适用于分类函数型数据的空间扫描统计量

Camille Frévent*, 1，Moustapha Sarr[1], 和 Sophie Dabo-Niang[2]

[1]里尔大学、里尔大学医院中心、ULR 2694 - METRICS：卫生技术与医疗实践评估，F-59000 里尔，法国[2]Paul Painlevé实验室、CNRS 8524联合研究单位、Inria-Datavers、里尔大学

arXiv:2603.01716v1 [stat.ME] 2 Mar 2026

摘要

我们开发并检验了一种适用于分类函数型数据的空间扫描统计量（CFSS）——这类数据结构使现有方法无法识别空间聚类。我们的方法将类别型函数观测的编码方案与非参数扫描统计量相结合。在包含三个不同情景的模拟研究中，CFSS 准确识别出模拟生成的空间聚类，且假阳性率极低、真阳性率较高，阳性预测值也较高。我们还使用 CFSS 识别并描述了法国2024年冬季空气污染数据中的空间聚类。

关键词 词： 类别数据、聚类检测、函数型数据、空间扫描统计量 s

1 引言

空间聚类检测长期以来在空间统计学中占据核心地位；其主要目标是开发统计方法和计算程序，以识别地理上局部集中的区域，在这些区域中，观测结果或特征与周边地区或整个研究区域的情况存在显著差异。聚类检测方法在许多领域（包括流行病学、环境科学、公共卫生和社会经济研究）中发挥着关键作用；在这些领域，识别和描述空间异质性对于阐明潜在机制以及制定循证决策至关重要。

空间扫描统计量是一种广泛使用且方法稳健的统计框架，用于检测空间聚类。通过使用位置和尺度各异的扫描窗口，对研究区域进行扫描，以寻找具有统计显著性的聚集。无需预先指定聚集的位置、空间范围或几何形状。这种以数据为驱动的方法使空间扫描统计特别适用于探索性空间分析和空间模式的发现。

Naus (1965a) 最初在一维情境下，随后在二维情境下（Naus, 1965b）提出了扫描统计量。随后，Kulldorff 和 Nagarwalla (1995) 将该技术应用于空间数据（针对伯努利模型），而 Kulldorff (1997) 则将其应用于泊松模型。这些奠基性研究结合使用似然比检验统计量和蒙特卡洛假设检验，来评估所识别聚集的统计显著性；这种方法能够检测不同大小和形状的聚集，同时校正多重

*通讯作者：camille.frevent@univ-lille.fr

检验。基于这项早期工作，空间扫描统计已推广至广泛的数据类型和用于连续观测值的概率模型，例如高斯模型（Kulldorff等人，2009）、指数模型（Huang等人，2007）和威布尔模型（Bhatt 和 Tiwari，2014）。许多研究者超越了单变量情形，并针对更复杂的数据结构开发了扩展方法。值得注意的是，Cucala等人 (2017) 基于多元正态模型开发了一种空间扫描统计量，明确纳入变量间的相关性；而 Cucala等人(2019) 则提出了一种基于 Wilcoxon–Mann–Whitney型检验统计量的多元非参数方法。

与此同时，传感技术、电子健康记录和数据存储基础设施的进步，使得人们能够获取日益丰富的数据集，其中的信息以连续的方式在时间和空间中被记录下来。这一转变推动了函数型数据分析（Ramsay andSilverman，1997）的发展。这是一种统计范式，将观测值建模为随机函数的实现，而非有限维向量。因此，许多经典统计技术经过重新审视并适用于函数型数据背景，其中包括针对函数型数据的空间扫描统计量（Frévent等人，2021，2023）。

然而，在许多实际场景中，观测数据并非实值数据。具体而言，健康科学领域经常涉及分类变量，例如疾病阶段、治疗方案或临床终点等记录数据。针对这些情形，人们已开发出多种空间聚类检测方法。尤为值得一提的是，Jung等人 （2007）和Jung等人（2010）分别基于序数和多项式建模框架提出了空间扫描统计量。这些方法能够有效检测纯分类空间数据集中的空间异质性。

相比之下，患者轨迹观测和纵向健康监测通常兼具类别性和函数性，因为对患者健康状态的重复记录会产生按时间索引的类别值函数。对此类数据进行统计分析，既涉及函数型数据分析的挑战，也涉及类别数据建模的挑战；必须同时处理时间依赖性、观测值固有的函数性质，以及离散且有限的结果空间。据我们所知，迄今开发的空间聚类检测方法中，没有一种能严谨且同时地考虑所有这些特征。当前的函数型空间扫描统计量（Frévent等人，2021，2023）是为实值函数型数据设计的，而类别空间扫描方法（Jung等人，2007，2010）则仅适用于标量、非函数型观测值。

为弥补这一不足，我们开发了一种新型空间扫描统计量，专门针对分类函数型数据设计，能够识别这样的空间聚类：其中类别状态随时间的变化与周边区域相比存在显著差异。我们的方法同时应对三项主要的方法学挑战。首先，它明确考虑了观测值的函数性质，从而捕捉个体轨迹中的时间依赖性。其次，它在统计上连贯的概率框架内对类别结果空间进行建模。第三，它纳入空间结构，用于检测地理位置明确的聚集，且无需预先知道其空间位置或范围。通过整合函数型数据表示、类别轨迹的概率建模以及严谨的基于蒙特卡罗的推断程序，这一新框架将经典空间扫描统计推广至更广泛且更具现实意义的一类问题。

该方法具有重要的实际意义，尤其是在卫生部门：捕捉患者轨迹或护理路径中的空间异质性，对于设计有针对性的干预措施和提升医疗质量至关重要。识别具有异常分类轨迹模式的区域，有助于公共卫生主管部门和临床医生发现新出现的问题、更有效地分配资源，并根据当地特点调整干预措施。本文其余部分安排如下。方法论框架

新型分类函数型数据空间扫描统计量（CFSS）的基础方法见第2节。扫描统计量在模拟研究中的表现及其在实证数据集上的应用见第3节。最后，第4节讨论总结性意见、可能的扩展方向以及未来研究方向。

2 方法

2.1 一般原理

设 $\{X(t)\ X(t) \in E = \{e_1, \ldots, e_J\},\ t \in [0,T]\}$ 为一个分类连续时间随机过程，其中 E 是状态集合，$[0,T]$ 是 R 中的一个区间。
设 s_1、……、s_K 为观测域 $S \subset \mathrm{R}^2$ 中的互不重叠位置，$X_{1,\ 1}$，……，$X_{1,\ n_1}$，……，$X_{K,\ 1}$，……，$X_{K,\ n_K}$ 为 X 在 s_1，……，s_K 中的观测值，其中 $X_{k,\ i}$ 表示空间单元 s_k 中的第 i 个个体。下文中，所有观测值均视为相互独立——这是扫描统计量中的经典假设。

空间扫描统计的目标是识别空间聚类并评估其统计显著性。这涉及一个原假设 $\mathcal{H}_0$（不存在聚类）和一个复合备择假设 $\mathcal{H}_1$（至少存在一个聚类 $w \subset S$，其 X 值异常）。

空间聚类定义为一个地理区域，其中个体的特征与其余总体中的个体不同。对于分类函数型数据，可以考虑多种类型的簇：(i) 簇内个体的状态序列不同于簇外个体的状态序列；(ii) 簇内个体改变状态的速度比簇外个体更快（或更慢）；(iii) 簇内个体在某些状态上花费的时间比簇外个体更多（或更少）。

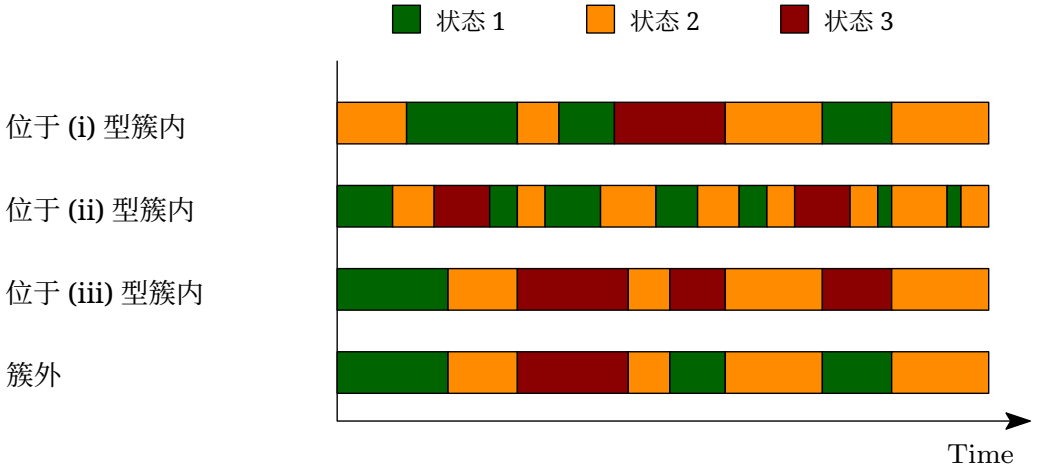

图 1：分类函数型数据中三种类型簇的示意图。在 (i) 型簇内，个体的状态序列不同于簇外个体的状态序列；在此示例中，第一个状态不同，且个体从状态 1 直接转移到状态 3，而不经过状态 2。在 (ii) 型簇中，个体改变状态的速度比簇外个体更快（或更慢）。在 (iii) 型簇中，个体在某些状态上花费的时间比簇外个体更多（或更少）。

根据Cressie（1977），扫描统计量定义为在一组潜在聚类上集中度指数的最大值 $\mathcal{W}$。在下文中，不失一般性，我们关注

可变大小的圆形聚类（如Kulldorff （1997）所提出）。潜在聚类集合$\mathcal{W}$ 是以某一位置为中心并经过另一位置的圆盘集合：

$$\mathcal{W} = \left\{ w_{k,k'} / 1 \le |w_{k,k'}| \le \frac{n}{2}, 1 \le k, k' \le K \right\},$$

其中 $w_{k,\ k'}$ 是以 s_k 为圆心并经过 $s_{k'}$的圆盘， $|w_{k,\ k'}|$是 $w_{k\text{中的个体数量},\ k'}$，而 $n = \sum_{k=1}^{K} n_k$ 是观测到的个体总数。因此，给定聚集所包含的研究对象不会超过研究个体总数的 50%，这符合Kulldorff 和 Nagarwalla(1995)的建议。文献中还提出了椭圆形聚类（Kulldorff等人，2006）和不规则形状聚类（Tango 和 Takahashi，2005；Yin 和 Mu，2018。

2.2 降维

分类函数型数据的主要问题在于缺乏处理这类数据的统计方法。将分类函数型数据转换为实值后，便于进行空间聚类检测。为此，我们采用函数型数据多重对应分析（由Preda等人（2021）提出），将$X_{k,\ i}$转换为实向量 $\boldsymbol{Z}_{k,\ i} =$ （$Z_{k\ i}^{(1)}$，，……， $Z_{k\ i}^{(M)}$，)$^\top$，同时尽可能多地保留信息。

令 $\mathbb{1}_{e_j}(t) = \begin{cases} 1 \text{ 当 } X(t)\ 0 \quad X\ t \quad = e_j \text{ 当 } 0 \neq e_j \text{。} \end{cases}$

我们假设 $(\mathcal{A}_1)$ X 依概率连续：$\lim_{h\to 0} \mathrm{P}(X(t+h) \neq X(t)) = 0$，并且 $(\mathcal{A}_2)$ $\forall t \in [0, T]$，$\forall e_j \in E$，$\mathrm{P}(X(t) = e_j) \neq 0$。

函数型多重对应分析寻找新的正交变量 $Z^{(1)}$，…， $Z^{(M)}$，使其最大化

$$\int_0^T \frac{\mathbb{V}(E_t(Z))}{\mathbb{V}(Z)}\, \mathrm{d}t \tag{1}$$

$_J$其中 $E_t(Z) = \sum \mathrm{E}(Z|X(t) = e_j)\mathbb{1}_{e_j}(t)$ 是 L^2 上的条件期望算子；$_{j=1}$表示与 $X(t)$ 相关的、具有有限二阶矩的实随机变量空间，$\mathrm{V}(E_t(Z))$ 是 $E_t(Z)$ 的方差。

命题 1。
E_t 是一个自伴、幂等算子。

最大化（1）等价于求解

$$\int_0^T E_t(Z)\mathrm{d}t = \lambda Z. \tag{2}$$

则 $Z^{(1)}$、……、 $Z^{(M)}$ 是 $\int_0^T E_t(\cdot)\, \mathrm{d}t$ 的正交特征向量，分别对应按降序排列的特征值$\lambda^{(1)}$、……、 $\lambda^{(M)}$，不包括与平凡特征值 $\lambda^{(0)} = T$对应的平凡解 $Z^{(0)} = 1$。

为了求出非平凡特征向量和特征值，Preda等（2021）将 $\zeta(t) = \frac{1}{\lambda} E_t(Z) \in H([0, T])$定义为具有有限总方差 $\int_0^T \mathrm{V}(\zeta(t))\, \mathrm{d}t$ 并配备标量积 $\langle \zeta, \zeta' \rangle_H = \int_0^T \mathrm{E}[\zeta(t)\zeta'(t)]\, \mathrm{d}t$的实值过程的希尔伯特空间。4

在考虑关于 $X(t')$的条件期望时，（2）可以改写为

$$\int_0^T K(t',t)\zeta(t)\ \mathrm{d}t = \lambda\zeta(t') \tag{3}$$

其中 $K(t',t) = E_{t'}E_t$。
$\zeta(t)^{(0)} = \frac{1}{T}, t \in [0,T]$ 是 K 的平凡特征函数，对应于平凡特征值$\lambda^{(0)} = T$，而 $\zeta(t)^{(1)},\ldots,\zeta(t)^{(M)}$ 是与特征值 $\lambda^{(1)},\ldots,\lambda^{(M)}$相关联的非平凡正交（对于 $\langle\cdot,\cdot\rangle_H$）特征函数，按降序排列。

为确保 $\zeta(t)^{(1)},\ldots,\zeta(t)^{(M)}$的唯一性，Saporta (1981)与Preda等人(2021)增加了以下约束：

$$\int_0^T \mathbb{E}\left[\left(\zeta(t)^{(m)}\right)^2\right]\ \mathrm{d}t = 1. \tag{4}$$

特征函数的性质在以下命题中给出。

命题 2 (Saporta(1981) 第107页)。
$\int_0^T \mathbb{E}\left[\zeta(t)^{(1)}\right]\ dt = \cdots = \int_0^T \mathbb{E}\left[\zeta(t)^{(M)}\right]\ dt = 0.$

命题 3 (Preda等人(2021))。
$\zeta(t)^{(m)}$ 可写为

$$\zeta(t)^{(m)} = \sum_{j=1}^{J} a_j(t)^{(m)}\mathbb{1}_{e_j}(t)$$

其中 $\{a_j(\cdot)^{(m)}、j = 1.....、J\}$ 是在 $[0,T]$上的确定性编码函数。因此，寻找满足（3）的 $\zeta(\cdot)^{(m)}$ 等价于求解更经典的问题

$$\int_0^T \sum_{j=1}^{J} \mathbb{P}(X(t') = e_{j'}, X(t) = e_j)a_j(t)^{(m)}\ dt = \lambda a_{j'}(t')^{(m)}\mathbb{P}(X(t') = e_{j'}), \tag{5}$$

且约束（4）变为

$$\int_0^T \sum_{j=1}^{J} \left[a_j(t)^{(m)}\right]^2 \mathbb{P}(X(t) = e_j)\ dt = 1.$$

命题 4（Saporta（1981））。
由 $\zeta(t)^{(0)}、.....、\zeta(t)^{(M)}$可推导出（2）的解 $Z^{(0)}、.....、Z^{(M)}$。 y

$$Z^{(m)} = \int_0^T \zeta(t)^{(m)}\,dt.$$

命题 5（Saporta（1981））。
(2) $Z^{(1)},\ldots,Z^{(M)}$ 的非平凡解是中心化且彼此不相关的随机变量，满足 $\mathbb{V}(Z^{(m)}) = \lambda^{(m)}, 1 \le m \le M.$

我们在补充材料中提供了Saporta（1981）和Preda等人(2021)的证明详细版本。

多重对应分析的实践应用

从观测数据X_1,1, . . . , X_K,n_K中实际计算 $\zeta_{1\,1}^{(1)}$, , . . . , $\zeta_{1\,1}^{(M)}$, , . . . , $\zeta_{K}^{(1)}$,n_K, . . . , $\zeta_{K}^{(M)}$,n_K， 进而计算 $Z_{1\,1}^{(1)}$, , . . . , $Z_{1\,1}^{(M)}$, , . . . ,$Z_{K}^{(1)}$,n_K, . . . , $Z_{K}^{(M)}$,n_K， 通常通过将编码函数 $\{a_{j,k,i}(\cdot)^{(m)}$, $1 \leq m \leq M, 1 \leq j \leq J, 1 \leq k \leq K, 1 \leq i \leq n_k\}$近似到任意有限函数基来实现。理论和计算细节见Saporta（1981）和Preda等（2021）， 该过程已在R软件包cfda中实现。

I接下来，我们选择 M ， 使得至少 90% 的方差得到解释 .

2.3 CFSS

目前， 数据 $X_{1,\,1}$， . . . ， $X_{1,\,n_1}$， . . . ， $X_{K,\,1}$， . . . ， $X_{K,\,n_K}$已被编码为 M 维实数（非函数型）数据 $\boldsymbol{Z}_{1,\,1}$， . . . ， $\boldsymbol{Z}_{1,\,n_1}$， . . . ， $\boldsymbol{Z}_{K,\,1}$， . . . ， $\boldsymbol{Z}_{K,\,n_K}$。
因此，我们可以将（由 Cucala等人 （2019）开发并在 R软件包**HDSpatialScan** （Frévent等人，2022; Frévent等人，2025）中实现的）非参数空间扫描统计量推广到多元数据的方法应用于当前情境。

我们考虑以下检验假设:

$\mathcal{H}_0$: $\boldsymbol{Z}_{1,\,1}$、.....、$\boldsymbol{Z}_{1,\,n_1}$、.....、$\boldsymbol{Z}_{K,\,1}$、.....、$\boldsymbol{Z}_{K,\,n_K}$具有相同的分布，无论它们对应的位置 s_1、.....、 s_K如何。

$\mathcal{H}_1$: 存在 $w \in \mathcal{W}$， 使得 $\boldsymbol{Z}$ 在 w 内的分布与其在 w 外的分布不同。

Let

$$R_{k,i} = \frac{1}{n} \sum_{k'=1}^{K} \sum_{i'=1}^{n_{k'}} \operatorname{sgn}\left[A_{\boldsymbol{Z}}(\boldsymbol{Z}_{k,i} - \boldsymbol{Z}_{k',i'})\right]$$

为多元秩，其中 sgn(·) 是多元符号函数

$$\begin{array}{rccc} \operatorname{sgn}: & \mathbb{R}^M & \to & \mathbb{R}^M \\ & \boldsymbol{z} & \mapsto & ||\boldsymbol{z}||_2^{-1} \boldsymbol{z}\, \mathbb{1}_{\boldsymbol{z} \neq \mathbf{0}_M} \end{array}$$

其中 $\mathbf{0}_M$ 是仅由 0 构成的 M 维向量， 而 $A_{\boldsymbol{Z}}$ 是一个基于数据的变换矩阵， 它使多元秩的表现仿佛是在单位 M 球面上呈球形分布:

$$\frac{M}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} R_{k,i} R_{k,i}^{\top} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} R_{k,i}^{\top} R_{k,i} I_M.$$

检验统计量

$$W^{(w)} = \frac{Mn}{\sum_{k=1}^{K} \sum_{i=1}^{n_k} R_{k,i}^{\top}} \left(|w| ||\bar{R}_w||_2^2 + |w^{\mathsf{c}}| ||\bar{R}_{w^{\mathsf{c}}}||_2^2 \right),$$

where $\bar{R}_g = \frac{1}{|g|} \sum_{\substack{k=1 \\ s_k \in g}}^{K} \sum_{i=1}^{n_k} R_{k,i}, g \in \{w, w^{\mathsf{c}}\}$ allows one to test (Oja and Randles, 2004):

$\mathcal{H}_0$: $\boldsymbol{Z}$ 的分布在 w 内外相同

vs.

$\mathcal{H}_1$: w 中 $\boldsymbol{Z}$ 的分布与 w之外 $\boldsymbol{Z}$ 的分布不同。

接下来，遵循Cucala等人（2019），我们将 $W^{(w)}$ 视为聚类检测的集中度指数，并将 CFSS 定义为

$$\Lambda_{\text{CFSS}} = \max_{w \in \mathcal{W}} W^{(w)}.$$

最可　　能的聚类（MLC）是获得该最大值的潜在聚类　　：

$$\text{MLC}_{\text{CFSS}} = \arg\max_{w \in \mathcal{W}} W^{(w)}.$$

2.4 计算 MLC 的统计显著性

确定 MLC 后，必须评估其统计显著性。由于统计量之间存在内在依赖关系，因此在 $\mathcal{H}_0$ 下扫描统计量的分布难以处理：当 $w \cap w' \neq \emptyset$ 时，统计量 $W^{(w)}$ 和 $W^{(w')}$。

为了估计观测到的空间扫描统计量对应的 p 值，我们采用一种基于随机置换的方法，通常称为“随机标记法”。设 P为对原始数据集执行的随机置换次数，$\Lambda^{(1)}{}_{\text{CFSS}} \cdots$， $\Lambda^{(P)}{}_{\text{CFSS}}$ 表示从这些模拟数据集中得到的空间扫描统计量。根据 Dwass （1957），原始数据中观测到的扫描统计量的估计 p 值计算如下：

$$\widehat{p} = \frac{1 + \sum_{p'=1}^{P} \mathbb{1}_{\Lambda^{(p')}_{\text{CFSS}} \geq \Lambda_{\text{CFSS}}}}{1 + P}.$$

最后，如果 MLC 对应的估计 p 值小于预先设定的 I 类错误率，则认为该 MLC 在统计上显著。

3 有限样本性质

3.1 模拟研究

我们使用法国本土94个省（县级行政区划）的地理坐标生成了人工数据集。每个省的位置由其质心确定。在每个数据集中定义并模拟了一个真实聚类（记为 w）。该聚类包含位于巴黎大都市区的8个省（在补充材料图6中以红色突出显示）。

3.1.1 人工数据集的生成

在94个空间单元中的每一个单元内，共模拟了 $n_k = 10$ 个体轨迹。每条轨迹对应于三状态过程 $\{X_{k,i}(t) \in E = \{e_1, e_2, e_3\}, t \in [0, 18]\}$ 并使用转移矩阵 $\boldsymbol{P}_k$ 和强度向量 $\boldsymbol{\lambda}_k = (\lambda_{1,k}, \lambda_{2,k}, \lambda_{3,k})^\top$生成。

对于每个个体，轨迹从 e_1开始。在该状态中停留的时间根据速率为λ_1的k 指数分布生成，随后根据 $\boldsymbol{P}_k$第一行中的概率选择下一个状态。然后对所有后续状态重复相同的过程（使用 $\boldsymbol{\lambda}_k$ 中的相应强度以及 $\boldsymbol{P}_k$中的概率），直到达到最大时间（$T = 18$）。

考虑了三种模拟情景：(i) 空间聚类内外的转移矩阵 $\boldsymbol{P}_k$ 没有差异，不存在吸收态，并且空间聚类与区域其余部分的强度不同，

(ii) 空间聚类内外的转移矩阵 $\boldsymbol{P}_k$ 相同，状态 e_3 是吸收态，并且空间聚类与区域其余部分的强度不同；(iii) 空间聚类内外的转移矩阵 $\boldsymbol{P}_k$ 不同，状态 e_3 是吸收态，并且空间聚类内外的强度相同。

对于情景 (i)，我们将所有 $\boldsymbol{P}_k$ 定义为

$$\boldsymbol{P}_k = \boldsymbol{P} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

以及 $\boldsymbol{\lambda}_k = (0.2 + \alpha \mathbb{1}_{s_k \in w},\ 0.2 + \alpha \mathbb{1}_{s_k \in w},\ 0.2 + \alpha \mathbb{1}_{s_k \in w})^\top$, $\alpha \in \{0, 0.15, 0.30, 0.45, 0.60, 0.75\}$。
对于情景 (ii),

$$\boldsymbol{P}_k = \boldsymbol{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

$\boldsymbol{\lambda}_k = (0.2 + \alpha \mathbb{1}_{s_k \in w},\ 0.2 + \alpha \mathbb{1}_{s_k \in w},\ 0)^\top$, $\alpha \in \{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$。对于情景 (iii)，$\boldsymbol{P}_k$ 定义为

$$P_k = \begin{pmatrix} 0 & 1 & 0 \\ 0.7 - \gamma \mathbb{1}_{s_k \in w} & 0 & 0.3 + \gamma \mathbb{1}_{s_k \in w} \\ 0.5 & 0.5 & 0 \end{pmatrix}, \gamma \in \{0, 0.10, 0.20, 0.30, 0.40, 0.50\}$$

以及 $\boldsymbol{\lambda}_k = \boldsymbol{\lambda} = (0.2,\ 0.2,\ 0)^\top$。
值得注意的是，$\alpha = 0$ 和 $\gamma = 0$ 也进行了测试，以评估名义第一类错误的维持情况。所有情景的数据示例见补充材料（图7）。

3.1.2 方法性能评估

对于每种情景及 α （或 γ）的每个取值，我们模拟了 1000 个人工数据集。通过对变换数据进行 999 次随机置换，估算与每个 MLC 相关的 p 值（详情请参阅第 2.2 节）。拒绝 $\mathcal{H}_0$ 时，第一类错误率设为 5%。我们从四项指标评估了 CFSS 的性能：检验功效、真阳性率、假阳性率和阳性预测值。检验功效定义为在第一类错误率设定下，导致拒绝 $\mathcal{H}_0$ 的模拟所占比例。对于导致拒绝 $\mathcal{H}_0$ 的模拟数据集，真阳性率为 w 中正确检测到的位点比例的平均值；假阳性率为 w^c 中包含在检测到的簇内的位点比例的平均值；阳性预测值则对应于检测到的簇内 w 中个体所占的比例。

我们比较了在近似编码函数 $a_{j,\ k,\ i}(\cdot)^{(m)}$ 时，采用 10、20 和 30 个基函数的不同方法性能（详情请参阅第 2.2 节）。

3.1.3 模拟研究结果

对于 $\alpha = 0$ 和 $\gamma = 0$，无论情景如何，也无论基函数的数量是多少，该方法均能保持正确的第一类错误率（0.05）（图2）。基函数的数量对性能的影响很小，这与Preda等人（2021）一致。

该方法表现令人满意，具体表现为假阳性率极低、真阳性率较高且阳性预测值较高。此外，该方法在所有考察的情景中均成功检测到了模拟簇。

3.2 在真实数据中的应用

我们考察了法国在1月至3月$_{10}$ 期间的每日平均PM 1st 浓度31^{st}（2024年），数据来自Geod'air数据库（https://www.geodair.fr/donnees/consultation）。279个监测站的位置及观测到的浓度曲线见图3 （左图）。随后，我们根据Atmo指数对浓度进行分类（见表1，https://www.atmo-nouvelleaquitaine.org/article/lindice-atmo-un-outil-precis-et-complet-de-la-qualite-de-lair）。为避免类别过于稀少，我们将Atmo指数中的"差""很差"和"极差"类别合并为一个"差"类别。

表1: 根据每日 PM_{10} 平均浓度范围划分的 Atmo指数。

每日 PM_{10} 平均浓度范围（$\mu g.m^{-3}$）	**Atmo指数**
$[0, 20]$	Good
$]20, 40]$	一般
$]40, 50]$	退化
$]50, 100]$	Bad
$]100, 150]$	很差
> 150	极差

编码数据时，我们使用了10个基函数应用CFSS。所考虑的潜在聚类集合 $\mathcal{W}$ 为第2节所述的集合。我们通过999次蒙特卡洛置换评估MLC的统计显著性；当相关p值低于0.05时，认为空间聚类具有统计显著性。

检测到的MLC（$\widehat{p} = 0.001$）位于法国东南部，包含71个监测站（见图4）。相关轨迹（图5）显示，与MLC以外的监测站相比，MLC内的监测站记录到更多向"一般""退化"和"差"状态的转变，并且在这些状态下花费的时间更多。

4 讨论

在此，我们介绍了 CFSS 的开发与测试。该方法能够识别轨迹与其他患者显著不同的患者群体，因此为公共卫生政策制定者提供了重要工具。此前发表的方法无法检测分类函数型数据中的空间聚类。

我们的模拟研究结果表明，CFSS 的性能非常令人满意。在所考虑的三种情景中，该方法均正确识别出了模拟的空间聚类，并且假阳性率非常低、真阳性率较高、阳性预测值较高。第一类错误维持在名义水平。模拟研究的结果还表明，编码数据时所使用的基函数数量对性能的影响非常小。这与 Preda等人(2021）的报告一致。

接下来，我们将 CFSS 应用于污染数据；该方法检测出了一个空间聚类，在该聚类内，监测站记录到的向"一般"、"退化"和"差"状态转变的次数多于其他地区，并且在这些状态中停留的时间也更长。

需要注意的是，我们仅考虑了圆形聚类。不过，CFSS 可以调整为检测非圆形聚类，例如椭圆形聚类（Kulldorff 等人，2006）和任意形状的聚类（Tango 和 Takahashi，2005）。然而，这些形状会产生许多

比圆形方法更多的潜在聚类，并相应地增加计算时间。

在此，我们仅考虑了MLC。然而，研究次级簇（即与MLC共存的具有统计显著性的空间簇）将很有意义。检测次级簇的一种初步方法由Kulldorff 和 Nagarwalla（1995）提出。该方法在于估计所有潜在聚类的p值，如同它们就是MLC，然后考虑所有不与更具统计显著性的聚类重叠的具有统计显著性的聚类。由于潜在聚类的集中度指数要与每个数据集在 $\mathcal{H}_0$下的集中度指数最大值进行比较，次级簇必须足够强，才能拒绝 $\mathcal{H}_0$ “本身”。

或者，Zhang等人（2010）的序贯方法是从数据集中移除与MLC对应的数据，然后迭代运行扫描统计量，以找出下一个最可能聚类。尽管已有人提出其他方法（Li等人，2011；Lin等人，2016；Xu和Gangnon，2016），但维持名义第一类错误始终是一个难点。

目前的 CFSS 可以基于单个分类变量检测聚集。不过，未来我们计划将 CFSS 扩展到涉及多个分类变量的情形。CFSS 也可能有助于检测由定性变量和定量变量组成的数据集中的聚集；这也将是未来工作的研究内容。

参考文献

Bhatt, V. and N. Tiwari (2014). 基于威布尔分布的生存数据空间扫描统计量. Statistics in Medicine 33(11), 1867–1876. Cressie, N. (1977). 关于圆和直线上的扫描统计量的一些性质. Journal of Applied Probability 14(2), 272–283. Cucala, L., M. Genin, C. Lanier, and F. Occelli (2017). 用于空间数据的多元高斯扫描统计量. Spatial Statistics 21, 66–74. Cucala, L., M. Genin, F. Occelli, and J. Soula (2019). 用于空间数据的多元非参数扫描统计量. Spatial Statistics 29, 1–14. Dwass, M. (1957). 非参数假设的修正随机化检验. The Annals of Mathematical Statistics, 181–187. Frévent, C., M.-S. Ahmed, S. Dabo-Niang, and M. Genin (2023). 多元函数型数据空间扫描统计量的研究. Journal of the Royal Statistical Society Series C: Applied Statistics 72(2), 450–475. Frévent, C., M.-S. Ahmed, M. Marbac, and M. Genin (2021). 函数型数据中的空间聚类检测：新的扫描统计量方法. Spatial Statistics 46, 100550. Frévent, C., M.-S. Ahmed, J. Soula, Z. Smida, L. Cucala, S. Dabo-Niang, and M. Genin (2022). R软件包HDSpatialScan：利用空间扫描统计量检测多元和函数型数据中的聚集. R Journal 14(3). Frévent, C., M.-S. Ahmed, J. Soula, Z. Smida, L. Cucala, S. Dabo-Niang, and M. Genin (2025). HDSpatialScan：多元与函数型空间扫描统计量. R软件包，版本1.0.5。

Huang, L., M. Kulldorff, and D. Gregorio (2007). 生存数据的空间扫描统计量. Biometrics 63(1), 109–118. Jung, I., M. Kulldorff, and A. C. Klassen (2007). 有序数据的空间扫描统计量. Statistics in Medicine 26(7), 1594–1607. Jung, I., M. Kulldorff, and O. J. Richard (2010). 多项数据的空间扫描统计量. Statistics in Medicine 29(18), 1910–1918. Kulldorff, M. (1997). 一种空间扫描统计量. Communications in Statistics-Theory and Methods 26(6), 1481–1496. Kulldorff, M., L. Huang, and K. Konty (2009). 基于正态概率模型的连续数据扫描统计量. International Journal of Health Geographics 8, 1–9. Kulldorff, M., L. Huang, L. Pickle, and L. Duczmal (2006). 一种椭圆空间扫描统计量. Statistics in Medicine 25(22), 3929–3943. Kulldorff, M. and N. Nagarwalla (1995). 空间疾病聚集：检测与推断. Statistics in Medicine 14(8), 799–810. Li, X.-Z., J.-F. Wang, W.-Z. Yang, Z.-J. Li, and S.-J. Lai (2011). 一种用于多个聚集的空间扫描统计量. Mathematical Biosciences 233(2), 135–142. Lin, P.-S., Y.-H. Kung, and M. Clayton (2016). 用于检测任意形状多个聚集的空间扫描统计量. Biometrics 72(4), 1226–1234. Naus, J. I. (1965a). 直线上最大点聚集规模的分布. Journal of the American Statistical Association 60(310), 532–538. Naus, J. L. (1965b). 二维随机点的聚集. Biometrika 52(1-2), 263–266. Oja, H. and R. H. Randles (2004). 多元非参数检验. Statistical Science 19(4), 598–605. Preda, C., Q. Grimonprez, and V. Vandewalle (2021). 分类函数型数据分析：cfda R 软件包. Mathematics 9(23), 3074. Ramsay, J. and B. Silverman (1997). 函数型数据分析. Springer 统计学丛书. Springer. Saporta, G. (1981). 时间数据探索性分析方法. 运筹学大学研究局论文集，研究系列 37, 7–194. Tango, T. and K. Takahashi (2005). 用于检测聚集的灵活形状空间扫描统计量. International Journal of Health Geographics 4, 1–15. Xu, J. and R. E. Gangnon (2016). 空间聚类检测的逐步与分阶段方法. Spatial and Spatio-Temporal Epidemiology 17, 59–74. Yin, P. and L. Mu (2018). 一种快速检测不规则形状空间疾病聚集的混合方法. GeoJournal 83, 693–705. Zhang, Z., R. Assunção, and M. Kulldorff (2010). 针对多个聚集进行调整的空间扫描统计量. Journal of Probability and Statistics 2010(1), 642379。

A 补充材料：第2节结果的证明

命题1的证明

回顾命题1。
E_t 是一个自伴且幂等的算子。

证明。
设 Z Z' 是 L^2中的两个随机变量。

$$\langle Z, E_t(Z')\rangle = \mathbb{E}\left[ZE_t(Z')\right] = \sum_{j=1}^{J} \mathbb{E}\left[Z\mathbb{E}(Z'|X(t)=e_j)\mathbb{1}_{e_j}(t)\right] = \sum_{j=1}^{J} \mathbb{E}(Z'|X(t)=e_j)\mathbb{E}(Z\mathbb{1}_{e_j}(t)).$$

Since $\mathbb{E}(Z\mathbb{1}_{e_j}(t)) = \mathbb{E}(Z|X(t)=e_j)\mathbb{P}(X(t)=e_j)$,

we deduce $\langle Z, E_t(Z')\rangle = \sum_{j=1}^{J} \mathbb{E}(Z'|X(t)=e_j)\mathbb{E}(Z|X(t)=e_j)\mathbb{P}(X(t)=e_j).$

Similarly, $\langle E_t(Z), Z'\rangle = \sum_{j=1}^{J} \mathbb{E}(Z'|X(t)=e_j)\mathbb{E}(Z|X(t)=e_j)\mathbb{P}(X(t)=e_j).$

This shows that E_t is self-adjoint.
此外，

$$\begin{aligned}
E_t(E_t(Z)) &= \sum_{j=1}^{J} \mathbb{E}\left[E_t(Z)|X(t)=e_j\right]\mathbb{1}_{e_j}(t) \\
&= \sum_{j=1}^{J} \mathbb{E}\left[\sum_{j'=1}^{J} \mathbb{E}(Z|X(t)=e_{j'})\mathbb{1}_{e_{j'}}(t)|X(t)=e_j\right]\mathbb{1}_{e_j}(t) \\
&= \sum_{j=1}^{J}\sum_{j'=1}^{J} \mathbb{E}(Z|X(t)=e_{j'})\mathbb{P}\left[X(t)=e_{j'}|X(t)=e_j\right]\mathbb{1}_{e_j}(t) \\
&= \sum_{j=1}^{J} \mathbb{E}(Z|X(t)=e_j)\mathbb{1}_{e_j}(t) = E_t(Z)
\end{aligned}$$

至此，证明完毕。 □

命题2的证明

回顾命题2。p

$$\int_0^T \mathbb{E}\left[\zeta(t)^{(1)}\right]\,\mathrm{d}t = \cdots = \int_0^T \mathbb{E}\left[\zeta(t)^{(M)}\right]\,\mathrm{d}t = 0.$$

证明。
设 $m \in \{1,\dots M\}$, $\zeta^{(m)}$ 与 $\zeta^{(0)}$ 正交。因此，$\int_0^T \mathbb{E}\left[\zeta(t)^{(m)}\right]\mathrm{d}t = T\int_0^T \mathbb{E}\left[\zeta(t)^{(m)}\zeta(t)^{(0)}\right]\mathrm{d}t = 0$。 □

命题的证明 3

回顾命题3。$\zeta(t)^{(m)}$ 可写为

$$\zeta(t)^{(m)} = \sum_{j=1}^{J} a_j(t)^{(m)}\mathbb{1}_{e_j}(t)$$

其中 $\{a_j(\cdot)^{(m)},\ j=1,\ \ldots\ldots,\ J\}$ 是定义在 $[0,T]$上的确定性编码函数。因此，求得$\zeta(\cdot)^{(m)}$ 使其满足（3）等价于求解一个更经典的问题

$$\int_0^T \sum_{j=1}^{J} \mathbb{P}(X(t')=e_{j'}, X(t)=e_j) a_j(t)^{(m)}\ \mathrm{d}t = \lambda a_{j'}(t')^{(m)} \mathbb{P}(X(t')=e_{j'}). \tag{6}$$

且约束条件（4）变为

$$\int_0^T \sum_{j=1}^{J} \left[a_j(t)^{(m)}\right]^2 \mathbb{P}(X(t)=e_j)\ \mathrm{d}t = 1.$$

证明。
由 $\zeta(t)^{(m)} = \frac{1}{\lambda}\sum_{j=1}^{J} \mathrm{E}(Z^{(m)}|X(t)=e_j)\mathbb{1}_{e_j}(t)$ 可知，$\zeta(t)^{(m)}$ 可写为

$$\zeta(t)^{(m)} = \sum_{j=1}^{J} a_j(t)^{(m)} \mathbb{1}_{e_j}(t)$$

其中 $a_j(t)^{(m)} = \frac{1}{\lambda}\mathrm{E}(Z^{(m)}|X(t)=e_j)$。

那么方程（3）变为

$$\int_0^T K(t',t) \sum_{j=1}^{J} a_j(t)^{(m)} \mathbb{1}_{e_j}(t)\ \mathrm{d}t = \lambda \sum_{j'=1}^{J} a_{j'}(t')^{(m)} \mathbb{1}_{e_{j'}}(t')$$

$$\int_0^T \sum_{j=1}^{J} a_j(t)^{(m)} E_{t'}\left[E_t(\mathbb{1}_{e_j}(t))\right]\ \mathrm{d}t = \lambda \sum_{j'=1}^{J} a_{j'}(t')^{(m)} \mathbb{1}_{e_{j'}}(t')$$

$$\int_0^T \sum_{j=1}^{J}\sum_{j'=1}^{J} a_j(t)^{(m)} \mathbb{P}(X(t)=e_j|X(t')=e_{j'}) \mathbb{1}_{e_{j'}}(t')\ \mathrm{d}t = \lambda \sum_{j'=1}^{J} a_{j'}(t')^{(m)} \mathbb{1}_{e_{j'}}(t') \tag{7}$$

若 $\{a_j(\cdot)^{(m)}\}$ 对所有 $1 \le j' \le J$ 均满足:

$$\int_0^T \sum_{j=1}^{J} a_j(t)^{(m)} \mathbb{P}(X(t)=e_j|X(t')=e_{j'})\ \mathrm{d}t = \lambda a_{j'}(t')^{(m)}, \tag{8}$$

则它满足(7)。并且由假设$(\mathcal{A}_2)$，(8)等价于

$$\int_0^T \sum_{j=1}^{J} a_j(t)^{(m)} \mathbb{P}(X(t)=e_j, X(t')=e_{j'})\ \mathrm{d}t = \lambda a_{j'}(t')^{(m)} \mathbb{P}(X(t')=e_{j'}). \tag{9}$$

约束(4)可以写成

$$\int_0^T \mathbb{E}\left[\sum_{j=1}^{J}\sum_{j'=1}^{J} a_j(t)^{(m)} a_{j'}(t)^{(m)} \mathbb{1}_{e_j}(t)\mathbb{1}_{e_{j'}}(t)\right]\ \mathrm{d}t = 1$$

$$\int_0^T \sum_{j=1}^{J} \left[a_j(t)^{(m)}\right]^2 \mathbb{P}(X(t)=e_j)\ \mathrm{d}t = 1.$$

证明到此结束。 □

命题的证明 4

回顾命题4。由 $\zeta(t)^{(0)}$、……、$\zeta(t)^{(M)}$，方程（2）的解 $Z^{(0)}$、……、$Z^{(M)}$ 可推导得出

$$Z^{(m)} = \int_0^T \zeta(t)^{(m)} \mathrm{d}t.$$

证明 令 $Z^{(m)} = \int_0^T \zeta(t)^{(m)} \mathrm{d}t$，则

$$\begin{aligned}\int_0^T E_t\left[Z^{(m)}\right] \mathrm{d}t &= \int_0^T E_t\left[\int_0^T \zeta(t')^{(m)} \mathrm{d}t'\right] \mathrm{d}t \\ &= \int_0^T \int_0^T E_t\left[\zeta(t')^{(m)}\right] \mathrm{d}t' \mathrm{d}t\end{aligned}$$

接下来，由 $\zeta(t')^{(m)} = \sum_{j=1}^{J} a_j(t')^{(m)} \mathbb{1}_{e_j}(t')$，

$$\begin{aligned}E_{t'}(\zeta(t')^{(m)}) &= \sum_{j=1}^{J} a_j(t')^{(m)} \sum_{j'=1}^{J} \mathbb{P}(X(t') = e_j | X(t') = e_{j'}) \mathbb{1}_{e_{j'}}(t') \\ &= \sum_{j=1}^{J} a_j(t')^{(m)} \mathbb{1}_{e_j}(t') = \zeta(t')^{(m)}.\end{aligned}$$

因此，

$$\begin{aligned}\int_0^T E_t\left[Z^{(m)}\right] \mathrm{d}t &= \int_0^T \int_0^T E_t E_{t'}\left[\zeta(t')^{(m)}\right] \mathrm{d}t' \mathrm{d}t \\ &= \int_0^T \lambda^{(m)} \zeta(t)^{(m)} \mathrm{d}t = \lambda^{(m)} Z^{(m)}.\end{aligned}$$

由方程（2），可得对于 $\zeta(t)^{(m)} = \frac{1}{\lambda} E_t(Z^{(m)})$ （3）的解，

$$\int_0^T E_t(Z^{(m)}) \mathrm{d}t = \lambda^{(m)} Z^{(m)}.$$

那么，$\int_0^T \lambda^{(m)} \zeta(t)^{(m)} \mathrm{d}t = \lambda^{(m)} Z^{(m)}$，并且可以写成

$$Z^{(m)} = \int_0^T \zeta(t)^{(m)} \mathrm{d}t.$$

□

命题的证明 5

回顾命题5。

(2) $Z^{(1)}, \ldots, Z^{(M)}$ 的非平凡解是中心化且互不相关的随机变量，使得 $\mathbb{V}(Z^{(m)}) = \lambda^{(m)}, 1 \leq m \leq M$.

证明。$\mathbb{E}(Z^{(m)}) = \int_0^T \mathbb{E}\left(\zeta(t)^{(m)}\right) \mathrm{d}t = 0$ 由命题2得出。于是，对于 $1 \leq m, m' \leq M, m \neq m'$,

$$
\begin{aligned}
Cov(Z^{(m)}, Z^{(m')}) = \mathbb{E}(Z^{(m)} Z^{(m')}) &= \mathbb{E}\left(\int_0^T \int_0^T \zeta(t)^{(m)} \zeta(t')^{(m')} \, \mathrm{d}t \, \mathrm{d}t'\right) \\
&= \mathbb{E}\left[\int_0^T \int_0^T E_t\left(\zeta(t)^{(m)}\right) E_{t'}\left(\zeta(t')^{(m')}\right) \, \mathrm{d}t \, \mathrm{d}t'\right] \text{ since } E_t(\zeta(t)^{(m)}) = \zeta(t)^{(m)} \\
&= \int_0^T \int_0^T \mathbb{E}\left[E_t\left(\zeta(t)^{(m)}\right) E_{t'}\left(\zeta(t')^{(m')}\right)\right] \, \mathrm{d}t \, \mathrm{d}t' \\
&= \int_0^T \int_0^T \mathbb{E}\left\{\zeta(t)^{(m)} E_t\left[E_{t'}\left(\zeta(t')^{(m')}\right)\right]\right\} \, \mathrm{d}t \, \mathrm{d}t' \text{ since } E_t \text{ is self-adjoint} \\
&= \int_0^T \mathbb{E}\left\{\zeta(t)^{(m)} \int_0^T E_t\left[E_{t'}\left(\zeta(t')^{(m')}\right)\right] \mathrm{d}t'\right\} \, \mathrm{d}t \\
&= \int_0^T \mathbb{E}\left(\zeta(t)^{(m)} \lambda^{(m')} \zeta(t)^{(m')}\right) \, \mathrm{d}t \text{ from Equation (3)} \\
&= \lambda^{(m')} \int_0^T \mathbb{E}\left(\zeta(t)^{(m)} \zeta(t)^{(m')}\right) \, \mathrm{d}t = 0 \text{ since } \zeta^{(m)} \text{ and } \zeta^{(m')} \text{ are orthogonal}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{V}(Z^{(m)}) &= \mathbb{E}\left[\left(Z^{(m)}\right)^2\right] \\
&= \mathbb{E}\left[\int_0^T \int_0^T \zeta(t)^{(m)} \zeta(t')^{(m)} \, \mathrm{d}t \, \mathrm{d}t'\right] \\
&= \mathbb{E}\left[\int_0^T \int_0^T E_t\left(\zeta(t)^{(m)}\right) E_{t'}\left(\zeta(t')^{(m)}\right) \, \mathrm{d}t \, \mathrm{d}t'\right] \text{ since } E_t(\zeta(t)^{(m)}) = \zeta(t)^{(m)} \\
&= \int_0^T \int_0^T \mathbb{E}\left[E_t\left(\zeta(t)^{(m)}\right) E_{t'}\left(\zeta(t')^{(m)}\right)\right] \mathrm{d}t \, \mathrm{d}t' \\
&= \int_0^T \mathbb{E}\left\{\zeta(t)^{(m)} \int_0^T E_t\left[E_{t'}\left(\zeta(t')^{(m)}\right)\right] \mathrm{d}t'\right\} \mathrm{d}t \text{ since } E_t \text{ is self-adjoint} \\
&= \int_0^T \mathbb{E}\left\{\zeta(t)^{(m)} \lambda^{(m)} \zeta(t)^{(m)}\right\} \mathrm{d}t \text{ from Equation (3)} \\
&= \lambda^{(m)} \int_0^T \mathbb{E}\left\{\left[\zeta(t)^{(m)}\right]^2\right\} \mathrm{d}t = \lambda^{(m)} \text{ from Constraint (4)}
\end{aligned}
$$

证明到此结束。 □

B 模拟研究的补充材料

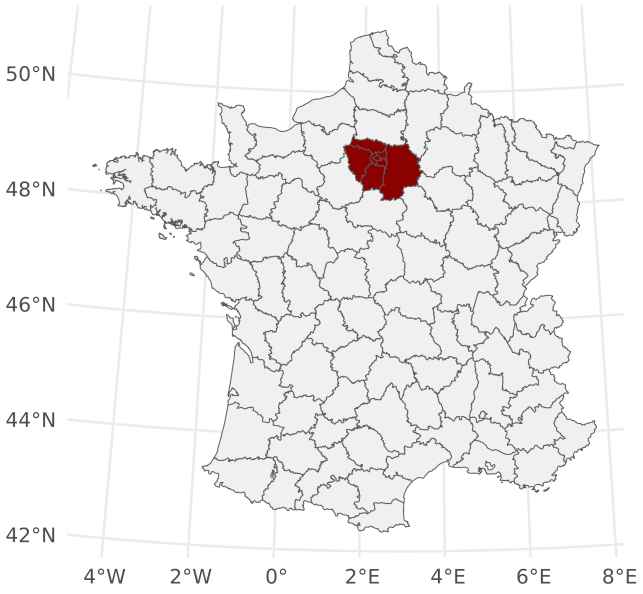

图6：法国本土的94个省，以及为每个人工数据集模拟的空间聚类（红色）。

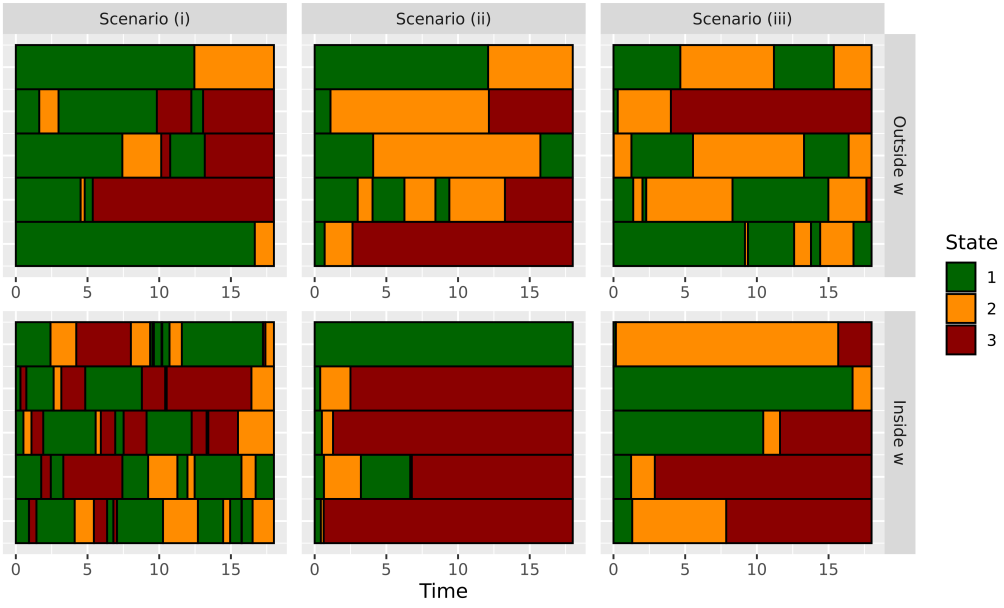

图7：对于三种情景（情景 (i)：$\alpha = 0.6$，情景 (ii)：$\alpha = 0.25$，情景 (iii)：$\gamma = 0.4$），模拟簇内外的五条个体轨迹 $w_\circ$

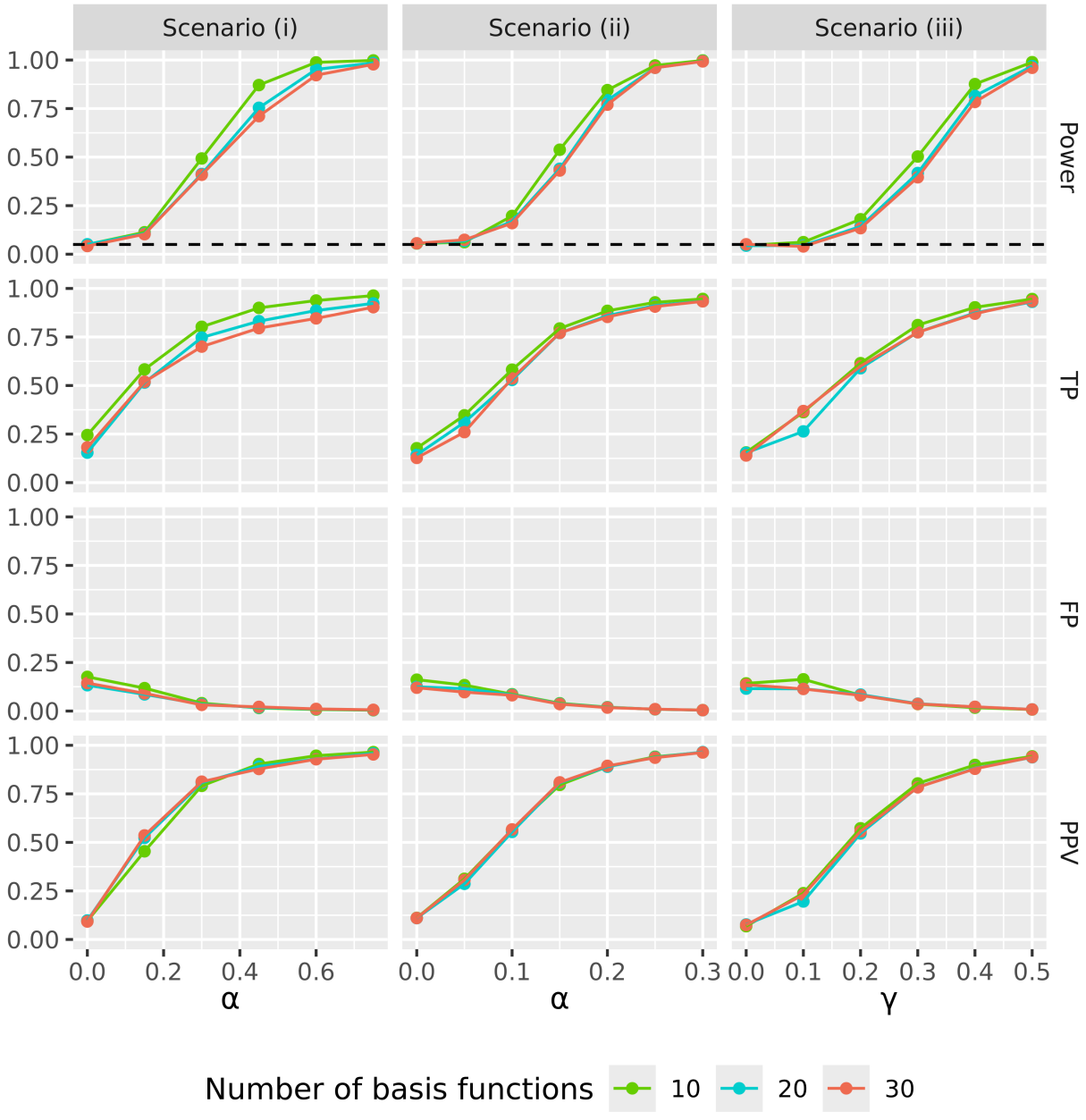

图2：模拟 CFSS 在分类函数型数据中检测空间聚类的能力。对于每种情景，图中展示了将该空间聚类检测为 MLC 时的功效曲线、真阳性率、假阳性率和阳性预测值。 α 和γ 是控制聚类强度的参数。

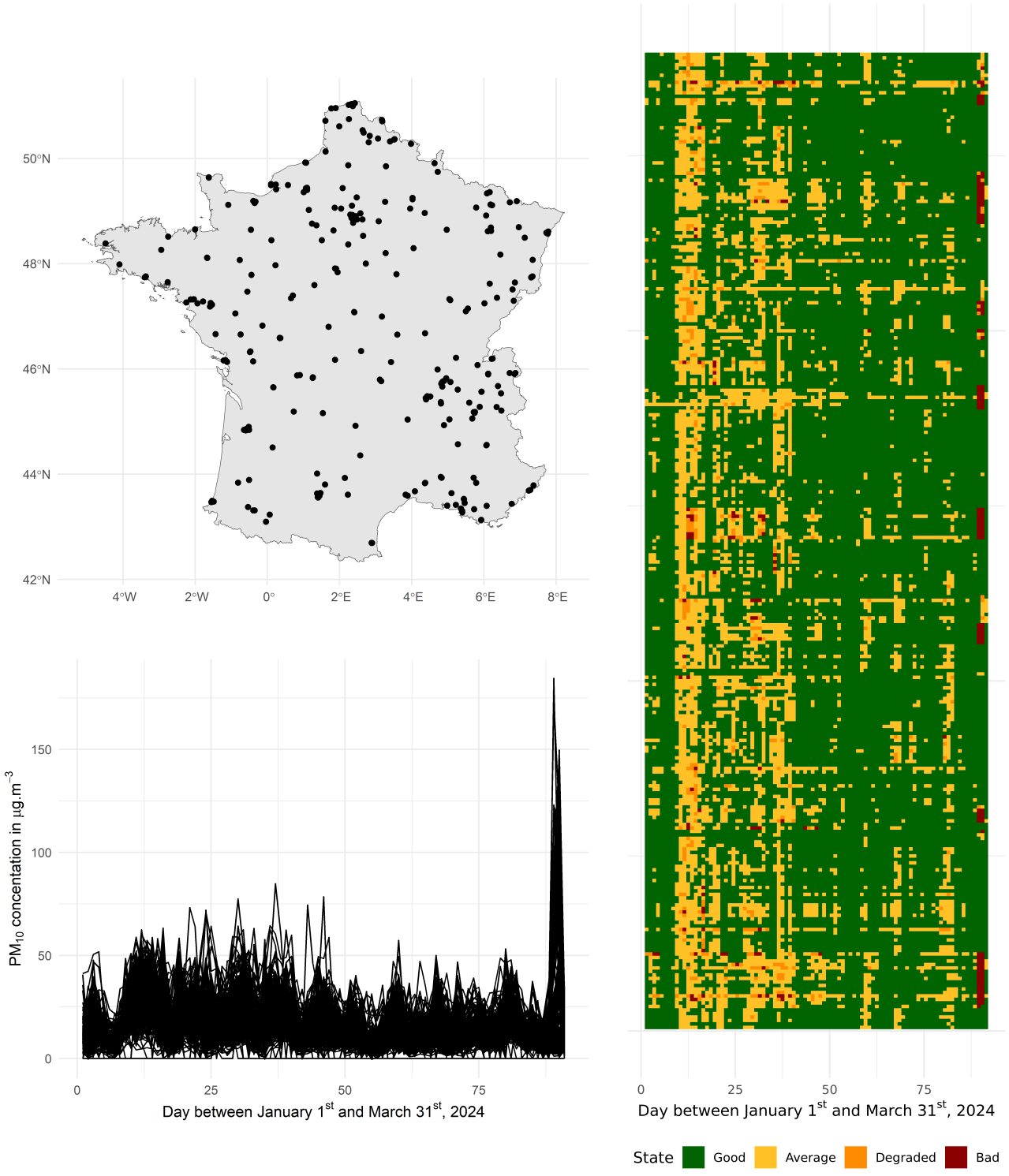

图3：监测站分布图，以及2024年1月 1st 至3月 31^{st}的日均 PM_{10} 浓度曲线（左图）和相应的类别轨迹（右图）。

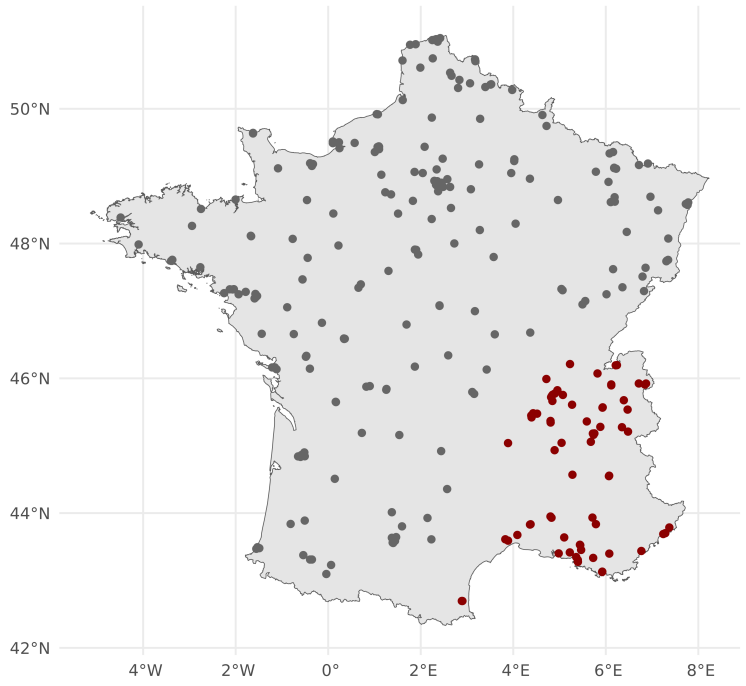

图 4：使用 CFSS 检测到的 MLC（红色），针对每日平均 PM_{10} 浓度（按 Atmo指数分类）。

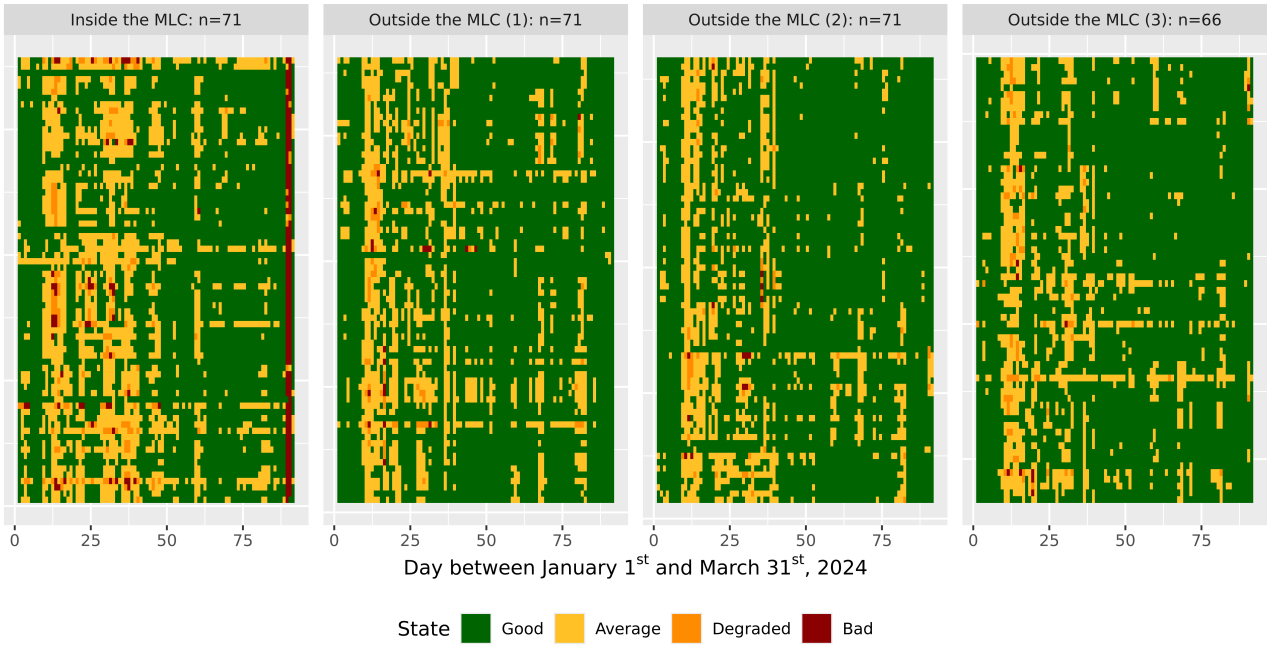

图 5：使用 CFSS 检测到的 MLC 内部和外部的轨迹，针对每日平均 PM_{10} 浓度（按 Atmo指数分类）。