DOI: 10.13758/j.cnki.tr.2019.01.021

基于随机森林模型的耕地表层土壤有机质含量空间预测①——以河南省辉县市为例

韩杏杏[1]，陈 杰[1*]，王海洋[1]，巫振富[2]，程道全[3]

(1 郑州大学水利与环境学院，郑州 450001；2 郑州大学公共管理学院，郑州 450001；3 河南省土壤肥料站，郑州 450002)

摘 要：耕地表层土壤有机质含量与作物生长发育密切相关，掌握土壤有机质空间分布对土壤肥力定向培养和农业生产指导具有重要意义。本研究以河南省辉县市 5 922 个耕地资源管理单元图斑中心点为基础数据，并分别按 8∶2、7∶3、6∶4 的比例随机划分训练数据集和验证数据集，以土壤类型作为辅助定性变量，利用随机森林模型模拟预测土壤有机质含量与自然环境变量(坡向、曲率、坡度、高程、土壤质地、归一化植被指数 NDVI)、社会经济因子(排水能力、灌溉状况)之间的复杂非线性关系。结果表明：①当训练集与检验集中样点数量的比例为 8∶2 时，对应的随机森林模型总体上预测精度较高；②选用 80% 基础数据作为训练集时，预测得到的地图与已有图件相比，相关性达到 0.859；③当用 303 个实地数据验证时，预测值与实测值的皮尔逊相关系数为 0.595。通过对影响因子的重要性排序，发现土壤质地是研究区农用地表层土壤有机质含量的最重要影响因子。因此，随机森林模型作为机器学习和数据挖掘的有效方法，能较好地模拟输入变量与有机质含量之间的关系，预测图件与实际情况相符，但对有机质含量精细的差异不能很好体现。

关键词：随机森林；土壤有机质；耕地预测制图；辉县市

中图分类号：S159-3 文献标识码：A

土壤有机质是土壤中最具有活力的成分，是土壤肥力的基础，是土壤质量最为重要的指标之一。首先，土壤有机质含有作物和微生物所需要的几乎各种营养元素，土壤有机质含量高低在很大程度上决定着土壤的养分供给能力；其次，土壤有机质能有效促进土壤结构发育，提高土壤吸附能力、缓冲能力、持水及渗透能力，是土壤肥力得以稳定发挥的基础[1-4]；另外，土壤有机质是一种稳定而长效的碳源，土壤有机碳库是地球陆地碳库的重要组成部分，在陆地碳循环中有着重要的作用。据估计土壤有机质的分解以及微生物和根系呼吸作用所产生的 CO_2 每年可达 1.35×10^{11} t，土壤有机碳库容的微小变化，都会对大气温室气体浓度及全球气候产生重大影响[5-7]。多尺度、全方位揭示土壤有机质，尤其是农田土壤有机质空间分异规律、动态演变趋势及其主要影响因素，对土壤质量培育、基础地力提升等农业生产实践具有重要的指导作用，可为制订农田土壤固碳减排、全球气候变化响应策略提供科学依据。

土壤有机质含量及其空间分布特征是多种自然与人为因素共同作用的结果，基于有限样点土壤数据，应用各种技术途径，在区域尺度上尽可能精确地预测土壤有机质，尤其是表层土壤有机质含量，并揭示其空间变异规律，一直以来都是土壤学的一个热点研究内容，也是数字化土壤制图的一个重要领域。以多元回归为核心的确定性模型[8-9]、基于地统计学手段的随机模型[10-12]及以神经网络(artificial neural network，ANN)、分类树与回归(classification and regression tree，CART)、模糊聚类、最大贝叶斯熵为代表的一系列机器学习(machine learning, ML)方法[7,13-16]，是过去几十年来土壤有机质空间预测实践中应用最为广泛的技术。随机森林(random forest，RF)模型是由 Breiman[8]于 21 世纪初在分类与回归树算法模型基础上开发的一种数据挖掘方法，是一种新的组合式自学习技术。其不仅同样具有 CART

①基金项目：国家自然科学基金项目(40971128)资助。

* 通讯作者(jchen@zzu.edu.cn)

作者简介：韩杏杏(1994—)，女，河南许昌人，硕士研究生，主要从事土地资源管理、土地资源评价、土壤空间预测与数字化制图方面的研究。E-mail: 1085310871@qq.com

与 ANN 等处理变量之间复杂非线性关系的优势，同时有效克服了上述模型存在的过度拟合、不稳定、计算复杂等缺点[9-11]。此外，随机森林模型在分类和预测时均支持多种数据类型如数值型、因子型、逻辑型等，方法简单易用。因此，随机森林模型一经问世，便迅速在包括数字化土壤制图在内的多个领域内得到推广应用[12-14]。目前在国内，将随机森林模型运用于土壤属性空间预测及数字制图研究还处于探索阶段，尤其是对土壤有机质这类时空可变性、影响因素复杂、人为活动影响深刻的土壤属性，相关的空间预测研究报道较少[11,15]。

本研究以河南省辉县市土壤为研究对象，基于研究区传统调查中的土壤属性数据、环境协变量以及与土地利用相关的基础信息，应用随机森林算法构建空间预测模型，模拟变量与有机质含量之间复杂的非线性关系，在县域尺度上对农用地表层土壤有机质含量实施空间预测并进行数字化制图，以期探索随机森林模型在土壤属性预测方面的有效性。

1　材料与方法

1.1　研究区概况及数据来源

辉县位于河南省西北部的太行山南麓，地理位置 35°17′~35°50′ N，113°20′~113°57′ E，属暖温带大陆性季风气候，总面积 2 007 km^2，地貌自北向南依次为山区、丘陵、平原 3 个类型，面积分别为 1 007、216、784 km^2。根据第二次全国土壤普查资料，辉县土壤包括 7 个土类：褐土(简育干润淋溶土)、潮土(淡色潮湿雏形土)、风砂土(干润砂质新成土)、砂姜黑土(砂姜潮湿雏形土)、棕壤(简育湿润淋溶土)、水稻土(潜育水耕人为土)、沼泽土(半腐正常有机土)，13 个亚类，29 个土属，69 个土种，其中褐土和潮土的分布最广，是主要的农业土壤。

本研究采用的主要数据源包括：河南省辉县市的测土配方施肥补贴项目(2009—2011 年)及其耕地地力评价专项项目(2011—2012 年)获取的耕地资源管理单元图、耕地地力调查点位图(图 1)等相关成果数据。其中，耕地资源管理单元图(共计 5 922 个图斑)制作过程中综合考虑了土壤类型、土地利用类型、地形地貌、灌排状况等自然条件和社会经济条件，每一个单元图斑均含有相对均质的上述相关信息；耕地地力调查点位图包含 2 445 个耕地表层(0 ~ 20 cm)土壤样点，样点布设和采集、土壤理化属性(包括土壤有机质、土壤全氮等)测定等严格执行《耕地地力调查与质量评价技术规程》[17]、《测土配方施肥技术规范》[18]的技术要求。此外，根据研究区土壤类型、地形地貌和土地利用类型的空间分布，野外补充采集 303 个耕地表层样点(图 2)作为验证数据，样点布设和采集以及土壤理化属性测定同耕地地力调查样点。本研究相关数据信息还包括第二次全国土壤普查数据以及 2008 年 10 月 Landsat 卫星影像等资料。

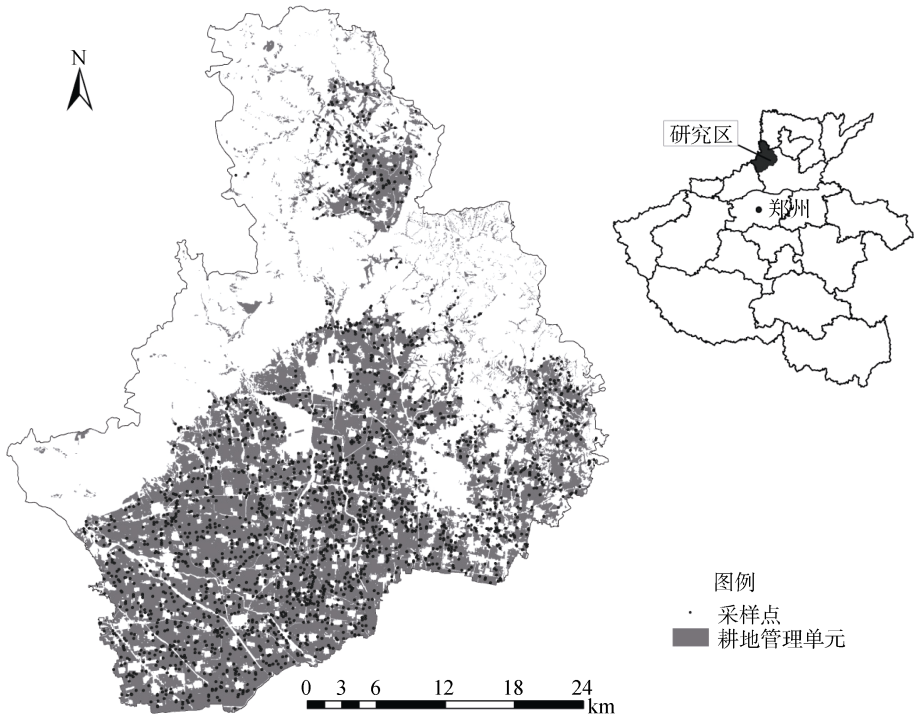


图 1　研究区耕地管理单元及样点分布图

Fig. 1　Cultivated land management units and sampling points of study area

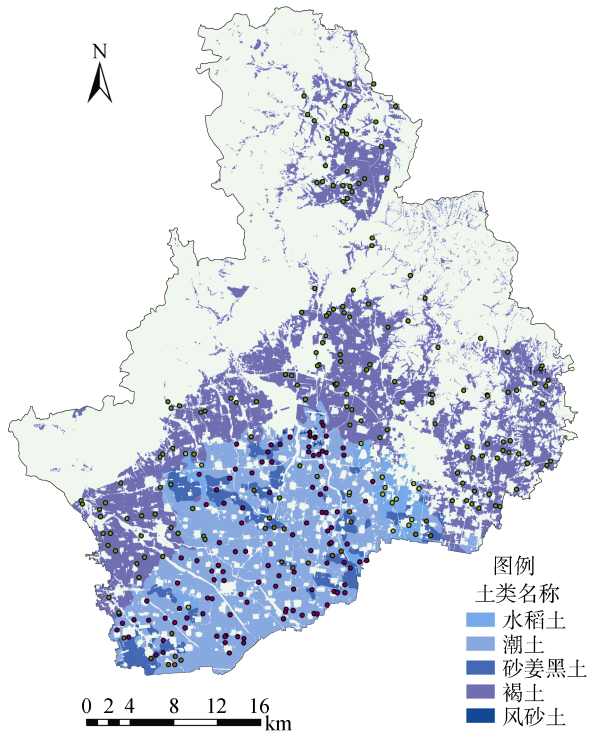


图 2 研究区不同土类实地采样点分布图
Fig. 2 Distribution of sampling sites in different soil types of study area

1.2 变量遴选与研究方法

本文从每个耕地管理单元图斑提取中心点(含图斑属性)，得到 5 922 个包含土壤有机质、高程、土类、灌溉水平等变量信息的样点，作为构建随机森林预测模型的基础数据。为了突显相关自然、人为因素对研究区农用地表层土壤有机质含量的影响在空间上的差异性，本研究将农用地表层土壤有机质含量分为两个构成部分[19]：

$$y_{ik}=m_k+r_i \quad (1)$$

式中：y_{ik} 为第 i 个样点的表层土壤有机质含量值；m_k 为其所属土类所有样点表层土壤有机质含量的平均值；r_i 为样点 y_{ik} 减去 m_k 后的残差。

以土壤质地、高程、坡度、坡向、地表曲率、归一化植被指数(NDVI)、灌溉水平以及排涝能力 8 个因子作为输入变量，构建随机森林模型表达上述影响因子与研究区农用地样点表层土壤有机质含量残差值 r_i 之间复杂的非线性映射特征：

$$r_i=f(x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6,\ x_7,\ x_8) \quad (2)$$

式中：x_1、x_2、x_3、x_4、x_5、x_6、x_7、x_8 分别代表坡度、高程、坡向、地表曲率、NDVI、灌溉水平、排涝能力和土壤质地。其中，x_1、x_2、x_3、x_4、x_5、x_6、x_7、x_8 为耕地管理单元图斑包含属性，x_2 通过公式 $x'=(x-x_{\min})/(x_{\max}-x_{\min})$ 实施标准化处理并映射在区间[0,1]中，以避免其较高的数值夸大对土壤有机质含量的贡献[20-21]；x_3 取坡向的余弦值来表征北半球坡向对接受太阳光的影响[22]；描述性变量 x_6、x_7 的量化借鉴《辉县市耕地地力评价》采用的特尔菲法对概念性因子的作用进行归纳、反馈、逐步收缩、集中，最后以隶属度确定相应的赋值(表 1)；x_8 以土壤黏粒含量的百分比赋值；x_5 基于 2008 年 10 月 Landsat 卫星影像在 Envi5.1 环境中计算。

表 1 研究区灌溉水平和排涝能力赋值
Table 1 Assignment of irrigation level and drainage capacity in study area

变量	灌溉水平 x_6				排涝能力 x_7				
	保灌区	能灌区	可灌区	无灌区	强	较强	中	较弱	弱
隶属度	1.0	0.75	0.5	0.1	1.0	0.8	0.6	0.4	0.2

随机森林建模通过 R 语言中的 Random Forest 包实现。随机森林是用随机手段建立的一个由多棵决策树构成的森林[8]。随机森林模型涉及 2 个关键参数：ntree(决策树的数量)和 m_{try}(分割节点的随机变量的数量，一般取自变量个数的平方根)。每棵树构建中，利用自助法抽样(bootstrap)生成新的样本，约 1/3 的原始训练集数据不会出现，对 n 个样本建立 n 个决策树模型。没被抽取的记录构成袋外数据，袋外误差的计算与交叉验证算法类似，因此，随机森林模型不需要单独再做交叉验证[23-27]。在分类树的每个节点上，从所有变量中随机选择 m 个预测变量，进而选择一个最优的进行节点分割，再汇总所有分类树的结果。随机森林算法能在运算量没有显著增加的前提下提高预测精度，且对多元共线性不敏感，有缺失数据和非平衡的数据时也比较稳健[25-28]。基于初步结果，本研究在随机森林实际建模操作中，设定 ntree 为 2 000，它不仅可以产生稳定的袋外误差率，数值也较小，能最大化提高计算效率，m_{try} 设置为 3。

1.3 模型验证

模型预测结果采用 3 种方式进行验证：第一，将本研究获取的原始数据划分为训练集和检验集，调整数据划入训练集和检验集的比例关系，然后计算模型输出结果的平均误差(ME)、平均绝对误差(MAE)、均方根误差(RMSE)和相关系数(r)，ME 越接近 0、MAE 和 RMSE 越小、r 越接近 1，则表明随机森林模型的预测精度越高；第二，为进一步验证随机森林模型预

测值的精度，将随机森林模型预测的非空间结果(耕地管理单元图斑中心点的有机质预测值)赋值给研究区耕地资源管理单元图斑，输出预测有机质含量分布的栅格地图，利用 Map Comparison Kit 3 软件与已有耕地管理单元有机质分布图(经辉县耕地地力评价专项项目验收的有机质分布图)进行对比[14]，计算两幅图的相关系数；第三，使用样点实测数据对模型进行检验，研究区不同土壤类型按面积比例共计选择 303 个农用地样点进行实地采样，测定表层土壤的有机质含量，与训练集和验证集的比例达到最好的预测精度时构建的随机森林模型预测的结果进行对比。

2　结果与讨论

2.1　表层土壤有机质含量描述性统计

研究区样点各类土壤表层有机质含量实测值描述性统计结果见表 2。5 922 个耕地管理单元中土壤有机质含量平均值为 21.53 g/kg，其中，水稻土平均含量最高，为 24.47 g/kg，其次是砂姜黑土，为 21.67 g/kg。就空间变异特征而言，各土类表层土壤有机质含量的变异系数均介于 10% ~ 100%，属中等程度变异，且各土类内部的变异系数比较接近。由于风砂土土类样点数量极少，相关数据的统计学意义较弱。

2.2　表层土壤有机质 r_i 值预测结果及精度

将研究区 5 922 个农用地表层土壤样点分为训练集和检验集，二者的比例关系分别设定为 8∶2、7∶3、6∶4，然后基于划分的训练集利用 R 语言 Random Forest 包建立随机森林模型，对研究区农用地表层土壤有机质 r_i 值进行预测，使用包含不同数量的训练集所得预测结果的精度如表 3 所示。

表 2　研究区不同土壤类型有机质含量描述性统计特征
Table 2　Descriptive statistics on SOM contents in different soil types in study area

土类	样本量	最小值(g/kg)	最大值(g/kg)	平均值(g/kg)	标准差(g/kg)	变异系数(%)	偏度
水稻土	207	12.20	44.30	24.47	6.27	25.62	1.58
潮土	1 011	4.90	49.40	20.84	3.44	16.51	0.49
砂姜黑土	243	11.20	39.50	21.67	3.88	17.90	1.51
褐土	4 450	4.10	49.90	21.58	2.52	11.68	1.84
风砂土	11	15.40	21.50	20.62	2.04	9.89	–2.24
全部	5922	4.10	49.90	21.53	3.03	14.07	1.87

表 3　随机森林模型使用不同训练集时表层土壤有机质 r_i 值预测精度
Table 3　Prediction accuracy of topsoil SOM contents under different training datasets

误差参数	训练集：验证集=8：2			训练集：验证集=7：3			训练集：验证集=6：4		
	训练集	验证集	总样本	训练集	验证集	总样本	训练集	验证集	总样本
r	0.388	0.404	0.829	0.386	0.418	0.783	0.372	0.420	0.671
ME	0.014	0.050	0.017	0.013	–0.030	–0.004	0.016	–0.080	0.014
MAE	1.489	1.470	0.916	1.520	1.430	0.964	1.480	1.590	1.057
RMSE	3.085	2.920	1.973	3.110	2.880	2.069	2.990	3.260	2.307

注：ME 为平均误差；MAE 为平均绝对误差；RMSE 为均方根误差；r 为皮尔逊相关系数。

从表 3 数据可以看出，当训练集与检验集样点数量的比例为 8∶2 时，用训练集运行随机森林模型可获得总体上精度较高的样点表层土壤有机质 r_i 值的预测结果。表 3 中的统计信息还显示，运行随机森林模型获得的预测结果，与训练集样点、验证集样点间的相关性均不显著(r = 0.372 ~ 0.420)，但与全区样点的对应值之间的相关性显著提高(r = 0.671 ~ 0.829)，随机森林模型在处理大数据方面的优势以及不过度拟合的特征得到体现。

2.3　表层土壤有机质含量空间预测结果

将随机森林模型输出的表层土壤有机质 r_i 值预测结果加上对应的有机质土类平均值，得到样点表层土壤有机质含量预测值，将其赋给样点所在的耕地资源管理单元多边形图斑，得到耕地管理单元有机质含量预测栅格图(图 3A)。在 GIS 环境中计算图 3A 与已有的耕地管理单元土壤有机质分布图(图 3B)之间栅格图像的相关性，显示二者的图像相关性达到 0.859。通过图 3A 和图 3B 的对比可以看出：①预测图对于表层土壤有机质含量的空间变异趋势有较好展示，高值区在预测图上同样出现在研究区域的东南部分，整体上预测图在一定程度上准确反映了研究区域农用地土壤有机质的变异特征；②在预测图中，有机质含量低值区

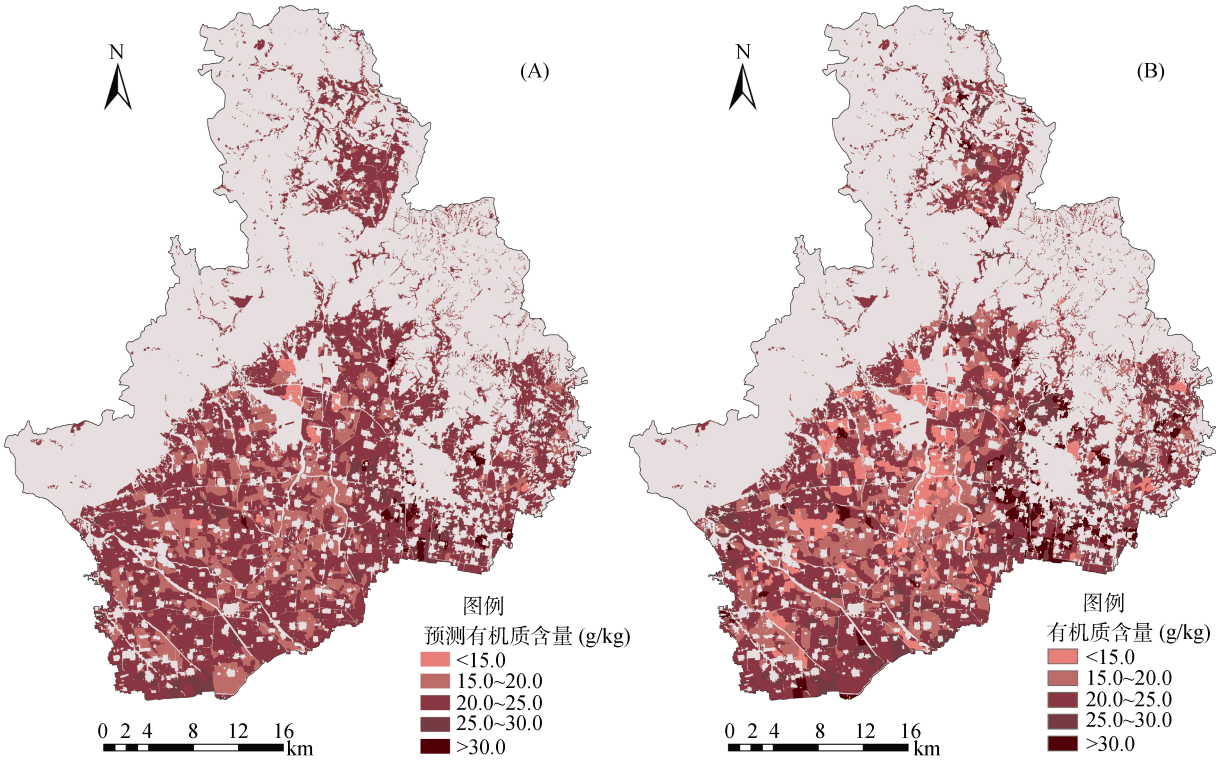


图 3　耕地管理单元有机质含量空间分布：预测图(A)和已有图(B)

Fig. 3　Spatial distribution of predicted SOM (A) and exiting SOM (B) for cultivated land management units

区域边缘不明显，中部本应出现的低值区在预测图中没有很好呈现，而是与周边土壤有机质含量无显著差异。这可能是因为随机森林模型在低值区域预测效果不佳，也可能与图中有机质含量设置梯度有一定的关系，以致图件可读性较差；③在预测图中，有机质含量的变化比较平缓，特别是在相邻地区差异不显著。从实际情况来看，在相邻的耕地区域中，土壤有机质的含量会有一定的空间自相关性，受到周边耕地的影响。在实现科学管田、耕地的集约化经营中，隐去较小的差异，对于管理上的指导更有其可行性和实践性。

此外，为更形象地对比两幅栅格地图，利用 Map Comparison Kit，对两幅图像的有机质含量数值采用公式：$Y=(b - a)/\max(\mathrm{abs}(b - a))$ 进行计算，式中，b 为耕地管理单元有机质含量预测值，a 为耕地管理单元有机质含量已有值，得到的结果如图 4 所示，有机质预测图和已有有机质含量分布图之间的标准化差异值在有机质东南侧高值区比较大(0.5 左右)，在研究区的其他地方预测效果比较理想(0 ~ 0.2)。

2.4　土壤有机质预测值与实地采样点实测值的对比

为了更进一步探讨在该研究区域基于所选变量构建的随机森林模型对耕地表层有机质含量预测结果的有效性，在研究区域实地采集 303 个表层土壤样本，在实验室测定各样点有机质实际含量，并将样点

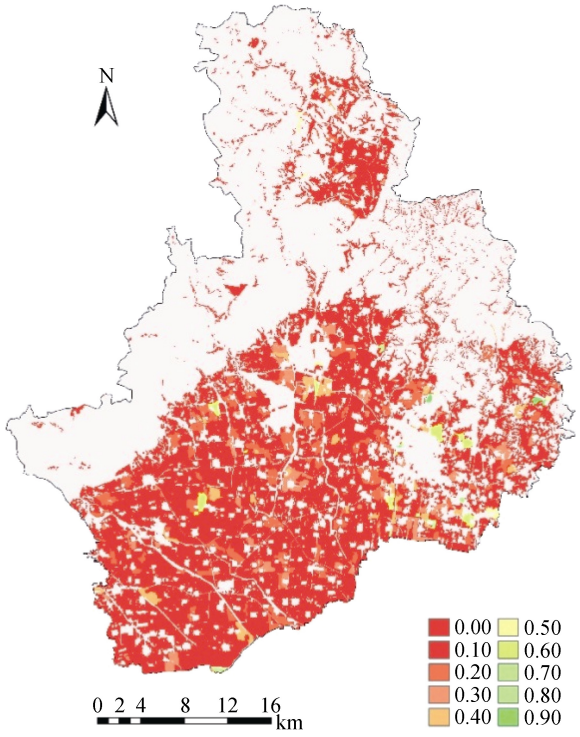


图 4　预测土壤有机质含量图与已有土壤有机质含量图的对比结果

Fig. 4　Comparison between predicted and existing SOM contents

的解释变量代入前述构建的最优模型(训练集 80%)，得到 303 个样点的有机质预测值。将实测值与预测值

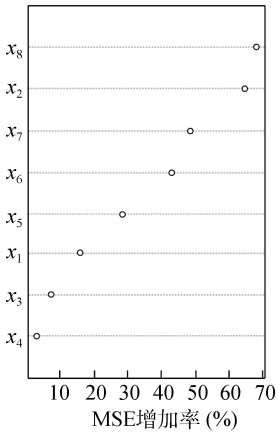
对比，计算得出皮尔逊相关系数为 0.595，ME 为 0.916，MAE 为 4.481，RMSE 为 5.958。

本研究中用训练样本构建的模型对全部样本预测时皮尔逊相关系数提高到 0.67 ~ 0.82，但用实测样点数据进行验证时，预测值与实测值的皮尔逊相关系数为 0.595，可能由于构建模型的样点为测土配方施肥补贴项目及耕地地力评价专项所得耕地管理单元图斑提取的中心点，其有机质含量与样点实测有机质含量有误差，影响了对实地样点有机质含量预测时模型的有效性。模型的预测精度还有进一步提高的空间，为得到更为精确的预测结果，可以从以下两方面入手：①进一步优化进入模型构建的训练样本的比例以及随机森林模型参数 ntree 和 m_{try} 的设置；②探究其他未在本研究体现的密切影响土壤有机质的变量，使所考虑的变量尽可能解释农用地表层土壤有机质含量的差异。

2.5　不同变量对有机质含量的影响

毋庸置疑，对于农用地表层土壤有机质含量而言，诸多影响因子的重要性各不相同。厘清农用地土壤有机质影响因子的相对重要性，对快速提升表层土壤有机质含量、定向培育农用地土壤质量等土壤利用与管理实践，具有极为关键的指导意义。随机森林算法一个很重要的功能，就是在对目标变量进行分类和回归的过程中，对解释变量的重要性进行评估。这种重要性评估并非通过统计显著性或赤池信息量准则(akaike information criterion，AIC)等指标进行间接估计，而是通过森林中每棵树生长过程中产生的袋外数据(out-of-bag, OOB)某一变量加入随机噪声后袋外误差的变化来判断的。这个误差的增加程度与预测变量的重要性具有特定比例关系。误差随加入的随机噪声增加越多，则该预测变量的重要性越高。图 5 为本研究在运行随机森林模型进行研究区农用地表层土壤有机质 r_i 值预测过程中，获得的相关解释变量相对重要性排序图。

由图 5 可以看出，土壤质地是研究区耕地表层土壤有机质含量最重要的影响因子。本研究中的土壤质地状况由土壤机械组成中黏粒含量百分比代表，因此土壤中黏粒含量的高低显著影响着土壤有机质的含量。一般而言，黏粒含量越高，表明土壤颗粒的比表面积越大，吸附能力越强，与有机颗粒组成有机-无机复合颗粒及阻止有机物分解的能力就越强；另一方面，黏粒含量高意味着土壤中小空隙比例增加，通气性较差且易为土壤水占据，好气性微生物活动受限，土壤有机质分解相对缓慢，有利于有机质在土壤中积累。在本研究区，高程是影响农用地表层土壤有机质含量的第二重要因素，这可能主要因为农用地海拔普遍较低，地势相对较高的区域，易于遭受风蚀、水蚀以及耕作侵蚀的影响，富含有机质的表层土壤流失，耕层变薄，有机质含量下降。灌溉水平和排涝能力很大程度代表了农田基本建设和田间管理水平的高低，致力于农田土壤质量培育、基础肥力提升的各种基础设施建设与耕作管理实践，同样有助于农用地表层土壤有机质含量的增加。坡向、坡度、曲率等地形因子对农用地表层土壤有机质含量的影响程度不高，主要是因为研究区农用地地势较为平坦，由上述地形因子空间变化导致的微域环境差异不足以对研究区域农用地表层土壤有机质含量产生较大影响。



(图中 x_1、x_2、x_3、x_4、x_5、x_6、x_7、x_8 分别代表坡度、高程、坡向、地表曲率、NDVI、灌溉水平、排涝能力和土壤质地；MSE 为均方误差)

图 5　自变量相对重要性排序

Fig. 5　Sequence of variables in relative importance

3　结论

本研究利用研究区耕地管理单元图斑含有的所选变量为基础数据，构建的随机森林模型在土壤有机质预测中具有有效性和可靠性，预测值在统计学意义上与实际值保持较高的一致性，可视化表达结果也有较好的体现，展示了研究区耕地土壤有机质的高低变化趋势。同时，本研究给出了所选解释变量对于辉县耕地表层有机质含量影响的相对重要性。主要结论如下：①本研究中用训练样本构建的模型对全部样本预测时皮尔逊相关系数提高到 0.67 ~ 0.82，证明了随机森林模型作为一种有效的机器学习方法在大的数据集上表现良好。②用实测样点数据进行验证时，预测值与实测值的皮尔逊相关系数值有所下降，为 0.595。③在选取的 8 个自变量中，土壤质地是辉县耕地表层土壤有机质含量最重要的影响因子。结合统计意义上

的预测精度以及最终的预测制图效果，可以认为，采用土类为定性辅助变量以及土壤质地、NDVI、高程、排灌能力等 8 个变量作为自变量，建立随机森林模型，对县域尺度农用地表层土壤有机质的含量进行预测，预测结果比较符合实际特征。

参考文献：

[1] Brady, Nyle C. The nature and properties of soils[M]. New York: MacMillan, 1984

[2] Skjemstad, J O, Reicosky D C, Wilts A R, et al. Charcoal carbon in U.S. agricultural soils[J]. Soil Science Society of America Journal, 2002, 66(4): 1249–1255

[3] 张枝枝, 张福平, 燕玉超, 等. 渭河两岸缓冲带的土壤有机质含量分布特征及其影响因子[J]. 土壤, 2017, 49(2): 393–399

[4] 向红英, 柳维扬, 彭杰, 等. 基于连续统去除法的南疆水稻土有机质含量预测[J]. 土壤, 2016, 48(2): 389-394

[5] 潘根兴, 李恋卿, 张旭辉, 等. 中国土壤有机碳库量与农业土壤碳固定动态的若干问题[J]. 地球科学进展, 2003, 18(4): 609–618

[6] 潘根兴. 中国土壤有机碳库及其演变与应对气候变化[J]. 气候变化研究进展, 2008, 4(5): 282–289

[7] 王岩松, 李梦迪, 朱连奇. 土壤有机碳库及其影响因素的研究进展[J]. 中国农学通报, 2015, 31(32): 123–131

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

[9] Li Q, Yue T, Wang C, et al. Spatially distributed modeling of soil organic matter across China: An application of artificial neural network approach[J]. CATENA, 2013, 104(Supplement C): 210–218

[10] Na X D, Zhang S Q, Zhang H Q, et al. Integrating TM and ancillary geographical data with classification trees for land cover classification of marsh area[J]. Chinese Geographical Science, 2009, 19(2): 177–185

[11] 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): 194–200

[12] Grimm R, Behrens T, Märker M, et al. Soil organic carbon concentrations and stocks on Barro Colorado Island-Digital soil mapping using Random Forests analysis[J]. Geoderma, 2008, 146: 102–113

[13] Wiesmeier M, Barthold F, Blank B, et al. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem[J]. Plant and Soil, 2011, 340(1/2): 7–24

[14] Heung B, Bulmer C E, Schmidt M G. Predictive soil parent material mapping at a regional-scale: A Random Forest approach[J]. Geoderma, 2014, 214/215(2): 141–154

[15] Guo P T, Li M F, Luo W, et al. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach[J]. Geoderma, 2015, 237/238: 49–59

[16] 郭治兴, 袁宇志, 郭颖, 等. 基于地形因子的土壤有机碳最优估算模型[J]. 土壤学报, 2017, 54(2): 331–343

[17] 中华人民共和国农业部. 耕地地力调查与质量评价技术规程: NY/T 1634－2008[S]. 北京: 中国标准出版社, 2008

[18] 中华人民共和国农业部. 测土配方施肥技术规范 NY/N 1118－2006[S]. 北京: 中国标准出版社, 2006

[19] 李启权, 王昌全, 岳天祥, 等. 基于定性和定量辅助变量的土壤有机质空间分布预测——以四川三台县为例[J]. 地理科学进展, 2014, 33(2): 259–269

[20] 张晋昕, 李河. 回归分析中定性变量的赋值[J]. 循证医学, 2005, 5(3): 169–171

[21] 马立平. 统计数据标准化──无量纲化方法──现代统计分析方法的学与用(三)[J]. 北京统计, 2000(3): 34–35

[22] 李立东, 陈杰, 宋轩, 等. 空间回归模型在区域数字化土壤制图中的应用——以河南封丘县为例[J]. 土壤学报, 2013, 50(1): 21–29

[23] 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(2): 342–354

[24] 张海阳, 齐俊传, 毛健. 基于 R 语言的数据挖掘算法研究[J]. 电脑知识与技术, 2016, 12(28): 16–19

[25] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011(3): 32–38

[26] 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

[27] 张良均, 谢佳标, 杨坦, 等. R 语言与数据挖掘[M]. 北京: 机械工业出版社, 2016

[28] 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

Spatial Prediction of SOM Content in Topsoil Based on Random Forest Algorithm: A Case Study of Huixian City, Henan Province

HAN Xingxing[1], CHEN Jie[1*], WANG Haiyang[1], WU Zhenfu[2], CHENG Daoquan[3]

(1 *School of Water Conservation and Environment*, *Zhengzhou University*, *Zhengzhou* 450001, *China*; 2 *School of Public Administration*, *Zhengzhou University*, *Zhengzhou* 450001, *China*; 3 *Soil and Fertilizer Station of Henan Province*, *Zhengzhou* 450002, *China*)

Abstract: The content of topsoil organic matter strongly influences the growth of crops, so understanding its spatial distribution is of great significance in guiding agricultural production and improving soil fertility. Taking 5 922 center points of polygons in the map of cultivated land management units of the Huixian City in Henan Province as the basic data, this study tried to evaluate the complex non-linear relationship between topsoil organic matter content and influential factors at the county scale by using the model of random forest (RF). Each point included soil types, which were the auxiliary qualitative variables, environmental variables (slope, curvature, slope, elevation, soil texture, NDVI) and socio-economic factors (drainage capacity, irrigation status), and in addition, 5 922 center points was randomly divided into the training data set and verification data set with the ratio of 8∶2, 7∶3 and 6∶4 separately. Then the accuracy of predicted map of SOM was evaluated by three ways according to the model. The results showed that when the ratio of the training data set and verification data was 8∶2, the prediction accuracy of RF model was generally higher, and the correlation was 0.859 between the predicted and the existing maps of SOM. Pearson correlation coefficient was 0.595 between the predicated and measured data of 303 field points. Based on the importance of the influential factors, it was found that soil texture was the most important variable affecting distribution of SOM in the agricultural land of the study area. The results demonstrate that the RF method, as a machine learning and data mining approach, can simulate relationships between the input variables and SOM content, meanwhile, the maps can show reliable predicted results of SOM but couldn't disclose the fine differences in SOM.

Key words: Random forest; Soil organic matter; Agricultural land predictive mapping; Huixian City