DOI: 10.13758/j.cnki.tr.2019.03.025

基于随机森林模型的安徽省土壤属性空间分布预测①

卢宏亮[1]，赵明松[1,2*]，刘斌寅[1]，张 平[1]，陆龙妹[1]

(1 安徽理工大学测绘学院，安徽淮南 232001；2 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)

摘 要：为探讨随机森林(random forest，RF)模型对土壤属性空间预测的精度，本文以安徽省为例，收集 140 个土壤样本，利用 GIS 和 RS 技术，获取相关的地形因子、遥感植被指数及气候数据，利用 RF 模型分析土壤有机碳(SOC)含量、土壤容重和土壤黏粒含量与地形因子、遥感植被指数及气候数据之间的关系，并进行空间分布预测。研究结果表明：①RF 建模预测中，当节点分裂次数(mtry)值为 1，决策树数量(ntree)值分别为 100、1 000 和 100 时，获得的 SOC 含量、土壤容重和土壤黏粒含量 RF 模型最优；②高程、归一化植被指数(NDVI)、地貌、多尺度山谷平坦指数(MrVBF)和土壤类型是 SOC 含量的重要预测因子；地貌、年均降水量(MAP)、MrVBF、高程和土壤类型是土壤容重的重要预测因子；高程、MAP、MrVBF 和平面曲率是土壤黏粒含量的重要预测因子；③RF 模型可以较好地进行土壤属性空间预测，多源环境变量组合可以分别解释 SOC 含量、土壤容重和土壤黏粒含量的 26%、23% 和 22%；同时 RF 模型对于土壤类型和地貌等类型变量的处理具有一定优势。研究表明，在大尺度研究区域内，利用 RF 模型进行土壤属性空间预测有一定的意义。

关键词：土壤属性预测；随机森林模型；环境变量；安徽省

中图分类号：S159 文献标识码：A

土壤有机碳(soil organic carbon，SOC)是陆地生态系统平衡的主要因子[1]，研究 SOC 含量的空间分布及其影响因素是陆地生态系统碳循环的基础。土壤容重(bulk density，BD)是土壤的基本物理性质之一，对土壤的透气性、入渗性能、持水能力、溶质迁移特征以及土壤的抗侵蚀能力有重要影响。土壤黏粒是土壤中最活跃的矿物成分[2]，研究不同阶段的土壤黏粒含量可以得出可靠的土壤相对年龄。研究上述土壤属性的空间变异及其分布特征和环境因子的关系，对于了解生态系统、制定农业政策、进行土壤管理和监测由于土地利用导致的环境变化有重要意义。

基于机器学习方法预测土壤属性的空间分布逐步成为近年来的研究热点。文献研究表明，已有研究中使用的主要机器学习技术有分类和回归树(classification and regression tree，CART)[3]、k 最近邻(k-nearest neighbors，KNN)算法[4]、树模型[5]和随机森林(random forest，RF)模型[6]等：RF 模型与大多数统计建模方法相比具有一些优势，它具有对多元共线性不敏感和不易出现过拟合问题等特点[6]，且在噪声和数据简化处理方面最准确和稳定[7]。国外研究中，Dharumarajan 等[8]利用 RF 模型对印度南部半干旱热带地区的 SOC、土壤 pH 等属性进行了预测，结果证明 RF 模型可以提高土壤属性空间预测的精度。Chagas 等[9]比较了 RF 模型和多元线性回归方法在半干旱地区土壤质地的空间预测制图的效果，结果表明 RF 模型可以避免过拟合且预测精度更高。国内的研究中，郭彭涛等[10]基于多源环境变量和 RF 模型预测了橡胶园土壤全氮含量的空间分布，结果证明 RF 模型相较于逐步回归、广义加性混合模型和分类回归树等模型具有更高的预测性。姜赛平等[11]比较了普通克里格、回归克里格、RF 等模型在海南岛土壤有机质(SOM)的预测研究中的精度，结果表明 RF 和回归克里格模型能够更好地描述 SOM 的局部变异信息。RF 模型最合适进行土壤属性空间的预测。研究 RF 模型在土壤属性空间预测中的应用对数字土壤制图具有一定的意义。

①基金项目：国家自然科学基金项目(41501226)、安徽省高校自然科学研究项目(KJ2015A034)和土壤与农业可持续发展国家重点实验室开发基金项目(Y412201431)资助。

* 通讯作者(zhaomingsonggis@163.com)

作者简介：卢宏亮(1993—)，男，安徽铜陵人，硕士研究生，主要从事数字土壤制图研究。E-mail: 15656232332@163.com

安徽省气候差异明显、地貌类型众多、土地利用存在明显的区域差异，这些条件的组合导致了多种环境因子共同影响土壤属性的空间分布及变异。本研究以安徽省为例，利用 GIS 和 RS 技术提取土壤景观环境因子，通过收集土壤野外采样数据，以 SOC 含量、土壤容重和土壤黏粒含量为预测目标，运用方差分析和相关性分析研究环境变量与预测目标的关系，通过 RF 建模选择最优环境变量组合和模型参数建立土壤属性的预测模型并进行空间分布预测，同时探讨省域尺度上 3 种土壤属性的主要影响因素。研究结果有望为安徽省生态系统研究、土壤质量管理和农业生产提供数据基础。

1　材料与方法

1.1　区域概况

安徽省(114°54' ~ 19°37'E，29°41' ~ 34°38'N)地处我国东部，跨长江、淮河中下游，东临以上海为中心的长江三角洲经济区，西接中原腹地。安徽省总面积 13.96 万 km^2，其中农田 69%，低山丘陵 14%，湖泊 17%。全省处于亚热带向温带过渡带。年均气温 14 ~ 16 ℃，年均降雨量 800 ~ 1 800 mm。除安徽西南和南部丘陵地区外，海拔一般不超过 100 m。安徽省从北到南分为淮河中游平原、江淮丘陵岗地、沿江平原、皖西大别山区、皖南丘陵地区等 5 个地理区域。主要的土壤类型有：水稻土、潮土、砂姜黑土、黄棕壤、黄褐土、棕壤、红壤、黄壤、紫色土、石质土、粗骨土、石灰土、山地草甸土等。

1.2　数据来源

本研究数据包括野外土壤调查数据、气候数据、地形数据和遥感植被指数。土壤调查数据来源于《中国土系志•安徽卷》[12]，该调查数据按照随机性、均匀性和代表性的原则在安徽省全省范围采集典型土壤剖面，数据集包含采样点位置、景观条件和土壤理化性质，采样时间为 2010—2011 年。本研究选择 140 个样点的表层土壤属性为预测的目标变量，包括 SOC 含量、土壤容重及土壤黏粒含量。气候数据主要包括：年均温(MAT)、年均降水量(MAP)。数据来自中国农业科学院农业资源与农业区划研究所中国生态环境背景层面建造项目完成的栅格数据(1 km 分辨率)，为 1980—1999 年的逐月平均值计算生成。在 ArcGIS 支持下，从上述环境变量的栅格数据中提取各样点的相应环境属性。地形数据来源于地理数据空间云(http://www.gscloud.cn) 的 SRTM 数字高程模型(DEM)，空间分辨率为 90 m。利用 ArcGIS 10.2 提取坡向、坡度、高程、平面曲率和剖面曲率；利用 SAGA GIS 6.3.0 提取多尺度山谷平坦指数(MrVBF)、多尺度脊顶平坦指数(MrRTF)、地形湿度指数(TWI)及地形位置指数(TPI)，其中坡向数据由 DEM 数据产品中的 SRTMTPI 坡位产品直接提取。归一化植被指数(NDVI)和增强植被指数(EVI)来源于 MODIS 陆地产品 16 d 合成植被指数(MOD13Q1)，空间分辨率为 250 m，时间为 2010 年 8 月。所有环境变量及土壤属性空间预测结果，分辨率统一为 250 m。

1.3　随机森林模型

随机森林(random forest，RF) 模型具有提高预测精度、减少过拟合、对缺失数据和多元共线性不敏感，且具有简单处理大量的定量和定性数据能力的优点[13]。对于土壤类型和地貌等类型变量，多数回归模型处理方式比较复杂，一些研究甚至找不到适合的定性指标进行定量化描述[14]，在 R 软件中编程建立的 RF 模型只需将定性变量转为因子(factor) 直接用于模型即可。

本研究使用 R 语言中的 Random Forest 4.6 软件包进行建模。140 个样点按 8:2 分为建模集和验证集。RF 模型采用 boostrap 的方法对于样本进行放回抽样。没有被抽取的记录会自动生成一个对照集，所以不需要进行交叉验证[6]。RF 模型中的两个可调参数决策树数量(ntree)和节点分裂次数(mtry)决定了模型的配置。

1.4　数据处理与分析

利用 SPSS 22 for windows 进行方差分析和相关性分析，研究环境变量对于土壤属性的影响[15]。对于定量环境变量，在 R 软件中，使用 scale() 函数进行归一化处理后，利用 SPSS 进行相关性分析，将 MAP、MAT、地貌和土壤类型进行方差分析，其中 MAP 分为 <800 mm、800 ~ 900 mm、900 ~ 1 000 mm、1 000 ~ 1 100 mm、>1 100 mm 5 个降雨带，MAT 分 8 ~ 10 ℃、10 ~ 12 ℃、12 ~ 14 ℃、14 ~ 16 ℃、>16 ℃ 5 个温度带，地貌数据根据安徽省地貌类型矢量图在 ArcGIS 中提取至采样点，根据 5 个地貌区域划分，在 SPSS 中以 1 ~ 5 重新编码后进行方差分析，土壤类型数据根据已有土壤数据在 SPSS 中以 1 ~ 13 重新编码后进行方差分析。

1.5　精度评价

模型精度评价选用均方根误差(RMSE)、平均绝对误差(MAE)[16]以及决定系数(R^2)[17]3 个指标，其中 MAE 和 RMSE 越小表明预测精度越高，建模集 R^2 用于评价建模的拟合精度，验证集 R^2 用于评价预测精度及模型泛化能力。计算方法如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i-o_i)^2} \quad (1)$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|o_i-p_i| \quad (2)$$

$$R^2=1-\frac{\sum_{i=1}^{n}(p_i-o_i)^2}{\sum_{i=1}^{n}(p_i-\hat{o}_i)^2} \quad (3)$$

式中：p_i 和 o_i 为预测值和观测值，$\hat{o}_i$ 是观测值的平均值。

2 结果与讨论

2.1 土壤属性统计特征

表 1 为安徽省土壤属性统计结果。SOC 含量介于 1.33 ~ 33.53 g/kg，平均含量为 14.57 g/kg；中等变异，变异系数为 52.06%。土壤容重范围为 0.59 ~ 1.56 g/cm³，中等变异，变异系数为 11.68%。土壤黏粒含量范围为 42.73 ~ 552.76 g/kg，中等变异，变异系数为 47.43%。

表 1 安徽省土壤属性基本统计特征
Table 1 Statistical characters of soil properties in Anhui Province

土壤属性	范围	均值	标准差	偏度(%)	峰度(%)	变异系数(%)
SOC (g/kg)	1.33 ~ 33.53	14.57	7.22	0.55	−0.29	52.06
BD (g/cm³)	0.59 ~ 1.56	1.25	0.15	–0.83	1.86	11.68
Clay (g/kg)	42.73 ~ 552.76	212.91	101.00	0.82	0.46	47.43

注：SOC：土壤有机碳；BD：土壤容重；Clay：土壤黏粒，下表同。

2.2 土壤属性的影响因素分析

相关性分析结果(表 2)表明，坡向、高程、MrRTF 和 MAP 与 SOC 含量显著相关(P<0.05)；土壤容重与 NDVI、坡向、高程、TPI、MrVBF、MrRTF、MAP 和 MAT 显著相关(P<0.05)；土壤黏粒含量则和 NDVI、EVI、坡度、MrRTF、MrVBF 和 MAP 都具有显著相关性(P<0.05)。

表 2 土壤属性与环境因子的相关性分析
Table 2 Correlation between soil properties and environmental factors

变量	SOC	BD	Clay
NDVI	0.157	–0.224**	–0.24**
EVI	0.028	0.060	0.282**
坡度	0.104	–0.101	–0.191*
坡向	–0.205*	0.172*	0.056
高程	0.242**	–0.267**	–0.132
剖面曲率	0.124	–0.166	–0.112
平面曲率	0.069	–0.012	0.067
TWI	0.119	–0.030	0.092
TPI	–0.147	0.201*	0.055
MrVBF	–0.155	0.193*	0.28**
MrRTF	0.196*	0.226**	0.222**
MAP	0.209*	–0.261**	–0.301**
MAT	–0.113	0.170*	0.021

注：*表示相关性达到 P<0.05 显著水平，**表示相关性达到 P<0.01 显著水平(双尾)。

方差分析结果(表 3)表明[18]，不同的 MAT、MAP 和土壤类型的 SOC 含量和容重均存在显著差异(P<0.05)，其他各因子对两者的变异性均有显著影响，地貌对于 SOC 含量的变异性没有显著影响(P = 0.18)；对于土壤黏粒含量，因子影响均显著。对于 SOC 含量，土壤类型的 F 值最大，说明土壤类型对 SOC 含量和容重的变异性影响最大；影响土壤黏粒含量变异性的最重要因素为 MAP。

表 3 安徽省各因子影响土壤属性的方差分析
Table 3 Variance analysis of effects of different factors on soil properties in Anhui Province

影响因子	SOC		BD		Clay	
	F	P	F	P	F	P
MAT	2.64	0.03	4.12	0.00	2.72	0.03
MAP	2.73	0.03	4.39	0.00	5.54	0.00
土壤类型	3.51	0.00	6.07	0.00	2.23	0.01
地貌	1.60	0.18	3.9	0.00	4.73	0.00

2.3 环境变量的筛选及重要性分析

利用 RF 模型进行变量重要性排序，对重要性较低的环境变量进行排除后重复建模，选取最优环境变量组合用于预测。最终确定高程和 NDVI 等 8 个环境变量作为自变量预测 SOC 含量；地貌、MAP 和土壤类型等 9 个环境因子作为土壤容重的预测因子；高程和 MAP 等 8 个环境因子用于土壤黏粒含量的预测。相关性分析和方差分析结果表明，NDVI 及地貌与 SOC 含量的相关性并不显著，RF 模型的重要性分析却表明 NDVI 和地貌是影响 SOC 含量重要的环境因素，这是由于 RF 模型对多元共线性不敏感。在进行土壤属性预测变量筛选时，应该结合土壤学专业知识选取。

预测因子重要性排序表明(图 1)，RF 模型以增长均方误差(increased in mean squared error，IncMSE)为变量重要性衡量指标，该值越大则变量重要性最高。对于 SOC 含量，高程、NDVI 和地貌等为主要影

响因子，高程和 NDVI 影响最大。有研究表明[8]，NDVI 与 SOC 含量呈极显著正相关关系，即该指数越大，SOC 含量越高，所以 NDVI 在 SOC 预测模型中为主要影响因子。影响土壤容重的环境因子中，地貌为最主要的影响因子。容重主要受土壤质地、结构的影响，不同地貌的土壤质地和结构区别显著，所以在影响因子重要性排序中地貌为首要因素。对于土壤黏粒含量，高程、MAP、MrVBF 和平面曲率是主要的预测因子。

2.4　随机森林模型参数设定

本文通过逐次试验，确定 RF 模型中 mtry 和 ntree 参数的最优值[19]。固定 mtry 值(分别设为 1、2 和 3)，逐次调整 ntree 值(分别设为 100、500 和 1 000)，进行 3 组 9 次试验。为避免过拟合问题，本文通过比较建模集和验证集的 R^2，选择两者最为接近的结果作为最优预测模型。结果表明(表 4)，当 mtry 值为 1，ntree 值为 100 时，SOC 预测模型的建模集和验证集 R^2 最为接近，表明此时的模型稳定性最好；当 mtry 值为 1，ntree 值分别为 1 000 和 100 时，容重和土壤黏粒含量的预测模型最为稳健。

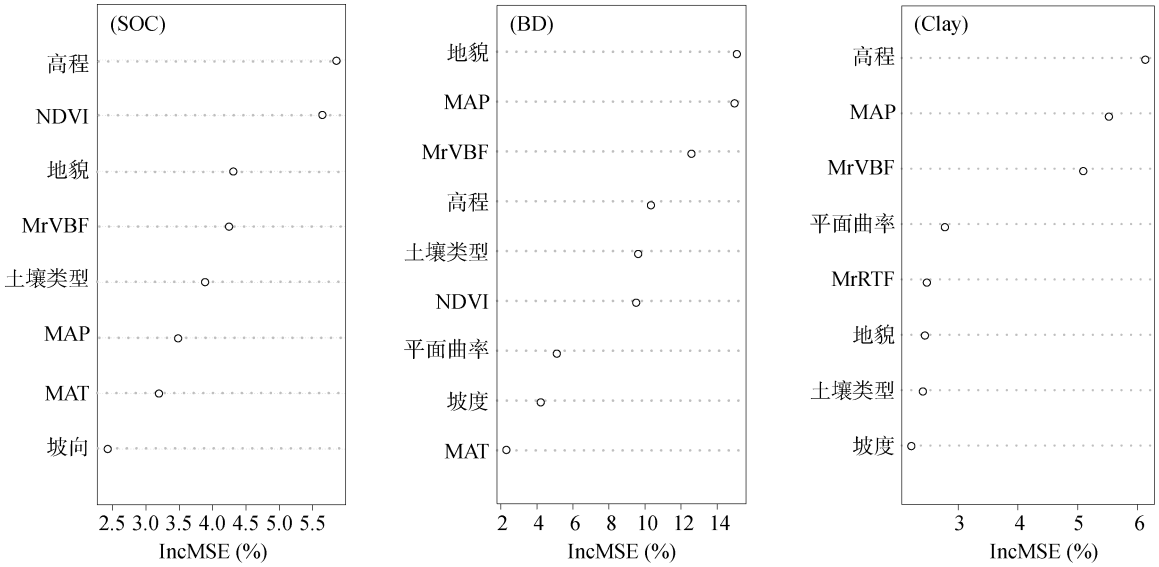

图 1　RF 模型中土壤属性预测因子的重要性排序

Fig.1　Importance sorting of predictors for soil properties in RF model

表 4　RF 模型中节点分裂次数(mtry)和决策树数量(ntree)的筛选

Table 4　Screening of splitting numbers of nodes (mtry) and numbers of decision trees (ntree) in RF model

	mtry	ntree	SOC		BD		Clay	
			建模集(R^2)	验证集(R^2)	建模集(R^2)	验证集(R^2)	建模集(R^2)	验证集(R^2)
试验组 1	1	100	0.26	0.27	0.23	0.20	0.22	0.21
	1	500	0.25	0.22	0.26	0.22	0.24	0.18
	1	1 000	0.25	0.22	0.23	0.22	0.22	0.18
试验组 2	2	100	026	0.23	0.25	0.21	0.24	0.21
	2	500	0.28	0.22	0.25	0.20	0.20	0.21
	2	1 000	0.29	0.23	0.25	0.18	0.22	0.19
试验组 3	3	100	0.22	0.21	0.26	0.18	0.19	0.18
	3	500	0.28	0.22	0.24	0.15	0.19	0.17
	3	1 000	0.29	0.22	0.24	0.19	0.21	0.19

2.5　预测精度分析及空间分布

RF 模型的性能通过计算 RMSE、MAE、R^2 等参数来进行评估，经过参数调优后采用最稳定的 RF 模型作为最终预测模型。结果(表 5、图 2)表明：①验证集中 SOC 含量、容重和黏粒含量的决定系数分别为 0.27、0.22 和 0.21。建模集中的决定系数与验证集相近，说明 RF 模型有效避免了过拟合的问题，这与前人的理论一致[6]；②SOC 含量预测效果最好，土壤容重次之，对土壤黏粒的预测效果最差；③对于 SOC 含量，建模集的 R^2 和验证集的 R^2 均高于 0.25 且整体水平相近，说明模型拟合度和泛化能力均较高，且模型较稳定；对于容重和土壤黏粒含量，建模集的

R^2 和验证集的 R^2 基本相同，说明模型稳定性极高，但是预测精度较低。④由 MAE 和 RMSE 可以看出，模型整体预测精度较高，说明在大尺度区域上，RF 模型对于土壤属性仍然有不错的预测效果。

利用 RF 模型分别对安徽省 SOC 含量、容重和土壤黏粒含量进行预测得到三者的空间分布图(图 3)，其中图 A～C 为实测值图，D～F 为预测值图。由图

表 5 土壤属性的 RF 建模精度评价

Table 5 Performance of RF model of soil properties

土壤属性	建模集			验证集		
	RMSE	MAE	R^2	RMSE	MAE	R^2
SOC	6.08	4.71	0.26	5.90	4.78	0.27
BD	0.11	0.09	0.23	0.12	0.09	0.22
Clay	90.50	72.70	0.22	78.40	60.40	0.21

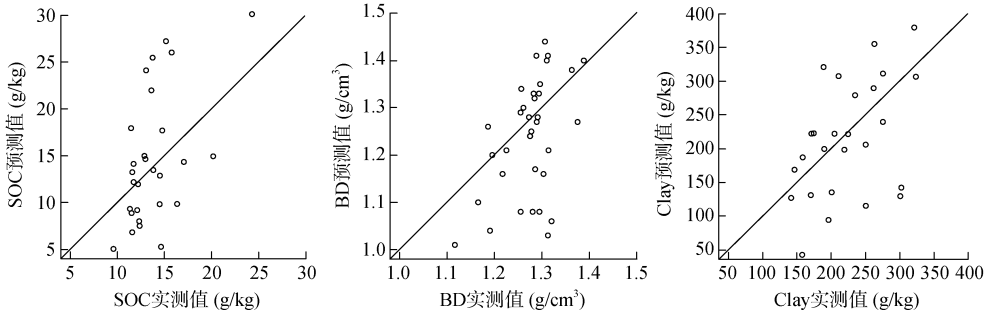

图 2 RF 模型预测散点图

Fig. 2 Scatter plots of prediction by RF model

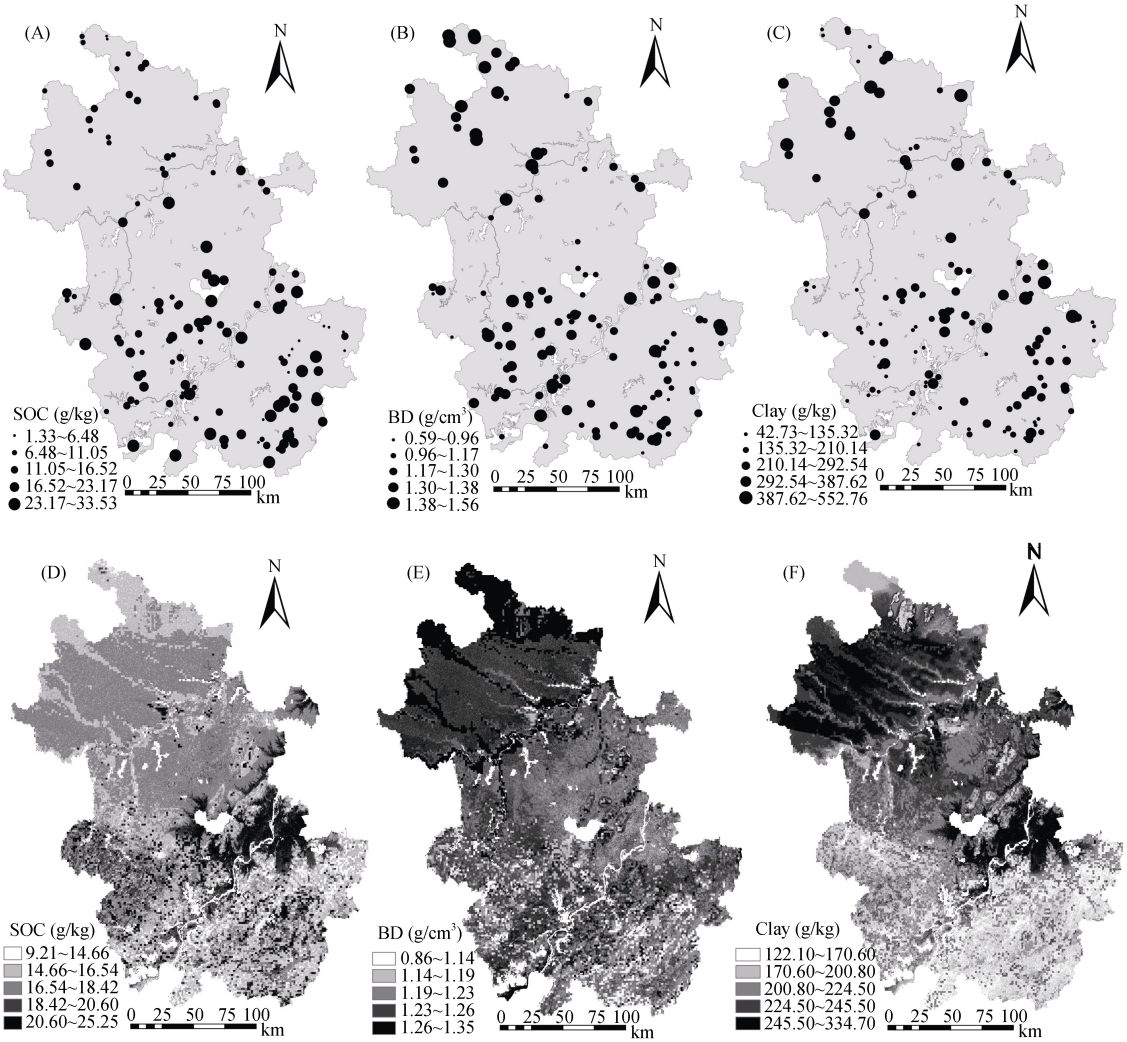

图 3 安徽省土壤属性实测值及预测值空间分布

Fig. 3 Spatial distributions of measured and predicted soil properties in Anhui Province

3 可知，安徽省 SOC 含量分布大致为由北向南逐渐增加，这基本符合以往的研究[20-21]，其中淮河中游平原地区 SOC 含量最低，沿江平原东部 SOC 含量最高。淮河中游平原地区土壤容重值最高，其他区域大致由北向南逐渐降低。安徽省土壤黏粒含量大致分布为由北向南逐渐降低。利用 RF 模型进行预测制图基本上能够反映大尺度区域上土壤属性的空间分布。

3　结论

1)安徽省内，对于土壤有机碳含量、土壤容重，土壤类型均是主要影响因子之一，可能是由于安徽省土地利用大部分为耕地，自然用地较少，导致人为因素对土壤属性影响较大；对于土壤黏粒含量，高程和年均降水量为最主要的影响因素。

2) RF 模型的建模结果表明，不同环境变量的组合分别解释了研究区域内土壤有机碳含量、容重和黏粒含量的 26%、23% 和 22%，建模集和验证集 R^2 相近，说明在大尺度区域内，RF 模型能够有效地减少过拟合问题且对于土壤属性空间分布的预测具有较高的稳定性。

3) 土壤容重和黏粒含量的预测精度不是很高，原因可能是由于研究区域面积过大，不同区域地貌和气候差异较大，以及一些可能影响土壤属性的环境变量并没有考虑到模型中。在以后的研究中可以增加采集样本数量并加入更多的环境因子作为预测变量以提高预测精度。

参考文献：

[1] Lal R, Kimble J M, Stewart B A, et al. Global climate change and pedogenic carbonate[J]. Geoderma, 1999, 104(1): 135–141

[2] Dixon J B. Roles of clays in soils[J]. Applied Clay Science, 1991, 5(5/6): 489–503

[3] Kuhn M. Building predictive models in R using the caret Package[J]. Journal of Statistical Software, 2008, 28(5): 1–26

[4] Mansuy N, Thiffault E, Paré D, et al. Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the *k*-nearest neighbor method[J]. Geoderma, 2014, s 235/236(4): 59–73

[5] Henderson B L, Bui E N, Moran C J, et al. Australia-wide predictions of soil properties using decision trees[J]. Geoderma, 2005,124(3): 383–398

[6] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5–32

[7] Rodriguez-Galiano V F, Chica-Rivas M. Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and Digital Terrain Models[J]. International Journal of Digital Earth, 2014, 7(6): 492–509

[8] Dharumarajan S, Hegde R, Singh S K, et al. Spatial prediction of major soil properties using Random Forest techniques—A case study in semi-arid tropics of South India[J]. Geoderma Regional, 2017, 10: 154–162

[9] Chagas C D S, Junior W D C, Bhering S B, et al. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions[J]. Catena, 2016, 139: 232–240

[10] 郭澎涛, 李茂芬, 罗微, 等. 基于多源环境变量和随机森林的橡胶园土壤全氮含量预测[J]. 农业工程学报, 2015, 31(5): 194–202

[11] 姜赛平, 张怀志, 张认连, 等. 基于三种空间预测模型的海南岛土壤有机质空间分布研究[J]. 土壤学报, 2018, 55(4): 1007–1017

[12] 李德成, 张甘霖, 王华, 等. 中国土系志·安徽卷[M]. 北京: 科学出版社, 2017: 3–24

[13] 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报, 2013, 50(4): 1190–1197

[14] 李龙, 姚云峰, 秦富仓, 等. 基于地理加权回归模型的土壤有机碳密度影响因子分析[J]. 科技导报, 2016, 34(2): 247–254

[15] 赵明松, 张甘霖, 李德成, 等. 江苏省土壤有机质变异及其主要影响因素[J]. 生态学报, 2013, 33(16): 5058–5066

[16] Chai T, Draxler R R. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature[J]. Geoscientific Model Development Discussions, 2014, 7(3): 1247–1250

[17] Miller F P, Vandome A F, Mcbrewster J. Coefficient of determination[J]. Alphascript Publishing, 2006, 31(1): 63–64

[18] Gelman A. Analysis of variance[J]. Quality control & applied statistics, 2006, 20(1): 295–300

[19] Sonobe R, Tani H, Shimamura H, et al. Parameter tuning in the support vector machine and random forest and their performances in cross-and same-year crop classification using TerraSAR-X[J]. International Journal of Remote Sensing, 2014, 35(23): 7898–7909

[20] 许信旺, 潘根兴, 曹志红, 等. 安徽省土壤有机碳空间差异及影响因素[J]. 地理研究, 2007, 26(6): 1077–1086

[21] 赵明松, 李德成, 王世航. 近 30 年安徽省耕地土壤有机碳变化及影响因素[J]. 土壤学报, 2018, 55(3): 595–605

Spatial Prediction of Soil Properties Based on Random Forest Model in Anhui Province

LU Hongliang[1], ZHAO Mingsong[1,2*], LIU Binyin[1], ZHANG Ping[1], LU Longmei[1]

(1 *School of Geodesy and Geomatics*, *Anhui University of Science and Technology*, *Huainan*, *Anhui* 232001, *China*; 2 *State Key Laboratory of Soil and Sustainable Agriculture*, *Institute of Soil Science*, *Chinese Academy of Sciences*, *Nanjing* 210008, *China*)

Abstract: It is important to study the spatial variability and distribution of soil properties for understanding ecosystems, formulating agricultural policies, conducting soil management and monitoring environmental changes caused by land use. The purpose of this paper is to explore the accuracy of the spatial prediction of soil properties at the provincial scale by the Random Forest (RF) model. Anhui Province in East China was selected as the study area, soil data obtained during the 2nd National Soil Survey and during 2010—2011 were used, the environmental variables were collected with GIS spatial analysis technique, and the correlation between environmental factors and soil properties was analyzed by RF model. The results showed that in the RF modeling process, SOC prediction model was the most robust and the prediction accuracy was the highest when the mtry value was 1 and the ntree value was 1 000; when the mtry value was 1 and the ntree value was 1 000 and 100 respectively, soil bulk density (BD) and clay content prediction models were the best. The elevation, NDVI, landform, muti-resolution index of valley bottom flatness (MrVBF) and soil type were the most important predictors of SOC content; Landform, mean annual precipitation (MAP), MrVBF, elevation and soil type were the most important prediction factors of soil BD; Elevation, MAP, MrVBF and plan curvature were the most important predictors of soil clay content; RF model can be used for spatial prediction of soil properties and has certain advantages in treating the qualitative variables such as soil type and landform; Multi-source environmental variable combinations explained 26% of SOC content, 23% of soil Bd and 22% of clay content, respectively. The use of machine learning for predicting soil properties and digital soil mapping is more efficient than traditional methods, it is of significance to use RF model in spatially predicting soil properties in the large-scale area.

Key words: Soil properties prediction; Random Forest model; Environmental variables; Anhui Province