

农业用地土壤重金属样本点数据精化方法——以北京市顺义区为例
唐炬彪, 朱庆伟, 董士伟, 高秉博, 潘瑜春, 王怡蓉, 郜允兵

引用本文:

唐炬彪, 朱庆伟, 董士伟, 等. 农业用地土壤重金属样本点数据精化方法——以北京市顺义区为例[J]. 农业环境科学学报, 2020, 39(10): 2288–2296.

在线阅读 View online: <https://doi.org/10.11654/jaes.2020-0323>

您可能感兴趣的其他文章

Articles you may be interested in

北京顺义区土壤重金属污染生态风险评估研究

韩平, 王纪华, 冯晓元, 马智宏, 陆安祥, 魏荔, 闫连波

农业环境科学学报. 2015, 34(1): 103–109 <https://doi.org/10.11654/jaes.2015.01.015>

农田土壤重金属空间变异多尺度分析——以北京顺义土壤Cd为例

刘伟, 郜允兵, 周艳兵, 潘瑜春, 戴华阳, 高秉博, 阎跃观

农业环境科学学报. 2019, 38(1): 87–94 <https://doi.org/10.11654/jaes.2018-0251>

土壤污染调查加密布点区域优化及效率研究

郝易成, 师华定, 白中科, 杨泽栋, 秦治恒

农业环境科学学报. 2019, 38(5): 1036–1042 <https://doi.org/10.11654/jaes.2018-0830>

黄土丘陵沟壑区耕层土壤重金属空间分异及影响因素

张敏, 陈海, 史琴琴, 张行, 刘迪, 赵岩

农业环境科学学报. 2019, 38(11): 2465–2475 <https://doi.org/10.11654/jaes.2019-0675>

基于空间差异的有机茶园环境适宜性评价的土壤采样方法

杜霞飞, 宗良纲, 张琪惠, 戴荣波, 潘含岳, 原强, 席运官, 王磊

农业环境科学学报. 2017, 36(1): 150–157 <https://doi.org/10.11654/jaes.2016-1001>



关注微信公众号, 获得更多资讯信息

唐炬彪, 朱庆伟, 董士伟, 等. 农业用地土壤重金属样本点数据精化方法——以北京市顺义区为例[J]. 农业环境科学学报, 2020, 39(10): 2288–2296.

TANG Gui-biao, ZHU Qing-wei, DONG Shi-wei, et al. Data refinement method for sampling sites of agricultural soil heavy metals: A case study in Shunyi district, Beijing, China[J]. *Journal of Agro-Environment Science*, 2020, 39(10): 2288–2296.



开放科学 OSID

农业用地土壤重金属样本点数据精化方法 ——以北京市顺义区为例

唐炬彪^{1,2}, 朱庆伟¹, 董士伟^{2,3*}, 高秉博⁴, 潘瑜春^{2,3}, 王怡蓉^{1,2}, 郜允兵^{2,3}

(1. 西安科技大学测绘科学与技术学院, 西安 710054; 2. 北京农业信息技术研究中心, 北京 100097; 3. 国家农业信息化工程技术研究中心, 北京 100097; 4. 中国农业大学土地科学与技术学院, 北京 100193)

摘要: 样本点空间分布是样点数据检测评价和挖掘分析的关键因素。以北京市顺义区为例, 研发了一种农业用地土壤重金属样本点数据精化方法: 首先构建样本点均匀变异指数和均匀因子离散图来共同检测样本点数据均匀性, 进一步将样本点类型分为均匀样本点、聚集样本点和稀疏样本点并确定其数量; 其次删除聚集样本点, 基于研究区历史数据加密稀疏样本点; 最后基于地理空间样本点均匀变异指数、特征空间偏离指数和插值误差共同评价数据精化效果。结果表明, 研究区样本点的均匀变异指数为 0.429, 存在一个聚集样本点和一个稀疏样本点, 空间偏离指数为 0.327, 空间属性插值误差为 6.538; 冗余数据精化后进行均匀性检测没有发现聚集样本点和稀疏样本点, 均匀变异指数下降到 0.406, 特征空间偏离指数微弱下降, 空间属性插值误差下降到 6.357。研究表明该方法可以对提高采样数据的均匀性和代表性提供理论指导, 可以服务于土壤污染防治行动计划(土十条)、土壤污染状况详查等, 为更加精确研究土壤空间信息变化提供一定的基础条件。

关键词: 样本点; 数据精化; 均匀变异指数; 空间分布; 偏离指数

中图分类号: X53 文献标志码: A 文章编号: 1672-2043(2020)10-2288-09 doi:10.11654/jaes.2020-0323

Data refinement method for sampling sites of agricultural soil heavy metals: A case study in Shunyi district, Beijing, China

TANG Gui-biao^{1,2}, ZHU Qing-wei¹, DONG Shi-wei^{2,3*}, GAO Bing-bo⁴, PAN Yu-chun^{2,3}, WANG Yi-rong^{1,2}, GAO Yun-bing^{2,3}

(1. College of Geomatics, Xi'an University of Science and Technology, Xi'an 710054, China; 2. Beijing Research Center for Information Technology in Agriculture, Beijing 100097, China; 3. National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China; 4. College of Land Science and Technology, China Agricultural University, Beijing 100193, China)

Abstract: The spatial distribution of sampling sites is the key factor for the detection, evaluation, and mining analysis of sampling data. By selecting a case study in Shunyi district, Beijing, a data refinement method for the sampling sites of agricultural soil heavy metals was developed in this study. First, the even variation indices and even factor discrete graphs of sampling sites were constructed to detect their data uniformity. The sampling sites were then divided into even, aggregate, and sparse sampling sites, and their corresponding amounts were determined. Second, aggregate sampling sites were deleted and sparse sampling sites were densified based on the historical sites of the study area. Third, the data refinement effect was evaluated based on the even variation index in the geographical space, deviation index in the fea-

收稿日期: 2020-03-21 录用日期: 2020-05-18

作者简介: 唐炬彪(1991—), 男, 贵州省黔西南人, 硕士研究生, 主要从事数据挖掘研究。E-mail: 908360494@qq.com

*通信作者: 董士伟 E-mail: dongsw@nercita.org.cn

基金项目: 国家重点研发计划课题(2016YFD0800904); 北京市自然科学基金项目(8192015); 北京市农林科学院青年科研基金项目(QNJJ201830); 国家自然科学基金项目(41801276)

Project supported: National Key R&D Program of China(2016YFD0800904); Beijing Natural Science Foundation(8192015); Youth Fund of the Beijing Academy of Agricultural and Forestry Sciences(QNJJ201830); National Natural Science Foundation of China(41801276)

ture space, and spatial interpolation error. The results showed that the even variation index of sampling sites in the study area was 0.429, with aggregate and sparse sampling sites; the deviation index in the feature space and spatial interpolation error were 0.327 and 6.538, respectively. After redundant data refinement, aggregate sampling sites and sparse sampling sites were not found by uniformity detection. The even variation index and interpolation error were reduced to 0.406 and 6.357, respectively; the deviation index slightly decreased. This study suggests that the method can provide theoretical guidance for improving the uniformity and representativeness of sampling sites. It can support soil pollution prevention and control action plans, soil pollution situation detailed investigation, etc., providing some basic conditions for further accurate study of soil spatial information changes.

Keywords: sampling sites; data refinement; even variation index; spatial distribution; deviation index

中国工农业的快速发展导致土壤重金属在污染水平、污染范围和持续时间方面都呈现出日益严重的趋势。为了应对这种趋势,国家发布一系列措施,比如,《全国农业可持续发展(2015—2030)》要求围绕农业污染防治和农村环境治理,开展水土资源保护重大工程建设;《土壤污染防治行动计划》(土十条)提出开展土壤污染调查,实施农用地分类管理,加强农业污染源监管。面对农业土壤重金属污染监测调查、分级分类防治、污染监管的新要求,研究农业用地土壤重金属样本点的空间分布特征和相应数据去冗精化方法非常关键。

土壤采样是调查土壤属性空间变异性及其统计参数的重要方式^[1-3],精确研究土壤空间信息分布和变异性必须以土壤采样点在空间均匀性及代表性为依托。因此,用样本点数据分析前必须先对样本集中样本点进行均匀性检测和冗余数据处理,空间样本点数据的均匀分布及代表性不仅是检测评价的关键因素,也是判断分析结果是否准确的重要依据。针对空间样本点数据的均匀性及代表性问题,国内外学者开展了大量的研究,许多学者采用随机选择样本点研究土壤重金属污染及评价^[4]和预测土壤重金属空间信息变化^[5-8],在一定程度上提高了土壤空间信息分布的预测结果精度,但未充分考虑样本点在地理空间均匀性和特征空间代表性,忽略了土壤空间信息相关性。基于格网采样方式进行土壤重金属污染评价^[9-11],考虑了样本点在地理空间的均匀性,但忽视其在特征空间的代表性。同时,有研究基于历史采样点预测土壤有机质含量^[12-13],基于插值方法验证样本点在特征空间代表性,但没有耦合样本点地理空间均匀性。除此之外,还有其他不少学者也从不同的角度分析,分别对不同的采样数量和布局研究样本点在土壤空间的均匀性及代表性问题^[14-20]。总体分析对于空间样本点数据在地理空间分布不均和特征空间的代表性问题,更多是将地理空间样本点的均匀性和特征

空间的代表性剥离计算,未耦合两者对空间数据精化提供具体方法及理论支撑。

综合分析国内外相关研究后发现主要存在两大问题:一是没有进行样本点的均匀性检测及非均匀样本点的去冗精化。这种数据处理模式可能会出现样本点代表性差的情景,增加采样点数据分析过程中的不确定性,进而影响采样点数据分析结果的精度和准确度;二是构建的样本点分布均匀性检测指标体系比较单一、缺乏系统性,致使检测结果容易出现偏差,也无法体现采样点数据局部均匀细节,不利于采样点数据的去冗精化和挖掘分析。具体在重金属样本点方面,以随机布点法、格网布点法和分区布点法为主的重金属样本点布设结果也存在上述两大问题,尤其缺乏重金属样本点分布非均匀时的数据精化处理方法。基于此,本文提供一种农业用地土壤重金属样本点数据精化方法,构建区域农业重金属样本点分布均匀性检测指标体系进行均匀性检测,并综合集成样本点分布非均匀化去冗精化方法。该方法可以为区域样本点数据质量评价和数据去冗精化提供一种新的技术方法,可直接服务于土壤污染防治行动计划(土十条)、土壤污染状况详查等,对提升农业环境监测体系和监管信息化水平具有指导作用。

1 材料与方法

1.1 研究区概况及数据来源

顺义区位于北京市城区东北方向,城区距市区中心 30 km,顺义区地势北高南低,地理位置 40°00'~40°18'N, 116°28'~116°58'E, 境域东西长 45 km, 南北宽 30 km, 总面积 1 021 km²。土壤为河流洪水携带沉积物质造成,表面堆积物主要是砂、亚砂土,北部山地最高点海拔为 637 m,境内最低点海拔为 24 m。本文研究该区域内的 4 个乡镇,分别是高丽营镇、赵全营镇、牛栏山地区和北石槽镇,农业用地面积为 114.379 km²,农业用地主要是菜地、水浇地、苗圃等。

土壤重金属样本点数据来源于北京市农林科学院农产品质量安全管理平台。为了分析北京市农田环境和农产品质量情况,北京市农林科学院定期组织下属单位开展北京市土壤样品采集。本研究数据采集于2009年春季,样本点的布局 and 数量根据田块的利用方式和面积进行确定,采用GPS定位记录样点中心位置,采样点主要分布于粮田、菜地、果园等农业用地。原始采样中,每个样本点在10 m×10 m格网范围内选择5个0~20 cm耕层土壤混合,按四分法选取分析样品1.0 kg。所有土样在室内自然风干,碾压磨碎后,过100目尼龙网筛,按照国家标准分析测定各重金属元素(Cu、Zn、Pb、Cd、As、Hg)。土壤样品测定采用20%样品平行样,并加入国家标准土壤样品(GSS-1和GSS-4)作为质量控制样品,质控样品相对误差小于10%。本研究区域内农业用地共分布95个样本点(图1),选择以重金属Cu为例进行数据精化研究。

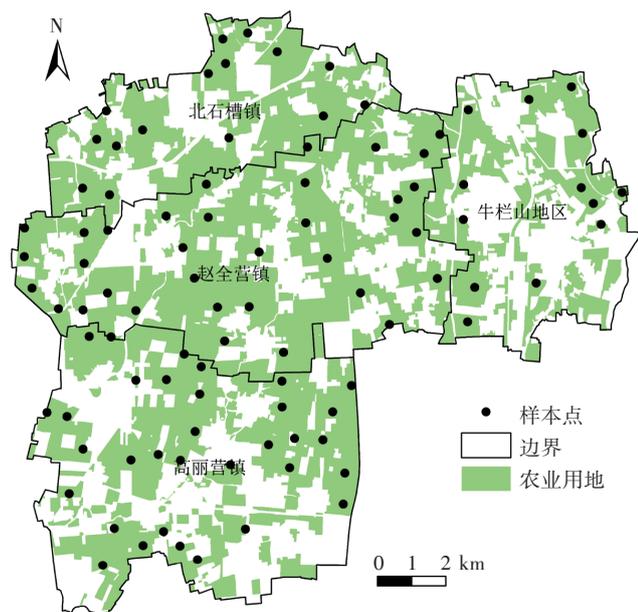


图1 研究区域
Figure 1 Study area

1.2 研究方法

空间样本点数据精化主要通过对其样本点的均匀性检测,找出其中存在的冗余数据,将其划分为聚集样本点和稀疏样本点。针对不同类型的样本点数据进行处理,从而改善其在地理空间分布均匀性以及特征空间的代表性,最终实现空间数据精化目的。

农业用地土壤重金属样本点数据精化方法主要包括3个部分:一是样本点数据均匀性表征方法;二是样本点数据去冗精化方法;三是样本点数据去冗精

化效果评价方法。

1.2.1 样本点数据均匀性表征方法

根据样本点在空间分布方式的不同,样本点的均匀性表征方法一般分为:地理空间样本点均匀性表征方法,特征空间样本点均匀性表征方法,耦合地理空间和特征空间样本点均匀性表征方法。地理空间数据的均匀性一般主要考虑地理空间采样位置均匀而忽视其在特征空间中属性分布的代表性,比较常用的方法为基于格网采样法^[21]、电荷排斥模拟法^[22]、空间模拟退火算法^[23]等;特征空间样本点均匀性表征方法一般主要考虑特征空间属性分布均匀而忽视样本点在地理空间位置的均匀性,比较常用的是目的性选择的方法^[4]。针对空间样本点的均匀性和代表性问题,全局代表性高的样本点不但在数值空间内很好地囊括了目标区域土壤特性的典型值,而且在地理空间和特征空间也可以极大限度地反映土壤属性的空间变异^[24]。因此耦合地理空间和特征空间样本点均匀性表征方法是理想选择,尽管许多学者耦合两个空间的采样布设尝试,比如超拉丁立方体采样方法^[25],但该方法很难操作。考虑样本点在地理空间的均匀性,又结合样本点在特征空间分布的代表性,现阶段很难找到统一的指标来衡量地理空间的均匀性和特征空间的代表性问题,而且还考虑各方面环境因素,又因为样本点在地理空间的均匀性和特征空间分布的代表性可能存在一定的矛盾,所以现阶段要实现该表征方法异常困难。

本文研究土壤重金属样本点数据均匀性检测和去冗精化局限在地理空间,但样本点数据去冗精化效果从地理空间和特征空间分别构建指标进行共同评价。地理空间样本点均匀性表征方法主要涉及两个重要指标:一是样本点的均匀因子,二是样本集的均匀变异指数;通过绘制均匀因子离散图,检测其存在的冗余数据。

(1) 样本点均匀因子

均匀因子表示样本点所在研究区域生成的泰森多边形面积与平均采样面积的比值,计算公式见式(1)。

$$V_i = \frac{S_i}{S_0} \quad (1)$$

式中: S_0 为平均采样面积, km^2 ; S_i 为第*i*个样本点所在泰森多边形面积, km^2 ; V_i 为第*i*个样本点的均匀因子。均匀因子表示单个样本点的均匀因子与标准值1的局部偏离程度,偏离程度越小表示样本点所在区域存

在冗余采样点数量越少。当均匀因子大于1时,表明样本点在所在区域比较稀疏,即稀疏样本点;当均匀因子等于1时,不需改善样本点在地理空间数据的均匀性,即均匀样本点;当均匀因子小于1时,表示在该区域采集的样本点处于聚集分布,即聚集样本点。

(2) 均匀变异指数

均匀变异指数表征样本集中全部样本点的整体均匀程度,可由公式(2)表达:

$$E_v = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (V_i - 1)^2}}{\sum_{i=1}^N V_i / N} \quad (2)$$

式中: E_v 为所有样本点的均匀变异指数; N 为采样区域中的样本点的个数; V_i 为第*i*个样本点的均匀因子。均匀变异指数越小,表示样本点在地理空间的分布越均匀。反之,样本点分布越趋向于聚集和稀疏。

1.2.2 样本点去冗精化方法

样本点数据去冗精化方法分为三种不同类型的场景:均匀样本点去冗精化、聚集样本点去冗精化和稀疏样本点去冗精化。

(1) 均匀样本点去冗精化

计算样本点的均匀因子,当均匀因子都等于1时,即采集的样本点在空间分布均匀,表示样本集中没有冗余数据,无需进行样本点冗余数据分析与处理。

(2) 聚集样本点去冗精化

计算样本点的均匀因子,当样本点均匀因子小于1时,表示样本点局部聚集。一般设置小于1的阈值或根据跳跃性来判断,当离散图中均匀因子小于阈值时或产生跳跃,就判断该样本点在样本集中为聚集样本点,处理方法为删除该样本点或样本点权重调整。然而,删除地理空间的聚集样本点时,还应该兼顾特征空间的特征点情景:①有先验知识或人为设置特征点情景,若聚集样本点不是特征空间的特征点,则删除该样本点;反之,则不能删除,可以采用样本点权重调整方法,减少该样本点的权重值;②没有先验知识或人为设置特征点情景,为防止误删特征点,在特征空间中引入局部 Moran's *I* 系数^[26]将土壤重金属含量的空间格局可视化,进一步研究其空间分布规律。Moran's *I* 散点图可以描述局部空间自相关性,将土壤重金属含量空间分布划分为5种类型:高值聚集(HH)、高值被低值包围(HL)、低值被高值包围(LH)、低值聚集(LL)和不显著。当聚集样本点属于高值被

低值包围(HL)或低值被高值包围(LH)时,表示该聚集样本点在特征空间的空间差异程度显著较大,可以剔除;若聚集样本点不属于HL或LH,则需要根据其他辅助数据和信息进行判断。

(3) 稀疏样本点去冗精化

计算样本点的均匀因子,当样本点的均匀因子大于1时,应该根据均匀因子离散图设置一个大于1的阈值或根据其跳跃性来判断。当超出阈值或产生跳跃时,即可判断对应的样本点在样本集中属于稀疏样本点。针对稀疏样本点数据处理有两种方法:样本点权重值调整方法和样本点添加方法。

样本点权重调整方法不需添加和删除样本点,只需对样本点的权重进行调整。样本点比较稀疏的区域,根据样本点影响的范围增加其权重值,样本点比较聚集的区域,减少其权重值。例如,可以选择样本点所在区域泰森多边形面积与该区域所有参与调整的样点泰森多边形面积总和的比值作为样本点调整权重。

样本点添加方法可分为3种模式,一是基于历史样本点添加方法,该方法对时间要求比较苛刻,目标样本点附近的历史样本点需要在一定时间间隔内采集,比如规定时间间隔为一年。否则,时间间隔太长,样本点的属性特征值会随之发生改变。二是现场补测方法,该方法适用于样本点近期采集作业,根据样本点数据处理结果或研究目的进行野外现场补测,时间间隔要求比历史样本点添加方法更短。三是基于样本点模型优化方法,耦合样本点地理空间分布和特征属性尝试建立添加样本点目标函数和优化方法,确定添加样本点的最佳位置。

上述不同方法具有各自的优缺点,适用不同的应用场景。而稀疏区域添加样本点数量确定方法,根据样本点在区域内生成的泰森多边形面积与平均采样面积比较确定添加样本点的个数,如公式(3)。

$$N_i = \frac{S_i}{S_0} - 1 \quad (3)$$

式中: S_0 为平均采样面积,km²; S_i 为第*i*个样本点所在泰森多边形面积,km²; N_i 四舍五入取整即为增加的样本点数。

1.2.3 样本点数据去冗精化效果评价

样本点数据去冗精化效果评价方法从地理空间和特征空间构建指标共同评价。以空间数据为研究对象对其均匀性检测,对存在的冗余数据处理结果进行评价,其中包含均匀变异指数、均匀变异指数变化

率、特征空间偏离指数和属性值的插值误差。以原始样本点计算结果为参考标准,验证处理后的样本点在地理空间的均匀性和特征空间的代表性。综合分析数据处理之后的均匀因子、均匀变异指数、偏离指数和空间插值误差,比较土壤样本点数据精化的结果对样本点的均匀性及代表性的改善程度进行综合评价。

(1) 地理空间评价指标: 均匀变异指数变化率

通过样本点的均匀因子计算其均匀变异指数,均匀变异指数越小,表示样本点在地理空间的分布越均匀。均匀变异指数变化率计算公式见式(4)。

$$V = \frac{(E_v - E_v')}{E_v} \times 100\% \quad (4)$$

式中: V 为均匀变异指数变化率; E_v 为原始样本集的统一变异指数; E_v' 为新样本集的统一变异指数。删除聚集样本点和加密稀疏样本点后,均匀变异指数变化率越大,表示样本点数据精化效果越好。反之,表示去冗精化效果不佳。

(2) 特征空间评价指标: 偏离指数

P-P 图 (Probability-probability plot) 和 Q-Q 图 (Quantile-quantile plot) 通过绘制样本点及相应总体的概率/分位数散点图来比较样本点及其总体的特征分布。为了量化样本的代表性,反映样本点在特征空间的偏离程度,定义特征空间偏离指数 (Deviation index, DI) 为以 P-P 图或 Q-Q 图中 $y=x$ 线为基准的标准残差^[27], 计算公式见式(5)。

$$DI = \sqrt{\frac{\sum_{i=1}^N (q_i - Q_i)^2}{N}} \quad (5)$$

式中: DI 是偏离指数; q_i 是第 i 个样本点属性值的分位数/概率; Q_i 是相应的总体分位数/概率; N 是样本点个数。偏离指数越小,表征样本点在特征空间中的分布代表性越好。

(3) 插值误差

样本点数据主要用途之一是空间制图,基于农业土壤重金属属性空间插值误差大小来定量表征去冗精化对样本点数据空间制图的影响。选择常用的均方根误差 (Root Mean Square Error, RMSE), 计算公式为:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (P_i - O_i)^2}{N}} \quad (6)$$

式中: $RMSE$ 是均方根误差; P_i 是第 i 个样本点的预测值, $\text{mg} \cdot \text{kg}^{-1}$; O_i 是第 i 个样本点的观测值, $\text{mg} \cdot \text{kg}^{-1}$; N 是样本点个数。空间插值方法主要推荐克里格插值,它是以变异函数理论和结构分析为基础,利用区域化变

量的原始数据和变异函数的结构特点,对未知样本点进行线性无偏、最优估计。首先对原始样本集属性空间插值计算插值误差;其次数据去冗精化后,计算去冗精化样本集的空间插值误差,并与原始样本集的空间插值误差比较。当计算出的样本点数据属性值的插值误差大于原始样本集属性值的插值误差,说明样本点的代表性更差;等于原始样本集属性值插值误差,则样本集数据精化对样本点的均匀性没有影响;小于原始样本集属性值的插值误差,说明样本集中的样本点在土壤特征空间信息更具代表性;插值误差越小表示样本点的代表性越好。反之,代表性越差。

2 结果与分析

2.1 样本点数据处理结果及分析

计算原始样本集样本点的均匀因子并绘制均匀因子离散图(图2),图中均匀因子是按照从小到大的顺序排列,计算样本点的均匀变异指数为0.429。基于 ArcGIS 10.1 软件平台,以样本集中重金属 Cu 为目标变量,利用地统计中探索性数据分析工具 Normal QQPlot 和普通克里格插值方法,分别计算样本点属性 Cu 的特征空间偏离指数为0.327 和插值误差为6.538。通过结果分析,当均匀因子为0.349,即最低点,与下一点均匀因子相比发生很大的跳跃,所以判断该均匀因子对应的样本点为聚集样本点。当均匀因子为2.744,即最高点,与上一点均匀因子相比也发生很大跳跃,所以判断该均匀因子对应的样本点为稀疏样本点。

2.2 样本点精化区域划定

根据数据处理结果及分析可知,样本集中存在1个聚集样本点和1个稀疏样本点。针对地理空间的聚集样本点,研究区内没有先验知识或人为设置特征

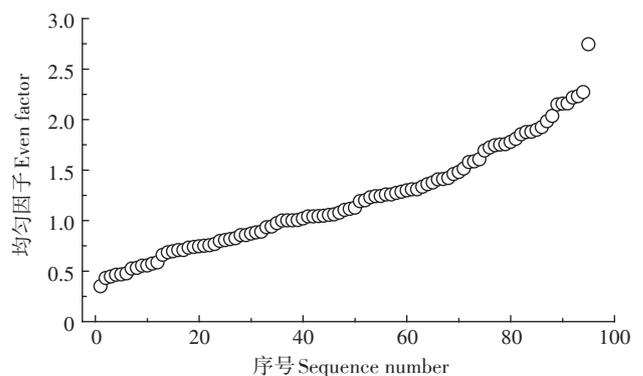


图2 均匀因子离散图

Figure 2 Discrete graph of even factor

点,为防止误删特征点,基于 ArcGIS 10.1 软件的 Spatial Autocorrelation 工具计算原始样本集的 Moran's I 系数,参数设置反距离进行空间关系的概念化,并绘制 Moran's I 散点图如图(3)所示,一个聚集样本点(图4)在特征空间属于低值被高值包围(LH),对该样本点进行删除。针对稀疏样本点,需要在稀疏样本点影响范围内加密样本点。通过计算添加样本点数量的结果可知,稀疏区域需要增加1个样本点。

本研究采用数据其样本采样时间为2009年,无法选取现场补测方法,而研究区存在2008年历史样本点数据,因此本研究采用基于历史样本点数据添加方法。在稀疏区域内共有3个2008年历史样本点(图4),分别记为1、2、3号,数据精化前已消除由于时间不一致带来的系统误差。

2.3 数据去冗精化结果及分析

通过对原始样本点的均匀性检测结果,剔除聚集样本点和添加不同历史样本点共分为3种不同方案:a方案具体内容为删除1个聚集点,添加1号样本点;b方案具体内容为删除1个聚集点,添加2号样本点;c方案具体内容为删除1个聚集点,添加3号样本点。不同处理方案的均匀因子离散图和均匀变异指数计算结果分别如图5和表1所示。结果表明,均匀变异指数下降明显,说明剔除聚集样本点和添加历史样本点能有效去除原始样本集中的冗余数据。

计算添加不同位置样本点处理结果,a方案均匀

变异指数由0.429下降到0.416,样本点均匀性整体得到改善,但该方案最小泰森多边形面积由0.349 km²下降到0.330 km²,重新生成1个聚集样本点(图5a);b方案和c方案与原始方案相比,都降低了样本点所在泰森多边形的最大面积,增大了最小面积,去冗精化后没有发现聚集样本点和稀疏样本点,改善了样本点整体均匀性。由于b方案的均匀变异指数0.406小于c方案的0.412,即不同方案样本点的均匀变异指数大小为:原始样本点>a方案>c方案>b方案。故b方案是研究区去冗精化的最优方案,即删除1个聚集样本点和添加2号历史样本点。

2.4 样本点数据精化评价

本文中聚集和稀疏样本点为数据集中均匀因子离散图两端的极值,根据地理空间均匀变异指数变化率、特征空间偏离指数和属性插值误差来共同评价样本点数据去冗精化效果。通过计算原始样本点均匀

表1 不同数据精化方案

Table 1 Different data refinement schemes

样本集 Sampling set	样本数量 Sample number	最小面积 Minimum area/ km ²	最大面积 Maximum area/ km ²	均匀变异指数 Even variation index
原始样本点	95	0.349	2.744	0.429
a方案	95	0.330	2.272	0.416
b方案	95	0.447	2.272	0.406
c方案	95	0.447	2.272	0.412

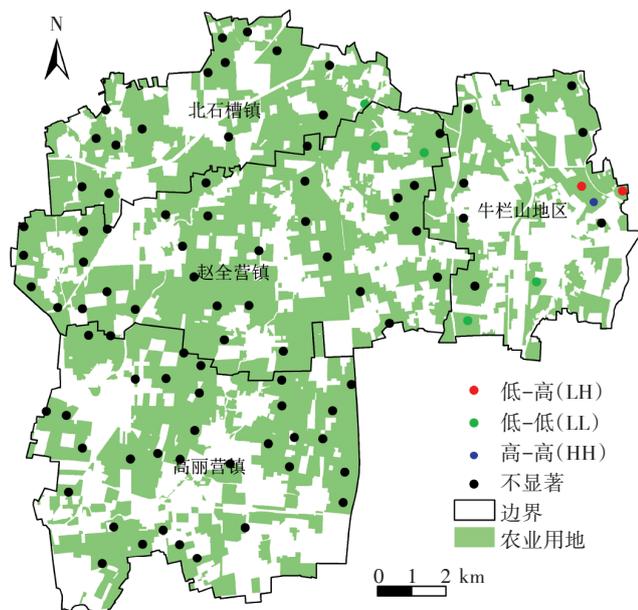


图3 样本点的Moran's I 散点图

Figure 3 Moran's I scatter diagram of sampling sites

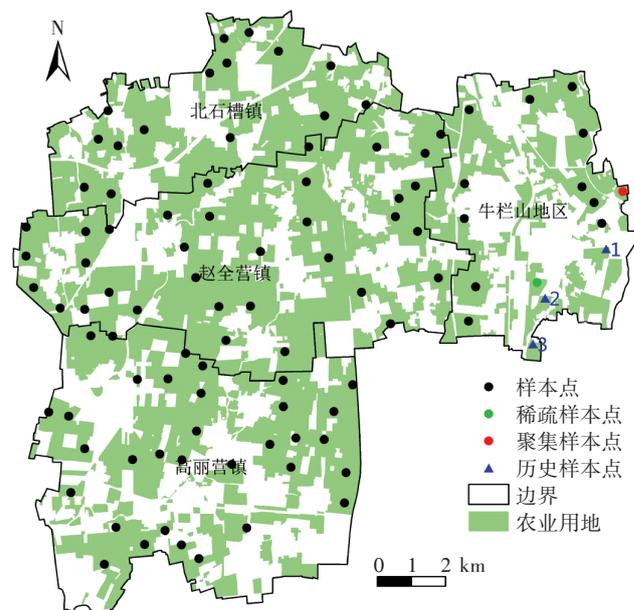


图4 不同类型样本点

Figure 4 Different types of sampling sites

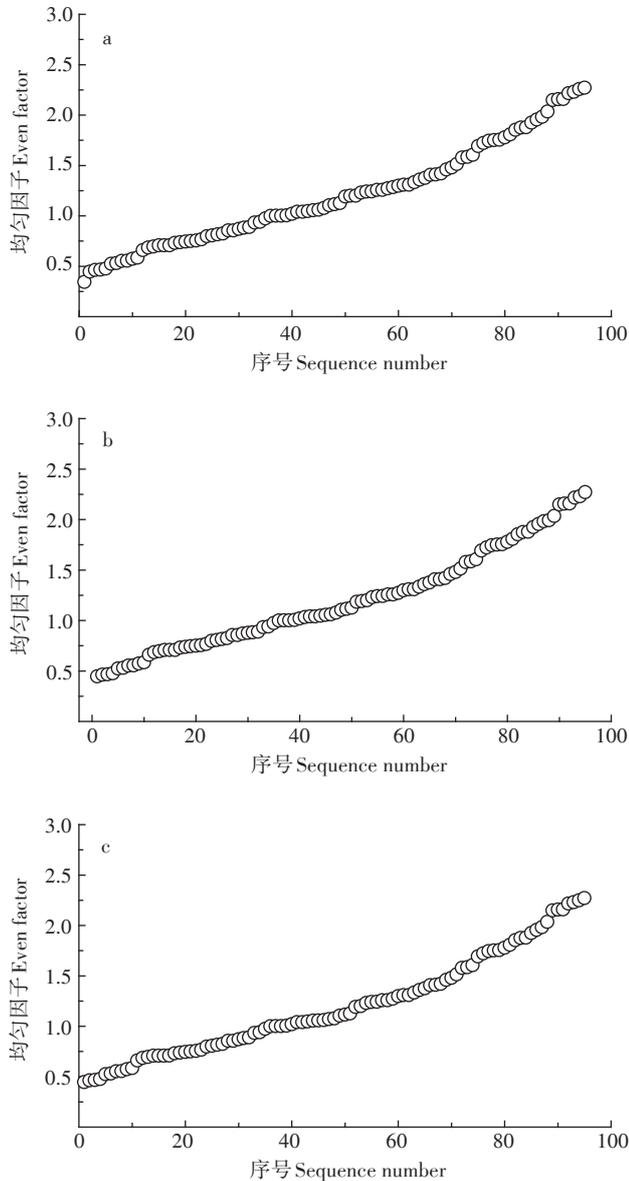


图5 不同数据精化方案的离散图

Figure 5 Discrete graphs of different data refinement schemes

变异指数为0.429。删除聚集样本点和稀疏样本点添加历史样本点数据之后均匀性检测的效果进行分析比较,结果如表2所示,a均匀变化率最小,则a方案样本点在地理空间均匀性改善效果最差。针对b方案和c方案比较分析,根据图5可知,添加2号和3号历史样本点未重新造成样本点冗余等情况。根据表2可知,不同方案样本点的均匀变异指数变化率比较:b方案>c方案>a方案。表明b方案在地理空间的均匀性最好,故添加2号历史样本点改善样本点的均匀性效果最佳。结果表明本研究数据精化方案在地理空间有效。

表2 均匀变异指数变化率

Table 2 Change rate of even variation index

样本集 Sampling set	均匀变异指数变化率 Change rate of even variation index/%
a方案	3.030
b方案	5.361
c方案	3.963

原始样本点与不同数据精化方案中,以土壤重金属Cu元素为目标变量,用Normal QQPlot分析工具分别计算该属性的特征空间偏离指数,如表3所示,不同方案的偏离指数大小为:原始样本点>c方案>a方案>b方案,表明经过冗余数据处理之后,各方案样本点的属性值特征空间偏离指数都在微弱减小,各方案都在一定程度上提高了样本点在特征空间的代表性。其中b方案偏离指数最小,表明在特征空间中的代表性最好。

用普通克里格空间插值方法分别计算Cu元素的插值误差,结果如表3所示,不同方案的插值误差大小为:原始样本点>a方案>c方案>b方案,表明经过冗余数据处理之后,各方案样本点属性值的插值误差都在明显减小,表明各方案都在一定程度上提高了样本点的代表性,其中b方案的效果最佳,与地理空间和特征空间评价结果一致。

综上所述,通过不同方案比较,b方案数据精化效果最好,提高了地理空间样本点的均匀性和特征空间的代表性。

表3 Cu特征空间偏离指数和插值误差

Table 3 The feature space deviation index and interpolation error of Cu

样本集 Sampling set	样本数量 Sample number	偏离指数 Deviation index	偏离指数差值 Deviation index error/ 10^{-4}	插值误差 Interpolation error
原始样本点	95	0.327	0	6.538
a方案	95	0.327	-0.008	6.468
b方案	95	0.327	-0.047	6.357
c方案	95	0.327	-0.003	6.418

3 讨论

3.1 样本点数据精化方法与相关研究的对比分析

对比分析本研究样本点数据精化方法与相关研究成果,可以归纳为3种类型:(1)样本点数据没有进行均匀性检测或进行了均匀性检测但没有进行相应的数据精化处理,例如,韩宗伟等^[13]不同采样尺度的

土壤有机质数据分析;(2)非均匀化样本点数据进行了简单数据精化处理,但没有系统评价数据精化效果,例如,吴丹等^[28]对农用机井的加密优化;(3)系统构建样本点分布均匀性检测指标体系进行均匀性检测,并综合集成样本点分布非均匀化去冗精化方法,例如,本研究土壤重金属样本点的数据精化。

3.2 样本点数据精化方法的不确定性分析

研发的样本点数据精化方法提高了样本点的均匀性和代表性,但研究过程中存在一定的不确定性。(1)地理空间极大值点(稀疏样本点)和极小值点(聚集样本点)不同于特征空间的特征点,二者之间没有明显的对应关系。由于缺乏数据进行真实性检验,本研究可以基于局部 Moran's I 系数进行判断,但对于其他应用场景,则需要根据其他辅助数据和信息进行判断;(2)添加历史采集的样本点方法存在局限性,在稀疏区域添加样本点的空间位置只是较优位置,并不一定是最佳位置,该目标函数和优化方法是个研究难点,正在考虑结合深度学习算法来进一步尝试;(3)本研究的数据精化处理过程中,聚集样本点数量和稀疏样本点数量恰好相等,使得不同数据精化方案和原始样本集中样本点总数相等。针对其他样本集,若出现数量不相等的情景,如何设置阈值准确判断和优化聚集样本点和稀疏样本点非常关键。此外,随着样本数、间距和分布格局不同,变异函数的理论模型可能会发生变化,影响去冗精化效果评估。

针对精化方法目前存在的不确定性,未来将进一步研究耦合地理空间和特征空间样本点去冗精化方法、不同样本点数量下的数据精化方案、样本点权重调整方法、样本点优化模型布设最佳加密点、目标变量空间非平稳情况下的去冗精化方法等。

3.3 样本点数据精化方法的应用推广

研发的农业用地土壤重金属样本点数据精化方法,以其兼顾样本点地理空间均匀性和特征空间代表性的优势,可以应用在面源与重金属污染、场地污染监测、耕地质量评价、气象和环境评估、海洋环境预警等样本点、监测点或监测站的优化布局,减少数据冗余,提高点位的代表性,具有很好的应用前景。

4 结论

(1)根据不同类型样本点去冗精化的结果评价,显示b方案(删除1个聚集点,添加2号历史样本点)的均匀变异指数变化率最大,特征空间偏离指数和

插值误差最小,提高了地理空间样本点的均匀性和特征空间的代表性,数据精化效果最佳。

(2)本研究数据精化方法在一定程度上可以兼顾样本点地理空间的均匀性及特征空间代表性,不仅为大数据去冗精化提供一种参考方法,而且可以用于样本点布设方案设计,在土壤污染防治行动计划(土十条)、土壤污染状况详查以及其他行业的点位优化布局中具有很好的应用潜力。

(3)在数据特征点判断、去冗精化方法选择和阈值设置方面存在一定的不确定性,未来将进一步研究不同适用条件下耦合地理空间和特征空间的样本点去冗精化方法。

参考文献:

- [1] McEwing K R, Fisher J P, Zona D. Environmental and vegetation controls on the spatial variability of CH₄ emission from wet-sedge and tussock tundra ecosystems in the Arctic[J]. *Plant and Soil*, 2015, 388(1/2):37-52.
- [2] Yao Z S, Wolf B, Chen W W, et al. Spatial variability of N₂O, CH₄ and CO₂ fluxes within the Xilin River catchment of Inner Mongolia, China: A soil core study[J]. *Plant and Soil*, 2010, 331(1-2):341-359.
- [3] Nelson P N, Banabas M, Goodrick I, et al. Soil sampling in oil palm plantations: A practical design that accounts for lateral variability at the tree scale[J]. *Plant and Soil*, 2015, 394(1/2):421-429.
- [4] Yang Q Q, Li Z Y, Lu X N, et al. A review of soil heavy metal pollution from industrial and agricultural regions in China: Pollution and risk assessment[J]. *Science of the Total Environment*, 2018, 642:690-700.
- [5] Wu J, Norvell W A, Hopkins D G, et al. Improved prediction and mapping of soil copper by Kriging with auxiliary data for cation-exchange capacity[J]. *Soil Science Society of America Journal*, 2003, 67(3):919-927.
- [6] 庞凤,李廷轩,王永东,等.县域农田土壤铜含量的协同克里格插值及采样数量优化[J]. *中国农业科学*, 2009, 42(8):2828-2836. PANG Su, LI Ting-xuan, WANG Yong-dong, et al. Spatial interpolation and sampling numbers of the concentration of copper in cropland soil on county scale using Cokriging[J]. *Scientia Agricultura Sinica*, 2009, 42(8):2828-2836.
- [7] 李润林,姚艳敏,唐鹏钦,等.县域耕地土壤锌含量的协同克里格插值及采样数量优化[J]. *土壤通报*, 2013, 44(4):830-838. LI Rui-lin, YAO Yan-min, TANG Peng-qin, et al. Optimization of spatial interpolation and sampling size for soil zinc content in cultivated land at county scale using Cokriging method[J]. *Chinese Journal of Soil Science*, 2013, 44(4):830-838.
- [8] 赵业婷,常庆瑞,李志鹏,等.基于Cokriging的耕层土壤全氮空间特征及采样数量优化研究[J]. *土壤学报*, 2014, 51(2):415-422. ZHAO Ye-ting, CHANG Qing-rui, LI Zhi-peng, et al. Study on spatial distribution of total nitrogen in arable soil layer and optimization of number of soil samples for the study based on Cokriging[J]. *Acta Pedologica Sinica*, 2014, 51(2):415-422.

- [9] 陈思萱, 邹滨, 汤景文. 空间插值方法对土壤重金属污染格局识别的影响[J]. 测绘科学, 2015, 40(1):63-67.
CHEN Si-xuan, ZOU Bin, TANG Jing-wen. Impact of spatial interpolation methods on identifying structure of heavy metal polluted soil[J]. *Science of Surveying and Mapping*, 2015, 40(1):63-67.
- [10] Fei X F, Christakos G, Xiao R, et al. Improved heavy metal mapping and pollution source apportionment in Shanghai City soils using auxiliary information[J]. *Science of the Total Environment*, 2019, 661:168-177.
- [11] Husein H H M, Kalkha M, Jrđi A A, et al. Urban soil pollution with heavy metals in Hama floodplain, Syria[J]. *Natural Resources*, 2019, 10(6):187-201.
- [12] 谢宝妮, 常庆瑞, 秦占飞. 县域土壤养分分离群样点检测及其合理采样数研究[J]. 干旱地区农业研究, 2012, 30(2):56-61.
XIE Bao-ni, CHANG Qing-rui, QIN Zhan-fei. Outlier identification and reasonable sampling number of soil nutrient at county level[J]. *Agricultural Research in the Arid Areas*, 2012, 30(2):56-61.
- [13] 韩宗伟, 黄魏, 罗云, 等. 基于路网的土壤采样布局优化: 模拟退火神经网络算法[J]. 应用生态学报, 2015, 26(3):891-900.
HAN Zong-wei, HUANG Wei, LUO Yun, et al. Application of simulated annealing method and neural network on optimizing soil sampling schemes based on road distribution[J]. *Chinese Journal of Applied Ecology*, 2015, 26(3):891-900.
- [14] 胡晓涛, 吴克宁, 马建辉, 等. 北京市大兴区耕地质量等级监测控制点布设[J]. 资源科学, 2012, 34(10):1891-1897.
HU Xiao-tao, WU Ke-ning, MA Jian-hui, et al. Control points for quality of arable land quality monitoring in Daxing district, Beijing[J]. *Resources Science*, 2012, 34(10):1891-1897.
- [15] 李启权, 岳天祥, 范泽孟, 等. 中国表层土壤有机质空间分布模拟分析方法研究[J]. 自然资源学报, 2010, 25(8):1385-1399.
LI Qi-quan, YUE Tian-xiang, FAN Ze-meng, et al. Study on method for spatial simulation of topsoil SOM at national scale in China[J]. *Journal of Natural Resources*, 2010, 25(8):1385-1399.
- [16] 卢岩君, 秦承志, 邱维理, 等. 基于少量典型样点土壤属性空间分布推测模型中的土壤属性参数敏感性分析[J]. 地理科学, 2011, 31(12):1549-1554.
LU Yan-jun, QIN Cheng-zhi, QIU Wei-li, et al. Sensitivity analysis of soil property parameter in typical-sample-based prediction model of digital soil mapping[J]. *Scientia Geographica Sinica*, 2011, 31(12):1549-1554.
- [17] 王珂, 沈掌泉, John S B, 等. 精确农业田间土壤空间变异与采样方式研究[J]. 农业工程学报, 2001, 17(2):33-36.
WANG Ke, SHEN Zhang-guan, John S B, et al. Spatial variants and sampling strategies of soil properties for precision agriculture[J]. *Transactions of the CSAE*, 2001, 17(2):33-36.
- [18] 潘瑜春, 刘巧芹, 阎波杰, 等. 采样尺度对土壤养分空间变异分析的影响[J]. 土壤通报, 2010, 41(2):257-262.
PAN Yu-chun, LIU Qiao-qin, YAN Bo-jie, et al. Effects of sampling scale on soil nutrition spatial variability analysis[J]. *Chinese Journal of Soil Science*, 2010, 41(2):257-262.
- [19] 范胜龙, 刘友兆. 不同布点密度条件下土壤有机碳的空间变异特性[J]. 土壤通报, 2012, 43(3):668-673.
FAN Sheng-long, LIU You-zhao. Spatial variability of soil organic carbon under different sampling densities[J]. *Chinese Journal of Soil Science*, 2012, 43(3):668-673.
- [20] 张忠启, 于法展, 李保杰. 土壤解氮空间变异与合理采样点数量研究[J]. 水土保持研究, 2013, 20(2):66-68, 72.
ZHANG Zhong-qi, YU Fa-zhan, LI Bao-jie. Spatial variation characteristics and reasonable sampling size of soil available nitrogen[J]. *Research of Soil and Water Conservation*, 2013, 20(2):66-68, 72.
- [21] 王宏斌, 杨青, 刘志杰, 等. 利用计算机模拟采样确定合理的土壤采样密度[J]. 农业工程学报, 2006, 22(8):145-148.
WANG Hong-bin, YANG Qing, LIU Zhi-jie, et al. Determining optimal density of grid soil-sampling points using computer simulation[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2006, 22(8):145-148.
- [22] Chen B S, Pan Y C, Wang J H, et al. Even sampling designs generation by efficient spatial simulated annealing[J]. *Mathematical & Computer Modelling*, 2013, 58(3/4):670-676.
- [23] 张淑杰, 朱阿兴, 刘京, 等. 基于模拟退火算法的土壤样点设计方法研究[J]. 土壤通报, 2013, 44(4):820-825.
ZHANG Shu-jie, ZHU A-xing, LIU Jing, et al. Soil sampling scheme based on simulated annealing method[J]. *Chinese Journal of Soil Science*, 2013, 44(4):820-825.
- [24] 王子龙, 陈伟杰, 付强, 等. 土壤优化采样策略研究进展[J]. 水土保持通报, 2017, 37(5):205-212.
WANG Zi-long, CHEN Wei-jie, FU Qiang, et al. Research progress of soil sampling strategy optimization[J]. *Bulletin of Soil and Water Conservation*, 2017, 37(5):205-212.
- [25] Minasny B, McBratney A B. A conditioned Latin hypercube method for sampling in the presence of ancillary information[J]. *Computers and Geosciences*, 2006, 32(9):1378-1388.
- [26] Anselin L. Local indicators of spatial association—LISA[J]. *Geographical Analysis*, 1995, 27(2):93-115.
- [27] Gao B B, Pan Y C, Chen Z Y, et al. A spatial conditioned Latin hypercube sampling method for mapping using ancillary data[J]. *Transactions in GIS*, 2016, 20(5):735-754.
- [28] 吴丹, 吉红霞, 刘玉, 等. 华北平原土地整治项目区农用机井的空间布局优化[J]. 经济地理, 2015, 35(6):154-160.
WU Dan, JI Hong-xia, LIU Yu, et al. Optimizing the distribution of pumping wells for agricultural land consolidation in the North China Plain[J]. *Economic Geography*, 2015, 35(6):154-160.