DOI：10.11766/trxb201708160188

基于成土环境地理邻域分析的历史土壤图训练样本筛选*

高 鸿[1, 2] 朱 娟[1, 3, 4†] 王良杰[5] 赵玉国[1, 2] 张甘霖[1]

（1 土壤与农业可持续发展国家重点实验室（中国科学院南京土壤研究所），南京 210008）

（2 中国科学院大学，北京 100049）

（3 安徽师范大学国土资源与旅游学院，安徽芜湖 241002）

（4 安徽省测绘局，合肥 230031）

（5 南京林业大学林学院，南京 210037）

摘 要 基于历史土壤图的知识挖掘和历史图更新对土壤资源调查、管理和利用有着重要的现实意义，而从历史土壤图中筛选代表性训练样本是进行知识挖掘和历史土壤图更新的关键步骤。以安徽省旌德县为研究区，提出一种新的土壤图训练样本筛选方法，包括样本数量确定和样本位置筛选。研究结果表明，面积分段线性缩放法确定的样本数量解决了已有研究未考虑同一类型多个图斑单元间样本数量分配的问题；采用邻域分析方法确定样本位置，当图斑位于地势平缓的区域时，基于高程因子和坡度因子确定的样本空间分布差异较小，而当图斑位于山区时，基于坡度因子确定的样本处于地形变化稳定的位置，全局代表性更高。通过与已有研究中环境因子直方图方法筛选样本进行对比，邻域分析方法确定的样本具有更高的差异比例和标准差，样本信息量更大。基于坡度因子采用邻域分析方法筛选出的图斑样本较高程因子样本拥有更高的全局空间代表性，邻域分析方法筛选的样本较相关研究中环境因子直方图方法筛选的样本拥有更高的信息量。

关键词 历史土壤图；样本数量；邻域分析；样本位置

中图分类号 P934 **文献标识码** A

20世纪我国进行的两次土壤普查产生的历史土壤图已应用至农业、生态环境等多个领域，成为土壤资源管理、生态水文模型等研究的主要数据来源[1-2]。然而，受当时制图者经验水平差异、传统制图技术以及数据支撑等的影响与限制，历史土壤图中存在边界错置的现象，空间分布精度受限，加之随着时间的推延土壤图的时效性降低，因而利用数据挖掘模型获取土壤—环境关系知识，进行土壤图的更新的研究越来越多[2-5]。此外，土壤图可以为缺少大量野外样点的地区提供土壤与环境关系知识[6]，从土壤图中挖掘土壤—环境关系知识需要对数据进行预处理和筛选[6-8]。而无论是土壤图更新中数据挖掘模型训练样本选取还是土壤图知识挖掘中对土壤图数据的预处理，关键点和难点均是如何在历史土壤图上选取具有代表性的训练样本。因而，如何从历史土壤图中有效选取代表性的训练样本具有较为重要的意义与价值。

土壤图更新和知识挖掘所需要的训练样本选取是指从不同土壤类型中筛选出样本数量与土壤图中面积大小相适应的、可代表土壤类型成土环境和图斑空间分布的训练样本，包括样本数量和样本空间位置的确定两个主要内容。对于样本数量的确定，

* 国家自然科学基金项目（41771251，41601209）、国家重点研发计划项目（2017YFC0803807）资助Supported by the National Natural Science Foundation of China（Nos. 41771251, 41601209），National Key Research and Development Plan of China（No.2017YFC0803807）

† 通讯作者Corresponding author，E-mail：zhujuannj@126.com

作者简介：高 鸿（1993—），男，甘肃天水人，硕士研究生，主要研究方向为数字土壤制图。E-mail: gaohong@issas.ac.cn

收稿日期：2017-08-16；收到修改稿日期：2017-11-24；优先数字出版日期（www.cnki.net）：2018-01-09

刘雪琦等[9]确定了一种基于土壤类型面积分级的训练样本数量确定的方法，Odgers等[10-11]在每个图斑里选择相同数目的训练样本。对于样本空间位置的确定，Odgers等[10-11]利用随机方法筛选样本，黄巍等[4]、朱阿兴等[6]、Qi和Zhu[7]通过建立环境因子直方图，在峰值区间进行土壤图样本选取，Yang等[3]利用模糊聚类法筛选土壤图样本并进行土壤图更新。

以上研究在确定样本数量时主要针对的是土壤类型，并未针对土壤类型所具有的多个图斑设计样点，这样可能会造成对较大面积图斑代表性不够或不同面积图斑之间样点数量失衡，而每个图斑均具有一定的地理空间意义，因而需要针对每个图斑设计样点数量。在确定样本位置时，通过环境因子直方图和聚类进行样本位置的选择只考虑了环境因子属性数值大小，忽略了环境因子在地理空间分布上的特征，而进行土壤类型空间预测所使用的样本应具有空间代表性，对环境因子进行空间分布变化的分析有助于筛选出成土环境稳定、空间代表性高的训练样本。基于此，本文提出一种基于环境因子邻域分析的土壤图训练样本筛选方法，包括根据土壤图图斑单元面积分段线性缩放的方法确定样本数量和基于环境因子地理空间分布特征来确定样本空间位置。

1 材料与方法

1.1 研究区概况

本文选取安徽省宣城市旌德县为研究区域（图1），118° 15′ ~118° 44′ E，30° 07′ ~30° 29′ N，位于安徽省南部。四面环山，地貌类型包括中山、低山、丘陵和山间盆地，海拔100 ~ 1 272 m；气候属于北亚热带湿润季风气候；植被以亚热带常绿阔叶林为主，包括常绿阔叶林、落叶阔叶林、马尾松林、灌丛、草丛等。旌德县第二次土壤普查历史土壤图包含红壤、黄壤、石灰岩土、潮土和水稻土等5种土类，9种土壤亚类，分别为山地黄壤（编号为S1）、棕色石灰土（编号为S2）、淹育型水稻土（编号为S3）、潜育型水稻土（编号为S4）、潴育型水稻土（编号为S5）、灰潮土（标号为S6）、红壤性土（编号为S7）、黄壤性土（编号为S8）、黄红壤（编号为S9），共1 364个土壤类型图斑。本文基于土壤亚类级别在土壤图斑单元中筛选训练样本。

1.2 数据来源

本文所使用数据包括第二次土壤普查形成的土壤图，将其数字化成矢量数据格式。此外，由于地形要素基本上可以代表土壤形成与发展过程中的主要影响要素[6]，所以本文所使用的另一主要数据为数字高程模型。高程数据来源于地理空间数据云平台（http://www.gscloud.cn），栅格空间分辨率为30 m × 30 m，在SAGA中计算坡度，采用土壤形成过程中高程和坡度这两个主要地形环境因子进行分析。本文采用ArcGIS 10.0和python2.7 GDAL进行空间数据处理。

1.3 样本数量计算

历史土壤图存在许多大小不一的图斑单元，其中面积较小的图斑单元受制作者主观影响较大，图斑边界及整体的可信度较低，因而本文设置面积阈值，将低于该阈值的图斑单元剔除。通过设置面积阈值选取历史土壤图中面积较大的土壤图图斑，再对所选择的图斑根据面积大小进行分段，得到不同面积子区间，不同的面积区间对应不同的样本数量区间，面积较小的区间样本数量小，反之样本数量大，在不同面积区间通过线性缩放法建立面积大小与样本数量的线性映射，计算得出所有图斑对应的具体样本数量。当存在过大面积图斑时，对面积进行分段可消除根据图斑面积比例直接确定样点数量造成的样本数量差异悬殊；在分段后的区间内通过线性缩放法建立

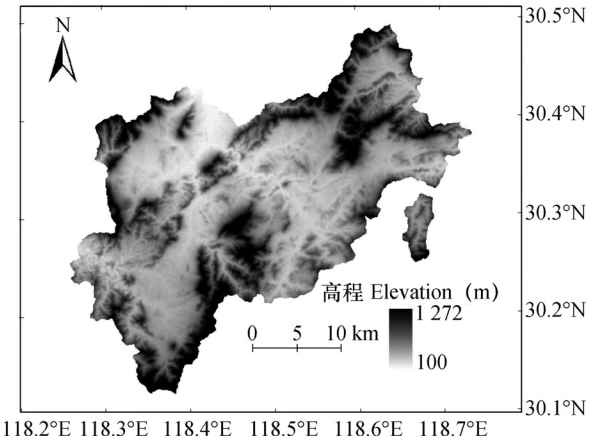

图1 研究区位置

Fig. 1 Location map of the study area

面积大小与样本数量的线性映射，基于面积大小计算得出样本数量，则会保证样本数量与整个研究区不同面积图斑相对应。

对于面积较大的山地黄壤、棕色石灰土、潜育型水稻土、潴育型水稻土、红壤性土、黄壤性土、黄红壤等土壤类型，图斑单元面积阈值为50 hm^2，面积较小的淹育型水稻土和灰潮土，图斑单元面积阈值为10 hm^2，面积阈值是结合不同土壤类型图斑的面积和数目确定的，主要目的是删除精度较低的细小图斑和减小计算量，具体阈值可随不同的研究数据而变化。筛选后图斑总面积占原始总面积的79.7%，图斑单元数占原始单元数的18.6%，通过面积阈值筛选保留了较少数量的分析图斑，但同时分析图斑的面积占研究区总面积的比重较大，减小了分析过程中的图斑计算量。

根据面积阈值筛选后的所有图斑面积分布，设置不同的面积区间和对应的样本数量区间（表1）。表中图斑面积分段区间值和对应的样本数量区间值是作者多次尝试确定的，具体的参数值可随不同的研究区域和不同的研究者而异。

表1　面积区间与样本数量区间对应表

Table 1　Reference between polygon area range and sample quantity range

图斑面积 Polygon area（hm^2）	图斑单元数 Number of polygons	样本数量 Number of samples（hm^2）
10 ~ 50	14	5 ~ 10
50 ~ 100	119	10 ~ 15
100 ~ 500	94	15 ~ 20
>500	27	20 ~ 30

对各面积区间，其中不同面积图斑单元所对应样本数量的计算公式为：

$$N_i = \left(\frac{N_{\max} - N_{\min}}{\mathrm{Area}_{\max} - \mathrm{Area}_{\min}} \right) \times (\mathrm{Area}_i - \mathrm{Area}_{\min}) + N_{\min}$$

式中，N_i表示不同区间第i个图斑样本数量；Area_i表示不同区间第i个图斑的面积；$N_{\max}$表示样本数量区间的上限；$N_{\min}$表示样本数量区间的下限；$\mathrm{Area}_{\max}$表示该区间图斑面积的最大值；$\mathrm{Area}_{\min}$表示该区间图斑面积的最小值。

1.4　样本空间位置确定

对于样本位置，本文假设对于某些主要成土环境因子，土壤图图斑训练样本的位置位于该因子空间变化稳定的位置。因为一方面当成土环境要素在地理空间某一尺度局部区域均质、稳定的分布时，其对土壤的影响也是持续、稳定的存在，更可能对应着特定的土壤类型；另一方面，从环境因子变化稳定的局部区域选择位于其地理中心的位置作为样本，样本对该局部区域成土环境具有较高的空间代表性。在土壤图图斑内，采用较小空间尺度，选择环境因子空间变化稳定的位置作为该图斑内土壤类型训练样本的位置，筛选出可能是土壤类型所需成土环境并且空间代表性较高的多个空间位置。通过设置指标对环境因子进行栅格窗口邻域分析，得出当前栅格窗口的尺度下环境因子空间变化程度的分布情况，进而筛选出环境因子空间变化程度最小也即空间变化最稳定的位置。具体思路是，在图斑单元内，按照环境因子邻域分析结果值从小到大的顺序依次筛选象元作为样本，第一次选出邻域分析结果值最小的样本，如果下一次选出的样本与已选出样本的距离在一个邻域内，则舍弃该样本继续选出下一个样本，当样本数量达到计算得出的该图斑单元目标数量时为止。

其中，邻域分析时采用的反映环境因子在空间分布上变化大小的指标定义如下：

$$M = \frac{1}{N-1} \times \sum_{t=1}^{N} (D_i - D_0)^2$$

式中，M表示邻域内环境因子空间变化的程度，N表示邻域内象元总个数，D_i表示第i个邻域象元因子值，D_0表示自身象元因子值，M表示某一位置象

元环境因子在当前空间邻域内的变化程度。邻域分析结果*M*越小表示该位置的环境因子数值与周围的差距越小，空间变化最稳定，因而该位置的土壤类型可能是图斑所对应的土壤类型。

基于本文的上述假设，在一个图斑内，土壤类型训练样本出现在环境因子邻域分析结果最小的位置。但对于不同的环境因子，这些位置可能存在较大差异，例如图2显示了黄壤性土类型海拔分布较大的某一图斑内，以3×3栅格窗口为邻域分别对高程因子和坡度因子进行邻域分析，空间变化最小5%、10%和20%象元的分布情况，其中空间变化最小5%是指图斑单元内邻域分析结果值小于5%分位数的象元，即占总数目5%的空间变化较小的象元位置，10%和20%同理。

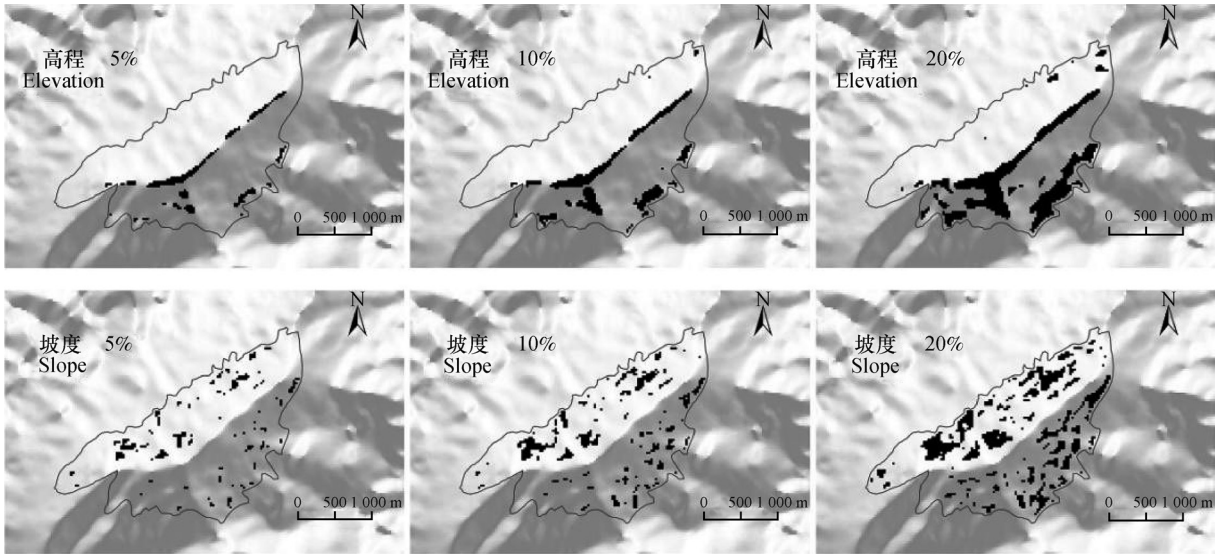

图2 不同因子空间变化最稳定象元位置分布

Fig. 2 Distribution of the most stable cells in spatial variation relative to covariate

由图2可知，高程因子空间变化最小的象元位于山脊、山脚等高程变化小、平缓的位置，空间分布高度集聚，图斑内绝大部分空间位置未被覆盖。对坡度因子进行邻域分析得出坡度的空间变化情况，空间变化最小的象元位于山脊与山脚中部的地形变化稳定的位置，并且覆盖到图斑大部分范围。所以，需要进行不同位置图斑、不同环境因子的讨论用以说明本文所提出方法的不同效果。

由于地形要素基本上可以代表土壤形成与发展过程中的主要影响要素[6]，本文对高程和坡度两个主要的地形要素环境因子进行讨论。本文基于高程和坡度两个重要的环境因子，选取3×3和7×7两种栅格窗口进行邻域分析筛选样本，由于研究区土壤图各土壤类型图斑最小面积的中数为36 250 m^2，而环境因子空间分辨率为30 m，所以不宜采用较7×7更大的栅格窗口作为邻域。

1.5 结果评价

评价样本筛选结果的指标之一就是样本的空间分布特征，不同数量的训练样本空间分布应覆盖大部分图斑以增加样本对图斑全局的代表性，即尽可能完备地筛选出包含一系列接近图斑土壤类型所对应的成土环境特征。此外，所采集的训练样本数值应该各不相同、具有较大的信息量，以保证样本对全局具有较高的代表性。

本文将各土壤图图斑所选出的训练样本进行位置分布制图以分析训练样本的空间分布，计算差异比例和标准差两个指标来评价样本的信息量。差异比例是指一组样本中除去重复值后的样本个数与该组样本总个数之间的比例，差异比例越大，说明样本重复程度低，差异程度高。在数理统计中，方差可作为评价信息量的一个简易指标，如在主成分分析中就是寻找方差最大的方向以保证较小的信息损失，方差越大表示样本的信息量越大，因而本文采用与方差相关的标准差这一统计指标。

本文还采用已有研究中环境因子直方图峰值区间的方法基于高程和坡度确定样本，通过对比来评价说明本文提出的邻域分析方法确定样本方法。环境因子直方图峰值区间的方法能够筛选出训练样本，被广泛用于从历史土壤图筛选训练样本，并进行历史土壤图知识挖掘[6-7]和历史土壤图的更新[3-4,9]，具有较高的更新精度。

2　结果与讨论

2.1　各土壤类型目标样本数量

通过上述面积分段线性缩放的方法，计算土壤图各个图斑单元的样本数量，最后对图斑所属土壤类型进行汇总，得出各土壤类型的样本数量（表2）。

表2　各土壤类型目标样本数量

Table 2　Quantity of target samples relative to soil type

土壤类型 Soil type	原始面积 Original area（hm^2）	原始图斑数 Original number of polygons	采样面积 Sampling area（hm^2）	采样图斑数 Number of sampling polygons	目标样本数量 Number of target samples
山地黄壤①	647.3	3	610.3	2	35
棕色石灰土②	4 247	22	4024	7	137
淹育型水稻土③	285.5	17	233.8	10	68
潜育型水稻土④	3 840	160	1 315	10	137
潴育型水稻土⑤	16 430	611	7 296	74	976
灰潮土⑥	137.6	5	137.7	5	38
红壤性土⑦	50 580	454	4 5090	106	1 671
黄壤性土⑧	7 890	34	7 597	18	309
黄红壤⑨	6 054	58	5 502	22	334

①Ali-Perudic Argosols，②Brown Carbonati-Udic Argosols，③Hapli-Stagnic Anthrosols，④Gleyi-Stagnic Anthrosols，⑤Fe-accumuli-Stagnic Anthrosols，⑥Ochri-Aquic Cambosols，⑦Ali-Udic Cambosols，⑧Ali-Perudic Cambosols，⑨Xanthic Ali-Udic Cambosols

其中原始面积和原始图斑数是在面积阈值筛选之前的土壤类型面积和土壤类型图斑单元数，采样面积和采样图斑数是面积阈值筛选之后的土壤类型面积和土壤类型图斑单元数，目标样本数量即为所属土壤类型所有采样图斑单元样本数量的总和。面积阈值筛选用较少的图斑单元数保留了较大的面积，而目标样本数量的大小大致与土壤类型采样面积和采样图斑数相匹配。对不同土壤类型的具体图斑单元而言，则具有与其面积对应的样本数量（表3），表中图斑为后文样本空间位置展示时所选的研究区典型土壤类型图斑单元。

表3　典型土壤类型图斑样本数量

Table 3　Quantity of samples of a typical soil map polygon

土壤类型 Soil type	图斑编号 Polygon code	图斑面积 Polygon area（hm^2）	目标样本数 Number of target samples
潴育型水稻土①	S5_63	110.7	15
红壤性土②	S7_68	842.1	20
黄壤性土③	S8_7	451.2	19

①Fe-accumuli-Stagnic Anthrosols，②Ali-Udic Cambosols，③Ali-Perudic Cambosols

2.2　典型土壤图斑单元样本空间位置

对于样本空间位置分布结果，本文选取研究区三种典型土壤类型的典型图斑进行展示分析。由于研究区山体、丘陵分布较广，所以选取研究区面积较大或者图斑较多、出现在山体不同位置的土壤类型作为典型土壤类型，即选取随海拔升高发育的土壤类型为序列，分

别为潴育型水稻土、红壤性土和黄壤性土。潴育型水稻土分布在河流两岸、低平丘陵的底部区域；红壤性土主要分布在600～700 m以下的低山地区；黄壤性土分布在600～700 m以上的山坡。三种典型图斑内，基于高程和坡度的不同环境因子，采用不同方法确定样本的空间位置（图3）。

相比于直方图峰值区间方法，邻域分析方法是根据环境因子空间变化，筛选出一系列成土环境稳定、对图斑具有全局代表性的空间位置。直方图峰值区间的方法只在直方图峰值一个数值区间内筛选样本，样本相似度高，筛选样本不全面，并且可能会出样本的局部集聚，造成样本冗余。采用邻域分

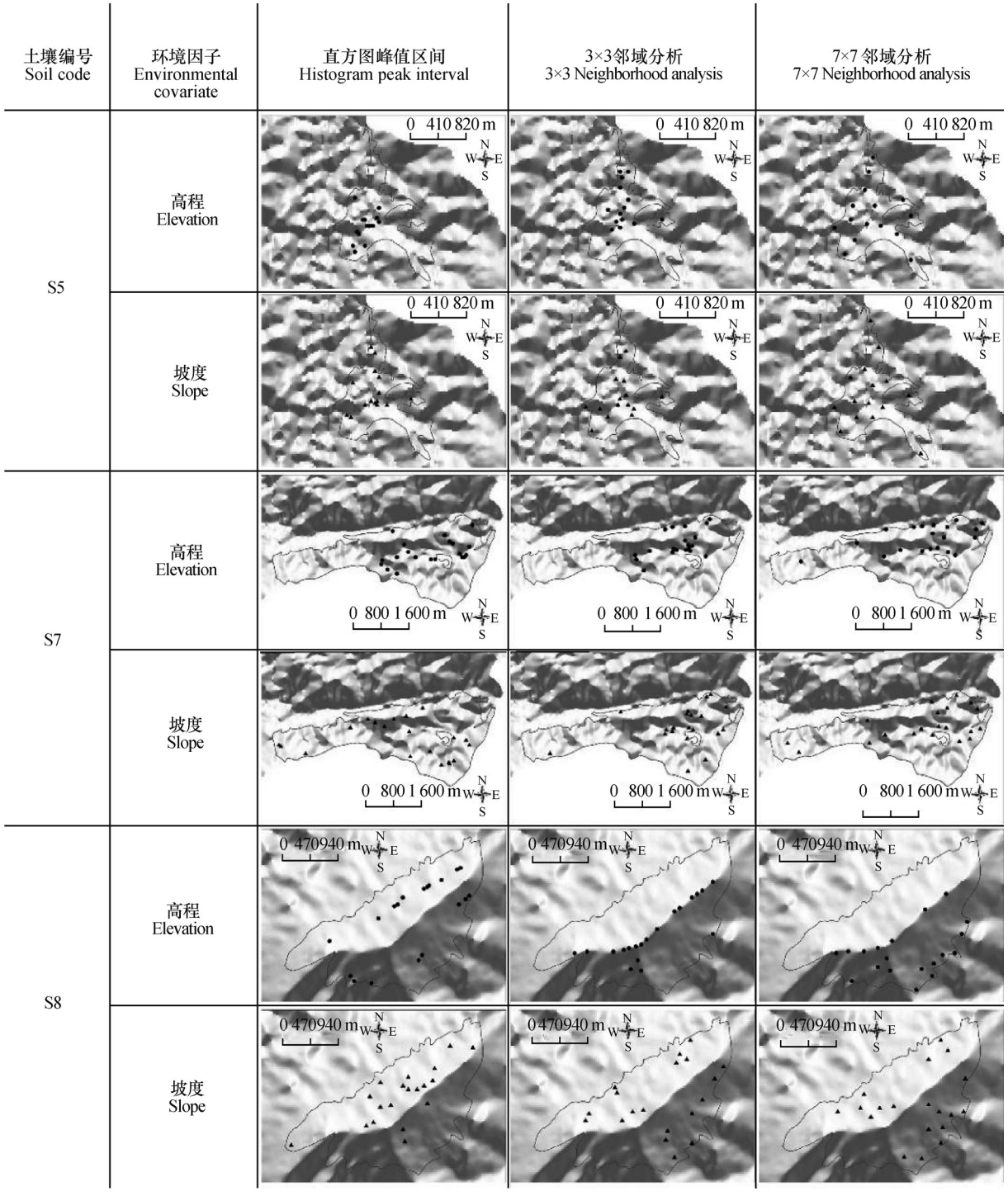

图3 典型土壤图斑单元训练样本空间分布

Fig. 3 Spatial distribution of training samples in typical soil map polygon

析方法确定样本位置，当图斑位于地势平缓的区域时，如潴育型水稻土，基于高程因子和基于坡度因子确定的样本位置空间分布差异较小，而当图斑位置位于山区时，如黄壤性土，基于高程因子采用邻域分析方法确定的样本位置会出现集聚，而基于坡度因子采用邻域分析方法确定的样本处于地形变化稳定的位置，样本分布在图斑范围内覆盖更大，样本对图斑分布的代表性更高。因而，当图斑位于山区时，基于坡度因子比基于高程因子邻域分析筛选的样本，对图斑整体分布范围具有更高的代表性。

2.3　典型土壤图斑单元样本信息量

同样以3种典型土壤为例，每种土壤选取5个图斑单元，分别计算不同方法基于高程和基于坡度确定各图斑单元样本高程信息和坡度信息的差异比例以及标准差，其中差异比例图所示（图4）。

各图斑单元样本标准差如表4所示：

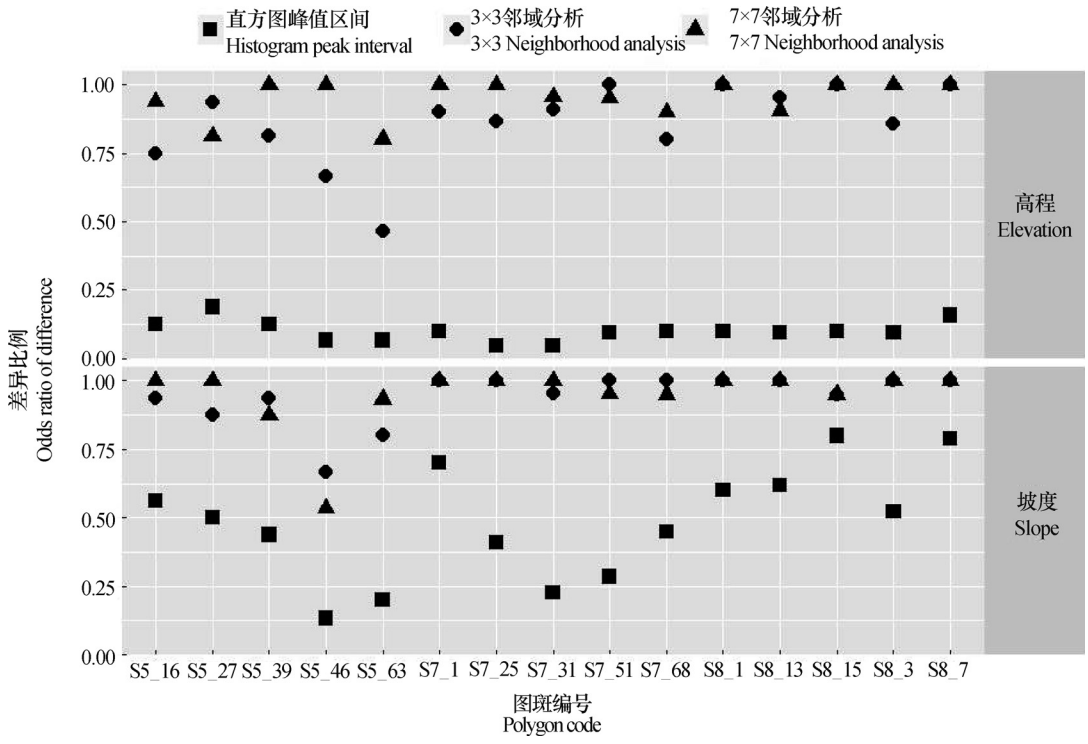

图4　不同土壤图图斑样本差异比例

Fig. 4　Odds ratio of difference in samples relative to soil map polygons

表4　不同土壤图图斑样本标准差

Table 4　Standard deviation of samples relative to soil map polygon

图斑编号 Polygon code	高程 Elevation			坡度 Slope		
	直方图峰值区间 Histogram peak interval	3×3邻域分析 3×3 Neighborhood analysis	7×7邻域分析 7×7 Neighborhood analysis	直方图峰值区间 Histogram peak interval	3×3邻域分析 3×3 Neighborhood analysis	7×7邻域分析 7×7 Neighborhood analysis
S5_16	0.48	14.74	17.78	0.09	2.47	1.84
S5_27	0.87	18.97	22.41	0.05	3.55	3.89
S5_39	0.48	19.2	28.88	0.1	3.23	3.2

续表

图斑编号 Polygon code	高程 Elevation			坡度 Slope		
	直方图峰值区间 Histogram peak interval	3×3邻域分析 3×3 Neighborhood analysis	7×7邻域分析 7×7 Neighborhood analysis	直方图峰值区间 Histogram peak interval	3×3邻域分析 3×3 Neighborhood analysis	7×7邻域分析 7×7 Neighborhood analysis
S5_46	0	10.68	13.29	0.03	0.83	0.43
S5_63	0	5.3	10.3	0.05	1	1.15
S7_1	0.46	40.99	64.97	0.05	4.4	4.27
S7_25	0	65.24	41.1	0.01	5.34	4.1
S7_31	0	91.53	66.45	0.01	4.91	5.82
S7_51	0.49	85.56	102.27	0.03	5.85	4.61
S7_68	0.5	17.08	39.51	0.03	9.25	9.09
S8_13	0.49	123.9	77.29	0.03	5.67	6.18
S8_15	0.5	179.5	177.9	0.06	5.94	4.78
S8_1	0.49	123.9	77.29	0.03	5.67	6.18
S8_3	0.45	106.98	97.71	0.03	3.04	3.47
S8_7	0.78	81.48	140.98	0.06	5.4	6.04

无论是基于高程因子还是坡度因子，邻域分析方法确定的样本其差异比例和样本标准差均大于环境因子直方图确定的样本，这是因为直方图峰值区间范围内的样本的环境因子值比较接近，样本差异比例较小，标准差也较小。邻域分析方法是从周围地理环境的角度确定样本位置，未从某一数值区间内筛选样本，因而确定的样本相比于直方图峰值区间样本具有较大的差异比例和标准差，具有较高的信息量，全局代表性更高，其中邻域大小对差异比例和标准差的影响随图斑而异。

但受历史土壤图空间分布精度的影响，从历史土壤图上筛选出的样本精度各异，因而历史土壤图精度是影响土壤图训练样本的主要因素。

3 结 论

对于历史土壤图图斑单元采用面积分段线性缩放的方法确定样本数量，可保证面积和图斑数与训练样本数量之间的对应，消除了过大面积土壤类型确定样本数量时造成的样本数量差异悬殊，图斑面积与样本数量直接匹配，利于从图斑直接筛选训练样本。采用邻域分析方法，当图斑位于地势平缓的区域时，基于高程因子和坡度因子确定的训练样本空间分布差异较小；当图斑位于山区时，基于坡度因子所确定的训练样本处于地形变化稳定的位置，对图斑的空间分布也具有更高的代表性，较高程因子更适合训练样本位置的筛选。此外，邻域分析方法确定的训练样本较环境因子直方图方法确定的样本具有更高的差异比例和标准差，样本信息更丰富，样本对全局的代表性更高。在土壤图不同土壤类型的图斑内，基于面积分段线性缩放法确定合适数量、基于环境因子邻域分析法确定特定位置的训练样本后，可以进行后续土壤图知识挖掘和土壤图更新的研究，其中不同土壤类型筛选样本时的环境因子的选取以及不同位置、大小图斑所需要的空间邻域大小的确定，均可进一步根据相关土壤地理学知识进行优化。

参 考 文 献

[1] 席承藩，章士炎. 全国土壤普查科研项目成果简介. 土壤学报，1994,31（3）：330—335
Xi C F，Zhang S Y. Brief introduction on achievements in national soil survey project since 1979 （In Chinese）. Acta Pedologica Sinica，1994，31（3）：330—335

[2] 杨琳，Fahmy Sherif，Jiao You，等. 基于土壤—环境关系的更新传统土壤图研究. 土壤学报，2010，47（6）：1039—1049
Yang L，Fahmy S，Jiao Y，et al. Updating conventional soil maps using knowledge on soil environment relationships extracted from the maps（In Chinese）. Acta Pedologica Sinica，2010，47（6）：1039—1049

[3] Yang L，Jiao Y，Fahmy S,et al. Updating conventional soil maps through digital soil mapping. Soil Science Society of America Journal，2011，75（3）：1044—1053

[4] 黄巍，罗云，汪善勤，等. 基于传统土壤图的土壤—环境关系获取及推理制图研究，土壤学报，2016,53（1）：72—80
Huang W，Luo Y，Wang S Q，et al. Knowledge of soil-landscape model obtain from a soil map and mapping（In Chinese）. Acta Pedologica Sinica，2016，53（1）：72—80

[5] Rad M R P，Toomanian N，Khormali F，et al. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma，2014，232/234：97—106

[6] 朱阿兴，李宝林，裴韬，等. 精细数字土壤普查模型与方法. 北京：科学出版社，2008
Zhu A X，Li B L，Pei T，et al. Fine digital soil census model and method（In Chinese）. Beijing：Science Press，2008

[7] Qi F，Zhu A X. Knowledge discovery from soil maps using inductive learning. International Journal of Geographical Information Science，2003，17（8）：771—795

[8] 杨琳，朱阿兴，李宝林，等. 应用模糊c-均值聚类获取土壤制图所需土壤—环境关系知识的方法研究. 土壤学报，2007，44（5）：784—791
Yang L，Zhu A X，Li B L，et al. Extraction of knowledge about soil-environment relationship for soil mapping using fuzzy c-means（FCM）clustering（In Chinese）. Acta Pedologica Sinica，2007，44（5）：784—791

[9] 刘雪琦，朱阿兴，杨琳，等. 土壤图更新中基于土壤类型面积分级的训练样点选择方法. 土壤学报，2017，54（1）：36—47
Liu X Q，Zhu A X，Yang L，et al. Training sample selection method based on grading of soil types by area for updating conventional soil maps（In Chinese）. Acta Pedologica Sinica，2017，54（1）：36—47

[10] Odgers N P，Sun W，Mcbratney A B，et al. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma，2014，214/215：91—100

[11] Odgers N P，McBratney A B，Minasny B. Digital soil property mapping and uncertainty estimation using soil class probability rasters. Geoderma，2015，237/238：190—198

[12] Yang L，Zhu A X，Qi F，et al. An integrative hierarchical stepwise sampling strategy and its application in digital soil mapping. International Journal of Geographical Information Science，2013，27（1）：1—23

[13] Silva S H G，Menezes M D D，Owens P R，et al. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. Geoderma，2016，267：65—77

[14] Malone B P，Minasny B，Odgers N P，et al. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma，2014，232/234：34—44

[15] Pahlavan-Rad M R，Khormali F，Toomanian N，et al. Legacy soil maps as a covariate in digital soil mapping：A case study from Northern Iran. Geoderma，2016，279：141—148

Screening of Training Samples Based on Environmental Covariate Geospatial Neighborhood Analysis of Historical Soil Maps

GAO Hong[1, 2] ZHU Juan[1, 3, 4†] WANG Liangjie[5] ZHAO Yuguo[1, 2] ZHANG Ganlin[1]

（1 *State Key Laboratory of Soil and Sustainable Agriculture*, *Institute of Soil Science*, *Chinese Academy of Sciences*, *Nanjing* 210008, *China*）

（2 *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*）

（3 *College of Territorial Resources and Tourism*, *Anhui Normal University*, *Wuhu*, *Ahhui* 241002, *China*）

（4 *Anhui Bureau of Surveying and Mapping*, *Hefei* 230031, *China*）

（5 *College of Forest*, *Nanjing Forestry University*, *Nanjing* 210037, *China*）

Abstract 【Objective】It is of great practical significance to investigate, manage and exploit soil resources based on knowledge mining and updating of historical soil maps, while screening of representative training samples out of the historical soil maps is a key step to accomplish the task. 【Method】Jingde County of Xuancheng City in Anhui Province was cited as the study area. In this paper, a new method for screening of training samples was developed, consisting of determining quantity of samples and specifying sample locations. The area segmented linear-scaling method, which builds a linear correspondence between the area and the number of samples in each area segment, was applied to determination of quantity of sample in each historical soil map polygon, and then after making geographical neighborhood analysis of elevation and slope, the two important environmental covariates, the stable cells in spatial variation of the environmental covariates were defined as sample locations. Geographical neighborhood analysis index indicates the degree of spatial variation of the environmental covariates in the neighborhood.【Result】Results show that the use of the area segmented linear-scaling method to determine quantity solved the problem that had never been pondered in past researches of how to distribute samples among units that were of the same type, but consisted of a number of polygons. When this method was used to define sample locations in polygons located in topographically flat areas, determination of spatial distribution of sample sites based on elevation or slope did not vary much, whereas in polygons located in mountainous areas, the sample sites defined based on slope were mostly located in places relatively stable in topography and more representative of the entire polygon. Compared with the environmental covariates histogram peak method used in most of the researches, this method is higher in odds ratio and standard deviation, and the samples defined with this method are bigger in volume of information.【Conclusion】The training samples defined with the geographical neighborhood analysis method based on slope are more representative of the entire polygon than those based on elevation, and contain more information than those defined with the environmental covariates histogram peak method in relevant researches.

Key words Historical soil map; Sample quantity; Geographical neighborhood analysis; Sample location

（责任编辑：檀满枝）