基于样本分级的土壤属性自适应回归拟合方法①

李志鹏，宋现锋，李润奎*

(中国科学院大学资源与环境学院，北京 100049)

摘 要：精细土壤属性信息在诸多领域均具有广泛的应用，历来倍受关注。现有土壤属性预测方法具有适用性不强或需要大量人工经验和专家知识等缺点，限制了这些方法在实际应用中的推广。本文提出了一种土壤属性自适应预测方法，可分为 4 步：①对采样点进行分组处理；②利用回归模型构建各分组内土壤属性与主导环境因子之间的典型关系；③对分组方案进行自动优化；④利用各组对应的土壤–环境因子典型关系对研究区进行优化拟合预测。为了验证方法的有效性，本文在我国东北典型黑土区以土壤有机质含量为例进行了应用研究，结果表明：所提方法可对环境因子做出自动选择，并可通过优化拟合对土壤属性空间分布进行自适应预测，预测精度较高。本方法初步实现了土壤属性预测的自动化，具有较好的适用性。

关键词：土壤属性；自适应预测；典型土壤–环境关系；优化拟合

中图分类号：P934

土壤是发育于地球陆地表面的具有肥力能够生长植物的疏松表层，是成土母质在一定水热条件和生物作用下，经一系列生化物理作用而形成的独立的历史自然体，众多领域对精细的土壤信息均有迫切的需求[1–3]。土壤属性制图在土壤研究中占据着重要地位，但由于土壤的成土影响因素较多，且目前尚无通用性的定量化表达，因此现代土壤属性制图(主要是数字土壤制图[4])大多借助各种可能影响土壤属性的环境因子(又称环境变量)[5]来间接推测土壤属性。现代土壤属性制图方法大致可以分为基于统计模型的方法和基于专家知识的方法两大类[6]。其中，基于统计模型的方法又包括基于经典统计学模型的方法[7–9]和基于地统计学模型的方法[10–12]。二者均尝试利用一些特定统计模型揭示土壤属性与影响土壤的环境因子之间的定量关系——这种关系是土壤学专家所迫切希望得到的。但由于上述关系目前尚不明确，且目前此类方法通常无法针对不同土壤属性对土壤–环境关系作出自适应调整，因此这类方法容易造成过度拟合的现象。在基于专家知识的模型中，贝叶斯网络模型[13]和模糊逻辑推理[14]的方法应用最为广泛，而贝叶斯模型方法实质上仍具有一定的统计特性。基于专家知识的模型在专家知识的指导下可以达到较高精度，专家知识的作用目前主要体现在环境变量的选取上；但是在缺乏专家知识的情况下，这类模型的适用性受到了限制，且难以实现自动化的土壤属性推理。本研究旨在建立一种可以结合统计模型与专家知识模型各自优势，且能实现自动化预测的方法——以土壤属性与环境因子之间关系的定量化方式实现环境变量的自适应选取和结果的自适应优化拟合。

1 方法

本文所述的“自适应”不仅体现在环境变量的自动选取上，而且体现在通过自动优化拟合预测结果，有了上述两方面的自适应过程即可实现土壤属性的自动预测。

1.1 基本假定

由于土壤是一个连续统一体，一般情况下其属性是渐变的，因此多数现代土壤属性制图方法均假定土壤属性在一定范围内是均质的[15]，这也是数字土壤制图的基本前提。此外，现代土壤属性制图方法大多基于土壤景观模型理论[16]，即特定的环境因子组合下形成特定类型的土壤。以广泛应用的“土壤–环境推理模型”为例[17]，其假定相似环境条件下的土壤具有相似的属性。上述假定在实际当中取得了广泛的认可，因此在现有研究的基础上本研究提出如下

①基金项目：国家自然科学基金青年基金项目(41201038)和中国博士后科学基金项目(2012M510588)资助。

* 通讯作者(lirk@ucas.ac.cn)

作者简介：李志鹏(1988—)，男，山西岢岚人，硕士研究生，主要从事地统计分析等。E-mail: lizhipeng428@gmail.com

基本假定：相似的土壤属性是由相似的环境因子造成的，且在一定范围之内土壤属性的变化存在主导环境因子。

假定的前半部分是土壤–环境推理模型假定的逆向推理；后半部分的假定主要基于以下推理：会对土壤属性产生影响的环境因子通常不只一种，但是存在两种或两种以上环境因子同时具有同等且主要影响的概率在实际中趋近于 0，因此通常会存在某种主导环境因子。这种关系类似于色彩学中原色的概念[18]。以叠加型原色 RGB 色彩空间为例，现实中色彩可用三原色来表达(计算机中三原色的值域为：0 至 255)，在某一实际颜色中存在两种色彩占比例相同且大于第三种的概率约为 0.0058，3 种颜色所占比例相同的概率则更小。而对于土壤属性而言，不仅变异性强，而且其变化范围通常要更大，因此存在两种或以上环境因子共同占主导地位的概率基本上为 0。

1.2　自适应预测方法

基于上述假定，土壤属性在一定的范围内的特征由一定的主导环境因子所决定或体现，因此本研究首先对采样点进行分组，以使同一组内的土壤属性较为相似，不同组之间属性差异较大。符合这一原则的分类方法常用 K-means 分类法[19]。经过分组之后，属性相似的样点会位于相同的组内，即可认为同组样点由某种共同的环境因子所主导。

接着，找出各组之内的主导环境因子。本研究采用通过对土壤属性与各种环境因子进行回归拟合的方式，选择其中相关性最强的环境因子作为本组的主导环境因子。由此，有多少组将对应有多少种拟合关系——本文称之为“典型关系”，即这些关系库具有足够的典型性和代表性以表达未知区域的土壤属性–环境关系。此过程类似于遥感图像处理中像元分解过程中纯像元的提取，此时的典型关系是抽象化的“纯像元”。

在利用环境因子对土壤属性预测时，未知区域的任一点理论上可用一种甚至几种典型关系来表达(极端情况与任何典型关系均不符)。本研究在对土壤属性预测的过程中，针对用到多种典型关系的情况进行了自适应的加权拟合。仍以遥感图像处理为对照，此过程类似于遥感图像处理中利用纯像元通过加权拟合的结果来表达研究区域内其他任意像元值的过程[20]。

综上，本研究实现自适应优化拟合的关键步骤为样点最优分组方案的确定、回归模型的选择和结果的优化拟合。

1.2.1　最优分组确定　　样本点分组数的确定在本方法中具有关键的作用。一般来讲，随着分组数的增加，各组内回归拟合的误差将越小：在不进行分组(亦可说只分一组)的情况下，拟合的误差将是最大的；相反，如组划分足够细则误差将极小，甚至不存在误差。但是值得注意的是：随着分组数的增加，各组内样点数将减少，该组对于整体研究区域的代表性亦将降低，这也会直接导致整体预测精度的降低，因此理论上最佳分组数应介于两种极端情况之间。本研究依照常规统计学模型的惯例，以建模点总数的 5%(针对非整数的情况，向上取整)为分组的下限，即认为样点数低于全部样点总数的 5% 的组别不足以构建典型关系。

分组的最优在本研究中体现为误差最小，因此本研究以误差指标为依据对分组方案进行评价。首先将采样点分为建模点和验证点(不少于样点总数的 10%)；然后通过在可行分组数范围内迭代寻找最佳分组方案；在分组方案确定后，再对未知区域进行预测。其流程如图 1 所示。

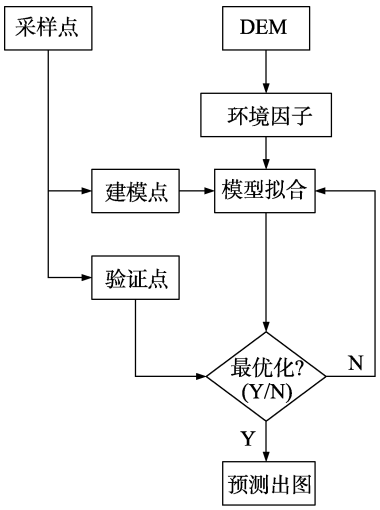

图 1　采样点分类流程图
Fig. 1　Flowchart of samples clustering

1.2.2　回归模型选择　　本研究中主导环境因子实质上是指那些能够体现土壤属性变化趋势的环境因子，即土壤属性随其属性变化呈单调趋势。为了确保这种因子的存在，本研究提取了大量可能与土壤属性相关的环境因子。

在上述条件下，各子组内部土壤属性应与主导因子之间呈现正(负)相关，用数学函数来表达可能为直线型、凹函数、凸函数或者凹凸不一致函数(如生长曲线)。非线性函数相对来说对样点数量和代表性要求更高，且在土壤属性与环境因子关系尚不明确的情况下，如果非线性模型的预测值与实际值出现偏差，

多种典型关系的拟合将会进一步加剧这种偏差；而线性模型一方面对样点数量要求较低，另一方面可以说是对不明确关系的一种折中表达。

因此，在对采样点进行了分组，并准备了大量环境变量，且会对多种典型关系进行拟合优化的情况下，线性关系可以说是一种风险较小的选择。

1.2.3 结果优化拟合　　由于未知点可能与不定数量的典型关系有关，因此各典型关系对未知点的贡献权重——典型性系数，需要得到确定。当未知区域某环境因子的值 X_i 位于一种相应典型关系 L 对应的环境因子值域(X_{i0}, X_{i1})对应的预测值域为(Y_{i0}, Y_{i1}))内时，其关系可示意为图 2。

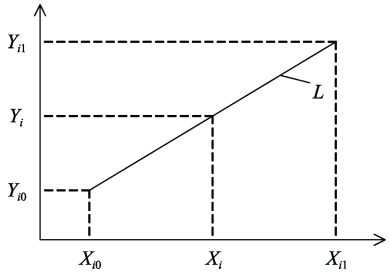

图 2 土壤属性与主导环境因子模拟关系图
Fig. 2 Relation between soil attributes and environmental variables

典型性系数由未知点的环境因子值与相应典型关系的取值范围的相对关系确定，计算方法如公式(1)所示：

$$w_i = \frac{X_i - X_{i0}}{X_{i1} - X_{i0}} \tag{1}$$

式中：w_i 为位于第 i 个典型关系取值区间内的典型性系数值，其他参数同上。在得到各典型性系数之后，则可对未知区域的土壤属性值进行综合预测，未知点的土壤属性预测值计算如公式(2)所示：

$$Y = \frac{\sum_{i}^{m} w_i Y_i}{\sum_{i}^{m} w_i} (m \in [1, n]) \tag{2}$$

式中：Y 为当前未知点的预测值，m 为当前点的所有环境变量满足的典型关系数，n 为典型关系的总数。按照同样方式遍历所有未知点即可得到当前研究区域内的最终预测结果图。

1.3 精度评价

本研究对模型精度的评价体现在分组方案的选择中，采用的评估指标为较为常用的平均绝对误差(mean absolute error，MAE)和均方根误差(root mean squared error，RMSE)[21]，MAE 代表预测值与观测值之间的恒定差，而 RMSE 代表预测值与观测值之间的总体差；二者均是越小代表预测精度越高，表达式分别如公式(3)和公式(4)所示。

$$\mathrm{MAE} = \sum_{j=1}^{l} \frac{\left| Y_j - Y_j' \right|}{l} \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{l} \sum_{j=1}^{l} \left[Y_j - Y_j' \right]^2} \tag{4}$$

式中：l 为验证点数目，Y_j 为第 j 个验证点的预测值，Y_j' 为第 j 点的观测值。

2 案例与分析

2.1 研究区域

研究区域位于黑龙江省黑河市嫩江县鹤山农场老莱河左岸(图 3)，属于典型的东北黑土区——我国

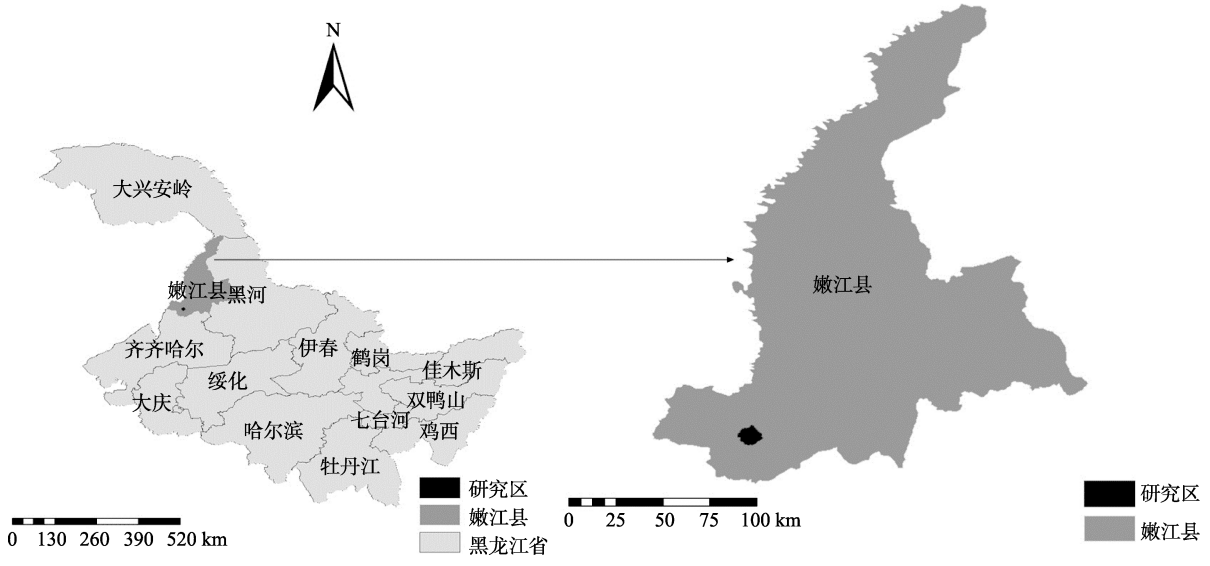

图 3 研究区位置示意图
Fig. 3 Location of study area

的“黑土粮仓”。面积约为 60 km^2，地形较缓，高差约为 100 m，年降水约 400 ~ 550 mm，成土母质基本上为黄土状亚黏土，原生植被为疏林草甸、灌丛草甸和杂草草甸，但现多被开垦为农田，作物以大豆为主[22]。由于长期受到不合理耕种等因素的影响，本区域水土流失较为严重，土壤结构遭到了破坏，有机质含量与未干扰地区相比显著降低。

2.2　数据与环境因子

有机质含量是土壤属性的一个重要指标，因此本研究以土壤有机质的详细空间分布为最终目标。为此，收集了 10 m×10 m 分辨率的 DEM 数据(经等高线转换而来)及覆盖全区的土壤 A 层有机质含量采样点 84 个(图 4)，以进行土壤有机质含量制图。

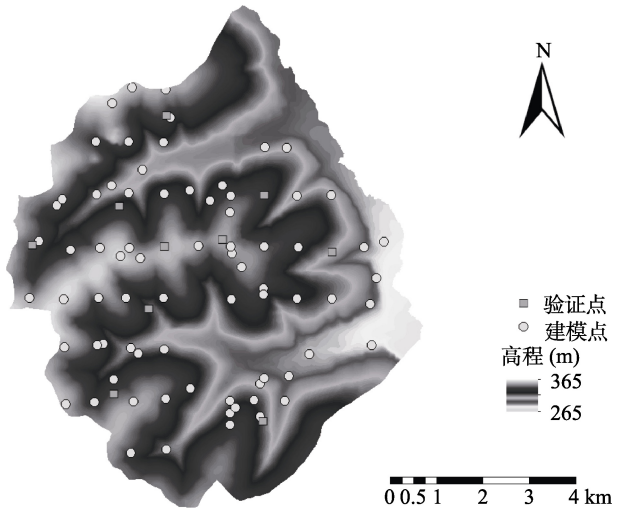

图 4　采样点(建模点与验证点)分布图
Fig. 4　Distribution of calibration and validation samples

目前，数字土壤制图中所用环境变量多提取自数字高程模型 DEM[23]，本文亦将 DEM 数据作为提取环境因子的数据源，根据前人经验，本文提取的可能表征土壤有机质含量的环境因子[24–25]列表及相关说明如表 1 所示。其中，为了解决量纲不统一的问题，

表 1　环境变量及提取方法汇总
Table 1　Environmental variables and their extraction methods

变量名	描述	提取方法
Curvature	地表总曲率	ArcGIS 10.1
Elevation	高程	
Hillshade	阴影	ArcGIS 10.1
LS-Factor	坡长因子	SAGA 2.0.8
planCurve	地表平面曲率	ArcGIS 10.1
profCurve	地表剖面曲率	ArcGIS 10.1
RPI	相对坡位指数	SimDTA 7.2[a]
Slope	坡度	ArcGIS 10.1
Solar	单位面积太阳辐射量	ArcGIS 10.1
streamPower	水流强度指数	SAGA 2.0.8
TWI	地形湿度指数	SAGA 2.0.8

注：a 简化数字地形分析软件[26]。

在具体运算前，本文首先对上述环境因子值进行了标准化处理，使其取值在[0,1]内。

2.3　预测过程与分析

本研究预测方法中最核心的步骤即为土壤样本点分组方案的确定。建模过程中，采样点先被分为两类：建模点 74 个，验证点 10 个；其中建模点用于预测时的分组。本研究区域的建模点分组时，各组的最少样点数至少为 4 个点(74×0.05，向上取整)。在此约束条件下，各分组方案中根据验证点和预测值关系得到的 RMSE 随着分组数增加的关系，如图 5 所示。

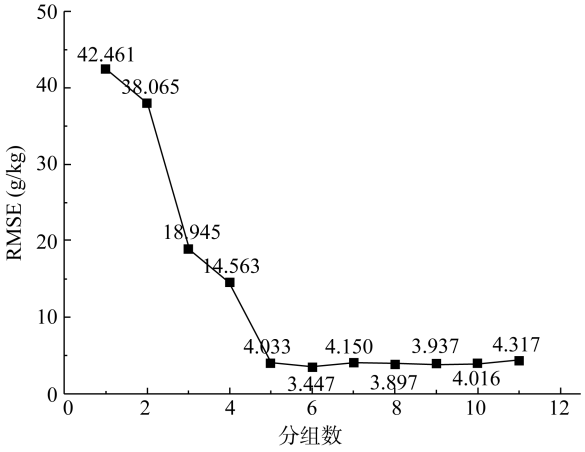

图 5　RMSE 与分组数关系
Fig. 5　Relation between RMSE and number of class

由图 5 可知，分组数多的情况误差小于分组数少的情况。分组数小于 5 时，RMSE 值随分组数的增加而急剧降低；分组数达到 5 之后，RMSE 值的变化变得平缓；在分组数为 6 时，达到全局小值 3.447；之后其值随分组数的增大，呈缓慢增大的趋势。因此，全局最优分组方案最终选择分组数为 6，其值随分组数的变化趋势基本符合本文的理论分析。

在分组数为 6 的情况下，各组的统计信息如表 2 所示(单位 g/kg)。从主导因子分布可以看出，面太阳辐射(Solar)及地形湿度指数(TWI)均出现两次：其中 Solar 在不连续的两组均占主导作用，可见本方法可根据不同土壤属性值域范围有效调整主导因子的选择；而 TWI 在连续两组起主导作用，但其拟合斜率不同(分别为 7.750 和 23.075)，表明在这两组中土壤属性随 TWI 变化幅度不同。

2.4　结果与讨论

本文的最终出图结果如图 6 所示，由图看出土壤有机质随地形的变化呈一定趋势，但是这种趋势在不同空间位置表现亦不尽相同；当然，本文结果也存在一定的条带特征，经对比发现，这与 DEM 数据源于

表 2 分组数为 6 时样点统计信息
Table 2 Descriptive statistics of samples when class number is 6

编号	最小值	最大值	平均值	标准差	主导环境因子
1	22.450	30.390	26.458	3.003	Solar
2	32.290	39.390	36.536	2.110	Elevation
3	40.810	44.290	42.471	1.199	Solar
4	45.190	49.810	46.895	1.216	TWI
5	50.220	64.740	56.184	5.161	TWI
6	76.080	91.800	83.323	5.617	RPI

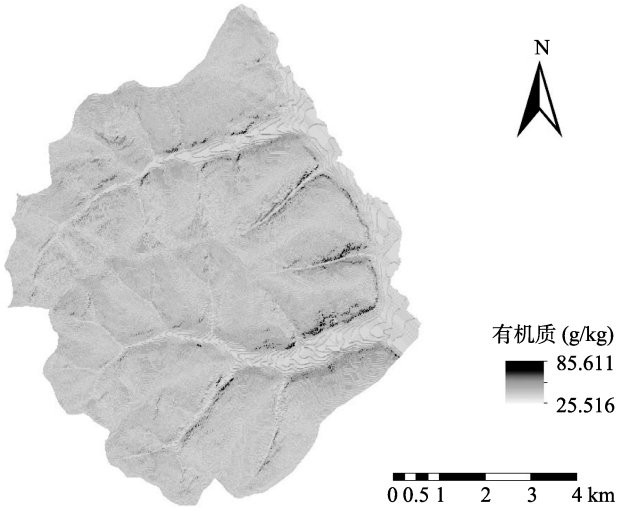

图 6 土壤有机质预测结果图
Fig. 6 Map of predicted soil organic matter

等高线可能存在一定的关联性。如分组方案确定部分所述，本文预测结果的 RMSE 值达到 3.447，此时对应的 MAE 为 2.526，相对而言[27]，本文模型具有较高的预测精度。

本文方法初步实现了土壤属性的自动化预测，且其精度较高；但是由于本方法始于对样本点属性的分级，且其核心流程是基于统计模型的，因此对采样点的代表性要求较高。例如，图 6 中存在不平滑的区域，主要为沟谷边缘，这可能与本文样点基本未覆盖沟谷有关；又如，上述预测结果中亦存在少量无值像元，即此部分点不满足任何一种典型关系，因此未给出预测。

3 结论与展望

本研究提出了一种基于样本点土壤属性分组，通过分段拟合土壤–环境关系，继而自适应选择主导环境因子并优化分组方案，最终自动预测土壤属性的数字土壤制图方法。相对于现有多数数字土壤制图方法，本研究实现了环境因子的自适应选取和预测结果的自适应优化拟合；达到了土壤属性预测中降低人工干预、提高预测过程自动化程度及预测效率的目标；且本方法能够给出较高的预测精度，表明其具有实用性。

本研究方法以对采样点进行分级为突破口，且是基于统计模型，因此对实测样点的代表性具有较高的要求；但反过来这在一定程度上可以作为检验现有样点代表性的指示。

此外，本研究环境因子主要提取自 DEM 数据，随着遥感技术的发展，一些学者已尝试将遥感图像及其派生产品用于土壤属性制图中[28–30]，这也是今后研究的一个主要方向和目标。

致谢： 感谢中国科学院地理科学与资源研究所朱阿兴研究员课题组对本文数据的支持，感谢中国科学院大学牛海山副教授在统计方法方面的指导；感谢京都大学 CSEAS 开放基金“Development of Web-GIS Framework for Soil Mapping and Modeling of Soil Dynamics for Sustainable Resource Management”在数字土壤制图技术方面的支持。

参考文献：

[1] 赵其国, 滕应. 国际土壤科学研究的新进展[J]. 土壤, 2013, 45(1): 1–7
[2] 海春心, 陈健飞. 土壤地理学[M]. 北京: 科学出版社, 2010
[3] Dobos E, Carré F, Hengl T, Reuter HI, Tóth G. Digital Soil Mapping as a Support to Production of Functional Maps[M]. Office for official publications of the European Communities, Luxemborg. EUR, 22123 (2006): 68
[4] 史学正, 于东升. “数字土壤”——21 世纪土壤学面临的机遇与挑战[J]. 土壤通报, 2000, 31(3): 104–121
[5] Jenny H. Factors of Soil Formation: A system of quantitative pedology[J]. New York: McGraw-Hill. 1941: 281
[6] 朱阿兴. 精细数字土壤普查模型与方法[M]. 北京: 科学出版社, 2008
[7] Moore ID, Gessler PE, Nielsen GA, Peterson GA. Soil Attribute Prediction Using Terrain Analysis[J]. Soil Science Society of America Journal, 1993, 57(2): 443–452
[8] McKenzie NJ, Ryan PJ. Spatial prediction of soil properties using environmental correlation[J]. Geoderma, 1999, 89(1/2): 67–94

[9] Lagacherie P, Holmes S. Addressing geographical data errors in a classification tree for soil unit prediction[J]. International Journal of Geographical Information Science, 1997, 11(2): 183–198

[10] Lark RM, Beckett PHT. A geostatistical descriptor of the spatial distribution of soil classes, and its use in predicting the purity of possible soil map units[J]. Geoderma, 1998, 83(3): 243–267

[11] Goovaerts P. Geostatistics in soil science: state-of-the-art and perspectives[J]. Geoderma,1999, 89(1/2): 1–45

[12] Harris P, Fotheringham AS, Crespo R, Charlton M. The Use of Geographically Weighted Regression for Spatial Prediction: An Evaluation of Models Using Simulated Data Sets[J]. Mathematical Geosciences. 2010, 42(6): 657–680

[13] Cook SE, Corner RJ, Grealish G, Gessler PE, Chartres CJ. A Rule-based System to Map Soil Properties[J]. Soil Science Society of America Journal, 1996, 60(6): 1 893–1 900

[14] Zhu AX, Hudson B, Burt J, Lubich K, Simonson D. Soil Mapping Using GIS, Expert Knowledge, and Fuzzy Logic[J]. Soil Science Society of America Journal, 2001, 65(5): 1 463–1 472

[15] Hengl T. A Practical Guide to Geostatistical Mapping[M]. Amsterdarn: University of Amsterdam, 2009

[16] Hudson BD. The soil survey as a paradigm-based science[J]. Soil Science Society of America Journal, 1992, 56: 836–841

[17] Zhu AX, Band L, Vertessy R, Dutton B. Derivation of soil properties using a soil land inference model (SoLIM) [J]. Soil Science Society of America Journal, 1997, 61(2): 523–533

[18] Grimley C, Love M. Color, Space, and Style: All The Details Interior Designers Need to Know but Can Never Find[M]. Rockport Publishers, 2007

[19] MacQueen JB. Some Methods for classification and Analysis of multivariate observations[A] // Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability[C]. Berkeley, CA, USA: University of California Press, 1967: 281–297

[20] 吕长春, 王忠武, 钱少猛. 混合像元分解模型综述[J]. 遥感信息, 2003(3): 55–58

[21] Willmott CJ. On the validation of models[J]. Physical Geography, 1981, 2(2):184–194.

[22] 杨琳, 朱阿兴, 李宝林, 秦承志, 裴韬, 刘宝元, 李润奎, 蔡强国. 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[J]. 土壤学报, 2007, 44(5): 784–791

[23] McBratney AB, Mendonça Santos M L, Minasny B. On digital soil mapping[J]. Geoderma, 2003, 117 (1/2): 3–52

[24] Pei T, Qin CZ, Zhu AX, Yang L, Luo M, Li B, Zhou C. Mapping soil organic matter using the topographic wetness index: A comparative study based on different flow-direction algorithms and kriging methods[J]. Ecological Indicators, 2010, 10(3): 610–619

[25] Qin CZ, Zhu AX, Qiu WL, Lu YJ, Li BL, Pei T. Mapping soil organic matter in small low-relief catchments using fuzzy slope position information[J]. Geoderma, 2012, 171–172: 64–74

[26] 秦承志, 卢岩君, 包黎莉, 朱阿兴, 邱维理, 程维明. 简化数字地形分析软件(SimDTA)及其应用——以嫩江流域鹤山农场的坡位模糊分类应用为例[J]. 地球信息科学学报, 2009, 11(6): 737–743

[27] 杨琳, 朱阿兴, 秦承志, 李宝林, 裴韬, 刘宝元.运用模糊隶属度进行土壤属性制图的研究——以黑龙江鹤山农场研究区为例[J]. 土壤学报, 2009, 46(1): 9–15

[28] Dobos E, Montanarella L, Nègre T, Micheli E. A regional scale soil mapping approach using integrated AVHRR and DEM data[J]. International Journal of Applied Earth Observation and Geoinformation,2001, 3(1): 30–42

[29] López-Granados F, Jurado-Expósito M, Peña-Barragán JM, García-Torres L. Using geostatistical and remote sensing approaches for mapping soil properties[J]. European Journal of Agronomy, 2005, 23(3): 279–289

[30] Shi X, Girod L, Long R, DeKett R, Philippe J, Burke T. A comparison of LiDAR-based DEMs and USGS-sourced DEMs in terrain analysis for knowledge-based digital soil mapping[J]. Geoderma, 2012, 170: 217–226

A Novel Self-adaptive Regression Model for Prediction of Soil Attributes Based on Sample Classification

LI Zhi-peng, SONG Xian-feng, LI Run-kui*

(*College of Resources and Environment*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract: Detailed and accurate soil attribute information has received growing attention and has increasingly been applied in various related fields. The majority of existing digital soil mapping approaches is effective only in specific condition while others require much expert knowledge and manual intervention. A new self-adaptive method for soil attribute mapping was presented in this paper: firstly, the samples were partitioned into several groups by their properties; secondly, the typical relation between soil attribute and key environmental factors in each class was derived through regression model, and then clustering method was optimized according to the residual information above, finally the non-sampled area was predicted by a weighted fitting of all the typical relations. A case study of soil organic matter mapping was taken in order to exam the performance of the approach. The result showed that the method was of wide suitability and could predict the soil attribute information with a high accuracy by choosing and fitting the key factors automatically.

Key words: Soil attributes, Self-adaptive, Typical relation, Weighted fitting