Kriging 算法在含水量三维属性模型构建中的应用研究*

姚凌青[1]，潘 懋[1]，屈红刚[1]，兰向荣[1]，丛威青[1]，刘学清[2]，于春林[2]
(1. 北京大学 造山带与地壳演化教育部重点实验室，北京 100871；2. 北京市地质矿产勘查开发局，北京 100050)

摘 要：Kriging 算法通过构造区域化变量的变异模型，并据此求取未知数据的最优线性无偏估计量。结合构建工程地质体含水量参数三维模型的应用，强调在三维环境下分析样本的空间变异结构特征，以 K-D 树建立样本数据的空间索引快速搜索插值邻域。针对插值过程中存在的负权值问题，采用线性规划的方法加以解决。研究表明，该算法综合考虑了各向异性以及不同变异尺度对算法的影响，比基于分层的二维 Kriging 方法更为合理。

关键词：Kriging 算法；各向异性；变异尺度；负权值；K-D 树；线性规划；含水量
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2008)08-2554-03

Research on application of Kriging in construction of three-dimensional property model about water content

YAO Ling-qing[1], PAN Mao[1], QU Hong-gang[1], LAN Xiang-rong[1], CONG Wei-qing[1], LIU Xue-qing[2], YU Chun-lin[2]
(1. Key Laboratory of Orogenic Belts & Crustal Evolution of Education Ministry, Peking University, Beijing 100871 China; 2. Beijing Bureau of Geology & Mineral Prospecting, Beijing 100050, China)

Abstract: Kriging algorithm builds varation model of regional variables and through which seeks for best linear unbiased estimator (BLUE) of unknown property. This paper combined with application of constructing 3D property model of water content within a certain engineering geological body, emphasized analyzing variation of regional variable under 3D circumstance, and adopted K-D tree to acquire neighborhoods quickly. Resolved negative weights during interpolation by linear program. The results show that 3D Kriging behaves better than 2D Kriging based on dividing layer for consideration of anisotropy and different scales on variation model.

Key words: Kriging algorithm; anisotropy; variation scale; negative weight; K-D tree; linear program; water content

0 引言

地质统计学在矿产勘查、石油勘探、环境监测、图像处理等科学领域都有着广泛的应用。Kriging 插值方法作为地质统计学中的核心算法，是一种综合考虑变量的结构性和随机性，对未采样点属性进行最优无偏估计的估值方法。

迄今为止，Kriging 方法在二维场合的应用非常普遍，相比之下三维环境下的实际应用较为少见。理论上地质统计学可以在任意的完备度量空间进行，当然也就包括了三维欧氏空间。而之所以出现 Kriging 方法在三维空间中应用的不足，究其原因，笔者认为可以归纳为三点：a) 二维形态的研究对象涵盖了大量的应用场合，如研究煤层在分布区域上的厚度变化；b) 对不同层位的研究区进行二维 Kriging 插值，形成三维空间的一系列截面，可间接地表达对象属性分布的三维形态；c) 计算机硬件条件的制约使三维地质信息系统发展相对滞后，间接影响了地质统计学在三维空间的应用。

利用 Kriging 方法构建研究对象的属性场，相关的文献根据应用领域的不同，研究的侧重点也有所差异[1~3]。文献[1]以油层温度场为研究对象，侧重分析回归模型和相关模型对 Kriging 算法的影响，变异结构以二维分析为主，插值结果通过曲面表示；文献[2]通过 Kriging 方法构造二维的钻孔温度等深面，通过等深面叠加形成规则体元数据并进行体视化；文献[3]利用 Kriging 方法建立水文地质层三维模型，思路与文献[2]类似，侧重于方法的实现过程以及体视化的表达。

上述文献在实现思路上仍然以二维空间的 Kriging 算法为主。本文以建立工程地质体含水量参数三维模型的应用为契机，重点研究 Kriging 方法在三维空间中的拓展，深入探讨随之产生的变异结构各向异性特征、变异尺度对算法的影响、三维邻域搜索算法以及插值运算中产生的负权值问题。

1 Kriging 算法思想

Kriging 算法的前提条件是构造区域化变量的变异模型，变异模型通过变异函数或协方差函数描述。算法的核心致力于满足无偏和最小估计方差的要求，求取局部邻域内样品的权值系数，数学上等价于一个求解条件极值的问题，通过附加一个拉格朗日乘数可列出求解权值的线性方程组，称为 Kriging 方程组。方程组的系数矩阵借助协方差函数确立。可以证明，在二阶平稳条件下，协方差函数和变异函数存在简单的线性关系[4]，因此 Kriging 方程组可以方便转换为以变异函数值来表

收稿日期：2007-09-17；修回日期：2007-12-08 基金项目：北京市多参数立体地质调查项目(200313000045)

作者简介：姚凌青(1980-)，男，江西余干人，博士，主要研究方向为三维地学建模、地质统计学(old_yao@163.com)；潘懋(1954-)，男，内蒙古兴和人，教授，博导，主要研究方向为信息地质；屈红刚，男，陕西澄城人，博士后，主要研究方向为三维 GIS、三维地质建模.

达系数矩阵。在协方差函数不存在的条件下, 根据变异函数建立的 Kriging 方程组依然成立。基于变异函数在 Kriging 算法中的通用性, 本文选择变异函数为工具对区域化变量进行变异结构分析。

2　Kriging 算法在含水量参数三维模型中的应用

实例的原始数据来源于工程地质体钻孔资料, 该地质体主要成分为粉粘土。地质体分布为薄层状结构, 水平方向展布约 16 ×12 km^2, 垂直方向的厚度仅有 5 ～10 m, 地势走向基本上为西高东低。展布形态如图 1 所示。

根据钻孔资料取样得到含水量参数的采样数据(计量单位为百分数), 利用 Kriging 插值结果可以构建出含水量参数在整个工程地质体分布的三维模型。通过分析地质体形态以及数据分布的特点, 将地质体离散化为分辨率 128 ×128 ×128 的规则体元数据。

2.1　采样数据特点

采样数据来源于工程钻孔资料, 共计 48 口钻孔, 取样数共计 345 个。钻孔分布在水平方向上分布稀疏, 间隔在 800 ～1 200 m, 垂向上采样较为稠密, 间距在 0.2 ～1 m。其分布形态如图 2 所示。参数分布近似为正态分布, 符合 Kriging 插值方法应用的条件。

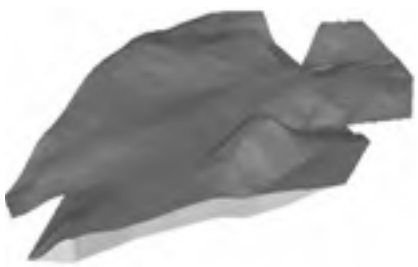

图 1　工程地质体形态图
(垂向放大比率 80∶1)

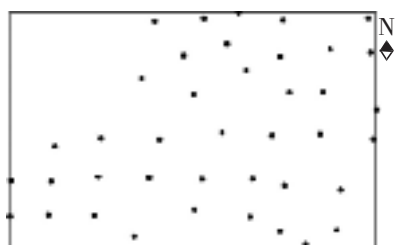

图 2　钻孔平面分布形态

2.2　应用流程与关键问题

本节重点论述 Kriging 方法构建工程地质体属性三维模型的流程以及存在的关键问题。在此基础上探讨其与常见二维场合中 Kriging 方法的差异, 并通过实验进行对比与分析。

2.2.1　变异结构分析

综合考虑方向和尺度因素对区域化变量变异性的影响, 获取一个相对合理的变异函数, 称为变异结构分析。

1) 各向异性分析　经过对不同方向的采样数据计算形成实验变异函数(式(1)), 发现实验数据在水平方向上的变异表现为各向同性, 在垂向上的变异趋势与水平方向不同。

$$\gamma^*(h) = 1/(2N(h)) \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \qquad (1)$$

2) 变异尺度分析　在进行变异分析的过程中务必要考虑变量在不同尺度上发生的变异, 而不能笼统地采用一种尺度对变异结构进行衡量。这一点可以借助采样理论[5] 加以解释, 低频的采样会丢失小尺度的变化。

本文将水平方向上的实验变异函数滞后距设为 800 m, 距离容许范围为 200 m; 垂直方向滞后距设为 0.5 m, 距离容许范围为 0.2 m; 不同方向上的角度容差范围都设为 15°角。对两个方向上的实验变异函数采用球形理论模型拟合, 拟合结果如图 3、4 所示。拟合后的理论模型在水平方向上为

$$r(h) = 13.51 + 12.01 \times sph(3.0 \times 10^3\ \text{m})$$

垂直方向的理论模型为

$$r(h) = 26.2 \times sph(4.48\ \text{m})$$

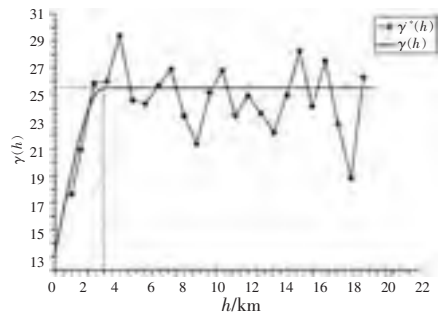

图 3　水平方向实验变异函数图

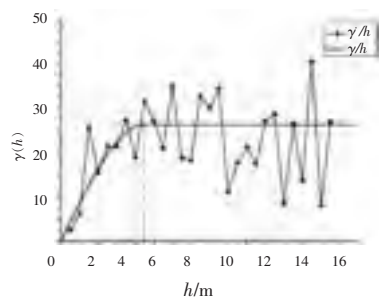

图 4　垂直方向实验变异函数图

2.2.2　基于 K-D 树索引搜索三维邻域

K-D 树索引最初由 Bentley 提出[6], 用于解决数据库检索的问题。K-D 树本质上是二叉树。具体而言, K 维空间数据每个维度分别对应一个关键字作为空间划分的标准, 小于该关键字的归类到左子树, 否则归类到右子树。依次按维度循环进行上述流程, 直到叶子节点包含的数据小于指定的个数则剖分完成。图 5 为一简单的二维空间 K-D 树数据组织示意图, 图中的矩形单元代表其中的节点。

文献[7] 是改进的 K-D 树, 选择节点中对应划分维度的关键字的中位数作为分界点, 可以优化索引的搜索效率。本文采用上述改进的 K-D 树作为采样数据邻域搜索的索引。构建 K-D树的算法可参考文献[7]。

插值的邻域设置为以不同轴向的变程为半径的长方体区域, 窗口搜索的算法描述如下(C ++ 语言):

```
struct KDNode_Base
{ enum { NODE, LEAF} tp; } //节点类型
struct KDNode : public KDNode_Base
{
size_t dividing_dim; //当前划分维度
float dividing_val; //当前划分界点
struct KDNode* lchild; //左子节点
struct KDNode* rchild; //右子节点
};
struct KDLeaf : public KDNode_Base
{ Point* pnts;} //样本数据
//邻域搜索函数主体
Range_Query( KDNode_Base* kdn, float lo[3], float hi[3])
{
if( kdn->tp = =KDNode_Base::LEAF){
//将叶子节点中满足要求的数据提取出来
GetNeighbors( KDBLeaf* (kdn), float lo[3], float hi[3]);
} else{ KDNode* tmp = (KDNode*) kdn;
size_t dim = tmp->dividing_dim;
float dval = tmp->dividing_val; //遍历左子节点
if( hi[dim] < dval ‖ lo[dim] < tmp-> dval)
Range_Query( tmp->lchild, lo, hi); //遍历右子节点
if( hi[dim] > =dval || lo[dim] > =dval)
Range_Query( tmp->rchild, lo, hi);}
}
```

2.2.3　负权值问题

关于负权值的问题, 已有相关文献进行过研究, 但在实际插值过程中却常被忽略。Kriging 算法应用的大量对象不容许出现负属性值, 如本文实例所研究的含水量参数。Kriging 插值的负权值除了可能产生无效的负属性值之外, 还可能导致属性值的畸高。

对于负权值的消除, 简单的方案可以将负权值系数置为零, 然后重新计算权值[4], 但缺乏合理的理论依据。Herzfeld[8] 提出基于 Kuhn-Tucher 条件利用二次规划解决负权值问题的方案, 但计算较为复杂。胡小荣[9] 提出利用线性规划解决权值非负约束问题的方法, 简化了运算过程, 但相对消元法求解计算复杂度依然偏高。

本文采用综合的方法:仅在属性估计值或期望方差为负时采用线性规划的方案重新求解,而其他情况则利用消元法。由于上述情形并不常见,可以节省大量的计算时间。其合理性在 2.3.3 节详细分析。

2.3 实验

2.3.1 二维分层 Kriging 方法与三维 Kriging 方法的比较

为了对比二维分层 Kriging 方法与三维 Kriging 方法,本文利用原始数据交叉验证对两者插值的结果进行了比较。采用二维分层的 Kriging 方法误差为[- 20.77, 13.49],误差绝对值均值为 3.99; 三维 Kriging 方法误差为[- 16.65, 15.50],误差绝对值均值为 3.42。显然三维 Kriging 方法从误差分布到误差平均绝对值的大小都优于二维分层的方法。进一步的比较结果如图 6 所示,可见两者的估计误差主要集中于(- 5,5),但是三维 Kriging 方法在较低误差处有更高的百分比。

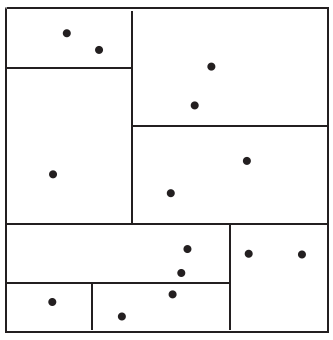

图 5 二维 K-D 树示意图

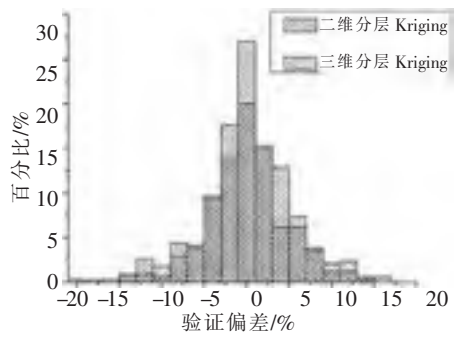

图 6 交叉验证误差比图

分析其原因,对于本实例数据,源于采样密度的缘故,参数的变异在垂直方向上有更显著的空间相关性,而在水平方向上却有明显的跃迁,存在块金效应[4],即随机性较强。按照垂直方向分层的方式对数据进行 Kriging 插值,由于未能考虑到相关性较强的垂向变化,不能充分体现 Kriging 算法根据待估计量与样本相关程度分布权值的优点。

少量数据点估计存在较大的偏差,达到 10 以上,经过与原始数据的对比观察发现,偏差较大的数据均位于边界或采样数据稀疏的地方,表明插值数据的构形对估计的结果有较大的影响,在采样数据充分的前提下三维 Kriging 方法可以取得理想的结果。

2.3.2 K-D 树搜索邻域与变异结构的关系

K-D 树在低维空间中有很好的查询性能[10],因此在二维和三维空间中都可以作为离散数据索引组织的方式。插值的结果显示,K-D 树能很好地应用于 Kriging 算法的邻域搜索中。

值得注意的是,搜索邻域范围指定应参考变异函数的变程来确定。位于变程之外的数据由于缺少相关性,对估计点的贡献不大,可排除在搜索范围之外。这样可减少 Kriging 方程组的元数,降低计算复杂度以及提高求解权值系数的稳定性。

2.3.3 负权值处理方法分析

在利用消元法求解 Kriging 方程组的过程中,负权值出现的情况较为常见。通过采用简单将负权值置零的方法或利用线性规划的方法均可消除其影响。

经进一步研究发现,负权值导致出现负的属性估计值或期望方差的情况较为少见。负权值在大多数情况下不影响插值结果的合理性,相反在这种情况下采用其他方法消除负权值反而带来更高的估计方差。而在消元法求取的插值结果无实际意义(估计值或方差小于 0) 的情况下,通过求取负权值处(消元法) 估计值的期望方差,发现线性规划求解法的方差普遍低于简单方法,说明后者在负权值处理上更为合理。对比结果如表 1 所示。

表 1 负权值消除方法对应方差

负权值处理	方差		
	最小值	最大值	平均值
简单置零(插值结果无实际意义时消除)	53.633	65.255	60.413
线性规划(插值结果无实际意义时消除)	22.519	30.504	28.205
消元法不作处理(插值结果有意义)	21.087	54.547	28.053
线性规划(插值结果有意义)	21.214	54.868	28.732

综上所述,在常规方法求解估计值无意义的前提下,利用线性规划消除负权值的影响对插值结果有明显的改进,比简单将负权值置零的方法表现出更低的方差。在权值都为正数的情况下,线性规划和常规方法取得相同的权值系数[9]。在其他情况下,线性规划消除负权值反而会带来略为更高的方差(插值结果相近),简单置零的方法尤为显著。

由于出现无意义插值结果的情况仅为少数,而线性规划运算量高于消元法。将两者综合运用,仅在消元法求解不能满足实际要求时采用线性规划方法,可以节省大量计算时间。测试硬件配置为 Mobile DualCore Intel Merom, 1866 MHz, 2 GB 内存, 显卡 NVIDIA GeForce 7900 GS (512 MB) 。本实例对比结果有: 消元法为 21 s; 线性规划法为 101 s; 综合法为 22 s. 可见综合法耗时与消元方法大致相同,但是比单纯利用线性规划法求解要节省约 4 倍的时间。

2.3.4 三维模型结果

综合利用前文提到的方法及分析结果,通过 Kriging 算法插值建立了工程地质体含水量的最终三维模型如图 7 所示。

为了探明含水量是否与地势存在一定关系,在模型体视化时有意加大了纵向的拉伸比例(150 1)。从整体趋势来看,含水量的变化与地势的走向有一定关系。地势较高的地方含水量较低,地势低洼处含水量较高,呈现出西低东高的走势。在水平方向上的变化呈区域性分布,如图 8 所示,代表了含水层不同物源迁移聚集的结果。

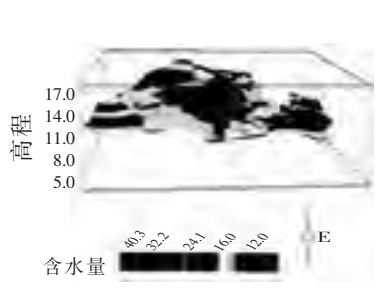

图 7 工程地质体含水量三维模型图 (z 轴适当放大, 三维效果更好)

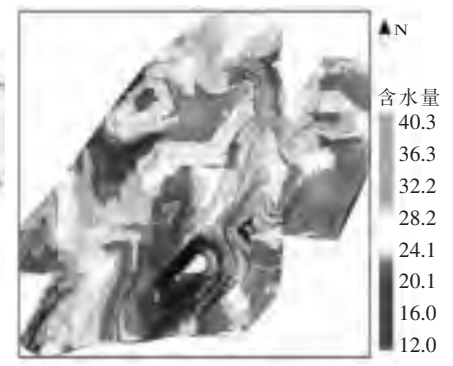

图 8 工程地质体含水量平面分布图

3 结束语

本文的研究表明,在三维环境下应用 Kriging 算法,可以充分发掘样本数据的空间变异特征,比传统二维分层的 Kriging 方法更具优势,在应用的过程中需要考虑各向异性以及观测尺度对算法的影响。本文选择 K-D 树索引用于离散样本数据的近邻搜索,搜索的范围参考不同方向的变程确定,对于常规的二维和三维数据均可以采用。Kriging 方程组求解中存在的负权值问题常被忽略,本文进行了详细的分析——利用线性规划可以消除负权值的影响,在负权值足以引起无实际意义的属性估计值或方差时,可以取得较好的结果; 反之其合理性有待商榷,计算时间也大大增加。关于负权值的数学意义还有待进一步的研究。最终的插值结果显示,三维 Kriging 方法用于三维工程参数属性模型的构建可以取得较好的效果。(下转第 2560 页)

结果分析:

a) 当 threshold 参数取 14 ~17 时,系统性能稳定在最优点。

b) 该参数的最优点取值随着特征数量的增减而上下浮动,大致范围在 50 以下。

7.5 实验组 5

本组实验对 Winnow 算法在线更新特点进行了评价。该组实验为 Winnow 算法应用中在线更新间隔选取提供了依据。

本组实验正常训练样本取 450 条,垃圾样本取 430 条;正常测试样本取 733 条,垃圾样本取 144 条;特征数量取 300,模型训练中 α取 8, θ取 10。更新测试中 α取 1.5, θ取 10。结果如表 2 所示。

表 2 更新间隔调整实验结果

参数	非更新测试	更新间隔:1	更新间隔:2
p_s	0.989 796	0.978 873	0.978 571
p_n	0.939 666	0.993 197	0.990 502
r_s	0.673 611	0.965 278	0.951 389
r_n	0.998 636	0.995 907	0.995 907
accuracy	0.945 268	0.990 878	0.988 598
F-score	0.801 653	0.972 028	0.964 789
details	*s_*97 *s_n* 47 *n_s* 1 *n_n* 732	*s_s* 139 *s_n* 5 *n_s* 3 *n_n* 730	*s_s* 137 *s_n* 7 *n_s* 3 *n_n* 730

更新间隔:3	更新间隔:4	更新间隔:5	更新间隔:10
0.978 261	0.978 571	0.978 417	0.984 962
0.987 821	0.990 502	0.989 16	0.982 527
0.937 5	0.951 389	0.944 444	0.909 722
0.995 907	0.995 907	0.995 907	0.997 271
0.986 317	0.988 598	0.987 457	0.982 896
0.957 447	0.964 789	0.961 131	0.945 848
s_s 135 *s_n* 9 *n_s* 3 *n_n* 730	*s_s* 137 *s_n* 7 *n_s* 3 *n_n* 730	*s_s* 136 *s_n* 8 *n_s* 3 *n_n* 730	*s_s* 131 *s_n* 13 *n_s* 2 *n_n* 731

更新间隔:15	更新间隔:20	更新间隔:25
0.984 375	0.984	0.983 333
0.975 968	0.972 074	0.965 654
0.875	0.854 167	0.819 444
0.997 271	0.997 271	0.997 271
0.977 195	0.973 774	0.968 073
0.926 471	0.914 498	0.893 939
s_s 126 *s_n* 18 *n_s* 2 *n_n* 731	*s_s* 123 *s_n* 21 *n_s* 2 *n_n* 731	*s_s* 118 *s_n* 26 *n_s* 2 *n_n* 731

表中: *s_s* 表示垃圾短信被分为垃圾短信的数量; *s_n* 表示垃圾短信被分为正常短信的数量; *n_s* 表示正常短信被分为垃圾短信的数量; *n_n* 表示正常短信被分为正常短信的数量。

结果分析:

a) 本组实验参数的选择是为了便于实验更新间隔,在错误率较高的情况下考察更新间隔调整对性能的影响。

b) 实验中的更新间隔为错误反馈个数。

c) 由实验结果可以看出,使用在线更新功能,能及时地修正模型,大大地减小错误率。

d) 更新间隔增大,垃圾判错率增高,但是正常短信召回率变化不大,说明系统在保证正常短信不丢失方面表现稳定。

e) 最优更新间隔与语料规模、参数选择有关,在当前实验规模下可以选择逐条错误反馈的方式。

8 结束语

本文提出了采用 Winnow 核心算法的嵌入移动终端式短信过滤系统。该算法具有分类速度快、性能好、支持在线更新的特点,对于终端上的过滤系统有着良好的应用前景。经过大量的实验和分析,本文对该算法在过滤系统中的应用提供了可靠的依据和参考。在未来的工作中,特征回退模块、特征提取模块等均有着较大的改进潜力;另外还在系统中添加用户定制类的规则控制模块,以取得更优良的应用效果。

参考文献:

[1] AUL G. A plan for spam [EB/OL]. (2002-12-10). http://www.Paulgraham.com/spam.html.

[2] 张国华. 汉语自动分词和词性标注方法研究及系统实现[D]. 北京: 中国科学院研究生院, 2007.

[3] 李凡, 鲁明羽, 陆玉昌. 关于文本特征抽取新方法的研究[J]. 清华大学学报: 自然科学版, 2001, 41(7): 98-101.

[4] SALTON G, WANG A, YANG C. A vector space model for information retrieval [J]. Journal of the American Society for Information Science, 1975, 18: 613-620.

[5] LITTLESTONE N. Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm [J]. Machine Learning, 1988, 4(2): 285-318.

[6] 王斌, 潘文峰. 基于内容的垃圾邮件过滤技术综述[J]. 中文信息学报, 2005, 19(5): 1-10.

(上接第 2556 页)

参考文献:

[1] 杜宇健, 萧德云. Kriging 算法在温度场计算中的应用分析[J]. 计算机辅助设计与图形学学报, 2004, 16(8): 1153-1158.

[2] 尚庆生, 郭建文, 李新, 等. 基于 Kriging 插值的钻孔地温数据体视化[J]. 遥感技术与应用, 2006, 21(4): 302-306.

[3] 颜辉武, 祝国瑞, 徐智勇, 等. 基于 Kriging 水文地质层的三维建模与体视化[J]. 武汉大学学报: 信息科学版, 2004, 29(7): 611-614.

[4] 孙洪泉. 地质统计学及其应用[M]. 徐州: 中国矿业大学出版社, 1990.

[5] BENEDETTO J, PAULO J S, FERREIRA G. Modern sampling theory: mathematics and applications [M]. Boston: Birkhauser Boston, 2001.

[6] BENTLEY J L. Multidimensional binary search trees used for associative searching [J]. Communications of the ACM, 1975, 18(9): 509-517.

[7] FRIEDMAN J H, BENTLEY J L, FINKEL R A. An algorithm for finding best matches in logarithmic expected time[J]. ACM Trans of Mathematical Software, 1977, 3(3): 209-226.

[8] HERZFELD U C. A note on programs performing Kriging with non-negative weights[J]. Mathmatical Geology, 1989, 21(3): 391-392.

[9] 胡小荣. 一种考虑权值非负约束的克立格算法[J]. 地质与勘探, 1999, 35(4): 29-32.

[10] CHEN Y S, HUNG Y P, YEN T F, *et al*. Fast and versatile algorithm for nearest neighbor search based on a lower bound tree [J]. Pattern Recognition, 2007, 40(2): 360-375.