每个样点的表层土壤样品。在实验室内，将采集的土壤样品自然风干，然后研磨、过筛，再对土壤的一些性质进行常规实验室理化测定。土壤 pH 值采用 pH 计测定；土壤有机质采用重铬酸钾—外加热法测定，除以土壤有机质含碳率 1.724，得到土壤有机碳含量（SOC）[24]。

同时，在实验室内利用 Cary 5000 分光光度计测定经过研磨、过 60 目筛后的土壤样本的光谱信息。采集的光谱波段范围在 350 ~ 2500 nm 之间，采样间隔为 1 nm，每个土壤样品的输出波段数为 2 151 个。光谱测量在密闭环境内进行，控制温度在 20° ~ 25°，空气湿度在 45% ~ 50% 范围内，测试前先校正白板，每个土壤样品采集两次，取平均值后得到该样品的光谱数据[21]。

对光谱数据的预处理包括三部分：（1）去除噪声较大波段：由于仪器在开始和结束测定时内部环境不稳定，土壤样品的原始光谱的边缘波段噪声较大，这些噪声会影响模型预测准确度，因此需要将 350 ~ 399 nm 和 2401 ~ 2500 nm 之间的光谱曲线予以去除。（2）主成分分析：土壤光谱数据具有数据量大、分辨率高的特点，但是数据冗余度大、计算量大，需要对其进行降维处理[25]。主成分分析是通过正交变换，以少数不相关变量代替原始可能存在相关性的多维变量的一种数据降维方法[26]。本研究采用主成分分析法对去除噪声后的 400 ~ 2400 nm 之间的光谱数据进行数据降维，在原始光谱数据的基础上提取了 7 个主成分（FAC1 ~ FAC7）。（3）提取光谱吸收峰特征参数：由于外界环境和仪器设备的误差等因素的影响，我们测量的光谱信号往往含有大量噪声，因此首先采用小波去噪法对光谱进行去噪。接着采用包络线去除法，将土壤反射光谱曲线上的吸收谷归一化到吸收谷的包络线上。包络线去除法的优点是能够扩大弱吸收波段的特征信息，有效突出土壤光谱的吸收和反射特性。经包络线去除后，根据吸收峰的波段宽度，找出每个土壤样品的光谱曲线中前 8 个波段范围最宽的吸收峰，并提取每个吸收峰的起、止位置（start_loc、end_loc）、宽度（width）、最大深度（depth）、最大深度处对应的位置（depthLoc）、最小反射率（R_min）、吸收峰的面积（area）、偏度（skew）、峰度（kurt）、斜率（s_trendline）。

将采集得到的土壤样本的 2 个土壤属性、7 个土壤光谱主成分和 80 个光谱曲线吸收峰的特征参数共计 89 个指标作为推测成土环境要素的辅助变量。

1.3　环境变量数据

环境变量的选取是以土壤发生学理论为依据，从成土因素中选出具有代表性的环境因素。由于母质和时间目前尚无统一的定量指标来描述，而地形和气候数据较易获取，且六月份平均的白天地表温度可近似代表平均最高气温这一气候指标，因此本研究从地形和气候类因子中选出了在地理环境方面比较关键的四个环境变量，分别是高程、年均温、年均降水量和地表温度（见图 2）。其中，高程数据（DEM）来自于资源环境数据云平台（http://www.resdc.cn/）；年均温（Tem）和年均降水量（Pre）数据来自于全球气候网站（http://chelsa-climate.org/）；地表温度（Lst）数据来源于美国地质调查局（USGS）官网（http://glovis.usgs.gov/），为长期（1970 ~ 2000 年）平均 6 月白天地表温度。所有环境变量均经栅格重采样至分辨率为 1 km。

1.4　模型构建及效果评价方法

分别采用神经网络和随机森林方法构建模型，对环境要素进行预测，并比较分析了两种方法的预测准确度。

BP 神经网络（Back propagation neural network，BPNN）是一种基于误差反向传播的多层前向反馈网络，由输入层、隐藏层 输出层组成，学习过程分为正向传播和反向训练[27−28]。把土壤属性、光谱信息作为输入信号，经由隐藏层处理后传向输出层，输出层给出预测值并与实测值进行比较，如果与实测值不符，则转向误差的反向传播阶段，在此过程中通过不断调整各神经元的权值，使得误差满足设定的阈值为止[27−28]。

在 R 软件环境下，从 746 个样本中随机选出 187 个作为测试集，剩下的 559 个样本作为训练集，分别对四个环境变量构建神经网络预测模型。采用 3 层神经网络，隐藏层神经元的激活函数设为 tansig，输出层传递函数设为 purelin，学习率为 0.001，通过对循环次数、隐藏层的神经元数、训练方法和训练次数的调整，比较模型预测结果的决定系数 R^2，不断优化模型，最终得到对各环境变量预测效果较好

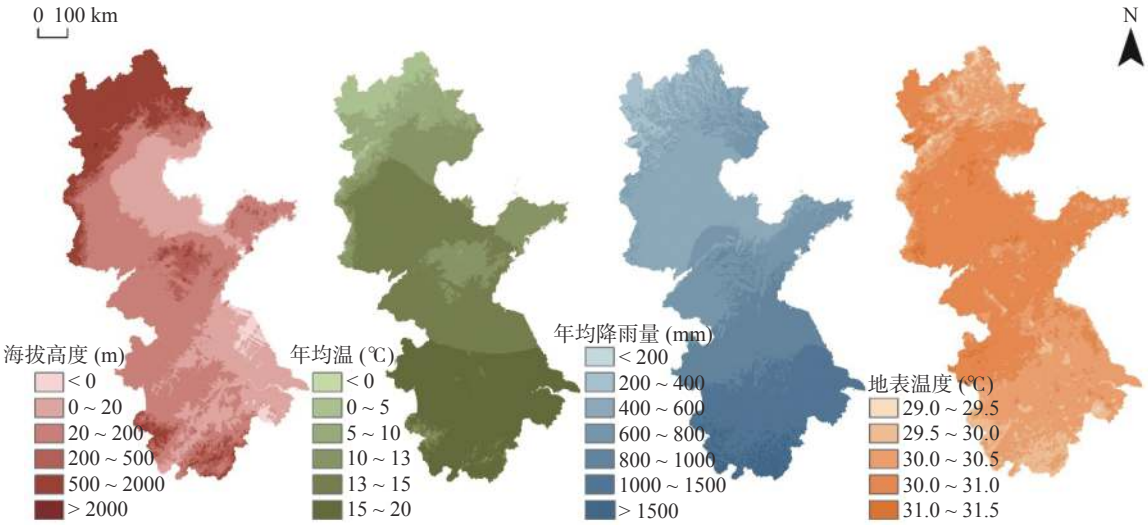


图 2 环境变量的空间分布

Fig.2 Spatial distribution of environmental variables

的 BP 神经网络，各变量的模型主要参数设置见表 1。

随机森林（Random Forest，RF）是一种将不同决策树组合到一起的集成学习算法，在构成随机森林的若干决策树中，每棵树都是基于随机样本产生的的一个独立集合，每棵树进行独立地学习和预测，最终通过所有决策树投票结果的均值来决定最终的结果[29]。本研究在 R 语言中通过调用 randomForest 包来建模，并通过不断调整节点处随机抽取的变量个数 mtry 和决策树数量 ntree 两个参数来优化模型，最终的模型参数设定见表 2。

表 1 各环境变量的 BPNN 模型参数

Table 1 BPNN model parameters of environmental variables

目标变量 Target variable	循环次数 Cycle times	隐藏层神经元数 Number of neurons	训练方法 Training method	训练次数 Training times
海拔高度	50	24	ADAPTgd	10000
年均温	100	28	ADAPTgd	1000
年均降水量	50	32	ADAPTgd	100
地表温度	100	45	ADAPTgd	1000

表 2 各环境变量的 RF 模型参数

Table 2 RF model parameters for environmental variables

目标变量 Target variables	节点变量数 Mtry	决策树数量 Ntree
海拔高度	3	1000
年均温	3	500
年均降水量	3	1000
地表温度	3	800

本文将土壤 pH、有机质含量、土壤光谱主成分和光谱曲线吸收峰的特征参数作为辅助变量，通过建立模型来推测成土环境信息。在提取的 89 个辅助变量数据集中，由于不同环境要素对土壤属性的影响程度不同，所以土壤属性对推测各环境信息的贡献度也有所差异，故有必要对不同环境要素预测模型的辅助变量进行筛选。通过计算随机森林模型的袋外误差来对辅助变量进行筛选，从辅助变量数据集中逐个剔除自变量，并观察袋外误差的变化，若误差增大则保留该自变量，反之则剔除，从而确定每个环境要素预测模型的最佳辅助变量组合。其中 Elev 的辅助变量组合是 SOC、光谱主成分 FAC1 ~ FAC7、吸收峰参数 width1、width2、area2、stat_loc5（注：吸收峰参数后的数字 n 代表第 n 个吸收峰的参数。）；Tem 的辅助变量组合是 pH、SOC、光谱主成分 FAC1 ~ FAC7、吸收峰参数 width1、area1、width2、area2、end_loc5；Pre 的辅助变量组合是 pH、SOC、光谱主成分 FAC1 ~ FAC7、吸收峰参数 width1、depth1、R_min1、area1、width2、depth2、area2；Lst 的辅助变量组合是 pH、SOC、光谱主成分 FAC1 ~ FAC7、吸收峰参数 start_loc1、depth1、depthLoc1、R_min1、area1、area2。

从原始 746 个样本中随机选出 25%（187 个）作

为测试样本。通过计算均方根误差（Root mean square error, RMSE）、决定系数（R square, R^2）和一致性相关系数（Concordance correlation coefficient, CCC）来评价模型的预测准确度。其中，RMSE 越小，R^2、CCC 越大，预测准确度越高。

1.5　变量重要性的计算方法

随机森林模型的一个优点是，在模型训练的同时可以评估输入变量的相对重要性。变量重要性可以反映特征变量在模型中的相对贡献大小，随机森林模型通过袋外误差来评价各变量对模型的重要性，其基本算法是对其中某一个变量随机加入噪声后，通过袋外误差是否增大来判断该变量是否重要[30]。若袋外误差大幅度增加，说明该变量对模型的预测结果影响较大，重要性较高，反之则重要性较低。本文通过对不同土壤信息在各成土环境预测模型中的重要性进行排序，识别出对各环境因子预测模型贡献率较大的土壤信息指标。

2　结果与讨论

2.1　环境变量描述性统计

对 746 个土壤剖面样点所对应的海拔高度 Elev、年均温 Tem、年均降水量 Pre、地表温度 Lst 四个环境变量的描述性统计特征如表 3 所示。

表 3　环境变量的描述性统计

Table 3　Descriptive statistics of environmental variables

环境变量 Environmental variables	最小值 Minimum	最大值 Maximum	均值 Mean	标准差 Standard deviation	变异系数（%） Coefficient of variable	偏度 Skewness	峰度 Kurtosis
海拔高度（m）	−1	2044	215.19	407.44	189.34	2.46	5.29
年均温（℃）	0	17.8	13.0	3.68	28.36	−1.56	2.29
年均降水量（mm）	289.79	2286.48	800.70	333.02	41.59	0.84	0.30
地表温度（℃）	29.5	31.1	30.49	0.31	1.00	−0.56	0.14

从表 3 可以看出，海拔高度 Elev 表现出极强的空间变异性，这是因为研究区内地形复杂多样，造成各采样点处的海拔高度差异较大；年均温 Tem 和年均降水量 Pre 呈中等程度变异，是由于气温和降水在我国南北方空间分布的差异造成的；地表温度 Lst 变异程度较弱，是因为 6 月日照时间长，整个研究区白天地表温度均达到较高值，空间差异不大。环境变量的统计结果说明研究区内的环境因素总体空间变异性较强。

2.2　模型准确度对比

通过计算 187 个测试样本的 RMSE、R^2 和 CCC，综合比较两种方法对四个环境变量的预测准确度（表 4）。

表 4　BPNN 与 RF 对各环境变量的预测准确度对比

Table 4　Comparisons of the predictive accuracy of environmental variables between BPNN and RF models

环境变量 Environmental variable	模型 Model	决定系数 R^2	均方根误差 RMSE	一致性相关系数 CCC
海拔高度（m）	BPNN	0.510	255.6	0.698
	RF	0.609	228.3	0.750
年均温（℃）	BPNN	0.394	2.582	0.600
	RF	0.559	2.205	0.707
年均降水量（mm）	BPNN	0.510	240.7	0.668
	RF	0.613	214.0	0.729
地表温度（℃）	BPNN	0.404	0.246	0.558
	RF	0.513	0.222	0.633

从表 4 可以看出，与 BPNN 相比，使用 RF 模型预测的四个环境变量的 RMSE 均比 BPNN 小，表明 RF 模型的预测结果与真实值之间的偏差相对较小；从决定系数 R^2 来看，使用 RF 模型预测的环境变量的 R^2 均达到 0.5 以上，相比于 BPNN，对海拔高度、年均温、年均降水量和地表温度预测结果的 R^2 分别提高了 9.9%、16.5%、10.3%、10.9%，表明 RF 模型对各环境变量的空间变异的解释度均高于 BPNN 模型；RF 模型的一致性系数 CCC 与 BPNN 相比明显提高，表明 RF 模型的预测值与真实值之间具有更好的一致性。准确度评价的结果表明，相比于 BPNN 模型，RF 模型对环境因素的预测准确度有显著提高，

因此，RF 模型可以更准确地反映土壤与各环境要素之间复杂的关系。其原因可能是：由于随机抽取的训练样本难以泛化到整个研究区的真实环境数据集上，而 BPNN 模型是网络状模型，模拟大脑皮层神经元结构，网络拟合准确度很高，如果学习了过多的特殊样本，容易使模型对其他样本的反映有偏差，造成过拟合[31]；RF 模型是集合式树状模型，融合了大量决策树进行训练，可以减轻过拟合问题[32]，这说明集合式树状机器学习模型表现更佳。

在四个目标环境变量中，RF 模型对 DEM 和 Pre 的预测效果最好，R^2 分别为 0.609 和 0.613。这是由于不同环境因素变量作用于土壤属性的空间尺度不同[33]。本研究区地理范围较广，在大尺度研究区域内，地形和降水对土壤属性的影响较为突出，因此模型在定量分析地形和降水与土壤属性的关系时更为精确。

BPNN 模型和 RF 模型对四个环境变量分别进行预测的验证样本实测值与预测值之间比较的散点图如图 3 所示。图中点越接近 1∶1 线，表示该点的预测值与实测值越接近，即预测效果越好。

从图 3 可以看出，BPNN 和 RF 对 Elev 的预测效果在海拔较低的区域均优于高海拔地区，且两种模型的预测效果在较高海拔地区差别不大，而在低海拔地区，RF 模型的预测值与实测值更趋近；对 Tem 的预测效果，RF 模型明显优于 BPNN 模型，且对 Tem 较高的样点预测准确度高于 Tem 较低的样点；从 Pre 的预测结果可以看出，RF 模型的验证样本相对更集中于 1∶1 线附近，并且在降雨量低的地区样本更趋近 1∶1 线，而在降雨量高的地区样本分布相对较为分散，即在降雨量低的地区的预测准确度优于降雨量高的地区；而对地表温度 Lst 的预测，两种模型的预测效果差异不大，且在高值和低值区预测准确度也没有显著差异。从 Lst 的散点图中还可以看到，散点呈平行于 Y 轴分布，这是由于地表温度的数值范围集中分布在 29.5 ℃ 至 31.5 ℃，并且数据分辨率不高的缘故。总体而言，RF 模型对各环境变量的预测效果优于 BPNN 模型，并且在 Elev、Tem、Pre 的高值区和低值区预测效果有显著差异。结合研究区的高程数据，发现气温较低的采样点分布在河北省北部燕山地区，降雨量较高的采样点分布在皖南山区，即山区的样点预测准确度要低于平原地区。

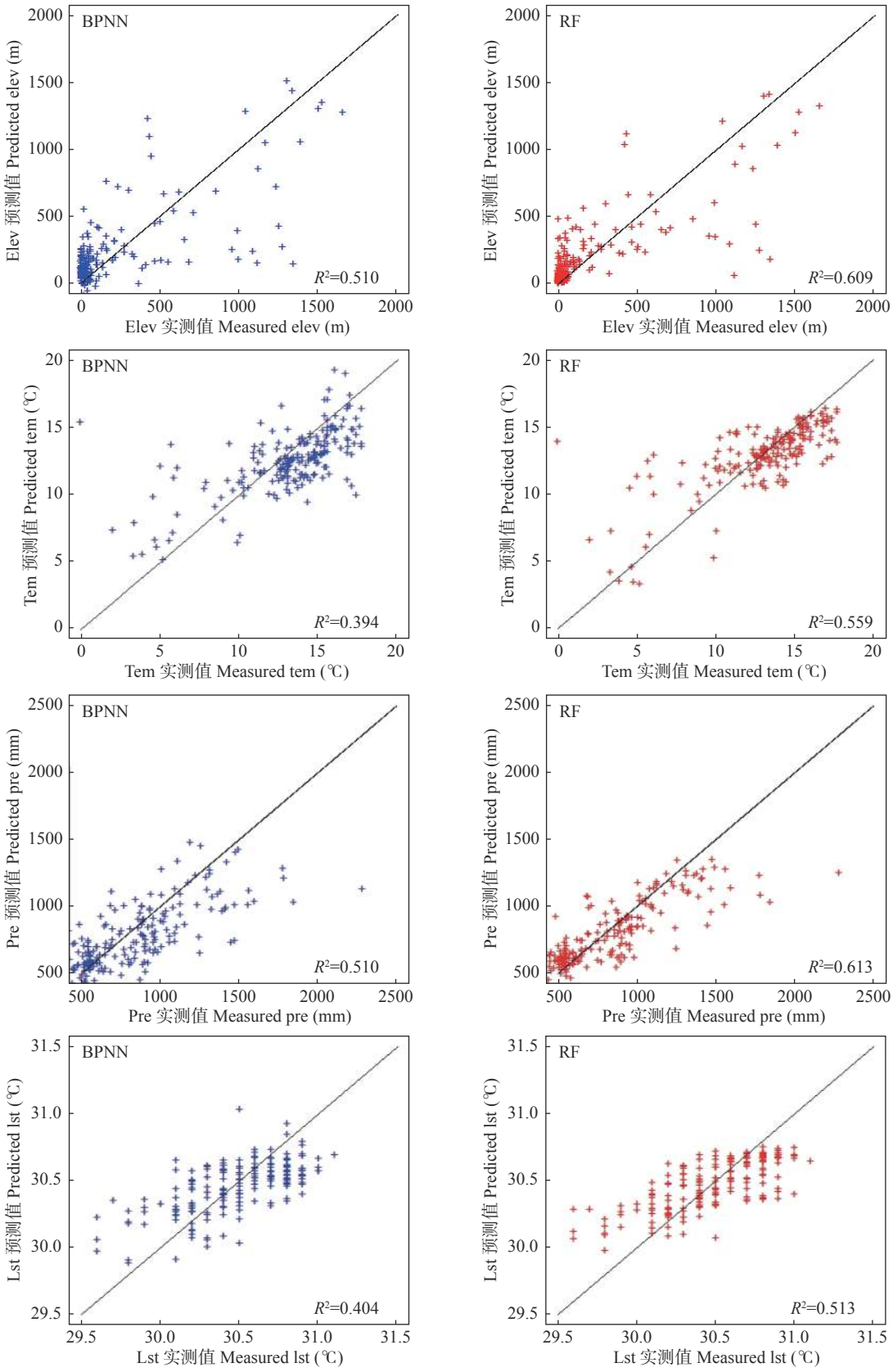
2.3 地形对预测准确度的影响

为进一步探究 RF 模型对 Elev、Tem、Pre 三个环境变量的预测准确度与地形之间的联系，将 RF 模型对 Elev、Tem、Pre 的预测值与实测值作差，其空间分布情况如图 4，蓝色调的样点代表预测值比实测值低，红色调的样点代表预测值比实测值高，颜色越深代表预测值与实测值的差值越大。从图中可以看出，各环境变量的预测值与实测值的差值空间分布呈现较强的规律性，颜色较深的点主要分布在河北的燕山、山东的泰山和皖南的黄山及丘陵地区（如图 4 中实线圈出部分），而颜色较浅的点主要分布在平原地区（如图 4 中虚线圈出部分），即 RF 模型在平原地区对环境变量的预测准确度明显优于山地和丘陵。其原因可能是地形起伏较大的丘陵和山区，各种环境因素变化差异较大，对土壤的作用较为复杂，土壤往往是由多种环境因素综合作用形成的，加上采样点较少，导致模型难以准确分析每一种环境因素与土壤属性之间的关系。另外，孙孝林等[34] 的研究表明，DEM 的分辨率影响土壤景观模型的准确度，尤其在山区，地形起伏大，低分辨率的 DEM 数据难以反映地形的局部差异，导致 DEM 在表达地形时的精确度受到一定限制，限制了模型准确度。可通过增加山区采样点、提高 DEM 分辨率等来提高模型在山区的预测准确度。

2.4 不同土壤属性对环境因子模拟的重要性分析

图 5 显示了各环境因子的预测模型中各土壤属性的重要性排序结果。从图中可以看出，对海拔高度和年均温的预测贡献率最高的是光谱第二主成分；对地表温度响应程度最高的是光谱第六主成分，说明光谱主成分是成土环境因子较为有效的协同变量，可以指示这些环境因子的空间变异。其原因在于土壤的反射特性与其内在结构密切相关，是其理化性质的综合反映，这在一定程度上间接地反映了气候、地形等环境因素对土壤的作用。以上结果说明，光谱主成分对环境因子预测的贡献率较大，在环境变量预测方面具有较大的潜力，今后可用于其他环境因子的预测模型中。土壤 pH 对年均降水量预测的贡献最高，南方地区降水相对充裕，淋溶作用较强，导致土壤酸性较高；而北部的河北等地降水量相对较少，淋溶作用不强，有少量碳酸钙沉积，导致土壤呈中性或微碱性。

尽管本文的建模研究取得了一定的效果，但仍存在一些问题需要深入研究：（1）Elev（图 4a）在海拔较高的山区预测值比实测值低，而在海拔较低



注: 图中“Elev”、“Tem”、“Pre”、“Lst”分别代表海拔高度、年均温、年均降水量、地表温度。

图 3　BPNN 和 RF 对各环境变量的实测值与预测值之间的比较

Fig.3　Comparison between the measured and predicted values of environmental variables by BPNN and RF

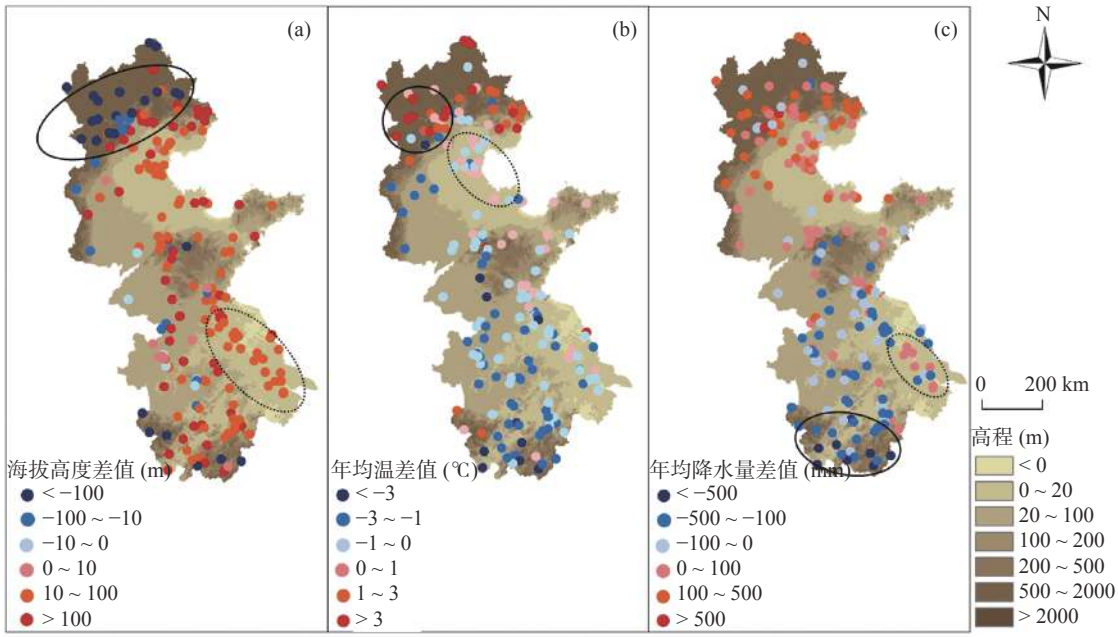


图 4 环境变量的预测值与实测值的差值空间分布

Fig.4 Spatial distribution of the difference between predicted and measured values of environmental variables

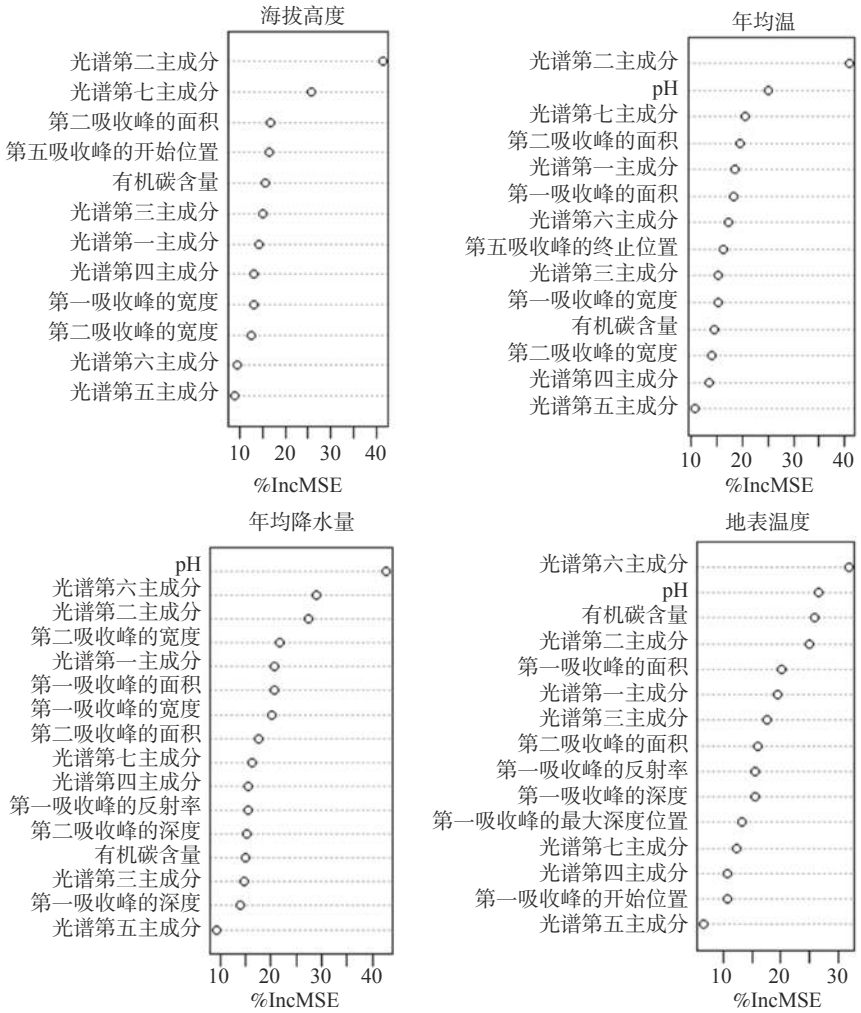


图 5 各土壤属性对各环境变量的重要性

Fig.5 The importance of various soil properties to various environmental variables

的平原地区预测值比实测值高；而 Tem（图 4b）和 Pre（图 4c）的预测误差均是以泰山为界，在泰山以北，气温较低，降水较少，Tem 和 Pre 的预测值均高于实测值，而在泰山以南，气温较高，降水较多，Tem 和 Pre 的预测值均低于实测值。可见，RF 模型对各环境变量的预测值呈现出“高值低估、低值高估”的现象，这是由于随机森林中将多棵树的结果取平均的算法，使预测值的值域比实测值的值域小，下一步应尝试其他模型方法提高模型预测准确度[35]。（2）本文侧重方法探讨，未对所有成土环境因素进行建模推测，进一步将增加对植被和母质等成土环境因素的预测，以较全面地获取土壤样本来源地的环境信息。（3）在空间范围上，本文仅在我国东部部分省市进行了建模分析，进一步可扩展至全国范围进行研究。（4）本文通过土壤属性反推成土环境信息在模型方法上进行了探讨，下一步将研究如何通过未知土壤样本的环境信息判断其空间来源，进而为土壤物证溯源提供技术参考。

3　结论

本文基于神经网络和随机森林模型研究了如何利用土壤信息准确推测环境要素信息的问题，以我国东部地貌和土壤条件复杂多样的大面积区域为研究区开展了研究，得到以下结论：（1）利用机器学习方法构建土壤与环境的关系，根据土壤属性反推环境信息是有效可行的；（2）相较于 BPNN 模型，RF 模型在预测各环境变量中均显示出了较好预测准确度，并且训练过程更加简单；RF 模型能更好地刻画土壤与环境之间复杂的关系，可以利用土壤属性解释 51% 以上的环境因素的空间变异，但是缩小了预测变量的值域范围，预测准确度有待进一步提高；（3）模型对海拔高度和年均降雨量的预测准确度比年均温度和地表温度要高，说明对于本案例区域，海拔高度和降雨的空间变异能更好地被解释。

参考文献：

[1] Moore I D, Gessler P E, Nielsen G A, et al. Soil Attribute Prediction Using Terrain Analysis[J]. Soil Science Society of America Journal, 1993, 57: 443 − 452.

[2] Thompson J A, Bell J C, Butler C A. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling[J]. Geoderma, 2001, 100: 67 − 89.

[3] 孙孝林, 赵玉国, 秦承志, 等. DEM栅格分辨率对多元线性土壤—景观模型及其制图应用的影响[J]. 土壤学报, 2008, 45: 971 − 977.

[4] Piccini C, Marchetti A, Francaviglia R. Estimation of soil organic matter by geostatistical methods: Use of auxiliary information in agricultural and environmental assessment[J]. Ecological Indicators, 2014, 36: 301 − 314.

[5] Zhang Y K, Ji W J, Saurette D D, et al. Three-dimensional digital soil mapping of multiple soil properties at a field-scale using regression kriging[J]. Geoderma, 2020, 366

[6] 赵　量, 赵玉国, 李德成, 等. 基于模糊集理论提取土壤—地形定量关系及制图应用[J]. 土壤学报, 2007: 961 − 967.

[7] 李启权, 王昌全, 张文江, 等. 基于神经网络模型和地统计学方法的土壤养分空间分布预测[J]. 应用生态学报, 2013, 24: 459 − 466.

[8] 齐雁冰, 王茵茵, 陈　洋, 等. 基于遥感与随机森林算法的陕西省土壤有机质空间预测[J]. 自然资源学报, 2017, 32: 1074 − 1086.

[9] 王茵茵, 齐雁冰, 陈　洋, 等. 基于多分辨率遥感数据与随机森林算法的土壤有机质预测研究[J]. 土壤学报, 2016, 53: 342 − 354.

[10] Liu F, Zhang G L, Song X D, et al. High-resolution and three-dimensional mapping of soil texture of China[J]. Geoderma, 2020, 361

[11] Liang Z Z, Chen S C, Yang Y Y, et al. High-resolution three-dimensional mapping of soil organic carbon in China: Effects of SoilGrids products on national modeling[J]. Science of the Total Environment, 2019, 685: 480 − 489.

[12] 黄　魏, 许　伟, 汪善勤, 等. 基于不确定性模型的土壤—环境关系知识获取方法的研究[J]. 土壤学报, 2018, 55(1): 54 − 63.

[13] Ramcharan A, Hengl T, Nauman T, et al. Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution[J]. Soil Science Society of America Journal, 2018, 82: 186 − 201.

[14] Rossel R A V, Chen C, Grundy M J, et al. The Australian three-dimensional soil grid: Australia’ s contribution to the GlobalSoilMap project[J]. Soil Research, 2015, 53: 845 − 864.

[15] Wald C. Forensic science: The soil sleuth[J]. Nature, 2015, 520: 422 − 424.

[16] Tighe M, Forster N, Guppy C, et al. Georeferenced soil provenancing with digital signatures[J]. Scientific Reports, 2018, 8: 3162.

[17] Jenny, Ha N S. Factors of Soil Formation[J]. Soil Science, 1941, 52(5): 415.

[18] 朱阿兴, 杨　琳, 樊乃卿, 等. 数字土壤制图研究综述与展望[J]. 地理科学进展, 2018, 37(1): 66 − 78.

[19] Rosero-vlasova O A, Vlassova L, Perez-cabello F, et al. Soil organic matter and texture estimation from visible-near infrared-shortwave infrared spectra in areas of land cover changes using correlated component regression[J]. Land Degradation & Development, 2019, 30: 544 − 560.

[20] Xu D Y, Ma W Z, Chen S C, et al. Assessment of important soil

properties related to Chinese Soil Taxonomy based on vis-NIR reflectance spectroscopy[J]. Computers and Electronics in Agriculture, 2018, 144: 1 − 8.
[21] 史 舟, 王乾龙, 彭 杰, 等. 中国主要土壤高光谱反射特性分类与有机质光谱预测模型[J]. 中国科学:地球科学, 2014, 44: 978 − 988.
[22] 纪文君, 史 舟, 周 清, 等. 几种不同类型土壤的VIS-NIR光谱特性及有机质响应波段[J]. 红外与毫米波学报, 2012, 31(3): 277 − 282.
[23] 龚子同, 黄荣金, 张甘霖. 中国土壤地理[M]. 北京: 科学出版社, 2014.
[24] 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.
[25] 杨诸胜. 高光谱图像降维及分割研究[D]. 西安: 西北工业大学, 2006.
[26] 傅 湘, 纪昌明. 区域水资源承载能力综合评价—主成分分析法的应用[J]. 长江流域资源与环境, 1999, 8(2): 168 − 173.
[27] 李 硕, 汪善勤, 张美琴. 基于可见—近红外光谱比较主成分回归、偏最小二乘回归和反向传播神经网络对土壤氮的预测研究[J]. 光学学报, 2012, 32(8): 297 − 301.
[28] 沈润平, 丁国香, 魏国栓, 等. 基于人工神经网络的土壤有机质含量高光谱反演[J]. 土壤学报, 2009, 46(3): 391 − 397.
[29] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5 − 32.
[30] 张 雷, 王琳琳, 张旭东, 等. 随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例[J]. 生态学报, 2014, 34(3): 650 − 659.
[31] 覃光华. 人工神经网络技术及其应用[D]. 四川: 四川大学, 2003.
[32] Claudia L, A B P, C I M, et al. Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1862 − 1874.
[33] 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.
[34] 孙孝林, 赵玉国, 赵 量, 等. 应用土壤—景观定量模型预测土壤属性空间分布及制图[J]. 土壤, 2008: 837 − 842.
[35] 李富富, 陈东湘, 王院民, 等. 基于随机森林与地统计预测城市土壤PAHs分布[J]. 中国环境科学, 2019, 39(12): 5240 − 5247.

Predicting of Key Environmental Factors from Soil Properties Based on Artificial Neural Network and Random Forest Learning Model

XU Jia[1, 2], LIU Feng[1*], WU Hua-yong[1], SONG Xiao-dong[1], ZHAO Yu-guo[1, 2], ZHANG Gan-lin[1, 2, 3]
(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. *Key Laboratory of Watershed Geographic Sciences*, *Nanjing Institute of Geography and Limnology*, *Chinese Academy of Sciences*, *Nanjing* 210008, *China)*

Abstract: Soil is closely related to its formative environment. How to use soil properties to accurately predict the associated environmental information is an important research problem in soil forensics. About 746 soil samples were selected from Beijing, Tianjin, Hebei, Shandong, Anhui and Jiangsu in eastern China. Four key environmental information (elevation, average annual temperature, average annual rainfall and surface temperature) were predicted based on basic soil properties and spectral data using two machine learning models (neural network and random forest). Root mean square error (RMSE), determination coefficients (R^2) and concordance correlation coefficient (CCC) were used to calculate the prediction accuracy. Results showed that the prediction accuracy of the two methods were between 0.39 and 0.61. Compared with the neural network model, the spatial variation of environmental variables using random forest model were increased by 9.9% (elevation), 16.5% (average annual temperature), 10.3% (average annual rainfall), and 10.9% (surface temperature). And altitude and rainfall in this study area showed a better prediction accuracy than the other environmental variables. This suggests that the machine learning methods can be effective for predicting environmental information based on soil properties. This study provided a technical support for identifying the source of unknown soil samples in soil forensics.

Key words: Soil forensics; Neural network; Random forest; Environmental factor; Soil attribute

[责任编辑：孙福军]