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*

杨　琳[1,2,3]　朱阿兴[1]　李宝林[1]　秦承志[1]　裴　韬[1]　刘宝元[2]
李润奎[1,3]　蔡强国[4]

(1 中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室,北京　100101)
(2 北京师范大学地理学与遥感科学学院,北京　100875)
(3 中国科学院研究生院,北京　100049)
(4 中国科学院地理科学与资源研究所陆地水循环与地表过程重点实验室,北京　100101)

摘　要　在没有土壤普查专家及土壤图的地区,获取土壤环境间关系的知识是基于知识进行预测性土壤制图中的关键问题。本文建立了一套应用模糊 *c* 均值聚类(Fuzzy *c*-means, FCM)获取土壤环境间关系知识的方法:得到对土壤形成发展具有重要作用的环境因子,建立环境因子数据库;对环境因子进行模糊聚类,得到环境因子组合隶属度分布图;根据隶属度值确定野外采样点;将环境因子组合与土壤类型对应,进而提取土壤 - 环境关系知识。为检验该方法的有效性,应用所得知识进行土壤制图,通过独立采样点对土壤图进行精度评价。本文在黑龙江鹤山农场一个研究区的应用结果表明,该方法仅需要少量的野外采样即可获得有效的土壤 - 环境关系知识,为预测性土壤制图提供必需的依据,同时也显著提高了野外采样的效率。

关键词　预测性土壤制图;土壤 - 环境关系知识;模糊 *c* 均值聚类方法(FCM);环境因子组合;土壤 - 环境推理模型(SoLIM)

中图分类号　P934　　**文献标识码**　A

现代遥感测量技术、地理信息系统(GIS)、数字地形模拟、人工智能及模糊推理等方法与技术的发展下,基于 GIS 的预测土壤制图应运而生[1,2]。预测性土壤制图以道库恰也夫和詹尼建立的成土因素学说[3,4]和土壤景观模型理论[5]为基础,将环境变量与土壤信息之间的关系或统计模型应用于地理空间数据库中进行推理,依据环境因子的空间分布来推测土壤的空间分布,从而生成土壤图。

目前预测性土壤制图方法主要包括两类:基于统计模型的方法[6~9]和基于专家知识的方法[2,10~14]。基于统计模型的方法多采用以回归为基础的统计方法或地统计模型的方法进行制图,由于土壤的形成是多要素制约的非线性过程并且土壤本身具有非常大的变异性,使得这类方法的应用受到制约[14],且这种方法也主要用于小范围内的土壤属性制图。基于专家知识的方法通过提取土壤与环境间关系的知识(也称为土壤景观模型),与环境信息相结合进行土壤推理制图。由于这类方法可以知识的形式描述复杂的土壤与环境间的关系,利于知识的积累和移植,具有广阔的应用前景,并得到越来越广泛的应用[1,14,15]。

土壤 - 环境关系知识的获取是基于专家知识的方法进行土壤制图的关键问题。目前获取土壤 - 环境关系知识主要有以下两种方式:(1)由土壤普查专家提供土壤 - 环境关系知识[11,13];(2)通过数据挖掘方法从已有数据中提取知识[12]。第一种方法获得的知识是土壤普查专家通过大量野外调查形成的经验性的知识,这种知识的形成需要长期经验知识的积累。然而,很多国家(包括中国)都缺乏专门进行野外调查的土壤普查专家,因此这种方法推广起来困难很大。第二种方法即数据挖掘的方法,如神经网络、决策树等,主要从传统土壤图或大量野外样

*中国科学院"百人计划"项目、中国科学院创新团队国际合作伙伴计划"人类活动与生态系统变化"、国家自然科学基金项目(40501056)、中国科学院地理科学与资源研究所领域前沿项目资助、国际重点基础研究发展计划(973)课题多尺度土壤侵蚀预报模型(2007CB407204)
作者简介:杨　琳(1982~),女,山东文登人,博士研究生,主要从事数字土壤制图研究。E-mail: yanglin@lreis.ac.cn
收稿日期:2006-08-01;收到修改稿日期:2007-03-28

点数据中获取土壤 - 环境关系知识,显然这种方法不适用于没有土壤图或野外样点数据的地区。

目前,在没有专门进行野外调查的土壤普查专家的情况下和没有传统土壤图或大量样点数据的地区,野外调查采样是获取土壤环境知识的主要手段。然而,野外采样通常需要大量的人力、物力、财力,没有经验的调查者不一定可以有效获取建立土壤景观模型所需的知识。因此,在这些地区如何获取知识为土壤制图提供必需的支持成为一个亟待解决的问题。本文以高分辨率的数字高程模型(Digital elevation model, DEM)为基础,通过模糊 c 均值聚类(Fuzzy c-means, FCM)方法对重要成土环境因子进行聚类分析,根据聚类结果进行有效的野外调查来建立土壤和环境之间的关系,以期为土壤制图提供必需的知识。

1　研究方法

根据土壤景观模型的理论[5],特定的环境因子组合下形成特定的土壤类型。基于此,我们假设特定土壤类型的空间位置可以由特定环境因子组合的位置来近似代替。因此,寻求土壤与环境之间的关系转变成寻找何种特定的环境因子组合对应于何种特定的土壤类型。

环境因子组合的典型位置为环境因子聚类中心所在的位置。寻找这些位置可通过模糊聚类的方法来实现。通过对环境因子进行模糊聚类可以得到各要素多维空间上的自然组合(即环境因子组合)的模糊隶属度分布图。根据隶属度图可以判断类别典型区(隶属度高值区),通过野外采样获取这些典型区的土壤信息,进而获取土壤环境关系知识。据此,本研究方法共包括以下四步:

1.1　建立环境因子数据库

在景观尺度下,决定土壤变化的主要因子是地形和水文状况。McSweeney 等研究认为在小流域尺度内,高程、坡度、沿等高线曲率、沿剖面曲率及湿度指数(Wetness index)这五个因子基本上代表了土壤形成与发展的主要影响因素[16]。对于不同的研究区,影响土壤形成发展最主要的环境因子可能会有所不同,因此在实际应用中应根据当地的情况选择不同的环境因子,建立研究区环境因子数据库。

1.2　模糊聚类获取环境因子组合

FCM 是一种应用较为广泛且有效的非监督模糊聚类方法,其基本原理是利用统计方法计算每个样本与每类原型在多属性空间中的距离,以隶属度进行加权,最终达到类内加权误差平方和目标函数最小化[17]。模糊目标函数可表达为:

$$J_m(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d_{ik}^2 \tag{1}$$

$$d_{ik}^2 = \| y_k - v_i \|_A^2 = (y_k - v_i)^{\tau} A (y_k - v_i)$$

$$y \in Y \tag{2}$$

其中,U 为模糊聚类的隶属度矩阵,v 为聚类中心集,Y 为环境要素数据集,n 为 Y 中数据的个数,c 为聚类类别的个数,m 为加权指数(亦称模糊度指标),d_{ik}为数据点 y_k到中心 v_i的加权距离,u_{ik}为第 k 个数据点属于第 i 类的隶属度,A 为距离权重矩阵,J_m为模糊分类误差。

FCM 用连续划分的模糊隶属度代替了传统模型中的二值分类,可以更好地表征连续地理现象的空间分布。以影响土壤形成的各环境因子为变量采用 FCM 聚类分析,可以得到各环境因子组合在属性空间的聚类中心和每一点环境条件与这些聚类中心环境条件的隶属度,隶属度高值区即为这种环境条件组合的典型区。

最佳类别数 c 和加权指数 m 是应用 FCM 的两个关键参数,这里采用划分系数(Partition coefficient, F)和标准化分类熵(Normalized entropy, H)来确定最优类别数 c[17]:

$$F_c(\hat{u}) = \sum_{k=1}^{n}\sum_{i=1}^{c}(\hat{u}_{ik})^2 / n \tag{3}$$

$$H_c(\hat{u}) = -\sum_{k=1}^{n}\sum_{i=1}^{c}(\hat{u}_{ik}\log_a(\hat{u}_{ik})) / n \tag{4}$$

F 度量不同类别间的重叠度,与聚类平均重叠度成反比。H 为聚类的模糊度测量。一般随着聚类类别数目增加,H 增大,F 减小。当类数由(c_1 - 1)变化到 c_1时,标准化分类熵变化(增量)(H_{c_1} - H_{c_1-1})较(c_1 - 2)到(c_1 - 1)的标准化分类熵变化(H_{c_1-1} - H_{c_1-2})和由 c_1到(c_1 + 1)的标准化分类熵变化(H_{c_1+1} - H_{c_1})小,即标准化分类熵值增大较小;同时,划分系数变化较前后两类变化小,可认为此时的聚类结果较稳定,c_1为可能的最佳聚类类别数。对应不同的 m 可得到多个最佳聚类类别数,其中出现频率最高的 c 值即认为是最佳聚类数。

加权指数 m 控制着聚类的模糊性。m 越接近于 1,聚类越趋向于突变(crisp),m 越大,结果越模糊、相对更易于反映空间的渐变性,但过大的 m 值将导致类别间的重叠太多,聚类结构不清晰,因此对

m 的选取需要在模糊度与清晰的聚类结构间进行权衡。一般 m 的有效值在 1～30 之间,部分文献根据实验建议最佳权重指数位于区间 1.5～2.5 之间[18,19]。在研究中可选择 1.5～2.5 之间多个 m 对所得相同聚类类别数的结果进行比较,以选择合适的 m。

1.3 野外采样解释环境因子组合

选择恰当的参数运行 FCM,可得到环境因子组合的自然聚类结果,即相对于每一环境因子组合聚类中心的隶属度分布。根据隶属度可判别典型的环境因子组合中心点以及环境因子组合的过渡区。隶属度值高的点为环境因子组合中心点,而隶属度值由高到低的部分为环境因子组合的过渡区。在环境因子组合的中心点进行野外采样,由土壤分类专家鉴定土壤类型,将环境因子组合与土壤类型对应起来,这些野外采样点称为解释集。

1.4 提取土壤 - 环境关系知识

在环境因子数据库中提取解释集的环境数据,得到研究区土壤类型的典型发生环境条件。通过环境因子组合的隶属度空间分布与野外调查,可以得到各土壤类型的空间分布规律,进而得到各土壤类型间随环境条件的过渡变化情况。在获得土壤类型的典型环境条件以及各土壤类型间的过渡变化情况之后,即可认为得到了研究区的土壤 - 环境间关系,土壤景观模型得以建立。

2 研究区概况及环境因子数据

研究区位于黑龙江省黑河市嫩江县鹤山农场老莱河左岸(图 1),位置为 48°53′24″～48°59′24″N,125°8′24″～125°16′12″E,面积为 60.2 km^2。本区处于寒温带季风草甸草原区,年降水量400～550 mm。该区的地貌主要是不同程度切割的山前洪积台地以及冲积湖积平原,当地称之为漫岗地,海拔高度为 265.8～365.7 m,地形较缓,坡度多在 4°以下,母质多为黄土状亚粘土。原生植被为疏林草甸、灌丛草甸和杂类草草甸,但近四十年多被开垦为农田,主要农作物为小麦和大豆等。

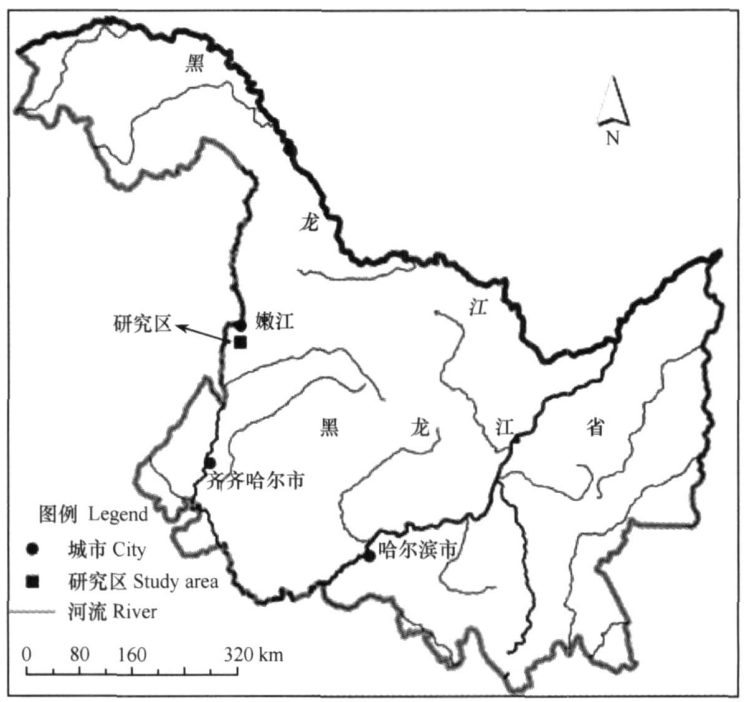

图 1 研究区示意图

Fig. 1 Location map of study area

由于研究区的相对高差小于 100 m,高程对土壤分布差异的影响很小,研究区母质基本一致,因此选择坡度、沿等高线曲率、沿剖面曲率及地形湿度指数四个因子作为影响该地区成土的环境因子数据。环境因子根据分辨率为 10m ×10m 的 DEM 计算得到。DEM 是用研究区 1∶10 000 地形图等高线在 Arc/ Info

中采用 TopoGrid 和 TINLATTICE 相结合的方法生成。

3　结果分析

3.1　环境因子组合聚类结果

对环境因子数据采用 5 种不同的加权指数(m = 1.5,1.75,2.0,2.25 和 2.5)进行 FCM 运算。根据前面所述参数选取原则,选择 m 为 2.0 时 c = 13为最佳类别数,获得各类别的聚类中心环境参数值和模糊隶属度分布图。模糊隶属度图示例见图 2。

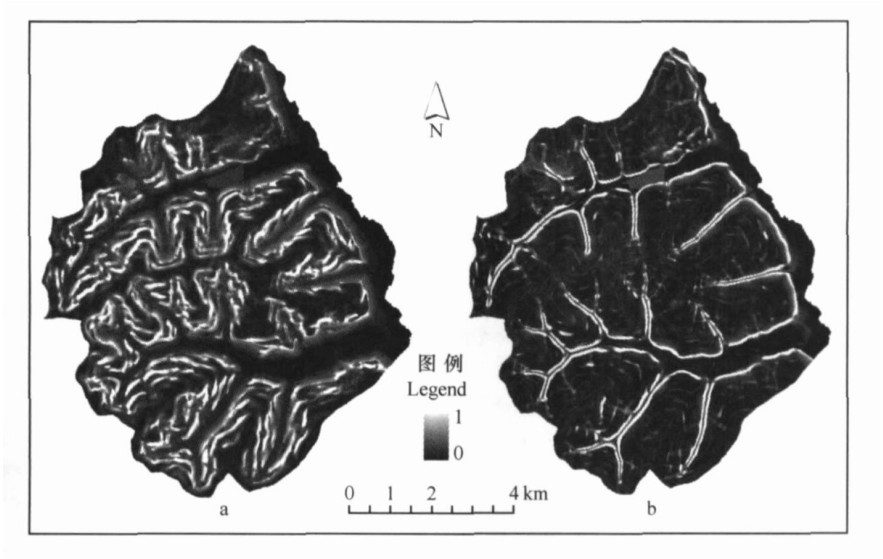

图 2　模糊隶属度分布图示例(a:class 10; b:class 13)

Fig. 2　Fuzzy membership maps (a: class 10; b: class13)

3.2　土壤 - 环境关系知识

根据环境组合的模糊隶属度图选择解释集,并在野外采集样点。原则上选择图中隶属度大于 0.85 的斑块中心点作为环境组合的典型位置,同时避开林带、道路、居民点及水域。对于每一样点,通过 GPS 在野外找到该点;然后,挖取剖面,进行分层并取样,将土样运回实验室进行理化分析;最后,由土壤分类专家结合野外剖面判断以及室内理化分析数据判断样点的土壤类型。目前,中国土壤系统分类取得了很大的发展[20],已应用在土壤制图中。在本实例中,选择中国土壤系统分类作为土壤分类系统,亚类作为基本单元,共获取 21 个典型的野外样点。

根据野外解释集采样将 FCM 获得的环境因子组合与土壤类型相对应,将对应相同土壤类型的 FCM 聚类类型进行合并,可概括出区域内各土壤类型发育的环境条件(见表 1)。

暗沃冷凉湿润雏形土主要发育在台地的顶部,坡度很小,地形平坦,发育了较厚的暗沃表层,约 30～40 cm,暗沃表层与其下相临土层水平过渡明显。普通简育湿润均腐土主要发育在两种环境条件下:(1)坡肩,背坡坡度较小,曲率微凸、平直或较凹的地形部位,其暗沃层厚度 25～50 cm 不等;(2)坡脚沿剖面曲率凹的地形部位,堆积作用强烈,暗沃土层较第一种环境条件下土壤的暗沃土层厚,一般可达 80～100 cm。普通冷凉湿润雏形土发育在背坡坡度较大的地形部位,曲率呈平直或较凸,在侵蚀作用下发育了较薄的暗沃表层,一般在 20 cm 左右。石质湿润正常新成土发育在背坡最底端的地形部位,坡度在整个研究区中最陡,大约在 3°以上。侵蚀相对较强烈,暗沃表层只有 10 cm 左右,有些陡的部位几乎没有暗沃表层。暗厚滞水湿润均腐土、普通简育滞水潜育土和纤维滞水潜育土是发育在河谷平原中的土壤。河谷平原地形平坦,湿度指数大,土壤发育受水分影响大。由于距离汇水线的远近不同,土壤水分不同,在河谷平原的中央部位因不同的植被类型相间分布发育着土壤普通简育滞水潜育土和纤维滞水潜育土,而在水分条件稍差一点的河谷平原外侧发育土壤暗厚滞水湿润均腐土。

表 1 各土壤类型发生的环境条件描述表[1)]

Table 1 Exhaustive environment descriptions of soil types

土壤类型 Soil types	环境条件 Environmental combination				
	坡度 Slope (°)	沿剖面曲率 Profile curvature	沿等高线曲率 Planform curvature	湿度指数 Wetness index	坡位 Slope position
暗沃冷凉湿润雏形土 Mollic Bori-Udic Cambosols	0～1.1	微凹,直,凸 Sl concave, lin, convex	凹,凸 Concave, convex	低,较低 Low, moderately low	山顶 Summit
普通简育湿润均腐土 Typic Hapli-Udic Isohumosols	1.1～2.4	微凹,直,微凸 Sl concave, lin, sl convex	直,微凸 Lin, sl convex	较低 Moderately low	背坡 Backslope
	1.65～2.65	凹,微凹 Convave, sl concave	微凹,直 Sl concave, lin,	较高 Moderately high	坡脚 Footslope
普通冷凉湿润雏形土 Typic Bori-Udic Cambosols	2.2～3.05	微凹,直,微凸 Sl concave, lin, sl convex	直,微凸 Lin, sl convex	较低 Moderately low	背坡 Backslope
石质湿润正常新成土 Lithic Udi-Orthic Primosols	>3.05	微凹,直,微凸 Sl concave, lin, sl convex	直,微凸 Lin, sl convex	低 Low	坡上最陡的位置 The steepest slope
暗厚滞水湿润均腐土 Pachic Stagni-Udic Isohumosols	0.6～1.65	直,微凹 Lin, sl concave	微凹,直,微凸 Sl concave, lin, sl convex	较高 Moderately high	河谷平原的边缘 Border of flood-plain
普通简育滞水潜育土,纤维滞水潜育土 Fibric Histic-Stagnic Gleyosols, Typic Haplic-Stagnic Gleyosols	<0.6	直 Lin	直 Lin	高 High	河谷平原 Floodplain

1) sl: slightly; lin: linear

综上所述,所有土壤类型的空间分布序列如图 3。

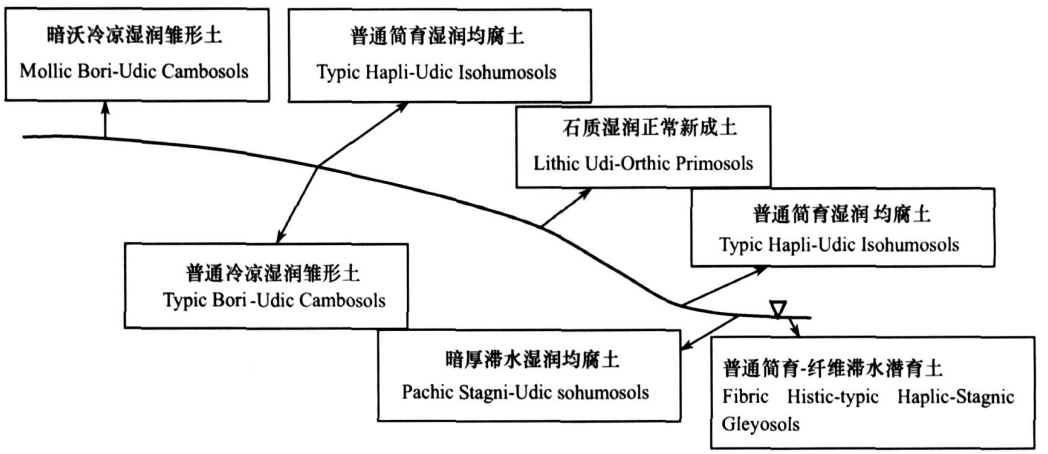

图 3 土壤类型空间序列表

Fig. 3 General catenary sequences of soil types

4　方法验证

将获得的土壤环境知识应用于预测性土壤制图,根据土壤类型图的精度来验证该方法的有效性。

4.1　预测性土壤制图

研究采用土壤 - 环境推理模型(SoLIM, Soil-Land Inference Model)进行预测性土壤制图。SoLIM 是一种基于知识的推理模型,它将土壤 - 环境关系知识与环境数据结合进行土壤推理制图,可得到每一土壤类型的隶属度分布图。通过对土壤类型的隶属度分布图的硬化(hardening)可得到土壤类型栅格分布图,硬化是指将每一像元对区域所有土壤类型隶属度中最大值所代表的土壤类型作为该像元的土壤类型[2]。

本文以土壤亚类作为土壤制图的基本单位。其中,由于普通简育滞水潜育土和纤维滞水潜育土在河谷平原中随植被类型的不同而相间分布,制图中将其作为组合处理。研究区的土壤类型硬化图如图 4 所示。

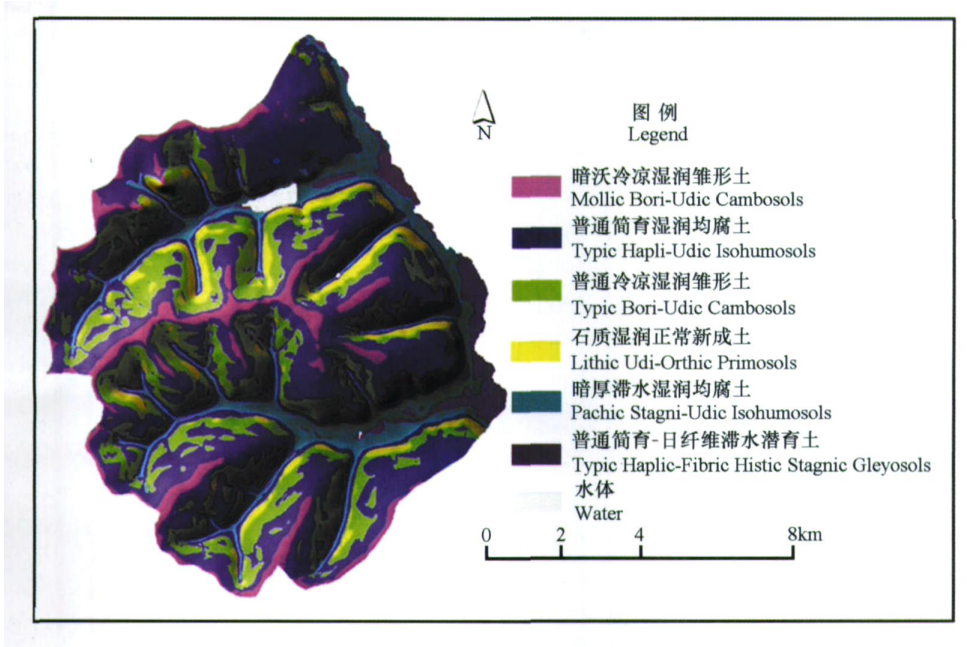

图 4　研究区硬化土壤类型图

Fig. 4　Harden soil map of the study site

4.2　土壤类型图的精度评价

用于土壤类型图精度检验的样点共 64 个,采用随机采样、均匀采样和按地形地貌采样三种采样方式。随机采样和均匀采样用来检验所获取的土壤信息的整体情况,其中随机采样 23 个,而均匀采样 31 个,以 0.6′×0.6′(纬度 ×经度)为间隔分布。按地形地貌采样旨在检验模型是否可以很好地捕捉土壤信息的空间变化,设计为横穿山坡、沟谷的线路,使得所布的 10 个采样点能在较短距离中穿越主要的景观类型变化。

土壤图中分类正确的点数总计为 46 个,因此土壤图的总精度为 46/64 ×100 % = 72 %。考察土壤图错分样点在空间的分布,发现这些采样点多分布在坡上较凹的地形中。这些采样点可能受局部地形的影响较大,由于通过 DEM 生成环境要素数据时经过了一定程度的平滑,使得所得环境数据和实际情况有所不同。因此,可能影响到了该点土壤的隶属度信息,从而导致分类的不正确。

中国第二次土壤普查一般所得最大比例尺的土壤图为县级的 1∶5 万,部分地区有乡级和更大比例尺的土壤图[13],本研究区尚没有大比例尺的土壤类型图,因此很难和国内的大比例尺土壤图的精度进行比较。对比美国同等比例尺(1∶24 000)的传统土壤制图大约为 50 %～60 %的精度[2],本方法获得的土壤图的精度更高。另外,微地貌采样分类精度达 80 %,可见土壤类型图在捕捉到土壤信息的整体空间分布的同时也可以捕捉到土壤信息在微地貌空间上的变化,因此通过模糊聚类获得土壤 - 环境知识

的方法是可行的。

5 结论与讨论

本研究提供了一种在没有土壤普查专家以及可利用的土壤图的地区利用 FCM 模糊聚类获取土壤 - 环境间关系知识的方法。研究将所建立的方法应用于黑龙江鹤山农场老莱河左岸的一个区域,得到研究区的土壤 - 环境关系知识,采用 SoLIM 模型将提取的土壤 - 环境知识与研究区的环境数据库结合进行土壤推理制图,得到土壤类型图。通过野外采样点进行验证,所得土壤类型图的总体精度达 72 %。由此可以证明通过 FCM 模糊聚类是一种获取土壤 - 环境关系知识的有效途径,可为自然资源(例如土壤)调查制图提供一种有效的方法。

野外采样是一个耗时耗力的过程,本研究中 FCM 对环境因子聚类的结果可以为野外目的性采样提供支持,即可以通过较少采样点(21 个)获取较高质量的知识,从而大大提高了采样效率,进而可以降低成本。此外,本文所建立的方法所获得的环境数据库和土壤 - 环境关系知识库均以文本和图件等数字形式进行保存,避免了因土壤专家流失而流失的问题,并且这种形式的数据库和知识库还可以因技术方法更新而得到不断的更新。

虽然 FCM 是一种在缺少土壤专家的情况下无需进行大量野外采样即可获取土壤 - 环境关系知识的有效途径,但在应用过程中也反映出 FCM 算法的一些问题,例如算法的稳定性等。其主要问题在于加权指数 m 和最佳聚类数 c 的选择目前并没有公认比较好的方法,研究中多带有一定的主观性。选择不同的 m 和 c 所得聚类结果不同。根据上图单元不同,采用越基层的分类单元进行制图通常需要更多的聚类类数。本研究中采用亚类作为基本单元,若要预测更详细的分类单元如土系的空间分布规律,则在应用 FCM 时应选择更多的环境组合聚类数,获取更多的采样点,以得到在小尺度空间中更详细的土壤信息。与此同时,土壤制图的比例尺也应随之增大。下一步的研究将针对 FCM 算法进行改进,以期得到更好的聚类结果,并采用更基层的分类单元如土系等进行土壤制图。本研究的后续工作还包括将基于 FCM 提取土壤 - 环境关系知识的方法推广到我国其他地区。

致 谢 本研究聘请中国科学院南京土壤研究所杜国华研究员作为本研究的土壤分类专家,在此特表示感谢。

参 考 文 献

[1] McBratney A B, Mendonca Santos M L, Minasny B. On digital soil mapping. Geoderma, 2003, 117: 3～52

[2] Zhu A X, Hudson B, Burt J E, *et al*. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci. Soc. Am. J., 2001, 65: 1 463～1 472

[3] Glinka K D. The great soil groups of the world and their development. Ann Arbor, MI: Edwards Bros., 1927

[4] Jenny H E, Hilgard W. The Birth of Modern Soil Science. Berkerley, CA: Farallo Publication, 1961

[5] Hudson B D. The soil survey as a paradigm-based science. Soil Sci. Soc. Am. J., 1992, 56: 836～841

[6] Odeh I O A, McBratney A B, Chittleborough D J. Fuzzy c-means and kriging for mapping soil as a continuous system. Soil Sci. Soc. Am. J., 1992, 56: 1 848～1 854

[7] Moore I D, Gessler P E, Nielsen G A, *et al*. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J., 1993, 57: 443～452

[8] 刘付程, 史学正, 于东升, 等. 基于地统计学和 GIS 的太湖典型地区土壤属性制图研究 ——以土壤全氮制图为例. 土壤学报, 2004, 41(1): 20～27. Liu F C, Shi X Z, Yu D S, *et al*. Mapping soil properties of the typical area of Taihu Lake watershed by geostatistics and geographic information systems —A case study of total nitrogen in topsoil (In Chinese). Acta Pedologica Sinica, 2004, 41(1): 20～27

[9] 杨胜天, 朱启疆, 李天杰. RS 和 GIS 支持下的土壤系统分类制图方法研究 ——以贵州贵阳市为例. 土壤学报, 2001, 38(1): 41～48. Yang S T, Zhu Q J, Li T J. The research on the cartography of soil taxonomy on the basis of RS and GIS, —Taking Guiyang City, Guizhou Province as example (In Chinese). Acta Pedologica Sinica, 2001, 38(1): 41～48

[10] Skidmore A K. Terrain position as mapped from gridded digital elevation data. Inter. J. GIS, 1990, 4(1): 33～49

[11] Zhu A X. A similarity model for representing soil spatial information. Geoderma, 1997, 77: 217～242

[12] Qi F, Zhu A X. Knowledge discovery from soil maps using inductive learning. Geographical Information Science, 2003, 17: 771～795

[13] Shi X, Zhu A X, Burt J E, *et al*. A case-based reasoning approach to fuzzy soil mapping. Soil Sci. Soc. Am. J., 2004, 68: 885～894

[14] 朱阿兴, 李宝林, 杨琳, 等. 基于 GIS、模糊逻辑和专家知识的土壤制图及其在中国应用前景. 土壤学报, 2005, 42(5): 844～851. Zhu A X, Li B L, Yang L, *et al*. Predictive soil mapping based on a GIS expert knowledge, and fuzzy logic framework and its application prospects in China (In Chinese). Acta Pedologica Sinica, 2005, 42(5): 844～851

[15] Hu Y M, Dai J, Wang R C. GIS-based red soil resources classification and evaluation. Pedoshphere, 1999, 9(2): 131～138

[16] McSweeney K, Slater B K, Hammer R D, *et al*. Towards a new framework for modeling the soil-landscape continuum. *In*: Amundson R. ed. Factors of Soil Formation: A Fiftieth Anniversary Publication. Madison, WI: Soil Science Society of America, 1994. 127～154

[17] Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy *c*-means clustering

algorithm. Computers and Geosciences, 1984, 10(2/3): 191～203

[18] 于剑. 论模糊 c 均值算法的模糊指标. 计算机学报, 2003, 26(8): 968～973. Yu J. On the fuzziness index of the FCM algorithms (In Chinese). Chinese Journal of Computers, 2003, 26(8): 968～973

[19] 高新波. 模糊聚类分析极其应用. 西安: 西安电子科技大学出版社, 2004. Gao X B. Fuzzy Cluster Analysis and Its Applications (In Chinese). Xian: Xidian University Press, 2004

[20] 龚子同. 面临新世纪挑战的土壤地理学. 土壤, 1999, 31(5): 236～243. Gong Z T. The prospect of soil geography under new challenges in the 21st century (In Chinese). Soils, 1999, 31(5): 236～243

EXTRACTION OF KNOWLEDGE ABOUT SOIL-ENVIRONMENT RELATIONSHIP FOR SOIL MAPPING USING FUZZY c-MEANS (FCM) CLUSTERING

Yang Lin[1,2,3]　Zhu Axing[1]　Li Baolin[1]　Qin Chengzhi[1]　Pei Tao[1]　Liu Baoyuan[2]　Li Runkui[1,3]　Cai Qiangguo[4]

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

(2 *College of Geography and Remote Sensing Sciences, Beijing Normal University, Beijing* 100875, *China*)

(3 *Graduate School of the Chinese Academy of Sciences, Beijing* 100049, *China*)

(4 *Key Laboratory of Water Cycle & Related Land Surface Processes, Institute of Geographical Sciences and Resources Research, Chinese Academy of Sciences, Beijing* 100101, *China*)

Abstract　For areas with no soil survey experts or soil maps available, knowledge about soil-environment relationships is a key to predictive soil mapping using knowledge-based approaches. An approach based on an unsupervised fuzzy clustering method (FCM, fuzzy *c*-means clustering) is recommended to acquire the knowledge. The method consists of four steps: 1) define those environmental factors which play decisive roles in formation and development of soil, then build up environmental database; 2) identify environmental niches (combination of environment conditions) by running FCM on the environmental database; 3) allocate field efforts to relate unique environmental factor combinations to soil types; 4) establish a soil-landscape model by analyzing the relationship established from the field samples. The model is then used to predict spatial distribution of soils. The method was applied in a study area in Northeast China and a soil map at the subgroup level was created for the area. The derived soil map was accurate up to 72 % when verified using an independently collected field sample set. Comparing this method with the conventional soil mapping method (50 %～60 %) in USA in accuracy, it could be concluded that the approach is effective for prognostic soil mapping over areas with no local soil experts available. At the same time, the clustering method can improve efficiency of the field sampling.

Key words　Predictive soil mapping; Knowledge about soil-environment relationships; fuzzy *c*-means (FCM); Environmental combination; SoLIM (Soil-Land Inference Model)