

# 时空点过程:一种新的地学数据模型、 分析方法和观察视角

裴 韬,李 婷,周成虎

(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:** 栅格计算因其具有简单的构架成为目前地学分析的主流模型,然而,由于栅格计算平均分配计算和存储资源的弱点,不仅容易产生冗余,更重要的是难以凸显研究对象的突变部分,从而使研究者有可能忽略地学现象的变化特征。为此,本文提出将时空点过程模型应用于地学研究。时空点过程不仅适用于模拟以点事件为基本单元的地学现象,而且由于大多数地学过程可以转化为时空点过程,故其具有更广泛的应用范围。因此,时空点过程不仅是一种数据模型,同时也是地学问题的分析方法,更是观察和理解地学问题的一种新视角。为了实现从点过程数据中提取模式,作者经过多年研究提出了时空点过程层次分解理论框架,该理论与信号处理理论中的谱分析思路类似,首先,假设任意点集为有限多个均匀点过程的叠加,然后,通过点局部密度表达工具K阶邻近距离,将空间点转换为混合概率密度函数,再应用优化方法将混合密度函数进行分解得到丛集点和噪声,最终利用密度相连原理从丛集中提取模式。该理论框架可适用于绝大多数点集数据,初步实现了点集数据的“傅里叶变换”。

**关键词:** 聚类;数据挖掘;K阶邻近距离;泊松过程;非均匀点过程

**DOI:** 10.3724/SP.J.1047.2013.00793

## 1 引言

遥感和DEM数据的广泛应用,使得栅格计算方法论在地学计算中盛行。将各类数据转化为栅格数据进行并行化处理、高性能计算已经成为地学计算的趋势。然而,就在栅格化给地学计算和空间分析带来极大便利的同时也埋下了一定的隐患:栅格计算框架中任何一个格网单元无论重要与否,其占用资源、表达方式、处理方法皆相同。这不仅会产生较大的计算资源浪费,也有可能因这种平均效应使研究者忽略了地学过程中重要的质变点和隐藏的模式。与栅格计算的策略不同,本文提出一种以事件为基本单元的地学数据处理和计算模型——时空点过程模型。

时空点过程模型的提出是用于刻画自然界的点事件,其实质为描述点(事件)时空分布的随机过程<sup>[1]</sup>。时空点过程在地学领域具有广泛的应用。一方面,地学过程中很多现象可抽象为点过程,如:地震的发生、疾病患者的住址、犯罪地点的分布等等;

另一方面,大多数地学过程通过转换也可以用点过程表达,如气温的时间序列可以转换为极端气温事件点过程<sup>[2]</sup>、元素的地球化学场也可转化为某种条件概率下的矿化点过程<sup>[3]</sup>、个人在一段时间内的迁移可以是不同居住地组成的点过程<sup>[4]</sup>等。由此可见,地学过程大多可以通过抽象得到时空点过程模型。在本文中,我们将点过程模型进行推广,不仅仅将其视为描述地学现象的模型,同时也将其作为观察和研究地学现象的一种新视角和方法论。与栅格计算将研究对象划分为大小相等网格的思路不同,点过程模型在观察地学现象时将研究对象抽象为离散的时空点集。这种抽象的优势是将地学现象的重要部分加以突出(即保留为点过程中的点事件),使研究者将注意力集中在地学过程中发生质变的节点上。这与平均分派资源的栅格计算的思路是截然不同的。

虽然点过程模型凸显了地学现象中的突变点,然而要将时空点过程模型应用于地学研究,还存在若干理论问题亟待解决,其中最为重要的有两点:

收稿日期:2013-11-14;修回日期:2013-12-03.

基金项目:国家自然科学基金面上项目(41171345);国家高技术研究发展计划项目(2012AA12A403)。

作者简介:裴 韬(1972-),男,江苏扬州人,博士,研究员,研究方向为时空数据挖掘。E-mail: peit@lreis.ac.cn

一是如何将地学现象转化为时空点过程,这种转化实际上是将地学过程抽象为点事件的集合;二是如何从点过程中提取模式。地学过程大都是各尺度地学作用的时空叠加,这种复杂的叠加是从地学过程中挖掘模式的瓶颈。要解决第一个关键问题,需要充分理解研究对象的本质和研究的目的及需求,从而完成对地学过程的抽象。而要突破第二个难点,一个重要的思路就是对地学过程进行系统的层次分解,而这正是本文所要论述的核心问题。本文首先分析时空点过程的主要特征;然后探讨时空点过程模式挖掘的基本内容及其关键问题;并在此基础上提出了时空点过程的分解理论;最后对本文予以总结并对未来的研究方向进行展望。

## 2 时空点过程的主要特征

与栅格计算模型中的基本单元为规则多边形不同,时空点过程模型的基本单元是点事件,从而让研究者将精力集中在地学现象的突变特征上。正因为此,时空点过程模型的主要特征集中表现为:抽象、离散和层次性。正是由于具备这些特征,时空点过程模型才成为有别于栅格计算的地学问题的研究利器。下面就对其主要特征作简要解释。

### (1) 抽象性

时空点过程的抽象性是指绝大多数地学现象可以在不同尺度上抽象为点事件集。例如,在较小的尺度上,地震震中、犯罪地点等可视为点事件,而城市则不能视为点;当我们将观察视角放在较大的尺度上时,城市则可抽象为点。这种对地学过程的抽象可以简化研究对象,将其转化为包含一系列事件的点集,从而凸显“与众不同”的事件本身。因此,这种抽象不是简单的细节删除,而是问题的聚焦和明晰。

### (2) 离散性

时空点过程的离散性是相对于连续性而言的,具体是指点事件之间的分布不连续。例如,在气温时间序列中,虽然气温值的变化具有空间连续性,但对于极端气温事件,它们的位置之间却是空间离散的。连续性和离散性是自然现象中同时存在的看似矛盾的两种性质。连续性可以让我们从相似性中发现不同,而离散性则让我们从不同中发现相似性。

### (3) 层次性

点过程的层次性是指:①点过程中的事件在更

高的时间(空间)分辨率下又可以视为点过程;②而一个点过程在更低的时间(空间)分辨率上又可以概化为一个事件。仍以地震为例,一个地震通常是某个断层的破裂造成的,而断层的破裂又可视为一系列的变形和位移事件所构成的点过程。而在更高的层次上,大的地震通常会伴有前震或余震,从而形成地震序列,而地震序列可以看成为点过程。不过,地震序列在更长的时间尺度上,又可被视为一个点事件。因此,在地学计算的领域中,不仅栅格模型具有层次性,点过程模型同样具有层次性。不过这种层次性与栅格模型的不同,是点事件与点过程在不同尺度之间的转换。

## 3 时空点过程模式挖掘的基本内容及其关键问题

与点过程相关的基本概念有强度函数 $\lambda(x)$ (intensity function)和支撑域(support domain)。点过程的强度函数 $\lambda(x)$ 可以近似理解为点过程的局部密度,这种局部密度可以随空间位置( $x$ )的不同而发生变化(下文中如未加特别说明则以密度指代强度)。支撑域定义为点过程所出现的区域,也可理解为点过程所“占领的领土”<sup>[5]</sup>。

根据强度函数的不同,点过程可以划分为多种类型。本文将点过程粗略地分为均匀过程(Homogeneous process)和非均匀过程(Nonhomogeneous process)。其中,均匀过程即为泊松过程(点在支撑域中完全随机分布),其强度函数在整个支撑域中为常数。非均匀过程的强度函数是一个连续函数,其值随空间点位置的不同而发生变化<sup>[5-6]</sup>。实际上,非均匀点过程还包括多种类型,如二项过程(Binomial process)、马尔科夫过程(Markov process)、Cox过程(Cox process)等<sup>[5]</sup>,限于篇幅不再详述。除了上述单变量过程之外,还有多个过程耦合在一起的复合过程,如复合泊松过程(即点事件的属性也服从另一种点过程)<sup>[1]</sup>。随着点过程研究的深入,会有更多的空间点过程类型不断产生。

时空点过程模式挖掘的目的是从点过程数据中提取出有意义的模式。相对于随机性而言,点过程的模式可分为丛集模式(clustering pattern)和规则模式(regularity pattern)<sup>[6]</sup>。本文讨论的重点是点过程的丛集模式,并将点过程数据丛集模式的识别分解为以下几个部分:(1)需要判别时空点数据中

是否存在丛集点(即数据是完全随机还是蕴含丛集点),这是进行后续模式挖掘的必要条件,即如果数据中蕴含丛集点才有进一步挖掘的需要。(2)在包含了噪声和丛集点的数据中,如何建立区分噪声和丛集点的模型。此处,噪声和丛集点之间的区别主要是密度上的差别。(3)如何确定区分噪声和丛集点的密度阈值。密度小于阈值的点划分为噪声,而大于阈值的点为丛集点(聚类)。(4)区分丛集点和噪声之后,如何获得相应的丛集模式。(5)当点具有时空属性时,如何挖掘其时空点模式。(6)随着尺度的细化,当点事件拉伸为时空序列(即下一层次的点过程)时,如何挖掘时空序列中的模式。

上述分析说明,从点过程数据中挖掘模式需要解决一系列关键问题。其中,第1部分针对点过程随机性的判别其关键是点随机性的尺度依赖性问题;第2部分区分噪声和丛集点的基础是需要寻找点过程局部密度的表达工具;第3部分确定区分噪声和丛集点的阈值其核心问题是阈值求取模型的优化;第4部分丛集模式的提取依赖于聚类模型的构建;第5部分时空耦合问题的解决取决于时空密度的定义和计算;第6部分挖掘时空序列模式的核心是如何衡量时空序列之间的相似性。下面逐一阐述各关键问题的内容和研究现状。

### 3.1 随机性判别中的尺度性

点过程数据随机性判别的研究始于生态学等相关领域,随后渗透到地学的多个领域。点数据的随机性判别方法主要分为指标判别和函数判别。指标判别实际上就是通过计算单一的指标值,参照空间点在随机假设下的分布,使用假设检验的方法判断点集是否为随机。随机性指标有多种类型,大体分为以下几类:基于样方(quadrat-based)<sup>[7-8]</sup>、基于角度(angle-based)<sup>[9-10]</sup>、基于距离(distance-based)<sup>[11-13]</sup>、角度与距离混合(angle-distance mixed)<sup>[14]</sup>、基于谱分析(spectral-based)<sup>[15]</sup>和基于二阶矩(second-order-momentmeasure-based)<sup>[16-18]</sup>等指标。

指标判别虽然简单易行,但由于指标仅为单一的数值,其中并未包含尺度信息,因此无法反映时空点过程丛集模式随尺度变化的特征。而函数判别由于包含了尺度信息,故可以判别丛集特征的尺度效应。判别随机性的函数主要有K函数、L函数和G函数及其各种改进类型等<sup>[6]</sup>,均可用于丛集模式尺度的估计(即聚类模式的大小)。即便如此,上

述函数仍无法估计丛集点模式中点的数目。为此,Pei等(2011)提出了K阶邻近距离(Kthnearest distance)方差比的判别函数<sup>[19]</sup>。该函数可以通过K阶邻近距离(3.2节中将介绍其概念)方差比值随K值变化的趋势确定丛集点模式中点的数目。而丛集点规模的确定为后续计算中参数的选择(如K值的选取)奠定了基础。

### 3.2 点过程局部密度的表达工具

如上所述,区分点集数据中噪声和丛集点的主要依据是密度,如何客观、简便、直观地刻画局部密度是区分噪声和丛集点的关键。用以表达局部点密度的方法主要有基于单元的方法、基于K阶邻近距离的方法和基于模型的方法。

基于单元的密度表达工具其思路是统计单元内点的数目,通过点数目与单元面积之比确定单元内的点密度,即局部点密度。此处的单元可为圆形(多用于空间扫描模型,如SaTscan模型)<sup>[20]</sup>、正方形单元(多用于格网模型)<sup>[21]</sup>,以及不规则多边形(如Voronoi多边形)<sup>[22]</sup>等。

K阶邻近距离方法另辟蹊径,用点与其第K近邻居点之间的距离表达该点的局部密度<sup>[23]</sup>(图1)。这样,某点的K阶邻近距离越大,其局部密度越小;反之则越大。除了K阶邻近距离之外,还有另外一些距离同样可以衡量局部密度,如共享邻近距离(shared nearest distance)<sup>[24]</sup>等。

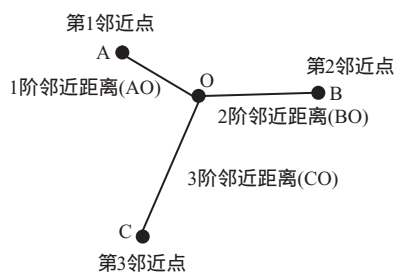


图1 K阶邻近距离法

Fig.1 The  $K^{\text{th}}$  nearest distance

基于模型的方法是利用事先定义的密度模型估算局部的点密度。这类模型有多种,如高斯模型、泊松模型、核密度函数等。而核函数模型又是应用较为广泛的一种,包括了均匀函数、三角函数以及径向基函数等多种类型<sup>[6,25]</sup>。

基于单元的方法计算虽然简单,但对于支撑域不规则的点过程,有效单元(即属于支撑域的单元)



的确定较为困难。基于模型的方法可以较好地刻画局部点密度,但需要有先验知识支持模型的选择(如,何时选择高斯模型、何时选择泊松模型等),同时对于噪声点和丛集点的区分仍依赖于先验知识。基于 $K$ 阶邻近距离的方法在计算局部密度时,所使用的参数较少(仅为 $K$ ),故计算简单且所受的主观性影响也较小。

### 3.3 丛集点与噪声区分阈值的求取模型

区分噪声和丛集点的阈值的求取过程其实质就是寻找两者之间的差异。不同的密度表达方法,其阈值计算的方法也是不同的。对于基于单元的表达方法,通常采用蒙特卡洛方法求取噪声和丛集点之间的阈值。其思路是:首先假设点事件在支撑域中呈随机分布(即将点视为噪声),然后通过模拟得到噪声的概率密度,并由此求出在一定置信度水平下推翻噪声假设的临界值<sup>[26]</sup>。该临界值即为区分噪声和丛集点的阈值。

基于模型的阈值求取思路为:首先判断数据中聚类模型的类型和数目,并据此选取相应的函数类型,然后建立求取阈值的目标函数,最后通过优化方法求取模型的参数。在优化目标函数的过程中,通常采用的优化模型包括 Expectation-Maximization(EM)算法和贝叶斯方法等<sup>[25]</sup>。

而基于 $K$ 阶邻近距离的方法,其阈值估计主要有以下几种思路:(1)通过可视化的方法,将所有点的 $K$ 阶邻近距离进行排序,然后通过先验知识估计区分二者的阈值,如 DBSCAN<sup>[27]</sup>和 OPTICS<sup>[28]</sup>聚类模型中的阈值估计方法。(2)建立点集 $K$ 阶邻近距离的混合概率密度模型,通过识别该模型中所包含的独立密度成分将点集分为噪声和丛集点。根据不同的情况, $K$ 阶邻近距离方法又可细分为3种类型:①当混合密度函数中成分的数目(即点过程的数目)已知时,可采用EM策略求取混合概率密度模型的参数,从而将噪声和丛集点分开<sup>[29]</sup>。②当点集数据中所包含的密度成分的数目未知时,可采用 reversible jump Markov Chain Monte Carlo(MCMC)方法。该方法的主要思路是首先通过蒙特卡洛方法生成不同密度成分数目下的参数,然后再利用马尔科夫方法生成混合点过程的现实,最终凭借出现概率确定数据中密度成分的数目(即取概率最大情况下所出现的密度成分的数目)<sup>[30]</sup>。③可采用逐步分解法,即每次将数据视为噪声和丛集点的集合,重

复使用EM策略,直到数据不可分为止<sup>[31]</sup>。

### 3.4 空间点的聚类模型

在确定区分噪声和丛集点的阈值之后,就需要通过聚类模型将丛集点进行分组,从而抽取其中的模式。由于点局部密度表达方式的不同,所使用的空间点聚类模型也不一样。基于单元的聚类,其聚类的机制是剔除密度较低的单元,并将密度较高的单元连接起来形成聚类,从而发现丛集模式<sup>[21,26]</sup>。而基于模型的聚类则是将统计上显著的模型提取出来作为聚类模式<sup>[25]</sup>。与上述两种聚类模型不同,基于 $K$ 阶邻近距离的聚类模型采用密度相连的机制将丛集点聚合在一起,同时将噪声剔除。其核心思路是将时空中密度一致的点聚为一类,从而实现每个点邻域内的点与自己一样都具有较高的密度。该思路类似于多米诺骨牌,点与点的相连类似于一个骨牌的倒下可以触发另一个骨牌倒下,而骨牌能否倒下则取决于其密度(此处以 $K$ 阶邻近距离表示)是否大于某一阈值。具有代表性的聚类方法包括 DBSCAN、OPTICS、DECODE 等<sup>[27-28,30]</sup>。

### 3.5 时空密度的计算

当点的属性中同时包含时空信息时就需要进行时空点过程分解。对于时空耦合点的密度聚类,其关键是如何定量计算“时空密度”。目前国内外的研究通常采用两种思路,其一是在三维空间内构建时空圆柱体,即 $X$ 轴和 $Y$ 轴代表平面二维空间,而 $Z$ 轴代表时间。时空密度的计算可以通过时空圆柱中点的数目与时空圆柱体积的比值计算得出。这种时空密度的表达方式已被应用于时空聚类的模型中,如时空扫描模型(SaTScan)<sup>[20,32]</sup>和时空密度聚类(ST-DBSCAN)<sup>[33]</sup>等。

虽然时空圆柱体有效解决了时空密度的计算,但在选择时空圆柱体时需要确定若干参数,如圆柱体的半径和高。这给相关的聚类算法引入了更多的主观性,增加了计算的复杂度和不确定性。为此,Pei等提出了加窗邻近距离<sup>[34]</sup>的概念,即将 $K$ 阶邻近距离的计算限制在一定的时间窗内,从而通过加窗邻近距离就可以估算点周围的局部时空密度。在建立加窗邻近距离的概念之后,就可以将空间点过程分解的思路进行移植并加以拓展,从而实现时空点集的丛集模式提取<sup>[34]</sup>。

### 3.6 序列之间相似性的度量

点过程模型的多尺度性一般表现为一个点事件可视为更低层次上的点过程。为了度量较低尺度上点过程之间的相似性,必须定义点过程序列之间的相似性指标。为此,Pei等提出了针对时间点过程的相似性度量方法及其丛集模式的提取思路<sup>[35]</sup>,即将某个时间点过程视为一个“对象”,利用Frechet距离<sup>[36]</sup>衡量“对象”之间的相似性,最终将丛集的对象类提取出来。该方法被应用于美国个人迁移的轨迹数据(个人每次迁移的时间点就构成了相应的时间点过程),最终发现了个人迁移若干个有趣的模式。当然,衡量点过程之间相似性的方法除了Frechet距离,还有DTW距离<sup>[37]</sup>、编辑距离<sup>[38]</sup>、Hausdorff距离<sup>[39]</sup>等。

## 4 时空点过程的分解理论

由于地学过程的成因十分复杂,因此,针对点

过程数据的模式提取也非常困难。而要提出一个较为通用的理论模型必须满足以下3个条件:(1)模型必须可以处理多个密度不同的点过程;(2)聚类模型可以挖掘任意形状的丛集模式;(3)聚类模型必须是客观的过程,并尽量减少对先验知识的依赖。目前针对空间点集数据的模式提取方法虽然不少(主要为针对点集的密度聚类方法,如前文所述的基于单元、基于模型和基于距离的聚类模型等),但或多或少都存在一些缺陷,以至于无法同时满足上述3个条件。具体地说,基于单元的模型对于单元的形状和大小较为敏感,不同形状和大小的单元会导致不同的聚类结果;基于距离的方法很难确定数据集中点过程的数目(每个点过程的密度均不相同),而这往往需要更多的先验知识;基于模型的方法借助于特定的模型(如高斯模型),因此对于任意形状的丛集模式往往显得力不从心。

本文提出的时空点过程的层次分解理论框架与前人的研究思路有所不同。该理论借鉴了傅里

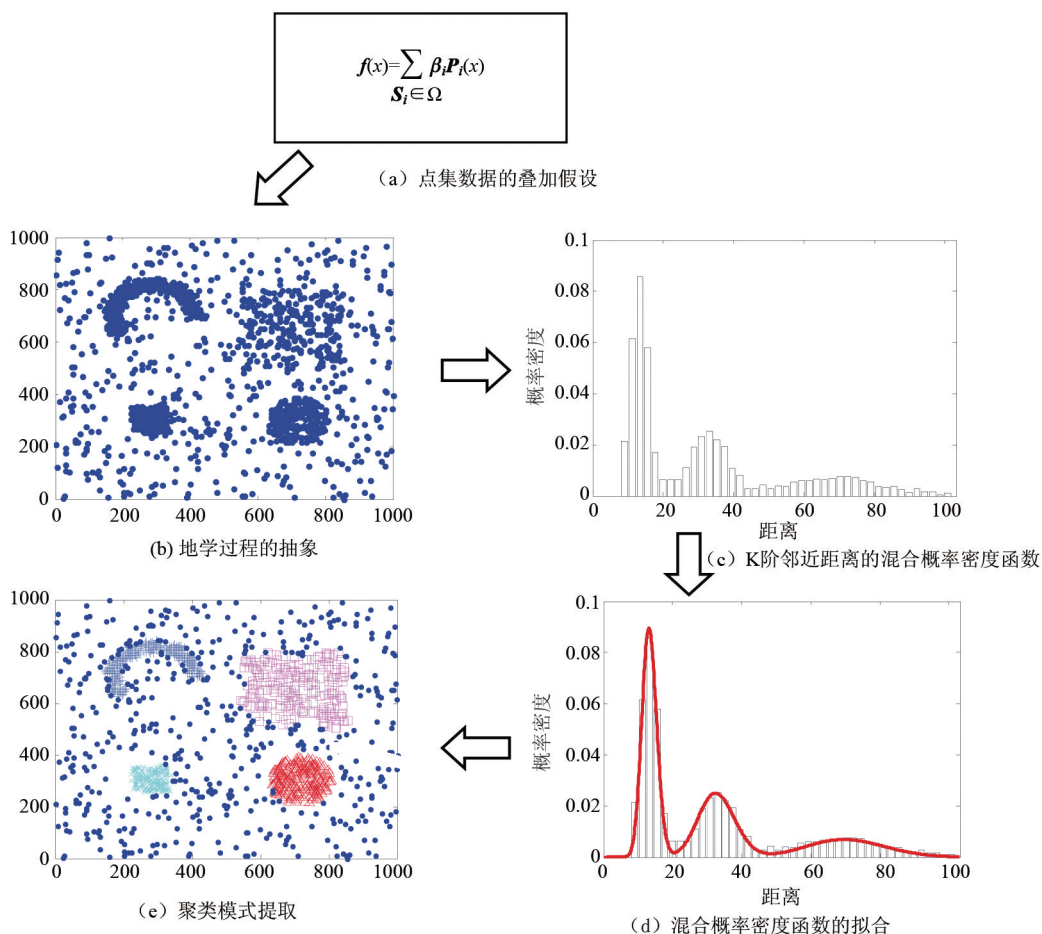


图2 点过程数据“傅里叶变换”的流程

Fig.2 Diagram of “Fourier transform” of point process data

叶变换和小波变换的思想处理点过程数据。具体为:首先,假设任意点集数据是有限多个均匀过程(类似傅里叶变换中的基函数)的叠加,例如:某地区的地震可视为板块背景地震过程、区域断裂带地震过程和某断层地震过程的叠加,图2(a)中,任意点集数据  $f(x)$  可表达为有限多个均匀点过程  $P_i(x)$  的叠加,即  $f(x) = \sum_{S_i \in \Omega} \beta_i P_i(x)$ , 其中,  $S_i$  为均匀过程  $P_i(x)$  的支撑域,  $\beta_i$  为  $P_i(x)$  在整个点集数据中所占的比重,  $\Omega$  为整个研究域;其次,将研究对象转换为点过程,并通过指标判别点过程中是否包含丛集点(如包含丛集点则进行后续的研究,否则就认为原数据是单一的均匀过程)(图2(b));再次,通过局部密度表达工具将时空点集转换为混合概率密度函数(图2(c));然后,利用噪声和丛集点的分解模型将混合密度分解为若干相对均匀的点过程(图2(d));最后,从所得的均匀点过程中抽取不同尺度上的丛集模式及其特征(图2(e))。上述理论可以视为时空点过程数据的“傅里叶变换”。它不仅可以确定点集数据中点过程的类型及其数目,同时输入参数达到最少(如采用K阶邻近距离作为点局部密度的表达工具,输入的参数仅为K),除此之外还可以胜任任意形状的丛集模式。所以,我们可以将该理论视为通用的点集数据的模式提取工具。

## 5 结论与展望

时空点过程模型是一种描述点事件现象的数据模型,并表现出抽象、离散和层次性等特征,正是由于具备这些特征使得点过程模型成为解决地学问题的有力工具。与栅格计算模型不同,点过程模型更加注重地学过程中的突变部分,这种特殊性又使其成为观察地学问题的一种新视角。为了将这一模型应用于地学现象的研究,我们根据多年的研究积累提出了点过程数据的层次分解理论。该理论吸纳了信号处理中谱分析的理论精髓,实现了从任意点集数据中提取模式,从而初步确立了点过程数据的“傅里叶变换”。

对于未来点过程模型的研究,有两个方向值得注意:其一,当点集中出现了性质不同的两种点过程时(例如,出租车的起点和终点可视为两种性质不同的点过程),如何提取包含这两种点过程的丛集模式;其二,当研究对象的基本单元本身就为时空点过程时(如居民使用手机的通话记录、个人迁

移历史所形成的轨迹等),如何提取其中(尤其是针对因无线通讯和网络技术的广泛使用而产生的海量时空轨迹数据)的模式具有非常重要的理论意义和应用前景。

### 参考文献:

- [1] 邓永录,梁之舜.随机点过程及其应用[M].北京:科学出版社,1998.
- [2] 杨萍,侯威,支蓉.利用空间点过程提取丛集点算法的适用性研究[J].物理学报,2009,58(3):2097-2105.
- [3] 李长江,徐有浪,蒋叙良.论矿床的分形性质[J].浙江地质,1994(2):25-31.
- [4] Pei T, Gong X, Shaw S L, *et al.* Clustering of temporal event processes[J]. International Journal of Geographical Information Science, 2013,27(3):484-510.
- [5] Illian J, Penttinen A, Stoyan H, *et al.* Statistical analysis and modelling of spatial point patterns[M]. West Sussex: John Wiley & Sons Ltd, 2008.
- [6] Cressie N. Statistics for spatial data[M]. New York: John Wiley & Sons,1993.
- [7] Lloyd M. Mean crowding[J]. The Journal of Animal Ecology, 1967,36(1):1-30.
- [8] Douglas J B. Clustering and aggregation[J]. Sankhyā: The Indian Journal of Statistics, Series B, 1975,37(4):398-417.
- [9] Assuncao R. Testing spatial randomness by means of angles [J]. Biometrics, 1994,50(2):531-537.
- [10] Trifković S, Yamamoto H. Indexing of spatial patterns of trees using a mean of angles[J]. Journal of Forest Research, 2008,13(2):117-121.
- [11] Eberhardt L L. Some developments in distance sampling [J]. Biometrics, 1967,23(2):207-216.
- [12] Johnson R B, Zimmer W J. A more powerful test for dispersion using distance measurements[J]. Ecology, 1985,66(5):1669-1675.
- [13] Prayag V R, Deshmukh S R. Testing randomness of spatial pattern using Eberhardt's index[J]. Environmetrics, 2000,11(5):571-582.
- [14] Lucio P S, Brito N L C. Detecting randomness in spatial point patterns: A “Stat-Geometrical” alternative[J]. Mathematical Geology, 2004,36(1):79-99.
- [15] Muggleston M A, Renshaw E. Spectral tests of randomness for spatial point patterns[J]. Environmental and Ecological Statistics, 2001,8(3):237-251.
- [16] Ripley B D. Modelling spatial patterns[J]. Journal of the Royal Statistical Society Series B (Methodological), 1977, 39(2):172-212.
- [17] Diggle P J, Gates D J, Stibbard A. A nonparametric esti-



- mator for pairwise-interaction point processes[J]. *Biometrika*, 1987,74(4):763-770.
- [18] Schiffrers K, Schurr F M, Tielbörger K, *et al.* Dealing with virtual aggregation—a new index for analyzing heterogeneous point patterns[J]. *Ecography*, 2008,31(5):545-555.
- [19] Pei T. A non-parameter index for differentiating between heterogeneity and randomness[J]. *Mathematical Geosciences*, 2011(43): 345-362.
- [20] Kulldorff M. A spatial scan statistic[J]. *Communications in Statistics-Theory and methods*, 1997,26(6):1481-1496.
- [21] Sheikholeslami G, Chatterjee S, Zhang A. Wavecluster: A multi-resolution clustering approach for very large spatial databases[C]. *Proceedings of the 24th International Conference on Very Large Data Bases*, New York City, 1998, 428-439.
- [22] Allard D, Fraley C. Nonparametric maximum likelihood estimation of features in spatial point processes using Voronoi tessellation[J]. *Journal of the American Statistical Association*, 1997,92(440):1485-1493.
- [23] Byers S, Raftery A E. Nearest-neighbor clutter removal for estimating features in spatial point processes[J]. *Journal of the American Statistical Association*, 1998,93(442): 577-584.
- [24] Guo D S, Zhu X, Jin H, *et al.* Discovering Spatial Patterns in Origin - Destination Mobility Data[J]. *Transactions in GIS*, 2012,16(3):411-429.
- [25] Fraley C, Raftery A E. How many clusters? Which clustering method? Answers via model-based cluster analysis[J]. *Computer Journal*, 1998(41):578-588.
- [26] Murtagh F, Starck J L. Pattern clustering based on noise modeling in wavelet space[J]. *Pattern Recognition*, 1998, 31(7):847-855.
- [27] Ester M, Kriegel H P, Sander J, *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise[C]. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, 1996,226-231.
- [28] Ankerst M, Breunig M M, Kriegel H-P, *et al.* OPTICS: Ordering points to identify the clustering structure[C]. // Delis A, Faloutsos C, Ghandeharizadeh S (Eds.). *Proc. ACM SIGMOD Int. Conf. on Management of Data*, June 1-3, 1999, Philadelphia, PA, USA. ACM Press, 1999, 49-60.
- [29] Pei T, Zhu A X, Zhou C H, *et al.* A new approach to the nearest - neighbour method to discover cluster features in overlaid spatial point processes[J]. *International Journal of Geographical Information Science*, 2006,20(2): 153-168.
- [30] Pei T, Jasra A, Hand D J, *et al.* DECODE: A new method for discovering clusters of different densities in spatial data[J]. *Data Mining and Knowledge Discovery*, 2009,18(3): 337-369.
- [31] Pei T, Zhu A X, Zhou C H, *et al.* Detecting feature from spatial point processes using Collective Nearest Neighbor [J]. *Computers, Environment and Urban Systems*, 2009,33 (6):435-447.
- [32] Kulldorff M. SaTScan user guide for version 9.0. 2011.
- [33] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial - temporal data[J]. *Data & Knowledge Engineering*, 2007,60(1):208-221.
- [34] Pei T, Zhou C H, Zhu A X, *et al.* Windowed nearest neighbour method for mining spatio-temporal clusters in the presence of noise[J]. *International Journal of Geographical Information Science*, 2010,24(6):925-948.
- [35] Pei T, Gong X, Shaw S L, *et al.* Clustering of temporal event processes[J]. *International Journal of Geographical Information Science*, 2013,27(3):484-510.
- [36] Alt H, Godau M. Computing the Fréchet distance between two polygonal curves[J]. *International Journal of Computational Geometry and Applications*, 1995,5(1): 75-91.
- [37] Sankoff D, Kruskal J B. Time warps, string edits, and macromolecules: The theory and practice of sequence comparison[M]. Addison-Wesley Publishing Company, 1983.
- [38] Crochemore M, Rytter W. Text algorithms[M]. New York, USA: Oxford University Press, 1994.
- [39] Alt H, Guibas L. Handbook on computational geometry [M]. Interpolation, and Approximation-A Survey, 1995, 251-265.

## Spatiotemporal Point Process: A New Data Model, Analysis Methodology and Viewpoint for Geoscientific Problem

PEI Tao\*, LI Ting and ZHOU Chenghu

*(State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China)*

**Abstract:** The gridding computation is a major model in current geoscientific research due to its simplicity in organizing data resources. However, because the gridding computation equally distributes computational resources, it brings redundancy to the computational process and neglects catastrophe points in geoscientific phenomena, which might overlook the important patterns and bring more uncertainties to the research result. To overcome this weakness, this paper proposes to use the spatial point process model in geoscientific research. The spatial point process model is used to model spatial point based geoscientific phenomenon, also is applied to most of the other geoscientific processes (because they can be transformed into spatial point processes). In this regard, the spatial point process is not only a data model, but also an analysis tool for geoscientific problems. Moreover, it provided a new angle of view for observing geoscientific problems. To extract patterns from point process data, the authors propose the frame of multilevel decomposition of spatiotemporal point process. This frame is similar to the basic idea of signal decomposition. We first assume that any point data set is the overlay of an unknown number of homogeneous point processes. Then, the points are transformed into a mixture probability density function of the  $K^{\text{th}}$  nearest distance of each point. After that, the optimization method is used to separate clustering points from noise. Finally, the patterns are extracted using the density connectivity mechanism. The theory can be used to any type of point process data. It can be considered as the “Fourier transform” of point process data.

**Key words:** clustering; data mining;  $K^{\text{th}}$  nearest distance; Poisson process; nonhomogeneous point process

\*Corresponding author: PEI Tao, E-mail: peit@lreis.ac.cn