



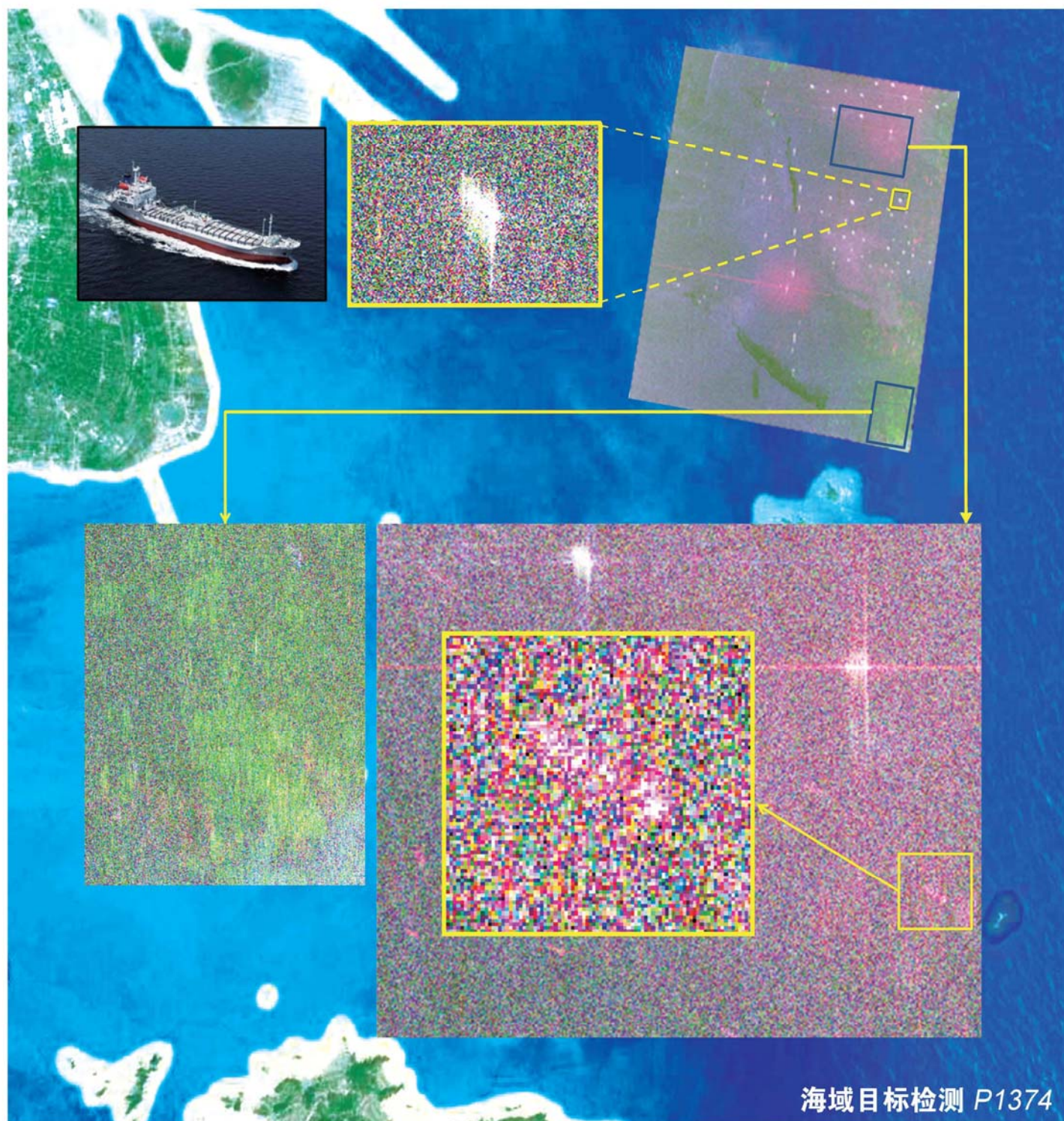
JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院遥感与数字地球研究所  
中国图象图形学学会  
北京应用物理与计算数学研究所

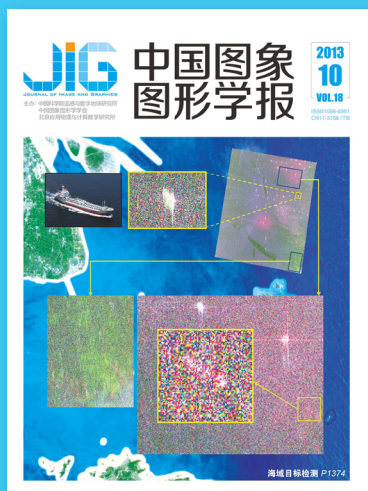
# 中国图象图形学报

2013  
10  
VOL.18

ISSN1006-8961  
CN11-3758/TB



海域目标检测 P1374



第18卷第10期（总第210期）

2013年10月16日

中国精品科技期刊  
中国国际影响力优秀学术期刊  
中国科技核心期刊

## 版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

## Copyright

2013 all rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院

主办单位 中国科学院遥感与数字地球研究所  
中国图象图形学学会  
北京应用物理与计算数学研究所

主 编 李小文

编辑出版《中国图象图形学报》编辑出版委员会

邮政信箱 北京9718信箱

邮 编 100101

电子信箱 jig@irsa.ac.cn

电 话 010-64807995 010-82614429

网 址 www.cjig.cn

广告经营许可证 京朝工商广字第0361号

总 发 行 北京报刊发行局

订 购 全国各地邮局

海外发行 中国国际图书贸易集团有限公司  
(邮政信箱: 北京399信箱 邮编: 100048)

印刷装订 北京科信印刷有限公司

**Journal of Image and Graphics**

Title inscription: Song Jian

Monthly, Started in 1996

Superintended by Chinese Academy of Sciences

Sponsored by Institute of Remote Sensing and Digital Earth, CAS  
China Society of Image and Graphics

Institute of Applied Physics and Computational Mathematics

Editor-in-Chief LI Xiaowen

Editor, Publisher Editorial and Publishing Board of Journal of  
Image and Graphics

P.O.Box 9718, Beijing, P.R.China

Zip code 100101

E-mail jig@irsa.ac.cn

Telephone 010-64807995 010-82614429

Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals

Domestic All Local Post Offices in China

Overseas China International Book Trading Corporation  
(P.O.Box 399, Beijing 100048, P.R.China))

Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB

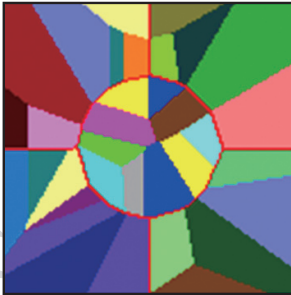
ISSN 1006-8961

CODEN ZTTXFZ

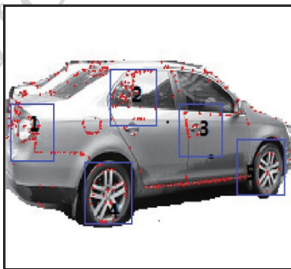
国外发行代号 M1406

国内邮发代号 82-831

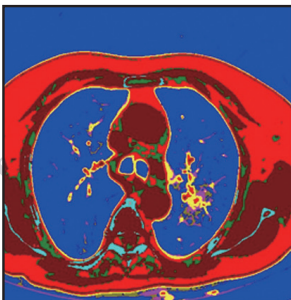
国内定价 50.00元



结合几何划分技术和最大期望值/最大边缘概率算法的彩色图像分割(第1270页)



联合生成与判别模型的目标检测与跟踪(第1293页)



聚类核值相似区特征点的医学影像分类(第1322页)

## 综述

### 手掌静脉图像识别技术综述

吴微, 苑玮琦 ..... 1215

### 视频检测烟雾的研究现状

罗胜, Jiang Yuzheng ..... 1225

## 图像处理和编码

### 自适应分数阶微分的复合双边滤波算法

胡伏原, 姒绍辉, 张艳宁, 孙瑾秋 ..... 1237

### 曲波域高斯混合尺度模型的图像压缩重构

纪建, 李晓, 许双星 ..... 1247

### 结合图像信号显著性的自适应分块压缩采样

王瑞, 余宗鑫, 杜林峰, 万旺根 ..... 1255

### 四阶各向异性扩散方程在图像放大中的应用

李鹏, 邹杨, 姚正安 ..... 1261

## 图像分析和识别

### 结合几何划分技术和最大期望值/最大边缘概率算法的彩色图像分割

赵泉华, 李玉, 何晓军 ..... 1270

### 自适应加权完全局部二值模式的表情识别

胡敏, 许艳侠, 王晓华, 黄忠, 朱弘 ..... 1279

### 面向局部光照突变的时间和空间中心对称局部二值模式算子

卞紫涵, 房胜, 徐田帅 ..... 1285

### 联合生成与判别模型的目标检测与跟踪

刘倩, 侯建华, 牟海军, 赵巍, 笪邦友 ..... 1293

### 目标不同视角下观察信息的迁移和再利用

张索非, 吴海洋, 吴镇扬 ..... 1302

## 图像理解和计算机视觉

### 张量分析在视差图生成中的应用

赵戈, 蔺兰, 唐延东, 王耀南 ..... 1307

### 采用旋转匹配的二进制局部描述子

卢鸿波, 孙愿, 张志敏 ..... 1315

## 医学图像处理

### 聚类核值相似区特征点的医学影像分类

李博, 曹鹏, 栗伟, 赵大哲 ..... 1322

### 异常宫颈细胞核的自适应局部分割

张灵, 李静立, 陈思平, 汪天富, 江少锋, 刘少雄 ..... 1329

## 地理信息技术

### Hilbert填充曲线与空间分布模式探测的点数据集空间划分方法

吴明光 ..... 1336

### 采用案例归纳推理进行道路网智能选取

郭敏, 钱海忠, 黄智深, 刘海龙, 王骁 ..... 1343

## 遥感图像处理

### 红外目标抗干扰跟踪算法

褚桐, 马惠敏 ..... 1354

### MSTAR图像的矩特征分析与多阈值分割

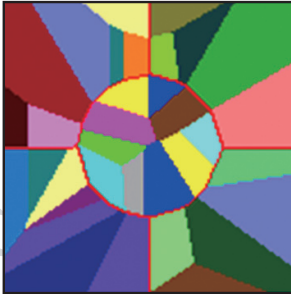
倪维平, 严卫东, 吴俊政, 郑刚, 芦颖 ..... 1364

### 改进Notch滤波的全极化SAR数据船舶检测方法

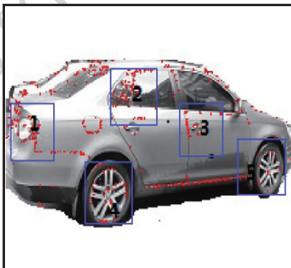
孙渊, 王超, 张红, 张波, 吴樊 ..... 1374

# CONTENTS

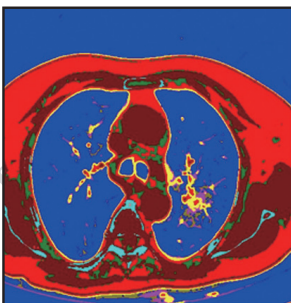
## JOURNAL OF IMAGE AND GRAPHICS



Color image segmentation using geometry tessellation technique and EM/MPM algorithm(P1270)



Object detection and tracking combining generative and discriminative model(P1293)



Medical image classification based on cluster univalue segment assimilating nucleus feature point(P1322)

### Review

- A survey of palm-vein image recognition  
Wu Wei, Yuan Weiqi ..... 1215
- State-of-art of video based smoke detection algorithms  
Luo Sheng, Jiang Yuzheng ..... 1225

### Image Processing and Coding

- Multi-bilateral filtering algorithm based on adaptive fractional order differentiation  
Hu Fuyuan, Si Shaohui, Zhang Yanning, Sun Jinqiu ..... 1237
- Image reconstruction algorithm based on the curvelet gaussian scale mixture model  
Ji Jian, Li Xiao, Xu Shuangxing ..... 1247
- Saliency-based adaptive block compressive sampling for image signals  
Wang Rui, Yu Zongxin, Du Linfeng, Wan Wanggen ..... 1255
- Fourth-order anisotropic diffusion equations for image zooming  
Li Peng, Zou Yang, Yao Zheng'an ..... 1261

### Image Analysis and Recognition

- Color image segmentation using geometry tessellation technique and EM/MPM algorithm  
Zhao Quanhua, Li Yu, He Xiaojun ..... 1270
- Facial expression recognition based on AWCLBP  
Hu Min, Xu Yanxia, Wang Xiaohua, Huang Zhong, Zhu Hong ..... 1279
- TSCS-LBP operator oriented to the local illumination mutation  
Bian Zihan, Fang Sheng, Xu Tianshuai ..... 1285
- Object detection and tracking combining generative and discriminative model  
Liu Qian, Hou Jianhua, Mou Haijun, Zhao Wei, Da Bangyou ..... 1293
- Transfer and reusing of object view information  
Zhang Suofei, Wu Haiyang, Wu Zhenyang ..... 1302

### Image Understanding and Computer Vision

- Disparity map generation with tensor analysis  
Zhao Ge, Lin Lan, Tang Yandong, Wang Yaonan ..... 1307
- Binary local descriptor based on rotative matching  
Lu Hongbo, Sun Yuan, Zhang Zhimin ..... 1315

### Medical Image Processing

- Medical image classification based on cluster univalue segment assimilating nucleus feature points  
Li Bo, Cao Peng, Li Wei, Zhao Dazhe ..... 1322
- Segmentation of abnormal cervical nuclei using an adaptive and local approach  
Zhang Ling, Li Jingli, Chen Siping, Wang Tianfu, Jiang Shaofeng, Liu Shaoxiang ..... 1329

### Geoinformatics

- Hilbert space-filling curve and spatial pattern detection-based spatial partitioning approach to point geospatial data  
Wu Mingguang ..... 1336
- Intelligent road network selection method based on cases inductive reasoning  
Guo Min, Qian Haizhong, Huang Zhishen, Liu Hailong, Wang Xiao ..... 1343

### Remote Sensing Image Processing

- Study of infrared anti-jamming tracking algorithm  
Chu Tong, Ma Huimin ..... 1354
- Moment feature analysis of MSTAR image and multiple thresholds based segmentation  
Ni Weiping, Yan Weidong, Wu Junzheng, Zheng Gang, Lu Ying ..... 1364
- Polarimetric SAR ship detection using improved notch filter  
Sun Yuan, Wang Chao, Zhang hong, Zhang bo, Wu Fan ..... 1374

中图法分类号: P208 文献标识码: A 文章编号: 1006-8961(2013)10-1336-07

论文引用格式: 吴明光. Hilbert 填充曲线与空间分布模式探测的点数据集空间划分方法[J]. 中国图象图形学报, 2013, 18(10): 1336-1342.  
[DOI:10.11834/jig.20131016]

## Hilbert 填充曲线与空间分布模式探测的 点数据集空间划分方法

吴明光

南京师范大学虚拟地理环境教育部重点实验室, 南京 210023

**摘要:** 空间数据划分是空间索引、并行 GIS 数据分解以及分布式数据管理与调度等问题的核心环节之一。针对点数据集多目标空间划分问题, 引入 Hilbert 空间填充曲线和空间分布模式探测过程, 提出针对规则、随机和聚集分布模式的点数据集空间划分方法。实验结果表明, 该方法能够在缺少覆盖范围信息的条件下准确判定空间分布类型, 该方法能够兼顾空间聚集性、数据量均衡与空间重叠度 3 种约束条件。

**关键词:** 空间数据划分; 空间分布模式; Hilbert 空间填充曲线

## Hilbert space-filling curve and spatial pattern detection-based spatial partitioning approach to point geospatial data

Wu Mingguang

Key Laboratory of Virtual Geographical Environment, Ministry of Education, Nanjing Normal University, Nanjing 210023, China

**Abstract:** Optimal partitioning of spatial dataset is an import concern for many data management context such as spatial indexing, parallelizing GIS and distributed data access. This paper aims to discuss an adaptive spatial data partitioning method based on Hilbert space-filling curve and spatial pattern detection. A spatial data partitioning approach is proposed for spatial data partitioning of regular, random, and clustering objects, respectively. We show that the proposed method can achieve optimal partitioning point dataset without point dataset boundary information. We also show that the proposed method can adapt to available information about data distributions, data sizes and clustering.

**Key words:** data partitioning; distribution pattern; Hilbert space-filling curve

## 0 引言

数据划分是空间数据管理与调度的关键环节之一<sup>[1-2]</sup>: 空间索引中采用空间划分方法构建层次结构的  
空间搜索树, 提高空间数据检索效率; 并行 GIS 中  
将大的数据集分为若干数据组, 与并行计算任务协  
同<sup>[3-4]</sup>; 分布式空间数据调度中采用空间划分方法将

一个大的数据集分解到若干数据节点上, 保证负载  
平衡, 提高系统运行效率<sup>[5]</sup>。

空间划分的指标主要包括: 划分后数据组的数据  
量、数据组内对象的个数、数据组之间的空间重叠  
度、数据组内对象的空间聚集性。Shekhar 等人证明  
在考虑负载均衡的情况下, 空间数据划分问题是一个  
NP 完全问题<sup>[2]</sup>。若再增加组间空间重叠度最小,  
则该问题求解更为复杂。现有的数据划分方法

收稿日期: 2012-10-25; 修回日期: 2013-04-01

基金项目: 国家自然科学基金项目(40901186, 41271446)

第一作者简介: 吴明光(1979—), 男, 副教授, 2007 年解放军信息工程大学获地图制图学与地理信息工程专业博士学位, 主要研究方向  
为空间数据模型与空间信息可视化。E-mail: wmg@njnu.edu.cn

主要包括映射函数、空间聚类以及空间索引划分 3 类。映射函数方法通过建立实体编号、空间位置等信息与待划分数据集之间的函数映射关系来实现数据划分。映射函数可以是随机函数(如 Round-Robin 法<sup>[1]</sup>)、条件函数(如 Oracle Range Partitioning 法<sup>[6]</sup>)、散列函数(如 Hash Partitioning 法<sup>[7]</sup>)、空间填充曲线函数(如 Hilbert 法<sup>[8]</sup>)。由于空间分布不规则以及空间实体非结构化,无法用通用的函数表示。所以基于映射函数的方法难以同时兼顾数据量平衡与查询效率均等。空间聚类数据划分方法基于最邻近距离<sup>[9-10]</sup>、密度、空间相似性<sup>[11]</sup>等原则,以聚类为中心将空间数据划分为若干组。该方法也存在若干问题:1)均匀分布或者随机分布的地理数据本身不具备聚类的前提。2)完全以聚类结果对数据进行划分容易导致数据倾斜(data skew);空间索引划分方法基于地物形态<sup>[12]</sup>或某类空间分割原则(如 B-tree、R-Tree<sup>[13-14]</sup>等)实现空间划分,但该类方法以空间查询为出发点,难以同时顾及空间聚集与数据量均衡。

文献[2]基于 Hilbert 填充曲线提出一种数据量均匀的数据划分方法,能够有效解决数据倾斜问题。文献[5]以节点平均数据量、网格累积数据量作为定量指标,提出一种基于 Hilbert 曲线层次分解的空间数据划分方法。本文认为 Hilbert 曲线方法能够很好地保持空间邻近性。但是,以数据量均匀作为划分依据,容易导致空间聚集的实体被硬分割,从而降低分组后实体的聚集性。针对点数据空间划分时多目标难以同时兼顾的问题,结合空间填充曲线与空间分布模式分析方法,提出一种同时兼顾数据均衡与空间聚集特征等条件的数据划分方法。线、面数据集的划分方法将另文论述。

## 1 基于 Hilbert 填充曲线的空间分布模式判断模型建立

德国数学家 Hilbert 于 1891 年构造出是一条处处连续但处处不可导的数学曲线(图 1)<sup>[15]</sup>。Hilbert 填充曲线已广泛应用于空间排序,其空间划分性质可归纳为

1) Hilbert 曲线采用四分模式递归划分填充区域。

2) Hilbert 曲线采用 1 维连续的直线填充 2 维甚至高维空间。基于 Hilbert 曲线的延展方向,可以对

递归的四分区域进行唯一编码。

3) Hilbert 码相近的实体在 2 维甚至高维空间中也是相近的(充分非必要条件)。

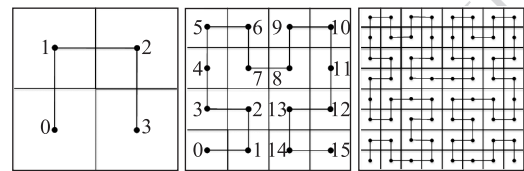


图 1 Hilbert 曲线

Fig. 1 Hilbert curves

由于 Hilbert 空间填充曲线具有类似四叉树的四分迭代空间分解的性质,使得空间划分较为均匀。与四叉树分割方法不同的是: Hilbert 曲线处处连续,在对空间进行划分的过程中能够保证空间上的连续,能够有效减少四叉树方法中连续对象被四象限硬分割的现象。与 Peano、Gray 等空间填充曲线相比, Hilbert 曲线具有更好的空间邻近性<sup>[15]</sup>,能够保证 Hilbert 编码连续的实体在空间上的最大程度的邻近。但是, Hilbert 曲线并不显式反映整个数据集的空间分布特征。若要实现多目标空间划分,还需要解决空间分布模式的识别和提取问题。

现有模式判定方法主要包括以聚集性为基础的基于密度的方法与以分散性为基础的基于距离的方法。上述两种方法均需要考虑实体集合的地理范围。其问题包括:1)一般情况下地理范围不容易直接获取,覆盖面积计算具有较大的模糊性;2)非规则的复杂边界(带岛、带洞)与实体之间空间关系计算量大。面向空间数据划分的分布模式判定本身并非为发现地理意义上的分布规律,本身可不考虑地理意义上的边界,只需判断覆盖范围内实体的聚集/离散程度。因此提出一种无需边界信息的分布类型判定方法,其基本思路是:采用 Hilbert 曲线填充实体集合的空间范围,对实体进行 Hilbert 编码,给出随机分布假设,然后通过离散程度及其显著性水平计算来否定或者肯定随机分布假设,得出具体的分布类型。

采用文献[16]中针对样方分析(quadrant analysis)方法的最优样方尺寸计算公式为  $Q = 2A/n$ , 式中  $Q$  为样方的尺寸,  $A$  为研究区域的面积,  $n$  为研究区域中点的数量。该公式表明每一个样方区间内的经验值为 2。假设该数据集包含  $n$  个呈随机分布的点实体。图 2 所示为用  $k$  阶( $2^k \approx n/2$ )。Hilbert 曲

线对实体集合最小外接矩形进行填充,得到等大的  $2^{2k}$  个规则正方形。按照 Hilbert 曲线填充顺序对每个数据块进行编号。每个实体均得到一个 Hilbert 编码。下文以点实体的 Hilbert 编码来论述空间分布模式判定过程。

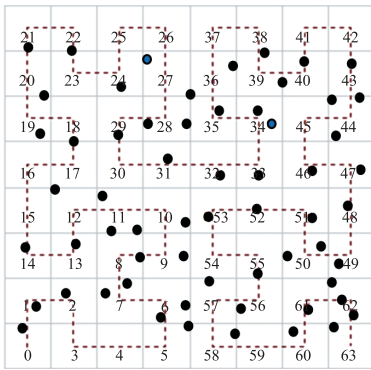


图2 Hilbert 曲线的层次分解与实体编码

Fig. 2 Hierarchical decomposition of Hilbert curve and entity coding

点集中任意一点  $P_i$  位于各子区域的可能性相等,则点编码为  $h_i$  的概率为  $1/2^{2k}$ 。因此,在  $n$  个点的情况下,对任意 Hilbert 区块内具有  $m$  个点的概率为

$$\frac{\binom{n}{m} \left(\frac{1}{2^{2k}}\right)^m \left(\frac{2^{2k}-1}{2^{2k}}\right)^{n-m}}{m! (n-m)!} \times \frac{(2^{2k}-1)^{n-m}}{(2^{2k})^n} \quad (1)$$

这是二项式分布的模型。 $m$  的方差统计值为

$$\frac{1}{2^{2k}} \sum_{i=1}^{2^{2k}} f_i (m_i - \bar{m})^2 \quad (2)$$

式中  $f_i$  表示  $m_i$  的频数,对于落在数据集分布范围之外的 Hilbert 区块不做统计。为避免空间范围计算,提出一种改进方法:对所有 Hilbert 区块内的实体个数进行统计。若处处相等则为规则分布。若不等则将个数从小到大排序,剔除统计个数为 0 的 Hilbert 区块(表示该 Hilbert 区块在空间分布范围之外,或者为空间分布范围内的空洞),形成  $S_{\text{sort}}$ ,  $S_{\text{sort}} = \{n_1, n_2, n_3, \dots, n_p\}$ ,  $n_i \leq n_{i+1}$ ,  $n_i > 0$ ;  $p$  为 Hilbert 区块内的实体个数不为零的区块数。对  $S_{\text{sort}}$  进行方差统计可以反映集合的离散程度,也避免了空间范围因素影响。要准确判断其空间分布模式还需计算其显著性水平。泊松定理表明,当  $1/2^{2k}$  较小时,二项式分布近似为泊松分布

$$\frac{\binom{n}{m} \left(\frac{1}{2^{2k}}\right)^m \left(\frac{2^{2k}-1}{2^{2k}}\right)^{n-m}}{m! (n-m)!} \approx \frac{\lambda^m e^{-\lambda}}{m!} \quad (3)$$

$$\lambda = \frac{n}{2^{2k}} \quad (4)$$

泊松分布中均值、方差相等且均为  $\lambda$ ,因此可以通过考察均值和方差的比值来判定点集是否为随机分布并进行显著检验。采用文献[16]中方差均值比之  $\chi^2$  检验法来计算显著性水平。定义分散性指数为

$$I = \frac{(m-1)s^2}{\bar{x}} = \frac{\sum_{i=1}^m (x_i - \bar{x})^2}{\bar{x}} \quad (5)$$

如果空间模式接近于泊松分布,则  $I \rightarrow 1$ 。 $I$  服从  $\chi^2$  分布,即  $I \sim \chi^2(m-1)$ 。 $I$  可以通过统计计算求取,通过与  $\chi^2(m-1)$  分位点的比较来计算显著水平。费希尔(Fisher)证明<sup>[16]</sup>,当  $m$  充分大时

$$\chi_a^2(m-1) \approx \frac{1}{2} (z_a + \sqrt{2(m-1)-1})^2 \quad (6)$$

式中,  $z_a$  是标准正态分布的上  $a$  分位点。利用式(6)可以求得  $I$  分布的上  $a$  分位点的近似值。当  $I$  的统计值大于显著性水平为  $a$  的  $\chi^2(m-1)$  时,数据集呈聚集分布,否则为随机分布。

## 2 基于分布模式探测的空间划分方法

每一个实体均包含一个 Hilbert 码,数据集中所有实体的 Hilbert 码形成一个 Hilbert 码集合:  $S_H = \{h_1, h_2, \dots, h_n\}$ , 式中  $n$  为数据集中实体个数。若实体呈规则分布或者随机分布,对  $S_H$  进行等差划分即可得到所求划分集合:定义  $S_{H_{\max}} = \max(S_H)$ ,  $\forall s_i \in S_{H_{\max}}, 1 \leq i \leq n$ ;  $S_{H_{\min}} = \min(S_H)$ ,  $\forall s_i \in S_{H_{\min}}, 1 \leq i \leq n$ 。对  $S$  进行等差划分所得  $S$  子集

$$S_i = \left\{ h_j \mid \left( \frac{S_{H_{\max}} - S_{H_{\min}}}{k} \right) (i-1) \leq \right.$$

$$\left. h_j < \left( \frac{S_{H_{\max}} - S_{H_{\min}}}{k} \right) (i), 1 < i \leq k, 1 \leq j \leq n \right\} \quad (7)$$

然后通过实体与 Hilbert 码的对应关系可获得实体的划分结果。

若实体呈聚集分布,则空间划分既要顾及实体之间的空间聚集特征,还要顾及分组后组间数据量的均衡。将实体的聚集特征描述为组内实体 Hilbert 离差平方和  $E_i$  及组间离差平方和  $E$  两个指

标,即

$$\begin{cases} E_t = \sum_{i=1}^{n_t} (h_i - h_t)^2, 1 \leq i \leq n_t, 1 \leq t \leq k \\ E = \sum_{t=1}^k E_t, 1 \leq t \leq k \end{cases} \quad (8)$$

式中,  $k$  为分组个数,  $t$  为组别标识,  $n_t$  表示第  $t$  组对象个数,  $h_t$  为  $t$  组 Hilbert 码均值,  $h_i$  为  $t$  组第  $i$  个 Hilbert 码。

呈聚集分布的点数据空间划分的具体步骤如下:

1) 首先对  $S_H$  进行等差划分, 将整个集合分解为  $k$  个组, 计算  $E_t$  和  $E$ 。

2) 对每分组下边界对象  $h_{n_t}$  进行调整: 若将  $h_{n_t}$  置入  $S_{t-1}$  后  $E$  减少, 则将  $h_{n_t}$  置入  $S_{t-1}$ , 完成一次组内上边界调整。若  $E$  不变, 则将  $h_{n_t}$  置入对象个数较小的一个分组。

3) 对分组下边界对象  $h_{n_t}$  进行调整: 若将  $h_{n_t}$  置入  $S_{t+1}$  后  $E$  减少, 则将置入  $S_{t+1}$ , 完成一次组内下边界调整。若  $E$  不变, 则将  $h_{n_t}$  置入对象个数较小的一个分组。

4) 对  $k$  个组分别进行局部上、下边界调整, 完成一次整体迭代寻优。

5) 重复进行步骤 4), 直至迭代寻优步骤使得  $E$  不再减少。

上述划分方法采用上、下边界调整的方法来进行局部寻优, 调整过程能够保持 Hilbert 码的大小顺序, 而 Hilbert 码的邻近性间接反映了实体之间的空间邻近性与聚集特征, 所以该方法能够顾及到实体之间的空间聚集特征; 在离差平方和相等的情况下, 采用数据数量补充的策略, 可以同时顾及到组间的数据量均衡; 采用组内离差平方和与组间离差平方和的定量指标, 能够同时顾及到整体与局部的空间划分质量。

### 3 实验分析

实验 1 是验证本文空间分布模式判断方法, 实验 2 的目标是定量分析本文空间划分方法在空间聚集性、数据量均衡等方面的情况。对于空间聚集性, 选用对象与分组中心的欧氏距离平方和来评价, 值越大说明聚集性越差。对于数据量均衡情况, 选择各组数据量的方差来衡量, 方差越大说明数据量均

衡性越差。

#### 实验 1

图 3 中实验数据是河流和山地分割下的城市边缘地带的 POI 数据, 包含 709 个点数据。采用 ArcGIS 平台中 Average Nearest Neighbor Distance 点模式分析工具对实验数据分析, 结果表明该数据呈聚集分布且显著性水平为 0.01。对于本文方法中 Hilbert 曲线阶数的选取, 对包含一个以上 POI 点的 Hilbert 区块 (347 个) 进行统计分析。  $I = (347 - 1) \times 2.076 / 2.043 = 351.59$ 。对其进行  $\alpha = 0.05$  的显著性检验。查表可得  $z_{0.05} = 1.64$ , 故

$$\chi_{0.05}^2(347 - 1) \approx \frac{1}{2}(z_{\alpha} + \sqrt{2 \times (347 - 1) - 1})^2 = \frac{1}{2}(1.645 + \sqrt{2 \times (347 - 1) - 1})^2 = 390.09$$

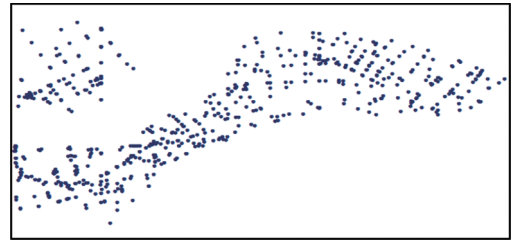


图 3 实验数据

Fig. 3 Sample data set

$I$  统计值小于  $\chi^2(347 - 1)$ , 所以该数据在受限区域内显著地呈随机分布, 可采用算法时间复杂度和空间复杂度都要小于迭代寻优划分方法的等差划分方法来进行数据划分。图 4 为实验数据按照 Hilbert 编码等差划分为 6 组的结果。

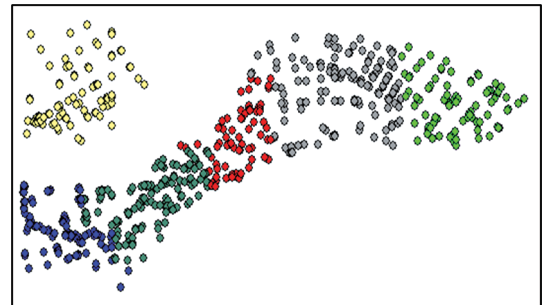


图 4 实验数据划分结果

Fig. 4 Result of partitioning of the sample dataset

#### 实验 2

采用南京市 POI 数据集 (包含 11 964 个实体) 进行实验。实验将本文方法与基于空间聚类、R-Tree 两类空间数据划分方法进行对比实验。采用

k-means 方法作为空间聚类数据划分方法的比较对象。k-means 聚类距离作为相似度测度,采用离差平方和作为准则函数进行聚类。R-Tree 及其变种均采用启发式数据分组方法来进行数据划分,典型的数据节点分裂方法包括 Guttman 提出的 QuadraticS-

plit 与 LinearSplit 方法以及 Beckmann 针对重叠度等多目标提出的 R\*-Tree 分裂方法<sup>[13-14]</sup>。实验分别采用 Linear R-Tree、Quadratic R-Tree 与 R\*-Tree 作为比较对象。实验中将原始数据划分为 7 组进行定量分析。实验结果如图 5 所示。

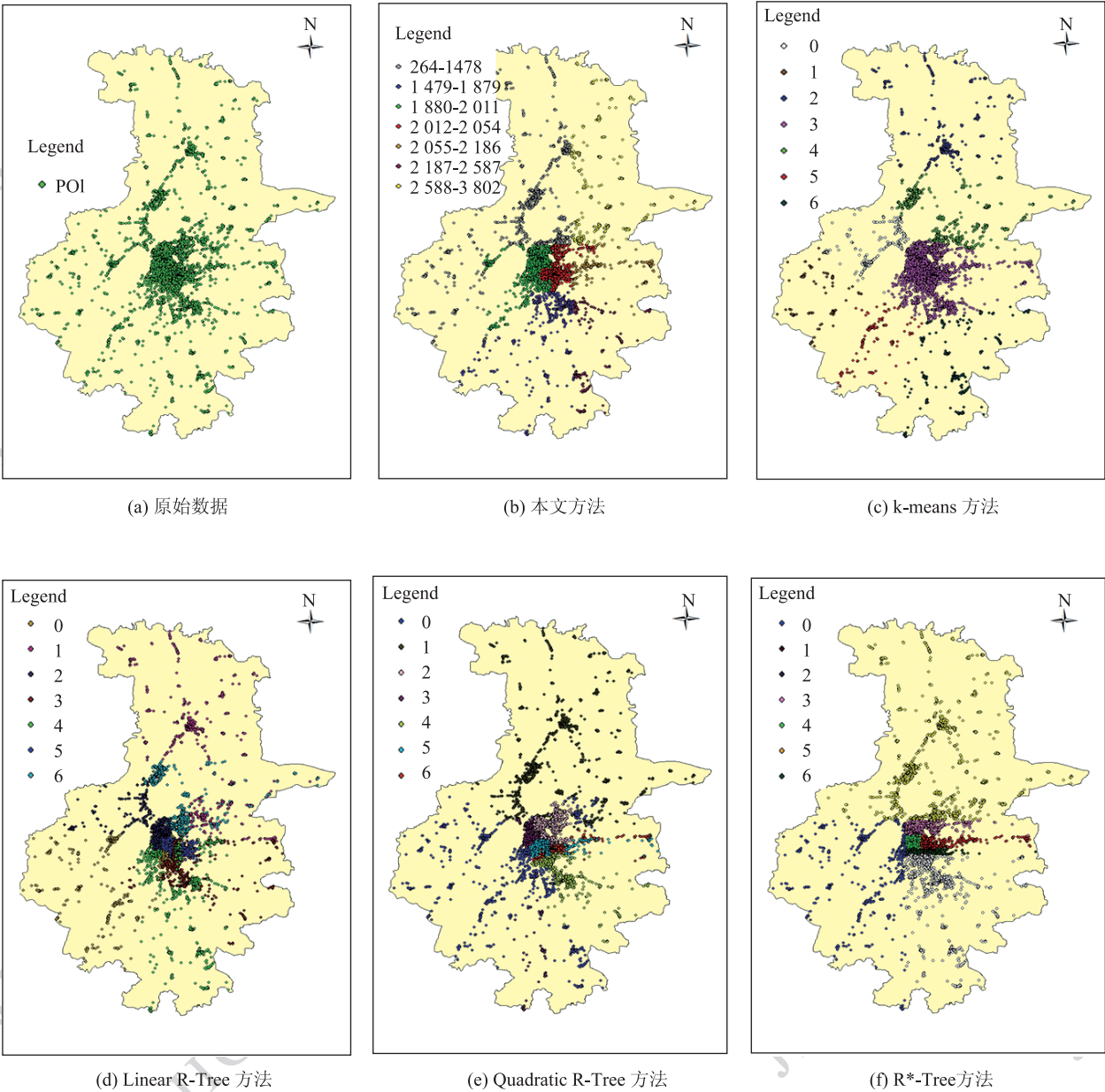


图 5 实验数据及划分结果

Fig. 5 Sample data set and its partitioning

分别统计 5 种方法所得分组结果的组内离差平方和组间离差平方和(表 1)。由表 1 可知,k-means 方法直接以离差平方和作为划分依据,其划分结果能够很好地反映数据集的聚集特征。Linear R-Tree、Quadratic R-Tree、R\*-Tree 作为 R-

Tree 的变种,以外接矩形的面积变化、重叠度作为划分依据,而不是将空间邻近性或者离散性指标作为划分依据。其划分结果没有很好地反映数据集的空间聚集特征,划分结果的离差平方比 k-means 方法划分结果的离差平方和大。本文方法

以局部和整体离差平方和作为划分指标,能够很好地反映空间聚集性,其离差平方和要小于 3 种 R-Tree 变种。在分组调整过程中为了防止数据倾斜而进行了数据量调整,所以本文方法的整体离差平方和要大于 k-means 方法。

表 1 5 种划分方法所得结果组间离差平方和对比

Table 1 The comparison of sum of squares by five kinds of partitioning approach

| 组别    | 本文方法    | k-means | Linear R-Tree | Quadratic R-Tree | R*-Tree   |
|-------|---------|---------|---------------|------------------|-----------|
| 1     | 428 734 | 24 778  | 195 999       | 444 875          | 195 107   |
| 2     | 32 170  | 11 018  | 311 949       | 480 563          | 577 387   |
| 3     | 109 321 | 92 345  | 75 856        | 21 936           | 206 185   |
| 4     | 34 563  | 285 022 | 94 603        | 124 770          | 20 083    |
| 5     | 62 732  | 135 772 | 300 316       | 48 311           | 3 528     |
| 6     | 35 832  | 12 660  | 10 352        | 36 822           | 53 795    |
| 7     | 150 763 | 54 710  | 135 540       | 42 677           | 10 826    |
| 离差平方和 | 854 115 | 616 305 | 1 124 615     | 1 199 954        | 1 066 911 |

分别记录 5 种方法所得分组结果的数据量并计算其数据量的方差(表 2)。由表 2 可知,Linear R-Tree、Quadratic R-Tree、R\*-Tree 作为 R-Tree 的变种,虽然节点分裂策略有所不同,但是均属于 R-Tree 的变种。R-Tree 作为一种平衡树,能够保证分组后组间数据量的均衡。k-means 方法采用离差平方和作为聚类依据,能够反映出组间的离散程度,但是不

表 2 5 种划分方法所得结果组间数据量对比

Table 2 The comparison of data amount by five kinds of partitioning approach

| 组别 | 本文方法      | k-means   | Linear R-Tree | Quadratic R-Tree | R*-Tree   |
|----|-----------|-----------|---------------|------------------|-----------|
| 1  | 1 932     | 472       | 1 041         | 1 637            | 1 045     |
| 2  | 965       | 173       | 1 084         | 1 955            | 2 339     |
| 3  | 3 860     | 765       | 2 411         | 2 216            | 1 832     |
| 4  | 3 047     | 8 705     | 1 462         | 1 204            | 1 818     |
| 5  | 890       | 1 198     | 2 538         | 1 117            | 1 221     |
| 6  | 690       | 247       | 1 847         | 2 290            | 1 940     |
| 7  | 580       | 404       | 1 581         | 1 545            | 1 769     |
| 方差 | 1 667 345 | 9 637 701 | 352 221.8     | 215 449.1        | 193 318.5 |

能顾及到组间数据量的均衡。本文方法采用离差平方和作为优化指标,迭代寻优过程中顾及了组间的数据量均衡。以空间分布模式作为出发点,而不是单纯以数据量均衡作为划分依据,在基于空间模式的划分过程中充分顾及数据量的均衡。所以本文划分方法的数据量均衡性要低于单纯以数据量作为划分依据的 3 种 R-Tree 变种,但要高于单纯以空间聚集性为划分原则的 k-means 方法。本文方法可以在顾及空间聚集特征的同时有效避免数据倾斜。

对于空间重叠度而言,在 3 种 R-Tree 变种中,Linear R-Tree 与 Quadratic R-Tree 以分割后叶子节点的面积变形作为依据,没有顾及到叶子节点间的重叠,所以划分结果组间存在多处的空间重叠。R\*-Tree 依据空间重叠度而进行叶子节点优化,其空间重叠度要小于 Linear R-Tree 与 Quadratic R-Tree。k-means 以组内离差平方和作为划分依据,组间没有空间重叠现象。本文方法在空间划分中也采用离差平方和作为分组依据,组间没有 Hilbert 码,也不存在空间重叠现象。

## 4 结 论

文献[2]已经证明空间划分问题是一个 NP 完全问题。划分过程中涉及空间聚集性、数据量均衡等指标。这些指标量纲不一致,不可公度。本文不试图解决多目标的同时优化求解。而是提出一个基于空间分布模式探测和 Hilbert 编码的空间划分方法。从空间排序和空间模式探测出发,给出不同分布类型数据集的不同划分策略:对于均匀分布或者随机分布数据,采用 Hilbert 编码等差划分方法。等差划分能够保证数据量均衡,采用 Hilbert 编码能够保持空间邻近性。对于聚集数据采用逐步寻优的划分方法。本文所提出的目标函数本身能够保持空间聚集特征,在寻优过程中通过数据量补偿策略兼顾数据量均衡。

两个实验结果表明:同时满足空间聚集性和数据量均衡是困难的,k-means 是满足空间聚集性而不满足数据量均衡的例子,R 树的 3 种变种分组方法是满足数据量均衡而牺牲空间聚集性的例子。本文方法空间聚集性要优于 R 树的 3 种变种,数据量均衡要优于 k-means 方法。对同时强调空间聚集性、空间重叠度与数据量均衡的应用场景(如分布式数据管理、并行计算等)有一定的应用价值。

## 参考文献 (References)

- [ 1 ] Shekhar S S, Ravada S, Kumar V P. Declustering and load-balancing methods for paralleling geographic information systems [J]. IEEE Transactions on Knowledge and Data Engineering, 1998, 10(4):632-655.
- [ 2 ] Zhao C Y, Meng L K, Lin Z Y. Spatial data partitioning towards parallel spatial database system [J]. Geomatics and Information Science of Wuhan University, 2006, 31(11):391-394. [赵春宇, 孟令奎, 林志勇. 一种面向并行空间数据库的数据划分算法研究[J]. 武汉大学学报:信息科学版, 2006, 31(11):391-394.]
- [ 3 ] Yan Z M, Liu R Y, Liu N. Multi-servers clustering of spatial data and realization strategies in high performance [J]. Journal of Remote Sensing, 2005, 9(4):421-428. [严志民, 刘仁义, 刘南. 空间数据多服务器群集及高性能实现策略[J]. 遥感学报, 2005, 9(4):421-428.]
- [ 4 ] Luo Y W, Wang X L, Xu Z Q. Collaborative and parallel processing of spatial information [J]. Journal of Computer-Aided Design & Computer Graphics, 2003, 15(10):1307-1314. [罗英伟, 汪小林, 许卓群. 空间信息合作与并行处理[J]. 计算机辅助设计与图形学学报, 2003, 15(10):1307-1314.]
- [ 5 ] Zhou Y, Zhu Q, Zhang Y T. A spatial data partitioning algorithm based on spatial hierarchical decomposition method of Hilbert space-filling curve [J]. Geography and Geo-Information Science, 2007, 23(4):13-17. [周燕, 朱庆, 张叶廷. 基于 Hilbert 曲线层次分解的空间数据划分方法[J]. 地理与地理信息科学, 2007, 23(4):13-17.]
- [ 6 ] Abugov D. Oracle spatial partitioning: best practices [R]. California: Oracle Corporation, 2010.
- [ 7 ] Zhang C X, Sun L Q. Multi-dimensional-data partition based on improved-Hash method [J]. Journal of Harbin University of Science and Technology, 2001, 6(1):24-27. [张春祥, 孙立鏖. 基于 Improved-Hash 方法的多维数据划分[J]. 哈尔滨理工大学学报, 2001, 6(1):24-27.]
- [ 8 ] Lu F, Zhou C H. A GIS spatial indexing approach based on Hilbert ordering code [J]. Journal of Computer-Aided Design & Computer Graphics, 2001, 13(5):424-429. [陆锋, 周成虎. 一种基于 Hilbert 排列码的 GIS 空间索引方法[J]. 计算机辅助设计与图形学学报, 2001, 13(5):424-429.]
- [ 9 ] Saita C A, Lirbat F. Clustering multidimensional extended objects to speed up execution of Spatial Queries [C]//Proceedings of the 9th International Conference on Extending Database Technology. Heraklion, Crete: Springer-Verlag, 2004:403-421.
- [ 10 ] Li G Q, Deng M, Liu Q L, et al. A spatial clustering method adaptive to local density change [J]. Acta Geodaetica et Cartographica Sinica, 2009, 38(3):255-263. [李光强, 邓敏, 刘启亮, 等. 一种适应局部密度变化的空间聚类方法[J]. 测绘学报, 2009, 38(3):255-263.]
- [ 11 ] Ai T H, Shuai Y, Li J Z. A spatial query based on shape similarity cognition [J]. Acta Geodaetica et Cartographica Sinica, 2009, 38(4):356-362. [艾廷华, 帅赞, 李精忠. 基于形状相似性识别的空间查询[J]. 测绘学报, 2009, 38(4):356-362.]
- [ 12 ] Zhao B B, Deng M, Li G Q. A new hierarchical spatial index for area entities based on urban morphology [J]. Acta Geodaetica et Cartographica Sinica, 2010, 39(4):435-440. [赵彬彬, 邓敏, 李光强. 基于城市形态学原理的面状地物层次索引方法[J]. 测绘学报, 2010, 39(4):435-440.]
- [ 13 ] Guttman A. R-trees: a dynamic index structure for spatial searching [C]//Proceedings of the ACM International Conference on Management of Data. New York, USA: ACM, 1984:47-57.
- [ 14 ] Zhu Q, Gong J, Zhang Y T. An efficient 3D R-tree spatial index method for virtual geographic environments [J]. Journal of Photogrammetry & Remote Sensing, 2007, 62(3):217-224.
- [ 15 ] Liu X, Schrack G. Encoding and decoding the hilbert order [J]. Software Practice and Experience, 1996, 26(12):335-346.
- [ 16 ] Wang Y F, He H L. Methods of Spatial Data Analysis [M]. Beijing: Science Press, 2007. [王远飞, 何洪林. 空间数据分析方法[M]. 北京: 科学出版社, 2007.]