基于分类树方法的土壤有机质空间制图研究*

周 斌 许红卫 王人潮
(浙江大学农业遥感与信息技术应用研究所, 杭州 310029)

摘 要 以浙江省龙游县研究区为例, 提供了一种推断和表达土壤有机质(OM)含量空间分布信息的方法, 通过一种数据挖掘方法——分类树建模方法将土壤 OM 含量与一些易于广泛观测的景观属性, 包括地形、地质、土地利用和遥感影像建立联系, 从而将有关土壤 OM 含量分布的知识转入一种清楚的、定量的、与景观因子相关联的规则系统中, 并以此来预测研究区土壤 OM 水平的连续空间分布。树分析选取了高程、岩石类型、土属类型、PC_4、PC_2、土地利用类型、PC_3、PC_1、上坡贡献面积、坡度、坡向、平面曲率和剖面曲率来预测研究区土壤 OM 等级的分布。其中, 高程、岩石类型、土属类型和反映植被覆盖度的 PC_4、PC_2 以及土地利用类型对于研究区土壤 OM 等级预测更为重要。从分析结果来看, 依据分类树所划分出的景观类型与土壤 OM 含量有着较好的关联性。

关键词 数据挖掘, 分类树, 土壤 OM, 空间推断, 景观建模
中图分类号 S159.92

对于许多水文、生态和土地管理应用而言, 仅仅知道土壤类型的分布状况是不够的, 我们还需要知道其性质的空间分布趋势。一些研究者在测量土壤性质的空间变异性时使用了地统计方法[1~4]。这些地统计方法对于简单的景观地区是很有用的, 因为它能满足地统计的平稳假设(Stationary assumptions)。然而, 这些定量的插值技术对于复杂地形应用则受到了限制, 因为土壤的发生的方式复杂, 无法达到地统计关于平稳假设的要求, 而且这些技术通常要求大量的野外采样数据, 这给许多应用带来困难。

对于复杂地区的土壤性质调查通常采用土壤景观建模方法。土壤景观模型表达了我们对于土壤类型的分布与地貌特征如水文网、地质构造或地质年代等之间关系的理解[5,6]。Skidmore 等[7,8]在首先建立了实地的土壤—植被—景观的相互关系后, 使用靠当地土壤专家打分的专家系统, 根据植被、坡度以及坡度曲率等信息来推测出其立地最有可能出现的土壤类型。Cook 等[9]则使用了贝叶斯概率方法计算了土壤有机质(OM)在坡地上的空间分布, 所使用的预测变量包括湿度指数(一种类似上坡贡献面积的水文参数)、坡向、坡度以及灌溉程度四种。Moore 等[10]和 Lark[11]使用多元线性回归分析将土壤性质与地形属性进行相关分析, 并利用这些关系来预测土壤性质。这项技术假设土壤和地形属性之间的关系是线性的, 而且也要求有大量的野外数据来提取这种关系。McKenzie 和 Ryan[12]将地形、气候以及航测 gamma 辐射数据与土壤性质进行相关分析, 结果可以解释在一个更大地区内测量的 3 种土壤性质的 40%~80% 的样本变异。

尽管以往研究取得了一定的研究效果, 但不难看出仍存在一些尚未解决的问题, 如需依赖领域专家的广泛参与[7,8]、无法满足地理数据的统计不相关假设[9], 以及难以把握存在于土壤性质和环境变量之间的非线性关系[6]等, 同时应用这些传统统计学方法还会面临处理字符型数据的困难。因此, 很有必要使用一种新的方法来获取和表达土壤性质与其景观因子之间的复杂对应关系。

人工智能是 20 世纪 50 年代兴起的一门学科。在数据库技术飞速发展的同时, 人工智能领域的一

* 国家自然科学基金项目(40101014 和 40001008)资助
收稿日期: 2002-04-10; 收到修改稿日期: 2002-12-14

个重要分支——数据挖掘的研究也取得很大进展。根据人类学习的不同模式,人们提出了很多数据挖掘方法,如分类树、神经网络、粗糙集理论和遗传算法等。其中某些常用且较成熟的算法已被人们运用于实际的应用系统及智能计算机的设计和实现中。本研究使用了一种数据挖掘方法——分类树建模方法,对浙江省龙游县已有土壤 OM 样点数据与其景观空间数据之间的对应关系进行了分析,将土壤样点数据、数字高程数据以及派生地形属性、数字地质图、数字土地利用现状图等进行复合分析,推导出土壤 OM 水平的预测模型。

1 研究区概况和数据

1.1 研究区概况

龙游县地处浙江省中西部,衢江沿岸。地理位置: 119°02′ ~ 119°20′ E, 28°44′ ~ 29°17′ N。全县总面积约1 138.72 km^2,土地利用现状以林业用地为主,其次为耕地22 753 hm^2(其中水田占90%以上)。龙游县海拔33~ 1 440 m,属于中亚热带季风气候区,全年四季分明,光照充足,气温适中,雨量充沛,全年降雨量为1 671.6 mm。

龙游县在地貌上形成了几个梯级,海拔高程分别为龙南、龙北向中部逐渐降低。根据地貌形态的差异,全县地貌可分为河谷平原、缓坡岗地和丘陵山地等三大类。由于地形地貌复杂多样,因此该县的土壤及其景观类型表现出了较强的异质性。根据第二次土壤普查结果,龙游县的土壤类型主要有 5 个土类,分别是红壤、黄壤、水稻土、潮土和岩性土。这些土类类型在龙游县可继续被划分为 13 个亚类和 45 个土属[1]。

1.2 所使用的数据

1.2.1 研究区的选择 由于龙游县的1∶50 000地质图目前尚未完全调查完毕,因此根据其现有的图件覆盖区域确定为本次研究的最终研究区域(图1),面积大约为777.8 km^2。

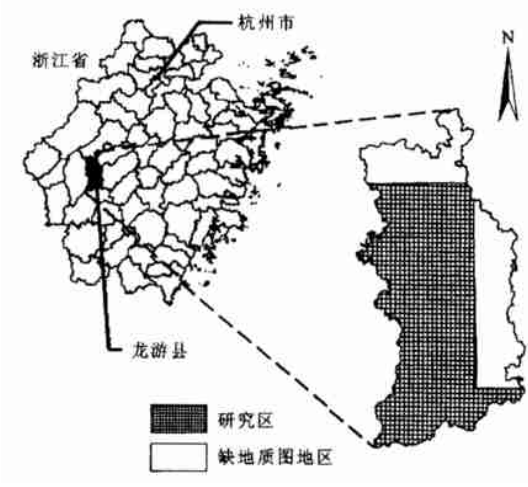

图 1 研究区的地理位置

Fig. 1 Location of the study area

1.2.2 实验材料 (1) 土壤 OM 数据。根据第二次土壤普查结果,研究区共有土壤样点 317 个,土壤图上所标明的土壤性质包括 OM、pH、速效 P 和速效 K 四种。这些土壤性质测量数据是经过归类的等级数据,针对土壤 OM 而言,共划分为 6 个等级,分别为: 1(>4%)、2(3%~4%)、3(2%~3%)、4(1%~2%)、5(0.6%~1%)和 6(≤0.6%)。将 317 个样本随机划分为两组: 一组约占 2/3,211 个样,作为训练数据;另一组占 1/3,106 个样,作为检验数据。(2) 景观数据。包括1∶50 000 土壤图、地质图、土地利用现状图、DEM 及其派生因子,也即坡度、坡向、剖面曲率(Profile curvature)、平面曲率(Plane curvature)、上坡贡献面积(Upslope contributing area)以及通过双时相的 TM 影像派生出的 4 个主成分波段作为预测因子来建立土壤 OM 的空间推断规则,各预测因子的数据类型及其描述列于表 1。

本研究所使用的遥感数据为两景 TM 卫星影像,一景为 1997 年 6 月 5 日获取,另一景为 2000 年 5 月 4 日获取。之所以要选用不同季节的两景卫星影像数据,主要是因为更加丰富的季节光谱信息有助于将农作物区(特别是水稻田)从其它土地覆盖类型中分离出来,从而更有利于把握土壤 OM 与农业利用类型之间的关系。为了减少波段数,我们对经过几何校正的两景TM 影像进行主成分分析,最后选用前 4 个主成分波段(占整个影像信息变异量的 99.4%)作为遥感光谱变量。

(1) 龙游县农业局。龙游县土壤志。1984

表 1　预测因子的数据类型及其描述

Table 1　Data types and description of predictive factors

预测因子 Predictive factors	缩写 Abbreviation	数据类型 Data format	因子数据描述 Discription
土壤	SOL	专题型	根据国家第二次土壤普查 1∶5 万土壤图数字化得到
土地利用	LDU	专题型	根据龙游县 1∶5 万土地利用现状图(1996)数字化得到
地质	GEO	专题型	根据浙江省地质调查 1∶5 万图件(龙游幅、灵山幅、沐尘幅和蕉川幅)数字化得到
高程	ELE	浮点型	根据国家 1∶5 万地形图数字化得到(10m 等高距)
坡度	SLP	浮点型	以下各地形因子均由高程数据派生而来
坡向	ASP	专题型	划分为 5 种坡向: 阴坡——0°~ 45° 方位角和 315°~ 360° 方位角 半阴坡——45°~ 90° 方位角和 270°~ 315° 方位角 半阳坡——90°~ 135° 方位角和 225°~ 270° 方位角 阳坡——135°~ 225° 方位角 平地——361° 方位角或坡度< 2°
剖面曲率	PRO	整型(0~ 255)	控制水流是加速还是减速, 因此可用来描述侵蚀或沉积程度。曲率越接近 255, 表明该地点越容易沉积; 曲率越接近 0, 则表明该地点越容易侵蚀
平面曲率	PLA	整型(0~ 255)	控制水流是集中还是分散。曲率越接近 255, 表明该地点越靠近山脊部; 曲率越接近 0, 则表明该地点越靠近山谷部
上坡贡献面积	UCA	整型(0~ 255)	对于同一个坡面而言, 上坡集水面积越小, 说明相对越接近坡顶, 也即上坡位; 上坡集水面积越大, 则相对越接近坡底, 也即下坡位
遥感	PC_1~ PC_4	整型(0~ 255)	PC_1——总体亮度; PC_2——绿度; PC_3——土壤裸露程度; PC_4——光谱季节变化

1.3　方　法

树模型分析利用了信息论中的熵值理论来寻找数据库中具有最大信息量的预测变量, 建立分类树的一个结点, 再根据该变量的不同取值建立树的分枝; 在每个分枝子集中重复建树的下层结点和分枝的过程, 树生成一直持续到每一个叶上的事件的数量达到最少或叶已经足够地同源。分类树的目的是用来依据上述选择的预测变量来预测土壤 OM 含量。预测模型语句中使用了连续定量参数高程、坡度、坡向、曲率、上坡集水面积以及遥感数据和分类参数如地质图、土壤图和土地利用图来预测研究区的土壤 OM 含量等级。为此, 首先使用事先提取的包含 211 个样本的训练样本来进行树模型分析, 研究土壤 OM 水平与景观因子之间潜在的关联性。

2　结果与讨论

2.1　土壤 OM 与全景观因子的分类树分析

树分析的结果最终选取了高程、岩石类型、土属类型、PC_4、PC_2、土地利用类型、PC_3、PC_1、上坡贡献面积、坡度、坡向、平面曲率和剖面曲率来预测研究区土壤 OM 等级的分布(图 2)。其中, 高程、岩石类型、土属类型和反映植被覆盖度及其季节变化信息的 PC_4、PC_2 以及土地利用类型被用于在相对较高结点处来分裂树, 因此对于研究区土壤 OM 等级预测而言是更为重要的控制变量。

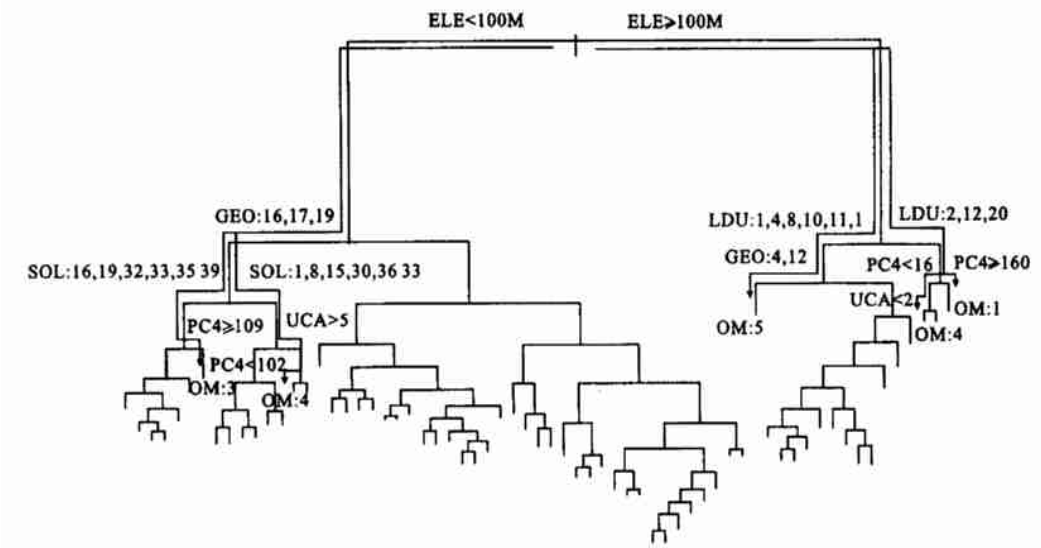

图 2 土壤 OM 预测的全景观因子分类树结构图(图中标注代表了沿箭头线路径进行推断时所使用的判据)

Fig 2 Structure of all landscape factor classification tree for soil OM prediction (the data in Fig. 2 are variables used for predicting soil OM along the tree)

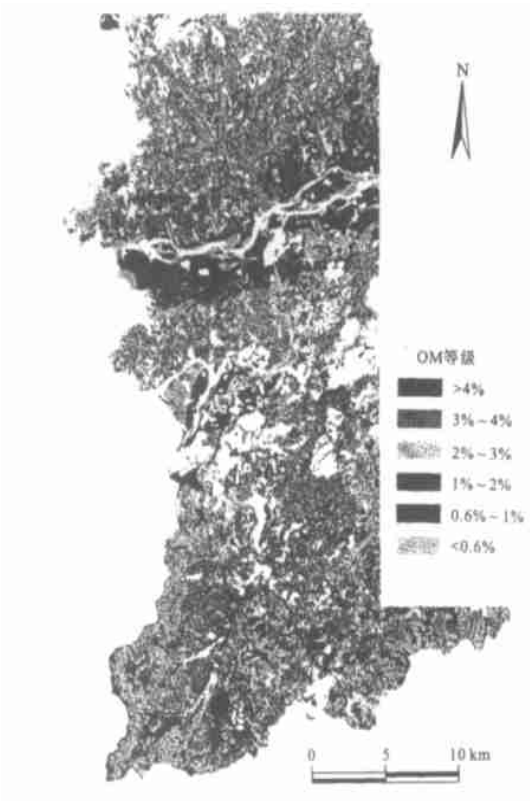

图 3 土壤OM 的全因子预测图(图中白色区域为规则空缺区)

Fig 3 Predicted Soil OM Map (White patches in the map refers to areas lacking information)

从分析结果来看,依据分类树所划分出的景观类型与土壤 OM 含量普遍有着较好的关联性。例如,沟谷部位要明显比山脊部位的土壤 OM 含量高;缓坡沟谷要比陡坡沟谷更富 OM;同样位于山脊的则明显受到不同利用类型的控制,如灌溉水田和竹林地就要比其它林地、疏林地和草地的土壤 OM 高出两个等级。平地和凹坡地带,相对而言要比凸坡更有利于土壤 OM 的积累,但又受到不同的利用类型、坡向和海拔高度对土壤OM 含量的明显影响;阴坡上的茶园和林地土壤 OM 要比非阴坡上的更高,处于高海拔地区的茶园和林地要比相对低海拔的茶园和林地拥有更高的土壤 OM 等。

利用 ERDAS Imagine 8.4 软件提供的知识工程器(Knowledge engineer),将分类树得出的结果加以规则化,继而对研究区的土壤 OM 进行空间预测。图 3 显示了根据全景观因子预测出的土壤 OM 空间分布情况。从结果图中可以发现存在不少“空白区”,导致这些”空白区”出现的主要原因是由于研究区的土壤样点并未覆盖所有的土属类别和土地利用类型,如所涉及到的土属类别仅 24 个,而研究区共有 42 个土属;另外,也没有完全涵盖研究区的土地利用类别。因此,在那些缺乏土壤采样点的土属和土地利用类型区域,就极可能出现一些规则未能覆盖的区域,从而导致无法对这些地区的土壤 OM 进行预测。

2.2　对“空白”区域的土壤 OM 等级的预测

对“空白”区域的土壤 OM 等级的预测可通过以下方式来进行: 通过只使用整型或浮点型等数值型变量而不考虑像土壤类型、土地利用类型等这样的专题型预测因子来进行预测, 然后将预测值填充到“空白区”。本研究主要是通过全地形因子来进行的, 也即选取数字高程模型及其派生地形因子(包括坡度、坡向、剖面曲率、平面曲率以及上坡集水贡献面积等) 作为预测因子参与预测, 最终生成一个具有 16 个叶节点的树。经过对划分参数进行分析, 我们可以将每一个输出节点都表示为一种地形单元(见表 2)。使用全地形因子预测出的土壤 OM 水平, 对全景观因子 OM 预测图进行了填补, 得到结果图 4。

表 2　不同地形单元的土壤 OM 水平

Table 2　Soil OM level on different topographic units

地形单元类型 Topographic unit	主要 OM 等级及频率[1] Dominating OM level and frequency	其它 OM 等级及频率(仅列出> 9%) Other OM levels and frequencies(only > 9% were listed as follow)	地形单元划分参数[2] Criteria of topographic unit
河谷平原	4(70%)	3(22%)	ELE< 50m
缓岗凸地	3(50%)	4(25%); 6(25%)	ELE: 50~ 64; PRO< 106
缓岗平地	3(44%)	4(34%); 2(15%)	ELE: 50~ 64; PRO ≥106; SLP< 2
缓岗半凸坡地	4(100%)	—	ELE: 50~ 64; PRO: 106~ 110; SLP ≥2
缓岗半凹坡地	3(56%)	4(25%); 5(19%)	ELE: 50~ 64; PRO: 110~ 119; SLP ≥2
缓岗凹地	4(75%)	3(25%)	ELE: 50~ 64; PRO: ≥119; SLP ≥2
低丘凸坡	5(67%)	4(17%); 6(17%)	ELE: 64~ 100; PRO< 106
低丘平地	3(65%)	4(16%); 2(14%)	ELE: 64~ 100; PRO ≥106
中高丘缓坡地	2(57%)	3(21%); 1(14%)	ELE: 100~ 206; SLP< 5
中高丘凸坡地	3(80%)	2(20%)	ELE: 100~ 206; SLP ≥5; PRO< 122
中高丘凹坡地	2(67%)	1(17%); 3(17%)	ELE: 100~ 206; SLP ≥5; PRO ≥122
山地凸坡谷地	1(71%)	3(14%); 6(14%)	ELE: ≥206; PRO< 110; PLA< 145
山地凸坡脊地	3(50%)	5(25%); 4(13%); 6(13%)	ELE: ≥206; PRO< 110; PLA ≥145
山地凹坡阴面	2(100%)	—	ELE: ≥206; PRO ≥110; ASP= 1
山地凹坡非阴面上坡位	3(75%)	2(25%)	ELE: ≥206; PRO ≥110; ASP= 2, 3, 4, 5; UCA< 2
山地凹坡非阴面中下坡位	1(73%)	2(18%); 3(9%)	ELE: ≥206; PRO ≥110; ASP= 2, 3, 4, 5; UCA ≥2

1) 土壤 OM 等级: 1(> 4%)、2(3%~ 4%)、3(2%~ 3%)、4(1%~ 2%)、5(0.6%~ 1%)和 6(≤0.6%)

2) ELE, 高程; PRO, 剖面曲率; SLP, 坡度; PLA, 平面曲率; ASP, 坡向; UCA, 上坡贡献面积

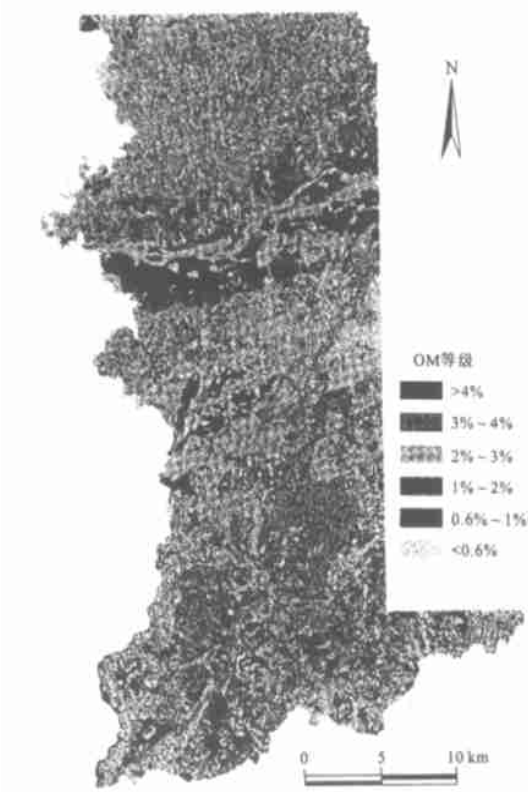

图 4 土壤 OM 预测图(使用地形预测值进行了规则空缺区的填补)

Fig. 4 Predicted Soil OM Map (This map shows that the white pitches in the previous map are filled up with the information obtained through the classification tree using all the topographic variables)

2.3 精度检验

用事先随机抽取出的 106 个样点作为检验数据,分别对预测结果进行了精度检验,结果精度如表 3 所示,其总体推断精度为 81.1%,总体 KAPPA 统计为 0.741。

2.4 误差分析

限于篇幅,在此只能简单列出导致预测出现偏差的几种可能因素。

2.4.1 所构建知识库的不完整性 由于知识、技术和数据等原因,预测模型中无法涵盖所有与土壤 OM 有关的因素,如气候、时间等,因此土壤 OM—环境模型中的一些信息也就无法被树分析的学习方法所获取,所获得的知识库也因而可能是不完整的。

2.4.2 GIS 数据层的不精确性 GIS 数据本身存在的误差会降低由机器学习方法推导出的土壤 OM 水平的精度。这些可能的误差包括土壤 OM 的测定和表示误差,其它地理图件如地质图、土地利用现状图、DEM 及其派生出的其它地形参数的属性和空间位置偏差等。

2.4.3 GIS 数据层的时相不匹配 导致机器学习方法产生误差的另一种原因是所使用的环境变量数据已非当初土壤调查时的环境数据。虽然一些变量如岩石类型和地形数据,可以被认为是时间上稳定的数据,但其它一些变量,特别是遥感和土地利用数据,则是时间敏感的数据。由于缺乏与土壤调查同期的遥感数据,本次工作所用的 TM 数据与第二次土壤普查在时间上相差了近 20 年。另外,土地利用现状资料也与土壤调查相差了 10 余年。因此,对预测结果图进行验证的最佳办法应当是进行实地采样分析,而非仅仅是与过去的土壤 OM 调查数据进行比较。

表 3 分类树结果的错误矩阵

Table 3 Cross-tabulation of reference vs classes predicted with the classification tree

土壤 OM[1)] Soil OM		参考数据 Reference						合计 Sum	生产精度(%) Producer's accuracy	用户精度(%) User's accuracy
		1	2	3	4	5	6			
预测数据	1	5	3	0	0	0	0	8	83.3	62.5
	2	0	12	3	0	0	2	17	66.7	70.6
	3	1	2	30	1	1	1	36	83.3	83.3
	4	0	1	2	34	1	0	38	94.4	89.5
	5	0	0	1	1	2	0	4	50	50
	6	0	0	0	0	0	3	3	50	100
合计		6	18	36	36	4	6	106	—	—

1) 土壤 OM 等级: 1(> 4%)、2(3%~4%)、3(2%~3%)、4(1%~2%)、5(0.6%~1%)和 6(≤0.6%)

3 结　论

树分析的结果最终选取了高程、岩石类型、土属类型、PC_4、PC_2、土地利用类型、PC_3、PC_1、上坡贡献面积、坡度、坡向、平面曲率和剖面曲率来预测研究区土壤 OM 等级的分布。其中, 高程、岩石类型、土属类型和遥感数据 PC_4、PC_2 以及土地利用类型对于研究区土壤 OM 等级预测而言是更为重要的控制变量。从分析结果来看, 依据分类树所划分出的景观类型与土壤 OM 含量普遍有着较好的关联性。精度评价结果表明, 本研究所采用的方法具有较好的预测精度。

参考文献

1. Webster R, Oliver M A. Optimal interpolation and isarithmic mapping of soil properties: VI. Disjunctive kriging and mapping the conditional probability. J. Soil Sci. , 1989, 40: 497~ 512
2. Goovaerts P. Geostatistics in soil science: State of the art and perspectives. Geoderma, 1999, 89(1/2): 1~ 45
3. 胡克林, 李保国, 林启美, 等. 农田养分的空间变异性特征. 农业工程学报, 1999, 15(3): 33~ 38
4. McBratney A B, Odeh I O A, Bishop T F A, *et al*. An overview of pedometric techniques for use in soil survey. Geoderma, 2000, 97: 293~ 327
5. Hall G F, Olson C G. Predicting variability of soils from landscape models. *In*: Mausbach M J, Wilding L P. ed. Spatial Variabilities of Soils and Landforms. SSSA Spec. Publ. 28. SSSA, Madison, WI. 1991. 9~ 24
6. Lammers R B, Band L E. Automated object description of drainage basin. Comput. Geosci. , 1990, 16: 787~ 810
7. Skidmore A K, Watford F, Luckananurug P, *et al*. An operational GIS expert system for mapping forest soils. Photogrammetric Engineering & Remote Sensing, 1996, 62(5): 501~ 511
8. Skidmore A K, Ryan P J, Dawes W, *et al*. Use of an expert system to map forest soils from a geographical information system. Int. J. Geographical Information Systems, 1991, 5(4): 431~ 445
9. Cook S E, Corner R J, Grealish G, *et al*. A rule-based system to map soil properties. Soil Sci. Soc. Am. J. , 1996, 60: 1 893~ 1 900
10. 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
11. Lark R M. Regression analysis with spatial autocorrelated error: Simulation studies and application to mapping soil organic matter. Int. J. Geographic Information Science, 2000, 14(3): 247~ 264
12. McKenzie N J, Ryan P J. Spatial prediction of soil properties using environmental correlation. Geoderma, 1999, 89: 67~ 94

SOIL ORGANIC MATTER MAPPING BASED ON CLASSIFICATION TREE MODELING

Zhou Bin Xu Hong-wei Wang Ren-chao

(Institute of Agricultural Remote Sensing and Information Technology Application, Zhejiang University, Hangzhou 310029, China)

Summary

Based on the case study of Longyou County, Zhejiang Province, an approach was introduced to deducing and expressing spatial distribution of soil organic matter. This is a kind of data mining method or classification tree modeling method, which associates soil OM content with some extensive easily observable landscape attributes, such as landform, geology, landuse and remote sensing images, thus transferring the soil OM-related information into a clear, quantitative, landscape factor-associated regular system. This system can be used to predict continuous soil OM spatial distribution. By analyzing the factors such as elevation, type of the rock, type of the soil, PC_4, PC_2, land uses, PC_3, PC_1, upslope contributing area, slope, exposure, plane curvature and profile curvature, the classification tree can predict distribution of soil organic matter levels. Among the factors, elevation, type of rock, type of soil, landuse, PC_4 and PC_2 (two indexes of vegetation coverage) are considered as the most important variables for predicting soil OM. Results of the prediction show a quite close relationship between soil OM contents and types of the landscape sorted by the classification tree with an accuracy of 81.1%.

Key words Data mining, Classification tree, Soil OM, Spatial prediction, Landscape modeling