引用格式:范协裕,陈瀚阅,刑世和.连续变量的自适应局部空间同位模式挖掘算法[J].地球信息科学学报,2016,18(7):902-909. [Fan X Y, Chen H Y, Xing S H. 2016. Self-adaptive local Co-location pattern mining algorithm for continuous variables. Journal of Geo-information Science, 18(7):902-909.] **DOI:**10.3724/SP.J.1047.2016.00902

连续变量的自适应局部空间同位模式挖掘算法

范协裕,陈瀚阅,刑世和*

福建农林大学资源与环境学院,福州 350002

Self-adaptive Local Co-location Pattern Mining Algorithm for Continuous Variables

FAN Xieyu, CHEN Hanyue and XING Shihe*

College of Resource and Environmental Science, Fujian Agriculture and Forestry University, Fuzhou 350002, China

Abstract: Existing approaches in finding the local co-location patterns have several shortcomings: (1) they depend on user predefining thresholds for proximity between the spatial feature instances and (2) the mining results miss the statistically significant explanation. In this paper, we proposed a new self-adaptive method for finding the local co-location patterns for spatial datasets containing continuous variables. The interestingness and indicator function and the proximity area that are defined based on the Voronoi diagrams are introduced. A proximity matrix is built to avoid user predefining thresholds for proximity. At last, the local Getis-Ord's G_i^* statistic quantity for the interestingness value is employed, which endowed the mining results with statistical significant. The actual datasets for cropland productivity surveying jointly with the land suitability evaluation results for tobacco planting and for water pollution are used to test the developed algorithm. The experimental results show that, the proposed approach is able to identify different local co-location patterns without the interference of user specified thresholds for proximity, and the captured local co-location patterns in the cropland productivity surveying datasets reveal the localized specified phenomenon in the experimental area. This approach has practical significances for cropland productivity surveying.

Key words: spatial co-location pattern; local; statistically significant; continuous variables

***Corresponding author:** XING Shihe, E-mail: fafuxsh@126.com

摘要:目前,局部空间同位模式挖掘方法存在需要预设定邻域范围、挖掘的结果无统计显著性意义而难以对结论进行科学地判定等问题,如当前常用的 K 近邻方法难以确定合适的搜索圆半径,而固定距离法由于空间数据集的多尺度特性,距离阈值的设定对结果的影响较大。因此,针对连续变量的空间采样点数据集,本文提出了一种自适应局部空间同位模式挖掘算法。首先,定义了连续变量的空间同位模式兴趣度函数、模式指示器函数及Voronoi邻域,并通过构建Voronoi邻域矩阵避免了预设定邻域阈值的问题,最后采用 G_i^* 统计量进行局部空间同位模式及其区域的发现,使挖掘的结果具有统计显著性意义,进而帮助专家对挖掘结果做出更科学的判定。通过使用真实的连接了烟草适应性评价结果的耕地地力样点调查数据和水污染数据,对开发的算法进行测试。实验结果表明,算法无需预设邻域范围,可查找同区域内的不同空间同位模式。实验所发现的局部空间同位模式发现了实验数据研究区域存在的特有现象,对耕地地力调查工作具有实际的指导作用。

关键词:空间同位模式;局部;统计显著性;连续变量

收稿日期:2015-07-27;**修回日期:**2015-11-09.

基金项目:福建省教育厅科技计划项目(JA14102);国家自然科学基金青年科学基金项目(41401399)。

作者简介:范协裕(1985-),男,福建永春人,博士,讲师,研究方向为空间数据挖掘、网络地理信息系统。
E-mail: xunbei100@aliyun.com

***通讯作者:**刑世和(1962-),福建连江人,博士,教授,研究方向为土地(壤)资源持续利用与评价。E-mail: fafuxsh@126.com

1 引言

空间同位现象指不同类型的空间实例或事件频繁地在一起或者邻近位置上出现，是一种不同空间特征要素类型间特殊的关联模式。在生态学领域、环境污染和社会经济生活中经常可以发现空间同位现象，如共生物种、疾病与污染源、购物广场与取款机等之间的关联关系。

空间数据集中很多关联知识都具有区域性特征，而全局统计方法很少提供有用的信息[1]。当前的空间挖掘算法普遍是挖掘全局的空间同位模式[2-5]，如Barua所述使用统一的全局参与度阈值存在诸多问题[6]，同时以上的算法均针对离散值空间特征数据。Ding和Qian等对区域的空间模式进行了挖掘，但都是以先进行区域发现，进而在发现的区域上使用传统同位模式算法进行挖掘的流程展开，而并非是发现区域同位模式进而确定其范围的模式[7-8]。Eick等的方法可以同时实现模式挖掘与区域发现，但是其使用全局的兴趣度函数作为聚类算法的目标函数，导致了其可能遗漏一些重要的模式[9]。Ord等提出了使用 G_i^* 统计量分别对空间数据集中的每个变量进行统计，并通过可视化方式发现变量之间的相关性，该方法对每个空间对象计算 G_i^* 统计量的过程仍然需要预先设置的尺度[10-11]，并且其只针对离散型变量。空间数据具有多尺度问题，不同的尺度下空间关联规则差异很大，因此在确定区域空间同位模式的区域时，有学者采用k近邻法或者距k离阈值等预先设定的方式来确定空间实例的邻域范围[8,12]。

针对常见的空间离散点采样数据（如从温度传感器，土壤采样点等获取的连续变量空间数据），本文提出了一种自适应区域空间同位模式挖掘算法，该方法针对连续值变量的空间特征数据，以候选模式的平均 Z 值作为兴趣度函数。通过定义Voroni邻域确定空间实例的邻域，避免了k近邻方式确定邻域实例方法中需要预先设定k值的问题及使用距离阈值需要预设定范围的问题。最后，利用 G_i^* 统计量，发现具有统计意义的空间同位模式的热点区域，最终实现区域空间同位模式的发现。

2 相关研究

空间同位模式挖掘算法的目的是找出空间特征中频繁邻近的空间特征子集[2]。最早的空间关联规则挖掘方法由传统的关联规则挖掘扩展而来，基于空间事物的方法[13]，将空间关系及属性离散化后，利用传统关联规则挖掘方法中的Apriori算法或FPfgrow算法[14-15]。Shekhar和Huang以邻域集合代替事物集合，提出了事物中心模型，并利用关联规则算法Apriori-gen算法从布尔型空间特征中挖掘频繁邻近的空间特征子集，算法采用全连接方法挖掘邻近的频繁子集。为了提高算法的效率，Yooh和Shekhar等先后提出不同改进算法[3-5]。Jin和Shekhar等利用空间分割，对空间邻域子集采用部分连接算法，通过跟踪被分割在不同部分之间的邻域实例的信息来保证算法的正确性，随后进一步提出了利用星形邻域查阅实例表的方法取代连接运算，并对候选同位模式进行粗过滤的无连接算法，进一步提高了算法的效率[3-5]。

Celik等使用了一种改进的四叉树索引来挖掘局部空间同位模式，并使用参与度阈值作为条件对树节点进行剪枝以算法的效率[16]。受区域知识发现的驱动，Ding等将聚类算法引入到热点区域的发现中，通过定义聚类的目标函数，使用聚类算法查找关心的热点区域，然后在热点区域上使用Apriori算法对空间同位模式进行挖掘。利用该方法，Ding等研究了Texas州水源附近重金属之间的关联规则[1]。利用相同的程序框架[7]，Eick等进一步使用候选模式的 Z 值累积作为兴趣度函数，迭代计算区域所有模式的兴趣度函数值，并以获取最大全局目标函数为目标，使用改进的类似CLARANS聚类算法的CLEVER算法，对空间特征数据集进行聚类运算，最后得到各个聚类区域及区域内使全局目标函数值最大的关联模式[7,17]。与Ding等提出的区域模式挖掘算法相比，该方法可以同时实现模式挖掘与区域查找，但是算法存在计算复杂度大，而且其在聚类的过程中使用全局的目标函数，导致算法在运算过程中虽然对所有的可能模式产生的兴趣度函数值进行了计算，但只取使兴趣度值最大的模式，因此有可能遗漏一些潜在的显著模式。

Huang等则将聚类的方法引入到空间同位模式的挖掘中，基于“如果某个空间要素A在B的邻域内的平均密度高于A的全局平均密度，则A存在与B具有同位关联的趋势”的假设，定义密度比率作为邻近性的度量函数，对所有空间特征要素计算相互之间的邻近矩阵，进而对特征要素运用聚类算法找

出存在同位模式关系的空间要素[18]。

边馥苓和Qian等从同位实例邻域关系度量方法上展开了研究。边馥苓利用空间实例的k近邻实例集合度量其与其他实例的相似度替代基于距离阈值的邻近度量方式，同时开发了基于格网索引的k邻近特征同位模式挖掘算法，算法对参数的设定和数据集的大小有很好的容忍度，并能解决在基于距离阈值的空间同位模式算法中难以解决的对稀少空间特征对象的同位模式的发现[12]。Qian等同样使用k近邻代替距离作为度量不同特征要素的空间实例的邻近关系，并引入了距离变异因子来衡量挖掘的地理空间中互为k近邻的特征实例组成的区域的内部邻域距离一致性。算法的流程可分为利用距离变异因子寻找、合并邻域距离一致的区域，以及在区域内使用基于连接的算法进行同位模式挖掘2个步骤。算法在邻域设定上，需在执行前设定k近邻的k初始值及其他几个阈值参数[8]。

目前，使用的空间模式挖掘算法中[2-5]，使用统一预设定的全局参与度阈值的方式仍存在不足。例如，在随机分布的空间数据集中，出现参与度值很高的子集并不少见，甚至可能出现一组相互关联的特征要素之间参与度特别低的情况[6]。Barua等通过引入统计检验方法，可以找出具有统计显著性意义的空间同位模式和空间分离模式[6]，但是算法使用仿真试验的方法进一步增加了其实现的时间复杂度。

3 相关概念定义及描述

首先对本文采用的一些相关概念及函数进行定义及描述。

给定由空间对象 o 组成的空间数据集 $F=\{o_1, o_2,\cdots,o_n\}$，空间数据 F 是一个基于空间关系数据库模式的空间数据集。空间对象 o_i 是 F 的一个元组。每个元组由表示空间位置的属性和非空间属性组成。

定义如下：

$S=\{S_1,S_2,\cdots,S_m\}$：空间属性

$N=\{A_1,A_2,\cdots,A_h\}$：非空间属性

$P=\{A_1\uparrow,A_1\downarrow,A_2\uparrow,A_2\downarrow,\cdots,A_h\uparrow,A_h\downarrow\}$：所有可能的同位模式属性，其中向上箭头表示属性值偏高，向下箭头表示属性值偏低。

$B\subseteq P$，且$\forall i\in[1,k]$，如果$A_i\uparrow\in B$，则$A_i\downarrow\notin B$：候选同位模式。

3.1 连续变量空间同位模式兴趣度度量函数

本文定义了候选空间同位模式的兴趣度函数，如式(1)所示。并使用了Eick等定义的兴趣度函数作为模式指示函数，用于指示模式中的所有属性是否都不违背模式的定义[9](式(2))。

$$i(B,o)=\begin{cases}\sum_{p\in B}z(p,o)/|B|, & ic(B,o)>0\\ 0, & \text{其它}\end{cases} \quad (1)$$

$$ic(B,o)=\prod_{p\in B}z(p,o) \quad (2)$$

模式属性成员用于表示空间实例中属性值的高低特性(式(3)-(5))。

$$z(A\uparrow,o)=\begin{cases}z-score(A,o), & z-score(A,o)>0\\ 0, & \text{其它}\end{cases} \quad (3)$$

$$z(A\downarrow,o)=\begin{cases}z-score(A,o), & z-score(A,o)>0\\ 0, & \text{其它}\end{cases} \quad (4)$$

$$z-score=(a-\mu A)/\sigma A \quad (5)$$

式中：a 为对应的空间属性 A 的取值；μ 为 A 在 F 中的的期望值；σ 为标准差。Z 值用于衡量某个属性高出或者低于其标准值的程度，备选模式 B 的组成属性值的平均 Z 值作为兴趣度函数 $i(B,o)$。兴趣度函数越大说明该模式的兴趣度越大，并且该值的大小直接指示该空间对象符合备选模式的程度。而当 B 中某个属性值 Z 值等于0时，表示该实例违背了模式 B 的定义，则其兴趣度值为0。

3.2 G_i^* 统计量及关键计算流程

在空间数据集中，高值要素往往容易引起注意，但可能不是具有显著统计学意义的热点。要成为具有显著统计意义的热点，要素应具有高值，且被其他同样具有高值的要素所包围。G_i^* 统计量是被广泛用于寻找空间数据集热点的局部统计量，G_i^* 统计值越高，高值(热点)的聚类就越紧密；反之，G_i^* 统计值低值，低值(冷点)就越紧密[10]。在对每个空间对象的属性计算 G_i^* 值时，周边对象的要素值是影响 G_i^* 值的关键，而对空间点数据而言，周边对象要素值的权重受到其与计算对象的距离影响，常用权重计算方式有反向距离权重法，临界距离法等，都需预设决定邻近的范围。如果使用k近邻方法，难以确定合适的搜索圆半径，并且无法保证各个象限都有邻域对象；而使用固定距离法，由于空间数据集的多尺度特性，距离阈值的设定对结果的影响非常大，若仅以距离为基础定义权重，微小变化可

能会使选点结果差别很大。Delaunay三角测量里自然邻域能够保证空间对象点的每个象限都有邻域对象[19]，同时其避免了k近邻法k值确定和固定距离法距离阈值确定的缺点，对所有空间对象点具有自适应确定邻域对象的特性。

3.3 Voronoi邻域及Voronoi邻域权重矩阵

Voronoi多边形又称泰森多边形，由荷兰气候学家Thiessen提出的一种根据离散分布的气象站的降雨量来计算平均降雨量的方法。其通过将所有相邻气象站连成三角形，然后做这些三角形各边的垂直平分线，进而得到泰森多边形。泰森多边形的特性[20]有：每个泰森多边形内仅含有一个离散点数据；泰森多边形内的点到相应离散点的距离最近；位于泰森多边形边上的点到其两边的离散点的距离相等。

（1）定义Voronoi邻域

对空间数据集F中的空间对象生成Voronoi图，$Vr(o_i)$是空间对象o_i所在的Voronoi多边形，定义空间对象的Voronoi邻域为式（6）所示，即与空间对象o_i所在的Voronoi多边形有共享边的其他Voronoi多边形中的空间对象。

$$Pv(o_i)=\{o_j, Vr(o_i)\cap Vr(o_j)\text{且}i\neq j\} \tag{6}$$

图1展示了空间对象A的自适应Voronoi邻域。由于其具有相邻Voronoi多边形的空间对象确定，而空间对象的Voronoi多边形在空间对象分布确定后即可确定，故无需预设定值。

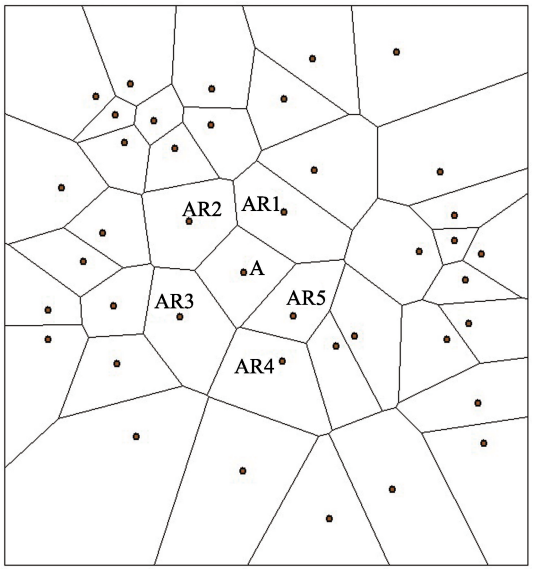


图1　Voronoi自适应邻域

Fig.1　Proximity area based on Voronoi diagrams

（2）Voronoi邻域权重矩阵

$$\boldsymbol{W}_{vr}=\begin{vmatrix} w_{00} & w_{01} & \cdots & w_{0n} \\ \vdots & & & \vdots \\ w_{n0} & w_{n01} & \cdots & w_{nn} \end{vmatrix} \tag{7}$$

其中，$w_{ij}=\begin{cases}1/d, & o_j\in Pv(o_i)\\ 0, & \text{其它}\end{cases}$。

当空间对象o_j与o_i互为Voronoi邻域时，W_{ij}权重为$1/d$，在计算G_i^*值时，d为o_i的相邻对象总数加1。

基于泰森多边形的特性，本文采用Voronoi邻域来定义空间对象的邻域，并定义了Voronoi邻域权重矩阵。与k近邻邻域和预设定距离阈值的圆形邻域不一样，在泰森多边形生成后，每个空间对象的邻域及其数量就已经确定，无需事先设定k值或其他阈值。另外，泰森多边形设计之初是为了获取空间离散点（如气象站、传感器、土壤样点等）空间数据的均值，因此特别适用于采样离散点数据。

4　SLCPMA算法

本文设计的自适应局部空间同位模式查找法（Self-adaptive Local Colocation Pattern Mining Algorithm，SLCPMA）的流程框架如图2所示，具体分为5个步骤：

（1）计算所有可能同位模式P中所有属性的Z值。

（2）根据P，生成选定的候选模式，并根据步骤（1）中的结果和式（1）计算所有候选模式的兴趣度值。

（3）对原始空间数据集生成Voronoi多边形，计算并生成Voronoi邻域矩阵。

（4）使用Voronoi邻域矩阵对所有备选模式的兴趣度值计算G_i^*统计量。

（5）针对所有候选模式，过滤给定置信度水平G_i^*统计量值小于设定值且兴趣度值大于设定阈值的模式的离散对象点；对所有剩下的G_i^*统计量达

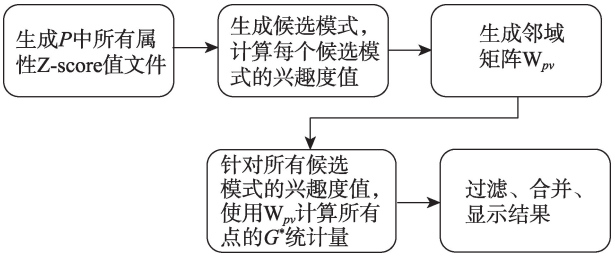


图2　自适应区域空间同位模式算法框架图

Fig.2　Framework of SLCPMA

到阈值的空间对象及其代表的模式进行合并、输出。

根据以上算法框架设计了SLCPMA算法，其详细描述和算法如下所示。

SLCPMA算法

Input

$F=\{o_1,o_2,\cdots o_n\}$, 空间离散点数据集

$S=\{S_1,S_2,\cdots,S_m\}$, 空间属性集

$N=\{A_1,A_2,\cdots,A_h\}$, 非空间属性集

P, 所有可能模式

$B=\{B_1,B_2,\cdots,B_k\}$, 从 P 中选取的候选模式集

α, 置信度阈值

g, G_i^* 统计量阈值

θ, 候选模式的兴趣度阈值

Output

在置信度水平阈值下，G_i^* 值和兴趣度达到设定阈值的离散点的空间同位模式及其范围。

Variables

W_{vr}, 空间邻域矩阵

V_r, F 的Voronoi图

$I_i(B_j)$, 空间离散点 o_i 上模式 B_j 的兴趣度值

$I(B)$, 所有空间对象的备选模式的兴趣度阈值表

$G^*_{i\ ij}$, 空间对象 o_i，对应的模式 B_j 的兴趣度阈值的 G_i^* 统计量值

$G_i^*(B)$, 空间数据 F 的 $G^*_{i\ ij}$ 集合

Procedure

Begin

1 $\boldsymbol{P}$ = GenerateAllPattrens(N);

2 $\boldsymbol{Vr}$ = CreateVoronoiPolygon(F);

3 $\boldsymbol{W_{vr}}$ = GenerageVoronoiProximityWeightMatrix(Vr);

4 for o_i in F do

5 for B_j in B do

6 $\boldsymbol{I_i(B_j)}$ =CaculatorInterstingness (B_j);

7 Insert $\boldsymbol{I_i(B_j)}$ into $\boldsymbol{I(B)}$;

8 end do

9 end do

10 for B_i in B do

11 for o_j in F do

12 $\boldsymbol{G_i}^*_{ij}$ =CaculatorGstarValue ($\boldsymbol{I(B)}$, $\boldsymbol{W_{vr}}$)

13 end do

14 end do

15 FilterAndMergePatternsGStar($G_i^*(B)$, α, θ, g)

End

算法中步骤1–15的含义如下：

1 GenerateAllPattrens用于生成所有的可能的模式，例如[A_High，B_Low，C_High]，其中A_High（或者A_H）表示A属性高于均值的模式。显然，一个属性的高值模式与低值模式不能同时出现在一个同位候选模式中；

2 CreateVoronoiPolygon函数对整个空间点数聚集创建Voronoi图；

3 GenerageVoronoiProximityWeightMatrix根据Voronoi图生成Voronoi邻域矩阵；

4–9 计算并生成所有空间对象的备选模式的兴趣度阈值表 $I(B)$；

6 CaculatorInterstingness (B_j) 根据函数定义计算空间对象的备选模式的兴趣度值；

10–14 计算所有空间对象的所有候选模式的兴趣度值得 G_i^* 统计量；

15 FilterAndMergePatternsGStar ($\boldsymbol{G_i^*(B)}$, $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, g) 根据输入的置信度水平和兴趣度阈值对10–14中计算出来的 G_i^* 值进行过滤、合并，并输出所有的候选模式的计算结果。

5 实验与应用

本文利用.NET平台及C#语言实现了SLCPMA算法，并采用2组真实数据进行了实验（表1）。一组数据来源于美国堪萨斯州水发展委员会（Texas Water Development Board）的地下水数据库（Groundwater Database，GWDB①）的水井所在蓄水层重要化学元素含量及水井深度等数据，并对其进行加工处理；

表1 实验数据集

Tab.1 Datasets used in the experiments

数据集	内容	数据量
GWDB水数据	砷（As），钼（Mo），钒（V），硼（Bo），氟（Fl），二氧化硅（Si），氯化物（Cl），硫酸盐（SO4），总溶解固体（TDS）和水井深度（WD）	1655
长汀耕地地力数据及烟草适宜性评价结果	烟草适应性评价得分（Score），有机质（Organic），PH，碱解氮（N），有效磷（P），速效钾（L）	475

① Texas Water Development Board Groundwater Database (GWDB), http://www.twdb.texas.gov/groundwater/data/index.asp

另一组来源于福建省长汀县耕地力调查样点数据。

5.1 对比实验

由于Ecik等的研究对象与本文类似，都是针对连续变量的局部空间同位模式的发现，因此，本实验使用了Ecik等研究中的数据和区域[6,9]，以该研究中挖掘得到的部分模式集作为本文算法候选集的一部分进行挖掘，对本文算法挖掘得到的模式和模式的影响区域进行对比研究。

图3依次列出了部分候选模式的挖掘结果的叠加效果，阈值设定为 ($\alpha=0.05, g=1.96, \theta=1.5$)，即在置信度为95%水平下且模式的兴趣度值（平均Z值）大于1.5的同位模式热点区域，图4单独列出了各个模式的区域。实验结果中，挖掘出的空间同位模式区域兴趣度高值聚集的区域与Ecik等的挖掘结果有一定的重叠[9]，如图4(a)的[As_High Bo_High Cl_High TDS_High]所在范围，在空间上与Eick等的实验在不同的参数设置下得到的模式在区域上重叠，其中模式中的后缀“High”表示该元素的含量相对均值较高。Eick等的运算结果受参数设定影响变化较大。本文得到的[As_High Bo_ High Cl_High TDS_High]模式区域面积较Eick的结果小[9]，然而，本文采用的方法能够使用统计显著性意义对得到的局部同位模式进行解释，并且图4(a)-(b)的同一个区域中，在满足了所有的设定阈值条件下，不同候选同位模式仍可被挖掘出来，避免了Eick等研究中使用全局的目标函数而导致一个区域只选出一个模式，从而产生可能遗漏的问题。

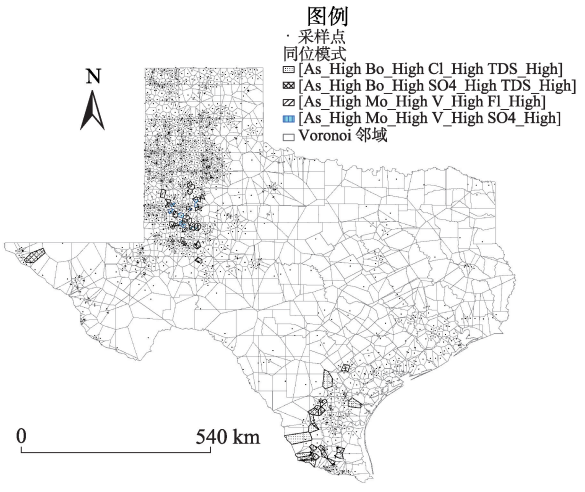


图3　Texas-GWDB数据同位模式挖掘结果
（$\alpha=0.05, g=1.96, \theta=1.5$）

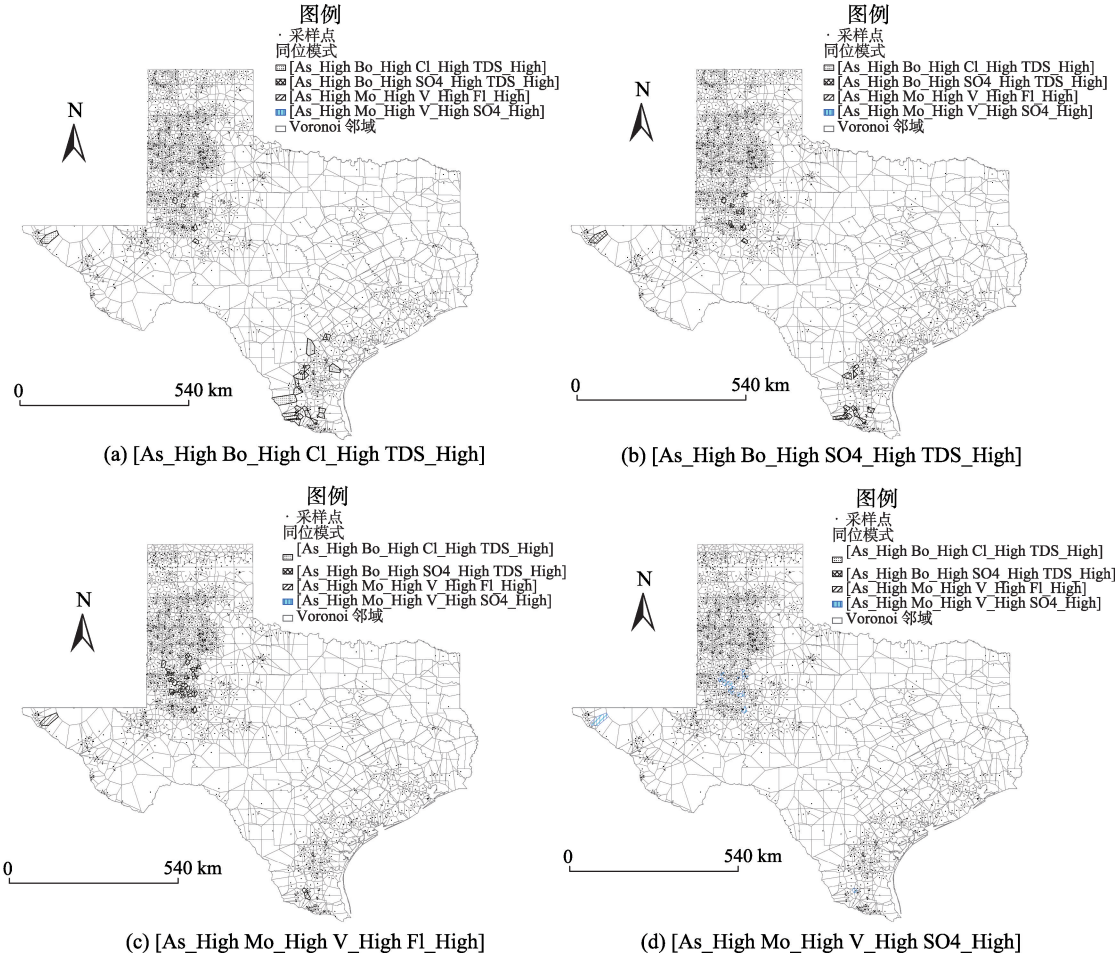
Fig.3　Experimental result of co-location mining for Texas-GWDB datasets ($\alpha=0.05, g=1.96, \theta=1.5$)

5.2 耕地地力调查应用实验

利用测土配方施肥调查数据开展耕地地力调查工作，是测土配方施肥补贴项目的一项重要内容，是摸清农村耕地资源状况、提高耕地利用效率、促进现代农业发展的重要基础工作。土壤养分和立地条件等因素使土壤耕地地力的主要评价指标，通过层次分析法构建相应评价指标体系即可实现耕地力及农作物适宜性的评价。

图5对长汀县耕地地力调查点养分数据及其基础上得到的烟草适宜评价得分进行了挖掘。长汀县耕地地力调查与质量评价的野外布点和采样根据《全国耕地地力调查与质量评价技术规程》和《福建省耕地地力调查与质量评价实施方案》的规定和要求，遵循以下原则：(1)布点要有广泛的代表性，要考虑土种类型、分布、地形地貌及种植作物的种类等；(2)原则上必须保证乡镇范围内每一土种类型至少布设一个样点，同时样点的分布应尽可能均匀；(3)耕地地力调查取样点应与测土配方施肥采样点相衔接；(4)原则上在第二次土壤普查取样点上布设样点；(5)采集样品点所在的评价单元应具有代表性，避免各种非调查因素的影响，选择具有代表性的一个农户的同一田块进行随机多点取样。

实验选择了烟草适宜性评价得分低作为开头的同位模式（Scroe_L）为候选模式（实验中H结尾表示该属性高）。实验结果选取了G_i^*值前五，且$\theta=2, \alpha=0.05$的同位模式。由图5可见，模式[Score_L,PH_L,N_H,P_H,K_H]（即适宜性低，PH低，碱解氮高，有效磷高，速效钾高）所在区域及其邻域聚集了评价得分低于均值，但是主要养分（碱解氮，有效磷，速效钾）远高于总体均值的样点。这是由于烟草的适宜性评价得分的评价指标除了土壤养分外还有其他重要指标，如灌溉条件、立地条件等。经查阅长汀县耕地地力报告发现，该区域所在的长汀县庵杰乡耕地面积495.75 hm^2，其中有效硼缺乏面积495 hm^2，干旱限制占地411.14 hm^2，酸性限制432.31 hm^{2}区域。另外，从长汀县不同坡度耕地面积的乡镇分布来看，15~25°的耕地主要分布于庵杰乡、红山乡和铁长乡等乡镇，合计面积386.83 hm^2，占全县15~25°耕地总面积的44.96%。从不同坡度耕地面积的乡镇分布来看，大于25°的耕地主要分布于庵杰乡和铁长乡等乡镇，合计面积19.15 hm^2，占全县大于25°耕地总面积的78.20%。可见，这些因素严



(a) [As_High Bo_High Cl_High TDS_High] (b) [As_High Bo_High SO4_High TDS_High]

(c) [As_High Mo_High V_High Fl_High] (d) [As_High Mo_High V_High SO4_High]

图4 Texas GWDB同位模式及其区域

Fig.4 The co-location patterns and region discovered through the experiment

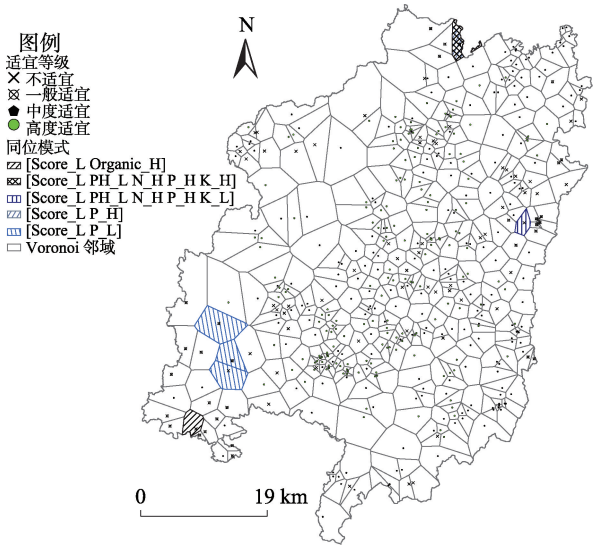


图5 长汀县烟草适宜性评价及 G_i^* 值前五的同位模式 ($\alpha=0.05, g=1.96, \theta=2$)

Fig.5 Land suitability evaluation results for tobacco planting in Chanting county and its co-location patterns with the top 5 G_i^* scores ($\alpha=0.05, g=1.96, \theta=2$)

重限制了庵杰乡烟草的种植适宜性。综上可见,挖掘的结果为有针对性地提高该地区的其他耕地条件水平提供了参考。

6 结论

本文提出的SLCPMA算法通过定义空间数据集连续变量的兴趣度函数,同位模式指示器函数及Voronoi邻域矩阵,进而计算空间对象的同位模式兴趣度函数的 G_i^* 统计量,最终实现区域空间同位模式的挖掘。实验结果显示,本文提出的SLCPMA算法框架能够在查找区域模式的同时,确定模式所在的聚集点,无需预先设定邻域的范围。同时,通过对长汀县耕地地力样点数据及烟草适宜性评价数据的实验发现,该方法对耕地地力调查研究具有一定的指导意义。今后,将进一步完善算法以提高其可用性及效率,并拓展算法的应用范围。

参考文献（References）：

[1] Ding W, Eick C F, Wang J, *et al*. A framework for regional association rule mining in spatial datasets[C]. The 6th IEEE International Conference on Data Mining (ICDM), 2006:851-856.

[2] Shekhar S, Huang Y. Co-location rules mining: A summary of results[C]. The 7th International Symposium on Spatio and Temporal Database (SSTD), New York, 2001.

[3] Yoo J, Shekhar S. A partial join approach for mining co-location patterns[C]. The 12nd Annual ACM International Workshop on Geographic Information Systems (ACM-GIS), Washington D C, USA, 2004.

[4] Yoo J, Shekhar S, Celik M, *et al*. A join-less approach for co-location pattern mining: A summary of results[C]. The 5th IEEE International Conference on Data Mining (ICDM' 05), Houston, USA, 2005.

[5] Xiong H, Shekhar S, Huang Y, *et al*. A framework for discovering co-location patterns in data sets with extended spatial objects[C]. 2004 SIAM International Conference on Data Mining (SDM), 2004.

[6] Barua S, Sander J. Mining statistically significant co-location and segregation patterns[J]. IEEE Transactions on Knowledge & Data Engineering, 2014,26(5):1-1.

[7] Bagherjeiran A, Celepcikay O U, Jiamthapthaksin R, *et al*. COUGAR^2: An open source machine learning and data mining development platform[EB/OL]. http://www.cs.uh.edu/ceick/kdd/BCJCRLTE09.pdf, 2010.

[8] Qian F, Chiew K, He Q, *et al*. Mining regional co-location patterns with kNNG[J]. Journal of Intelligent Information Systems, 2013,42(3):485-505.

[9] Eick C F, Parmar R, Ding W, *et al*. Finding regional co-location patterns for sets of continuous variables in spatial datasets[C]. Proceedings of 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2008.

[10] Getis A, Ord, J K. Local Spatial statistics: An overview[A]. In: Spatial analysis: modeling in a GIS environment[M]. Cambridge, UK: GeoInformation International, 1996:261-277.

[11] Ord J K, Getis A. Local spatial autocorrelation statistics: distributional issues and an application[J]. Geographical Analysis, 1995,27(4):286-306.

[12] 边馥苓,万幼.k-邻近空间关系下的空间同位模式挖掘算法[J].武汉大学学报·信息科学版,2009,34(3):331-334. [Bian F L, Wan Y. A novel spatial co-location pattern mining algorithm based on k-nearest feature relationship[J]. Geomatics and Information Science of Wuhan University, 2009,34(3):331-334.]

[13] Koperski K, Han J. Discovery of spatial association rules in geographic information databases[A]. In: Advances in Spatial Databases[M]. Berlin, Germany: Springer Berlin Heidelberg, 1995:47-66.

[14] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]. The 20th International Conference on Very Large Databases, Santiago, Chile, 1994.

[15] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[C]. 2000 ACM SIGMOD International Conference on Management of Data, 2000:1-12.

[16] Celik M, Kang J M, Shekhar S. Zonal co-location pattern discovery with dynamic parameters[C]. Proceedings of the 2007 Seventh IEEE International Conference on Data Mining. IEEE Computer Society, 2007:433-438.

[17] Ng R. Efficient and effective clustering methods for spatial data mining[J]. Proceedings of the VLDB Conference, 1994,88(9):144-155.

[18] Huang Y, Zhang P. On the Relationships between Clustering and Spatial Co-location Pattern Mining[C]. 18th IEEE International Conference on Tools with Artificial Intelligence, 2006:513-522.

[19] 张祖勋,张剑清.数字摄影测量学[M].武汉:武汉测绘科技大学出版社,1996. [Zhang Z X, Zhang J Q. Digital photogrammetry[M]. Wuhang: Wuhan Technical University of Surveying and Mapping Press,1996.]

[20] Watson D. Spatial tessellations: concepts and applications of voronoi diagrams: by Atsuyuki Okabe, Barry Boots, and Kokichi Sugihara, 1992, John Wiley & Sons, New York, 532 p., ISBN 0 471 93430 5, US $112.00[J]. Computers & Geosciences, 1993,19:1209-1210.