DOI：10.11766/trxb201604210130

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*

刘雪琦[1] 朱阿兴[1，2，3，4] 杨 琳[2†] 缪亚敏[1] 曾灿英[1]

（1 南京师范大学地理科学学院，南京 210023）

（2 资源与环境信息系统国家重点实验室（中国科学院地理科学与资源研究所），北京 100101）

（3 虚拟地理环境教育部重点实验室（南京师范大学），江苏省地理环境演化国家重点实验室培育建设点，江苏省地理信息资源开发与利用协同创新中心，南京 210023）

（4 Department of Geography，University of Wisconsin-Madison，Madison，WI 53706，USA）

摘 要 基于数据挖掘模型的土壤图更新是一项重要的研究。数据挖掘模型构建中训练样点的质量不仅决定其对研究区土壤-环境关系表达的充分程度，而且会对推理制图的结果产生至关重要的影响。本文提出一种基于土壤类型面积分级的典型训练样点选择方法，即依据土壤面积对土壤类型分级，并按照等级之间的比例关系基于典型点选择训练样点。将方法应用于更新美国威斯康星州Raffelson流域的传统土壤图，并与另外两种训练样点选择方法对比，以验证该方法的有效性。结果表明，500次重复实验中，本研究方法与另外两种训练样点选择方法相比，能够更新传统土壤图的比例分别为79.5%、71.8%和63.6%，而且其推理制图结果更符合研究区土壤分布的特征。本研究所提方法是一种有效的训练样点选择方法。

关键词 训练样点；数据挖掘模型；传统土壤图更新；土壤—环境关系

中图分类号 P934 **文献标识码** A

长期以来，土壤专家通过土壤普查技术积累了大量的土壤图数据（传统土壤图）[1-2]。然而，受制图技术、数据质量以及人为主观性（如：制图者“经验”、人工目视解译差别）等影响，土壤多边形的边界很可能产生错置[3-4]，造成传统土壤图的精度通常不高[5-6]。但是，土壤图中蕴含着土壤专家们对当地土壤—环境关系的理解和探究[7]，随着可获取的环境数据的增多以及数据挖掘方法的推进，更新传统土壤图以提高其精度成为可能[8-11]。

目前利用数据挖掘模型更新传统土壤图是获取土壤空间分布信息的一种重要方法[8，12-13]，其原理是通过训练样点挖掘每种土壤类型发育或存在的环境条件，即获取各土壤类型的土壤—环境关系，然后将这种关系应用于土壤类型（属性）的推测与制图[14]。训练样点的质量对土壤—环境关系表达的充分与否，以及对推理制图的精度高低均会产生至关重要的影响[15]。当前训练样点的获取途径主要有两种：第一种是野外采集，其获取的样点实时性较好，但野外采样耗时耗力，成本高昂[16]；第

* 国家自然科学基金项目（41431177，41471178）、江苏省高校自然科学研究重大项目（14KJA170001）、江苏省高校研究生科研创新计划项目（KYLX15_0715）、国家重点基础研究发展计划973项目（2015CB954102）和千人计划资助 Supported by the National Natural Science Foundation of China（Nos.41431177 and 41471178），the Natural Science Research Program of Jiangsu（No.14KJA170001），the Graduate Research Innovation Program of Jiangsu（No.KYLX15_0715），the National Basic Research Program of China（No.2015CB954102），and the “One-Thousand Talents” Program of China

† 通讯作者 Corresponding author：杨 琳，女，博士，副研究员，硕士生导师。E-mail：yanglin@ lreis.ac.cn

作者简介：刘雪琦（1992—），女，内蒙古包头人，硕士研究生，主要从事数字土壤制图研究。E-mail：xueqiliu@yeah.net

收稿日期：2016-04-21；收到修改稿日期：2016-07-13；优先数字出版日期（www.cnki.net）：2016-07-28

二种是基于传统土壤图的采集，虽然受限于土壤图的精度，但传统土壤图资源丰富，从土壤图中选择训练样点可极大地降低样点采集成本，因此，这种方法得到广泛应用[3, 8, 12]。

基于土壤图选择训练样点的方法主要有三种：（1）每种土壤类型选择相同数目的训练样点[17]；（2）每个土壤多边形选择相同数目的训练样点[18]；（3）按照每种土壤类型的面积所占比例选择相应数目的训练样点[19]。一般而言，对于土壤图中像元数目少的土壤类型或斑块面积小的土壤多边形，只需数量较少的训练样点就能充分表达它们的土壤—环境关系，但对于像元数目多的土壤类型，它们发育的环境条件相对复杂，而且可能包含多种土壤—环境关系，所以需要数量较多的训练样点对其进行表达和体现[15, 17, 19]。但是方法一与方法二针对不同像元数目的土壤类型或不同面积的土壤多边形均选择相同数目的训练样点，这在很大程度上降低了训练样点的质量。而方法三是完全依照土壤类型的面积比选择相应数目的训练样点，对于面积相差较大的土壤类型可能会得到失衡的训练样点集[20-21]，即利用过多的训练样点描述面积大的土壤类型的土壤—环境关系，但对于面积小的土壤类型却不能通过极少的训练样点充分反映其环境特征，这很可能降低推理制图的精度。

此外，上述方法运用时大多是在确定选取数量后直接在土壤多边形中随机选择训练样点。然而，由于土壤图中存在边界错置等问题，这种随机选择样点的方式可能会造成训练样点无法充分表达土壤类型的土壤—环境关系或含有冗余数据的情况。已有研究者尝试在土壤类型的典型像元中选择训练样点[5, 12]。结果表明，基于典型像元选取训练样点用于制图的精度高于随机选择训练样点的推理制图精度。但该研究选择每种土壤类型的训练样点数目是采用与其面积成正比的方法，并未考虑土壤类型面积与该类型下所选择的样点数目之间的平衡关系。

本文针对以上问题，提出一种基于土壤类型面积分级的训练样点选择方法，首先依据土壤类型面积对土壤类型分级，然后按照等级之间的比例关系基于典型样点选择训练样点。以美国威斯康星州Raffelson流域为研究区，选择在传统土壤图更新方面广泛使用的随机森林模型（Random forest model，RF）为推理制图模型[22-24]，对比本文方法与其他两种训练样点选择方法，利用野外独立验证样点检验不同训练样点选择方法用于推理制图的结果与精度，以评价本研究方法的有效性。

1　研究方法

1.1　训练样点获取方法

（1）获取各土壤类型的典型点集。训练样点的选择应尽量减少因土壤多边形的边界错置所产生的“噪音”像元[3]。假设传统土壤图中土壤多边形覆盖的大部分面积或范围是准确的，那么某种土壤类型的所有土壤多边形内某个环境因子直方图（即土壤—环境直方图）的峰值即代表了该土壤发育或存在的典型环境特征，可以认为某土壤类型的环境条件接近或落入峰值区域内的像元即为典型像元[3, 12]。

针对每个土壤—环境直方图，其横轴代表某种环境因子的值域，纵轴代表环境条件在对应区间内的像元数量。直方图的区间数量会影响直方图峰值区间的确定，也会影响对应峰值区间的典型像元的确定。本文通过设定落在每个区间内的像元数目来确定区间数量，公式如下[4]：

$$N_i=\mathrm{int}\left(N_p / r_i\right) \quad (1)$$

式中，N_i表示土壤类型i的土壤—环境直方图中划分的区间数量；N_p表示该土壤类型的所有像元数目；r_i表示该土壤类型选择训练样点的数目，即每个区间内的像元数目。

对于每种土壤类型典型像元的确定均通过以下两个步骤完成，第一步是一次对一种环境因子进行像元采样，即只采用某环境条件落入土壤—环境直方图峰值区的像元；第二步是将某土壤类型的所有环境因子的像元汇总，对于可能不止一次落入土壤—环境直方图峰值的像元只保留一次，其余像元直接合并进该土壤类型的典型点集，以此类推得到每种土壤类型的典型点集。根据直接合并所得每种土壤类型的典型点集在一定程度上降低了样点中“噪音”像元的出现概率，但随着所选总样点数目的增加，这种减少“噪音”像元的能力也逐渐弱化。

（2）基于土壤类型面积分级的典型训练样点选择方法。基于每种土壤类型的典型点集进行训练样点的选择，以提高训练样点的质量。为平衡土壤类型面积与训练样点数量之间的关系，本文所采用

的方法是对土壤类型的面积进行分级，利用不同等级之间的比例关系选择训练样点，以适当缩小不同土壤类型面积之间的差距。

土壤类型面积分级方法的基本思路是，首先对各土壤类型所占面积比取对数，以缩小不同土壤类型面积之间的绝对差异，然后，将对数结果转换为正值再进行取整分级，即可以将各土壤类型所占面积比转换为不同的等级，例如，1、2、3三个等级，最后根据等级值的比值得到各等级中训练样点数目的比例关系，在此考虑到仍需体现土壤类型面积之间的差异，可利用对等级值取平方的方式拉伸等级之间的比例关系，例如，1、2、3三个等级内的土壤类型所选训练样点数目之比为1∶4∶9，各土壤类型训练样点数目的比例关系具体求算方式如下：

$$L=(\mathrm{int}(R-\lg(N_p/N)))^2 \quad (2)$$

式中，L表示该种土壤类型所在等级对应的训练样点比值；N_p表示某种土壤类型的所有像元数目；N表示土壤图的所有像元数目；R表示一个实数，其值的确定需保证最小等级Ⅰ所对应的比值至少为1；（int（$R-\lg(N_p/N)$））代表各土壤类型的取整分级结果。

利用式（2）选择训练样本，只需设定等级I的训练样点数目，便可根据比例关系确定其他等级的训练样点数目，之后基于每种土壤类型的典型点集选择相应数目的样点构成训练样点集。

1.2 利用随机森林模型推理制图

随机森林模型的基本原理是以随机的方式建立起一棵棵决策树，然后由这些决策树组成一个森林，其中每棵决策树之间没有关联，当有一个新的样本输入时，每棵树会独立地做出判断，按照投票原则决定该样本的分类结果[25-27]。其中“随机”包括两次随机选择的过程，第一次是按比例随机选择训练样本集中的子样本集；第二次是对变量，即环境因子的随机选择。随机森林模型不易出现过拟合现象，同时因其最终结果是以投票原则决定，所以它对离群值不敏感，抗噪能力强[28]，推理制图精度较高。

模型建立过程涉及到2个关键的参数：mtry和ntree。其中mtry是每次树模型重建时随机选择环境因子的个数，模型推测结果为类别变量时mtry=$\sqrt{E}$[29]，E表示输入到模型的环境因子个数；ntree是随机森林中决策树的个数，利用R语言中的random Forest包在确定mtry的前提下可以自动推荐ntree的最优数值。

训练样点和最优参数输入随机森林模型中建模，并应用到整个研究区，对土壤图中每个像元的土壤类型进行分类预测，得到推理土壤图。为更加准确地探究训练样点的选择方法与土壤图更新之间的关系，避免个别训练样点集对制图精度的影响，本文针对提出的训练样点选择方法重复采样500次，即利用所选训练样点和模型最优参数，基于随机森林模型的推理制图重复实验500次。

1.3 精度验证

本文将提出的训练样点选择方法A与另外两种训练样点选择方法B和C进行对比，验证方法A的有效性。方法B是按照方法A所得每种土壤类型等级之间的比例关系，在每种土壤类型多边形内的所有像元中随机选择样点构成训练样点集，方法C是完全按照土壤类型面积所占比例，在每种土壤类型多边形内的所有像元中随机选择样点构成训练样点集（即引文中提到的第三种方法）。方法A与方法B的区别在于方法A 是在典型像元中进行选点而方法B是在土壤类型多边形内的所有像元中进行随机选点，二者对比针对基于典型像元选择训练样点与多边形内随机选择训练样点的差异；方法B与方法C均是在土壤类型多边形内的所有像元中进行选点，但区别在于方法B是按照土壤类型面积分级确定样点数量而方法C是完全依照土壤类型的面积比确定样点数量，二者对比针对基于土壤类型面积分级选择训练样点与基于土壤类型面积比选择训练样点的差异。

本研究按地形地貌采样获取92个野外独立验证样点（图1），旨在检验利用不同训练样点所生成的土壤类型图是否可以很好地表达土壤信息的空间变化。地形地貌采样策略是在横穿山坡、沟谷的线路上设计样点，使所布设的样点能在较短距离内穿越主要的景观类型变化[30]；验证样点中各土壤类型的分布情况为土壤类型面积越大则该类型对应的验证样点数目越多，比如41号土壤类型因其面积最小仅有2个验证样点，而503号土壤类型因其面积最大所以有13个验证样点。

精度验证包括两个验证指标，第一个是平均推理制图精度，本文为检验每种训练样点选择方法的稳定性而分别进行了500次重复采样，计算推理制图精度时首先通过对比验证样点的实际土壤类型

与推理土壤图的土壤类型获得单次推理制图精度，再求取500次推理制图精度的平均值获得平均推理制图精度；第二个是更新传统土壤图的比例，即在500次重复实验中，更新后的土壤图精度高于原始传统土壤图精度的次数与总实验次数的比值。本文利用这两个验证指标评价三种训练样点选择方法的优劣及对研究区的适用性。

2　案例研究

2.1　研究区概况与传统土壤图

研究区位于美国威斯康星州La Crosse县的Raffelson流域，面积约4 km^2。该区位于威斯康星州无冰渍作用的边缘地区，未直接受到更新世大陆冰川的影响，该流域是明显的山脊–沟谷地形，即具有相对平缓的、狭窄的山脊与相对宽平的沟谷。研究区的高程由254m变化至416m（图1），坡度由0°变化至60°。土地利用类型主要是耕地和林地，其中还有少部分林地被改造为牧场。耕地作物主要为玉米、小粒谷类作物和紫花苜蓿等；林地作物主要为南方落叶林、橡树、山胡桃树、枫树和椴木等。

研究区的传统土壤图（图2）是由美国农业部制作[31-32]，包含12种土壤类型（土

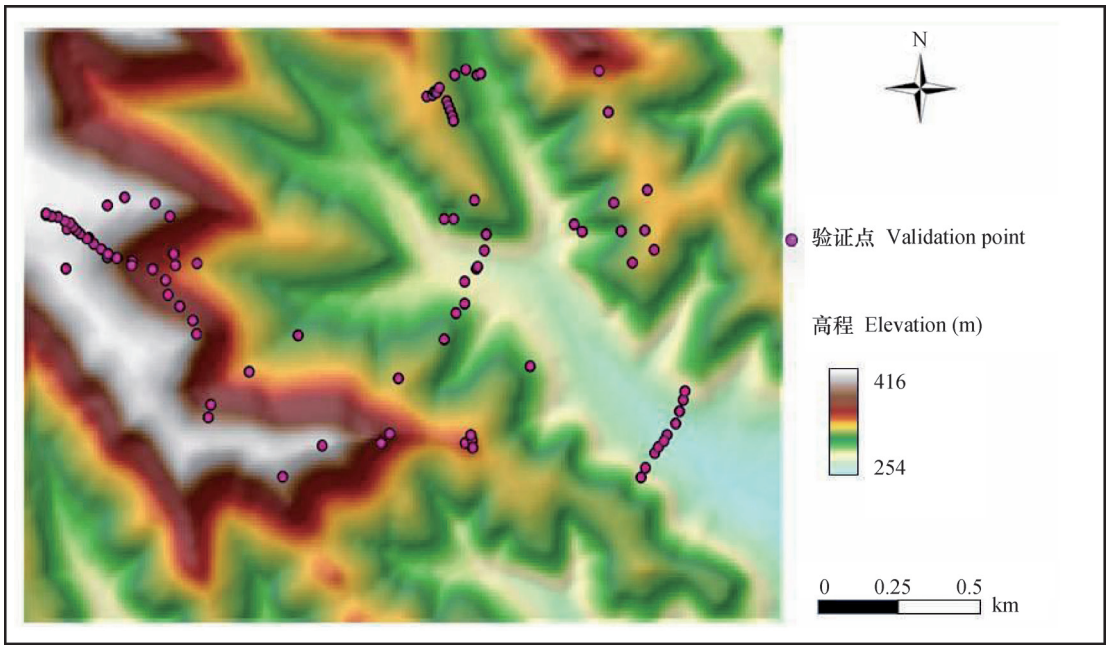


图1　Raffelson流域数字高程模型（DEM）

Fig. 1　DEM of the Raffelson watershed

系），其中Valton（1）与Lamoile（3）分布在坡顶；Dorerton–Elbaville（501）分布在坡肩；Churchtown（7）、Greenridge（21）与Gaphill–Rockbluff（502）主要分布在背坡；Norden（20）、Council（23）与Council–Elevasil–Norden（503）主要分布在坡脚；Kickapoo（30）、Orion（31）与Hixton（41）主要分布在沟谷。土壤类型501号、502号和503号为土壤复区，复区的存在是由于其内部两种或多种土壤类型的环境特征相似，且制图过程中土壤制图者无法在立体镜下较好地将其区分开所造成的。研究区土系的名称多以当地县/乡名为依据命名，将其进行中文译名意义不大，因而将土系的上一等级—亚类列出，其中，1号属残存湿淋溶土；3号、7号、20号、21号、23号、41号、501号、502号与503号属薄层干淋溶土；30号与31号属湿润冲积新成土。

2.2　环境因子数据

利用八个环境因子（表1）来刻画研究区的地理环境[5, 32-33]，其分辨率均为10m。本文对美国地质调查局（USGS）提供的分辨率为10 m的地形图数字化，生成10 m分辨率的数字高程模型（Digital elevation model，DEM），在SimDTA[34]中基于DEM派生出坡向、坡度、平面曲率、剖面曲率和地形湿度指数；通过数字化当地的地质图获

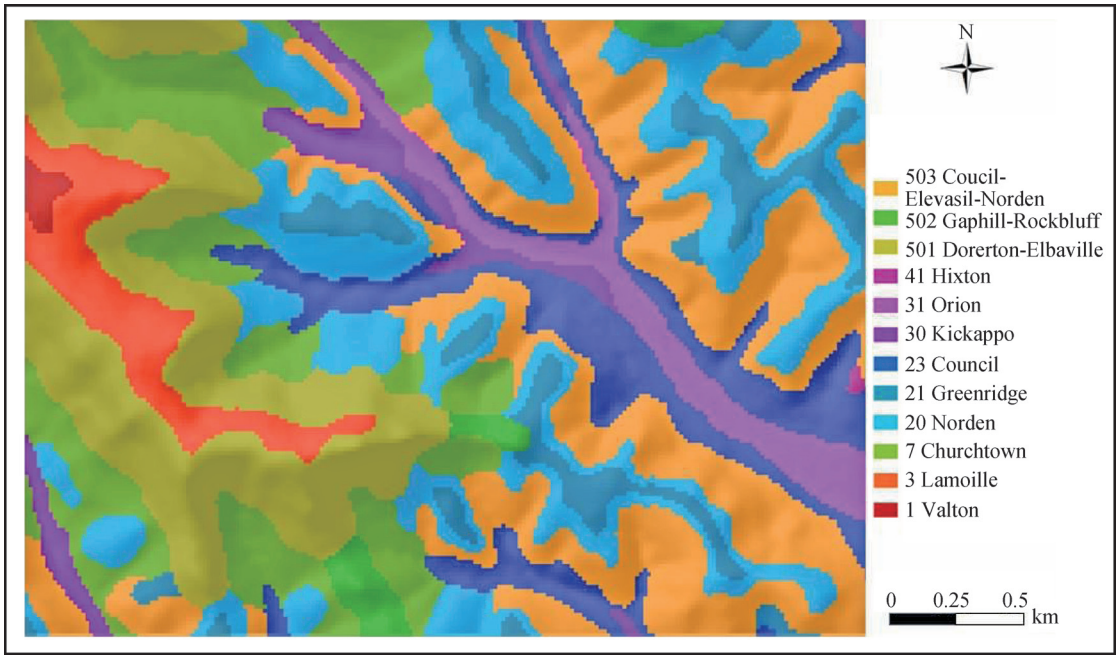


图2 Raffelson流域传统土壤图

Fig. 2 Conventional soil map of the Raffelson watershed

表1 影响Raffelson流域成土过程的环境因子

Table 1 Environmental factors affectingsoil development in the Raffelson watershed

环境因子 Environmental parameters		土壤学意义 Significance for pedology	数据源 Data resource
地形属性 Topographic attribute	高程	气候、水文、植被分布等影响土壤发育的因素均与高程密切相关	地形图
	坡向	重要的局部地形属性，影响地表水流的排聚能力、土壤侵蚀的强度、生物分布与土壤属性分布	
	坡度		
	平面曲率		
	剖面曲率		
	地形湿度指数	利用水流方向和斜坡动力来定量表达土壤排水与蓄水的综合状况	
母质条件 Material attribute	母质	影响成土过程的重要因素	母质图
	冲积物成分百分比	某点所接受的来自某种母岩类型分布区的冲积物的百分含量，本文采用该点上游各母岩汇流面积的百分比来近似表示	

得母质信息，在ArcInfo中基于母质信息派生出冲积物成分百分比，以描述地形地貌条件。

2.3 各土壤类型的典型点集

获取研究区传统土壤图中12种土壤类型的典型点集，以土壤类型Valton为例，它的高程环境因子直方图如图3所示，该土壤类型的像元数目为202个，根据式（1），本文设定r_1=5（即土壤类型Valton的训练样点数目为5个），那么它的土壤—环境直方图的区间数量为41，可以确定每个土壤—环境直方图中峰值区间内所对应的像元，汇总所选的像元即得到土壤类型Valton的典型点集。以此类推，根据每种土壤类型的训练样点数目分别设定相应的r_i来获取各自的典型点集，所得各土壤类型的典型点数目如表2所示。

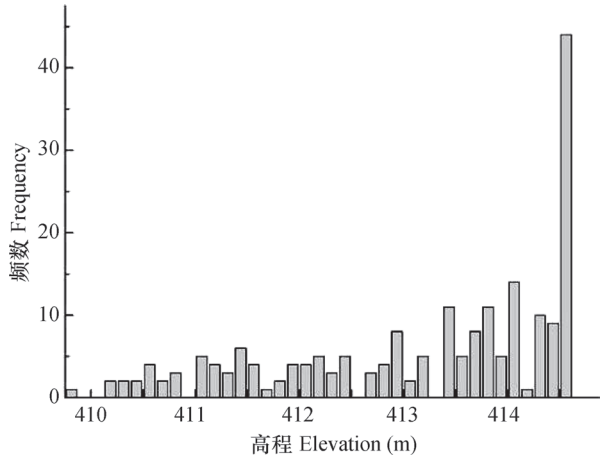


图3　Raffelson流域基于Valton土壤类型的高程环境因子直方图

Fig. 3　Histogram of the environmental factor of elevation in Raffelson watershed based on the soil type of Valton

2.4　训练样点的选择

传统土壤图中各土壤类型的像元数目及土壤面积取对数的比值数如表2所示，为满足最小等级Ⅰ所对应的比值至少为1，R的取值为4。根据式（2）将12种土壤类型划分为三个等级Ⅰ、Ⅱ、Ⅲ，其中等级I的比值数为1，包含2种土壤类型，ID为1号和41号；等级Ⅱ的比值数为4，包含5种土壤类型，ID分别为3号、21号、30号、31号和502号；等级Ⅲ的比值数为9，包含5种土壤类型，ID分别为7号、20号、23号、501号和503号。等级Ⅰ、Ⅱ、Ⅲ的训练样点数目的比例关系为1∶4∶9。本文设定等级Ⅰ中土壤类型的训练样点数目为5个，依照比例关系，等级Ⅱ与等级Ⅲ中土壤类型的训练样点数目分别为20个和45个。基于每种土壤类型的典型点集选择训练样点，最终在研究区利用方法A选择335个训练样点。

表2　基于传统土壤图所选训练样点在每种土壤类型中的分布信息及相关参数

Table 2　Distribution information and related parameters contained in the training samples selected for each soil type based on conventional soil maps during the process of generating training samples

土壤类型ID ID of soil type	土壤类型的像元数目 Number of pixels of the soil type	土壤类型的典型点数目 Numbers of typical samples of the soil type	面积比值 Proportion in area	土壤类型所在等级 Grade of the soil type
1	202	153	2.23	Ⅰ
3	1 780	833	1.29	Ⅱ
7	4 839	3 938	0.85	Ⅲ
20	5 685	4 627	0.78	Ⅲ
21	2 060	1 870	1.22	Ⅱ
23	3 481	2 571	0.99	Ⅲ
30	1 177	698	1.47	Ⅱ
31	1 439	829	1.38	Ⅱ
41	33	26	3.00	Ⅰ
501	5 388	5 345	0.80	Ⅲ
502	706	630	1.69	Ⅱ
503	7 970	6 734	0.63	Ⅲ

三种方法（A、B、C）的训练样点中各土壤类型的样点数目如表3所示，方法C中每种土壤类型的样点数目均不相同，而且土壤类型面积相差较大的41号与503号的训练样点数目相差75个；与之相比，方法A与方法B中不同土壤类型的训练样点数目之间的差异不大。

2.5　推理制图及精度验证

为更加准确地探究不同训练样点的选择方法对土壤图更新的影响，三种方法（A、B、C）分别重复采样500次，即对每种方法利用所选训练样点和模型最优参数，基于随机森林模型的推理制图重复实验500次。

采用92个野外独立验证样点检验传统土壤图

表3 三种训练样点选择方法所包含的样点数目

Table 3 Number of training samples relative to training sample selection method

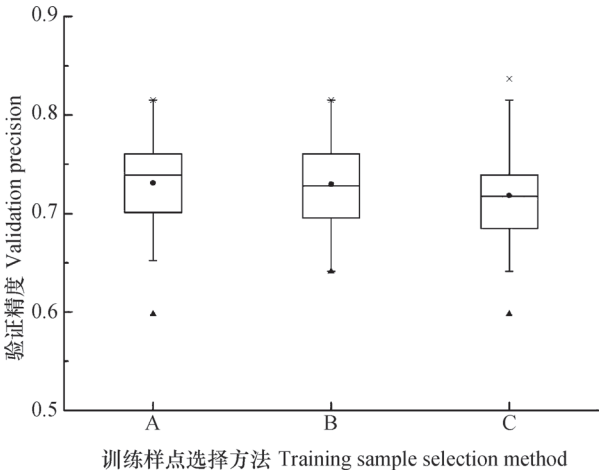
土壤类型ID ID of soil type	不同选择方法的训练样点个数 Number of training samples		
	方法A Method A	方法B Method B	方法C Method C
1	5	5	2
3	20	20	17
7	45	45	47
20	45	45	55
21	20	20	20
23	45	45	34
30	20	20	11
31	20	20	14
41	5	5	1
501	45	45	52
502	20	20	6
503	45	45	76

精度、基于不同训练样点选择方法的推理土壤图平均制图精度与更新传统土壤图比例。其中，传统土壤图的精度为69.6%，三种训练样点选择方法500次重复实验后的平均制图精度与更新传统土壤图比例如表4和图4所示，三种方法的平均制图精度均高于71.5%，更新传统土壤图比例均高于63.6%，表明三种方法均基本可以达到更新土壤图的目的。具体来看，方法C验证精度的平均值为71.5%，明显低于其他两种训练样点选择方法，而且其精度的稳定性较差，500次重复实验中能够更新传统土壤图的比例是最低的，制图精度多数集中在箱线图的中值以下，说明方法C在更新土壤图中结果不稳

表4 三种不同训练样点选择方法的验证精度及更新比例

Table 4 The validation precision and updating ratio of three training sample selection methods

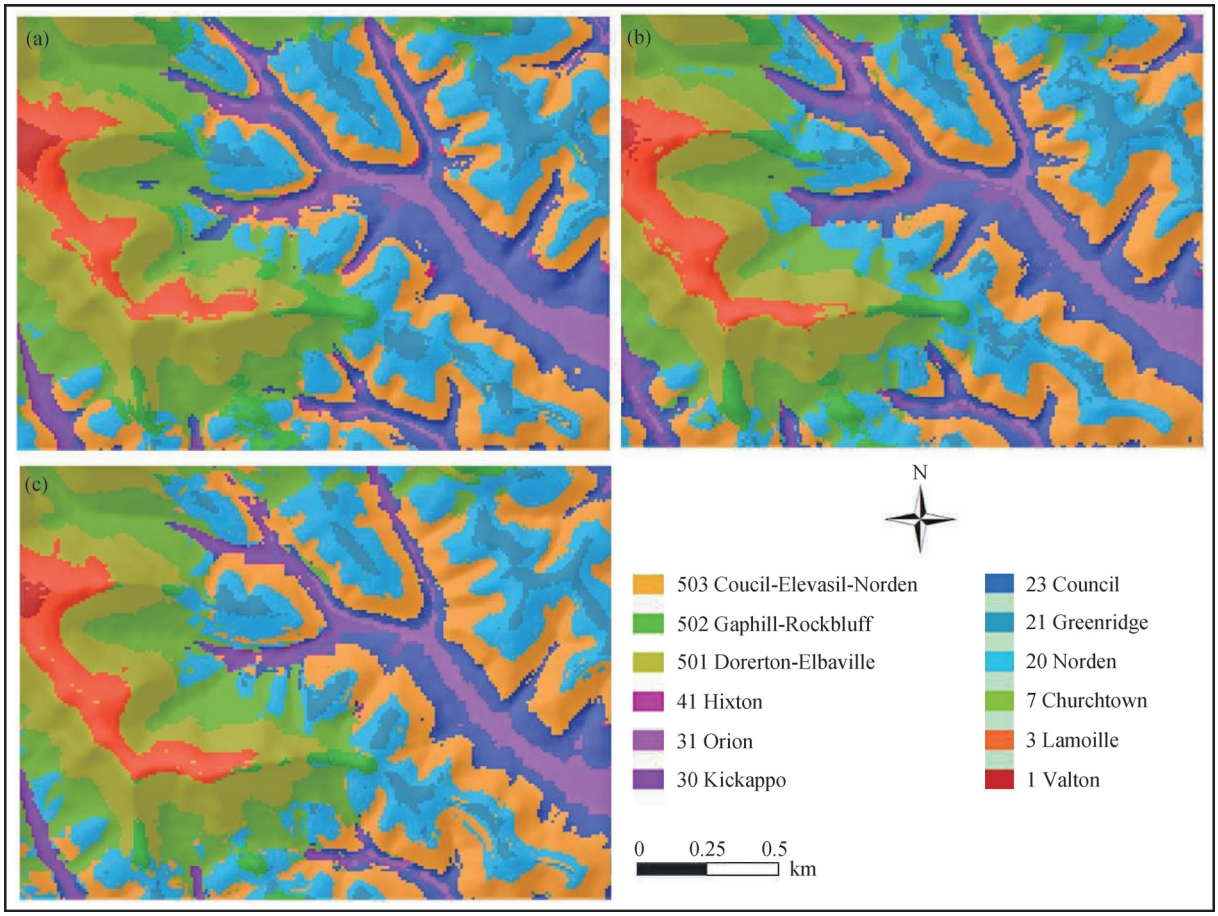
训练样点集编号 Serial number of training samples	平均制图精度 Mean mapping precision（%）	传统土壤图更新比例 Updating ratio of conventional soil map（%）
A	73.2	79.5
B	72.6	71.8
C	71.5	63.6



注：A：土壤类型分级后按比例选择典型点构成训练样点；B：土壤类型分级后按比例选择样点构成训练样点；C：按照土壤类型斑块面积比选择样点构成训练样点 Note：A：Training sample set consisting of the typical samples selected according to the proportion of the soil type in grading；B：Training sample set consisting of samples selected according to the proportion of the soil type in grading；C：Area-weighted sampling approach generate the training sample set consisting of samples selected according to the proportion of the soil polygons of the soil type in area

图4 三种不同训练样点选择方法的验证精度

Fig. 4 Validation of precision of the three training sample selection methods



注：（a）方法A、（b）方法B、（c）方法C所得推理土壤图 Note：（a）Method A；（b）Method B；（c）Method C

图5　Raffelson流域分别应用三种训练样点选择方法

Fig. 5　Inferential soil maps of the Raffelson watershed based on three training sample selection methods

定，更新效果较差。与之相比，方法A和方法B的结果均表明，在土壤图更新方面，对土壤类型分级按比例选择训练样点的方法较完全依照土壤类型的面积比选择训练样点的方法能够更加稳定地提高传统土壤图精度。但是，在对土壤类型分级的前提下，方法A与方法B验证精度的平均值分别为73.2%和72.6%，而且应用方法A，79.5%的推理土壤图可以达到土壤图更新的效果，与之相比方法B只有71.8%的推理土壤图可以达到土壤图更新的效果，此外A的箱线图中制图精度大多分布在中值以上，B的箱线图中制图精度较多分布在中值以下，这表明基于典型点选择训练样点的方式可以提高训练样点的质量，从而得到更加稳定的土壤图更新结果。

对比三种训练样点选择方法用于制图的精度均为81.5%的推理土壤图，如图5、图6和图7所示，它们与传统土壤图相比均表现了更详细的空间分布，具有更高的精度。但是，应用方法C的推理土壤图中训练样点数目多的土壤类型503号和20号，它们的面积明显被夸大，然而在沟谷地带训练样点少的土壤类型41号和30号，它们的面积则被压缩。此外，对比方法A与方法B所得推理土壤图发现，方法B的推理土壤图中存在较多的破碎化斑块，尤其是在土壤—环境关系较复杂的背坡和沟谷地带，其破碎化现象更加明显。尽管应用三种训练样点选择方法的制图精度均相同，但方法A的推理制图结果更符合研究区土壤分布的特征。

3　讨　论

本文所提方法的适宜性可能会受到以下5个因素的影响：（1）传统土壤图的质量。训练样点的选择基础是传统土壤图，因此传统土壤图的质量会对训练样点的质量产生很大影响，可尝试使用不同精度的传统土壤图来检验该方法的有效性；

（2）研究区环境因子的完备性。针对不同研究区特征，选择完备的环境因子数据是确保所选训练样点能够准确表达该区土壤—环境关系的重要前提，如：针对我国南方平原地区，除应用传统的环境因子来刻画土壤的空间变化之外，还应添加可反映平区土壤空间分布信息的环境因子，例如，已有研究表明利用遥感数据生成的遥感动态反馈模式是一种适用于平区的环境因子[35]；（3）土壤类型图斑的形状。本研究更多的关注土壤类型面积之间的差异对训练样点选择的影响，在选择训练样点时还应考虑土壤类型的图斑形状这一因素，特别是针对成土条件复杂的地区，土壤类型图斑形状的差异也在一定程度上反映了环境因子的空间异质性，此外，有些特定的土壤类型尽管面积比例较小，但很有特色，针对这些情况，对土壤类型图斑形状的考虑显得更加必要；（4）训练样点数量。样点数量同样会影响本文所提方法的适宜性以及土壤图的更新效果，本文是限定了一个土壤类型最低等级的训练样点数量，未来可探索不同数量的训练样点对本文方法与更新结果的影响；（5）数据挖掘模型。不同的数据挖掘模型对训练样点中各土壤类型的分布与数量均有各自的要求，未来可尝试使用其他数据挖掘模型来检验方法的有效性与土壤图更新的效果。

需要指出的是，本文是在确定归属于最小等级的土壤类型需选取的训练样点数（本文中γ_1=5）的基础上所得结果。针对每种土壤类型选择不同数目的训练样点，可能会影响土壤图更新的结果与精度。对此，将在后续工作中进一步研究。

4 结 论

本研究提出了一种基于土壤类型面积分级的典型训练样点选择方法，通过与其他两种训练样点选择方法对比检验其有效性。结果表明：500次重复实验中，该方法在推理制图精度和传统土壤图被更新的比例方面均明显优于另外两种选择方法，同时本文所提方法的推理制图结果更符合研究区土壤分布的特征，此外，应用该方法推理制图，79.5%的推理土壤图可以达到土壤图更新的效果。因此，本研究针对数量较少的训练样点提供了一种有效的样点选择方法，可为其他数据挖掘模型中训练样点的选择提供参考。

参 考 文 献

[1] 土壤普查办公室. 中国土壤普查技术. 北京：农业出版社，1992
National Soil survey Office. Soil survey and technology of China（In Chinese）.Beijing：Agriculture Press，1992

[2] 潘剑君. 土壤调查与制图. 第3版. 北京：中国农业出版社，2010
Pan J J. Soil survey and mapping（In Chinese）. 3rd ed. Beijing：China Agriculture Press，2010

[3] Qi F. Knowledge discovery from area-class resource maps：Data preprocessing for noise reduction. Transactions in GIS，2004，8（3）：297—308

[4] 朱阿兴，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，et al. Model and method of detail digital soil survey（In Chinese）. Beijing：Science Press，2008

[5] 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

[6] 辛文文，刘建立. 数字土壤及其应用研究进展. 土壤，2013，45（5）：800—808
Xin W W，Liu J L. Advance in digital soil and its application（In Chinese）. Soils，2013，45（5）：800—808

[7] Hudson B D.The soil survey as paradigm-based science. Soil Science Society of America Journal，1992，56：836—841

[8] Grinand C，Arrouays D，Martin M P，et al. Extrapolating regional soil landscapes from an existing soil map：Sampling intensity，validation procedures，and integration of spatial context. Geoderma，2008，143（1）：180—190

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

[10] Kempen B，Brus D，StoorvogelJ，et al. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Science Society of America Journal，2012，76（6）：2097—2115

[11] Kempen B，Brus D J，Heuvelink G B M. Soil type mapping using the generalised linear geostatistical model：A case study in a Dutch cultivated peatland.

Geoderma，2012，189/190（6）：540—553

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

[13] Kempen B，Brus D J，Vries F D. Operationalizing digital soil mapping for nation wide updating of 1∶50，000 soil map of the Netherlands. Geoderma，2015，241/242：313—329

[14] McBratney A B，Mendonça Santos M L，Minasny B.On digital soil mapping.Geoderma，2003，117：3—52

[15] Heung B，Ho H C，Zhang J，et al. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma，2016，265：62—77

[16] 刘京，朱阿兴，张淑杰，等. 基于样点个体代表性的大尺度土壤属性制图方法. 土壤学报，2013，50（1）：12—20
Liu J，Zhu A X，Zhang S J，et al. Large-scaled soil attribute mapping method based on individual representativeness of sample sites（In Chinese）. Acta Pedologica Sinica，2013，50（1）：12—20

[17] Moran C J，Bui E N. Spatial data mining for enhanced soil map modeling. International Journal of Geographical Information Science，2002，16：533—549

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

[19] Heung B，Bulmer C E，Schmidt M G. Predictive soil parent material mapping at a regional-scale：A random forest approach. Geoderma，2014，214/215：141—154

[20] Galar M，Fernández A，Barrenechea E，et al. A review on ensembles for the class imbalance problem：Bagging-，boosting-，and hybrid-based approaches. IEEE Transactions on Systems Man& Cybernetics Part C，2011，42：463—484

[21] Van Hulse J，Khoshgoftaar TM，Napolitano A.Experimental perspectives on learning from imbalanced data. Proceedings of the 24th Annual International Conference on Machine Learning（ICML 2007），Corvalis，OR，2007：935—942

[22] 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（12）：97—106

[23] Stum A K，Boettinger J L，White M A，et al. Random forests applied as a soil spatial predictive model in arid Utah. Gradworks，2010，2：179—190

[24] Barthold F K，Wiesmeier M，Breuer L，et al. Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia. Journal of Arid Environments，2013，88（1）：194—205

[25] Breiman L. Randomforests. Machine Learning，2001，45（1）：5—32

[26] Gislason P O，Benediktsson J A，Sveinsson J R. Random forests for land cover classification. Pattern Recognition Letters，2006，27（4）：294—300

[27] Peters J，Baets B D，Verhoest N E C，et al. Random forests as a tool forecohydrological distribution modelling. Ecological Modelling，2007，207（2/4）：304—318

[28] Díaz-Uriarte R，Andrés S A D. Gene selection and classification of microarray data using random forest. BMC Bioinformatics，2006，7（1）：1—13

[29] 李亭，田原，邬伦，等. 基于随机森林方法的滑坡灾害危险性区划. 地理与地理信息科学，2014，30（6）：25—30
Li T，Tian Y，Wu L，et al. Landslide susceptibility mapping using random forest（In Chinese）. Geography and Geo-Information Science，2014，30（6）：25—30

[30] 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 from soil mapping using fuzzy c-means（FCM）clustering（In Chinese）. Acta Pedologica Sinica，2007，44（5）：784—791

[31] Zhu A X，Hudson B，Burt J，et al. Soil mapping using GIS，expert knowledge，and fuzzy logic. Soil Science Society of America Journal，2001，65（5）：1463—1472

[32] Qi F，Zhu A X，Harrower M，et al. Fuzzy soil mapping based on prototype category theory. Geoderma，2006，136（3/4）：774—787

[33] Qi F，Zhu A X，Pei T，et al. Knowledge discovery from area-class resource maps：Capturing prototype effects. Cartography & Geographic Information Science，2008，35（4）：223—237

[34] 秦承志，卢岩君，包黎莉，等. 简化数字地形分析软件（SimDTA）及其应用——以嫩江流域鹤山农场区的坡位模糊分类为例. 地球信息科学，2009，11（6）：737—743
Qin C Z，Lu Y J，Bao L L，et al. Simple digital terrain analysis software（SimDTA 1.0）and its application in Fuzzy classsification of slope positions（In Chinese）. Journal of Geo-information Science，2009，11（6）：737—743

[35] Zhu A X, Liu F, Li B L, et al.Differentiation of soil conditions over flat areas using land surface feedback dynamic patterns extracted from MODIS. Soil Science Society of America Journal, 2010, 74(3): 861—869

Training Sample Selection Method Based on Grading of Soil Types by Area for Updating Conventional Soil Maps

LIU Xueqi[1] ZHU A-Xing[1, 2, 3, 4] YANG Lin[2†] MIAO Yamin[1] ZENG Canying[1]

(1 *School of Geographical Science, Nanjing Normal University, Nanjing* 210023, *China*)

(2 *State Key Laboratory of Environment and Resources Information System, Institute of Geographical Sciences and Resources Research, Chinese Academy of Sciences, Beijing* 100101, *China*)

(3 *Key Laboratory of Virtual Geographic Environment* (*Nanjing Normal University*), *Ministry of Education*; *State Key Laboratory Cultivation Base of Geographical Environment Evolution* (*Jiangsu Province*); *Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing* 210023, *China*)

(4 *Department of Geography, University of Wisconsin-Madison, Madison, WI* 53706, *USA*)

Abstract 【Objective】Traditional soil surveyshave turned out huge piles of conventional soil maps various in scale and nature. Although these maps are not very high in spatial detail or accuracy, they contain large volumes of valuable expertise concerning soil-environment relationships in relevant regions. Data mining models can be used to extract from these maps information useful to updating of the conventional soil maps. In using data mining models to extract the information of soil spatial distribution, selection of training samples is an essential step. Quality of training samples will affect to a great extent full expression of soil-environmental relationships and accuracy of the updatedsoil maps. The area-weighted proportion method was a common method for selecting of training samples. However, this method usually assigns too much weight to those soil types large in area, so that too many training samples would be selected. Meanwhile, random selection of training samples from polygons of the same soil type may bring in some "noise" samples, occurring on transition areas between soil types, which make the accuracy of the updated soil maps not high.【Method】 In this paper, a new method was developed to select training samples from conventional soil maps based on grading of soil types by area. The method consists of the following two steps. The first step is to specify typical (representative) samples of each soil type based on conventional soil map, so as to avoid generation of "noise pixels" due to misplacement in delineating boundaries between soil polygons. It is assumed that most of the boundaries of the soil polygons of a certain soil type are correctly delineated, and then the peak of the histogram of a certain environmental factor enclosed in the polygons of the soil type represents the typical environmental conditionunder which the soil develops or exists. The pixels close to the selected environmental conditions or within the peak zone of the histogram are considered as representative samples. All the representative samples selected through histograms of various environmental conditions of a certain soil type are combined into a typical sample set of the soil type. The second step is to select training samples based on grading of soil type by area, with a view to keep the numbers of samples of each soil type in balance. Soil types in the same grade should have the same number of training samples out of the typical sample set of each of the soil types.The random forest model adopted in this study is to update conventional soil maps based on the selected training samples. To evaluate the above-proposed method, comparison was made between this method and two other training sample selection methods. One is to randomly select training samples from polygons of each soil type and the number of training samples for each soil type depended on proportion of the grade

the soil type is in, while the other is the common area-weighed proportion method, which randomly selects training samples form the soil polygons of a soil type and the number of training samples for each soil type depended on the area-weighted proportion of the soil type. The study area was a small watershed in Raffelson, Wisconsin of USA. The three selection methods were tried repeatedly, each for 500 times, and validate mean precision of the inferential mapping and proportion of the updated conventional soil maps with 92 independent verification samples in the field.【Result】Results show that based on the 500 trails, comparison of this method with the other two reveals that about 79.5%, 71.8% and 63.6% of the conventional soil maps could be updated, respectively. Meanwhile, the updated soil maps based on the proposed training sample selection method are more consistent with the actual soil distribution in the Raffelson watershed.【Conclusion】It is concluded that the proposed method is an effective training sample selection method for data mining model to update conventional soil maps.

Key words　Training sample; Data mining model; Update conventional soil map; Soil-environmental relationships

（责任编辑：檀满枝）