DOI:10.13203/j.whugis20140425　　文章编号:1671-8860(2016)06-0810-08

一种基于多层次专题属性约束的空间异常探测方法

杨学习[1]　石　岩[1]　邓　敏[1]　唐建波[1]

1　中南大学地球科学与信息物理学院地理信息系,湖南 长沙,410083

摘　要:针对现有空间异常探测方法在构建空间邻近域以及准确探测各类空间异常模式方面的局限性,本文提出了一种基于多层次专题属性约束的空间异常探测方法(spatial outlier detection by considering multi-level thematic attribute constraints,MTACSOD)。首先采用层次约束 Delaunay 三角网构建合理、稳定的空间邻近域;进而根据空间邻接实体间的专题属性距离,针对各空间连通子图施加全局和局部约束;最后通过一个异常模式识别指标提取各类空间异常模式。该方法不需要人为输入参数,通过模拟实验比较和实际应用验证了本文方法的优越性和有效性。

关键词:空间异常;Delaunay 三角网;专题属性;多层次约束

中图法分类号:P208　　**文献标志码**:A

空间异常探测是空间数据挖掘的重要手段之一,旨在从空间数据集中挖掘得到偏离整体或局部分布模式的少部分空间实体(这些异常实体可能隐含了地理实体和地理现象的突变过程或潜在的发展规律[1]),并在地质灾害监测、环境保护、公共卫生、极端气候事件探测等领域得到广泛应用。

异常的定义最初来源于 Hawkins 的研究工作,即"严重偏离其他对象的观测数据,以至于令人怀疑它是由不同机制产生的"[2]。Shekhar 等人将传统异常在空间数据中进行了扩展,将空间异常定义为"专题属性与其邻近空间实体显著不同,而在整体数据范围内差异可能不明显的空间实体"[3]。基于 Shekhar 对于空间异常的定义,相关学者发展了一系列空间异常探测方法,这些方法可大致分为基于距离的方法[3-5]、基于密度的方法[6,7]、基于聚类的方法[8,9]和基于模型的方法[10,11]。基于距离的方法采用专题属性值与空间邻近域内实体专题属性平均值(或中值)的差值来度量实体的异常度,进而通过统计测试的方法识别异常实体。这类方法仅适合发现全局空间异常,无法准确识别局部异常实体,代表性方法有文献[3]的 SLZ (Shekhar Lu Zhang)、文献[5]的空间异常探测(spatial outlier measure,SOM)等。基于密度的方法引入局部密度的概念,进而通过不同的密度估计来度量实体的局部异常度,局部异常度较大的空间实体被识别为异常。此类方法严重依赖空间邻近域的选择,且邻近域内存在的异常实体会对探测结果造成较大影响,因此缺乏一定的准确性和稳健性[6,7]。基于聚类的方法是把不隶属于任何簇的空间实体识别为异常,然而聚类的主要目的在于发现空间簇,因此探测异常的能力有限,且探测结果依赖于聚类算法的选择。基于模型的方法利用统计模型(例如高斯随机场)[10]、机器学习模型(例如自组织图)[11]等数学工具进行空间异常探测;然而此类方法并非数据驱动,而是需要首先满足模型的假设条件,例如假设空间数据服从某种分布,但在实际应用中又难以准确获得,可能导致探测结果偏离实际情况。

通过对现有方法进行分析总结可以发现,现有空间异常探测方法主要存在两方面问题:(1)定义空间邻近域时大多需要人为设定邻近域范围(如 k-邻域、ε 邻域等),导致探测结果严重依赖于所输入的参数;(2) 无法准确、稳健、全面地识别各类空间异常模式。如图 1 所示,空间异常模式

收稿日期:2014-10-15

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);湖南省自然科学杰出青年基金(14JJ1007);中南大学中央高校基本科研业务费专项资金(2015zzts256)。

第一作者:杨学习,博士生,主要从事空间/时空异常探测及其应用研究。studyang@sina.cn

通讯作者:邓敏,博士,教授。dengmin208@tom.com

可以分为全局异常点、全局异常区域、局部异常点和局部异常区域,现有方法大都仅能探测全局和局部异常点,而无法准确识别全局和局部异常区域,并且这些空间异常模式的存在使得现有方法出现大量误判和漏判现象。针对上述问题,本文借助层次约束 Delaunay 三角网构建合理、稳定的空间邻近域,进而发展了一种基于多层次专题属性约束的空间异常探测方法(spatial outlier detection by considering multi-level thematic attribute constraints,MTACSOD)。

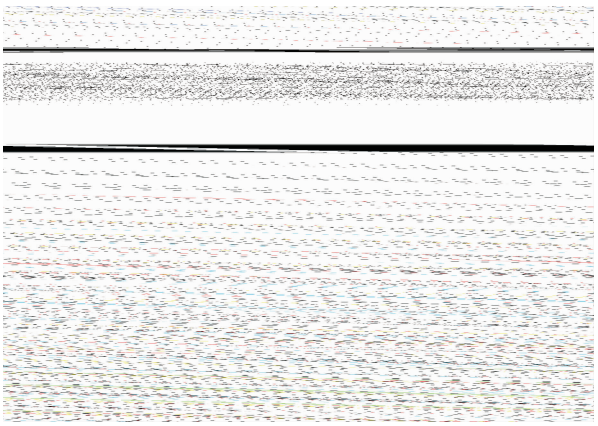

图 1　空间异常分类

Fig. 1　Classification of Spatial Outlier

1　基于多层次专题属性约束的空间异常探测方法

本文空间异常探测方法主要包括三个步骤:(1)针对原始数据建立的 Delaunay 三角网施加层次约束以构建合理、稳定的空间邻近域;(2)根据空间邻接实体间的专题属性距离,针对各空间连通子图施加全局和局部两层约束,从而获得内部实体间空间邻近、专题属性相似的连通子图;(3)利用一个异常识别指标进行统计判断,得到各类空间异常模式。

1.1　空间邻近域构建

Delaunay 三角网是一种边相互邻接、互不重叠的三角形集合,且满足最大最小角特性、外接圆特性和唯一性的三角剖分。Delaunay 三角网能够自然地反映空间实体间的邻近关系,是一种构建空间邻近域的有效工具[13,14],并已成功应用于空间聚类分析[12,13-15]。由于空间采样数据集大都呈不均匀的离散分布,因此建立的原始 Delaunay 三角网在空洞、边界以及不同密度的过渡区域存在明显误差,如图 2(a)所示,实体 A 与 B、C 与 D 之间的空间邻接关系是不准确的。通过分析空间实体构成的 Delaunay 三角网可以发现,三角网中发生错误邻接关系区域处的边明显较长。因此,本文借鉴文献[14,15]的研究策略,对原始 Delaunay 三角网施加层次约束,删除整体长边和局部长边,从而有效地消除原始 Delaunay 三角网构建空间邻近域在边界处和分布不均匀处存在的误差,从而获得合理、稳定的空间邻近域。首先给出几个重要定义。

定义 1　空间距离全局约束指标。给定空间数据集(spatial database,SDB),建立 Delaunay 三角网 DT,针对 Delaunay 三角网的所有边定义全局约束指标,表达为:

$$\mathrm{GS\_CI}(E_i) = \mathrm{mean}(\mathrm{DT}) + \frac{\mathrm{mean}(\mathrm{DT})}{L(E_i)} D(\mathrm{DT}) \tag{1}$$

式中,E_i 为原始 Delaunay 三角网中的任一边;$L(E_i)$为 E_i 的边长;mean(DT)为 Delaunay 三角网所有边的平均边长;D(DT)为 Delaunay 三角网所有边的边长标准差。通过调节系数数$\frac{\mathrm{mean}(\mathrm{DT})}{L(E_i)}$可有效删除原始 Delaunay 三角网中的全局长边,从而更新空间邻接关系,如图 2(b)所示。分析结果可以发现,虽然全局长边所导致的全局误差已有效消除,但在某些局部区域(如实体 P 与 Q、R 与 S 之间的邻接关系)仍存在误差,需要进一步处理以获得更加准确的空间邻近域。

定义 2　空间距离局部约束指标。基于边长全局约束得到的空间邻接关系,针对每个空间实体 P_i 相应的局部边定义局部约束指标,表达为:

$$\mathrm{LS\_CI}(E_{P_i}) = \mathrm{mean}^2(E_{P_i}) + k\frac{\sum_{j=1}^{n} D(E_{P_j})}{n} \tag{2}$$

式中,E_{P_i} 为实体 P_i 与其邻域实体构成的局部边;$P_j \in G_i$;G_i 为删除整体边长后得到的更新邻接关系构成的任一子图,如图 1(b)中 G_1 和 G_2;P_j 为子图 G_i 中任一实体;$\mathrm{mean}^2(E_{P_i})$为实体 P_i 的 2 阶邻域内所有边长的平均值;$D(E_{P_j})$为实体 P_j 与其邻域实体构成的局部边边长标准差;k 为调节系数,用于调节局部边长约束指标的敏感度,一般取值 1～3,默认为 2[13]。通过此策略删除所有长度大于局部边长约束指标的边,从而获得最终合理、精确的空间邻接关系,如图 2(c)所示。

1.2　专题属性差异层次约束

地理现象通常由全局(大尺度)趋势和局部(小尺度)随机误差两方面因素共同作用,全局趋势首先对地理现象产生作用,其次是局部随机误

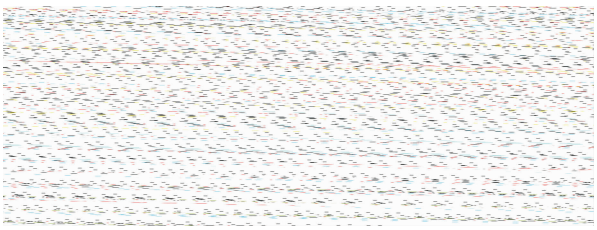

图2 空间邻近域构建

Fig. 2 Construction of Spatial Neighborhood

差影响[12]。空间异常为不符合普适性规律,表现出不同特性的地理实体或现象,通常空间异常的专题属性值呈现出明显的全局或局部差异。因此,在获取准确、合理的空间邻域基础上,采取一种与§1.1类似的层次约束策略继续对空间邻近实体间的专题属性距离进行约束分析。

定义3 空间邻域。对于SDB中任一空间实体P_i,通过层次约束Delaunay三角网的边与P_i直接相连的空间实体构成P_i的空间邻域NN(P_i)。

定义4 连通子图。将SDB中任一空间实体P_i与其邻域实体构成的邻接关系作为传递路径进行递归扩展,路径上的所有实体构成一个连通子图S_{G_i},如图2(c)中S_{G_1}、S_{G_2}和S_{G_3}。

定义5 专题属性距离。给定SDB中任意两个邻近空间实体P_i和P_j,对应的专题属性向量分别为($f(P_{i1})$, $f(P_{i2})$,…, $f(P_{id})$)、($f(P_{j1})$, $f(P_{j2})$,…, $f(P_{jd})$),那么P_i和P_j间的专题属性距离可表达为:

$$\mathrm{Dist}_{\mathrm{Attr}}(P_i,P_j)=\sqrt{\sum_{k=1}^{d}(f(P_{ik})-f(P_{jk}))^2} \tag{3}$$

其中,$f(P_{ik})$和$f(P_{jk})$分别为P_i和P_j的第k维专题属性值。

定义6 专题属性距离全局约束指标。给定任一连通子集S_{G_k},针对S_{G_k}内各实体与邻近实体间的专题属性距离定义全局约束指标为:

$$\mathrm{GA\_CI}(E_i)=\mathrm{mean}(S_{G_k})+\alpha_i D(S_{G_k})$$
$$\alpha_i=\frac{\mathrm{mean}(S_{G_k})}{L_{\mathrm{Attr}}(E_i)} \tag{4}$$

其中,E_i为S_{G_k}内的任一边;$L_{\mathrm{Attr}}(E_i)$为边E_i对应空间实体间的专题属性距离;mean(S_{G_k})和D(S_{G_k})分别为S_{G_k}内所有邻接实体间专题属性距离的平均值和标准差;α为适应系数。

连通子图S_{G_k}内所有邻接实体间的专题属性距离平均值和标准差可以很好地反映S_{G_k}内的专题属性全局差异分布,较大的专题属性距离对应的边可能隐藏着全局空间异常实体。为了无遗漏地识别此类不一致边,引入适应系数α,其中较大专题属性距离对应较小α。通过这种策略可从全局角度有效删除较大专题属性距离对应的边,如图3(a)所示,然而在某些局部区域存在的较大专题属性差异仍然蕴含着局部异常实体,如图4(b)中红色虚线框所示。为此,构造一个局部约束指标识别此类局部不一致区域。

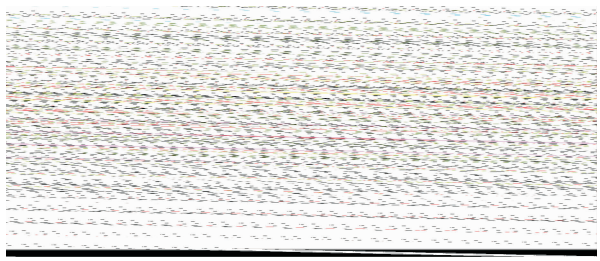

图3 专题属性距离层次约束

Fig. 3 Schematic of Multi-level Constraints on the Non-spatial Attribute Distance

定义7 专题属性距离局部约束指标。在施加专题属性距离全局约束后所得到的更新邻接关系基础上,针对各子图内的空间实体连接的局部边定义专题属性距离局部约束指标,可表达为:

$$\mathrm{LA\_CI}(E_i)=\mathrm{mean}(E)+\beta_i\frac{\sum_{j=1}^{n}D(P_j)}{n},$$
$$P_j\in S_{G_k},\beta_i=\frac{\mathrm{mean}(E)}{L_{\mathrm{Attr}}(E_i)} \tag{5}$$

其中,E_i为与空间实体P_j连接的局部边集合E的成员;$L_{\mathrm{Attr}}(E_i)$为边E_i对应的空间实体间的专题属性距离;mean(E)为局部边的专题属性距离平均值;$D(P_j)$为与空间实体P_j相关的局部边的专题属性距离的标准差;β为适应系数。

局部专题属性距离的平均值和标准差反映了各子图内专题属性的局部差异分布,通过附加适应系数β构成的局部约束指标可以很好地识别局部较大专题属性距离所对应的边,如图3(b)中红色虚线所示。通过删除此类局部不一致边可以获得实体间的最终邻接关系,即满足空间邻近、专题属性相似,如图4(c)所示,进而通过最终邻接关系获得的连通子图可进一步统计识别各类空间异常模式。

1.3 空间异常的自动识别

空间异常模式为严重偏离全局或局部分布的空间孤立实体或少量空间聚集簇。以上对原始Delaunay三角网所进行的空间距离和专题属性距离层次约束操作所获得的最终一系列连通子图,准确地描述了空间数据集的分布情况,并将正常聚

集模式和异常模式进行了分离。针对空间异常为“孤立实体”或“少量空间聚集簇”这一特性,下面构造一个异常识别指标探测各类空间异常模式。

定义 8　空间异常识别指标。记连通子图 S_{G_i} 包含空间实体的数目为 N_i,将数目相同的子图作为同一对象参与运算达到分类的目的,从而构成对象数据集 $N=\{\forall N_i: N_i \neq N_j, 1 \leqslant j < i \leqslant n\}$,其中 n 为连通子图数目。例如对原数据集 $N'=\{1,1,2,2,5,8,10\}$ 进行分类得到新数据集 $N=\{1,2,5,8,10\}$,进而空间异常识别指标可表达为:

$$\mathrm{SO\_I}(N_i) = \mathrm{mean}(N) - \gamma_i D(N)$$
$$\gamma_i = \frac{N_i}{\mathrm{mean}(N)} \tag{6}$$

其中,mean(N)和 $D(N)$分别为集合 N 中成员数值的平均值和标准差,描述了连通子图中实体数目的整体分布情况;γ 为适应系数,通过 γ 能够有效识别容量较小的子图。N_i 越小则 SO_I 越大,反之越小,通过此策略将包含空间实体的数目小于相应 SO_I 的连通子图识别为空间异常。例如对数据集 N,SO_I 分别设为[4.46, 3.72, 1.51, −0.70, −2.17],则 1 和 2 识别为异常。

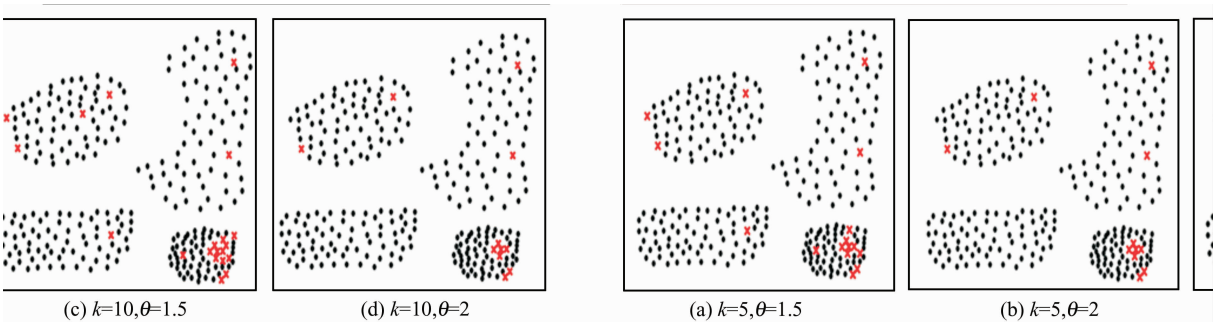

图 4　专题属性差异层次约束

Fig. 4　Multi-level Constraints on the Non-spatial Attribute Difference

1.4　算法描述和复杂度分析

基于以上定义,MTACSOD 算法可描述为:

输入:包含 N 个实体的空间数据集 SDB(包含 d 维专题属性);

输出:空间异常数据集;

(1) 对空间数据集 SDB 构建 Delaunay 三角网,获取粗糙空间邻近域;

(2) 根据空间距离全局约束指标及局部约束指标删除不一致边,获取精确的空间邻近关系以及连通子图 S_{G_k},构建空间邻域;

(3) 对专题属性进行归一化处理,根据连通子图中各实体间的空间邻近关系计算实体间的专题属性距离;

(4) 根据专题属性距离全局约束指标及局部约束指标对各连通子图内实体间的专题属性距离施加层次约束,从而得到实体间最终邻接关系和连通子图;

(5) 采用空间异常识别指标从各空间连通子图中探测空间异常模式,获取空间异常数据集。

通过对 MTACSOD 算法进行分析,其复杂度主要包括:(1) 建立 Delaunay 三角网的复杂度约为 $O(N\log N)$;(2) 对 Delaunay 三角网施加空间距离全局、局部约束,时间复杂度均与 N 呈线性关系;(3) 专题属性归一化处理时间复杂度为 $O(dN)$;(4) 对连通子图 S_{G_k} 施加专题属性距离全局和局部约束,时间复杂度均与 N 呈线性关系;(5) 根据空间邻域进行递归扩展的复杂度亦近似线性。因此,当 $d \ll N$ 时,本文方法总的时间复杂度约为 $O(N\log N)$,这种较高的效率能够有效地适应分析大量空间数据集。

2　实验分析及应用

本文设计两组实验以验证算法的有效性和实用性。实验一采用 ArcGIS10.0 模拟生成一组空间数据集 SDB,并对其空间分布及预设异常区域局部放大,如图 5 所示。将本文算法探测结果分别与 SLZ[3]、SLOM[6]、DDBSC(dual distance based spatial clustering)[9] 三种方法进行对比分析;实验二采用我国陆地区域 527 个气象站点 1998 年夏季(6～8 月)平均降水数据进行分析。

2.1　模拟实验

模拟数据集 SDB 设置了任意形状且密度不均匀的复杂空间簇分布,每个空间实体含有一维专题属性,并预设了位于不同空间簇中的异常点和异常簇的复合分布情况,异常实体共 26 个(异常点 9 个,异常簇 5 个)。实验结果中所有方法所得到的异常空间点用×表示;另外,本文方法用不同符号

标示所探测得到的各类异常空间聚集簇模式。

图6为本文方法对模拟数据SDB的探测结果，图7～9分别为SLZ、SLOM、DDBSC法探测结果。通过分析可以得出以下结论。

(1) MTACSOD算法能准确探测识别预设空间的9个异常点和5个异常空间聚集簇模式；

(2) SLZ方法在不同参数下只探测出一个异常簇，在参数$k=5$、$\theta=1.5$时准确探测出了9个异常点，其余参数均出现漏判现象。该方法只能探测全局异常，局部异常的探测结果不够理想，尤其是对内部异常聚集模式；

(3) SLOM方法在不同参数下均可探测出预设的9个异常点，但均存在误判现象，且没有探测出异常簇。该方法计算局部异常度所采用的波动参数β由对称分布状况来决定，在空间邻域较少或波动幅度较小时难以准确表现波动情况，易出现误判或漏判现象；

(4) DDBSC方法在不同参数下均探测出了9个异常点，但均有异常簇漏判现象，如图9中红色虚线框标识。该方法采用全局的阈值设置策略，难以适应专题属性差异分异特性，在探测局部异常簇方面存在不足。

2.2 实际应用

本文采用的实际数据集来源于中国气象科学数据共享中心气象资料室，包括1982～2011年中国陆地区域527个气象站点的月平均降水数据，气象站点的空间分布如图10(a)所示。实验采用1998年6～8月的降水均值数据。首先对该数据进行趋势性分析，如图10(b)所示，其中X轴表示东西方向、Y轴表示南北方向、Z轴表示降水值。从降水趋势图可以看出从北到南有降水逐渐上升的趋势，但在中南部地区有明显异常突变(红色虚线框所示)，这与1998年夏季长江中下游的极端降水气候事件高度吻合。

利用MTACSOD算法探测空间异常过程如图11所示，其中图11(a)和图11(b)分别表示构建空间邻近域和专题属性距离层次约束所得到的空间连通子图，图11(c)表示空间异常探测可视化结果。通过对结果进行分析可以发现：在1998年我国夏季降水分布中探测出20个异常单站点，27个异常空间小簇构成5个明显异常区域，O_1为松花江上游的嫩江区域，O_2、O_3对应于长江中下游区域，O_4对应于珠江下游区域，O_5对应于我国西南地区(分布着众多长江上游支流)。其中，区域O_1～O_4与1998年夏季特大洪灾事件最严重的区域高度吻合；O_5区域对应我国西南地区，该地区的降水由于受季风环流和复杂地理环境的影响，局部差异较明显，所探测的异常降水站点位于文献[16]中所划分的异常区域中。另外，探测得到的异常区域与降水空间分布图(图10(c))中深蓝色区域相吻合。其他三种方法需要通过不断调整参数来得到相对较佳的探测结果，如图(12)所示。分析发现SLZ和SLOM两种方法探测出一系列零散分布的异常站点，而对探测局部异常区域存在明显不足；而DDBSC法主要用于聚类，探测异常能力有限，尤其是探测异常区域。这也充分证明了本文算法在实际应用中的有效性和正确性。另外，将探测结果与相关领域知识有效结合，对深入研究我国降水的时空演变规律以及极端气候事件的准确预测具有重要的参考价值。

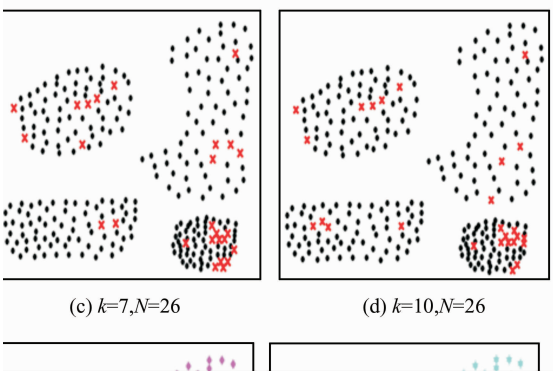

图5 模拟数据集

Fig. 5 Simulated Dataset

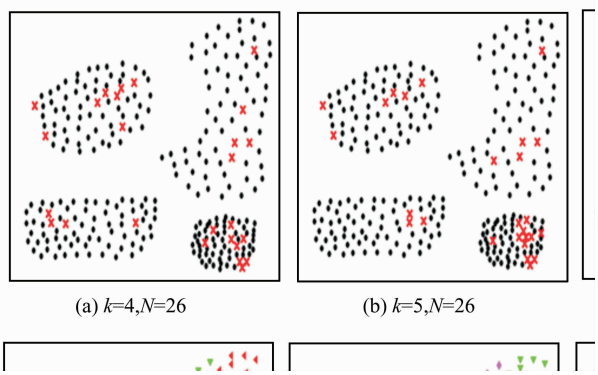

图6 本文方法探测结果

Fig. 6 Spatial Outlier Detecting Results of MTACSOD

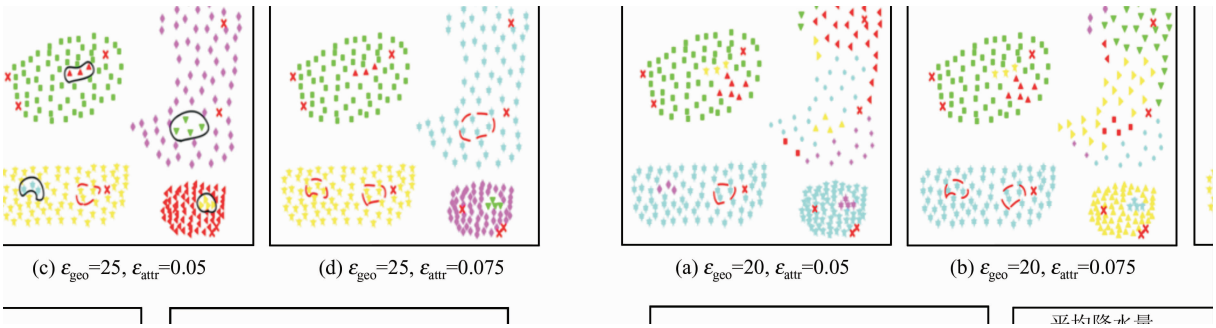

图 7　SLZ 方法异常探测结果(采用 k-NN 邻域)

Fig. 7　Spatial Outlier Detecting Results of SLZ

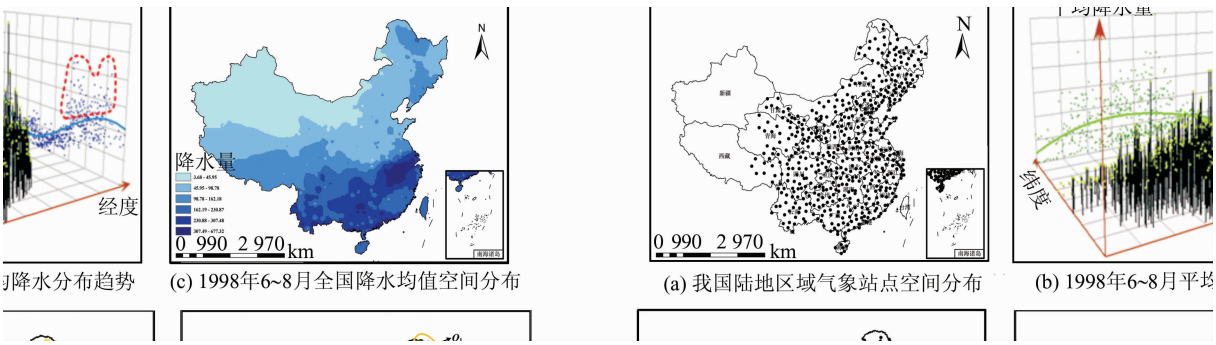

图 8　SLOM 方法异常探测结果(采用 k-NN 邻域)

Fig. 8　Spatial Outlier Detecting Results of SLOM

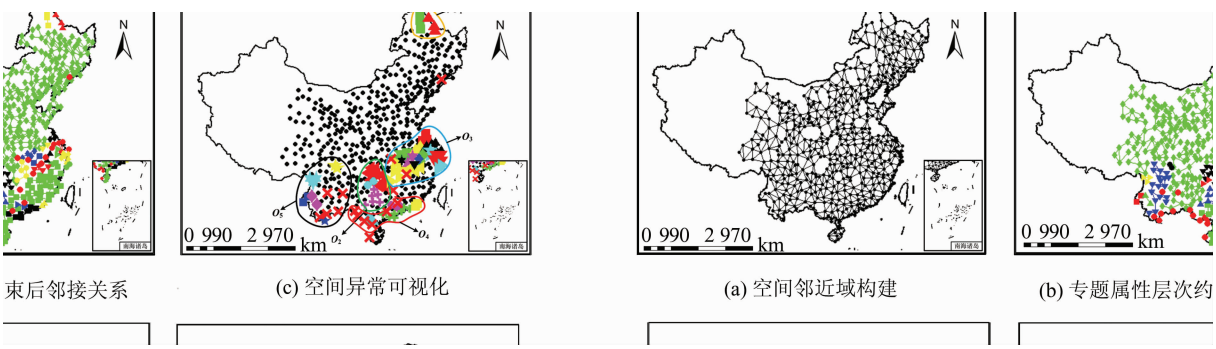

图 9　DDBSC 方法异常探测结果

Fig. 9　Spatial Outlier Detecting Results of DDBSC

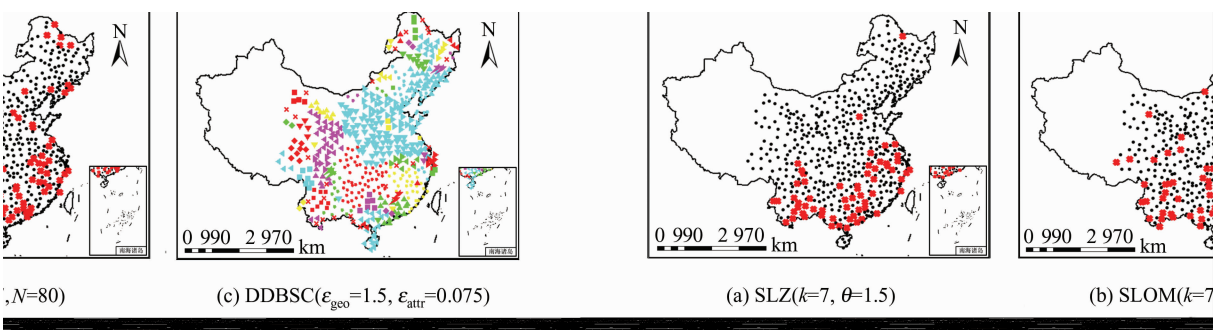

图 10　实际数据

Fig. 10　Real-World Data

图 11 空间异常探测结果

Fig. 11 Result of Spatial Outlier Detecting

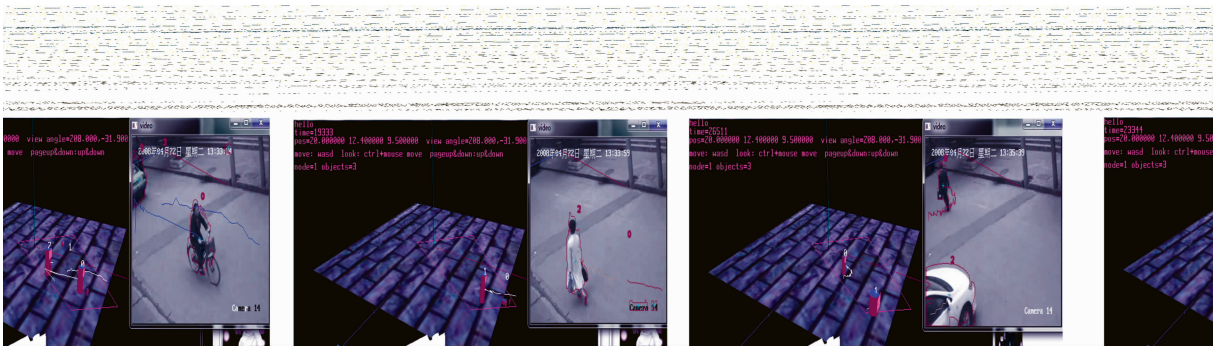

图 12 其他方法的空间异常探测结果

Fig. 12 Results of Spatial Outlier Detecting by Other Algorithms

3 结 语

针对现有空间异常探测方法多需要人为设定空间邻近域、无法准确探测各类空间异常模式等问题，本文提出了一种基于多层次专题属性约束的空间异常探测方法。通过模拟实验分析和实际应用得出：(1) 层次约束 Delaunay 三角网可自动获取合理、稳定的空间邻近域；(2) 能够准确探测任意形状及密度不均匀复杂空间数据集中的全局、局部异常点和异常区域；(3) 不需要输入任何参数，具有很好的自适应能力和实用性。

进一步研究工作主要集中在以下方面：对空间异常模式的有效性进行定量评估和深入解释；研究时空异常探测方法，并深入分析时空异常模式在多尺度效应中的变化规律。

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

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

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

[5] Li G Q, Deng M, Zhu J J, et al. Spatial Outlier Detection Considering Distances among Their Neighbors[J]. *Journal of Remote Sensing*, 2009, 13(2): 197-202

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

[7] Xue Anrong, Ju Shiguang. Outlier Mining Based on Spatial Constraint[J]. *Computer Science*, 2007, 34(6): 207-209+230(薛安荣，鞠时光. 基于空间约束的离群点挖掘[J]. 计算机科学，2007，34(6)：207-209+230)

[8] Deng M, Liu Q L, Li G Q. Spatial Outlier Detection Method Based on Spatial Clustering[J]. *Journal of Remote Sensing*, 2010, 14(5): 944-958

[9] Li Guangqiang, Deng Min, Cheng Tao, et al. A Dual Distance based Spatial Clustering Method[J].

Acta Geodaetica et Cartographica Sinica, 2008, 37(4): 482-488(李光强,邓敏,程涛,等. 一种基于双重距离的空间聚类方法[J]. 测绘学报,2008,37(4): 482-488)

[10] Chen F, Lu C T, Boedihardjo A. GLS-SOD: A Generalized Local Statistical Approach for Spatial Outlier Detection[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2010

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

[12] Deng Min, Liu Qiliang, Li Guangqiang, et al. Spatial Clustering Analysis and Application[M]. Beijing: Science Press, 2011(邓敏,刘启亮,李光强,等. 空间聚类分析及应用[M]. 北京:科学出版社, 2011)

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

[14] 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)

[15] Liu Qiliang, Deng Min, Shi Yan, et al. A Novel Spatial Clustering Method Based on Multi-constraints[J]. *Acta Geodaetica et Cartographica Sinica*, 2011, 40(4): 509-516(刘启亮,邓敏,石岩,等. 一种基于多约束的空间聚类方法[J]. 测绘学报, 2011,40(4): 509-516)

[16] Liu Xiaoran, Li Guoping, Fan Guangzhou, et al. Spatial and Temporal Characteristics of Precipitation Resource in Southwest China during 1961-2000[J]. *Journal of Natural Resources*, 2007, 22(5): 783-792(刘晓冉,李国平,范广洲,等. 我国西南地区 1961-2000 年降水资源变化的时空特征[J]. 自然资源学报,2007,22(5): 783-792)

A New Method of Spatial Outlier Detection by Considering Multi-level Thematic Attribute Constraints

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

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

Abstract: Spatial outlier detectionis an important approach in spatial data mining and knowledge discovery. Spatial outliers are entities whose non-spatial attributes are significantly different from the value of other entities in their spatial neighborhoods. The current methods have limitations in capturing spatial neighborhoods and detecting small abnormal clusters. In order to solve this problem, we develops a new method of spatial outlier detection that considers thematic attributes, named MTACSOD. Firstly, a constrained Delaunay triangulation is used to construct reasonable and stable spatial proximity relationship. Then, for the thematic attribute distance between adjacent spatial entities, global and local constraints are imposed consecutively to refine spatial adjacency. Finally, a spatial outlier identification index is proposed to detect spatial outliers. Both simulated and real-life datasets are used to illustrate the advance and practicability of the MTACSOD proposed in this paper.

Key words: spatial outlier; Delaunay triangulation; thematic attribute; multi-level constraints

First author: YANG Xuexi, PhD candidate, specializes in spatio-temporal data mining and analysis. E-mail: studyang@sina.cn

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

Foundation support: The National High-Tech R&D Program of China (863), No. 2013AA122301; the Program for New Century Excellent Talents in University, No. NECT-10-0831; the Hunan Provincial Science Fund for Distinguished Young Scholars, No. 14JJ1007; the Fundamental Research Funds for the Central Universities, No. 2015zzts256.