

基于河马优化算法与机器学习的土壤有机质制图采样策略

连振翔^{1,2}, 费徐峰², 任周桥^{2*}

(1. 浙江农林大学数学与计算机科学学院, 浙江杭州 311300, 中国;

2. 浙江省农业科学院数字农业研究所, 浙江杭州 310021, 中国)

摘要: [目的/意义] 土壤有机质 (Soil Organic Matter, SOM) 是土壤质量的核心表征指标, 开展SOM制图研究具有重要意义。尽管机器学习已成为提升数字土壤制图 (Digital Soil Mapping, DSM) 精度的重要手段, 但其性能依赖于输入采样数据的质量。因此, 合理的采样点布局是DSM的关键前提。本研究旨在剔除冗余采样点、降低采样成本, 并进一步提升SOM的预测精度。[方法] 构建基于河马优化算法 (Hippopotamus Optimization Algorithm, HO) 随机森林残差克里金插值的最优采样策略。以浙江省兰溪市已布设调查的1 080个土壤样点为基础, 结合环境协变量, 优化生成多组采样方案, 确定最优的采样密度和样点位置, 用于SOM的空间预测与制图。[结果和讨论] HO优化的多组采样方案中, 当采样密度为2.3点/km² (629个采样点) 时效果最佳, 均方根误差和平均绝对误差分别降低至5.11和3.79 g/kg, 决定系数为0.49, 林氏一致性相关系数达到0.63, 优化后采样成本较原始方案下降41.8%。[结论] 综上所述, 兼顾采样成本和预测精度, HO是一种潜在有效的采样优化方法, 能为类似区域的土壤有机质空间预测与制图采样提供参考。

关键词: 河马优化算法; 采样优化; 随机森林; 克里金插值; 数字土壤制图; 遥感制图

中图分类号: TP18; S159

文献标识码: A

文章编号: SA202508027

引用格式: 连振翔, 费徐峰, 任周桥. 基于河马优化算法与机器学习的土壤有机质制图采样策略[J]. 智慧农业(中英文), 2026, 8(2): 48-58. DOI: 10.12133/j.smartag.SA202508027

LIAN Zhenxiang, FEI Xufeng, REN Zhouqiao. Optimal Sampling Strategy for Soil Organic Matter Based on Hippopotamus Optimization Algorithm and Machine Learning[J]. Smart Agriculture, 2026, 8(2): 48-58. DOI: 10.12133/j.smartag.SA202508027 (in Chinese with English abstract)

0 引言

土壤是人类生存和发展的基础, 其质量直接关系到粮食安全和人类可持续发展^[1]。当前, 由于不合理的开发利用, 土壤退化问题日益严重^[2], 影响土壤健康和农产品安全。因此, 土壤质量的监测和评估对于制定土地利用策略、保护土壤资源、确保人类健康具有重要意义。随着数字土壤制图 (Digital Soil Mapping, DSM) 的发展, 机器学习模型因其强大的非线性拟合能力与多源数据融合优势, 正逐步取代传统线性统计模型, 成为高精度土壤属性空间预测的核心工具^[3]。其中, 随机森林 (Random Forest, RF) 因其能够有效捕捉土壤属性与环境协变量之间复杂的非线性关系, 对输入数据中的

噪声、缺失值和冗余变量具有较强的鲁棒性, 适用于遥感、地形、气候等多源异构数据的融合建模, 成为数字土壤制图中最具代表性的机器学习方法之一^[4]。尽管随机森林模型能够有效建模土壤属性与环境因子之间的复杂关系, 但其通常将采样点视为独立样本, 可能忽略残差中存在的空间自相关性, 导致预测结果在空间过渡上不够平滑。为克服这一局限, 随机森林残差克里金插值 (Random Forest Residual Kriging, RFRK)^[5]的集成策略被证明能有效结合两者的互补优势, 通过克里金法对RF模型的预测残差进行空间插值, 再利用插值结果对原模型预测结果进行残差修正, 来提高模型的预测精度。然而, 机器学习模型的预测性能高度依赖于训练样本的质量与代表性。土壤样点的合理布局不仅

收稿日期: 2025-08-27

基金项目: 国家重点研发计划项目 (2023YFD1902900)

作者简介: 连振翔, 硕士研究生, 研究方向为农林资源大数据与智能决策。E-mail: lianzx2025@163.com

*通信作者: 任周桥, 博士, 研究员, 研究方向为数字耕地。E-mail: renzq@zaas.ac.cn

copyright©2026 by the authors

(登录 www.smartag.net.cn 免费获取电子版全文)

是土壤调查的基础，更是决定数字土壤制图精度的关键环节，既要满足一定精度的土壤属性数字制图在技术方法上的要求，又要在经济上、效率上具有可行性^[6-8]。因此，针对不同目标要求开展土壤采样布局优化研究具有重要意义。

在早期的研究中，常用的采样方法主要包括简单随机采样、系统网格采样和分层随机采样^[9]等，这些方法简单易行，但往往忽视了土壤属性的空间相关性，可能导致估计结果的不确定性和不精确；相比之下，条件拉丁超立方体采样（Conditional Latin Hypercube Sampling, cLHS）^[10-12]通过整合多种环境协变量，在环境特征空间中优化采样点分布，提升了采样点的代表性。然而，由于其主要追求环境空间中的均匀覆盖，可能忽视地理空间上高变异区域的采样密度需求，导致在土壤属性空间异质性强的局部区域代表性不足，难以兼顾全局覆盖与局部细节。此外，协变量的选择具有区域依赖性，如何根据不同地理单元确定关键环境变量仍具挑战。随着优化计算技术的发展，模拟退火算法（Simulated Annealing, SA）^[13,14]、遗传算法（Genetic Algorithm, GA）^[15,16]等启发式优化算法也先后被引入用于土壤采样布局研究，但SA、GA等优化算法在局部范围内替换样本点以逐步改进采样方案，可能限制了算法的全局搜索能力，容易陷入局部最优，在面对复杂的计算时，存在收敛速度慢等问题。因此，需要寻找一个能够兼顾全局探索与局部开发的优化算法来解决采样优化问题。

河马优化算法（Hippopotamus Optimization Algorithm, HO）^[17]因其3阶段优化设计，能够增强全局探索并逃离局部最优，具有收敛速度快、求解精度高的特点，在土壤采样布局优化中的应用尚未见报道。因此，本研究以浙江省兰溪市为研究区，以土壤有机质含量预测制图为目标，探究利用HO优化研究区土壤采样布局，以期减少长期土壤有机质（Soil Organic Matter, SOM）含量监测的采样成本。

1 研究区与数据

1.1 研究区域

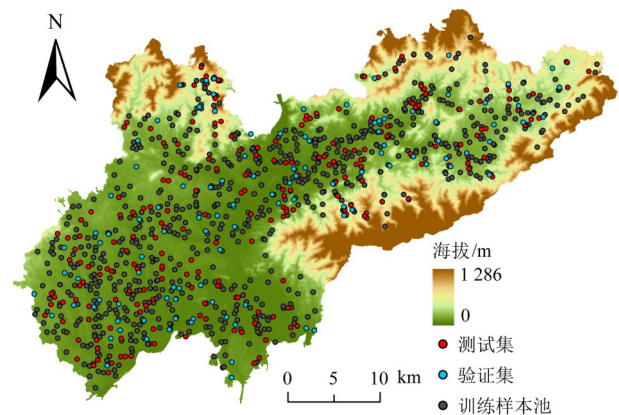
研究区位于浙江省中西部的兰溪市，地理范围为29°5'20"~29°27'30"N，119°13'30"~119°53'50"E，总面积约1 313 km²，耕地面积273.91 km²。境域地貌为丘陵盆地，平均海拔约为102 m，属东亚副热带季风区，温暖湿润，年均气温为17.7℃，年平均降水量为1 439 mm。研究区

主要土壤类型包括红壤、黄壤、岩性土、潮土和水稻土，主要农作物包括水稻、小麦、油菜等。

1.2 样本点数据

根据测土配方施肥3年一轮回设计。2021—2023年，兰溪市在全市耕地园地上按照平均20 hm² 1个点（4.5个/km²）的密度，共取土调查了1 080个样点，有机质含量为3.63~47.05 g/kg，平均值为23.92 g/kg，变异系数为30.82%。

对1 080个样点按照7:1:2分为训练样本池（ $n=756$ ）、验证集（ $n=108$ ）和测试集（ $n=216$ ）。验证集用于计算样本优化过程中HO算法的目标函数均方根误差（Root Mean Square Error, RMSE），以评估基于所选样本的SOM预测精度。图1显示了研究区域的数字高程模型（Digital Elevation Model, DEM），以及验证集、测试集和训练样本池的空间分布。



注：该图基于自然资源部标准地图服务网站下载的审图号为GS(2019)3266号标准地图制作，底图无修改。

图1 兰溪市概况以及土壤样本数据的空间分布

Fig.1 An overview of Lanxi city and the spatial distribution of soil sample data

1.3 特征变量集

根据DSM-SCORPAN框架^[18]，参考以往的土壤有机质含量数字制图研究^[19-21]，共选择了28种特征变量，如表1所示。DEM来源于地理空间数据云网站（<http://www.gsccloud.cn/>），地形因子数据基于DEM由ArcGIS Pro提取得到^[22,23]；气象因子通过资源环境科学数据平台（<https://www.resdc.cn>）获取^[24]；植被因子和盐分指数通过Sentinel-2遥感卫星的波段计算得到^[25]；其余土壤因子均来自中国科学院南京土壤研究所（<https://www.issas.ac.cn>），为多年平均值。所有特征变量在建模前均使用ArcGIS Pro重采样至30 m空间分辨率。

表 1 SOM的特征变量
Table 1 Feature variable of soil organic matter (SOM)

变量类别	变量名称	原始分辨率/m	年份
土壤因子	土壤酸碱度(pH)	90	2010—2018
	全磷(Total Phosphorus, TP)		
	容重(Bulk Density, BD)、黏粒含量(Clay Content, CLY) 粉粒含量(Silt Content, SLT)、砂粒含量(Sand Content, SND)		
	盐分指数(Salinity Index 2, SI2)	30	2023
位置因子	经度(Longitude, LON)	—	—
	纬度(Latitude, LAT)		
地形因子	坡向(Aspect, ASP)、坡度(Slope, SLP)	30	2009
	高程(Digital Elevation Model, DEM)		
	剖面曲率(Profile Curvature, Kv)、水流强度指数(Stream Power Index, SPI)		
气象因子	年降水量(Precipitation, PRE)	1 000	2022
	年均气温(Temperature, TEM)		
	年蒸发量(Evaporation, EVP)		
	年均地温(Ground Surface Temperature, GST)		2012—2022
	多年平均降水量(Mean Annual Precipitation, MAP)		
	多年平均气温(Mean Annual Temperature, MAT)		
植被因子	归一化植被指数(Normalized Difference Vegetation Index, NDVI)	30	2022
	增强型植被指数(Enhanced Vegetation Index, EVI)		
	植物总初级生产力(Gross Primary Productivity, GPP)		
	叶面积指数(Leaf Area Index, LAI)		
	植物净生产力(Net Primary Productivity, NPP)		
	总潜在蒸散量(Potential Evapotranspiration, PET)		
	总蒸散量(Evapotranspiration, ET)		
	光合有效辐射(Fraction of Photosynthetically Active Radiation, FPAR)		

2 研究方法

2.1 方法概述

本研究以剔除冗余采样点、降低采样成本和提高 SOM 预测精度为目标，构建基于 HO 并结合 RFRK 的最优采样策略。首先，在数据预处理阶段整合土壤样本与特征变量。随后，进入迭代采样优化，其中 HO 算法负责生成候选采样方案，而 RFRK 模型则作为评估器，通过计算验证集预测结果的 RMSE 来评估每个方案的优劣。该 RMSE 作为适应度值反馈给 HO 以引导种群优化。最终，算法输出最优采样方案，并基于此方案训练模型，以林氏一致性相关系数 (Lin's Concordance Correlation Coefficient, LCCC)、决定系数 (Coefficient of Determination, R^2)、平均绝对误差 (Mean Absolute Error, MAE) 和 RMSE 评估其在测试集上的泛化能力。本研究采用 Python 语言进行编码实现，整体技术路线如图 2 所示。

2.2 河马优化算法

HO 是一种新型的自然启发式优化算法，由 AMIRI 等^[17,26]共同开发。该算法的设计灵感来源于河马的天生行为，包括它们在河流或池塘中的位

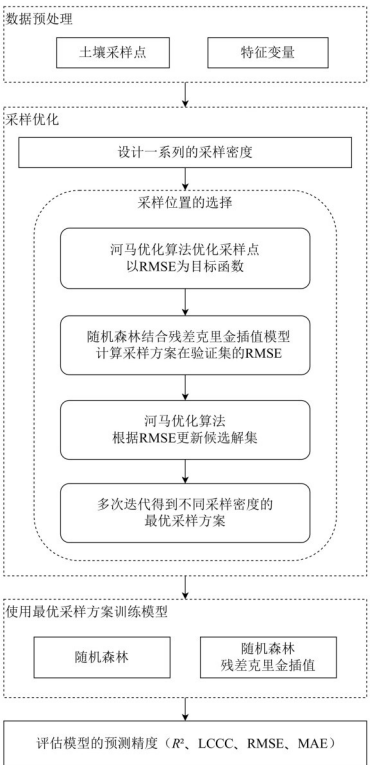


图 2 基于 HO-RFRK 的土壤采样优化技术路线图
Fig.2 Technical roadmap for soil sampling optimization based on hippopotamus optimization algorithm and random forest residual kriging (HO-RFRK)

置更新、对抗捕食者的防御策略，以及逃避方法。HO的参数主要包括河马数量 (N)、最大迭代次数 (MaxIter)，以及随机向量。经网格参数调优后，本研究设置HO的河马数量为5，最大迭代次数为1 000。HO土壤采样优化的流程如图3所示。

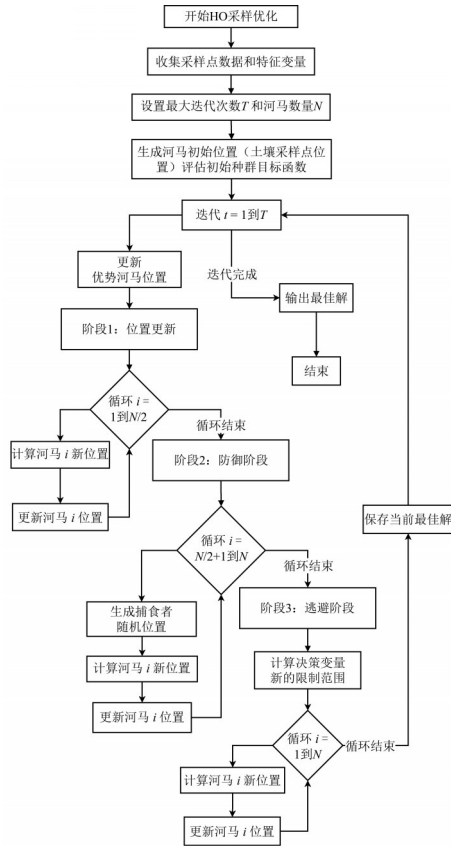


图3 HO土壤采样优化流程图

Fig.3 Flow chart of the hippopotamus optimization (HO) algorithm in soil sampling optimization

算法开始时，首先随机生成1个河马群体（候选解集）如公式（1）所示，每头河马代表1个潜在的解决方案，河马的位置在搜索空间中随机初始化，其中位置由决策变量的值决定。本研究利用该模式从训练样本池中选择1组采样点当作1个候选解（1头河马），多组不同的采样点组成1个河马群体。

$$X(:, i) = X_{\min}(i) + r \cdot (X_{\max}(i) - X_{\min}(i)) \quad (1)$$

式中： $X(:, i)$ 为第*i*个候选解的位置； r 在 $[0, 1]$ 区间内， $X_{\min}(i)$ 、 $X_{\max}(i)$ 分别为第*i*个决策变量的下限和上限。

2.2.1 位置更新

河马群体由数只成年雌性河马、幼崽、多只成年雄性河马和优势雄河马（河马群的首领）组成，幼崽因天性好奇可能脱离群体而面临被捕食的风

险。在位置更新阶段，群体中雄性河马的位置更新如公式（2）所示，群体中雌性或未成年河马的位置如公式（3）~公式（5）所示。

$$X_P_1(i, :) = X(i, :) + y_1 \cdot (D_{\text{hippo}} - I_1 X(i, :)) \quad (2)$$

式中： $X_P_1(i, :)$ 为雄性河马的位置； $X(i, :)$ 为第*i*个河马的位置； D_{hippo} 为优势雄性河马的位置； y_1 为随机数；随机数 $I_1 \in [1, 2]$ 。

$$h = \begin{cases} I_2 \cdot \vec{r}_1 + (\sim Q_1) \\ 2 \cdot \vec{r}_2 - 1 \\ \vec{r}_3 \\ I_1 \cdot \vec{r}_4 + (\sim Q_2) \\ r_5 \end{cases} \quad (3)$$

$$X_P_2(i, :) = \begin{cases} X(i, :) + h_1 \cdot (D_{\text{hippo}} - I_2 \cdot MG_i) & T > 0.6 \\ E & \text{else} \end{cases} \quad (4)$$

$$E = \begin{cases} X(i, :) + h_2 \cdot (MG_i - D_{\text{hippo}}) & r_6 > 0.5 \\ X_{\min}(i) + r_7 \cdot (X_{\max}(i) - X_{\min}(i)) & \text{else} \end{cases} \quad (5)$$

式中： $X_P_2(i, :)$ 为雌性或未成年河马在群体中的位置；随机数 $I_2 \in [1, 2]$ ； MG_i 为随机选择一些河马的平均值； T 为选择概率； r_7 为随机数； h_1 和 h_2 在公式（3）的 h 中随机选择，其中 $\vec{r}_1 \sim \vec{r}_4$ 为随机向量， r_5 为随机数， Q_1 和 Q_2 为整数随机数。

2.2.2 防御阶段

当河马面临捕食者威胁时，会采取防御行为。在算法中，这被模拟为河马对捕食者（即当前解的威胁）的响应，河马会尝试通过调整位置来躲避或对抗捕食者。面对捕食者的入侵，算法利用Levy分布进行位置突变，如果捕食者的目标函数小于当前方案的目标函数，则迅速调整采样点的位置以逼近最优采样点，否则进行有限的位置调整，如公式（6）所示，这个阶段能保护已经找到的较好的采样布局方案，同时对新的潜在方案进行评估。

$$X_P_3(i, :) =$$

$$\begin{cases} \overrightarrow{RL} \oplus \text{Predator}_j + \left(\frac{b}{c - d \cos(2\pi g)} \right) \left(\frac{1}{\vec{D}} \right), & F_{\text{Predator}_j} < F_i \\ \overrightarrow{RL} \oplus \text{Predator}_j + \left(\frac{b}{c - d \cos(2\pi g)} \right) \left(\frac{1}{2\vec{D} + r_9} \right), & F_{\text{Predator}_j} \geq F_i \end{cases} \quad (6)$$

式中： $X_P_3(i, :)$ 为捕食者入侵河马群体的领

地；随机向量 $\vec{RL}$ (Levy分布) 为攻击河马时捕食者的位置突变； $Predator_j$ 表示捕食者的位置； $\vec{D} = |Predator_j - X(i, :)|$ 为第*i*只河马到捕食者的距离；*b*、*c*、*d*、*g*为均匀分布随机数， $b \in [2, 4]$ ， $c \in [1, 1.5]$ ， $d \in [2, 3]$ ， $g \in [2, 4]$ ；随机变量 $r_9 \in [0, 1]$ ； $F_{Predator_j}$ 为河马保护自己免受捕食者的伤害的因子； F_i 为目标函数值。

2.2.3 逃避阶段

当防御行为无法抵御捕食者时，河马会选择逃离并寻找安全的水域以避免被捕食。这对应于防御阶段土壤采样布局未能产生良好效果时，算法选择放弃当前解，转而寻找一个更优的新位置以确保安全，如公式(7)所示。

$$X_{P_4}(i, :) = X(i, :) + r_{10} \left[\frac{X_{\min}(i)}{t} + r_{11} \left(\frac{X_{\max}(i)}{t} - \frac{X_{\min}(i)}{t} \right) \right] \quad (7)$$

式中： $X_{P_4}(i, :)$ 为河马逃往最近安全地点的位置；随机数 $r_{10} \in [0, 1]$ ；随机数 r_{11} 符合正态分布；*t*为当前迭代次数。

2.3 随机森林残差克里金插值

RFRK^[27,28]是一种混合模型，利用普通克里金法，对RF模型的预测残差进行空间插值后将其与RF模型的预测结果相加，从而提高模型预测精度^[5]。HO算法的每次迭代中，基于当前候选采样点集，利用RFRK模型预测土壤属性，并计算验证集上的RMSE作为目标函数。通过最小化预测误差，逐步搜索得到最优的采样密度和样本位置，用于土壤有机质含量空间预测。

RFRK需要调整超参数来平衡模型复杂度与泛化能力，本研究采用网格搜索的方法选择最优的参数组合，最终设置 $n_estimators=90$ 控制决策树数量以避免过拟合， $max_depth=8$ 限制单棵树深度防止过度复杂化。此外，将 $min_samples_split$ 和 $min_samples_leaf$ 均设为2，分别约束内部节点分裂和叶节点形成所需的最小样本量。半变异函数模型是克里金插值的核心部分，它用于描述数据的空间相关性，本研究的半变异函数模型为指数模型(Exponential)，设置参数 $n_closest_points=8$ ，这一参数表示在计算目标点的插值时，仅考虑距离目标点最近的8个已知数据点。

2.4 评估指标

本研究采用LCCC、 R^2 、MAE和RMSE评估模型的解释能力和预测精度。其中，当 R^2 、LCCC越趋近于1，RMSE、MAE越趋近于0时，模型的预测精度越高^[29]。

2.5 对比方法

为全面评估本研究提出的HO优化方法在采样布局中的性能，本研究选取了3种具有代表性的采样优化方法作为对比基准：cLHS^[10-12]、GA^[15,16]和粒子群优化算法(Particle Swarm Optimization, PSO)^[30]。为进行公平比较，其采样点规模及所依据的特征变量集与HO优化方法完全相同。

3 结果与分析

3.1 特征变量与目标变量的相关性分析

本研究通过计算SOM与各个特征变量之间的皮尔逊相关系数，并结合显著性检验($P < 0.05$)，判断其相关性的显著程度。如图4所示，所选的13个变量与SOM存在显著相关性。为进一步识别预测SOM的关键特征变量，本研究基于RF算法，采用平均不纯度减少(Mean Decrease Impurity, MDI)指标对特征重要性进行排序，结果如图5所示。

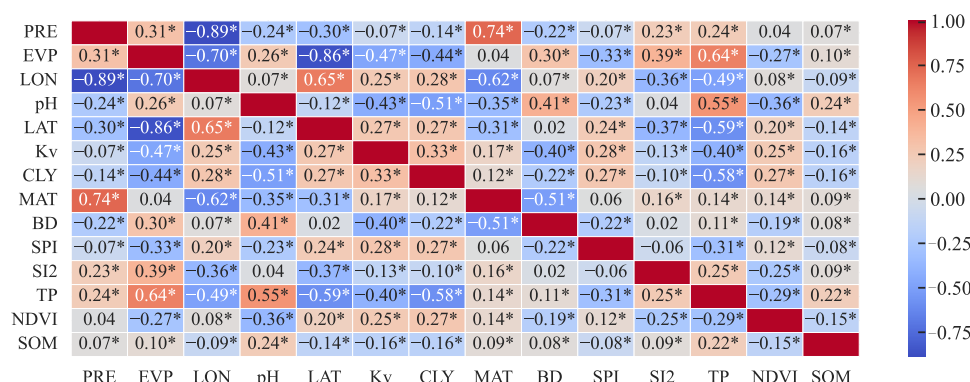
3.2 土壤有机质空间统计分析

表2提供了4个数据集的统计指标，包括训练样本池、验证集、测试集和全样本。这些数据集的统计指标包括样本数量、最大值、最小值、平均值、中位数和变异系数。训练样本池的取值范围与全样本一致，4个数据集的变异系数相差不大，适合参与模型训练。

3.3 HO的优化过程

根据采样密度的设计，利用本研究采用的HO采样优化方法，从训练样本池(756个样本)中优化出一系列样本大小，即 $n=136, 164, 191, 219, 246, 273, 301, 328, 356, 383, 410, 438, 465, 493, 521, 547, 575, 601, 629, 657$ (相应的采样密度分别为0.5, 0.6, 0.7, 0.8, ..., 2.3, 2.4点/km²)，作为土壤有机质预测的训练样本，通过河马优化算法的3个阶段，不断迭代优化，选出最具代表性的样本集。

图6以383个采样点(1.4点/km²)为例，展示了HO迭代优化过程中，RFRK模型在验证集上



注：*表示 $P < 0.05$ 。

图4 兰溪市特征变量与SOM含量的相关性分析

Fig. 4 The correlation of feature variables and SOM contents in Lanxi city

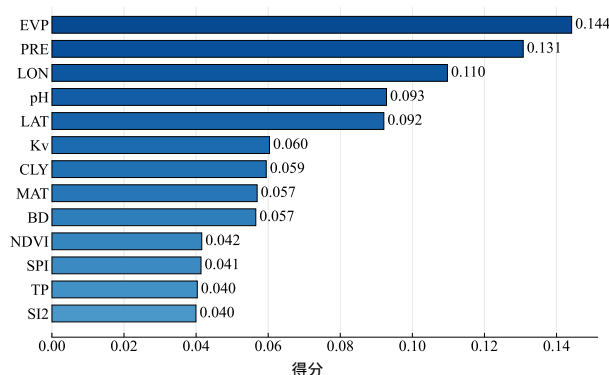


图5 兰溪市特征变量MDI重要性得分

Fig.5 The importance score of feature variables

MDI in Lanxi city

表2 兰溪市SOM含量统计分析

Table 2 Statistical analysis of SOM content in Lanxi city

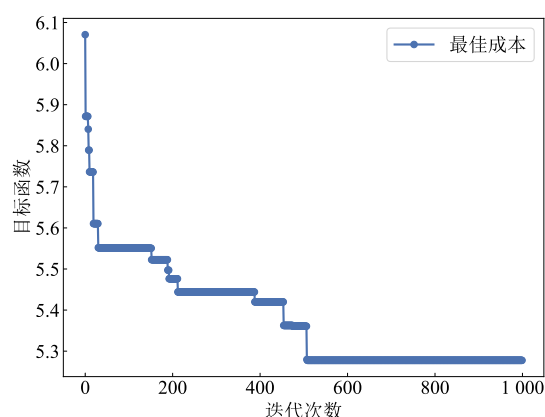
数据集	样本量	最大值/ (g/kg)	最小值/ (g/kg)	平均值/ (g/kg)	中位数/ (g/kg)	变异系数 CV/%
全样本	1 080	47.05	3.63	23.92	24.26	30.82
训练样本池	756	44.48	3.63	24.01	24.23	30.75
验证集	108	40.90	4.81	23.20	23.56	32.72
测试集	216	47.05	6.48	23.97	24.61	30.03

RMSE的变化趋势。随着迭代次数的增加，RMSE的值从6.07 g/kg下降到5.28 g/kg，起初0~100次迭代范围，目标函数值迅速下降，随后下降速度变慢，趋于平稳，陷入局部最优；在200~500次迭代范围，算法继续探索更优的方案，RMSE的值继续下降，最终在510~1 000次迭代范围，目标函数值收敛，找到最优值。

3.4 最佳采样密度及不同优化方法的性能对比

3.4.1 HO优化方法的最佳采样密度

为了分析采用HO优化的不同采样密度样本点



注：训练样本数=383。

图6 HO迭代优化过程中目标函数的收敛过程

Fig.6 Convergence process of the objective function during the iterative optimization of hippopotamus optimization algorithm

训练的模型在测试集上预测精度和解释能力的差异。本研究以SOM作为预测目标，使用RF和RFRK两种方法作为预测模型，绘制了不同采样密度条件下各评估指标（RMSE、MAE、 R^2 和LCCC）的变化图。

根据实验结果图7a和图7b，随着样本量增加，两种预测模型RF_HO和RFRK_HO的RMSE和MAE都在下降，RFRK_HO的误差小于RF_HO。在采样密度为0.5~2.3点/km²（136~629个采样点）时，RFRK_HO测试集样本的RMSE从6.04 g/kg显著降低到了5.11 g/kg，下降比例约为15.4%。值得注意的是，在采样密度为2.3点/km²（629个采样点）时，测试集上的预测误差达到最优水平，RMSE降低至5.11 g/kg，MAE为3.79 g/kg。当采样密度逐步递增至2.4点/km²的过程中，RMSE的变化趋于平稳，在0.1 g/kg左右轻微波动。观察图7c和

图 7d 不同采样密度的 SOM 预测拟合精度 R^2 和 LCCC 可见, RFRK_HO 的 R^2 从初始的 0.29 逐步上升至 0.49, LCCC 整体呈上升趋势。当采样密度为

2.3 点/ km^2 时, LCCC 值达到 0.63, 此时 R^2 为 0.49。此外, 采样密度大于 2.3 点/ km^2 之后, 随着采样点的继续增加, 性能并没有提升。

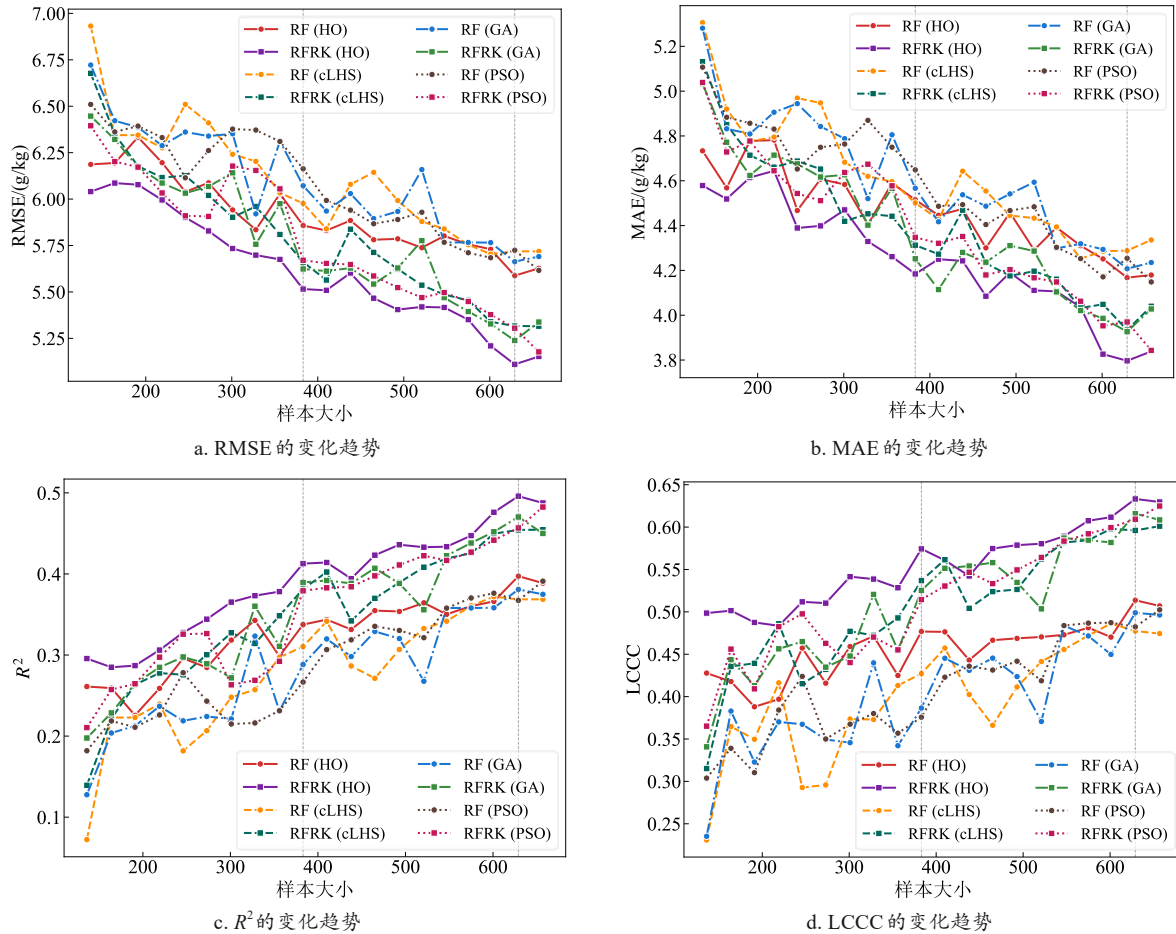


图 7 不同优化算法性能评估与比较

Fig.7 Performance evaluation and comparison of different optimization algorithms

总体来看, 随着采样密度的增加, RMSE、MAE 呈现出下降趋势, 而 LCCC、 R^2 则随之上升, 这与 BARCA 等^[6]、ŽIŽALA 等^[7]、SHAO 等^[14]、LIU 等^[31]的研究结果相似。通过 HO 算法的优化, 在 2.3 点/ km^2 的采样密度下, 不仅实现了最低的 RMSE 值, 同时也获得了最高的 LCCC 值。综合考虑 R^2 、LCCC、RMSE、MAE 等评估指标, 最优方案是 2.3 点/ km^2 , 此时的采样点为 629 个, 相比于原始的 1 080 个采样点, 采样点数量减少 451 个, 同时采样成本下降 41.8%。然而, 从成本效益的角度分析, 性价比最高的采样密度是 1.4 点/ km^2 , 此时仅需 383 个采样点即可实现 R^2 为 0.41、LCCC 为 0.57 的预测精度, 采样成本下降 64.5%。

3.4.2 HO 与 cLHS、GA、PSO 优化方法的比较

如图 7a~图 7d 所示, 本研究利用测试集样本,

对 HO 与 cLHS、GA、PSO 这 4 种采样优化方法在不同采样密度下的 SOM 空间预测性能进行了比较。

在采样密度为 0.5~1.3 点/ km^2 (136~356 个采样点) 时, 4 种方法的预测性能整体都偏低, 表现为 R^2 和 LCCC 较低, 而 RMSE 和 MAE 较高, 均未达到理想水平。以 $R^2=0.40$, LCCC=0.55 为分割线, 在 1.4 点/ km^2 (383 个采样点) HO 优化方法率先到达 $R^2=0.41$, LCCC=0.57 的预测精度, 而 cLHS 优化方法在该采样密度下 $R^2=0.38$, LCCC=0.53、GA 方法的 $R^2=0.39$, LCCC=0.52、PSO 方法的 $R^2=0.38$, LCCC=0.51 均低于 HO 方法。在 1.4~2.3 点/ km^2 (383~629 个采样点) 的采样密度下, HO 方法表现出更优的预测性能, 其 R^2 均高于 cLHS、GA、PSO 方法, 而 RMSE 均低于 cLHS、GA、PSO 方法。以采样密度 2.3 点/ km^2 (629 个采样点) 为例, 使用

RFRK作为预测模型, HO方法的 R^2 和LCCC分别达到0.49和0.63, 相较于cLHS方法($R^2=0.45$, LCCC=0.59)分别提升了约8.9%和6.8%; 相较于GA方法($R^2=0.47$, LCCC=0.61)分别提升了约4.3%和3.3%; 相较于PSO方法($R^2=0.45$, LCCC=0.60)分别提升了约8.9%和5.0%。同时, 其误差指标RMSE和MAE分别为5.11 g/kg和3.79 g/kg, 低于cLHS方法(RMSE=5.31 g/kg, MAE=3.93 g/kg)、GA方法(RMSE=5.24 g/kg, MAE=3.93 g/kg)和PSO方法(RMSE=5.30 g/kg, MAE=3.97 g/kg)。

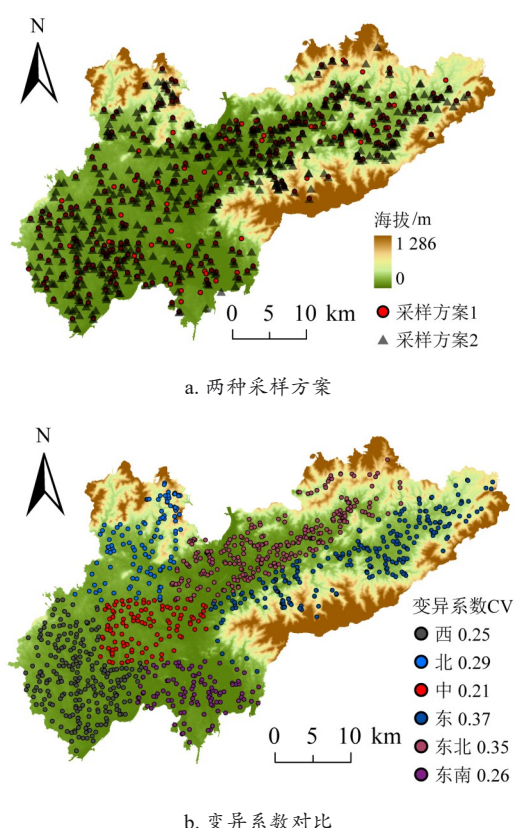
综合结果表明, HO优化方法在整体性能上优于cLHS、GA、PSO方法, 进一步验证了其在不同采样密度下优化土壤采样点布局的适用性与有效性。

3.5 HO优化后的采样点位置分析

在不同的采样密度下, 使用HO采样优化方法得到的样本点呈现出不一样的空间分布。以最优采样密度2.3点/ km^2 和最具成本效益采样密度1.4点/ km^2 为例。如图8a所示, 采样方案1代表采样密度1.4点/ km^2 , 采样方案2代表采样密度2.3点/ km^2 。两种方案的采样点分布均覆盖研究区整体, 但并不是均匀分布。HO算法根据目标函数的计算结果进行优化, 考虑采样点的空间异质性, 继而调整采样位置的分布, 总体上表现出在变异性强的区域采样点密集, 变异性弱的区域采样点稀疏。1.4点/ km^2 采样密度下, 在区域中部布设的采样点较少, 因为中部SOM变异系数最低, 如图8b所示; 随着采样密度的增加, 在2.3点/ km^2 时, 东部和东北部高海拔区域增加较多的点, 因为其SOM变异系数较高。

3.6 SOM的制图结果

不同的采样密度的样本点使用RFRK方法的制图结果如图9所示, 图9a和图9b的样本点数分别为383个和629个, 分别对应最具成本效益采样密度1.4点/ km^2 和最佳采样密度2.3点/ km^2 , 图9c和图9d的样本点数分别为756个和1 080个, 分别为训练集全部采样点和研究区所有采样点。所有采样方案的SOM含量分布格局基本相似, 东部和南部SOM含量高, 西部次之, 中部较低。629个点的预测范围和全样本的预测范围较为相似, 表明通过HO优化, 能去除冗余采样点, 以更少的成本获取较为精确的SOM含量图。



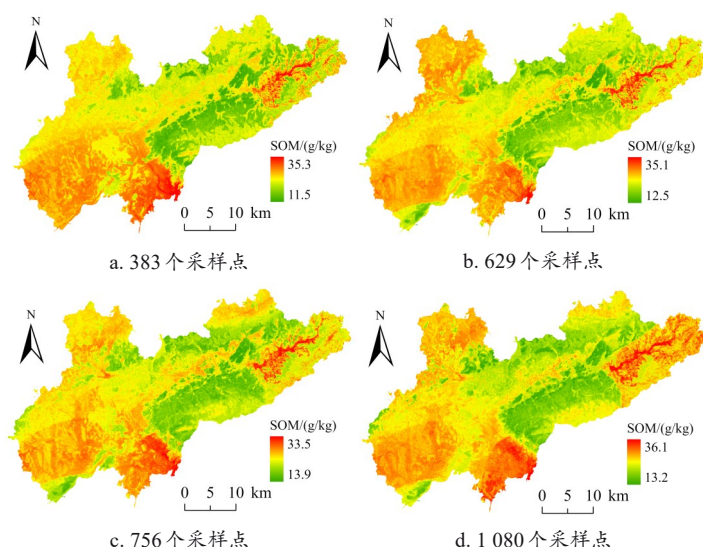
注: 该图基于自然资源部标准地图服务网站下载的审图号为GS (2019)3266号标准地图制作, 底图无修改。

图8 两种土壤采样密度的空间位置 and 不同区块的变异系数
Fig. 8 Spatial distribution of soil sampling points under two densities and coefficients of variation for different blocks

4 讨论

在本研究所研究的区域, HO算法能够根据所需的采样密度优化得出相应的采样点, 在高采样密度下土壤有机质的预测误差RMSE能降低至较小的值。在到达一定采样密度时, RMSE的下降趋于平稳, 单纯增加样本量不能显著地提升模型的效果, 过多的采样点可能会造成数据冗余, 继而影响模型的预测效果, 在前人的研究中也存在类似情况^[5,6]。HO在土壤采样点优化中具有充足的潜力, 一方面是因为HO算法本身的全局探索能力很强, 能够跳出局部最优; 另一方面依赖于RFRK模型相较于传统的克里金模型能够更好地处理非线性关系、融合多源辅助数据, 并降低对空间平稳性假设的依赖, 从而在复杂环境下的预测精度和适应性上更具优势^[32]。

特征变量是保证模型预测精度的重要训练参数, 合理选择具有代表性的特征变量十分关键。如果本身特征变量与目标变量的相关性就不强, 那么无论优化算法设计得多么精巧, 最终的效果都是不



注:该图基于自然资源部标准地图服务网站下载的审图号为GS(2019)3266号标准地图制作,底图无修改。

图9 兰溪市不同采样方案的SOM制图结果

Fig.9 Mapping results of SOM under different sampling schemes in Lanxi city

理想的。此外,本研究在特征变量种类的选择上仅仅参照前人在相似土壤环境中的研究,可能存在地理环境上的偏差。特征变量的选择多基于自然环境因子,随着土地利用的复杂多样化,人类活动因子对土壤养分累积的作用越来越强,但是这类变量的收集比较困难。在今后的研究中,可以探究人类活动因子与土壤养分的关联性和可记录性,如使用夜间灯光数据、兴趣点(Point of Interest, POI)密度、交通可达性等替代指标来间接表征。

关于算法在不同地貌区的适用性问题。本研究基于丘陵盆地地貌所构建的HO最优采样框架,其参数配置和性能表现可能受区域地理特征影响,在推广至平原、山地等其他地貌类型时,需谨慎评估其适应性。由于不同地貌区在地形复杂度、土壤异质性及环境驱动因子空间格局上存在显著差异,直接迁移当前参数可能导致采样效率下降或模型预测偏差。因此,在应用于新区域时,建议根据具体地貌特征对HO算法的种群规模、迭代次数及RFRK模型的参数、协变量组合进行重新优化与验证。

5 结论

本研究提出了一种基于HO的新型土壤采样优化策略,以兰溪市土壤有机质预测作为测试,结论如下:

1) RFRK作为预测模型时,HO优化的采样方案,在2.3点/km²获得最优,测试集LCCC为0.63,

RMSE降至5.11 g/kg,优化后采样成本较原始方案下降41.8%;当采样密度为1.4点/km²时, RMSE从最初(0.5点/km²)的6.04 g/kg下降至5.51 g/kg, LCCC为0.57,采样成本下降64.5%。兼顾采样成本和预测精度,土壤采样密度可以设置为1.4~2.3点/km²。

2) 对比两种预测模型,在一系列采样密度下, RFRK的预测精度均高于RF,证明在SOM预测中RFRK具有优势,将HO与RFRK结合的采样策略是有效的。

综上,本研究以较少的采样点能够达到较高的预测精度,能减少土壤有机质含量长期监测的采样成本,可为兰溪市以及相似环境的丘陵盆地地区采样点优化提供参考。

利益冲突声明: 本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献:

- [1] 杨贵军,赵春江,杨小冬,等.粮食生产大数据平台研究进展与展望[J].智慧农业(中英文),2025,7(2):1-12.
YANG G J, ZHAO C J, YANG X D, et al. Grain production big data platform: Progress and prospects[J]. Smart Agriculture, 2025, 7(2): 1-12.
- [2] 何发坤,蒲生彦,肖胡萱,等.遥感技术在土壤退化中的应用研究进展[J].农业资源与环境学报,2021,38(1):10-19.
HE F K, PU S Y, XIAO H X, et al. Review of remote sensing application in soil degradation[J]. Journal of Agricultural Resources and Environment, 2021, 38(1): 10-19.
- [3] RADOČAJ D, JUG D, JUG I, et al. A comprehensive evaluation of machine learning algorithms for digital soil organic carbon mapping on a national scale[J]. Applied Sciences, 2024, 14(21): 9990.
- [4] TOBORE A O, NKWUNONWO U C, ABDUSSALAAM S A, et al. Random forest algorithm and remote sensing techniques for wetland soil organic carbon prediction towards environmental sustainability[J]. Discover Environment, 2025, 3(1): 222.
- [5] 周洋,赵小敏,郭熙.基于多源辅助变量和随机森林模型的表层土壤全氮分布预测[J].土壤学报,2022,59(2):451-460.
ZHOU Y, ZHAO X M, GUO X. Prediction of total nitrogen distribution in surface soil based on multi-source auxiliary variables and random forest approach[J]. Acta Pedologica Sinica, 2022, 59(2): 451-460.
- [6] BARCA E, DE BENEDETTO D, STELLACCI A M. Optimization of sampling design for soil total organic carbon assessment in the precision agriculture framework: Impact of different variogram models and potentiality of ground penetrating radar (GPR) covariate information[J]. Computers and Electronics in Agriculture, 2024, 226: 109470.
- [7] ŽÍŽALA D, PRINC T, SKÁLA J, et al. Soil sampling design matters-Enhancing the efficiency of digital soil mapping at the field scale[J]. Geoderma Regional, 2024, 39: e00874.
- [8] PETROVSKAIA A, GASANOV M, NIKITIN A, et al. Maximizing dataset variability in agricultural surveys with spatial sampling based on MaxVol matrix approximation[J]. Precision Agriculture, 2024, 26(1): 9.
- [9] 黄思华,濮励杰,解雪峰,等.面向数字土壤制图的土壤采样设计研究进展与展望[J].土壤学报,2020,57(2):259-272.
HUANG S H, PU L J, XIE X F, et al. Review and outlook of de-

- signing of soil sampling for digital soil mapping[J]. *Acta Pedologica Sinica*, 2020, 57(2): 259-272.
- [10] KHAN A, AITKENHEAD M, STARK C R, et al. Optimal sampling using conditioned Latin hypercube for digital soil mapping: An approach using Bhattacharyya distance[J]. *Geoderma*, 2023, 439: 116660.
- [11] SAURETTE D D, BISWAS A, HECK R J, et al. Determining minimum sample size for the conditioned Latin hypercube sampling algorithm[J]. *Pedosphere*, 2024, 34(3): 530-539.
- [12] 李维友, 段良霞, 谢红霞, 等. 基于条件拉丁超立方抽样的县域耕地土壤有机质空间插值合理样本密度的确定[J]. *土壤通报*, 2022, 53(3): 505-513.
- LI W Y, DUAN L X, XIE H X, et al. Determination of reasonable sample density for spatial interpolation of soil organic matter in cultivated land of county region based on conditional Latin hypercube sampling[J]. *Chinese Journal of Soil Science*, 2022, 53(3): 505-513.
- [13] MOLLA A, ZUO S D, ZHANG W W, et al. Optimal spatial sampling design for monitoring potentially toxic elements pollution on urban green space soil: A spatial simulated annealing and k-means integrated approach[J]. *The Science of the Total Environment*, 2022, 802: 149728.
- [14] SHAO S S, SU B W, ZHANG Y L, et al. Sample design optimization for soil mapping using improved artificial neural networks and simulated annealing[J]. *Geoderma*, 2022, 413: 115749.
- [15] DADA B A, NWULU N I, OLUKANMI S O. Bayesian optimization with Optuna for enhanced soil nutrient prediction: A comparative study with genetic algorithm and particle swarm optimization[J]. *Smart Agricultural Technology*, 2025, 12: 101136.
- [16] WU X Y, LI Y, WU K N, et al. GA-optimized sampling for soil type mapping in plain areas: Integrating legacy maps and multi-source covariates[J]. *Agronomy*, 2025, 15(4): 963.
- [17] AMIRI M H, MEHRABI HASHJIN N, MONTAZERI M, et al. *Hippopotamus* optimization algorithm: A novel nature-inspired optimization algorithm[J]. *Scientific Reports*, 2024, 14(1): 5032.
- [18] RUSS A, RIEK W, WESSOLEK G. Three-dimensional mapping of forest soil carbon stocks using SCORPAN modelling and relative depth gradients in the north-eastern Lowlands of Germany[J]. *Applied Sciences*, 2021, 11(2): 714.
- [19] SINDHUSHREE T S, KAVYA D, JITENDRA G H, et al. Digital soil mapping: A review of techniques, applications and emerging trends[J]. *Journal of Scientific Research and Reports*, 2025, 31(7): 1151-1158.
- [20] ZHANG Y, LUO C, ZHANG W Q, et al. Mapping soil organic matter in black soil cropland areas using remote sensing and environmental covariates[J]. *Agriculture*, 2025, 15(3): 339.
- [21] LIU X N, WANG M C, LIU Z W, et al. Improving spatial prediction of soil organic matter in typical black soil area of Northeast China using structural equation modeling integration framework[J]. *Computers and Electronics in Agriculture*, 2025, 236: 110404.
- [22] JO Y, PANJA P, KIM H, et al. Soil organic carbon (SOC) prediction using super learner algorithm based on the remote sensing variables[J]. *Environmental Challenges*, 2025, 19: 101160.
- [23] 李安琪, 杨琳, 蔡言颜, 等. 基于递归特征消除-随机森林模型的江浙沪农田土壤肥力属性制图[J]. *地理科学*, 2024, 44(1): 168-178.
- LI A Q, YANG L, CAI Y Y, et al. Digital mapping of soil fertility attributes in croplands in Jiangsu, Zhejiang and Shanghai based on recursive feature elimination-random forest model[J]. *Scientia Geographica Sinica*, 2024, 44(1): 168-178.
- [24] 郭静, 龙慧灵, 何津, 等. 基于 Google Earth Engine 和机器学习的耕地土壤有机质含量预测[J]. *农业工程学报*, 2022, 38(18): 130-137.
- GUO J, LONG H L, HE J, et al. Predicting soil organic matter contents in cultivated land using Google Earth Engine and machine learning[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2022, 38(18): 130-137.
- [25] 张晓婷, 黄魏, 傅佩红, 等. 基于特征筛选算法的数字土壤制图研究[J]. *土壤学报*, 2024, 61(3): 635-647.
- ZHANG X T, HUANG W, FU P H, et al. Research on digital soil mapping based on feature selection algorithm[J]. *Acta Pedologica Sinica*, 2024, 61(3): 635-647.
- [26] WANG H B, BINTI MANSOR N N, MOKHLIS H B. Novel hybrid optimization technique for solar photovoltaic output prediction using improved *Hippopotamus* Algorithm[J]. *Applied Sciences*, 2024, 14(17): 7803.
- [27] 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.
- WANG Y X, YANG K, GAO B B, 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*, 2022, 38(12): 65-73.
- [28] HO V H, MORITA H, BACHOFER F, et al. Random forest regression Kriging modeling for soil organic carbon density estimation using multi-source environmental data in central Vietnamese forests[J]. *Modeling Earth Systems and Environment*, 2024, 10(6): 7137-7158.
- [29] CHEN Z X, WANG Z, WANG X, et al. Including soil spatial neighbor information for digital soil mapping[J]. *Geoderma*, 2024, 451: 117072.
- [30] 张世文, 朱曾红, 王维瑞, 等. 基于粒子群-随机森林模型的采样布局优化[J]. *安徽理工大学学报(自然科学版)*, 2023, 43(6): 37-44.
- ZHANG S W, ZHU Z H, WANG W R, et al. Sample layout optimization based on particle swarm-random forest model[J]. *Journal of Anhui University of Science and Technology (Natural Science)*, 2023, 43(6): 37-44.
- [31] LIU Y Q, JIANG C L, FENG A P, et al. A causal prediction method for soil organic carbon storage change estimation, with Shaanxi province as a case study[J]. *Computers and Electronics in Agriculture*, 2025, 234: 110271.
- [32] 徐英, 谢若禹, 沈丽佳, 等. 基于回归克里格法的土壤盐分采样点布局优化[J]. *农业机械学报*, 2022, 53(8): 275-282.
- XU Y, XIE R Y, SHEN L J, et al. Layout optimization of soil salt sampling points based on regression Kriging[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2022, 53(8): 275-282.

Optimal Sampling Strategy for Soil Organic Matter Based on Hippopotamus Optimization Algorithm and Machine Learning

LIAN Zhenxiang^{1,2}, FEI Xufeng², REN Zhouqiao^{2*}

(1. School of Mathematics and Computer Science, Zhejiang A & F University, Hangzhou 311300, China;

2. Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China)

Abstract: [Objective] Soil quality is crucial for food security, ecosystem health, and sustainable development, but faces degradation

due to intensive land use. Accurate soil quality assessment is therefore essential for informed land management and ecological protection. Machine learning has enhanced digital soil mapping (DSM) by improving modeling accuracy through multi-source data integration. Within DSM, soil sampling design is a foundational step that directly influences prediction accuracy, cost, and efficiency. An ideal scheme must balance mapping precision with economic and operational feasibility. This study focuses on soil organic matter (SOM), a core indicator of soil quality affecting fertility, carbon sequestration, and environmental regulation. Precisely mapping its spatial variability is vital for sustainable soil management. To address the need for efficient sampling, the aim of this research is to develop an optimal sampling design method for regional-scale SOM mapping, reduce sampling redundancy and cost while improving spatial prediction accuracy. **[Methods]** A sampling optimization framework was proposed that integrated intelligent optimization algorithms with a hybrid spatial interpolation model. The framework was built upon the hippopotamus optimization algorithm (HO) and incorporated the random forest residual kriging (RFRK) method to construct an optimal sampling strategy for the spatial prediction of SOM. At the initialization stage, a population of candidate solutions, referred to as "hippopotamuses", was randomly generated, with each individual representing a potential sampling layout. The HO was employed to select subsets of sampling points from the training sample pool, with each subset forming a candidate solution. Collectively, these solutions constituted the initial hippopotamus population. The study area was located in Lanxi city, Zhejiang province, where a total of 1 080 field-measured soil samples were collected. These samples were partitioned into a training set ($n=756$), a validation set ($n=108$), and a test set ($n=216$) at a ratio of 7:1:2. Environmental covariates, including terrain attributes, vegetation indices, and climate factors, were extracted from multi-source remote sensing datasets. Using these covariates, the HO optimized sampling schemes across varying densities and spatial configurations. The resulting designs were then evaluated using the RFRK model to assess their SOM prediction performance. This process enabled the identification of the optimal sampling density and spatial layout that balanced accuracy and cost-efficiency. **[Results and Discussions]** When the HO-RFRK framework was applied, the prediction accuracy of SOM improved significantly as sampling density increased from 0.5 to 2.3 points/km² (136–629 points). The root mean square error (RMSE) on the test set decreased from 6.04 to 5.11 g/kg, representing a reduction of approximately 15.4%. The lowest prediction errors were observed at a sampling density of 2.3 points/km², with the RMSE and mean absolute error (MAE) reaching their minimum values of 5.11 and 3.79 g/kg, respectively, beyond which further increases yielded only marginal gains, indicating diminishing returns. To assess the effectiveness of HO, its performance was compared with three established methods: conditioned Latin hypercube sampling (cLHS), genetic algorithm (GA), and particle swarm optimization (PSO). At lower densities (0.5–1.3 points/km²), all methods showed limited predictive power. However, at 1.4 points/km² (383 points), the HO method was the first to exceed predefined accuracy thresholds (coefficient of determination, $R^2>0.40$; Lin's concordance correlation coefficient, $LCCC>0.55$), achieving $R^2=0.41$ and $LCCC=0.57$, outperforming cLHS ($R^2=0.38$, $LCCC=0.53$), GA ($R^2=0.39$, $LCCC=0.52$), and PSO ($R^2=0.38$, $LCCC=0.51$). Across the range of 1.4–2.3 points/km², HO consistently delivered superior results. At 2.3 points/km², the HO-RFRK combination achieved $R^2=0.49$ and $LCCC=0.63$, surpassing cLHS, GA, and PSO in both metrics. **[Conclusions]** Based on the cultivated land of Lanxi city as a test case, a novel sampling optimization strategy was proposed based on the HO. First, the strategy successfully identified an optimal sampling density that maximizes prediction accuracy, as well as a lower, cost-effective density that maintains robust predictive performance with substantially reduced survey costs, defining a practical density range that balances precision and economic feasibility. Second, the RFRK model consistently demonstrated superior prediction accuracy compared to the standard random forest (RF) model across all tested sampling schemes, validating the effectiveness of the integrated HO-RFRK approach. In summary, this optimized strategy achieves high mapping accuracy with greater sampling efficiency, offering a scientifically grounded and practical methodology for reducing long-term soil monitoring costs. It provides a valuable reference for optimizing soil surveys in Lanxi city and other regions with similar environmental settings.

Keywords: hippopotamus optimization algorithm; sampling optimization; random forest; kriging interpolation; digital soil mapping; remote sensing mapping

Foundation items: National Key Research and Development Program (2023YFD1902900)

First author: LIAN Zhenxiang, E-mail: lianzx2025@163.com

***Corresponding author:** REN Zhouqiao, E-mail: renzq@zaas.ac.cn