

空间扫描统计量方法中候选聚集区域生成的快速算法

李小洲¹, 王劲峰^{2*}

(1. 武汉科技大学医学院公共卫生学院, 武汉 430065;

2. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室, 北京 100101)

摘要: 空间扫描统计量方法是公共卫生监测领域应用非常广泛的空间聚集探测快速算法。其利用传染病监测数据可探测到病例异常增多的局部区域, 对可能的传染病暴发出早期预警。候选聚集区域的预先生成是该方法的一个关键步骤。将现有的候选聚集区域生成方法应用到包含子区域较多的大区域时, 可能导致大量候选聚集区域的遗漏, 影响探测结果的准确性; 或可能生成大量重复的候选聚集区域, 导致随后空间扫描计算时间的延长。本文在原有候选聚集区域生成方法的基础上, 提出了一种新的快速算法。它以格网点间隔的优化选择, 可减少可能对候选聚集区域的遗漏; 同时, 基于多重排序算法可在较短的时间之内, 删除掉原始候选聚集区域集合中的大量重复。通过山东省西南部 608 个乡镇点的候选聚集区域生成测试, 改进的方法可减少候选聚集区域的遗漏, 并在较短的时间内删除掉所有的重复候选聚集。

关键词: 空间扫描统计量; 候选聚集区域; 格网; 多重排序

DOI: 10.3724/SP.J.1047.2013.00505

1 引言

空间扫描统计量方法是由哈佛大学 kuldorff 教授提出的一种空间聚集探测方法^[1-2]。它已成为公共卫生监测领域的经典方法。该方法以研究区域内每一个最小单元子区域的风险人口数和监测病例数, 在此空间分布数据中探测是否存在患病风险统计显著性高于其他区域的病例聚集区域。该方法现已有了多个扩展模型^[3-9]。

在传染病监测的实际工作中, 空间扫描统计量方法已得到了广泛的应用^[10-11]。我国的传染病自动预警系统就应用了该方法, 对可能的传染病暴发事件做出早期预警^[12-13]。该系统每天都需要针对收集到的全国各乡镇传染病监测数据, 按不同的区域分别进行空间扫描探测, 以寻找可能的病例聚集区域。因此, 该方法针对大量数据时的运算效率, 在一定程度上决定了传染病自动预警系统是否能及时地发挥作用。

运用空间扫描统计量方法时, 对于全部的最小单元子区域, 需要预先生成所有的候选聚集区域, 并构成一个集合, 此集合的快速准确生成是空间扫描统计量方法的一个关键步骤。首先, 最终探测到的聚集区域是一定来自于此候选聚集区域集合, 该集合是否完备, 是否存在遗漏的候选聚集区域, 决定了最终探测结果的准确性; 其次, 为了计算探测到的最大可能聚集区域的统计显著性 P 值, 需要进行大量的(大约 5000 左右)蒙特卡洛模拟运算, 每一次模拟都需要针对每一个候选聚集区域分别计算其统计量值。此时, 该集合中候选聚集区域的数目, 在很大程度上决定了该方法执行的时间复杂度。因此, 如果该集合中存在大量的重复候选聚集, 将严重影响空间扫描统计量方法的时间效率。

现有的生成候选聚集区域集合的方法虽然简单直接, 但应用到较大的研究区域时, 可能会因为参数选择不当, 遗漏部分可能的候选聚集区域, 导致最终探测结果的准确性降低; 同时, 该方法没有

收稿日期: 2013-01-14; 修回日期: 2013-03-20.

基金项目: 国家科技重大专项子课题“艾滋病和病毒性肝炎等重大传染病防治/传染病病原谱时空信息分析和预报”(2012ZX10004-201); 卫生行业科研专项项目“传染病时空预警模型及关键参数研究”(201202006)。

作者简介: 李小洲(1974-), 男, 湖北麻城人, 讲师, 主要研究方向为空间数据分析及空间流行病学。E-mail: lixiaozhou@wust.edu.cn

*通讯作者: 王劲峰(1965-), 男, 上海人, 博士生导师, 研究员, 主要研究方向为空间数据分析及空间统计。E-mail: wangjif@lreis.ac.cn

考虑到删除候选聚集区域集合中的大量重复候选聚集,会导致算法的计算时间大大延长。本文针对原有方法进行了修正,避免了候选聚集区域的遗漏,并提出了一种快速的重复候选聚集区域删除算法,它通过多重排序操作,可以在较短的时间内删除候选聚集区域集合中的重复子区域。通过在理论上计算该重复删除方法的时间复杂度,以及通过山东省西南部608个乡镇的空间分布数据作为实际案例,说明了本文所提出新方法的优越性。

2 空间扫描统计量方法中候选聚集区域的生成

应用空间扫描统计量方法时,整个研究区域完全划分为若干个最小单元子区域。在具体的公共卫生监测应用中,整个的监测区域可以按照地区、县或者乡镇的行政区域范围进行完全划分。然后,对于每一个最小单元子区域分别得到一个代表点,此代表点可以是子区域的质心点,或者是该子区域中人口的分布重心点。候选聚集区域是由一系列在整个研究区域上滑动的,不同圆心和不同半径的搜索圆来生成的。若某个搜索圆恰好包含了若干个最小单元子区域的代表点,则这些代表点就构成了一个候选聚集区域。最大候选聚集区域的上限值需要用户预先指定,它可以是候选聚集区域内所有最小单元子区域的患病风险人口总数,或者是候选聚集区域内所有最小单元子区域的个数。

空间扫描统计量方法的提出者,哈佛大学kull-dorff教授曾给出了通过预先建立的格网点来找到全部的候选聚集区域的算法^[14]。现将其基本过程简述如下:

(1)为了保证搜索圆能在整个研究区域范围内移动,在整个研究区域内建立规则的格网点,作为搜索圆的圆心。假设建立的格网点共有 n 个,分别为 G_1 到 G_n 。

(2)对于每一个格网点,将所有最小单元子区域的代表点,按照与当前格网点的距离从近到远进行排序。这样。每一个格网点就得到了一个所有子区域代表点的序列。

假设整个研究区域内共有 m 个最小单元子区域,即一共有 m 个子区域代表点,分别为 P_1 到 P_m 。对于格网点 G_i ,其对应的子区域代表点序列为 $P_{s_{i,1}}, P_{s_{i,2}}, \dots, P_{s_{i,m}}$,其中, $s_{i,1}$ 为距离格网点 G_i 最

近的那个子区域代表点的下标序号, $s_{i,m}$ 为距离格网点 G_i 最远的那个子区域代表点的下标序号。

(3)以每一个格网点作为搜索圆的圆心,不断扩大搜索圆的半径,根据其包含的最小单元子区域代表点,就可以生成不同的候选聚集区域。

例如,对于格网点 G_i ,通过第2步骤已知该圆心与所有子区域代表点从近到远的序列。则可以生成的候选聚集区域应该为该序列的前面若干个子区域代表点的集合,即分别是 $(P_{s_{i,1}})$ 、 $(P_{s_{i,1}}, P_{s_{i,2}})$ 、 \dots 、 $(P_{s_{i,1}}, P_{s_{i,2}}, \dots, P_{s_{i,k}})$ 等 k 个候选聚集区域, k 的大小由预先设定的最大候选聚集区域的上限来确定。

这样,每一个格网点就生成了 k 个候选聚集区域,可以在生成候选聚集区域的同时分别计算该候选聚集区域的似然比统计量值。对于全部的 n 个格网点,一共可以生成 $n \times k$ 个候选聚集区域。

3 大数据集中空间扫描统计量方法存在的问题

空间扫描统计量方法在公共卫生监测领域也得到了广泛的应用。例如,在我国的国家传染病自动预警系统中,依地区分病种进行了空间聚集探测。

在大数据集上应用空间扫描统计量方法时,探测到聚集区域的准确性,以及算法执行的时间效率显得尤其重要。解决这个问题的关键在于生成完备的候选聚集区域,不要遗漏可能的候选聚集区域;并尽可能地控制所生成的候选聚集区域的数目,即预先删除掉重复的候选聚集区域。其在处理最小单元子区域不是太多时,可快速直接地得到所有的候选聚集区域,并依次计算其似然比统计量值。但是,当最小单元子区域的数目较大时,该算法存在以下2个问题:

(1)不合适的格网点可能导致某些候选聚集区域的遗漏。图1与图2有同样的6个最小单元子区域代表点,它们的空间分布位置完全一致,但图2中格网点间隔小于图1。可见,处于研究区域中央成倒三角的3个代表点可以构成一个可能的候选聚集区域。但是,如果限定搜索圆的圆心只能在图1中的格网点上,则不存在这样的一个搜索圆,它仅仅包括了中央的3个子区域代表点而不包括任何其他子区域代表点。只有当搜索圆的圆心限定在图2中

的格网点上时,才存在这样一个搜索圆,恰好包括了这3个代表点。这说明,当格网点的间隔过大时,以此为圆心产生的搜索圆会漏掉某些可能的候选聚集区域。当研究区域包含的子区域代表点较多时,产生的遗漏也会增多,大大影响最终探测结果的准确性。

经上分析可知,只有当格网点的间隔与所有最小单元子区域代表点间的最小距离一致时,才能够保证搜索圆圆心在研究区域上的每一步移动距离,不会大于任意2个子区域代表点间的距离,也就不会有遗漏的候选聚集区域了。

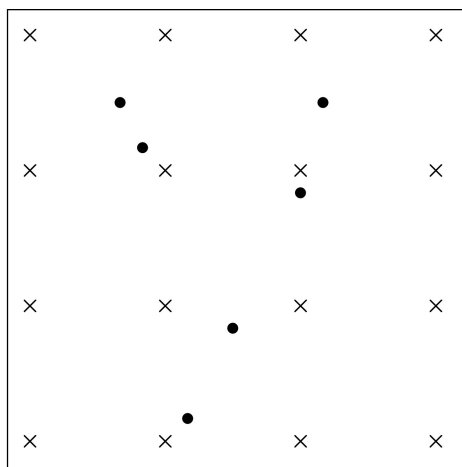


图1 6个子区域代表点与稀疏格网点

Fig.1 Six sub-regions representing points and sparse grid points

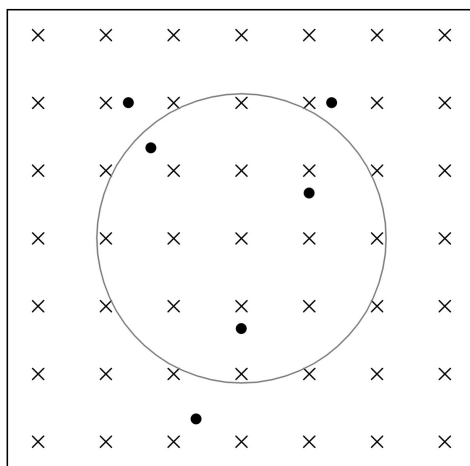


图2 6个子区域代表点与较密格网点

Fig.2 Six sub-regions representing points and intensive grid points

(2)对包含最小单元子区域较多的研究区域,格网点的密度也会相应变大,格网点的数目 n 也较

大。由于每个格网点对应的最大候选聚集区域可以包含的代表点最大数目 k ,与整个研究区域内的代表点数目成比例,一般设为 20%~50%,因此,对于较大的研究区域,按照原始算法所生成的候选聚集区域数目 $n \times k$ 将是一个非常巨大的数字。对于不同圆心不同半径的两个搜索圆,只要它们包含了相同的若干个代表点,就应该认为它们是重复的候选聚集区域。因此,对于较大的研究区域,在初始按照格网点依次生成的候选聚集区域中,包含大量的重复聚集,如果不预先剔除掉重复聚集区域,就会导致大量的重复计算时间,极大影响空间扫描统计量方法的时间效率。

Kulldorff教授的原始算法实现过程中,每通过格网点得到一个候选聚集区域后,马上计算其统计量值,并没有考虑到不同的格网点可能产生重复的候选聚集区域。下面将介绍一种优化的重复候选聚集区域删除算法。

4 优化的重复候选聚集区域删除算法

若按照格网点依次生成候选聚集区域,不同的格网点会产生大量重复的候选聚集区域,则不应在通过格网点生成候选聚集区域时,马上计算其似然统计量值,而应先记录下所有生成的候选聚集区域,再删除其中相互重复的部分,以获得一个较小的候选聚集区域集合,从而减少随后的空间扫描统计量算法的运算时间。

假设,按照格网点依次生成的候选聚集区域数目为 l ,我们已知 $l = n$ (格网点个数) $\times k$ (候选聚集区域上限)。如果想剔除掉这 l 个候选聚集区域中重复的部分,按照简单直接的办法,需要将这 l 个候选聚集区域两两比较,则比较次数的时间复杂度为 $O(l^2)$ 。每一次进行比较的两个候选聚集区域,都可能包含最多 k 个代表点,则每一次比较的时间复杂度为 $O(k^2)$ 。这样,总的时间复杂度为 $O(l^2 \times k^2)$ 。

假设研究区域内全部的最小单元子区域代表点分别编号为 $1, 2, \dots, m$ 。这样,每一个候选聚集区域都可以表示为一个 k 元素的向量,若某候选聚集区域包含不足 k 个代表点,可用一个预先设定的特殊的编号(例如,用一个远大于 m 的编号)来填补空缺位置。这样,全部的 l 个候选聚集区域 k 元向量,就可以构成一个的 $l \times k$ 矩阵,每一行表示一个候选聚集区域。如果其中两行包含了相同的代表点编

号组合,而只是排序方式不同,则认为是重复的候选聚集区域,应该删除其中的一个。

对于一系列的元素删除其重复,首先对其进行排序显然可以大大节省计算时间。这里的每一个元素是一个行向量,应该借用多关键字基数排序^[15]的思想来进行行向量的排序。值得注意的是,这里的任两个行向量是按照集合相同的方式进行比较,即只要包含了相同的元素,不管其排列次序是否一致,都认为是相同。因此,在排序之前,需要对每个行向量进行一次内部的排序操作,便于以后相互的比较。

鉴此分析,本文提出了一个优化的候选聚集区域剔除重复算法。下面以一个简化的例子来演示该算法的执行过程。假设只有2个格网点,一共4个最小单元子区域代表点。按照与格网点1的距离从近到远,所有代表点排序为(4,2,3,1);按照与格网点2的距离排序为(2,4,1,3)。再设定9代表不存在的代表点编号。为了简化问题,便于演示算法,这里没有考虑最大候选聚集区域的上限值。因此,按照格网点1生成的4个候选聚集区域依次分别为{4}、{4,2}、{4,2,3}、{4,2,3,1};按照格网点2生成的4个候选聚集区域依次分别为{2}、{2,4}、{2,4,1}、{2,4,1,3}。这8个候选聚集区域就构成了一个 8×4 的矩阵,如矩阵(1)所示。在不必进行两两比较的情况下,优化的重复剔除算法可以删除其中包含元素相同的行,例如,第2行和第6行,表示了2个同样的候选聚集区域,应该删除其中一行。

$$\begin{pmatrix} 4 & 9 & 9 & 9 \\ 4 & 2 & 9 & 9 \\ 4 & 2 & 3 & 9 \\ 4 & 2 & 3 & 1 \\ 2 & 9 & 9 & 9 \\ 2 & 4 & 9 & 9 \\ 2 & 4 & 1 & 9 \\ 2 & 4 & 1 & 3 \end{pmatrix} \quad (1)$$

$$\begin{pmatrix} 4 & 9 & 9 & 9 \\ 2 & 4 & 9 & 9 \\ 2 & 3 & 4 & 9 \\ 1 & 2 & 3 & 4 \\ 2 & 9 & 9 & 9 \\ 2 & 4 & 9 & 9 \\ 1 & 2 & 4 & 9 \\ 1 & 2 & 3 & 4 \end{pmatrix} \quad (2)$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 4 & 9 & 9 & 9 \\ 2 & 4 & 9 & 9 \\ 2 & 3 & 4 & 9 \\ 2 & 9 & 9 & 9 \\ 2 & 4 & 9 & 9 \\ 1 & 2 & 4 & 9 \end{pmatrix} \quad (3)$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 9 \\ 1 & 2 & 4 & 9 \\ 4 & 9 & 9 & 9 \\ 2 & 4 & 9 & 9 \\ 2 & 9 & 9 & 9 \\ 2 & 4 & 9 & 9 \end{pmatrix} \quad (4)$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 9 \\ 2 & 3 & 4 & 9 \\ 2 & 4 & 9 & 9 \\ 2 & 4 & 9 & 9 \\ 4 & 9 & 9 & 9 \\ 2 & 9 & 9 & 9 \end{pmatrix} \quad (5)$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 9 \\ 2 & 3 & 4 & 9 \\ 2 & 4 & 9 & 9 \\ 2 & 4 & 9 & 9 \\ 2 & 9 & 9 & 9 \\ 4 & 9 & 9 & 9 \end{pmatrix} \quad (6)$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 9 \\ 2 & 3 & 4 & 9 \\ 2 & 4 & 9 & 9 \\ 2 & 9 & 9 & 9 \\ 4 & 9 & 9 & 9 \end{pmatrix} \quad (7)$$

优化的候选聚集区域剔除重复算法主要包括以下3个步骤:

(1)对原始矩阵(1)的每一个行向量,分别按照代表点编号从小到大次序进行排序,可以得到矩阵(2)。表示相同候选聚集区域的两行(例如,第2行和第6行)也成为完全相同的两个行向量,便于以后比较和删除其中一行。

(2)需要对矩阵(2)的8个行向量在矩阵中的排列顺序进行重新排列。按照所有行向量的第4个元素,即矩阵的最后一列,从小到大的次序对所有行向量的行号进行重新排列。对于排序元素相同的行向量,应保持其相对位置不变。按第4个元素排

序,得到矩阵(3);再按第3个元素排序,得到矩阵(4);再按第2个元素排序,得到矩阵(5);最后按第1个元素排序,得到矩阵(6)。此时,完全相同的行向量,即相同的候选聚集区域,已经处在相邻的行号。

(3)对于矩阵(6),从第2个行向量开始依次扫描到最后一个行向量。如果当前的行向量与前一个行向量相同,则删除当前的行向量。最后,获得已删除所有重复行向量的结果矩阵(7),该矩阵的每一行对应一个候选聚集区域。

此外,优化重复删除算法对于 $l \times k$ 矩阵的时间复杂度的分析,主要包括3个步骤:(1)需要进行 l 次对 k 维行向量的排序,总的时间复杂度为 $O(l \times k \times \log(k))$; (2)一共需要进行 k 次对 l 维列向量的排序,总的时间复杂度为 $O(k \times l \times \log(l))$; (3)一共需要进行 l 次 k 维行向量的比较,总的时间复杂度为 $O(l \times k)$ 。该算法总的时间复杂度为 $O(k \times l \times (\log(l) + \log(k) + 1))$ 。由上分析可知,如直接将生成的候选聚集区域两两比较删除重复,时间复杂度将为 $O(l^2 \times k^2)$ 。因此,新的优化重复删除算法具有更好的时间复杂度。

5 山东省乡镇代表点的候选聚集区域生成测试

山东省西南部的菏泽市、临沂市、枣庄市、济宁市和泰安市,一共有608个乡镇代表点(见图3),本

测试将以此为研究区域,生成所有的候选聚集区域。设定最大的候选聚集区域上限为20%的乡镇代表点,即本测试中候选聚集区域最多可以包含121个乡镇代表点。本次测试的计算机实验平台为DELL-Vostro 270s。

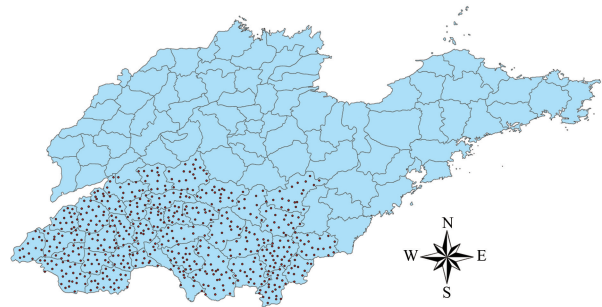


图3 山东省西南部的608个乡镇代表点
Fig.3 Representing points of 608 counties in southwest Shandong Province

该研究区域内所有乡镇代表点间的最短距离大约为1000m。以此为基础,A测试方案选定两倍最短距离2000m为搜索圆圆心格网点的间距,B方案和C方案选定一倍最短距离1000m为搜索圆圆心格网点的间距;D方案选定0.5倍最短距离500m为搜索圆圆心格网点的间距。4个方案对应的格网点数目分别如表1第3列所示。按照原始的候选聚集区域生成算法,对于每一个搜索圆圆心格网点都可以产生121个候选聚集区域,4个方案产生的原始候选聚集区域数目分别如表1第4列所示。

表1 测试方案比较
Tab.1 Comparison of the testing schemes

测试方案	格网点间隔 (m)	格网点数目	候选聚集区域数目 (原始)	重复删除算法	候选聚集区域数目 (删除重复后)	计算时间 (s)
A	2000	16 600	2 008 600	两两比较	1 974 137	1275
B	1000	66 400	8 034 400	两两比较	2 322 164	3265
C	1000	66 400	8 034 400	优化算法	2 322 164	265
D	500	265 600	32 137 600	优化算法	2 322 853	678

这样生成的候选聚集区域中包含大量的重复,如果直接应用到空间扫描统计量方法中,重复的候选聚集区域将浪费掉大量的时间开销,需要先删除其中的重复候选聚集区域。方案A与方案B选择简单直接的两两比较删除重复算法,方案C与方案D选择本文中提出的优化重复删除算法。每一种方案所最终得到的,去除重复后候选聚集区域数

目,以及耗费的计算时间,都展示在表1的最后两行。

可以看出,对于A方案,当搜索圆圆心格网点的间隔两倍于乡镇代表点间的最短距离时,所得到的删除重复后的候选聚集区域数目显著少于其他3种方案,说明遗漏掉了不少可能的候选聚集区域。对于D方案,格网点的间距过小,只有乡镇代表点

间最短距离的一半。虽然按照格网点依次产生的原始候选聚集区域数目较大,但删除重复后的数目与B、C方案几乎相同。这说明,对于B、C方案,当格网点间距等于代表点间最短距离时,已产生几乎所有的候选聚集区域。

对于应用优化后重复删除算法的后两种方案,在计算时间上显著好于应用简单两两比较重复删除算法的前两种方案。更加重要的是不管对于哪种方案,删除重复后的候选聚集区域数目,都大大少于初始得到的候选聚集区域数目,这样可在之后空间扫描统计量方法中极大节省计算的时间开销。

6 结论

当空间扫描统计量方法应用于包含最小单元子区域较多的研究区域时,会因搜索圆圆心格网点选择不当,而遗漏掉一些可能的候选聚集区域;同时,按照格网点产生的初始候选聚集区域中会存在大量的重复候选聚集区域。本文在原有的候选聚集区域生成算法的基础上,提出了一个合适的格网点间距选择方案,与子区域代表点间最短距离一致,并提出了一个优化的重复候选聚集区域删除算法。通过对山东省包含643个乡镇代表点的研究区域上生成候选聚集区域的测试,证明了本文提出格网点间距设置方案,以及优化的重复候选聚集区域删除算法,既不会遗漏掉可能的候选聚集区域,也可以在较短的时间内,删除初始生成的候选聚集区域中的大量重复。同时,应用此方法可以生成一个准确完备,且不含有重复的候选聚集区域集合,保证在下一步的空间扫描统计量方法运算中,得到准确的探测结果,大大减少运算时间。

参考文献:

- [1] Kulldorff M, Nagarwalla N. Spatial disease clusters: Detection and Inference[J]. *Statistics in Medicine*, 1995(14): 799-810.
- [2] Kulldorff M. A spatial scan statistic[J]. *Communications in Statistics: Theory and Methods*, 1997(26):1481-1496.
- [3] Lawson A B, Kleinman K. Spatial and syndromic surveillance for public health[M]. New York: Wiley, 2005, 115-131.
- [4] Pfeiffer D U, Robinson T P, Stevenson M, *et al.* Spatial Analysis in Epidemiology[M]. Oxford: Oxford University Press, 2008, 51-56.
- [5] Jung I, Kulldorff M, Klassen A. A spatial scan statistic for ordinal data[J]. *Statistics in Medicine*, 2007(26): 1594-1607.
- [6] Huang L, Tiwari R, Tiwari R, *et al.* Weighted normal spatial scan statistic for heterogeneous population data[J]. *Journal of the American Statistical Association*, 2009(32): 1034-1042.
- [7] Kulldorff M, Mostashari F, Duczmal L, *et al.* Multivariate spatial scan statistics for disease surveillance[J]. *Statistics in Medicine*, 2007(26):1824-1833.
- [8] Li X Z, Wang J F, Yang W Z, *et al.* A spatial scan statistic for nonisotropic two-level risk cluster[J]. *Statistics in Medicine*, 2012,31(2):177-187.
- [9] Li X Z, Wang J F, Yang W Z, *et al.* A spatial scan statistic for multiple clusters[J]. *Mathematical Biosciences*, 2011, 233(2):135-142.
- [10] Mostashari F, Kulldorff M, Hartman J J, *et al.* Dead bird clustering: A potential early warning system for West Nile virus activity[J]. *Emerging Infectious Diseases*, 2003 (9):641-646.
- [11] Ghebreyesus T A, Byass P, Witten K H, *et al.* Appropriate tools and methods for tropical microepidemiology: A case-study of malaria clustering in Ethiopia[J]. *Ethiopian Journal of Health Development*, 2003(17):1-8.
- [12] 杨维中,兰亚佳,李中杰,等.国家传染病自动预警系统的设计与应用[J].*中华流行病学杂志*,2010,31(11):13-18.
- [13] Yang W Z, Li Z J, Lan Y J, *et al.* A nationwide web-based automated system for outbreak early detection and rapid response in China[J]. *Western Pacific Surveillance and Response*, 2011,2(1):1-6.
- [14] Kulldorff M. Spatial scan statistics: Models, calculations and applications[M]. // Balakrishnan N and Glaz J (eds). *Recent Advances on Scan Statistics and Applications*. Boston, USA: Birkhäuser, 1999, 303-324.
- [15] 严蔚敏,吴伟民.数据结构[M].北京:清华大学出版社, 2007, 284-286.

A Fast Method for Making Candidate Clusters in Spatial Scan Statistic Method

LI Xiaozhou¹ and WANG Jinfeng^{2*}

(1. Medical School, Wuhan University of Science and Technology, Wuhan 430065, China; 2. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China)

Abstract: Spatial scan statistic method is a widely adopted spatial cluster detection method in the field of public health surveillance. It can detect a sub-zone where the number of disease cases rises abnormally, based on infectious disease surveillance data, and thus is able to make early warning on possible outbreak of infectious disease. Chinese Center for Disease Control and Prevention (China CDC) launched China Infectious Disease Automated-alert and Response System (CIDARS) in 2004, which handles the infectious disease surveillance data of all of the counties of China to detect possible case clusters. The making of candidate clusters is a key step to this method, which to some extent determines the accuracy and time efficiency of the spatial scan statistic method. There are two deficiencies if the existing candidate clusters making method is applied to a very big research area with a lot of sub-regions. The first is that, the inappropriate separation distance of grid points might miss a lot of possible candidate clusters, which affects the accuracy of detected result. The second is that, the existing method might duplicate a great number of candidate clusters, which could prolong the computing time of subsequent spatial scan operation. In this paper a new efficient method is proposed according to the former existing candidate clusters making method. Based on the correct setting to the separation distance of grid points, the new method could greatly reduce the possibility of missing of some possible candidate clusters. At the same time, applying multiple-sort arithmetic, the proposed new method could find and delete a great number of duplicate clusters in the original-making candidate clusters in a shorter time. Finally, the paper applies and tests the proposed method for the making of candidate clusters in 608 counties in southwest Shandong Province and proves that the method works satisfactorily in both two aims, that is, it reduced the computing time and reduced the missing of candidate clusters.

Key words: spatial scan statistic; candidate cluster; grid; multiple-sort

*Corresponding author: WANG Jinfeng, E-mail: wangjf@lreis.ac.cn