文章编号：0253-2697(2017)12-1425-09 DOI:10.7623/syxb201712010

基于 p-stable LSH 的多点地质统计建模算法

喻思羽[1] 李少华[1] 王端平[2] 王 军[2] 张以根[2] 于金彪[2]

(1. 长江大学地球科学学院 湖北武汉 430100； 2. 中国石油化工股份有限公司胜利油田分公司勘探开发研究院 山东东营 257015)

摘要：SIMPAT 将图像重建思想引进储层地质建模领域，借助于弱化概率的相似性判别指标，用最相似地质模式替换待估点处的数据事件完成预测。当模型较大且数据样式较多时，海量的数据样式相似度计算使得 SIMPAT 的计算效率较低。为了有效平衡多点地质统计建模算法效率和内存的矛盾，基于 SIMPAT 算法，提出基于 p-stable 局部敏感哈希的多点地质统计建模算法 LSHSIM，该方法使用局部敏感哈希将数据样式的特征向量映射到哈希表。建模时从哈希表里取出与数据事件的特征向量具有相同哈希值的数据样式，用最相似的数据样式替换覆盖待估区的数据事件完成建模。利用实例对比新算法与 SIMPAT 等现有方法的结果表明，LSHSIM 算法计算效率高，并节省了内存空间，对算法的关键参数进行了敏感性分析、非条件和条件模拟，能较好再现训练图像的先验地质模式。

关键词：储层建模；局部敏感哈希；SIMPAT；多点地质统计学；训练图像

中图分类号：TE19 **文献标识码**：A

Multipoint geo-statistical modeling algorithm based on p-stable LSH

Yu Siyu[1] Li Shaohua[1] Wang Duanping[2] Wang Jun[2] Zhang Yigen[2] Yu Jinbiao[2]

(1. *College of Geosciences, Yangtze University, Hubei Wuhan* 430100, *China*; 2 *Research Institute of Exploration and Development, Sinopec Shengli Oilfield Company, Shandong Dongying* 257015, *China*)

Abstract: Image reconstruction idea is introduced into reservoir geological modeling field by SIMPAT, of which the most similar geological mode is used to replace the data event at a to-be-estimated point for prediction using the similarity evaluation index of weakening probability. When the model is larger with more data patterns, the SIMPAT computational efficiency is lower due to the humongous data pattern similarity calculation. To effectively balance the contradiction between the efficiency and memory of multipoint geo-statistical modeling algorithm, based on the SIMPAT algorithm, the multipoint geo-statistical modeling algorithm LSHSIM based on p-stable locality sensitive hashing was put forward, and in this method, the locality sensitive hashing was used for mapping the eigenvector of data pattern on the hash table. During modeling, the data pattern with the same hash value to the eigenvector of data event was extracted from the hash table, and the most similar data pattern was applied to replace the data events at the to-be-estimated area. Through the instance-based comparison between the new algorithm and existing methods such as SIMPAT, and the new algorithm has a high computational efficiency in saving the memory space, and the sensibility analysis, non-conditional and conditional simulations are conducted on the key parameters of algorithm, which can better reproduce the prior geo-mode of training image.

Key words: reservoir modeling; local sensitive hashing; SIMPAT; multipoint geo-statistic; training image

引用：喻思羽，李少华，王端平，王军，张以根，于金彪. 基于 p-stable LSH 的多点地质统计建模算法[J]. 石油学报，2017，38(12)：1425-1433.

Cite: YU Siyu, LI Shaohua, WANG Duanping, WANG Jun, ZHANG Yigen, YU Jinbiao. Multipoint geo-statistical modeling algorithm based on p-stable LSH[J]. Acta Petrolei Sinica, 2017, 38(12): 1425-1433.

早期的储层地质建模方法主要包括两点地质统计学方法和基于目标建模方法。两点地质统计学建模方法属于基于象元的模拟方法[1]，模拟对象是单个网格节点，容易实现井数据条件化模拟。由于变差函数局限于统计空间两点的相关性，因此难以再现复杂的几何形态特征。基于目标的模拟方法[2-3]能够模拟复杂的储层结构单元(如弯曲的河道)，但是难以满足条件化模拟，尤其当地质体单元的尺寸大于井的平均间距或者需要融合多种软数据的情况[4]。多点地质统计学(MPS)建模方法结合了两点地质统计学和基于目标建模方法的优点，既能做到易于数据条件化，又可以较好地再现具有复杂几何形态的地质结构单元。在当前的油气田勘探开发中，多点地质统计建模方法越来越多被应用于研究项目，例如裂缝性储层建模[5]、辫状河储层建模[6]等，多点地质统计建模方法优点体现在较好地再现储层形态，在多数据融合方面也具有优势[7-8]。

基金项目：国家自然科学基金项目(No. 41572121)、国家重大科技专项(2016ZX05011-001)和湖北省科技创新群体项目(2016CFA024)资助。

第一作者：喻思羽，男，1987 年 6 月生，2009 年获长江大学学士学位，现为长江大学博士研究生，主要从事地质统计学算法的研究。Email: 573315294@qq.com

通信作者：李少华，男，1972 年 8 月生，1994 年获江汉石油学院学士学位，2003 年获中国石油勘探开发研究院博士学位，现为长江大学地球科学学院教授、博士生导师，主要从事地质统计学、地质建模方面的研究和教学。Email: 534354156@qq.com

空间多点统计信息属于高维数据，如何从海量的高维度数据样式中快速找到与数据事件最相似的数据样式是当前多点地质统计建模方法的研究热点。局部敏感哈希技术是海量高维数据检索技术中最流行的方法之一，拥有深厚的理论依据并且在高维数据空间中表现优异[9]。按照局部敏感哈希原理，哈希表里相似的数据样式(事件)属于同一个哈希桶的概率远大于不相似的数据样式(事件)。基于这种思想，笔者提出了一种新的多点地质统计建模算法 LSHSIM，该算法将数据样式的线性查询过程转换为哈希检索方式，能有效提高建模效率并节省内存空间。

1 多点地质统计建模算法

多点地质统计学方法由 Guardiano 和 Srivastava 最早提出，并创建了第一种 MPS 算法 ENESIM[10]，该算法每次模拟一个网格节点都需要扫描一次训练图像，模拟计算效率较低。Strebelle 等[11-12]提出 SNESIM 方法，只需扫描一次训练图像，将所有扫描到的数据事件及其概率存储在一种称为搜索树的动态数据结构中。面对几何形态复杂、相类型较多的训练图像和节点数量较多的数据样板，搜索树对内存的需求急剧增加，建立千万网格级别的精细模型非常耗时[13]。为解决 MPS 的计算效率及内存开销方面的问题，国内外学者的做法可以归纳为以下 4 种[14-18]：

(1) 不再一次模拟一个网格点，而是一次模拟由多个网格单元组成的数据样式(pattern)，也就是基于样式(pattern-based)建模，将空间多点相关性数据存于数据样式，不仅能减小内存开销，同时一次模拟多个网格点可以提高计算速度。Arpat 等[14,19]提出了基于数据样式相似度建模的 SIMPAT 算法，该方法从样式数据库里查找与数据事件最相似的数据样式，覆盖冻结数据事件的所有节点实现对未知区的预测，但是海量数据样式的相似度计算使得建模效率很低。针对 SIMPAT 算法计算效率低的问题。Zhang 等[15]提出了 Filtersim 算法，采用样式聚类和两步建模的思想提高了建模效率。Honarkhah 等[17]提出 DisPat 算法，应用多维尺度分析和 K-means 聚类对样式数据库进行聚类，极大提高建模速度，但是多维尺度分析中建立距离矩阵的内存负担较大。

(2) 模拟对象还是单个网格节点，一次模拟一个网格，但是采用不同的数据结构存放训练图像中扫描的数据事件概率。Straubhaar 等[20]提出的 IMPALA，采用列表与索引树组合结构，不仅计算更快，也降低了内存开销。

(3) 借助于数据样式的思想，但是一次还是模拟一个网格节点。如 Mariethoz 等[18]改进了 Guardiano 和 Srivastava 的 ENESIM 方法，提出了 DS 方法。该方法在扫描训练图像进行样式匹配时，不需要扫描整个训练图像，只要扫描到与数据事件最相似的数据样式就停止扫描，然后将该数据样式的中心节点赋值到模拟网格。DS 方法计算更快，但是定义样板相似度的参数难以调节。

(4) 引入新的技术(如纹理图像生成技术)。Mahmud 等[21-22]提出了基于改进 IQ(image quilting)的建模方法，借助 GPU 并行计算提高效率，但是 IQ 算法难以条件模拟。喻思羽等[23]提出一种基于样式降维聚类的多点地质统计建模算法，采用邻近等间距取样法对所有数据样式进行降维聚类处理，将相似的数据样式聚为一类，较大程度地提高计算效率，缺点是难以直接进行连续型变量的建模。除了算法层面的改进，随着计算机硬件的发展，国内外学者也提出了一些基于 CPU、GPU 并行计算的多点地质统计建模算法[24-25]。

2 基于样式的 MPS 算法

2.1 基本原理

基于样式的 MPS 将图像重建思想引入储层地质建模领域。基于样式 MPS(如 SIMPAT)与基于概率 MPS(如 SNESIM)的关键差异是基于样式的 MPS 在预测未知区域时，以预测点为中心、数据样板范围内的全部节点都参与相似度计算，将最相似数据样式的所有节点整体替换数据事件。目前主流的基于样式 MPS 有 SIMPAT、Filtersim 和 DisPat 等。

基于样式 MPS 算法的核心思想是相似度(距离)计算。相似度是一种评价对象间相似(异)性程度的指标，主要用于模式识别、机器视觉等领域。相似度与距离的关系是相似度越大，距离越小；反之相似度越小，距离越大。相似度(距离)函数以两个对象为输入变量来确定一个非负实数表示相似度(距离)。常用的相似度(距离)函数有 Minkowski 函数(范数)、Hsim 函数、皮尔森系数及 Jaccard 相似系数等。基于样式 MPS 通常用 Manhattan 距离(L_1 范数)计算数据样式(事件)之间的相似性(距离)，Manhattan 距离可表示为：

$$L_1(x,y)=\sum_{i=1}^{d}|x_i-y_i| \tag{1}$$

如图 1 所示，图 1(f)是数据样式 x 与 y 的 Manhattan 距离计算原理，距离值等于 x 与 y 的同样位置节点的差的绝对值之和。x 和 y 的距离加 1 求倒数即为相似度。距离越大，相似度越小，相似度取值在 0～1，其表达式为：

$$s(x,y)=1/[d(x,y)+1] \tag{2}$$

图 1(a)—图 1(e)是 5 组数据样式的 Manhattan 距离。图 1(a)中的数据样式 x 和 y 的距离等于 1，图 1(c)中的数据样式 x 和 y 距离为 9，在 5 组数据中相似性最小。

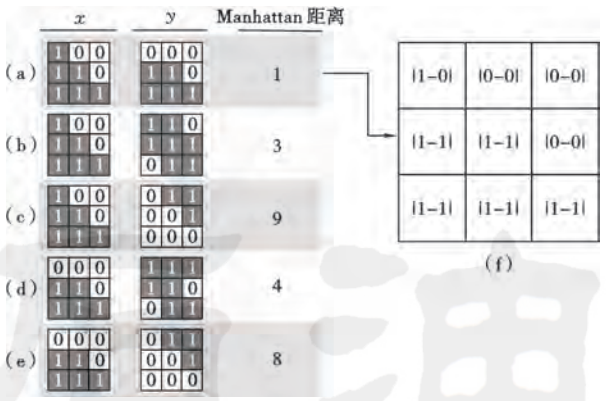


图 1　基于 Manhattan 距离的数据样式相似度计算原理[14]

Fig. 1　Data pattern similarity calculation principle based on Manhattan distance

SIMPAT 算法属于序贯模拟方法,基本思想是基于相似度(距离)实现储层地质模型的重建。该算法用数据样板扫描训练图像获取数据样式,作为样式数据库存储于计算机内存;建模时计算数据样式与数据事件的相似度,从样式数据库中查询与数据事件最相似的数据样式,用该数据样式替换数据样式直到模拟完工区全部网格节点。

Filtersim 改进了 SIMPAT,使用不同方向的线性过滤器计算数据样式的加权得分,然后依据加权得分对数据样式进行聚类分析。相似数据样式分为同一类,同一类数据样式的均值称为原型模型(prototype)。Filtersim 建模过,程分 2 步:①计算与数据事件最相似的原型模型;②从原型模型对应的样式聚类中得到最相似的数据样式,覆盖数据事件完成模拟。

DisPat 算法的特点是采用多维尺度分析(MDS)、高斯核变换技术降低数据样式的维度,进而用 K-means 聚类将数据样式聚成多个类,同一类数据样式的相似性高。与 Filtersim 相同,建模时先对比待估点的数据事件与样式聚类的原型模型,然后从原型模型的样式聚类中得到最相似的数据样式,覆盖数据事件完成模拟。

2.2　存在问题

传统 MPS 算法在计算效率和内存占用两方面各自存在不足之处。SIMPAT 计算效率与训练图像的数据样式数量密切相关,数据样式的数量越多,建模的时间越长。实际油田建模应用中,模型网格的节点数量可达千万级别,SIMPAT 建模的时间长达数小时乃至数天。基于过滤器进行样式聚类的 Filtersim 比 SIMPAT 计算效率高,但 Filtersim 依赖于数据样式分类的线性过滤器,对不同训练图像选择合适的过滤器较为困难。DisPat 比 Filtersim 和 SIMPAT 的计算效率有了极大提高,但是数据样式的相似度(距离)矩阵要求极大的内存开销。虽然 DisPat 采用了等间距抽样方式减少数据样式的数量,一定程度上降低内存开销,但是排除数据样式导致丢失部分空间多点统计信息。

为实现在合理内存开销的情况下提高 MPS 的计算效率,笔者基于 p-stable LSH 提出新的多点统计算法 LSHSIM,将耗时的数据样式相似度计算过程转换为局部哈希检索计算过程。

3　基于“Block Grid”计算地质模式的特征向量

LSHSIM 建模前,必须将数据样式转换为适于局部敏感哈希映射的计算形式。笔者提出“分块网格(block grid)”策略提取数据样式的特征向量——统计分块网格的每个“块(block)”内的所有节点之和,进而将特征向量散列至哈希表。

图 2 展示了基于“分块网格”提取数据样式(事件)特征的原理。图 2(a)—图 2(c)是 3 个数据样式,图 2(d)是数据事件,图 2(e)的蓝色网格代表分块网格的网格结构。分块网格中每个块的值等于块内所有节点的和:

$$B_{\mathrm{Grid}}(m,n)=\mathrm{Sum}(G_{i,j}) \tag{3}$$

$$M_{\mathrm{Size}}=\mathrm{INT}\left(\frac{I_{\mathrm{Count}}}{M_{\mathrm{Count}}}\right),N_{\mathrm{Size}}=\mathrm{INT}\left(\frac{J_{\mathrm{Count}}}{N_{\mathrm{Count}}}\right) \tag{4}$$

其中,$m\in[0,M_{\mathrm{Count}}-1]$,$n\in[0,N_{\mathrm{Count}}-1]$,$i\in[m\cdot M_{\mathrm{Size}},(m+1)\cdot M_{\mathrm{Size}}]$,$j\in[n\cdot N_{\mathrm{Size}},(n+1)\cdot N_{\mathrm{Size}}]$。

如图 2 所示,数据样板的 I_{Count} 和 J_{Count} 均为 9,M_{Count} 和 N_{Count} 都为 3,M_{Size} 与 N_{Size} 等于 3。图 2(a)的数据样式与图 2(b)的数据样式的几何形态非常相似,相应地,二者的特征向量(表 1)也极为相近;反之,图 2(a)的数据样式与图 2(c)的数据样式的形态差异较大,二者的特征向量也有明显差异。

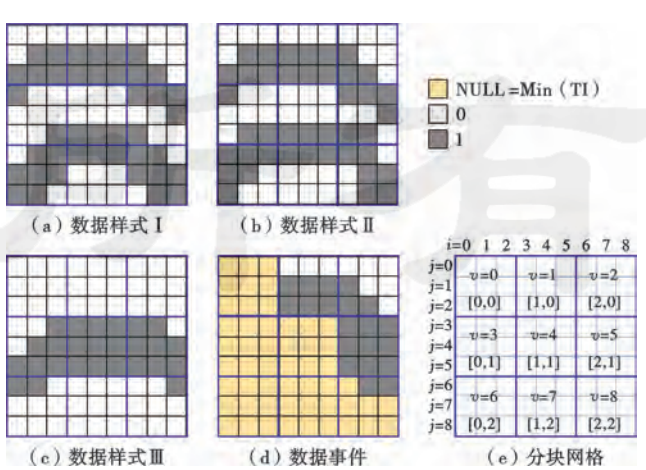


图 2　基于分块网格计算数据样式(事件)特征向量的原理

Fig. 2　Principle of calculating feature vector of pattern(data event) based on block grid

表 1 基于分块网格的数据样式(事件)特征向量

Table 1 Feature vector of pattern(data event) based on block grid

数据样式(事件)	基于分块网格的特征向量 v								
	$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$
数据样式Ⅰ	5	6	3	2	3	5	8	3	7
数据样式Ⅱ	5	6	3	3	3	6	6	3	6
数据样式Ⅲ	0	0	0	6	9	8	2	0	1
数据事件	0	6	3	0	0	9	0	0	2

分块网格的线性形式称为特征向量 $\boldsymbol{v}$,是局部敏感哈希计算的输入变量:

$$\boldsymbol{v}_i = B_{\text{Grid}}(m,n) \tag{5}$$

其中,i 是分块网格的线性索引,m 和 n 是分块网格的二维索引,$i = m \cdot N_{\text{Count}} + n$。

建模过程中,数据事件的节点可能是空值(NULL)。此时以训练图像变量的最小值替代空值:

$$c = \begin{cases} \text{Min(TI)} & c = \text{NULL} \\ c & c \neq \text{NULL} \end{cases} \tag{6}$$

4 LSHSIM 算法的基本原理

4.1 基于 p-stable LSH 检索地质模式

在多点地质统计学领域中,数据样式是多个空间点组合结构,用于反映复杂的地质模式,属于高维数据信息。SIMPAT 通过计算数据事件与数据样式的相似度,进而预测井间的储层地质体。如何从丰富且复杂的样式数据库中快速地检索与数据事件最相似的一个或多个数据样式成为研究的难点。为了解决该问题,须引进一些类似索引的技术来加快查找过程,这类技术包括最邻近查找(NN),例如 K-d tree[26],或相似最邻近查找(ANN),例如 K-d tree with BBF。局部敏感哈希[27](LSH)属于相似最邻近查找的一类方法。

p-stable LSH 方法是基于 p-stable 思想的局部敏感哈希方法。p-stable LSH 计算每个数据样式的特征向量 $\boldsymbol{v}$[式(5)]的哈希值 h,由于该哈希函数具有局部敏感性,若两个特征向量 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 较近,那么其哈希值 h_1 和 h_2 映射到相同桶中的概率会较大,反之则较小。p-stable LSH 的哈希函数为:

$$h_{a,b}(\boldsymbol{v}) = \text{INT}\left(\frac{\boldsymbol{a} \cdot \boldsymbol{v} + b}{W}\right) \tag{7}$$

通过式(7)将特征向量 $\boldsymbol{v}$ 映射至一个整数,称为哈希值(或哈希桶号)。

举例说明 p-stable LSH 的基本原理如图 3 所示。2 维空间平面上有 5 个黄色点,传统查找与蓝色点距离最近的黄色点的方法是计算蓝色点与所有黄色点的距离,然后对距离排序,取距离最小的黄色点作为查询结果,计算时间复杂度 O(n = 5)。采用p-stable LSH 查询时,第 1 步为预处理,随机构建 3 条基轴 X_1、X_2 和 X_3,以桶宽将每条轴划分为若干相邻的哈希桶,采用式(2)计算全部点的哈希值,并根据对应哈希值将点投影每条轴的哈希桶;第 2 步为查找过程,首先检索与蓝色点桶号相同的黄色点,与蓝色点位于相同哈希桶的黄色点包含有基轴 X_2 里的点 1、基轴 X_3 中的点 2,再从点 1、点 2 中取距离最小的点作为查询结果。对比查询时间,传统查询时间复杂度为 O(5),基于 p-stable LSH 的查询时间复杂度 O(2)。在海量高维数据检索中,p-stable LSH 的查询效率远远高于传统方法。

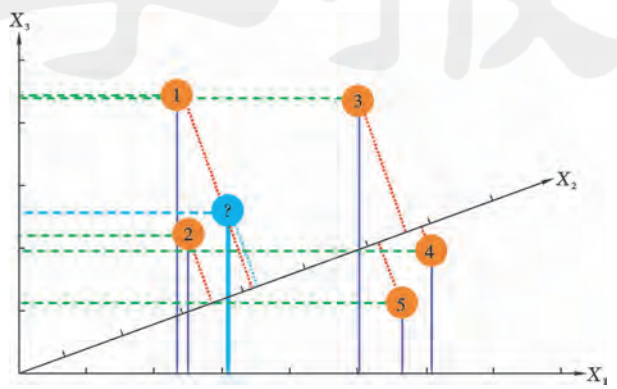


图 3 基于 p-stable LSH 算法检索对象的原理

Fig. 3 Principle of searching object based on p-stable LSH algorithm

4.2 实现步骤

LSHSIM 利用 p-stable LSH 将数据样式(事件)的特征向量散列到若干哈希表。根据 p-stable LSH 原理,哈希表中相似的数据样式(事件)位于相同哈希桶的概率远大于不相似的数据样式(事件)。

建模前,首先提取样式数据库 P_{DB}(pattern database)里所有数据样式的特征向量,然后用 p-stable LSH 计算特征向量的哈希值,建立数据样式哈希库 P_{LSHLIB}。

建模时,首先根据数据事件 dev(data event)的特征向量得到相应哈希值 h_{dev},然后从数据样式哈希库里查找哈希值等于 h_{dev} 的所有数据样式,得到目标样式数据库 T_{PatDB}(target pattern database);最后从目标样式数据库里找到最相似数据样式进而实现单步模拟。

基于 p-stable 局部敏感哈希检索数据样式的多点地质统计建模方法的具体步骤为:

(1) 输入训练图像 TI,定义模拟实现 R 和数据样板 T 的尺寸;

(2) 设置分块网格 B_{Grid} 的尺寸,输入 p-stable LSH 的参数,包括哈希桶宽 W 和哈希表数量 N;

(3) 以数据样板 T 扫描训练图像 TI,建立样式数据库 P_{DB};

(4) 基于分块网格计算数据样式 pat 的特征向量 $\boldsymbol{v}_{\text{pat}}$,进一步使用 p-stable LSH 计算特征向量 $\boldsymbol{v}_{\text{pat}}$ 的哈

希值 h_{pat},建立数据样式的哈希库 P_{LSHLIB};

(5) 根据模拟实现 R 创建随机路径 P_{Random};

(6) 如果随机路径 P_{Random} 里有未模拟节点 u,进入步骤(7)的路径;否则进入步骤(11)的路径;

(7) 提取节点 u 处的数据事件 dev,统计数据事件的分块网格块内变量之和,得到数据事件的特征向量 $\boldsymbol{v}_{dev}$,进行 p-stable LSH 计算,得到数据事件的哈希值 h_{dev};

(8) 从数据样式哈希库 P_{LSHLIB} 中查询所有哈希值等于 h_{dev} 的数据样式,构成目标数据样式库 T_{PatDB};

(9) 从目标数据样式库 T_{PatDB} 查找与数据事件 dev 最相似的数据样式 pat;

(10) 用数据样式 pat 整体覆盖并冻结模拟实现 R 的节点 u 区域;返回步骤(6);

(11) 模拟结束,输出模拟实现 R。

5　实例分析

5.1　非条件模拟测试

应用 LSHSIM 进行沉积相的非条件模拟。如图 4 所示,图 4(a)是相模型的训练图像[14],其网格维度为 250×250,网格单元大小为 10 m×10 m。其中网格重数为 3,数据样板的网格维度为 15×15。图 4(b)是相模型的一个随机模拟实现,模拟实现的河道分布、形态、连通性与图 4(a)的训练图像保持较高一致性。

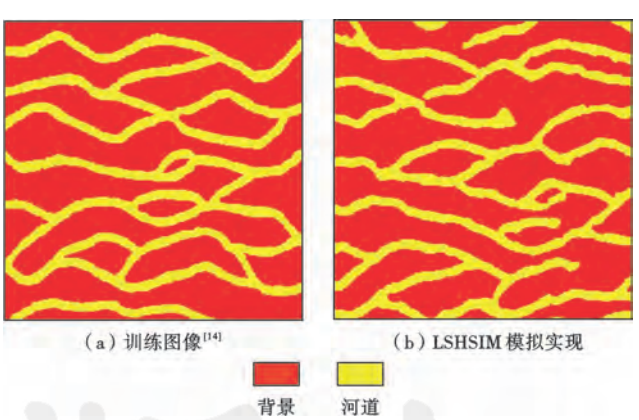


图 4　LSHSIM 的沉积相建模

Fig. 4　Facies modeling of LSHSIM

数据事件与先验地质模式的相似度计算是基于样式 MPS 建模算法的核心思想。因此,分析模拟实现的数据样式对应训练图像中最相似数据样式的相似度统计规律,能够比较 LSHSIM 和 Filtersim 及 DisPat 对训练图像先验地质模式的再现能力。图 5 为再现最相似数据样式的相似度累积频率曲线,LSHSIM 再现最相似数据样式的相似度均值为 0.963 1、标准差为 0.018 1、中位数是 0.964 4;Filtersim 的均值为 0.951 2、标准差为 0.021 7、中位数为 0.953 3;DisPat 的均值是 0.958 0、标准差为 0.018 3、中位数是 0.096 22。LSHSIM 的均值、标准差和中位数均大于其他两种算法。从图 5 可知,LSHSIM 再现最相似数据样式的相似度分布集中于高值区,与 Filtersim 和 DisPat 同样能较好地再现训练图像先验地质模式。

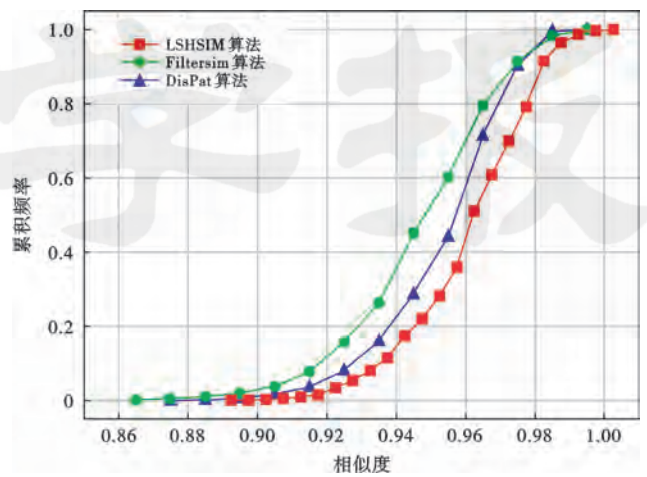


图 5　LSHSIM 与 Filtersim、DisPat 再现训练图像最相似地质模式的累积频率曲线对比

Fig. 5　Comparison of cumulative frequency curves of the most similar geological patterns reappeared in training image by using LSHSIM and Filtersim, DisPat

5.2　条件模拟测试

以井数据为条件数据,用加权的相似度计算方法[20]进行 LSHSIM 的条件模拟:

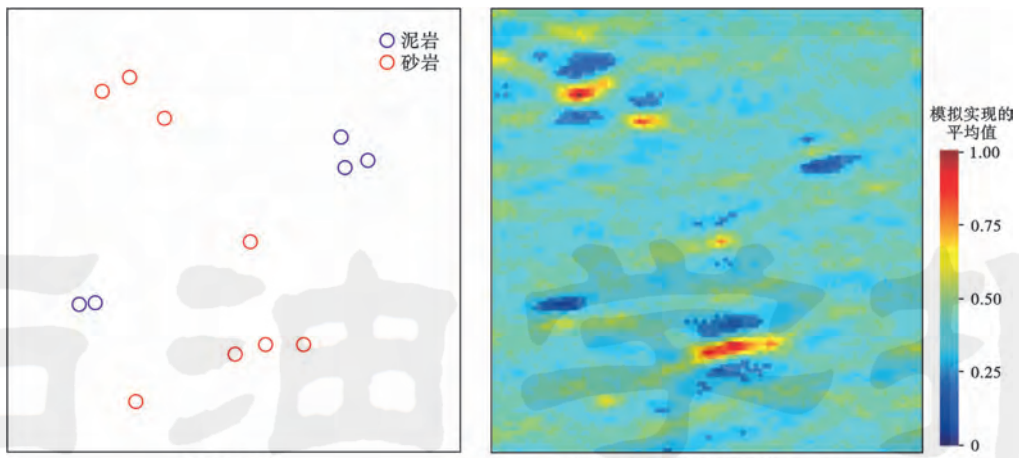
$$D(\mathrm{dev},\mathrm{pat}) = \sum_{i=1}^{d} w_i \mid \mathrm{dev}(i) - \mathrm{pat}(i) \mid \qquad (8)$$

其中,i 是数据样板里节点的索引序号。

以图 4(a)作为训练图像进行 500 次条件模拟,其中图 6(a)是取心收集的岩相数据,条件数据的权重为 0.8,模拟数据的权重为 0.2。计算全部随机模型的 E-Type模型,即点对点平均值。如图 6(b)所示,模拟实现里越靠近硬数据的网格节点,其预测结果的不确定性越小,反之越大。随着开发深入,油田井位密度增加,不断丰富的先验地质信息对勘探剩余油十分有利。在密集井网的建模应用中,提出方法能较好符合井位数据并表现出随机性[14]。

5.3　效率与内存占用比较

通过实例对比 LSHSIM 与其他 MPS 方法 SIMPAT、DisPat、Filtersim 和 SNESIM 算法的效率和内存开销两方面。如图 7 所示,以三维地质模型作为训练图像[28],训练图像的网格维度为 70×70×40,数据样板的网格维度为 11×11×7,模拟实现的网格维度为 70×70×40,网格单元尺寸为 20 m×20 m×0.5 m。如图8所示,LSHSIM的桶宽等于0.01时,LSHSIM



(a) 硬数据 (b) LSHSIM 条件模拟的 E-Type

图 6 LSHSIM 算法基于“硬数据”的条件模拟

Fig. 6 Condition simulation base on “hard data”of LSHSIM algorithm

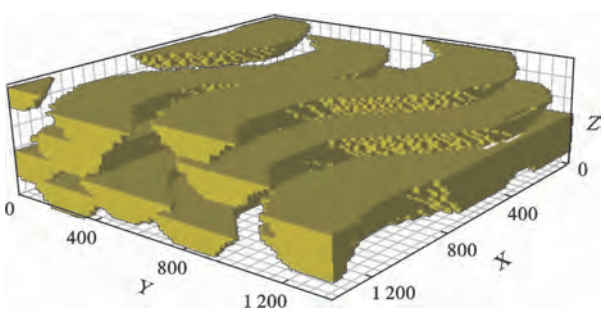


图 7 三维训练图像

Fig. 7 3D training image

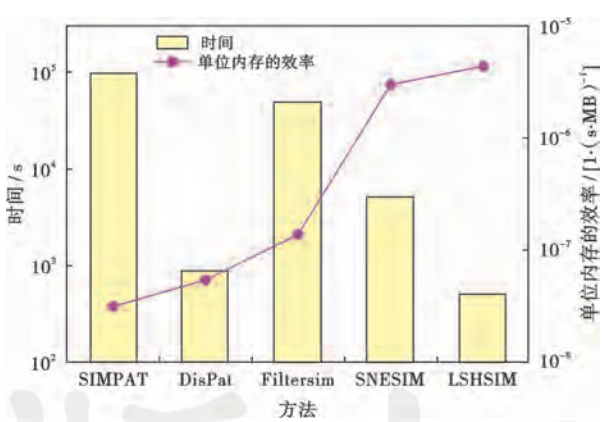


图 8 LSHSIM 与其他 MPS 方法建立 100 个模拟实现的时间、内存综合对比

Fig. 8 Comparison of consume time and RAM between LSHSIM and some other MPS

模拟 100 个随机实现的时间为 512 s，在所有参与比较的 MPS 中效率最高，DisPat、SNESIM、Filtersim 和 SIMPAT 的计算时间分别为 900 s、5 170 s、50 938 s 和 96 588 s；综合考虑效率、内存两方面，LSHSIM 的单位内存效率相对其他 MPS 方法优势明显。

p-stable LSH 的检索效率与桶宽大小相关。如图 9 所示，当桶宽较小，单个哈希桶内的数据样式数量较少，查询速度较快，查询准确率相对较低；当桶宽较大，单个哈希桶的数量较多，查询速度相对较慢，查询准确性较高。哈希桶宽的选取通常依据数据样式的数量和分块网格的维度进行选取。

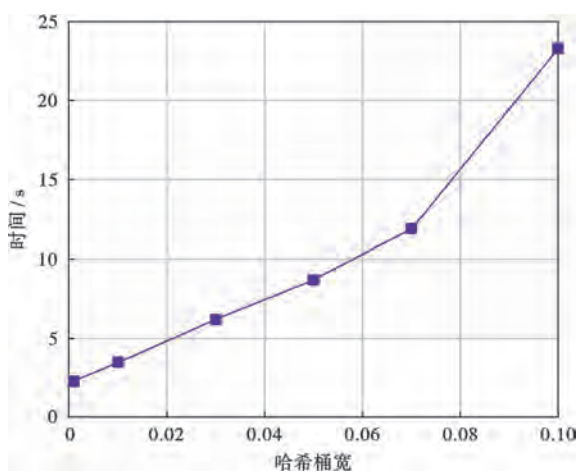


图 9 计算效率与哈希桶宽的关系

Fig. 9 Relationship between calculation efficiency and hash bucket wide

5.4 参数敏感性分析

与经典地质统计建模方法类似，LSHSIM 的建模效果受多种因素影响。LSHSIM 的常用参数包括数据样板的尺寸、多重网格的数量、分块网格的尺寸、哈希表的数量和哈希桶宽。

数据样板的尺寸大小直接影响空间多点相关性的范围。如图 10 所示，以图 4(a)为训练图像，网格重数等于 3，随着数据样式的尺寸增加，模拟实现的河道的连续性和结构形态逐渐变好；另外，通过分块网格的尺寸敏感性分析可以看出，分块网格的尺寸对建模的影响较小。

局部敏感哈希的哈希表数量与哈希桶宽是影响检索高维数据的查全率和查准率的两个关键因素。如图 11(a)、图11(d)、图11(g)所示，通过哈希计算，

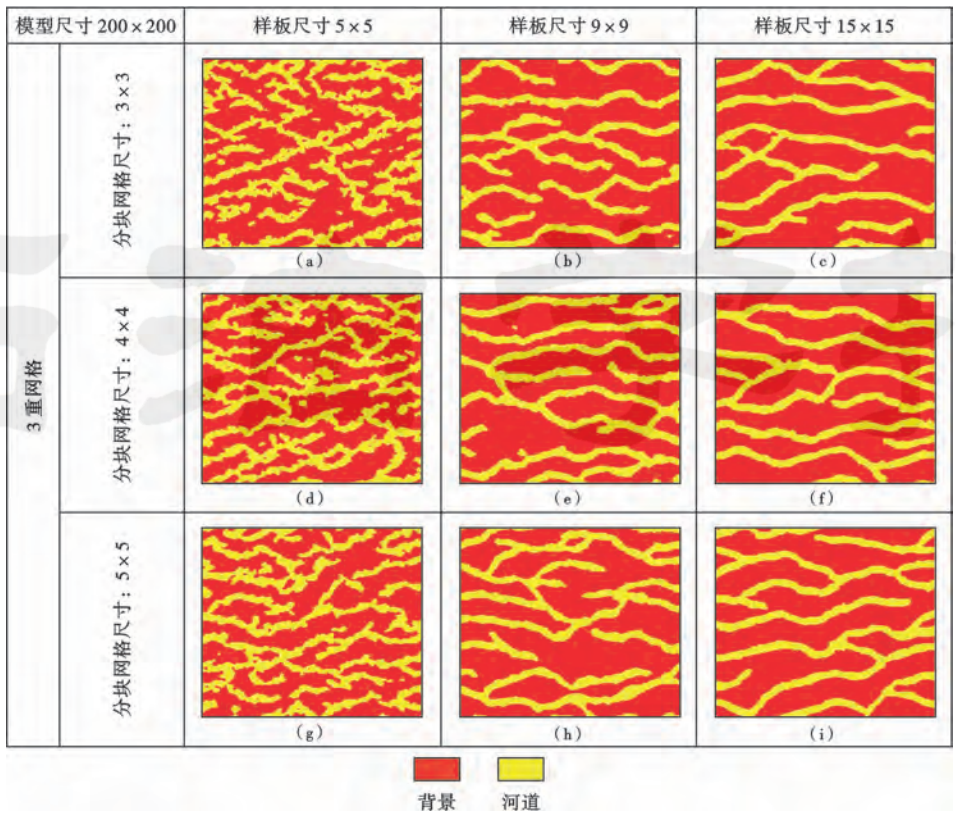


图 10　数据样板及分块网格的尺寸敏感性测试

Fig. 10　Sensitivity test of template size and block grid size

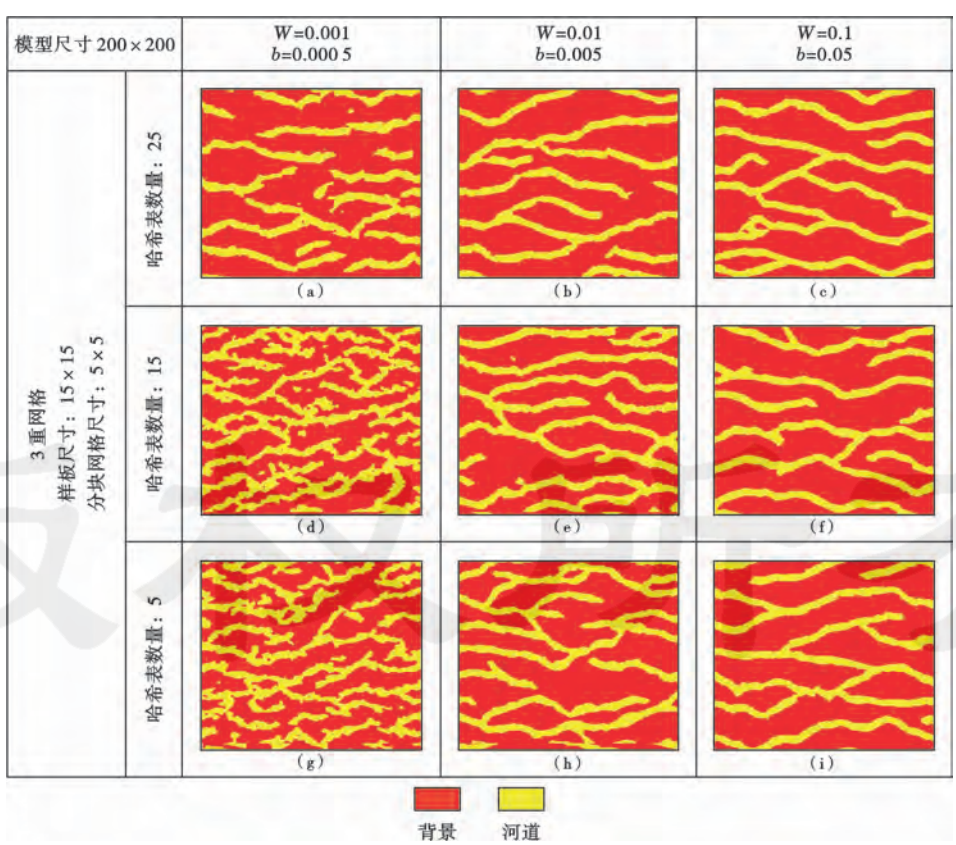


图 11　桶宽和哈希表数量敏感性测试

Fig. 11　Sensitivity test of hash bucket width and number of hash table

哈希表数量越多，数据样式的特征向量在哈希表中的分布越稳定，建模时检索的数据样式准确率越高，模拟实现对训练图像的再现性越好。哈希桶的宽度同时影响计算效率和建模质量，从图 11(a)—图 11(c)可以看出，随着哈希桶宽的增加，模型的河道形态连贯性越来越好。

6 结 论

(1) 剖析传统的基于样式 MPS 算法 SIMPAT 的原理，分析 SIMPAT 计算效率的瓶颈——数据事件与训练图像中所有数据样式进行相似度计算。在 SIMPAT 算法基础上，基于局部敏感哈希检索技术提出一种新的多点地质统计建模算法 LSHSIM。

(2) 通过比较 LSHSIM 与 SIMPAT、Filtersim、SNESIM 和 DisPat 等当前主流 MPS 算法的计算效率和内存占用情况，结果表明 LSHSIM 算法不仅大幅度提高建模效率，并且有效地控制了内存开销。

(3) 以实例对比 Filtersim 和 DisPat，LSHSIM 可以很好地再现训练图像先验地质模式；在条件数据约束下，LSHSIM 能够建立更精确的储层地质模型；通过参数敏感性分析，为选择合适参数提供了参考依据。

符号注释： a—服从 p-stable 分布的独立随机向量；b—[0，W]范围内的随机数；W—哈希桶宽；R—模拟实现；N—哈希表数量；T—数据样板；P_{DB}—样式数据库；P_{LSHLIB}—数据样式的哈希库；P_{Random}—随机路径；T_{PatDB}—目标数据样式库；h_{pat}—数据样式的哈希值；h_{dev}—数据事件的哈希值；$\boldsymbol{v}_{dev}$—数据事件的特征向量；m、n—分块网格的二维索引；$h_{a,b}(\boldsymbol{v})$—p-stable LSH 的哈希函数；B_{Grid}—分块网格；$B_{Grid}(m,n)$—索引[m，n]的分块内所有节点值之和；d—x 和 y 的维度；dev—数据事件；D—数据事件与数据样式的距离；G—网格；$G_{i,j}$—数据样式(事件)网格里索引等于[i，j]的节点值；h—特征向量的哈希值；I_{Count}—数据样板在水平方向上的网格数量；J_{Count}—数据样板在垂直方向上的网格数量；L_1—曼哈顿距离；M_{Count}—分块网格在水平方向上的数量；Min—数据集合的最小值；M_{Size}—分块网格在水平方向上的网格数量；N_{Count}—分块网格在垂直方向上的数量；N_{Size}—分块网格在垂直方向上的网格数量；NULL—空值；pat—数据样式；s—相似度；Sum—求和函数；TI—训练图像；$\boldsymbol{v}$—分块网格的特征向量；c—数据事件的节点值；w—权重；x、y—数据样式(事件)；x_i、y_i—数据样式事件的第 i 个节点；INT—取整函数；O—时间复杂度；u—节点。

参 考 文 献

[1] DEUTSCH C V, JOURNEL A G. GSLIB: geostatistical software library and user's guide[M]. 2nd ed. New York: Oxford University Press, 1998.

[2] DEUTSCH C V, TRAN T T. FLUVSIM: a program for object-based stochastic modeling of fluvial depositional systems[J]. Computers & Geosciences, 2002, 28(4): 525-535.

[3] PYRCZ M J, BOISVERT J B, DEUTSCH C V. ALLUVSIM: a program for event-based stochastic modeling of fluvial depositional systems[J]. Computers & Geosciences, 2009, 35(8): 1671-1685.

[4] PYRCZ M J, DEUTSCH C V. Geostatistical reservoir modeling[M]. 2nd ed. New York: Oxford University Press, 2014.

[5] 黄小娟，李治平，周光亮，等. 裂缝性致密砂岩储层裂缝孔隙度建模——以四川盆地平落坝构造须家河组二段储层为例[J]. 石油学报，2017，38(5)：570-577.
HUANG Xiaojuan, LI Zhiping, ZHOU Guangliang, et al. Fracture porosity modeling of fractured tight sandstone reservoir: a case study of the reservoir in Member 2 of Xujiahe Formation, Pingluoba structure, Sichuan Basin[J]. Acta Petrolei Sinica, 2017, 38(5): 570-577.

[6] 牛博，高兴军，赵应成，等. 古辫状河心滩坝内部构型表征与建模——以大庆油田萨中密井网区为例[J]. 石油学报，2015，36(1)：89-100.
NIU Bo, GAO Xingjun, ZHAO Yingcheng, et al. Architecture characterization and modeling of channel bar in paleo-braided river: a case study of dense well pattern area of Sazhong in Daqing oilfield[J]. Acta Petrolei Sinica, 2015, 36(1): 89-100.

[7] 谭学群，廉培庆，张俊法，等. 基于"二次震控"的三维油藏建模新方法[J]. 石油学报，2016，37(12)：1518-1527.
TAN Xuequn, LIAN Peiqing, ZHANG Junfa, et al. A new method for 3D oil reservoir modeling based on "double seismic constraint"[J]. Acta Petrolei Sinica, 2016, 37(12): 1518-1527.

[8] 林承焰，陈仕臻，张宪国，等. 多趋势融合的概率体约束方法及其在储层建模中的应用[J]. 石油学报，2015，36(6)：730-739.
LIN Chengyan, CHEN Shizhen, ZHANG Xianguo, et al. Probability constraint method based on multiple trend integration and its application in reservoir modeling[J]. Acta Petrolei Sinica, 2015, 36(6): 730-739.

[9] 唐俊华，阎保平. 基于 LSH 索引的快速图像检索[J]. 计算机工程与应用，2002，38(24)：20-21.
TANG Junhua, YAN Baoping. Fast image retrieval based on LSH indexing[J]. Computer Engineering and Applications, 2002, 38(24): 20-21.

[10] GUARDIANO F B, SRIVASTAVA R M. Multivariate geostatistics: beyond bivariate moments[M]// SOARES A. Geostatistics Tróia'92. Dordrecht: Springer, 1993: 133-144.

[11] STREBELLE S B, JOURNEL A G. Reservoir modeling using multiple-point statistics[R]. SPE 71324, 2001.

[12] STREBELLE S. Conditional simulation of complex geological structures using multiple-point statistics[J]. Mathematical Geology, 2002, 34(1): 1-21.

[13] STREBELLE S, CAVELIUS C. Solving speed and memory issues in multiple-point statistics simulation program SNESIM [J]. Mathematical Geosciences, 2014, 46(2): 171-186.

[14] ARPAT G B. Sequential simulation with patterns[D]. Stanford, CA, USA: Stanford University, 2005.

[15] ZHANG Tuanfeng, SWITZER P, JOURNEL A. Filter-based classification of training image patterns for spatial simulation [J]. Mathematical Geology, 2006, 38(1): 63-80.

[16] MARIETHOZ G, RENARD P. Reconstruction of incomplete data sets or images using direct sampling[J]. Mathematical Geosciences, 2010, 42(3): 245-268.

[17] HONARKHAH M, CAERS J. Stochastic simulation of patterns using distance-based pattern modeling[J]. Mathematical Geosciences, 2010, 42(5): 487-517.

[18] MARIETHOZ G, RENARD P, STRAUBHAAR J. The direct sampling method to perform multiple-point geostatistical simulations[J]. Water Resources Research, 2010, 46(11): 1-14.

[19] ARPAT G B, CAERS J. Conditional simulation with patterns[J]. Mathematical Geology, 2007, 39(2): 177-203.

[20] STRAUBHAAR J, RENARD P, MARIETHOZ G, et al. An improved parallel multiple-point algorithm using a list approach [J]. Mathematical Geosciences, 2011, 43(3): 305-328.

[21] MAHMUD K, MARIETHOZ G, CAERS J, et al. Simulation of earth textures by conditional image quilting[J]. Water Resources Research, 2014, 50(4): 3088-3107.

[22] MARIETHOZ G, LEFEBVRE S. Bridges between multiple-point geostatistics and texture synthesis: review and guidelines for future research[J]. Computers & Geosciences, 2014, 66: 66-80.

[23] 喻思羽,李少华,何幼斌,等. 基于样式降维聚类的多点地质统计建模算法[J]. 石油学报, 2016, 37(11): 1403-1409.
YU Siyu, LI Shaohua, HE Youbin, et al. Multiple-point geostatistics algorithm based on pattern scale-down cluster[J]. Acta Petrolei Sinica, 2016, 37(11): 1403-1409.

[24] 黄涛. 基于 GPU 的多点地质统计逐点模拟并行算法的研究[D]. 合肥:中国科学技术大学, 2013.
HUANG Tao. Research on the GPU-based parallel algorithms for point-based multiple-point geostatistical simulation [D]. Hefei: University of Science and Technology of China, 2013.

[25] 李少华,喻思羽,王军,等. 并行布尔模拟算法[J]. 物探与化探, 2013, 37(4): 723-725.
LI Shaohua, YU Siyu, WANG Jun, et al. The parallel boolean simulation method[J]. Geophysical and Geochemical Exploration, 2013, 37(4): 723-725.

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

[27] INDYK P, MOTWANI R. Approximate nearest neighbors: towards removing the curse of dimensionality: proceedings of the 13th annual ACM symposium on theory of computing[C]. Dallas, Texas, USA: ACM, 1998, 604-613.

[28] LIU Yuhong. Using the Snesim program for multiple-point statistical simulation[J]. Computers & Geosciences, 2006, 32(10): 1544-1563.

(收稿日期 2017-03-24　改回日期 2017-10-09　编辑　王培玺)