arXiv:2303.02050v1 [stat.ME] 3 Mar 2023

利用低成本传感技术进行环境场预测的自适应空间抽样设计

Eun-Hye Yoo，纽约州立大学布法罗分校地理系，美国

Andrew Zammit-Mangion，澳大利亚卧龙岗大学数学与应用统计学院

Michael G. Chipeta，牛津大学大数据研究所地理空间流行病学研究员，英国

摘要

过去十年间，可用于监测和预测环境现象的数据源呈爆炸式增长。尽管已有多种推断方法通过整合这些数据，对底层过程进行预测，但当需要补充来自其他异质数据源的数据时，尚未开发出相应的最优抽样设计。本文提出一种基于效用函数的自适应空间设计策略，该函数综合考虑预测不确定性和风险因素准则。预测不确定性通过一种基于固定秩克里金法的空间数据融合方法获得，该方法能够处理空间支持域和信噪比各异的数据。本文重点探讨将低成本便携式传感器用于空气污染监测的应用；这类传感器往往噪声相对较大，同时还可获得监管监测站及数值建模系统的数据。我们发现，空间自适应采样设计有助于改进预测并降低预测不确定性，但只有在低成本便携式传感器数量和质量足够时，它们才可能发挥作用。我们的结论基于一项多因素模拟实验，以及对纽约州西部伊利县和尼亚加拉县污染物进行的逼真模拟。

关键词：自适应空间抽样设计、支持域变换问题、固定秩克里金法、低成本便携式空气传感器、测量不确定性

1 引言

新技术的发展正在迅速改变环境科学的面貌。环境传感硬件和软件的成本、尺寸和重量不断降低，可靠性不断提高，推动了这一发展（Rundel 等，2009）。公民科学和机器人系统的兴起，也促进了传感技术作为基础数据收集工具在环境科学中的应用。例如，机器人系统如今已用于探索深海、追踪有害藻华及污染扩散、监测气候变量，以及研究偏远地区的火山（Dunbabin 和 Marques，2012）。尽管低成本便携式传感器日益普及，观测覆盖不完整和数据质量不一致的问题仍然制约着环境监测（美国环境保护署，2018）。

过去十年间，人们为解决数据稀疏问题付出了大量努力，其中包括引入遥感仪器提供的辅助数据（例如气溶胶光学厚度（AOD）测量值），以及数值模型提供的数据（例如社区多尺度空气质量（CMAQ）建模系统）。近来，先进的传感技术使研究人员和公众都能够使用低成本便携式传感器，在感兴趣的位置采集实时空气质量测量值。有些传感器可穿戴，另一些则安装在公共汽车或无人驾驶飞行器上（Kim et al. 2010, Villa et al. 2016）。这些高度便携的空气质量传感器灵活且价格实惠，因此能够以精细的空间和时间尺度，以移动或静态配置部署可扩展的高密度空气质量传感器网络（Mead et al. 2013, Apte et al. 2017）。这些新技术拓展了联邦和州级监管监测以外的诸多应用领域，不过目前仍处于发展初期。此外，各类仪器的准确度差异很大（美国环境保护署 2018）。

传感技术的日益普及可能会给空气污染暴露评估带来重大变化，但关于传感器部署的一些关键问题仍有待解决，例如针对特定研究区域需要多少传感器，以及如何布设传感器。环境网络设计领域已有大量文献，并提供了一些解决方案；其中包括最大熵设计方法（Bernardo 1979, Zidek 等，2000, Le & Zidek 2006, Fuentes 等，2007）；旨在最优估计变异函数模型参数的空间设计（Bras & Rodríguez-Iturbe 1976, Cressie 等，1990）；以及旨在最小化感兴趣区域内平均克里金方差或最大克里金方差的设计。后者还可进一步分为两类：一类是在几何空间填充设计中假设变异函数模型已知（Nychka & Saltzman 1998, Müller 2007），另一类则考虑随空间或时间变化的变异函数模型（Wikle & Royle 1999, Romary 等，2011）。然而，现有的空间和时空抽样方法是为州级或联邦级监管监测站（SLAMS，https:// www3.epa.gov/airquality/montring.html）设计的，这些监测站能够提供准确的测量值，并满足联邦参考方法等效性要求。另一方面，近期现场研究表明（Castell 等，2017, Kelly 等，2017, Zimmerman 等，2018），由低成本便携式传感器采集的新空气质量样本，其质量可能不够一致和可靠。考虑到测量不确定性的低成本便携式传感器网络最优设计

对于数据收集是必要的。

Fan- shawe & Diggle（2013）提出了一种灵活的空间采样方法——“自适应抽样设计”，用于基于模型的地统计学框架下的空间预测。最近，Chipeta 等人（2016）展示了该方法在资源受限情境中的应用，用于绘制马拉维南部马杰特野生动物保护区周边社区的基于调查的疟疾患病率地图。自适应空间或时空抽样设计会在一系列抽样时间内选取固定数量的采样位置；在每个步骤中，数据分析都基于扩充样本数据进行。灵活的自适应设计使研究人员能够通过概率模型，使用经过精心挑选且动态变化的监测地点子集。这些模型通过利用数据以及所研究过程中存在的空间和／或时间相关性来构建。Fan- shawe & Diggle（2013）和 Chipeta 等人（2016）均采用了基于模型的地统计学，但更通用的空间和时空预测模型也同样适用。

在空气质量建模中，已有多种方法用于整合来自不同数据源的数据，这些数据源在空间分辨率和质量方面各不相同。具体而言，监管空气监测站提供的点测量通常非常准确；而空间分辨率为 1 至 36 km 的网格化代理变量（如气溶胶光学厚度或 CMAQ 输出），以及来自低成本便携式传感器的点测量，则可靠性较低，或容易产生偏差和误差。空气污染数据整合的一项关键挑战，在于多源数据具有不同的空间分辨率，这通常被称为支持域变换问题。例如，贝叶斯融合（Fuentes & Raftery 2005）假设，来自 SLAMS 网络的观测数据与网格化代理数据均源自同一连续潜在场；通过同时考虑这两个数据源，该方法能够处理空间错位和支撑域转换问题（Gelfand 等人 2001，Gotway & Young 2002，Wikle & Berliner 2005）。这种方法足够灵活，可以考虑多个代理变量和多种污染物，但所提出的模型往往很快就会出现参数过多的问题（Chang 2016）。另一方面，统计校准将网格化代理变量用作线性模型中的预测变量。这种建模选择显著降低了计算负担，因为只需考虑与地面观测位置重合的部分代理变量。Berrocal 等人（2010）和 Paciorek（2012）的研究表明，与贝叶斯融合相比，统计校准产生的预测误差相对较小，不过代理变量中不允许存在缺失值。在实际应用中，这往往难以实现；例如，气溶胶光学厚度产品可能因恶劣天气条件而出现若干缺失值。但需要注意的是，大多数支撑域转换问题的解决方案仅用于对两个数据源建模，并未纳入稳定性和灵敏度都值得质疑的低成本便携式传感器测量值（Lewis & Edwards 2016）。

此外，迄今为止，这些用于解决多源空间数据融合中支持域转换问题的统计模型尚未用于环境网络设计。本文提出了一种抽样设计框架，用于利用低成本传感器监测空气污染物的长期（数月至数年）浓度；这些传感器可补充来自 SLAMS 网络的监测数据，以及来自 CMAQ 的网格化代理数据，并采用支撑域变换模型。具体而言，我们采用空间固定秩克里金法（FRK）框架（Cressie & Johannesson 2008），生成细尺度空气污染物的预测标准误差表面，并据此对低成本便携式传感器进行最优布设。不同

数据源的空间分辨率差异以及不同测量变量的不确定性，均由模型隐式纳入考量。在多尺度时空数据融合框架下提出的这种空间网络设计方法具有通用性，有望适用于多种应用场景，助力利用低成本便携式传感器及其他数据源加深我们对环境的认识。

2 方法

考虑一个均值为 $E(Y(\mathbf{s})) = \mu_Y(\mathbf{s})$、协方差函数为 $C_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{s}') \equiv \text{cov}(Y(\mathbf{s}), Y(\mathbf{s}') \mid \boldsymbol{\theta})$ 的空间高斯过程 $\{Y(\mathbf{s}), \mathbf{s} \in D \subset \mathbb{R}^2\}$，其中 $\boldsymbol{\theta}$ 是参数向量，且$\mathbf{s},\mathbf{s}' \in D$。这里， D 表示空间域，过程 $Y(\cdot)$ 在该空间域上通过以下三类空气污染数据被观测：

1. 在 n_Z 州和地方空气质量监测站使用监管仪器测得的污染物浓度。我们将这些测量值记为 $\mathbf{Z} \equiv (Y(\mathbf{s}_{Z_i}) + \epsilon_{Z_i} : i = 1, \ldots, n_Z)'$，其中 $\{\epsilon_{Z_i}\}$ 是一组独立同分布（i.i.d.）的高斯测量误差。

2. 来自数值大气模型（如 CMAQ）的网格化代理数据，表示网格单元上空气污染的区域平均浓度。设 Y_B 表示 Y 在网格单元 B上的区域平均值，它与过程 $Y(\cdot)$ 的关系如下：

$$Y_B \equiv Y(B) = \frac{1}{|B|} \int_B Y(\mathbf{s}) \mathrm{d}\mathbf{s}.$$

大气模型输出可能受到模型参数设定或输入数据相关误差的影响。因此，我们对可用的区域平均模型输出进行建模，并将测量误差纳入其中。我们将模型输出记为 $\mathbf{Q} \equiv (Y(B_i) + \epsilon_{Qi} : i = 1, \ldots, n_Q)'$ 其中 $\{\epsilon_{Qi}\}$ 是一组高斯测量误差，这些误差不一定彼此不相关。

3. 使用低成本便携式传感器网络获取的经校准（即无偏）污染物浓度测量值。我们将这些测量值记为 $\mathbf{X} \equiv (Y(\mathbf{s}_{X_i}) + \epsilon_{X_i} : i = 1, \ldots, n_X)'$ 其中 $\{\epsilon_{X_i}\}$ 是一组独立同分布的高斯测量误差。

我们假设便携式传感器采样点的数量大于或等于监管监测站的数量；即 $n_X \geq n_Z$。我们还假设原始数据经过了转换，使其经验边缘分布近似服从高斯分布。从我们的模型可以看出，给定底层过程 $Y(\cdot)$ 时，数据 $\mathbf{Z}$ 和 $\mathbf{X}$ 条件独立。这并不意味着它们在边缘上不相关，而是说它们所呈现的任何相关性都源于过程 $Y(\cdot)$。

2.1 自适应空间采样设计

在环境传感器网络设计问题中，目标是从潜在抽样地点中选择一个有限的位置集合 $\mathcal{S}^X \equiv \{\mathbf{s}_{X_1}、\mathbf{s}_{X_2}、\ldots\ldots、\mathbf{s}_{X_{n_X}}\}$，使其在某种意义上对于潜在污染场的空间预测而言达到最优，

潜在采样位置在下文也称为“候选站点”，记为 $\mathcal{S}_0^c = \{\mathbf{s} \in D \backslash \mathcal{S}^Z\}$，位于研究区域 D 内，并且与固定监测站点 $\mathcal{S}^Z \equiv \{\mathbf{s}_{Z_1}、\mathbf{s}_{Z_2}、\ldots\ldots、\mathbf{s}_{Z_{n_Z}}\}$互不重合。候选站点的定义具有灵活性，既可以定义为无限位置集合，也可以定义为有限位置集合。在选择候选站点时，可以考虑研究的可行性及其多重目标。这些目标可能包括人体健康影响评估（Baldauf 等，2001）、近道路空气质量监测（Baldauf 等，2009），或检测空气污染物浓度的极端值（Chang 等，2007；Chang 与 Lee，2007）。

我们在此采用的算法与 Chipeta 等人（2016）的方法密切相关，但针对从不同来源获得异质数据时的空气污染物场预测问题进行了调整。算法的初始步骤（我们将其称为 $k = 0$）是根据当前的信息状态获取预测不确定性表面。对于这一初始步骤，我们采用第 2.2 节所讨论的空间数据融合方法，在考虑空间自相关的同时，将具有不同不确定性和不同支撑域的多个空间数据集进行合并。在此，对过程 $Y(\cdot)$ 的推断以 $\mathbf{Z}$ 和 $\mathbf{Q}$ 为条件，并随后用于选择低成本便携式传感器的采样位置。具体而言，根据同时取决于预测不确定性以及可能的其他信息的效用函数（见下文公式（2）），从候选集 $\mathcal{S}_0^c$ 中选出一组新的 b 位置。这组在初始步骤中新选定的抽样地点记为 $\mathcal{S}^{X^{(1)}}$，而在这些抽样地点由低成本便携式传感器采集的污染物浓度测量值记为 $\mathbf{X}^{(1)}$。

如果批次大小 b 小于低成本便携式传感器测量总数（$b < n_X$），则自适应抽样算法通过使用 $\mathbf{Z}$，$\mathbf{Q}$ 和 $\mathbf{X}^{(1)}$对 $Y(\cdot)$ 进行推断，进入步骤 $k = 1$。随后，根据效用函数从候选集合 $\mathcal{S}_1^c = \mathcal{S}_0^c \setminus \mathcal{S}^{X^{(1)}}$ 中的剩余位置中选取一组新位置 $\mathcal{S}^{X^{(2)}}$，并重复该过程，直到没有剩余便携式传感器可供放置。在步骤 k，从剩余候选站点中选择额外的采样位置，使效用在以下约束下达到最大值：$\mathcal{S}^{X^{(k+1)}}$ 中的任意两个新站点之间的距离均不小于 δ_d。在抽样设计中施加最小距离约束，是为了避免传感器网络在效用较高的区域内聚集。自适应抽样设计在每个步骤所选抽样地点数量方面具有灵活性，即每个步骤都可以考虑不同数量的抽样地点 b。请注意，如果将批次大小设为 1，则抽样过程会重复 n_X 次；而如果 $b = n_X$，则只需一个步骤（初始步骤）。

算法步骤如下。设 $U_k(\mathbf{s})$表示位置 $\mathbf{s}$ 在步骤k($k = 0$，1、……、n_K)处的效用，其中 $n_K = \lceil \frac{n_X}{b} - 1 \rceil$，算子 $\lceil \cdot \rceil$定义为对其参数向上取整。这里，U_0 表示初始步骤的效用。步骤 k、U_k处的效用通过综合截至步骤 k、$\{\mathbf{Z}、\mathbf{Q}、\mathbf{X}_k\}$的全部信息来确定，其中 $\mathbf{X}_k \equiv (\mathbf{X}^{(1)\prime}、\ldots\ldots、\mathbf{X}^{(k)\prime})'$，且 U_0 仅是 $\mathbf{Z}$ 和 $\mathbf{Q}$ 的函数（因为 $\mathbf{X}_0 = \mathbf{X}^{(0)} = \emptyset$）。步骤 k、$\mathcal{S}^{X^{(k+1)}}$处的新位置集合从所有由大小为 b 或更小的位置集合构成的可能集合中选取

˜自 $\mathcal{S}_k^c$中的位置，该集合记为 $\mathcal{S}_k^c$，并表示为

$$\begin{aligned} \mathcal{S}^{X^{(k+1)}} &= \underset{\mathcal{S}\in\tilde{\mathcal{S}}_k^c}{\arg\max} \left(\sum_{\mathbf{s}\in\mathcal{S}} U_k(\mathbf{s}) \right) \\ \text{subject to} \quad & d(\mathbf{s},\mathbf{s}') < \delta_d,\ \mathbf{s},\mathbf{s}' \in \mathcal{S}, \end{aligned} \tag{1}$$

其中， $|\mathcal{S}|$ 表示 $\mathcal{S}$ 的大小，而 d（s，s′）是s 与s′之间的欧几里得距离。方程（1）通常无法获得闭式最优解。因此，我们通过单点自适应抽样（Chipeta 等，2016），以贪心方式逐一向 $\mathcal{S}^{X^{(k+1)}}$ 添加位置，从而求得次优解。

本文将感兴趣区域内的预测方差作为效用依据。不过，采样位置的效用也可以定义为决策者关注的其他相关变量的函数，或者是针对不同利益相关方的多个目标的组合。例如，Kabaghe 等人（2017）采用超越概率准则，而非预测方差，因为该研究的首要目标是划定疟疾发病率可能超过政策干预阈值或国家阈值的子区域。在空气污染监测领域，研究区域内可能危害居民健康、健康风险较高的区域，可能需要比其他区域更密集的监测，因此，在步骤 k，位置s 的效用形式为

$$U_k(\mathbf{s}) = \sqrt{\text{var}(Y(\mathbf{s})|\mathbf{Z},\mathbf{Q},\mathbf{X}_k)} + \lambda T(\mathbf{s}), \tag{2}$$

其中， λ 是用于确定监测地点s相对重要性的权重， T(s) >0 是风险因素的函数，例如与主要公路的距离。例如， T(s) $= I(R$(s) $< c)$，其中 R(s)是位置s 处的风险函数，而 c 是阈值。

公式（2）中的效用函数很有吸引力，因为它只涉及已知的 $T(\cdot)$以及预测方差；对于我们所考虑的线性高斯模型，在给定 $\boldsymbol{\theta}$的情况下，预测方差仅是数据位置和支撑域的函数，而与数据值本身无关（例如，Rasmussen 和 Williams，2006，第2章）。在设计用于长期空气质量暴露评估的网络时，这一特性非常重要，因为在抽样设计中选择下一组位置之前，反复等待数据收集完成是不可行的。对于此应用，适合首先基于Z 和Q估计 $\boldsymbol{\theta}$，然后通过自适应抽样算法从 $k=0$ 运行至$k=n_K$来获取 $\mathcal{S}^X$，最后一次性部署传感器。

2.2 空间数据融合

公式（2）中效用函数的评估，需要利用来自多个数据源的数据，在具有不同空间支持域且测量误差各不相同的情况下，计算任意位置的预测方差。由于这一预测方差将在优化过程中反复计算，因此计算必须高效。固定秩克里金法（FRK）是一种能够轻松处理支撑域转换问题并满足计算要求的空间建模框架。下面简要介绍

这一计算框架；更多细节请参见 Zammit-Mangion & Cressie (2019)。

首先，考虑过程 $Y(\cdot)$。在 FRK 中，$Y(\cdot)$ 被建模为以下各项之和：（大尺度）固定效应、由随机系数加权的 r 预先指定基函数，以及一个细尺度变异项。具体而言，

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\beta + \boldsymbol{\phi}(\mathbf{s})'\boldsymbol{\eta} + \xi(\mathbf{s}), \quad \mathbf{s} \in D, \tag{3}$$

其中，大尺度变异（或趋势）通过 $\mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$（其已知的 p维协变量向量为 $\mathbf{x}(\mathbf{s}) \equiv (x_1(\mathbf{s}) \ldots, x_p(\mathbf{s}))'$，$\mathbf{s} \in D$）来刻画；中尺度变异通过 r 基函数 $\boldsymbol{\phi}(\mathbf{s}) \equiv (\phi_1(\mathbf{s}) \ldots, \phi_r(\mathbf{s}))'$，$\mathbf{s} \in D$ 来刻画；细尺度变异则通过 $\xi(\mathbf{s})$，$\mathbf{s} \in D$ 来刻画，该项在空间上几乎不相关。系数随机向量 $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_r)'$ 的均值为零，协方差矩阵为 $\mathbf{K}$，用于刻画基函数之间的依赖关系。基函数具有局部性、紧支撑和多分辨率特性：不同分辨率下的基函数系数被建模为相互独立，而同一分辨率下的系数之间具有协方差，该协方差随基函数质心之间距离的增加而呈指数衰减；更多详情请参见 Zammit-Mangion & Cressie (2019)。过程 $\phi(\cdot)'\eta + \xi(\cdot)$ 有时也称为空间随机效应模型。

为处理支撑域转换，方便起见，可考虑定义在由 N 个小区域 $\{A_i$ 组成的细网格上的过程 $Y(\cdot)$（其中 N 很大）：$i = 1, \ldots, N\}$，这些区域定义在 D 上，通常称为基本面积单元（基本分析单元）。在基本分析单元上求得的过程（我们将其收集在向量 $\mathbf{Y} \equiv (Y(A_i) : i = 1, \ldots, N)'$ 中）表示为

$$Y_i \equiv Y(A_i) = \frac{1}{|A_i|}\int_{A_i} Y(\mathbf{s})\mathrm{d}\mathbf{s} = \mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\phi}_i'\boldsymbol{\eta} + \xi_i, \tag{4}$$

对于 $i = 1, \ldots, N$，其中

$$\mathbf{x}_i = \left(\frac{1}{|A_i|}\int_{A_i} x_l(\mathbf{s})\mathrm{d}\mathbf{s} : l = 1, \ldots, p\right)',$$

是一个由对基本分析单元上的协变量取平均值而形成的 $(p \times 1)$ 维协变量向量，

$$\boldsymbol{\phi}_i' = \left(\frac{1}{|A_i|}\int_{A_i} \phi_l(\mathbf{s})\mathrm{d}\mathbf{s} : l = 1, \ldots, r\right)',$$

是对基本分析单元上的基函数取平均值后得到的向量，且

$$\xi_i = \frac{1}{|A_i|}\int_{A_i} \xi(\mathbf{s})\mathrm{d}\mathbf{s},$$

是一个细尺度变异项，假定其在空间上不相关且服从均值为 0、方差为 σ^2_ξ 的正态分布。

考虑点位数据向量 $\mathbf{Z} = (Z(\mathbf{s}_j) : j = 1, \ldots, n_Z)'$。在基本面积单元设定中，我们假设每个数据点都是对单个基本面积单元的一次含噪观测。因此，我们可以写作

$$\mathbf{Z} = \mathbf{H}_Z\mathbf{Y} + \boldsymbol{\varepsilon}_Z, \tag{5}$$

其中， $\boldsymbol{\varepsilon}_Z$ 服从均值为 **0** 、协方差矩阵为 $\sigma^2_{\boldsymbol{\varepsilon}_Z}\mathbf{I}$ 的正态分布，而 $\mathbf{H}_Z$ 是一个每行仅含一个非零元素的关联 矩阵。类似地，当纳入低成本便携式传感器测量值时，

$$\mathbf{X}_k = \mathbf{H}_{X_k}\mathbf{Y} + \boldsymbol{\varepsilon}_{X_k}, \quad k > 0, \tag{6}$$

其中 $\boldsymbol{\varepsilon}_{X_k}$， $k > 0$，服从均值为 **0**、协方差矩阵为 $\sigma^2_{\boldsymbol{\varepsilon}_X}\mathbf{I}$和$\mathbf{H}^{X_k}$， $k > 0$ 的正态分布，且矩阵中每行也只有一个非零元素。回想一下， k 表示自适应空间设计中的步骤，$\mathbf{X}_k$ 表示截至步骤 k（含该步骤）收集的数据，$\mathbf{H}^{X_k}$ 表示与步骤 k中的低成本传感器相对应的关联矩阵，而 $\boldsymbol{\varepsilon}_{X_k}$ 表示相应的测量误差。因此，该观测模型会针对空间设计中的每个 k 进行更新。

现在考虑网格化代理模型Q的输出，例如 CMAQ。网格化输出将覆盖一个或多个基本分析单元（BAU），因此我们可以写成

$$\mathbf{Q} = \mathbf{H}_Q\mathbf{Y} + \boldsymbol{\varepsilon}_Q, \tag{7}$$

此时，$\mathbf{H}^Q$ 每行可以包含多个非零元素，且$\mathbf{H}^Q$ 的每行之和均为 1。和为 1 的约束确保我们将每个网格化模型输出建模为过程均值。在规则网格域中，$\mathbf{H}^Q$ 的非零元素值都相同，且等于模型输出中每个网格单元内嵌套的基本分析单元数量的倒数。

所有观测值均可合并，得到步骤 k 的观测模型，

$$\widetilde{\mathbf{Z}}_k = \mathbf{H}_k\mathbf{Y} + \boldsymbol{\varepsilon}_k, \quad k \geq 0,$$

~其中 $\mathbf{Z}_k \equiv (\mathbf{Z}', \mathbf{Q}', \mathbf{X}_k')'$, $\mathbf{H}_k \equiv (\mathbf{H}_Z', \mathbf{H}_Q', \mathbf{H}_{X_k}')'$, $\boldsymbol{\varepsilon}_k \equiv (\boldsymbol{\varepsilon}_Z', \boldsymbol{\varepsilon}_Q', \boldsymbol{\varepsilon}_{X_k}')'$，并且在情形 $k = 0$ 下，这些定义也有相应的显而易见的修改。因此，对于每个 k， $p(\mathbf{Z}_k \mid \mathbf{Y})$ 服从高斯分布， $p(\mathbf{Y})$ 也如此。利用标准高斯恒等式，可以求得条件期望 ^~~ ~Y $_k \equiv E(\mathbf{Y} \mid \mathbf{Z}_k)$ 和条件方差 $\text{var}(\mathbf{Y} \mid \mathbf{Z}_k)$。模型中出现的未知参数，即 σ^2_ξ 以及 K 中的参数，也需要估计；在参数的最大似然估计与 Y 的预测之间迭代的过程由期望最大化（EM）算法加以形式化；完整细节见 Zammit-Mangion & Cressie (2019)。

当 EM 算法在步骤 k 收敛时，即可得到预测值，以及更重要的预测方差 $\text{var}(\mathbf{Y} \mid \mathbf{Z}_k)$。位置 s 处的预测方差 $\text{var}(Y(\mathbf{s}) \mid \mathbf{Z}_k)$，如公式（2）中所用，仅为包含 s 的 BAU 的预测方差，即 $\text{var}(Y(\mathbf{s}) \mid \mathbf{Z}_k) = \text{var}(Y_i \mid \mathbf{Z}_k)$ for $\mathbf{s} \in A_i$。如第2.1节所述，如果无法进行重复部署和数据收集，则可使用 EM 算法，仅根据 **Z** 和 **Q** 估计参数；随后运行自适应算法，以获得整个集合 $\mathcal{S}^X$ ，而无需收集额外数据。

2.3 性能评估

我们通过考察一组 n_p 验证位置 $\mathcal{S}^* \equiv \{\mathbf{s}_1^*, \ldots, \mathbf{s}_{n_p}^*\}$处的预测准确度和不确定性量化表现，来评估自适应采样方案的效率。作为仅用于评估预测准确度的指标，我们使用均值

绝对预测误差（平均绝对百分比误差，MAPE），其定义为 $MAPE = \frac{1}{n_p}\sum_{p=1}^{n_p}|Y(\mathbf{s}_p^*) - \hat{Y}(\mathbf{s}_p^*)|$，均方根预测误差（RMSPE），其定义为

$$RMSPE = \sqrt{\frac{1}{n_p}\sum_{p=1}^{n_p}(Y(\mathbf{s}_p^*) - \hat{Y}(\mathbf{s}_p^*))^2},$$

以及由下式给出的平均预测误差 (MPE)

$$MPE = \frac{1}{n_p}\sum_{p=1}^{n_p}(Y(\mathbf{s}_p^*) - \hat{Y}(\mathbf{s}_p^*)),$$

其中，$Y(\mathbf{s}_p^*)$ 和 $\hat{Y}(\mathbf{s}_p^*) = E(Y(\mathbf{s}_p^*) \mid \mathbf{Z}^{n_K})$ 分别表示验证位置 $\mathbf{s}_p^*$ 处底层过程的实现，以及自适应采样完成时根据三种不同数据源得到的预测。我们使用连续排序概率评分（CRPS）作为评估不确定性量化性能的指标；有关这一适当评分规则的详情，请参见 Gneiting 和 Raftery（2007）。在第3节中，我们使用这些指标比较第2.1节所述的自适应采样策略与纯随机采样策略。

3 案例研究

本节通过两个实验案例研究，评估低成本便携式传感器网络设计对预测质量的影响。在第一个案例中，我们考察在底层过程平稳且各向同性、区域由正方形围定的理想环境下，预测性能对信噪比 (SNR) 和待部署低成本便携式传感器数量的灵敏度。在第二个案例研究中，我们开展观测系统模拟实验，其中过程为 CMAQ 在美国纽约州西部伊利县和尼亚加拉县（布法罗-尼亚加拉地区）的模拟输出。在此案例中，区域的几何形状更为复杂，并且在自适应空间抽样设计中明确考虑了多个风险因素。

统计分析和自适应空间采样均在 R (v.3.5.1) 中进行（R 核心团队 2018）。我们使用 geosample 软件包 (v.0.2.1) 进行自适应空间采样；使用 FRK 软件包 (v.0.2.2) 进行数据融合；使用 verification 软件包计算连续排名概率得分；并使用 rsm 软件包 (v.2.10) 分析实验设计因素之间的交互效应（Lenth 等人 2009）。

3.1 模拟实验

我们考虑的研究区域为 $D \equiv [0,1] \times [0,1]$。我们假设底层过程具有零均值和各向同性指数协方差函数，定义为 $C_{\boldsymbol{\theta}}(\mathbf{h}) \equiv \text{cov}(Y(\mathbf{s})\ Y(\mathbf{s}+\mathbf{h}) \mid \boldsymbol{\theta}) = \sigma_Y^2 \exp(-||\mathbf{h}||/\tau)$，其中 $\boldsymbol{\theta} \equiv (\sigma_Y^2, \tau)'$。具体而言，我们假设该过程的方差为 1（$\sigma_Y^2 = 1$），e折叠长度为 0.3（$\tau = 0.3$）。我们在 D 上的 100×100 规则网格上获得了该过程的多个实现，并且

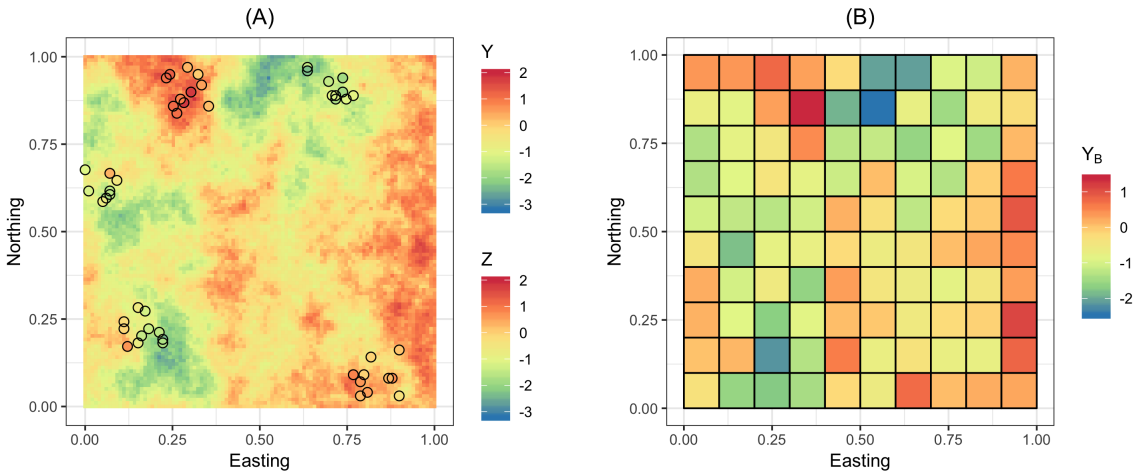


图1：（A）在 $D = [0,1] \times [0,1]$ 上的 100×100 网格上模拟的过程，以及监管监测地点的测量值**Z** 。（B）在更粗的 10×10 网格上计算的过程平均值$\mathbf{Y}_B$。

对于底层过程 **Y** 的每个实现，我们分别生成了两个合成 $PM_{2.5}$浓度数据集，记为 **Z** 和 **Q**。

我们假设生成的数据是无偏的，但包含特定来源水平的测量误差。具体而言，监管固定监测数据**Z**通过获取 $n_Z = 50$ 位置处的过程值，并加上由均值为零、协方差矩阵为 $\sigma^2_{\epsilon_Z}\mathbf{I} \equiv \frac{\sigma^2_Y}{SNR_Z}\mathbf{I}$的高斯分布生成的测量误差 ϵ_Z 来生成。**Z** 的信噪比设为9（$\text{SNR}_Z = 9$）。50个监管固定监测站点的位置以非规则形式选取，以模拟现有的监测网络。具体而言，我们在五个空间聚类中选取了50个站点，如图1（A）所示。

为生成网格数据，我们首先在 D 上将过程值平均到较粗的 10×10 网格中，从而得到一个过程平均值向量，记为 $\mathbf{Y}_B$ （即 $\mathbf{Y}_B \equiv \mathbf{H}_Q\mathbf{Y}$）。随后，按照公式（7），通过向 $\mathbf{Y}_B$ 添加 ε_Q 得到数据 **Q** ；为简单起见，假设 ε_Q 不相关、均值为零，且信噪比为 1（$\text{SNR}_Q = 1$）。研究区域内底层过程**Y** 的一次实现如图1(A)所示，背景为曲面，50 个监测站的地面观测以其上的圆形符号表示；图1(B)则显示了在较粗网格下对同一过程取平均后的结果。选定新的采样位置后，我们以类似于数据 **Z** 的方式生成测量值 **X** ，即提取所选位置处的过程值，并添加均值为零、协方差矩阵为 $\sigma^2_{\epsilon_X}\mathbf{I} = \frac{\sigma^2_Y}{SNR_X}\mathbf{I}$ 的测量误差 ε_X。在此，我们考虑一系列信噪比值，以评估低成本传感器测量值的数据质量的影响：较大的信噪比（$\text{SNR}_X = 9$）对应测量误差方差较小的数据，例如通过监管仪器获得的测量值；较小的信噪比（$\text{SNR}_X = 1$）则对应由低成本便携式传感器获得的精度较低的测量值。

我们考虑了两种采样策略——批次大小为$b \in \{3, 15, 30\}$的自适应空间采样，以及随机空间采样，用于选取 n_X 个位置。对于后一种方法，采样位置是随机选取的；而对于前一种方法，采样位置则根据

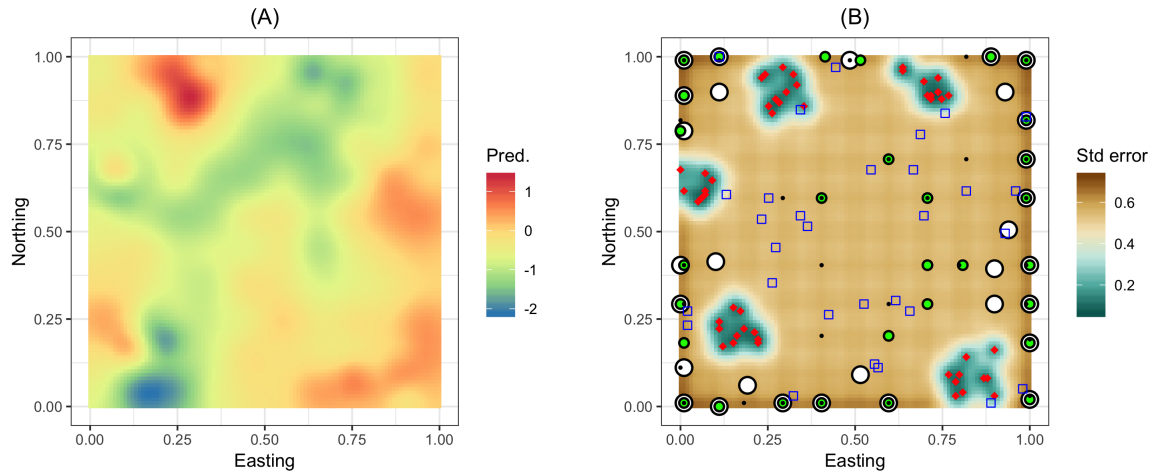


图2：（A）图1中所示基于Z 和Q 得到的FRK预测值。（B）相应的FRK预测标准误差。在（B）图中，我们还展示了监管监测站和低成本传感器的采样位置。FRK预测标准误差较低区域中的红色圆圈表示50个监管监测站。不同大小和填充方式的黑色圆圈表示通过自适应采样选出的采样位置，其批次大小各不相同。白色填充的最大圆圈表示批次大小为 $b = 30$时的采样位置，绿色填充的中等大小圆圈表示批次大小为 $b = 15$时的采样位置，最小的黑点表示批次大小为$b = 3$时的采样位置。随机采样选出的位置以蓝色方块表示。

FRK预测标准误差表面和 $\delta_d = 0.1$的最小间距约束进行选取。基于图1所示实现中仅使用 Z 和Q （即步骤 $k = 0$）所得的预测值和预测标准误差表面分别见图2(A)和图2(B)。图2(B)中还显示了监管监测站的50个监测点（以红色圆圈表示），以及使用随机采样（蓝色方块）和采用不同批量大小 $b \in \{3, 15, 30\}$的自适应采样为便携式传感器选取的额外 $n_X = 30$ 个监测点。

我们的结果表明，在三种批次大小中， $b = 3$ （总样本量的10%）在对 D （总体不确定性最低）取平均后，公式（2）的效用函数取得了最小值。我们进一步考察了批次大小对不同 n_X 和 b情况下总体预测准确度和不确定性的影响；结果汇总于附录A。正如预期， n_X 越大、 b 越小，预测性能越好。值得注意的是，低成本便携式传感器选出的抽样地点集中在高不确定性区域；在本例中，无论选择何种批次大小，这些区域都位于 D 的边界处。Krause等（2008）讨论了对效用函数进行修改以避免这一问题的方法。当效用函数考虑空间分辨风险因素等其他约束时，这种现象通常会消失；参见第3.2节。

上述结果基于单个过程实现。为了得到更具普遍性的结果，我们在多种设置下、针对多个实现，使用第 2.3 节介绍的指标评估预测性能。我们生成了一组 100 次蒙特卡

洛（MC）模拟来模拟底层过程，并在每次模拟中考虑低成本便携式采样因子的每种成对组合，适用于 $n_X \in \{10, 30, 50, 70\}$ 和 $\text{SNR}_X \in \{1, 3, 5, 7, 9\}$。进行这 100 次蒙特卡洛模拟，是为了考虑底层过程可能对预测质量及自适应采样结果造成的影响。对于 100 次过程模拟中的每一次，我们都生成了 $\mathbf{Z}$ 和 $\mathbf{Q}$，并据此使用 FRK 得到预测标准误差表面；对于自适应采样方法，则根据方程（1）和公式（2），结合 $T(\cdot) = 0$，确定 $\mathcal{S}^{X^{(1)}}$。在此实验中，我们固定了 $b = 1$，因为附录A的结果表明，b 越小，预测性能越好。对于每次模拟，都使用$\widetilde{\mathbf{Z}}_1 = [\mathbf{Z}', \mathbf{Q}', X^{(1)}]'$重新运行 FRK，并再次应用自适应采样来选择 $\mathcal{S}^{X^{(2)}}$。这一过程重复进行，直到选出所有 n_X 个样本。选定这些位置后，生成 30 个$\mathbf{X}$ 的随机实现，并针对每个实现计算第 2.3 节所述的诊断指标。最终，每种诊断指标以及 n_X 和 SNR_X的每种组合都得到一张 100×30 表。对于随机采样的情况也得到了这些表，其中 n_X 个样本在 D 中随机放置，并针对每个 SNR_X额外进行了 30 次蒙特卡洛模拟。对于每种情况，均使用 FRK 得到预测值和预测误差曲面。

图3展示了过程的一次单独实现所得到的结果（对应于我们的 100×30 表中的一行）。从第一列（$n_X = 10$）可以看出，当部署的低成本便携式传感器数量较少时，无论信噪比$_X$如何，采样策略都不会影响预测质量。相比之下，如图3第四列所示，当部署大量传感器（$n_X = 70$）时，两种策略在全部四项指标上的差异都很显著。还可以明显看出，当测量值更加可靠（信噪比$_X = 9$）时，相较于随机采样，采用自适应采样的相对优势会增大。因此，正如预期，增加低成本传感器的数量会降低预测误差，无论选择哪种采样策略；不过，自适应采样通常比随机采样更有效，尤其是在低成本便携式传感器数据的不确定性相对较低时。

请注意，图3仅展示了底层过程的一次实现。为评估我们的发现对于所有其他实现的灵敏度，我们按诊断指标比较了自适应采样相较于随机采样所带来的性能提升幅度。具体而言，我们计算了所有实现中与各采样策略（随机采样 – 自适应采样）相关的指标平均差值。当两种采样策略得出的结果相近时，差值应接近于零。图4展示了这项分析的结果，其中平均差值的汇总统计量按指标（行）和样本量（列）组合呈现。总体而言，自适应采样优于随机采样，因为所有平均差值均大于或等于零（位于虚线之上或线上），这表明自适应采样的预测误差小于或等于随机采样的预测误差。正如对单次过程实现的分析所预期的那样，当样本量较大时（$n_X = 70$），采样策略的效果十分明显。例如，图4第四列中的指标显示，无论信噪比$_X$取何值，平均差值都大于零。另一方面，同样明显的是，当样本量较小时（$n_X = 10$）（见图4第一列），低成本传感器的质量（信噪比$_X$）对采样策略所带来的性能增益几乎没有影响，甚至毫无影响。即使样本量

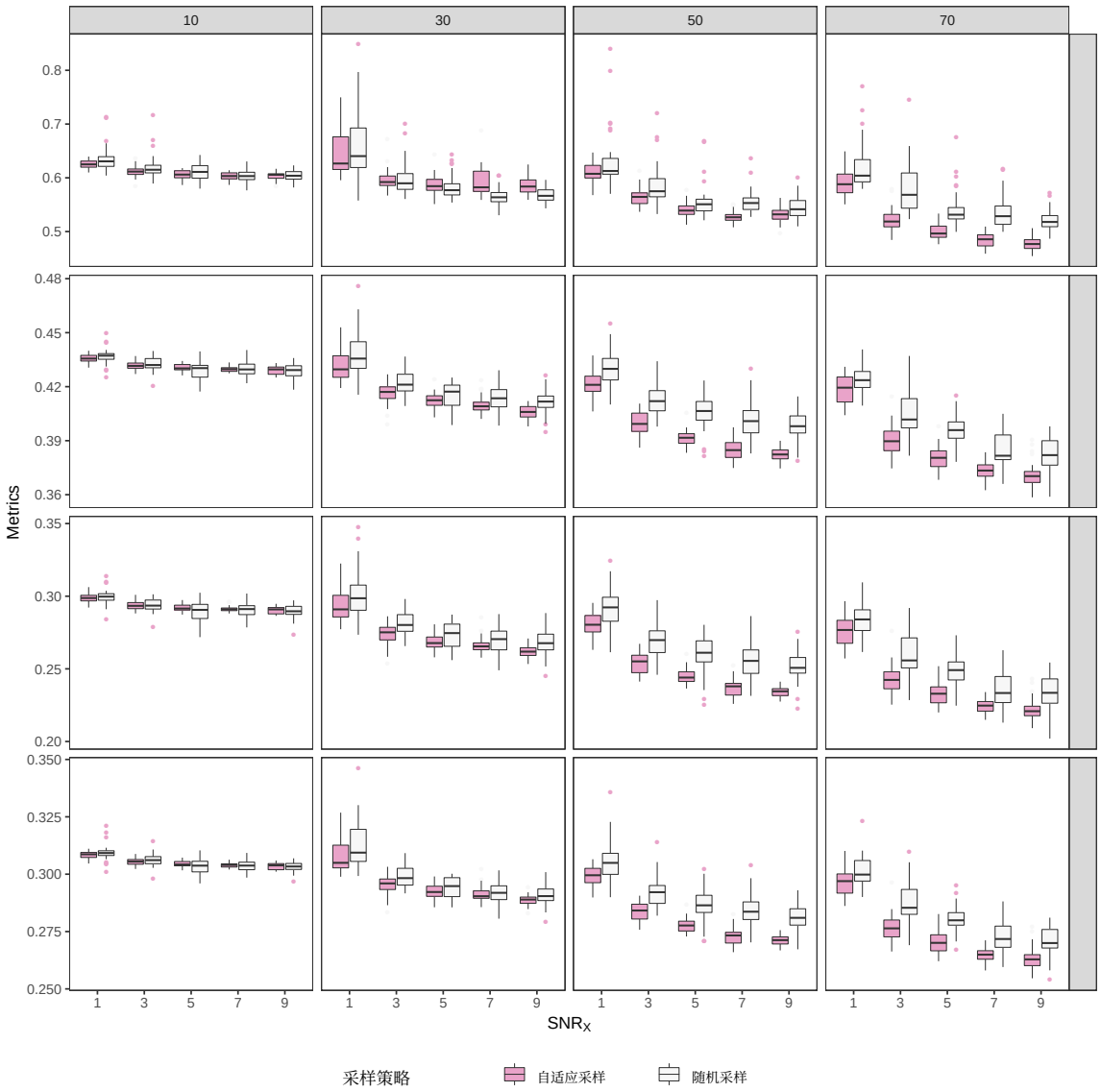


图3：对于 $Y(\cdot)$ 的单次实现，低成本传感器数据的规模和质量对预测准确度的影响。在每个子图中，通过 30 次蒙特卡洛模拟的箱线图，评估了信噪比$_X$ 五个不同水平下的采样效果。左侧的箱线图（阴影部分）对应自适应采样，右侧的箱线图（无阴影部分）对应随机采样。

足够大（$n_X = 70$），测量值质量（信噪比$_X$升高）对自适应采样所带来的性能提升的影响也并非线性。随着信噪比$_X$ 升高（例如从1升至3），性能增益会迅速增加；但尽管信噪比$_X$进一步升高，性能提升仍只是逐步增加。

最后，我们通过响应面分析，评估了样本量（n_X）与低成本传感器测量值的数据质量（信噪比$_X$）之间的交互作用对自适应采样性能的影响，评估依据为四项指标。图5以等高曲面图的形式展示了这四项指标的响应面。显然，随着传感器数量 n_X 增加，

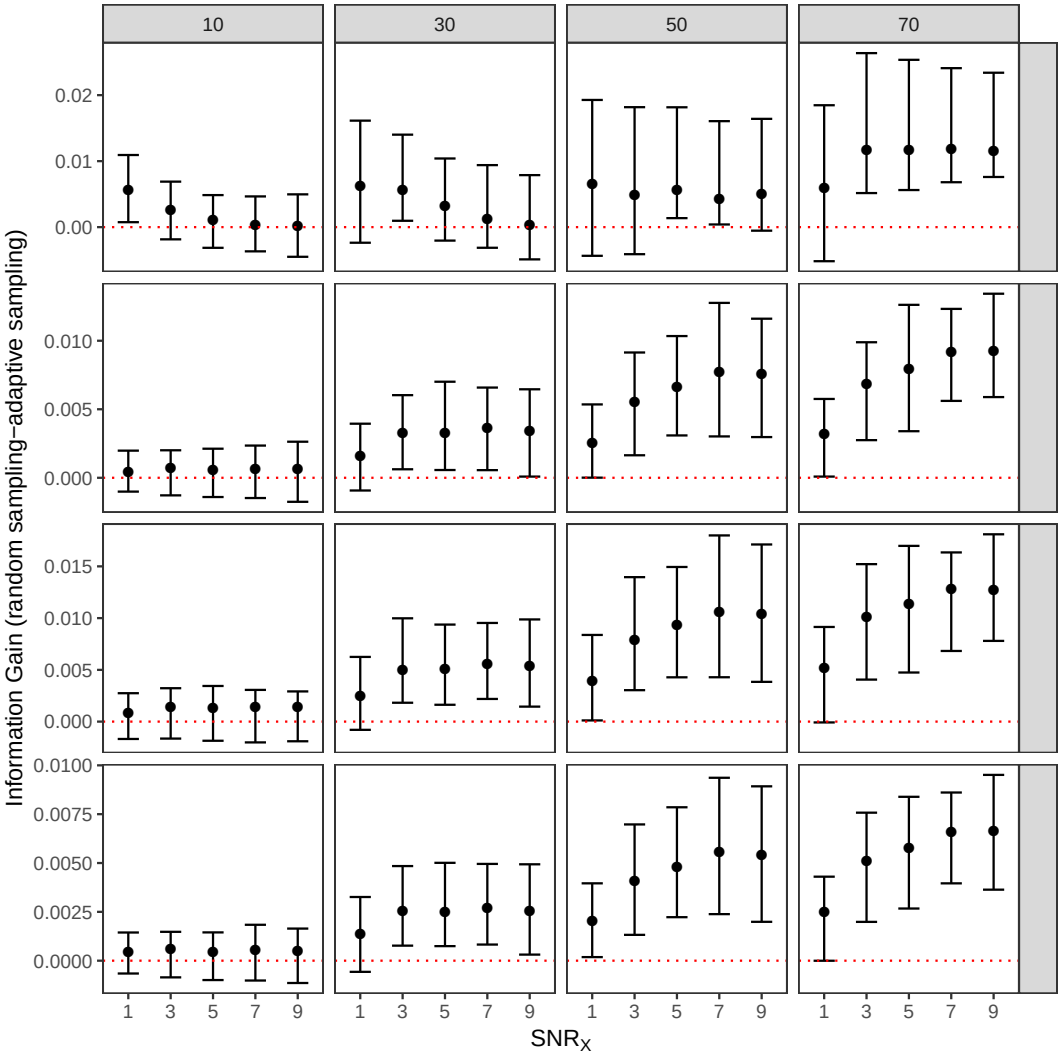


图4：两个采样策略相关指标的均值差值（随机采样 – 自适应采样），基于 100×30次模拟，并按 n_X和 SNR_X展示。圆点表示指标差值的均值，上下误差线表示差值的四分位距。

数据质量也会提高（信噪比$_X$ 增加），四项指标均有所改善。有趣的是，这些函数的形状相当相似，因此也可以考虑利用这些曲面，在成本效益分析中预测增加 n_X和信噪比$_X$ 所能带来的潜在性能提升。

3.2 伊利县和尼亚加拉县 PM2.5 浓度的观测系统模拟实验

在本实验中，我们在美国纽约州西部的伊利县和尼亚加拉县上方的规则 $1 \times 1\,\text{km}^2$网格上模拟了过程值。然而，与第3.1节中使用的无条件模拟不同，此处的过程值是在给定每日 PM2.5

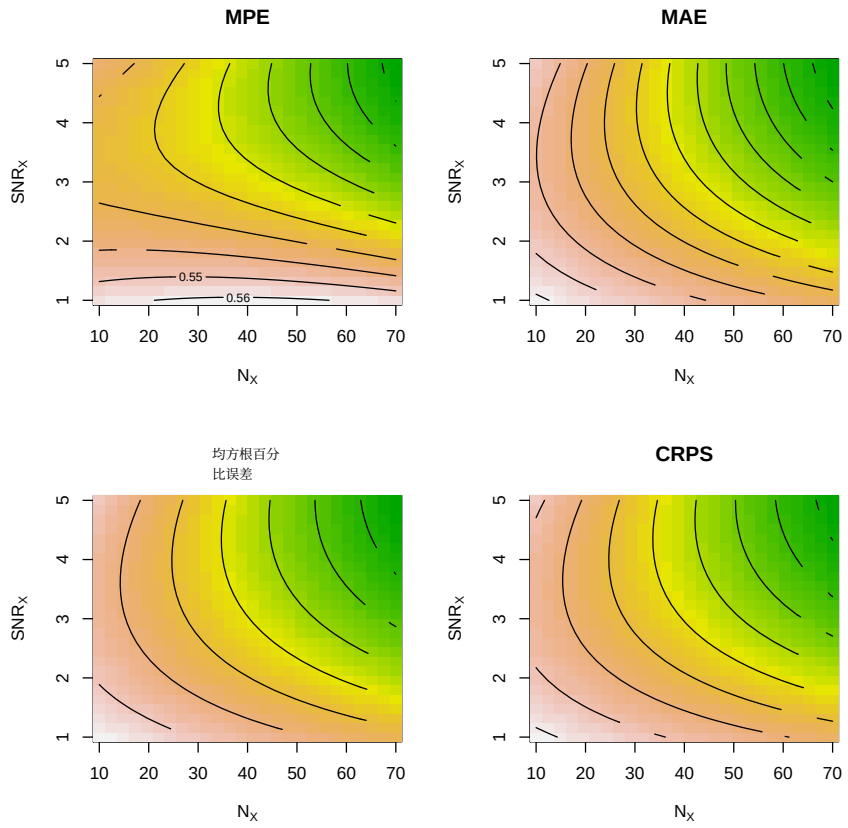


图5：低成本传感器测量的样本量（n_X）和数据质量（信噪比$_X$）对四项指标的交互作用。

质量浓度的条件下模拟的，这些浓度来自 2011年9月2日的一次 CMAQ 模型模拟。具体而言，首先将 CMAQ输出转换为点参考数据，将数据集中到 CMAQ网格单元的几何质心上。随后，以点参考的 CMAQ输出为数据，使用普通克里金法生成$1\times1\,\text{km}^2$网格上的 PM2.5浓度面。采用了常数均值和指数协方差函数 $C_{\boldsymbol{\theta}}(\mathbf{h}) = \sigma_Y^2 \exp(-||\mathbf{h}||/\tau)$，其中 $\sigma_Y^2 = 0.76$ 和 $\tau = 35.8$ km。该过程的一个条件实现（我们将其视为真实过程）如图 6(A) 所示。

我们将 $1\times1\,\text{km}^2$ 网格单元视为基本分析单元（BAU），因此，空气质量监测站和低成本便携式监测点的所有点测量值均归属于某个网格单元。2011 年，研究区域内共运行了七个现有环境空气监测站 $\{$个z_i： $i = 1$、.....、 $7\}$，如图6(A)所示。我们通过获取与七个监测站位置重合处的过程值，并加入均值为零、协方差矩阵为 $\sigma^2_{\boldsymbol{\varepsilon}_Z}\mathbf{I} \equiv \frac{\sigma_Y^2}{SNR_Z}\mathbf{I}$、信噪比为 $_Z = 7$ 的测量误差 ϵ_Z，生成这七个监测站的地面PM2.5测量值。通过对 12×12 km^2 规则网格 $\{B_i$ 上的过程值取平均，并加入均值为零且方差随空间变化的不相关测量误差 $\boldsymbol{\varepsilon}_Q$，生成代表CMAQ测量值的模拟面域数据： $i = 1$、.....、 $51\}$。鉴于CMAQ是基于大气科学和空气质量建模技术的确定性模型，且其纳入了多个排放源及大气化学过程，

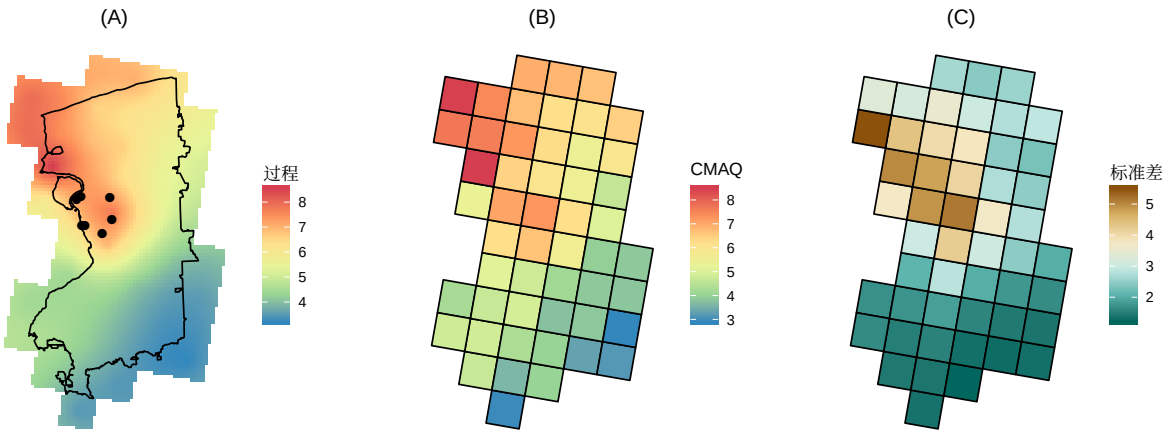


图 6：(A) 伊利县/尼亚加拉县 $PM_{2.5}$ 浓度的模拟过程，以及州级环境空气监测站的位置（黑点）。(B) 代表 CMAQ 输出的模拟 $PM_{2.5}$浓度。(C) 与模拟 CMAQ 输出相关的不确定性。

表1：过程值和模拟数据的汇总统计量

	Min.	第一四分位数	中位数	Mean	第三四分位数	Max.	SD	N
过程值	3.26	4.08	5.26	5.37	6.66	8.52	1.41	4990
监管监测站	6.25	6.68	7.41	7.16	7.53	7.65	0.51	7
CMAQ 输出	2.95	4.09	5.36	5.42	6.56	8.45	1.46	51

模拟输出的不确定性无法直接获得。因此，我们改用过程的时间变异性来表示不确定性，方法为 ε_Q。具体而言，我们汇总了每个像素连续15天的每日CMAQ值的变异性——即2011年9月2日前后各七天的数据——以获得不确定性度量。图 6(B) 和图 6(C) 分别展示了每日CMAQ结果及其对应的时间变异性（以经验标准差表示）。

表1汇总了模拟过程值的统计量，以及在七个空气质量监测站和 CMAQ 网格上获得的模拟浓度。整个区域和空气监测站的过程均值分别为 5.37 和 7.16；后者的取值范围（6.25 至 7.65）比模拟值的范围（3.26 至 8.52）窄。这在一定程度上是意料之中的，因为监管监测站集中在区域的一小部分。另一方面，CMAQ $PM_{2.5}$ 浓度的均值和标准差分别为 5.42 和 1.46，与模拟过程的相应数值（分别为 5.37 和 1.41 ）非常接近。

自适应抽样算法的第一步要求我们使用Z 和Q通过 FRK 获取预测不确定性。对这些数据运行 FRK 得到的预测值和预测标准误分别见图 7(A) 和 7(B)。

七个现有监测站的地面观测值对预测不确定性的影响显而易见：在研究区域中部偏西、七个美国环境保护署监测点所在区域之外，预测误差方差相对较高且分布均匀。

对于自适应采样，我们在站点选择过程中纳入了其他风险因素，以表示空间变化需求。这些风险因素包括一组老年人口高度集中的邮政编码单元（如图 7(C) 所示），以及与主要公路的邻近程度（公路周围 2 km 的缓冲区，如图 7(D) 所示）。具体而言，我们选取了 75 岁及以上人口比例位于前十个百分位的邮政编码单元。令 $R_1(\cdot) = 1$ 位于高风险邮政编码区域内，否则为零；令 $R_2(\cdot) = 1$ 位于距公路不超过 2 km 的区域内，否则为零。我们令 $T(\cdot) = R_1(\cdot) \times R_2(\cdot)$，并将公式（2）中的 $\lambda = 1.0$ 设定为相应值，以使便携式传感器几乎不可能被放置在风险区域之外。$T(\cdot) > 0$ 所在的空间位置在图 7(E) 中以红点下方的彩色网格单元表示。网格单元的颜色表示高风险区域内的预测不确定性。

自适应空间采样是在 $n_X = 20$ 条件下进行的，距离限制为$\delta_d = 3$km。采用 $b = 1$ 的自适应采样流程确定了 20 个低成本传感器采样位置；这些位置在图 7(E) 中以红点表示。在算法的每个步骤中，都会在新选定的位置获取过程值，并加入信噪比为$_X = 4$的测量误差，以生成新的低成本传感器测量值。

对全部三个数据集运行 FRK 后得到的预测表面和预测标准误差表面分别见图 8(A) 和图 8(B)。预测值介于 3.19 和 7.86 之间，范围略窄于过程值的范围（3.26 至 8.52）。预测值的均值（5.57）和标准差（1.26）均与过程的相应统计量一致（见表1）。与此同时，使用和不使用低成本传感器测量值所得预测之间的差异似乎并不显著；比较图 7(A) 与图 8(A)。这可能是因为底层过程相对平滑。不过，从预测标准误差来看，新样本数据的影响十分明显——现在，在较大区域内（包括放置便携式传感器的区域），预测标准误差较低（约为 0.2 和 0.3）；比较图 8(B) 与图 7(B)。我们还通过计算各预测位置处的预测误差，评估了使用三种数据源进行预测的准确度。图 8(C) 所示的预测误差表明，预测值与底层过程总体吻合良好（均值为 -0.09，标准差为 0.17）。

为了更好地了解低成本便携式传感器测量对预测准确度的贡献，同时考虑低成本便携式传感器测量中测量误差所带来的不确定性，我们针对与 $\mathbf{X}$ 相关的测量误差进行了 100 次不同模拟，并开展了上述实验。这里，我们固定了通过初始自适应采样策略获得的采样位置，如图 7(E) 所示。对于每次实现，我们使用第 2.3 节讨论的四项指标评估预测准确度，并将结果与仅基于 $\mathbf{Z}$ 和 $\mathbf{Q}$ 两个数据集的结果进行比较。如表 2 所示，纳入低成本传感器测量值的三个数据集所得到的 FRK预测大幅提升了预测质量，但并非总是如此。例如，100 次蒙特卡罗运行所得平均预测误差的 0.75 分位数高于仅使用两个数据集时所得的平均预测误差。另一方面，纳入额外数据进行预测后，平均绝对误差和均方根百分比误差均显著改善。

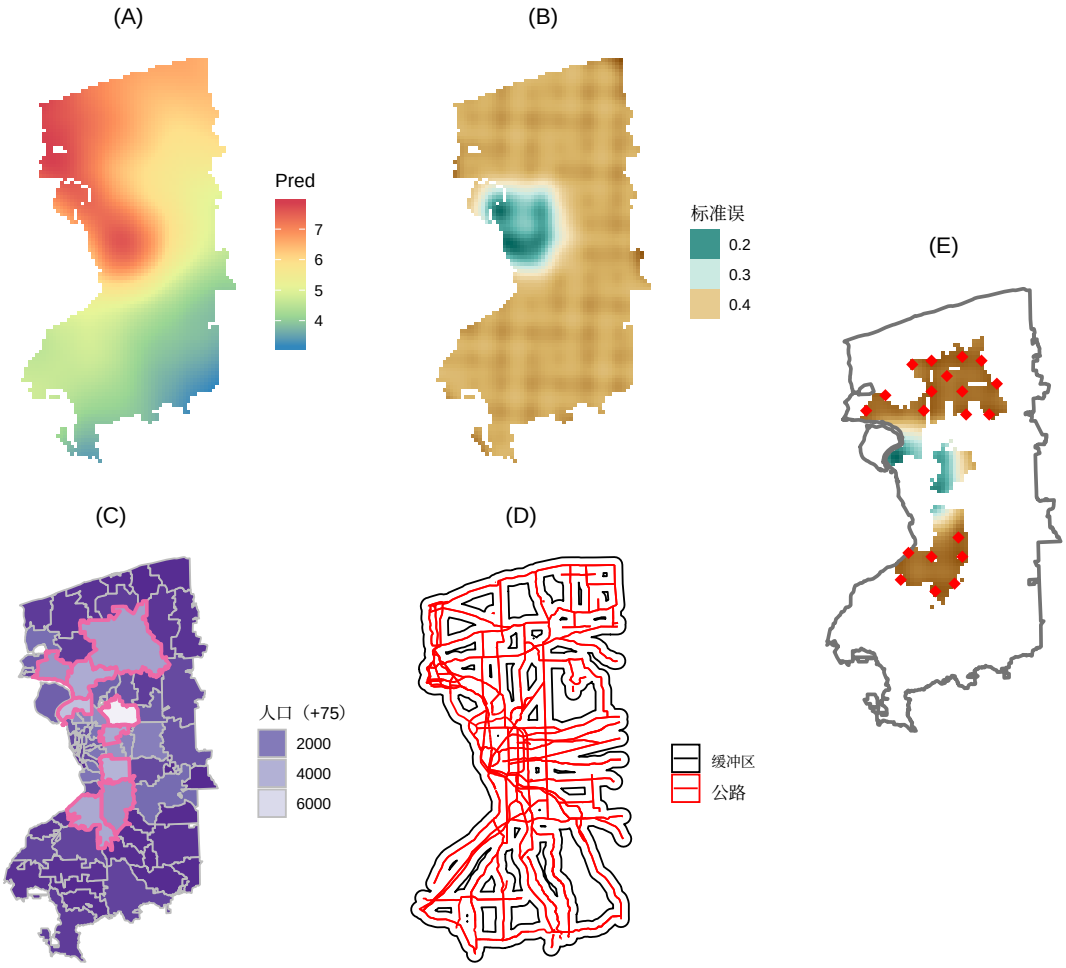


图 7：（A）使用模拟数据在监管监测站位置和模拟 CMAQ 输出时，FRK 得到的预测值。（B）相应的预测标准误差表面。（C）邮政编码单元中，70 岁及以上人群占比位于前 10% 的单元以粉色粗线标出。（D）主要公路（红色）和 2 km 缓冲区（黑色）。（E）预先确定的高风险位置处，（B）的预测标准误差表面。红点表示通过自适应采样选定的位置。

4 讨论

近年来，传感技术的进步使研究人员和社区能够收集更多数据用于环境监测，但这些新增数据的质量及其空间分辨率，尚无法与现有监管仪器测量相媲美。关于空间抽样设计策略的研究十分丰富，但这些策略未必会考虑支持域转换问题，也未必会考虑多个数据源之间异质的数据质量。本文将自适应空间抽样设计与通用空间数据融合方法相结合，以确定在低成本便携式传感器部署数量固定时的最优抽样地点。

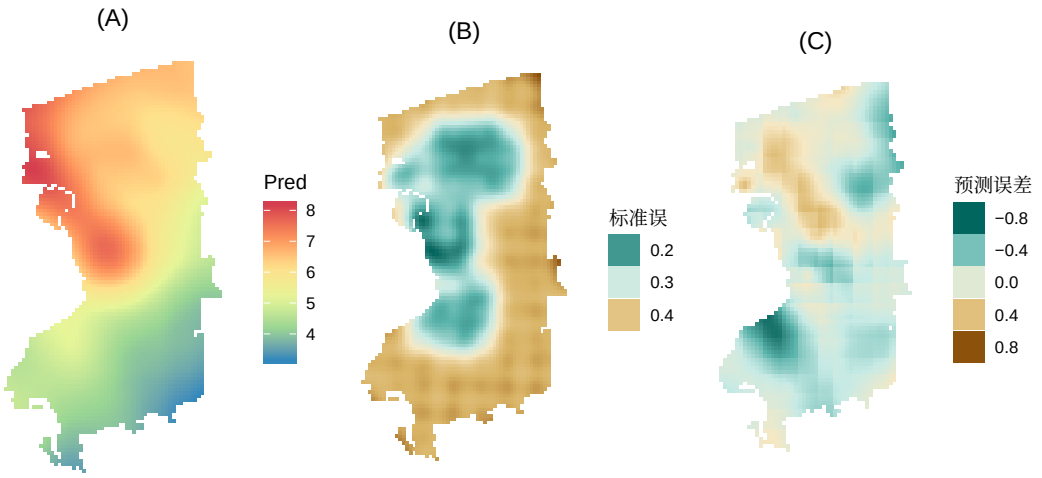


图 8：（A）使用全部三个数据源时的 FRK 预测。（B）相应的预测标准误。（C）预测误差（图 8(A) – 图6(A)）。

表 2：通过蒙特卡洛模拟获得的低成本便携式传感器数据（X）对空间预测准确度的影响。所使用指标的定义见第 2.3 节。

指标	不使用X的预测	使用X的100次蒙特卡洛模拟预测				
		Min	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	Max
MAE	0.19	0.12	0.14	0.14	0.15	0.17
MPE	0.43	0.40	0.42	0.43	0.51	0.65
均方根百分比误差	0.06	0.02	0.02	0.02	0.02	0.03
CRPS	0.15	0.13	0.14	0.15	0.15	0.19

本文的主要贡献有两方面。我们采用最先进的统计空间数据融合方法——固定秩克里金法（FRK），在量化设计准则的基础上进行最优选址，并明确考虑了一种新兴数据源——低成本便携式空气传感器测量值的质量。基于FRK提出的框架具有灵活性，既能适应空间数据之间的支撑域差异以及各数据源质量的不同，也能够处理潜在的大规模数据。在低成本便携式传感器的最优采样点选址算法中，我们在明确考虑不确定性的前提下采用了自适应采样方法，并纳入了决定决策者实际需求的多种风险因素。

通过两个案例研究，我们使用两个合成数据集考察了采样策略对预测准确度的影响。在第一个案例研究中，我们采用基于模拟数据的实验设计，在一个理想但相当简化的场景下评估传感器总数和传感器质量对预测质量的影响。在该场景中，底层过程是平稳且各向同性的，区域由一个正方形围成。我们使用四项指标量化预测质量。结果表明，总体而言，自适应采样比随机采样更有效，但其效果

在部署大量传感器时更为显著。此外，低成本便携式传感器测量值的质量会提升预测准确度，但只有在样本量足够大时，这种提升才较为明显。具体而言，如果样本量过小，自适应采样带来的相对性能增益可能并不明显。同样，我们不应期待预测性能会随着数据质量（低成本传感器的质量）线性提升。这些结果表明，在资源匮乏的条件下，当必须在数量与质量之间作出取舍时，所提出的设计框架有望为决策提供帮助。

需要指出的是，我们的实验设计并未探讨需要增加多少数据点才足够，也未探讨低成本便携式传感器的测量误差达到何种程度才可接受，以满足预先设定的标准。相反，我们关注的是在传感器数量和质量已知的情况下，设计策略所产生的影响。在第二个案例研究中，我们展示了如何将风险因素等其他条件纳入自适应采样，这对实际应用具有现实意义。然而，结果证实了我们在第一个案例研究中的发现：低成本传感器测量值并不能普遍保证提升预测性能。只有使用大量便携式传感器，才能确保这一点。

我们的实验使用了合成数据，但研究动机来自现实世界的政策需求。研究结果表明，增加低成本传感器测量值有望显著增进我们对个体空气污染暴露的认识，因为这些测量值使研究人员能够量化并描述城市区域内的暴露梯度，从而用于流行病学研究。此外，高空间分辨率的空气质量信息可能有助于监管机构制定政策和规划。例如，美国环境保护署支持的 Village Green 项目（https://www.epa.gov/air-research/village-green-project）旨在利用低成本空气传感器，向公众和社区提供实时空气质量信息。重要的是，我们的结果引发了这样的疑问：何时以及是否应使用便携式传感器，因为我们发现，除非部署足够数量的传感器，或传感器质量达到一定标准，否则其使用效率可能不如预期。例如，如果底层浓度场在空间上较为平滑，那么增加一两个质量相对较低的便携式传感器就是一种浪费。低成本传感器往往对环境条件敏感且需要现场校准，这一事实使上述问题更加突出（Zimmerman 等，2018）。

所提出的抽样设计旨在提升空气污染物长期预测的质量。对于这类污染物，其小尺度（城市内部）空间变异十分显著，现有监测网络无法捕捉这种变异。然而，该方法所采用的监测在时间上可能受到限制。为尽量减少潜在偏差，可在研究期间多次实施所提出的抽样设计，以考虑空气污染物浓度的季节变化，或避免异常天气及节假日、野火等事件的影响；类似研究也提出了这一做法（Hoek 等，2002；Henderson 等，2007；Madsen 等，2007）。另一项重要局限是将 CMAQ 用作代理变量。CMAQ 有望填补地面观测在空间和时间上的空缺，并使研究人员能够估计特定排放源的空气污染物浓度，但其计算成本高昂，因此无法用于实时估算。另一种选择是将气溶胶光学厚度或扩散模型输出作为区域空气质量数据的代理变量（Pu 和 Yoo，2019）。

未来研究有几个方向。首先，在我们的两个案例研究中，我们假设底层过程的空间结构已知，具体而言，假设其具有常数均值和指数协方差函数。然而，环境过程往往会随时间变化，其空间模式也可能存在异质性。针对室外空气污染物长期暴露评估所提出的抽样设计，可以扩展为更新均值和协方差，以适应短期暴露研究中空气污染的时间变化。其次，我们假设所有数据源的测量误差标准差均已知，但这在实际应用中并不现实。最后，我们发现，自适应空间采样中使用的批次大小 b 会影响最优抽样结果。因此，有必要进一步研究优化效用函数时批次大小 b 的影响。

5 结论

新兴技术为科学界带来了机遇与挑战，大气科学和环境暴露科学领域也不例外。低成本空气监测传感器价格低廉、便于携带，使研究人员和社区能够在感兴趣的时间和位置采集更多空气质量测量值，以解决数据稀疏问题；然而，这些测量值的准确度仍令人担忧。此外，在较小地理区域内部署大量传感器的需求也因地而异：人口密度较低、交通拥堵程度较轻区域的空气污染水平，与人口密度较高、交通拥堵严重的大型城市区域相比，可能存在显著差异。为充分发挥低成本便携式传感器的效用，本研究建议采用高效的空间抽样策略进行传感器部署，并将低成本便携式传感器的测量值与现有数据源的数据相结合，例如监管监测站的点测量，或大气模型生成的网格化代理数据。具体而言，我们在灵活的空间数据融合框架——固定秩克里金法——中提出一种自适应空间数据抽样设计方法，以高效收集低成本便携式传感器数据。基于一项多因素模拟实验，我们得出结论：只有在数量和质量均达到一定水平时，低成本便携式传感器才可能发挥效益。本研究以空间场景为例，但通过考虑观测值背后动态演变的过程，该方法还可扩展至时空场景。除了空气污染之外，我们预计，所提出的设计框架也可直接应用于其他环境科学问题，例如整合具有不同支撑域和测量误差特征的异质空间数据集。

致谢

EHY感谢计算研究中心以及布法罗大学能源、环境与水研究和教育项目（RENEW）种子项目资金的支持。AZM获得了澳大利亚研究理事会（ARC）探索早期职业研究奖（DE180100203）的资助。

参考文献

Apte, J. S., Messier, K. P., Gani, S., Brauer, M., Kirchstetter, T. W., Lunden, M. M., Marshall, J. D., Portier, C. J., Vermeulen, R. C. & Hamburg, S. P. (2017), '利用谷歌街景汽车进行高分辨率空气污染制图：挖掘大数据', 《环境科学与技术》**51**(12), 6999–7008. Baldauf, R. W., Lane, D. D. & Marote, G. A. (2001), '用于评估空气传播污染物暴露所致人类健康影响的环境空气质量监测网络设计', 《环境监测与评估》**66**(1), 63–76. Baldauf, R., Watkins, N., Heist, D., Bailey, C., Rowley, P. & Shores, R. (2009), '近道路空气质量监测：影响网络设计和数据解读的因素', 《空气质量、大气与健康》**2**(1), 1–9. Bernardo, J. M. (1979), '作为期望效用的期望信息', 《统计年鉴》**7**(3), 686–690. Berrocal, V. J., Gelfand, A. E. & Holland, D. M. (2010), '数值模型输出的时空降尺度模型', 《农业、生物与环境统计学杂志》**15**(2), 176–197. Bras, R. L. & Rodríguez-Iturbe, I. (1976), '用于估计降雨事件区域平均值的网络设计', 《水资源研究》**12**(6), 1185–1195. Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D. & Bartonova, A. (2017), '商用低成本传感器平台能否为空气质量监测和暴露估计作出贡献？', 《国际环境》**99**, 293–302. Chang, H. (2016), 环境污染场的数据同化，载于 A. Lawson, S. Banerjee, R. P. Haining & M. D. Ugarte, 编，《空间流行病学手册》，Chapman & Hall/CRC, pp. 289–302. Chang, H., Fu, A. Q., Le, N. D. & Zidek, J. V. (2007), '用于测量极端值的环境监测网络设计', 《环境与生态统计》**14**(3), 301–321. Chang, S.-C. & Lee, C.-T. (2007), '1994至2003年台湾台北空气质量趋势评估', 《环境监测与评估》**127**(1-3), 87–96. Chipeta, M. G., Terlouw, D. J., Phiri, K. S. & Diggle, P. J. (2016), '患病率调查的自适应地统计设计与分析', 《空间统计学》**15**, 70–84. Cressie, N., Gotway, C. A. & Grondona, M. O. (1990), '基于网络的空间预测', 《化学计量学与智能实验室系统》**7**(3), 251 – 271. Cressie, N. & Johannesson, G. (2008), '超大空间数据集的固定秩克里金法', 《英国皇家统计学会会刊：*B*辑（统计方法论）》**70**(1), 209–226. Dunbabin, M. & Marques, L. (2012), '用于环境监测的机器人：重大进展与应用', 《*IEEE*机器人与自动化杂志》**19**(1), 24–39.

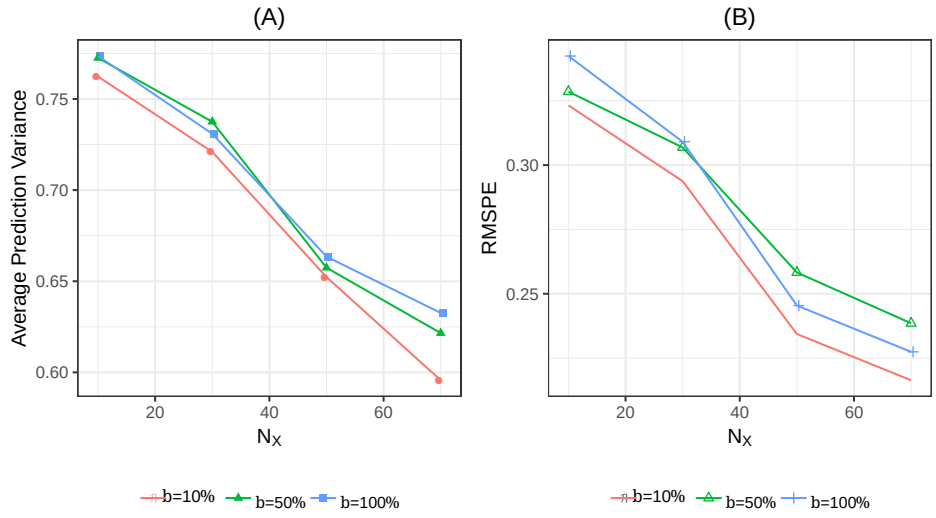
Fanshawe, T. R. & Diggle, P. J. (2013), 时空预测的自适应抽样设计，载于 J. Mateu & W. G. Müller, 编，《时空设计：高效数据获取的进展》，Wiley，第249–268页。Fuentes, M., Chaudhuri, A. & Holland, D. M. (2007), ‘环境数据空间抽样设计的贝叶斯熵’，《环境与生态统计》14(3), 323–340。Fuentes, M. & Raftery, A. E. (2005), ‘通过贝叶斯组合观测值与数值模型输出进行模型评估和空间插值’，《生物统计学》61(1), 36–45。Gelfand, A. E., Zhu, L. & Carlin, B. P. (2001), ‘时空数据的支撑域变更问题’，《生物统计学》2(1), 31–45。Gneiting, T. & Raftery, A. E. (2007), ‘严格适当评分规则、预测与估计’，《美国统计学会会刊》102(477), 359–378。Gotway, C. & Young, L. (2002), ‘组合不相容空间数据’，《美国统计学会会刊》97(458), 632–648。Henderson, S. B., Beckerman, B., Jerrett, M. & Brauer, M. (2007), ‘应用土地利用回归估算交通相关氮氧化物和细颗粒物的长期浓度’，《环境科学与技术》41(7), 2422–2428。Hoek, G., Meliefste, K., Cyrys, J., Lewné, M., Bellander, T., Brauer, M., Fischer, P., Gehring, U., Heinrich, J., van Vliet, P. 等 (2002), ‘欧洲三个地区细颗粒物浓度的空间变异性’，《大气环境》36(25), 4077–4088。Kabaghe, A. N., Chipeta, M. G., McCann, R. S., Phiri, K. S., Van Vugt, M., Takken, W., Diggle, P. & Terlouw, A. D. (2017), ‘自适应地统计抽样可有效识别马拉维农村地区重复横断面调查中的疟疾热点区域’，《公共科学图书馆·综合》12(2), e0172266。Kelly, K., Whitaker, J., Petty, A., Widmer, C., Dybwad, A., Sleeth, D., Martin, R. & Butterfield, A. (2017), ‘低成本颗粒物传感器的环境与实验室评估’，《环境污染》221, 491–500。Kim, S., Paulos, E. & Gross, M. D. (2010), Wearair：用于空气质量感测的表达性T恤，载于《第四届有形、嵌入式与具身交互国际会议论文集》，ACM，第295–296页。Krause, A., Singh, A. & Guestrin, C. (2008), ‘高斯过程中的近最优传感器布设：理论、高效算法与实证研究’，《机器学习研究杂志》9(Feb), 235–284。Le, N. D. & Zidek, J. V. (2006), 环境时空过程的统计分析，Springer Science & Business Media。Lenth, R. V. 等 (2009), ‘在R中使用rsm进行响应面方法’，《统计软件杂志》32(7), 1–17。

Lewis, A. & Edwards, P. (2016), ‘验证个人空气污染传感器’, *Nature News*535(7610), 29。Madsen, C., Carlsen, K. C. L., Hoek, G., Oftedal, B., Nafstad, P., Meliefste, K., Jacobsen, R., Nystad, W., Carlsen, K.-H. & Brunekreef, B. (2007), ‘对挪威奥斯陆室外交通污染的城市内部变异进行建模——一项GA2LEN项目’, 《大气环境》41(35), 7500–7511。Mead, M. I., Popoola, O., Stewart, G., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J., McLeod, M., Hodgson, T., Dicks, J. 等 (2013), ‘利用电化学传感器在低成本高密度网络中监测城市空气质量’, 《大气环境》70, 186–203。Müller, W. G. (2007), 空间数据采集：随机场的最优试验设计，施普林格。Nychka, D. & Saltzman, N. (1998), 空气质量监测网络的设计，载于 D. Nychka, W. W. Piegorsch & L. H. Cox, 编，《环境统计案例研究》，Springer，第51–76页。Paciorek, C. J. (2012), ‘在考虑代理指标系统误差的同时组合空间信息源’, 《皇家统计学会会刊：*C*辑（应用统计）》61(3), 429–451。Pu, Q. & Yoo, E.-H. (2019), ‘针对缺失数据问题的PM2.5浓度时空建模：中国北京案例研究’, 《国际地理信息科学杂志》第1–25页。R Core Team (2018), *R*：统计计算语言与环境，R 统计计算基金会，奥地利维也纳。**网址：** *https://www.R-project.org/*Rasmussen, C. E. & Williams, C. (2006), 麻省理工学院出版社，马萨诸塞州剑桥。Romary, T., De Fouquet, C. & Malherbe, L. (2011), ‘空气质量测量调查的抽样设计：一种优化方法’, 《大气环境》45(21), 3613–3620。Rundel, P. W., Graham, E. A., Allen, M. F., Fisher, J. C. & Harmon, T. C. (2009), ‘生态学研究中的环境传感器网络’, *New Phytologist* 182(3), 589–607。美国环境保护署 (2018), ‘空气传感器工具箱’。**网址：** *https://www.epa.gov/air-sensor-toolbox/how-use-air-sensors-air-sensor-guidebook*（访问日期：*2019*年*1*月*9*日）Villa, T. F., Gonzalez, F., Miljievic, B., Ristovski, Z. D. & Morawska, L. (2016), ‘用于空气质量测量的小型无人驾驶航空器概述：当前应用与未来展望’, 传感器16(7), 1072。Wikle, C. & Berliner, L. M. (2005), ‘跨空间尺度组合信息’, 《计量经济学》47(1), 80–91。

Wikle, C. K. & Royle, J. A.（1999），‘环境监测网络的时空动态设计’，《农业、生物与环境统计学杂志》4(4)，489–507。Zammit-Mangion, A. & Cressie, N.（2019），‘FRK：用于大型数据集空间及时空预测的 R 软件包’，《统计软件杂志》（待刊）。Zidek, J. V.、Sun, W. & Le, N. D.（2000），‘为监测多元高斯污染场设计并整合复合网络’，《皇家统计学会会刊：*C*辑（应用统计）》49(1)，63–79。Zimmerman, N.、Presto, A. A.、Kumar, S. P.、Gu, J.、Hauryliuk, A.、Robinson, E. S.、Robin-son, A.L. & Subramanian, R.（2018），‘使用随机森林的机器学习校准模型，以提升低成本空气质量监测中的传感器性能’，《大气测量技术》11，291–313。

附录A 批次大小对预测方差和准确度的影响

基于从Z 和Q 获得的 FRK预测标准误差面（见图2），我们为低成本传感器选择了四组采样位置，分别为 $n_X = 10$、30、50、70。在每种情况下，我们选择了三个批次大小级别，分别对应总样本量 n_X的 10%、50% 和 100%。为评估批次大小对自适应空间采样的影响，我们针对不同样本量 n_X和批次大小 b的每种情况获得了 FRK预测。通过计算 D 上的平均预测方差和平均均方根预测误差（均方根百分比误差），对模型性能进行了汇总，结果见图A1。结果清楚表明，随着 n_X 增加，平均预测误差和均方根百分比误差均会降低；同时，在四种不同样本量 n_X下，较小的批次大小 b 都能得到更低的平均预测误差。



图A1：批次大小分别为 b =3、15和30的批次自适应采样，初始样本量为 $n_Z = 50$，如图1所示。对于每个 n_X，红色圆圈、绿色三角形和蓝色方块分别表示批次大小为 n_X 的10%、50%和100%，对应（A）预测误差方差的均值和（B）均方根预测误差（RMSPE）。