第 58 卷 第 3 期 2009 年 3 月
1000-3290/2009/58(03)/2097-09

物 理 学 报
ACTA PHYSICA SINICA

Vol.58,No.3,March,2009

利用空间点过程提取丛集点算法的适用性研究*

杨 萍[1)2)†] 侯 威[1)3)] 支 蓉[1)3)]

1)(兰州大学大气科学学院,兰州 730000)

2)(中国科学院大气物理研究所,东亚区域气候环境重点实验室,北京 100029)

3)(国家气候中心,中国气象局气候研究开放实验室,北京 100081)

(2008 年 2 月 28 日收到;2008 年 9 月 18 日收到修改稿)

以二维泊松过程理论为基础,结合空间点过程理论,引入 k 阶最近邻距离的概念,介绍了基于 k 阶最近距离的丛集点的提取算法,对该算法的适用范围进行了讨论和分析,发现丛集区域和背景区域疏密程度差异以及所研究的数据点数目对该方法有影响,疏密程度差异较小时,该算法的有效性不强,疏密程度差异较大时,该算法较为适用,同时,数据点的总数不同时,算法的适用范围有所差异,但差异不大.此外,引入权重的思想,对算法中理想数据点的设置进行了一定程度的拓展,将所研究区域内的数据点赋予不同的权重,进行丛集数据点的提取,从而扩展了该算法的使用范围.

关键词:丛集点,最近邻距离,EM 算法,分布参数比值

PACC:9260X

1. 引 言

天气和气候极端事件的频率和强度的变化对人类社会经济和环境很可能产生巨大影响,因而受到越来越多的关注.近几年,不少学者利用多种方法从不同空间和时间角度对极端气候事件的变化展开了一系列研究[1—7].在客观世界中,存在很多自然现象属于空间点过程的范畴,如极端气候事件,地震,火山爆发等.利用空间点过程的理论,我们可以将每一个事件的发生看做是空间中理想化的点.通常某一区域中的一个事件集往往不是完全随机分布的,部分事件可能较为集中在某一区域,其他事件则分散在剩余的区域中.前者具有较高的分布参数,可以看做是丛集点或者特征点,在同样的状态空间中遵循着相似的动力学机理,而后者的分布参数较低,可以看做是噪声点或背景点,其在状态空间中的动力学机理有着较大差异或受到的外界扰动较大.因此,利用空间点过程理论进行丛集事件的提取可以帮助我们提取出系统控制机理的外在集中表现形式,并根据这种表现形式的差异反推出控制机理的改变.在很多自然过程中,丛集的事件往往可以认为是重大事件发生的前兆或者是某些重大事件的产物.因此,进行丛集事件的提取可以帮助我们预测重大事件或者分析某些重大事件如极端天气气候事件产生的原因,从而为极端天气气候事件的研究,尤其是其发生发展的空间变化提供新的思路和方法.

关于丛集事件提取方法的研究已有很多[8—12],目前用的较多的一种方法是引入 k 阶最近邻距离的概念,结合空间点过程理论,采用最大期望算法来求出研究区域中每个事件发生的概率,通过一定的判断条件确定出具有丛集性质的点,从而达到丛集点提取的目的[13—22].该方法对数据点的区域分布形状以及数据点的数目要求较低,因此得到较为广泛的运用.该方法的使用中,对丛集点的提取结果产生重要影响的是最近邻距离阶数 k 的选择,已有研究中,对 k 的选择均有所涉及,文献[13]认为 k 较大时可能会带来较大的误差,文献[15]讨论了丛集点的误分率与 k 的关系,文献[17]给出了不同 k 下的 k 阶距离的模拟数据的直方图,并认为最佳距离应选择在介于总点数的 3% —15% 之间.已有研究主要是将该算法应用到丛集地震的研究中.对于算法的研究,主要是对 k 的研究,而对算法的有效性和适用性的研究不多.丛集点提取的正确与否是该算

* 国家自然科学基金(批准号:40675044)和科技部支撑项目资助课题.

† E-mail:yangp_edu@sohu.com

法是否有效的标志. k 是影响算法有效性的重要因素,这已经是该算法研究中所达成的共识,在算法的实际应用中,丛集区域和背景区域的分布参数的比值以及所研究数据点的数目也可能对算法的有效性产生影响,本文首先介绍了该算法的主要思想,然后就算法的有效性进行了详细讨论,并对数据点的设置进行一定程度地拓展,从而使该算法能够更好地更大范围地运用到实际研究中去.

2. 基于 k 阶最近邻距离的丛集区域的提取算法

2.1. 二维泊松过程理论

假设一定范围内的单一的二维泊松过程点集 Y,对于任一点 s_i 的 k 阶最近邻距离 W_k,其概率分布问题可以转化为求以 s_i 点为中心,以 x 为半径的圆内,存在 0,1,2,…,$k-1$ 个点的概率分布函数

$$P(W_k \geqslant x) = \sum_{m=0}^{k-1} \frac{\mathrm{e}^{-\lambda\pi x^2}(\lambda\pi x^2)^m}{m!} = 1 - G_{W_k}(x), x \in [0,\infty). \quad (1)$$

如果 W_k 大于 x,则必定存在 0 或 1 或 2 或…$k-1$ 个点,那么其概率密度函数 $G_{W_K}(x)$ 为

$$g_{w_k}(k,\lambda) = \frac{\mathrm{e}^{-\lambda\pi x^2} 2(\lambda\pi)^k x^{2k-1}}{(k-1)!}. \quad (2)$$

根据伽马分布密度函数的定义,该密度函数为伽马型的复合密度函数,即 $Y \propto \Gamma(k,\lambda\pi)$,其中 $Y = (W_k)^2$.

在实际研究中,丛集点和噪声点分布代表不同参数的空间泊松过程. 它们在空间上互相叠加,因此,以两个叠加的二维泊松过程为例,它们的 W_k 分布服从以下的混合分布:

$$W_k \propto p\Gamma^{1/2}(k,\lambda_1\pi) + (1-p)\Gamma^{1/2}(k,\lambda_2\pi), \quad (3)$$

其中,p 为比例系数,λ_1 和 λ_2 分别为丛集区域和背景区域的分布参数,k 为距离阶数.

2.2. 期望最大化算法(EM 算法)计算分布参数

Byers 和 Raftery[13] 提出用期望最大化算法(EM 算法)计算特征事件和噪声事件的分布参数. 主要思想如下.

以两个叠加的二维泊松过程为例,结合(2)式和(3)式,它们的 W_k 服从以下分布:

$$W_k \propto pg_{W_k}(k,\lambda_1) + (1-p)g_{W_k}(k,\lambda_2), \quad (4)$$

其中,p 为比例系数,λ_1 和 λ_2 分别为分布参数,k 为距离阶数.

EM 算法包括 E 步骤和 M 步骤(具体可见文献(21)).

E 步骤:

$$E(\hat{\delta}_i^{(t+1)}) = \frac{\hat{p}^{(t)} g_{w_k}(k;\hat{\lambda}_1^{(t)})}{\hat{p}^{(t)} g_{w_k}(k;\hat{\lambda}_1^{(t)}) + (1-\hat{p}^{(t)}) g_{w_k}(k;\hat{\lambda}_2^{(t)})}.$$

M 步骤:

$$\hat{\lambda}_1^{(t+1)} = \frac{k\sum_{i=1}^{n}\hat{\delta}_i^{(t+1)}}{\pi\sum_{i=1}^{n} w_{k,i}^2 \hat{\delta}_i^{(t+1)}},$$

$$\hat{\lambda}_2^{(t+1)} = \frac{k\sum_{i=1}^{n}(1-\hat{\delta}_i^{(t+1)})}{\pi\sum_{i=1}^{n} w_{k,i}^2 (1-\hat{\delta}_i^{(t+1)})},$$

$$p^{(t+1)} = \sum_{i=1}^{n}\delta_i^{(t+1)}, \quad (5)$$

其中,n 代表事件总数,$w_{k,i}$ 表示第 $i(i=1,2,\cdots,n)$ 个事件的 k 阶最近邻距离,t 表示迭代次数. 若 λ_1 代表丛集点的分布参数,则 $\hat{\delta}_i^{(t+1)} \geqslant 0.5$ 的点为丛集点,$\hat{\delta}_i^{(t+1)} < 0.5$ 的点为背景点.

3. 算法适用范围的研究

为了量化所确定的丛集点的准确性,引入了错误率这个概念[9]. 错误率是指所研究区域内错分点的个数与数据点总数的比值. 其中,错分点的个数包括丛集点错分为背景点的个数以及将背景点错分为丛集点的个数. 错误率越小,意味着利用该算法所提取出的丛集点和背景点越准确,算法的有效性越强. 已有研究表明[9—12],错误率影响要素大体包括三个方面:一是 k 的取值,算法引入 k 阶最近邻距离,他人现有研究均已说明 k 的取值对错误率有较大影响;二是丛集区与背景区的疏密程度差异,即 λ_1/λ_2 的值,为表述简单起见,我们将 λ_1/λ_2 用 R 表示,即 $R = \lambda_1/\lambda_2$;三是所研究区域数据点的总数,用 n 表示. 其中,相对于疏密程度差异 R 和距离阶数 k 而言,数据点对错误率影响相对最小.

3.1. 数据点数目保持一定

由于数据点总数对错误率的影响较小,暂不考

虑.已有研究表明，R 较小时，错误率较高，k 较小时，错误率较高.我们的研究重点是进一步讨论 R 和 k 对错误率的影响.前提是保持数据点一定，这里，取数据点为 600. k 从 1 取至 30，R 的范围在 1 至 10 之间，不同 R 下，根据最小错误率的值以及错误率与 k 的曲线关系，可以将 R 分为算法不适用和算法适用这两大类.

3.1.1 算法的不适用范围

1. 最小错误率大于 50%

$R<2.1$ 时，最小错误率均达到 50% 以上.图 1(a)是 $R<2.1$ 时部分 R 的错误率随 k 的关系图.

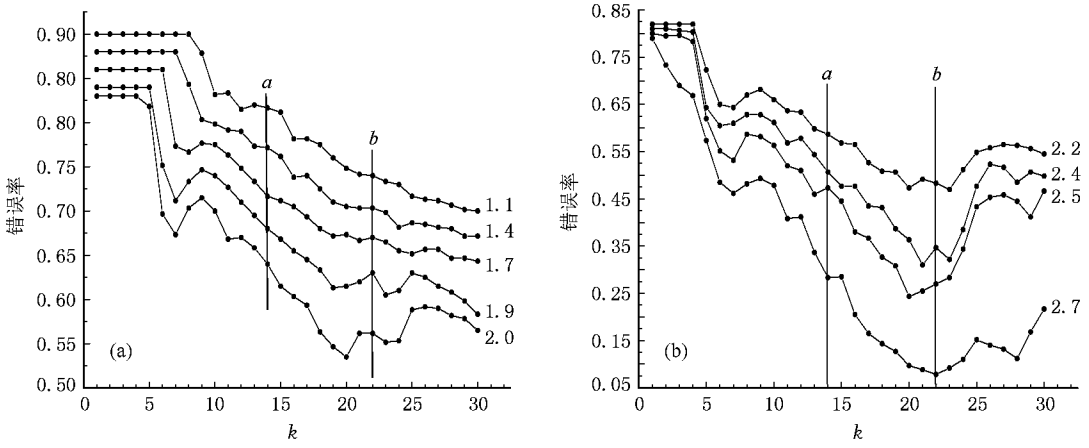

图 1　不同 R 的错误率与 k 的关系图　(a)最小错误率大于 50%；(b)最小错误率大于 5% 小于 50%

图 1(a)中，各曲线尾部的数值表示该曲线所对应的分布参数比值 R，从图中可以看到它们具有相似的特征.k 较小时，错误率保持不变，随着 k 的逐渐增大，错误率呈减小趋势.这些曲线的共同特征是错误率的最小值均大于 50%.此外，当 k 取任一定值时(如图中线 a 表示 $k=14$，线 b 表示 $k=22$)，其错误率随着 R 的变大而呈递减状态.综合而言，当 $R<2.1$ 时，利用该算法所提取出的丛集点和背景点的错误率很高，已经失去了提取的意义.因此，$R<2.1$ 时，该算法不适用.

2. 最小错误率小于 50%，大于 5%

当 $2.1<R<2.9$ 时，最小错误率在 5% 到 50% 之间.图 1(b)是 $2.1<R<2.9$ 部分 R 下错误率与 k 的关系图，与图 a 类似，各曲线尾部的数值表示该曲线所对应的分布参数比值 R，可以看到它们具有相似的特征.k 较小时，错误率保持不变，随着 k 的逐渐增大，错误率随着 k 的增大呈减小趋势，错误率均在 5% 以上.当 k 达到 22 左右时，错误率随着 k 的增大有上升趋势，此外，k 在 22 附近处，各曲线的错误达到最小值.此外，当 k 取任一定值时(如图中线 a 表示 $k=14$，线 b 表示 $k=22$)，其错误率随着 R 的变大而呈递减状态.总体而言，R 在该范围内，错误率没有稳定在一个较小的值上，因此，该算法在此范围内也不适用.

3.1.2. 算法的适用范围

当 $R\geqslant 2.9$ 时，不同 R 下，各自错误率随 k 的变化遵循相似的特征，即先急速衰减，再围绕一较小值平稳波动.我们将错误率刚刚达到稳定时所对应的 k 定义为 k 的初始有效值，不同的 R 其初始有效值不尽相同.不同 R 下，k 大于各自的初始有效值后，错误率稳定在一较小值上，说明所提取出的丛集点具有较高的准确性，因此算法是有效的.针对 k 的初始有效值的位置，可将其细分为以下两种情况.

1. k 的初始有效值不稳定

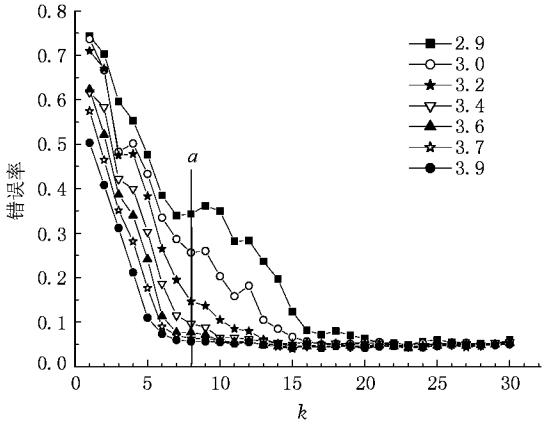

图 2　错误率与 k 的关系图(k 的初始有效值不稳定)

当 $2.9\leqslant R<4.0$ 时，错误率随 k 的变化关系遵

循相似的规律,即急速衰减,后平缓变化.当 $k>16$ 左右,不同 R 的曲线几乎重合,且错误率重合在一小值上,图中约为 5% .也就是说,$2.9\leqslant R<4.0$ 时,$k>16$ 左右时,可以保证该提取出的丛集点具有较高的准确性,即算法有效.此外,我们还可以看到,上图中,各曲线未重合时,相同 k 下(例如图中竖线 a 表示 k 取 8),所对应的错误率随着 R 的增大而减小.尽管该图中 R 的变化范围很小,最大值与最小值的差值仅在 1 左右,但是,不同 R 下 k 的初始有效值却不相同.表 1 给出了不同 R 所对应的 k 的初始有效值.

表 1 不同 R 相应的 k 的初始有效值一览表

R	k 的初始有效值
2.9	17
3.0	14
3.2	12
3.4	9
3.6	8
3.7	6
3.9	6

从表 1 中可以看到,k 的初始有效值随 R 的增大而减小.R 的变化范围虽小,k 的初始有效值的移动却较大,从 16 前移至 6.由于 k 在有效值范围内,取不同值时,其错误率的变化很小,即 k 大于初始有效值后,用算法所提取出的丛集点的准确率均较高.因此,k 在有效值范围内,取不同值时所提取出的丛集点的差别很小.具体运算中,距离阶数 k 取值越大,程序的运算量越大,为避免增大程序的运算量,一般 k 值取在初始有效值附近.

2.k 的初始有效值稳定

图 3 中,竖线所对应的 k 值是为 6.可以看到,当 $R>4.0$ 时,首先,不同的 R 其错误率与 k 的关系曲线均遵循同一变化规律,即错误率随着 k 的增大先急速增大,再围绕某一小值稳定波动.其合适 k 的初始范围也基本相同,在 4—6 之间,也就是说,$R>4.0$ 时,不同 R 下 k 的初始有效值的移动很小,较为稳定.

本节关于 R 和 k 的进一步讨论中可以得到,当数据点保持一定时具有以下性质.

1)当区域内丛集区域与背景区域的疏密差异程度没有足够大时,即 $R<3$ 左右时,该算法不适用.

2)当 $R>3$ 左右时,错误率与 k 的关系类似,均为先急速减小,减小到某一值后,稳定在一近乎水平的线上波动.

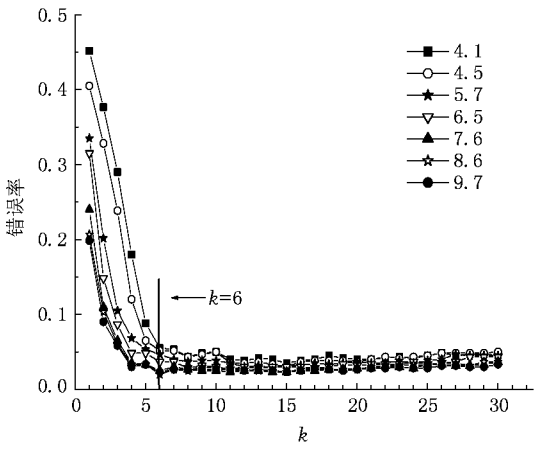

图 3 错误率与 k 的关系图(初始有效值稳定)

3)当 $3<R<4$ 时,其转折点的 k 值有所不同,R 越小,k 越大.当 $R>4$ 时,转折点的位置基本保持在一个稳定的值(6 左右)上.

从适用范围中各 R 下错误率随 k 的变化关系可以看到,当取邻界值下 k 的初始有效值时,适用范围内其余各 R 下的错误率均很小,即均满足错误率在较小值的稳定变化这个条件.前面已经说明,为了降低运算量,一般 k 在初始有效值附近取值.在实际研究中,R 未知的前提下,k 在邻界 R 下的 k 的初始有效值附近取值最为合适,一般,我们可以给出一个小幅变化的范围,如定义初始有效值至其至 $+5$ 以内作为 k 的取值范围.如 $n=600$ 时,k 可以取 17—21 之间.当然,具体研究中可以适当调整.一般,R 较大时,k 可取较小值.

3.2. 数据点数目与算法适用范围的研究

上一部分详细讨论了 n 为 600 时,不同 R 下错误率随 k 的变化关系,揭示了不同 R 下错误率随 k 的变化特征.在实际研究中,并不需要详细了解所有 R 下错误率与 k 的关系,而是希望知道算法是否有效.因此,上部分所讨论的 R 的适用范围是实际研究中最关心的问题.

当数据点为 600 时,适用范围是 $R\geqslant 2.9$ 时,即邻界值为 2.9.在此范围内,错误率与 k 的关系遵循一个较为稳定的关系,即先急速减小,到达某一转折点后,平稳变化,且平稳变化的错误率的值较小,一般在 10% 以下,最小错误率一般可达到 5% 以下.当然,不同 R 下错误率随 k 的变化关系本身存在渐近

变化的过程，此外，理想数据点设置的不同也可能导致邻界值的微小变化．因此，我们所讨论的邻界值并不是一个精确值，而是揭示邻界值的大体位置．当数据点数目为 600 时，R 的适合范围是 $R \geqslant 2.9$．当数据点不同时，邻界值是否会发生变化，本部分着重讨论了该问题．

研究中，分别取不同的数据点总数，采用与第 1 部分类似的方法，画出各数据点下不同 R 时错误率与 k 的关系图，找出邻界值 R_1，从而得到不同数据点时该算法的适用范围．在适用范围内的 R 具有以下特征，1）最小错误率达到 5%，2）错误率随着 k 具有一段平稳变化的过程．上述两个条件都符合，才能将其归为该算法的适用范围内．我们从 $R = 1.1$ 开始取值，逐渐增大 R，观察其错误率随 k 的变化关系，从而确定出不同 n 下的邻界值 R_1，表 2 给出数据点个数不同时各自的邻界值 R_1 的值以及相应 k 的初始有效值和 k 的合适取值范围．

表 2　不同数据点总数的邻界值 R_1 以及相应 k 的初始有效值和合适取值范围

n	邻界值 R_1	k 的初始有效值	k 的取值范围
100	3.6	12	12—16
140	3.6	11	11—15
180	3.2	9	9—13
200	3.2	10	10—14
240	3.2	13	13—17
280	3.0	14	14—18
300	3.4	13	13—17
350	3.2	12	12—16
380	3.7	12	12—16
400	2.9	13	13—17
500	2.9	17	17—21
600	2.9	17	17—21
700	2.9	23	23—27
800	2.9	23	23—27
900	2.9	22	22—26

从表 2 可以看到：第一，数据点在 400 以上时具有类似的性质：1）数据点个数对邻界值的影响不大；2）邻界值相同，均为 2.9；3）邻界状态下 k 的初始有效值不同，具体而言，k 的初始有效值随着 n 的增大而后移．第二，数据点在 400 以下时，邻界值虽然没有数据点总数在 400 以上那么固定，但还是在一定范围内波动，基本是在 3.0—3.7 之间波动，k 的初始有效值没有明显的规律，基本在 12 附近波动．

3.3. 总　结

关于算法的适用范围的研究，总结如下：

1）该算法存在一定的适用范围，并不是所有区域内随机分布的数据点都适合使用该算法．

2）当所研究区域内丛集点相对于背景点而言，丛集点的密集程度不是那么明显，该算法不适用．相对密集程度可以用分布参数比值 R 表示．当 R 小于某一邻界值 R_1 时，该算法不适用．

3）邻界值 R_1 受到区域内数据点总数 n 一定程度的影响．数据点总数较大一般在 400 以上时，邻界值 R_1 相对偏小，且比较稳定，在 3 左右．数据点总数较小一般在 400 以下时，邻界值 R_1 相对偏大，且在一定范围内波动，波动范围基本在 3 至 4 之间．

4. 理想数据点设置的拓展

4.1. 数据点分布形态的研究

前面所有数据点形状的设置中，丛集点均是集中在 300×300 规则矩形内．事实上，在实际研究中，丛集区域不可能如模拟数据一样，完全集中在某一块形状规则的区域内，文献[6]指出，数据点的形状不会给结果带来影响，为了验证此结论，我们利用该算法画出了不同形状的丛集点，分别为圆环形，十字形和 I 形，用该算法进行丛集点的提取．模拟数据分布在为 1000×1000 区域内，总数均为 600 个点，k 均取 20．分布图如图 4，其中(a)，(c)，(e)为模拟图，叉形表示背景点，圆点表示丛集点；(b)，(d)，(f)为算法提取图，叉形表示提取出的背景点，圆点表示提取出的丛集点．

图 4 中，从提取图与模拟图的对比可以看到，丛集区域的形状对该算法没有影响，不管丛集区域的形状如何，该算法均能够较为准确将丛集点提取出来．

4.2. 数据点的权重研究

在前面所讨论的理想数据的设置中，不会在相同的位置出现重复的数据点，即所研究区域中的某一位置，数据点的可能性最多只有一次．前部分研究表明，如果丛集点和背景点的疏密差异程度不够，即 R 小于邻界值，那么就无法正确地将理论上的丛集

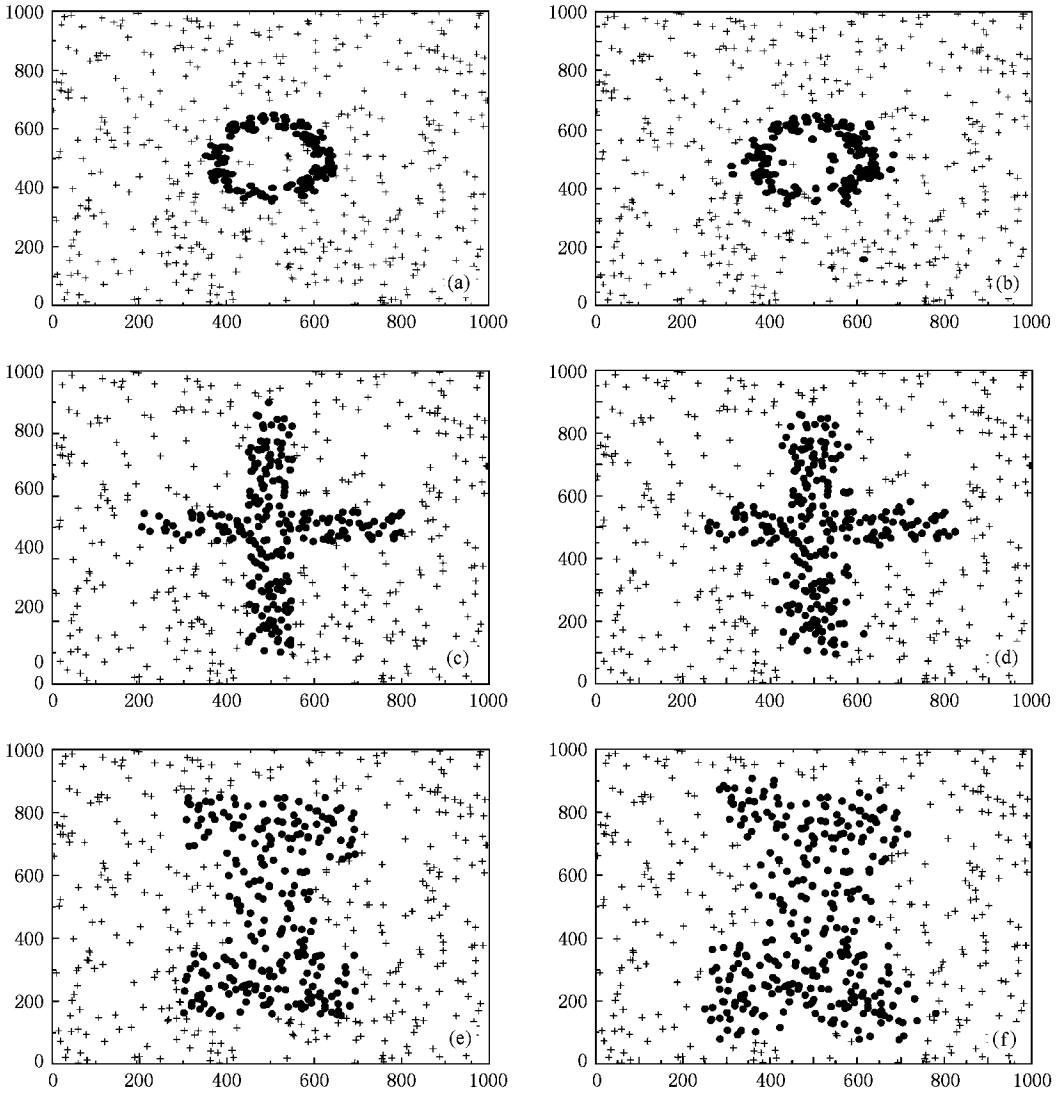

图 4 不同形状分布下丛集点的提取图 (a),(b)圆环形;(c),(d)十字形;(e),(f)I形

点提取出来.例如:设置的模拟数据为横坐标和纵坐标均在(0,1000)之间的 1000×1000 正方形区域内二维泊松过程数据点的数据点 600 个,其中丛集点的所在区域范围是横坐标和纵坐标均在在(300,600)之间的 300×300 正方型区域,具体分布图如图 5.

图 5 中,中心部分的虚框所包含的区域为设定的丛集区域,黑色圆点为所设置的丛集点.该区域内分布的数据点数为 90 个,利用 $R=(n_1/s_1)/(n_2/s_2)$ 公式,可求出该理想数据的分布参数比值 R 的理论值为 1.8,前面研究了数据点大于 400 左右时,其算法的适用范围的邻界值 R 为 2.9 左右,根据前面的研究结果,该算法无法有效地提取出丛集点.为了表示出该算法在此处的不适用性,我们画出了利用该算法所得到了丛集点的分布图,如图 6(a),其中 k

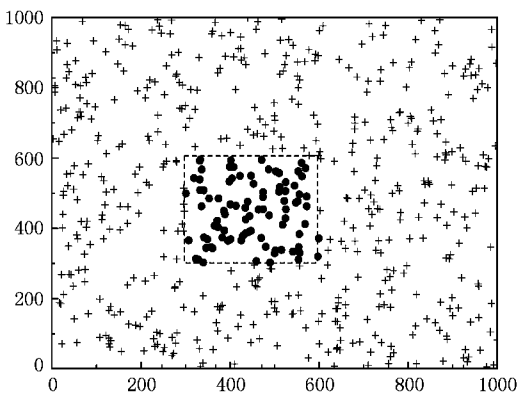

图 5 模拟数据的丛集点和背景点的空间分布

取 23.图 6(a)中,黑色圆点为利用算法所得到的丛集点,大大超出了中心部分虚框所包含的点,计算错

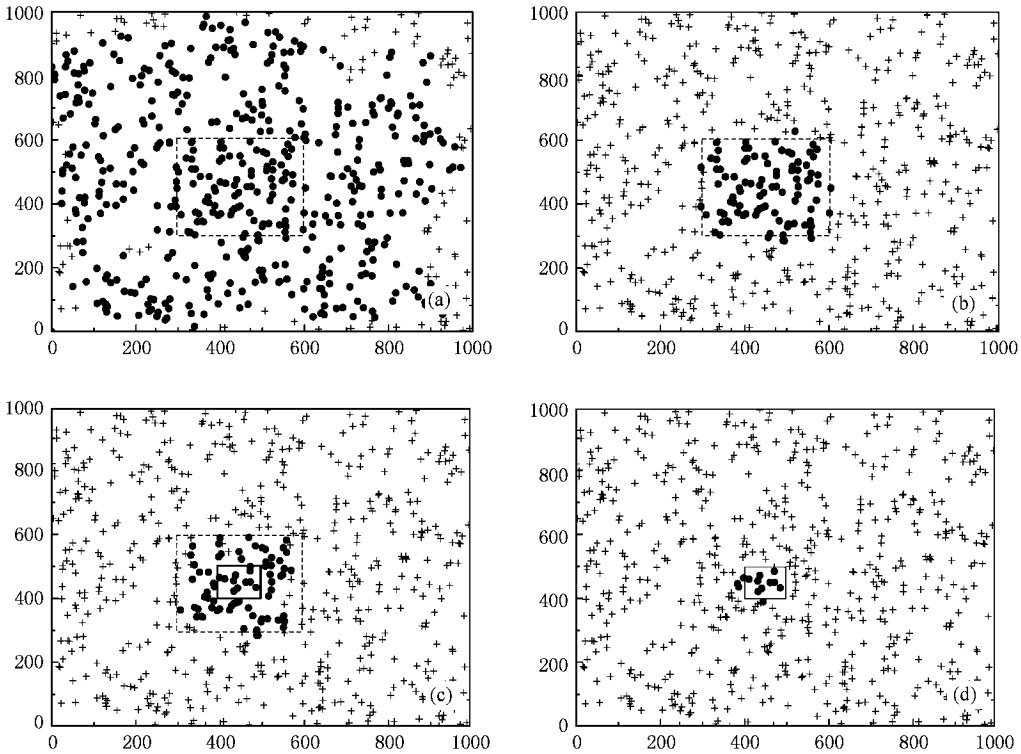

图 6　不同权重下模拟数据的分解结果

误率，达到 66.7% . 该图更加直观地反映了当 R 不在适用范围内时，提取出的丛集点是毫无意义的.

以上所讨论的数据点，在所研究区域内均出现了一次，如果某些数据点出现的次数不是一次，而是两次甚至更多次，我们利用该算法能否把出现次数多的那些提取出来呢？为了验证我们的想法，我们作了如下设置 :图 5 中横坐标和纵坐标均在在(300，600)之间的 300×300 正方形区域的点保持原来的位置不变，但是点的次数均改为两次. 由于丛集区域的点数增加了一倍，即丛集区域内增加了 90 个点，保持背景区域内点的位置和个数不变，总点数增至 690. 也就是说，在本次研究的区域中，从点的分布图来看，和图 5 相同，但是实际上，中间黑色圆点均出现了两次. 我们利用该算法将丛集点提取出来，k 取 23，得到图 6(b)：

图 6(b)中，虚框所包含的区域代表丛集区域，可以看到，提取出的丛集点的准确性很高，基本在虚框内. 因此在这里，该算法显然是适用的. 究其原因，也很容易理解. 在该具有重复点的丛集区域内，虽然数据点的分布比较特殊，即每个点都出现两次，但是，丛集点的个数为 180 个，背景点的个数为 510，利用 $R = (n_1/s_1)/(n_2/s_2)$ 公式，可求出其 R 值为 3.6，显然在适用范围之内，因此丛集点是可以提取出来的.

从上面的讨论中，我们可以引入权重的概念来反映每一个数据点出现的次数. 如出现次数为 1，权重为 1，出现次数为 2，权重为 2，出现次数为 m，权重为 m. 权重越高，说明该点的重复的次数越多，在具体问题中，可以说明该点的重要性或级别越大. 此时，用算法所提取出的丛集点不仅仅是指空间分布密集的点，而且反映了该位置上点的权重，或者称之为重要性或级别. 因此，加入权重后的点的丛集性是空间丛集与点的级别的结合，即空间分布越密集，该区域内点的丛集性越高，点的权重越大，该处点的丛集性越大.

在图 6(b)的研究基础上，进行如下改进，将横坐标和纵坐标均在在(400，500)之间的 100×100 正方形区域的点保持原来的点，但是出现次数增加到 6 次，其余数据点的分布以及出现次数保持不变. 利用算法所提取出的丛集点如图 6(c)，其中，虚框内的数据点包含两类，一类是虚框中心部分的实框内的数据点，权重为 6，其余部分为第二类数据点，权重为 2，整体区域内十字形数据点的权重均 1. 从所提取出的数据点可以看到，所提取出的丛集点中，虚

框内权重为 2 的数据点中空间分布较为稀疏的点就被剔除.这说明,提取的丛集点是权重和空间分布两者的结合.

若继续增加实框内 100×100 正方形区域各点的权重,将其权重增加到 14,所提取出的丛集点如图 6(d),该图中,所挑出的丛集点基本都为权重为 14 的点.可见,当权重增大到足够大的时候,尽管权重为 2 的数据点和权重为 1 的数据点有所差别,尽管数据点的空间分布上疏密程度有所差异,但当权重处于主导地位的时候,就会将权重大的点挑出来,此时,空间上的疏密程度就被忽视了.反之,如果空间疏密程度的地位远远大于权重之间的差异,权重的作用就会被忽略.

5. 结 论

不少学者利用多种方法从不同空间和时间角度对极端气候事件的变化展开了一系列的研究.利用空间点过程的理论,我们可以将每一个事件的发生看做是空间中理想化的点,部分事件可能较为集中在某一区域,具有较高的分布参数,可以看做是丛集点或者特征点,在同样的状态空间中遵循着相似的动力学机理;其他事件则分散在剩余的区域中,其分布参数较低,可以看做是噪声点或背景点,其在状态空间中的动力学机制有着较大差异或受到的外界扰动较大.因此,利用空间点过程理论进行丛集事件的提取可以帮助我们提取出系统控制机理的外在集中表现形式,并根据这种表现形式的差异反推出控制机理的改变.进行丛集事件的提取可以帮助我们预测重大事件或者分析某些重大事件产生的原因,从而为分析极端天气气候事件的空间分布及其变化提供新的思路和方法[23—26].通过以上研究,可以得到如下结论:

1. 丛集点提取算法具有一定的适用范围,并不是所有区域内随机分布的数据点均适合用该算法.当所研究区域内的丛集点相对于背景点的密集程度不那么明显,该算法不适用.相对于疏密程度差异 R 和距离阶数 k 而言,数据点对错误率影响相对最小,即数据点对算法的有效性的影响相对较小.

2. 根据分布参数比值 R,距离阶数 k 和数据点总数 n 与错误率的关系,定性地给出了不同数据点下算法的适用范围 数据点数目较小时,所要求的丛集点与背景点的疏密差异程度要更明显一些,即 R 要相对大些,且在一定范围内波动,波动范围基本在 3—4 之间.数据点较多时,一般大于 400 左右,R 相对小一些,且比较稳定,一般在 3 左右.

3. 对理想数据点的设置进行一定程度的拓展研究,算法的有效性并不受数据点分布形状的影响,数据点分布形状不同时,如丛集性质的数据点分布为环状、十字形等形状时,在算法的适用范围内,该算法均能有效地将丛集性质的点提取出来.

4. 研究中,对数据点的设置进行了一定程度的拓展,若研究区域内存在重叠的数据点,可以引入权重的思想,给各重叠的数据点赋予不同的权重,同样可以利用该算法提取出丛集的数据点,其中,当权重增大到足够大的时候,尽管其他权重的数据点如权重为 2 的数据点和权重为 1 的数据点有所差别,尽管数据点的空间分布上疏密程度有所差异,但当权重处于主导地位时,就会将权重很大的点挑出来,此时,权重较小的数据点以及空间的疏密程度就被忽视了.反之,如果空间疏密程度的地位远远大于权重之间的差异,权重的作用就会被忽略.

[1] Cao H X 1993 *Science in China* (Series B) **23** 104(in Chinese) [曹鸿兴 1993 中国科学(B 辑) **23** 104]

[2] Li J P, Chou J F 2003 *Chinese Science Bulletin* **48** 703 (in Chinese) [李建平、丑纪范 2003 科学通报 **48** 703]

[3] Dai X G, Fu C B, Wang P 2005 *Chin. Phys.* **14** 850

[4] Shi N, Gu J Q, Yi Y M *et al* 2005 *Chin. Phys.* **14** 844

[5] Yang J P, Ding Y J 2005 *Scientia Gegraphica Sinica* **25** 441 (in Chinese) [杨建平、丁永建 2005 地理科学 **25** 441]

[6] Li D L, He J H 2007 *Plateau Meteorology* **26** 39(in Chinese) [李栋梁、何金海 2007 高原气象 **26** 39]

[7] Ma Z G, Shao L J 2006 *Chinese Journal of Atmospheric Sciences* **30** 46(in Chinese) [马柱国、邵丽娟 2006 大气科学 **30** 464]

[8] Allard D, Fraley C 1997 *Journal of the American Statistical Association* **92** 1485

[9] Ester M, Kriegel H P, Sander J, Xu X 1996 *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Portland: Oregon) p324

[10] Jin H D, Leung K S, Wong M L, Xu Z B 2005 *Pattern Recognition* **38** 637

[11] Dasgupta A, Raftery A E 1998 *Journal of the American Statistical*

Association **93** 294

[12] Zhou M，Sun S D 1999 *The theory of Genetic Algorithm and its application*（Beijing：National Defence Industry Press）p1［周　明、孙树栋 1999 遗传算法原理及应用（北京：国防工业出版社）第 1 页］

[13] Byers S D，Raftery A E 1998 *Journal of the American Statistical Association* **93** 577

[14] Pei T，Zhu A X，Zhou C H，Li B L，Qin C Z 2006 *International Journal of Geographical Information Science* **19** 153

[15] Wang M，Leung Y，Zhou C H，Pei T，Luo J C 2006 *Data Mining and Knowledge Discovery* **12** 97

[16] Pei T，Zhu A X，Zhou C H，Li B L，Qin C Z 2007 *Computers & Geosiences* **33** 952

[17] Cressie N A C 1991 *Statistics for spatial Data*（New York：first ed. Wiley）p900

[18] Pei T，Zhou C H，Yang M，Luo J C，Li Q L 2004 *Acta Seismologica Sinca* **26** 53［裴　韬、周成虎、杨　明、骆剑承、李全林 2004 地震学报 **26** 53］

[19] Deng Y L，Liang Z S 1998 *Radom point process and application*（Beijing：Nature Process）p10［邓永录、梁之舜 1998 随机点过程及其应用（北京：科学出版社）第 10 页］

[20] Sheng Z，Xie S Q，Pan C Y 1999 *The Theory of Probability and the Mathematical Statistic*（Beijing：Higher Education Press）p304［盛　骤、谢式千、潘承毅 1999 概率论与数理统计（北京：高等教育出版社）第 304 页］

[21] Celeux G，Govaert G 1992 *Computational Statistics and Data Analysis* **14** 315

[22] Moon T K 1996 *IEEE Signal Processing Magazine* **13** 47

[23] Feng G L，Dong W J，Li J P 2006 *Acta Phys. Sin.* **55** 2664（in Chinese）［封国林、董文杰、李建平 2006 物理学报 **55** 2664］

[24] Feng G L，Hou W，Dong W J 2006 *Acta Phys. Sin.* **55** 962（in Chinese）［封国林、董文杰 2006 物理学报 **55** 962］

[25] Ning D，Lu J A 2005 *Acta Phys. Sin.* **54** 4590（in Chinese）［宁　娣、陆君安 2005 物理学报 **54** 4590］

[26] Yuan G Y，Yang S P，Wang G R，Chen S G 2005 *Acta Phys. Sin.* **54** 1510（in Chinese）［袁国勇、杨世平、王光瑞、陈式刚 2005 物理学报 **54** 1510］

Research on the validility of cluster delineation based on spatial point processes*

Yang Ping[1)2)†]　Hou Wei[1)3)]　Zhi Rong[1)3)]

1)（*Department of Atmospheric and Sciences*，*Lanzhou University*，*Lanzhou* 730000，*China*）

2)（*Key Laboratory of Regional Climate Environment Research for Temperature East Asia*，*Institute of Atmospheric Physics*，*Chinese Academy of Sciences*，*Beijing* 100029，*China*）

3)（*Laboratory for Climate Studies of China Meteorological Administration*，*National Climate Center*，*Beijing* 100081，*China*）

（Received 28 February 2008；revised manuscript received 18 September 2008）

Abstract

On the basis of the two-dimensional Poisson process theory and spatial point process theory，we introduced the algorithm of cluster extraction which is based on k-th order distance，aiming at determining the valid range of the algorithm. We found that the ratio of cluster and cluster's densities and the number of data both have effect on the method. But the effect of the former much bigger than that of the latter. Furthermore，we introduced the concept of weight to extend the range of the simulated data，and gave the data different weights for delineating clusters，so that the range of this algorithm can be extended.

Keywords：cluster，nearest-distance，EM algorithm，ratio of densities

PACC：9260X

* Project supported by the National Natural Science Foundation of China（Grant No. 40675044）and the Ministry of Science and Technology.

† E-mail：yangp_edu@sohu.com