用于疾病制图的可靠事件率

HarrisonQuick* **和 Guangzi Song**流行病学与《生物统计学》系，德雷塞尔大学，费城，宾夕法尼亚州 19104

*电子邮件：hsq23@drexel.edu

arXiv:2302.04582v1 [stat.ME] 9 Feb 2023

SUMMARY. 在分析空间定位事件数据时，将事件率认定为“可靠”的标准仍存在争议。然而，这些不同标准的共同点是：在小区域分析情境下，粗略估计值很少能够满足这些标准，因此促使人们使用空间模型来提高可靠性。尽管这看似合理，但近期研究已量化表明，空间统计学文献中的常用模型在多大程度上可能压倒数据所包含的信息，从而导致过度平滑。本文首先为事件率的“可靠”估计值提供一个定义，该定义既适用于粗略估计值，也适用于基于模型的估计值，并允许对可靠性作出离散和连续的表述。随后，我们构建了一个空间贝叶斯框架，使用户能够将先验信息融入模型中，以提高可靠性，同时防止过度平滑。我们将该方法应用于宾夕法尼亚州的县级出生数据，重点展示空间模型中的过度平滑效应，以及我们的方法如何帮助用户将注意力更好地集中于数据充足、能够支持推断决策的区域。最后，我们简要讨论了如何将这一可靠性定义用于小区域研究的设计。

KEY WORDS: 贝叶斯推断、信息性先验、早产、小区域分析、空间统计学

1 引言

在估计小区域内不良健康事件（例如心脏病相关死亡）的发病率或风险因素（例如肥胖）的患病率时，对于何种估计值可被视为“可靠”，目前尚未形成共识。例如，美国癌症统计（USCS）工作组建议，如果估计值基于16个或更多病例数，则将其认定为可靠（USCS，2022年）；而纽约州卫生部（1999年）等机构则要求病例数达到20个或更多。罗得岛州卫生部（2016年）和犹他州卫生部（2009年）提供了详细的流程图，说明在何种情况下应报告估计值、何时应在报告估计值时附加可靠性警告，以及何时应在其发布的报告中抑制估计值或汇总数据，这体现了该问题的复杂性。近期，国家卫生统计中心（NCHS；Parker等，2017年）发布了一份报告，详细说明了其报告比例的新标准；这些标准综合采用了基于样本量（大于30）的指导原则以及估计值置信区间宽度的指导原则。许多此类方法都基于变异系数（CV；也称为相对标准误）——即率估计值的标准误除以估计值本身——但这些规则背后的依据仍相当模糊。USCS的标准是一个例外：该标准要求估计值与其95%置信区间宽度之比（假设具有正态性）大于1，这等价于要求CV $< 1/4$。然而，无论采用何种标准，小区域数据往往无法满足必要的标准，尤其是在按照种族/族裔、性别和年龄等人口统计学因素进行分层时。

缓解对报告高不确定性估计值的担忧的一种方法，是将相邻空间区域和/或相邻时间段的数据进行聚合。例如，县级健康排名与路线图 项目（CHR&R；Remington 等，2015 年）将数据聚合到最长七年的时间段内，以便在构建用于确定各县排名的指标时，为尽可能多的县提供可靠估计。虽然聚合应能增加病例数——从而得出更可靠的估计值——

但它也可能妨碍对细粒度地理差异和/或时间变化进行推断。更糟糕的是，将基础发生率截然不同的相邻区域进行空间聚合，可能会产生无法准确反映各个单独区域情况的估计值（例如，Spielman 和 Folch，2015年；Bradley 等，2017年）。这些担忧构成了开展本研究的动机。

在分析空间参照数据时，除了数据聚合，一种颇具吸引力的替代方案是在贝叶斯框架下报告基于模型的估计值。例如，Besag 等人（1991）提出的条件自回归（CAR）模型及其多变量扩展（如 Gelfand 和 Vounatsou，2003；Quick 等人，2018），已被用于生成按年龄、种族和性别划分、覆盖逾40年的县级心脏病死亡率估计值（Vaughan 等人，2019）、2005—2015年的县级自杀率估计值（Khan 等人，2018），以及人口普查区级肥胖率估计值（Quick 等人，2020）。使用贝叶斯方法有双重好处：通过融入先验信息，我们不仅能获得更精确（因而也更可靠）的估计值，还能借助贝叶斯收缩获得更准确的估计值（Stein，1956）。不过，这种方法的主要缺点在于，量化并明确模型相对于数据所贡献的信息量等工作并不总是直截了当（例如 Morita 等人，2008）。例如，Quick 等人（2021）近期的研究表明，在泊松分布死亡率数据的情境下，Besag 等人提出的 CAR 模型对每个空间区域的贡献相当于超过30个事件；Song 等人（2021）在二项分布出生数据的情境下也发现了类似结果。在这两种情况下，若任由 CAR 模型框架自行发挥，其贡献在大多数区域都远 超数据本身的贡献，这可能导致过度平滑，并使推断的可信度受到影响。

本文的目标有两个。首先，我们基于统计点估计和区间估计，为发病率或患病率的“可靠”估计给出定义。该定义不仅适用于粗估计和基于模型的估计，也允许以离散或连续的方式表述可靠性——例如，区分可靠与不可靠，并能够指出某个估计更

可靠。其次，通过将该定义置于贝叶斯框架中，我们允许用户将先验信息（例如空间结构）融入模型，以提高估计值的可靠性。然而，在此过程中，我们也必须意识到这些模型可能对估计值产生的影响。因此，拟议工作的基础是需要“拥抱不确定性”。更具体地说，我们不会设计模型来确保所有估计值都可靠，而是会提供指导，限制常用的 Besag 等人的条件自回归模型的信息量，同时着眼于要求观察到足够数量的病例数，才能将估计值视为“可靠”。

2 方法

2.1 可靠性的定义

我们首先假设 $y_i \sim \text{Bin}(n_i, \pi_i)$，其中 y_i 表示区域 i 中的病例数，该区域的总体（若为调查数据，则为样本）规模为 n_i，而 π_i 表示潜在的事件率，其中 $i = 1, \ldots, I$。为了确定可靠事件率的定义，我们进一步假设 $\pi_i \sim \text{Beta}(a_i, b_i)$，由此得到如下形式的后验分布

$$\pi_i \mid y_i \sim \text{Beta}\left(y_i + a_i, n_i - y_i + b_i\right), \tag{1}$$

并由此将 a_i 和 b_i 分别解释为先验病例数和先验非病例数，先验总体总规模为 $a_i + b_i$。在此框架下，USCS 对可靠事件率的标准要求

$$\text{CV}\left(\pi_i \mid y_i\right) = \frac{\sqrt{V\left[\pi_i \mid y_i\right]}}{E\left[\pi_i \mid y_i\right]} = \sqrt{\frac{n_i - y_i + b_i}{\left(y_i + a_i\right)\left(n_i + a_i + b_i + 1\right)}} < 1/4. \tag{2}$$

为确定 y_i 在哪些条件下，式 (1) 中的后验分布能够给出可靠估计，我们首先注意到，式 (2) 中的要求可以重新表述为对后验分布的要求

病例数：

$$(y_i + a_i) > 16 \times \frac{n_i - y_i + b_i}{(n_i + a_i + b_i + 1)} \approx 16\,(1 - E\,[\pi_i \,|\, y_i])\,, \tag{3}$$

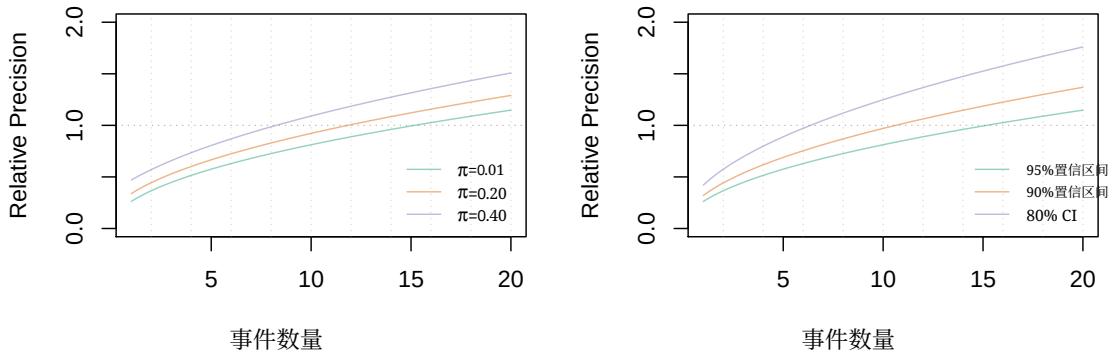
当 $n_i + a_i + b_i$ 较大时。条件（2）有一个实用之处：若我们将 π_i 的95%可信区间（CI）近似为 $E\,[\pi_i \,|\, y_i] \pm 2\sqrt{V\,[\pi_i \,|\, y_i]}$，那么满足（3）也会使 $E\,[\pi_i \,|\, y_i]$ 大于其95% CI的宽度。最后，虽然（1）–（3）中的关系基于二项分布数据，但当数据被建模为泊松分布时，也有类似结果；在分析出生率和死亡率（Brillinger，1986年）及其他稀有事件数据时，通常采用这种建模方式。网络附录A给出了泊松模型设定下的类似推导；当 π 较小时，其中的条件与二项数据的条件一致。

因此，我们将USCS可靠性标准推广为后验分布分位数的函数，如下所示：

定义 1。 率参数的估计值 π 由后验分布获得$p(\pi \,|\mathbf{y})$ 在 $1-\alpha$ 水平上是 可靠的，如果 π 及其相反情况 $1-\pi$的后验中位数均大于其各自 $(1-\alpha) \times 100\%$ 等尾可信区间的宽度。我们还将某个估计值在 $1-\alpha$ 水平上的相对精度 定义为其后验中位数与其 $(1-\alpha) \times 100\%$ 可信区间宽度之比，其中相对精度大于 1 表示该估计值在 $1-\alpha$ 水平上是可靠的。最后需要注意的是，定义 1 旨在确保 π 及其相反情况 $1-\pi$获得相同的可靠性区分——例如，对吸烟人群患病率的估计值，与对不吸烟人群患病率的估计值具有相同的可靠性水平。

2.2 信息性先验的影响

根据这一可靠性定义、式 (3) 的要求，以及针对 π_i 的相对非信息先验——即 $a_i = 1/2$ 和 $b_i = a_i(1-\pi_{i0})/\pi_{i0}$，使得 $\pi_{i0} = a_i/(a_i + b_i)$——我们



（a）相对精度 × 事件率　　（b）相对精度 × 可靠性水平

图1：相对精度随事件数量变化的比较。面板（a）展示了各种基础事件率下0.95水平的相对精度，面板（b）展示了事件率为 $\pi_i = 0.01$时不同可靠性水平下的相对精度。

需要观测到 16 例病例，才能在 $\pi_i = 0.01$ 时以 0.95 的水平获得可靠估计；当 $\pi_i = 0.20$ 时需要 12 例，当 $\pi_i = 0.40$ 时需要 9 例，如图1(a)所示。类似地，如果我们降低期望的可靠性水平，并假设 $\pi_i = 0.01$，那么要以 0.95 的水平将某个率判定为可靠，大约需要 16 个事件；以 0.90 的水平需要 11 个事件；以 0.80 的水平则需要 7 个事件，如图1(b)所示。除非另有说明，以下提及的可靠性和相对精度均指 0.95 的水平。

虽然上述结果很方便，但当我们考虑针对 π_i 的信息性 先验设定时，便能看出这一定义的可靠性真正的价值。例如，假设我们正在分析一个数据集，其中包含一个或多个小区域内早产婴儿（即孕周不足 37 周）的数量，以及这些区域的出生总数；先验信息表明，10% 的婴儿为早产。根据（3），我们的早产儿出生数后验值 $y_i + a_i$ 必须超过 $16 \times (1 - 0.10) = 14.4$，这些区域的估计值才会被视为“可靠”。因此，与其明确地 要求 y_i 超过某个阈值，才能将估计值视为可靠，不如将这类要求理解为对先验所能贡献的信息量上限的限制。例如，如果 $a_i < 16 \times (1 - 0.10)/2 = 7.2$，那么数据所提供的信息必须多于先验所提供的信息——即 $y_i > a_i$ ——估计值才会被视为可靠。

话虽如此，如上所述明确地 纳入先验信息并不常见。此外，在疾病制图中，常见的模型设定是

$$\operatorname{logit}\left(\pi_i\right) \mid \boldsymbol{\beta}, \mathbf{z}, \sigma^2 \sim \operatorname{Norm}\left(\mathbf{x}_i^T \boldsymbol{\beta}+z_i, \sigma^2\right), \tag{4}$$

其中，$\mathbf{x}_i$ 表示区域特定协变量的 p维向量，其对应的回归系数向量为 $\boldsymbol{\beta}$；z_i 则是基于 Besag 等人（1991）的条件自回归（CAR）模型框架定义的空间随机效应，记为$\mathbf{z} \sim \text{CAR}\left(\tau^2\right)$。尽管衡量式（4）中模型设定的信息量并非易事，Song 等人（2021）的近期研究建立了式（1）的后验分布与式（4）所得后验分布之间的关系。具体而言，他们首先使用德尔塔法令贝塔分布与逻辑正态分布的均值和方差相等，然后将这一关系推广到式（4）下给定其余 π_j、$j \neq i$时 π_i 的条件分布，并将空间随机效应$\mathbf{z}$从模型中积分消去，从而得到如下形式的近似：

$$\widehat{a}_i=\frac{1+\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{\sigma^2+\left(\sigma^2+\tau^2\right)/m_i}-\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T \boldsymbol{\beta})}, \tag{5}$$

其中，m_i 表示与区域 i相邻的区域数量。由于每个区域都可以有自己的协变量向量$\mathbf{x}_i$和自己的邻居数量 m_i，Song 等人（2021）建议将模型的基准信息量定义为

$$\widehat{a}_0=\frac{1+\exp(\mathbf{x}_0^T \boldsymbol{\beta})}{\sigma^2+\left(\sigma^2+\tau^2\right)/m_0}-\frac{\exp(\mathbf{x}_0^T \boldsymbol{\beta})}{1+\exp(\mathbf{x}_0^T \boldsymbol{\beta})}, \tag{6}$$

其中，（5）中的$\mathbf{x}_i$和 m_i分别替换为$\mathbf{x}_0$和 m_0；其中$\mathbf{x}_0=\left(x_{01}, \ldots, x_{0 p}\right)'$表示平均协变量值向量——即 $x_{0 j}=\sum_i x_{i j} / I$对于 $j=1, \ldots, p$而言——而 m_0 表示基准邻居数量，建议将 $m_0=3$ 作为经验法则，以便比较不同数据集。按照类似的流程，Quick 等人（2021年）

估计了类似的泊松模型设定的信息量，结果如下

$$\widehat{a}_0 = \frac{1}{\exp\left[\sigma^2 + \left(\sigma^2 + \tau^2\right)/m_0\right] - 1}. \tag{7}$$

网络附录B中提供了如何在 $\widehat{a}_0 < A$ for some $A > 0$ 的约束下拟合模型（4）的简要说明。

3 案例研究：宾夕法尼亚州各县的早产率

为说明这种方法在实践中的应用方式，我们对2010–2019年宾夕法尼亚州各县的早产开展了一项案例研究。数据按年份以及母亲的种族/族裔（白人、黑人、西班牙裔和亚裔）进行分层，数据来源于宾夕法尼亚州卫生部基于网络的企业数据发布信息交换系统（EDDIE）（宾夕法尼亚州卫生部，2020年）。在 EDDIE 系统中，种族/族裔被编码为白人、黑人、西班牙裔或亚裔，但需注意，这些群体可能并非相互排斥（例如，被编码为“白人”的母亲可能是西班牙裔或非西班牙裔白人）。不过，就我们的分析而言，我们将分别分析这些群体。同样，尽管考虑使用能够纳入不同种族/族裔背景母亲之间相关性和/或潜在发病率时间关系的模型可能颇具吸引力（例如，采用 Quick 等人（2017）的多变量时空条件自回归模型结构），但分别分析这些情形将作为展示可复现性的一种方式。

表1总结了数据。在这里，我们可以看到若干重要现象。首先，宾夕法尼亚州的人口种族构成十分明显：该州约70%的新生儿由白人母亲所生；此外，种族/族裔少数群体母亲——尤其是黑人和亚裔母亲——在地理上更多集中于费城和匹兹堡这两个城市中心。因此，种族/族裔少数群体母亲的数据较为稀疏，超过30%的

	白人	黑人	亚裔	西班牙裔
年平均出生人数	97,934	19,848	5,970	14,809
年平均早产人数	8,550	2,618	470	1,474
平均早产率	8.7%	13.2%	7.9%	10.0%
早产儿出生数为 < 10 的县所占百分比	4.5%	67.2%	83.7%	70.4%
早产数为零的县的百分比	0.3%	38.4%	50.3%	31.3%
城乡出生总数比	0.44	3.27	1.95	0.64

表1：宾夕法尼亚州早产出生数据摘要，以十年期间（2010—2019年）的平均值呈现。城乡出生比的计算方法是：将人口密度超过每平方英里1,000人的县定义为“城市”，这对应于费城大都市区（费城县以及邻近的巴克斯县、特拉华县和蒙哥马利县）和匹兹堡大都市区（阿勒格尼县）。

宾夕法尼亚州各县没有黑人、亚裔和西班牙裔母亲所生的零 早产儿出生数，且有三分之二的县早产儿出生数不足10。尽管如此，表1也显示早产发生率存在显著的种族/族裔差异：宾夕法尼亚州黑人母亲的早产发生率比非黑人母亲高出近50%。因此，尽管面临人口统计学挑战，按种族 考察县级早产趋势仍具有流行病学意义，这也促使我们采用空间模型，以获得比单靠数据所能提供的更稳定的估计值。

在此背景下，令 y_{irt} 和 n_{irt} 分别表示 i 县中种族为 r 的母亲在 t 年所生的早产儿出生数和出生总数。然后，我们假设 $y_{irt} \,|\, \pi_{irt} \sim \text{Bin}\,(n_{irt}, \pi_{irt})$，并采用与（4）类似的两种模型设定，按种族和年份的每种组合进行分层，以对 π_{irt} 建模。首先，我们令

$$p\left(\boldsymbol{\pi}_{\cdot rt}, \beta_{0;rt}, \mathbf{z}_{\cdot rt}, \sigma^2_{rt}, \tau^2_{rt} \,|\, \mathbf{y}_{\cdot rt}\right) \propto \prod_{i=1}^{I} \left[\text{Bin}\left(y_{irt} \,|\, n_{irt}, \pi_{irt}\right) \times \text{Norm}\left(\text{logit}\ \pi_{irt} \,|\, \beta_{0;rt} + z_{irt}, \sigma^2_{rt}\right)\right] \times$$
$$\times\ \text{CAR}\left(\mathbf{z}_{\cdot rt} \,|\, \tau^2_{rt}\right) \times \text{IG}\left(\sigma^2_{rt} \,|\, 1, 1/100\right) \times \text{IG}\left(\tau^2_{rt} \,|\, 1, 1/7\right), \quad (8)$$

表示一种标准CAR 建模方法，其中上述关于 σ^2_{rt} 和 τ^2_{rt} 的先验分布与既往研究一致（例如，Bernardinelli 等人，1995年；Waller 等人，1997年），而关于 $\beta_{0;rt}$ 的先验（即平坦的非正规先验）可表示为 $p(\beta_{0;rt}) \propto 1$。由于该方法

（8）中的方法可能倾向于产生过度信息化的模型，我们还考虑了一种模型设定，其中我们限制 $\widehat{a}_{0;rs} < 5$；如第2.2节所述，预计这一限制会导致这样的要求：只有当 $y_{irt} \geq 10$ 时，π_{irt} 的估计值才能在0.95的水平上被视为可靠。为此，我们遵循 Song 等人（2021年）的方法，并令

$$p\left(\boldsymbol{\pi}_{\cdot rt}, \beta_{0;rt}, \mathbf{z}_{\cdot rt}, \sigma_{rt}^2, \tau_{rt}^2 \,|\, \mathbf{y}_{\cdot rt}\right) \propto \prod_{i=1}^{I} \left[\mathrm{Bin}\left(y_{irt} \,|\, n_{irt}, \pi_{irt}\right) \times \mathrm{Norm}\left(\mathrm{logit}\pi_{irt} \,|\, \beta_{0;rt} + z_{irt}, \sigma_{rt}^2\right)\right] \times$$
$$\times \mathrm{CAR}\left(\mathbf{z}_{\cdot rt} \,|\, \tau_{rt}^2\right) \times \mathrm{IG}\left(\sigma_{rt}^2 \,|\, 1, 1/100\right)$$
$$\times \mathrm{IG}\left(\tau_{rt}^2 \,|\, 1, 1/7\right) \times I\left\{\widehat{a}_{0;rt} < 5\right\}, \tag{9}$$

表示我们的 受限 CAR 建模方法。标准模型和受限模型均使用马尔可夫链蒙特卡罗（MCMC）算法在 `R` （R核心团队，2020年）中拟合，每次运行总计进行 100,000 次迭代，并针对每种种族/族裔与年份组合分别运行。为保持一致并减轻事后分析的计算负担，每条链的前 50,000 次迭代作为预热期舍弃，最后 50,000 次迭代则按 10 倍抽稀，因此标准模型和受限模型针对每种种族/族裔与年份组合均获得相当于 5,000 次迭代的样本。

在讨论率估计值本身（及其可靠性程度）之前，我们先考察图2所示标准 CAR 模型的信息量估计值（基于 $\widehat{a}_{0;rt}$）。可以看到，Besag 等人（1991）的标准 CAR 模型框架对每个少数族裔群体的每个县、每年贡献了相当于 13–43 例早产儿出生数的信息量，这远高于这些县中大多数县观察到的早产儿出生数（如表1中所述）。白人母亲的情况虽不完全如此——模型似乎为每个县贡献的早产儿出生数少于数据所示——但模型贡献的数量仍超过每县 40 例早产儿出生数，因此，所有纳入考察群体的估计值都可能受到过度平滑的影响。

为说明过度平滑的潜在影响，我们考察了2010年麦基恩县白人母亲的估计值，以及亚当斯县黑人母亲的估计值，

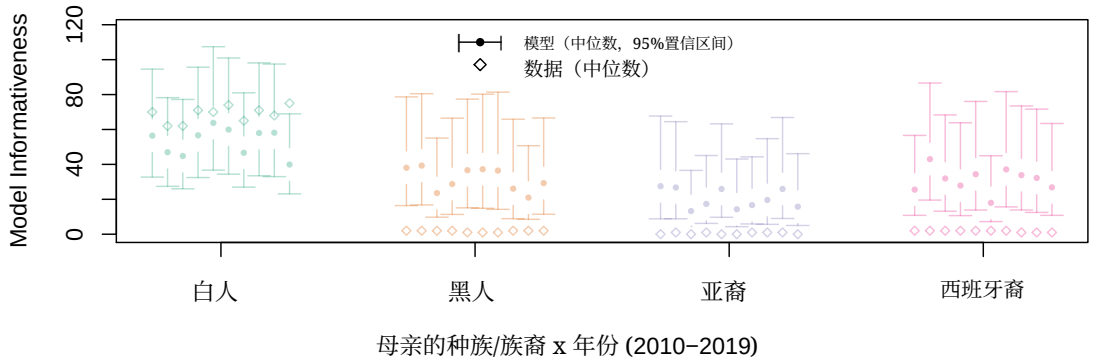
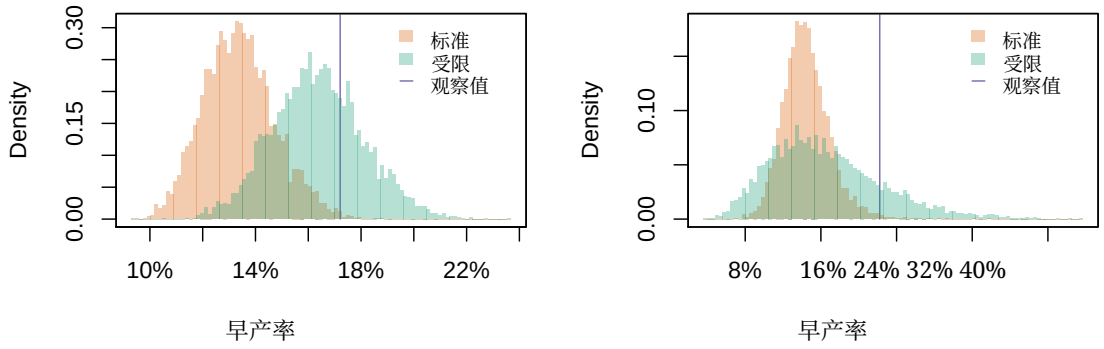


图2：宾夕法尼亚州早产分析中估计的模型信息量参数，$\widehat{a}_{0;rt}$；提供县级计数中位数供参考。

如图3所示，后者为2019年的估计值。在图3(a)中，我们看到，尽管受限 CAR 模型对麦基恩县白人母亲给出的估计值与观测数据（17.2%）相当一致——这反映了数据中观察到的早产儿出生数相对较多（74例）——但标准 CAR 模型更强的信息性使该估计值偏离其观测率，转而趋近于邻近县的比率（11.4%）。正如我们在网络附录C（尤其是图C.1）中更详细地说明的那样，这种极端的过度平滑可能会妨碍我们识别异常地区，从而阻碍州和地方卫生部门的干预工作。同时，图3(b)展示了标准 CAR 模型倾向于过度平滑估计值的第二个缺点：精度（无依据地）提高。具体而言，亚当斯县是一个以白人为主的农村县，2019年该县黑人母亲中仅观察到两例早产儿出生（出生总数为9例）。尽管该县的观测率与其两个邻近县以及全州黑人母亲平均值均一致，标准 CAR 模型所产生估计值的相对精度却是受限 CAR 模型所产生估计值的两倍，并且与一个早产儿出生数超过30例的县的估计值精度一致。

讨论了过度平滑对单个县估计值可能产生的影响后，我们现在将注意力转向本文的主要关注点：估计值和可靠性地图。为此，我们考察图4所示的地图以及

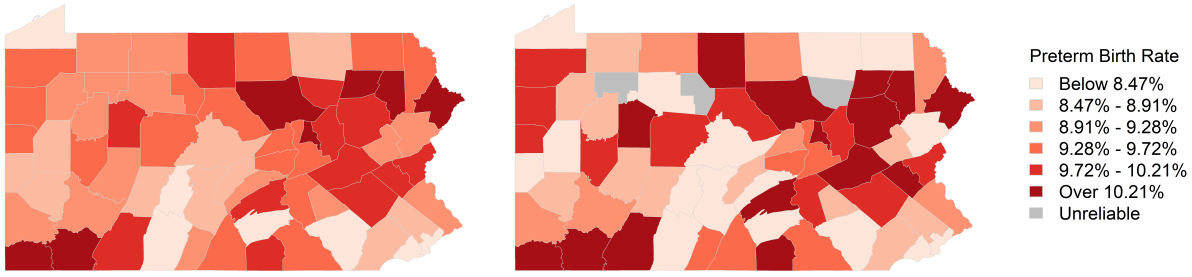


(a) 白人母亲；麦基恩县　　(b) 黑人母亲；亚当斯县

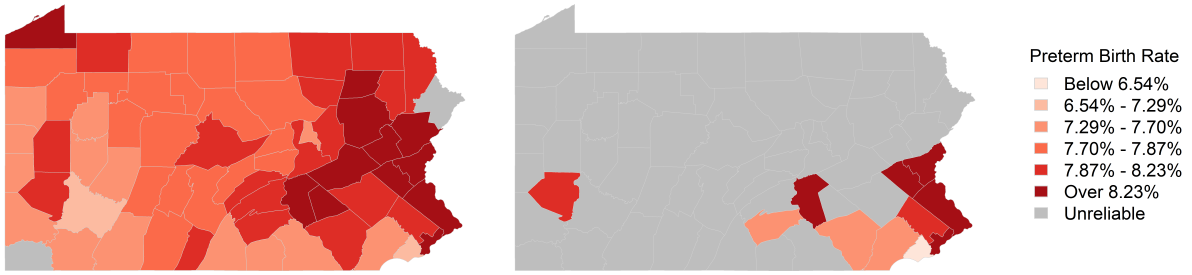
图3：比较选定县中白人和黑人母亲在标准 CAR 模型和受限 CAR 模型下率参数 π_{irt} 的后验分布。

图5所示的相对精度图。我们首先比较图4(a)中基于标准 CAR 模型得出的2019年白人母亲估计值与图4(b)中基于受限 CAR 模型得出的估计值。虽然总体地理分布模式相似，但受限模型的估计值更为极端，这体现了图3(a)中观察到的现象。此外，受限模型中有若干县的估计值被判定为不可靠；相比之下，标准 CAR 模型得出的所有估计值均被判定为可靠。如图5(a)的相对精度图所示，受限模型下不可靠的估计值（即相对精度小于 1 的估计值）都对应于 $y_{irt} < 10$ 的情况，这表明我们的 $\widehat{a}_{0;rt} < 5$ 限制达到了预期效果。

虽然图4(a)和图4(b)中白人母亲的结果对标准 CAR 模型与受限 CAR 模型之间的差异进行了较为细微的说明，但图4(c)和图4(d)中亚裔母亲的结果则展现出更为明显的差异。具体而言，与宾夕法尼亚州的白人人口不同，宾夕法尼亚州的少数族裔在地理上更加集中于该州的主要大都市区。因此，标准 CAR 模型更高的信息量不仅使除两个县以外的所有县都获得了可靠估计，还产生了一幅空间上非常平滑的地图，因为只有极少数县的早产儿出生数足以推翻模型的空间结构。相比之下，受限



(a) 白人母亲：2019年（标准）　(b) 白人母亲：2019年（受限）

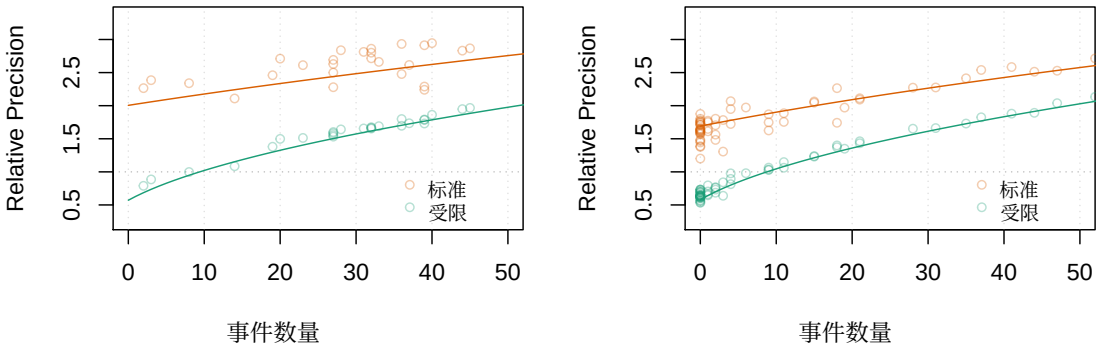


(c) 亚裔母亲：2019年（标准）　(d) 亚裔母亲：2019年（受限）

图4：标准和受限 CAR 模型中 2019 年白人母亲和亚裔母亲早产率的比较。

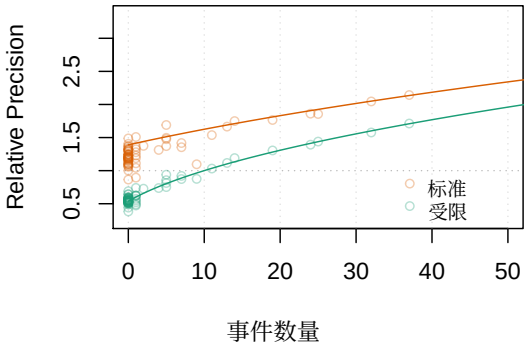
CAR 模型再次导致事实上必须 $y_{irt} \geq 10$ 才能获得可靠估计（如图5(c)所示），并在该州67个县中的56个县产生了不可靠的估计值。黑人和西班牙裔母亲以及整个十年研究期间也观察到类似结果，如网络附录中的图C.1和图C.2所示。

最后，至此为止的讨论一直将“可靠性”视为估计值“可靠或不可靠”的二元属性，但我们对可靠估计的定义有一个关键特点：通过确定$\alpha \in (0, 1)$中使其后验中位数大于其 $(1-\alpha) \times 100\%$ 等尾可信区间宽度的数值，我们可以判定估计值的可靠性水平。例如，图6展示了标准 CAR 模型和受限 CAR 模型对2019年亚裔母亲的估计值可靠性水平地图。正如图4(c)中首先观察到的，图6(a)显示，在标准 CAR 模型下，除两个县外，所有县的可靠性水平均高于0.95。相比之下，图6(b)

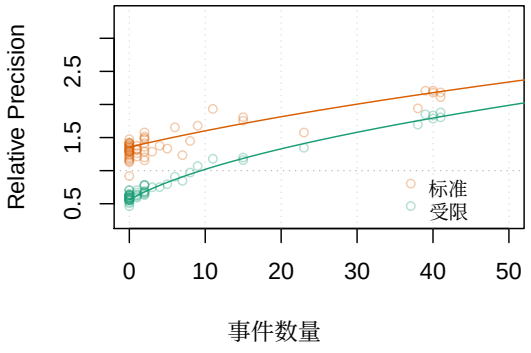


(a) 白人母亲：2019年

(b) 黑人母亲：2019年



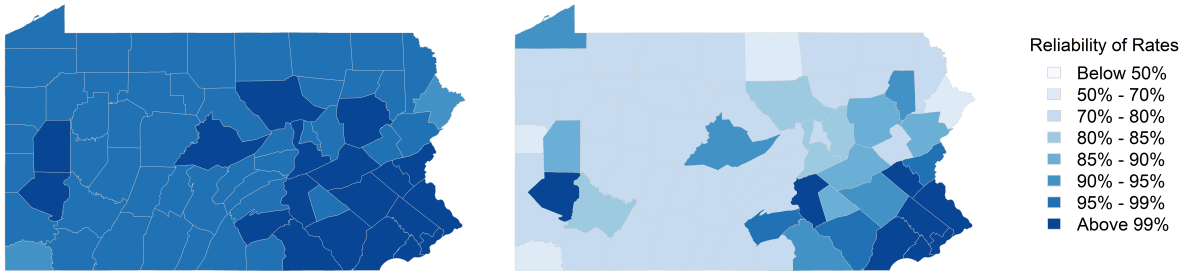
(c) 亚裔母亲：2019年



(d) 西班牙裔母亲：2019年

图5：比较标准 CAR 模型和受限 CAR 模型下 2019 年种族/族裔特定估计的相对精度。

凸显了宾夕法尼亚州亚裔人口的地理聚集程度，这体现在估计值可靠性水平的差异上。其他种族/族裔群体和其他年份的可靠性地图见网络附录图C.3。在该图中，我们不仅观察到标准 CAR 模型与受限 CAR 模型所产生的可靠性水平存在显著差异，也观察到年度间变异的差异。具体而言，受限模型对每个种族/族裔群体产生的可靠性水平逐年保持一致——这反映出基础数据的时间行为也具有类似的一致性——而标准 CAR 模型产生的估计值可靠性水平则表现出不规则的变化。这表现为实现可靠性所需的数据要求逐年不同，如网络附录图C.1所示。



(a) 亚裔母亲：2019年（标准）　　(b) 亚裔母亲：2019年（受限）

图6：比较标准和受限 CAR 模型下2019年亚裔母亲早产率估计值的可靠性水平。

4 讨论

本文的动机在于，统计学和流行病学文献对于事件率估计值要满足哪些条件才能被视为“可靠”，尚未达成共识。第2节提出的定义既适用于粗略估计值，也适用于基于模型的估计值；既支持离散的可靠性表述（依据某个预先设定的水平，判定为可靠或不可靠），也支持连续的可靠性表述。同样重要的是，我们提出的可靠性定义可直接与后验分布中的事件数量相关联——也就是观察到的事件数量与先验事件数量之和——从而使用户能够限制模型设定所包含的信息量，确保只有观察到最低数量的事件后，才预期能获得可靠估计。此外，尽管通过共轭贝塔-二项分布和泊松-伽马分布建模框架，最能清晰地阐释我们所定义的可靠性的性质，但 Quick 等人（2021）和 Song 等人（2021）提出的近似方法，使我们能够将这些性质扩展至疾病制图文献中常用的 Besag 等人（1991）框架，在这一框架下，可靠性问题经常出现。未来的工作旨在将这些限制扩展到其他较近期提出的疾病制图方法（例如 Leroux 等人，2000；Datta 等人，2019），以及用于多变量空间和时空情境的方法（例如 Gelfand 和 Vounatsou，2003；Quick 等人，2017），特别是用于根据按年龄分层的事件数据估计年龄标准化率。

本研究的一个重要立足点是，将某个估计值称为“可靠”，传达的是一种可信度，而不仅仅是表示该估计值精度高。因此，我们认为，使用基于模型的估计值，不应以生成“可靠”的估计值为目的，而应有意且审慎地提高估计值的精度。因此，本文旨在阐述研究者如何保持克制，并运用可靠性标准来指导统计模型的设计，从而获得具有理想精度水平的估计值。此外，此处给出的可靠性定义的一个关键特点是，对可靠性进行连续量化，为研究者呈现研究结果提供了一定的灵活性。例如，网络附录的图 C.4 展示了如何利用动态和/或交互式数据可视化工具，呈现可靠性水平各异的估计值；这可以作为一种替代方案，避免生成仅有少数“可靠”估计值的地图，或为了让更多估计值被视为“可靠”而人为提高估计值的精度（例如，通过采用信息量过高的模型）。

我们还应承认，在某些情况下，除人口最多的区域外，几乎所有区域在标准时间段（例如一年）内都不会发生足够多的事件，因而无法被视为可靠，无论是否有先验信息的贡献；尤其是在对罕见结果进行估计和推断时更是如此。在这些情况下，与其提高模型的信息量，不如将可靠性要求作为跨多个时间段聚合数据的依据。例如，根据第3节所考察的各个种族/族裔的预期早产率，将四年的数据与贡献了 $\widehat{a}_0 \approx 5$先验事件的模型设定相结合，足以为宾夕法尼亚州三分之一的县中的亚裔母亲，以及该州一半的县中的黑人和西班牙裔母亲，生成可靠估计；与使用单年数据相比，每个群体均增加了10多个县。可是，正如采用较宽松的可靠性水平一样，各机构需要权衡通过聚合提高可靠性的益处，以及在估计结果发布之前就可能被视为过时的风险。

参考文献

Bernardinelli, L.、Clayton, D. 和 Montomoli, C.（1995）。“疾病地图的贝叶斯估计：先验分布有多重要？”《医学统计学》，14，2411–2431。

Besag, J., York, J. 和 Molli´e, A.（1991年）。“贝叶斯图像复原及其在空间统计学中的两项应用。”《统计数学研究年刊》，43，1–59。

Bradley, J. R., Wikle, C. K. 和 Holan, S. H.（2017年）。“利用空间聚合误差准则对多尺度空间过程进行区域化。”《皇家统计学会会刊》*B*辑，79，815–832。

Brillinger, D. R.（1986年）。“生命率的自然变异及相关统计量。”《生物统计学》，42，693–734。

Datta, A., Banerjee, S., Hodges, J. S. 和 Gao, L.（2019年）。“使用有向无环图自回归（DAGAR）模型进行空间疾病制图。”《贝叶斯分析》，14，1221–1244。

Gelfand, A. E. 和 Vounatsou, P.（2003年）。“用于空间数据分析的适当的多元条件自回归模型。”《生物统计学》，4，11–25。

Khan, D.、Rossen, L. M.、Hedegaard, H. 和 Warner, M.（2018年）。“利用 R-INLA 对次国家级地区死亡率地理变异进行制图的贝叶斯时空建模方法。”《数据科学杂志》，16，147–182。

Leroux, B. G.、Lei, X. 和 Breslow, N.（2000年）。“小区域疾病率估计：一种用于空间依赖性的全新混合模型。”载于《流行病学、环境与临床试验中的统计模型》，M. E. Halloran 和 D. Berry 编，179–191。纽约州纽约市：Springer New York。

Morita, S., Thall, P. F., 和 M¨uller, P.（2008年）。“确定参数先验的有效样本量。”《生物统计学》，64，594–602。

纽约州卫生部（1999）。“基于小数目的比率——统计教学工具。”可在 https://www.health.ny.gov/diseases/chronic/ratesmall.htm 查阅。

Parker, J.、Talih, M.、Malec, D. 及数据抑制工作组（2017年）。《国家卫生统计中心比例数据展示标准》。国家卫生统计中心，生命与健康统计，2。

宾夕法尼亚州卫生部（2020年）。https://www.health.pa.gov/topics/HealthStatistics/EDDIE/Pages/EDDIE.aspx#。

Quick, H.、Song, G. 和 Tabb, L.（2021年）。《评估 Besag-York-Molli´e CAR 模型的信息量》。空间与时空流行病学，37，100420。

Quick, H.、Terloyeva, D.、Wu, Y.、Moore, K. 和 Diez Roux, A. V.（2020年）。《费城按种族/族裔、空间和时间划分的普查区层面肥胖患病率趋势》。流行病学，31，15–21。

Quick, H.、Waller, L. A. 和 Casper, M.（2017年）。“年龄特异性卒中死亡率的多变量时空建模。” 应用统计年鉴，11，2170–2182。

—（2018年）。“用于分析按种族和性别划分的县级心脏病死亡率的多变量时空模型。” 英国皇家统计学会会刊*C*辑，67，291–304。

R核心团队（2020年）。*R*：一种用于统计计算的语言和环境。R Foundation for Statistical Computing，维也纳，奥地利。

Remington, P. L.、Catlin, B. B. 和 Gennuso, K. P.（2015年）。“县级健康排名：理论依据和方法。”人口健康指标，13。doi：10.1186/s12963-015-0044-2。罗得岛州卫生部（2016年）。“罗得岛州小数值报告政策。”可访问 http://www.health.ri.gov/publications/policies/SmallNumbersReporting.pdf。

Song, G.、Tabb, L. P. 和 Quick, H.（2021年）。“宾夕法尼亚州低出生体重发病率的地理和种族差异。”arXiv 预印本 arXiv:2106.10571。

Spielman, S. E. 和 Folch, D. C.（2015年）。“通过数据驱动的区域划分减少美国社区调查中的不确定性。”*PLOS ONE*，10，2，1–21。

Stein, C.（1956年）。“多元正态分布均值的通常估计量不可容许。”载于《第三届伯克利数学统计与概率研讨会论文集》第*1*卷：统计理论的贡献，197–206。加利福尼亚州伯克利：加利福尼亚大学出版社。

美国癌症统计工作组（2022年）。“美国癌症统计数据可视化工具技术说明。”美国卫生与公众服务部、美国疾病控制与预防中心和美国国家癌症研究所。可访问：https://www.cdc.gov/cancer/uscs/pdf/uscs-data-visualizations-tool-technical-notes-508.pdf。

犹他州卫生部数据抑制决策规则工作组（2009年）。“数据结果抑制指南报告。”可访问：https://ibis.health.utah.gov/ibisph-view/pdf/resource/DataSuppression.pdf。

Vaughan, A. S., Schieb, L., Kramer, M. R., Quick, H., Taylor, H., and Casper, M. (2019年)。“种族与性别心脏病死亡率的率序变化：县级种族与性别差异探析。” *SSM — 人口健康*，7。doi：10.1016/j.ssmph.2018.100334。

Waller, L. A., Carlin, B. P., Xia, H., andGelfand, A. E. (1997)。“疾病率的分层时空制图。”美国统计学会杂志，92，607–617。