·农业水土工程·

基于两点机器学习方法的土壤有机质空间分布预测

王雨雪[1,2]，杨 柯[3,4]，高秉博[1,2※]，冯爱萍[5]，田 娟[1,2]，姜传亮[1,2]，杨建宇[1,2]

（1. 中国农业大学土地科学与技术学院，北京 100193；2. 农业农村部农业灾害遥感重点实验室，北京 100083；3. 中国地质调查局哈尔滨自然资源综合调查中心，哈尔滨 150080；4. 中国地质科学院地球物理地球化学勘查研究所，廊坊 065000；5. 生态环境部卫星环境应用中心，北京 100094）

摘 要：准确预测土壤有机质（Soil Organic Matter，SOM）空间分布对精细农业、耕地质量建设、生态环境保护以及固碳减排等均具有重要的意义。该研究探讨了基于两点机器学习方法（Two-point Machine Learning，TPML）提高 SOM 空间分布预测的可行性。以黑龙江省海伦市为研究区，以气候、地形地貌、社会经济和空间位置信息等因素作为辅助变量，充分利用空间位置信息和属性相似关系，有效处理 SOM 空间分布异质性及其与辅助变量间关系异质性，以提高 TPML 方法进行 SOM 空间分布预测的精度。采用随机森林、基于随机森林的回归克里格、反距离权重法和普通克里格（Ordinary Kriging，OK）方法作为对比，以平均绝对误差（Mean Absolute Error，MAE）、均方根误差（Root Mean Square Error，RMSE）、预测值与真实值相关系数（r）和决定系数（R^2）作为评价指标，进行不同样本量下的多组对比试验，评价不同方法的预测精度。结果表明：1）研究区 SOM 含量在 1.775～7.188 g/kg 之间，平均值为 3.179 g/kg，空间分布不均匀，呈东高西低的分布趋势。2）在不同样本量条件下，与其他模型相比，TPML 的预测精度均最高，其 MAE（0.088～0.097 g/kg）和 RMSE（0.116～0.139 g/kg）均为最小，r（0.992～0.996）和 R^2（0.971～0.985）均为最高。3）预测值的误差标准差（理论误差）与实际误差具有相似的空间模式，说明 TPML 可以为预测结果提供合理的不确定性估计。综上，TPML 模型可以通过同时利用空间自相关性和属性相似性来提高预测精度，该模型适用于预测具有一定空间自相关性且具有可用辅助数据的资源环境变量。

关键词：土壤；有机质；随机森林；空间分布预测；空间自相关性；属性相似性；两点机器学习

doi：10.11975/j.issn.1002-6819.2022.12.008

中图分类号：S159.9 **文献标志码**：A **文章编号**：1002-6819(2022)-12-0065-09

王雨雪，杨柯，高秉博，等. 基于两点机器学习方法的土壤有机质空间分布预测[J]. 农业工程学报，2022，38(12)：65-73. doi：10.11975/j.issn.1002-6819.2022.12.008 http://www.tcsae.org

Wang Yuxue, Yang Ke, Gao Bingbo, et al. Prediction of the spatial distribution of soil organic matter based on two-point machine learning method[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(12): 65-73. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2022.12.008 http://www.tcsae.org

0 引 言

土壤有机质（Soil Organic Matter，SOM）是土壤固相部分的重要组成成分，也是反映土壤肥力水平的重要指标。精确的 SOM 分布图是精细农业变量施肥、耕地质量建设、农田土壤固碳减排核算等工作的基础[1]。然而土壤调查只能获得离散点位上的 SOM 含量，需要借助数字土壤制图方法将样点上的有机质信息推算到整个研究区域，预测未采样位置的有机质含量以获得 SOM 分布图。目前，围绕 SOM 空间分布预测，已经有较多的研究[2-7]。学者从数据和预测方法等方面优化预测精度和结果。马重阳等[8]使用随机森林模型（Random Forest，RF）对河南省许昌市耕地表层 SOM 进行空间分布预测，预测结果的平均绝对误差（Mean Absolute Error，MAE）为 2.687 g/kg、均方根误差（Root Mean Square Error，RMSE）为 3.572 g/kg 和决定系数（R^2）为 0.339。尉芳等[9]使用普通克里格（Ordinary Kriging，OK）、地理加权回归模型、偏最小二乘回归模型、地理加权回归扩展模型和 RF 模型对陕西渭北旱塬区农田 SOM 空间分布预测。RF 的 RMSE 和 MAE 最小，分别为 1.316 g/kg 和 0.958 g/kg。由于空间预测精度取决于总体变异程度、样点数目、样点布设方式和采用的预测方法，因此不同研究的预测精度差异较大，难以直接比较。

按照理论基础，目前的土壤属性空间分布预测方法可以分为基于空间位置关系的方法、基于属性相似关系的方法和结合空间位置关系和属性相似关系的方法三类[10-15]。第一类空间分布预测方法以地理学第一定律和第二定律为理论基础，基于空间位置和距离预测未采样位置的土壤属性，如基于空间自相关性的反距离权重法（Inverse Distance Weighting，IDW）、普通克里格[16-18]方法，以及考虑空间异质性的各向异性克里格[19]、三明治[20]、分层异质表面点均估计（Point Mean of Surface with Non-homogeneity，P-MSN）[21]等方法。第二类空间分布预测方法以地理学第三定律为基础，基于辅助变量与土壤属性的相似性预测未采样位置的土壤属性，包括多元

收稿日期：2022-04-06 修订日期：2022-05-22

基金项目：国家重点研发计划项目（2021YFE0102300）；松嫩平原海伦地区黑土地地表基质层调查项目（DD20211589）

作者简介：王雨雪，博士生，研究方向为空间统计与土地资源管理。Email：B20213211001@cau.edu.cn

※通信作者：高秉博，博士，副教授，博士生导师，研究方向为空间统计建模、空间因果推断等。Email：gaobingbo@cau.edu.cn

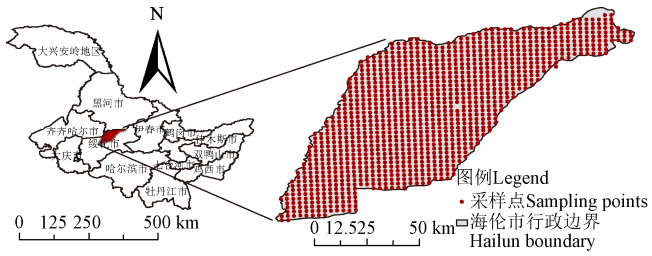
线性回归[22]、Lasso 回归、Ridge 回归、支持向量机、随机森林、梯度提升决策树[23]和神经网络[24]等。第三类方法同时借助空间位置关系信息和属性相似信息预测未采样位置的土壤属性，如空间自回归（例如空间误差模型和空间滞后模型）[25]、地理加权回归[26]、协同克里格[27]、线性回归克里格[28-29]、空间随机森林（Random Forest Spatial Interpolation，RFSI）[30]和基于随机森林的回归克里格（Random Forest Regression Kriging，RFRK）[31]。

在 SOM 预测中，早期主要以属于第一类的确定性预测方法和横跨第一类和第三类的地统计方法为主。如江厚龙等[32]使用克里格，Zhang 等[33]分别采用多元线性逐步回归法和克里格回归法，陈琳等[34]采用地理加权回归克里格法预测了 SOM 含量的空间分布。近些年随着观测数据不断丰富，属于第二类的机器学习方法，由于其在处理多维连续型和类别型变量方面的优势，已逐渐发展成主流的土壤制图方法[35-36]。虽然机器学习方法能够充分挖掘多维变量与土壤属性的相似性，提高预测精度，但是不能充分利用空间位置关系信息。由于 SOM 的扩散和迁移使得空间自相关性成为其空间分布固有特性，同时影响因素的遗漏、辅助变量数据不准确性、以及机器学习模型的拟合精度（正则化等防过拟合手段导致的拟合不足）等因素，会导致具有空间自相关性的预测误差，因此在机器学习模型中，充分利用空间位置信息能够进一步提高预测精度。为解决这个问题，众多学者尝试将空间自相关性融合到机器学习中，如 RFSI 和 RFRK，但是融合方法均存在不足之处。RFSI 分别将空间上近邻观测点的观测值与相应距离添加到预测变量中，丢失了土壤属性值变化与空间距离之间的关联关系；RFRK 使用 RF 对趋势进行建模，并使用克里格法对残差进行建模，难以一体化估计预测的误差方差。两点机器学习法（Two-point Machine Learning，TPML）[37]融合了空间位置关系与属性相似关系信息进行一体化建模，并从因变量角度出发搜索近邻点，破解了机器学习局部建模面临的维度灾难，能够有效处理 SOM 空间分布异质性及其与辅助变量间关系异质性，提高预测精度，并给出合理的不确定性度量。同时，两点机器学习法是基于随机森林模型开发的，能够处理具有一定相关性的解释变量，破解了回归预测模型可能存在的因子共线性问题[38-40]。Gao 等[37]使用 TPML 模型预测了土壤重金属含量的空间分布，验证了 TPML 模型在不同采样数量下均可以极大地提高预测精度。

因此，本文对海伦市 SOM 含量进行空间分布预测，研究充分利用 SOM 的辅助变量信息及其空间自相关性，以提高 TPML 模型的预测精度的可能性。首先使用皮尔逊相关系数和地理探测器[41]筛选与 SOM 含量相关性较高的辅助变量，其次采用 TPML、IDW、RF、RFRK 和 OK 进行 SOM 预测并对比分析预测效果，最后进行 SOM 空间分布预测并对预测结果进行不确定性分析，以期为耕地质量建设、生态环境保护等研究提供更为准确的数据支撑。

1 研究区概况

海伦市位于黑龙江省中部，绥化市北部，地处松嫩平原东北端，小兴安岭西麓，48°58′～47°52′ N，126°14′～127°45′ E 之间（图 1）。地势东北高西南低，平均海拔 239 m。总面积 4 667 km^2，其中耕地面积 2 940 km^2，占土地总面积的 63%[42]。海伦市属寒温带大陆性季风气候，年降水量 500～600 mm，年平均气温 1～2 ℃[43]。海伦市位于典型黑土区，土壤类型以黑土和草甸土为主，暗棕壤、沼泽土、白浆土、水稻土亦有少量分布，适宜作物生长，主要作物种类包括水稻、玉米、大豆等，是国家重要的商品粮基地之一[44-45]。



a. 研究区 a. Study area　　b. 采样点分布 b. Distribution of sampling points

图 1 研究区位置和采样点分布

Fig.1 Location of study area and distribution of sampling points

2 材料与方法

2.1 研究数据

SOM 数据来源于 2021 年松嫩平原海伦地区黑土地地表基质层调查项目中海伦地区 1∶25 万黑土地地表基质层调查结果，包含 1 170 个采样点空间分布及 SOM 含量（图 1b）。SOM 的迁移和积累受自然因素和人为活动的影响。为了表征相应的自然和人为因素，选择了表 1 中关于土壤、气候、地形、人类活动以及地理坐标等的 14 个代理变量[33,36,46]。数据分辨率均为 1 km。

表 1 数据类型与来源

Table 1 Data type and source

序号 No.	数据名 Data name	类型 Type	年份 Year	数据来源 Data source
1	地貌类型	CD	2009	《中华人民共和国地貌图集（1∶100 万）》
2	植被类型	CD	2001	https://www.resdc.cn/
3	土壤类型	CD	1995	https://www.resdc.cn/
4	土地利用类型	CD	2015	https://www.resdc.cn/
5	数字高程	LD	2015	https://www.resdc.cn/
6	坡度	LD	-	由数字高程计算得到
7	地形湿度指数	LD	-	由单位等高线长集水面积和坡度计算得到
8	人口	LD	2015	https://data.worldbank.org.cn/
9	国内生产总值	LD	2015	https://data.worldbank.org.cn/
10	归一化差分植被指数	LD	2015	https://ladsweb.modaps.eosdis.nasa.gov/search/
11	年平均气温	LD	2020	https://www.worldclim.org/
12	年均降水量	LD	2020	https://www.worldclim.org/
13	经度	LD	-	ArcGIS 几何计算
14	纬度	LD	-	ArcGIS 几何计算

注：CD 为类别型变量；LD 为连续型变量。

Note: CD is categorical variables and LD is continuous variables.

土壤类型和地貌类型用于反映自然资源的差异；人口（Population，POP）和国内生产总值（Gross Domestic

Product，GDP）反映了社会经济要素；坡度（Slope）、数字高程（Digital Elevation Model，DEM）、年均降水量（Mean Annual Precipitation，MAP）、年平均气温（Mean Annual Air Temperature，MAAT）、地形湿度指数（Topographic Wetness Index，TWI）、植被类型和归一化差分植被指数（Normalized Difference Vegetation Index，NDVI）代表自然影响因素，土地利用类型代表人为影响因素。使用 ArcMap10.3，多值提取到点工具将各辅助变量提取到采样点上，进而生成试验数据。

2.2　影响因素关系分析

探索 SOM 含量的空间分布情况及其与各影响因素间的相关性和异质性关系，进而选择合适的影响因素和方法进行 SOM 空间分布预测。使用皮尔逊相关系数来计算连续型变量与 SOM 含量之间的关系，使用地理探测器[41]分析类别型变量与 SOM 含量之间的关系（q 统计量）。其中，皮尔逊相关系数采用 R 语言 cor（）函数计算，地理探测器采用 R 语言 geodetector 包（http://www.geodetector.cn/）计算。使用 GeoDa 软件 SOM 的莫兰指数，以反映 SOM 含量的空间自相关性。选择与 SOM 存在较高相关性的要素作为 SOM 预测模型的输入。

2.3　TPML 方法

TPML 借鉴地统计学中半变异函数的构建和利用方式，对点对之间的关系进行建模，将空间位置信息和协变量信息融合进行一体化建模，通过挖掘空间位置关系和多维辅助变量数据，充分利用空间自相关性和属性相似性，提高空分布预测精度[37]。TPML 模型通过点对双向组合的方式，能够将样本量扩展到原样本量的平方倍，解决多维辅助变量建模中样本量不足的问题，同时实现无偏估计。除此之外，TPML 模型从因变量角度出发搜索近邻点，不仅能够为空间位置变量和其他辅助变量赋予合适的权重，而且能够破解机器学习局部建模面临的维度灾难。TPML 模型不仅能够给出较高精度的空间预测结果，还能估算对应的不确定性。

TPML 方法将空间位置变量和其他辅助变量组织到一个高维空间中，基于高维空间中的差异建模，将空间自相关性和属性相似原理都统一为高维空间中的自相关性，即在高维空间中加权距离越近的 2 个点的 SOM 含量越接近。TPML 方法通过以下 5 个步骤实现（如图 2 所示）：

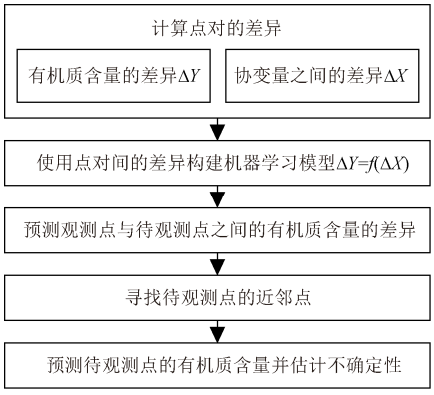


图 2　两点机器学习流程图

Fig.2　The flowchart of two-point machine learning method

1）计算配对的两点之间差异，包括 SOM 含量的差异和相应的协变量之间的差异；

2）以 SOM 含量的差异为目标变量，相应的协变量之间的差异为预测变量，建立机器学习模型；

3）预测观测点和待观测点之间的 SOM 含量的差异；

4）根据预测的 SOM 含量的差，得到 SOM 含量相近的观测点用来参与最终预测；

5）预测待观测点的 SOM 含量并使用选定的相近的观测点估计不确定性。

点对之间 SOM 含量的差异使用式（1）计算。

$$\Delta y_{ij} = y_i - y_j \tag{1}$$

式中Δy_{ij}表征第 i 和第 j 个点之间 SOM 含量的差异。同理，点对之间连续型和类别型协变量之间的差异分别由式（2）和式（3）计算得到。

$$\Delta xco_{kij} = xco_{ki} - xco_{kj} \tag{2}$$

式中 xco_{ki} 和 xco_{kj} 分别是点 i 和 j 的第 k 个连续型变量值，Δxco_{kij} 是它们的差异。

点 i 和 j 之间第 k 个类别型变量的差异使用式（3）计算。

$$\Delta xca_{kij} = (xca_{ki}, xca_{kj}) \tag{3}$$

式中 xca_{ki} 和 xca_{kj} 分别表征点 i 和 j 的第 k 个类别型变量的值，Δxca_{kij} 是它们的组合，(xca_{ki}, xca_{kj}) 是 xca_{ki} 和 xca_{kj} 之间的向量。协变量中所有类别型变量的处理方式都是通过将 1 个变量转换为 2 个变量获得的。以土壤质地（Soil Type，ST）为例，第 1 个 ST 为“黏土”，另一个 ST 为“沙土”，它们之间的差异有 2 个字段，ST1 为“黏土”和 ST2 为“沙土”。即对于一个类别型协变量 ST，在计算点对差异时，它变成了 2 个类别协变量 ST1 和 ST2，在模型训练中它们是不同的辅助变量。

由于不同点的 SOM 含量和协变量的差异，在目标变量和协变量之间构建了监督机器学习，如式（4）所示：

$$\Delta Y = f\left(\Delta Xco, \Delta Xca\right) \tag{4}$$

式中 ΔY 是 SOM 含量的差异（响应变量），Δy_{ij} 是它的值；ΔXco 和 ΔXca 分别是连续型和类别型协变量的差异。Δxco_{kj} 是 ΔXco 的值，Δxca_{kij} 是 ΔXca 的值；f 是监督机器学习模型。TPML 方法使用数字土壤制图中常用的随机森林模型。

建立随机森林模型后，用它来预测观测点和待观测点之间的 SOM 含量差异，如式（5）：

$$\Delta\hat{y}_{0i} = f(\Delta xco_{10i}, \Delta xco_{20i}, \ldots, \Delta xco_{m0i}; \Delta xca_{20i}, \ldots, \Delta xca_{l0i}) \tag{5}$$

式中 $\Delta\hat{y}_{0i}$ 表征待观测点 0 和观测点 i 之间的 SOM 含量差异，Δxco_{10i} 是待观测点 0 和观测点 i 之间的第一个连续型变量的差异，Δxca_{10i} 是待观测点 0 和观测点 i 之间的第一个类别型变量的差异，m 和 l 分别是连续型和类别型协变量的数量。

基于式（6），可以使用任何观测点对待观测点进行预测。

$$\hat{y}_{0i} = y_i + \Delta\hat{y}_{0i} \tag{6}$$

式中 y_i 是第 i 个观测点的 SOM 含量，$\hat{y}_{0i}$ 是基于式（6）计算得到的待观测点 0 的预测值。待观测点最终的预测值是几个近邻采样点预测值的线性加和，如式（7）所示：

$$\hat{y}_0=\frac{1}{\eta}\sum_{i=1}^{\eta}\hat{y}_{0i},\quad \eta\leqslant n \ \text{且}\left|\Delta\hat{y}_{0i}\right|<\left|\Delta\hat{y}_{0(i+1)}\right| \tag{7}$$

式中 $\hat{y}_0$ 是待观测点 0 的最终预测值，所有观测点的预测值按照与点 0 对应的差值的绝对值升序排序，排序列表中的前 η 个预测用于最后的预测，n 是观测点的总数，η 不大于 n。这样只有邻近观测点的预测才参与 $\hat{y}_0$ 的最终预测。η 通过交叉验证方法设置，其中所有观测结果重复分为训练部分和验证部分，以搜索最优 η 给出最佳准确度。

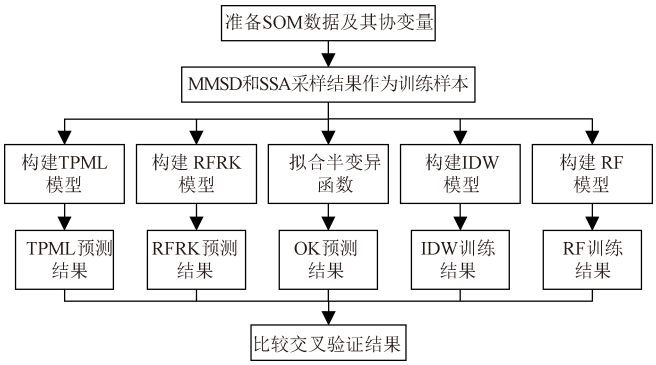
基于百分位回归树计算最终预测的不确定性，如式（8）。其中 V 是方差的统计量。通过假设预测误差的对称分布，可以得到最终预测 $\hat{y}_0$ 的方差。

$$\begin{aligned}\delta_0^2&=\mathrm{V}\left(\hat{y}_0-y_0\right)\\&=\mathrm{V}\left(\frac{1}{\eta}\sum_{i=1}^{\eta}\hat{y}_{0i}-y_0\right)\\&=\mathrm{V}\left[\frac{1}{\eta}\sum_{i=1}^{\eta}\left(y_i+\Delta\hat{y}_{0i}\right)-\frac{1}{\eta}\sum_{i=1}^{\eta}\left(y_i+\Delta\mathrm{y}_{0i}\right)\right]\\&=\frac{1}{\eta^2}\mathrm{V}\left(\sum_{i=1}^{\eta}\Delta\hat{y}_{0i}-\Delta\mathrm{y}_{0i}\right)=\frac{1}{\eta^2}\sum_{i=1}^{\eta}\Delta{\delta_{0i}}^2\end{aligned} \tag{8}$$

式中 δ_0^2 是点 0 的预测值 $\hat{y}_0$ 的方差，$\Delta{\delta_{0i}}^2$ 是第 i 个点与点 0 之间含量差异的预测误差方差，即式（7）所得预测值的不确定性。

2.4 模型对比试验

实施以下 5 个步骤，对比 TPML 模型与 OK、RF、IDW 和 RFRK 模型进行 SOM 预测的精度，如图 3 所示。



注：SOM 为土壤有机质；MMSD 为最短距离均值模型；SSA 为空间模拟退火算法；TPML 为两点机器学习模型；RFRK 为基于随机森林的回归克里格模型；IDW 为反距离权重法；RF 为随机森林模型；OK 为普通克里格模型。
Note: SOM is the Soil Organic Matter; MMSD is the Mean Most Shortest Distances model; SSA is the Spatial Simulated Annealing method; TPML is the Two-point Machine Learning model; RFRK is the Random Forest Regression Kriging model; IDW is the Inverse Distance Weighting model; RF is the Random Forest model; OK is the Ordinary Kriging model.

图 3 试验线路图

Fig.3 The road map of experiment

1）准备 SOM 数据并提取所有 1 170 个点的协变量值。

2）从 1 170 个点中抽取 50、100、200、300、400 和 500 个样本，以最短距离均值（Mean Most Short Distance，MMSD）和空间模拟退火（Spatial Simulated Annealing，SSA）得到的采样结果作为训练样本。

3）构建 TPML、RF、RFRK，OK 和 IDW 模型。

4）使用 TPML、OK、RF、IDW 和 RFRK 预测验证集的 SOM 含量。

5）比较 5 种方法的预测精度和结果。

为了更好地评估每个模型的性能，使用十折交叉验证方法来测试模型的预测能力。在十折交叉验证中，所有样本被随机分成 10 组，大小大致相等。每组依次保留作为验证数据集以评估模型性能，而其余 9 组用于模型训练。这个过程重复 10 次，直到每组都作为验证集被测试了 1 次并得到了相应的预测。计算 MAE、RMSE 和预测值与真实值之间的相关系数（r），通过将估计结果与实际对应的采样点观测数据进行比较来评估模型的准确性。使用 R^2 来评估模型的稳定性[38-40]。

2.5 空间分布预测

使用 2.3 节描述的两点机器学习方法，以 2.1 节得到的土地利用类型、土壤类型、植被类型、气候因子（降水和气温）、地形地貌因子、社会经济因子和空间位置信息（经纬度）作为模型输入，预测 SOM 含量空间分布，并基于式（8）计算预测结果的不确定性。

3 结果与分析

3.1 样本描述性统计

分析 1 170 个采样点 SOM 含量数据，结果表明，最小值为 1.360 g/kg，最大值为 9.170 g/kg，极差为 7.810 g/kg，中位数为 2.990 g/kg，均值为 3.226 g/kg，标准差为 1.065 g/kg。现有采样点的 SOM 含量空间分布如图 4 所示。SOM 含量主要集中在 2.200～4.590 g/kg 之间，占全部采样点的 80%，SOM 含量的均值为 3.226 g/kg，说明采样点 SOM 含量整体呈中等水平。图中橙色和红色的点代表 SOM 含量较高，主要分布在海伦市东北部。蓝色代表 SOM 含量较低，主要分布在海伦市南部和西南部，中部 SOM 含量处于中等水平。SOM 含量空间分布不均匀，整体呈东高西低的分布趋势。

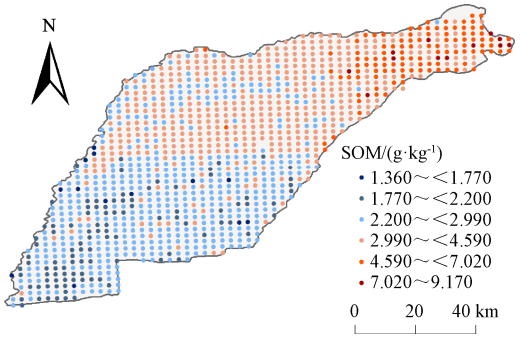


图 4 海伦市 SOM 含量空间分布

Fig.4 Spatial distribution of SOM content of Hailun city

使用 Geoda 软件计算样本点的莫兰指数表征 SOM 含量的空间自相关性。结果表明，样本点的 Moran’s I 指数为 0.779，说明样本点的 SOM 含量存在高度空间自相关性。

3.2 影响因素分析

土壤类型、土地利用类型、地形指标、气候条件和

社会经济要素等的显著差异可能导致 SOM 含量的变异性。分别使用皮尔逊相关系数和地理探测器计算 SOM 含量与土地利用类型、土壤质地、地形因子、气候条件和社会经济要素中连续型变量间的相关性以及类别型变量间的异质性。表 2 为连续型变量与 SOM 含量之间的相关性。从表 2 可知，SOM 含量与 NDVI、DEM 和 MAP 呈正相关关系，与 GDP 和 MAAT 呈现负相关关系。同时，SOM 含量与 DEM、MAP、NDVI、GDP 以及 MAAT 均呈现较强的相关性关系。

表 2　连续型变量相关性分析

Table 2　Analysis of correlation for continuous variables

变量名 Variables	相关系数 Correlation coefficient
数字高程 Digital elevation model	0.708***
坡度 Slope	0.137***
地形湿度指数 Topographic wetness index	-0.038
人口 Population	-0.698***
国内生产总值 Gross domestic product	-0.697***
归一化差分植被指数 Normalized difference vegetation index	0.482***
年平均气温 Mean annual air temperature	-0.804***
年均降水量 Mean annual precipitation	0.710***

注（Note）：‘***’ P<0.001; ‘**’ P< 0.01; ‘*’ P<0.05。

表 3 为类别型变量与 SOM 含量的空间异质性检验结果（q）。其中，q 统计量为空间分异性度量值，表征各类别型变量多大程度上解释了 SOM 的空间分异。从表 3 可知，土地利用类型、地貌类型、植被类型和土壤类型对 SOM 含量有显著影响。

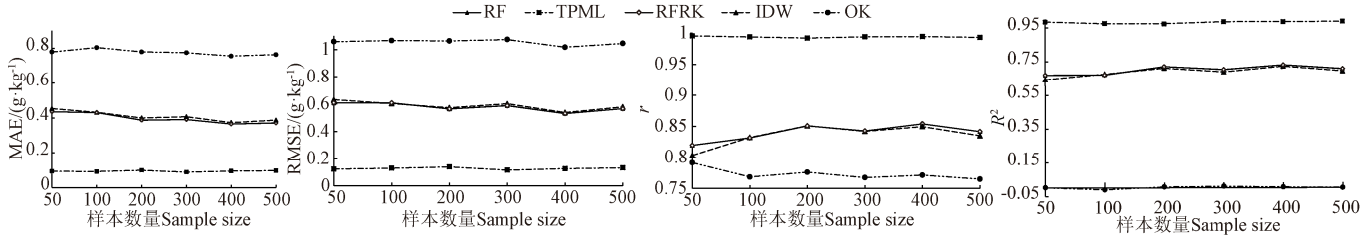
表 3　类别型变量空间异质性检验结果

Table 3　Spatial heterogeneity test results for categorical variables

变量名 Variables	q 值 q value
土地利用类型 Land use type	0.355***
地貌类型 Landform type	0.417***
土壤类型 Soil type	0.204 ***
植被类型 Vegetation type	0.306 ***
黏土 Clay soil	0.243 ***
沙土 Sandy soil	0.205 ***
泥土 Silt soil	0.178***

3.3　SOM 预测精度分析

对不同 SOM 预测模型进行对比，结果如图 5 所示。在不同采样数量下，TPML、OK、RF、IDW 和 RFRK 的 MAE 的均值分别为 0.094、0.777、0.397、0.409 和 0.397 g/kg，RMSE 的均值分别为 0.127、1.053、0.577、0.588 和 0.577 g/kg，r 的均值分别为 0.994、0.773、0.840、0.836 和 0.840，R^2 的均值分别为 0.980、0.009、0.700、0.689 和 0.700。可见，在不同采样数量下，TPML 模型在 MAE、RMSE、r 和 R^2 4 个模型精度评价指标上的表现均优于 RF、RFRK、IDW 和 OK 4 种方法，进而验证了本文使用的两点机器学习方法的准确性和有效性。



a. 平均绝对误差　a. Mean Absolute Error (MAE)

b. 均方根误差　b. Root Mean Square Error (RMSE)

c. 预测值与真实值之间的相关系数　c. Correlation coefficient between predicted and true value (r)

d. 决定系数　d. Coefficient of determination (R^2)

图 5　不同模型 SOM 预测精度评估

Fig.5　Accuracy evaluation of different models for SOM prediction

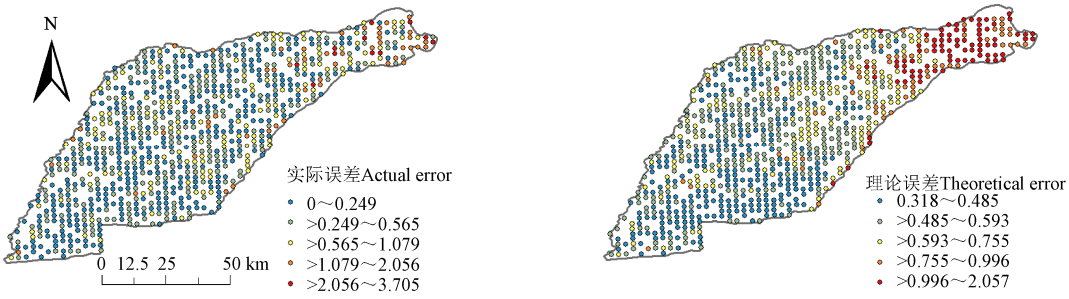
以样本数量为 300 时的验证结果为例（图 6），参考文献[37]，使用验证集预测结果的绝对误差的绝对值来表征实际误差（图 6a），使用验证集预测结果的误差标准差来表征理论误差（图 6b）。从图 6 中可以看出，高估和低估并存。此外，理论误差与实际误差具有相似的空间模式。证明了误差标准差可以用来反映预测误差，并且 TPML 可以为预测结果提供合理的不确定性。不确定性对于评估预测结果的可靠性，指导制作的 SOM 含量空间分布图的使用具有重要意义，可以为未来采样的布局设计提供依据。

3.4　SOM 预测结果

以 1 170 个采样点及其协变量作为各方法的输入数据，得到的预测结果以相同分辨率的栅格在 ArcGIS10.3 中进行显示，并导出各方法得到的研究区 SOM 含量空间分布图（图 7）。从图中可以看到，5 种方法所得到的研究区 SOM 含量空间分布的总体趋势相似，不同方法预测结果的高值区和低值区分布位置大致相同。其中，除 OK 方法，其他 4 种方法得到的预测结果高值区与低值区变化特征明显，呈现渐变趋势。其中，使用 OK 方法进行预测的精度较低（图 5），得到的 SOM 含量也相对较低（图 7a）。IDW 方法预测精度相对较低（图 5），但由于 IDW 是根据样本点之间的欧式距离加权，选择近邻的点进行插值预测，得到的 SOM 含量相对较高（图 7d）。从图 7b 和图 7c 中可以看出，引入随机森林方法进行 SOM 空间分布预测，其预测精度有所提升（图 5），得到的 SOM 含量相对于普通克里格有所提升。图 7e 为使用 TPML 方法进行 SOM 含量预测的结果，其预测精度较高（图 5），且得到的 SOM 含量有所提升。图 7f 为使用 TPML 方法进行 SOM 含量空间分布预测的理论误差的分布情况，与图 6b 具有相似的分布情况，东北部误差标准差较高，中部和西南部误差标准差较低。

使用 TPML 方法得到的研究区 SOM 含量在 1.775～

7.188 g/kg 之间，平均值为 3.179 g/kg，整体处于中等水平，SOM 空间分布不均匀，呈东高西低的分布趋势。精度评价结果（图 5）表明在 50～500 个采样点数量下，TPML 的预测精度均最高，其 MAE 和 RMSE 均为最小，预测结果图中无明显块状区域，高值区域中会零星出现低值，在低值区域中会零星出现高值，更加贴近真实情况。

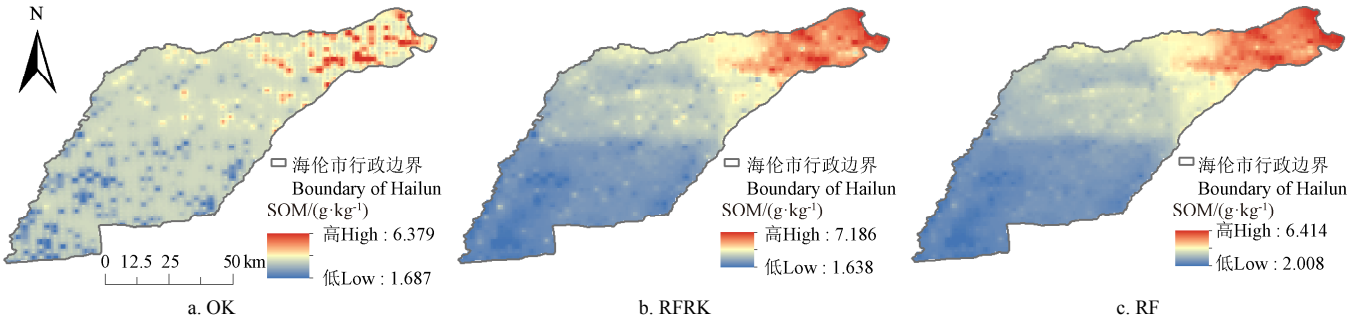


a. 实际误差
a. Actual error

b. 理论误差
b. Theoretical error

图 6 样本数量为 300 时 TPML 验证集预测 SOM 结果的实际误差与理论误差

Fig.6 Actual and theoretical error of the SOM prediction results for sample size 300 with TPML



a. OK　b. RFRK　c. RF

d. IDW　e. TPML　f. TPML 预测误差的标准差
f. Standard deviation of TPML prediction error

图 7 基于不同模型的 SOM 预测结果

Fig.7 The SOM prediction results by different models

4 讨 论

研究结果表明，海伦市 SOM 既存在空间相关性，同时各影响因素对 SOM 含量有显著影响。仅基于空间相关性的插值方法 OK 和 IDW，RMSE 分别为 1.053 和 0.588 g/kg，MAE 分别在 0.777 和 0.409 g/kg，能够达到一定的空间预测精度。在协变量难以收集或相关性不高时，可以降低精度要求，采用仅基于空间相关性的插值方法。

仅引入协变量的 RF 模型的预测精度优于 OK 和 IDW 模型（图 5）。其原因在于 SOM 与气候、地形地貌、NDVI 等环境因素以及人类活动密切相关，具有较强的空间异质性格局[30]。通过引入与 SOM 相关的 DEM、GDP 和年均降水量等多种影响因素，利用属性相似性建模，能够取得高于仅考虑空间相关性方法的精度。因此，近年能够处理多维多类型协变量的机器学习方法在土壤制图中应用较多。

RFRK 方法虽然同时考虑了空间相关性和属性相似性，但是对两者分别建模。因此，虽然空间精度比 RF 有所改进，但改进幅度非常有限。综合考量空间异质性和相关性进行建模的思想在相关领域也得到了学者的认可，刘艳芳等[47]证明使用考虑残差空间结构的回归克里格模型预测土壤有机碳密度的表现优于其他模型。同样的，徐占军等[48]使用分区 OK 插值法进行采煤沉陷区土壤有机碳含量空间预测，得到的预测结果优于直接进行 OK 插值得到的，验证了同时考虑空间异质性以及空间自相关性建模，其精度会提高。

TPML 方法的预测精度高于地质统计学方法（OK 和 IDW）或同时考虑多个协变量（RFRK 和 RF）的方法。在 50～500 个采样点数量下其误差曲线、相关系数曲线和决定系数曲线均较为平稳，浮动较小，可见 TPML 可以显著提高预测精度（图 5）。TPML 模型的 RMSE 和 MAE 均在 0.09～0.2 g/kg 之间，相比于其他 4 种方法精

度有所提高。预测结果与真实值间的相关系数和 R^2 均在 0.9 以上，说明 TPML 模型稳定性较优，得到的预测结果接近真实情况。TPML 的预测效果优于 IDW 等其他 4 个模型的原因在于其同时融合了属性相似关系和位置邻近关系进行一体化建模。

TPML 借鉴地统计建模点对关系的方式，建模点对之间辅助变量和 SOM 含量之间的关系。通过点对双向组合的方式，能够将样本量扩展到原样本量的平方倍，解决多维辅助变量建模中样本量不足的问题，并实现无偏估计。通过从因变量角度出发搜索近邻点，TPML 不仅能够为空间位置变量和其他辅助变量赋予合适的权重，而且能够破解机器学习局部建模面临的维度灾难。此外，TPML 的误差标准差（理论误差）和实际误差具有相似的空间模式，证明了误差标准差可以用来反映预测误差，为预测结果提供合理的不确定性估计。不确定性对于评估预测结果的可靠性，指导制作的 SOM 含量空间分布图的使用具有重要意义，也可以为未来采样的布局设计提供依据。

5　结　论

本文使用两点机器学习方法预测海伦市土壤有机质（Soil Organic Matter，SOM）空间分布，得出的结论如下：

1）采样点上的 SOM 含量空间分布不均匀，呈东高西低的趋势，SOM 含量整体呈中等水平。同时，SOM 含量与归一化差分植被指数、数字高程和年均降水量呈正相关关系，与国内生产总值、年平均气温呈负相关关系，与土地利用类型、地貌类型、植被类型和土壤类型均具有显著的空间异质性关系。

2）将两点机器学习模型（Two-point Machine Learning，TPML）与反距离权重（Inverse Distance Weighting，IDW）、普通克里格、随机森林、基于随机森林的回归克里格模型进行对比，结果表明，在不同采样点数量下，TPML 模型的均方根误差和平均绝对误差均在 0.09～0.2 g/kg 之间，高于其他模型精度。TPML 模型的预测结果与真实值间的相关性和决定系数（R^2）均在 0.9 以上，精度较高，说明了 TPML 模型的有效性和准确性。

3）TPML 模型预测结果的理论误差与实际误差具有相似的空间分布，说明 TPML 可以为预测结果提供合理的不确定性，基于此给出了预测结果和不确定性分析结果。基于 TPML 的 SOM 空间分布预测结果表明，研究区 SOM 含量在 1.775～7.188 g/kg 之间，平均值为 3.179 g/kg，整体处于中等水平，SOM 含量空间分布不均匀，呈东高西低的分布趋势，预测结果图中无明显块状区域，更加贴近真实情况。

本研究证明了 TPML 模型能够有机融合空间位置关系和多维辅助数据，有效处理 SOM 空间分布异质性及其与辅助变量之间关系异质性，充分利用空间相关性和属性相似性提高空间预测精度，适用于呈现一定程度的空间自相关性并具有丰富辅助数据的资源环境变量空间预测。未来可以进一步探索 TPML 方法在不同尺度不同情景下的空间预测性能，准确评估其适用条件与可用范围。

［参　考　文　献］

[1] Wang Y Q, Zhang X C, Zhang J L, et al. Spatial variability of soil organic carbon in a watershed on the Loess Plateau[J]. Pedosphere, 2009, 19(4): 486-495

[2] Tziachris P, Aschonitis V, Chatzistathis T, et al. Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters[J]. Catena, 2019, 174: 206-216.

[3] Zhang C T, Yang Y. Can the spatial prediction of soil organic matter be improved by incorporating multiple regression confidence intervals as soft data into BME method?[J]. Catena, 2019, 178: 322-334.

[4] Dou X, Wang X, Liu H, et al. Prediction of soil organic matter using multi-temporal satellite images in the Songnen Plain, China[J]. Geoderma, 2019, 356: 113896.

[5] Long J, Liu Y, Xing S, et al. Optimal interpolation methods for farmland soil organic matter in various landforms of a complex topography[J]. Ecological Indicators, 2020, 110: 105926.

[6] Meng X, Bao Y, Ye Q, et al. Soil organic matter prediction model with satellite hyperspectral image based on optimized denoising method[J]. Remote Sensing, 2021, 13(12): 2273.

[7] Nikou M, Tziachris P. Prediction and uncertainty capabilities of quantile regression forests in estimating spatial distribution of soil organic matter[J]. ISPRS International Journal of Geo-Information, 2022, 11: 130.

[8] 马重阳，孙越琦，巫振富，等. 基于不同模型的区域尺度耕地表层土壤有机质空间分布预测[J]. 土壤通报，2021，52(6)：1261-1272.
Ma Chongyang, Sun Yueqi, Wu Zhenfu, et al. Spatial prediction of topsoil organic matter of arable land by different models at the regional scale[J]. Chinese Journal of Soil Science, 2021, 52(6): 1261-1272. (in Chinese with English abstract)

[9] 尉芳，刘京，夏利恒，等. 陕西渭北旱塬区农田土壤有机质空间预测方法[J]. 环境科学，2022，43(2)：1097-1107.
Wei Fang, Liu Jing, Xia Liheng, et al. Spatial prediction method of farmland soil organic matter in Weibei dryland of Shaanxi province[J]. Environmental Science, 2022, 43(2): 1097-1107. (in Chinese with English abstract)

[10] Qqla B, Txy B, Cqw A, et al. Spatially distributed modeling of soil organic matter across China: An application of artificial neural network approach[J]. Catena, 2013, 104: 210-218.

[11] Liu Y, Guo L, Jiang Q, et al. Comparing geospatial techniques to predict SOC stocks[J]. Soil and Tillage Research, 2015, 148: 46-58.

[12] Hengl T, Heuvelink G, Stein A. A generic framework for spatial prediction of soil variables based on regression-kriging[J]. Geoderma, 2004, 120: 75-93.

[13] Jafari A, Khademi H, Finke P A, et al. Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran[J]. Geoderma, 2014, 232: 148-163.

[14] Kumar A, Lal R, Liu D. A geographically weighted regression kriging approach for mapping soil organic carbon stock[J]. Geoderma, 2012, 189: 627-634.

[15] Zhu Q, Lin H S. Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes[J]. Pedosphere, 2010, 5: 594-606.

[16] Jin L, Heap A D. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors[J]. Ecological Informatics, 2011, 6(3/4): 228-241.

[17] Jin L, Heap A D. Spatial interpolation methods applied in the environmental sciences: A review[J]. Environmental Modelling & Software, 2014, 53: 173-189.

[18] Stein A, Varekamp C, Egmond C V. Zinc Concentrations in groundwater at different scales[J]. Journal of Environmental Quality, 1995, 24(6): 1205-1214.

[19] Stein A, Hoogerwerf M, Bouma J. Use of soil-map delineations to improve (Co-)kriging of point data on moisture deficits[J]. Geoderma, 1988, 43(2/3): 163-177.

[20] Liao Y, Li D, Zhang N, et al. Application of sandwich spatial estimation method in cancer mapping: A case study for breast cancer mortality in the Chinese mainland, 2005[J]. Statistical Methods in Medical Research, 2019, 28(12): 3609-3626.

[21] Gao B, Hu M, Wang J, et al. Spatial interpolation of marine environment data using P-MSN[J]. International Journal of Geographical Information Science, 2020, 34(3): 577-603.

[22] Zhou Y, Chen S, Zhu A X, et al. Revealing the scale- and location-specific controlling factors of soil organic carbon in Tibet[J]. Geoderma, 2021, 382: 114713.

[23] Breiman L. Statistical modeling: The two cultures[J]. Statistical Science, 2001, 16: 199-215.

[24] Tan Z, Yang Q, Zheng Y. Machine learning models of groundwater arsenic spatial distribution in Bangladesh: Influence of Holocene sediment depositional history[J]. Environmental Science & Technology, 2020, 54: 9454-9463.

[25] Zhu A X, Liu J, Du F, et al. Predictive soil mapping with limited sample data[J]. European Journal of Soil Science, 2015, 66: 535-547.

[26] Darmofal D. Spatial Analysis for the Social Sciences(Analytical Methods for Social Research)[M]. Cambridge: Cambridge University Press, 2015: 158-199.

[27] Xiao M Y, Zhang G H, Breitkopf P, et al. Extended Co-Kriging interpolation method based on multi-fidelity data[J]. Applied Mathematics and Computation, 2018, 323: 120-131.

[28] Georganos S, Grippa T, Niang G A, et al. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling[J]. Geocarto International, 2021, 36(2): 121-136.

[29] Hengl T, Heuvelink G B M, Rossiter D G. About regression-kriging: From equations to case studies[J]. Computers & Geosciences, 2007, 33: 1301-1315.

[30] Sekulić A, Kilibarda M, Heuvelink G B M, et al. Random Forest Spatial Interpolation[J]. Remote Sensing, 2020, 12: 1687.

[31] Xu J, Zhang F, Ruan H, et al. Hybrid modelling of random forests and kriging with sentinel-2A multispectral imagery to determine urban brightness temperatures with high resolution[J]. International Journal of Remote Sensing, 2021, 42: 2174-2202.

[32] 江厚龙，刘国顺，杨夏孟，等. 精准农业中不同取样间距下Kriging插值精度对比研究[J]. 土壤通报，2011，42(4)：879-886.
Jiang Houlong, Liu Guoshun, Yang Xiameng, et al. Comparison of kriging interpolation precision in different soil sampling interval in precision agriculture[J]. Chinese Journal of Soil Science, 2011, 42(4): 879-886. (in Chinese with English abstract)

[33] Zhang S, Huang Y, Shen C, et al. Spatial prediction of soil organic matter using terrain indices and categorical variables as auxiliary information[J]. Geoderma, 2012, 171/172: 35-43.

[34] 陈琳，任春颖，王宗明，等. 基于克里金插值的耕地表层土壤有机质空间预测[J]. 干旱区研究，2017，34(4)：798-805.
Chen Lin, Ren Chunying, Wang Zongming, et al. Prediction of spatial distribution of topsoil organic matter content in cultivated land using kriging methods[J]. Arid Zone Research, 2017, 34(4): 798-805. (in Chinese with English abstract)

[35] 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: 49-59.

[36] Lorenzo G, Marta C, Luca F, et al. Mapping soil organic carbon in Tuscany through the statistical combination of ground observations with ancillary and remote sensing data[J]. Geoderma, 2021, 404: 115386.

[37] Gao B B, Stein A, Wang J, et al. A two point machine learning method for spatial prediction of soil pollution[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 108: 102742.

[38] 刘焕军，张美薇，杨昊轩，等. 多光谱遥感结合随机森林算法反演耕作土壤有机质含量[J]. 农业工程学报，2020，36(10)：134-140.
Liu Huanjun, Zhang Meiwei, Yang Haoxuan, et al. Invertion of cultivated soil organic matter content combining multi-spectral remote sensing and random forest algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2020, 36(10): 134-140. (in Chinese with English abstract)

[39] 李德，陈文涛，乐章燕，等. 基于随机森林算法和气象因子的砀山酥梨始花期预报[J]. 农业工程学报，2020，36(12)：143-151.
Li De, Chen Wentao, Le Zhangyan, et al. Forecast method for the first flowering date of Dangshansu pear based on random forest algorithm and meteorological factors[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2020, 36(12): 143-151. (in Chinese with English abstract)

[40] 刘峻明，和晓彤，王鹏新，等. 长时间序列气象数据结合随机森林法早期预测冬小麦产量[J]. 农业工程学报，2019，35(6)：158-166.
Liu Junming, He Xiaotong, Wang Pengxin, et al. Early prediction of winter wheat yield with long time series meteorological data and random forest method[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2019, 35(6): 158-166. (in Chinese with English abstract)

[41] Wang J F, Li X H, Christakos G, et al. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China[J]. International Journal of Geographical Information Science, 2010, 24(1): 107-127.

[42] 赵军，张久明，孟凯，等. 地统计学GIS在黑土区域土壤养分空间异质性分析中的应用—以海伦市为例[J]. 水土保持通报，2004，24(6)：53-57.
Zhao Jun, Zhang Jiupeng, Meng Kai, et al. Spatial heterogeneity of soil nutrients in blacksoil, China-A Case Study at Hailun County[J]. Bulletin of Soil and Water Conservation, 2004, 24(6):

53-57. (in Chinese with English abstract)

[43] 李欣宇，宇万太，李秀珍. 遥感与地统计方法在表层土壤有机碳空间格局研究中的应用比较[J]. 农业工程学报，2009，25(3)：148-152.
Li Xinyu, Yu Wantai, Li Xiuzhen. Comparison and application of remote sensing and geostatistics methods to spatial distribution of surface soil organic carbon[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2009, 25(3): 148-152. (in Chinese with English abstract)

[44] 黑龙江省海伦县土壤普查办公室. 海伦县土壤志[M]. 海伦：黑龙江省海伦县土壤普查办公室，1985.

[45] 王建华，陶培峰，袁月，等. PSR 框架下的黑龙江省海伦市耕地质量评价[J]. 地质与资源，2020，29(6)：525-532.
Wang Jianhua, Tao Peifeng, Yuan Yue, et al. PSR-Based evaluation of the cultivated land quality in Hailun city of Heilongjiang province[J]. Geology and resources, 2020, 29(6): 525-532. (in Chinese with English abstract)

[46] Wager S, Hastie T, Efron B. Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife[J]. Journal of Machine Learning Research: JMLR, 2014, 15: 1625-1651.

[47] 刘艳芳，宋玉玲，郭龙，等. 结合高光谱信息的土壤有机碳密度地统计模型[J]. 农业工程学报，2017，33(2)：183-191.
Liu Yanfang, Song Yuling, Guo Long, et al. Geostatistical models of soil organic carbon density prediction based on soil hyperspectral reflectance[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(2): 183-191. (in Chinese with English abstract)

[48] 徐占军，张媛，张绍良，等. 基于 GIS 与分区 Kriging 的采煤沉陷区土壤有机碳含量空间预测[J]. 农业工程学报，2018，34(10)：253-259.
Xu Zhanjun, Zhang Yuan, Zhang Shaoliang, et al. Spatial prediction of soil organic carbon content in coal mining subsidence area based on GIS and partition Kriging[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(10): 253-259. (in Chinese with English abstract)

Prediction of the spatial distribution of soil organic matter based on two-point machine learning method

Wang Yuxue[1,2], Yang Ke[3,4], Gao Bingbo[1,2※], Feng Aiping[5], Tian Juan[1,2], Jiang Chuanliang[1,2], Yang Jianyu[1,2]

(1. *College of Land Science and Technology*, *China Agricultural University*, *Beijing* 100193, *China*; 2. *Key Laboratory of Remote Sensing of Agricultural Disasters*, *Ministry of Agriculture and Rural Affairs*, *Beijing* 100083, *China*; 3. *Harbin Natural Resources Comprehensive Survey Center*, *China Geological Survey*, *Harbin* 150080, *China*; 4. *Institute of Geophysical and Geochemical Exploration*, *Chinese Academy of Geological Sciences*, *Langfang* 065000, *China*; 5. *Ministry of Ecology and Environment Center for Satellite Application on Ecology and Environment*, *Beijing* 100094, *China*)

Abstract: An accurate prediction of the spatial distribution of Soil Organic Matter (SOM) is of great importance for precision agriculture, farmland quality construction, ecological environment protection, and soil carbon sequestration. However, the accuracy of prediction dominates by the heterogeneity of SOM spatial distribution and its relationship with auxiliary variables. Taking Hailun City, Heilongjiang Province (126°14′-127°45′ E, 48°58′-47°52′ N) of northeast China as the study area, this study aims to accurately and rapidly predict the SOM spatial distribution using a Two-Point Machine Learning Method (TPML) with the climate, topography, socio-economic, and spatial location as the auxiliary variables. The spatial location and auxiliary variables were also integrated to effectively deal with the heterogeneity of SOM spatial distribution and the heterogeneity of its relationship with auxiliary variables. The performance of TPML was then evaluated using the Random Forest (RF), RF regression kriging, inverse distance weighting, and Ordinary Kriging (OK) models. The performances of the models with samples of different sizes were also evaluated using the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), correlation coefficient between the predict and true value (r), and the coefficient of determination (R^2). The results reveal that: 1) The SOM was predicted to range from 1.775 to 7.188 g/kg in the study area, with an average value of 3.179 g/kg. The spatial distribution of SOM spatially varied, with a trend of the high in the east and the low in the west. Meanwhile, the SOM content was positively correlated with the normalized difference vegetation index (NDVI), digital elevation, and mean annual precipitation, whereas, negatively correlated with the gross domestic product, mean annual air temperature, and topographic wetness index, particularly significantly related to the land use, landform, vegetation, and soil type. 2) The TPML presented the highest accuracy of prediction under different sample sizes, with the lowest MAE (0.088-0.097 g/kg) and RMSE (0.116-0.139 g/kg), while the highest r (0.992-0.996) and R^2 (0.971-0.985). The MAE and RMSE of the TPML model were improved much more than 0.7 g/kg, while the r and R^2 were improved by more than 0.2, and 0.9, respectively, compared with the most frequently-used OK. 3) There is a similar spatial pattern between the standard deviation of prediction errors (theoretical errors) and the actual errors, indicating that the TPML provided reasonable uncertainty estimates for the prediction. Consequently, the TPML can be expected to employ spatial autocorrelation and attribute similarity at the same time for higher spatial prediction accuracy. Anyway, the TPML spatial prediction of variables is feasible for the resource and environment with a certain degree of spatial autocorrelation and available auxiliary data.

Keywords: soils; organic matter; random forest; spatial distribution prediction; spatial auto-correlation; attribute similarity; two-point machine learning