·基础理论与方法·

扫描统计量的理论及其在空间流行病学中的应用

李秀央　陈坤

【导读】 介绍扫描统计量的基本概念、计算方法及在空间流行病学中的用途。以近年来杭州市监测点的心血管疾病急性发作资料为实例，说明回顾性时空重排扫描统计量（采用 SaTScan 7.0.3统计分析软件）在疾病监测中评价其时空聚集性的应用。结果表明，监测点建德的乾潭镇监测人群心血管疾病急性发作在 1997 年 1 月 1 日至 2007 年 2 月 28 日具有聚集性（$P=0.001$），富阳的鹿山街道、灵桥镇监测人群心血管疾病的急性发作在 1997 年 1 月 1 日至 1999 年 2 月 28 日具有聚集性（$P=0.003$），富阳的大源镇、新义镇和受降镇监测人群心血管疾病的急性发作在 2001 年 3 月 1 日至 2004 年 2 月 29 日具有聚集性（$P=0.004$）和杭州市城站街道、紫阳街道、湖滨街道、清波街道、小营街道、望江街道、潮鸣街道、长庆街道、武林街道、天水街道、文晖街道、石桥街道监测人群心血管疾病的急性发作在 2004 年 3 月 1 日至 2006 年 2 月 28 日具有聚集性（$P=0.005$）。回顾性时空重排扫描统计量是一种能分析疾病的时空聚集性的工具，扫描统计量是空间流行病学中可以用来评价或预测疾病聚集性的一种行之有效的方法。

【关键词】 扫描统计量；空间流行病学；疾病；聚集性

Scan statistic theory and its application in spatial epidemiology　*LI Xiu-yang, CHEN Kun. Department of Epidemiology & Health Statistics, Zhejiang University, Hangzhou* 310058, China

Corresponding author: *CHEN Kun*, Email: ck@zju.edu.cn

【Introduction】 To introduce the basic concept of scan statistic, its computation method and application in the area of spatial epidemiology. Retrospective space-time permutation statistics for evaluating the clustering of disease monitoring program is illustrated, using data on recent acute onset of cardiovascular disease in Hangzhou, China. Calculations were performed with SaTScan Version 7.0.3. With 999 Monte Carlo replications, the program took 5 seconds to run on a 100-MHz Pentium PC. The geographical surveillance program on acute onset clusters of cardiovascular disease, data which showed statistical significance, would include: a) from January 1, 1997 to February 28, 2007 in Qiantan township, Jiande county ($P=0.001$); b) highly significant between January 1, 1997 and February 28, 1999 for Lushan street, Lingqiao township in Fuyang county ($P=0.003$); c) between March 1, 2001 and February 29, 2004 for Dayuan town, Xinyi town, Shouxiang town in Fuyang ($P=0.004$); d) between March 1, 2004 and Feb 28, 2006 for Chengzhan street, Ziyang street, Hubin street, Qinbo street, Xiaoying street, Wangjiang street, Chaoming street, Changqing street, Wulin street, Tianshui street, Wenhui street and Shiqiao street in Hangzhou ($P=0.005$), respectively. The retrospective space-time permutation statistics seems useful as a screening tool for identifying the cluster of disease. Scan statistics are practical and effective method for deciding which cluster alarms would merit further investigation and which clusters are probably chance occurrences in the study of spatial epidemiology.

【Key words】 Scan statistic; Spatial epidemiology; Disease; Cluster

扫描统计量在流行病学中通常可以用来监测和评价疾病在时间、空间或时空方面的聚集性。疾病时间聚集性的常规统计分析方法是图示法、两率比较的 u 检验或 χ^2 检验，多个率比较的 χ^2 检验及均匀分布的拟合优度检验等都有很大的局限性，对于小样本资料和未分组资料，以上方法的检验效能较低且难以避免时间分组上的主观性。Naus 提出了扫描统计量的概念，即用事先选定的时间区间扫描整个观察期所得到的病例数的最大值。由于该方法消除了人为地按年、月分组造成的主观性而且检验

基金项目：杭州市科技局重点创新项目（20051323B44）

作者单位：310058 杭州，浙江大学公共卫生学院流行病与卫生统计学系

通讯作者：陈坤，Email：ck@zju.edu.cn

效能较高，目前已成为疾病时间聚集性或区域聚集性分析研究中的热点，根据资料性质不同，目前常用的主要方法有 Bernoulli 模型的扫描统计量、Poisson 模型的扫描统计量、时空重排模型的扫描统计量、Ordinal 模型的扫描统计量和指数模型的扫描统计量等[1]。

基本原理

历史上一次经典试验把扫描统计量与流行病学中疾病的聚集性联系在一起。Glaz 等[2] 记载了 Revsez 设计的著名试验。Revsez 在课堂上让部分学生各自抛一枚均匀的硬币 200 次（多重 Bernoulli 试验），并记录观测值，出现正面的记为 H，出现反面的记为 T，这样就分别形成多个长度为 200 的 HT 序列。同时，Revsez 让另外部分学生在草稿纸上臆造一份他们自认为随机的序列。以下两个序列，一个是真实的，一个是臆造的。为了简洁起见，只给出前 50 个记录值：

序列一：THTHHTTTTHHTHTHHHHHHHHTTTH
HTHTTTHTTTHTTHHTHHTHTHT

序列二：HHTHTTTHTTTTHHTHTTTHTTHHHH
HTHTHTHTTTTHHTTHHTTHHHT

把连续出现正面（H）的最长长度称为最长链，由此可知序列一的最长链为 8，序列二的最长链为 5。连续出现 8 次 H 是否正常，连续出现 5 次 H 的概率有多大，哪个序列是真实的？利用扫描统计量的知识，可以方便地解决类似问题。从后文可看到，根据扫描统计量，对于 200 次试验，最长链≤5的概率不到 3%，也就是说，序列二是臆造的。

扫描统计量的概念：设观察期（0，T）的长度为 T，扫描统计量（Scan Statistics，S_w）是以长度为 w 的扫描窗口从时点 $t(1\leqslant t\leqslant T-w)$ 开始，扫描整个观察期 T 所得到的各窗口病例数的最大值，若记为 Y_t 区间 $(t, t+w)$ 内的病例数，则 $S_w=\max(Y_t|0\leqslant t\leqslant T-w)$。

若观察期内的总病例数 N 已知，可假设 N 个病例的发生属均匀分布，借助扫描统计量可回顾性地分析病例发生有无时间聚集性。若观察期内的总病例数 N 未知，可假设 N 为随机变量，借助 Poisson 分布下的扫描统计量模型前瞻性地分析病例发生有无时间聚集性，即监测疾病是否“流行”。

［情形 1］ Poisson 模型（此时为 Poisson 分布扫描统计量[3]）

当观察期（0，T）内的总病例数 N 未知，若单位时间的长度为 w，H_0：单位时间内的病例数符合均数为 λ 的 Poisson 分布，记为 $Y_t\sim Poisson(\lambda)$，即 $N\sim Poisson(\mu)(\mu=\lambda T/w)$，$H_1$：在观察期（0，T）内，某些长度为 w 的区间内的发病数高于其他区间，即在某一区间 $(\tau,\tau+w)$ 内，$Y_t\sim Poisson(\theta\lambda)$，$\theta>1$，而在其余所有区间内都有 $Y_t\sim Poisson(\lambda)$。

$P^*_{(n,\mu,w)}$ 表示 N 符合均数为 μ 的 Poisson 分布时扫描统计量 S_w 等于或大于某一特定值 n 的概率，即 $P^*_{(n,\mu,w)}=P(S_w\geqslant n)$，$N\sim Poisson(\mu)$（$N$ 较大，w 相对于 T 较小时，确切概率的计算非常复杂）。Wallenstein & Neif 提出了 $P^*_{(n,\mu,w)}$ 近似计算公式：

$$P^*_{(n,\mu,w)}=(n-\lambda)\left(1-\frac{w}{T}\right)\left(\frac{w}{T}+1\right)\frac{e^{-\lambda}\lambda^n}{n!}+2\left(1-\sum_{x=1}^{n}\frac{e^{-\lambda}\lambda^x}{x!}\right),\lambda=\frac{\mu w}{T}$$

陈滔等[4] 编制了观察期（0，T）内预期总病例 $N<50$，扫描间隔 w 为观察期的 $\frac{1}{4},\frac{1}{6},\frac{1}{8},\frac{1}{12},\frac{1}{24}$ 时 S_w 的临界值。

Poisson 分布下的扫描统计量模型可用于符合 Poisson 分布的稀有疾病的动态监测，监测方法：事先选定一个观察期（0，T）和以长度为 w 的扫描窗口，用前瞻性的方式考察扫描统计量 S_w 的值是否超过 H_0 下的临界值 $n_{(N,w,\alpha)}$［N 为观察期（0，T）预期总病例，w 为扫描间隔，α 为检验水准］，当 $S_w\geqslant n_{(N,w,\alpha)}$，$P\leqslant\alpha$，则可以认为该窗口内的病例有聚集现象，即疾病发生“流行”。

模型的适用条件是基线期的数据没有增加、减少的长期趋势，即有一个比较稳定的基线；理论上 λ 是该病基线期长度为 w 区间内的实际发病数。其缺点是在计算过程中需要提供各区域的人口数据。

［情形 2］ 时空重排模型（此时为时空重排扫描统计量[5]）

时空重排扫描统计量基本思想是扫描窗口为圆柱形窗口，圆柱形的底对应一定的地理区域，圆柱形的高对应一定的时间长度。由于疾病的发生会在何时何地以何种规模发生都是未知的，因此在运用时空重排扫描统计量进行分析时，圆柱形扫描窗口的大小和位置是处于动态变化之中。

计算的具体过程：假定在区域 z 在 d 天中发病数 C_{zd}，则所有区域在所有时间的总发病数 C 为：

$$C=\sum_{z}\sum_{d}C_{zd}$$

对于每区域和每天，预期发病数为：

$$\mu_{zd}=\frac{\sum_{z}C_{zd}\sum_{d}C_{zd}}{C}$$

因此每个圆柱 A 的预期发病数 μ_A 为：$\mu_A=\sum_{(z,d)\in A}\mu_{zd}$

设 C_A 为每个圆柱 A 中的实际发病数，C_A 服从均数为 μ_A 的超几何分布，其概率函数为：

$$P(C_A)=\frac{\binom{\sum_{z\in A}C_{zd}}{C_A}\binom{C-\sum_{z\in A}C_{zd}}{\sum_{d\in A}C_{zd}-C_A}}{\binom{C}{\sum_{d\in A}C_{zd}}}$$

当 $\sum_{z\in A}C_{zd}$ 和 $\sum_{d\in A}C_{zd}$ 相对于 C 非常小时，C_A 近似服从均数为 μ_A 的 Poisson 分布。基于这一近似，采用广义似然函数（generalized likelihood ratio, GLR）来衡量圆柱 A 中发病数是否异常：$\left(\frac{C_A}{\mu_A}\right)^{C_A}\left(\frac{C-C_A}{C-\mu_A}\right)^{(C-C_A)}$，然后利用 Monte Carlo method 产生模拟数据集，计算 P 值。

相对于普通时空扫描统计量（基于 Poisson 分布的时空扫描统计量），时空重排扫描统计量概率模型的优点是建模过程中不需人口数据。

［情形 3］ 其他模型扫描统计量

Ordinal 模型扫描统计量适用于每个观察结果是一个等级资料，即有序多分类的情况，若观察结果是二分类的，则等同于 Bernoulli 模型扫描统计量[3,6]。Ordinal 模型可以分析单纯时间、单纯空间或时空扫描统计量[7]。指数模型扫描统计量适用于生存时间数据[8]，而正态模型扫描统计量适用于连续型分布的资料。

实例分析

以 2002 年 3 月至 2007 年 2 月杭州市上城区、下城区、富阳和建德 4 个监测点心血管疾病的急性发作资料为实例，采用回顾性时空重排扫描统计量来分析杭州市监测区心血管疾病急性发作情况的时空聚集性。

具体分析方法如基本原理中时空重排扫描统计量部分，有关疾病是否具有聚集性的推断使用 Monte Carlo 假设检验。如果是从 999 个模拟数据集中计算而得的最大 GLR 是真实数据中前 50 个最大 GLR 值，那么，在检验水准 $\alpha=0.05$ 疾病聚集有统计学意义。一般 $P=R/(S+1)$，其中 R 是真实数据中最大 GLR 的秩次，S 是模拟实验的次数。

回顾性时空重排扫描统计量分析结果表明：建德的乾潭镇监测人群心血管疾病急性发作在 1997 年 1 月 1 日至 2007 年 2 月 28 日期间具有聚集性（$P=0.001$），富阳的鹿山街道、灵桥镇监测人群心血管疾病急性发作在 1997 年 1 月 1 日至 1999 年 2 月 28 日具有聚集性（$P=0.003$），富阳的大源镇、新义镇和受降镇监测人群心血管疾病急性发作在 2001 年 3 月 1 日至 2004 年 2 月 29 日具有聚集性（$P=0.004$）和杭州的城站街道、紫阳街道、湖滨街道、清波街道、小营街道、望江街道、潮鸣街道、长庆街道、武林街道、天水街道、文晖街道、石桥街道监测人群心血管疾病急性发作在 2004 年 3 月 1 日至 2006 年 2 月 28 日具有聚集性（$P=0.005$）（表 1）。

讨 论

扫描统计量消除了人为地按年、月分组造成的主观性，适用于已经确切发病（疾病急性发作或死亡等事件）时间的小样本资料。该方法可以分析区域数据，也可以直接分析个体数据；但是要求没有缺失值，故在建立数据库时，为了充分利用原有信息，零数据必须填上，否则会作为系统缺失值处理。

选用扫描统计量时需要注意适用条件，由于模型比较多，在分析时应该正确判别其具体资料的特点来判别该选哪种模型；至于时空重排扫描统计量

表1 1997 年 1 月 1 日至 2007 年 2 月 28 日杭州市心血管疾病的回顾性时空分析结果

监测点	定位/半径(km)	时间（年/月/日）	病例数	期望例数	观察例数/期望例数	统计量	Monte Carlo rank	P 值
乾潭	(29.6191N,119.5762E)/0.00	1997/1/1－2007/2/28	17	2.12	8.05	20.69	1/1000	0.001
鹿山、灵桥	(29.9681N,119.8146E)/3.52	1997/1/1－1999/2/28	3	0.07	42.65	8.33	3/1000	0.003
大源、新义、受降	(30.2202N,119.8608E)/15.55	2001/3/1－2004/2/29	17	5.28	3.22	8.29	4/1000	0.004
城站、紫阳、湖滨、清波、小营、望江、潮鸣、长庆、武林、天水、文晖、石桥	(30.2436N,120.1744E)/0.00	2004/3/1－2006/2/28	252	200.76	1.26	8.11	5/1000	0.005

选用回顾性的还是前瞻性的模型，应根据研究的侧重点，若对历史资料某现象的出现是否具有聚集性或随机性进行评价，则选用回顾性时空重排扫描统计量，若利用历史资料对某现象的出现是否具有聚集性或随机性进行预测，则该选用前瞻性时空重排扫描统计量。另外，在分析前应充分熟悉相关软件功能和各种模型所需的数据库格式和要求。

目前，SaTScan 7.0.3软件只能计算各种模型扫描统计量的各种参数，不能实现将其结果可视化，否则，需要将分析结果导入地理信息系统，在 ArcGIS 软件或 Google Earth 软件中才能三维可视化。所以，进一步开发兼具分析功能和可视化功能的相关应用软件也是一大重要课题。

扫描统计量在质量控制、计算机病毒、通讯工程、图像识别、可靠性、天文学和应力工程等科学或行业中发挥了重要的作用。目前，此方法也逐步应用于金融领域和流行病学研究[4,9,10]。但是国内的研究和实际应用还不充分，其中文献[4,9]介绍方法仅局限于疾病的时间聚集性的应用研究，文献[10]仅局限于利用历史资料对传染病的出现是否具有聚集性或随机性进行预测，实际上扫描条件量在空间流行病学中的时空聚集性评价和预测有广泛的应用前景。

参 考 文 献

[1] Kulldorff M. SatScan™ User Guide for version 7.0. http://www.satscan.org/.
[2] Glaz J, Joseph Naus L, Sylvan W. Scan statistics. New York: Springer-Verlag, 2001.
[3] Kulldorff M. A spatial scan statistic. Communication in Statistics: Theory and Metods, 1997, 26: 1481-1496.
[4] 陈滔，杨树勤，吴艳乔，等. 扫描统计量在稀有疾病监测中的应用. 中国公共卫生，1997，13(5)：301-304.
[5] Kulldorff M, Hefferman R, Hartman J, et al. A space-time permutation scan statistic for disease outbreak detection. Plos Medicine, 2005, 2: 216-224.
[6] Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Stat Med, 1995, 14(8): 799-810.
[7] Jung I, Kulldorff M, Klassen AC. A spatial scan statistic for ordinal data. Stat Med, 2006, 26(7): 1594-1607.
[8] Huang L, Kulldorff M, Gregorio D. A spatial scan statistic for survival data. Biometrics, 2007, 63(1): 109-118.
[9] 陈滔，杨树勤，吴艳乔. 扫描统计量在疾病时间聚集性分析中的应用. 中国卫生统计，1996，13(2)：7-9.
[10] 殷菲，冯子健，李晓松，等. 基于前瞻性时空扫描统计量的传染病早期预警系统. 卫生研究，2007，36(4)：455-458.

（收稿日期：2008-02-21）
（本文编辑：张林东）

·疾病控制·

石家庄市 2006－2007 年流感疫情暴发分析

张国平 白萍 史艳 徐保红 侯书恒

对石家庄市 2006－2007 年 19 起流感疫情暴发的监测资料进行分析，探讨全市流感活动状况。所有咽拭子标本均来自 19 个疫点的流感样患者，流感样病例均根据“全国流感/人禽流感监测实施方案”确定，并在 24 h 内进行流感病毒分离。检测方法参照“流感及其实验技术”和“全国流感监测实施方案”进行。标本采用狗肾传代细胞（MDCK）分离培养，每份标本接种 2 瓶细胞，传两代，特殊情况传三至四代。连传两代都为血凝阴性者，视为病毒分离阴性弃掉。血凝滴度≥1:8时，进行分型鉴定。病毒分型鉴定采用国家统一的微量血凝抑制法，分型鉴定后的毒株送往国家流感中心进行复核确认。疫情监测结果显示，2006 年 4 月至 2007 年 6 月，石家庄市共报告流感样疾病暴发疫情 19 起，经实验室核实诊断全部为流感暴发疫情（表 1）。

2006－2007 年石家庄市的流感样疾病暴发疫情均发生在学校和托幼机构。从监测结果看，流感疫情暴发疑似有两个高峰期，第一个高峰为 12 月至翌年 1 月，这一时期的流行优势株为 A 型流感病毒；第二个高峰期为春末夏初季节，即 4－6 月份，这一时期的流行优势株为 B 型流感病毒。一般认为，北方的流感流行在季节上表现为冬季形成流行高峰，其他季节散发，但从石家庄市这 2 年的监测结果看，在春末夏初季节存在着一个小的流行高峰。由于监测年份有限，对于北方是否存在着两个流行高峰还有待以后的监测研究。共检测 B 型流感 8 起：B 型 Victoria 系 6 起、B 型 Yamagata 系 1 起、B 型 Victoria 系＋Yamagata 系 1 起，均发生在春末夏初，可见石家庄市近 2 年的春季流感流行主要是以 B 型 Victoria 系为主。但是，在 2007 年 5 月份，暴发了一起由 B 型 Yamagata 系引起的疫情，且同年 6 月的一起流感疫情中同时分离到 Victoria 系＋Yamagata 系，这是否提示本市 B 型流感亚型有转化为 Yamagata 系的趋势，有待进一步监测。

表1 2006－2007 年石家庄市流感暴发监测结果

时间（年.月）	暴发起数	流感病毒分型					
		H1N1	H3N2	H1＋H3	BY	BV	BY＋BV
2006.05	3	0	0	0	0	3	0
2006.06	1	0	0	0	0	1	0
2006.12	6	0	5	1	0	0	0
2007.01	5	0	5	0	0	0	0
2007.05	3	0	0	0	1	2	0
2007.06	1	0	0	0	0	0	1

作者单位：050011 石家庄市疾病预防控制中心（张国平、白萍、史艳、徐保红）；石家庄市中医院（侯书恒）

（收稿日期：2007-12-21）
（本文编辑：尹廉）