

空间点格局分析和社会研究

社会
2009 · 5
Society
第 29 卷

时培建

摘要:空间因素长期以来在社会学研究中被忽视。造成这种现象的原因很多,一个可能的原因在于缺少适合于研究空间影响的研究方法。幸运的是,由于许多统计学家卓越的工作,在环境流行病学的研究中,空间点格局分析的方法日渐成熟起来。因为诸多社会学研究对象在一个较大的尺度上存在空间性的特点,所以我们认为可以借用这些分析方法来研究社会问题。本文向社会学界介绍几种空间点格局分析中的经典方法。

关键词:*K* 函数 核估计 空间聚集 空间变异 风险趋势面

空间点格局分析源于植物生态学,用于分析一定距离尺度下植物的空间分布情况。上个世纪 50 年代末至 60 年代初,这种方法被推广到其他研究领域(Gatrell et al., 1996: 256 - 274)。上个世纪 70 年代以来,经过一些学者的发展,这一方法日臻完善。它的基本思路是:划分一定面积的研究区域,在平面图中标出在此研究区域内的所有点事件,通过一定的计算方法,分析点事件在一定距离尺度下是否存在空间聚集的现象。在环境流行病学中,空间点格局分析还被用来探索疾病风险的空间变异情况。由于点事件是对研究对象的一种抽象概括,点事件既可以代表植物、疾病,也可以代表一些社会研究命题,如越轨事件、贫民窟问题,对于充足社会学量化研究是十分有意义的。“时间和空间是社会生活‘环境’,这一观点在某些方面帮助加强了学科的划分,因而时间可能受到历史学家们的极大关注,空间可能受到地理学家们的关注,而社会科学的其他部分则极大地忽略了这些方面。我认为时间和空间对于社会科学是极为基本的问题。”(吉登斯,[2000]2003: 155) 本文旨在对空间点格局分析中的一些具有代表性的方法作一介绍,为

时培建 江西农业大学昆虫学研究所 研究生

英国兰开斯特大学的 Peter Diggle 教授在本文准备过程中提供了极富价值的评论和建议,在此表示诚挚的谢意。

社会学的学者提供一种可供选择的研究方法。

一、基本定义和基本定理¹

定义 1:一个空间点过程的一阶强度函数

$$\lambda(x) = \lim_{|dx| \rightarrow 0} \left\{ \frac{E[N(dx)]}{|dx|} \right\}$$

其中, x 表示平面上任意一点; dx 表示包含点 x 的一个小区域。

定义 2:一个空间点过程的二阶强度函数

$$\lambda_2(x,y) = \lim_{\substack{|dx| \rightarrow 0 \\ |dy| \rightarrow 0}} \left\{ \frac{E[N(dx)N(dy)]}{|dx||dy|} \right\}$$

定义 3:一个空间点过程的协方差密度

$$\gamma(x,y) = \lambda_2(x,y) - \lambda(x)\lambda(y)$$

如果在整个研究区域 A 内, 一阶强度函数是常数 $\lambda(x) = \lambda = E[N(A)/|A|]$, 则称这个点过程是稳定的; 如果二阶强度函数 $\lambda_2(x,y) = \lambda_2(\|x - y\|) = \lambda_2(u)$, 说明其只依赖于事件的方向和距离, 而不依赖于事件的绝对位置。如果进一步假设 u 仅仅表示距离, 则称这个点过程是同向性的。对于一个具有稳定性和同向性的空间点过程而言, 则有:

$$\gamma(u) = \lambda_2(u) - \lambda^2$$

定义 4: 一个稳定的、同向性的空间点过程, 其减少的二阶矩函数为:

$$K(s) = 2\pi\lambda^{-2} \int_0^s \lambda_2(u) u du$$

定理 1: 对一个稳定的、同向性的空间点过程而言, 有

$$K(s) = \lambda^{-1} E(\text{距离任意一事件的长度小于 } s \text{ 的若干事件的数量})$$

其中, s 是距离尺度; $E(\cdot)$ 表示在整个研究区域内距离任意一点的长度小于 s 的其他点数量的数学期望(也可以理解为均值)。

定理 2: 对于一个同质性的平面泊松过程而言, 则有:

$$K(s) = \pi s^2$$

1 此部分内容主要参阅了 Diggle, P.2000 年兰开斯特大学的内部讲稿 Spatial Statistics for Environmental Epidemiology。

同质性的假设很重要,它表示事件的发生是一个完全的随机过程, $K(s)$ 可直观地理解为以一点为圆心,半径为 s 的圆中所有点的集合。在空间点格局分析中,事件究竟是不是完全随机的,就要看它等不等于 πs^2 。于是需要对事件的 $K(s)$ 进行估计。现在,我们以被研究区域内所有事件的数量为 n ,则有:

$$\hat{\lambda} = n / |A|$$

$$\hat{E}(s) = n^{-1} \sum_{i=1}^n \sum_{j \neq i} w_{ij} I(d_{ij} \leq s)$$

其中,字母上的帽子表示估计, $I(\cdot)$ 是指示器函数,如果研究区域内其他点 j 距离一点 i 的距离 d_{ij} 大于指定的距离尺度 s ,则 $I(\cdot) = 0$;反之,则有 $I(\cdot) = 1$; w_{ij} ,是边缘校正权重,它等于整个圆周长与以 i 点为圆心、 d_{ij} 为半径的圆落在研究区域内的弧长的比值。它表示校正圆以外还可能存在未被考虑的点,于是假设校正圆上点发生的概率和所在弧长成比例。因为对于不规则的研究区域求出边缘校正的数学表达式较难,于是在传统研究中,一般把研究区域选为矩形,这样边缘校正权重精确的数学表达式能够得到(Diggle, 1983: 72; 汤孟平等, 2003: 1533 - 1538)。Rowlingson 和 Diggle(1993: 627 - 655)、时培建等(《生态学报》待刊稿)提出了对研究区域为任意多边形边缘校正权重的计算机解法。

根据上式,则有:

$$\hat{K}(s) = \frac{|A|}{n^2} \sum_{i=1}^n \sum_{j \neq i} w_{ij} I(d_{ij} \leq s) \quad (1)$$

二、完全空间随机化方法

完全空间随机化方法是建立在空间同质性的假设基础之上的,在一定距离尺度 s 上,事件的 $\hat{K}(s) - \pi s^2$ 理论值应该为0,然而由于存在估计误差,这个值一般围绕0的两侧波动,利用 Monte-Carlo 方法在研究区域内模拟 r 组与被研究事件数量相同的 n 个随机事件,求出它们的 $\hat{K}(s) - \pi s^2$ 值,保留在一定距离尺度 s 下 r 组中的 $\hat{K}(s) - \pi s^2$ 的最大值和最小值,作为被研究事件隶属完全空间随机的上下限。如果在此距离尺度上被研究事件的 $\hat{K}(s) - \pi s^2$ 位于此区间内,则认为被研究事件是完全随机分布;如果高于此区间的最大值,则认为被研究事件是聚集分布;如果低于此区间的最小值,则认为被研究事件是均匀分布。这

种判断源于 $K(s)$ 的估计式,对于聚集分布而言, $\hat{K}(s)$ 明显大于 πs^2 。在研究中,也可以使用 $\hat{L}(s) = \sqrt{\hat{K}(s)/\pi} - s$ 进行判断(张金屯,1998: 344 - 349)。我们使用俄克拉何马城白人和黑人所进行财产偷窃的案发地点数据(参看图 1)进行示例分析。

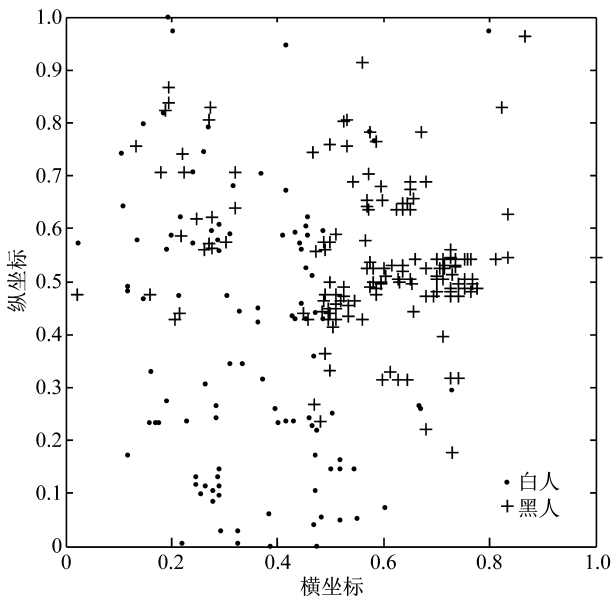


图 1:俄克拉何马城白人和黑人所进行财产偷窃的案发地点¹

三、考虑人口分布不均后的空间聚集的二阶分析

在环境流行病学、社会学等与人口密切相关的空间点格局研究中,事件的发生与人口的空间分布密切相关,然而人口分布并不是均质的,而是呈空间异质性的。显然,完全空间随机化方法对于这些研究是不可行的,于是 Cuzick 和 Edwards (1990: 73 - 104)、Diggle 和 Chetwynd (1991:1155 - 1163)分别设计了基于 case-control 数据的二阶分析方法。他们的方法大致分如下几个步骤。

1 数据来源于 Splancs 软件包自带的数据表 okwhite 和数据表 okblack,原始数据区域横坐标范围[111,382],纵坐标范围[64,341]。为了分析方便,本文将它们都转化在[0,1]区间。这些数据可进一步追溯到 Bailey 和 Gatrell(1995:122 - 125)的文献。

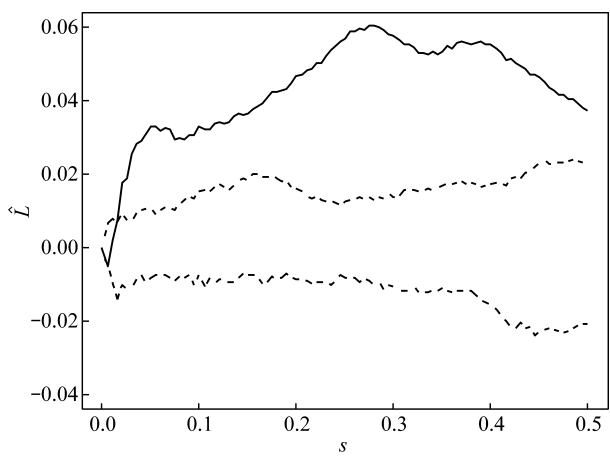


图 2:完全空间随机化方法

注:在完全空间随机化假设条件下仅对白人偷窃的案发地点的分析; s 表示距离尺度,从 0 到 0.5,步长选取为 0.005;Monte-Carlo 模拟的次数为 99 次,虚线部分为包迹线,即由随机标定模拟 99 次得到 $L(s)$ 估计值的上下限;实线部分为观测组 $L(s)$ 的估计值。按照完全空间随机化方法,其结论是白人偷窃的案发地点呈聚集分布。

第一步:划定研究区域,把研究区域内被研究对象点集通称为观测组(cases),同时根据研究区域内分布的人口随机抽取一定的点集作为控制组(controls)。为了保证选取的随机性,选取控制组的方法一般分两种:一是使用研究区域内人口出生登记簿随机抽取他们的居住地,标出位置,即控制组;二是使用研究区域的邮政编码,随机选定邮政编码,因为英国一个邮政编码代表了 1 到 20 投递地址(Prince *et al.*, 2001: 1083 - 1088),划分较细,对于大尺度的研究区域,这种选取方法在英国具有可行性。

第二步:分别计算观测组和控制组的 $K(s)$ 之差,即 $D(s) = K_{11}(s) - K_{22}(s)$ 。根据零假设,如果观测组的空间分布也是随机的,则由观测组和控制组来自同一总体,它们不过都是这一总体中的样本,那么两组的 $K(s)$ 应该是相同的,并且它们和两组混合构成的点集的 $K(s)$ 也应该是相同的。

$$K_{ij}(s) = \lambda_j^{-1} E(\text{距离来自 } i \text{ 组中任意一事件的} \\ \text{长度小于 } s \text{ 的 } j \text{ 组中若干事件的数量})$$

其中, $j = 1, 2$ (分别表示观测组和控制组的点); $K_{11}(s)$ 表示观测组的 $K(s)$; $K_{22}(s)$ 表示控制组的 $K(s)$; $K_{12}(s)$ 表示混合组的 $K(s)$ 。如果零假设成立, 则有:

$$K_{11}(s) = K_{22}(s) = K_{12}(s)$$

它们的估计值分别为:

$$\hat{K}_{11} = \frac{|A|}{n_1(n_1 - 1)} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} w_{ij} I(d_{ij} \leq s) \quad (2)$$

$$\hat{K}_{22} = \frac{|A|}{n_2(n_2 - 1)} \sum_{i=n_1+1}^n \sum_{j=n_1+1}^n w_{ij} I(d_{ij} \leq s) \quad (3)$$

$$\hat{K}_{12} = \frac{|A|}{n(n_1 - 1)(n_2 - 1)} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^n (n_2 w_{ij} + n_1 w_{ji}) I(d_{ij} \leq s) \quad (4)$$

其中, n_1 表示在研究区域内观测组中点的数量; n_2 表示在研究区域内控制组中点的数量; $n = n_1 + n_2$ 。

对于给定的距离尺度 s 和 u , 如果令 $W_{ij} = \frac{1}{2}(w_{ij} + w_{ji}) I_{ij}(d_{ij} \leq s)$, $V_{ij} = \frac{1}{2}(w_{ij} + w_{ji}) I_{ij}(d_{ij} \leq u)$, $W = \sum_{i=1}^n \sum_{j=1}^n W_{ij}$, $V = \sum_{i=1}^n \sum_{j=1}^n V_{ij}$, $X = \sum_{i=1}^n \sum_{j=1}^n W_{ij} V_{ij}$, $Z = \sum_{i=1}^n (\sum_{j=1}^n W_{ij}) (\sum_{k=1}^n V_{ik})$ 。再令 $m^{(r)} = m(m-1)\Lambda(m-r+1)$, 则有:

$$\begin{aligned} Cov\{\hat{D}(s), \hat{D}(u)\} &= \frac{|A|^2}{(n_1^{(2)} n_2^{(2)})^2} \{ (n_2^{(2)})^2 \mu_1 - 2(n_1^{(2)} n_2^{(2)}) \mu_2 \\ &\quad + (n_1^{(2)})^2 \mu_3 \} \end{aligned} \quad (5)$$

其中,

$$\begin{aligned} \mu_1 &= (n_1^{(4)} / n^{(4)}) (WV - 4Z + 2X) + 4(n_1^{(3)} / n^{(3)}) (Z - X) \\ &\quad + 2(n_1^{(2)} / n^{(2)}) X \end{aligned}$$

$$\mu_2 = (n_1^{(2)} n_2^{(2)} / n^{(4)}) (WV - 4Z + 2X)$$

$$\begin{aligned} \mu_3 &= (n_2^{(4)} / n^{(4)}) (WV - 4Z + 2X) + 4(n_2^{(3)} / n^{(3)}) (Z - X) \\ &\quad + 2(n_2^{(2)} / n^{(2)}) X \end{aligned}$$

第三步: 随机标定观测组和控制组, 即在由观测组和控制组共同构成的点集中随机抽取 n_1 个点作为新的观测组, 剩余的 n_2 个点作为新的控制组, 分别计算它们的 $\hat{K}_{11}(s) - \hat{K}_{22}(s)$, 反复进行 r 次, 保留一定距

离尺度上的上下限,分别作为判定第二步结果是否存在空间聚集的标准。如果结果高于上限,则观测组在一定距离尺度上存在空间聚集;如果结果低于下限,则观测组在一定距离尺度上为均匀分布;结果位于上下限之间,则观测组在一定距离尺度上随机分布。

另外,Diggle 和 Chetwynd 还提供了一种对观测组是否存在空间聚集的总体检验方法,即:

$$D = \sum_{k=1}^m \frac{\hat{D}(s_k)}{\sqrt{\text{Var}\{\hat{D}(s_k)\}}}$$

其中, $\text{Var}\{\hat{D}(s_k)\}$ 根据协方差公式求出。检验时,首先将距离尺度 s 划分为 m 个离散点 s_k , 求出已知的观测组和控制组的 D , 不妨令其为 D_0 。通过随机标定的方法, 设随机标定了 r 次, 计算这 r 次的 D , 令其为 $D_i (i = 1, 2, \Delta, r)$, 设 D_0 在由其自身和 $D_i (i = 1, 2, \Delta, r)$ 构成的降序数列中排在第 r_0 位, 则规定总体检验的 p 值等于 $r_0 / (r + 1)$ (Barnard, 1963: 294; Besag & Diggle, 1977: 327 - 333)。如果 p 值 ≤ 0.05 , 则认为观测组

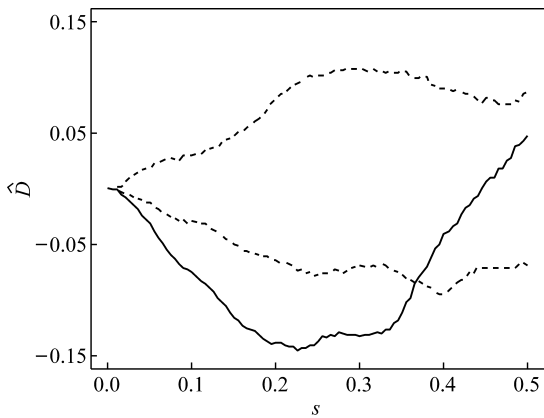


图 3: 基于 case-control 数据的二阶分析方法

注: 以白人偷窃的案发地点为观测组, 以黑人偷窃的案发地点为控制组; s 表示距离尺度, 从 0 到 0.5, 步长选取为 0.005; Monte-Carlo 模拟的次数为 99 次, 虚线部分为包迹线, 即由随机标定模拟 99 次得到 $D(s)$ 估计值的上下限; 实线部分为观测组 $D(s)$ 的估计值。从图中可以看出在 $s < 0.37$ 的距离尺度下, 白人偷窃案发地点呈均匀分布, 位于包迹线之下; 而在 $s \geq 0.37$ 的距离尺度上, 白人偷窃的案发地点呈随机分布, 位于包迹线之内。这个结论和完全空间随机化方法得到的结论显然是不同的, 产生的原因在于完全空间随机化方法没有使用控制组。

空间聚集显著,否则认为这种空间聚集不显著。

四、事件发生风险的空间变异之衡量方法

在空间点格局研究中,除了建立在距离尺度上的方法之外,建立在面积上事件发生的风险分析也同样备受关注(Gatrell *et al.*, 1996:256-274),因为通过这种方法,事件发生风险趋势面(risk surface)可以直观地反映这种风险。假设观测组和控制组在研究区域内的强度函数分别为 $\lambda_1(x)$ 和 $\lambda_2(x)$,¹ 用对数风险函数来衡量事件发生风险在研究区域内的空间变异,规定其为:

$$\rho(x) = \log\{\lambda_1(x)/\lambda_2(x)\}$$

这个公式中 $\rho(x)$ 很难通过 $\lambda_1(x)$ 和 $\lambda_2(x)$ 直接求出,因为后者都是不知道的。需要估算 $\rho(x)$ 的方法很多(参见 Kelsall & Diggle, 1998: 559-573)。本文主要介绍二项回归(binary regression)的方法。

使用标签 y_1, y_2, Λ, y_n 对由观测组和控制组构成的点集 x_1, x_2, Λ, x_n , 规定:

$$y_i = \begin{cases} 1 & i \in [1, n_1], \\ 0 & i \in [n_1 + 1, n] \end{cases}$$

依赖于点 x_i , 则 y_i 实际上是相互独立的伯努力随机变量 Y_i 的结果, 而 $P(Y_i = 1 | X_i = x) = p(x)$, 此处

$$p(x) = \frac{q_1 \lambda_1(x)}{q_1 \lambda_1(x) + q_2 \lambda_2(x)}$$

那么,继而能够得到:

$$\ln\{p(x)/1-p(x)\} = \rho(x) + \ln(q_1/q_2)$$

可以看出,除了常数项 $\ln(q_1/q_2)$, $\text{logit}\{p(x)\}$ 近似等于 $\rho(x)$, 而 $p(x)$ 可以通过核方法进行估计(Hastie *et al.*, [2001]2003:115-134), 选用 Nadaraya-Watson 核回归估计值, 则有:

$$\hat{p}_h(x) = \sum_{i=1}^n [\kappa_h(x-x_i) y_i] / \sum_{i=1}^n \kappa_h(x-x_i)$$

此处,核估计函数 $\kappa_h(u) = \frac{1}{2\pi h^2} \exp\left(-\frac{\|u\|^2}{2h^2}\right)$ 。 h 为带宽,它对求解

1 此处 $\lambda_2(x)$ 不再表示二阶强度函数,而是指控制组的一阶强度函数。

$p_h(x)$ 具有明显的影响, h 的根据最大似然函数进行选择。因为 0—1 分布的分布律是已知的(薛毅、陈立萍, 2007: 11), 所以概率函数的似然为:

$$L\{p(\cdot)\} = \prod_{i=1}^n p(x_i)^{y_i} \{1 - p(x_i)\}^{1-y_i}$$

通过极大似然就可得到最优带宽 h 。但当出现 $p(\cdot) = 1$ 或 $p(\cdot) = 0$, 这种方法难以达到满意的效果。为了解决这个问题, Kelsall 和 Diggle (1998: 559 - 573) 提出用交叉验证来改进最大似然。

$$CV_2(h) = \left[\prod_{i=1}^n \hat{p}_h^{-i}(x_i)^{y_i} \{1 - \hat{p}_h^{-i}(x_i)\}^{1-y_i} \right]^{-1/n}$$

其中, $\hat{p}_h^{-i}(x_i)$ 表示在计算过程中不计算 $p(x)$ 在点 x_i 的值, 也就是说, 对于连乘积中当运行到第 i 次时, 在计算 $p_h(x)$ 时“挖掉”其中 $j = i$ 时的 $x_j (j = 1, 2, \Delta, n)$, 即:

$$\hat{p}_h^{-i}(x_i) = \sum_{j=1, j \neq i}^n [\kappa_h(x_i - x_j) y_i] / \sum_{j=1, j \neq i}^n \kappa_h(x_i - x_j)$$

当 $h \rightarrow +\infty$ 时, 由上式明显可以得到:

$$\hat{p}_h^{-i}(x_i) = \begin{cases} (n_1 - 1)/n & i \in [1, n_1] \\ n_1/(n - 1) & i \in [n_1 + 1, n] \end{cases}$$

于是, 当 $h \rightarrow +\infty$ 时:

$$CV(+\infty) = (n - 1)(n_1 - 1)^{-n_1/n} (n_2 - 1)^{-n_2/n}$$

如果规定:

$$CV(h) = CV_2(h) / CV(+\infty) \quad (6)$$

显然如果 h 增大时, $CV(h)$ 最终会等于 1。而如果存在较小的 h 使得 $CV(h) < 1$, 那么存在一个 h 使得 $CV(h)$ 最小, 此时即为最优带宽。

根据上述结论, 使用核二项回归的方式, 可以用 $\logit\{p_h(x)\}$ 代替 $\hat{\rho}(x)$, 形成风险趋势面。为了更清晰地反映风险的空间变异, 令

$$\hat{s}(x) = \hat{\rho}(x) - n^{-1} \sum_{i=1}^n \hat{\rho}(x_i)$$

这样, $\hat{s}(x)$ 同样能够反映风险的空间变异, 并且它在观测位置的均值为 0。然而由于存在随机波动, 在风险趋势面上会产生局部的波峰和凹槽, 为了避免过度解释这些随机产生的波峰和凹槽, 一种检验方法被提了出来。

同样对观测组和控制组进 r 次随机标定,零假设认为 $H_0: \rho(x) = c$ 。将上面求出的最优带宽 h 代入,求出 $\hat{s}_i(x) (i = 1, 2, \Delta, r)$ 。根据观测数据求出的 \hat{s} 定义为 $\hat{s}_0(x)$, 根据 $\hat{s}_0(x)$ 在由 $\hat{s}_i(x)$ 构成的有序数列的位置, 将风险表面划分为 3 个部分: p 值小于等于 0.025 的部分; p 值大于 0.975 的部分; 介于前两个 p 值之间的部分。在风险趋势面中画出 0.025 和 0.975 的 p 值等值线(需要在研究区域内每个点上计算 $\hat{s}_i(x)$)。这样, 随机产生的波峰和凹槽对于分析风险表面的空间变异的影响将大大降低。

除此而外, 还需要执行 Monte-Carlo 检验来衡量风险总体上偏离零假设的显著性程度, 根据下式:

$$t_j = n^{-1} \sum_{i=1}^n \hat{s}_j(x_i)^2$$

使用 $p = r_0 / (r + 1)$ 来求 p 值, 其中 r_0 表示 t_0 在由其自身和 $t_j (j = 1, 2, \Delta, r)$ 构成的降序数列中的位置。

五、含有位置隐含变量的空间点格局分析

在以上几种分析方法中都忽略了点事件在一定位置受到其他因素的影响, 影响事件发生的变量未被充分利用, 而观测事件的空间变异性极有可能受到这些变量的影响。例如把某一城市作为研究区域, 把一定时间内在该城市发生的自杀行为作为事件, 上述介绍的方法仅仅利用了自杀事件发生的空间地理位置, 并没有考虑在自杀地点存在的位置隐含变量, 例如自杀者的家庭收入、职业声望、家庭组成等因素。因此, 如果在空间点格局分析中能把这些位置隐含变量考虑进去, 似乎更为合理。对于许多社会学学者而言, 他们可以根据自己的职业判断, 分析这些事件发生的潜在影响因素, 发掘位置隐含变量, 使分析的结果更具解释力。幸运的是, 近年来空间点格局分析方法进一步发展, 通过分析中引入广义线性模型或广义相加模型, 这些可能影响事件发生的位置隐含变量得到了重视。

Baddeley 等人(2000: 329 - 350)假设观测组的一阶强度函数已知, 减少的二阶矩函数的估计值为:

$$\hat{K}_l(s, \lambda) = \frac{1}{|A|} \sum_{i=1}^n \sum_{j \neq i} \frac{w_{ij} I(d_{ij} \leq s)}{\lambda(x_i) \lambda(x_j)} \quad (7)$$

其中, $\lambda(x)$ 的估计式如下:

$$\hat{\lambda}_h(x) = \sum_i^n \kappa_h(x - x_i) / C_h(x)$$

$\kappa_h(\cdot)$ 和上边提到的 $\kappa_h(u)$ 函数形式相同; $C_h(x) = \int_A \kappa_h(x - u) du$, 表示边缘校正, 由 Berman 和 Diggle (1989) 提出。Baddeley 等人 (2000) 的研究显示, 使用上式去估计 K 函数是严重有偏的, 然而通过下式再配合一定精挑细选的带宽, 这种估计的有偏性可以被消除。

$$\hat{\lambda}_h(x_i) = \sum_{i \neq j} \kappa_h(x_j - x_i) / C_h(x_j)$$

Diggle 等人 (2007: 550 - 557) 扩展了 Baddeley 等人的研究, 建立起一种既包含对控制组一阶强度函数核估计 (非参数估计), 又包含对位置隐含变量的参数 (如果使用广义线性模型来估计) 或非参数估计 (如果使用广义相加模型来估计) 的分析方法。

假设 λ_1 代表观测组的一阶强度函数, λ_2 代表控制组的一阶强度函数。首先根据控制组的数据使用核方法估计 λ_2 , 但是 λ_1 并不使用观测组的数据直接估计, 而是寻找 λ_1 和 λ_2 之间的关系, 通过 λ_2 间接估计出 λ_1 。一个最简单的假设是两者呈比例模型, 即:

$$\lambda_1 = b\lambda_2$$

b 反映了观测组和控制组点数量之间的比例。这个简单的假设可以进一步扩展, 使其能够包含对位置隐含变量的讨论。把比例常数换成含有位置隐含变量的函数。广义线性模型 (Nelder & Wedderburn, 1972: 370 - 384) 和广义相加模型 (Hastie & Tibshirani, 1986: 297 - 318) 被建议采用。如果使用广义线性模型, 则有:

$$\lambda_1 = \lambda_2 \exp\{\alpha + z(x)\beta\} \quad (8)$$

其中 $z(x)$ 代表位置隐含变量的诸个变量构成的变量数组 (z_1, z_2, Λ, z_r), 相应的 $\beta = (\beta_1, \beta_2, \Lambda, \beta_4)$ 。如果只考虑一个位置隐含变量 $z_1(x)$, 则有 $z(x) = z_1(x), \beta = \beta_1$ 。可以看出, 使用广义线性模型来估计 λ_1 , 所有位置隐含变量的解释性较好, 因为它是一种参数估计, 参数 α 和 β 都可以拟合得到。但是一般而言, 广义线性模型的拟合效果较广义相加模型差。如果使用广义相加模型, 则有:

$$\lambda_1 = \lambda_2 f(z(x); \beta) \quad (9)$$

其中, $f(\cdot)$ 是未指定的光滑函数, 此处必须是非负的。由于广义相加

模型是一种非参数估计,虽然拟合效果较好,但是模型的解释性不强。比较广义线性模型和广义相加模型来说,各有长短。它们对于拟合 logit 函数都十分有效。现在将观测组和控制组合为一个点集,根据此点集中的点是否属于观测组,设定标志变量 Y_i 等于 1 或 0。假设这个点集取自异质性的泊松过程, Y_i 彼此独立且有:

$$P(Y_i = 1 \mid z_i) = \frac{\exp(\alpha + z'_i \beta)}{1 + \exp(\alpha + z'_i \beta)} \quad (10)$$

$P(Y_i = 1 \mid z_i)$ 表示某点隶属观测组的概率。此处使用的是广义线性模型。把求得的参数 α 和 β 代入(8)式,再根据(7)式求得 K 函数。如果使用的是广义相加模型, $P(Y_i = 1 \mid z_i) = f(z_i) / \{1 + f(z_i)\}$, 则把估计的 $f(\cdot)$ 代入(9)式,再根据(7)式求得 K 函数。

当我们要判断观测组在研究区域内一定距离尺度下是否存在聚集,同样需要建立零假设,此处零假设认为观测组是一个异质性的泊松过程。如果观测组较零假设而言具有更强的聚集性,直接使用观测组数据进行核估计则可能仅仅把这种聚集性归咎于在强度函数上表现出的空间变异,而忽略了位置隐含变量导致的聚集性。因此,间接估计 $\lambda_1(x)$ 较为合理。为了分析观测组在异质性空间的聚集性,我们同样需要对零假设进行多次 Monte-Carlo 模拟。在由观测组和控制组合并的点集中,按照公式(10)求出的概率任选 n_1 个点作为新的观测组,其余 n_2 个点作为新的控制组。再根据公式(8)和已经得到的相关参数,求出若干次模拟得到的 K 函数。现在我们设这若干次模拟求得的 K 函数的均值和方差分别为 $E(s)$ 和 $V(s)$, 则检验观测组聚集显著程度的公式如下:

$$T_q = \sum_{k=1}^m \frac{\hat{K}_I^q(s_k; \hat{\lambda}_2) - E(s_k)}{\sqrt{V(s_k)}}$$

其中, $q = 1, 2, \Delta, r$; $\hat{K}_I^q(s_k; \hat{\lambda}_2)$ 表示第 q 次模拟中在距离尺度 s_k 上的 K 函数值。如果距离尺度是连续的, 则有 $T_q = \int_0^{s_0} \{\hat{K}_I^q(s; \hat{\lambda}_2) - E(s)\} / \sqrt{V(s)} ds$ 。我们将未做模拟之前求得的 T 为 T_0 , p 值等于 $r_0 / (r + 1)$, r_0 表示 T_0 在由其自身和 T_q 构成降序数列中的排序。仍以 0.05 为判定标准, 此处不再赘述。

六、空间点格局的时空分析

上述方法均假定研究区域内的事件是在一定时间中发生,在这段时间中不存在时间效应,即事件发生的时间先后对事件聚集的空间构成不产生影响,在合适的时间段内,这种假定是没有问题的,但是如果事件发生的时间先后差异很大,需要在空间点格局分析中考虑时间因素的影响。Diggle 等人(1995:124-136)定义了“时空” K 函数:

$$K(s, t) = \lambda^{-1} E(\text{距离任意一事件长计小于 } s, \\ \text{时间小于 } t \text{ 的若干事件的数量})$$

其中,强度 λ 被定义为在单位空间、单位时间上预期事件的数量。在有限的时空 $A \times (0, T)$ 中,如果时间和空间是相互独立的,则有:

$$K(s, t) = K_1(s)K_2(t) \quad (11)$$

此处, $K_1(\cdot)$ 和 $K_2(\cdot)$ 分别指空间过程的 K 函数和时间过程的 K 函数。它们分别为:

$$K_1(s) = \lambda_1^{-1} E(\text{距离任意一事件长度小于 } s \text{ 的若干事件的数量})$$

$$K_2(s) = \lambda_2^{-1} E(\text{距离任意一事件长度小于 } s \text{ 的若干事件的数量})$$

λ_1 和 λ_2 分别是空间强度和时间强度, $\lambda_1 = \lambda T$, $\lambda_2 = \lambda |A|$ 。这三个 K 函数相应的估计式为:

$$\begin{aligned} \hat{K}(s, t) &= \frac{|A| T}{n(n-1)} \sum_i^n \sum_{j \neq i} \omega_{ij} I(d_{ij} \leq s) v_{ij} I(u_{ij} \leq t) \\ \hat{K}(s) &= \frac{|A|}{n(n-1)} \sum_i^n \sum_{j \neq i} \omega_{ij} I(d_{ij} \leq s) \\ \hat{K}(s, t) &= \frac{T}{n(n-1)} \sum_i^n \sum_{j \neq i} v_{ij} I(u_{ij} \leq t) \end{aligned} \quad (12)$$

这里, ω_{ij} 是用以估计空间二阶属性的边缘校正权重,与上文规定的相同,计算中十分重要; v_{ij} 是用以估计空间二阶属性的边缘校正权重,它等于 $2u_{ij}$ 除以以 i 为中点长度为 $2u_{ij}$ 的线段落在 $(0, T)$ 中的长度。因为后者在实际计算中不太重要,所以可以近似计算,只要以 i 为中点长度为 $2u_{ij}$ 的线段完全落在 $(0, T)$ 中,规定 $v_{ij} = 1$, 否则, $v_{ij} = 2$ 。

现在面临的问题是如何诊断时间和空间是否存在着依赖关系。为了解决这个问题, Diggle 等人首先定义了下边的两个重要函数:

$$\begin{aligned}\hat{D}(s, t) &= \hat{K}(s, t) - \hat{K}_1(s)\hat{K}_2(t) \\ \hat{D}_0(s, t) &= \hat{D}(s, t) / \{\hat{K}_1(s)\hat{K}_2(t)\}\end{aligned}\quad (13)$$

显然, $\hat{D}(s, t)$ 可以看做是 s 与 t 的函数, 那么通过绘制 $\hat{D}(s, t)$ 的三维图 (或者关于两个变量的等值图) 就可以用以诊断时空聚集 (space-time clustering)。十分明显 $\hat{D}(s, t) = 0$, 当 s 或者 t 增加时, 如果时间空间存在依赖关系, 则 $\hat{D}(s, t)$ 也会随之增加; 而此时 $\hat{D}(s, t)$ 的样本波动也会随之增加, 于是这个函数只有在 s 与 t 较小时才能用来判断这种依赖关系。注意此处 $\hat{D}(s, t)$ 的物理意义不是十分明确, 然而 $\hat{D}_0(s, t)$ 的物理意思则十分明确。因为 $\lambda K(s, t)$ 代表的是距离任意一个给定点长度小于 s 时间小于 t 的所有点期望的数量, 当存在时空互动时, 对比于不存在时空互动时的时空结构, $\hat{D}(s, t)$ 呈比例增长, $\hat{D}_0(s, t)$ 就表示了时空互动导致的过量风险 (excess risk)。

现在设 $Q(s, t) = \sum_{i \neq j} w_{ij} v_{ij} I(d_{ij} \leq s) I(u_{ij} \leq t)$ 。对于给定的 s 与 t , 现规定 $W_{ij} = \frac{1}{2}(w_{ij} + w_{ji}) I(d_{ij} \leq s)$, $V_{ij} = \frac{1}{2}(v_{ij} + v_{ji}) I(u_{ij} \leq t)$; 同时规定所有的 $W_{ii} = 0$, 所有的 $V_{ii} = 0$ 。如果用 s', t' 分别代替 s, t , 那么就用 W'_{ij}, V'_{ij} 分别替代 W_{ij}, V_{ij} 。再规定 $W_1 = \sum_{i=1}^n \sum_{j=1}^n W_{ij}$, $W'_1 = \sum_{i=1}^n \sum_{j=1}^n W'_{ij}$, $W_2 = \sum_{i=1}^n (\sum_{j=1}^n W_{ij}) (\sum_{k=1}^n W'_{ik})$, $W_3 = \sum_{i=1}^n \sum_{j=1}^n W_{ij} W'_{ij}$ 。与之相似, 根据 V_{ij} 和 V'_{ij} 可以规定出 V_1, V'_1, V_2, V_3 。最后, 规定 $n_r = n(n-1)\cdots(n-r+1)$ 。对时间点进行随机排列, 在零假设 (即不存在时空互动) 下可以得到:

$$E\{\hat{D}(s, t)\} = 0$$

$$\begin{aligned}Cov\{\hat{D}(s, t), \hat{D}(s', t')\} &= \frac{|A|^2 T^2}{n_2^2} \\ &\left\{ \frac{(W_1 W'_1 - 4W_2 + 2W_3)(V_1 V'_1 - 4V_2 + 2V_3)}{n_4} \right. \\ &\quad \left. + \frac{4}{n_3} (W_2 - W_3)(V_2 - V_3) + \frac{2W_3 V_3}{n_2} - \frac{W_1 V_1 W'_1 V'_1}{n_2^2} \right\} \quad (14)\end{aligned}$$

既然 $\hat{D}(s, t)$ 和 $\hat{D}(s', t')$ 的协方差公式已知, 那么很容易就可以得到 $Var\{\hat{D}(s, t)\}$ 。现在以 $\hat{K}_1(s)\hat{K}_2(t)$ 为横坐标, 以 $R(s, t) = \hat{D}(s, t)/Var\{\hat{D}(s, t)\}$ 为纵坐标绘制图形, 进而用来分析时空互动, 这种方

法十分类似于回归模型诊断中使用拟合值与标准化残差作图的方法。它的优点是更加直观,因为分析所使用的图形是二维图,而用 $\hat{D}(s, t)$ 、 $\hat{D}_0(s, t)$ 分析所使用的图形则是三维图。

利用 $R(s, t)$, 还可以从总体上判断时空互动的影响为正还是为负。现在我们设:

$$U = \sum_s \sum_t R(s, t) \quad (15)$$

其中, s 表示规定的空间距离尺度, t 表示规定的时间尺度, U 的正负就代表时空互动的正负。基于上述公式(15), 我们同样可以通过 Monte-Carlo 模拟来检验零假设(即不存在时空互动)。具体方法是, 通过对时间点观测值的随机排列, 假设共进行了 r 次排列, 根据公式(14)和公式(15), 计算出这 r 次排列得到的 U 值, 让其构成降序数列 $U_i (i=1, 2, 3, \dots, r)$, 再令使用观测值(包含观测的空间数据和时间数据)计算的 U 值为 U_0 。现在看 U_0 在降序数列中的位置, 如果排在第 r_0 位, 则 p 值就为 $r_0/(r+1)$, 进而就可以根据 p 值来判断零假设是否成立了。

七、讨论

谢宇(2006: 9-28)认为, 社会科学研究存在着三个基本原理: 变异性原理, 社会分组原理和社会情境原理。而在空间点格局分析中, 对变异的研究, 可以说是研究的主线, 而社会分组和社会情境也得到了体现。在简单的完全空间随机化方法中, 通过空间完全随机化假设, 假定观测数据在一定距离尺度下是随机分布的, 通过 Monte-Carlo 方法设定了完全随机分布的上下包迹线(即模拟结果的上下限), 落在包迹线之内与假设不背离, 符合空间完全随机化的假设, 然而如果落在包迹线的上方则是对零假设的背离, 聚集分布对随机分布而言, 显然是一种变异, 变异通过曲线之间的位置关系得到证实, 最后对事件聚集进行判断。值得注意的是, 空间点格局分析中对聚集的判断, 并不是直接设定判断聚集的指标, 即零假设不直接认为事件呈空间聚集状态, 而是假定事件是随机分布, 聚集只是对随机的一种变异。其他几种分析方法也都体现了通过研究变异而间接研究聚集的这种思想, 在对事件发生风险的空间变异分析中, 变异思想得到了更为直接的体现, 设计了非常精

巧的检验方法就是去消除随机波动对分析的影响。

社会分组的原理在空间点格局分析中也得到了体现。由于在一定面积的研究区域内人口分布往往不是均匀的,即人口密度分布不均,单纯使用空间完全随机化假设就显得不太合理,因为这种假设忽略了事件与人口分布的联系。诸如在社会学研究中,社会事实的分析离不开对人的分析,一定的社会学命题(尤其是在城市社会学研究中)显然需要考虑人口分布的特点。于是巧妙地安排零假设成为决定这种分析方法能否适用于分析与人口分布相关联的事件空间聚集状况的关键。根据现有的人口分布,随机在其中抽出一定的位置作为控制组,零假设认为观测组和控制组来自同一总体,于是如果观测组在考虑了人口分布的因素以后还存在空间聚集,则结果一定会超出根据随机标定方法计算了 $\hat{D}(s)$ 值包迹线的上限。当然,这种社会分组较一般研究中的社会分组更为复杂,因为一般研究中的社会分组可能是根据人种、国籍、年龄、性别、阶层、收入、职业等指标进行的,而在空间点格局分析中,这种分组主要考虑的是事件可能发生的空间分布,控制组来自可能发生与观测组相同事件的人口中。

“群体变异性的模式会随社会情境的变化而变化,这种社会情境常常是由时间和空间来界定的。”(同上)在空间点格局分析中,根据研究对象的特点,需要划分好研究区域。聚集是在一定距离尺度上而言的,离开了对空间情境的界定,这种研究方法的基础将不复存在。如果说对变异的研究是空间点格局分析方法的主线、灵魂的话,对空间情境的界定则是这种分析方法的基础。在上述介绍的方法中,也都暗含对时间情境的界定,所有事件都是在一定时间内产生的,除最后介绍的方法外,其余方法都假设事件在界定的时间内不存在受时间因素影响的变异。在使用这些方法时,需要从事社会科学研究的学者对时间段进行理性的划分,这将使分析渗透着研究者的职业判断,这也很正常,因为在社会科学研究中,没有纯粹的、脱离人为价值判断的机械性研究。

需要指出,本文介绍的几种空间点格局分析方法,只是为了起到丰富社会科学研究方法的目的,它们并不是万能的,它们很可能只是对于研究发生在较大空间尺度上的一些社会学命题有效,对于发生在微观组织内的社会学命题可能会显得苍白无力。然而由于这种方法考虑了时空情境和事件的空间性变异,在分析许多社会问题时无疑会起到重

要的作用。由于空间点格局在环境流行病学上已经成熟,因此毫无疑问,这种分析模式至少应该在医学社会学研究中得到重视。Prince 等人(2001: 1083 - 1088)和 Diggle 等人(2007)的研究都显示,原发性胆汁性肝硬化(primary biliary cirrhosis)在一定距离尺度上的聚集性反映了这种流行性的起因很可能来自一种未被觉察到的“社会-经济”风险因子。

另外在空间点格局分析中还有其他一些重要的方法,例如个体相搭配观测组-控制组(individually matched case-control)的空间聚集研究方法(Chetwynd *et al.*, 2001: 277 - 293)、使用相搭配观测组-控制组数据的点-源模型方法(Diggle *et al.*, 2000: 89 - 105),有兴趣的读者可以参见相应的文献,学习、使用这些方法,以期繁荣空间点格局分析在社会学研究中的普及。

此外,实现上述方法一般都有相应的软件包,绝大多数可以在网上免费获取。本文介绍的大部分方法可以使用软件包 *Splancs*¹或软件包 *Spatstat*²来实现,它们是在 R 软件³中使用的,但事件发生风险的空间变异之衡量方法目前需要下载专门的软件包 *epiS*⁴,但这个软件包是在 S-plus 中使用的。不过 S-plus 和 R 都是基于 S 语言开发的优秀统计软件,稍作调整,许多函数命令就可以互用。有关这些软件包的使用方法,可以在下载这些软件包的网络地址上同时获得,还可以参考 Rowlingson 和 Diggle(1993: 627 - 655)的论文以及 Jarner 和 Diggle(2002)的技术报告。

参考文献

- 黑斯蒂等. [2001] 2003. 统计学习基础—数据挖掘、推理与预测[M]. 范明、柴玉梅、咎红英, 等, 译. 北京: 电子工业出版社.
- 吉登斯, 安东尼. [2000]2003. 社会理论与现代社会学[M]. 文军、赵勇, 译. 北京: 社会科学文献出版社.
- 时培建、戈峰、杨清培, 等. 点格局分析中边缘校正的一种新算法及其应用[J]. 生

1 <http://cran.r-project.org/web/packages/splancs/index.html>.

2 <http://cran.r-project.org/web/packages/spatstat/index.html>.

3 可在 <http://www.r-project.org/> 免费下载 R 软件。

4 <http://www.maths.lancs.ac.uk/Software/epiS/>.

- 态学报 (待刊稿).
- 汤孟平、唐守正、雷相东,等. 2003. Ripley's $K(d)$ 函数分析种群空间分布格局的边缘校正[J]. 生态学报 (23).
- 谢宇. 2006. 社会学方法与定量研究[M]. 北京: 社会科学文献出版社.
- 薛毅、陈立萍. 2007. 统计建模与 R 软件[M]. 北京: 清华大学出版社.
- 张金屯. 1998. 植物种群空间分布的点格局分析[J]. 植物生态学报 (22).
- Baddeley, A., J. Moller and R. Waagepetersen, 2000. "Non- and Semi-parametric Estimation of Interaction in Inhomogeneous Point Patterns." *Statistica Neerlandica* (54).
- Bailey, T. C. and A. C. Gatrell. 1995. *Interactive Spatial Data Analysis*. Harlow: Longman.
- Barnard, G. A. 1963. "Contribution to the Discussion of Professor Bartlett's Paper. " *Journal of the Royal Statistical Society. Series B* (25).
- Berman, M. and P.J. Diggle. 1989. "Estimating Weighted Integrals of the Second-order Intensity of a Spatial Point Process." *Journal of the Royal Statistical Society. Series B* (51).
- Besag, J. and P.J. Diggle. 1977. "Simple Monte Carlo Tests for Spatial Patterns." *Applied Statistics* (26).
- Chetwynd, A. G., P. J. Diggle, A. Marshall, and R. Parslow. 2001. "Investigation of Spatial Clustering from Individually Matched Case-control Studies." *Biostatistics* (2).
- Diggle, P. J. 1983. *Statistical Analysis of Spatial Point Patterns*. London: Academic Press.
- Diggle, P. J. and A.G. Chetwynd. 1991. "Second-order Analysis of Spatial Clustering for Inhomogeneous Populations." *Biometrics* (47).
- Diggle, P. J., A. Chetwynd, R. Haggkvist, and S. Morris. 1995. "Second-order Analysis of Space-time Clustering." *Statistical Methods in Medical Research* (4).
- Diggle, P. J., S. E. Morris, and J. C. Wakefield. 2000. "Point Source Modeling Using Matched Case-control Data." *Biostatistics* (1).
- Diggle, P.J., V. Gomez-Rubio, P.E. Brown, A.G. Chetwynd and S. Gooding, 2007. "Second-order Analysis of Inhomogeneous Spatial Point Process Using Case-control Data." *Biometrics* (63).
- Jarner, M.F. and P.J. Diggle. 2002. "An S+ Library on Risk Estimation and Cluster Detection in Case-control Studies." [Technical Report of Lancaster University].
- Gatrell, A.C., T.C. Bailey, P.J. Diggle, and B.S. Rowlingson. 1996. "Spatial Point

- Pattern Analysis and Its Application in Geographical Epidemiology.” *Transactions of the Institute of British Geographers. New Series* (21).
- Hastie, T. and R. Tibshirani. 1986. “Generalized Additive Models.” *Statistical Science* (1).
- Kelsall, J.E. and P.J. Diggle. 1998. “Spatial Variation in Risk of Disease: a Nonparametric Binary Regression Approach.” *Applied Statistics* (47).
- Nelder, J.A. and R.W.M. Wedderburn. 1972. “Generalized linear Models.” *Journal of the Royal Statistical Society. Series A (General)* (135).
- Prince, M.I., A. Chetwynd, P.J. Diggle, M. Jarner, J. Metcalf, and O.F.W. James. 2001. “The Geographical Distribution of Primary Biliary Cirrhosis in a Well-defined Cohort.” *Hepatology* (34).
- Rowlingson, B.S. and P.J. Diggle. 1993. “SplanCS: Spatial Point Pattern Analysis Code in S-plus.” *Computers in Geosciences* (51).

责任编辑:路英浩