arXiv:2304.05933v1 [stat.AP] 12 Apr 2023

一种用于对区间删失事件时间下时空犯罪风险进行建模的贝叶斯犯罪时序逻辑回归

Álvaro Briz-Redón[1*]

[1]西班牙瓦伦西亚大学统计与运筹学系*alvaro.briz@uv.es

摘要

从统计学角度来看，犯罪数据具有某些特殊性，因此人们对其分析的兴趣日益浓厚。具体而言，一些财产犯罪经常呈现的一个特征是，其发生的确切时间存在不确定性，通常只能确定一个界定事件发生时间的时间窗口。处理这类区间删失观测值的方法有多种，其中大多数基于对事件时间进行插补。另一种方法是开展模糊时间分析，其依据是对界定事件时间不确定性的区间内每个时间单位赋予相同的权重。然而，这种方法也有其局限性。本文提出一种基于逻辑回归的时空模型，可依照时态概率分析的思路，对存在时间不确定性的犯罪数据进行分析。该模型从贝叶斯视角构建，因此能够纳入观测值的时间不确定性。我们将该模型应用于西班牙瓦伦西亚记录的住宅入室盗窃数据集，并将模型结果与完整案例模型的结果进行比较；后者会舍弃时间不确定事件。

关键词 — 贝叶斯统计；删失数据；犯罪数据；数据插补；时空模型；时间不确定性

1 引言

过去十年间，先进统计技术在犯罪分析中的应用显著增长。特别是，分析犯罪在时空中的分布，以及研究时空交互作用的存在，尤其受到关注。因此，人们开发并改进了不同的建模方法，以寻找与较高犯罪风险相关的环境因素，并对短期和中期犯罪进行预测。其中包括不同版本的自激模型（Mohler et al., 2011; Zhuang and Mateu, 2019）、具有非线性结构的空间模型（Briz-Redón 等人，2022），以及贝叶斯单变量/多变量时空模型（Chung and Kim, 2019; Law et al., 2014; Li et al., 2014; Quick et al., 2018）。

犯罪数据通常具有一些特殊特征，可能使分析变得复杂，进而推动新方法的发展。例如，众所周知，犯罪数量通常被低估，甚至存在偏差（Buil-Gil 等人，2021a,b）。此外，在空间、时间或时空犯罪数据集的背景下，空间精度和时间精度往往都令人关注。空间精度方面的问题通常是指无法确定事件发生所在的空间单元（例如，研究区域的某个行政区划）。特别是，如果事件的空间位置以文本信息的形式提供（即位置的可读地址），则常会出现地理编码错误，或有事件未能完成地理编码。此前已有研究探讨了这类问题，旨在确定最低可接受的地理编码率（Andresen 等人，2020；Briz-Redón 等人，2020；Ratcliffe，2004）。至于时间精度方面的问题，对于某些犯罪事件，我们也常常无法以所需的精度（分钟、小时、日期等）观测其时间位置。事实上，如果涉及财产盗窃等某些类型的犯罪，大多数事件都会出现这种情况（Ashby 和 Bowers，2013）。在这种情况下，我们掌握的是事件时间位置的下限和上限，也就是说，每个事件对应一个时间区间或时间窗口。这类时间观测通常称为区间删失。

区间删失事件时间的存在，是定量犯罪学领域一个众所周知的问题。尽管这一问题可能未得到足够重视，但针对犯罪记录中的区间删失事件时间，已有多种处理策略。最简单的方法是选择合适的时间单位，以消除时间不确定性。例如，如果现有的不确定性

处于日级别，我们就可以按周级别进行处理。尽管在某些情况下，这种方法能让我们避免大多数（或全部）观测值中的时间不确定性，但根据观测值所面临的不确定性来设定分析的时间分辨率并不可取。

处理不确定性的另一种方法是对事件时间进行插补，这可能是最常见的策略，并且有多种实现方式。一种做法是将这些事件分配到可用不确定性时间窗口的中点对应的时间单位（小时、日期等）。考虑时间窗口的起始或结束时间位置也是一种选择，但采用中点的方法所产生的误差（和偏差）较小（Ashby 和 Bowers，2013）。然而，正如 Ashby 和 Bowers（2013）所示，这些确定性方法不如一种名为时段概率法的非确定性方法。

时段概率法主要由犯罪学家 J. Ratcliffe 提出（Ratcliffe and McCullag h, 1998；Ratcliffe，2000，2002），其基础是为划定犯罪事件不确定性的区间内所包含的每个时间单位赋予相同的权重。因此，当采用时段概率方法时，时间不确定事件不会获得单一的插补值，而是会在事件所在的每个时间单位上获得一个概率分数。具体而言，所有时间单位都会获得相同的分数或权重，使其总和为1。正因如此，时段概率法并不完全对应于一种插补方法。然而，遵循时段概率程序，例如，可以通过将确定的事件数量与时间段内每个日期对应的不确定事件的分数相加，推导出一组犯罪的时间分布。尽管时段概率方法能够对包含区间删失时间观测值的犯罪数据集开展一些探索性分析，但根据分析目的的不同，这可能并不足够。

最后，也可以采用基于模型的方法来处理区间删失观测。例如，冯·米塞斯分布或狄利克雷过程通常用于分析圆形数据，最近也有人提出将其用于时态概率分析数据（Mulder，2019）。此外，处理观测值时间不确定性的一种方法是将所有观测值纳入同一个模型，并假设这些观测值的精确时间位置服从某一概率分布。一般而言，这会导向在贝叶斯框架下处理缺失数据的模型。事实上，Reich and Porter（2015）采用贝叶斯建模框架对犯罪事件进行聚类，该框架还具有其他特性，其中包括能够处理区间删失事件时间。在这种情况下，作者将区间删失事件时间视为潜变量，其完整条件分布

服从截断正态分布，均值等于簇均值（即属于同一簇的事件的平均时间位置）。

本文采用后一种方法，分析西班牙瓦伦西亚记录的入室盗窃数据集。与住宅入室盗窃的常见情况一样，该数据集包含很大比例的时间不确定事件。具体而言，本文提出一种时空逻辑回归模型，用于刻画时空中的入室盗窃风险。该模型在贝叶斯框架下进行估计，从而能够纳入时间不确定事件。在将这种不确定性引入模型时，本文借鉴了时段概率方法。因此，本文有两个目标。第一，描述一种建模框架，在考虑时间不确定事件的同时，估计时空中的入室盗窃风险。第二，强调在分析中纳入此类犯罪事件的适用性，以获得更可靠的参数估计值，并尝试还原犯罪潜在的时空分布。

本文结构如下。第 2 节介绍分析所用的数据，重点说明事件中存在时间不确定性。第 3 节介绍在存在区间删失时间观测值的情况下，用于估计时空入室盗窃风险的建模框架。随后，第 4 节展示分析得出的主要结果。最后，第 5 节展开讨论并作出总结。

2 数据

2.1 研究设置

本案例研究在西班牙人口第三多的城市瓦伦西亚开展，该市人口约为80万。具体而言，分析范围为城市核心城区，排除了部分外围城区，这些区域仅占总人口的5%。此外，为研究入室盗窃在全市范围内的空间分布，分析还涵盖了瓦伦西亚的各个行政区。研究划定的区域内共有70个行政区。

2.2 入室盗窃数据

分析使用了西班牙国家警察提供的数据集，其中包含2016年1月1日至2017年12月31日期间在瓦伦西亚市记录的2626起已地理编码的入室盗窃信息。Briz-Redón 等人（2020）已使用该数据集研究近重复

现象。在该数据集中，每起事件都提供了地理坐标，因此可以按照任何所需的空间尺度开展分析。相比之下，部分入室盗窃事件的时间位置存在一定程度的不确定性。具体而言，只有60.9%的事件有确切的入室盗窃发生日期。关于数据时间不确定性的更多说明，请参阅下文小节。

2.3 事件时间不确定性

本研究所用数据集的一个重要特点是存在区间删失事件时间。具体而言，对于每起入室盗窃，都有一个起始日期 和一个截止日期 变量，用于根据有关该起入室盗窃的现有信息界定事件的时间位置（起始日期表示屋主能够确定住宅尚未遭入室盗窃的最后一个日历日期，而“截止日期”则是屋主、警方或任何公民确认入室盗窃已经发生的日期）。尽管原始日期变量采用 YYYY-MM-DD 格式，但为便于使用，已将其转换为数值，并将 2016 年 1 月 1 日（研究期间的起始日）设为 1。除非另有说明，本文其余部分均将日期视为数值。

因此，虽然由于能够获得研究期间每处遭入室盗窃住宅的坐标（当然，仅限已向警方报案的案件），数据不存在空间不确定性，但事件时间不确定性不容忽视。在 Briz-Redón 等人（2020）的一项先前研究中，研究人员通过中点日期插补法解决了部分观测值的时间不确定性。该研究背景下开展的一项初步分析聚焦于近重复现象，结果表明，插补方法（中点日期法或概率日期）对结果影响不大。因此，鉴于计算上的便利，研究人员优先采用中点日期法，而非时态概率分析。

在本文中，旨在明确处理区间删失事件，而不直接插补缺失的事件日期，正如后续章节将展示的那样。

3 方法论

3.1 病例对照研究设计

为了遵循逻辑斯蒂建模框架（将在下一小节中介绍），需要一个表示是否发生入室盗窃事件的二元响应变量。因此，

可获得的入室盗窃事件的时空位置（具有事件时间不确定性）记为$\{\boldsymbol{x}_i, t_i^{from}, t_i^{event}, t_i^{to}\}_{i=1}^{2626}$，被视为病例/事件，其中 $\boldsymbol{x}_i$ 是与事件 i对应的地理坐标，而 t_i^{from} 、 t_i^{event}和 t_i^{to}分别表示起始日期、事件实际发生的日期以及截止日期。对于没有时间不确定性的事件，有$t^{from} = t^{event} = t^{to}$。对于时间不确定事件，我们有 $t^{from} < t^{to}$ 以及t^{event}的缺失值。换言之，我们只知道 $t^{event} \in [t^{from}, t^{to}]$。

为生成对照组，我们考虑了研究区域内所有住宅的位置所构成的点模式。具体而言，该模式包含研究区域内的 28682 个位置，共有 382539 个住宅单元（对应于 2016 年在瓦伦西亚登记的住宅）。通过有放回抽样，共抽取了 13130 个对照位置，并将抽中概率设定为与该位置的住宅数量成正比。因此，对照组数量设为病例数的五倍，即对照组与病例的比例为 5:1。文献表明，进行病例-对照研究设计时，4:1 的病例-对照比例通常已足够（Gail 等，1976；Hong 和 Park，2012）。该比例的选择会影响逻辑回归模型得出的概率估计，因此，在进行依赖于截断概率的二元（事件/无事件）预测时，应将其纳入考虑。每个对照位置都被赋予了一个 2016 年 1 月 1 日至 2017 年 12 月 31 日之间的随机日期。对照数据未假设存在时间不确定性，因此 $t^{from} = t^{event} = t^{to}$ 适用于所有这些时空位置。图 1 展示了研究区域内用于分析的病例和对照组的空间位置。

3.2 逻辑回归模型

若 $\{\boldsymbol{x}_i, t_i^{from}, t_i^{event}, t_i^{to}\}_{i=1}^{15756}$表示分析所考虑的完整时空位置集合，则令 y_i 为一个二元变量，用于指示每个位置代表病例（入室盗窃）还是对照。按照惯例，若 i 是一个事件，则设 $y_i = 1$ ，否则设为 0。为了对研究窗口内每个时空位置发生入室盗窃的风险进行建模，逻辑回归建模框架是自然的选择。在逻辑回归模型下，时空位置 i 发生入室盗窃事件可通过伯努利随机变量 $Y_i \sim Ber(\pi_i)$进行描述，其中 π_i 表示发生入室盗窃的概率。本文尝试通过若干固定效应和随机效应（空间和时间）来解释这一参数，从而得到以下表达式：

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_{DoW(i)} + \delta_{w(i)} + \varepsilon_{w(i)} + u_{b(i)} + v_{b(i)} \tag{1}$$

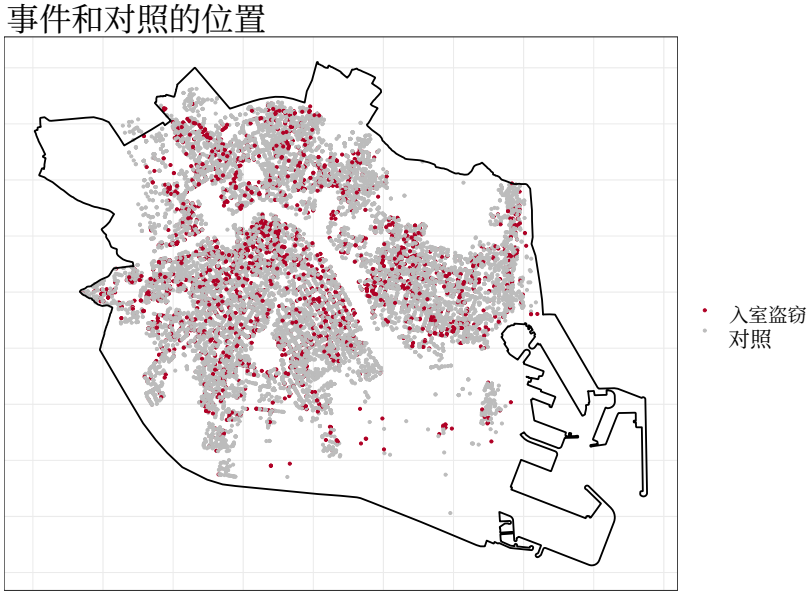


图 1：分析中纳入的事件（入室盗窃）和对照在研究区域内的位置

其中各项定义如下。首先，α 是逻辑回归模型的全局参数。其次，$\beta_{DoW(i)}$ 表示与事件 i ($DoW(i)$) 对应的星期几对 logit(π_i) 的影响。具体而言，以星期一作为该变量的参考水平，因此需要估计六个 β_{DoW} 参数，分别对应其余六个星期几。对 α 和各个 β_{DoW} 赋予宽松的高斯先验 $N(0, 1000)$。模型中涉及的其余参数表示空间和时间随机效应，接下来的内容将对此进行说明。

时间结构化的周效应 δ_w ($w = 1,..., 104$) 通过二阶随机游走 $\delta_w|\delta_{w-1}, \delta_{w-2} \sim N(2\delta_{w-1} + \delta_{w-2}, \sigma_\delta^2)$ 指定，其中 σ_δ^2 为方差分量；对于时间非结构化的周效应 $\varepsilon_w \sim N(0, \sigma_\varepsilon^2)$（$w = 1, ..., 104$），则采用独立同分布的高斯先验。方差分量 σ_δ^2 和 σ_ε^2 服从伽马分布先验 $Ga(1, 0.5)$。

关于空间随机效应，研究采用了 Besag–York–Mollié（BYM）模型（Besag 等，1991）。在该模型下，行政区 b 上空间结构化效应的条件分布为 u_b（$b = 1$，...，70），

$$u_b|u_{\tilde{b}\neq b} \sim N\left(\sum_{\tilde{b}\neq b=1}^{n} w_{b\tilde{b}}u_{\tilde{b}}, \frac{\sigma_u^2}{N_b}\right)$$

˜其中，N_b 为区域 b 的邻居数量，$w_{b\tilde{b}}$ 为行标准化邻域矩阵的（b，b）元素（如果行政区 b 和 b 共享地理边界，则 $w_{b\tilde{b}} = 1/N_b$，否则为 0），σ_u^2 表示该随机效应的方差。各区域上的空间非结构化效应记为 v_b（$b = 1$，...，70），服从高斯分布 $v_b \sim N$（0，σ_v^2），其中 σ_v^2 为该效应的方差。对于 σ_u^2 和 σ_v^2，均假设其先验分布为伽马分布 Ga（1，0.01）。

此时需要指出，选择瓦伦西亚行政区构成的空间结构来衡量入室盗窃风险的空间变化，在一定程度上是任意的。在本例中，选择行政单位较为方便，因为这能让从业者以简单的方式设计监控策略。此外，从计算角度来看，这种选择也有优势，因为空间单位的数量（70 个）并不算多。无论如何，所提出的建模框架可以调整，以适用于研究区域的其他类型划分，无论这些划分是规则还是不规则、粗略还是精细。

3.3 处理时间不确定性

由 (1) 表示的逻辑回归模型隐含地假设，对于每个事件/对照地点，星期几（DoW）和一年中的周都确切已知，i。此时，$DoW(i)$ 和$w(i)$是两个已知值，相应的固定效应和随机效应可以估计。如果事件 i 的发生确切日期未知，则 $DoW(i)$ 未知；除非起始日期 和截止日期 属于同一个日历周，否则 $w(i)$ 也未知。在这种情况下，一种做法是将所有日期未知的事件从分析中剔除。这样可以避免处理缺失数据，对应于完整病例分析（仅考虑不含缺失值的数据记录）。从分析中删除所有这些观测值会降低精度和统计功效，这是我们不希望看到的。

因此，为了将所有这些时间不确定事件纳入建模框架，可以像贝叶斯统计中通常处理缺失数据那样，将每个缺失日期视为一个随机变量。具体而言，按照时段概率方法，为每个日期指定一个均匀先验分布，并考虑起始日期 和截止日期 变量所提供的信息，即 $t_i^{event} \sim U(t_i^{from} - 0.5, t_i^{to} + 0.5)$。随后，在马尔可夫链蒙特卡洛（MCMC）过程的每次迭代中，根据该分布抽取一个数值日期，并将其四舍五入（至最接近的整数），以便计算 $DoW(i)$ 和 $w(i)$，进而在估计模型的固定效应和随机效应时纳入所有可用数据。对于已知日期，由于$t_{to} = t_{from}$，t_i^{event}的所有抽样值都与精确的（已知）日期一致。此外，请注意

分别从 t_i^{from} 中减去 0.5、向 t_i^{to}中加上 0.5 是必要的，以避免降低先验分布中与该事件对应区间两端的权重。

因此，通过将 t_i^{event}视为随机变量，分析时不会丢弃任何数据。在本文余下部分中，该模型称为完整模型。此外，我们还可以计算后验分布 $p(t_i^{event} = t|D)$，其中 D 从现在起表示用于拟合模型的数据集；对于$t \in \{t_i^{from}, t_i^{from}+1, ..., t_i^{to}-1, t_i^{to}\}$，则可据此估计警方划定的发生时间区间内各个日期发生时间不确定事件的概率。

前文所述的逻辑回归模型已在用于贝叶斯推断的 NIMBLE 系统中实现（de Valpine et al., 2017），并基于 MCMC程序。

3.4 模型批评

在贝叶斯分析中，模型评估与比较通常通过一些众所周知的拟合优度度量来进行，例如 Spiegelhalter 等人（2002）提出的偏差信息准则（DIC），或 Watanabe 和 Opper（2010）提出的瓦塔纳贝–赤池信息准则（WAIC）。然而，这些指标仅适用于比较具有相同似然函数的模型，因此无法使用这些指标比较上述完整案例模型和完整模型（由于每个模型拟合的数据集不同，它们的似然函数也不同）。

因此，需要采用不同的策略。一种可能的方法是通过分析 π_i 的点估计值（记为 $\hat{}\pi_i$）的分布来进行模型批评，其中每个点估计值均计算为后验分布 $p(\pi_i|D)$ 的均值。具体而言，我们首先比较不同模型和位置类型（病例与对照组）下 $\hat{}\pi_i$ 的分布。这样可以帮助我们判断每个模型是否能够区分病例与对照组，以及是否存在显著差异。其次，针对完整模型，我们考察对应于确定的事件和时间不确定事件的 $\hat{}\pi_i$ 的分布是否存在差异。

此外，研究模型作为分类工具的质量也很有意义。为了评估拟合优度，我们可以考察模型的样本内预测能力。本研究选用 F1分数和马修斯相关系数（MCC）（Matthews，1975）来评估模型的样本内预测质量。分析中未采用准确率等其他常见指标，因为它们在不平衡数据集上的表现并不可靠。事实上，这些指标往往会过于乐观地估计分类器对多数类的分类能力（Chicco and Jurman,

2020）。F1分数定义如下：

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

近期研究表明，MCC 的表现优于其他更常见的指标（Chicco and Jurman, 2020）。MCC 的取值范围为 -1（最差值）至 1（最佳值），其定义如下：

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}},$$

其中，TP 是真阳性的数量（阳性预测且实际事件发生），FP 是假阳性的数量（阳性预测但未发生事件），FN 是假阴性的数量（阴性预测但实际事件发生），TN 是真阴性的数量（阴性预测且未发生事件）。为了将一个观测值标记为阳性或阴性，需要使用截断概率 $c \in]0, 1[$, 作为阈值。然后，如果 $\hat{}\pi_i > c$，则该观测值被分类为阳性（入室盗窃）；而如果 $\hat{}\pi_i \leq c$，则该观测值被分类为阴性（未发生入室盗窃）。如果我们使用的是平衡数据集（病例数与对照组数相同），$c = 0.5$ 将是自然的选择。然而，所分析的数据集偏向对照组，呈不平衡状态（按构造，对照组与病例的比例为 5:1）。因此，较低的 c 值更为合适；否则，大多数观测值都会被分类为阴性，分类效果也会远非最优。基于此，我们分析了 MCC 随 c 变化的情况。

MCC 的公式提供了模型预测质量的点估计。按照 Gilardi et al. (2022) 提出的模型平衡准确率分析方法，本例估计了 MCC 的分布。具体而言，后验分布 $p(\pi_i|D)$ 中的抽样值可用于模拟 MCC 值，从而得出 MCC 的分布，进而更全面地比较所拟合的两个模型的预测能力。

3.5 软件

分析使用了R编程语言（R Core Team，2021）。具体而言，使用了R软件包ggplot2（Wickham，2016）、lubridate（Grolemund和Wickham，2011）、nimble（de Valpine等，2017）、rgdal（Bivand等，2019）、rgeos（Bivand和Rundel，2020）、spatstat（Baddeley等，2015）和spdep（Bivand等，2008）。

4 结果

4.1 一项模拟研究

在分析瓦伦西亚住宅入室盗窃数据集之前，我们先开展模拟研究，以检验在对数据作出某些假设的情况下，所提模型是否适用。具体而言，我们选取了真实数据分析中也会使用的对照观测，并加入一个包含3000个元素的样本，使其充当病例。该样本通过对照组重抽样获得，从而使感兴趣事件在星期几上的强度保持恒定，并使其时空强度保持恒定（以各行政区的人口规模为条件）。我们通过两种不同机制，对一定比例的这些病例引入时间不确定性，以反映区间删失观测的存在可能取决于时间因素。具体而言，我们模拟了三个数据集，其中区间删失观测的概率取决于星期几（情景1、2和3）；另模拟了三个数据集，其中该概率取决于年度周（情景4、5和6）。模拟的目的是验证完整模型能否处理这种不确定性，并恢复真实的基线风险；如上所述，该风险在空间和时间上均为恒定。为了模拟后续将要分析的真实入室盗窃数据集的基本特征，我们通过从速率为 $\lambda = 0.2$ 的指数变量中抽样，并将模拟值加1，来获得这些事件各自的时间不确定性大小（以天为单位），从而确保时间不确定性至少为1天，否则便不存在不确定性。接着，若 t_i 是模拟事件 i的实际时间位置，我们计算 $t_i^{from} = t_i - Round（u + 1）$ 和 $t_i^{to} = t_i + Round（u + 1）$，其中 $u \sim Exp（\lambda = 0.2）$。下文将详细说明如何为每种情景选取区间删失观测。

设 p_{DoW} 为某个事件实际发生在一周中的 DoW 时，呈现某种时间不确定性的概率（p_1 对应星期一的此类概率， p_2 对应星期二的此类概率，依此类推）。我们假设 p_{DoW} 有三种情景。在第一种情景中，我们假设一周内的概率相等，即 $p_{DoW} = 0.4$、 $DoW = 1$，.....，7。在第二和第三种情景中，我们假设，如果某个事件实际发生在星期五至星期日，那么它呈现时间不确定性的概率更高。具体而言，在情景 2 中，我们考虑 $\{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$ =$\{0.3, 0.3, 0.3, 0.3, 0.5, 0.5, 0.5\}$；在情景 3 中，我们考虑 $\{p_1, p_2, p_3, p_4, p_5, p_6, p_7\} = \{0.2, 0.2, 0.2, 0.2, 0.6, 0.6, 0.6\}$。因此，在这方面，情景 3 比情景 2 更为极端。是的。

milarly, let p_w be the专业版 ba性 y that 一个事件t that actually 发生在 wee k w of th以及每iod

2016—2017（$w = 1, ..., 104$）呈现出时间不确定性。我们还对 p_w的情况作出三种情景假设。在情景4中，我们假设 $p_w = 0.4$, $w = 1, ..., 104$。在情景5和情景6中，我们假设观测值为区间删失的概率在一年最后一个季度较高，该季度大约对应第40周至第52周。我们用 $W_{Q4} = \{40, ..., 52, 92, ..., 104\}$ 表示2016年和2017年最后一个季度的周集合。因此，在情景5中，我们为 $w \in W_{Q4}$ 选择 $p_w = 0.5$，为 $w \notin W_{Q4}$ 选择 $p_w = 0.3$；在情景6中，我们为 $w \in W_{Q4}$ 选择 $p_w = 0.6$，为 $w \notin W_{Q4}$ 选择 $p_w = 0.2$。

图 2展示了在模拟情景1、2和3下估计的星期效应，其中考虑了完整案例和完整模型。在情景2和3中，从星期五到星期日存在更多时间不确定观测，导致这些日期在用于拟合完整模型的数据集中代表性不足，因为这些区间删失事件被从分析中移除了。这导致星期五、星期六和星期日的估计值较低，从而得出这些日期发生的事件较少这一错误结论。在情景1中不会出现这种情况，因为在该情景下，所有星期几被移除的事件比例相同，因此各效应的估计值相近。另一方面，可以看出，完整模型处理了某些观测值中存在的时间不确定性，能够正确估计出所有星期几大致呈现相同的效应。

图 3 显示了在模拟情景 4、5 和 6 下，通过完整案例模型和完整模型估计的结构化周效应。正如预期，由于模拟中日历年最后一个季度的时间不确定事件比例较高，完整案例模型检测到该季度的一些周风险较低。相比之下，完整模型给出的结构化时间随机效应估计值围绕 0 波动，这与不存在周效应的情况一致。因此，即使存在时间不确定观测，完整模型仍能捕捉潜在的风险水平。

4.2 区域数据分析

4.2.1 完整案例分析与完整分析对比

本文的一个主要目标是比较使用完整案例模型（舍弃时间不确定事件）和完整模型所得到的结果，研究对象是瓦伦西亚在2016年和2017年期间记录的入室盗窃数据集。具体而言，目标是检验完整案例模型是否会导致参数估计产生偏差，以及完整模型能否捕捉事件的真实时间分布。因此，表1从.....方面总结了结果

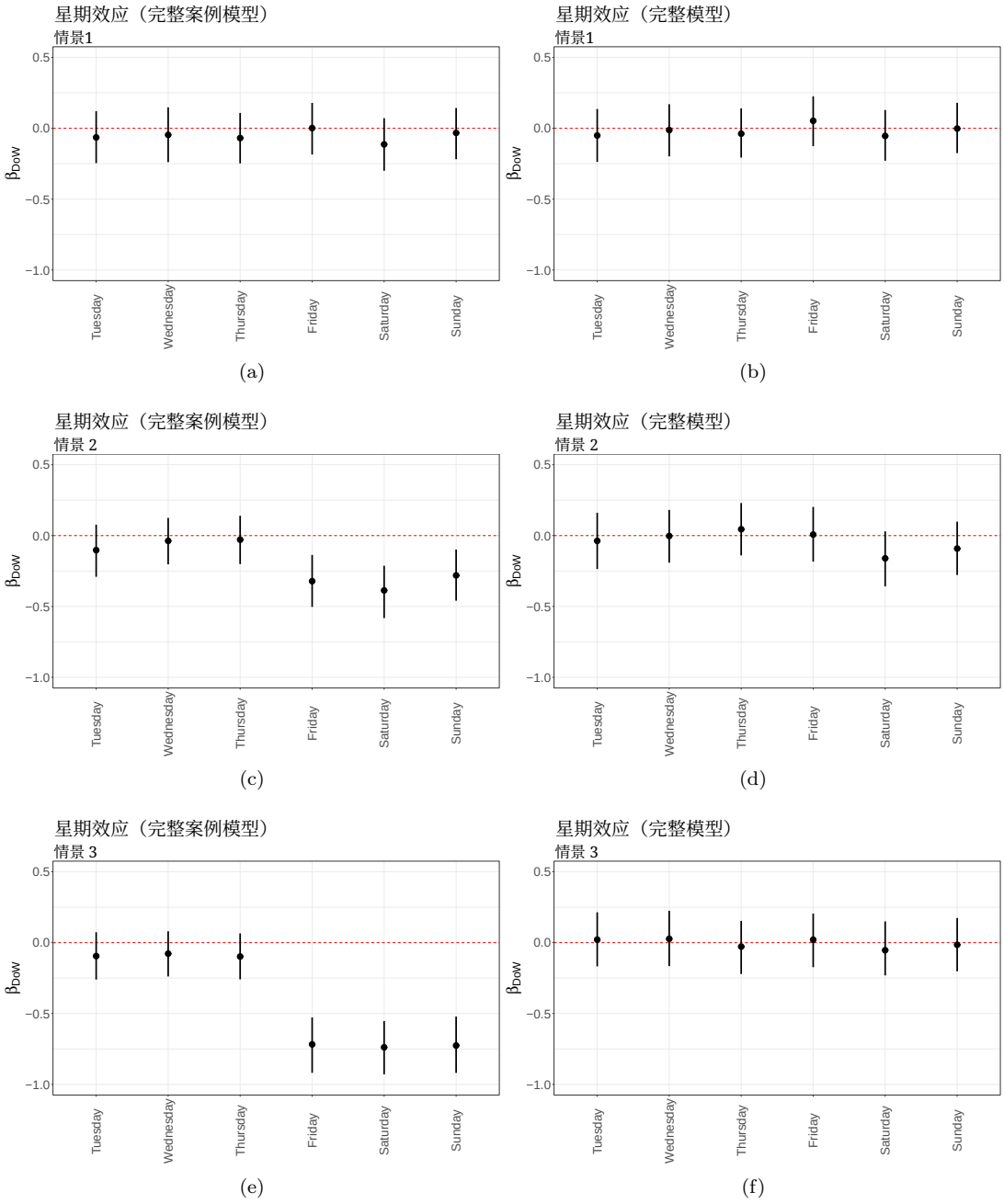
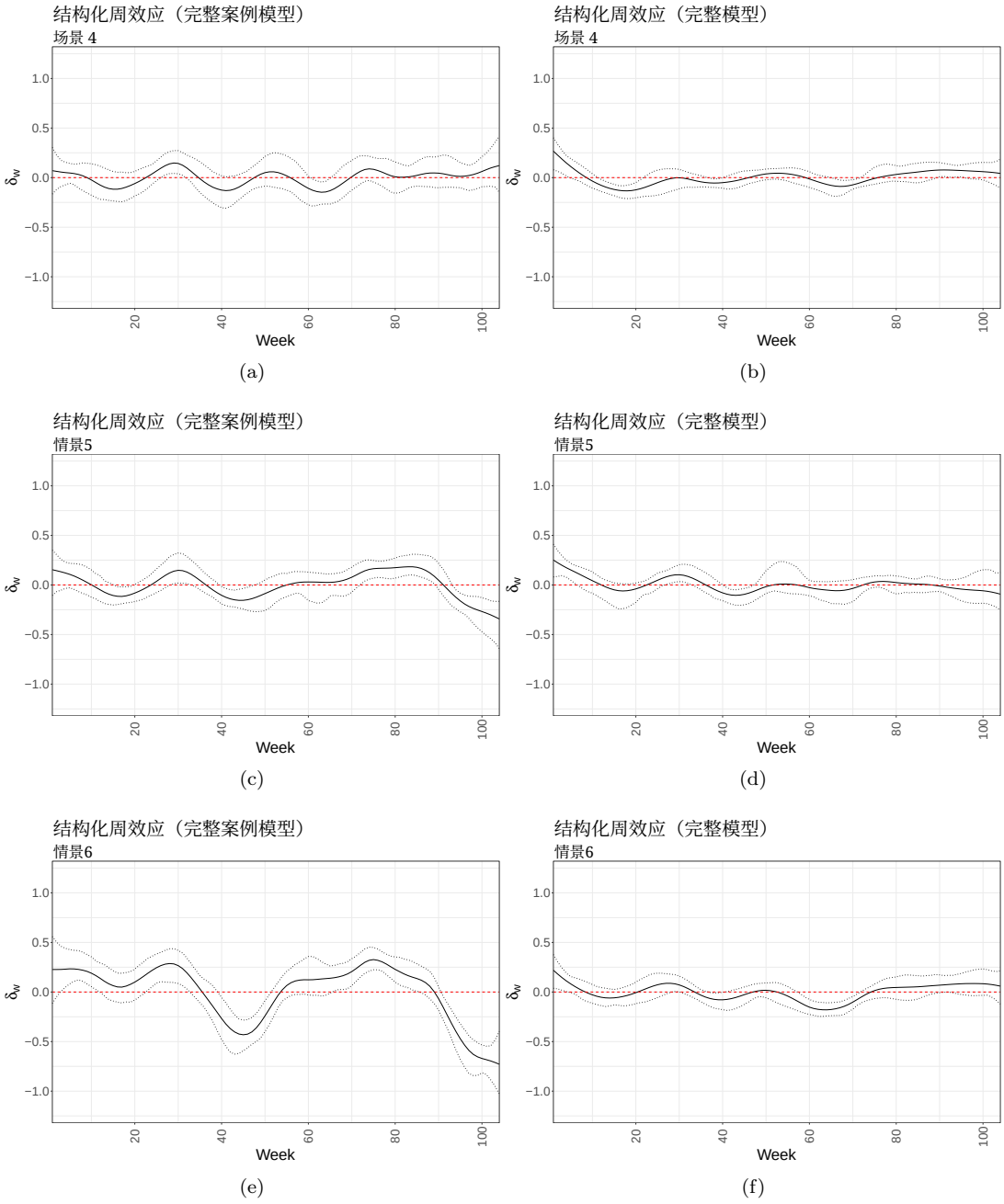


图 2：模拟情景 1、2 和 3 下，完整案例模型和完整模型得出的星期效应估计



(a) (b) (c) (d) (e) (f)

图 3：在模拟情景 4、5 和 6 下，完整案例模型和完整模型得出的星期效应估计

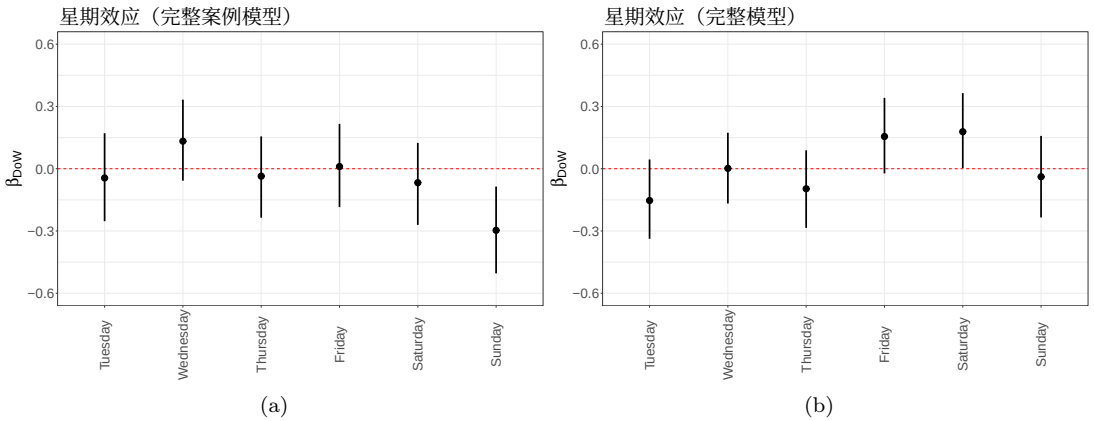
两个模型中相关参数的点估计和区间估计，可供直接比较。首先，我们注意到，完整案例模型中的 α 估计值较小。这是因为用于拟合该模型的数据集中病例所占比例低于完整模型所用数据集，因此事件发生的基线概率估计值也较低。无论如何， α 参数在解释方面并无重要意义。

完整案例模型				完整模型			
参数	**Est.**	**Lo.**	**Up.**	**参数**	**Est.**	**Lo.**	**Up.**
α	-2.17	-2.34	-2.01	α	-1.74	-1.89	-1.60
$\beta_{Tuesday}$	-0.04	-0.25	0.17	$\beta_{Tuesday}$	-0.15	-0.34	0.04
$\beta_{Wednesday}$	0.13	-0.06	0.33	$\beta_{Wednesday}$	0.00	-0.17	0.17
$\beta_{Thursday}$	-0.04	-0.24	0.16	$\beta_{Thursday}$	-0.10	-0.28	0.09
β_{Friday}	0.01	-0.18	0.22	β_{Friday}	0.16	-0.02	0.34
$\beta_{Saturday}$	-0.07	-0.27	0.12	$\beta_{Saturday}$	0.18	0.00	0.36
β_{Sunday}	-0.30	-0.50	-0.09	β_{Sunday}	-0.04	-0.23	0.16
τ_δ	1748.60	543.62	3924.02	τ_δ	880.02	187.52	2172.18
τ_ε	35.99	13.94	91.83	τ_ε	31.35	12.72	75.96
τ_u	2.33	1.21	4.33	τ_u	2.11	1.18	3.60
τ_v	111.07	12.66	529.44	τ_v	181.80	13.90	1036.27

表1：所考虑的两个模型给出的参数估计值（Est.），以及各参数95%可信区间的上限（Up.）和下限（Lo.）。每个 τ参数对应于相应随机效应的精度，定义为其方差的倒数。例如， $\tau_\delta = 1/\sigma_\delta^2$ 是结构化时间随机效应的精度

分析星期几的影响更有意义，这些影响由 β_{DoW} 参数表示（即与参考水平星期一相比的风险变化）。在这种情况下，差异十分显著，从实际角度来看也具有重要意义。如图4所示，根据完整案例模型，星期日的犯罪风险明显较低，而犯罪风险的最高估计值出现在星期三。相比之下，完整模型显示，星期五和星期六是入室盗窃风险较高的两天（尽管与 β_{Friday} 相关的95%可信区间下限略低于0），而星期日的入室盗窃风险并不低。因此，所采用的模型不同，结果也存在显著差异。事实上，如图5所示，对住宅入室盗窃按星期几的分布进行模糊时间分析，可以理解模型之间的差异。可以看出，在分析中纳入时间不确定事件，会使犯罪比例在周末达到最高，尤其是星期六。换言之，这表明涵盖（部分或全部）周末的时间不确定事件比例较高，这可能是因为部分人口在这些日子会住在第二住所，也可能只是因为

周末日常作息的变化可能使窃贼更容易在某些时段作案。可以通过开展小时级分析来研究这类合理的假设，不过这也可能使模型估计更加复杂。无论如何，完整案例模型完全遗漏了这类信息，而完整模型则将其纳入，即使数据存在区间删失也是如此。



(a) (b)

图4：完整案例模型（a）和完整模型（b）得出的星期效应估计

分析完模型中唯一的固定效应（星期几）后，我们比较了两个模型给出的时间和空间随机效应估计。首先，图6展示了结构化时间随机效应的估计， δ_w。尽管两个模型所捕捉的时间趋势总体表现相似，但仍存在一些差异。例如，完整案例模型确定了研究期间初期入室盗窃风险的一个峰值，而完整模型并未确定这一峰值。此外，完整模型检测到第70至90周左右犯罪风险出现双峰，而完整案例模型则在这一期间确定了单个峰值。最后，两个模型在第30周左右检测到的峰值中，完整模型给出的估计明显更高。实际上，与完整案例模型相比，完整模型能够检测到更多的周级变异。这一点可从图6中看出，也可通过比较随机效应 δ_w的精度估计得知。如表1所示，在完整模型中， τ_δ 的估计值明显更小，这证实了随机效应 δ_w 在后一个模型中捕捉到更多变异，因为 $\tau_\delta = 1/\sigma_\delta^2$。

按周分析住宅入室盗窃的模糊时间分布，使我们能够再次确认，完整模型充分捕捉了入室盗窃的时间分布。因此，如图

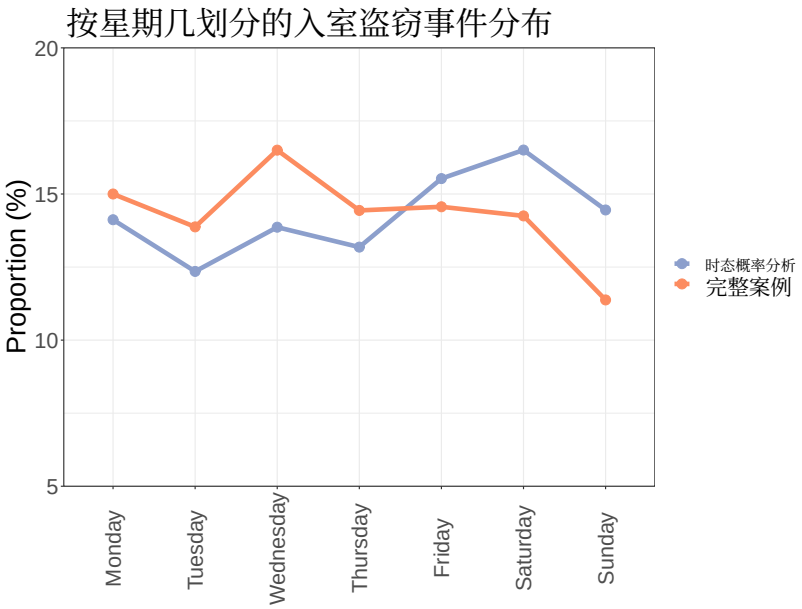


图5：按星期几（DoW）划分的入室盗窃事件分布，仅考虑确切日期已知的事件（完整案例分析），以及通过对时间不确定事件进行模糊时间分析得到的整个可用数据集

图7所示，夏季月份（七月和八月）时间不确定事件的数量明显较多，此时大多数居民都在度假，住宅空置数天甚至数周的可能性随之增加。此外，模糊时间分析还显示，2017年五月的入室盗窃数量达到峰值，这可能是由一个不同于夏季峰值所对应过程的独立过程造成的。2017年五月的这一峰值是完整案例模型检测到的唯一峰值，因为如图7所示，这一时期时间不确定事件的比例相当低。矛盾的是，完整案例模型表明2017年夏季是低风险时期，但这只是由于这些夏季月份中存在大量时间不确定事件，导致数据代表性不足。

最后，图8使我们能够比较两个模型所得的空间随机效应估计值。在本例中，差异很小，仅在市中心周边的某些邻域更为明显。换言之，这些结果表明，对于所分析的数据集，时间不确定事件的出现并不遵循某种空间模式。

4.2.2 事件时间插补

所提出的完整模型的主要优势在于，它允许按照时段概率方法将时间不确定事件纳入分析。这样可以避免缩减样本量

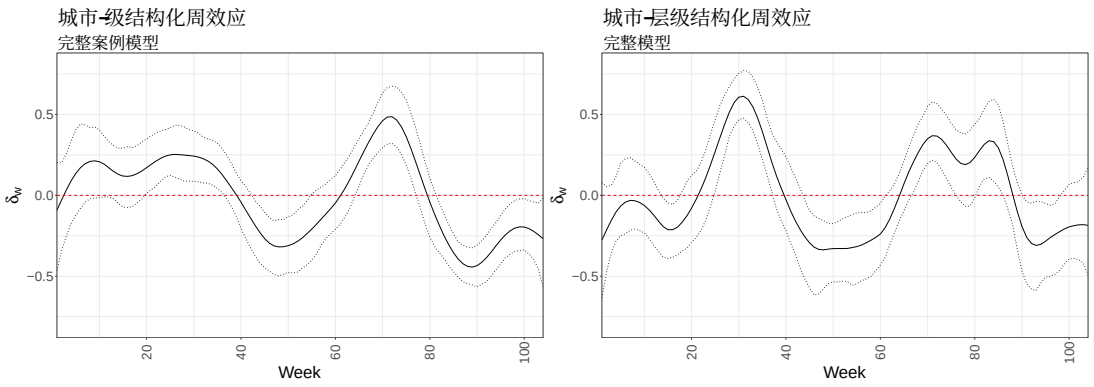


图6：完整案例模型（a）和完整模型（b）得出的周效应估计（结构化组分）。虚线界定了与每个点估计相关的95%可信区间。

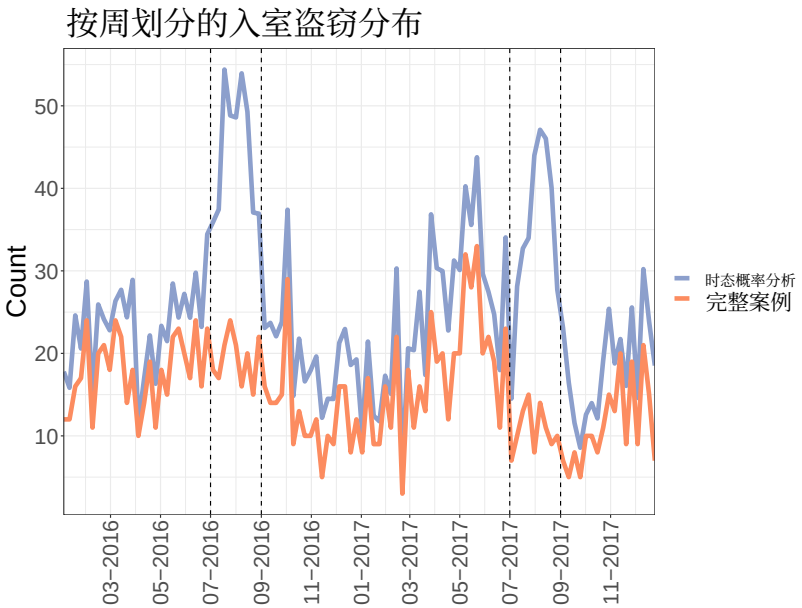
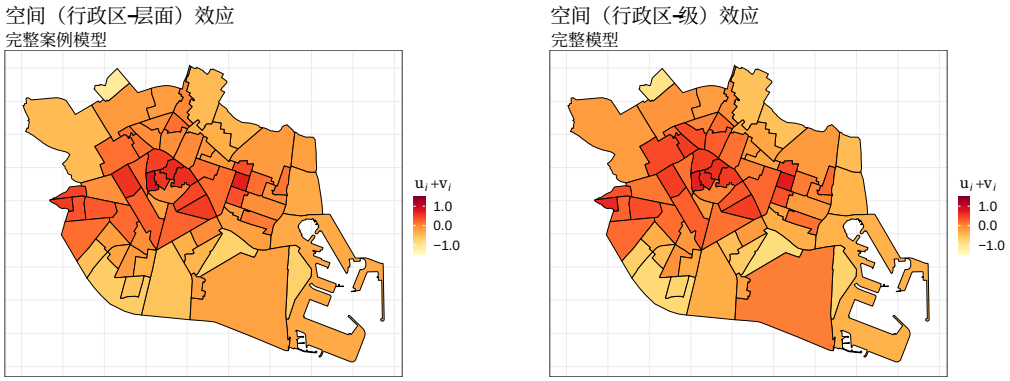


图7：仅考虑确切日期已知的事件（完整案例分析）时按周划分的入室盗窃分布，以及通过对时间不确定事件进行模糊时间分析得到的整个可用数据集。虚线界定了两年的夏季月份（七月和八月）。



(a) (b)

图8：完整案例模型（a）和完整模型（b）得出的空间效应估计（行政区层级）

（如完整案例模型中所述），并避免了插补事件发生时间可能带来的误差。此外，该模型的另一项优势是，可以根据每个时间单位（在本研究中为天）的后验概率，对事件时间进行插补；这些时间单位位于界定各事件不确定性范围的区间内。因此，图9显示了2016年1月6日至2016年1月28日期间发生的一组时间不确定的入室盗窃案对应的 $p(t_i^{event} = t|D)$值。每个案例中， t 的取值范围从 t_i^{from} 到 t_i^{to}，因此区间内的每个时间单位都被赋予了一定的发生概率。 $p(t_i^{event} = t|D)$值依据模型中包含的所有信息计算得出，涵盖所考虑的固定效应和随机效应。具体而言，时间不确定性与星期效应和时间随机效应相关（另一方面，空间随机效应不受时间不确定性的影响）。因此， $p(t_i^{event} = t|D)$值反映了研究期间每个星期几和每周的估计犯罪风险。具体而言，与风险较高的星期几（星期五和星期六）相对应的 $p(t_i^{event} = t|D)$值往往较高。图9所示的部分事件清楚地体现了这一点：这些事件的时间不确定性为2天、3天或4天，不确定性窗口涵盖了整个或部分周末。对于其他不确定性更大的事件（时间窗口更宽），这一效应则不那么明显。

因此，所描述的建模方法也可用作插补技术，只需考虑数值 $\text{argmax}_t\, p(t_i^{event} = t|D)$。要评估这种插补的质量，需要一个数据集，其中事件根据警方记录具有时间不确定性，同时外部来源提供了确切的时间位置。

方法。

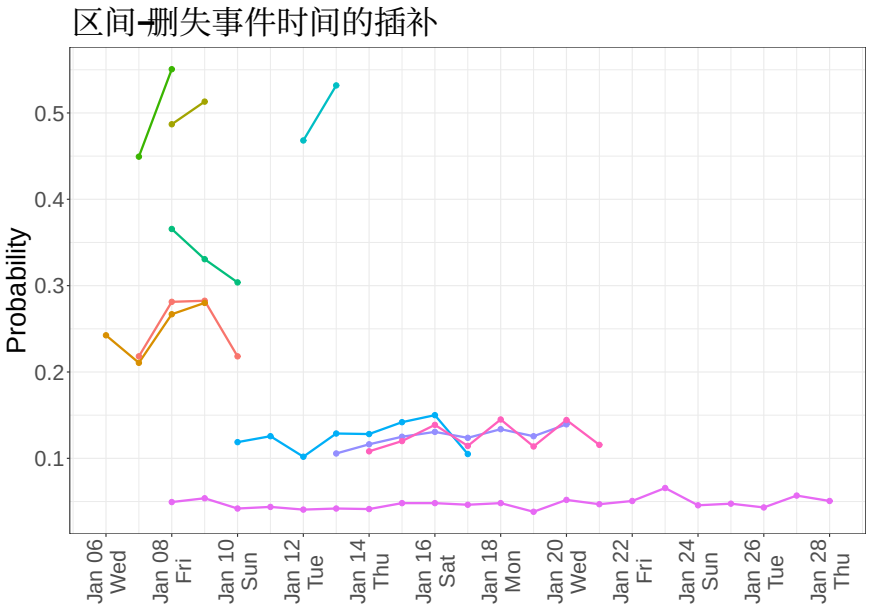


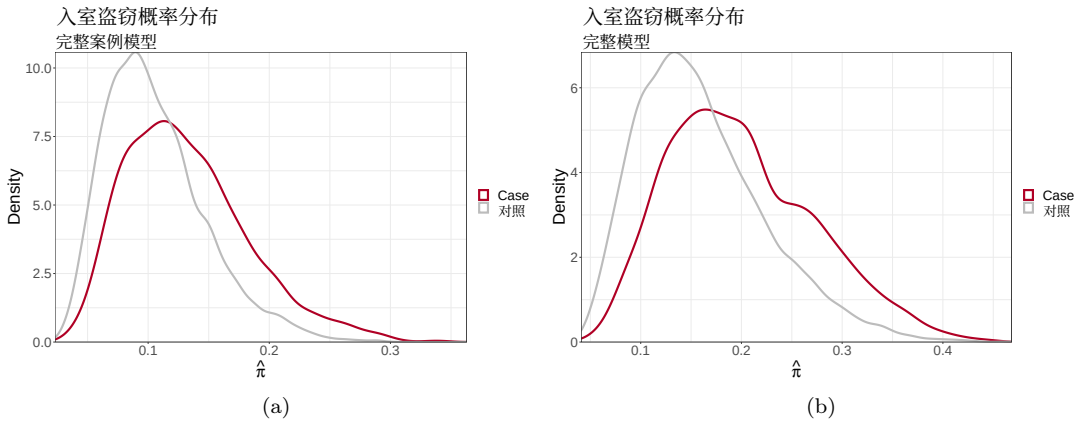
图9：2016年1月6日至2016年1月28日期间发生的部分时间不确定的入室盗窃案的 $p(t_i^{event} = t|D)$ 估计值

4.2.3 模型评估

评估模型质量的第一步，是分析 $\hat{}\pi_i$'s 的分布；该值是后验分布 $p(\pi_i|D)$ 的均值。图10展示了完整案例（图10a）和完整模型（图10b）中病例与对照组的 $\hat{}\pi_i$'s 分布。可以看出，两种模型都能恰当地区分病例与对照组，也就是说，病例的 $\hat{}\pi_i$'s 分布取值更高（分布整体更偏向右侧）。此外，由于完整模型中病例占比更高，该模型的 $\hat{}\pi_i$'s 分布覆盖的数值范围更大（病例与对照组均如此）。在比较模型的 α parameters 时已经讨论过这一点；稍后评估模型的分类能力时，这一点还会再次发挥重要作用。

此外，图 11 比较了通过完整模型估计的时间不确定事件与确定事件（精确到日期）的 $\hat{}\pi_i$ 分布。与时间不确定事件对应的 $\hat{}\pi_i$ 分布表现更好，因为该分布更偏向于较大的 π_i 估计值。通过这项分析，我们可以确认时间不确定观测已被妥善纳入模型，因为它们的表现并不比

从模型拟合的角度来看，优于那些未能做到这一点的模型（事实上，它们的表现似乎更好）.



(a) (b)

图10：完整案例模型（a）和完整模型（b）中病例（事件）与对照组对应的$\hat{\pi}_i$分布

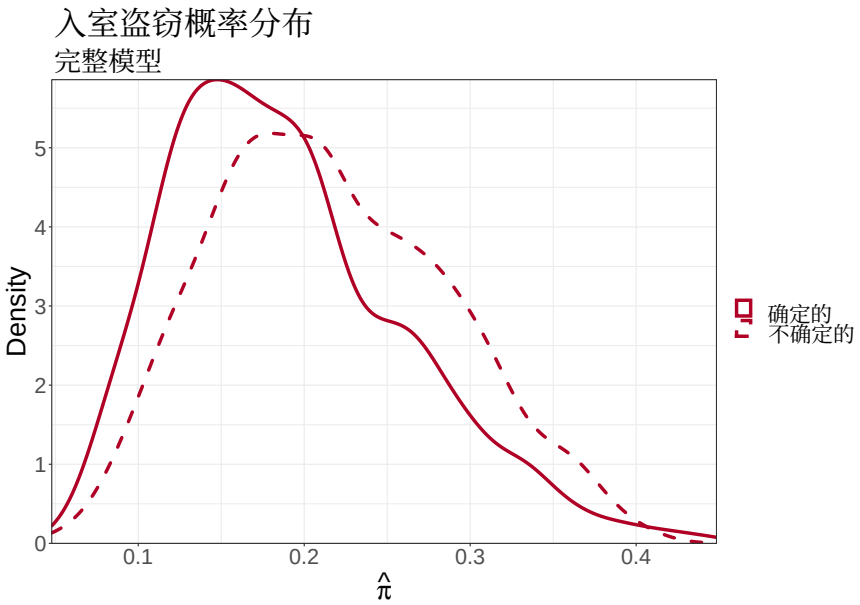
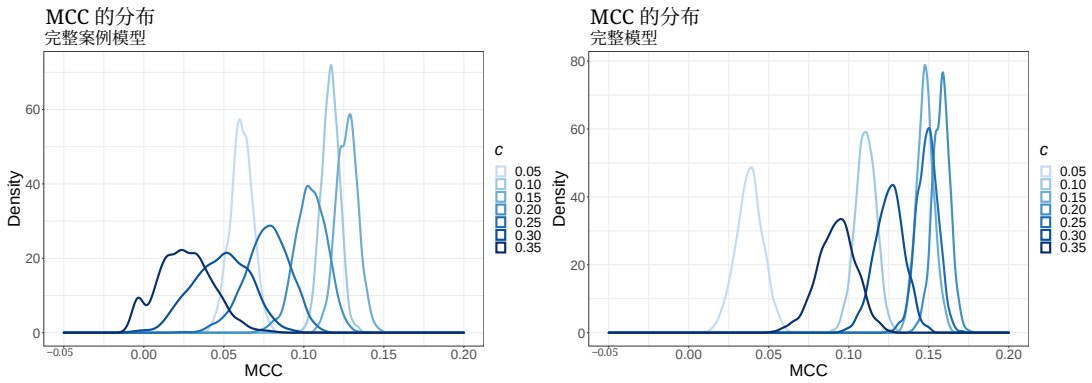


图 11：完整模型中，时间不确定事件与确定的（已知确切日期）事件所对应的 $\hat{\pi}_i$ 的分布

关于模型的分类能力，图 12 展示了根据后验分布 $p(\pi_i|D)$ 的抽样值推导出的 MCC 分布。截断概率 c 从 0.05 到 0.35 变化，步长为 0.05。由于阳性预测数量会变得过少，因此舍弃了较高的 c 值。这是因为数据集不平衡，病例观测值较少（与对照观测值的数量相比），导致模型

预测值小于 0.5。这并不是问题，我们只需考虑低于 0.5 的 c 值（对于平衡数据集而言，0.5 是常用阈值）。图 12 展示了两个值得关注的结论。首先，完整案例模型和完整模型的 c 最优值分别为 0.15 和 0.20（若对 c 值进行更细致的划分，就能更准确地估计每种情况下的 c 最优值）。完整模型的 c 最优值更高，这一点或许在意料之中，因为完整数据集中的病例比例更高（此前在讨论各模型的 α 参数估计值时已提及这一点）。其次，也是更重要的一点，图 12 表明完整模型的 MCC 值往往更高。具体而言，在考虑最优 c 值时，完整案例模型的 MCC 取值范围为 0.114 至 0.140（可信度为 95%）（对于$c = 0.15$），而完整模型的 MCC 取值范围为 0.147 至 0.168（可信度为 95%）（对于 $c = 0.20$）。在这些 c取值下，完整案例模型和完整模型对应的混淆矩阵分别为（$^{\text{TP}}_{\text{FP}}{}^{\text{FN}}_{\text{TN}}$）=（$^{513}_{19031}{}^{086}_{11148}$）和（$^{\text{TP}}_{\text{FP}}{}^{\text{FN}}_{\text{TN}}$）=（$^{1190}_{31511}{}^{434}_{9900}$）。同样，如图 13 所示，完整模型的 F1分数也更优。具体而言，无论选择哪种模型，F1分数在 $c = 0.15$时均达到最优。完整案例模型的 F1分数范围为 0.227 至 0.254（可信度为 95%），而完整模型的 F1分数范围为 0.328 至 0.339（可信度为 95%）。



(a) (b)

图 12：根据后验分布$p(\pi_i|D)$的抽样值计算得出的 MCC 分布，分别考虑完整案例模型（a）和完整模型（b）。对截断概率 c的多个取值进行了测试和比较

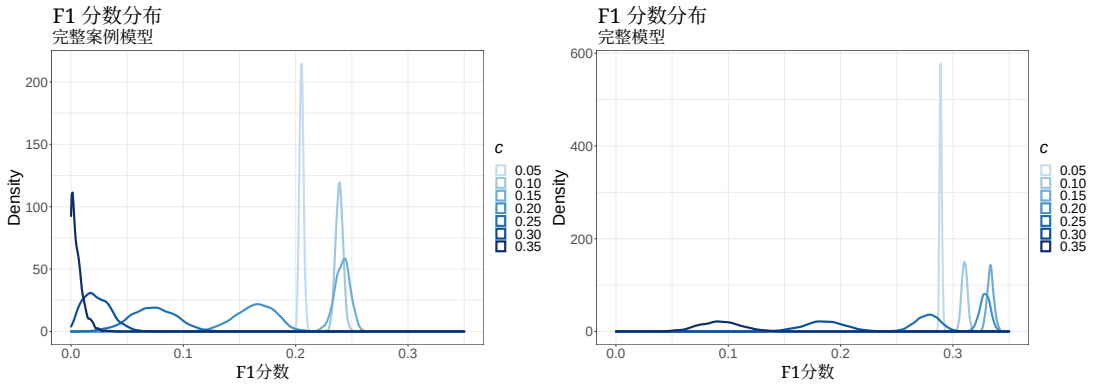


图 13：根据后验分布$p(\pi_i|D)$的抽样值计算得出的 F1 分数分布，分别考虑完整案例模型 (a) 和完整模型 (b)。对多个截断概率 c进行了测试和比较

5 讨论与结论

本文提出了一种逻辑回归模型，用于在存在时间不确定观测的情况下分析犯罪数据；这类观测在某些类型的犯罪中十分常见。本文考虑了时段概率法，该方法可在此情境下对数据进行探索性分析，并将此类时间不确定性纳入模型。鉴于缺失数据的贝叶斯处理，这是一种自然的方法。所实施的模型使我们得以发现，在分析中舍弃时间不确定观测可能会导致错误结论。尽管文献中已经提出了用于处理区间删失事件观测的这类建模方法（Reich and Porter，2015），但据作者所知，本文的创新之处在于将时段概率方法用于建模情境，并将所提出的模型与完整案例对应模型（这是一种典型的选择）进行了全面比较。

所提出的模型仍有改进空间。实际上，可以通过加入协变量信息和交互项来增强模型。例如，可以考虑时空交互随机效应，或星期几与一年中周次之间的交互。加入这类项或许能使模型提供更多信息，从而使基于 t_i^{event}后验分布对事件时间进行的插补更加贴近实际。其缺点是，纳入这些项会增加模型的复杂性，可能使模型估计更加困难，尤其是在可用犯罪数据集规模不大的情况下。此外，对时间

不确定性的处理方式还可以扩展到其他类型的模型中，例如空间效应并非基于区域结构的点过程模型（Mohler et al., 2011; Shirota and Gelfand, 2017）。

另一个重要方面是，在讨论完整案例和完整模型所提供的结果时，我们实际上不加说明地假定了对数据进行的模糊时间分析能够真实反映数据的时间分布。然而，正如 Mulder（2019）所指出的，模糊时间分析可能倾向于使事件的时间分布过度分散。因此，通过为观测窗口内的每个时间单位赋予相同权重，我们可能假定了过多的不确定性（变异性）；尽管如此，在我们不了解事件真实时间位置的情况下，这仍然是一种自然的方法。所提出的模型允许按照贝叶斯推断中采用的标准方式纳入关于事件的先验信息。例如，可以为那些可能有人（警方或财产所有者本人）凭直觉判断其实际时间位置的事件指定特定的先验分布，或者可以检验某种更接近现实的非均匀分布。事实上，未来的研究可以采用截断正态分布，将均值设在区间的中点位置，甚至设在更靠近区间起点（或终点）的位置。这或许能够揭示，在某些情况下，事件是否倾向于在时间窗口的前一部分聚集；如果窃贼一直在观察住户，并趁其离开住宅之机实施盗窃，这种现象便可以得到解释。

最后，为了更好地评估完整模型预测时间不确定事件真实时间位置的潜力，需要一个同时包含警方记录的区间删失时间位置（依据所有者或其他住户提供的信息）以及从其他来源获取的实际（精确）时间位置的数据集。例如，在 Ashby 和 Bowers（2013）进行的案例研究中，研究人员利用闭路电视摄像头的影像确定了所研究事件的精确时间位置。遗憾的是，这类数据集确实十分稀缺。

参考文献

Andresen, M. A., Malleson, N., Steenbeek, W., Townsley, M., and Vandeviver, C.（2020）。最低地理编码匹配率：数据与区域单元大小影响的国际研究。国际地理信息科学杂志，34(7):1306–1322。

Ashby, M. P. 和 Bowers, K. J.（2013）。时态概率分析犯罪的时间分析方法比较。《犯罪科学》，2(1)：1–16。

Baddeley, A.、Rubak, E. 和 Turner, R.（2015）。《空间点模式：方法论与 *R* 应用》。CRC 出版社。

Besag, J.、York, J. 和 Mollié, A.（1991）。贝叶斯图像复原及其在空间统计学中的两个应用。《统计数学研究所年鉴》，43(1)：1–20。

Bivand, R.、Keitt, T. 和 Rowlingson, B.（2019）。*rgdal*：‘地理空间’数据抽象库的绑定。R 软件包版本 1.4-6。

Bivand, R. 和 Rundel, C.（2020）。*rgeos*：*Geometry Engine - Open Source*（*‘GEOS’*）接口。R 软件包版本 0.5-3。

Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V., and Pebesma, E. J. (2008). 使用 *R* 进行应用空间数据分析，第 747248717 卷。Springer。

Briz-Redón, Á., Martínez-Ruiz, F., and Montes, F. (2020). 通过考虑时空犯罪风险异质性来调整 Knox 检验，以分析近重复犯罪。欧洲犯罪学杂志，页码 1477370820905106。

Briz-Redón, Á., Mateu, J., and Montes, F. (2022). 通过非线性模型识别犯罪发生器及空间上重叠的高风险区域：对瓦伦西亚地区（西班牙）三座城市的比较。荷兰统计学，76(1):97–120。

Briz-Redón, A.、Martinez-Ruiz, F.和 Montes,F.（2020）。重新估计开展空间分析所需的最低可接受地理编码命中率。《国际地理信息科学杂志》，34(7)：1283–1305。

Buil-Gil, D., Medina, J. 和 Shlomo, N.（2021a）。衡量地理区域内犯罪的暗数：基于英格兰和威尔士犯罪调查的小区域估计。《英国犯罪学杂志》，61(2):364–388。Buil-Gil, D., Moretti, A. 和 Langton, S. H.（2021b）。犯罪统计的准确性：评估警方数据偏差对地理犯罪分析的影响。《实验犯罪学杂志》，第 1–27 页。Chicco, D. 和 Jurman, G.（2020）。在二元分类评估中，马修斯相关系数（MCC）相较于 F1分数和准确率的优势。《*BMC* 基因组学》，21(1):6。Chung, J. 和 Kim, H.（2019）。犯罪风险地图：犯罪数据的多变量空间分析。《地理分析》，51(4):475–499。

de Valpine, P.、Turek, D.、Paciorek,C.J.、Anderson-Bergman, C.、Lang,D. T. 和 Bodik, R.（2017）。使用模型进行编程：利用 NIMBLE为通用模型结构编写统计算法。计算与图形统计学杂志，26(2):403–413。

Gail, M., Williams, R., Byar, D. P., Brown, C. 等人（1976）。需要多少对照组？*Journal of ChronicDiseases*，29(11)：723–731。

Gilardi, A.、Mateu, J.、Borgoni,R.和 Lovelace, R.（2022）。考虑空间网络格局的车祸数据多变量层次分析。皇家统计学会杂志 *A*辑，第1–28页。

Grolemund, G. 和 Wickham, H.（2011）。借助 lubridate 轻松处理日期和时间。《统计软件杂志》，40：1–25。Hong, E.-P. 和 Park, J.-W.（2012）。遗传关联研究中的样本量与统计功效计算。《基因组学与信息学》，10(2)：117–122。Law, J.、Quick, M. 和 Chan, P.（2014）。用于分析小区域尺度犯罪局部模式随时间变化的贝叶斯时空建模。《定量犯罪学杂志》，30(1)：57–78。Li, G.、Haining, R.、Richardson, S. 和 Best, N.（2014）。入室盗窃风险的时空变异：一种贝叶斯时空建模方法。《空间统计》，9：180–191。Matthews, B. W.（1975）。T4噬菌体溶菌酶预测二级结构与观测二级结构的比较。《生物化学与生物物理学学报（*BBA*）—蛋白质结构》，405(2)：442–451。

Mohler, G. O.、Short, M. B.、Brantingham, P. J.、Schoenberg, F. P. 和 Tita, G. E.（2011）。犯罪的自激点过程建模。美国统计学会期刊，106(493)：100–108。Mulder, K. T.（2019）。贝叶斯圆统计学：基于冯·米塞斯分布的实际问题解决方案。博士论文，乌得勒支大学。Quick, M.、Li, G. 和 Brunton-Smith, I.（2018）。犯罪通用与犯罪特定空间模式：小区域尺度下四种犯罪类型的多变量空间分析。《刑事司法杂志》，58：22–32。R Core Team（2021）。R：一种用于统计计算的语言与环境。Ratcliffe, J. H.（2000）。模糊时间分析：对时间不确定事件的空间解读。《国际地理信息科学杂志》，14(7)：669–679。Ratcliffe, J. H.（2002）。概率时段特征与高发犯罪模式的时空分析。《定量犯罪学杂志》，18(1)：23–43。

Ratcliffe, J. H. (2004). 犯罪地理编码与最低可接受命中率的初步估算。《国际地理信息科学杂志》，18(1):61–72.

Ratcliffe, J.H. 和 McCullagh,M. J.（1998）。概率时段犯罪分析。《国际地理信息科学杂志》，12(7):751–764。

Reich, B. J. 和 Porter, M. D.（2015）。用于入室盗窃犯罪序列识别的部分监督时空聚类。《皇家统计学会杂志：*A*辑（社会统计）》，178(2):465–480。

Shirota, S. 和 Gelfand, A. E.（2017）。空间与圆周时间对数高斯考克斯过程及其在犯罪事件数据中的应用。《应用统计年鉴》，第481–503页。Spiegelhalter, D. J.、Best, N. G.、Carlin, B. P. 和 Van Der Linde, A.（2002）。模型复杂度与拟合的贝叶斯度量。《皇家统计学会杂志：*B*辑（统计方法学）》，64(4)：583–639。Watanabe, S. 和 Opper, M.（2010）。奇异学习理论中贝叶斯交叉验证与广泛适用信息准则的渐近等价性。《机器学习研究杂志》，11(12)。

Wickham, H.（2016）。《*ggplot2*：数据分析的优雅图形》。Springer-Verlag New York。

Zhuang, J. and Mateu, J. (2019). 一种用于犯罪数据的、具有周期性背景的半参数时空Hawkes型点过程模型。《皇家统计学会杂志：*A*辑（社会统计）》，182(3):919–942。