厘清因果推断中的空间干扰与空间混杂偏倚。

Isqeel Ogunsola[1,2] 和 Olatunji Johnson[1]

[1]曼彻斯特大学统计学系，英国。[2]联邦农业大学统计学系，阿贝奥库塔，尼日利亚

arXiv:2602.02777v1 [stat.ME] 2 Feb 2026

摘要

在处理观测空间数据时，空间干扰和空间混杂是阻碍获得精确因果估计的两大主要问题。此外，文献中对空间混杂的定义和解释仍存在争议。本文旨在从有向无环图（DAG）的视角，以一种新颖的方式澄清围绕空间混杂的误解与问题，并根据其对因果估计造成的偏倚，厘清直接空间混杂、间接空间混杂与空间干扰。此外，现有关于空间混杂偏差的分析通常依赖于处理和混杂因素服从正态分布的假设，而这些假设在实践中往往不成立。我们放宽这些假设，以泊松分布为例，在更一般的分布设定下推导出空间混杂偏差的解析表达式。我们发现，空间权重的选择、处理的分布以及干扰的程度，都会显著决定空间干扰所致偏倚的大小。我们进一步证明，直接空间混杂和间接空间混杂可以加以区分，其中权重矩阵和暴露的性质在决定间接偏倚的大小方面起着关键作用。模拟研究和真实世界空间数据的应用支持了理论结果。未来将开发参数框架，以便在估计直接效应和中介效应时，同时调整空间干扰以及直接和间接空间混杂。

关键词：观测空间数据、偏差、有向无环图、间接空间混杂、空间权重

1 引言

1.1 空间干扰

空间干扰是指一个位置的处理或暴露会影响另一个位置的响应。这种现象常被研究者忽视，因此为便于识别和处理效应估计，研究者通常假设不存在空间干扰（例如，Papadogeorgou 等，2019；Hern´an 和 Robins，2010）。这一假设与无多重处理假设共同构成了 Rubin 所称的稳定单位处理值假设（Rubin，1980）。然而，在空间设置中，无干扰假设可能不成立，因为一个位置的暴露可能会影响

空间依赖可能导致附近位置的结局受到影响。空间干扰有多种不同形式，其中最常见的是聚类（或部分）干扰和一般干扰。在聚类干扰中（Sobel，2006），单位被划分为若干组，并假设只有同一组内的单位会相互作用或产生干扰。相比之下，一般干扰（Giffin，2020）不设结构性约束，允许单位之间自由产生干扰。虽然在某些条件下可以识别此类干扰效应（Tchetgen 和 VanderWeele，2012），但现实应用中所需的数据很少能够获得（Giffin，2020）。

研究者提出了多种方法，用于估计观察性数据中的空间干扰效应。Giffin 等人（2023年）提出了一种方法，将倾向得分推广为在倾向得分模型中纳入干扰效应作为协变量，以估计直接效应和溢出效应，但他们假设不存在混杂效应。他们采用贝叶斯样条回归，以避免模型维度问题。Giffin 等人（2021年）也采用空间工具变量来估计干扰效应。不过，他们虽然考虑了混杂效应，但若要找到合适的空间工具变量，并检验／验证某些假设，即使并非不可能，也可能是一项艰巨的任务。在存在干扰的情况下，Zigler 和 Papadogeorgou（2020年）针对两种不同的干扰情形，提出了不同的可能因果估计目标。第一种情形是某个单位接受的处理不同于测量结局的单位所接受的处理。第二种情形是单位之间存在相互作用，即某一位置的单位会影响另一位置的结局。他们采用逆概率加权方法，估计了这些估计目标的一个简化子集。Tchetgen 和 VanderWeele（2012年）在存在空间干扰效应时推导出了新颖的逆概率加权估计量。Torrione 等人（2024年）还采用断点回归，在存在干扰的情况下估计因果效应。Torrione 等人（2024年）提出了一种基于距离的因果估计目标非参数估计量，并推导了其性质。该模型假设满足条件无混杂性。估计存在干扰时的处理效应并不局限于观察性数据，在随机试验中也有相关研究（例如 Basse 等人（2019年）；Athey 等人（2018年）；Sobel（2006年））。

这些研究人员尝试估计干扰效应，但未能量化和评估空间干扰存在却未加以考虑时所引发的偏差对因果估计的影响，而这正是本研究的一项贡献。

1.2 空间混杂

另一方面，文献中对空间混杂的定义存在争议。它主要被用来指两种不同的情况（Papadogeorgou and Samanta (2023), Urdangarin et al. (2023)）。首先，它被认为是回归模型中协变量与空间随机效应相关的情形（Dupont et al. (2022), Reich et al. (2006), Khan and Calder (2022)）。其次，它也被用来指因果推断中的一种情形：测得的协变量未能控制遗漏的混杂因素，且未测混杂因素具有空间结构（Gilbert et al. (2021), Papadogeorgou and Samanta (2023)）。与此同时，Gilbert et al. (2024) 指出，遗漏混杂偏倚、随机效应共线性、正则化偏倚和共曲线性有时都

会在文献中被称为空间混杂。最近，Khan and Berrett (2024) 尝试区分所有被称为空间混杂的术语，将这一现象分为两类：前者称为分析模型空间混杂，后者称为数据生成空间混杂。为作出贡献并进一步澄清概念，我们将在后文从有向无环图（DAG）的视角明确区分因果推断中的空间混杂与其他情形。Khan and Berrett (2024) 提出的这两种情形多年来受到了广泛关注。对于前一种情形，较有代表性的方法包括：空间+ 方法（Dupont et al. (2022)）、限制空间回归方法（Reich et al. (2006), Khan and Calder (2022)），以及变换高斯马尔可夫随机场（TGMRF）方法（Banerjee et al. (2014) and Prates et al. (2015)），仅举几例。这些方法通过用其他形式的随机效应替代空间随机效应，旨在消除协变量与新随机效应的方差-协方差矩阵之间的共线性。作者还比较了其中一些方法的表现，并确定在存在空间混杂因素时，哪种方法能产生最佳的固定效应估计值（在特定条件下）。有关这些比较的更多详情，请参见 Urdangarin et al. (2023)。然而，本文的重点是后一种情形，即因果推断中的空间混杂。在因果推断中，无混杂因素这一假设对于观察性数据并不成立；在空间设置下，情况则更为复杂。

另有多位研究者提出了缓解因果推断中这一挑战的方法。近期，Gilbert 等人（2021）提出了一种不依赖分布假设的因果效应估计方法。他们利用空间坐标调整未测量的混杂因素，并额外要求混杂因素可测性是空间函数，且暴露中不存在空间成分。他们采用双重机器学习程序来估计所关注的处理效应。Bobb 等人（2022）还提出了暴露惩罚样条，以减少空间混杂效应。他们通过空间平滑程度来解释暴露的空间变异性。在一些因果推断文献中，空间随机效应用于表示空间混杂因素（例如 Reich 等人（2006））。然而，有作者指出，用空间随机效应替代空间混杂因素并不能缓解未观测空间混杂问题（Schnell 和 Papadogeorgou（2020）；Paciorek（2010））。随后，Schnell 和 Papadogeorgou（2020）基于常用线性最小二乘估计量的偏差项，开发了一种用于处理空间混杂的仿射估计量。该方法采用贝叶斯方法实现。Thaden 和 Kneib（2018）也提出了地理加性结构方程模型，以缓解空间混杂。具体而言，他们同时量化响应和协变量中的空间效应，并在方法开发中纳入离散空间信息。Papadogeorgou 等人（2019）在倾向得分计算中采用调整距离，将单元之间的空间邻近性信息纳入其中，以改善空间混杂问题。研究表明，他们提出的调整方法优于其他在倾向得分调整中纳入空间信息的方法。

由于人们对空间混杂存在若干误解，且对其解决方案缺乏共识，研究者也尝试量化并评估偏差，以提出可能的解决方案。Dupont 等人（2023）量化了空间混杂因素引入的偏差，并指出

当频率较高（即权重接近 1）时，这种偏差不可忽视，这意味着空间相关性较强。Narcisi 等人（2024）运用二次型理论考察并评估了空间混杂偏差。其研究揭示，边际协变量变异性以及混杂因素与协变量之间的边际协方差，都是决定混杂所致偏差大小的关键因素。Keller 和 Szpiro（2020）提出了一种基于样条的方法，利用傅里叶滤波和小波滤波进行混杂调整，从而避开了样条自由度所带来的空间尺度解释困难。Paciorek (2010) 也指出，混杂偏差取决于残差和协变量的空间尺度。他还推断，当协变量的空间尺度小于未测量混杂的尺度时，拟合随机效应模型有助于减小偏差。与此同时，现有研究均未量化间接空间混杂所致的偏差，并假设其与直接空间混杂偏差相同，但事实并非如此。此外，当处理和混杂因素分布不服从正态分布时，空间混杂偏差的量化也尚未得到充分研究。

1.3 从有向无环图（DAG）的角度厘清空间混杂。

有向无环图通常简称为 DAG，是一种用于表示和可视化变量间因果关系的图。在图 1 中，A、Y 和 U 分别表示处理、响应和混杂因素。节点代表变量；之所以称为有向图，是因为从一个变量（例如 A）指向另一个变量（ Y ）的箭头表示 A 对 Y 有因果效应，而反向则不成立。图中的“无环”意味着 DAG 中不存在环。这表示 A例如不能以任何方式影响自身。除了用于对因果关系建模外，DAG 还能帮助我们更轻松地确定应对哪些变量进行条件化（即调整能够阻断非因果路径的变量），厘清因果假设，并找出暴露和响应可能存在的混杂因素。请注意， U 上的方框表示 U 缺失且具有空间结构。

本节从有向无环图的视角进一步阐明了关于空间混杂的问题和误解。



(a) 因果推断中的空间混杂　　(b) 其他空间混杂

图 *1*：空间混杂

图1a和图1b展示了因果推断与非因果推断中空间混杂的有向无环图。两幅图中从 U 指向 Y 以及从 A 指向 Y 的箭头表明，U 和 A 都会对 Y产生因果效应。然而，在非因果推断中，空间缺失变量 U 与处理 A 之间存在关联，但这种关联并非因果关系。

如图1b所示，两者之间以非箭头端的连线表示。另一方面，图1a中 U 与 A 之间的关系具有因果性，以从 U 指向A的箭头表示。这里所说的因果性意味着，这不仅仅是关联，而且第2节中所述的条件/假设对于 U 对 A的影响均成立。例如，在图1a中， U 必须先于A 发生，但在图1b中则不必如此。简而言之，因果推断中的空间混杂与非因果推断中的空间混杂之间的明显区别在于：缺失空间变量与 Y 和 A 之间的关系具有因果解释，而非因果推断中的空间混杂则不具备这种因果解释。

与此同时，空间混杂可能出现在两种情景中：（i）直接空间混杂，即未测量的空间变量影响同一位置上的处理和结局；（ii）间接空间混杂，即某一位置上的未测量空间变量会影响其他位置的结局和/或暴露，这通常是由空间干扰造成的。例如，在研究气候变化对健康的影响时，空间分布不均的未测量污染可能会影响同一位置上的气候变化和健康（直接空间混杂）。另一方面，一个位置上的污染可能导致气候变化，并影响附近位置的健康（间接空间混杂）。尽管人们已广泛关注量化和评估直接空间混杂造成的偏差，但间接空间混杂和空间干扰的影响仍未得到充分探讨。图2和图3展示了直接和间接空间混杂之间的细微差异，其中 U， A， Y分别为空间混杂因素、暴露和响应，下标1和2表示位置1和位置2。

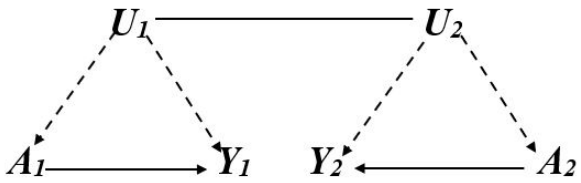

图*2*：直接空间混杂

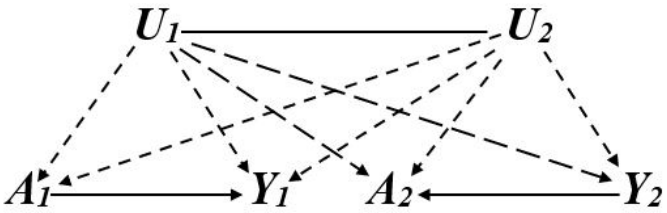

图*3*：间接空间混杂

在图2中，空间混杂因素 U_1 仅同时影响位置1处的响应和暴露，而 U_2 仅影响位置2处的响应和暴露。 U_1 与 U_2 之间的粗线表示混杂因素之间的空间依赖。相比之下，在图3中，混杂因素 U_1 影响两个位置的暴露和响应， U_2亦是如此。从 U_1 和 U_2 指向响应与暴露的带箭头虚线表明该混杂因素未被测量。

为此，我们的目标是：(i) 量化空间干扰和间接空间混杂所导致偏差的大小；(ii) 评估不同

情景下的偏差（采用模拟研究）；(iii) 推导处理和混杂因素服从不同抽样分布时的偏差；(iv) 提出一种校正空间干扰所致偏差的简单方法；(v) 从有向无环图的角度阐明空间混杂问题。接下来，我们将研究方法扩展到空间干扰和空间混杂同时存在的情况。

简而言之，本文的贡献如下：

（1）本研究从有向无环图（DAG）的视角阐明了空间混杂问题。（2）据我们所知，这是首篇在空间和非空间情境下对因果推断中的空间干扰所导致的偏差进行量化和评估的论文。（3）本文将直接空间混杂偏倚与间接空间混杂偏倚区分开来，纠正了现有因果文献中将二者视为同一现象的假设；此外，本文还推导了存在间接空间混杂时的偏差，这一结果此前尚未见于因果推断文献。（4）本文还推导了处理和混杂因素的分布不服从正态分布时的偏差。具体而言，本文考察了二者均服从泊松分布的情形，而此前尚无人对此进行研究。（5）最后，本研究表明，若忽略三种现象中的任何一种（空间干扰、直接空间混杂或间接空间混杂），处理效应估计值就会产生偏差，基于这些估计值作出的决策也将不可靠。

第 1 节介绍了本研究的背景和贡献；第 2 节和第 3 节分别阐述了使我们的估计值成为有效因果估计所需的假设以及分析框架。第 4 节推导了不同情景下的偏差。第 5 节通过模拟评估偏差并展示结果，第 6 节则给出了真实数据应用。第 7 节讨论了研究结果，并简要提出了仅针对空间干扰所致偏差的校正建议，同时阐述了本研究的局限性和未来工作方向。最后一节总结了本研究的发现。

2 因果效应识别假设

为确保直接估计值具有因果意义且可识别，必须满足以下假设。

假设 Y, A, X and 分别表示结局、暴露和协变量，而“a”表示处理实施时的处理水平。

假设 1（正值性）： $Pr[A = a|X = x] > 0$，其中 $f_X(x) > 0$。这意味着研究中的每个单位被分配至暴露组的概率均大于零。
假设2（一致性）： 若 $A = a$，则 $Y^a = Y$。此处的一致性是指，如果某个单位/受试者接受了特定水平的暴露/处理，则观测到的响应必须等于该暴露下的响应。这意味着处理应当定义明确，并且在所有单位/受试者之间保持一致。

假设3（可交换性）： $Y^a \perp\!\!\!\perp A|X$。该假设意味着，所比较各组的基线特征不应存在显著差异（例如，在二元情形中，即治疗组与未治疗组），从而使治疗组与未治疗组之间平均潜在结局的差异可归因于暴露效应。简言之，这仅表示在给定观测协变量的条件下，处理应独立于潜在结果。这也被称为无混杂假设。可交换性假设必须成立，观察性数据才能模拟实验数据中所用的随机化技术，从而进行有效的因果推断。

假设 4（时间顺序）： 这一假设明确区分了因果关系与关联。为了使因果效应估计值有效，我们假设此处的暴露先于结局发生，而非相反。

请注意，在本节中，我们假设无混杂假设成立，但无干扰假设被违反。在后面的章节中，我们还将放宽无混杂假设，使这两个假设都被违反。

3 分析框架

首先，我们从一个不具有空间结构、但包含空间处理的数据生成模型开始，模型如下：

$$\mathbf{M1} \qquad Y(\mathbf{s_i}) = \beta_{\mathbf{0}} + \beta_{\mathbf{a}}\mathbf{A}(\mathbf{s_i}) + \epsilon_{\mathbf{i}} \tag{1}$$

其中，$Y(\mathbf{s_i})$ 表示位置 $\mathbf{s_i} = \mathbf{s1}, ..., \mathbf{s_n} \in \mathbb{R}^{\mathbf{2}}$ 处的响应，$A(\mathbf{s_i})$ 表示在位置 $\mathbf{s_i}$ 随空间变化的处理，ϵ_i 是非空间误差向量，假设其服从均值为 0、方差为 σ^2 的正态分布，即 $\epsilon_i \sim N(0, \sigma^2)$；$\beta_0$ 和 β_a 分别是截距和我们感兴趣的处理效应。为简化起见，我们省略了协变量，仅考虑感兴趣的处理。

由于处理呈现出空间依赖，邻近的处理很可能相互干扰，从而违反独立性假设并引入干扰效应。接下来，我们考虑一个存在干扰但假设不存在混杂因素效应的模型，如下所示：

$$\mathbf{M2} \qquad Y(\mathbf{s_i}) = \beta_{\mathbf{0}} + \beta_{\mathbf{a}}\mathbf{A}(\mathbf{s_i}) + \beta_{\tilde{\mathbf{a}}}\tilde{\mathbf{A}}(\mathbf{s_i}) + \epsilon_{\mathbf{i}}, \tag{2}$$

其中，$\beta_{\tilde{a}}$ 为干扰效应，$\tilde{A} = \Psi A = \sum_{i \neq j} \psi_{ij} A_j$ 是权重$\Psi \in \mathbb{R}^{n \times n}$ 和处理 A（$\mathbf{s_i}$）的函数。权重 Ψ 是一个矩阵，用于表示处理之间的依赖程度，但其对角线元素为 0，因此处理不会与自身相互作用。其他符号的含义与上一节中的定义相同。

我们在模拟中进一步考虑了空间响应但仅存在干扰效应的情形，即当误差为 $N(0, \Sigma)$ 时，其中 $\Sigma = \sigma_\epsilon^2 \Omega_\epsilon(\omega) + \sigma_\epsilon^2 I$ 且 $\Omega_\epsilon(\omega)$ 是带有空间范围参数 ω 的空间相关函数。同时，误差中的空间相关性可能源于某些未测量混杂因素。作为扩展，我们考虑了同时包含空间干扰和空间混杂因素的模型。该模型如下：

$$\mathbf{M3} \qquad Y(\mathbf{s_i}) = \beta_{\mathbf{0}} + \beta_{\mathbf{a}}\mathbf{A}(\mathbf{s_i}) + \beta_{\tilde{\mathbf{a}}}\tilde{\mathbf{A}}(\mathbf{s_i}) + \beta_{\mathbf{u}}\mathbf{U}(\mathbf{s_i}) + \epsilon_{\mathbf{i}} \tag{3}$$

其中 $U(\mathbf{s_i})$ 是一个未知的空间混杂因素，假定其服从均值为 0、方差-协方差为 $\sigma_u^2\Omega_u$ 的正态分布（该假设的违背将在后文中探讨），且 β_u 为直接混杂效应。下标 u 和 ϵ 分别表示与混杂因素和误差相关的参数，并且我们假定 Ω 采用 Mat´ern 相关函数建模，其形式如下：

$$\Omega(d;\omega,v) = \frac{2^{1-v}}{\Gamma(v)}\left(\frac{2\sqrt{v}d}{\omega}\right)^{v} K_v\left(\frac{2\sqrt{v}d}{\omega}\right)$$

其中，ω 是空间范围参数，$d = |s_i - s_j|$、v 是平滑度参数，而K_v 是贝塞尔函数。具体而言，我们考虑了 $v = 0.5$ 的情形，因此得到的指数相关函数为 $\exp\left(\frac{-d}{\omega}\right)$。

—

最后，据我们所知，在因果推断领域关于偏差量化与评估的文献中，研究者仅考虑了这样一种情形：一个位置上的混杂因素会影响同一位置上的响应，即模型形式如下：

$$\mathbf{M4} \qquad Y(\mathbf{s_i}) = \beta_{\mathbf{0}} + \beta_{\mathbf{a}}\mathbf{A}(\mathbf{s_i}) + \beta_{\mathbf{u}}\mathbf{U}(\mathbf{s_i}) + \epsilon_{\mathbf{i}} \tag{4}$$

我们指出，空间混杂效应也可能是间接的，即一个位置上的空间混杂因素会影响另一个位置上的响应。若将空间混杂视为对空间变量的遗漏（因果推断中的定义），那么即使只考虑了直接空间混杂，而实际存在间接空间混杂，估计值仍会有偏差。我们从理论上表明，直接空间混杂与间接空间混杂并不相同，正如空间干扰与直接空间混杂并不相同一样。我们通过模拟对这一点进行了评估。在这种情况下，模型（假设不存在干扰效应）如下：

$$\mathbf{M5} \qquad Y(\mathbf{s_i}) = \beta_{\mathbf{0}} + \beta_{\mathbf{a}}\mathbf{A}(\mathbf{s_i}) + \beta_{\mathbf{u}}\mathbf{U}(\mathbf{s_i}) + \beta_{\tilde{\mathbf{u}}}\tilde{\mathbf{U}}(\mathbf{s_i}) + \epsilon_{\mathbf{i}} \tag{5}$$

˜其中，$\beta_{\tilde{u}}$ 是间接空间混杂效应，$U = \Phi U = \sum_{i\neq j}\phi_{ij}U_j$ 是空间混杂因素 U 上若干权重 $\Phi \in \mathbb{R}^{n\times n}$ 的函数。请注意，这与处理 A 上的权重 Ψ 不同

我们综合了所有结果，并使用完整模型评估偏差，即包含所关注的处理效应、干扰效应、直接空间混杂和间接空间混杂的模型。因此，完整模型如下：

$$\mathbf{M6} \qquad Y(\mathbf{s_i}) = \beta_{\mathbf{0}} + \beta_{\mathbf{a}}\mathbf{A}(\mathbf{s_i}) + \beta_{\tilde{\mathbf{a}}}\tilde{\mathbf{A}}(\mathbf{s_i}) + \beta_{\mathbf{u}}\mathbf{U}(\mathbf{s_i}) + \beta_{\tilde{\mathbf{u}}}\tilde{\mathbf{U}}(\mathbf{s_i}) + \epsilon_{\mathbf{i}} \tag{6}$$

4 偏差推导

4.1 空间干扰导致的偏差量化

如前所述，关于存在直接空间混杂时的偏差，已有文献进行过研究（见 Paciorek (2010)；Page et al. (2017)；Dupont et al. (2023)；Narcisi et al. (2024)；

Khan and Berrett (2024)），但在因果推断中，针对存在空间干扰和间接空间混杂时偏差的量化与评估研究仍然有限。这正是本研究试图填补的空白之一。本节将推导空间和非空间情形下由空间干扰导致的偏差的解析表达式。对于非空间情形，我们采用最小二乘法；对于空间情形，则采用广义最小二乘法。

命题1。 给定公式*2*中的模型形式，并假设只有 β_a 未知，则非空间情形和空间情形下空间干扰导致的偏差分别为 $Bias\left(\hat{\beta}_a^*|A\right) = \beta_{\tilde{a}} \frac{A^T\Psi A}{A^TA}$ 和 $Bias\left(\hat{\beta}_a|A\right) = \beta_{\tilde{a}} \frac{A^T\Omega^{-1}\Psi A}{(A^T\Omega^{-1}A)}$。

证明。 在非空间情形中，考虑上文给出的公式2:

$$Y(\mathbf{s_i}) = \beta_{\mathbf{a}}\mathbf{A}(\mathbf{s_i}) + \beta_{\tilde{\mathbf{a}}}\tilde{\mathbf{A}}(\mathbf{s_i}) + \epsilon_{\mathbf{i}}$$

注意，假设数据已中心化，模型中省略了截距。回顾一下，$(\hat{\beta}_a)$ 的偏差为 $= E\left(\hat{\beta}_a\right) - \beta_a$，而 β_a 的普通最小二乘估计量为$\hat{\beta}_a = \left(A^TA\right)^{-1}A^TY$。

Now;

$$\begin{aligned}
E\left(\hat{\beta}_a^*|A\right) &= \left(A^TA\right)^{-1}A^T\left(\beta_a A(\mathbf{s_i}) + \beta_{\tilde{\mathbf{a}}}\mathbf{E}\left(\tilde{\mathbf{A}}(\mathbf{s_i})|\mathbf{A}\right)\right) \\
&= \beta_a\left(A^TA\right)^{-1}A^TA + \beta_{\tilde{a}}\left(A^TA\right)^{-1}A^TE\left(\tilde{A}|A\right) \\
&= \beta_a + \beta_{\tilde{a}}\left(A^TA\right)^{-1}A^T\tilde{A} \\
&= \beta_a + \beta_{\tilde{a}}\left(A^TA\right)^{-1}A^T\Psi A \\
&= \beta_a + \beta_{\tilde{a}}\frac{A^T\Psi A}{A^TA}
\end{aligned}$$

因此，非空间情形下由空间干扰导致的处理效应偏差为:

$$Bias\left(\hat{\beta}_a^*|A\right) = \beta_{\tilde{a}}\frac{A^T\Psi A}{A^TA} \tag{7}$$

同样，在空间设置中,

$$\begin{aligned}
E\left(\hat{\beta}_a|A\right) &= \left(A^T\Omega^{-1}A\right)^{-1}A^T\Omega^{-1}\left(\beta_a A(\mathbf{s_i}) + \beta_{\tilde{\mathbf{a}}}\mathbf{E}\left(\tilde{\mathbf{A}}(\mathbf{s_i})|\mathbf{A}\right)\right) \\
&= \beta_a\left(A^T\Omega^{-1}A\right)^{-1}A^T\Omega^{-1}A + \beta_{\tilde{a}}\left(A^T\Omega^{-1}A\right)^{-1}A^T\Omega^{-1}E\left(\tilde{A}|A\right) \\
&= \beta_a + \beta_{\tilde{a}}\left(A^T\Omega^{-1}A\right)^{-1}A^T\Omega^{-1}\tilde{A} \\
&= \beta_a + \beta_{\tilde{a}}\left(A^T\Omega^{-1}A\right)^{-1}A^T\Omega^{-1}\Psi A \\
&= \beta_a + \beta_{\tilde{a}}\frac{A^T\Omega^{-1}\Psi A}{A^T\Omega^{-1}A}
\end{aligned}$$

因此，空间设置中由空间干扰引起的偏差如下:

$$Bias\left(\hat{\beta}_a|A\right) = \beta_{\tilde{a}}\frac{A^T\Omega^{-1}\Psi A}{A^T\Omega^{-1}A} \tag{8}$$

□

可以看出，由空间干扰引起的偏差是 A的分布特性、干扰效应的强度以及权重矩阵Ψ选取方式的函数。因此，在模拟部分，我们使用两种不同的权重、处理的两种分布（离散和连续）以及干扰效应的两种不同情形（即干扰效应低于和高于处理效应的情形）评估了干扰效应。

4.1.1 选择权重矩阵，Ψ

如前一节所示并指出，空间干扰所导致的偏差在很大程度上取决于权重矩阵 Ψ。因此，如何选择权重矩阵对于确定偏差的大小至关重要。这里，我们简要介绍在实践中选择权重矩阵的一些方法。

权重可以基于邻接关系，即根据相邻单元来选择。邻接方式可以是车式邻接（邻居共享一条边）或后式邻接（邻居彼此位于垂直或水平方向）。也可以根据位置之间的距离来选择。该距离可以根据专家意见、研究需求或便利性预先设定，并将该区域内距离在此范围内的单元视为邻居。还可以使用反距离；在这种情况下，距离越大，权重越小。例如，假设距离 $d_{ij} = 20km$ 是一个预先设定的距离，那么两个单元 i 和 j 之间的权重可以表示为 $w_{ij} = \frac{1}{d_{ij}^p}$，其中 p 是幂参数，通常取 1 或 2。选择邻居时，也可以使用核函数或平滑衰减函数，例如高斯函数或 Epanechnikov 函数（$w_{ij} = e^{(d_{ij}^2/2)}$）。这种方法比距离的倒数更加灵活。此外，无论单元之间的距离如何,—也可以使用 k近邻（K-NN）来构建权重。每个单元都与其 k 个最近邻相关联。此外，还可以使用基于区块的权重来构建权重矩阵。当单元被划分到不同区域时，这种方法很有用。随后根据组内或组间关系分配权重。最后，也可以依据非空间标准构建权重，例如共同特征（如语言或社会经济地位等）。不过，本研究仅评估了 k近邻和基于距离的权重。

构建基于 k-NN 的权重矩阵

- 步骤 1：对于 k 的选择（k=1、2、3、4.....），找出距离最小的 k 个邻居 e
  对于每个单元 i
- 步骤 2：对于每个单元 i，为已识别的 k 个邻居分配权重，如下

$$w_{ij} = \begin{cases} \frac{1}{k}, & \text{if } j \in \text{k-NN}(i) \\ 0, & \text{otherwise} \end{cases}$$

构建基于距离的权重矩阵

- 步骤 1：构建单位 i 与 j 之间成对距离的完整矩阵 d.

- 步骤 2：确定或计算距离的百分位数阈值或截止值（例如 95%、50%）。

- 步骤 3：对于每一对 $(i,\ j)$，计算权重如下：

$$w_{ij} = \begin{cases} \dfrac{1}{d_{ij}}, & \text{if } d_{ij} \leq \text{threshold and } where \quad i \neq j \\ 0, & \text{otherwise} \end{cases}$$

这两种方法都可以对权重进行行标准化，使得 $w_{ij} = \frac{w_{ij}}{\sum_k w_{ik}}$. 矩阵的每一行之和为 1，以便于比较权重并防止过拟合。

4.2 处理和混杂因素服从泊松分布时的直接空间混杂与偏差

我们关注的是引言部分所定义的、与因果推断相关的空间混杂。我们也已确认，使用空间随机效应并不能解决空间混杂问题（Schnell and Papadogeorgou (2020); Paciorek (2010)）。文献中已有的偏差结果（假设处理和混杂因素均服从正态分布；例如 Paciorek (2010); Khan and Berrett (2024)）会因混杂因素和处理的分布不同而发生变化。作为示例，我们提出一种情形：未观测混杂因素服从泊松分布（例如，在健康研究中，若就诊频率／次数是一个未测量的潜在混杂因素），但具有空间结构，从而引入空间混杂。随后，我们推导了在存在直接空间混杂时，暴露和缺失的混杂因素均具有泊松空间结构的情况下所产生的偏差。

命题*2*。设处理 $A(s_i)$ 和混杂因素 $U(s_i)$ 按空间分布，使得 $A(s_i) = A_a(s_i) + A_u(s_i)$，其中 $A_u(s_i)$ 的分布与 $U(s_i)$ 相同，且两个分量 $A_a(s_i)$ 和 $A_u(s_i)$ 是相互独立的泊松过程，其速率分别为 $\delta_1(s_i)$ 和 $\delta_2(s_i)$。那么，在假设 δ_i 在各个位置 s_i. 上均为常数的情况下，给定 $A(s_i)$ 时，$U(s_i)$ 的条件分布为二项分布$(A, \frac{\delta_2}{\delta_1+\delta_2})$

——

证明。 假设 $U(\mathbf{s}_i)$ 服从泊松分布(δ_2)，且 $A_a(\mathbf{s}_i)$ 服从泊松分布(δ_1)，二者相互独立。那么它们的和

$$A(\mathbf{s}_i) = A_a(\mathbf{s}_i) + U(\mathbf{s}_i)$$

也服从速率为 $\delta_1 + \delta_2$ 的泊松分布，即，

$$A(\mathbf{s}_i) \sim \text{Poisson}(\delta_1 + \delta_2).$$

令 $A = A(\mathbf{s}_i)$ 和 $U = U(\mathbf{s}_i)$。 U 和 A 的联合分布可以用两个相互独立的泊松分布变量的卷积表示：

$$f(U = u, A = a) = f_U(u) \cdot f_{A_a}(a - u),$$

其中

$$f_U(u) = \frac{\delta_2^u e^{-\delta_2}}{u!}, \quad f_{A_a}(a-u) = \frac{\delta_1^{a-u} e^{-\delta_1}}{(a-u)!}.$$

A的边际分布为:

$$f_A(a) = \frac{(\delta_1+\delta_2)^a e^{-(\delta_1+\delta_2)}}{a!}.$$

因此，条件分布为:

$$\begin{aligned} f(U=u \mid A=a) &= \frac{f(U=u, A=a)}{f_A(a)} \\ &= \frac{\frac{\delta_2^u e^{-\delta_2}}{u!} \cdot \frac{\delta_1^{a-u} e^{-\delta_1}}{(a-u)!}}{\frac{(\delta_1+\delta_2)^a e^{-(\delta_1+\delta_2)}}{a!}} \\ &= \frac{\delta_2^u \delta_1^{a-u} e^{-(\delta_1+\delta_2)} \cdot a!}{u!(a-u)!(\delta_1+\delta_2)^a e^{-(\delta_1+\delta_2)}} \\ &= \binom{a}{u} \frac{\delta_2^u \delta_1^{a-u}}{(\delta_1+\delta_2)^a} \\ &= \binom{a}{u} \left(\frac{\delta_2}{\delta_1+\delta_2}\right)^u \left(\frac{\delta_1}{\delta_1+\delta_2}\right)^{a-u}. \end{aligned}$$

因此， $U \mid A \sim$ 服从二项分布$(A, \frac{\delta_2}{\delta_1+\delta_2})$。 □

命题 3. 由非正态空间混杂因素（服从泊松分布）在估计与混杂因素于两个尺度上变化的、服从泊松分布的空间暴露之因果效应时所引起的偏差为 $\beta_u \left[\frac{\delta_2}{\delta_1+\delta_2}\right]$，其中 δ_1、 δ_2 为泊松参数， β_u 为混杂效应。

证明。假设 $U(\mathbf{s_i}) \sim$ 泊松分布$(\delta_2(\mathbf{s_i}))$，且 $A(\mathbf{s_i})$在两个尺度上变化，其中一个依赖于混杂因素（$A_a(\mathbf{s_i})$），另一个独立于混杂因素 $A_u(\mathbf{s_i})$，使得 $A(\mathbf{s_i}) = \mathbf{A_a(s_i)} + \mathbf{A_u(s_i)}$。若 $A_a(\mathbf{s_i}) \sim$ 泊松分布$(\delta_1(\mathbf{s_i}))$与 $A_u(\mathbf{s_i}) \sim$ 泊松分布$(\delta_2(\mathbf{s_i}))$相互独立，其中 $A(\mathbf{s_i}) \sim$ 泊松分布$(\delta_1(\mathbf{s_i}) + \delta_2(\mathbf{s_i}))$，则 $U(\mathbf{s_i})$与 $A(\mathbf{s_i})$条件独立。现在，利用方程 4，其形式为:

$$Y(\mathbf{s_i}) = \beta_\mathbf{a}\mathbf{A}(\mathbf{s_i}) + \beta_\mathbf{u}\mathbf{U}(\mathbf{s_i}) + \epsilon_\mathbf{i}$$

这种情况下，非空间情形中的偏差推导如下:

$$\begin{aligned} E\left(\hat{\beta}_a^* | A\right)_C^{Poi} &= \left(A^T A\right)^{-1} A^T \left(\beta_a A(\mathbf{s_i}) + \beta_\mathbf{u}\mathbf{E}\left(\mathbf{U}(\mathbf{s_i})|\mathbf{A}\right)\right) \\ &= \beta_a + \beta_u \left(A^T A\right)^{-1} A^T E\left(U(\mathbf{s_i})|\mathbf{A}(\mathbf{s_i})\right) \end{aligned}$$

利用命题2，我们有；

$$\begin{aligned} E\left(\hat{\beta}_a^*|A\right)_C^{Poi} &= \beta_a + \beta_u \left(A^T A\right)^{-1} A^T \left[A\frac{\delta_2}{\delta_1+\delta_2}\right] \\ &= \beta_a + \beta_u \left(A^T A\right)^{-1} A^T A \left[\frac{\delta_2}{\delta_1+\delta_2}\right] \\ &= \beta_a + \beta_u \left[\frac{\delta_2}{\delta_1+\delta_2}\right] \end{aligned}$$

因此，非空间情形下的偏差为:

$$Bias\left(\hat{\beta}_a^*|A\right)_C^{Poi} = \beta_u \left[\frac{\delta_2}{\delta_1+\delta_2}\right]$$

同样，空间设置下的偏差推导如下:

$$\begin{aligned} E\left(\hat{\beta}_a|A\right)_C^{Poi} &= \left(A^T\Omega^{-1}A\right)^{-1} A^T\Omega^{-1}\left(\beta_a A(\mathbf{s_i}) + \beta_{\mathbf{u}}\mathbf{E}\left(\mathbf{U}(\mathbf{s_i})|\mathbf{A}\right)\right) \\ &= \beta_a + \beta_u \left(A^T\Omega^{-1}A\right)^{-1} A^T\Omega^{-1} E\left(U(\mathbf{s_i})|\mathbf{A}(\mathbf{s_i})\right) \\ &= \beta_a + \beta_u \left(A^T\Omega^{-1}A\right)^{-1} A^T\Omega^{-1}\left[A\frac{\delta_2}{\delta_1+\delta_2}\right] \\ &= \beta_a + \beta_u \left(A^T\Omega^{-1}A\right)^{-1} A^T\Omega^{-1} A\left[\frac{\delta_2}{\delta_1+\delta_2}\right] \end{aligned}$$

因此，空间设置下的偏差为:

$$Bias\left(\hat{\beta}_a|A\right)_C^{Poi} = \beta_u \left[\frac{\delta_2}{\delta_1+\delta_2}\right]$$

□

请注意，当处理和混杂变量均服从泊松分布时，空间情形和非空间情形下的偏差相同，且与方差-协方差矩阵无关。

4.3 量化直接和间接空间混杂导致的偏差。

如方程 4 所示，直接空间混杂情况下的偏差已得到广泛推导和研究（参见 Paciorek (2010) 和 Khan and Berrett (2024)）。本文重点推导并评估间接空间混杂导致的偏差。

考虑如下方程 5:

$$Y(\mathbf{s_i}) = \beta_a A(\mathbf{s_i}) + \beta_{\mathbf{u}}\mathbf{U}(\mathbf{s_i}) + \beta_{\tilde{\mathbf{u}}}\tilde{\mathbf{U}}(\mathbf{s_i}) + \epsilon_{\mathbf{i}}$$

假设处理 A可分解为两个部分：一部分与混杂因素相关，另一部分是独立的高斯过程。具体而言，

$$A = A_a + A_u, \qquad A \sim N(0, \sigma_a\Omega_a(\omega) + \sigma_u\Omega_u(\omega)),$$

$$U \sim N(0, \sigma_u\Omega_u(\omega)), \qquad Cov(A, U) = \rho\sigma_a\sigma_u\Omega_u(\omega),$$

其中，$\Omega_u(\omega)$ 和 $\Omega_a(\omega)$ 是空间协方差矩阵， ρ是 A 与 U 之间的相关性， σ_a 和 σ_u 分别是它们各自的边际标准差。

$$
\begin{aligned}
E\left(\hat{\beta}_a^*|A\right)_{SC} &= \left(A^TA\right)^{-1}A^T\left(\beta_a A(\mathbf{s_i}) + \beta_{\mathbf{u}}\mathbf{E}\left(\mathbf{U}(\mathbf{s_i})|\mathbf{A}\right) + \beta_{\tilde{\mathbf{u}}}\mathbf{E}\left(\tilde{\mathbf{U}}(\mathbf{s_i})|\mathbf{A}(\mathbf{s_i})\right)\right) \\
&= \beta_a + \beta_u\left(A^TA\right)^{-1}A^TE\left(U|A\right) + \beta_{\tilde{u}}\left(A^TA\right)^{-1}A^TE\left(\tilde{U}|A\right)
\end{aligned}
$$

第二部分 $\beta_u\left(A^TA\right)^{-1}A^TE(U|A)$已由Paciorek (2010)在非空间情境下、并由Khan and Berrett (2024)在空间设置下得到（详见附录）。

此处的主要创新在于第三部分，该部分涉及间接混杂因素$\beta_{\tilde{u}}\left(A^TA\right)^{-1}A^TE\left(\tilde{U}|A\right)$即由间接空间混杂引起的偏差。

命题4. 由间接空间混杂因素 $U = \Phi U$引起的偏差（其中 U和 A 在非空间情境和空间情形下均为高斯过程）由下式给出： $\beta_{\tilde{u}}M(\omega)A^T\Omega^{-1}\phi^T(A-\mu_a 1)\ A^T\Omega^{-1}A$和 $\beta_{\tilde{u}}M(\omega)(A^TA)^{-1}A^T\phi^T(A-\mu_a 1)$ 分别

其中： $M(\omega) = \frac{\rho\sigma_a\sigma_u}{\Omega_u(\omega)\sigma_a^2\Omega_a(\omega)+\sigma_u^2\Omega_u(\omega)}$ 。注意， A 已进行中心化。

证明。从以下内容开始:

$$
\begin{aligned}
\beta_{\tilde{u}}\left(A^TA\right)^{-1}A^TE\left(\tilde{U}|A\right) &= \beta_{\tilde{u}}\left(A^TA\right)^{-1}A^TE\left(\Phi^TU|A\right) \\
&= \beta_{\tilde{u}}\left(A^TA\right)^{-1}A^T\Phi^TE\left(U|A\right) \\
&= \beta_{\tilde{u}}\left(A^TA\right)^{-1}A^T\Phi^T\left(\mu_u\mathbf{1} + \frac{\rho\sigma_{\mathbf{a}}\sigma_{\mathbf{u}}\mathbf{\Omega_u}(\omega)}{\sigma_{\mathbf{a}}^{\mathbf{2}}\mathbf{\Omega_a}(\omega)+\sigma_{\mathbf{u}}^{\mathbf{2}}\mathbf{\Omega_u}(\omega)}(\mathbf{A}-\mu_{\mathbf{a}}\mathbf{1})\right)
\end{aligned}
$$

聚焦于第二个元素，我们有;

$$
\begin{aligned}
\beta_{\tilde{u}}\left(A^TA\right)^{-1}A^TE\left(\tilde{U}|A\right) &= \beta_{\tilde{u}}\left(A^TA\right)^{-1}A^T\phi^T\left(\frac{\rho\sigma_a\sigma_u\Omega_u(\omega)}{\sigma_a^2\Omega_a(\omega)+\sigma_u^2\Omega_u(\omega)}(A-\mu_a 1)\right) \\
&= \beta_{\tilde{u}}M(\omega)(A^TA)^{-1}A^T\phi^T(A-\mu_a 1)
\end{aligned}
$$

其中 $M(\omega) = \rho\sigma_a\sigma_u\Omega_u(\omega)[\sigma_a^2\Omega_a(\omega) + \sigma_u^2\Omega_u(\omega)]^{-1}$，其他所有符号均采用此前定义的通常含义。

因此，在非空间情境中，由间接空间混杂引起的偏差记为 $Bias\left(\hat{\beta}_a^*|A\right)_{Ind.SC}$，其表达式为:

$$
Bias\left(\hat{\beta}_a^*|A\right)_{Ind.SC} = \beta_{\tilde{u}}M(\omega)(A^TA)^{-1}A^T\phi^T(A-\mu_a 1) \tag{9}
$$

同样地，在空间设置中，由间接空间混杂引起的偏差记为$Bias\left(\hat{\beta}_a|A\right)_{Ind.SC}$，其表达式为:

$$
Bias\left(\hat{\beta}_a|A\right)_{Ind.SC} = \frac{\beta_{\tilde{u}}M(\omega)A^T\Omega^{-1}\phi^T(A-\mu_a 1)}{A^T\Omega^{-1}A} \tag{10}
$$

□

Hen中，由直接和间接空间混杂导致的偏差，适用于存在空间和非空间

情境分别由方程（11）和（13）给出：

$$Bias\left(\hat{\beta}_a^*|A\right)_{direct+ind.SC} = \beta_u\rho\frac{\sigma_u}{\sigma_a}\left[\left(A^{*T}A^*\right)^{-1}A^{*T}K\left(A-\mu_a 1\right)\right]_2 + \beta_{\tilde{u}}M(\omega)(A^TA)^{-1}A^T\phi^T(A-\mu_a 1) \tag{11}$$

and

$$Bias\left(\hat{\beta}_a|A\right)_{direct+ind.SC} = \beta_u\rho\frac{\sigma_u}{\sigma_a}\left[\left(A^{*T}\Omega^{-1}A^*\right)^{-1}A^{*T}\Omega^{-1}K\left(A-\mu_a 1\right)\right]_2 \tag{12}$$

$$+\frac{\beta_{\tilde{u}}M(\omega)A^T\Omega^{-1}\phi^T(A-\mu_a 1)}{A^T\Omega^{-1}A} \tag{13}$$

4.4 存在空间干扰和空间混杂（直接和间接）时的偏差

本节放宽了无混杂和无干扰的假设，并推导了这两种现象所导致的偏差。

利用方程 6，存在空间干扰的模型、直接混杂和空间混杂如下：

$$Y(\mathbf{s_i}) = \beta_{\mathbf{a}}\mathbf{A}(\mathbf{s_i}) + \beta_{\tilde{\mathbf{a}}}\tilde{\mathbf{A}}(\mathbf{s_i}) + \beta_{\mathbf{u}}\mathbf{U}(\mathbf{s_i}) + \beta_{\tilde{\mathbf{u}}}\tilde{\mathbf{U}}(\mathbf{s_i}) + \epsilon_{\mathbf{i}}$$

根据我们在方程 7 和 11 中推导偏差的结果，在非空间情境下，对于正态分布的混杂因素和治疗变量，空间干扰和空间混杂所导致的偏差如下：

$$Bias\left(\hat{\beta}_a^*\right)_{SI_SC} = \beta_{\tilde{a}}\left(A^TA\right)^{-1}A^T\Psi A + \beta_u\rho\frac{\sigma_u}{\sigma_a}\left[\left(A^{*T}A^*\right)^{-1}A^{*T}K\left(A-\mu_a 1\right)\right]_2$$
$$+\beta_{\tilde{u}}M(\omega)(A^TA)^{-1}A^T\phi^T(A-\mu_a 1)$$

类似地，在空间设置中，利用方程（8）和（13）的结果，空间设置下空间干扰和空间混杂所导致的偏差如下：

$$Bias\left(\hat{\beta}_a\right)_{IC} = \beta_{\tilde{a}}\left(A^T\Omega^{-1}A\right)^{-1}A^T\Psi A + \beta_u\rho\frac{\sigma_u}{\sigma_a}\left[\left(A^{*T}\Omega^{-1}A^*\right)^{-1}A^{*T}\Omega^{-1}K\left(A-\mu_a 1\right)\right]_2$$
$$+\beta_{\tilde{u}}M(\omega)\left(A^T\Omega^{-1}A\right)^{-1}A^T\Omega^{-1}\phi^T(A-\mu_a 1)$$

请注意，在求干扰效应和间接空间混杂所导致的偏差时，我们通过去掉截距 β_0 对治疗变量进行了中心化；但在求直接空间混杂所导致的偏差时，则保留变量未中心化，以复现 Khan and Berrett (2024) 研究中的结果。

5 模拟研究

在本节中，我们开展一系列模拟实验，以考察在各种空间数据生成机制下估计处理效应时所引入的偏差。本模拟研究围绕四个关键目标展开：

1. 评估由空间干扰引入的偏差，并关注不同权重结构、干扰效应大小以及处理分布的影响。

2. 考察处理与混杂因素的分布假设如何影响偏差的估计。3. 探讨直接和间接空间混杂对偏差的贡献是相似还是不同。4. 量化当三种空间现象——干扰、直接混杂或间接混杂——中的任何一种存在但在分析中未被考虑时所产生的偏差。

5.1 评估指标

本文始终采用均方根误差和覆盖概率作为评估由所有现象产生的模拟估计值的指标。

• **均方根误差（RMSE）：** 均方根误差是一项指标，用于评估估计值与真实参数值之间的平均差距。尽管它对离群值敏感，但会赋予较大的误差更高的权重，而且易于解释。RMSE 越低，表示模型拟合效果越好或估计值越准确。

β 的 RMSE 为：

$$RMSE\ (\hat{\beta}) = \sqrt{Var(\hat{\beta}) + Bias(\hat{\beta})^2}$$

其中，偏差（$\hat{}\beta$） $= E(\hat{}\beta) - \beta$，方差（$\hat{}\beta$） $= E\left[(\hat{}\beta - E(\hat{}\beta))^2\right]$

使用模拟，

$$\hat{RMSE}\ (\hat{}\beta) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\beta}_i - \bar{\hat{\beta}}\right)^2 + \left(\bar{\hat{\beta}} - \beta\right)^2}$$

其中

$$\bar{\hat{\beta}} = E(\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_i$$

$\hat{}N$， β 和 β_i 分别是模拟次数、实际处理效应以及第 i次模拟的估计处理效应。如果估计量无偏，则均方根误差恰好等于方差的平方根。

• **覆盖概率：** 覆盖概率是指在多个样本或试验中重复进行时，真实参数值落入置信区间的次数，或简言之其所占的比例。

其表达式如下：

$$Coverage\ Probability = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(\beta \in C_i)$$

其中，1 是一个指示函数：如果参数 β确实落在置信区间 C_i 和 O 内，则取值为 1；否则取值为 0，而 n 是置信区间的数量。它通常以百分比表示。例如，如果进行 1000 次

模拟，并在每次模拟中为估计参数构造一个 90%置信区间，那么 90% 的覆盖概率意味着，在构造的 1000 个置信区间中，真实参数有 900 次落在区间内。如果覆盖概率高于置信区间的置信水平，则该区间称为保守的；如果低于置信水平，则称为宽松的。

5.2 空间干扰引起的偏差

我们首先根据公式2中的模型模拟数据:

$$Y(\mathbf{s}_i) = \beta_a A(\mathbf{s}_i) + \beta_{\tilde{a}} \tilde{A}(\mathbf{s}_i) + \epsilon_i$$

这里， $A(\mathbf{s}_i)$ 表示空间变化的处理，而 $\tilde{A}(\mathbf{s}_i)$ 则表示空间干扰。我们分别在非空间和空间情景下进行模拟。在非空间情景中，$\epsilon_i \sim N(0,\ 1)$。在空间情景中， ϵ_i 服从均值为零的高斯过程，其协方差结构为指数协方差函数，范围参数为 2。

我们在网格 $[0, 10] \times [0, 10]$ 内随机抽取 100 个空间位置，并根据均值为零、范围参数为 1 的指数协方差函数高斯过程构造$A(\mathbf{s}_i)$。

我们考虑四种情景:

- **情景1：** 生成数据时存在干扰，拟合模型时不考虑干扰。
- **情景2：** 生成的数据包含干扰，拟合模型时考虑了干扰。
- **情景3：** 生成的数据不包含干扰，拟合模型时未考虑干扰。
- **情景 4：** 生成的数据不包含干扰，拟合模型时考虑了干扰。

对于每种情景，我们评估两种情况:

- **情况1：** 处理效应强于干扰效应（$\beta_a > \beta_{\tilde{a}}$）。
- **情形 2：** 处理效应弱于干扰效应（$\beta_a < \beta_{\tilde{a}}$）。

干扰结构使用两种权重类型生成：k近邻（$k = 4$）和基于距离的权重（以第95百分位数为阈值）。每种模拟设置重复迭代 1,000 次。

5.2.1 非空间情形的模拟结果

表1报告了非空间情形下离散处理和连续处理的结果。图4展示了相应的箱线图及不确定性。

*Ta*表 *1*：非空间情境下离散处理和连续处理的模拟结果

离散处理		基于 k-NN 的权重 (k=4)			基于距离的权重 - 95% 阈值		
情形	**情景**	**Bias**	**RMSE**	**覆盖率**	**Bias**	**RMSE**	**覆盖率**
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	0.5054	0.5163	0.436	-0.0125	0.157	0.946
	情景2	-0.0076	0.1702	0.953	0.0054	0.1677	0.949
	情景3	-0.0050	0.1530	0.954	0.0070	0.1549	0.951
$(\beta_a > \beta_{\hat{a}})$	情景 4	-0.0076	0.1702	0.953	-0.0054	0.1677	0.949
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	2.2921	2.3052	0.117	-0.0808	0.2073	0.871
	情景2	-0.0076	0.1702	0.953	0.0054	0.1677	0.949
	情景3	-0.0050	0.1530	0.954	0.0070	0.1549	0.951
$\beta_a < \beta_{\hat{a}}$	情景 4	-0.0076	0.1702	0.953	-0.0054	0.1677	0.949
连续处理		基于 k-NN 的权重 (k=4)			基于距离的权重 - 95% 阈值		
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	0.7338	1.0470	0.894	0.1568	3.6976	0.952
	情景2	-0.0389	0.9834	0.957	0.0902	4.1877	0.950
	情景3	-0.0118	0.8086	0.951	0.1651	3.6958	0.953
$(\beta_a > \beta_{\hat{a}})$	情景 4	-0.0389	0.9834	0.957	0.0902	4.1877	0.950
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	3.3435	3.37432	0.270	0.1278	3.7109	0.951
	情景2	-0.0389	0.9834	0.957	0.0902	4.1877	0.950
	情景3	-0.0118	0.8086	0.951	0.1651	3.6958	0.953
$\beta_a < \beta_{\hat{a}}$	情景 4	-0.0389	0.9834	0.957	0.0902	4.1877	0.950

表1展示了非空间情形下离散处理和连续处理中由空间干扰引起的偏差评估模拟结果。结果显示，情景1（存在干扰但未予以考虑）的平均偏差最高，均方根误差也最大。这表明，当干扰效应存在却被忽略时，会影响对估计值的推断。不过，这种偏差对所用权重的选择较为敏感：采用k近邻选择权重时，偏差高于采用基于距离的权重时的偏差。此外，对于两种权重选择方法，在大多数情形下，当干扰效应大于处理效应时，偏差更高。结果还表明，就权重方法和所考察的干扰效应大小而言，连续处理和离散处理的偏差表现出相似的模式。一个重要区别是，连续处理的偏差高于离散处理。

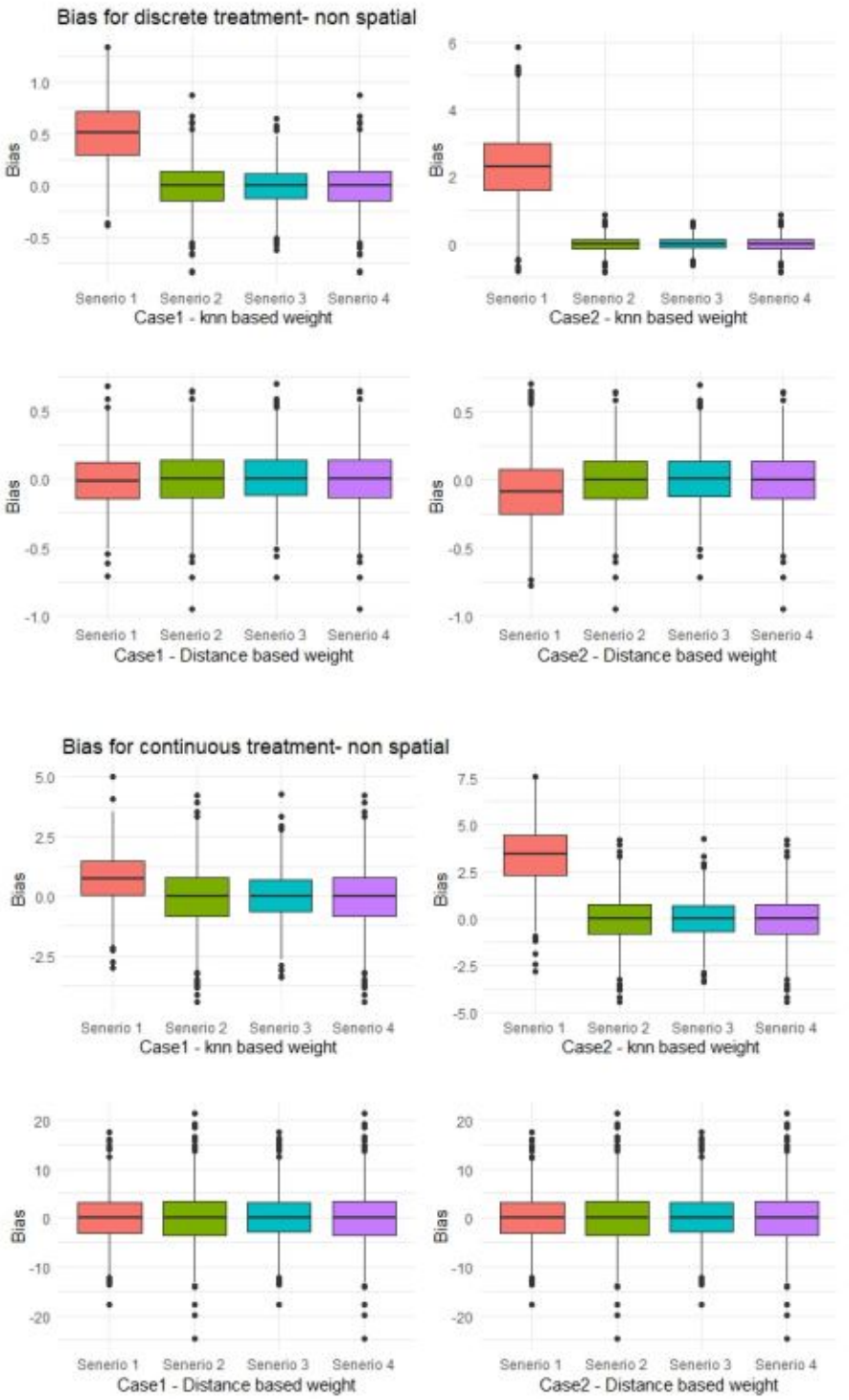

图*4*：箱线图展示了非空间情境下，使用k近邻和基于距离的权重时离散处理与连续处理的偏差

图4展示了所考察的各种情形及所用两种权重下，离散处理和连续处理的箱线图。图中显示，在存在干扰却未加以考虑时，所有考察情形和所用权重下的偏差幅度都是最大的。事实上，图中显示，这种情况下的偏差范围可能为负，也可能为正，从而改变因果解释。当干扰效应高于处理效应（情形 2）且使用基于 k-NN 的权重时，偏差范围最宽。不过，与离散处理相比，连续处理的偏差更大

处理。

5.2.2 空间设置的模拟结果

在空间设置中，我们保留与之前相同的情景和情形，但现在将 ϵ建模为空间相关噪声。离散处理和连续处理的结果见表 2。图5展示了相应的箱线图，以呈现不确定性。

表 *2*：空间设置下离散和连续处理的模拟结果

离散		基于 k-NN 的权重（k=4）			基于距离的权重- 95% 阈值		
情形	**情景**	**Bias**	**RMSE**	**覆盖率**	**Bias**	**RMSE**	**覆盖率**
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	-0.2550	0.284	0.630	-0.0161	0.128	0.945
	情景2	0.0080	0.134	0.939	0.0057	0.144	0.938
	情景3	0.0044	0.127	0.943	0.0044	0.127	0.943
$(\beta_a > \beta_{\hat{a}})$	情景 4	0.0080	0.134	0.939	0.0057	0.144	0.938
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	-1.3700	1.380	0.140	-0.0881	0.148	0.912
	情景2	0.0080	0.134	0.939	0.0057	0.144	0.938
	情景3	0.0044	0.127	0.943	0.0044	0.127	0.943
$\beta_a < \beta_{\hat{a}}$	情景 4	0.0080	0.134	0.939	0.0057	0.144	0.938
连续		基于 k-NN 的权重 (k=4)			基于距离的权重 - 95% 阈值		
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	-0.0887	0.1610	0.830	-0.0141	0.0832	0.936
	情景2	0.00451	0.0829	0.942	0.0063	0.0887	0.933
	情景3	0.00540	0.0827	0.936	0.0054	0.0827	0.936
$(\beta_a > \beta_{\hat{a}})$	情景 4	0.00451	0.0829	0.942	0.0063	0.0887	0.933
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	-0.5720	0.6820	0.687	-0.0859	0.113	0.854
	情景2	0.0045	0.0829	0.942	0.0063	0.0887	0.933
	情景3	0.00540	0.0827	0.936	0.0054	0.0827	0.936
$\beta_a < \beta_{\hat{a}}$	情景 4	0.00451	0.0829	0.942	0.0063	0.0887	0.933

同样，对于离散处理和连续处理，表 2 展示了空间设置下的结果。情景1中的偏差最高，均方根误差也较高。相反，与非空间情境下的结果不同，使用基于距离的权重时，处理效应的向下偏差更大，且数值也高于使用基于 k-NN 的权重时。此外，对于离散处理，情形 2（干扰效应高于处理效应的情形）中的偏差更高。

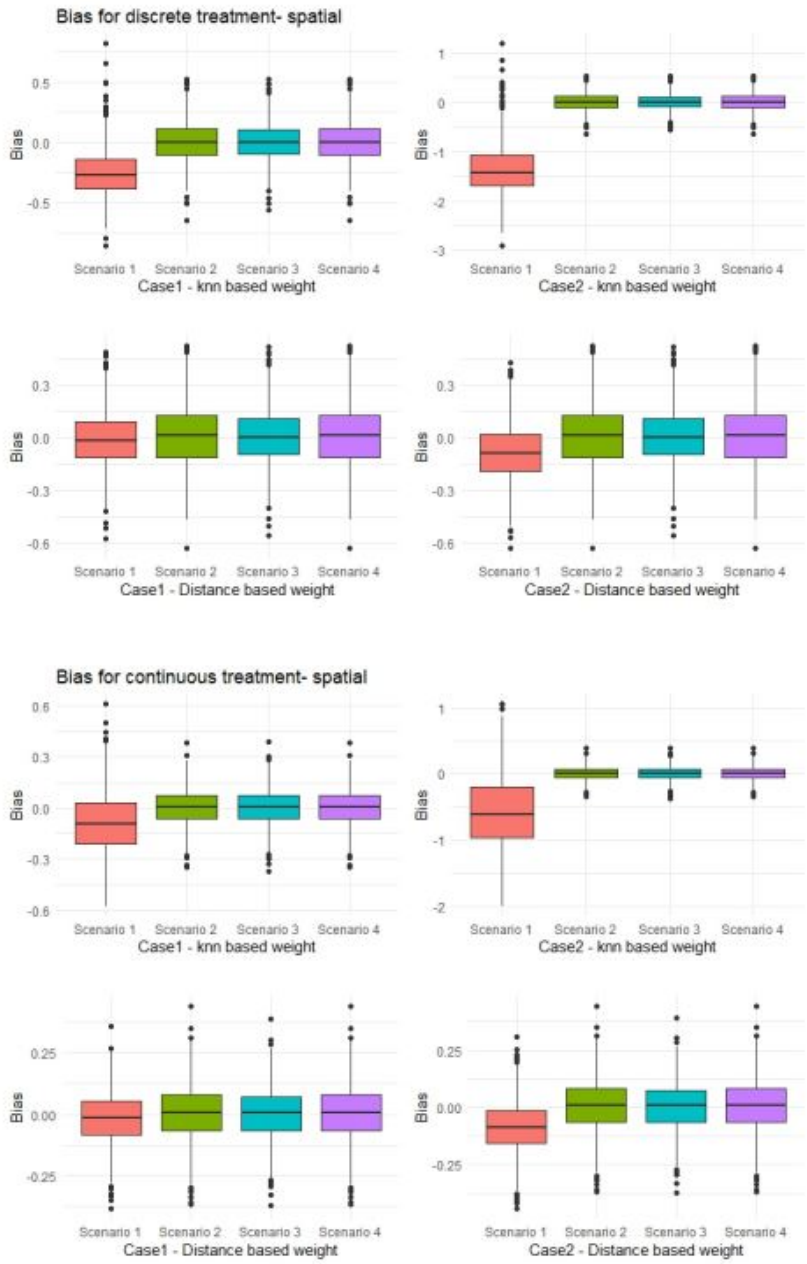

图*5*：展示空间设置中采用 k近邻和基于距离的权重时，离散处理与连续处理偏差的箱线图

图5展示了所考虑的空间设置下离散处理和连续处理的箱线图。对于离散处理，结果呈现出与非空间情境下相似的模式，但由于空间误差的结构，此处偏差的范围更小。情景1、情形 2中的偏差范围最大。此外，对于连续处理，情景1中的偏差最高，但与非空间情境下的结果相反，使用基于距离的权重时，处理效应高于使用 k近邻时的处理效应。令人意外的是，在空间设置下，连续处理的偏差低于离散处理。

5.3 处理和混杂因素具有不同分布时空间混杂因素引起的偏差。

在此，我们考察了处理和混杂因素的分布发生变化时，空间混杂因素引起的偏差。我们的目标是探究当处理和混杂因素均不服从正态分布时，空间混杂导致的偏差如何变化。我们模拟了方程 4 所示形式的模型，如下：

$$Y(\mathbf{s_i}) = \mathbf{2A}(\mathbf{s_i}) + \mathbf{1.5U}(\mathbf{s_i}) + \epsilon_\mathbf{i}$$

其中， $A\mathbf{s_i}$、 $U\mathbf{s_i}$分别为空间处理变量、空间混杂因素和空间随机误差。 $A\mathbf{s_i}$与 $U\mathbf{s_i}$相关，并且我们将 $U\mathbf{s_i}$中的因果假设施加于 $A\mathbf{s_i}$。

我们从网格 $[0,\,10]\times[0,\,10]$内随机抽取 100 个空间位置，并基于均值为零、范围参数为 2 的指数协方差函数的高斯过程构造 $A\mathbf{s_i}$ 和 $U\mathbf{s_i}$。在模拟中，我们改变了处理和混杂因素的分布。具体而言，我们模拟了二者均服从正态分布、泊松分布和二元分布的情形，但 ϵ_i 仅具有空间特性。随后，我们拟合不含混杂因素的模型，即$Y\mathbf{s_i} = \beta_\mathbf{a}\mathbf{A(s_i)} + \epsilon_\mathbf{i}$，以考察不同分布下处理效应所产生的偏差。为确保精度，每次模拟重复迭代 10,000 次。模拟结果列于表3，并见图10。我们以图示方式展示了模拟的正态、泊松分布和二元空间处理变量及空间混杂因素，以及相应的结局，见图 9。

5.3.1 模拟结果

正态分布、泊松分布和二项分布下模拟混杂因素和处理变量及其相应结局的图示或表示见附录B.3。

*T*表 *3*：处理变量和混杂因素服从不同分布时的处理效应偏差

混杂因素和处理变量	Bias	均方根误差	覆盖率
正态 ，正态	0.0049	0.1732	0.93
二元、二元	-0.0005	0.2288	0.894
泊松分布，泊松分布	0.0038	0.1796	0.93

进一步比较混杂因素和处理在不同分布下的偏差，表3显示，当混杂因素和处理的分布发生变化时，偏差也会有所不同。事实上，偏差可能改变因果解释，因为它可能为负（二元情形），也可能为正（正态和泊松分布情形）。当空间处理和空间混杂因素均服从正态分布时，偏差最大，均方根误差最低；其次是空间处理和混杂因素均服从泊松分布的情形。当混杂因素和处理均服从二元分布时，偏差最小，覆盖概率也最低。这里的关键发现是，处理和混杂因素的分布发生变化时，偏差也会显著变化。这进一步支持了我们的论点，并说明了在混杂因素和处理均不服从正态分布时推导偏差的动机。

二元处理和混杂因素、正态处理和混杂因素，以及泊松处理和混杂因素的偏倚箱线图见附录 B.4。

5.4 存在空间干扰、直接和间接空间混杂时的偏差

本小节旨在验证：在因果推断中忽略间接空间混杂或其他存在的效应时产生的偏差。具体而言，我们将探究直接空间混杂与间接空间混杂是否如因果推断中此前假设的那样相同，以支持我们的理论推导。我们模拟了一个完整模型，即包含处理、空间干扰、直接和间接空间混杂的模型（T ＋ I ＋ DSC ＋ISC），其形式如方程 6 所示。因此，模拟模型如下：

$$Y(\mathbf{s_i}) = \mathbf{5A}(\mathbf{s_i}) + \mathbf{3\tilde{A}}(\mathbf{s_i}) + \mathbf{2.5U}(\mathbf{s_i}) + \mathbf{2\tilde{U}}(\mathbf{s_i}) + \epsilon_\mathbf{i}$$

随后拟合了若干嵌套模型，包括包含处理和空间干扰的模型（T + I）、包含处理和直接空间混杂的模型（T ＋ DSC）、包含处理和间接空间混杂的模型（T ＋ ISC）、包含处理、空间干扰和直接空间混杂的模型（T ＋I ＋ DSC）、包含处理、空间干扰和间接空间混杂的模型（T ＋ I ＋ ISC），以及包含处理、直接和间接空间混杂的模型（T ＋ DSC + DSC），以考察缺少这些因素时产生的偏差。

我们在网格内随机抽取 100 个空间位置，并根据均值为零、范围参数为 2 的指数协方差函数所定义的高斯过程构造 $[0,\,10]\times[0,\,10]$ and$A(\mathbf{s_i})$ 和 $U(\mathbf{s_i})$。对于干扰效应，我们还考察了 k近邻和基于距离的权重； ϵ_i 在此处仅为空间变量，并按照第 5.1 节前述方法进行模拟。估计使用 R 中 nlme 软件包的 gls 函数完成。

5.4.1 模拟结果

表 *4*：存在空间干扰、直接和间接空间混杂以及基于 k-NN 和基于距离的权重时的偏差

S/N	模型	基于 k-NN 的权重 (k=4)			基于距离的权重 (95%)		
		Bias	均方根误差	覆盖率	Bias	RMSE	覆盖率
1	T + I	-0.0244	0.3980	0.97	-0.0125	0.2530	0.93
2	T + DSC	0.5890	0.7260	0.54	-0.0078	0.0913	0.91
3	T + ISC	1.1900	1.2100	0.23	-0.0166	0.2450	0.93
4	T + I + DSC	0.0077	0.1690	0.94	0.0151	0.0885	0.96
5	T ＋I ＋ISC	0.0018	0.1990	0.98	-0.0113	0.2400	0.93
6	T + DSC ＋ISC	-0.2560	0.4240	0.45	-0.0061	0.0872	0.92
7	T ＋I ＋DSC ＋ ISC	0.0218	0.0815	0.93	0.0177	0.0832	0.96

表 4 展示了在存在空间干扰、直接空间混杂和间接空间混杂的情况下，使用 k近邻和基于距离的权重时，正态分布的处理和混杂因素所产生的偏差。结果表明，在以下模型中，偏差更高：

忽略干扰效应的模型具有更高的均方根误差，无论采用何种加权方法都是如此。这进一步证实，在存在空间干扰时，进行空间因果推断不应忽视干扰效应。

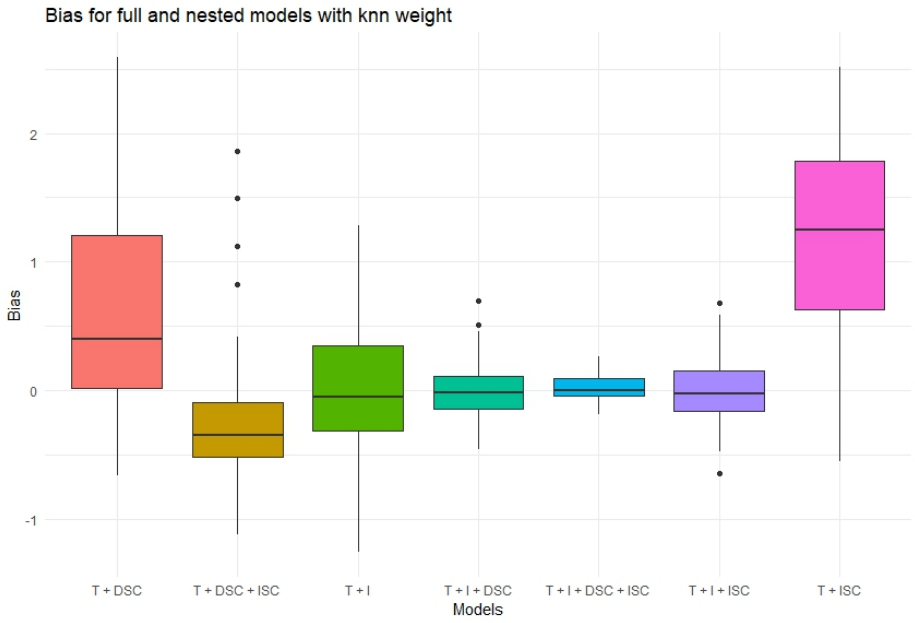

图*6*：展示采用 k近邻权重的完整模型和嵌套模型偏差的箱线图

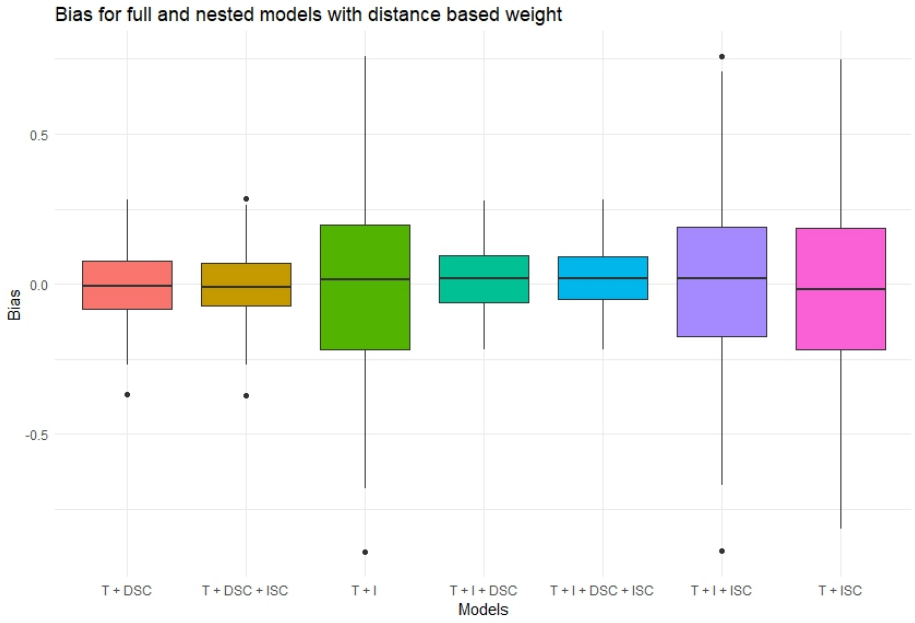

图*7*：展示采用基于距离的权重的完整模型和嵌套模型偏差的箱线图

图6和图7展示了完整模型（包含处理、干扰、直接空间混杂和间接空间混杂的模型）及嵌套模型的偏差。同时，采用基于 k近邻的权重时，结果表明，如图6所示，仅考虑干扰效应和直接空间混杂时，处理效应估计偏差的范围比考虑间接空间混杂时更大。与此相反，采用基于距离的权重时，偏差几乎相同，且落在相同范围内，如

图7所示。因此，我们得出结论：间接空间混杂并非总是可以忽略，其影响取决于混杂因素如何干扰其他位置的响应和处理。

6 应用

为进行说明并区分空间干扰以及直接和间接空间混杂偏倚，我们使用曼彻斯特气候与地表日面板（2018年）中2018年1月1日至31日期间的曼彻斯特气温、地表温度和降水数据进行了应用。数据通过谷歌地球引擎（哥白尼气候变化服务（C3S））获取，开放访问（2022年），并可从（U˜noz Sabater, J., 2019）获取。

气温是指空气的冷热程度，是测量最多的环境因素之一，会影响地表温度等其他气候变量。因此，本研究将气温视为处理变量，将地表温度视为研究结局。此外，降水/降雨可能是一个混杂因素，同时影响气温和地表温度。例如，降雨可能因云量或湿度增加而使气温降低。此外，降雨后土壤湿润度和水分往往会增加，进而不可避免地影响地表温度。因此，降水被视为混杂因素。气温在地面以上2米处测量；为确保稳健性，地表温度在夜间测量，而降水则在白天测量。

我们拟合了T ＋ I、T ＋ DSC、T ＋ ISC、T ＋ I ＋ DSC、T ＋ I ＋ ISC、T+ DSC ＋ ISC 和 T ＋I ＋ DSC ＋ ISC 模型，与表 4 中的做法相同，以考察空间干扰和空间混杂因素同时存在但未加以控制时产生的偏差，并区分直接和间接空间混杂。结果见表5。构建干扰效应权重时，采用了 k近邻和基于距离的权重方法。重点关注处理效应估计值。展示处理效应估计值及其置信区间的图见图8。模型使用 geoR 软件包中的 likfit() 拟合（Ribeiro Jr 等人，2020年）

表*5*：应用数据的处理效应估计值

S/N	**模型**	**基于 k-NN 的权重（k=4）**		**基于距离的权重（50%）**	
	.	**估计值**	**置信区间**	**估计值**	**置信区间。**
1	T + I	0.1786	[0.1680, 0.1891]	0.1426	[0.1396， 0.1459]
2	T + DSC	-0.2068	[-0.2104， -0.2031]	-0.2068	[-0.2104, -0.2031]
3	T + ISC	-0.2045	[-0.2081, -0.2008]	-0.2066	[-0.2102, -0.2029]
4	T + I + DSC	-0.1667	[-0.1764, -0.1569]	-0.1817	[-0.1853, -0.1782]
5	T ＋I ＋ISC	0.1890	[0.1601, 0.2178]	0.4012	[0.3941, 0.4083]
6	T + DSC ＋ISC	-0.2070	[-0.2106, -0.2033]	-0.2072	[-0.2108， -0.2035]
7	T ＋I ＋DSC ＋ ISC	-0.1898	[-0.2020, -0.1776]	-0.2037	[-0.2074, -0.2000]

表5展示了在存在空间干扰、直接空间混 ct

杂和间接空间混杂的情况下，使用 k近邻和基于距离的权重对相关数据进行的处理效应估计值。首先，结果表明，空间干扰和空间混杂是两种不同的现象，因为在两种情景下，估计值均存在显著差异。结果显示，估计值存在显著差异。尤其值得关注的是，当存在直接空间混杂和间接空间混杂但未对其加以控制时，处理效应估计值也会出现显著差异，这表明二者彼此不同。结果发现，同时考虑这三种现象的模型是最佳模型，因为其赤池信息准则最低。

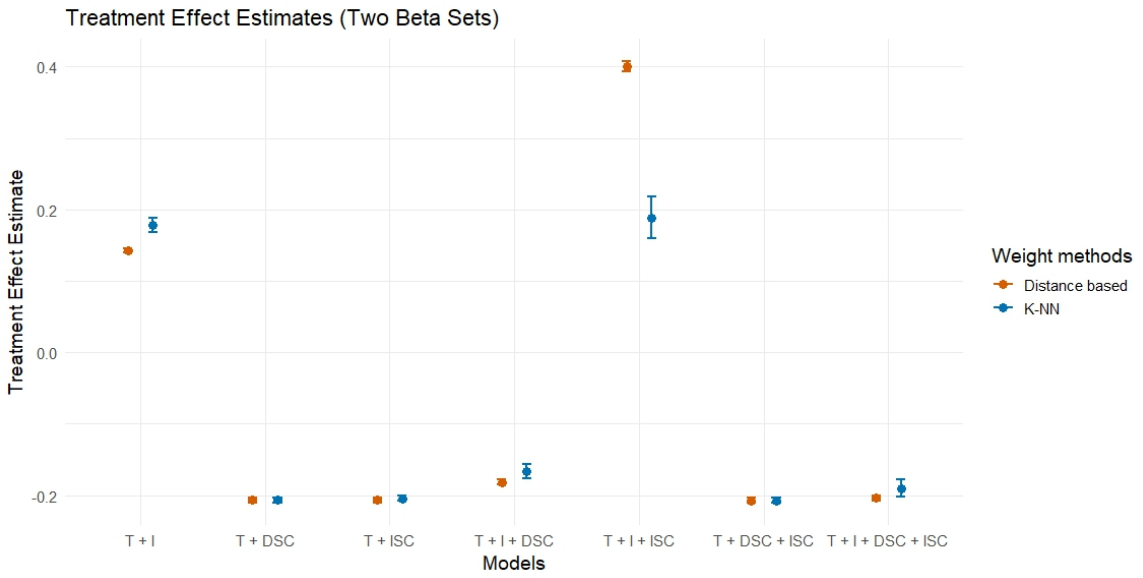

图*8*：带有95%置信区间界限的模型处理效应估计值图

图8展示了处理效应估计值的图形表示。蓝点表示使用基于 k-NN 的权重构建空间干扰和间接空间混杂因素所得到的处理效应估计值，而红点表示使用基于距离的权重得到的处理效应估计值。该图清楚地表明，空间干扰和空间混杂是两种不同的现象，因为两者模型下的估计值彼此不同。此外，该图还显示，直接空间混杂和间接空间混杂也并非相同现象，因为仅考虑其中一种现象时得到的估计值同样存在差异。因此，与现有文献中认为二者是一体的假设相反，空间混杂可以被区分开来。此外，还可以推断出，当所有现象同时存在但忽略其中任何一种时，估计值会产生显著差异。

7 讨论

忽略存在的空间干扰，可能造成严重或轻微的后果，这取决于干扰的大小，而干扰大小由处理之间空间相关性的强度决定。与未能明确何时需要关注空间干扰效应的现有文献不同，本研究发现，存在干扰却未加以考虑时，平均偏差最高，同时均方根误差最大、覆盖概率较低。本研究还发现，即使考虑了干扰效应，处理的分布也会显著影响偏差的大小。

事实上，这有时会改变因果解释。此外，响应是否具有空间性也会造成偏差。在非空间情境下，离散处理的偏差往往高于连续处理；而在空间情境下，情况则恰好相反。简言之，在基于观测空间数据进行因果推断时，需要仔细考察干扰效应的大小、处理的分布，以及响应是否具有空间性。

此外，我们采用 k近邻权重和基于距离的权重构建权重矩阵，该矩阵取决于处理之间的空间相关性。本研究还考察并分析了最近邻数量以及距离阈值或截断值的变化对处理效应估计的影响。模拟结果表明，若权重基于距离构建，则在空间和非空间情境下，干扰效应导致的偏差都会随着阈值距离增大而减小。相比之下，使用 k近邻权重时，偏差会随着最近邻数量增加而增大；但如附录B.1和B.2所示，空间设置中的情况表现不规则。

此外，以往关于空间混杂偏差的研究假设处理和混杂因素均服从正态分布（例如 Khan and Berrett (2024)），但实际情况并非总是如此。我们研究了处理和混杂因素的分布变化所导致的偏差。研究发现，在因果推断中处理空间混杂偏差时，混杂因素和处理的分布特性也是一个关键因素。这里的主要发现是，当处理和混杂因素的分布发生变化时，偏差也会显著变化。具体而言，我们的模拟结果表明，当处理和混杂因素均服从正态分布时，偏差更大。这进一步强化了我们的论点，也说明了研究处理和混杂因素分布均发生变化时偏差的动机，而现有因果推断文献此前尚未对此进行研究。

此外，在存在空间干扰、直接空间混杂和间接空间混杂的情况下，针对正态分布的处理与混杂因素，使用k近邻和基于距离的权重评估偏差后发现：无论采用何种加权方法，在忽略干扰效应的模型中，偏差更高，均方根误差也更大。这进一步印证了：在空间因果推断中，当存在干扰效应时，不应将其忽略。然而，基于k近邻权重的结果表明，仅考虑干扰效应和直接空间混杂时，处理估计值的偏差大于同时考虑三者（干扰效应、直接空间混杂和间接空间混杂）时的偏差。但与之相反，使用基于距离的权重时，两种情形下的偏差几乎相近，且落在同一范围内。因此，我们得出结论：间接空间混杂并非总能被忽略，而是取决于混杂因素如何干扰另一位置的响应与处理。此处的推论是，直接混杂与间接混杂并不相同，这与以往研究相反——那些研究考虑了空间混杂，但将直接空间混杂与间接空间混杂视为同一事物，如Dupont等人（2023）以及Khan和Calder（2022年）的研究。不过，当间接混杂的构建方式使干扰效应变得不那么显著时，二者可以被解释为同一事物，但这

并非总是如此。

最后，对空气表面温度、地表温度和降水的真实数据应用表明，空间干扰与空间混杂并不相同，因此空间混杂也可以拆分为直接空间混杂和间接空间混杂。考虑了这三种现象的模型对数据的拟合效果最佳。总之，当这些现象同时存在而其中任何一种被忽略时，处理效应就会产生偏差，基于此类估计值作出的推断可能并不可靠。

7.1 空间干扰导致的偏差调整

在此，我们根据量化得到的偏差，介绍一种校正空间干扰所致偏差的简单方法。假设 $\beta_{\tilde{a}}$的值已知，可通过从估计得到的有偏处理效应 β_a中减去方程7和8分别针对空间设置和非空间情境推导出的偏差ˆ，从而校正偏差。因此，空间设置和非空间情境下校正后的处理效应分别为:

$$\begin{aligned}\text{Corrected } \beta_a &= \hat{\beta}_a - Bias\left(\hat{\beta}_a^*\right) \\ &= \hat{\beta}_a - \beta_{\tilde{a}}\frac{A^T\Psi A}{A^T A}\end{aligned}$$

and

$$\begin{aligned}\text{Corrected } \beta_a &= \hat{\beta}_a - Bias\left(\hat{\beta}_a\right) \\ &= \hat{\beta}_a - \beta_{\tilde{a}}\frac{A^T\Omega^{-1}\Psi A}{A^T\Omega^{-1}A}\end{aligned}$$

偏倚调整算法

可按以下步骤调整偏差:

(i) 选择一个 Ψ 类别并计算偏差ˆ (ii) 估计 β_aˆ (iii) 计算调整后的 β_a

7.2 局限性与未来研究

我们揭示了在因果推断中忽略空间干扰时所产生偏差的影响，但只考察了两种不同的权重。不过，也可以探索其他形式的权重。此外，我们区分了直接和间接混杂，并量化了空间干扰、直接空间混杂和间接空间混杂同时存在时的偏差，但也没有考虑因果推断中由空间异质性引起的偏差。我们还仅根据量化得到的偏差，提出了一种针对空间干扰偏差的简单偏倚校正方法，却没有提供在空间干扰和空间混杂因素同时存在时精确估计处理效应的方法。未来，我们计划研究因果推断中的异质性，并开发一个框架，以同时处理空间干扰、直接和间接空间混杂因素，从而估计直接效应和中介效应。

8 结论

我们以一种新颖的方式，从有向无环图的视角澄清了有关空间混杂的误解和问题，并厘清了空间干扰、直接空间混杂和间接空间混杂之间的区别。我们还推导并评估了空间干扰和间接空间混杂所引起的偏差。结果表明，权重的选择、分布特性以及干扰效应的大小，是决定空间干扰所致偏差大小的关键因素；而在空间设置中，空间协方差结构也发挥着关键作用。此外，模拟结果表明，如果权重是基于距离构建的，那么在空间和非空间情境下，干扰效应所引起的偏差都会随着阈值距离的增大而减小。相比之下，使用 k近邻权重时，偏差会随着最近邻数量的增加而增大，但空间设置下的情况除外，其变化并不规则。另一方面，我们指出，文献中先前得到的直接空间混杂所致偏差，是基于处理和缺失混杂因素均服从正态分布这一假设。我们指出，如果混杂因素和处理的分布发生变化，偏差也会随之变化。我们通过使用服从泊松分布的混杂因素和处理对此进行了说明。模拟结果和实际数据应用表明，直接空间混杂和间接空间混杂并不相同。如果空间干扰、直接空间混杂和间接空间混杂同时存在，而其中任何一种被忽略，所关注的处理效应就会产生偏差；基于此类估计值进行的推断也可能产生误导。

A 直接空间混杂偏差结果

Paciorek (2010) 指出，非空间情境下由空间混杂导致的偏差，即 $\beta_u \left(A^T A\right)^{-1} A^T E(U|A)$，为：

$$\beta_u \rho \frac{\sigma_u}{\sigma_a} \left[\left(A^{*T} A^*\right)^{-1} A^{*T} K \left(A - \mu_a 1\right) \right]_2$$

类似地，对于空间设置，Khan and Berrett (2024) 推导出由空间混杂导致的偏差为：

$$\beta_u \rho \frac{\sigma_u}{\sigma_a} \left[\left(A^{*T} \Omega^{-1} A^*\right)^{-1} A^{*T} \Omega^{-1} K \left(A - \mu_a 1\right) \right]_2$$

其中 $K = p_c \left(p_c I + (1 - p_c)\Omega(\omega_u)\Omega(\omega_c)^{-1}\right)^{-1}$，$p_c = \frac{\sigma^2}{\sigma_c^2 + \sigma_u^2}$ 和 $A^* = [1A]$。这些推导在第 4.4 节中采用。完整推导请参见 Khan and Berrett (2024)。

B 补充模拟结果

B.1 k = 1、2、3 和 5 时的模拟结果

表 6：非空间情境下离散处理的模拟结果

权重		k近邻（k=1）			k近邻（k=2）		
情形	**情景**	**Bias**	**RMSE**	**覆盖率**	**Bias**	**RMSE**	**覆盖率**
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	-0.0437	0.3178	0.832	0.3054	0.3611	0.705
	情景2	-0.0056	0.1620	0.944	-0.0008	0.1651	0.956
($\beta_a > \beta_{\hat{a}}$)	情景3	-0.0057	0.1600	0.948	0.0068	0.1586	0.950
	情景 4	-0.0056	0.1620	0.944	-0.0008	0.1651	0.956
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	-0.1768	1.2610	0.744	1.3717	1.5190	0.497
	情景2	-0.0056	0.1620	0.944	0.0000	0.1651	0.956
$\beta_a < \beta_{\hat{a}}$	情景3	-0.0057	0.1601	0.948	0.0007	0.1586	0.950
	情景 4	-0.0056	0.1620	0.944	0.0000	0.1651	0.956
		k近邻 (k=3)			k近邻 (k=5)		
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	0.4377	0.4607	0.544	0.5560	0.5641	0.366
	情景2	0.0015	0.1700	0.942	-0.0112	0.1833	0.944
($\beta_a > \beta_{\hat{a}}$)	情景3	0.0038	0.1574	0.953	-0.0053	0.1611	0.953
	情景 4	0.0015	0.1700	0.942	-0.0112	0.1833	0.944
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	1.9564	1.9935	0.257	2.5204	2.5257	0.089
	情景2	0.0015	0.1700	0.942	-0.0112	0.1833	0.944
$\beta_a < \beta_{\hat{a}}$	情景3	0.0038	0.1574	0.953	0.0053	0.1611	0.953
	情景 4	0.0015	0.1700	0.942	-0.0112	0.1833	0.944

表7：非空间情境下连续处理的模拟结果

权重		k近邻（k=1）			k近邻（k=2）		
情形	**情景**	**Bias**	**RMSE**	**覆盖率**	**Bias**	**RMSE**	**覆盖率**
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	-0.0498	0.4753	0.926	0.4447	0.7074	0.905
	情景2	-0.0173	0.4046	0.949	-0.0121	0.5885	0.952
($\beta_a > \beta_{\hat{a}}$)	情景3	-0.0143	0.4027	0.947	-0.0062	0.5566	0.951
	情景 4	-0.0173	0.4046	0.949	-0.0121	0.5885	0.952
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	-0.1742	1.2414	0.796	2.0336	2.1311	0.478
	情景2	-0.0173	0.4046	0.949	-0.0121	0.5885	0.952
$\beta_a < \beta_{\hat{a}}$	情景3	-0.0143	0.4027	0.947	-0.0062	0.5566	0.951
	情景 4	-0.0173	0.4046	0.949	-0.0121	0.5885	0.952
		k近邻 (k=3)			k近邻 (k=5)		
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	0.6472	0.9143	0.881	0.8030	1.1619	0.898
	情景2	-0.0085	0.7701	0.951	-0.0639	0.1693	0.946
($\beta_a > \beta_{\hat{a}}$)	情景3	0.0000	0.6826	0.945	-0.0131	0.9171	0.943
	情景 4	-0.0085	0.7701	0.951	-0.0639	1.1693	0.946
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	2.9122	2.9595	0.302	3.6595	3.6901	0.267
	情景2	-0.0085	0.7701	0.951	-0.0639	0.1693	0.946
$\beta_a < \beta_{\hat{a}}$	情景3	0.0000	0.6826	0.945	-0.0131	0.9171	0.943
	情景 4	-0.0085	0.7701	0.951	-0.0639	1.1693	0.946

表 8：空间设置中离散处理的模拟结果

权重		k近邻（k=1）			k近邻（k=2）		
情形	**情景**	**Bias**	**RMSE**	**覆盖率**	**Bias**	**RMSE**	**覆盖率**
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	-0.4240	0.518	0.506	-0.3680	0.404	0.502
	情景2	0.0077	0.134	0.947	0.0065	0.135	0.946
($\beta_a > \beta_{\hat{a}}$)	情景3	0.0044	0.127	0.943	0.0044	0.127	0.943
	情景 4	0.0077	0.134	0.947	0.0065	0.135	0.946
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	-1.0200	2.680	0.348	-1.8200	1.900	0.236
	情景2	0.0077	0.134	0.947	0.0065	0.135	0.946
$\beta_a < \beta_{\hat{a}}$	情景3	0.0044	0.127	0.943	0.0044	0.127	0.943
	情景 4	0.0077	0.134	0.947	0.0065	0.135	0.946
		k近邻（k=3）			k近邻（k=5）		
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	-0.3030	0.333	0.551	-0.2200	0.251	0.679
	情景2	0.0067	0.136	0.943	0.0083	0.134	0.943
($\beta_a > \beta_{\hat{a}}$)	情景3	0.0044	0.127	0.943	0.0044	0.127	0.943
	情景 4	0.0067	0.136	0.943	0.0083	0.134	0.943
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	-1.580	1.610	0.167	-1.200	1.210	0.133
	情景2	0.0067	0.136	0.943	0.0083	0.134	0.943
$\beta_a < \beta_{\hat{a}}$	情景3	0.0044	0.127	0.943	0.0044	0.127	0.943
	情景 4	0.0067	0.136	0.943	0.0083	0.134	0.943

表9：空间设置下连续处理的模拟结果

权重		k近邻（k=1）			k近邻（k=2）		
情形	**情景**	**Bias**	**RMSE**	**覆盖率**	**Bias**	**RMSE**	**覆盖率**
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	0.2490	0.4040	0.655	-0.0300	0.2060	0.818
	情景2	0.0054	0.0826	0.937	0.0049	0.0825	0.938
($\beta_a > \beta_{\hat{a}}$)	情景3	0.0054	0.0827	0.936	0.0054	0.0827	0.936
	情景 4	0.0054	0.0826	0.937	0.0049	0.0825	0.938
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	2.2600	2.9100	0.522	-0.2860	0.8760	0.774
	情景2	0.0054	0.0826	0.937	0.0049	0.0825	0.938
$\beta_a < \beta_{\hat{a}}$	情景3	0.0054	0.0827	0.936	0.0054	0.0827	0.936
	情景 4	0.0054	0.0826	0.937	0.0049	0.0825	0.938
		k近邻 (k=3)			k近邻 (k=5)		
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	-0.0765	0.1710	0.833	-0.0891	0.148	0.826
	情景2	0.0046	0.0828	0.940	0.0051	0.0829	0.943
($\beta_a > \beta_{\hat{a}}$)	情景3	0.0054	0.0827	0.936	0.0054	0.0827	0.936
	情景 4	0.0046	0.0828	0.940	0.0051	0.0829	0.943
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	-0.5320	0.7110	0.755	-0.5630	0.632	0.647
	情景2	0.0046	0.0828	0.940	0.0051	0.0829	0.943
$\beta_a < \beta_{\hat{a}}$	情景3	0.0054	0.0827	0.936	0.0054	0.0827	0.936
	情景 4	0.0046	0.0828	0.940	0.0051	0.0829	0.943

B.2 距离阈值为 $=$ 90%、80%、75%和50%时的模拟结果

表*10*：非空间情境下离散处理的模拟结果

权重		90%阈值			80%阈值		
情形	**情景**	**Bias**	**RMSE**	**覆盖率**	**Bias**	**RMSE**	**覆盖率**
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	0.0036	0.1703	0.936	0.0855	0.1961	0.892
	情景2	-0.0007	0.1750	0.945	-0.0064	0.1787	0.944
($\beta_a > \beta_{\hat{a}}$)	情景3	0.0000	0.1629	0.947	-0.0039	0.1586	0.953
	情景 4	-0.0007	0.1750	0.945	-0.0064	0.1787	0.944
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	0.0163	0.2805	0.772	0.3980	0.5572	0.492
	情景2	-0.0007	0.1750	0.945	-0.0064	0.1787	0.944
$\beta_a < \beta_{\hat{a}}$	情景3	0.0000	0.1639	0.947	-0.0039	0.1586	0.953
	情景 4	-0.0007	0.1750	0.945	-0.0064	0.1787	0.944
		75% 阈值			50%阈值		
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	0.1527	0.2406	0.829	0.4451	0.4789	0.450
	情景2	0.0046	0.1903	0.936	0.0131	0.2034	0.947
($\beta_a > \beta_{\hat{a}}$)	情景3	0.0074	0.1615	0.943	0.0090	0.1591	0.949
	情景 4	0.0046	0.1903	0.936	0.0131	0.2034	0.947
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	0.6615	0.7741	0.375	1.9719	2.0205	0.159
	情景2	0.0046	0.1903	0.936	0.0131	0.2034	0.947
$\beta_a < \beta_{\hat{a}}$	情景3	0.0074	0.1615	0.943	0.0090	0.1591	0.949
	情景 4	0.0046	0.1903	0.936	0.0131	0.2034	0.947

表*11*：非空间情境下连续处理的模拟结果

权重		90%阈值			80%阈值		
情形	**情景**	**Bias**	**RMSE**	**覆盖率**	**Bias**	**RMSE**	**覆盖率**
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	0.1003	11.7803	0.963	0.5100	8.4193	0.956
	情景2	-0.3877	13.6483	0.953	0.0223	10.7109	0.950
($\beta_a > \beta_{\hat{a}}$)	情景3	0.0510	11.7832	0.963	0.3292	8.4155	0.955
	情景 4	-0.3877	13.6483	0.953	0.0223	10.7108	0.950
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	0.2731	11.7739	0.963	1.1425	8.4925	0.954
	情景2	-0.3877	13.6483	0.953	0.0223	10.7109	0.950
$\beta_a < \beta_{\hat{a}}$	情景 3	0.0510	11.7832	0.963	0.3292	8.4155	0.955
	情景 4	-0.3877	13.6483	0.953	0.0223	10.7108	0.950
		75% 阈值			50%阈值		
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	0.6089	7.6306	0.949	0.7649	4.3006	0.951
	情景2	0.0609	10.0416	0.938	0.2478	6.1634	0.953
($\beta_a > \beta_{\hat{a}}$)	情景3	0.3603	7.6179	0.949	0.1818	4.2446	0.954
	情景 4	0.0609	10.0416	0.938	0.2478	6.1634	0.953
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	1.4787	7.7689	0.951	2.808	5.1064	0.881
	情景2	0.0609	10.0416	0.938	0.2478	6.1634	0.953
$\beta_a < \beta_{\hat{a}}$	情景3	0.3603	7.6179	0.949	0.1818	4.2446	0.954
	情景 4	0.0609	10.0416	0.938	0.2478	6.1634	0.953

表 *12*：空间设置中离散处理的模拟结果

权重		90%阈值			80%阈值		
情形	**情景**	**Bias**	**RMSE**	**覆盖率**	**Bias**	**RMSE**	**覆盖率**
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	-0.0173	0.129	0.945	-0.0186	0.129	0.944
	情景2	0.0056	0.138	0.945	0.0063	0.133	0.948
($\beta_a > \beta_{\hat{a}}$)	情景3	0.0044	0.127	0.943	0.0044	0.127	0.943
	情景 4	0.0058	0.138	0.945	0.0063	0.133	0.948
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	-0.0949	0.153	0.896	-0.1060	0.160	0.882
	情景2	0.0058	0.138	0.945	0.0063	0.133	0.948
$\beta_a < \beta_{\hat{a}}$	情景3	0.0044	0.127	0.943	0.0044	0.127	0.943
	情景 4	0.0058	0.138	0.945	0.0063	0.133	0.948
		75% 阈值			50%阈值		
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	-0.0196	0.129	0.945	-0.0286	0.131	0.946
	情景2	0.0058	0.134	0.945	0.0023	0.132	0.944
($\beta_a > \beta_{\hat{a}}$)	情景3	0.0044	0.127	0.943	0.0044	0.127	0.943
	情景 4	0.0058	0.134	0.945	0.0023	0.132	0.944
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	-0.115	0.163	0.879	-0.182	0.208	0.775
	情景2	0.0058	0.134	0.945	0.0023	0.132	0.944
$\beta_a < \beta_{\hat{a}}$	情景3	0.0044	0.127	0.943	0.0044	0.127	0.943
	情景 4	0.0058	0.134	0.945	-0.0023	0.132	0.944

表*13*：空间设置下连续处理的模拟结果

权重		90%阈值			80%阈值		
情形	**情景**	**Bias**	**RMSE**	**覆盖率**	**Bias**	**RMSE**	**覆盖率**
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	-0.0143	0.0834	0.932	-0.0152	0.0847	0.935
	情景2	0.0063	0.0852	0.933	0.0057	0.0842	0.932
$(\beta_a > \beta_{\hat{a}})$	情景3	0.0054	0.0827	0.936	0.0054	0.0827	0.936
	情景 4	0.0063	0.0852	0.933	0.0057	0.0842	0.932
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	-0.0962	0.121	0.825	-0.1170	0.135	0.800
	情景2	0.0063	0.0852	0.933	0.0057	0.0842	0.932
$\beta_a < \beta_{\hat{a}}$	情景3	0.0054	0.0827	0.936	0.0054	0.0827	0.936
	情景 4	0.0063	0.0852	0.933	0.0057	0.0842	0.932
		75% 阈值			50%阈值		
$\beta_a = 8, \beta_{\hat{a}} = 2$	情景1	-0.0151	0.0852	0.938	-0.0238	0.0884	0.922
	情景2	0.0055	0.0848	0.940	0.0033	0.0843	0.936
$(\beta_a > \beta_{\hat{a}})$	情景3	0.0054	0.0827	0.936	0.0054	0.0827	0.936
	情景 4	0.0055	0.0848	0.940	0.0033	0.0843	0.936
$\beta_a = 3, \beta_{\hat{a}} = 9$	情景1	-0.1230	0.1410	0.787	-0.1970	0.207	0.642
	情景2	0.0055	0.0848	0.940	0.0033	0.0843	0.936
$\beta_a < \beta_{\hat{a}}$	情景3	0.0054	0.0827	0.936	0.0054	0.0827	0.936
	情景 4	0.0055	0.0848	0.940	0.0033	0.0843	0.936

B.3 正态、泊松分布和二项分布下模拟的混杂因素与处理及其对应结局的图

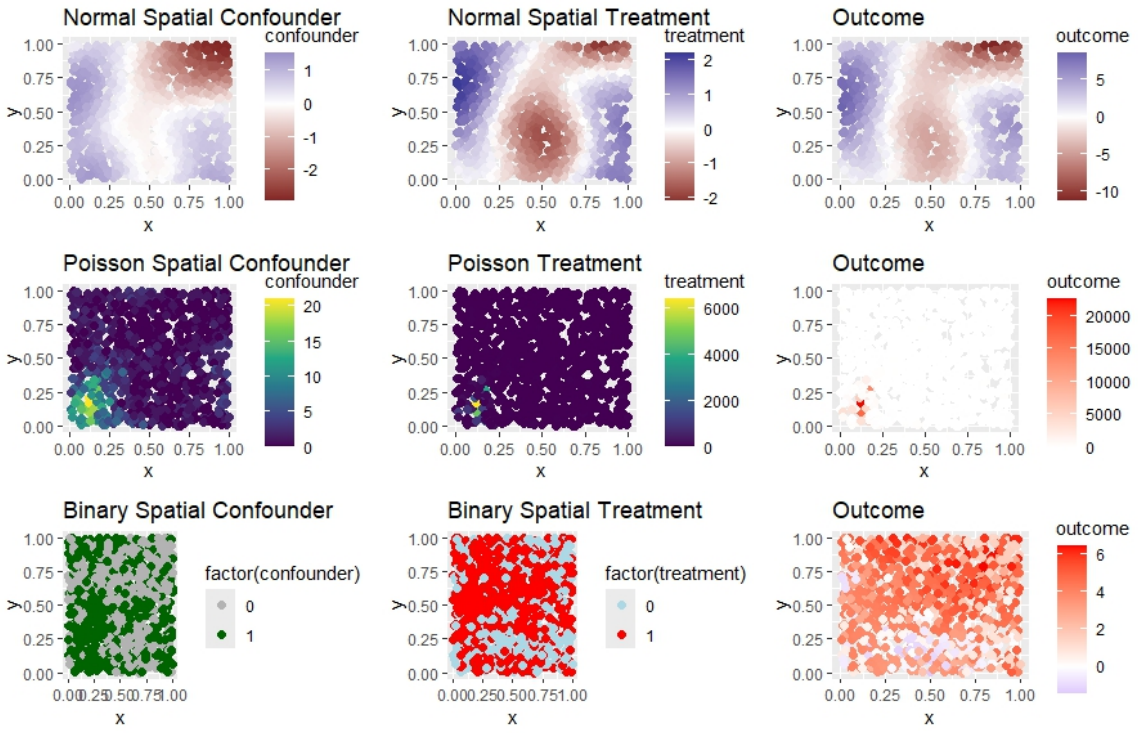

图*9*：正态、泊松分布和二元空间处理及其对应结局的空间混杂因素

B.4 二元处理与混杂因素、正态处理与混杂因素以及泊松处理与混杂因素的偏倚箱线图

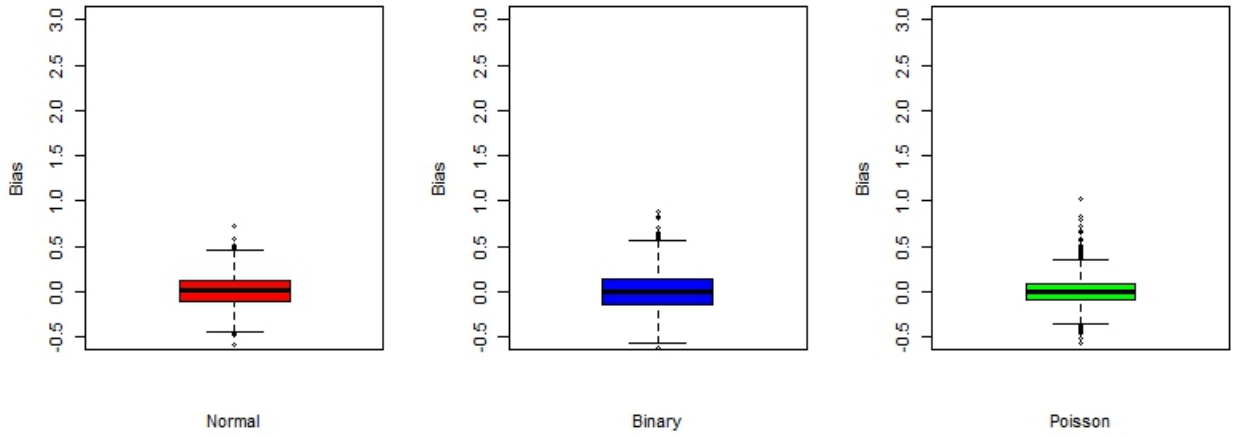

图*10*：正态处理、泊松处理、二元处理及混杂因素的偏差

参考文献

Athey, S.、Eckles, D.和Imbens, G. W.（2018年）。网络干扰的精确p值。《美国统计学会杂志》，113(521):230–240。Banerjee, S.、Carlin, B. P.和Gelfand, A. E.（2014年）。空间数据的层次建模与分析。Chapman and Hall/CRC。Basse, G. W.、Feller, A.和Toulis, P.（2019年）。干扰下因果效应的随机化检验。*Biometrika*，106(2):487–494。Bobb, J. F.、Cruz, M. F.、Mooney, S. J.、Drewnowski, A.、Arterburn, D.和Cook, A. J.（2022年）。在具有个体水平暴露的流行病学研究中考虑空间混杂：一种暴露惩罚样条方法。《英国皇家统计学会会刊*A*辑：社会统计》，185(3):1271–1293。哥白尼气候变化服务（C3S）)（2022年）。1950年至今的ERA5-Land逐小时数据。哥白尼气候变化服务（*C3S*）气候数据存储库（*CDS*）*DOI: 10.24381/cds.68d2bb30*（访问日期：*2025*年*10*月*8*日）。Dupont, E.、Marques, I.和Kneib, T.（2023年）。揭秘空间混杂。*arXiv*预印本*arXiv:2309.16861*。Dupont, E.、Wood, S. N.和Augustin, N. H.（2022年）。Spatial+:一种应对空间混杂的新方法。《生物计量学》，78(4):1279–1290。Giffin, A.（2020年）。面向环境与公共卫生应用的空间数据因果推断方法。北卡罗来纳州立大学。Giffin, A.、Reich, B.、Yang, S.和Rappold, A.（2023年）。存在空间干扰时因果推断的广义倾向评分方法。《生物计量学》，79(3):2220–2231。

Giffin, A.、Reich, B. J.、Yang, S.和Rappold, A. G.（2021年）。工具变量、空间混杂与干扰。*arXiv*预印本 *arXiv:2103.00304*。Gilbert, B.、Datta, A.、Casey, J. A.和Ogburn, E. L.（2021年）。空间混杂的因果推断框架。*arXiv*预印本 *arXiv:2112.14946*。Gilbert, B.、Datta, A.、Casey, J. A.和Ogburn, E. L.（2024年）。空间混杂的因果推断框架。*arXiv*预印本 *arXiv:2112.14946v12*。Hern´an, M. A.和Robins, J. M.（2010年）。因果推断。Imbens, G. W.（2020年）。因果推断的潜在结果与有向无环图方法：对经济学实证实践的意义。《经济文献杂志》，58(4):1129–1179。Keller, J. P.和Szpiro, A. A.（2020年）。空间混杂调整的尺度选择。《英国皇家统计学会会刊*A*辑：社会统计》，183(3):1121–1143。Khan, K.和Berrett, C.（2024年）。重新思考空间线性混合模型中的空间混杂。*arXiv*预印本 *arXiv:2301.05743*。Khan, K.和Calder, C. A.（2022年）。受限空间回归方法：对推断的启示。《美国统计学会杂志》，117(537):482–494。Narcisi, M.、Greco, F.和Trivisano, C.（2024年）。论线性回归模型中混杂的影响：一种基于二次型理论的方法。《环境与生态统计》，31(2):433–461。Paciorek, C. J.（2010年）。尺度对空间混杂偏差及空间回归估计量精度的重要性。《统计科学：数理统计学会综述期刊》，25(1):107。Page, G. L.、Liu, Y.、He, Z.和Sun, D.（2017年）。存在空间混杂时空间线性模型的估计与预测。《斯堪的纳维亚统计学杂志》，44(3):780–797。Papadogeorgou, G.、Choirat, C.和Zigler, C. M.（2019年）。利用距离调整倾向评分匹配对未测量的空间混杂进行调整。《生物统计学》，20(2):256– 272。Papadogeorgou, G.和Samanta, S.（2023年）。存在未测量混杂与干扰时的空间因果推断。*arXiv*预印本 *arXiv:2303.08218*。Prates, M. O.、Dey, D. K.、Willig, M. R.和Yan, J.（2015年）。变换高斯马尔可夫随机场与物种丰度的空间建模。《空间统计学》，14:382–399。Reich, B. J.、Hodges, J. S.和Zadnik, V.（2006年）。残差平滑对疾病制图模型中固定效应后验分布的影响。《生物计量学》，62(4):1197–1206。Ribeiro Jr, P. J.、Diggle, P. J.、Ribeiro Jr, M. P. J.和Imports, M.（2020年）。“geor”软件包。*R*软件包版本*1.8*，1。

Rubin, D. B.（1980年）。实验数据的随机化分析：关于 Fisher 随机化检验的评论。《美国统计学会杂志》，75(371):591–593。Schnell, P. M. 和 Papadogeorgou, G.（2020年）。估计超市可达性对心血管疾病死亡的影响时缓解未观测空间混杂。《应用统计年鉴》，14(4):2069 – 2095。Sobel, M. E.（2006年）。住房流动随机研究证明了什么？存在干扰时的因果推断。《美国统计学会杂志》，101(476):1398–1407。Tchetgen, E. J. T. 和 VanderWeele, T. J.（2012年）。论存在干扰时的因果推断。医学研究中的统计方法，21(1):55–75。Thaden, H. 和 Kneib（2018年）。用于处理空间混杂的结构方程模型。美国统计学家，72(3):239–252。Torrione, E. D., Arduini, T. 和 Forastiere, L.（2024年）。存在干扰时的回归断点设计。*arXiv* 预印本 *arXiv:2410.02727*。Urdangarin, A., Goicoa, T. 和 Ugarte, M. D.（2023年）。评估克服空间混杂的近期方法。*Revista Matem´atica Complutense*，36(2):333–360。U˜noz Sabater, J.（2019年）。1950年至今的 ERA5-Land 月平均数据。哥白尼气候变化服务（*C3S*）气候数据存储库（*CDS*）*DOI*：*10.24381/cds.68d2bb30*（访问日期：*2025*年*10*月*8*日）。Zigler, C. M. 和 Papadogeorgou, G.（2020年）。存在干扰时的二分因果推断。《统计科学：数理统计学会评论期刊》，36(1):109。