李坤,陈宇昊,李文岳,等.基于核密度估计的土壤样本代表性修正研究[J].华中农业大学学报,2025,44(1):94-104.
DOI:10.13300/j.cnki.hnlkxb.2025.01.010

基于核密度估计的土壤样本代表性修正研究

李坤,陈宇昊,李文岳,王子影,傅佩红,黄魏

华中农业大学资源与环境学院,武汉 430070

摘要 为充分利用历史样点数据获取更可靠的土壤-环境知识,进而获取更高精度的土壤预测推理图,采用样本代表性修正方法获取更高的知识精度,利用样本空间与总体空间环境协变量的空间相似度关系,以核密度估计为基础,采用3种不同算法对每个土壤采样点探寻最优权重,并以土壤表层有机质含量预测制图为例验证方法的科学性和有效性。结果显示,该修正方法最高可将多元线性回归制图的RMSE和MAE分别降低10.30%和12.74%,证实了该方法的可行性与有效性。

关键词 环境协变量;空间偏差修正;样本代表性;启发式算法;数字土壤制图;历史样点

中图分类号 S159.9 **文献标识码** A **文章编号** 1000-2421(2025)01-0094-11

土壤采样是数字土壤制图研究的重要内容,从采样点获取数据的准确性直接影响土壤制图的精确度,如何从已存在的历史样点获取更可靠的土壤-环境知识已成为土壤制图中的重要科学问题。野外采样点是数字土壤制图研究过程中的重要数据,并且在很多情况下成为数字土壤制图可靠的数据源[1],其中,具有代表性的样点在应用、成本、准确度等几个方面均有明显优点[2]。为了提高样点的代表性,一般需要提前设计采样方案,合理设计的采样点能够充分发挥土壤样点的价值[3],常见的采样设计方法主要有经典采样法、目的性采样法、辅助变量法、样点代表性等级法等[4-7]。对于提高土壤样点代表性的研究大多集中在采样前的采样方案设计,或对现有土壤采样点进行补充采样[8]以达到提高土壤制图精度的目的。但是对于已经采集到的样点,研究人员基本不可能再去还原或重建当时样点获取过程的地理场景信息,这就导致样点在空间代表性上的欠缺无法弥补,进而对土壤属性空间预测结果造成不同程度的影响[9]。因此,为了更加充分地利用花费大量人力物力采集土壤的样点,将其进行空间代表性修正就显得尤为重要。

空间偏差修正在多个领域都有所发展,如统计学、机器学习、预测制图等领域,研究结果表明,通过合理的空间偏差修正可提高所需结果的可靠性[10-12]。基于土壤样本代表性的空间偏差修正方法起源和发展于志愿者地理信息(volunteered geographic information,VGI),VGI是一种由任何公民,可以在任何地点、任何时间自愿地将包含地理信息特征的数据进行编辑并通过网络形式上传到网站,由志愿审稿人审查数据的准确性和重要性的方法[13]。目前空间偏差修正研究已取得一定进展,如局部预测训练模型、样本过滤、基于累积可视化的样本加权以及根据因子重要性程度选择性建模等[14-17]。这些研究成果依赖于潜在的抽样或样本获取过程的信息,在一定程度上可以修正样本代表性,然而,对于数字土壤制图而言,其在地理环境空间信息的考虑上仍有一定欠缺。

地形信息和遥感数据等作为环境协变量不仅可以用于辅助采样设计,而且可以用于直接或间接制图,两者均被证明是可行且精度较高的[18-19]。因此,本研究基于土壤景观模型和VGI的样点修正理论,利用低维的土壤样点与整体区域高维的环境协变量的相似度关系进行空间代表性修正,并使用不同的算法对不同土壤样本预测土壤有机质(soil organic matter,SOM)含量分布的修正结果进行比较,从而验证土壤样本代表性修正方法的科学性和有效性,

收稿日期:2023-12-05
基金项目:国家自然科学基金项目(42171056;41877001)
李坤,E-mail:2601277070@qq.com
通信作者:黄魏,E-mail:ccan@mail.hzau.edu.cn

更加充分地利用和挖掘已有样点数据获取更高精度的土壤预测推理图，旨在为土壤样本的数据处理提供技术支撑。

1　材料与方法

1.1　研究区概况

江夏区位于湖北省武汉市南部，地处114°01′～114°35′E、29°58′～30°32′N，属于江汉平原向鄂南丘陵过渡地段，东临梁子湖，中部由107国道贯穿南北，西靠长江，丘陵地形主要呈条带状分布在区境北部。全区属中亚热带过渡的湿润季风气候，地面高程20～40 m，年总降水量889.2～1 862.6 mm。江夏区北部拢冈平原以黄棕壤和潮土为主，中部低丘拢冈平原皆为红壤，西部平原以潮土为主，其面积占比大小分别是水稻土＞红壤＞黄棕壤＞潮土，全区的六大主导产业包括“菜、瓜、莲、鱼、畜、林”等。

1.2　数据准备

研究区环境协变量数据包括高程数据和遥感数据。数字高程数据来源于地理空间数据云（https://www.gscloud.cn/），遥感数据来源于欧洲航天局官网（https://scihub.copernicus.eu/dhus/#/home），经过对比不同时期的遥感影像，选取2021年12月10日哨兵二号10 m×10 m分辨率的MSIL1C遥感数据产品，其云量覆盖度较低、植被覆盖度适中、地物特征明显，经处理后得到最终遥感数据。

本研究在江夏区行政区划图（江夏区自然资源和规划局2021年土地调查与变更数据）的基础上，依据GB/T 33469—2016《耕地质量等级》耕地地力评价规范，兼顾土壤类型与土地利用类型获取研究区92个土壤采样点，测定土壤有机质含量，根据随机抽样原则将92个采样点中的60个点作为样本点，32个点作为验证点，其中将样本点平均分为A、B两组各30个点进行代表性修正研究和制图，并均用验证点进行精度检验，研究区验证点位置分布图如图1所示。

1.3　环境协变量处理

1）基础数据。由于研究区面积范围较小，气候和母质较为单一，所以在进行有机质含量制图时，选取DEM、坡度、坡向、地形湿度指数（TWI）、平面曲率、剖面曲率、归一化植被指数（NDVI）等较为合理的地形因子和遥感因子进行制图研究[20]。利用渔网工具将每个网格的中心点作为采样点得到总体样点分布，通过比较300、600、900、1 200、1 500 m为间隔的采样规则并计算样本代表性误差，最终选择以900 m为间隔提取的总体样点（即900 m以下的采样间隔对样本代表性计算的误差＜0.000 1），样本点和总体样点分布如图2所示。

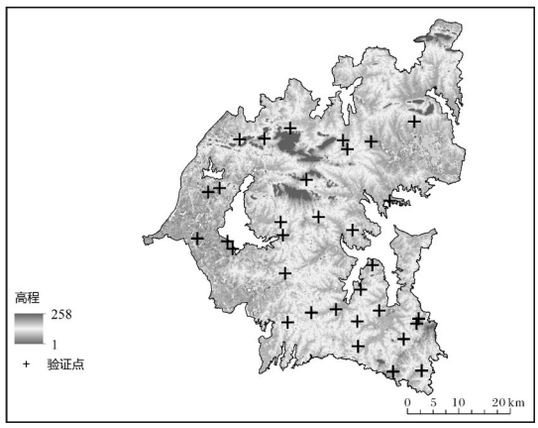


图1　研究区验证点位置分布图

Fig.1　Verification point location distribution map in the study area

2）样本空间和总体空间。多维环境协变量数据存在冗余性或可能的相关性，本研究将7个协变量数据分别提取到总体样点，并根据拉依达准则剔除异常值[21]得到1 832个总体样点进行主成分分析。结果显示，环境协变量数据的前3个主成分分别占有64.20%、18.95%、12.39%的信息，涵盖了协变量图层95.54%的信息积累，故将第一、二、三主成分图层（PC1、PC2、PC3）作为新的协变量图层用于后续的分析，新的协变量数据如图3所示。将3个协变量图层数据提取至样本点和总体样点，分别构成样本空间A、B和总体空间数据集。

1.4　样本代表性计算

1）核密度估计。核密度估计（kernel density estimation，KDE）是一种非参数统计方法，基本思想是将每个观测值周围一定范围内的数据计算权重后加权平均，形成估计的概率密度函数[22]。本研究通过核密度估计模拟样本空间和总体空间的概率密度分布，核函数选择只有带宽1个参数的高斯核[23]。对于少量样本的样本空间，采用网格搜索法估计每组样本协变量带宽最优解；对于大量样本的总体空间，基于样本服从正态分布的性质，通过经验法则法估计每组协变量空间的带宽最优解，通过上述方法计算得到的最优带宽如表1所示。本研究所有计算均通过Python代码实现。

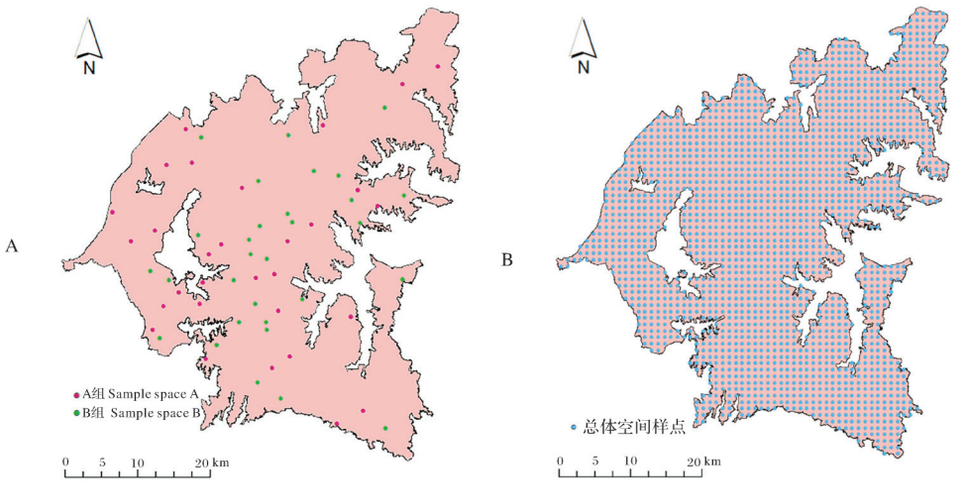


图2 样本空间(A)和总体空间(B)的采样点分布

Fig.2 Sample points distribution in sample space(A) and population space(B)

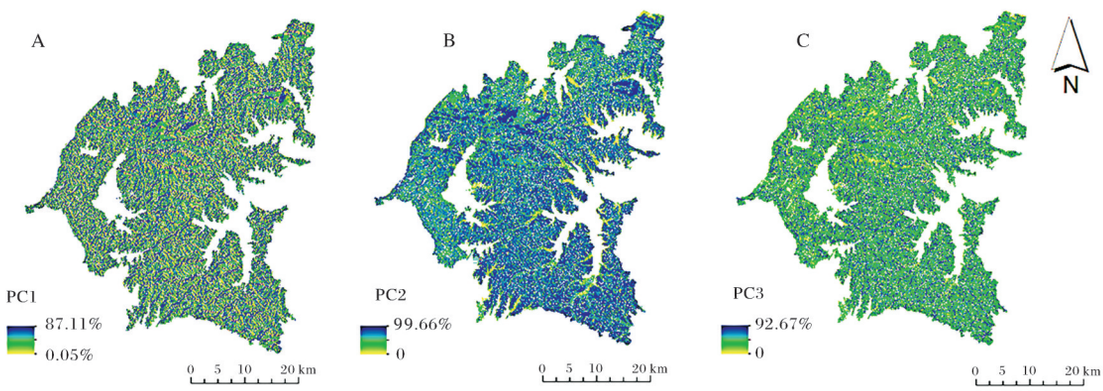


图3 第一(A)、二(B)、三(C)主成分协变量图层

Fig.3 First(A), second(B), and third(C) principal component covariate layers

表1 协变量空间最优带宽

Table 1 Optimal bandwidth of the covariate space

分组 Group	PC1	PC2	PC3
样本空间A Sample space A	2.595 0	4.328 8	4.534 9
样本空间B Sample space B	6.280 3	6.579 3	2.595 0
总体空间 Global space	5.779 7	2.729 2	2.374 3

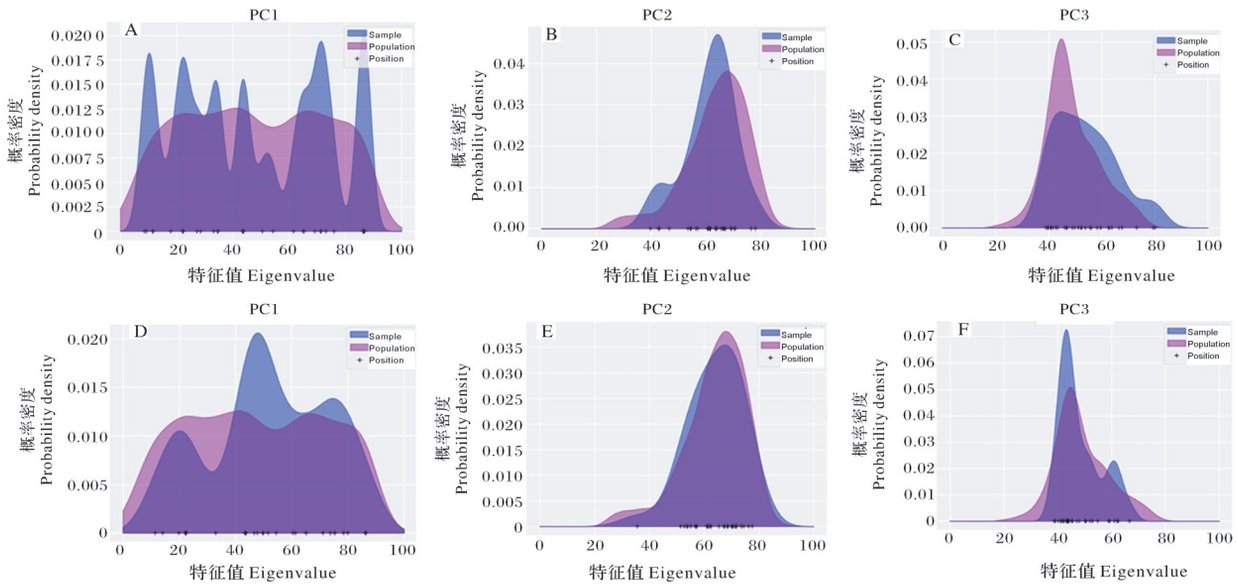
核密度估计使用Scikit-learn包的Neighbors模块实现,通过将每个数据点周围的概率密度权重化估计该点附近的密度,从而模拟样本空间和总体空间的概率密度分布。在数据处理阶段将主成分分析的特征值范围缩放到[0,100],主成分分析的特征值指的是每个主成分坐标轴对应的主成分变量能解释多少原始数据中的变异(方差)。在[0,100]范围等距生成10 000个观测点用于拟合函数曲线,拟合的函数曲线如图4所示,拟合结果显示,样本空间在第一主成分上存在明显的多峰状况,且更多的覆盖坐标轴区域,这不仅表明第一主成分包含样本空间绝大部分的信息量,更直观地体现了样本空间与总体空间的偏差性,第二、第三主成分所占信息量较少,但是依然反映了多维协变量空间样本的偏差程度。

2)样本代表性计算。数字土壤制图中可以选用总体的子集-样本来代替总体进行制图,但是选取的样本往往要具有代表性[24]。本研究中样本代表性可以解释为样本和总体的相似程度。利用选取的L个环境协变量空间相似度的加权平均值计算样本分布和总体分布之间的相似度,即为样本空间的代表性[25]。每组协变量的相似度和总体相似度的计算公式为:

$$S_{IM}{}^{l}=\frac{2\times A_{Q}{}^{l}\cap A_{P}{}^{l}}{A_{Q}{}^{l}+A_{P}{}^{l}} \tag{1}$$

$$R=S_{IM}{}^{overall}=\frac{\lambda^{i}}{\sum_{j=1}^{L}\lambda^{j}}S_{IM}{}^{i} \tag{2}$$

其中,$A_{Q}{}^{l}$和$A_{P}{}^{l}$分别是样本空间和总体空间概



A-C：样本空间 A Sample space A；D-E：样本空间 B Sample space B. A，D：PC1；B，E：PC2；C，F：PC3.

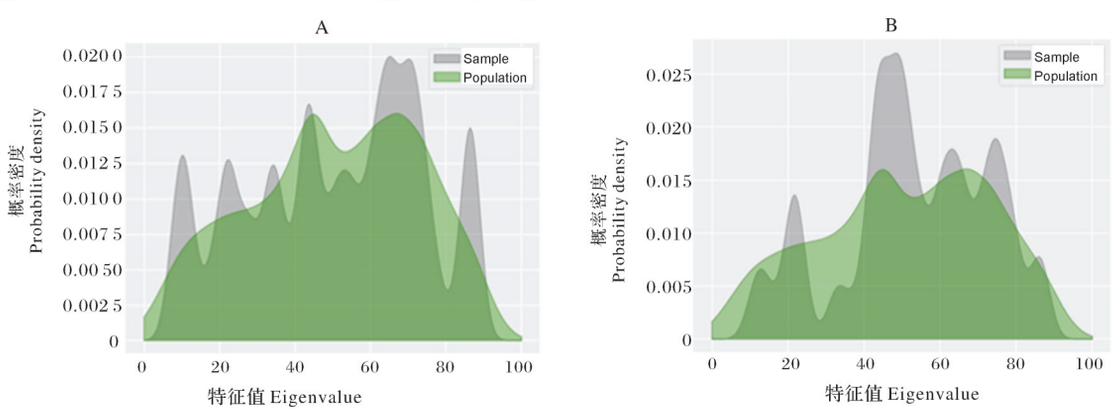
图 4　每个协变量组分样本空间和总体空间的概率密度曲线

Fig.4　Probability density curves of sample space and population space for each covariate component

率密度函数与x轴构成的面积，$A_Q^l \cap A_P^l$为函数重叠的面积，S_{IM}^l表示第L个协变量的相似度。样本的代表性R即为总体相似度$S_{IM}^{overall}$，R越大表明样本代表性越高，其取值范围为0～1。S_{IM}^i为分量i的相似度，对于L个分量，λ^i表示了i组分总体空间的方差，其归一化方差比值代表了i组分所占的权重比例，即较大方差的分量可以反映更多的协变量信息。

概率密度函数的积分是概率分布函数，其曲线下与x轴构成的面积为1，将公式(1)约分后，每组协变量空间的相似度等于样本空间和总体空间概率密度函数重叠部分的面积。比较图4中样本空间和总体空间概率密度函数值(10 000对数据)，取比较后的最小值作为重叠面积概率密度取值，构成重叠面积概率密度曲线，对其进行积分得到每组协变量的相似度，继而通过公式(2)计算总相似度，即样本代表性值。

通过上述方法计算样本空间A、B组代表性值分别为0.813 8和0.866 0。绘制样本空间A、B组分别与总体空间的总概率密度曲线，如图5所示，由数据和图5可知，样本空间A组环境协变量的样本代表性低于样本空间B组，说明样本空间A、B组在数据分布上存在一定差异。



A：样本空间 A Sample space A；B：样本空间 B Sample space B.

图 5　样本空间和总体空间的总概率密度曲线

Fig.5　Total probability density curve of sample space and population space

1.5　样本代表性修正

1)样本权重处理。样本代表性的修正是通过计算1组基于每个样本点的最优权重，并将其加权于样本空间的概率密度分布，使样本的代表性最大化来

实现的。在样本空间估计的概率密度函数中，每个样本点均具有1个标准化权重，且与核函数相乘。在初始的核密度估计中，每个样本的标准化权重是相等的，且和为1，而在最优权重确定时，是通过优化算法进行加权来寻找总相似度最大值，代表性较高的样点得到较小的权重，代表性较低的样点得到较大的权重，每个样点均有1个权重值，且权重和为1。

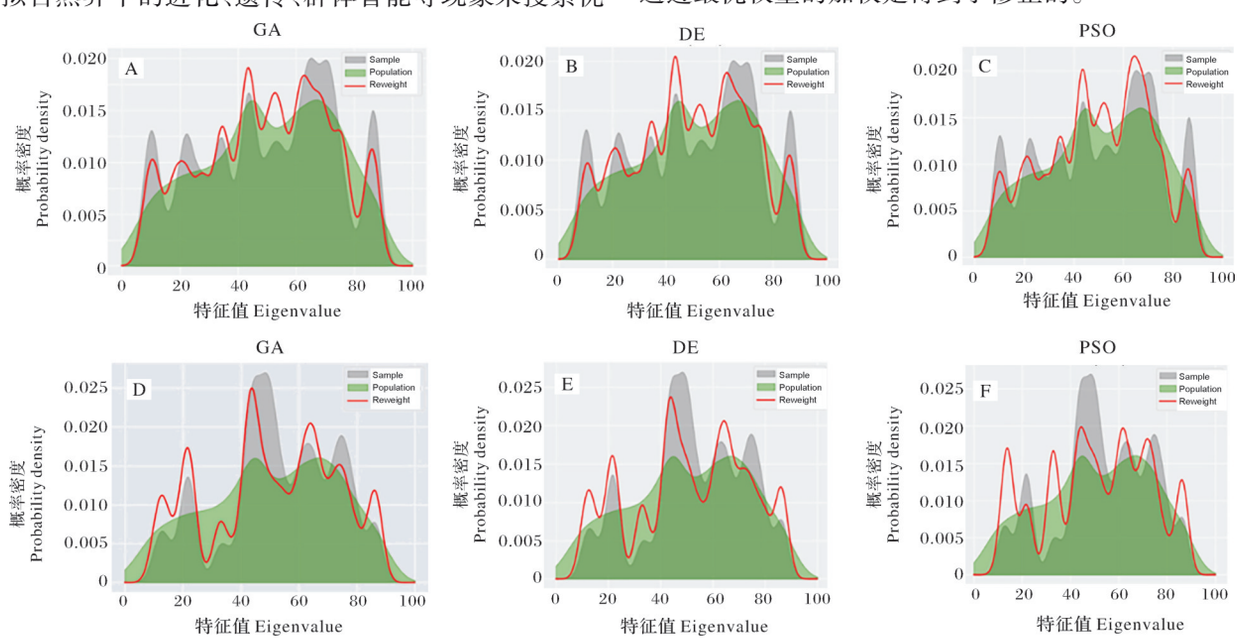
在样本代表性修正时，将权重值范围从[0,1]更改为[1,10]，主要原因有以下三点：①更换后的数值范围可以保证每个样本都对训练预测模型有贡献，即保证每个样本的权重最少为1，避免排除掉权值为0的样本，这可以更加充分的利用数据。②更换后的数值范围可以更方便地计算2个样本之间的相对重要性之比，若是采用[0,1]的权值范围，则样本的相对重要性之比可以无限大。③更换后的权值范围可以使探寻最优权重的启发式算法在计算时更加灵活。但是，在利用加权样本预测模型时可以将其进行归一化处理。本研究将计算总相似度的函数体作为待优化对象，在计算中通过更新样本权重数组进行代表性的修正。

2)启发式算法优化权重。启发式算法是通过模拟自然界中的进化、遗传、群体智能等现象来搜索优化问题的最优解，其可以模拟人的逻辑思维能力和学习能力[26]。本研究使用遗传算法[27](genetic algorithm，GA)、差分进化算法[28](differential evolution algorithm，DE)、粒子群优化算法[29](particle swarm optimization，PSO)3种启发式算法作为优化算法，3种算法均通过Python中Scikit-opt(sko)包相应的算法库实现。本研究利用启发式算法将样本点的环境协变量朝着总体样本环境协变量的方向修正，每个样本点均获得1个优化权重，最终实现土壤样本的空间代表性修正。

2 结果与分析

2.1 空间代表性修正

在PyCharm中对3种启发式算法均设置每代搜索次数为100，迭代次数为200，解空间范围为[1,10]，3种算法通过设置参数得到对应的最优解时获取最优权重。最终，得到样本空间A、B两组样本代表性修正的最优加权曲线结果如图6所示。由图6以及概率密度函数的定义可知，曲线下的总面积为1，因此修正曲线在减少过度(增加缺失)与总体分布曲线的重叠面积时，会带来曲线其他位置的波动，但是在总体上重叠面积是在增加的，说明样本代表性通过最优权重的加权是得到了修正的。

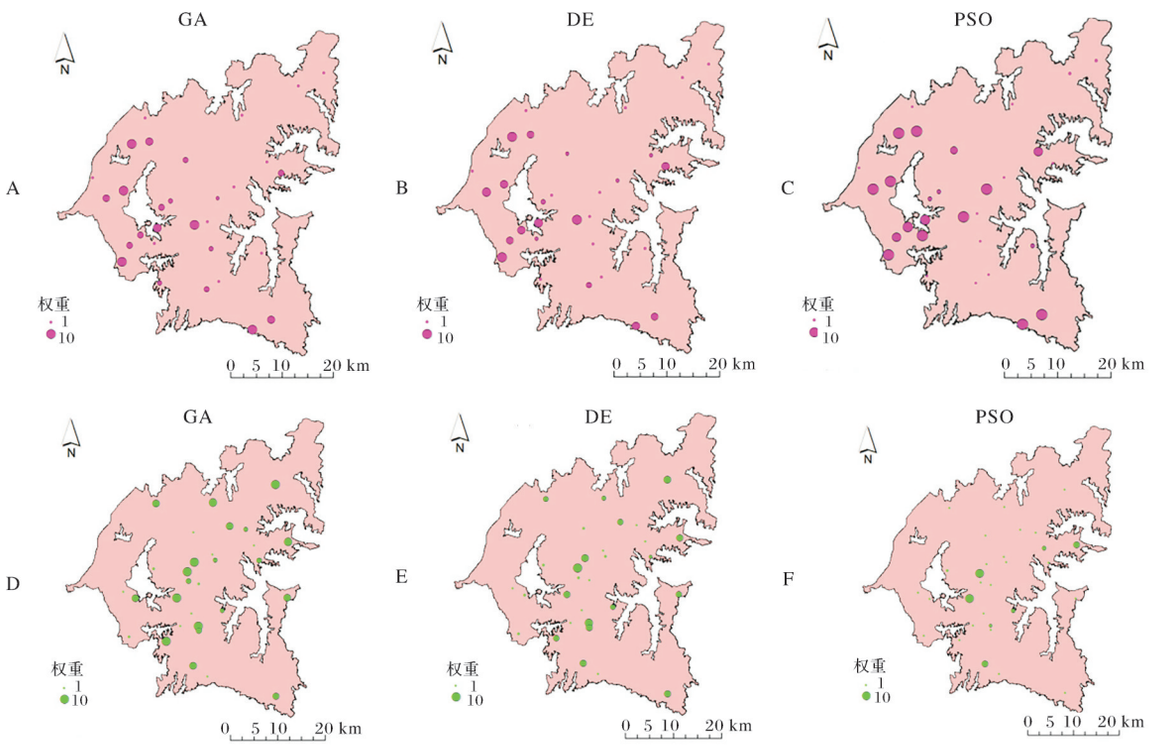


A-C：样本空间A Sample space A；D-E：样本空间B Sample space B；A，D：遗传算法 Genetic algorithm；B，E：差分进化算法 Differential evolution algorithm；C，F：粒子群优化算法 Particle swarm optimization.

图6 遗传算法、差分进化算法、粒子群优化算法样本代表性修正的最优曲线

Fig.6 Optimal curve of sample representativeness correction of genetic algorithm，differential evolution algorithm，particle swarm optimization algorithm

对于修正结果，不同算法得到的最优权重如图7所示。从整体上看，GA和DE算法对于样本空间A、B组样点修正的最优权重值接近，样点在研究区东部和中部获得的权重值较高，表明该区域采样点的代表性偏低；而在PSO算法中，样本空间A、B组均反映出较大的变异性，对于A组样本，PSO算法赋予研究区东部样点过多的权重，导致其相似度拟合较低（相似度仅为0.874 3），但是对于B组样本算法则赋予研究区中部和南部样点较大的权重，但是相似度拟合较好，故从适合程度上看，PSO算法不如GA和DE算法。通过3种算法的优化结果，可以认为A组样本数据在研究区东部和中部的样本代表性偏低；B组样本数据样本代表性偏低的点分布较为离散，但大致集中分布在107国道以及鲁湖附近位置。



A-C：样本空间A Sample space A；D-E：样本空间B Sample space B；A，D：遗传算法 Genetic algorithm；B，E：差分进化算法 Differential evolution algorithm；C，F：粒子群优化算法 Particle swarm optimization.

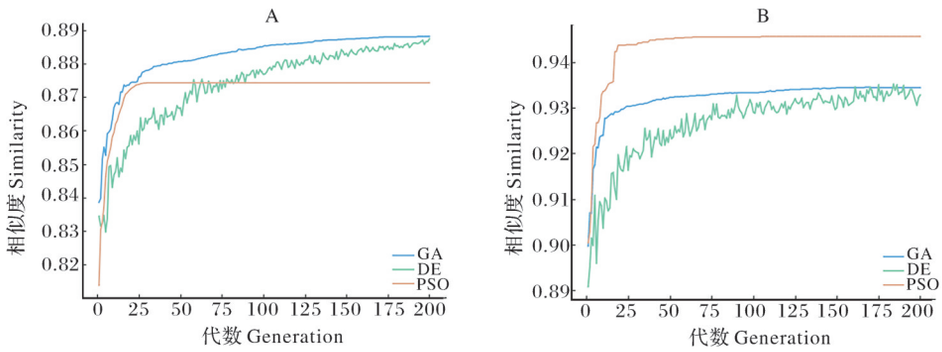
图 7　基于不同算法的最优权重点地理分布

Fig.7　Optimal weight distribution based on different algorithms

3种算法的相似度演化如图8所示，由图8可以看出GA算法的相似度提升效果最为稳定，在25代左右相似度增长趋于稳定；DE算法体现出了明显的震荡性和变异性，但是相似度优化的总体方向趋于稳定；PSO算法通常在15代左右快速收敛，继而趋于稳定。其中，GA、DE算法迭代较为稳定，且GA算法在效率上基本上完全超过DE算法，PSO算法在收敛速度上超过DE算法，但是也容易过早陷入局部最优解。这3种算法无法严格界定使用哪种算法计算最优，因为在求最优解的结果上，样本空间A组GA>DE>PSO，样本空间B组PSO>DE>GA。这可能是由于算法的算子随机性导致的，参数一致的优化程序运行出的结果也会不一致，PSO算法在快速找到解空间以后其变异方式导致其难以跳出局部最优解。因此在考虑算法的稳定性上，GA>DE>PSO，在快速收敛的性能上，PSO>GA>DE。总体而言，3种算法均可提升样本的代表性，表明运用启发式算法修正样本代表性是有效的。

2.2　多元线性回归法制图

本研究通过多元线性回归法中的最小二乘法对研究区表层土壤有机质含量进行预测制图，将协变量空间的前3个主成分作为自变量，预测位置的土壤表层有机质含量作为因变量，通过多元线性回归建立土壤有机质与环境协变量之间的关系模型，将图7所示每个样本点的权重乘以个体平方残差，通过最小化平方残差和求得最优参数[30]，并与未加入权重的制图方法比较，利用均方根误差和绝对平均误差来评价制图精度[31]。对权重修正样本和未修正样本



A:样本空间A Sample space A;B:样本空间B Sample space B.

图8 不同算法迭代200代相似度演化

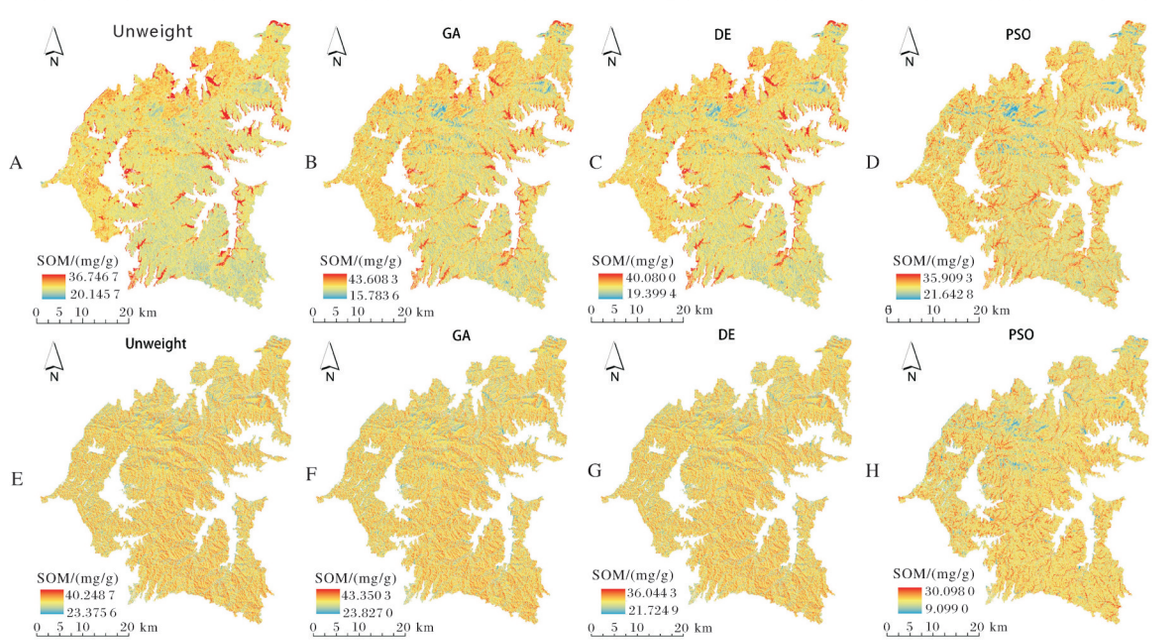
Fig.8 Iterative similarity evolution based on different algorithms for 200 generations

进行模型拟合,最终将样本空间A、B两组样本通过3种算法拟合的模型进行土壤有机质含量预测制图,制图结果如图9所示。

从图9可以看出,基于样本空间A、B两组样本3种算法的有机质含量制图,每组样本从样点变异位置上可以更加直观看出制图精度的提升效果,并且在运用3种不同算法进行相似度拟合时,GA和DE算法表现出了较高的一致性,这也与图7中样本点权重优化的表现结果一致。由图9可知,有机质含量在江夏区中心、乡镇主干道、国道和青龙山坡等位置明显减少,这也验证了人类活动对土壤表层有机质含量的消极影响[32]。制图结果的可靠性也侧面反映了本研究修正方法的有效性。

2.3 制图精度评价

运用采样得到的32个验证点对不同的制图结果进行验证,不同分组算法下最优相似度和预测精度如表2所示。由表2可知,样本空间A组RMSE和MAE最高分别降低4.27%、7.87%;样本空间B组RMSE和MAE最高分别降低10.30%、12.74%,且对于2组样本来说,GA算法对于降低制图的RMSE和MAE均是最优的。对比样本空间A、B组数据发现,样本空间B组的初始代表性优于样本空间A组,且样本空间B组的修正效果整体上优于样本空间A组,这表明初始的样本空间B组数据可以更好地拟



A-D:样本空间A Sample space A;E-H:样本空间B Sample space B;A,E:未加权Unweight;B,F:遗传算法Genetic algorithm;C,G:差分进化算法Differential evolution algorithm;D,H:粒子群优化算法 Particle swarm optimization.

图9 加权与未加权的土壤有机质含量预测制图

Fig.9 Mapping of soil organic matter content prediction based on initial weight and optimal weight

合整体区域的制图模型，即修正后就能更好地代表整体区域的有机质分布状况；而样本空间A组，因其初始代表性较低，故拟合模型所代表的整体区域制图模型比B组差，这可能是由于数据本身含有较多的干扰误差，通过此方法修正后，制图精度虽有所提升，但是仍然低于B组。从整体上看，不论是样本空间A组还是B组，使用最优权重修正样本代表性的方法制图精度均比运用未修正样本代表性的方法制图精度高，表明了此修正方法的可行性。

2.4　制图精度与样本代表性关系显著性分析

为了验证利用算法迭代为样本空间寻最优权重进而使其达到与总体空间最相似的这种修正制图方法的有效性，将不同算法迭代200代的预测制图误差和其每代对应的相似度，利用线性回归的方法拟合为一次函数，拟合后函数的相关参数如表3所示，其中，R值表示相关系数，P值表示假设检验中线性关系是否显著，可以认为，当$P<0.05$时，模型的线性关系显著。从表3可以看出，样本空间A组中各种算法的相似度和预测误差的显著性并不如B组明显，但是样本空间A、B组数据拟合的函数斜率均为负值，表明随着总体相似度的提升，预测制图的误差总体呈不断减小趋势，且对于不同组样本，相似度和预测误差斜率的绝对值均为GA>DE>PSO，正由于此相关性具有显著关系，即可以解释表2中样本空间B组最优加权相似度结果为PSO>DE>GA，但精度提升结果是GA>DE>PSO的现象。

表 2　基于初始权重和最优权重的制图精度对比

Table 2　Comparison of mapping accuracy based on initial weights and optimal weights

组 Group	加权方式 Weighting mode	相似度 SIM	相似度提升/% Similarity enhancement	均方根误差 RMSE	精度提升/% Accuracy enhancement	平均绝对误差 MAE	精度提升/% Accuracy enhancement
A	未加权 Unweight	0.813 8	—	8.050 8	—	6.227 9	—
	GA	0.888 1	9.13	7.706 9	4.27	5.737 6	7.87
	DE	0.887 4	9.04	7.788 0	3.26	5.930 7	4.77
	PSO	0.874 3	7.43	7.913 7	1.70	6.102 5	2.01
B	未加权 Unweight	0.866 0	—	9.013 9	—	6.541 6	—
	GA	0.934 5	7.91	8.086 0	10.30	5.708 5	12.74
	DE	0.935 2	7.99	8.249 4	8.48	5.772 8	11.75
	PSO	0.945 6	9.19	8.452 4	6.23	5.854 5	10.50

注 Note：A：样本空间 A Sample space A；B：样本空间 B Sample space B. 下同 The same as below.

表 3　基于不同算法预测制图显著性关系

Table 3　Significance relationship of predictive mapping based on weighting of different algorithms

项目 Item	组 Group	加权方式 Weighting mode	斜率 Slope	截距 Intercept	R	P	标准差 Standard deviation
RMSE	A	GA	−1.722 4	9.301 9	−0.339 7	8.62×10^{-7}	0.338 9
		DE	−0.772 9	8.982 8	−0.073 8	0.299 1	0.742 4
		PSO	−0.295 5	8.164 1	−0.049 7	0.484 2	0.421 5
	B	GA	−12.745 2	20.063 7	−0.905 3	1.55×10^{-75}	0.425 1
		DE	−9.504 3	17.516 4	−0.810 7	6.53×10^{-48}	0.487 8
		PSO	−2.969 3	11.325 5	−0.349 7	3.87×10^{-7}	0.565 3
MAE	A	GA	−5.463 3	10.715 2	−0.614 0	4.09×10^{-22}	0.499 0
		DE	−0.981 0	6.676 9	−0.073 4	0.301 8	0.947 7
		PSO	−2.354 6	8.157 7	−0.354 6	2.57×10^{-7}	0.441 1
	B	GA	−11.154 6	16.179 6	−0.918 8	7.36×10^{-82}	0.340 6
		DE	−8.016 0	13.481 3	−0.758 9	9.84×10^{-39}	0.488 8
		PSO	−3.101 3	8.838 5	−0.385 5	1.73×10^{-8}	0.527 5

对于不同算法，样本空间A、B组中GA算法相关系数R的绝对值均为同组最大，P值也均小于0.05，且结合表2可知，不论初始样本相似性的好坏，GA算法修正制图的精度提升率均为最高。因此，可以认为GA算法更适合对土壤表层有机质含量预测制图的样本代表性进行修正，且具有可行性。

3 讨 论

本研究以充分利用已采集样点数据为导向，围绕现有样本空间偏差修正问题，基于核密度估计，利用3种启发式算法迭代优化样本空间的概率密度分布，计算样本点数据的最优权重，最终通过土壤表层有机质含量制图来验证此方法的有效性及可行性。制图结果显示，3种不同的启发式算法对土壤表层有机质含量制图的精度均有提升（最高可将RMSE和MAE分别降低10.30%、12.74%），GA在修正样本代表性的性能和稳定性上更优，且此方法对于初始样本代表性高的样本制图精度提升更多，产生这种问题的原因可能是在低相似度下样点复杂程度导致了加权模式的随机性，因此，不能很好地在整体分布中拟合样本代表性。

主成分分析可以将复杂的数据降维，并且得到的新变量之间是相互正交的。因此，本研究使用主成分分析的原因主要有两方面，一方面在进行样本优化权重时，平均1种算法优化的时间大概在6～10 h，对数据降维能够进一步缩短优化时间（如果使用并行计算改进代码可能会降低运行时间）；另一方面，主成分分析可以降低变量之间的相关性，如果变量之间存在相关性，它们可能会具有相似的特征从而使算法陷入局部最优解。

基于重要性加权的偏差修正方法和本研究的方法具有一定的相似性，其最优加权函数为测试数据特征与训练数据特征的概率密度函数之比，Mathelin等[33]使用机器学习算法对这种加权方式进行了改进使其能够更快速、准确的修正不同的数据集，然而，估计最优函数需要大量的样本数据，对于数字土壤制图而言，采样点是宝贵且少量的，因此本研究使用启发式算法根据样点和整体区域的环境相似度关系来修正样本代表性，不仅使其适用于少量样本，而且融入了整体区域的地理环境信息。本研究仅使用土壤有机质制图验证此方法的可行性及有效性，对于其他土壤属性制图是否合适还需进一步探索，在实际研究中，应根据计算的速度和精度需求选择不同的算法进行尝试。从研究过程来看，应尽量选择初始代表性较好的土壤样本进行修正以达到更高的制图精度，因此，本研究的方法能够更好的为精确获取土壤表层有机质的空间分布提供技术支持，让技术服务于国家的农业事业。

参考文献 References

[1] 张淑杰，朱阿兴，刘京，等. 基于样点的数字土壤属性制图方法及样点设计综述[J]. 土壤，2012，44(6)：917-923.ZHANG S J，ZHU A X，LIU J，et al.Sample-based digital soil mapping methods and related sampling schemes[J].Soils，2012，44(6)：917-923(in Chinese with English abstract).

[2] AN Y M，YANG L，ZHU A X，et al.Identification of representative samples from existing samples for digital soil mapping [J].Geoderma，2018，311：109-119.

[3] 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 designing of soil sampling for digital soil mapping[J].Acta pedologica sinica，2020，57(2)：259-272(in Chinese with English abstract).

[4] STERBA S K. Alternative model-based and design-based frameworks for inference from samples to populations：from polarization to integration[J].Multivariate behavioral research，2009，44(6)：711-740.

[5] 黄亚捷，李菊梅，马义兵. 土壤重金属调查采样数目的确定方法研究进展[J]. 农业工程学报，2019，35(24)：235-245. HUANG Y J，LI J M，MA Y B.Research progress of methods for determining sampling numbers of soil heavy metals survey [J].Transactions of the CSAE，2019，35(24)：235-245(in Chinese with English abstract).

[6] QIN C Z，AN Y M，LIANG P，et al.Soil property mapping by combining spatial distance information into the soil land inference model (SoLIM)[J].Pedosphere，2021，31(4)：638-644.

[7] 张河川. 基于样点代表性等级与道路网信息的采样设计研究[D]. 武汉：华中农业大学，2018.ZHANG H C.Research on sampling design based on representative grade of sample points and road network information [D]. Wuhan：Huazhong Agricultural University，2018 (in Chinese with English abstract).

[8] 黄魏，许伟，汪善勤，等. 基于不确定性模型的土壤：环境关系知识获取方法的研究[J]. 土壤学报，2018，55(1)：54-63. HUANG W，XU W，WANG S Q，et al.Extraction of knowledge about soil-environment relationship based on an uncertainty model[J].Acta pedologica sinica，2018，55(1)：54-63(in Chinese with English abstract).

[9] 巫振富，赵彦锋，程道全，等. 样点数量与空间分布对县域尺度土壤属性空间预测效果的影响[J]. 土壤学报，2019，56(6)：1321-1335.WU Z F，ZHAO Y F，CHENG D Q，et al.In-

fluences of sample size and spatial distribution on accuracy of predictive soil mapping on a county scale[J]. Acta pedologica sinica, 2019, 56(6): 1321-1335(in Chinese with English abstract).

[10] BETHLEHEM J. Using response probabilities for assessing representativity [M]. Netherlands: Statistics Netherlands, 2012.

[11] PASSOW C, DONNER R V. Regression-based distribution mapping for bias correction of climate model outputs using linear quantile regression[J]. Stochastic environmental research and risk assessment, 2020, 34(1): 87-102.

[12] ZHANG G M, ZHU A X. Sample size and spatial configuration of volunteered geographic information affect effectiveness of spatial bias mitigation[J]. Transactions in GIS, 2020, 24(5): 1315-1340.

[13] GOODCHILD M F. Citizens as sensors: the world of volunteered geography[J]. GeoJournal, 2007, 69(4): 211-221.

[14] FINK D, DAMOULAS T, DAVE J. Adaptive spatio-temporal exploratory models: hemisphere-wide species distributions from massively crowdsourced eBird data[J]. Proceedings of the AAAI conference on artificial intelligence, 2013, 27(1): 1284-1290.

[15] VARELA S, ANDERSON R P, GARCÍA-VALDÉS R, et al. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models[J]. Ecography, 2014, 37(11): 1084-1091.

[16] ZHU A X, ZHANG G M, WANG W, et al. A citizen data-based approach to predictive mapping of spatial variation of natural phenomena[J]. International journal of geographical information science, 2015, 29(10): 1864-1886.

[17] ZHANG G M, ZHU A X, HUANG Q Y. A GPU-accelerated adaptive kernel density estimation approach for efficient point pattern analysis on spatial big data[J]. International journal of geographical information science, 2017, 31(10): 2068-2097.

[18] 韩宗伟, 黄魏, 张春弟, 等. 基于土壤养分-景观关系的土壤采样布局合理性研究[J]. 华中农业大学学报, 2014, 33(1): 56-61. HAN Z W, HUANG W, ZHANG C D, et al. Rationality of sampling strategies based on soil-landscape relationships[J]. Journal of Huazhong Agricultural University, 2014, 33(1): 56-61(in Chinese with English abstract).

[19] 周紫燕, 黄魏, 许伟, 等. 基于随机森林算法的原始土壤图更新研究[J]. 华中农业大学学报, 2019, 38(3): 53-59. ZHOU Z Y, HUANG W, XU W, et al. Updating traditional soil maps based on random forest algorithm[J]. Journal of Huazhong Agricultural University, 2019, 38(3): 53-59(in Chinese with English abstract).

[20] XIONG X, GRUNWALD S, MYERS D B, et al. Holistic environmental soil-landscape modeling of soil organic carbon[J]. Environmental modelling & software, 2014, 57: 202-215.

[21] 王小凯, 朱小文. 计量检定中3种判别和剔除异常值的统计方法[J]. 中国测试, 2018, 44(S1): 41-44. WANG X K, ZHU X W. Three statistical methods for distinguishing andeliminating outliers in metrological verification[J]. China measurement & test, 2018, 44(S1): 41-44(in Chinese with English abstract).

[22] SILVERMAN B W. Density estimation for statistics and data analysis[M]. London: Chapman and Hall, 1986.

[23] 刘晓金, 陈文武, 王庆锋. 基于优化核函数带宽SVDD的机械振动预警模型[J]. 机电工程, 2023, 40(11): 1641-1654. LIU X J, CHEN W W, WANG Q F. Mechanical vibration warning model based on optimized kernel bandwidth SVDD[J]. Journal of mechanical & electrical engineering, 2023, 40(11): 1641-1654(in Chinese with English abstract).

[24] ZHANG G M, ZHU A X. The representativeness and spatial bias of volunteered geographic information: a review[J]. Annals of GIS, 2018, 24(3): 151-162.

[25] ZHU A X. A personal construct-based knowledge acquisition process for natural resource mapping[J]. International journal of geographical information science, 1999, 13(2): 119-141.

[26] 盛亮, 包磊, 吴鹏飞. 启发式方法在机器人路径规划优化中的应用综述[J]. 电光与控制, 2018, 25(9): 58-64. SHENG L, BAO L, WU P F. Application of heuristic approaches in the robot path planning and optimization: a review[J]. Electronics optics & control, 2018, 25(9): 58-64(in Chinese with English abstract).

[27] 李腾辉, 周德强, 何冯光, 等. 基于遗传算法优化模糊PID的甘蔗收获机切割器控制系统[J]. 华中农业大学学报, 2023, 42(2): 243-250. LI T H, ZHOU D Q, HE F G, et al. Control system of sugarcane harvester cutter based on fuzzy PID optimized by genetic algorithm[J]. Journal of Huazhong Agricultural University, 2023, 42(2): 243-250(in Chinese with English abstract).

[28] STORN R, PRICE K. Differential evolution: a simple and efficient heuristic for global optimization over continuous spaces[J]. Journal of global optimization, 1997, 11(4): 341-359.

[29] 刘振超, 苑迎春, 王克俭, 等. 融合特征权重与改进粒子群优化的特征选择算法[J]. 计算机工程与科学, 2024, 46(2): 282-291. LIU Z C, YUAN Y C, WANG K J, et al. Feature selection algorithm based on feature weights and improved particle swarm optimization[J]. Computer engineering & science, 2024, 46(2): 282-291(in Chinese with English abstract).

[30] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: machine learning in python[J]. Journal of machine learning research, 2011, 12: 2825-2830.

[31] BOUASRIA A, IBNONAMR K, RAHIMI A, et al. Evaluation of Landsat 8 image pansharpening in estimating soil organic matter using multiple linear regression and artificial neural networks[J]. Geo-spatial information science, 2022, 25(3): 353-364.

[32] 文鑫, 王艺惠, 钟聪, 等. 贵州表层土壤有机质空间变异特征

及其影响因素分析[J].水土保持学报,2023,37(3):218-224.
WEN X, WANG Y H, ZHONG C, et al. Spatial variation of surface soil organic matter and its influencing factors in Guizhou Province[J]. Journal of soil and water conservation, 2023, 37(3):218-224(in Chinese with English abstract).

[33] DE MATHELIN A, DEHEEGER F, MOUGEOT M, et al. Fast and accurate importance weighting for correcting sample bias[C]//AMINI M R, CANU S, FISCHER A, et al. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer, 2023: 659-674.

Representative revision of soil samples based on estimation of kernel density

LI Kun, CHEN Yuhao, LI Wenyue, WANG Ziying, FU Peihong, HUANG Wei

College of Resources and Environment, Huazhong Agricultural University, Wuhan 430070, China

Abstract How to obtain more reliable soil-environment knowledge from existing historical samples has become an important scientific issue in digital soil mapping. This article used the method of revising the representativeness of samples to obtain higher accuracy of knowledge. Three different algorithms and the spatial similarity relationship between the covariates of the sample space and the overall spatial environment were used to identify the optimal weights for each sampling point of soil based on the estimation of kernel density. The prediction mapping of the content of organic matter on the surface of soil was used as an example to verify the scientific and validity of the method. The results showed that the revised method reduced RMSE and MAE of multiple linear regression mapping by 10.30% and 12.74%, confirming the feasibility and validity of this method. It will provide technical support for processing the data from sampling points of soil to make full use of historical data and improve the accuracy of mapping soil.

Keywords environmental covariates; spatial deviation revision; representativeness of samples; heuristic algorithm; digital soil mapping; historical samples

(责任编辑:陆文昌)