基于贝叶斯最大熵与历史数据的土壤属性空间预测

杨勇，李卫东，贺立源
（华中农业大学资源与环境学院，武汉 430070）

摘要：高精度的土壤属性图件是精准农业实施和土壤质量评价的基础，而搜集和合理利用尽可能多的信息/数据是提高土壤属性空间预测精度的有效途径，BME 方法在综合利用多源数据，特别是具有模糊性质的数据方面较经典地统计方法更有优势。本文将土壤属性的历史分布状况作为软数据，将具有现势性的土壤采样数据作为硬数据，利用 BME 方法将两类数据综合起来用于土壤属性空间预测中，通过与不同样点密度下的经典地统计方法对比，表明本文所提方法有较高的预测精度。文章最后对 BME 方法的优势和不足作了评价。

关键词：贝叶斯最大熵；软数据；土壤属性；空间预测
中图分类号：S159

Bayesian Maximum Entropy Prediction of Soil Properties Using Legacy Data as Soft Information

YangYong, Li Weidong, HeLiYuan
(Department of Resource and Environmental Information, College of Resources and Environment, Huazhong Agricultural Unviersity, WuHan 430070)

Abstract: High-precision maps of soil properties are basis for precision agriculture and soil quality assessment. Collection and rational use of more information/data is an effective way to improve the prediction accuracy of soil properties. BME has advantages over classical statistical methods in comprehensive utilization of multi-source data, in particular when data is fuzzy. In this paper, BME was used to estimate spatial distribution of soil continuous properties with historical map data as soft data and soil sampling data as hard data. We compared the prediction performance of BME with that of kriging using hard data of different densities. It is found that BME was more accurate. Finally, we discussed the pros and cons of the BME method.

Keywords: Bayesian Maximum Entropy; Soft Data; Soil Properties; Spatial Prediction

0 引言

土壤属性（如土壤中养分含量，污染性重金属含量等）的空间分布特征和定量分布信息是进行土壤质量评价和区域环境综合评估的基础。精准农业战略的实施和各种区域生态评价也都需要更详细更精确的土壤属性信息作为依据。因此，土壤属性空间预测一直是土壤学研究的热点问题，经典地统计学是目前应用于土壤属性空间预测上的常用方法，但其方法本身仍存在着一些不足，如缺乏对辅助信息的有效利用导致预测精度降低[1-3]（特别是在采样点比较稀时），预测结果具有一定的平滑效应[4-6]，方法要求的单点到多点高斯分布假设不易满足[7]等。本文将使用贝叶斯最大熵（Bayesian Maximum Entropy，以下简称 BME）方法，将土壤中有机质含量为分析对象，以同一属性的历史分布数据为软数据，以不同分布密度的土壤样品为硬数据，对土壤连续属性进行空间预测，并与经典地统计学方法所得结果进行对比。

基金项目：教育部新教师基金（20100146120018）国家自然科学基金 (41101193) 数字制图与国土信息应用工程国家测绘局重点实验室开放研究基金资助项目。
作者简介：杨勇，（1980-），男，副教授。主要从事地统计学及其在土壤中的应用方面的研究. E-mail: yangyong@mail.hzau.edu.cn

1 BME 方法

与经典地统计方法相比，BME 是一个相对较新的空间预测方法，使用该方法进行空间预测时，可以综合多种来源与多种形式的数据，使得预测结果的精度较单独使用被预测属性本身样品数据的方法高[8,9]。该方法进行空间预测时使用两方面数据：（1）专用数据（KS）：按照数据的精确与否分为硬数据（hard data）和软数据（soft data）两类，两类数据均定量表示被研究属性的含量，区别在于硬数据为确定性的值，而软数据的值具有模糊性质，形式为值域区间或概率分布，如对某个点位的田间观测近似数据，从过去的土壤图的图斑中获取的某养分含量范围等。相对于硬数据而言，软数据具有模糊性，获取容易，成本低等特点；（2）普遍知识\数据（KG）：用来描述空间随机域的整体特征的数据或知识，如一般自然规律、经验知识和基于硬数据任何阶的统计动差（如数学期望，协方差，方差等）。基于这两方面数据，BME 方法分为两个步骤：（1）使用 KG，基于最大熵原理，计算研究区域内未测点变量分布的先验概率密度函数（probability density function ,以下简称 pdf ）；（2）使用 KS，基于贝叶斯条件概率，更新上一步获取的先验 pdf，得到研究区域内未测点的后验 pdf。根据最终得到的后验 pdf，可以方便地制作多种土壤数字地图，如预测图、超越某个阈值的概率分布图等。到目前为止，BME 方法在国外已被用于多个自然资源与社会环境领域[10]，均取得了较经典地统计方法更精确的预测结果，但在软数据的来源与数据处理方法上还需要更多的探索和尝试。本文将历史预测结果作为软数据源，辅以恰当的数学手段，将软数据源转化为 BME 方法可用的软数据格式，在引入 BME 方法的同时，也为 BME 方法中软数据获取提供了一种思路。

2 研究区域与土壤样品

2.1 研究区域及历史数据

研究区域为湖北省武汉市汉南区，地处武汉市西南部，东经 113° 45′ 0″—114° 06′ 15″，北纬 30° 11′ 03″—30° 21′ 20″，全区面积 287 平方公里，地势平坦，气候温润，水资源丰富，土壤以黄棕壤和灰潮土为主，主要种植棉花、玉米、花生、水稻、油菜。2005 年，对研究区域进行了土壤样品的采集，并用经典克里格方法对土壤多个属性进行了空间预测，得到了若干土壤图件，其中有机质的分布如图 1 所示：

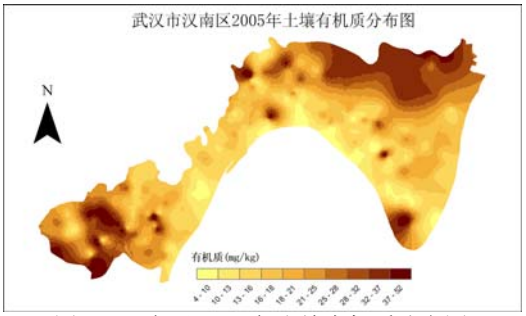


图 1：汉南区 2005 年土壤有机质分布图
Fig.1 Soil organic matter distribution in Hannan

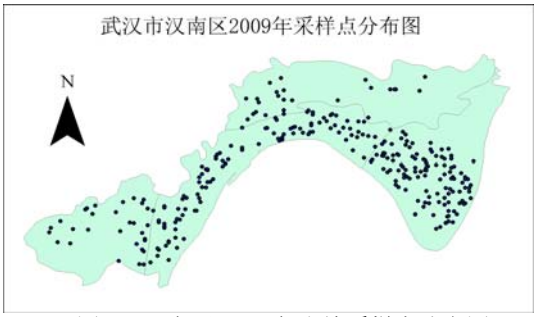
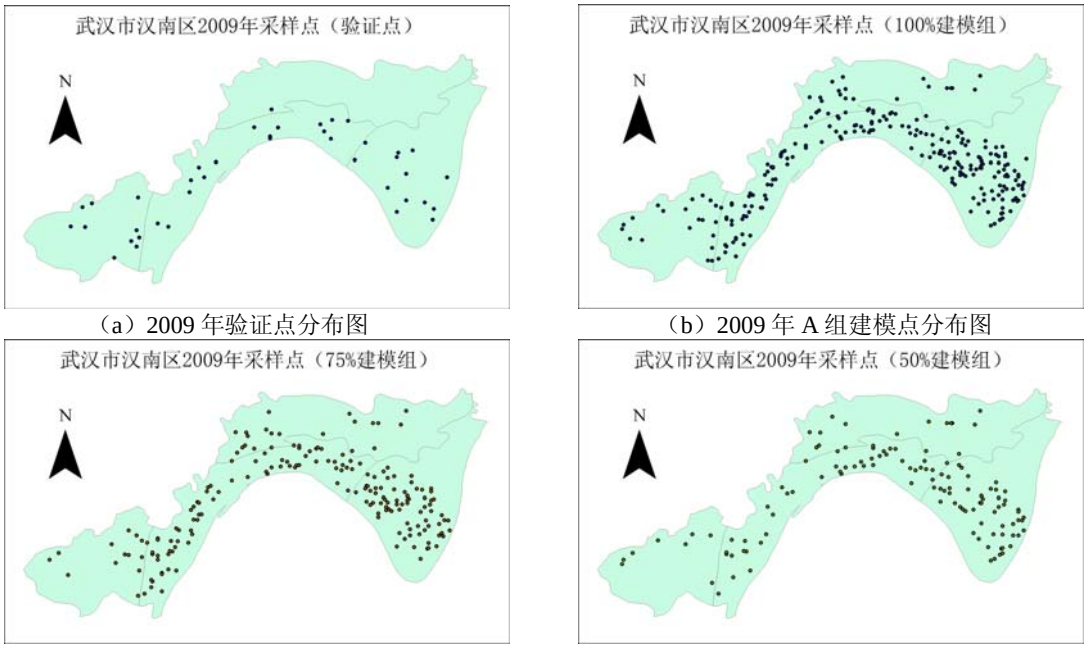


图 2：汉南区 2009 年土壤采样点分布图
Fig.2 Soil sampling sites in Hannan

2.2 土壤样品及分组

2009 年秋，又对该区域进行了土壤采样，共取得 283 个样点（如图 2）。为了分析不同土壤采样密度下，BME 方法的预测精度，我们将采样点分为建模点（243 个点）和验证点

（40 个点）两部分，其中建模点又分为 A、B、C 三个建模组，分别为全部建模点（243 个），75%的建模点（182 个）和 50%的建模点（121 个），选点方式均为随机，如图 3 所示：



（a）2009 年验证点分布图　（b）2009 年 A 组建模点分布图

（c）2009 年 B 组建模点分布图　（d）2009 年 C 组建模点分布图

图 3：验证点及各组建模点分布图

Fig.3 Soil verify sampling sites and diverse modeling sampling sites distribution in Hannan

3 计算方法

根据 BME 方法和本研究的目的，本文将 2009 年的各组采样点作为硬数据，考虑到土壤属性的含量有历史遗留性，即过去的土壤属性含量分布对当前的土壤属性含量是有影响的，所以将 2005 年的结果作为软数据源，以基于硬数据的统计动差作为普遍知识\数据。

3.1 先验 pdf 计算

定义空间随机变量（R.V）$Z_{map}=(Z_{hard},Z_{soft},Z_0)$，其中 Z_{hard},Z_{soft} 和 Z_0 分别表示待预测点 x_0 周围一定范围（变程）内硬数据、软数据和待预测位置的未知值。K_G 和 K_S 分别表示普遍知识和专用知识，其中 K_S 由硬数据和软数据组成。用 $f_G(z_{map})$ 表示基于 K_G 的 pdf。根据贝叶斯最大熵原理，当 K_G 仅考虑硬数据的各阶统计动差时，所得结果与普通克里格法结果一样，因此，本文中先验 pdf 可用普通克里格方法计算，结果符合正态分布，表示为：

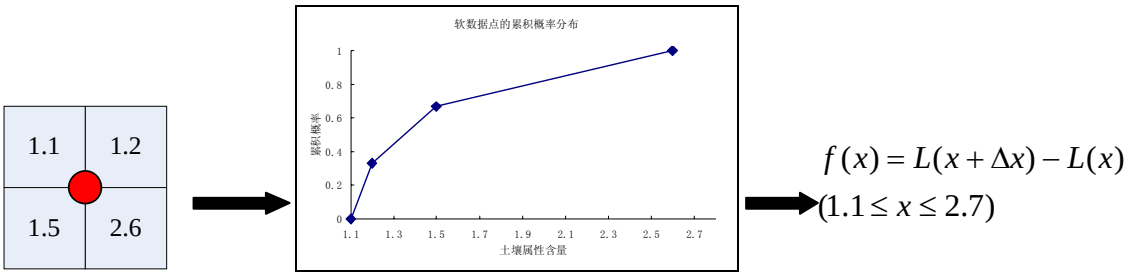
$$f_G(z_{map})=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(z_{map}-\mu)^2}{2\sigma^2}}$$

（其中 μ 为普通克里格预测结果，σ^2 为预测方差）。

3.2 软数据获取及表达

软数据表示了被预测变量的模糊分布，可用概率密度函数或在一定范围内的均一分布来表达。本文中软数据来源于 2005 年的属性分布图，表达了历史分布状况对现在的影响，这种影响存在着不确定及模糊性，因此，本文将 2005 年预测分布图中相邻的 $n\times n$ 个栅格的中心作为软数据点的位置，将 $n\times n$ 个栅格中对应的含量值由小到大排序，由排序结果得出软数据点土壤属性的累积概率分布函数，再由累积概率分布函数得到软数据点土壤属性含量的

分布概率密度，其过程如图 4 所示（本例图中 $n=2$）：



（a）软数据点　　（b）累积概率分布图　　（c）各含量的概率密度函数

图 4：软数据获取过程

Fig.4 calculate method of soft data

根据上述软数据获取方法，本文中取 $n=10$，利用 2005 年所得预测结果（原图每个栅格的大小为 $100m\times100m$），生成了 243 个软数据点，其分布如图 5 所示：

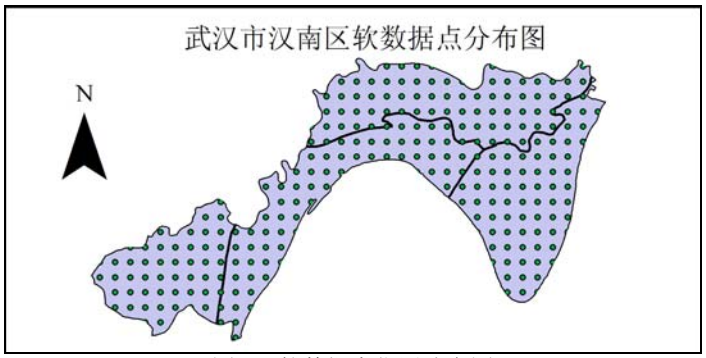


图 5：软数据点位置分布图

Fig.5　Soft data points in Hannan

3.3 后验 pdf 计算及预测结果

根据贝叶斯条件概率公式，考虑硬数据和软数据，修正先验 pdf，则定义变量 Z 在预测位置 x_0 处的后验 pdf 为：$f_K(z_0)=f_G(z_0\mid z_{hard},z_{soft})=\dfrac{f_G(z_0,z_{hard},z_{soft})}{f_G(z_{hard},z_{soft})}$。其中 $z_{hard}=[x_1,\cdots\cdots,x_h]'$，$z_{soft}=[x_{h+1},\cdots\cdots x_m]'$，$h$ 为待预测点周围变程范围内硬数据个数，$m-h$ 为软数据个数。本研究中，软数据以 pdf 的方式给出，则 $f_k(z_0)=\dfrac{\int f_G(z_0,z_{hard},z_{soft})f_S(z_{soft})dz_{soft}}{\int f_G(z_{hard},z_{soft})f_S(z_{soft})dz_{soft}}$ 为本方法最后得出的未测点土壤连续属性含量 pdf，一般为非高斯分布。它描述了待估计位置变量的完全分布特征，取其数学期望 $\hat{z}_k=\int f_k(z_k)z_k dz_k$ 作为预测结果。

4 结果分析

基于上述 BME 方法，使用三组建模采样点和软数据对土壤中有机质进行了空间预测，对比只使用三组建模采样点的普通克里格方法所得结果，如图 6 所示。

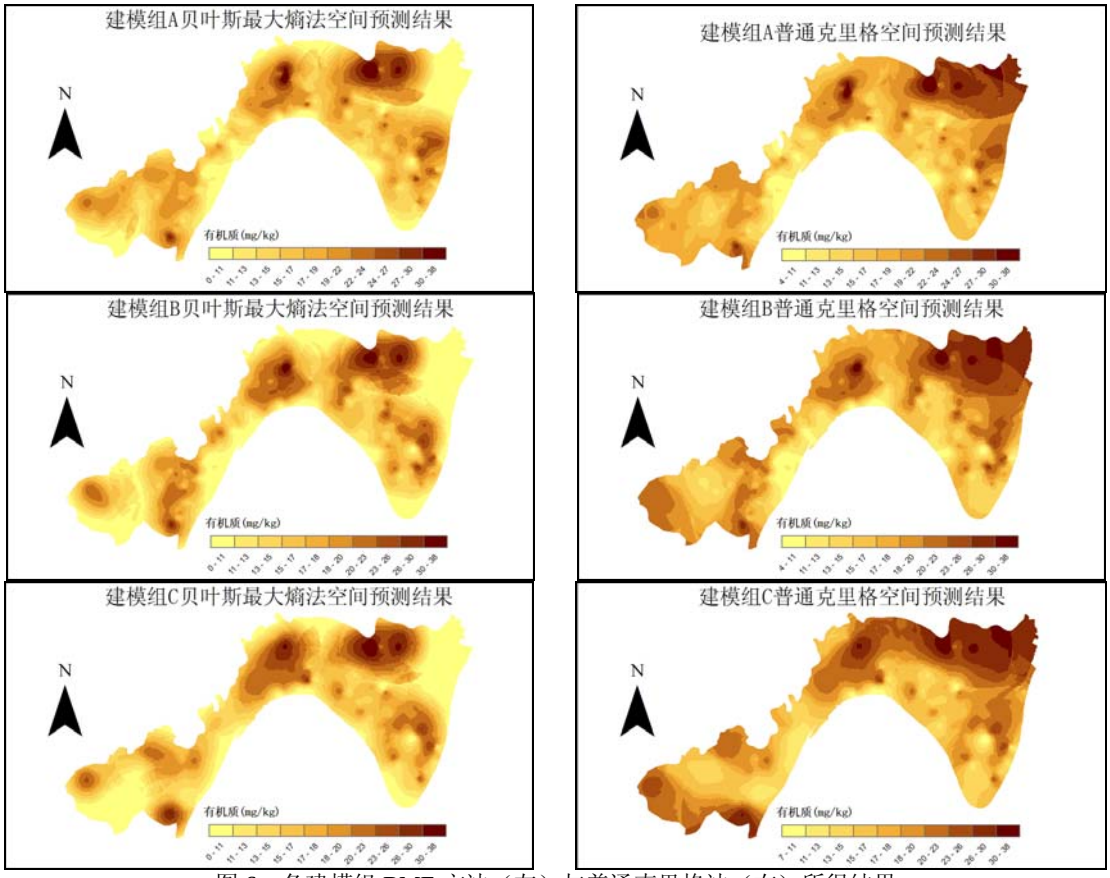


图 6：各建模组 BME 方法（左）与普通克里格法（右）所得结果

Fig.6 Result of BME(left) and OK(right) using different modeling data set

从图中可看出，两种方法所得预测结果在分布上是大致类似的，但在区域东北角和西南角差别较大，这是因为这些区域属人口聚集区，未布置采样点，而按照 BME 的计算方法，这些地方距离硬数据点较远，其值接近于 0。除这两个区域外，BME 方法所得预测结果的值域范围较广，而普通克里格法所得结果值域范围较窄，特别是 C 组建模点所得结果。从图中分布的连续性上看，BME 方法所得结果较破碎，而普通克里格法所得结果较平滑，特别是 C 组建模点所得结果存在大片同色块区域，这说明 BME 方法在采样点较少的情况也能克服平滑性。使用验证点，对比平均绝对误差（MAE），相关系数（R）和一致性系数（AC，取值范围为 0 到 1，1 代表预测值和观测值完全一致，而 0 代表预测值和观测值完全不同）三个指标，结果如表 1 所示。

表 1：MAE，R，AC 三个指标的对比结果

Table.1 Compare result of MAE, R, AC

	A 组建模点			B 组建模点			C 组建模点		
	MAE	R	AC	MAE	R	AC	MAE	R	AC
BME	2.064	0.774	0.779	2.182	0.756	0.743	2.714	0.623	0.698
Kriging	2.371	0.693	0.775	2.518	0.674	0.717	3.371	0.372	0.566

通过对比可知，BME 方法所得结果要比只使用采样点数据的普通克里格方法所得结果精确，特别是在采样密度减少到原密度的一半时，克里格法精度大大降低，而 BME 方法的精度并未明显降低（表 1）。可见，在软数据的支持下，BME 方法的预测精度对采样数据的依赖性明显弱于克里格方法。

5 总结与展望

5.1 总结

本文在对土壤属性进行空间预测时，综合使用采样数据和过去的预测结果信息，以样点数据 Kriging 插值的正态分布结果作为预测位置属性的先验概率分布，用累积概率分布函数转换的方法，将一片区域的历史信息转化为区域中心点上的模糊分布信息，以此来获得软数据。最后将贝叶斯条件概率公式运用于先验概率分布和软数据之上，得到预测位置属性的后验概率分布，并以其数学期望作为最终预测结果，该结果既体现当前采样点之间的空间相关性，又体现土壤属性空间分布的时序稳定性。通过最后与 Kriging 方法的对比结果可知，本文所提方法有较高的预测精度，且在对采样数据的依赖性上明显弱于 Kriging 方法。

5.2 BME 方法的优势

作为一种非线性空间插值方法，BME 比经典地统计学方法在多源数据利用和预测精度方面有着一定的优势：

（1）在数据利用方面，当前大多数地学统计模型都只能处理硬数据，而不能直接有效地处理和利用软数据，对软数据的利用是将其“硬化”为硬数据，然后将其作为辅助信息集成到空间变量进行插值估计或将软数据作为分类或者分层的根据对采样点进行分类，如协同克里格[11]、分层克里格[12,13]、具有局部先验值的指示克里格等[14]。这样的利用方式无疑减少了软数据所带来的信息量，弱化了软数据对空间变量的影响。而 BME 作为一种非线性估计方法，可以将硬数据和软数据集成到一起，计算目标变量值分布的 pdf，从而实现空间预测和不确定行评价的目的，整个过程中软数据不需要被“硬化”，也不止是一个分类依据。

（2）在软数据的支持下，BME 能够在一定程度上提高空间预测精度，特别是在采样点较少时。因此，在辅助信息完备的情况下，使用 BME 方法能够节约采样成本，而不影响预测精度。

（3）在软数据的来源方面，BME 方法具备更多的开放性，只要符合方法给出的软数据形式，都可以被用来进行空间预测。因此，只要是和被预测属性有关的信息，都有机会被作为软数据来利用，如与被预测属性有关的环境信息、各种专家经验、其他相关性强的土壤属性等。

（4）适用面广，结果具有传承性：计算结果为每个被预测位置属性的 pdf，根据不同的制图目的，可依据 pdf 提取不同的值赋予每个点作为土壤专题图件中像素值，如：pdf 中最大概率处的值；数学期望（$\hat{z}_k = \int f_k(z_k) z_k dz_k$）；超过或小于某个阈值的累积概率等。达到空间预测或不确定性评价的目的，体现了方法适用面的广泛。而且本次的研究结果，可作为软数据被下一时期同一地域的类似研究所用，亦可为相似地理环境条件的其他地域所用，即结果具有传承性。

5.3 展望

本文以土壤连续属性为例，将 BME 方法的利用方式引入国内，并希望近期在以下方面进行进一步探索：

（1）探索如何将更多的相关信息（特别是环境信息）转化为软数据，并形成一套空间软信息自动生成算法。

（2）在空间不确定性评价方面，将 BME 方法与其他方法（如指示克里格、序贯指示

模拟（SIS）和序贯高斯模拟（SGS））进行对比。

（3）将 BME 方法运用于更多的自然资源与环境或社会经济领域的研究中去。

[参考文献] (References)

[1] Orton, T.G., R.M. Lark. Account for the uncertainty in the local mean in spatial prediction by Bayesian Maximum Entropy[J]. Stochastic Environmental Research and Risk Assessment, 2007, 21(6):773-784.

[2] Lee, S.J., R. Balling, P. Gober. Bayesian Maximum Entropy Mapping and the Soft Data Problem in Urban Climate Research[J]. Annals of the Association of American Geographers, 2008, 98(2):309-322.

[3] 檀满枝，陈杰，徐方明等. 基于模糊集理论的土壤重金属污染空间预测[J]. 土壤学报，2006，43(3)：390-396.

[4] Goovaerts, P. Geostatistics for Natural Resources Evaluation[M]. New York: Oxford University Press. 1997.

[5] Wang, G., G. Gertner, P. Parysow. Spatial prediction and uncertainty analysis of topographic factors for the revised soil loss equation (RUSEL) [J]. Soil Water Conservation, 2000, 55(3):374-384.

[6] 史舟，李艳，程街亮，水稻土重金属空间分布的随机模拟和不确定评价[J]. 环境科学，2007， 28(1)：209-214.

[7] Phillips, S.J., R.P. Anderson, R.E. Schapire. Maximum entropy modeling of species geographic distributions[J]. Ecological Modelling, 2006, 190(3-4):231-259.

[8] Christakos, G. A Bayesian/Maximnu-Entroy View to the Spatial Estimation Problem[J]. Mathematical Geology, 1990, 22(7):763-777.

[9] Christakos, G. Modern Spatiotemporal Geostatistics[M]. New York: Oxford University Press. 2000.

[10] 张贝，李卫东，杨勇，汪善勤，蔡崇法. 贝叶斯最大熵地统计学方法及其在土壤和环境科学上的应用状况。土壤学报，2011，48（4）。

[11] Liu, T.L., K.W. Juang, D.Y. Lee. Interpolating Soil Properties Using Kriging Combined with Categorical Information of Soil Maps. Soil Science Society of America Journal, 2006, 70(4):1200-1209.

[12] Lagacherie, P., M. Voltz. Predicting Soil Properties over a Region Using Sample Information from a Mapped Reference Area and Digital Elevation Data: A Conditional Probability Approach[J]. Geoderma, 2000, 97(3-4):187-208.

[13] Lyon, S.W., A.J. Lembo, M.T. Walter. Defining Probability of Saturation with Indicator Kriging on Hard and Soft Data[J]. Advances in Water Resources, 2006, 29(2):181-193.