arXiv:2602.02277v1 [stat.ME] 2 Feb 2026

一种用于人群层面的流行病学风险评估的空间随机森林算法

Duncan Lee*格拉斯哥大学数学与统计学院与Vinny Davies格拉斯哥大学数学与统计学院

2026年2月3日

摘要

空间流行病学识别导致群体层面疾病风险升高的驱动因素，所依据的是区域单元层面的疾病计数、暴露因素和已知混杂因素。推断通常采用泊松回归模型，该类模型包含线性/加性回归部分，并通过一组空间自相关随机效应来处理未测量混杂。这一方法要求预先指定混杂变量交互作用及其与疾病风险的函数关系，而不是从数据中学习得到。因此，本文提出SPAR-Forest- ERF算法，首次将用于捕捉非线性及交互性混杂因素-反应效应的随机森林，与能够在完整不确定性量化下估计可解释的暴露-响应函数（ERF）的贝叶斯空间自相关模型相融合。在方法论上，我们通过在机器学习模型与统计模型之间传播不确定性，扩展了现有以预测为情境的方法，提出旨在确保主要推断目标稳定性的新停止准则，并纳入一系列不同的暴露-反应函数，以实现最大的模型灵活性。该方法论的提出源于一项新研究，该研究利用最近发布的2022年全国人口普查数据，量化空气污染浓度对苏格兰自评健康的影响。

关键词：地理关联研究、混合堆叠模型、空间自相关。

*通讯作者邮箱 - Duncan.Lee@glasgow.ac.uk

1 引言

空间流行病学的一个重要领域是量化暴露因素对疾病风险的影响，例如空气污染对新冠肺炎住院率的影响（Lee 等，2022），以及绿地可及性与心理健康之间的关联（Nutsford 等，2013）。在这些研究中，研究区域被划分为 K 互不重叠的小型区域单元，每个单元的数据汇总了居住于该区域的人群中的疾病病例总数、平均暴露水平及相关混杂因素。由于这类研究立足于群体层面，无法估计个体层面的因果关系；但其优势在于地理覆盖范围广，涵盖大量风险人群，而且所需的非披露性数据易于获取，因此实施起来快速且成本低廉。

贝叶斯层级模型常用于从此类数据中进行推断，其均值函数依赖人口偏置项，以控制人口统计特征、关注的暴露因素、一组混杂因素以及空间自相关随机效应。后者用于捕捉数据中未被协变量解释的剩余空间自相关，这可能是未测量混杂等问题所致。条件自回归（CAR，Besag 等，1991）先验分布通常赋予这些随机效应；同时，假设混杂因素与疾病风险之间呈线性或简单加性关系。该建模框架能够提供具有完整不确定性量化（UQ）的、易于解释的暴露-反应函数（ERF），并能够捕捉残差空间自相关。然而，混杂变量交互作用及其与疾病风险之间的函数关系必须由分析人员预先指定，这需要大量假设，也使得模型无法从数据中学习这些模式。

随机森林和神经网络等机器学习（ML）算法能够捕捉复杂且相互作用的混杂变量-响应关系，但它们假设观测值之间相互独立，因此不适用于空间数据。它们提供的推断结果也较为有限，例如变量重要性图；与统计模型自然产生的、基于相对风险的暴露-反应函数相比，这类结果在流行病学上的相关性低得多。许多作者已将机器学习模型扩展至空间数据领域，近期综述见 Wikle （2023年）和 Patelli 等 （2024年）。最简单的方法是用基函数等空间平滑量表示数据中的空间自相关（Chen 等，2024），或者在神经网络模型中使用空间滞后协变量／节点（Zhu 等，2022）。另一种方法是将空间自相关嵌入机器学习模型的优化函数中（Saha 等，2023、Zhan 和 Datta，2025），或者在应用机器学习算法之前进行去相关预处理（Heaton 等，2024）。然而，这些方法基于机器学习模型，因此推断能力有限，包括无法提供具有完整不确定性量化的、基于相对风险的暴露-反应函数。

因此，我们在此采用残差学习方法，使用堆叠建模策略，依次将机器学习和空间平滑模型应用于数据，并将一个模型的输出作为下一个模型的输入。该方法兼具机器学习算法的灵活性，可捕捉非线性和相互作用的混杂变量-响应关系；同时也利用贝叶斯空间平滑模型捕捉残差自相关，并提供具有完整不确定性量化的、可解释的暴露-反应函数。Hengl 等人提出了首个此类两阶段方法（Hengl 等人（2015），该方法在第一阶段结合随机森林模型，并在第二阶段进行平滑

模型，对随机森林的残差进行克里金插值。然而，该方法在假设数据相互独立的情况下应用随机森林；众所周知，存在空间自相关时，这会导致推断效果欠佳。因此，MacBride 等人（2025年）和Figueira 等人（2025年最近提出了迭代估计算法，依次对数据拟合随机森林和空间自相关模型组件，并分别根据优化模型的样本外预测性能和库尔贝克-莱布勒散度来确定终止条件。

现有的空间机器学习文献几乎完全聚焦于在未测量位置预测空间过程，而这并非流行病学关联研究中的推断目标。因此，本文提出 `SPAR-Forest-ERF` 算法，这是首个连贯的空间机器学习框架，旨在存在具有非线性和交互效应的已知混杂因素以及残差空间自相关的情况下，估计一系列暴露-反应函数。该算法扩展了MacBride等人（2025年）的迭代残差学习策略，并通过三项关键进展加以实现。首先，它引入了一种基于暴露-反应函数的新型停止规则。其次，它在两个模型组成部分之间传播不确定性，从而提供比Figueira等人（2025年）更为全面的不确定性量化。最后，它估计一系列线性和非线性暴露-反应函数，以最大限度地提高模型灵活性。这一方法论源于一项新研究，该研究利用最近发布的2022年全国人口普查数据，考察空气污染浓度对苏格兰自评健康的影响。这些数据在第2节中加以描述，方法论的发展则在第3节中概述。评估模型性能的模拟研究在第4节中加以描述，而作为动机的研究结果在第5节中呈现，随后在第6节给出结论。

2 动机研究

最近一次苏格兰人口普查于2022年进行，这是一项每10年对全体苏格兰人口开展的调查。本研究利用这些人口普查数据，调查苏格兰空气污染与自评健康之间的关联，数据可从 https://www.scotlandscensus.gov.uk/获取。继德国的近期研究（Liao等人，2025年）和英国的近期研究（Abed Al Ahad等人，2023年）之后，本研究是苏格兰此类研究中的首例。苏格兰被划分为 $K = 6{,}972$ 数据区（数据区），平均包含738人（$1st$ 百分位数 - 396， $99th$ 百分位数 - 2,032 ），并且是数据可公开获取的最小面状单元地理。

2.1 疾病数据

疾病结局为自评总体健康，数据来自人口普查问题总体而言，您的健康状况如何？。可选回答包括非常好、良好、一般、差和非常差；数据包含每个数据区中各类别的回答人数。本文考虑两种健康结局来评估结论的稳健性：(i) 认为自己健康状况不良的人数（回答差或非常差）；以及 (ii) 认为自己健康状况良好的人数（回答良好 或非常好）。这些数据区级人数取决于当地人口的年龄-性别人口结构。为此，我们采用间接标准化，计算每个数据区健康状况不良／良好的人数期望值。

标准化患病率比（SMR）是用于探索各数据区健康状况不良／良好风险的指标，通过观测人数除以相应的期望人数计算得出。SMR为1表示苏格兰平均风险，高于／低于1的数值分别表示风险高于／低于全苏格兰平均水平。补充材料中的表1汇总了这些数据区级SMR值的分布。结果显示，健康状况不良的分布变异性远大于健康状况良好的分布，这是因为后者基于的人数多得多。最后，为了展示SMR的空间模式，补充材料第1.1节还给出了苏格兰两座最大城市（爱丁堡和格拉斯哥）的SMR值地图。

2.2 暴露数据

我们重点关注二氧化氮（NO_2）以及空气动力学直径小于或等于 $10\mu m$ 的颗粒物（PM_{10}）和 $2.5\mu m$（$PM_{2.5}$）。在本研究中关注这些污染物，是因为它们是苏格兰26个空气质量管理区中大多数区域的主要原因，详情请见 http://www.scottishairquality.scot/laqm/aqma。这些污染物由稀疏的固定站点监测器网络测量；2022年该网络约有85个空间位置，主要分布在四大城市中心（见补充材料中的图3）。由于该网络对 $K = 6{,}972$个构成该国的数据区的空间覆盖率极差，我们使用污染气候制图（PCM）模型提供的2022年估计年均浓度来估算污染物水平（ https://uk-air.defra.gov.uk/data/pcm-data）。该模型是为环境、食品与乡村事务部（DEFRA）开发的，并用于按照欧盟指令2008/50/EC的要求监测空气污染物浓度。

PCM 估计值以 $1km^2$ 网格提供，其空间位置与形状不规则的数据区并不吻合。我们通过空间加权平均来解决这种空间支撑变换：将质心位于每个数据区内的所有网格浓度的平均值，作为该数据区的估计浓度。对于不包含任何网格质心的数据区，则使用最近网格的估计浓度。补充材料中的表1汇总了这些数据区层级浓度的分布，显示苏格兰的污染水平总体较低，中位数分别为 $6.69\mu gm^{-3}$ (NO_2)、$4.66\mu gm^{-3}$ ($PM_{2.5}$) 和 $8.73\mu gm^{-3}$ (PM_{10})。作为参考，英国空气质量战略规定，苏格兰这些污染物的年平均浓度目标应分别低于 $40\mu gm^{-3}$ (NO_2)、$10\mu gm^{-3}$ ($PM_{2.5}$) 和 $18\mu gm^{-3}$ (PM_{10})。与预期一致，三种污染物在城市地区的浓度均高于农村地区；同时，受来自欧洲大陆的跨境污染影响，东海岸的 $PM_{2.5}$ 和 PM_{10} 浓度也较高。

2.3 混杂因素数据

本研究中可用的主要混杂因素汇总于补充材料的表1。其中最突出的是社会经济剥夺，在苏格兰由苏格兰多重剥夺指数（SIMD，https://simd.scot/）衡量。SIMD是一个由27项独立指标组成的综合指数，其中与健康相关的

指标被剔除，因为用健康数据对健康数据进行建模属于循环论证。此外，与犯罪率、到校率以及连续型学业成绩评分相关的指标也因缺失值数量过多（分别为501、567和189）而被剔除。最后，对于绝对相关系数超过0.85的任意一对指标，剔除其中的一个变量，最终得到12项指标，汇总于补充材料的表2。

我们还获取了有关族裔的数据，具体而言，是每个 DZ 中属于以下族裔群体的人口比例：（i）白人；（ii）印度人／巴基斯坦人／孟加拉国人；（iii）华人；（iv）非洲裔；以及（v）加勒比裔或黑人。由于第一类变量与其余变量存在共线性，因此将其剔除；这是因为在所有 DZ 中，这些变量的总和都接近 100%。下一组协变量涉及 DZ 的城乡属性，并通过人口密度（每 1km^2 的人数）以及两个二元指示变量来衡量，分别表示每个 DZ 是否属于城市地区（人口超过 10,000 人的聚居地的一部分）或农村地区（人口少于 3,000 人的聚居地的一部分）。最后一个混杂因素用于衡量孤独感，我们使用人口普查数据进行近似，即统计每个 DZ 中居住在单人家庭中的人口比例。

2.4 探索性分析

我们使用相关性分析考察了疾病风险（$\ln$(标准化死亡比)）与协变量之间的关系（见补充材料第1.2节），发现：（i）由于污染物之间存在共线性，应将其纳入不同的模型；（ii）最终混杂因素集合中不太可能还存在共线性。随后，我们分别拟合简单的泊松对数线性模型和随机森林（以对数标准化死亡比为尺度）对数据进行建模，考察两种健康结局中是否存在残差空间自相关。随机森林的调优参数通过最小化袋外误差的均方根误差来选定，详情见第3节。我们采用 Moran's I置换检验评估空间自相关（Moran, 1950年），该检验基于由共享边界规则定义的邻接矩阵，并进行了10,000次随机置换。对于不良健康，观测到的 I 统计量分别为0.239（泊松广义线性模型）和0.110（随机森林）；对于良好健康，分别为0.170（泊松广义线性模型）和0.110（随机森林）。这些结果表明，与较简单的对数线性模型相比，随机森林通过协变量捕捉了两组数据中更多的空间自相关。然而，针对独立性零假设的所有 p 值均小于0.0001，这表明尽管随机森林模型具有更强的解释能力，但单独使用该模型不足以捕捉两种健康结局中的空间结构，因此下一节将其与空间平滑模型融合。

3 方法论

用于估计暴露-响应函数的空间自回归随机森林算法记为 `SPAR-Forest-ERF`。该算法将随机森林用于捕捉灵活的混杂因素-响应关联，并与贝叶斯空间平滑模型相结合，在允许存在残差空间自相关的同时，估计一系列暴露-响应函数并进行完整不确定性量化。第3.1节介绍了这一灵活的模型类别，而推断算法则概述于第

3.2.

3.1 SPAR-Forest-ERF模型

将研究区域划分为 K 个互不重叠的区域单元 S $=$ {S1, . . . , S$_K$}，并令 Y $=[Y$(S$_1$) . . . , Y(S$_K$)]$^\top$和 e $=[e$(S$_1$) . . . , e(S$_K$)]$^\top$分别表示包含观测疾病计数和预期疾病计数的向量。此外，令 x(S$_k$) $=[x_1$(S$_k$) . . . , x_q(S$_k$)]$_{1\times q}$ 表示区域单元 S$_k$ 的暴露因素向量，而 z(S$_k$) $=[z_1$(S$_k$) . . . , z_p(S$_k$)]$_{1\times p}$表示混杂因素向量。在这里，暴露因素对疾病风险的影响是推断目标，而混杂因素则代表并非直接关注的干扰协变量。所提出的 SPAR-Forest-ERF模型的一般形式为:

$$\begin{aligned} Y(\mathcal{S}_k) &\sim \text{Poisson}[e(\mathcal{S}_k)\theta(\mathcal{S}_k)] \quad \text{for } k = 1, \ldots, K \\ \ln[\theta(\mathcal{S}_k)] &= \beta_0 + g[\mathbf{x}(\mathcal{S}_k)] + m[\mathbf{z}(\mathcal{S}_k)] + \phi(\mathcal{S}_k), \end{aligned} \tag{1}$$

其中， $\theta\mathcal{S}_k$表示$\mathcal{S}_k$ 相对于 $e\mathcal{S}_k$的疾病风险。对数风险面中的空间模式由感兴趣的暴露因素{x($\mathcal{S}_k$)}、混杂因素{z($\mathcal{S}_k$)} 以及一组随机效应{ϕ($\mathcal{S}_k$)} 建模，这些随机效应用于解释任何残差空间自相关；截距项则采用弱信息先验 $\beta_0 \sim$ N(0,100000)。

3.1.1 暴露-反应函数 $g[\mathbf{x}(\mathcal{S}_k)]$

根据分析目标，可以为暴露-反应函数 g[x($\mathcal{S}_k$)] 指定多种函数形式。下文将针对单一暴露因素，即x($\mathcal{S}_k$) $= x(\mathcal{S}_k)$ where $q = 1$，概述三种可能的选项。

I `Linear ERF` - $g[x$(S$_k$)$] = \alpha x$(S$_k$)，这是最简单的设定，可用于量化暴露与疾病风险之间的总体线性关联 α 。II `Non-linear ERF` - $g[x$(S$_k$)$]$, ，该设定允许这种关联呈平滑非线性函数，其大小取决于暴露水平。该平滑函数由贝叶斯 p 样条表示，并按照 Wang 等（2018）概述的方法，使用 INLA 软件将其拟合为二阶随机游走潜在场模型（Rue等，2009）。III `Measurement error ERF` - $g[x$(S$_k$)$] = \alpha x$(S$_k$) 和 x(S$_k$) $\sim$ N($\tilde{x}$ (S$_k$) σ_x^2)，该设定使用Berkson测量误差模型，将真实但未知的暴露 x(S$_k$)与现有的含误差估计值 $\tilde{x}$(S$_k$)关联起来。在示例应用中，可根据补充材料第2.1节所述，利用建模得到的污染数据和实测污染数据估计误差方差σ_x^2 。

For 对于选项 I 和 III，我们假设采用弱信息先验 $\alpha \sim$ N(0,100000).

3.1.2 混杂因素分量 $m[\mathbf{z}(\mathcal{S}_k)]$

混杂因素与疾病风险自然对数之间的关联 m[z($\mathcal{S}_k$)]由随机森林（RF）估计，因为它易于快速实现，能够捕捉非线性

且存在交互作用的混杂因素-响应关联，同时通过自助法构建并利用袋外预测，为数据生成近似样本外预测。此外，现有研究表明，随机森林在表格数据（Grinsztajn 等，2022）和空间数据（MacBride 等，2025）上的表现通常优于神经网络，而广义加性模型则要求预先指定混杂因素之间的具体交互作用，无法从数据中学习这些交互作用。补充材料第 2.2 节简要介绍了随机森林及其调优参数。

3.1.3 随机效应分量 ϕ（$\mathcal{S}_k$）

随机效应向量 $K \times 1$ $\boldsymbol{\phi} = [\phi(\mathrm{S}_1) \ldots, \phi(\mathrm{S}_K)]^\top$ 捕捉协变量调整后疾病风险中剩余的空间自相关，以及相对于泊松数据似然的任何过度离散。通常为它们指定条件自回归（CAR）先验分布，其空间自相关结构取决于邻接矩阵 W $K \times K$ $(=)w_{kj}$：当区域单元 $(\mathrm{S}_k, \mathrm{S}j)$ 共用一条边界时，$w_{kj} = 1$；否则，$w_{kj} = 0$（其中 $w_{kk} = 0\ \forall k$）。对于没有任何邻居的区域单元，则根据其与其余区域单元之间最小的质心间距离，为其指定一个邻居。已有多种不同的 CAR 先验被提出，也可纳入 SPAR-Forest-ERF。此处我们采用 Riebler 等（2016）提出的修正的 Besag-York-Mollie 模型，因为它包含一组随机效应 $\mathbf{u} = [u(\mathrm{S}_1) \ldots, u(\mathrm{S}_K)]^\top$，用于捕捉空间自相关；此外还包含第二组相互独立的随机效应 $\mathbf{v} = [v(\mathrm{S}_1) \ldots, v(\mathrm{S}_K)]^\top$，用于捕捉过度离散，具体如下。

$$
\begin{aligned}
\boldsymbol{\phi} &= \frac{1}{\sqrt{\tau_\phi}} \left[\sqrt{(1-\rho)}\mathbf{v} + \sqrt{\rho}\mathbf{u} \right] \\
\mathbf{v} &\sim \mathrm{N}(\mathbf{0}, \mathbf{I}) \\
\mathbf{u} &\sim \mathrm{N}(\mathbf{0}, \tilde{\mathbf{Q}}(\mathbf{W})^{-}).
\end{aligned} \tag{2}
$$

此处，将 v 建模为独立的零均值、单位方差高斯随机变量，$\boldsymbol{\phi}$ 中的变异量由全局精度参数 τ_ϕ控制。该变异中呈空间自相关的比例由 ρ控制，其中 $\rho = 0$对应于独立，而 $\rho = 1$ 对应于强空间自相关。空间自相关随机效应 u 被建模为零均值多元˜高斯分布，其奇异精度矩阵 Q(W)（此处 $^-$ 表示广义逆）是内在 CAR 模型精度矩阵的缩放形式，以确保边际方差的几何均值等于 1。最后，精度参数 τ_ϕ 在标准差尺度上被赋予精度为 0.001 的半正态先验（即 $\tau_\phi^{-1/2} \sim$ Half-Normal$(\tau_0 = 0.001)$），而 ρ 则在 $\ln(\rho/(1-\rho))$ 尺度上被赋予惩罚复杂度先验，因为 Riebler 等（2016）建议如此。

3.2 模型拟合与推断

模型（1）采用一种新颖的迭代两步残差学习策略进行拟合，该策略在以下方面扩展了 MacBride 等人（2025年）和 Figueira 等人（2025年）：（i）允许在算法的两个步骤之间充分传播不确定性；（ii）提出新的停止准则

该准则基于暴露-反应函数（主要推断目标）；（iii）通过其后验分布，对暴露-反应函数进行完整不确定性量化。拟合过程概述如下文 `Algorithm1`所示，初始化时设定 $g[\mathbf{x}(\mathcal{S}_k)] = \phi(\mathcal{A}_k) = 0 \ \ \forall k$。随后，算法迭代进行以下步骤：使用随机森林（**步骤 1**）估计 $\{m[\mathbf{z}(\mathcal{S}_k)]\}$ ，并使用贝叶斯空间平滑模型（**步骤 2**）估计$\{g[\mathbf{x}(\mathcal{S}_k)], \phi(\mathcal{A}_k)\}$ ；每个组成部分拟合时，均会根据另一组成部分的当前值及相关不确定性进行调整。请注意，为加快算法运行，在选择随机森林调优参数的最优值时不考虑不确定性。当连续两次迭代之间 $\{g[\mathbf{x}(\mathcal{S}_k)]\}$ 的平均绝对差小于 $\epsilon = 0.0005$时，算法终止；模拟研究表明，这一条件能带来良好的模型性能。

在**步骤 1** $\{m[\mathbf{z}(\mathcal{S}_k)]\}$ 中，使用袋外预测$\{\hat{m}[\mathbf{z}(\mathcal{S}_k)]\}$进行估计；由于真实值$\{m[\mathbf{z}(\mathcal{S}_k)]\}$ 相较于平滑后的袋外估计值$\{\hat{m}[\mathbf{z}(\mathcal{S}_k)]\}$很可能存在额外的随机变异，因此我们在贝叶斯空间平滑模型中加入伯克森测量误差分量，以传播这些预测中的不确定性。这便得到**步骤 2**模型：

$$
\begin{aligned}
Y(\mathcal{S}_k) &\sim \text{Poisson}[e(\mathcal{S}_k)\theta(\mathcal{S}_k)] \quad \text{for } k = 1, \ldots, K \\
\ln[\theta(\mathcal{S}_k)] &= \beta_0 + g[\mathbf{x}(\mathcal{S}_k)] + m[\mathbf{z}(\mathcal{S}_k)] + \phi(\mathcal{S}_k) \\
m[\mathbf{z}(\mathcal{S}_k)] &= \hat{m}[\mathbf{z}(\mathcal{S}_k)] + \varepsilon(\mathcal{S}_k), \quad \varepsilon(\mathcal{S}_k) \sim \text{N}(0, \hat{\sigma}^2_m),
\end{aligned}
\tag{3}
$$

其余模型组成部分的定义与第3.1节相同。预测误差$\{\varepsilon(\mathcal{S}_k) = m[\mathbf{z}(\mathcal{S}_k)] - \hat{m}[\mathbf{z}(\mathcal{S}_k)]\}$ 的方差记为$\hat{\sigma}^2_m$；按照Figueira 等人 (2025年)，用袋外预测误差的均方根误差进行估计。

与 Figueira 等人（2025年）和 MacBride 等人（2025年）不同，我们还将式（3）的后验不确定性纳入步骤 1 中随机森林的下一轮迭代。这是通过从式（3）生成 $q = 1, \ldots Q = 100$ 后验样本 $\{g[\mathbf{x}(\mathcal{S}_k)]_q + \phi(\mathcal{S}_k)_q\}_{q=1}^{Q}$ 来实现的，并用它们计算调整后的响应变量集合 $\{R(\mathcal{S}_k)_q = \ln[Y(\mathcal{S}_k)/e(\mathcal{S}_k)] - g[\mathbf{x}(\mathcal{S}_k)]_q - \phi(\mathcal{S}_k)_q\}_{k=1}^{K}$ 中的 Q 个样本，以用于随机森林的下一轮迭代。这些调整后的响应对应于式（3）的线性预测量尺度，并根据暴露-反应函数和随机效应的当前值进行了调整；混杂因素正是在这一尺度上进入模型。随后，对这些 Q 组响应变量分别拟合一个包含 10 棵树的随机森林，再将这 100 个随机森林合并，构建一个包含 1,000 棵树的总体随机森林，从而纳入 $\{g[\mathbf{x}(\mathcal{S}_k)] + \phi(\mathcal{S}_k)\}_{k=1}^{K}$ 的后验不确定性。该随机森林的袋外预测是由未使用相关数据点进行拟合的子森林中的树所产生的平均预测值，其误差方差则通过以下方式计算：

$$
\hat{m}[\mathbf{z}(\mathcal{S}_k)] = \frac{1}{\sum_{q=1}^{Q} n_{qk}} \sum_{q=1}^{Q} n_{qk} \tilde{m}[\mathbf{z}(\mathcal{S}_k)]_q \quad \text{and} \quad \hat{\sigma}^2_m = \frac{1}{K} \sum_{k=1}^{K} (\tilde{R}(\mathcal{S}_k) - \hat{m}[\mathbf{z}(\mathcal{S}_k)])^2.
$$

这里，$\{\tilde{m}[\mathbf{z}(\mathcal{S}_k)]_q, n_{qk}\}$ 分别表示基于第 q 个数据样本构建的随机森林中的袋外预测，以及该预测所依据的树的数量

对应第 k 个观测值。平均响应变量的计算方式类似，为$\tilde{R}(\mathcal{S}_k) = \frac{1}{\sum_{q=1}^{Q} n_{qk}} \sum_{q=1}^{Q} n_{qk} R(\mathcal{S}_k)_q$。

算法 1 SPAR-Forest-ERF

初始化

初始化 $g[\mathbf{x}(\mathcal{S}_k)]^{(0)} = \phi(\mathcal{A}_k)^{(0)} = 0 \ \ \forall k$，并选择停止阈值 ϵ .

迭代

迭代**步骤 1** 和**步骤 2**，遍历 i = 1、2、3、.....，直到 $\frac{1}{K}\sum_{k=1}^{K}\left|g[\mathbf{x}(\mathcal{S}_k)]^{(i)} - g[\mathbf{x}(\mathcal{S}_k)]^{(i-1)}\right| < \epsilon$。

步骤 1。生成 $q = 1, ..., Q$ 个样本，这些样本取自经调整的响应变量集合 $\{R(\mathbf{S}_k)_q^{(i)} = \ln[Y(\mathbf{S}_k)/e(\mathbf{S}_k)] - g[\mathbf{x}(\mathbf{S}_k)]_q^{(i-1)} - \phi(\mathbf{S}_k)_q^{(i-1)}\}_{k=1}^{K}$；其中，若 $i = 1$，则对所有 q，都有 $g[\mathbf{x}(\mathbf{S}_k)]_q^{(i-1)} = \phi(\mathbf{S}_k)_q^{(i-1)} = 0$。使用协变量 $\{\mathbf{z}(\mathbf{S}_k)\}$ 和响应变量的均值（对 $q = 1, ..., Q$ 取平均）作为响应变量，对随机森林进行调参，以获得 `mtry` 和 `minnode` 的最优值。基于这些最优值，针对响应变量的 $q = 1, ..., Q$ 个样本分别拟合随机森林，并构造袋外预测 $\{\hat{m}[\mathbf{z}(\mathbf{S}_k)]^{(i)}\}$ 及其方差 $\hat{\sigma}_m^{2(i)}$。

步骤 2。使用 INLA 拟合贝叶斯空间平滑模型 (3)，其中真实的混杂因素分量 $\{m[\mathbf{z}(\mathbf{S}_k)]\}_{k=1}^{K}$ 由 Berkson测量误差模型表示，该模型基于步骤 1 中计算的估计值 $\{\hat{m}[\mathbf{z}(\mathbf{S}_k)]^{(i)}\}_{k=1}^{K}$ 和误差方差 $\hat{\sigma}_m^{2(i)}$。从该模型的后验分布中生成 $q = 1, \ldots, Q$ 个 $\{g[\mathbf{x}(\mathbf{S}_k)]_q^{(i)} + \phi(\mathbf{A}_k)_q^{(i)}\}_{k=1}^{K}$ 样本。

结束迭代

推断

暴露反应函数 $\{g[\mathbf{x}(\mathcal{S}_k)]\}_{k=1}^{K}$ 和随机效应$\{\phi(\mathcal{A}_k)\}_{k=1}^{K}$ 的推断基于在**步骤 2**中拟合的最终贝叶斯空间平滑模型的后验分布，而有关混杂因素的推断则由**步骤 1**。中得到的最终袋外估计$\{\hat{m}[\mathbf{z}(\mathcal{S}_k)]^{(i)}\}_{k=1}^{K}$ 给出。

3.3 软件

用于实现SPAR-Forest-ERF算法的软件可从 https://github.com/vinnydavies/SPARforest 免费获取，并附有相应的文档。该软件以R编写，随机森林组件通过 `ranger` (Wright & Ziegler 2017)软件包中的递归二元划分算法进行拟合，而层次贝叶斯模型则通过 `INLA`(Rue等，2009)软件包，使用积分嵌套拉普拉斯近似和变分贝叶斯近似进行拟合。

4 模拟研究

本研究量化了所提出的 SPAR-Forest-ERF 算法及其竞争算法在一系列不同条件下估计暴露-反应函数的准确性，这些条件包括不同的疾病患病率、暴露-反应函数形状、混杂类型以及残差空间自相关程度。

4.1 数据生成

本研究基于苏格兰中部地区的 $K = 3{,}066$ 个数据区，其中包括格拉斯哥和爱丁堡；由于苏格兰全境的数据量过大，对数百个模拟数据集反复进行模型拟合在计算上难以承受，因此未纳入全境数据。疾病计数数据由相互独立、均值为 $\{e(\mathcal{S}_k)\theta(\mathcal{S}_k)\}$ 的泊松分布生成；在不同情景下改变预期计数 $\{e(\mathcal{S}_k)\}$ ，以确定疾病患病率对估计准确性的影响。空间风险面 $\{\theta(\mathcal{S}_k)\}$ 在自然对数尺度上生成，作为暴露变量、已知混杂因素和一组随机效应的函数；后者代表现有协变量未能解释的疾病风险残余空间模式。针对每个模拟数据集，构造两个空间自相关程度不同的暴露变量$\{x_1(\mathcal{S}_k)\ x_2(\mathcal{S}_k)\}$ ，以考察其依赖结构对各自暴露-反应函数估计的影响。第一个变量在空间上独立生成，服从标准正态分布。第二个变量假定具有空间自相关性，以模拟空气污染浓度，并从均值为零的多元正态分布中抽取。其协方差矩阵由指数型自协方差函数确定，该函数基于数据区质心之间的两两距离；范围参数经过设定，使相距 4.45 千米的位置之间的相关性为 0.75（质心间距离分布的第 $5th$ 百分位数），以确保存在较强的空间自相关。随后将两个暴露变量都重新缩放至单位区间，确保它们具有相同的正暴露范围。

每个模拟数据集生成八个不同的混杂因素，但并非所有因素都会影响真实风险面。前四个在空间上相互独立，按照上述方法从标准正态分布中生成，然后重新缩放至单位区间。后四个具有空间自相关性，生成时取自均值为零的多元正态分布，其球状自相关函数对于间距超过 5 km 的数据区为零；随后同样重新缩放至单位区间。这些混杂因素与疾病风险之间的关系为 $m[\mathbf{z}(\mathcal{S}_k)] = 3/\{1+\exp[6-12z_1(\mathcal{S}_k)]\}+2[z_3(\mathcal{S}_k)z_6(\mathcal{S}_k)-0.4]^2+\cos[2\pi z_7(\mathcal{S}_k)]+z_8(\mathcal{S}_k)$，其中包含线性和非线性关系以及交互作用。最后，残差空间自相关由另一个均值为零的多元正态分布生成，其协方差矩阵对应于 Leroux 等人（2000）提出的条件自回归模型，空间依赖参数固定为 $\rho = 0.95$ ，以产生相对较强的空间自相关。

4.2 研究设计

在 9 种不同情景下分别生成 100 个模拟数据集，其中前 8 种情景涵盖了以下三个因素的所有可能组合。此外，第九种情景具有完全线性的混杂因素-响应关联，且不存在交互作用，其形式为 $m[\mathbf{z}(\mathcal{S}_k)] = 0.5z_1(\mathcal{S}_k) - z_3(\mathcal{S}_k) + 2z_7(\mathcal{S}_k) - 0.1z_8(\mathcal{S}_k)$，这使我们能够观察 SPAR-Forest-ERF 在这种简化情形下的表现。

- **疾病患病率** - (A) `Rare` - 其中 $e(\mathcal{S}_k) \sim$ 均匀分布$[10, 25]$；以及 (B) `Common`- 其中 $e(\mathcal{S}_k) \sim$ 均匀分布$[100, 200]$。这使我们能够评估当数据中的信息量不同时，各模型的表现如何。

- **真实ERF形状** - (A) `Linear` - $g[x_j(\mathcal{S}_k)] = x_j(\mathcal{S}_k)\alpha$; 以及 (B) `Sigmoidal` -

$g[x_j(\mathcal{S}_k)] = \alpha/[1+\exp(6-12x_j(\mathcal{S}_k))]$, 对于每个协变量 $x_j(\mathcal{S}_k)$ 对于 $j = 1,2$。这些暴露-反应函数的形状显示于补充材料的图5中。尽管为简便起见，流行病学空气污染研究通常假设线性形状，但近期一项对3项大型队列研究进行的协调分析发现了S形形状（Chen 等，2023）。此处我们固定 $\alpha = 0.2$，因为这样得到的暴露-反应函数范围与动机研究中观察到的$PM_{2.5}$ 相近。

- **控制混杂**- (A) `Good` - $m[\mathbf{z}(\mathcal{S}_k)]$ 捕获了数据中未被暴露因素解释的95%变异，因此由 $\phi(\mathcal{S}_k)$表示的未测量混杂仅占该变异的5%；(B) `Poor` - $m[\mathbf{z}(\mathcal{S}_k)]$ 捕获了数据中未被暴露因素解释的60%变异，因此由 $\phi(\mathcal{S}_k)$表示的未测量混杂占该变异的40%。

本研究比较了三种不同的模型。第一种是标准空间流行病学风险模型，记为 `GLMM`，它是式（1）的简化形式，具有线性的混杂因素-响应关联，即 $m[\mathbf{z}(\mathbf{S}_k)] = \mathbf{z}(\mathbf{S}_k)^\top \boldsymbol{\delta}$。第二种是本文提出的SPAR-Forest-ERF算法（记为 `SPAR-Forest-ERF`），第三种是该算法的简化版本：它不采用迭代方式（两个模型组件均只拟合一次），且不在两个模型组件之间传播不确定性（记为 `SPAR-Forest-ERF-1`）。纳入这一简化模型是为了确定完整迭代算法的复杂性是否必要。通过模型估计空间自相关暴露因素和独立暴露因素对应暴露反应函数$\{g[x_j(\mathbf{S}_k)]$的准确度，衡量这些模型的相对性能。分别用$\{g[x_j(\mathbf{S}_k)], \hat{} g[x_j(\mathbf{S}_k)]\}_{k=1}^K$ 表示真实和估计的（后验均值）暴露反应函数，同时还计算逐点95%可信区间。通过为100个模拟数据集中的每一个计算以下指标，量化各模型的估计准确度：（i） $\texttt{Bias} = \frac{1}{K}\sum_{k=1}^K\{\hat{} g[x_j(\mathbf{S}_k)] - g[x_j(\mathbf{S}_k)]\}$；（ii）均方根误差（RMSE） $= \sqrt{\frac{1}{K}\sum_{k=1}^K\{\hat{} g[x_j(\mathbf{S}_k)] - g[x_j(\mathbf{S}_k)]\}^2}$；（iii）覆盖率 $=$ ——包含$\{g[x_j(\mathbf{S}_k)]$真实值的逐点95%可信区间所占百分比；以及（iv）~~平均区间宽度（MIW） $=$ ——~~$\{g[x_j(\mathbf{S}_k)]$逐点95%可信区间的平均宽度。偏差概括了各模型平均而言是高估还是低估了暴露反应函数，而均方根误差则量化了点估计的准确度。最后，覆盖率和平均区间宽度指标概括了模型量化这些估计值不确定性的准确度。

4.3 结果

研究结果如下所示，补充材料第3.3节也展示了相关结果；SPAR-Forest-ERF算法的迭代次数，以及其最后两次迭代所得估计的暴露-反应函数之间的平均绝对差汇总，则见补充材料第3.2节。在全部100个模拟数据集的所有情景中，三种模型对暴露-反应函数的估计基本无偏，因此为简洁起见，此处不展示这些结果。简而言之，自相关暴露对应的偏差介于（-0.006，0.006）之间，独立暴露对应的偏差介于（-0.001，0.001）之间；无论暴露-反应函数是线性的还是S形的，均如此。

线性暴露-反应函数估计值的均方根误差对应于

空间自相关暴露的结果见图1，图中展示了三种模型的结果。本文主要关注这种暴露，因为在作为研究动机的研究中，它模拟的是空气污染。上方四个面板对应罕见与常见疾病、混杂控制良好与较差这两组因素的所有两两组合；下方面板则对应一个简单情景，即所有已知混杂因素与疾病风险之间都恰好呈线性关系。为便于观察，图中同时以箱线图和实际数值（水平抖动显示）的形式展示了100个模拟数据集的均方根误差，并在每个箱线图上方标出各情景下的均方根误差中位数。图中显示，在混杂因素-响应关联并非恰好呈线性的所有情况下（上方四个面板），SPAR-Forest-ERF算法的表现都大幅优于常用的广义线性混合模型；四种情景下均方根误差中位数的降幅为49%至83%。简化版SPAR-Forest-ERF-1算法的均方根误差介于另外两种模型之间；与广义线性混合模型相比，其中位数降幅为16%至73%。最后，在所有混杂因素与疾病风险之间都恰好呈线性关系这一简单情景中，尽管数据生成机制与广义线性混合模型相符，因而有利于该模型，但广义线性混合模型与SPAR-Forest-ERF模型的结果仍相当。独立暴露的相应均方根误差结果，以及真实暴露-反应函数呈S形时两种暴露的结果，见补充材料第3.3节。

表1展示了空间自相关暴露在线性暴露-反应函数下95%可信区间的逐点百分比覆盖率和平均区间宽度，其中结果为每种情景下所有100个模拟数据集的平均值。请注意，由于真实暴露-反应函数是线性的，因此每个95%可信区间要么对所有 K 观测值均包含真实暴露-反应函数，要么均不包含，从而对100个模拟数据集取平均后得到整数覆盖率。关于覆盖率和平均区间宽度的其余结果见补充材料第3.3节。表中显示，在几乎所有情况下，SPAR-Forest-ERF算法的覆盖率都接近名义95%水平（范围为80%–99%），而广义线性混合模型的相应覆盖率则低得多，介于56%和94%之间。简化的SPAR-Forest-ERF-1算法的覆盖率大多介于另外两个模型之间，范围为63%–94%。除覆盖率有所提高外，与广义线性混合模型相比，SPAR-Forest-ERF算法还具有更精确的可信区间，因为在存在非线性混杂的8种情景中，其平均区间宽度窄了9%–52%。在线性混杂情景中，区间略宽，为0.023，而广义线性混合模型为0.019。当空间随机效应占疾病风险变异的较大比例时（即当$\{\phi(\mathcal{S}_k)\}$占混杂的40%时），并且疾病较为常见时，三种模型的覆盖率均最低；此时，SPAR-Forest-ERF和广义线性混合模型的覆盖率分别低至80%和61%。出现这些较低覆盖率的原因在于，在这种情况下：（i）大量疾病病例意味着数据中包含大量信息，从而导致相对较窄的可信区间；（ii）空间自相关随机效应具有很强的影响力，并对空间平滑暴露的估计效应造成混杂（见Hodges与Reich（2010）），从而导致估计相对较差。

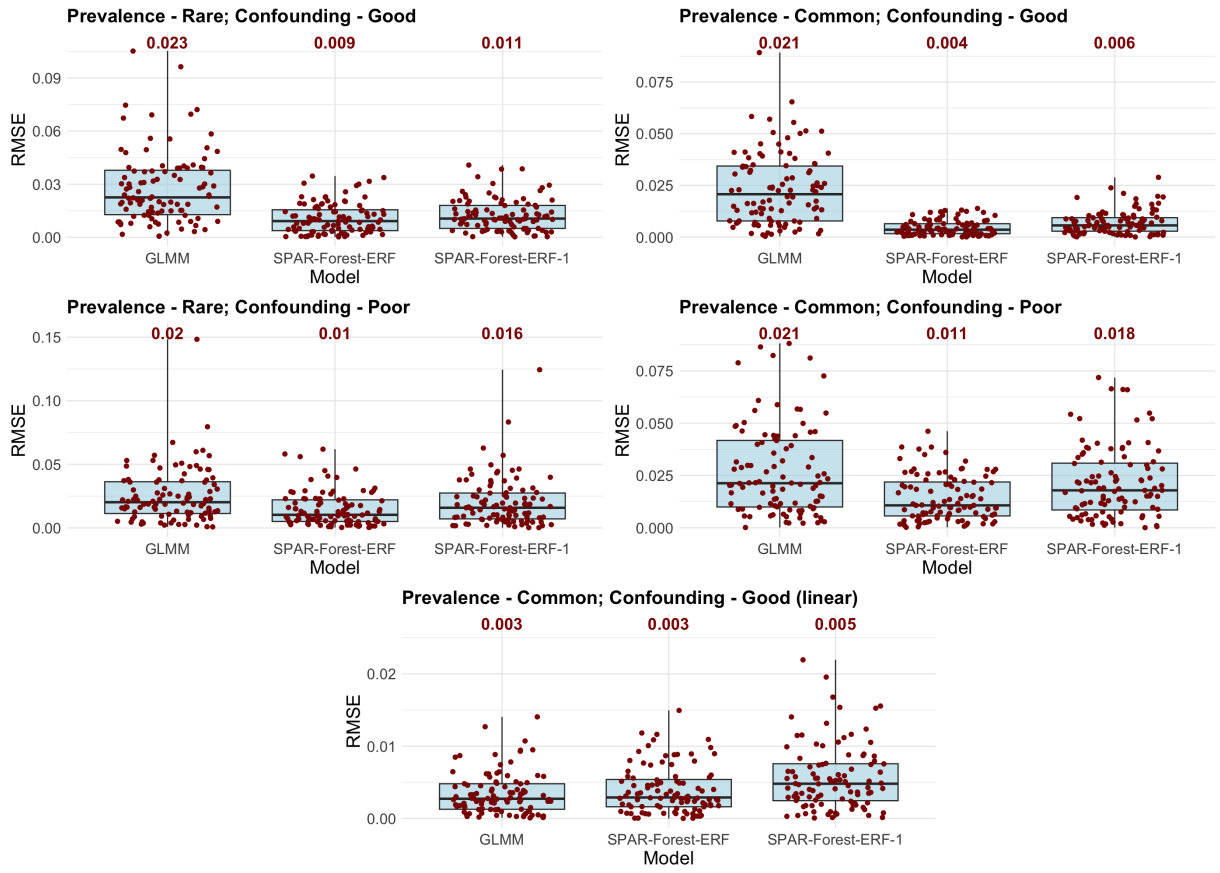

图1：空间自相关暴露对应的线性暴露-反应函数估计值的均方根误差（箱线图和点）。上方四个面板分别对应不同的疾病患病率（罕见与常见）及混杂控制情况（良好与较差）；下方面板对应所有已知混杂因素与疾病风险均呈简单线性关系的情形。数字表示100个模拟数据集的平均均方根误差。

流行率混杂		广义线性混合模型	SPAR-Forest-ERF	SPAR-Forest-ERF-1
覆盖率				
Rare	Good	84	99	94
常见	Good	56	98	84
Rare	Poor	83	93	88
常见	Poor	61	80	63
常见	良好（线性）	94	99	91
MIW				
Rare	Good	0.090	0.069	0.054
常见	Good	0.048	0.023	0.024
Rare	Poor	0.086	0.068	0.075
常见	Poor	0.050	0.045	0.045
常见	良好（线性）	0.019	0.023	0.022

表1：空间自相关暴露对应的线性暴露-反应函数的覆盖率（%）和平均区间宽度（MIW）。

5 动机研究的结果

此处考虑模拟研究中比较的相同三种模型，并使用第 2 节所述的混杂因素集合，分别将每种模型应用于自评健康状况不佳和自评健康良好这两种结局。由于流行病学文献中普遍采用这种做法，以避免潜在的共线性问题，因此将三种污染物 NO_2, $PM_{2.5}$ 和 PM_{10} 分别纳入不同模型。对于每个污染物-结局对和模型，我们估计三种不同的暴露-反应函数（ERF），包括：(I) 简单的线性暴露-反应函数，可得出单一且易于解释的相对风险；(II) 非线性暴露-反应函数，以确定其真实形状；以及 (III) 通过伯克森测量误差模型考虑模型化浓度中可能存在的误差的线性暴露-反应函数。线性（I）和非线性（II）暴露-反应函数的结果见下文第 5.2 和 5.3 节；测量误差模型（III）的结果见补充材料第 4.1 节。

5.1 模型比较

采用线性暴露-反应函数（案例 I）的每个模型的性能总结于表2 中，评估指标包括其对观测数据的拟合程度、预测性能以及捕捉数据中空间自相关的能力。对观测数据的拟合程度通过偏差$D(\boldsymbol{\Theta}) = -2\ln[l(\boldsymbol{\Theta}|\mathbf{Y})]$（其中 $\boldsymbol{\Theta}$ 表示模型参数）进行度量，具体通过偏差后验分布的均值 $\hat{}D = \mathbf{E}_{\boldsymbol{\Theta}|\mathbf{Y}}[D(\boldsymbol{\Theta})]$)以及在参数后验均值处计算的偏差 $\hat{}\boldsymbol{\Theta}$（记为 $D_{\hat{\boldsymbol{\Theta}}}$）进行评估。表中显示，SPAR-Forest-ERF算法对两种健康结局均提供了最佳拟合，在两种偏差指标下的数值均最低。对于不良健康结局，SPAR-Forest-ERF还表现出最佳的交叉验证预测性能，因为其广泛适用信息准则（WAIC）值最低，而该指标在渐近意义上等价于贝叶斯交叉验证损失（Watanabe 2010）。然而，对于良好健康结局，其WAIC值最差（最高），这是由于其大得多的

模型	偏差 - $\hat{D}$	偏差 -$D_{\hat{\Theta}}$	WAIC (p_w)	莫兰指数
Bad self-rated health				
GLMM	46,885	43,021	50,528 (2,767)	-0.0339
SPAR-Forest-ERF	45,926	41,092	49,464 (2,634)	0.0320
SPAR-Forest-ERF-1	46,915	43,558	50,224 (2,553)	-0.0205
Good self-rated health				
GLMM	62,030	61,560	62,369 (315)	0.0289
SPAR-Forest-ERF	61,606	58,732	63,170 (1,244)	0.0626
SPAR-Forest-ERF-1	61,824	61,619	61,964 (136)	0.0224

表2：比较各模型针对每种健康结局的相对性能。模型拟合度量包括：(i) 通过偏差后验均值（$\hat{D}$）以及在参数后验均值处计算的偏差 $D_{\hat{\Theta}}$)评估其对观测数据的拟合程度；(ii) 通过WAIC及其有效独立参数数量（p_w）评估预测性能；以及通过其残差的莫兰 I 统计量评估其捕捉空间自相关的能力。

估计的独立参数数量 p_w。最后，对于两种健康结局，所有模型都能较好地捕捉数据中的空间自相关，因为残差的莫兰指数接近于零，介于 -0.035 和 0.063 之间。相比之下，简单的仅含协变量的泊松对数线性模型所得残差的相应数值为 0.239（不良健康）和 0.170（良好健康），原始标准化死亡比（SMR）值则为 0.550（不良健康）和 0.551（良好健康）。

5.2 线性暴露反应函数- $g[x（\mathcal{S}_k）]$

简单的线性暴露反应函数是流行病学文献中的标准方法，因为它用一个易于解释的相对风险（RR）来表示污染与健康的关联。这些相对风险量化了每种污染物的年平均浓度每固定增加一定幅度所对应的疾病风险增幅。此处我们给出了每种污染物在以下合理增幅下的相对风险（这些增幅接近其空间标准差）：NO_2 - $4\mu gm^{-3}$；$PM_{2.5}$ - $1\mu gm^{-3}$；PM_{10} - $2\mu gm^{-3}$。这些相对风险的估计值和95%可信区间列于表3中，其中上半部分对应不良自评健康，下半部分对应自评健康良好。

GLMM和SPAR-Forest-ERF模型的结果一致表明，颗粒物空气污染（$PM_{2.5}$ 和PM_{10}）在5%的显著性水平上会显著损害自评健康，因为所有95%可信区间均不包含为1的无效相对风险。估计的相对风险表明，如果颗粒物浓度升高，不良自评健康的风险将增加约4%（$PM_{2.5}$）和2.7%（PM_{10}）。自评健康良好的估计降幅则小得多，两种颗粒物污染物对应的相对风险均降低约0.4%。尽管GLMM和SPAR-Forest-ERF模型的相对风险点估计（后验均值）相近，但后者的估计精度更高，颗粒物相关结果的95%可信区间窄了0%至24%。相比之下，NO_2 的结果不一，四个相对风险中仅有一个在5%的显著性水平上具有统计学显著性。

模型	相对风险（RR）		
	NO_2	$PM_{2.5}$	PM_{10}
Bad self-rated health			
GLMM	1.005（0.988，1.022）	1.040（1.023，1.057）	1.027（1.013，1.041）
SPAR-Forest-ERF	1.032（1.021，1.043）	1.041 (1.028, 1.054)	1.027（1.015，1.039）
SPAR-Forest-ERF-1	1.002（0.991，1.012）	1.002（0.991，1.013）	1.007（0.997，1.018）
Good self-rated health			
GLMM	1.000（0.998，1.003）	0.996（0.993，0.999）	0.996（0.994，0.998）
SPAR-Forest-ERF	0.998（0.996，1.000）	0.996（0.993，0.998）	0.997（0.995，0.999）
SPAR-Forest-ERF-1	0.999 (0.998, 1.001)	1.000（0.998，1.002）	0.999（0.997，1.001）

表3：各污染物-结局对在线性暴露-反应函数下相对风险（RR）尺度上的汇总。相对风险对应于以下污染物浓度增幅：NO_2 - $4\mu gm^{-3}$；$PM_{2.5}$ - $1\mu gm^{-3}$；PM_{10} - $2\mu gm^{-3}$。

SPAR-Forest-ERF算法的单次迭代简化版本在所有情况下都产生了被削弱且不显著、接近1的相对风险。这可能是因为在该算法中，混杂因素通过高度灵活的随机森林组件，尽可能多地解释了疾病风险的变异，因此留给暴露（和随机效应）解释的剩余变异相对较少。这一假设得到有效独立参数数量的佐证 p_w，见表2，简化版SPAR-Forest-ERF算法的这一数值最低。最后，为了更清晰地展示SPAR-Forest-ERF算法的工作原理，补充材料第4.2节展示了该算法各次迭代中估计的线性暴露-反应函数的演变。

5.3 非线性暴露-反应函数- $g[x（\mathcal{S}_k）]$

为确定线性暴露-反应函数能否准确描述污染物与自评健康之间的关联，我们使用灵活的贝叶斯 p 样条估计了平滑的暴露-反应函数，该样条通过第3.1.1节所述的二阶随机游走潜场模型实现。我们使用 SPAR-Forest-ERF 模型估计这些可能呈非线性的暴露-反应函数，因为模拟研究表明该模型的估计值最准确。针对 $PM_{2.5}$ 和 PM_{10} 的非线性暴露-反应函数，分别在自评健康不佳和自评健康良好的结局下进行估计；而未考虑 NO_2，因为其大多数线性暴露-反应函数均未显示显著关联（见表 3）。

图中展示了四个估计的暴露-反应函数及其逐点95%可信区间 2，其中左列和右列分别对应 $PM_{2.5}$ 和 PM_{10}，上排和下排则分别对应不良健康和良好健康结局。每个暴露-反应函数均以研究中观测到的最低污染物浓度为参照，以相对风险表示；因此，该浓度下的相对风险等于 1，且没有后验不确定性。最后，相对风险为 1 的红色虚线表示无关联。图中显示，在低至中等浓度下，暴露-反应函数总体呈线性形态，随后在研究中观测到的最高浓度处趋于平稳。然而，一个完全线性的暴露-反应函数落在

所有四种情况下的逐点95%可信区间内，这表明没有确凿证据可以否定线性暴露-反应函数。针对不良健康的两个暴露-反应函数在几乎所有浓度下都与零假设相对风险 1 存在显著差异，表明有一致的关联证据。相比之下，对于良好健康结局，虽然数据不支持相对风险恒为 1，但在低至中等浓度以及最高浓度处，逐点95%可信区间包含 1，这表明两者之间的关系要弱得多。

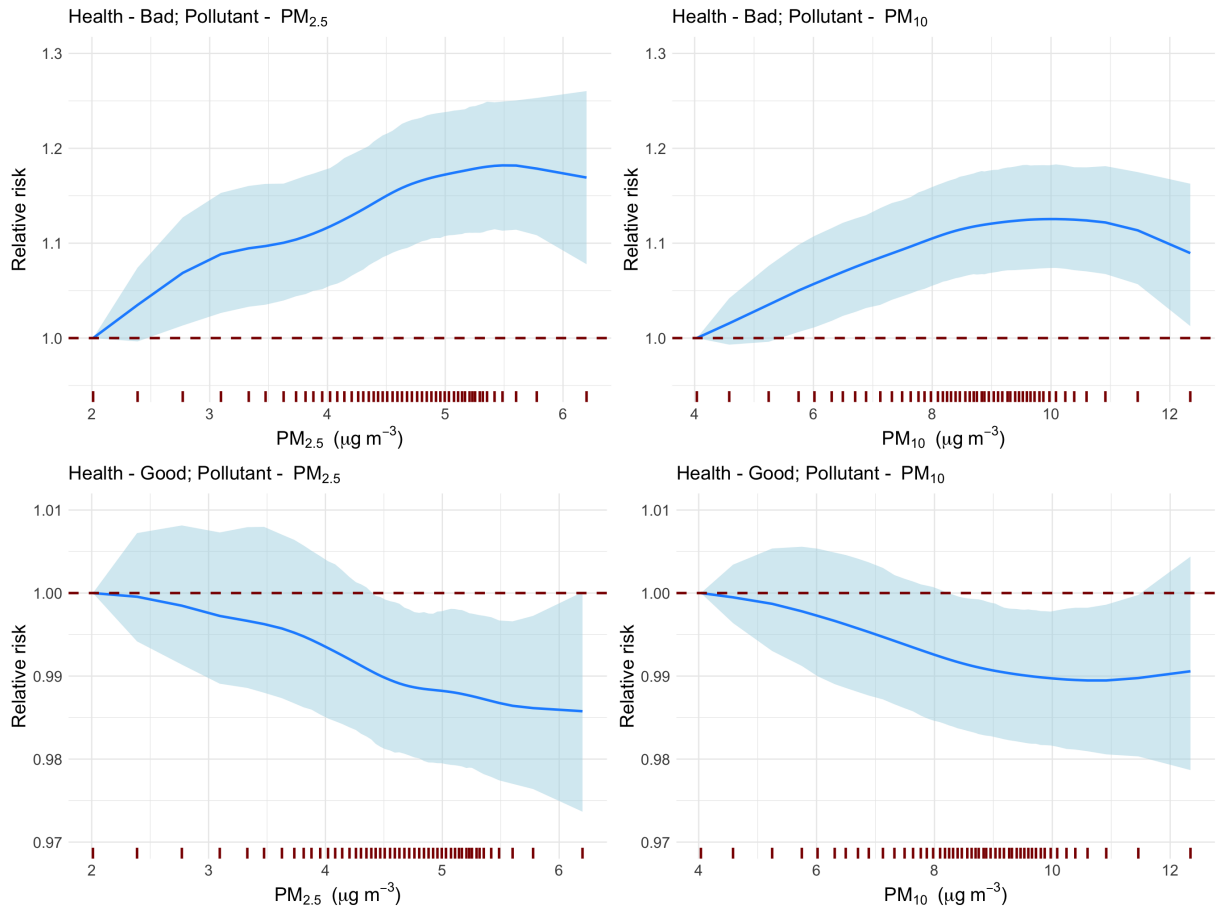

图2：SPAR-Forest-ERF模型对自评健康较差（上）和良好（下）人群中 $PM_{2.5}$ （左）和 PM_{10} （右）的暴露反应函数$g[x(S_k)]$ 的估计值及逐点95%可信区间。暴露-反应函数以相对于研究中观测到的最低污染物浓度的相对风险表示。相对风险为1处的红色虚线表示不存在关联。

6 讨论

本文提出了 SPAR-Forest-ERF算法，用于人群层面的流行病学风险评估。该算法将随机森林捕捉非线性混杂因素-反应效应及其交互作用的能力，与贝叶斯空间平滑模型相结合；后者能够估计具有完整不确定性量化、且易于解释的基于相对风险的暴露反应函数，同时捕捉此类数据中普遍存在的残差空间自相关。该算法

首次将机器学习算法与贝叶斯空间平滑模型结合起来，用于分析目标是效应估计而非预测的区域单元数据。从方法上说，该算法建立在MacBride 等人(2025年) 和 Figueira 等人(2025年) 的研究基础上，通过在机器学习和统计模型之间正确传递不确定性，开发了旨在确保主要推断目标稳定性的新停止准则，并纳入多种不同的暴露-反应函数，以实现最大的模型灵活性。配套软件使该算法能够得到广泛应用；鉴于其性能优于现有方法，且空间生态关联研究广泛应用于经济学、环境流行病学和社会科学等领域（仅举三例），其影响可能十分显著。

模拟研究揭示了 SPAR-Forest-ERF 相较于一种常用于流行病学风险评估、且假设混杂因素与响应之间存在线性关联的广义线性混合模型的相对效能，并得出了若干重要结论。首先，在更贴近现实的情形中，暴露在空间上存在自相关（例如空气污染），且混杂因素与疾病风险之间并非完全呈线性关系、也不存在交互作用，此时所提出的迭代算法明显更优：以均方根误差衡量，点估计的表现提升了49%至83%。不确定性量化方面的表现也更优，具体而言，95%可信区间更窄，同时其百分比覆盖率也更接近名义95%水平。对于形状各异的暴露-反应函数（例如线性和S形），这些结果似乎都一致，这使 SPAR-Forest-ERF 成为一种出色的通用工具，可利用空间区域单元数据量化暴露对响应的影响。值得注意的是，即便在不太可能出现的理想化情形下，即：（i）所有混杂因素与响应之间均完全呈线性关系，且不存在交互作用；或（ii）暴露在空间上完全独立，且与混杂因素相互独立；SPAR-Forest-ERF 的表现也与常用的广义线性混合模型相近，尤其是在点估计方面。我们还强调，算法的迭代特性不可或缺，因为模拟研究结果表明，在几乎所有情景下，将算法简化为仅迭代一次都会导致表现变差。

动机性研究的结果表明，以 $PM_{2.5}$ 和 PM_{10} 衡量的颗粒物空气污染与总体自评健康呈现一致且显著的关联，这与大量流行病学文献的结论一致；这些文献量化了颗粒物对特定发病率（例如，Forastiere 等人（2024））和死亡率（例如，Orellano 等人（2024））终点指标的影响。本研究的证据具有较强的稳健性，因为一系列分析得出的发现一致；这些分析包括线性和非线性暴露-反应函数、对暴露测量误差的考虑，以及对其对自评不良健康和自评健康良好两方面影响的评估。本研究还为不断积累的证据添砖加瓦，例如 Boogaard 等人 （2024年）的研究表明，即使空气污染浓度较低，仍会造成危害，尤其是低于英国国家空气质量战略目标和世界卫生组织空气质量指南所规定水平的浓度。例如，研究发现，长期浓度每增加 $_{2.5}$ ，自评不良健康的风险会显著增加 2.3% 至 4.1%；尽管苏格兰所有数据区的浓度均低于英国目标值 $1\mu gm^{-3}$ ，且 68% 的数据区低于世界卫生组织指南值 $10\mu gm^{-3}$ 。 $5\mu gm^{-3}$

此外，空气污染与自评健康良好相关的估计相对风险，其幅度要小得多，约为自评不良健康相关风险的十分之一。

（变化幅度为0.4%，而非4%）。造成这些差异性关联的一个可能原因是，与不良健康风险相比，良好健康风险的空间变异要小得多，这一点可以从补充材料中的表1所列标准化死亡比中观察到。从本质上说，由于两种结局的暴露相同，但响应的尺度不同，因此估计的斜率系数的幅度也必然不同。

最后，SPAR-Forest-ERF 可视为一种通用算法，它将用于捕捉非研究重点的复杂混杂因素-响应关联的机器学习组件，与用于估计具有完整不确定性量化且易于解释的暴露-反应函数、同时捕捉数据中任何残差自相关的贝叶斯平滑模型相结合。因此，该算法可以轻松修改，以适应其他应用的需求，因而是利用空间数据开展风险评估的一种非常灵活的工具。例如，可以将此处使用的空间自相关模型替换为适用于点级数据的地统计模型，或者在数据涵盖多个时间段时替换为时空模型。此外，还可以纳入不同类型的暴露-反应函数，例如变系数模型；而随机森林组件则可以替换为神经网络等其他机器学习算法。

7 利益冲突声明

无利益冲突

8 数据可用性声明

用于说明该方法的研究所用数据均可公开获取，数据来源见第 2 节。

补充材料

补充材料：补充材料可应要求提供。因篇幅限制，未纳入正文的额外数据分析和方法说明。（.pdf）

参考文献

Abed Al Ahad, M., Demsar, U., Sullivan, F. & Kulu, H.（2023年），《空气污染对个体报告健康状况的时空效应及其在英国不同族群中的差异：一项多水平纵向分析》，*BMC Public Health* **23**，897。

Besag, J., York, J. &Mollié, A.（1991年），《贝叶斯图像复原及其在空间统计学中的两项应用》，*Annals of the Institute of Statistics and Mathematics* **43**，1–59。

Boogaard, H., Crouse, D., Tanner, E., Mantus, E., van Erp, A., Vedal, S. & Samet, J.（2024），《长期暴露于低浓度环境空气污染的不良健康效应评估：健康效应研究所的经验与未来方向？》，*Environmental Science & Technology***58**，12767–12783。

Chen, J., Braun, D., Christidis, T., Cork, M., Rodopoulou, S., Samoli, E., Stafoggia, M., Wolf, K., Wu, X., Yuchi, W., Andersen, Z., Atkinson, R., Bauwelinck, M., Hoogh, K., Janssen, N., Katsouyanni, K., Klompmaker, J., Kristoffersen, D., Lim, Y., Oftedal, B., Strak, M., Vienneau, D., Zhang, J., Burnett, R., Hoek, G., Dominici, F., Brauer, M. & Brunekreef, B.（2023），《长期暴露于低浓度PM2.5与死亡率：通过协调加拿大、美国和欧洲大型队列研究的分析，探究异质性》，《环境健康展望》**131**，12。Chen, W., Li, Y., Reich, B. & Y, S.（2024），《DeepKriging：用于空间预测的空间依赖深度神经网络》，《中国统计学报》**34**，291–311。Figueira, M., Cameletti, M. & Patelli, L.（2025），《INLA-RF：时空环境数据的混合建模策略》，*arXiv*预印本 *arXiv:2507.18488*。Forastiere, F., Spadaro, J., Ancona, C., Jovanovic, Z., Cozzi, I., Gumy, S., Loncar, D., Mudu, P., Medina, S., Perez-Velasco, R., Walton, H., Zhang, J. & Krzyzanowski, M.（2024），《长期暴露于空气污染的健康风险评估中发病结局和浓度-反应函数的选择》，《环境流行病学》**8**，e314。Grinsztajn, L., Oyallon, E. & Varoquaux, G.（2022），为什么基于树的模型在典型表格数据上仍优于深度学习？载于《第36届神经信息处理系统大会》。Heaton, M., Millane, A. & Rhodes, J.（2024），《在机器学习和深度学习中调整空间相关性》。**网址：** *https://arxiv.org/abs/2410.04312*Hengl, T., Heuvelink, G., Kempen, B., Leenaars, J., Walsh, M., Shepherd, K., Sila, A., MacMillan, R., Mendes de Jesus, J., Tamene, L. & Tondoh, J.（2015），《以250米分辨率绘制非洲土壤属性图：随机森林显著改善了当前预测》，《公共科学图书馆·综合》 **10**，1–26。Hodges, J. & Reich, B.（2010），《加入空间相关误差可能会扰乱你所重视的固定效应》，《美国统计学家》**64**，325–334。Lee, D., Robertson, C., McRae, C. & Baker, J.（2022），《量化空气污染对苏格兰新冠肺炎住院率和死亡率的影响》，《空间与时空流行病学》**42**，100523。Leroux, B., Lei, X. & Breslow, N.（2000），《小区域疾病率：一种用于空间依赖的新型混合模型》，施普林格出版社，纽约，章节《流行病学、环境与临床试验中的统计模型》，Halloran, M 和 Berry, D（编），第135–178页。Liao, M., Zhang, S., Wolf, K., Bolte, G., Laxy, M., Schwettmann, L., Peters, A., Schneider, A. & Kraus, U.（2025），《环境空气污染与自我感知健康状况的长期关联：基于人群的KORA-FIT研究结果》，《国际卫生与环境健康杂志》**264**，114513。

MacBride, C., Davies, V. & Lee, D.（2025年），《面向小区域空间预测的空间自回归随机森林算法》，《应用统计年鉴》**19**，485–504。Moran, P.（1950），《连续随机现象札记》，《生物统计学》**37**，17–23。Nutsford, D., Pearson, A. & Kingham, S.（2013），《一项探讨城市绿地可及性与心理健康之间关联的生态学研究》，《公共卫生》**127**，1005–1011。Orellano, P., Kasdagli, M., Pérez-Velasco, R. & Samoli, E.（2024年），《颗粒物长期暴露与死亡率：世界卫生组织全球空气质量指南系统综述与荟萃分析的更新》，《国际公共卫生杂志》**69**。Patelli, L., Cameletti, M., Golini, N. & Ignaccolo, R.（2024年），回归随机森林中的空间依赖探索路径：分类体系与系统综述，第467–489页。Riebler, A., Sørbye, S., Simpson, D. & Rue, H.（2016），《一种直观的贝叶斯空间疾病制图模型，并考虑了尺度调整》，《医学研究统计方法》**25**，1145–1165。Rue, H., Martino, S. & Chopin, N.（2009），《使用积分嵌套拉普拉斯近似对潜在高斯模型进行近似贝叶斯推断（附讨论）》，《皇家统计学会杂志》*B*辑**71**，319–392。Saha, A., Basu, S. & Datta, A.（2023年），《空间相关数据的随机森林》，《美国统计学会杂志》**118**，665–683。Wang, X., Yue, Y. & Faraway, J.（2018），使用*INLA*进行贝叶斯回归建模，Chapman and Hall/CRC。Watanabe, S.（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**，3571–3594。Wikle, C和Zammit-Mangion, A.（2023年），《空间和时空数据的统计深度学习》，《统计学及其应用年度评论》**10**，247–270。Wright, M. & Ziegler, A.（2017），《ranger：一种在C++和R中快速实现高维数据随机森林的方法》，《统计软件杂志》**77**，1–17。Zhan, W. & Datta, A.（2025年），《地理空间数据的神经网络》，《美国统计学会杂志》**120**，535–547。Zhu, D., Liu, Y., Yao, X. & Fischer, M.（2022年），《空间回归图卷积神经网络：空间多元分布的深度学习范式》，《地理信息学》**26**，645–676。