DOI:10.13203/j.whugis20140831 文章编号:1671-8860(2016)12-1584-09

利用层次约束 Delaunay 三角网探测空间点事件离群模式

石 岩[1] 杨学习[1] 邓 敏[1]

1 中南大学地球科学与信息物理学院,湖南 长沙,410083

摘 要:空间离群模式探测是空间数据挖掘的一个研究热点。以带有空间位置属性的点事件为研究对象,针对现有方法的局限性,在扩展了空间离群模式定义的基础上引入层次约束 Delaunay 三角网,发展了一种空间点事件离群模式探测方法(简称层次约束 TIN 法)。首先,借助 Delaunay 三角网粗略地构建空间点事件间的邻接关系;然后,利用统计学方法针对 Delaunay 三角网的边长特性进行三个层次约束分析,以精化空间点事件的邻近域;最后,对具有空间邻接关系的点事件集合进行统计分析,以形成一系列空间簇,并通过一个统计约束指标提取数量较少的空间簇,即空间点事件离群模式。该方法不需要人为输入参数,通过模拟数据和实际数据实验,证明该方法可以有效、稳健地识别各类空间点事件离群模式。

关键词:空间点事件;空间属性;空间邻近域;空间离群模式;层次约束 Delaunay 三角网

中图法分类号:P208 **文献标志码**:A

空间离群模式探测旨在从海量空间数据中挖掘出偏离整体或局部分布模式的小部分空间实体,已成为空间数据挖掘的一个重要研究分支。在地理信息科学、环境学、社会经济学等诸多领域,空间离群模式并非是无意义的噪声数据,很可能蕴含着潜在的、未知的重要知识和规律,并在异常气候事件探测、环境监测、犯罪和疾病异常分布等方面发挥着重要作用。

Hawkins 最先提出离群点的概念,将其定义为"严重偏离其他对象的观测点,以至于令人怀疑它是由不同机制产生的"[3]。针对空间数据的特性,Shekhar 等将空间离群点定义为"非空间属性与空间邻近域内其他实体差异显著,而与整体数据集相比差异可能不明显的空间实体"[4]。

传统离群点探测方法可大致分为基于距离的方法[5]、基于密度的方法[6-8]、基于聚类的方法[9-11]等。基于距离的方法将数据集中远离其他实体的对象识别为离群点,侧重于发现全局离群点,而忽略了偏离局部分布的离群点。基于密度的方法主要根据实体与其邻域实体的密度差异定义其局部离群度(local outlier factor,LOF),可在非均匀分布的数据中有效识别全局和局部离群点和离群小簇,但需要输入大量参数对实体的离群度进行深入分析才能识别异常实体。基于聚类的方法是借助聚类手段(例如 K-means)来探测离群模式,严重依赖聚类方法的选择。

进而,许多学者基于 Shekhar 对空间离群点的定义发展了一系列探测方法,可大致分为基于距离的方法[4,12,13]、基于密度的方法[14,15]、基于聚类的方法[16,17]以及基于模型的方法[18-20]。其中,基于距离和密度的方法是传统离群点探测方法在空间数据集的扩展,未能避开传统方法的缺陷。基于聚类的方法是将聚类结果中不隶属于任何簇的空间实体识别为空间离群点,主要目的是发现空间簇,探测离群模式的能力有限[16]。基于模型的方法并非数据驱动,需要首先满足模型的假设条件,如空间数据服从某种分布,这在实际应用中难以实现,可能导致探测结果偏离实际情况。

通过对现有研究的总结分析可以发现,传统方法可经过扩展来探测空间点事件离群模式,但这些方法并非专门针对空间数据集,因此缺乏空间邻近关系的精确度量;并且,这些方法仅将离群模式分为全局和局部离群,缺乏全面性。空间离群探测方法核心在于分析非空间属性差异,无法

收稿日期:2015-07-24

项目资助:国家 863 计划(2013AA122301);教育部新世纪优秀人才支持计划(NECT-10-0831);湖南省自然科学基金(14JJ1007)。

第一作者:石岩,博士,主要从事时空数据挖掘分析及其应用研究。csu_shiy@126.com

通讯作者:邓敏,博士,教授。dengmin208@tom.com

直接用于空间点事件离群模式的探测。图 1 给出了三类空间点事件离群模式,图 1(a)为全局离群模式,即明显偏离整体分布的空间点和空间簇;图 1(b)为局部离群模式,即明显偏离局部分布的空间点和空间簇;图 1(c)为内部离群模式,这类离群模式以密集小簇的形式存在于空间大簇的内部,对这类模式进行研究,具有重要的实际应用。例如,某区域犯罪事件的发生呈均匀分布,而在此区域内存在较小规模且不易被发现的密集区域,重点研究此类区域可有助于深入分析犯罪分布的发展规律,从而有效抑制潜在大规模犯罪事件的发生。

针对此,本文首先对空间点事件离群模式进行重新定义,进而利用层次约束 Delaunay 三角网发展了一种空间点事件离群模式的探测方法——层次约束 TIN 法。

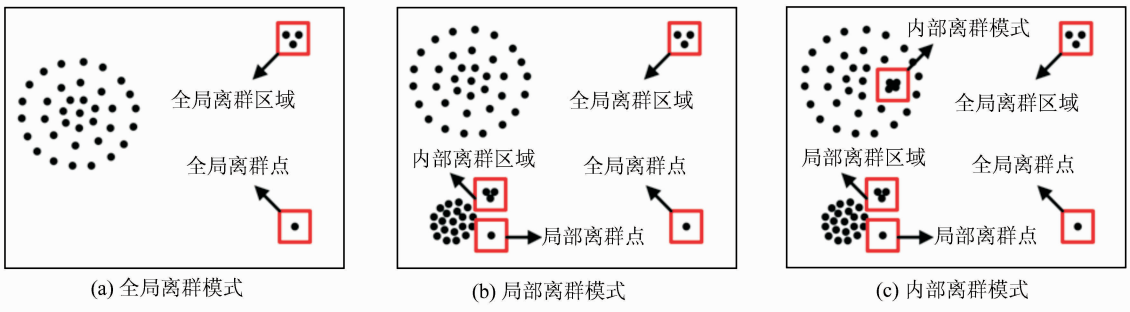


图 1　空间离群模式类型

Fig. 1　Types of Spatial Outlier Patterns

1　基于层次约束 Delaunay 三角网的空间点事件离群模式探测

本文中,定义空间点事件数据集中偏离整体或局部普适聚集分布模式的孤立点、稀疏或密集极小丛簇为空间离群模式。顾及空间点事件数据集的特性,借助 Delaunay 三角网进行空间邻近域的构建和表达。由于连接全局离群模式、局部离群模式、内部离群模式与其他正常聚集分布模式的边特性互不相同,本文对原始 Delaunay 三角网施加 3 个层次约束,并分别用来提取各类离群模式。具体过程主要包括 3 个步骤。

1.1　空间邻接关系粗表达

初始空间数据集呈离散状态,缺乏对空间实体间邻接关系的表达。可通过构建空间邻近域来判断某空间点事件的空间离群程度。Delaunay 三角网是一种建立空间点实体间邻接关系的有效工具[21,22],但不能准确描述复杂的空间数据集中的实体间邻接关系。如图 2 所示,对模拟空间数据集建立 Delaunay 三角网,其中与红色虚线相交的边均为明显误差边,通过误差边建立的邻接关系是不准确的。鉴于此,本文在利用 Delaunay 三角网粗略表达空间点事件间邻接关系的基础上,通过施加层次约束以实现空间邻接关系的逐步精化。

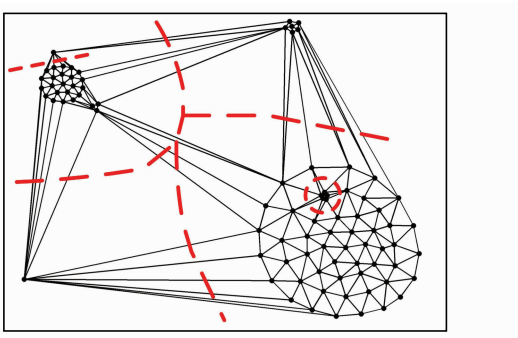


图 2　空间邻接关系的 Delaunay 三角网表达

Fig. 2　Spatial Adjacency Relationships Described by Delaunay Triangulation

1.2　空间邻接关系精描述

针对 Delaunay 三角网在表达空间点事件间邻接关系时存在的误差问题,本文采取一种三阶层次约束策略对原始 Delaunay 三角网进行修复以精确描述空间邻接关系。一些学者利用不同的边长约束指标对 Delaunay 三角网进行修复以用于空间聚类分析[21,22],并证明了其有效性。这些指标可以归纳为:

$$CI = Mean + c * Std \tag{1}$$

式中,Mean 和 Std 分别表示边集合中的平均边长和边长标准差;c 为可调节系数,或为预设常数,或为随边长变化的量值。仿照这种形式,本文构造三个层次约束指标和一个离群簇提取指标。

定义 1　一阶约束指标,是指针对空间点事件数据集建立的 Delaunay 三角网中所有边定义

的约束指标，表达为：

$$\begin{cases} \mathrm{CI}_1(E_i) = \mathrm{G\_Mean(TIN)} + \alpha \cdot \mathrm{G\_STD(TIN)} \\ \alpha = \dfrac{\mathrm{G\_Mean(TIN)}}{l(E_i)} \end{cases} \tag{2}$$

式中，E_i 和 $l(E_i)$ 分别为 Delaunay 三角网中任一边及其边长；G_Mean(TIN)和 G_STD(TIN)分别为三角网的平均边长和边长标准差；α 为适应系数。

Delaunay 三角网的平均边长和边长标准差能从宏观层次较好地反映空间点事件的整体分布，并大致区分其中较长和较短的边。为了精确识别一阶长边，引入适应系数 α。若某条边的长度大于一阶整体平均边长，则 $\alpha<1$，相应的一阶约束指标越小；反之，则反之。据此删除所有长度大于一阶约束指标的边，可以得到更新后的空间邻接关系，如图 3(a)所示。可以看出，一阶长边已有效删除，并且分离得到全局离群点 G_2 和离群簇 G_4，但在某些局部仍存在误差需要消除，其中蕴含着局部和内部离群模式，如图 3(a)虚线框所示。

定义 2 二阶约束指标，即经过一阶边长约束后，每个空间点事件与其更新后的邻接点事件构成的二阶局部边，进而定义的二阶局部边约束指标，表达为：

$$\begin{cases} \mathrm{CI}_2(E_i) = \mathrm{L\_Mean}(E_{P_i}) + \beta \cdot \dfrac{\sum_{j=1}^{n} \mathrm{L\_STD}(E_{P_j})}{n} \\ \beta = \dfrac{\mathrm{L\_Mean}(E)}{l(E_i)} \end{cases} \tag{3}$$

式中，E_i 为与空间实体 P_i 连接的局部边集合 E_{P_i} 的成员；$\mathrm{L\_Mean}(E_{P_i})$ 和 $\mathrm{L\_STD}(E_{P_j})$ 分别为局部边的平均边长和边长标准差，P_i、$P_j \in G_k$；G_k 为施加一阶约束后的子图集合，如图 3(a)中 $G_1 \sim G_4$；β 为适应系数。

以一阶约束后的各子图为分析单元，局部边的平均边长和边长标准差从局部层次表征了空间点事件在局部的分布模式，通过附加适应系数 β 构成的局部约束指标能很好地识别局部较长边引起的误差。进而，删除边长大于二阶约束指标的局部边，继续更新空间邻接关系，如图 3(b)虚线框所示。可以发现，由局部长边引起的误差已基本消除，并且分离得到了局部离群点 G_5 和离群簇 G_6，但某些局部仍然存在误差，因此无法分离内部离群模式。接下来以局部边长标准差为分析对象，首先识别局部边长标准差较大的分布非均匀区域，进而对此类区域施加三阶约束实现空间邻接关系的精确化。

定义 3 三阶约束指标。经过二阶约束，每个空间点事件获得分布较为均匀的三阶局部边。为进一步准确识别三阶局部长边，首先根据空间点事件的邻域边长标准差识别全局和局部分布非均匀邻域，分别为：

$$\begin{cases} \mathrm{GSI}(\mathrm{Local\_STD}_{P_i}) = \mathrm{Mean(Local\_STD)} + \alpha' \cdot \mathrm{SD(Local\_STD)} \\ \alpha' = \dfrac{\mathrm{Mean(Local\_STD)}}{\mathrm{Local\_STD}_{P_i}} \end{cases} \tag{4}$$

$$\begin{cases} \mathrm{LSI}(\mathrm{Local\_STD}_{P_i}) = \mathrm{L\_Mean}(P_i) + \beta' \cdot \dfrac{\sum_{j=1}^{n} \mathrm{L\_STD}(P_j)}{n},\ P_j \in G_k \\ \beta' = \dfrac{\mathrm{L\_Mean}(P_i)}{\mathrm{Local\_STD}_{P_i}} \end{cases} \tag{5}$$

式中，$\mathrm{Local\_STD}_{P_i}$ 为空间点事件 P_i 连接的邻域边长标准差；Local_STD 为 P_i 所在子图中所有点事件的邻域边长标准差集合；α' 为适应系数。式(5)中，$\mathrm{L\_Mean}(P_i)$ 和 $\mathrm{L\_STD}(P_j)$ 分别为空间点事件 P_i 的邻域边长标准差均值和标准差；β' 为适应系数。分布非均匀邻域所对应的实体需要满足 $\mathrm{Local\_STD}_{P_i} \geqslant \mathrm{GSI}(\mathrm{Local\_STD}_{P_i})$（简称“条件一”）或者 $\mathrm{Local\_STD}_{P_i} \geqslant \mathrm{LSI}(\mathrm{Local\_STD}_{P_i})$（简称“条件二”）。

需要注意的是，为了避免全局分布不均匀邻域的影响，在计算局部分布不均匀邻域识别指标时，忽略满足条件一的点事件。进而，根据二阶约束指标对此类点事件的邻域边长施加约束以删除残余局部误差，通过以上约束，成功分离得到内部离群模式 G_7，并获得空间点事件间的精确邻接关系，如图 3(c)所示。

1.3 空间离群模式自动识别

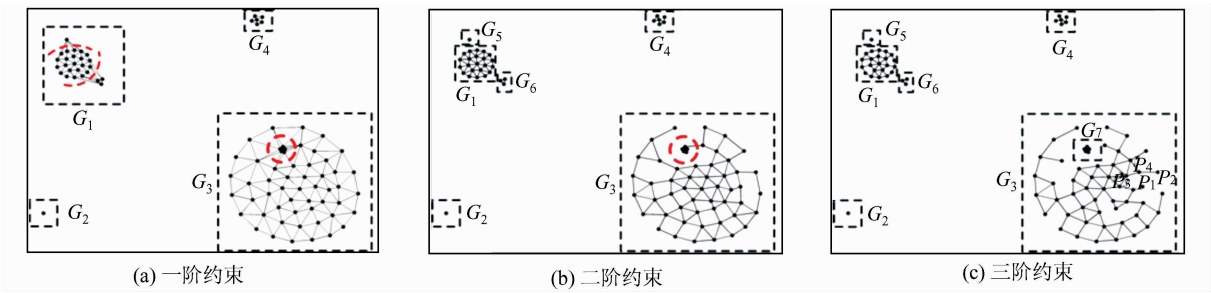
通过空间点事件间的精确邻接关系可得到一系列连通子集，下面探测其中的孤立点和离群簇。

定义 4 连通子集，即针对任一空间点事件，以其空间邻域作为传递路径进行递归扩展，经过路径上的所有点事件构成一个连通子集 G，如图 3(c)中 $G_1 \sim G_7$，各类聚集和离群分布模式蕴含于其中。

根据本文对空间点事件离群模式的定义，构造以下指标识别空间点事件离群模式。

定义 5 空间离群模式识别指标。记各连通子集中含空间点事件的数目为 N_i 并构成集合 N，

为避免孤立点和小簇对探测结果的影响,对 N 中相同数值进行合并以构成集合 N'。进而,定义空间点事件离群模式指标为:



(a) 一阶约束　(b) 二阶约束　(c) 三阶约束

图 3　空间邻接关系的精描述

Fig. 3　Precise Spatial Adjacency Relationships

$$\left.\begin{aligned} \mathrm{SOI}(N_i') &= \mathrm{Mean}(N') - \gamma \cdot \mathrm{STD}(N') \\ \gamma &= \frac{N_i'}{\mathrm{Mean}(N')} \end{aligned}\right\} \quad (6)$$

式中,$\mathrm{Mean}(N')$和 $\mathrm{STD}(N')$分别为各连通子集包含空间点事件个数分类后的平均值和标准差;γ 为适应系数。N_i'越小,SOI 越大,反之越小。将连通子集中容量小于此指标的子集视为空间离群簇,如图 4 所示。

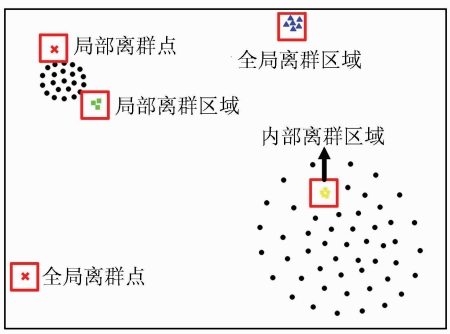


图 4　各类空间点事件离群模式的识别

Fig. 4　Identification of all Types of Outlier Patterns from Spatial Point Events

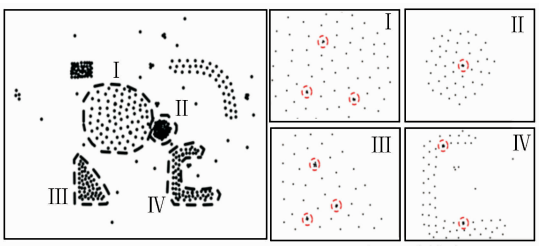
2　实验分析

本节分别采用模拟数据和实际数据进行实验验证分析。首先采用 AcrGIS 10.0 人工生成的模拟数据,进而采用 MATLAB 程序随机生成的复杂模拟数据来研究本文方法的稳健性和运行效率。另外,分别采用我国主要气象监测站点数据和黑龙江省洪河自然保护区毛果苔草群落空间分布数据进行实例应用,以验证方法的实用性。

2.1　模拟实验

2.1.1　有效性分析

图 5(a)为一组通过 AcrGIS 10.0 人工生成的空间点事件模拟数据集,包含了全局、局部以及内部离群模式,各类离群模式包含任意形状和密度,具有很好的代表性和说明性。图 5(b)为包含内部离群模式的 4 个子区域的放大显示。本节通过对比实验分析来说明本文方法的有效性。



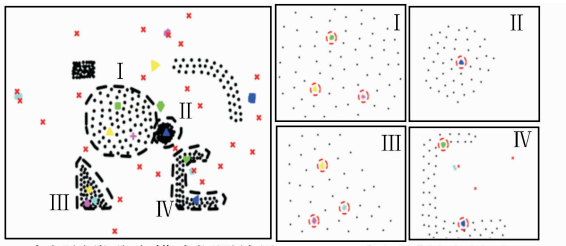
(a) 模拟数据　(b) 局部区域放大显示

图 5　模拟数据集 1

Fig. 5　Simulated Dataset 1

图 6 和图 7 分别为本文方法探测结果、基于 K 邻近距离法以及基于 K-邻域的 LOF 法的探测结果。从图 6 可以发现,本文方法不仅能够从不均匀分布数据中有效识别出全局和局部空间离群模式,而且能够探测隐藏在空间聚集簇中异常密集分布的内部离群模式。虽然在某些离群簇模式中,极少数离群簇结构被打乱而形成了离群点和子簇,但最终得到的离群模式与预先设置的结果完全一致。其余两种经典方法的 K 值分别设置为 10、15。离群点数目设置为两类,一类为除内部离群模式外其他离群模式中点的数目 41,另一类为所有预设离群模式中点的数目 85。如图 7 所示,当离群点数目设置为 41 时,基于 K 邻近距离法可以有效探测数据中的全局离群模式,而在分布复杂、密度不均匀的某些局部区域则忽略了一些局部离群模式,如数据分布右下角密集圆形簇和"[”状簇附近的局部离群点和离群簇无法识别。另外,对于某些聚集模式(如"⟍”状簇两段位置)的边界处则误判为离群模式。基于 K-邻域的 LOF 法可以有效探测全局离群点、离群簇模式和

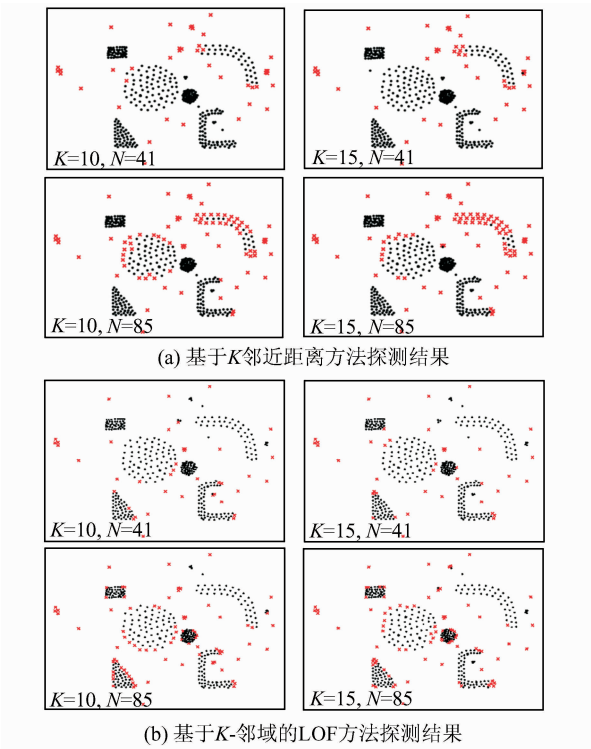
局部离群点模式，但无法全面有效地识别局部离群簇模式。当离群点数目设置为 85 时，两种方法均无法识别内部离群模式，而将聚集模式（如中间的大球形簇）的边界点识别为离群点。



(a) 空间异常分布模式探测结果 (b) 局部区域放大显示

图 6 本文方法探测结果

Fig. 6 Results Obtained by Method Proposed in this Paper



(a) 基于K邻近距离方法探测结果

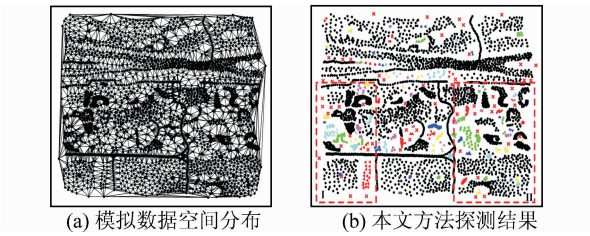
(b) 基于K-邻域的LOF方法探测结果

图 7 各方法探测结果比较

Fig. 7 Results Obtained by Different Methods

2.1.2 稳健性分析

本节通过采用 MATLAB 随机生成一组复杂模拟数据来分析本文方法的稳健性和运行效率，该数据分布和相应的 Delaunay 三角网如图 8(a)所示，图 8(b)给出了本文方法的探测结果。可以发现，探测结果具有以下特征。整体看来，本文方法可以从中探测离群点以及形成空间小簇的离群区域；局部看来，如图 8(b)中区域 I 和 II，可以探测得到圆形、条状、稀疏、密集等形状和密度各异的各类空间离群簇，这些离群模式可以分布于其他空间簇的外部、内部等任何区域；图 8 中的部分离群簇在局部放大看来更像一种正常聚集模式，这是由于该数据包含实体数目较多，在进行各类模式的分离时，各正常空间簇亦包含较多实体，因此利用离群模式识别指标提取的离群簇所包含的实体数目可能与数据量较小的数据中的空间正常簇相当，即离群簇的大小是个相对值，该值与正常簇的大小相关。

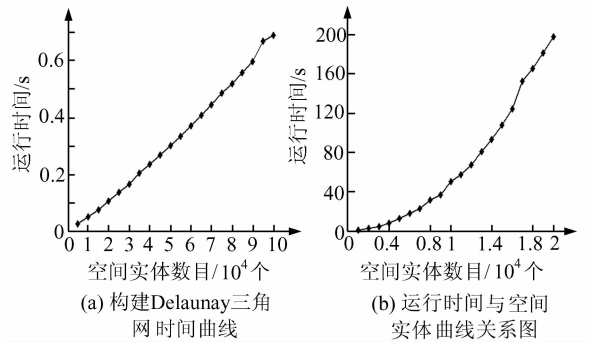


(a) 模拟数据空间分布 (b) 本文方法探测结果

图 8 模拟数据集 2 空间分布及其探测结果

Fig. 8 Simulated Dataset 2 and Detection Result

本文方法的核心在于构建 Delaunay 三角网。下面首先验证构建 Delaunay 三角网的效率，其时间曲线如图 9(a)所示。可以看出，构建 Delaunay 三角网所需时间与数据量之间大致呈线性正相关关系，且运行时间较短，例如 100 000 个空间实体的运行时间接近 0.7 s，表明运行大数据量对运行时间不会产生严重影响。另外，本文方法运行时间与空间实体数目间的关系如图 9(b)所示。由于其主要分析对象为 Delaunay 三角网的各条边，随着数据量的增大，Delaunay 三角网中边的数目增长更快，因此运行时间的增长率呈现增大趋势，即运行时间变化曲线的斜率逐渐增大，但该方法同样具有较高的运行效率，可以稳健地适应海量空间数据集。



(a) 构建Delaunay三角网时间曲线 (b) 运行时间与空间实体曲线关系图

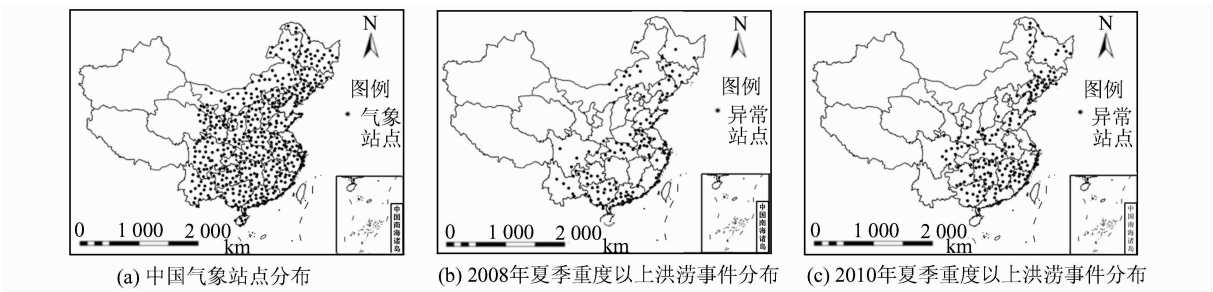
图 9 本文方法构建 Delaunay 三角网与运行效率图

Fig. 9 Running Time of Constructing Delaunay Triangulation and Implementing the Proposed Algorithm

2.2 实际应用

本节分别采用我国主要气象监测站点数据和黑龙江省洪河自然保护区的毛果苔草群落空间分

布数据来验证本文方法的实用性。其中,气象监测站点数据来源于国家气象信息中心气象资料室,包括 1982～2011 年间中国陆地区域 486 个站点的降水月均值数据,其空间分布如图 10(a)所示,气象站点相对均匀地分布于中国中部和东部地区。本文首先针对每个气象站点所记录的降水时间序列数据计算标准化降水指数(standardized precipitation index, SPI),然后根据 NDMC(National Drought Mitigation Center)对 SPI 指数的分类[23](表 1),设置时间尺度为 3 个月,即考虑 3 个月累计降水量,从各站点中提取重度以上洪涝事件(即 SPI≥1.5)发生的时间点,可以得到每年每月发生重度以上洪涝事件的气象站点。下面以夏季洪涝事件为研究对象,即若某站点在某年的 6、7、8 月任意一个月的 SPI≥1.5,则认为此站点在该年夏季发生了重度以上洪涝事件。进而,选取近年来夏季发生重度以上洪涝事件的区域分布较为广泛的 2008 年(图 10(b))和 2010 年(图 10(c))作为离群模式探测的空间点事件集。

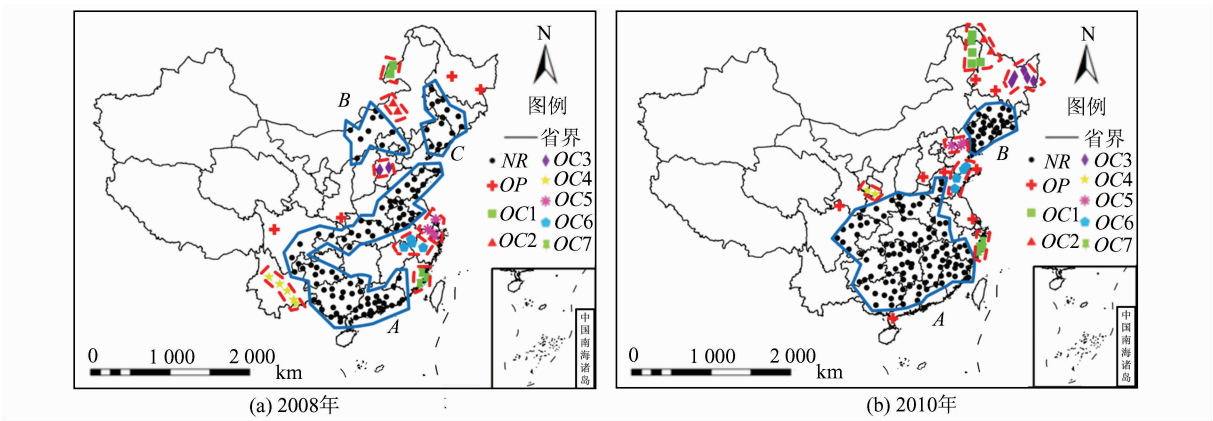


(a) 中国气象站点分布　(b) 2008年夏季重度以上洪涝事件分布　(c) 2010年夏季重度以上洪涝事件分布

图 10　实际数据集

Fig. 10　Real Dataset 1

图 11(a)、11(b)分别为不同年份本文方法对实际数据集的探测结果,其中 *NR* 表示正常分布区域,*OP* 表示离群点,*OC* 表示各离群空间簇。针对 2008 年夏季重度以上洪涝事件分布数据,本文方法探测得到 4 个离群点和 7 个离群空间簇,这些离群模式分布于黑龙江中部、内蒙古东部、河北中部、长江下游、福建沿海以及云南中部等地区;另外,正常分布区域可明显大致分为 *A*、*B*、*C* 三个子区域,而各离群模式分大都分布于这三个子区域的过渡区域。在这些离群模式中,云南地区具有极其复杂的地形;长江下游为河流入海口,且以平原为主,因此地形、海洋同时影响降水量;福建沿海通常为台风登陆我国的区域。综合以上可能的因素,可对本文方法探测得到的洪涝事件离群模式作出合理解释。针对 2010 年夏季重度以上洪涝事件分布数据,本文方法探测得到 7 个离群点和 7 个离群空间簇,这些离群模式分布于黑龙江北部和东部、环渤海湾、河北中部、浙江东北部等地区。另外,正常分布区域可大致分为 *A*、*B* 两个子区域,而各离群模式分大都分布于这三个子区域的过渡区域及外围区域。其中,黑龙江北部和东部具有山脉阻隔,环渤海湾地区夏季受到海风强烈影响,而浙江东北部沿海地区与其相邻内陆相比,受海洋、台风影响明显。这些都可解释为何这些区域的强降水具有离群模式。进一步,可以结合气象领域知识更深入地分析空间离群模式的成因,这对于研究我国气候变化规律、预测极端气候事件具有重要的研究意义。



(a) 2008年　(b) 2010年

图 11　异常降水事件空间离群模式探测结果

Fig. 11　Results of Spatial Outlier Patterns of Abnormal Precipitation Events

表 1 标准化降水指数 SPI 分类

Tab. 1 Classification of SPI Index

事件类型	极端干旱	重度干旱	轻度干旱	正常	轻度洪涝	重度洪涝	极端洪涝
SPI 指数	$(-\infty, -2)$	$(-2, -1.5)$	$(-1.5, -1)$	$(-1, 1)$	$[1, 1.5]$	$[1.5, 2]$	$[2, +\infty)$

进而，利用本文方法探测黑龙江省洪河自然保护区毛果苔草群落的空间离群分布模式。研究区域位于三江平原东北部，范围为(133°34′38″E～133°46′29″E，47°42′18″N～47°52′00″N)。本节所采用的数据来源于文献[24]，其原始数据为该区域 2010 年的 TM 影像和 2012 年的 SPOT 影像。首先通过融合、配准、几何校正、镶嵌和裁剪等处理获取研究区域标准的遥感影像数据，进而根据野外采集的植被类型样本建立解译标志，并利用监督分类方法进行影像解译，从而获得各种植被类型的空间分布，其中毛果苔草群落的空间分布如图 12(a)所示。可以发现，毛果苔草群落大致位于研究区域的边缘地带，具有块状和条状的分布特点。需要注意的是，该数据的数据量较大(共包含 14 738 个空间点)且分布复杂，因此通过肉眼无法准确获取其中隐藏的空间离群模式。利用本文方法对其进行空间离群探测可以得到如图 12(b)所示结果。从图 12 中可以直观地发现，该区域毛果苔草群落分布包含了大量空间离群模式。其中条状分布区域主要由空间离群簇构成，块状分布区域的周围也蕴含了大量离群模式。从图 12(b)对区域 I 和 II 的局部放大显示可以明显发现这一特点；另外，在聚集簇模式的内部也探测出一系列密集的离群簇，即内部离群模式；该实例也说明了本文方法可以从实际的海量空间点事件分布数据中稳健有效地进行空间离群模式探测。

在湿地中影响植被类型空间分布的因素有很多，例如高程、水位、水环境和土壤中各种元素的含量等，其中高程通常认为保持不变，因此各种植被的空间分布可以大致反映该区域水位以及水环境和土壤中各元素含量随空间变化的规律。另外，植被的空间离群分布模式未构成较大的分布规模，而代表了一类局部区域内更加细致的特殊分布，相应区域内的水位、水环境和土壤所含各元素含量也可能呈现出局部特殊的分布规律。进一步，通过在不同的时间段探测各种植被类型的空间离群模式，结合领域知识可以深入分析该区域的湿地类型空间分布随时间的动态变化规律，从而进一步帮助地方政府部门有效遏制破坏湿地生态环境的行为。

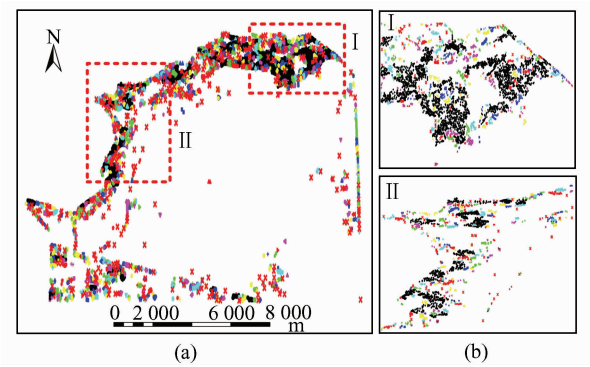


图 12 毛果苔草群落空间异常模式探测

Fig. 12 Detection of Spatial Outlier Patterns from Spatial Distribution of Carex Lasiocapa

3 结 语

本文提出了一种基于层次约束 Delaunay 三角网的空间点事件离群模式探测方法——层次约束 TIN 法。通过实验分析发现，本文方法可以有效地顾及空间点事件之间的邻近关系，稳健有效地探测各种类型空间离群模式，不需要人为输入参数。通过模拟数据和实际数据实验证明了本文方法的自适应性、有效性、稳健性和实用性。

进一步的工作将主要集中在顾及多维非空间属性的空间离群模式探测，对空间离群模式的有效性进行定量评估(本文对探测结果的评估主要源于已有先验知识)以及扩展到时空维的时空离群模式探测。

参 考 文 献

[1] Li Deren, Wang Shuliang, Li Deyi, et al. Theories and Technologies of Spatial Data Mining and Knowledge Discovery [J]. *Geomatics and Information Science of Wuhan University*, 2002, 27(3): 221-233(李德仁，王树良，李德毅，等. 论空间数据挖掘和知识发现的理论和方法[J]. 武汉大学学报·信息科学版，2002，27(3)：221-233)

[2] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques[M]. 3rd Edition. San Francisco: Morgan Kaufman, 2012

[3] Hawkins D M. Identification of Outliers [M]. London: Chapman and Hall, 1980

[4] Shekhar S, Lu C T, Zhang P S. A Unified Approach to Detecting Spatial Outliers [J]. *GeoInformatica*, 2003, 7(2): 139-166

[5] Ramaswamy S, Rastogi R, Shim K. Efficient Algorithms for Mining Outliers from Large Data Sets [C]. 2000 ACM SIGMOD International Conference on Management of Data, Dallas, USA, 2000

[6] Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying Density-based Local Outliers [C]. 2000 ACM SIGMOD International Conference on Management of Data, Dallas, USA, 2000

[7] Chiu A L M, Fu A C. Enhancements on Local Outlier Detection[C]. 7th International Database Engineering and Applications Symposium, Hong Kong, China, 2003

[8] Jin W, Tung A K H, Han J W, et al. Ranking Outliers Using Symmetric Neighborhood Relationship[C]. 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Singapore, 2006

[9] Jiang M F, Tseng S S, Su C M. Two-phase Clustering Process for Outliers Detection [J]. *Pattern Recognition Letters*, 2001, 22(6): 691-700

[10] Al-Zoubi M B, Al-Dahoud A A, Yahya A. New Outlier Detection Method Based on Fuzzy Clustering [J]. *WSEAS Transaction on Information Science and Applications*, 2010, 7(5): 681-690

[11] Gu Ping, Liu Haibo, Luo Zhiheng. Multi-clustering Based Outlier Detect Algorithm [J]. *Application Research of Computers*, 2013, 30(3):751-753,756 (古平, 刘海波, 罗志恒. 一种基于多重聚类的离群点检测算法 [J]. 计算机应用研究, 2013, 30(3): 751-753,756)

[12] Chen D C, Lu C T, Kou Y F, et al. On Detection Spatial Outliers [J]. *Geoinformatica*, 2008, 12: 455-475

[13] Li Guangqiang, Deng Min, Zhu Jianjun, et al. Spatial Outliers Detection Considering Distances Among Their Neighbors [J]. *Journal of Remote Sensing*, 2009, 2:197-202(李光强, 邓敏, 朱建军,等. 一种顾及邻近域内实体间距离的空间异常检测新方法 [J]. 遥感学报, 2009, 2: 197-202)

[14] Chawla S,Sun P. SLOM: A New Measure for Local Spatial Outliers [J]. *Knowledge and Information Systems*, 2006, 9(4): 412-429

[15] Ma Ronghua, He Zengyou. Fast Mining of Spatial Outliers from GIS Database [J]. *Geomatics and Information Science of Wuhan University*, 2006, 31(8): 679-682(马荣华, 何增友. 从 GIS 数据库中挖掘空间离群点的一种高效算法 [J]. 武汉大学学报·信息科学版, 2006, 31(8): 679-682)

[16] Liu Qiliang, Li Guangqiang, Deng Min. A Local Distribution Based Spatial Clustering Algorithm [J]. *Geomatics and Information Science of Wuhan University*, 2010, 35 (3): 373-377(刘启亮, 李光强, 邓敏. 一种基于局部分布的空间聚类算法 [J]. 武汉大学学报·信息科学版, 2010, 35 (3): 373-377)

[17] Telang A, Deepak P, Joshi S, et al. Detecting Localized Homogeneous Anomalies over Spatio-temporal Data [J]. *Data Mining and Knowledge Discovery*, 2014, 28(5-6): 1 480-1 502

[18] Chen F, Lu C T, Boedihardjo A P. GLS-SOD: A Generalized Local Statistical Approach for Spatial Outlier Detection[C]. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC, USA, 2010

[19] Cai Q, He H B, Man H. Spatial Outlier Detection Based on Iterative Self-organizing Learning Model [J]. *Neurocompuing*, 2013, 117: 161-172

[20] Albanese A, Pal S K, Petrosino A. Rough Sets, Kernel Set and Spatio-temporal Outlier Detection [J]. *IEEE Transactions on Knowledge & Data Engineering*, 2014, 26(1): 194-207

[21] Shi Yan, Liu Qiliang, Deng Min, et al. A Hybrid Spatial Clustering Method Based on Graph Theory and Spatial Density[J]. *Geomatics and Information Science of Wuhan University*, 2012, 37 (11): 1 276-1 280(石岩, 刘启亮, 邓敏,等. 融合图论与密度思想的混合空间聚类方法[J]. 武汉大学学报·信息科学版, 2012, 37(11): 1 276-1 280)

[22] Deng M, Liu Q L, Cheng T, et al. An Adaptive Spatial Clustering Algorithm Based on Delaunay Triangulation [J]. *Computer, Environment, Urban and Systems*, 2011, 35(4): 320-332

[23] Hayes M. Drought Indices [OL]. http://www.drought.unl.edu/whatis/indices.htm,2013

[24] Zheng Mingyue. The Application of Spatial Clustering Rules in the Honghe Wetland Vegetation Form Distribution Change of Gradient [D]. Harbin: Harbin Normal University, 2013(郑明月. 空间聚类规则在洪河湿地类型分布梯度变化中的应用 [D]. 哈尔滨: 哈尔滨师范大学, 2013)

Detection of Spatial Outlier Patterns from Point Events Based on Multi-constrained Delaunay Triangulation

SHI Yan[1] *YANG Xuexi*[1] *DENG Min*[1]

1 Department of Surveying and Geo-informatics, Central South University, Changsha 410083, China

Abstract: In recent years, spatial outlier detection has become a research hotspot in the domain of spatial data mining. The aim of spatial outlier detection is to discover those small parts of spatial entities deviating from the global or local distribution in massive spatial datasets. Spatial outliers may indicate potential, unknown, and important knowledge instead of noise in many application domains, e.g., environmental science, meteorology, urban traffic, and so on. Existing spatial outlier detection methods focus on detecting spatial outliers in the spatial datasets with non-spatial attributes. There is still a lack of detection methods specifically designed for spatial point event datasets, in particular, for complicated spatial point event datasets with clusters having arbitrary shapes and/or different densities. Therefore, we developed a method of detecting outlier patterns for spatial point events by considering spatial locations; the definition of a spatial outlier is extended and a multi-level constrained Delaunay triangulation is employed. Spatial adjacency relationships are roughly obtained from Delaunay triangulation. Then, three-level constraints are described and utilized for precise spatial adjacency relationships with the consideration of statistical characteristics. Finally, those spatial point events connected by the remained edges are gathered to form a series of clusters. The clusters containing very few point events are regarded as spatial outlier patterns. This algorithm does not involve any parameters. Experiments on both synthetic and real-world spatial datasets demonstrate that this algorithm can detect all kinds of spatial outlier patterns efficiently and robustly.

Key words: spatial point events; spatial attributes; spatial proximity relationships; spatial outlier patterns; multi-level constrained Delaunay triangulation

First author: SHI Yan, PhD, specializes in spatio-temporal data mining analysis. E-mail: csu_shiy@126.com

Corresponding author: DENG Min, PhD, professor. E-mail: dengmin208@tom.com

Foundation support: The National High Technology Research & Development Program of China (863 Program), No. 2013AA122301; Program for New Century Excellent Talents in University, No. NECT-10-0831; the Hunan Provincial Science Fund for Distinguished Young Scholars, No. 14JJ1007.