DOI:10.13203/j.whugis20150237　　文章编号:1671-8860(2018)03-0364-08

一种基于场论的空间异常探测方法

杨学习[1]　徐　枫[1]　石　岩[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:从空间数据场的角度,借鉴高斯势函数发展了一种新的空间异常度度量指标。进而,提出了一种基于场论的空间异常探测方法。该方法通过空间聚类获得局部相关性较强的空间簇,并构建合理、稳定的空间邻近域。在此基础上,采用专题属性变化梯度修复策略减弱空间邻近域中潜在异常的影响,并利用空间异常度度量指标计算实体的异常度,从而探测空间异常。实验结果及实例证明了此方法的正确性。

关键词:空间异常探测;空间聚类;场论;空间异常度

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领域一项重要研究内容[1-3],旨在从海量空间数据中挖掘小部分偏离普遍模式的空间实体。这些异常实体通常蕴含着难以预知的知识,可能代表着地理现象或地理过程的特殊发展规律。近年来,空间异常探测受到学者们的广泛关注,并已应用于地质灾害监测、环境监测与保护、公共卫生、遥感图像数据处理等领域,具有重要的研究价值。

异常(亦称离群点)的定义最初来源于文献[4]的研究工作,描述为"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"。文献[5]将传统异常在空间上进行了有效扩展,指出空间异常是指专题属性与其空间邻近域内其他参考实体的专题属性显著不同,而在整体数据范围内差异可能不明显的空间实体。空间异常通常根据空间属性(即位置)确定空间邻近关系,进而借助专题属性确定异常程度。现有的空间异常探测方法可分为:(1)基于统计的方法[4];(2)基于图形的方法[6];(3)基于距离的方法[5,7-8];(4)基于密度的方法[9-11];(5)基于聚类的方法[12-13];(6)基于模型的方法[14-15]。基于统计的方法要求数据服从一定的分布,适用性不强。基于图形的方法主要是采用可视化的方式(如变量云和散点图等)来显现空间异常点,利用人眼进行主观性的识别使这类方法已很少使用。基于距离的方法采用专题属性值与其空间邻域的专题属性的均值(或中值)的差值来度量实体的异常程度,继而通过统计测试的方法识别异常,这类方法采用全局策略,仅适合探测全局异常,不适合探测局部异常实体。基于密度的方法采用不同的密度估计策略度量实体的局部异常度,异常度较大的实体识别为空间异常。这类方法依赖于空间邻近域的选择,缺乏一定的准确性和稳定性。基于聚类的方法旨在发现空间簇,探测空间异常的能力有限,且探测结果依赖于聚类算法的选择。基于模型的方法采用统计模型、机器学习模型等数学工具进行空间异常探测。这类方法需满足模型的假设条件,如数据分布,这在实际运用中很难准确获得,可能使得探测结果偏离实际情况。

对现有方法进行分析归纳可以发现,当前方法存在的主要问题是:(1)现有方法多认为所有空间实体之间具有等同的相关性,而实际上空间实体间具有局部的相关性,在全局更呈现异质性。如在一个大区域内,经常出现专题属性差异较明显的几个子区域,在探测时若不加以区分,一方面可能导致空间邻域内实体不满足相关性假设,导致异常度量的偏差;另一方面可能造成由于局部出现较多异常实体,使得一些在整体上不明显而局部差异明显的空间异常难以被发现。如图 1 所示,采用一个模拟数据集来分析异常度量的差异[9],其中Ⅰ、Ⅱ、Ⅲ各自区域内的实体间具有较强的相关性,而不同区域内实体间的相关性较弱,

收稿日期:2016-01-25

项目资助:国家高技术研究发展计划(863 计划)(2013AA122301);湖南省自然科学杰出青年基金(14JJ1007);国家自然科学基金(41471385);中南大学中央高校基本科研业务费专项资金(2016zzts085)。

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

甚至是独立的。根据空间局部异常度(spatial local outlier measure,SLOM)法[9]共探测出 5 个异常,如图 1(b)中浅绿色标识。然而,在区域Ⅱ、Ⅲ中 3 个浅蓝色背景标识 SLOM 值明显偏大,很可能是空间异常。但由于Ⅰ区出现较多明显的异常点,导致Ⅱ、Ⅲ区域内的异常被忽略。因此,需要加以区分。(2)现有方法多采用空间属性确定空间邻近关系、专题属性度量异常度,缺乏一种耦合两类属性的异常度度量的指标。针对上述问题,本文受物理学中场论思想的启发,从数据场的角度对空间异常探测进行解释,采用似高斯势函数度量空间异常度,发展了一种基于场论的空间异常探测方法(field-theory based spatial outlier detecting method, FTSOD)。

	0	1	2	3	4	5	6	7	8	9
0	-16	-9	-16	4	8	25	-2	20	-11	9
1	-3	-1	9	-12	1	1	-1	-2	-4	-2
2	14	1	11	2	-13	15	4	3	11	19
3	-10	16	-11	-2	-10	-11	-17	4	8	-15
4	-5	20	-11	4	-5	8	6	6	-2	-1
5	15	10	-9	7	12	-9	-18	16	8	-6
6	0	0	0	0	-21	-5	12	-15	-5	11
7	0	0	0	0	5	6	1	1	-9	3
8	0	8	0	0	-9	8	-1	-2	9	5
9	0	0	0	0	19	-1	-2	-7	-3	-12

	0	1	2	3	4	5	6	7	8	9
0	0.09	0.06	0.05	0.03	0.08	0.43	0.06	0.21	0.10	0.11
1	0.03	0.02	0.03	0.03	0.06	0.08	0.05	0.01	0.02	0.04
2	0.05	0.02	0.08	0.02	0.11	0.25	0.05	0.03	0.02	0.14
3	0.17	0.03	0.03	0.05	0.11	0.07	0.15	0.02	0.05	0.17
4	0.06	0.05	0.03	0.02	0.06	0.10	0.02	0.05	0.10	0.11
5	0.04	0.09	0.09	0.02	0.03	0.03	0.05	0.04	0.07	0.03
6	0.02	0.02	0.02	0.01	0.06	0.08	0.03	0.04	0.02	0.05
7	0.00	0.00	0.00	0.02	0.02	0.06	0.01	0.01	0.03	0.03
8	0.00	0.17	0.00	0.02	0.03	0.02	0.01	0.02	0.09	0.03
9	0.00	0.00	0.00	0.02	0.10	0.02	0.02	0.02	0.05	0.12

图 1　模拟空间数据集[9]

Fig.1　Simulated Spatial Dataset[9]

1　基于场论的空间异常探测方法

本文提出的空间异常探测方法主要分为 4 个步骤:(1)采用空间聚类技术获取空间簇,针对各空间簇生成 Delaunay 三角网,进而约束获得空间邻近域;(2)采用专题属性变化梯度修复策略处理空间邻域中潜在异常的影响;(3)引入空间数据场概念,采用似高斯势函数度量空间异常度;(4)针对各空间簇分别统计识别空间异常,并进行评价分析。

1.1　空间邻近域构建

本文采用一种基于多约束的自适应空间聚类方法[16](adaptive spatial clustering algorithm based on Delaunay triangulation, ASCDT)进行空间聚类。该方法通过施加不同层次、不同类型的约束可以适应空间数据分布不均匀、空间簇形态各异、位置邻近等复杂情况下的聚类,且输入参数较少,自适应能力强。通过聚类,将所有空间实体划分为若干空间相关性较强的空间簇。针对各空间簇,借助 Delaunay 三角网来构建空间邻近关系。Delaunay 三角网是一种满足最大最小角特性、外接圆特性和唯一性的三角剖分,能自然地反映空间实体间的邻接关系。但从图 2(b)可以发现,原始 Delaunay 三角网在边界和空洞处的边长明显偏长,如实体 A 与 B、C 与 D 空间邻近是不合理的。文献[17]通过实验证明,不合理的边可通过删除超过平均边长一定倍数的边来有效移除。本文亦采用一种稳健的平均边长来处理不合理的边。

定义 1　稳健的平均边长:给定各空间簇 S,S 中所有实体构成 Delaunay 三角网的 N 条边,构成边长集合 E,E 中所有边长按升序排列,序列中位于上四分位数和下四分位数之间所有边长的均值称为稳健的平均边长,记为 $R_A(E)$,表示为:

$$R_A(E)=\frac{\sum E_i}{n},\quad Q_1\leqslant E_i\leqslant Q_3 \tag{1}$$

式中,Q_1 为上四分位数;Q_3 为下四分位数;n 表示上下四分位数之间所有边的数量。

定义 2　不合理的边:边长集合 E 中,与稳健平均边长相比明显偏大的边定义为不合理的边,所有不合理的边构成不合理的边集合 E_I,表示为:

$$E_I=\{E_i\,|\,E_i>\alpha\times R_A(E)\},\quad E_i\in E \tag{2}$$

式中,α 为调节因子,用于调整不合理边的判断阈值,其取值范围可通过启发式策略确定[17]。本文

通过大量实验发现,α 取值[2, 3]时较为合适。

在空间簇 Delaunay 三角网中打断不合理的边后,可以获得每个实体的空间邻近实体。如图2(c)所示,经打断操作后($\alpha=2$)空洞和边界处的不合理边被有效移除,据此建立的实体间邻近关系更为合理、稳定。没有隶属于任何簇的实体识别为空间位置孤立点,不参与接下来的检测。

定义3 空间邻域:对于空间簇中任一空间实体 P_i,与打断不合理的边后的 Delaunay 三角网的边直接相连的空间实体构成 P_i 的空间邻域 $N_S(P_i)$,如图2(c)中空间簇 S_1 实体 P 的空间邻域为 $N_S(P)=\{P_1,P_2,P_3,P_4,P_5,P_6\}$。

图2 空间邻近域构建

Fig.2 Construction of Spatial Neighborhood

1.2 专题属性变化梯度修复

在度量空间异常度时,首先需要有效消除空间邻近实体中异常值的影响,借鉴文献[12]的研究策略,本文提出了一种专题属性变化梯度修复策略消除异常值。

定义4 专题属性变化梯度:给定空间实体 P,其空间邻域为 $N_S(P)=\{X_1, X_2\cdots X_n\}$,$f(X_i)$ 表示实体 X_i 的专题属性值,专题属性变化梯度定义为空间实体 P 与其某一邻近域实体 X_i 的专题属性差值的绝对值与二者间欧氏距离的比值,记为 $G(P, X_i)$:

$$G(P,X_i)=\frac{|f(P)-f(X_i)|}{D(P,X_i)},\forall X_i\in N_S(P) \tag{3}$$

式中,$D(P,X_i)$ 表示 P 与 X_i 的欧氏距离。

针对任一空间实体 P 进行邻域异常值修复的步骤为:

1)令 $f(P)=0$,分别计算实体 P 与其空间邻域实体 X_i 的专题属性变化梯度 $G(P, X_i)$,按升序排列获取序列 $G(P)$,并计算 $G(P)$ 的中位数 $M(P)$。

2)针对任一空间邻域 X_i,计算专题属性变化梯度偏离 $G_D(X_i)=|G(P,X_i)-M(P)|$,按升序排列获取序列 $G_D(P)$。

3)将邻域实体按专题属性变化梯度偏离划分为大、中、小3个等级,处于最大等级的[$(n+1)/3$]个实体组成待修复集合 $R(P)$。进而,采用专题属性变化梯度序列的中位数 $M(P)$ 进行修复:

$$f_R(X_i)=M(P)\times D(P,X_i),\forall X_i\in R(P) \tag{4}$$

式中,$f_R(X_i)$ 表示专题属性修复值。

异常值修复策略旨在消除空间邻域内潜在异常值对度量异常度的影响,保证空间邻域内实体间专题属性的局部平稳性假设。且这种修复是暂时的,仅在空间异常度量的过程中进行,并不改变实体的固有专题属性值。

1.3 基于场论的空间异常度量

场的概念最早由英国物理学家法拉第于1837年提出,用于描述物质粒子间的非接触相互作用。随着场论思想的发展,人们将其抽象为一个数学概念,用于描述某个物理量或数学函数在空间内的分布规律,分为矢量场和标量场。势场是一种重要的标量场,势函数是位置或距离的函数,可以叠加。数据场是指数据通过辐射将其数据能量从样本空间辐射到整个母体空间,接受数据能量并被数据辐射所覆盖的空间[18-19]。如图3所示,是一个借助 Voronoi 图和 Delaunay 三角网得到的空间数据场[20]。数据场已应用于数据挖掘[20-22]和图像分割[23]等领域。实验研究表明,短程场作用更有利于揭示数据分布的凝聚特性,因此,数据场的作用范围必须在有限范围内迅速衰减[18,22]。高斯函数则可以满足迅速衰减的特性,在充分考虑空间数据自身的特点以及空间数据辐射的特性,本文通过约束 Delaunay 三角网获取空间邻域的基础上,采用似高斯势函数来定义空间

实体的空间异常度。

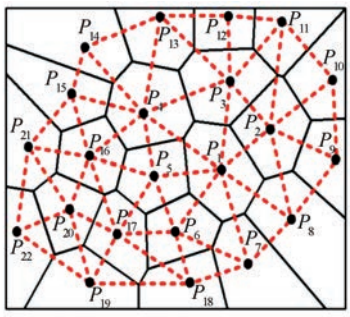


图 3　空间数据场示意图

Fig.3　Schematic of Spatial Data Field

定义 5　空间异常度:给定空间实体 P,其空间异常度 $M_{S_O}(P)$表达为:

$$M_{S\_O}(P)=\frac{\sum_{i=1}^{|N_S(P)|} D_{\mathrm{Attr}}(P,X_i)\mathrm{e}^{-\frac{D^2_{\mathrm{Geo}}(P,X_i)}{2\sigma^2}}}{|N_S(P)|},\quad \forall X_i\in N_S(P) \tag{5}$$

式中,$N_S(P)$为实体 P 的空间邻域;$|N_S(P)|$为空间邻域数目;$D_{\mathrm{Geo}}(P,X_i)$表示实体 P 与 X_i 之间的欧氏距离;$D_{\mathrm{Attr}}(P,X_i)$表示实体 P 与 X_i 之间的专题属性距离,本文采用归一化的闵氏距离;σ 为辐射因子。

1.4　空间异常识别

针对空间聚类获取的若干空间相关性较强的空间簇,采用统计判别法进行识别,分别探测异常。

定义 6　空间异常:给定一空间簇数据集 S,其空间异常集 S_{outlier}为:

$$S_{\mathrm{outlier}}=\{X_i\,|\,M_{S\_O}(X_i)-\mu>k\sigma\},\quad \forall X_i\in S \tag{6}$$

式中,μ 为异常度平均值;σ 为标准差;k 为调节因子。通常 k 取值 2 或 3,文献[24]通过实验结果表明,当 $k=1.645$ 时判断异常的结果较合理可靠。因此,本文中 k 取值 1.645。

所有簇的异常集 S_{outlier}的集合构成空间数据集的空间异常集。

空间聚类后,各空间簇异常探测的样本减少。直接采用式(6)可能降低异常探测的稳定性。因此,针对小样本数据(样本数小于 30)采用稳健统计量中位数 median 和中位数绝对偏差(median absolute deviation, MAD)[25]替代 μ 和 σ。其中,$\mu=\mathrm{median}(M_{S\_O})$,$\sigma=D_M(M_{S\_O})$。

$$D_M(M_{S\_O})=\mathrm{median}\{|M_{S\_O}(X_1)-\mathrm{median}(M_{S\_O})|\cdots|M_{S\_O}(X_n)-\mathrm{median}(M_{S\_O})|\} \tag{7}$$

1.5　算法描述和复杂度分析

基于以上定义,基于场论的空间异常探测算法可描述为如下所示。

输入:包含 N 个实体的空间数据集

输出:空间异常数据集

1)对空间数据集进行空间聚类,获取若干空间相关性较强的空间簇;

2)针对各空间簇分别生成 Delaunay 三角网,采用稳健平均边长移除不合理边,获取空间邻域;

3)采用专题属性变化梯度修复策略消除空间邻域内潜在异常值的影响;

4)采用异常度量指标计算各空间实体的空间异常度;

5)针对各空间簇分别采用统计判别准则识别异常,获取空间异常集。

对于包含 N 个实体的空间数据库,非空间属性维数为 d,其复杂度主要包括:ASCDT 聚类算法的复杂度为 $O(N\lg N)$;构建 Delaunay 三角网、打断不合理边并获取空间邻域的复杂度约为 $O(N\lg N)+O(6N)+O(6N)$;专题属性归一化的复杂度为 $O(dN)$,异常度计算的复杂度为 $O(N\lg N)+O(6dN)$,异常判别的复杂度为 $O(N\lg N)$。于是,该算法的复杂度为:$O(N\lg N)+O(N\lg N)+O(6N)+O(6N)+O(dN)+O(N\lg N)+O(6dN)+O(N\lg N)$。当 $d\ll N$ 时,算法复杂度近似为 $O(N\lg N)$。

2　实验与分析

本文设计了两组实验来验证所提出的异常探测算法的可行性。实验 1 采用模拟数据,模拟数据是由文献[13]中 3 组经典数据库的一部分组成,进一步增加了一维专题属性,专题属性服从正态分布,对其空间分布及部分预设异常局部放大,如图 4 所示。实验 2 采用华南某市土壤重金属 Cr 浓度监测数据,进行实际应用分析。两组实验结果均与经典的 SLOM 算法[9]进行比较,SLOM 算法需要两个输入参数,即空间邻域数目 k 和空间异常数目 M。实验环境为 acer Aspire V5 笔记本电脑(处理器 A10-5757M APU,内存 4 GB,系统 Windows 8.1 中文版 64-bit)和 Matlab R2011b 编程环境。

2.1　模拟算例分析与比较

本文采用的模拟数据集具有不同形状、不同密度的空间簇分布,且专题属性值服从正态分布。按照 Shekhar-Outlier[5]定义在空间簇中预设了 25 个空间异常点,包含全局和局部异常点。实验结果中空间异常点用“X”表示。针对模拟数据

集，本文算法调节因子 α 取值2，辐射因子 σ 取值0.05；针对SLOM算法，采用 k-邻域搜索空间邻域，其中 k 取值为7，空间异常数目 M 取值25，即模拟数据预设空间异常点数目。

图6、7分别给出了本文方法和SLOM算法的异常度空间分布及空间异常探测结果。对比探测结果可以发现：(1)本文提出的空间异常探测算法顾及了实体间的局部相关性以及邻近域内空间异常点的影响，能够更全面地发现局部的异常现象，正确识别了预设的25个空间异常点。(2)SLOM算法在度量局部异常度时不能有效消除邻域内潜在异常的影响，使得异常度量不准确，且从全局去识别异常，没有顾及到实体的空间分异特性。其探测正确率为60%，误判率为40%，漏检率为40%。(3)SLOM算法的复杂度为 $O(Nk\lg N+kdN)$[9]，而本文算法需先聚类，再约束Delaunay三角网获取空间邻近域，进而度量空间异常并统计识别，算法复杂度高，且执行效率低于SLOM算法。而SLOM算法在确定异常数目时需要较多先验知识，且邻域参数的选择对异常度量影响明显，故本文算法在正确性及易用性方面略优于SLOM算法。

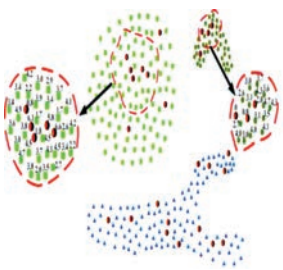


图4 模拟数据集SDB空间分布

Fig.4 Simulated Dataset SDB

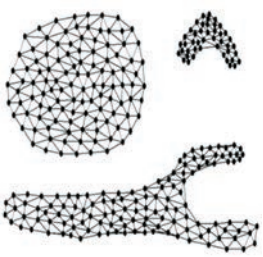
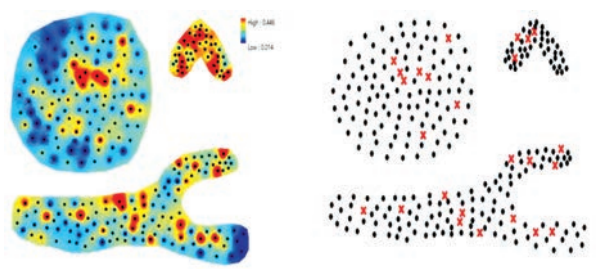


图5 模拟数据集SDB的空间邻近域

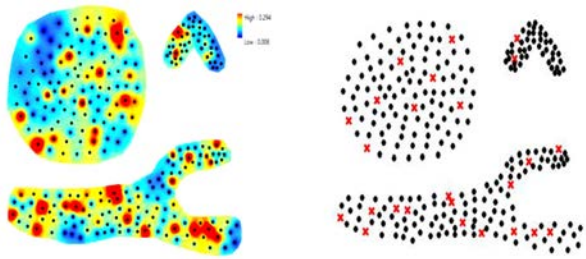
Fig.5 Spatial Neighborhood of SDB



(a) FTSOD算法空间异常度分布图 (b) 异常点空间分布σ=0.05

图6 本文算法的异常探测结果

Fig.6 Spatial Outlier Detection Results Obtained by the Proposed Algorithm



(a) SLOM算法空间异常度分布图 (b) 异常点空间分布k=7,M=25

图7 SLOM算法异常探测结果

Fig.7 Spatial Outlier Detection Results Obtained by SLOM Algorithm

2.2 实例分析与比较

采用中国华南某市环保数据中土壤重金属铬(Cr)浓度监测数据验证本文算法的可行性。采样点空间分布如图8(a)所示，采样点共103个。首先进行空间聚类。图8(b)为ASCDT算法聚类结果，所有空间实体均加入空间簇中，共获得11个空间簇。针对各空间簇分别生成Delaunay三角网并处理不合理的边，α 取值2.5，空间邻域结果图如图8(c)所示。当空间簇个数小于或等于6个时，簇中所有空间实体互为空间邻域。进而，采用本文提出的基于场论的方法在各个空间簇中探测空间异常。

FTSOD算法在11个空间簇中共探测出16个空间异常采样点，其空间分布如图9(a)所示。SLOM算法空间邻域数目 k 取值为5，异常数目 M 取值16，探测结果图9(b)所示。比较分析可以发现：(1)SLOM算法虽能顾及局部特性，但仅能发现整体上异常度较大的实体，对于局部异常现象，如簇9中的86号采样点(见表1)，在整体上异常程度不明显，但其专题属性严重偏离空间邻域的其他实体，表现为局部异常现象；(2)FTSOD算法在异常识别时所采用的小样本识别策略十分稳健(簇2、7中包含4个实体，簇8、9中包含5个实体)。

表1 空间簇9中实体专题属性值

Tab.1 Thematic Attribute of Objects in Cluster 9

空间簇	采样点编号	土壤Cr实测值/(mg·kg^{-1})
9	82	20.63
	83	12.05
	84	28.35
	85	24.58
	86	3.9

通过对数据分布及来源进行分析，空间异常产生的原因可从土地使用类型(见表2)、高程差

异和污染源等 3 方面进行分析。本文把重金属污染企业(如电镀厂)视为主要污染源,空间分布如图 9(a)三角形所示。以异常采样站点为中心,建立半径 L 为 5 km 的缓冲区。

综合以上信息,对采样点空间异常产生的原因进行分析,可以发现:(1)从表 2 可以发现绝大多数空间异常发生在菜地和水稻土,这可能与化肥、农药的过度、不合理使用有密切关系;(2)从图 10 可以看出,1、22、40、42、70、71、86、97、98 号采样点与其邻近域实体间的高程差异明显,进而影响土壤重金属含量的分布,这可能是产生空间异常的主要因素;(3)从图 11 可以发现 1、31、35、42、51、55、71、102 号采样点与污染源联系比较紧密,极有可能是受附近污染企业的影响。

(a) 采样点空间分布　(b) 空间聚类结果　(c) 空间邻近域构建(α=2.5)

图 8　实际数据

Fig.8　Real-World Data

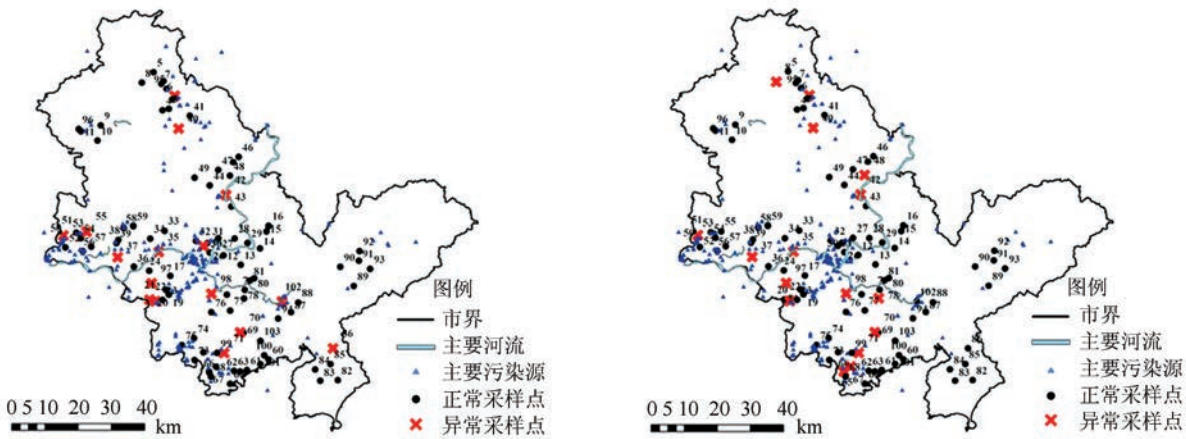


图 9　空间异常探测结果

Fig.9　Results of Spatial Outlier Detection

表 2　空间异常采样点土地使用类型

Tab.2　Land-Use Types of Spatial Outliers

土地使用类型	采样点编号	土地使用类型	采样点编号	土地使用类型	采样点编号
	21		70		1
	22		71	水稻土	35
菜地	31	菜地	86		42
	40		97	香蕉地	37
	51		98,102	荔枝地	55

3　结　语

空间异常探测对于揭示地理实体或地理现象的潜在发展规律具有重要价值。针对现有空间异常探测方法大多没有顾及空间实体之间的局部相关、整体分异的特性,且缺乏一种耦合专题属性和空间属性度量空间异常度的指标,本文首先采用空间聚类技术获得空间相关性较强的空间簇,借助 Delaunay 三角网并打断不合理的边以获取合理、稳定的空间邻域,进而采用似高斯函数度量空间异常度,并在每个空间簇中探测空间异常。本文方法具有两方面的优势:(1)顾及了空间实体间的局部相关性,能更全面地探测局部空间异常;(2)空间异常度量指标耦合了专题属性和空间属性,使得探测的空间异常更具有物理意义。

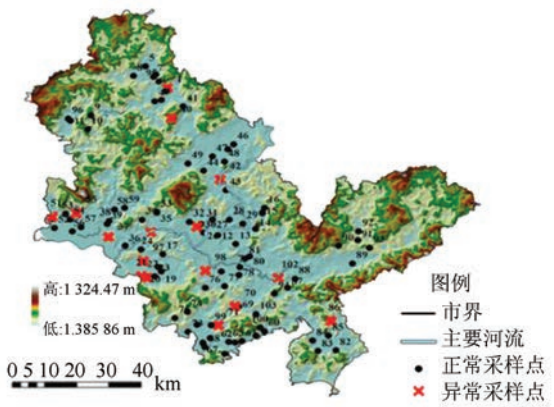


图10 空间异常采样点与高程关系图

Fig.10 Relationship Between Spatial Outliers and Elevation

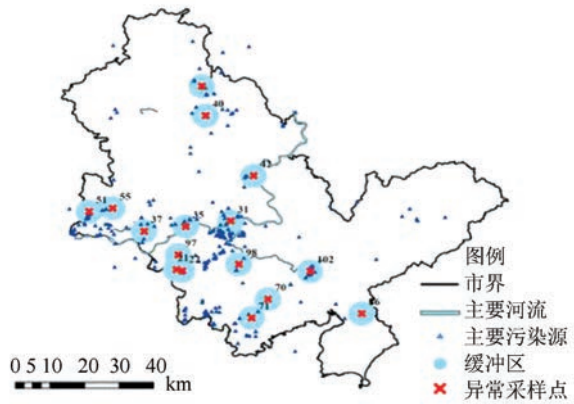


图11 异常采样站点缓冲区 $L=5$ km

Fig.11 Buffer of Spatial Outliers

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] Li Deren, Wang Shuliang, Li Deyi. Spatial Data Mining Theories and Applications[M]. Beijing: Science Press, 2013(李德仁, 王树良, 李德毅. 空间数据挖掘理论及应用 [M]. 北京:科学出版社, 2013)

[3] Liu Dayou, Chen Huiling, Qi Hong, et al. Advance in Spatiaotemporal Data Mining [J]. *Journal of Computer Research and Development*, 2013, 50(2): 225-239 (刘大有, 陈慧灵, 齐红, 等. 时空数据挖掘研究进展[J]. 计算机研究与发展, 2013, 50(2): 225-239)

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

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

[6] Haslett J, Brandley R, Craig P, et al. Dynamic Graphics for Exploring Spatial Data with Application to Locating Global and Local Anomalies [J]. *The American Statistician*, 1991, 45(3): 234-242

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

[8] 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)

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

[10] Schubert E, Zimek A, Kriegel H P. Local Outlier Detection Reconsidered: A Generalized View on Locality with Applications to Spatial, Video, and Network Outlier Detection [J]. *Data Mining and Knowledge Discovery*, 2014, 28(1): 190-237

[11] 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)

[12] Deng Min, Liu Qiliang, Li Guangqiang. Spatial Outlier Detection Method Based on Spatial Clustering[J]. *Journal of Remote Sensing*, 2010, 14(5): 944-958

[13] Ester M, Kriegel H P, Sander J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]. The 2nd International Conference on Knowledge Discovery and Data Mining, Portland, O R, 1996

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

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

[16] 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

[17] Kolingerova I, Zalik B. Reconstructing Domain

Boundaries within A Given Set of Points Using Delaunay Triangulation[J]. *Computers & Geosciences*, 2006, 32(9): 1 310-1 319

[18] Li Deyi, Du Yi. Artificial Intelligence with Uncertainty (Second Edition) [M]. BeiJing: National Defense Press, 2014 (李德毅, 杜鹢. 不确定性人工智能 (第 2 版) [M]. 北京:国防工业出版社, 2014)

[19] Wang Shuliang. Data Field and Cloud Model Based Spatial Data Mining and Konledge Discovery[D]. Wuhan: Wuhan University, 2002(王树良. 基于数据场与云模型的空间数据挖掘和知识发现 [D]. 武汉:武汉大学, 2002)

[20] Deng Min, Liu Qiliang, Li Guangqiang, et al. Field-Theory Based Spatial Clustering Method[J]. *Journal of Remote Sensing*. 2010, 14(4): 694-709 (邓敏, 刘启亮, 李光强,等. 基于场论的空间聚类算法[J]. 遥感学报. 2010, 14(4): 694-709)

[21] Deng Min, Peng Dogliang, Liu Qiliang, et al. A Hierarchical Spatial Clustering Algorithm Based on Field Theory[J]. *Geomatics and Information Science of Wuhan University*, 2011, 36(7): 847-852 (邓敏,彭东亮, 刘启亮,等. 一种基于场论的层次空间聚类算法[J]. 武汉大学学报·信息科学版, 2011, 36(7): 847-852)

[22] Gan Wenyan, Li Deyi, Wang Jianmin. An Hierarchical Clustering Method Based on Data Fields[J]. *Acta Electronica Sinica*, 2006, 34(2):258-262 (淦文燕, 李德毅, 王建民. 一种基于数据场的层次聚类算法[J]. 电子学报, 2006, 34(2): 258-262)

[23] Wu Tao, Qin Kun. Image Segmentation Using Cloud Model and Data Field[J]. *PR&AI*, 2012, 25(3): 397-405(吴涛, 秦昆. 利用云模型和数据场的图像分割方法[J]. 模式识别与人工智能, 2012, 25(3): 397-405)

[24] Jiang Shengyi, Li Qinghua. GLOF: A New Approach for Mining Local Outlier[C]. The 2nd International Conference on Machine Learning and Cybernetics, Xi'an, China, 2003

[25] Rousseeuw P J, Hubert M. Robust Statistics for Outlier Detection[J]. *WIREs: Data Mining and Knowledge Discovery*, 2011, 1(1):73-79

Field-Theory Based Spatial Outlier Detecting Method

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

1　School of Geosciences and Info-physics, Central South University, Changsha 410083, China

Abstract: Spatial outlier detection is one of the major data mining methods. Detection of outliers will contribute to the discovery of implicit knowledge, significant changes, surprising patterns, and meaningful insights. In the field of geography, a spatial outlier is an object whose non-spatial attribute value is significantly different from the values of its spatial neighbors. Most current spatial outlier detection methods primarily consider that all the objects for outlier detection are correlated. Actually, spatial correlation decreases with the increase of distance. At the same time, the objects could be potentially wrongly identified as spatial outliers when there are several real outliers in their spatial neighborhoods. From the viewpoint of the spatial data field, a similar Gaussian potential function is utilized to measure the degree of spatial outlier degree. Further a field-theory based spatial outlier detecting algorithm is proposed. Firstly, the spatial clustering is employed to extract the local autocorrelation patterns, called clusters. Then the clusters were utilized to construct the reasonable and stable spatial neighborhoods using the constraint Delaunay triangulation. Finally, a robust spatial outlier measure is proposed to determine spatial outliers in each cluster. Experimental results show that the proposed method is effective for determining detecting spatial outliers in spatial point datasets.

Key words: spatial outlier detection; spatial clustering; field theory; spatial outlier measure

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

Foundation support: The National High-Tech R & D Program of China(863 Program), No.2013AA122301; the Hunan Provincial Science Fund for Distinguished Young Scholars, No. 14JJ1007; the National Natural Science Foundation of China, No.41471385; the Fundamental Research Funds for the Central Universities of Central South University, No. 2016zzts085.