DOI:10.13203/j.whugis20170008 文章编号:1671-8860(2018)10-1538-08

显著局部空间同位模式自动探测方法

徐 枫[1] 蔡建南[1] 刘启亮[1] 何占军[1] 邓 敏[1]

1 中南大学地理信息系,湖南 长沙,410083

摘 要:局部空间同位模式挖掘旨在揭示多类地理事件在异质环境下的共生共存规律。已有的方法一方面需要模式筛选的频繁度阈值参数,另一方面需要区域探测的划分参数或聚类参数,参数的不合理设置会导致挖掘结果不可靠甚至出现错误。因此,提出了一种显著局部空间同位模式自动探测方法。首先,基于空间统计思想,采用非参数模式重建方法对空间同位模式进行显著性判别,将全局非显著空间同位模式作为进一步局部探测的候选模式;然后,借助自适应空间聚类方法提取每个候选模式的热点区域;最后,通过不断生长并测试每个热点区域,界定显著局部空间同位模式的有效边界,即空间影响域。通过实验与比较发现,该方法能够客观且有效判别空间同位模式的显著性,并且自适应地提取局部同位模式的空间分布结构,降低了现有方法参数设置的主观性。

关键词:空间异质性;局部空间同位模式;非参数检验;模式重建;自适应空间聚类

中图分类号:P208 **文献标志码**:A

空间同位模式挖掘是空间数据挖掘的一个重要分支,能够有效发现多类地理事件间的共生关联关系,现已被广泛应用于生态环境、公共安全、商业选址、移动通信和交通运输等领域[1-2]。然而,由于地理事件具有空间异质特性[3-5],不同的地理事件经常仅在特定子空间内的邻近位置上频繁并发,发现此类空间模式(即局部空间同位模式)有助于深入理解不同空间现象在微观层次上的空间作用关系[6]。

近年来,在全局空间同位模式挖掘模型[7-8]的基础之上,通过特定的区域划分策略或空间聚类手段,发展了一系列的局部空间同位模式挖掘方法。区域划分的策略旨在将全局空间预先划分为一系列的子区域,进而在子区域内采用全局模型提取局部空间同位模式,主要的区域划分方法有四叉树分区法[9]和 k-邻近图分区法[10]。此类策略能够发现一些从全局视角难以发现的局部空间同位模式,但是其挖掘结果依赖于区域划分方法的选择,且人为划分的区域难以真实反映空间同位模式的空间分布结构。为此,一些学者进一步借助空间聚类的思想,分别探测每个候选模式频繁出现的局部热点区域[11-13]。此类方法能够有效区分不同局部空间同位模式分布区域的差异,但是大多需要对所有候选模式执行聚类操作,当空间变量种类众多时,将面临巨大的计算量。为此,文献[14]结合区域划分和空间聚类的优点,提出一种混合的策略,首先利用某个兴趣事件的热点对全局空间进行划分,进而在每个划分中提取与兴趣事件相关的局部同位模式,然后采用多分辨率格网聚类的方法界定每个局部模式的有效分布范围。该方法虽然可以首先剔除一些无效模式,但是其空间划分结果仍然会割裂局部空间同位模式原有的分布结构,进而可能导致某些有效模式的遗漏。

上述方法能够从一定程度上缓解空间异质性对空间同位模式挖掘带来的挑战,但是在挖掘过程中涉及过多的参数设置,在实际应用中由于缺乏相应的领域知识,将难以获得客观的挖掘结果。主要体现在以下两个方面:①现有局部空间同位模式挖掘的研究工作中几乎所有方法都需要预先设置频繁度阈值来评定空间同位模式的频繁度,阈值设置较高将可能遗漏某些有效模式,反之将

收稿日期:2017-11-26

项目资助:国家自然科学基金(41730105,41601410);湖南省科技计划项目(2015SK2078);资源与环境信息系统国家重点实验室开放基金;中南大学研究生自主探索创新项目(2017zzts174)。

第一作者:徐枫,博士生,主要从事空间数据挖掘方法及应用研究。xufengcsu@163.com

通讯作者:蔡建南,博士生。jncai@outlook.com

可能得到某些无效模式;②局部空间同位模式的分布区域提取需要设置空间划分参数或聚类参数,不合理的参数设置会破坏局部空间同位模式自身潜在的空间分布结构。针对以上问题,本文基于非参数统计和自适应聚类的思想,提出一种显著局部空间同位模式自动探测方法。

1　局部空间同位模式研究策略

为了减少人为参数设置对局部空间同位模式的频繁度度量和分布区域提取这两个过程的影响,分别给出相应的研究策略。

1)空间同位模式的频繁度实际上描述的是多类地理事件间的空间依赖关系,空间统计学中常通过建立两类事件分布相互独立的零假设,对多元点模式的空间依赖关系进行测试[15-17]。仿照该统计思想,将空间同位模式的频繁程度建模为显著性水平,通过非参数模式重建方法[18]构建模拟数据,进而识别统计上显著的空间同位模式。

2)由于空间同位模式包含多类地理事件,不同空间同位模式的分布各异,传统聚类方法仅能探测单类地理事件的分布热点,且参数设置困难,难以处理分布复杂的空间数据。为此,首先对空间同位模式的实例位置进行建模,将其作为空间聚类的对象,进而借助自适应空间聚类方法[19]自动提取空间同位模式的分布区域。

2　显著局部空间同位模式挖掘

2.1　基于模式重建的空间同位模式显著性判别

空间同位模式显著性判别的零模型需要在消除多类地理事件间分布依赖性的同时,保持单类地理事件自身的分布特征[17, 20]。本文借助一种非参数模式重建方法[18]构建零模型,相比于其他已有方法(如空间点过程方法[15, 21]和环形移动方法[22]),模式重建方法无需对数据的零分布做先验性假设,且不会破坏原始数据的分布结构。

首先,针对每个地理事件,采用多个空间统计量来刻画其原始数据 OD(original data)的分布特征。进而生成与原始数据 OD 实例个数相同的随机数据 SD(stochastic data),通过不断优化随机数据,使其与原始数据的分布特征尽可能的相似,优化过程的目标函数 $E(\mathrm{SD})$ 表达如下:

$$E(\mathrm{SD})=\sum_{i=1}^{I} w_i X\int_0^{R_i}(f_i(\mathrm{OD},r)-f_i(\mathrm{SD},r))^2\mathrm{d}r \tag{1}$$

式中,$f_i(\mathrm{OD},r)$ 和 $f_i(\mathrm{SD},r)$ 分别表示原始和随机数据中第 i 个空间统计量在邻域距离 r 上的统计值;w_i 和 R_i 分别表示第 i 个空间统计量的权重和自变量取值范围;I 表示空间统计量的个数。

为了在对数据分布特征详尽描述的同时,减少不同统计量间描述信息的冗余,本文共选取对相关函数 $g(r)$、最邻近分布函数 $D(r)$ 和球面接触分布函数 $H_s(r)$ 进行模式重建,分别用于描述数据的二阶统计特征、最邻近统计特征和形态学统计特征[24]。如图 1 所示,分别对事件 A 和 B 进行模式重建,模拟数据中很好地保持了原始数据中每类事件的分布特征。

进一步,用参与指数(participate index, PI)[1]作为检验统计量,判别空间同位模式的显著性。参与指数是空间同位模式的频繁度度量指标,具体表达为:

$$\mathrm{PI}(\mathrm{CP})=\min_{i=1}^{I}\left\{\frac{\sharp(\pi_{f_i}(\mathrm{instances}(\mathrm{CP})))}{\sharp(\mathrm{instances}(f_i))}\right\} \tag{2}$$

式中,$\sharp(\mathrm{instances}(f_i))$ 表示事件 f_i 的实例个数;$\sharp(\pi_{fi}(\mathrm{instances}(\mathrm{CP})))$ 表示事件 f_i 参与同位模式 CP(colocation pattern)的实例个数。进而,通过大量的模拟数据计算零假设下同位模式 CP 参与指数的实验分布,由此可以计算出空间同位模式 CP 参与指数的显著性 p 值:

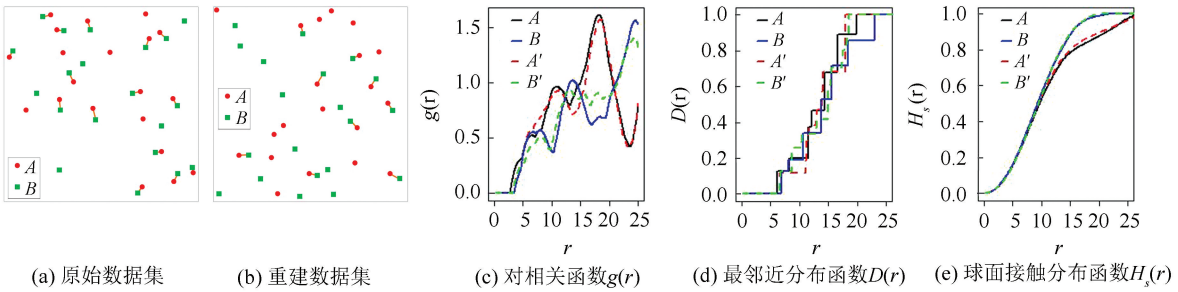
$$p_V=\frac{\sharp(\mathrm{PI}_{\mathrm{SD}n}(\mathrm{CP})\geqslant\mathrm{PI}_{\mathrm{OD}}(\mathrm{CP}))+1}{n+1},\quad n=1,2\cdots N \tag{3}$$

式中,$\mathrm{PI}_{\mathrm{SD}n}(\mathrm{CP})$ 和 $\mathrm{PI}_{\mathrm{OD}}(\mathrm{CP})$ 分别表示第 n 组模拟数据集和原始数据集中同位模式 CP 的参与指数;n 表示模式重建次数。给定显著性水平 α,若同位模式 CP 的显著性 $p_V\leqslant\alpha$,则拒绝零假设,将该模式识别为显著空间同位模式。

2.2　基于自适应聚类的候选局部模式热点探测

针对每个同位模式,首先采用 §2.1 方法检测其全局显著性,若不显著,则将其视为候选局部模式。进而采用自适应空间聚类方法[19]提取候选局部模式的分布热点。空间同位模式每个实例虽然包含多个空间点,但是各空间点彼此邻近,因此,如图 2(a)所示,可用同位模式实例中各空间点的平均位置将该模式建模为特殊的单类地理事件。

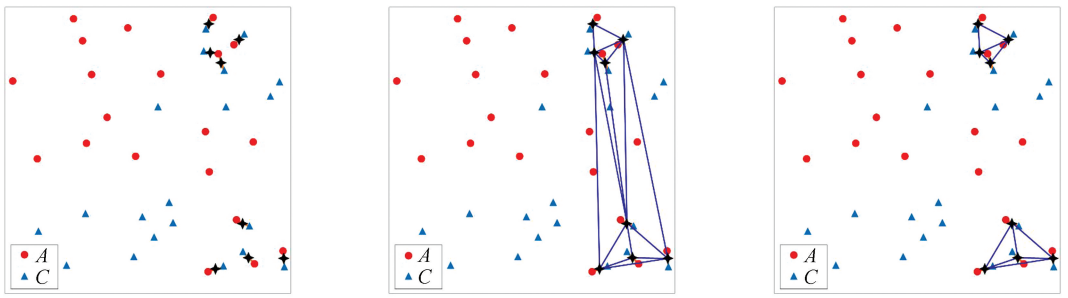
如图 2(b)所示,对此特殊地理事件的空间位置构建 Delaunay 三角网 DTN(delay tolerant network)。先从全局层次对三角网 DTN 的边长施加约束,对于每个空间点 P_i,删除与其直接相



(a) 原始数据集 (b) 重建数据集 (c) 对相关函数$g(r)$ (d) 最邻近分布函数$D(r)$ (e) 球面接触分布函数$H_s(r)$

图1 基于模式重建的零模型构建

Fig.1 Construction of Null Model Based on Pattern Reconstruction



(a) 空间同位模式实例的空间位置 (b) Delaunay三角网 (c) 整体和局部约束

图2 基于自适应聚类的空间同位模式热点探测

Fig.2 Detection of Hot Spots of a Spatial Colocation Pattern Based on Adaptive Clustering Method

连的边中长度大于全局边长统计量 GET(P_i)的整体长边,表达式为:

$$\mathrm{GET}(P_i)=\mathrm{mean}(\mathrm{DTN})+\frac{\mathrm{mean}(\mathrm{DTN})}{\mathrm{mean}(NN^1(P_i))}\cdot\mathrm{Variation}(\mathrm{DTN}) \quad (4)$$

其中,mean(DTN)和 Variation(DTN)分别表示三角网 DTN 中所有边长的平均值和标准差;mean($NN^1(P_i)$)表示与点 P_i 直接相连的所有边的平均长度。进一步对所剩的每个子图 SG_i,从局部层次删除每个空间点 P_i 二阶邻域内边长大于局部边长统计量 LET(P_i)的局部长边,表达式为:

$$\mathrm{LET}(P_i)=\mathrm{mean}(NN^2(P_i))+\frac{\sum_{k=1}^{\#(\mathrm{SG}_i)}\mathrm{Variation}(NN^1(P_k))}{\#(\mathrm{SG}_i)} \quad (5)$$

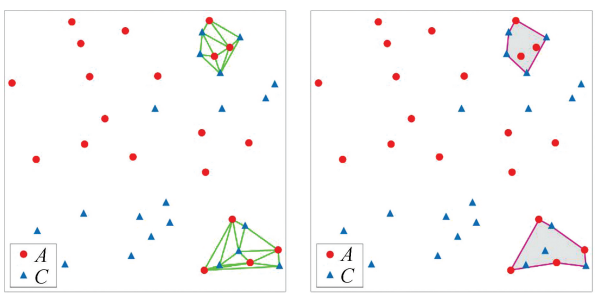
式中,mean($NN^2(P_i)$)表示空间点 P_i 二阶邻域内所有边长的平均值;Variation($NN^1(P_k)$)表示子图 SG_i 中与空间点 P_k 直接相连的所有边的长度标准差;#(SG_i)表示子图 SG_i 中空间点的个数。如图2(c)所示,删除整体长边和局部长边后,三角网 DTN 被划分为一系列的子图,每个子图即为该候选局部模式的分布热点。

2.3 显著局部空间同位模式的有效区域提取

为进一步检验候选模式的局部显著性,需要描绘候选模式的热点区域。如图3(a)所示,对于候选局部模式的每个分布热点,分别构建 Delaunay 三角网 DTN_i 连接所有空间点。根据三角网中的边长统计量定义长边,表达式为:

$$ET=\mathrm{mean}(\mathrm{DTN}_i)+3\mathrm{Variation}(\mathrm{DTN}_i) \quad (6)$$

式中,mean(DTN_i)和 Variation(DTN_i)分别表示三角网 DTN_i 中所有边长的平均值和标准差。通过删除包含任一长边的三角形,对三角网 DTN_i 进行修剪;将修剪后的三角网中的非公共边视为边界边,由边界边包围的区域即为该候选模式的热点区域,如图3(b)所示。



(a) Delaunay三角网 (b) 热点区域

图3 基于 Delaunay 三角网的热点区域描绘

Fig.3 Delineation of Hot Spots Based on Delaunay DTN

不断将热点区域向外扩展,在局部区域内对该模式的显著性进行测试。如果发现任一显著性

小于等于给定显著性水平 α 的区域,则将该候选模式识别为显著局部空间同位模式,并将相应的区域定义为显著区域。各个显著区域继续扩展至该模式的局部显著性消失,最终将最大的显著区域识别为该显著局部模式的有效区域。

3　实验分析与应用

为了验证本文方法的有效性,分别采用包含预设模式的模拟数据与实际生态群落数据进行实验分析,并与 Ding 等人提出的方法[14](简称 MRG)进行比较。为了使 MRG 算法适用于本文的实验数据和目的,实验中首先对研究区域施加规则格网,进而将包含任意事件的格网定义为空间事务,并按原文建议设置算法参数。本文方法中全局和局部的模式重建次数均设为 99,空间同位模式的显著性水平设为 0.05。实验测试环境为 Windows 10 系统,CPU 2.50 GHz,内存 8 GB。

3.1　模拟实验与比较

模拟数据集如图 4 所示,其中事件 A、B 和 C 均具有预设的聚集结构,且不同类型之间包含相互重叠的空间簇,事件 D 为随机分布的干扰事件。模拟实验中采用 Yoo 等人[25]的建议,借助 L 函数估计合适的邻域距离,估计结果见图 5。

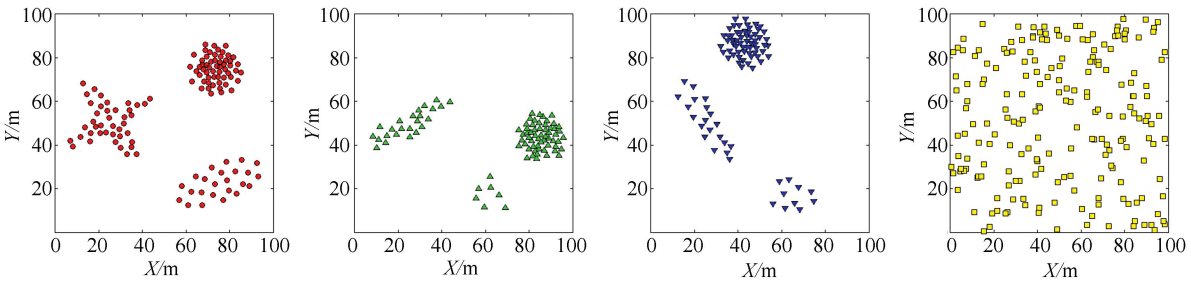


图 4　模拟数据集

Fig.4　Simulated Dataset

本文方法自动探测的结果如表 1 所示,包含 3 个显著局部空间同位模式{A,B}、{A,C}、{B,C}和 1 个显著全局空间同位模式{A,B,C}。以空间同位模式{A,B}为例,进一步采用 cross-K 函数[24]验证挖掘结果的正确性。如图 5(a)所示,全局范围内事件 A 和 B 的 cross-K 函数计算结果表明两者之间没有显著的空间依赖关系,本文方法全局判别结果与其吻合。如图 6(a)所示,本文方法进一步自适应提取了模式{A,B}的热点区域及其有效范围。如图 5(b)所示,采用 cross-K 函数对有效范围内两类事件的空间依赖性进行验证,发现事件 A 和 B 之间具有显著的空间依赖性。因此,本文方法能够准确识别局部空间同位模式,且能有效剔除随机事件对结果的影响。

表 1　本文方法对模拟数据集的自动探测结果

Tab.1　Results of Our Method on the Simulated Dataset

同位模式	全局		局部			
	PI	p_V	$\underline{\text{PI}}$	$\overline{\text{PI}}$	$\underline{p_V}$	$\overline{p_V}$
{A, B}	0.32	0.17	0.50	0.78	0.01	0.04
{A, C}	0.33	0.20	0.67	0.71	0.01	0.04
{B, C}	0.19	0.38	0.78	1.00	0.02	0.03
{A, B, C}	0.19	0.05	—	—	—	—

注:$\underline{\text{PI}}$和$\overline{\text{PI}}$为最小和最大局部参与指数;$\underline{p_V}$和$\overline{p_V}$为最小和最大局部 p 值

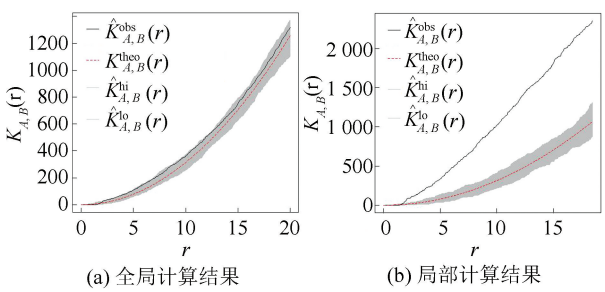


图 5　事件 A 和 B 的 cross-K 函数计算结果

Fig.5　Results of cross-K Function for Features A and B

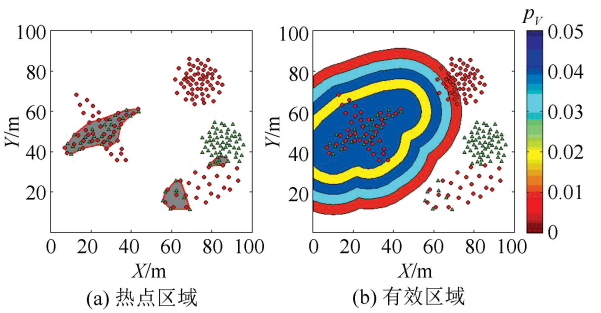


图 6　本文方法对局部空间同位模式{A,B}的探测结果

Fig.6　Regional Pattern {A, B} Detected by Our Method

为了与本文方法进行对比分析,进一步采用 MRG 算法进行实验。对于 MRG 算法,用 10×10 的规则格网构建空间事务,采用初始单元大小

为 50×50、终止单元大小为 12.5×12.5 的多分辨率格网探测各类事件的热点区域，但未探测出任何事件的热点。因为对于每类事件，模拟数据中都不存在出现概率大于其全局概率的人造格网，从而 MRG 算法不能从本文模拟数据中发现任何局部空间同位模式。

以空间同位模式{A，B}为例，从效率和稳定性两个方面测试模式重建次数设置对本文方法挖掘结果的影响，每种重建次数都实验 20 次，取其运行时间的平均值衡量算法效率，并以全局 p 值和有效区域内局部 p 值的标准差衡量算法稳定性。如表 2 所示，运行时间与重建次数呈线性增长关系，当重建结果大于等于 99 次时，算法挖掘结果趋于稳定。现有研究亦发现，在显著性水平为 0.05 时，99 次模拟次数能够保证多数应用的可靠性[26]。

表 2 模式重建次数对本文方法性能的影响

Tab.2 Effects of the Simulation Times on Our Method

算法性能	模式重建次数				
	49	99	499	999	4 999
运行时间/s	3.010	5.858	29.763	59.642	291.312
全局 p 值标准差	0.025	0.016	0.013	0.012	0.010
局部 p 值标准差	0.020	0.011	0.009	0.010	0.008

3.2 实际应用与分析

进一步采用本文方法探测湿地生态数据集中的局部共生关系，以验证本文方法的实际应用效果。在湿地生态系统中，存在复杂的植被种间关系，且易受生长环境的不同而发生显著变化[27-28]，探测湿地物种间的局部共生关系对于研究生态群落结构、维持生态系统平衡、保护物种多样性和促进环境可持续发展都有着重要的现实意义。本文选取中国东北地区某湿地的 5 种沼泽植被（毛果苔草、漂筏苔草、狭叶甜茅、小叶章和沼柳）进行实验分析，5 种沼泽植被的空间分布如图 7 所示，分别有 666、1 039、1 660、387 和 2 555 个实例。

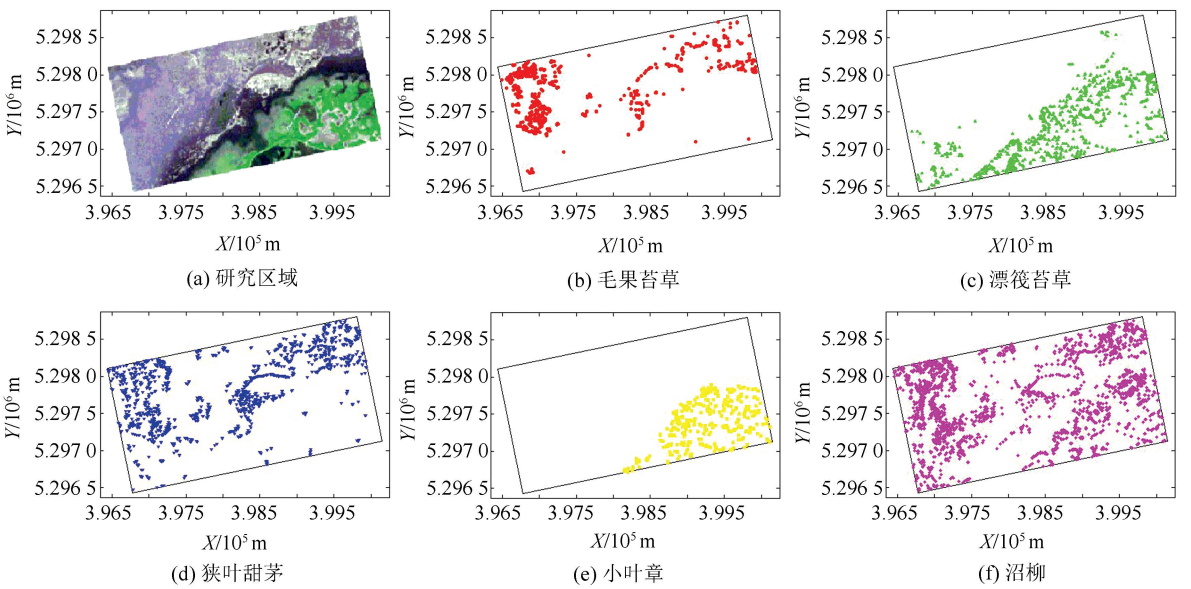


图 7 研究区域及 5 种沼泽植被的空间分布

Fig.7 Study Area and Locations of Five Types of Wetland Species

生态学中，不同生态物种间的相互作用关系可能存在显著差异[28]，因此采用单一邻域距离挖掘空间同位模式难以准确发现共生物种。本文采用 Barua 等人[29]的建议，设置了多个邻域距离（50 m、100 m 和 150 m）进行实验分析，探测结果如表 3 所示。分析实验结果，可以发现：①部分显著空间同位模式的参与指数很低，现有方法设置较高的参与指数阈值可能会遗漏这些模式，本文方法通过对空间同位模式的显著性进行非参数检验，能够更加客观地评价空间同位模式的频繁程度；②在 50 m 的邻域距离下主要发现了一些显著全局空间同位模式；随着邻域距离的增加，部分全局模式会逐渐退化为局部模式，甚至消失，有些局部模式也会逐渐消失，同时也出现了一些新的局部模式；另外，也有一些显著同位模式的空间层次不会随着邻域距离的变化而变化，说明这些植被之间具有稳定的共生关系，对湿地生态系统构成起主导性作用。

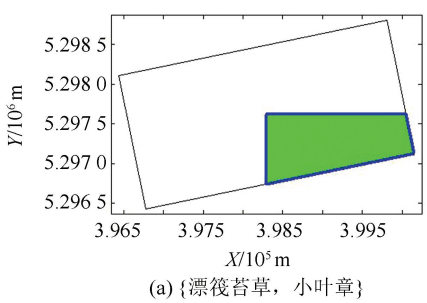
表 3　本文方法探测的显著空间同位模式

Tab.3　Significant Spatial Colocation Patterns Detected by Our Method

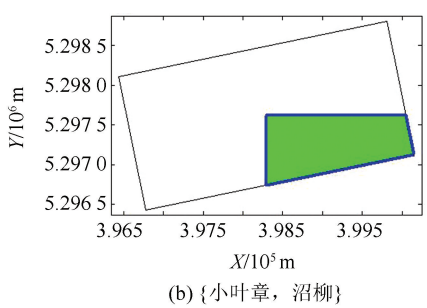
显著空间同位模式	邻域距离 50 m						邻域距离 100 m						邻域距离 150 m					
	全局		局部				全局		局部				全局		局部			
	PI	p_V	$\underline{PI}$	$\overline{PI}$	$\underline{p_V}$	$\overline{p_V}$	PI	p_V	$\underline{PI}$	$\overline{PI}$	$\underline{p_V}$	$\overline{p_V}$	PI	p_V	$\underline{PI}$	$\overline{PI}$	$\underline{p_V}$	$\overline{p_V}$
{小叶章,沼柳}	0.20	0.54	0.47	0.92	0.01	0.05	—	—	—	—	—	—	—	—	—	—	—	—
{毛果苔草,狭叶甜茅}	0.65	0.01	—	—	—	—	0.86	0.01	—	—	—	—	0.93	0.01	—	—	—	—
{毛果苔草,沼柳}	0.37	0.01	—	—	—	—	0.60	0.01	—	—	—	—	0.68	0.20	0.86	0.94	0.03	0.05
{漂筏苔草,狭叶甜茅}	—	—	—	—	—	—	0.26	1.00	0.32	0.57	0.01	0.04	0.39	1.00	0.40	1.00	0.02	0.02
{漂筏苔草,小叶章}	0.31	0.03	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
{漂筏苔草,沼柳}	0.33	0.83	0.65	0.94	0.01	0.05	0.46	1.00	0.99	1.00	0.03	0.04	0.54	1.00	1.00	1.00	0.02	0.02
{狭叶甜茅,沼柳}	0.70	0.01	—	—	—	—	0.80	0.15	0.29	0.99	0.01	0.05	0.88	0.95	0.41	1.00	0.01	0.05
{小叶章,沼柳}	0.20	0.54	0.47	0.92	0.01	0.05	—	—	—	—	—	—	—	—	—	—	—	—
{毛果苔草,漂筏苔草,狭叶甜茅}	—	—	—	—	—	—	0.17	0.99	1.00	1.00	0.03	0.03	0.29	1.00	0.35	0.99	0.03	0.04
{毛果苔草,漂筏苔草,沼柳}	—	—	—	—	—	—	—	—	—	—	—	—	0.25	1.00	0.33	0.33	0.04	0.04
{毛果苔草,狭叶甜茅,沼柳}	0.35	0.01	—	—	—	—	0.60	0.01	—	—	—	—	0.67	0.02	—	—	—	—
{漂筏苔草,狭叶甜茅,小叶章}	—	—	—	—	—	—	—	—	—	—	—	—	0.03	1.00	1.00	1.00	0.02	0.03
{漂筏苔草,狭叶甜茅,沼柳}	0.08	1.00	0.16	0.29	0.01	0.04	0.25	1.00	0.28	1.00	0.01	0.04	0.39	1.00	0.38	1.00	0.02	0.05
{漂筏苔草,小叶章,沼柳}	0.16	0.01	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
{狭叶甜茅,小叶章,沼柳}	—	—	—	—	—	—	0.02	1.00	1.00	1.00	0.02	0.02	—	—	1.00	1.00	0.02	0.02

注:$\underline{PI}$ 和 $\overline{PI}$ 为最小和最大局部参与指数;$\underline{p_V}$ 和 $\overline{p_V}$ 为最小和最大局部 p 值

采用 MRG 算法进一步比较分析。采用 100 m×100 m 的规则格网定义空间事务,结果仅发现了小叶章的热点区域,并挖掘出 2 个包含小叶章的局部同位模式{漂筏苔草,小叶章}和{小叶章,沼柳},其有效边界分别如图 8(a)和 8(b)所示。可见,MRG 算法受限于人工格网的划分,导致探测的局部模式不完整,且局部模式的有效区域被人为切边,难以反映局部同位模式自然的分布结构。



(a) {漂筏苔草, 小叶章}



(b) {小叶章, 沼柳}

图 8　实际应用中 MRG 算法探测结果

Fig.8　Results of MRG Method on the Real Dataset

实际应用中,本文方法的运行效率主要取决于 3 个部分:①所有可能模式的全局显著性检验,其时间复杂度约为 $O(N \cdot 2^K)$,其中 N 为模式重建次数,K 为地理事件类型数目;②每个候选局部模式的热点探测,其时间复杂度约为 $O(M \cdot \log M)$,其中 M 为该候选模式的实例个数;③每个候选局部模式的局部显著性检验,针对每个局部区域,其时间复杂度约为 $O(N \cdot X^2)$,其中 X 为该局部区域内该候选模式的实例个数。在较大的邻域距离下,每个可能模式的实例个数均会增加,且需要在局部层次进行检验的候选局部模式个数亦会增加,因此,如图 9 所示,本文方法运行时间会随邻域距离的增加而显著增加。相比于现有方法,本文方法虽然计算量大,但是很大程度上降低了现有方法中参数设置的主观性。还可以结合空间索引和高性能计算等技术改善本文方法在实际应用中的计算效率。

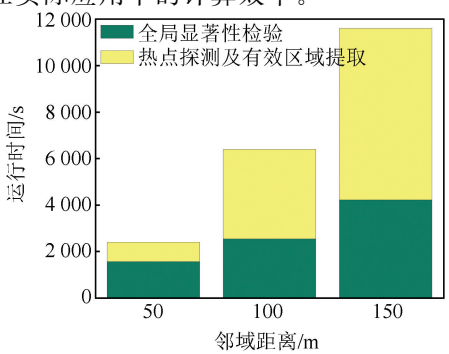


图 9　实际应用中本文方法运行时间

Fig.9　Running Time of Our Method on the Real Dataset

4 结 语

为了降低人为参数对局部同位模式挖掘结果的影响,本文提出了一种显著局部空间同位模式的自动探测方法。通过实验分析和比较发现,本文方法不仅能够有效识别统计上显著的局部空间同位模式,还能自适应地提取局部同位模式的分布区域和有效边界,从而能够更加客观地揭示地理事件间的相互作用关系。

进一步的研究工作主要集中于:①本文多类事件间邻域距离的选择具有一定的主观性,需要研究多元事件邻域的自适应构建方法;②局部空间同位模式热点区域生长采用的是各向等距离扩展的策略,还需要研究热点区域的有向扩展方法。

参 考 文 献

[1] Shekhar S, Huang Y.Discovering Spatial Colocation Patterns: A Summary of Results[C]. International Symposium on Spatial and Temporal Databases, Redondo Beach, USA, 2001

[2] Yoo J S, Shekhar S, Smith J, et al. A Partial Join Approach for Mining Colocation Patterns[C]. The 12th Annual ACM International Workshop on Geographic Information Systems, Washington D C, USA, 2004

[3] Openshaw S. Geographical Data Mining: Key Design Issues[C]. Proceedings of GeoComputation, Virginia, USA, 1999

[4] Goodchild M F. The Fundamental Laws of GI-Science[R]. University Consortium for Geographic Information Science, University of California, Santa Barbara, 2003

[5] Shekhar S, Evans M R, Kang J M, et al. Identi-fying Patterns in Spatial Information: A Survey of Methods [J]. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2011, 1(3): 193-214

[6] Sha Zongyao, Li Xiaolei. Algorithm of Mining Spatial Association Data Under Spatially Heterogeneous Environment [J]. *Geomatics and Information Science of Wuhan University*, 2009, 34(12): 1 480-1 484(沙宗尧, 李晓雷. 异质环境下的空间关联规则挖掘[J]. 武汉大学学报·信息科学版, 2009, 34(12): 1 480-1 484)

[7] Yoo J S, Shekhar S. A Joinless Approach for Mining Spatial Colocation Patterns [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2006, 18(10): 1 323-1 337

[8] Xiao X, Xie X, Luo Q, et al. Density Based Colocation Pattern Discovery [C]. The 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Irvine, CA, USA, 2008

[9] Celik M, Kang J M, Shekhar S. Zonal Colocation Pattern Discovery with Dynamic Parameters [C]. The 7th IEEE International Conference on Data Mining, Omaha, NE, USA, 2007

[10] Qian F, Chiew K, He Q, et al. Mining Regional Colocation Patterns with KNNG[J]. *Journal of Intelligent Information Systems*, 2014, 42(3): 485-505

[11] Eick C F, Parmar R, Ding W, et al. Finding Regional Co-location Patterns for Sets of Continuous Variables in Spatial Datasets[C]. The 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Irvine, California, 2008

[12] Mohan P, Shekhar S, Shine J A, et al. A Neighborhood Graph Based Approach to Regional Colocation Pattern Discovery: A Summary of Results[C]. The 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Chicago, USA, 2011

[13] Wang S, Huang Y, Wang X S. Regional Colocations of Arbitrary Shapes[C]. International Symposium on Spatial and Temporal Databases, Munich, Germany, 2013

[14] Ding W, Eick C F, Yuan X, et al. A Framework for Regional Association Rule Mining and Scoping in Spatial Datasets[J].*Geoinformatica*, 2011, 15(1): 128

[15] Illian J, Penttinen A, Stoyan H, et al. Statistical Analysis and Modelling of Spatial Point Patterns[J]. *Technometrics*, 2008, 47(4):516-517

[16] Gelfand A E. Handbook of Spatial Statistics[M]. UK: CRC Press, 2010

[17] Wiegand T, Moloney K A. Handbook of Spatial Point Pattern Analysis in Ecology[M]. UK: CRC Press, 2013

[18] Wiegand T, He F, Hubbell S P. A Systematic Comparison of Summary Characteristics for Quantifying Point Patterns in Ecology [J]. *Ecography*, 2013, 36(1): 92-103

[19] 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)

[20] Barua S, Sander J. Mining Statistically Significant Colocation and Segregation Patterns [J]. *IEEE Transactions on Knowledge & Data Engineering*, 2014, 26(5):1 185-1 199

[21] Neyman J, Scott E L. Statistical Approach to Problems of Cosmology[J].*Journal of the Royal Statistical Society*, 1958, 20(1):143

[22] Lotwick H W, Silverman B W. Methods for Analysing Spatial Processes of Several Types of Points [J].*Journal of the Royal Statistical Society. Series B (Methodological)*, 1982, 44(3): 406-413

[23] Diggle P J. Statistical Analysis of Spatial Point Patterns [M]. London: Edward Arnold Publishers, 2003

[24] Ripley B D. The Second Order Analysis of Stationary Point Processes[J]. *Journal of Applied Probability*, 1976,13(2): 255-266

[25] Yoo J S, Bow M. Mining Spatial Colocation Patterns: A Different Framework [J]. *Data Mining and Knowledge Discovery*, 2012, 24(1): 159-194

[26] Besag J, Diggle P J. Simple Monte Carlo Tests for Spatial Patterns [J].*Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1977, 26(3): 327-333

[27] Zimmer K D, Hanson M A, Butler M G. Interspecies Relationships, Community Structure, and Factors Influencing Abundance of Submerged Macrophytes in Prairie Wetlands[J].*Wetlands*, 2003, 23(4): 717-728

[28] Keddy P A. Wetland Ecology: Principles and Conservation[M]. UK: Cambridge University Press, 2010

[29] Barua S, Sander J. Mining Statistically Sound Colocation Patterns at Multiple Distances[C]. The 26th International Conference on Scientific and Statistical Database Management, Aalborg, Denmark, 2014

An Automatic Method for Discovering Significant Regional Spatial Colocation Patterns

XU Feng[1]　*CAI Jiannan*[1]　*LIU Qiliang*[1]　*HE Zhanjun*[1]　*DENG Min*[1]

1　Department of GeoInformatics, Central South University, Changsha 410083, China

Abstract: Discovery of regional spatial colocation patterns facilities understanding of the spatial dependency of different spatial features at the regional scale. However, two challenges remain: ①appropriate thresholds for prevalence measures are difficult to specify without prior knowledge; and ②natural localities of regional spatial colocation patterns with different densities and shapes can hardly be automatically detected. On that account, an automatic method for discovering significant regional spatial colocation patterns is proposed in this paper. First, a nonparametric statistical model is developed to test for significance of spatial colocation patterns. Then, an adaptive spatial clustering method is modified to detect hot spots of each candidate regional spatial colocation pattern that is not identified as a statistically significant spatial colocation pattern at the global scale. At last, all hot spots are iteratively expanded until no larger statistically significant localities can be detected. Comparison between this automatic method and an existing method is carried out with both simulated and ecological datasets. Experiments show that the regional spatial colocation patterns can be effectively detected with less subjectivity and prior knowledge by this automatic method.

Key words: spatial heterogeneity; regional spatial colocation patterns; nonparametric test; pattern reconstruction; adaptive spatial clustering

First author: XU Feng, PhD candidate, specializes in the methods and applications of spatial data mining. E-mail: xufengcsu@163.com

Corresponding author: CAI Jiannan, PhD candidate. E-mail: jncai@outlook.com

Foundation support: The National Natural Science Foundation of China, Nos. 41730105,41601410; the Science and Technology Foundation of Hunan Province, No. 2015SK2078; Open Research Fund of the State Key Laboratory of Resources and Environmental Information System; the Postgraduate Research and Innovation Foundation of Central South University, No. 2017zzts174.