城市警务的共享成分点过程模型

Claire Kelling卡尔顿学院，美国明尼苏达州诺斯菲尔德。电子邮件：ckelling@carleton.edu 数学与计算中心 225，美国明尼苏达州诺斯菲尔德 55057 Murali Haran宾夕法尼亚州立大学，美国宾夕法尼亚州斯泰特学院。

arXiv:2303.00206v1 [stat.ME] 1 Mar 2023

摘要。
新近可得的点级数据集使我们能够将警察使用武力与描述警察行为的其他事件联系起来。当前用于关联两个点过程的方法通常依赖于对其中一个点过程进行空间聚合。我们研究了结合共享成分模型与病例对照方法的新方法，在保留两个点过程点级特征的同时刻画它们之间的关系。我们发现，共享成分方法在灵活关联两个点过程方面尤其有用，并通过模拟示例以及对芝加哥警务数据的应用展示了这种灵活性。

关键词：病例对照、犯罪学、高斯过程、标记点过程、警务、共享成分模型

1. 引言

如今，许多城市地区越来越容易获取带有精确位置信息的事件级数据，涵盖警察使用武力事件、警察拦截和暴力犯罪等现象。随着警察过度使用武力的情况持续存在，确定警察使用武力的原因及其后果，正给犯罪学家和公共政策制定者带来越来越大的挑战。我们考察了芝加哥丰富的警务数据集。现有研究常常采用较粗略的空间信息来研究警察行为，例如事件发生辖区的指示变量，而非事件的确切位置（参见 Antonovics and Knight，2009）。点过程方法能够让我们纳入事件精确位置的详细信息。此外，将相关现象结合起来，也有助于研究警察使用武力。例如，暴力犯罪和警方拦截的频率及空间分布，或许有助于描述警察使用武力事件的空间分布及其影响因素。警方拦截能够提供一些关于警察巡逻地点的初步信息，而暴力犯罪的发生率则是描述警察巡逻社区状况的一部分。

分析不同点过程之间关系的方法有很多，从描述性统计到参数和非参数方法皆有。例

如，病例对照模型常用于分析两类事件之间的关系，而不对任一类点进行聚合。具体而言，一个点过程（“对照”过程）的强度用于缩放第二个点过程（“病例”过程）的强度。本文提出一种新的点过程共享成分模型，为分析两种点模式之间的空间关系提供了丰富的框架，并避免了对病例与对照过程进行指定。我们将共享成分方法与现有的点过程关系模型进行比较，特别是病例对照模型。

我们在下文概述本文的主要贡献。

- 我们提出了一个用于两个点过程的共享成分模型，该模型既允许每个点过程具有独特的空间模式，也允许两个点过程之间存在共享的模式。该模型建立在为区域数据开发的共享成分模型（Knorr-Held 和 Best，2001年）之上。通过模拟示例及其在芝加哥警务数据中的应用，我们发现该模型具有灵活性且易于解释。

- 我们通过模拟研究和对芝加哥警务数据的应用，研究了点过程病例对照模型的使用。我们考虑了在此背景下估计强度函数和回归系数的两种方法：逻辑回归（参见 Diggle 等人，2007）和空间强度函数的贝叶斯估计。我们发现，对于这类模型，必须谨慎选择估计程序及相应的解释。

- 我们将共享成分模型与病例对照模型进行了比较，发现共享成分模型在计算上更为复杂，但在建模两个点过程之间的关系时，能够提供额外的灵活性和空间结构。

- 尽管我们的结果仍是初步的，但我们发现，共享分量方法能够有效研究芝加哥警察使用武力与警方拦截之间的关系。我们展示了一个为两个点过程所共有的空间模式，该模式位于芝加哥市中心以南；同时也展示了分别影响这两个过程的独特因素。在考虑共享空间模式后，芝加哥北部的使用武力事件具有更高的空间强度，而该市南部的警方拦截则具有更高的强度。

本文结构安排如下。我们将在第 2 节介绍当前将警察使用武力与其他事件数据集（如警察拦截和暴力犯罪）联系起来的活跃研究领域，以及所采用的空间分析方法。在第 3 节，我们将介绍用于关联两个点过程的统计方法，概述点过程的病例对照方法，并介绍我们提出的新型点过程共享成分模型。第 4 节将详细介绍本文使用的芝加哥警察拦截和警察使用武力数据，并介绍模拟示例。第 5 节将病例对照模型和共享成分模型应用于芝加哥警务数据及模拟示例，最后在第 6 节展开讨论并总结全文。

2. 背景

我们首先概述本研究所依据的警务数据类型，以及犯罪学中相关的研究问题，并介绍为分析这些数据集而开发的现有方法。研究问题例如：哪些因素可能影响警察使用武力？如何借助更多丰富的点级数据集，更好地理解警察使用武力的空间分布以及更广泛意义上的警察行为？第一小节介绍将警察使用武力与警察拦截数据联系起来的研究；第二小节则介绍将警察使用武力与另一类点级数据集——暴力犯罪数据集——联系起来的研究。最后，我们展示当前针对这些数据集开展空间分析的方法。这些研究大多依赖空间聚合，不过近期也有一些研究致力于开发能够减少聚合程度的方法。

2.1. 警察使用武力与警察拦截

分析警方拦截的空间分布，有助于确定警察使用武力事件的基线预期。尽管警方拦截并不代表警方的所有接触，但我们可以利用警方拦截，进一步了解武力使用事件发生率与所有警方拦截的总体发生率不同的地点。某个社区的武力使用事件可能更多，并非因为该社区的任何特征，而是因为警察在该面积内巡逻得更多。Ba 等（2021a）指出，在时空范围内将数据汇总到较大的单元时，“观察到的行为差异可能仅仅反映了不同的巡逻环境，而非警务方式的差异。”警察存在的代理指标（例如警察实施的拦截）可以帮助确定警察的巡逻地点，因此有助于阐明哪些警察使用武力事件的发生超出了警察存在增加所预期的范围。Weisburst（2019）将武力使用事件视为逮捕的延伸，并比较逮捕和同一事件中随后发生的使用武力事件的人口统计模式。与逮捕相比，警方拦截能更贴近地作为警察存在的代理指标，因为并非所有拦截都会导致逮捕。

我们指出，警方拦截并未提供警方观察到的所有个体的完整信息，因为并非每次警民接触都会导致拦截。在一次巡逻中，警员面对所有接触对象时是否决定实施拦截，可能存在偏差。研究表明，在分析警察使用武力偏差时，考虑警察拦截偏差十分重要（参见 Knox 等，2020a）。未将平静无事的轮班纳入考量，可能会导致对警务偏差作出不准确的推断（Knox，2021）。不过，本研究旨在分析警察拦截与警察使用武力之间的空间分布，而非武力使用结果中的偏差。未来可以扩展这一框架，将有关警员和平民的信息纳入空间模型，例如采用 Kelling 和 Haran（2022）提出的两阶段框架。

2.2. 警察使用武力与暴力犯罪

暴力犯罪数据是确定警察使用武力的潜在原因和后果的另一项有用工具。特定区域内的暴力犯罪数量会形成

一种邻里情境，可能会对警察使用武力产生重要影响，因为警员可能掌握有关该社区的先验知识。Phillips 和 Sobol（2011）通过情境短文法和问卷调查发现，在暴力犯罪更多、且暴力犯罪率占人口比例更高的地区，与其他地区相比，警员更可能认为使用不必要武力是可以接受的；这一结论是在控制其他因素后得出的。在犯罪学文献中，有假设认为，暴力犯罪更多且处境不利的地点更可能发生更多警察使用武力事件（Lawton，2007；Terrill 和 Reisig，2003）。具体而言，许多现有研究利用暴力犯罪的空间分布，发现暴力犯罪与警察使用武力之间存在微弱的正相关关系（Lawton，2007；Lee 等人，2010；Terrill 和 Reisig，2003；Lee 等人，2014）。然而，另有一种相互矛盾的假设认为，暴力犯罪的发生率不会导致警察使用武力增加（Slovak，1988）。其他近期研究表明，要确定社区内的暴力犯罪究竟是警务的结果还是原因，或许“是不可能的”（Simckes 等人，2021）。我们开发了相关方法，以分析不同点过程之间的关系，例如暴力犯罪、警方拦截和使用武力，同时尽可能保留空间和事件层面信息，并避免指定病例对照设计所隐含的因果方向。

可用于与使用武力事件进行比较的“基线过程”有很多种，例如暴力犯罪、逮捕和拦截；它们各自提供的信息略有不同。本研究关注警方拦截与警察使用武力之间的关系，但未来可以进一步拓展研究范围。我们提出一种方法，用于研究警务数据集对警察使用武力的影响，同时保留两个数据集的点级特性。

2.3. 空间聚合的使用

研究暴力犯罪与使用武力之间的关系时，各项研究对暴力犯罪所采用的空间分析层级往往各不相同。许多研究采用社区层面的暴力犯罪指标，例如每一定数量居民对应的警察分区、指挥辖区或城市内暴力犯罪或杀人案件数量（Reisig and Parks，2003；Lawton，2007；Lee 等人，2010；Terrill 和 Reisig，2003）。图1展示了芝加哥警方拦截所采用的这种方法，其中每起使用武力事件都与其所在人口普查区的警方拦截次数相关联。除警方拦截所在的人口普查区外，不保留任何有关该次拦截的信息。

邻域层面暴力犯罪率存在局限，因为社区在空间上并非均质，这表明有必要在更低的聚合层级进行分析。为解决这些局限，Lee 等人（2014）使用径向缓冲区，统计每起武力使用事件一定半径范围内的暴力犯罪数量。径向缓冲区的半径从 500 英尺到 3,000 英尺不等。随后，在模型中将这些一定半径范围内的暴力犯罪数量用作协变量，而非按社区统计的数量。图 2 展示了这一通用方法：图中以红色标出芝加哥某个区域内部分武力使用事件。蓝色径向缓冲区用于统计每起武力使用事件周围 1.5 公里范围内发生的警方拦截（黑色）数量。

Lee 等人（2014）在径向缓冲区的四个层级上分别运行了四个多项逻辑回归模型，并发现微观层面的场所使用（使用径向

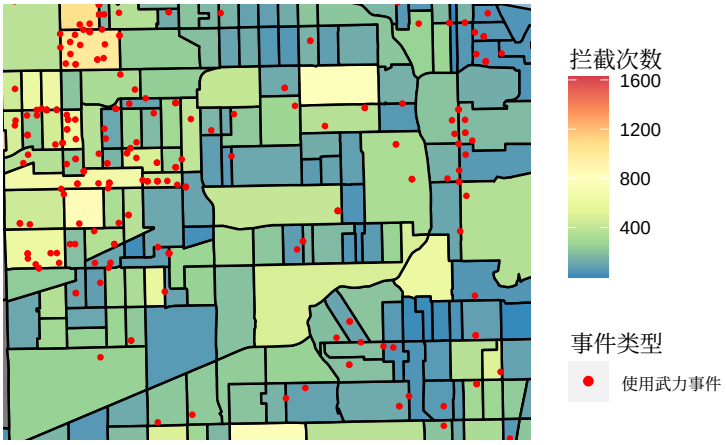

图 1. 芝加哥各人口普查区的警察使用武力事件及警方拦截次数

缓冲区）在对暴力犯罪与警察使用武力之间的关系进行建模时非常重要。结果表明，较小的径向缓冲区会导致暴力犯罪与警察使用武力之间更大的正向关系，而较大的径向缓冲区则可能会“模糊”暴力犯罪对警察使用武力的真实影响。这项研究的发现表明，在社区层面进行考察时，这种关系几乎可能完全消失。总体而言，这说明使用尽可能小的聚合层级，可以降低信息丢失的可能性。

径向缓冲区在确定暴力犯罪／警方拦截地点与警察使用武力之间的关系时，能够保留部分空间细节。我们可以创建小于人口普查区的缓冲区，且所有缓冲区的大小可以相同，如 Lee 等人（2014）所述；也可以根据人口密度等变量调整缓冲区大小。然而，这种方法仍属于聚合，因此会丢失许多关于暴力犯罪事件或警方拦截的精确信息。例如，我们必须固定径向缓冲区的范围，因此一旦设定大小，就无法得知某起使用武力事件周围 10 英尺和 500 英尺范围内分别发生了多少起犯罪。我们也无法保留关于暴力犯罪事件或警方拦截的精确点位信息或事件层面信息。这些信息可能包括警员和／或平民的种族与性别、是否涉及帮派、是否涉及武器，以及该事件是否属于仇恨犯罪。将数据聚合到径向缓冲区后，我们便无法在点过程模型中使用暴力犯罪事件或警方拦截的这类事件层面信息。

3. 方法

我们按如下方式构建一种空间点过程方法。设警察使用武力事件的 n 随机位置 x_1，...，$x_n \in D \subset \mathbb{R}^2$ 构成一个点过程，其强度函数为 $\lambda_1(s)$ $s \in D$。 D 是研究区域，例如芝加哥市。第二个点过程的m 随机位置 y_1，.....，$y_m \in D$（在我们的应用中为警方拦截）可通过多种方法与第一个点过程关联。在病例对照

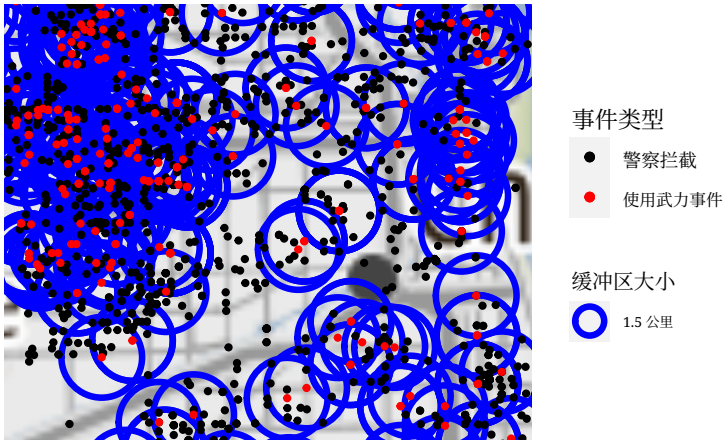

图 2. 芝加哥径向缓冲区示意图

模型中，该点过程的基线强度函数为 $\lambda_0(s)$；而在我们的共享成分模型中，它具有第二个强度函数 $\lambda_2(s)$。在我们的共享成分框架中，我们通过回归系数 β，将两个点过程都与空间协变量 $z(s)$关联。在病例对照模型中，只有病例过程（即我们数据中的武力使用事件）与空间协变量相关联。

我们的目标是在保留所有数据集点级信息的同时，关联两个或多个点过程。下文将介绍文献中的相关模型，重点关注标记点过程模型以及汇总统计量或检验统计量。接下来，我们将详细介绍两类模型：第一类是成熟的病例对照模型，第二类则是将共享成分模型创新性地用于点过程。我们将介绍一种贝叶斯推断方法，用这两类模型关联两个点过程。

3.1. 关联多元或多类型点过程的现有方法

多类型点过程的标记为类别标签，是描述两个或多个点过程之间关系的一种方法。在这类点过程中，标记决定给定位置上点的类型。由分类标记决定的不同类型点之间的关系，通常通过双变量点过程中的交叉协方差函数，或通过检验一阶特征是否相似来建模。例如，Fuentes-Santos 等（2017）提出了一种基于一阶性质对多类型点过程进行非参数比较的方法，这些性质描述了感兴趣区域内的空间分布。该方法用于分析西班牙不同类型的野火，依据的两个特征是规模（小型、常规型、大型）和起因（纵火、自然原因、疏忽、复燃和未知）。Berman（1986）检验空间点过程与其他空间随机过程之间的关系。研究使用新颖的空间数据类型，旨在分析铜矿床（一种点过程）与该区域内观测到的线性特征（称为“线性构造”）之间的关系；这些线性构造通常是道路，或由卫星观测到的地壳断裂带。Berman（1986）

提出了一种检验这两个过程之间关系的检验统计量，该统计量基于点过程（铜矿床）中的点到随机过程（线性构造）中最近点的距离。Fuentes-Santos 等（2017）和 Berman（1986）提出的检验方法与近期开发的其他多类型检验统计量类似（参见 Illian 等人，2008年；Møller 和 Waagepetersen，2003）。这些检验是实用的探索性工具，还可启发人们进一步通过点过程强度的参数模型开展研究。

Mohler（2014）使用标记点过程模型创建热点地图。此点过程框架中的类别标记用于指定不同的犯罪类型，包括凶杀以及可能先于凶杀发生的犯罪类型。Mohler（2014）通过自激点过程框架，确定事件 i 触发某一凶杀事件的概率，其中事件 i 可能是另一宗凶杀，也可能是另一种犯罪。所有非凶杀类犯罪都涉及手枪；这些数据用于分析枪支犯罪与预测未来凶杀之间的关系。这种创新的时空方法纳入了两个或更多标记，在考虑许多标记时很有用。该方法依赖短期和长期核密度估计，而我们采用参数方法来评估两个点过程之间的关系。触发函数或许是未来分析中纳入两个点过程时间维度的有用工具。

Liang 等人（2008）开发了一种双变量标记点过程模型，允许点过程模型中标记不同层级之间存在依赖关系。该模型包含空间变量及其回归系数，$\boldsymbol{z}(\boldsymbol{s})$ 和 $\boldsymbol{\beta}$，非空间变量及其回归系数，$\boldsymbol{\nu}$ 和 $\boldsymbol{\alpha}$，以及一个高斯过程 $\omega(s)$。点过程标记 k 的对数空间强度函数定义为 $\log(\lambda_{\boldsymbol{\theta}_k}(s,\nu)) = \boldsymbol{z}(\boldsymbol{s})'\boldsymbol{\beta_k} + \boldsymbol{\nu}'\boldsymbol{\alpha_k} + \omega_k(s)$。Liang 等人（2008）考察的标记是癌症类型，其中考虑了两种癌症类型。该模型通过高斯过程的交叉协方差函数 $\omega_k(s)$，允许两个标记层级之间存在灵活的空间依赖关系，$k = 1,2$。Kelling 和 Haran（2022）发现，由于空间强度函数中使用了非空间变量，模型的解释较为困难。此外，两个标记层级之间依赖关系的分析仅限于相关高斯过程的 ρ 参数，其交叉协方差函数 Λ定义如下。参数 ρ提供了有关空间残差模式中依赖强度的信息，但我们的目标是更详细地分析两个点过程之间的关系。

$$\Lambda = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2. \end{bmatrix}$$

3.2. 点过程的病例对照方法

病例对照方法是用于比较两个点过程的一类常见模型。Guan 等（2008）采用病例对照点过程方法，将1990年鸟类调查（基线/对照过程）与2004年较新的调查（主要/病例过程）联系起来。结果显示，金鸽的空间分布随时间发生了变化，且与坡度和棉草覆盖率等空间协变量有关。Diggle 等人（2007）采用病例对照方法研究热带雨林中幼树与成年树之间的关系。幼树被视为对照组，

它们代表环境变异的基础；成年树则被视为病例组，其空间分布会受到影响生存的空间条件（例如海拔）的影响。Guan 等（2010）也研究了热带雨林中树木之间的关系，其中将死亡的幼树视为病例组，将新生树木视为对照组，以研究幼树的死亡率。这项分析中使用的空间协变量包括海拔、坡度以及土壤含量信息。Guan 等（2010）研究的第二个例子包含了 Guan 等（2008）的金斑鸽数据，并使用海拔、坡度以及石楠和棉草的覆盖率等空间变量。Diggle 等（2000）和 Chetwynd 等（2001）将病例对照方法扩展到匹配病例-对照数据，根据可能相关的混杂因素，为每个病例分配一组匹配的对照组。

为引入病例对照方法的记号，在位置 s处，所关注的主要点过程的强度函数记为 $\lambda_1(s)$，称为病例过程。在位置 s处基线过程的强度，即对照，记为 $\lambda_0(s)$。空间变量记为 $\boldsymbol{z}(\boldsymbol{s})$ ，相应的回归系数为 $\boldsymbol{\beta}$ ，而截距项记为 α。病例-对照点过程中位置 s 的完整强度函数如下（参见 Diggle 等人，2007年）：

$$\lambda_1(s) = \lambda_0(s) \exp\left(\alpha + \boldsymbol{z}(\boldsymbol{s})'\boldsymbol{\beta}\right) \tag{1}$$

通常还采用一种略有不同的参数化方式，其中截距项按如下方式缩放总体强度函数：$\lambda_1(s) = \alpha\lambda_0(s) \exp(\boldsymbol{z}(\boldsymbol{s})'\boldsymbol{\beta})$（参见 Diggle，1990年；Diggle 和 Rowlingson，1994年；Guan 等人，2008年）。在这种表述中，$\lambda_0(s)$ 是风险人口的强度，α 是一个尺度参数，表示病例相对于对照的发生率，通常不被视为主要的科学研究对象（Diggle 和 Rowlingson，1994 年）。$\lambda_1(s)$ 的其余部分，即 $\exp(\boldsymbol{z}(\boldsymbol{s})'\boldsymbol{\beta})$，根据空间变量确定病例的升高风险。在这些病例对照模型中，不对基线强度 $\lambda_0(s)$.假设任何参数形式。

在许多情况下，由于估计基线强度 $\lambda_0(s)$会带来复杂问题，因此通常避免直接对病例点过程 $\lambda_1(s)$的完整空间强度函数进行参数估计。Guan 等（2008）估计主要/病例过程的成对相关函数，以刻画主要过程在考虑基线过程后呈现的二阶结构，例如聚集。Guan 等（2010）提出了研究主要点过程二阶结构的非参数和参数方法，并提出了无需估计对照强度$\lambda_0(s)$即可估计主要过程回归系数的方法。模拟研究固定了回归系数 $\boldsymbol{\beta}$，但没有评估估计程序恢复回归参数的能力。相反，该模拟研究着重考察成对相关函数估计量的偏差和标准差，以评估聚集情况。我们还注意到，研究假设病例过程与对照过程相互独立。Diggle 和 Rowlingson（1994）也通过采用非线性二元回归模型来估计 $\boldsymbol{\beta}$，从而避免估计 $\lambda_0(s)$。

一些病例对照方法通过核密度估计对基线强度函数 $\lambda_0(s)$ 进行非参数估计，再用这些估计值估计病例过程的强度 $\lambda_1(s)$（Diggle 等人，2000，2007）。当我们

希望将完整的空间强度函数 $\lambda_1(s)$ 可视化并加以解释时，这种方法很有用。Diggle 等人（2007）提出了一种简单的估计程序，无需估计对照强度 $\lambda_0(s)$，即可估计病例对照模型中的回归系数。系数 α 和 $\boldsymbol{\beta}$ 通过逻辑回归进行估计，其中 $Y(s)$ 是二元变量，用于指示给定位置 s 属于病例（1）还是对照（0）。逻辑回归形式如下：$P(Y(s) = 1|\boldsymbol{z}(s)) = \exp(\alpha+\boldsymbol{z}(s)'\boldsymbol{\beta})\ 1+\exp(\alpha+\boldsymbol{z}(s)'\boldsymbol{\beta})$。如果需要估计病例强度 $\lambda_1(s)$，则通过核密度估计来估计对照强度 λ_0 。结合逻辑回归和核密度估计的结果，即可得到强度函数 $\lambda_1(s)$ 的估计值。Diggle 等人（2007）将回归系数 $\boldsymbol{\beta}$ 解释为 $\boldsymbol{z}(\boldsymbol{s})$ 对位置 s 处病例相对于对照的相对强度所产生的影响。截距α 则解释为病例与对照之间的“选定比例”。

Hessellund 等人（2021）探讨了 Diggle 等人（2007）所阐述的病例对照逻辑回归模型的多变量扩展。重要的是，这种多变量病例对照模型并不假设不同点过程之间相互独立。Hessellund 等人（2021）分析了华盛顿特区的六类街头犯罪：抢劫、汽车盗窃、车辆盗窃、袭击、入室盗窃和其他盗窃。每种点格局 X_i 的强度定义为 $\lambda_i(\boldsymbol{u};\ \boldsymbol{\gamma}_i) = \lambda_0(\boldsymbol{u})\ \exp\left(\boldsymbol{\gamma}_i{}'\boldsymbol{z}(u)\right)\ \ i = 1,\ \ldots\ldots,\ p$。空间协变量 $\boldsymbol{z}(u)$ 定义在位置 $\boldsymbol{u}$ ，其回归系数为 $\boldsymbol{\gamma}_{i}$。背景强度定义为 $\lambda_0(\boldsymbol{u})$，其含义是“潜在因素（如城市结构和人口密度）的空间效应，并假设对所有点类型均相同”（Hessellund 等人，2021）。研究表明，参数 $\gamma_{\mathbf{i}}$ 不可识别，因此回归系数 $\boldsymbol{\beta}_i$的定义满足 $\boldsymbol{\beta}_i = \boldsymbol{\gamma}_i - \boldsymbol{\gamma}_p,\ \ i = 1,\ \ldots p-1$，其中 X_p为基准点过程。基准点过程的回归系数 $\boldsymbol{\gamma}_p$ 设为 0，因此 $\lambda_0(\boldsymbol{u})$ 即为基线过程的强度。在对华盛顿特区街头犯罪的应用中，犯罪类别“其他盗窃”被设为基准类别，其他所有类别均与这一点过程进行比较。

Hessellund 等人（2021）再次通过多项逻辑回归来估计回归系数 $\boldsymbol{\beta}$ ，从而避免估计基线强度 $\lambda_0(s)$；这与 Diggle 等人（2007）的方法类似，该方法不依赖于 $\lambda_0(s)$。为了解释华盛顿特区的街头犯罪，Hessellund 等人（2021）绘制了基于该多项逻辑回归的条件概率图。为了绘制 $\lambda_i(s)$，作者随后通过核估计来估计 $\lambda_0(s)$。

Xu 等人（2019）提出了病例对照模型的一种修订版本，其中空间变化的对照过程按抽样方案 $\alpha(s)$ 进行缩放，而该方案通常假定为已知。在 Xu 等人（2019）对北京餐馆的分析中，假定对照组的抽样率为 6%，且采用均匀抽样。对照组的强度 $\lambda_M(s) = \alpha(s)\lambda_0(s)$ 和病例组的强度 $\lambda_N(s) = \lambda_0(s)\ \exp\left(\boldsymbol{z}(\boldsymbol{s})'\boldsymbol{\beta}\right)$ 均依赖于过程 $\lambda_0(s)$。同样，由于策略性地利用比例强度函数估计回归系数， $\lambda_0(s)$ 的估计得以避免。由于担心 $\lambda_0(s)$ 的不一致估计及其对回归系数推断的影响，项 $\lambda_0(s)$ 被称为“无限维干扰参数”。假定病例过程与对照过程相互独立。

在对模拟示例和芝加哥警务数据进行分析时，我们考虑了病例对照模型的两种估计方法：逻辑回归，以及将其作为非齐次泊松过程的空间强度函数。本文中的所有模型

均通过贝叶斯分层框架实现，但逻辑回归例外；逻辑回归通过最大似然和 R 中的广义线性模型 函数进行估计。

3.3. 点过程共享成分模型

我们基于 Knorr-Held 和 Best（2001年）的面域数据框架，构建了一个用于点过程的共享成分模型。与病例对照模型类似，我们用一个项 $\lambda(s)$ 对强度函数进行缩放。不过，在我们的模型中，这个共享成分由参数 δ 加权，并同时作用于两个点过程，而不是表示其中一个点过程的强度。我们根据数据推断这一共享成分，而非通过非参数方法计算它，或将其忽略；而在病例对照方法中，对照过程通常会被如此处理。Knorr-Held 和 Best（2001年）研究了口腔癌和食管癌这两种不同癌症的空间分布；研究表明，它们具有共同和各自独特的风险因素。类似地，我们希望推断多种警务数据之间共享的空间模式，包括警察使用武力、警方拦截和暴力犯罪事件。

我们的共享成分模型如方程2所示，其中 $\lambda_1(s)$表示一个点过程的空间强度， $\lambda_2(s)$ 表示第二个点过程的空间强度。每个点过程的空间变量 $\boldsymbol{z}(s)$及其对应的回归系数 $\boldsymbol{\beta_1}$ 和 $\boldsymbol{\beta_2}$ ，使我们能够了解各自独有而非彼此共享的空间模式。例如，我们可以假设人口效应主要由共享成分 $\lambda(s)$捕捉，而其他邻里效应可能是每个点过程所独有的。通过模拟研究，我们发现，在两个强度函数中纳入相同的空间变量，仍能准确估计相应的回归系数，不过这一点还应针对每种应用和具体情境进行更详细的检验。

$$\begin{aligned} \lambda_1(s) &= \lambda(s)^{\delta} \exp\left(\boldsymbol{z}(\boldsymbol{s})'\boldsymbol{\beta_1}\right) \\ \lambda_2(s) &= \lambda(s)^{1-\delta} \exp\left(\boldsymbol{z}(\boldsymbol{s})'\boldsymbol{\beta_2}\right) \end{aligned} \tag{2}$$

方程2中给出的共享成分模型有许多可以考虑的变体。该模型最简单的版本是考虑一个不随空间变化的共享成分，即对所有 s， $\lambda(s) = \lambda$ 。通过模拟研究，我们发现，在这种情况下，无法在任一点过程的强度中加入截距项，因为它们会与共享成分混淆；共享成分实质上是两个点过程共有的截距项。在我们的分析中，我们重点研究空间变化的共享成分，其中 $\log(\lambda(s))$ 是一个高斯过程。由于估计一个 $n \times n$ 协方差矩阵的计算负担较重，我们采用预测过程变换来估计高斯过程（参见 Banerjee 等人，2008年）。我们在少量结点上估计高斯过程，然后利用结点与数据点之间的协方差函数，将该高斯过程变换到数据点上。在本研究中，我们使用了82个在研究区域内均匀分布的结点，如图7所示。利用结点之间以及数据点与结点之间的协方差函数，将结点上的高斯过程（$\lambda^*(s)$）变换为数据点上的高斯过程估计值（$\tilde{\lambda}(s)$），如下所示：

$$\tilde{\lambda}(s) = \text{cov}(\lambda(s), \lambda^*(s))\text{var}^{-1}\left(\lambda^*(s)\right)\lambda^*(s)$$

加权参数 δ 的参数化方式也有多种选择。我们重点研究两种参数化方式：如方程2所示，对每个点过程分别赋予 δ 和 $1-\delta$ 权重；以及分别对两个点过程赋予 δ 和 $1/\delta$ 权重。第一种情况下，δ 的取值范围为0到1，因此我们为 δ 使用均匀先验（0,1）。也可以使用其他有界分布，例如贝塔分布。第二种情况下，δ 的取值不受0到1的限制。按照 Knorr-Held 和 Best（2001）的建议，δ 的对数服从正态先验分布。总体而言，我们发现，与将第二个点过程加权为 $1/\delta$ 相比，将共享成分的权重设为 δ 和 $1-\delta$ 的参数化方式能够得到可靠的参数估计，且更易于解释。需要指出的是，在两个例子中，我们都对 δ 加以约束，使共享成分对两个点过程均作正向贡献（≥ 0）。尽管共享成分 $\lambda(s)$ 必须为正，但 δ 的其他参数化方式可以允许共享成分对点过程强度函数作负向贡献。

共享成分模型的似然如下所示。要估计该似然，我们必须估计感兴趣区域 W 内各点过程强度函数 $\lambda_1(s)$ 和 $\lambda_2(s)$ 的积分。我们使用蒙特卡洛平均值来估计该积分，即计算各强度函数在积分点处的取值平均值，相关内容见第8.1节。随后，分别针对两个点过程的所有数据点，将两个点过程的强度函数相乘（第一个点过程的 n 个数据点，第二个点过程的 m 个数据点）。

$$\mathcal{L}(\boldsymbol{\theta}; \{s_{1:n}\}, \{s_{1:m}\}, s_i \in W) \propto \exp\left(-\int_W \lambda_1(s)ds - \int_W \lambda_2(s)ds\right) \times \prod_{i=1}^{n} \lambda_1(s_i) \times \prod_{j=1}^{m} \lambda_2(s_j)$$

最后，我们可以考虑对强度函数中各点过程特有的部分采用不同的参数化方式。例如，可以考虑两个强度函数中空间变量的不同组合，包括两个点过程使用不同或相同的变量。这是我们的共享成分方法的一项优势；病例对照方法不包含对照过程强度的参数估计。在方程2所示的参数化中，对于两个点过程强度函数中不共享的部分，我们均采用非齐次泊松过程（NHPP）。未来可以探索该强度函数的其他形式，例如对数高斯考克斯过程形式。在这种情况下，必须注意避免共享成分与各点过程特有的高斯过程之间产生混杂。我们注意到，在两个过程均采用NHPP的框架下，当截距项 β_0 仅包含在其中一个点过程中，而非同时包含在两个点过程中时，我们得到的估计值最为可靠。

我们的共享成分模型并不假设一个过程是“基线过程”，而另一个过程是“病例”或“主要”过程。这有许多优点，尤其是在因果关系方向未知时。例如，在警务方面，上文提到的许多研究都指出，暴力犯罪构成了社区的背景情况，可能会影响警务行为。然而，也有一些警务研究，例如破窗理论，认为警务行为也会影响暴力犯罪的发生率。具体而言，破窗理论指出，如果

警方容忍“轻微”犯罪，那么在特定区域内可能会有更多罪犯实施犯罪（Wilson 和 Kelling，1982年）。一个社区警察存在增加，也与犯罪向其他区域发生空间转移有关（参见 Ratcliffe，2002年）。鉴于警务行为可能影响犯罪，而犯罪也可能影响警务行为，我们的共享成分模型适用于这项分析。我们指出，共享成分模型可以写成病例对照模型的一种形式，其中缩放过程现在取决于第二个过程的强度函数、上文定义的参数 δ，以及第二个点过程的空间协变量和系数。第8.4节将对此进行更详细的讨论。

我们介绍两种有趣的基于病例对照的方法，并将其结果与我们所构建的共享成分模型进行比较。在 Xu 等人（2019）提供的热带雨林数据示例中，没有明显可用于研究这三个树种的对照过程。因此，使用齐次泊松过程作为对照过程，并将这三个树种均视为病例组，与均匀分布的对照进行比较。Xu 等人（2019）指出，齐次泊松过程的规模由变化的 α 确定，其中 $\lambda_0(s) = 1$，实际上会影响病例过程的系数估计。我们在此构建的共享成分模型避免了将某一过程指定为对照过程。在 Guan 等（2010）提出的模拟病例对照示例中，作者假设对照强度采用以下形式：$\log \lambda_0(s) = \log \eta + X^*(s)$。假设病例过程强度采用以下形式：$\log \lambda_1(s) = \log \gamma + \beta X(s) + X^*(s)$。假设过程 $X(s)$ 已知，而$X^*(s)$未知。项 $X^*(s)$类似于我们共享成分框架中两个点过程之间的未加权共享成分。

4. 数据

我们通过模拟数据以及来自芝加哥的警务数据集，评估病例对照方法和共享成分方法。本节将介绍模拟数据集的生成过程，以及来自芝加哥的警务数据集。

4.1. 模拟数据

本文使用空间稀疏化过程生成所有模拟点过程。首先，我们在空间窗口上模拟一个齐次泊松过程，其参数对应于窗口面积以及空间窗口内可能的最大强度。随后对该点过程进行稀疏化，保留某个点的概率等于该点处的强度除以区域内的最大强度。空间强度函数取决于一个或多个空间变量，如图 3 所示。所有空间变量和点过程均在单位正方形上模拟。

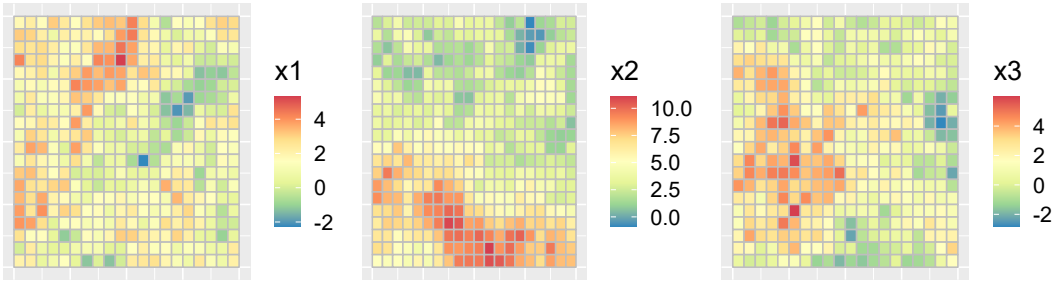

图 3. 我们在单位正方形上的格点中模拟三个空间变量，随后用它们在病例对照模型和共享成分模型下模拟点过程。该格点是区域数据的一个简单示例；在芝加哥案例中，我们将把它扩展到人口普查区。模拟示例中，我们使用这些空间变量的不同组合来确定一个或两个点过程的空间分布。

对于病例对照模型，我们使用参数化强度函数模拟基线/对照过程，以使点过程在空间上呈现非齐次分布。我们在公式 1 中将该参数化强度函数记为 $\lambda_0(s)$。第一个点过程的强度 $\lambda_0(s)$和空间变量用于模拟第二个点过程，其依据的是完整的空间强度 $\lambda_1(s)$，如公式 1 所示。重要的是，我们使用空间强度模型设定而非逻辑回归来模拟病例对照数据，因此预计该空间强度函数在恢复参数方面表现良好。

对于共享成分模型，两个点过程是同时模拟的，而非依次模拟。首先，我们模拟一个高斯过程，作为共享成分的对数，因此共享成分始终大于零。我们将 $\log(\lambda(s)) \sim GP(0, \Sigma)$ 记作其中 Σ 是单变量指数协方差函数，两个位置 s_i, s_j 之间的协方差定义为 $\Sigma = \sigma^2 \exp\left(\frac{-|s_i - s_j|}{\phi}\right)$，其中$\sigma^2, \phi > 0$。在模拟研究和真实数据应用中，$\phi$ 参数是固定的，不根据数据进行估计；由于可识别性问题，其他研究中也通常如此（参见 Liang 等人，2008年）。我们计算 ϕ 的取值，使所有点之间距离的 95th 百分位数对应的相关性为 0.05；并计算 ϕ 的取值，使距离的 5th 百分位数对应的相关性为 0.95。我们将 ϕ 固定为这两个值的平均值。参数 σ 在模拟研究中也固定不变，但在估计模型参数时则作为未知量处理。模拟高斯过程后，我们固定权重参数 δ，它决定共享成分对两个点过程各自的贡献，同时固定每个点过程的回归系数。最后，如方程2所示，我们使用空间稀疏化，根据相应的空间强度函数生成两个点过程。

4.2. 芝加哥警务数据

接下来，我们介绍本研究使用的警察使用武力数据集和警察拦截数据集。Ba 等（2021a）通过信息公开申请及申诉，从芝加哥警察局获取了详细数据。我们使用发布在

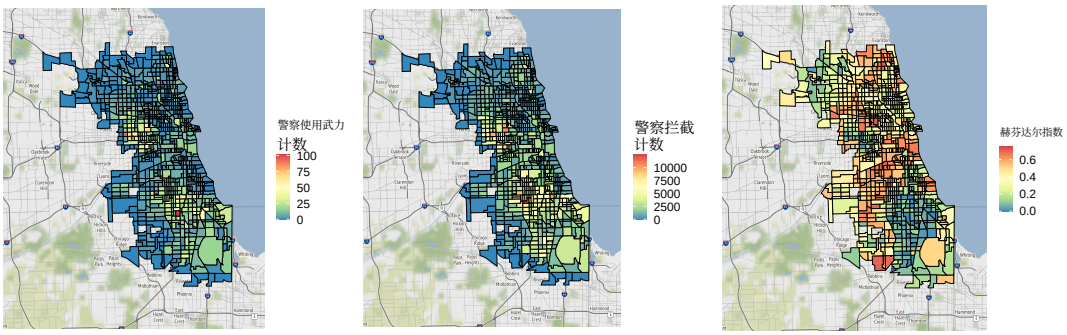

图4。按人口普查区划分的警察使用武力次数、拦截次数和赫芬达尔指数

Code Ocean capsule 中的复现数据（Ba 等，2021b）。武力使用数据和警察拦截数据均涵盖 2012—2015 年。需要指出的是，许多事件没有提供坐标（纬度/经度）。具体而言，18.8% 的使用武力事件和 14.6% 的警方拦截没有可用坐标。在本分析中，我们剔除了这些点，但未来的研究可能需要进一步调查这些点缺失的情况。

芝加哥警察局的武力使用数据集包含2012年1月至2015年12月期间的9,293起事件，其中7,539起具有完整的位置信息。该数据还包括事件涉及人员的信息，如平民种族、性别、年龄和受伤状况，以及警员ID。警方拦截数据集包含2012年1月至2015年12月期间的1,703,158起事件，其中1,453,832起具有完整的空间信息。该数据集还包括拦截类型、平民种族、性别和年龄以及警员ID等信息。为便于可视化，我们在图4中按区域层面绘制了这两个数据集。我们注意到，这两个结果具有共同的空间模式：城市边缘的数量较少，而城市南部中心的数量较多。请注意，为了便于对数十万个点进行可视化，我们仅在区域层面绘制了这些数据——我们的分析是在连续空间中的点层面进行的。

除了分析芝加哥警务数据集之外，我们还从美国人口普查和美国社区调查中收集社会经济信息。我们使用完全包含在警察辖区内的人口普查区；这些警察辖区数据从芝加哥开放数据门户下载。我们重点关注从人口普查数据中收集或计算得出的三个社会经济变量：年龄中位数、失业率，以及衡量社区多样性的赫芬达尔指数。图 4 展示了芝加哥各人口普查区的赫芬达尔指数。我们指出，本文介绍的模型可灵活适应不同的协变量选择，具体取决于应用场景和研究问题。

5. 结果

我们使用模拟数据集和芝加哥的实际警务数据，对病例对照模型和共享成分模型进行比较。我们针对病例对照模型测试了许多不同的参数设置，以说明对常用估计方法应谨慎解读。我们还将共享分量方法应用于模拟数据和芝加哥警务数据，发现它为分析两个点过程之间的关系提供了灵活的框架。

对于所有参数的估计，我们使用 R 中 NIMBLE 软件包实现的马尔可夫链蒙特卡罗（MCMC）（de Valpine 等人，2017）。我们的方法采用 Metropolis-Hastings自适应随机游走采样器，并为回归系数使用单变量正态分布提议分布（Normal(0,100)）（de Valpine 等人，2017）。与 Liang 等人（2008）一样，我们对共享成分协方差函数相关的参数 σ 使用 Inverse-Gamma($\alpha = 2, \beta = 0.5$) 先验分布。我们通过轨迹图、有效样本量（Gong 和 Flegal，2016）以及蒙特卡罗标准误（Haran 和 Hughes，2020；Flegal 等人，2021）评估收敛性。

5.1. 模拟数据的结果

通过模拟研究，我们发现，病例对照模型的解释会因模型假设而面临一些挑战。我们从病例对照模型中进行模拟，将其视为空间强度函数，并为回归系数设定许多不同的模拟参数值。为此，首先我们从一个点过程中模拟数据。随后，我们根据方程 1 中所述的空间强度函数，使用基准点过程的真实强度函数 $\lambda_0(s)$，模拟第二个非齐次泊松过程（NHPP）。在实际应用中，我们通常不知道 $\lambda_0(s)$ 的函数形式，因此必须使用核密度估计（KDE）对其进行估计，或完全避免估计。下面我们讨论两种可能的估计程序。

根据与病例对照模型对应的空间强度函数模拟病例和对照非齐次泊松过程（NHPP）后，我们使用空间非齐次泊松过程和逻辑回归这两种估计方法，检验病例强度函数的参数恢复情况。对于第一种情况，我们通过贝叶斯方法，利用以下似然函数估计空间非齐次泊松过程的参数，其中 $\lambda_1(s)$ 是病例强度函数，包含回归系数 α 和 $\boldsymbol{\beta}$ 以及对照强度 $\lambda_0(s)$。我们为回归系数使用均值为 0 且相互独立的正态先验。我们不假设已知基线强度函数，因此在估计病例强度函数参数时，使用对照强度函数 $\lambda_0(s)$ 的核密度估计（KDE）估计值。

$$\mathcal{L}(\alpha, \boldsymbol{\beta}; s_1, ..., s_n, s \in W) \propto \exp\left(-\int_W \lambda_1(s)ds\right) \times \prod_{i=1}^{n} \lambda_1(s_i)$$

对于第二种方法，我们采用第 3.2 节所述的逻辑回归。估计回归系数时，这种方法无需使用基线强度函数 $\lambda_0(s)$。如果这两种方法基于相同数据估计参数

表1。 单位正方形上模拟数据集的带置信区间/可信区间的估计回归系数（分别对应逻辑回归/非齐次泊松过程）。

模型	β_0	β_1 (x_1)	β_2 (x_2)
真实值	-10.0	1.7	0.8
非齐次泊松过程估计（积分）	-12.93 (-13.17, -12.70)	1.75（1.71，1.79）	0.84（0.82，0.86）
逻辑回归估计（积分）	-9.84 (-10.26, -9.42)	1.70 (1.63, 1.79)	0.78（0.74，0.82）
真实值	-10.0	2.1	-0.1
非齐次泊松过程估计（积分）	-12.40 (-12.59, -12.22)	2.08（2.05，2.12）	-0.10 (-0.12, -0.09)
逻辑回归估计值（截距）	-5.29（-5.58，-5.01）	2.14 (2.06，2.23)	-0.14 (-0.16, -0.11)

表2。 单位正方形上模拟数据集的估计回归系数及可信区间，其中估计了 $\lambda_0(s)$个参数。

模型	$\beta_0 + \beta_{0,\text{对照}}$	β_1 (x_1)	β_2 (x_2)	$\beta_{1,\text{对照}}$
真实值	-5.8	2.1	-0.1	0.7
非齐次泊松过程估计（积分）	-5.75（-5.96，-5.54）	2.09（2.04，2.13）	-0.10（-0.11，-0.08）	0.69（0.66，0.71）

得到的参数估计相近，那么我们就可以灵活选择估计方法及其对应的解释。

如表1所示，当所有参数均为正时，逻辑回归方法能够很好地恢复回归参数。使用 NHPP 方法估计回归参数时则面临一些挑战，这是因为该方法使用了基线强度函数的 KDE 估计值。纳入负参数后（如第二个模拟示例所示），逻辑回归无法很好地恢复截距参数，其他一些参数估计值也会受到影响。我们检验了这些结果对缩放的敏感性（例如，将截距参数设为 -1 而非 -10），但发现使用逻辑回归估计 β_0 仍然存在挑战。我们给出了使用 MCMC 对 NHPP 进行估计所得的可信区间，以及通过 R 中的 glm函数估计的逻辑回归置信区间。根据这些模拟数据的结果，逻辑回归是 NHPP 估计的一种颇具前景的替代方法，但在某些情形下仍然会面临挑战。

在第二组模拟示例中，我们假设已掌握一些关于基线强度函数的信息，并检验在这些假设下，NHPP 方法能否准确估计参数。我们从强度函数为 $\lambda_0 = \exp(\beta_{0,\text{control}} + \beta_{1,\text{control}} x_3)$ 的对照过程，以及强度函数为$\lambda_1(s) = \lambda_0(s)\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ 的病例过程中模拟生成数据。我们假设已知基线强度函数的形式，并估计病例强度和对照强度对应的回归系数。需要指出的是，截距无法识别，因此我们估计两个点过程的截距参数之和。如表2所示，假设我们了解基线强度结构的部分信息时，使用 NHPP 方法能够准确估计系数。在实际应用中，可能无法为基线强度指定结构。

根据这些发现，如果假设应将结果解释为空间强度函数，我们建议在解读 NHPP 和逻辑回归估计方法 Diggle 等人（2007）的结果时保持谨慎。我们发现，

表 3. 单位正方形上的模拟数据集的估计回归系数和共享分量参数，及其 95% 可信区间。

模型	点过程1 β_1 (x_1)	β_2 (x_2)	点过程2 β_0	β_1 (x_3)	共享成分 δ	σ
真实值	0.12	0.06	0.1	0.25	0.3	1.7
估计值, Unif 可信区间	0.13 (0.11, 0.15)	0.06 (0.05, 0.07)	0.06 (-0.07, 0.20)	0.25 (0.23, 0.28)	0.30 (0.26, 0.34)	1.92 (1.48, 2.40)
估计值, Norm 可信区间	0.12 (0.10, 0.15)	0.05 (0.04, 0.06)	-0.76 (-1.21, -0.36)	0.29 (0.26, 0.31)	0.29 (0.27, 0.32)	2.18 (1.81, 2.58)

这两种估计程序可能得出不同的结果，因此应考虑这些参数的预期解释。逻辑回归具有许多优点，包括在估计回归系数时无需估计基线强度函数 $\lambda_0(s)$，且计算效率高。与对非齐次泊松过程强度函数进行完整的贝叶斯估计相比，我们展示了这种方法的一些优势。我们还展示了非齐次泊松过程方法的一个优点：当我们能够假设基线强度函数具有某种结构时，该方法便可发挥优势，但这并非总是可行的。在考虑这两种方法的估计与解释能力及目标时，应仔细权衡它们各自的优势。

接下来，我们通过模拟数据示例评估本文提出的共享成分模型。对于权重参数 δ 的分布，我们考虑两种情况。第一种情况下，第一点过程强度中的共享成分以 δ 加权，第二点过程强度中的共享成分以 $1-\delta$ 加权。在这种情况下，我们对 δ 使用均匀分布（0,1）作为先验分布。第二种情况下，共享成分对第二点过程强度的贡献由 $1/\delta$ 加权，而非 $1-\delta$，且 $\log(\delta)$ 的先验分布为正态分布。第二种情况选用正态分布，是因为 Knorr-Held 和 Best（2001）在区域数据的共享成分模型中采用了这一分布。

通过共享成分模型的模拟病例，我们发现，对加权参数 δ 采用均匀先验的共享成分模型更容易解释，且参数估计更可靠。我们根据这两种加权方案进行模拟，并分析参数的后验估计。附录第 8.2 节中，我们展示了这些点过程的图，其中包含参数相同但分布不同的情况。在表 3 中可以看到，对于使用 δ 和 $1-\delta$ 作为共享成分权重的情况，所有参数的可信区间均包含真实参数。而在使用 δ 和 $1/\delta$ 的情况下，一些参数未能通过模拟研究准确恢复，具体而言，参数 σ（共享成分的协方差函数参数）和 β_0（第二点过程的截距）即是如此。因此，我们建议在共享成分模型中使用均匀先验，并采用 δ 和 $1-\delta$ 的加权方案。

在分析共享成分病例的模拟数据时，我们还注意到，当两个截距均被采用（两个点过程都使用β_0 ）时，这两个参数之间会产生混杂，因而无法准确估计。因此，我们在其中一个点过程中使用一个截距项，并在另一个点过程中省略截距

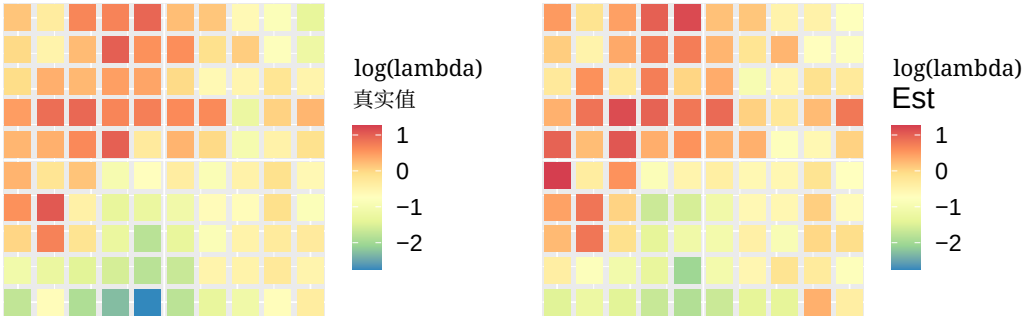

图 5. 共享成分：真实值与估计值（均匀分布）

项。我们还注意到，我们测试了在两个点过程的强度函数中纳入相同空间协变量的情况；即使两个点过程的强度函数中都纳入了相同的空间协变量，我们仍能得到准确的参数估计，不过这一点还需要更深入的研究。在图5中，我们将通过模拟生成的真实共享成分，与采用均匀先验以及权重 δ 和 $1-\delta$得到的共享成分后验均值估计进行比较。我们发现，能够相对准确地恢复共享成分。

5.2. 芝加哥警方拦截与使用武力的结果

来自芝加哥的点级数据集使我们能够运用病例对照模型和共享成分模型，详细分析警察使用武力与警方拦截之间的关系。Ba 等（2021a）将我们在此研究的警务数据汇总为警员轮班的面板数据集，并按警区进行空间汇总。比较不同“月份、星期、班次和警区”下的警察行为，可以对人口统计特征不同但巡逻任务分配相似的警员进行比较。Ba 等（2021a）采用带有月份、班次、星期和警区固定效应的普通最小二乘法，确定警员和公民特征对使用武力结果的影响。我们对 Ba 等（2021a）使用的芝加哥数据进行空间分析，但保留警方拦截和武力使用事件的精确空间信息，而不是按警区汇总。这一建模框架使我们能够纳入宏观制度因素，例如是否决定向特定社区增派警员；此前一些研究有意未考虑此类因素（Knox 等，2020a、b）。

首先，我们采用两种估计程序——逻辑回归和利用核密度估计的空间非齐次泊松过程模型——将病例对照模型拟合到芝加哥的警察使用武力和拦停数据上。我们还估计了两组不同的空间协变量：一组包含赫芬达尔指数和失业率，另一组则以年龄中位数替代失业率。我们发现，这两种估计程序得出的结果不同，如表4所示。对于逻辑回归，年龄中位数的回归系数的置信区间包含0；而使用非齐次泊松过程模型时，该回归系数的估计值为负。我们列出了逻辑回归的95%置信区间，以及非齐次泊松过程模型的95%可信区间。

表4. 将病例对照模型应用于芝加哥警察使用武力和警察拦截数据时的估计回归系数及其95%置信区间/可信区间（分别对应逻辑回归/非齐次泊松过程）。

模型	β_0	β_1 (赫芬达尔指数)	β_2 (失业率)	WAIC/AIC
NHPP	-2.10（-2.17，-2.03）	-0.88（-1.01，-0.76）	-0.49（-0.74，-0.24）	145,244
逻辑回归	-4.89（-4.96，-4.82）	-0.62（-0.74，-0.50）	-1.19（-1.45，-0.93）	94,338
模型	β_0	β_1 (赫芬达尔指数)	β_2 (中位年龄)	WAIC/AIC
NHPP	-1.75（-1.89，-1.62）	-0.83（-0.94，-0.72）	-0.01 (-0.02, -0.01)	145,210
逻辑回归	-5.21（-5.35，-5.07）	-0.34 (-0.44, -0.23)	0.001 (-0.003, 0.005)	94,420

我们注意到，在大多数病例中，可信区间并不重叠。我们还使用广义威布尔信息准则评估非齐次泊松过程模型的拟合效果，并使用赤池信息准则评估逻辑回归的拟合效果；两种准则的值越低，都表示模型拟合效果越好。我们发现，模型拟合检验显示，在这两种估计程序中，另一种模型的拟合效果更好。使用两种估计程序拟合相同模型，对芝加哥的使用武力和拦停数据进行分析后，仍有许多尚待解答的问题。我们认为，或许最稳妥的做法是将此处通过拟合逻辑回归得到的回归系数解读为影响某个事件成为病例而非对照的因素。这与将结果解读为影响病例空间分布、并以对照组为基准进行缩放的因素不同。

接下来，我们使用共享成分模型分析芝加哥警察拦截和武力使用数据集。我们采用 Xu 等人（ 2019 ）的方法，对其中一个点过程进行抽样。由于拦截数据集中的点数量较多，我们对其进行空间稀疏化，使每个点保留在数据集中的概率相同。在我们的案例中，稀疏化概率设为 10%。我们在附录第 8.3 节详细介绍了这一方法。在分析中，我们只关注描述武力使用事件与警方拦截的空间分布及其空间关系。如果我们还关心拦停数据的其他特征，例如涉事公民或警员的人口统计特征，则应考虑使用完整停车数据，或采用一种能够顾及这些感兴趣变量的其他抽样方法。

在表 5 中，我们列出了两个点过程的强度函数估计系数，以及与共享成分相关的参数 σ 和 δ。我们采用了表 5 中描述的权重项 δ 的两种参数化方式，其先验分布分别为正态分布或均匀分布。可以看到，共享成分模型的两种不同参数化方式在协变量效应 β 的估计结果上相似。两种参数化方式的点估计和可信区间都非常相近。值得注意的是，两种参数化方式下的 δ 参数均表明，共享成分对武力使用数据集的贡献最大，而非警察拦截数据。与 Knorr-Held 和 Best（2001）一样，我们假设共享成分模型反映了影响警察使用武力和拦截事件的因素所造成的空间变异，例如警务活动增加或人口变化。因此，我们将两种参数化方式下的δ 解读为：由警务活动或人口等因素导致的共享空间模式

表5. 将共享成分模型应用于芝加哥警察使用武力数据和警察拦截数据所得的回归系数估计值及95%可信区间。

模型	点过程1 - 使用武力		点过程2 - 拦截		共享成分	
	β_1 (赫芬达尔指数)	β_2 (失业率)	β_0	β_1 (中位年龄)	δ	σ
估计值，Unif	-0.52	-1.4	4.44	-0.017	0.54	6.31
可信区间	(-0.64, -0.39)	(-1.7, -1.07)	(4.24, 4.67)	(-0.018, -0.016)	(0.53, 0.55)	(4.91, 7.87)
估计值，Norm	-0.51	-1.38	4.44	-0.017	1.09	3.14
可信区间	(-0.65, -0.37)	(-1.67, -1.06)	(4.15, 4.66)	(-0.019, -0.016)	(1.07, 1.11)	(2.46, 3.92)

对武力使用事件分布的影响大于对警察拦截事件分布的影响。两种参数化方式中，唯一明显不同的参数是 σ，它来自高斯过程的协方差函数；当 δ 在均匀先验下被限制在 0 到 1 之间时，该参数的估计值更高。

我们还希望将估计得到的空间共享成分与点过程特异性成分进行比较。图6展示了共享成分模型各个方面所推断出的空间分布。需要指出的是，在本例中，模型中点过程特异性部分的空间分布估计采用的是非齐次泊松过程，因此估计的强度仅随人口普查区而变化。图6中间和右侧的图展示了这一点。在未来的工作中，可以考虑将对数高斯考克斯过程纳入这一框架，不过空间变化共享成分与各点过程特有的高斯过程之间可能存在混杂问题。图6左侧展示了共享成分的后验均值估计，该估计已转换至分析中使用的积分点上。可以看出，两个点过程之间存在明显的共享空间模式。需要指出的是，我们的模型并未将人口用作空间协变量。因此，如前所述，这一共享空间成分可能主要由人口和/或这些区域警务活动增加所决定。今后，我们还可以考虑用人口密度对两个强度函数进行缩放，这也是点过程强度函数中常见的做法（参见 Liang 等人，2008年；Walder 等人，2020年）。接下来，我们将关注各点过程特有的空间成分，它们表明存在仅与警察使用武力或警方拦截相关、而非两者都相关的因素。我们分析了回归系数的后验均值估计，并发现，除共享成分之外，警察使用武力事件的分布在城市北部最为显著，城市南部也有一些强度较高的区域。另一方面，在考虑共享成分后，警方拦截强度在城市最南部和中部最高。这些信息有助于我们在考虑另一种共享空间模式后，分析哪些地方的武力使用和警方拦截更为频繁。未来的研究可帮助确定其他可纳入模型的空间变量和社会经济信息，从而揭示影响某一个点过程、但不同时影响两个点过程的因素。这有助于解释为何某些地区发生警方拦截的频率高于仅根据警务活动所预期的水平，也能回答警察使用武力方面的同类问题。

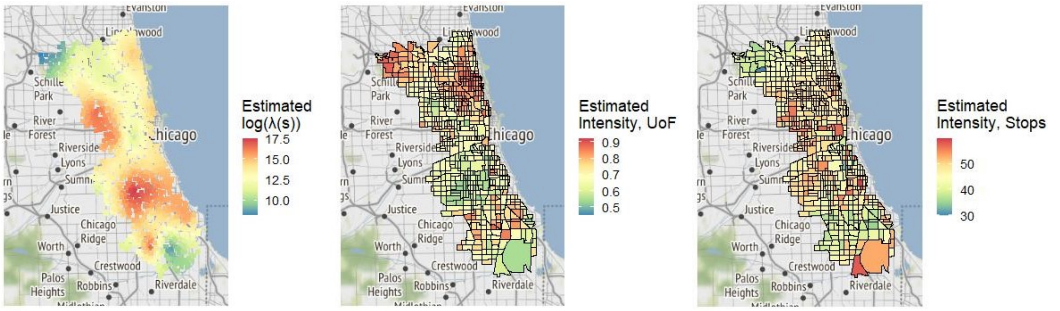

图 6. 基于后验均值估计的共享成分（左）与点过程特异性成分（使用武力—中，警方拦截—右）

6. 讨论

通过芝加哥警务数据和模拟示例，我们证明了本文提出的共享成分模型在关联两个点过程时能够提供新的见解。在实践中，通常难以证明这两个过程之间存在因果方向。我们的模型提供了一种比较两个点过程的方法，无须指定哪个过程是病例或对照，从而避免建模者必须设定因果方向。我们能够全面刻画两个点过程共有的空间模式，以及各自独有的空间模式。将该模型应用于芝加哥警务数据，尤其有助于辨析警察使用武力与警方拦截共有的空间模式，以及各点过程独有的模式。我们的方法采用高斯过程，其计算复杂度高于病例对照模型，但灵活性更强。我们还提出了降低计算负担的方法，例如按照 Xu 等人（ 2019 年）的建议，对其中一个点过程进行抽样。

共享成分模型使我们能够全面刻画两个点过程之间的关系。我们不再是用一个点过程的强度去缩放另一个点过程的强度，而是在确定了人口等两个点过程的驱动因素后，将空间/社区变量纳入考量，以确定这些点过程可能具有的独特效应。这样，我们便能够分析两个点过程的潜在驱动因素，而不只是其中一个点过程的驱动因素。该模型还可以直观呈现两个点过程的空间趋势，以及根据数据推断出的两个点过程的驱动因素。

在本研究中，我们重点比较两种点模式之间的空间分布：警察使用武力和警方拦截。未来，可以通过使共享成分的权重参数 δ 适应两个以上的点过程，将这一框架扩展到两个以上的点过程。本分析并不试图分析警员和公民特征与警察行为空间模式之间的关系，不过，可以将本方法与现有方法相结合来实现这一点，例如采用 Kelling和 Haran（2022）提出的方法。我们还注意到，共享成分也可以采用高斯过程以外的形式，例如使用聚类模型（Knorr-Held 和 Best,

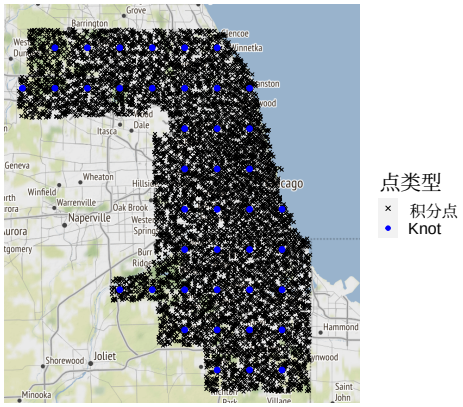

图 7. 积分点、预测过程点

2001年）。

7. 致谢

作者谨向 Peter Diggle 教授和 Ephraim Hanks 教授致谢，感谢他们进行了富有助益的讨论，大大改进了本文。本项目由美国司法部司法项目办公室国家司法研究所颁发的奖项编号 2020-R2-CX-0033 提供资助。本出版物中表达的观点、调查结果、结论或建议均为作者所有，不一定反映司法部的立场。

8. 附录

8.1. 积分点

与其他城市地区不同，芝加哥有许多面积很小的人口普查区。为了准确计算芝加哥的强度函数积分，我们必须确保每个人口普查区至少有一个点，并且每个人口普查区的积分点数量必须与该区的面积成正比，从而使计算该区域的强度函数积分时，各人口普查区都按其面积加权（参见 Liang 等人，2008年）。我们在芝加哥总共使用了7,304个积分点。由于小面积人口普查区数量众多，采用其他参数化方式（例如令点的数量与面积的平方根成正比）会导致芝加哥的积分估计值不准确。图7展示了空间变化共享成分的预测过程中使用的积分点和结点。

这种方法使每个人口普查区的积分点数量与其面积严格成正比，这一点对芝加哥尤为重要，因为该市最小人口普查区之间的面积差异非常悬殊。举例来说，我们比较得克萨斯州达拉斯和伊利诺伊州芝加哥，这两座城市的人口普查区面积大小不一。达拉斯城市辖区内最大的人口普查区面积是最小人口普查区的115倍。芝加哥最大的人口普查区面积则是最小人口普查区的3,573倍。

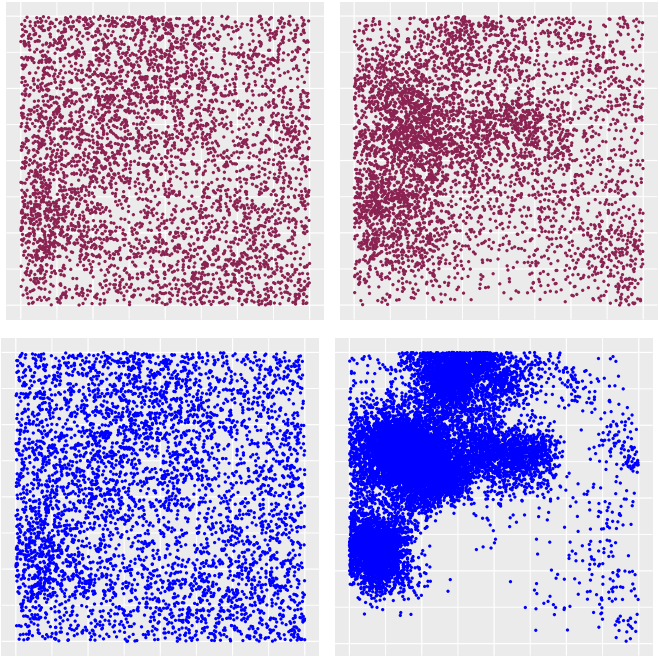

图 8. 共享成分采用均匀分布（上）和正态分布（下）的比较，其他所有参数均相同

我们还通过人口普查区面积的第一和第三四分位数，分析更具代表性的数值，而非极端值。达拉斯人口普查区面积的第三四分位数是第一四分位数的2.8倍，而芝加哥人口普查区面积的第三四分位数是第一四分位数的5.3倍。

8.2. 共享成分加权分布

我们使用相同的参数模拟两个点过程，如表 3 所示，但采用不同的加权方案和分布。这些方案和分布基于图 3 所示的空间协变量。图 8 中的洋红色部分展示了对两个点过程分别赋予权重 δ 和 $1-\delta$ 后得到的点过程，其中 δ 参数采用均匀先验（0,1）。图 8 下方的蓝色部分展示了采用权重 δ 和 $1/\delta$ 后得到的两个点过程，其中 δ 参数采用正态先验。在两次模拟中，δ 均设为 0.3。

8.3. 芝加哥拦截样本

对于共享成分模型，我们必须将共享成分（涉及高斯过程）转换为与拦停数据和武力使用数据维度一致的形式。警察武力使用数据包含7,539个具有完整空间信息的事件，在计算上是可处理的。另一方面，拦停数据包含1,453,832个具有完整空间信息的事件。这在大规模矩阵计算方面构成了计算挑战，

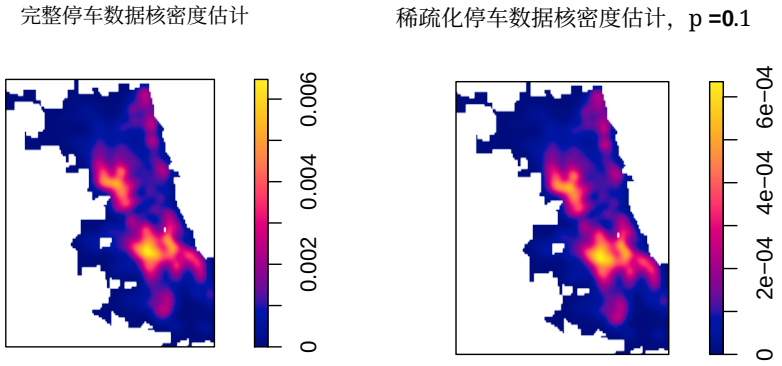

图 9. 完整停车数据的核密度估计（左）与稀疏化停车数据（右）的比较

。我们注意到，如果对所有点进行均匀稀疏化，完整停车数据与稀疏化数据之间点的空间分布不会发生太大变化。在图9中，我们展示了完整停车数据和稀疏化停车数据的核密度估计（KDE），其中每个点被保留的概率为10%。空间稀疏化产生了一个包含145,783个点的数据集。我们注意到，除了图右侧所示的尺度不同之外，这两个点过程的KDE几乎完全相同。我们采用了与 Xu 等人（2019）类似的方法，对其中一个点过程进行抽样。在未来的研究中，当我们关注拦停事件的不同特征时，可能需要考虑完整点过程，或采用基于其他变量进行抽样的更复杂抽样机制。

8.4. 病例对照模型与共享成分模型之间的关系

本节将详细阐述病例对照模型（参见 Diggle 等人，2007）与我们用于点过程的共享成分模型之间的关系。病例对照模型根据对照过程的强度 $\lambda_0^{CC}(s)$，定义病例过程的强度 $\lambda_1^{CC}(s)$，如下所示：

$$\lambda_1^{CC}(s) = \lambda_0^{CC}(s)\exp(\boldsymbol{z}(s)'\boldsymbol{\beta}). \tag{3}$$

作为提醒，下面展示的共享成分模型分别描述两个点过程的强度（记为 $\lambda_1(s)$ 和 $\lambda_2(s)$），但依赖于一个对两个过程的强度都有贡献的过程，我们将其记为 $\lambda(s)$。

$$\begin{aligned}\lambda_1(s) &= \lambda(s)^{\delta}\exp\left(\boldsymbol{z_1}(s)'\boldsymbol{\beta_1}\right)\\ \lambda_2(s) &= \lambda(s)^{1-\delta}\exp\left(\boldsymbol{z_2}(s)'\boldsymbol{\beta_2}\right)\end{aligned} \tag{4}$$

这种由共享过程同时作用于两种强度的结构，使我们能够将共享成分模型写成病例对照模型的一种形式。

我们的目标是确定共享成分模型中“缩放过程”的参数化形式。对于典型的病例对照模型，该过程就是对照过程 $\lambda_0^{CC}(s)$

。在方程5中，我们重写共享成分模型中第一个点过程的强度，以展示病例对照模型结构下的缩放过程。具体而言，我们先从第二个点过程的强度中解出同时影响两个点过程的共享成分 $\lambda(s)$，再将其代入第一个点过程的强度中。我们将空间协变量记为 $\boldsymbol{z_1}(\boldsymbol{s})$ ，其对应的回归系数记为 $\boldsymbol{\beta_1}$ ；第一个点过程的空间协变量记为 $\boldsymbol{z_2}(\boldsymbol{s})$ ，回归系数记为 $\boldsymbol{\beta_2}$ 。与病例对照模型中的情况不同，此处的缩放过程并非单纯估计第二个空间强度函数（$\lambda_0^{CC}(\boldsymbol{s})$），而是对于重新参数化的共享成分模型 $\left(\left[\frac{\lambda_2(s)}{\boldsymbol{z_2}(\boldsymbol{s})\ \exp((\boldsymbol{s})'\boldsymbol{\beta_2})}\right]^{\delta/(1-\delta)}\right)$ ，它取决于第二个空间强度、第二个空间强度的协变量和回归系数，以及参数 δ ；该参数定义了共享成分对两个点过程的贡献。

$$\lambda_1(s) = \left[\frac{\lambda_2(s)}{\exp(\boldsymbol{z_2}(\boldsymbol{s})'\boldsymbol{\beta_2})}\right]^{\delta/(1-\delta)} \exp(\boldsymbol{z_1}(\boldsymbol{s})'\boldsymbol{\beta_1}). \tag{5}$$

参考文献

Antonovics，K. 和 Knight，B. G.（2009）重新审视种族画像：来自波士顿警察局的证据。《经济学与统计学评论》，**91**，163–177。

Ba, B. A., Knox, D., Mummolo, J. 和 Rivera, R.（2021a）警员种族与性别在芝加哥警民互动中的作用。《科学》，**371**，696–702。—（2021b）警员种族与性别在芝加哥警民互动中的作用[源代码]。网址：https://doi.org/10.24433/CO.4519202.v1。

Banerjee, S., Gelfand, A. E., Finley, A. O. 和 Sang, H.（2008）用于大型空间数据集的高斯预测过程模型。《皇家统计学会杂志：*B*辑（统计方法论）》，**70**，825–848。

Berman, M.（1986）检验点过程与另一随机过程之间的空间关联。《皇家统计学会杂志：*C*辑（应用统计学）》，**35**，54–62。Chetwynd, A. G., Diggle, P. J., Marshall, A. 和 Parslow, R.（2001）对个体匹配病例对照研究中的空间聚类进行调查。《生物统计学》，**2**，277–293。de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D. 和 Bodik, R.（2017）使用模型编程：为一般模型结构编写统计算法并应用于 NIMBLE。《计算与图形统计学杂志》，**26**，403–413。Diggle, P. J.（1990）采用点过程建模方法研究预先指定点附近罕见现象发病率升高。《皇家统计学会杂志：*A*辑（社会统计学）》，**153**，349–362。

Diggle, P. J., G´omez-Rubio, V., Brown, P. E., Chetwynd, A. G. 和 Gooding, S.（2007）利用病例对照数据对非齐次空间点过程进行二阶分析。《生物统计学》，**63**，550–557。Diggle, P. J., Morris, S. E. 和 Wakefield, J. C.（2000）利用匹配病例对照数据进行点源建模。《生物统计学》，**1**，89–105。Diggle, P. J. 和 Rowlingson, B. S.（1994）一种针对升高风险的点过程建模条件方法。《皇家统计学会杂志：*A*辑（社会统计学）》，**157**，433–440。Flegal, J. M., Hughes, J., Vats, D., Dai, N., Gupta, K. 和 Maji, U.（2021）*mcmcse*：用于马尔可夫链蒙特卡罗的蒙特卡罗标准误。美国加利福尼亚州里弗赛德和印度坎普尔。R软件包版本 1.5-0。Fuentes-Santos, I., Gonz´alez-Manteiga, W. 和 Mateu, J.（2017）一种比较空间点过程一阶结构的非参数检验。《空间统计学》，**22**，240–260。Gong, L. 和 Flegal, J. M.（2016）一种实用的高维马尔可夫链蒙特卡罗序贯停止规则。《计算与图形统计学杂志》，**25**，684–700。Guan, Y., Waagepetersen, R. 和 Beale, C. M.（2008）利用比例强度函数对非齐次空间点过程进行二阶分析。《美国统计学会会刊》，**103**，769–777。—（2010）利用病例对照数据对非齐次空间点过程的二阶结构进行非参数与参数估计。技术报告。Haran, M. 和 Hughes, J.（2020）*batchmeans*：蒙特卡罗标准误的一致批均值估计。美国马里兰州弗雷德里克。R软件包版本 1.0-4。Hessellund, K. B., Xu, G., Guan, Y. 和 Waagepetersen, R.（2021）多变量点模式数据的半参数多项逻辑回归。《美国统计学会会刊》，1–16。Illian, J., Penttinen, A., Stoyan, H. 和 Stoyan, D.（2008）空间点模式的统计分析与建模，第70卷。约翰·威利父子公司。Kelling, C. 和 Haran, M.（2022）一种包含空间与非空间协变量的两阶段 Cox过程模型。《空间统计学》，**51**，100685。Knorr-Held, L. 和 Best, N. G.（2001）一种用于检测两种疾病联合与选择性聚类的共享成分模型。《皇家统计学会杂志：*A*辑（社会统计学）》，**164**，73–85。Knox, D.（2021）揭示种族偏见。《科学》，**374**，701–702。Knox, D., Lowe, W. 和 Mummolo, J.（2020a）行政记录掩盖了种族偏见警务。《美国政治科学评论》，**114**，619–637。

Knox, D., Mummolo, J. 等（2020b）《构建警务中种族偏见研究的一般因果框架》。《政治制度与政治经济学杂志》，**1**，341–378。Lawton, B. A.（2007）《非致命武力的层级：对个体、情境及背景因素的考察》。《犯罪与违法行为研究杂志》，**44**，163–184。Lee, H., Jang, H., Yun, I., Lim, H. 和 Tushaus, D. W.（2010）《利用警察培训和邻里情境因素考察警察使用武力：多层次分析》。《警务：警察策略与管理国际期刊》，**33**，681–702。Lee, H., Vaughn, M. S. 和 Lim, H.（2014）《邻里犯罪水平对警察使用武力的影响：微观与中观层次的考察》。《刑事司法杂志》，**42**，491–499。Liang, S., Carlin, B. P. 和 Gelfand, A. E.（2008）《利用空间与非空间协变量信息分析明尼苏达州结肠癌和直肠癌点模式》。《应用统计年鉴》，**3**，943。Mohler, G.（2014）《用于预测芝加哥凶杀和枪支犯罪的标记点过程热点地图》。《国际预测学杂志》，**30**，491–497。Møller, J. 和 Waagepetersen, R. P.（2003）《空间点过程的统计推断与模拟》。CRC出版社。Phillips, S. W. 和 Sobol, J. J.（2011）《关于使用不必要武力的警察态度：一项生态学考察》。《警察与犯罪心理学杂志》，**26**，47–57。Ratcliffe, J.（2002）《入室盗窃减少与转移效应神话》。《犯罪与刑事司法趋势与问题》，1–6。Reisig, M. D. 和 Parks, R. B.（2003）《邻里情境、警察行为与对警察的满意度》。《司法研究与政策》，**5**，37–65。Simckes, M., Willits, D., McFarland, M., McFarland, C., Rowhani-Rahbar, A. 和 Hajat, A.（2021）《警务对人口健康的不利影响：一个概念模型》。《社会科学与医学》，**281**，114103。Slovak, J.（1988）《城市警务风格：美国部分城市的组织、环境与警务风格》。纽约大学出版社。Terrill, W. 和 Reisig, M. D.（2003）《邻里情境与警察使用武力》。《犯罪与违法行为研究杂志》，**40**，291–321。Walder, A., Hanks, E. M. 和 Slavkovi´c, A.（2020）《空间点过程数据的隐私》。Weisburst, E. K.（2019）《警察使用武力是逮捕的延伸：考察平民与警员种族之间的差异》。载于《美国经济学会论文与会议录》，第109卷，152–56。

Wilson, J. Q. 和 Kelling, G. L.（1982）《破窗理论》。《大西洋月刊》，**249**，29–3 8.

Xu, G., Waagepetersen, R. 和 Guan, Y.（2019年）病例-对照点格局数据的随机准似然。《美国统计学会会刊》，**114**，631–644。