1482

库尔德多夫

窗口中的点数被记录并与其在纯随机泊松过程零假设下的分布进行比较。

一维问题已经向各个方向扩展。当点被分组为几个子区间之一时，我们有聚合数据。这已经被Wallenstein等人(1989年)等研究过，当我们有某些事件的每月计数时，这很感兴趣。Weinstock(1981年)研究了在零假设下，基本泊松过程的强度具有已知非均匀性的问题。许多作者，如Saperstein(1972年)和Naus(1974年)，研究了一个相关的伯努利模型，其中包含一系列二进制结果。Loader(1991年)允许非固定窗口大小。Glaz和Naus(1983年)研究了寻找多个簇的扫描统计量。对于所有这些扩展，并且取决于应用，扫描统计量可能取决于观察到的总点数，也可能不取决于。

本文提出了一种空间扫描统计方法。尝试在尽可能广泛的背景下处理这个问题，除了分析总是基于观察点的总数之外。窗口可以采用任何预定义的形状，并且允许在扫描研究区域时窗口大小发生变化。当我们对聚类覆盖区域的面积缺乏先验知识时，后者非常有用。该方法还允许-个任意但已知的基线强度，该基线强度控制着零假设下点的分布。这可以采取许多不同的形式取决于应用程序。在地理空间G中，t是一个度量u。当G是一条直线，p是[a，6]上的均匀度量时，我们得到传统的-维问题作为一个特殊情况。利用平面上的勒贝格度量，我们有一个均匀的空间泊松过程。其他可能的措施包括:

1在林业中，树木的空间聚集现象被研究。一个潜在感兴趣的问题是在补偿所有树木的不均匀空间分布后，是否存在特定种类或具有某种特征的树木聚集。也就是说，我们想了解某种树木的比例在某些地方是否特别高。在这种情况下，一个区域的衡量标准是那里生长的树木总数

2 在天文学中，如果我们想要在补偿所有恒星的非规则空间分布后探测特定类型恒星的星团，存在一个等效的三维问题。

3 流行病学家对疾病的地理集群感兴趣。这里需要补偿整个人口密度不均的情况。当数据汇总到人口普查区时，该测量值将集中在这些区的中心坐标上。

为了寻找铀矿，飞机在大面积平行线飞行时测量盖革计数。某一特定区域的计数数量过高

表示可能存在沉积物。该措施将在飞行线沿线保持一致，其他地方为零。

5动物学家可能会研究海鸥巢的空间分布。在一个群岛中，巢将位于岛屿上。适当的措施是陆地上的均匀分布，其他地方为零。

6 如果我们对一种疾病的时空集群感兴趣，那么这个度量仍然会集中在地理维度上，就像例子3一样，但它也会扩展到第三个维度，反映人口普查区中随时间变化的人口规模。

7调整人口分布不均的情况并不总是足够的，无论是人类、树木、恒星还是其他事物。我们可能还需要考虑各种混杂因素。例如，在流行病学中，我们可以让测量值反映年龄标准化的预期发病率。

在一维情况下，测试统计量的确切分布仅在特殊情况下已知。许多文献关注于找到好的近似值。在高维情况下，统计理论变得更加复杂Naus(1965b)为均匀背景测量和固定但任意大小的矩形窗口下的二维扫描统计量获得了分布限制。Loader(1991)处理了相同的问题但允许可变窗口大小。Turnbull等人。(1990年)，使用上述示例(3)中描述的基础测量值，使用具有恒定测量值的圆形窗口。由于无法确定测试统计量的确切分布，因此使用蒙特卡罗模拟进行假设检验，上述所有模型都是本文概述的特例。另外两个特例可以在Kulldorff和Nagarwalla(1995年)和Hialmars等人(1995年)中找到，他们分别将空间扫描统计应用于纽约上州的白血病数据集和瑞典的数据集

当扫描窗口的大小固定时，在任何给定时间，测试统计量总是取窗口中的最大点数。使用可变窗口大小不再可能，取而代之的是使用似然比检验统计量(Loader，1991)。

在第2节中描述了泊松和伯努利模型，然后在第3节中介绍了似然比检验统计量，第4节描述了它的某些理论性质，分别在第5和第6节中讨论了计算问题和给出了一个实际例子。

2.泊松模型和伯努利模型

设N表示一个空间点过程，其中N(A)是集合A、C、G中点的随机数。当窗口在学习区域上移动时，它定义了一个

4.测试统计量的性质

4.1检测与推理

大多数用于空间点过程的聚类分析的统计方法要么是描述性的，因为它们可以检测集群的位置，但不涉及任何推理，要么是它们做推理，但不能检测集群的位置。空间扫描测试的一个重要特征是，它两者都有，因此，当零假设被拒绝时，我们可以定位导致拒绝的地图的特定区域。精确地说，让x={qi，i= 1，.，ng}表示数据集中ne点的坐标集，其中Z是最有可能的簇，并让x‘={c}，i=1，...，ng}是一个具有完全相同的点数的替代配置。下面的定理适用于伯努利模型和泊松模型。

定理1如果零假设在x下被拒绝，那么它也在x‘下被拒绝，如果z’= xi为所有；。

换句话说，该定理指出，只要构成最有可能集群的区域内的点位于它们所在的位置，无论其他的点如何被打乱，我们仍然会拒绝零假设。例如，如果零假设由于西雅图的疾病聚集而被拒绝，那么无论美国东海岸的病例如何变化，零假设仍然会被拒绝。这听起来可能像是一种不言而喻的特性，但它并不适用于大多数其他的空间聚类测试，如诺克斯（1964）、惠特莫尔等人（1987）、库齐克和爱德华兹（1990）。或Diggle和Chetwynd（1991）。因此，如果我们想知道集群的位置，这些测试就不合适了。相反，他们是愿意回答这个问题的

1488

聚类现象是否发生在整个研究区域，例如一种疾病是否具有传染性，这是一个空间扫描统计不适合的问题。

证明：设A (x)和A（x‘）表示两个不同数据集的检验统计量的值。由于这两个数据集具有相同数量的点数，因此A在原假设下的分布将是相同的，因此这足以证明X (x)> A (x)。在伯努利的情况下

L(2)×)

L (2)x)VI

入（x）=

Lo Lo

第一个不等式成立，因为x在区域Z内的点至少和x一样多。对于泊松模型，如果A(x)=为1，则是平凡的正确的。当A (x)>1时，我们从方程4中得到了那个

nZ

nZ y

nG-T2Z )-μ()μG)nc--n2 p (G)-v(△)nG-n/z μ(G)-μ(2))-μ()

A(x)=支持

X1

μ（2）20

邢宝1

T22至AM μg (2)n/a

心心-VI 心

μg（2）

n'z n'2 一≤SUPK

nc-nz nd-mz μG)

H(Z)

库尔多夫

supzL (Z)x)

=A(x).

Lo 一

nG-TZ

nc-Bz

na一nz

=A(x).

其中K=（nc/μ(G)）nc。第一个不等式对于任何常数a，B都成立

当一个> β（N-n）时，N-n是n的递增函数。

和N，（an）“（B（N-n）4.2功率

Wallen-stein等人（1993,1994）和Sahu等人的a1研究了一维扫描统计量的威力。（1993年）等。对于空间扫描统计量，我们不能期望找到统一的最强大的检验，除非在特殊情况下，只有一个区域。相反，我们证明它满足了标准，使它成为我们所说的个体最强大的测试。

为了定义一个单独的最强大的测试，我们将复合的改变本机假设划分为不同的子集。参数空间被划分为一个可数的子集{A，}，这样a，=，a，=0，并且UA构成了整个备选假设的全部。同样地，使用相同的索引，重新原假设的临界区域C被划分为不相交子集{C，其中UC；=C。设C‘= UC’；表示一个具有相应不相交子集的备选临界区域。

1489

空间扫描统计量

定义1对于一个特定的显著性级别a，对于参数空间的一个分区{a；}和一个分区，一个测试分别是最强大的

103

y这样

[C]}的临界区域，如果对每个Ap没有序列C‘和那个

1.C;=C;forally=ks. 2.P (w∈C"Ho)=a

3.P（w∈C\EC\（Z，p，Q））>P（w∈CGe|（Z，p，9））对于任何（Z，P.9）∈Ae这意味着如果我们修复关键区域除了其子集Cre由声明l，那么测试是一致最强大的相比所有剩余的选择的关键区域和对所有参数（Z，p，q）∈Ak。这个属性在任何多重测试类型中都非常重要，其中有一个复合替代假设，我们希望知道它的哪一部分导致了拒绝。如前所述，扫描统计量有能力识别负责拒绝零假设的区域，如果我们不能检测到一个真正的簇，如果基于研究区域的另一个不真实的簇而拒绝零假设，那就没有什么安慰了。事实上，这通常比不仅仅拒绝零假设更不可取。这个问题类似于其他多重比较的情况，我们可能不是同时测试多个集群位置，而是测试几个新的农业作物品种，看看其中是否比目前使用的更好，或者我们可能同时测试几个潜在的癌症风险因素。

如果我们只关心拒绝和不关心拒绝，而对集群的位置不感兴趣，那么作为一个单独的最强大的测试的属性就没有什么价值了。对于这种问题，基于似然比的空间扫描统计量将是一个次优选择。

现在，让Az=[（Z，p，9）：p>q}和Ao={（Z，P，9）：p=q}。设Ca表示临界区域C和Z是最有可能的簇的样本空间的子集的交集。

定理2基于A的测试形成了一个关于分区的单独的最强大的测试。这也适用于伯努利模型和波克松模型。

证明：我们证明，如果定义中的语句(1)和(2)是true，那么(3)就不能成立。对于任意的Z，让D=：w∈Cz，和D+=：w∈C‘z，w∈Cz}。让

L(Z.P.gleo)M = sup L (Hol)weD+

5.计算和蒙特卡罗抽样

为了找到检验统计量的值，我们需要一种方法来计算似然比，因为它在备择假设的区间集合上被最大化。这可能看起来像是一项艰巨的任务，因为区间的数量很容易是无限的。两个属性使我们能够将其简化为有限的问题。观察点的数量总是有限的，对于固定数量的点，似然随着移动窗口的度量增加而减小。考虑第2节中示例(2)的扫描窗口。如果我们让圆的大小增加，并保持焦点不变，我们只需要在任何一个新点进入圆时重新计算概率。由于点的数量是有限的，我们需要为每个焦点计算概率的次数也是有限的，由于焦点的数量也是有限的，因此总计算次数也是有限的。假设使用伯努利模型或均匀泊松模型，对于第2节中给出的其他四个示例，类似的论点也成立。

一旦计算出了检验统计量的值，就很容易进行推断。我们无法期望找到检验统计量的分布的封闭解析形式。相反，我们依赖于蒙特卡罗模拟。最初由Dwass(1957年)提出，该技术最初由Turnbull等人(1990年)在扫描统计学的背景下使用。由于我们知道基础测量p，因此当我们对总点数nG进行条件时，我们可以获得在零假设下生成的数据集的复制。如果有9999次这样的复制，如果实际数据集中的检验统计量的值在来自复制的检验统计量的前500个最高值中，则该检验在5%的显著性水平上是显著的。

除了最有可能的聚类，我们可能还想查看具有高可能性值的次要聚类。其中一些与最有可能的聚类相关，因为它们包含相同的一组点，各自的区域相互重叠。这样的次要聚类通常没有多大意义，尽管它们提醒我们，所获得的检测到的聚类的位置和大小只是估计值。

更有趣的次级集群类型是那些位于研究区域另一部分的集群我们定义这些集群为与更可能的集群不重叠的集群。通常，报告这些集群与最可能的集群一起是有意义的。

对于推断，我们可以选择一个次要集群，并将其似然值与蒙特卡罗重复的最大似然比进行比较和排名。任何排名低于显著水平的次要集群本身都会导致拒绝原假设，即使数据集中没有其他更可能的集群。这为我们提供了对次要集群的推断程序，但由于我们比较的是数据集中的次要集群与重复中的最可能集群，因此这种测试有些保守。

6.例子

我们使用北卡罗来纳州婴儿猝死综合症(SIDS)的数据来说明这些模型。这些数据是由M.Symmons, D.Atkinson汇编的。

以及北卡罗莱纳州人力资源部门的州卫生统计中心。此前，克雷西和陈（1989）等人曾对它们进行过分析。

在北卡罗来纳州的100个县中，每一个县的数据都包括1974-1984年的活产婴儿总数和婴儿猝死人数（SIDs）数。该县的活产人数在567人至52345人之间。县城的所在地被用作地理坐标。在753354例活产婴儿中，艾滋病患者总数为1503例。这使得全州的发病率为2.0/1000。“总计”每个县的出生人数，以及全州范围内的sid的人数，也被分为白人和非白人。完整的数据由Cressie和Chan（1989）提供。

每个县坐标点的测量值为该县的活产数。该措施在其他地方为零。这就如第1节的例子(3)所示。作为窗口的区域，我们使用所有以一个县坐标点为中心的圆，其中最多包括总人口的一半。下面是第2节的例子(2)。

请注意，仅相对于聚合的数据，这些区域是圆形的。当我们在一个县周围画圈时，其他县要么完全是一个区域的一部分，要么不全部，这取决于它的县是否在这个圈内。因此，我们得到了一个紧凑但不规则形状的区域。这可以在图l和图2中看到。6.1伯努利模型

伯努利模型是对这个数据集最自然的模型。我们有出生计数，每一次出生最多只能对应一个婴儿猝死。表1总结了分析的结果。

最有可能的集群是A，包括位于该州南部的布拉登、哥伦布、霍克、罗布森和苏格兰等县。排名为1/10000，i.e.a p值为0.0001。

还有另一个重要的集群，B，由东北部的哈利法克斯、赫特福德和北安普顿县组成。当排名为5/10000时，它的p值为0.0005。后一种测试是保守的，因为我们正在将数据集中的次要集群与来自副本的最有可能的集群进行比较。6.2泊松模式！

由于我们处理的是一种罕见的疾病，泊松模型应该给出一个接近于伯努利模型。这个数据集的结果确实相似，请见表1。

当我们有我们希望在分析中包含的协变量时，泊松近似特别有用。对于小岛屿发展中国家，一个可能的协变量是种族（Cressie和Chan，1989），这可能通过住房质量或获得卫生保健的机会等变量与单一服务的小岛屿发展中国家有关。北卡罗来纳州的各县的种族分布差异很大，这可能解释之前检测到的集群。我们可能想看看，在根据种族因素进行调整后，是否仍然有地理集群。这可能会导致我们的其他空间相关的风险因素，否则就会被隐藏起来。

白人儿童的SIDS总发病率为1.512，非白人儿童的总发病率为2.970（Cressie和Chan，1989）。每个县坐标x处的基本测量现在可以定义为

μ(x)=white births×1.512+nonwhite births×2.970

这与零假设下的sid的预期数量成正比。请注意，我们不需要知道在按种族细分的每个县的sid的数量。该似然比检验结果见表2。

将这个分析与没有纳入种族的分析进行比较，我们观察到三件事。

1南部的a组的排名为36/10000（p=0.0036），仍然很显著，不能仅仅用该地区非白人出生的高比例来解释。

2.东北方向的集群本已不再显著，排名为3336/10000（p=0.3336）。