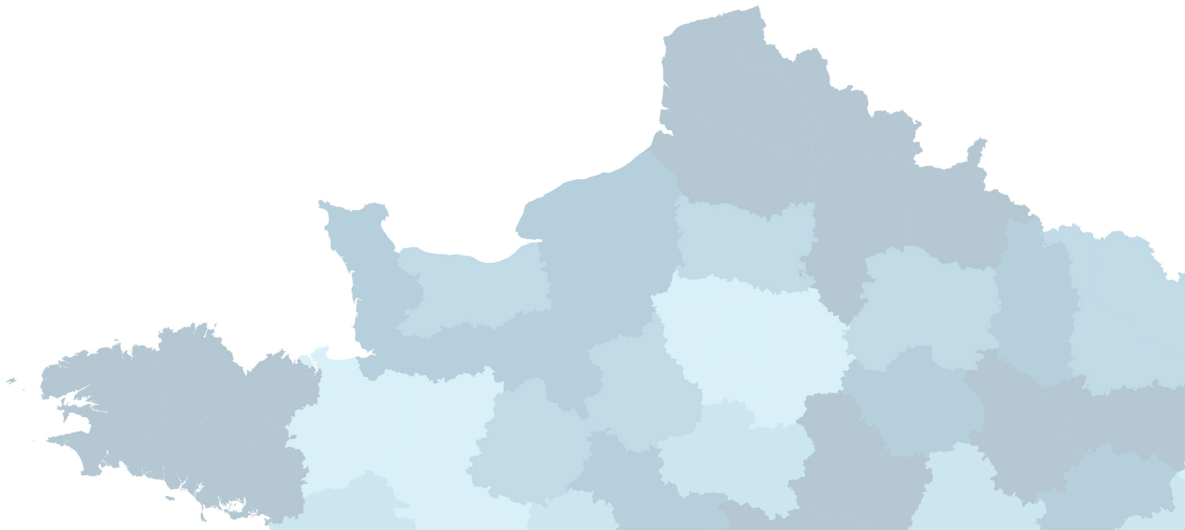
Méthodes d'analyses spatiales pour données fonctionnelles

Michaël Génin

Univ. Lille – ULR 2694 METRICS | CHU de Lille – SEED

Webinar QuanTIM, 13 octobre 2023



Plan

Introduction

Université de Lille – ULR 2694 METRICS
Evaluation des technologies de santé et des pratiques médicales

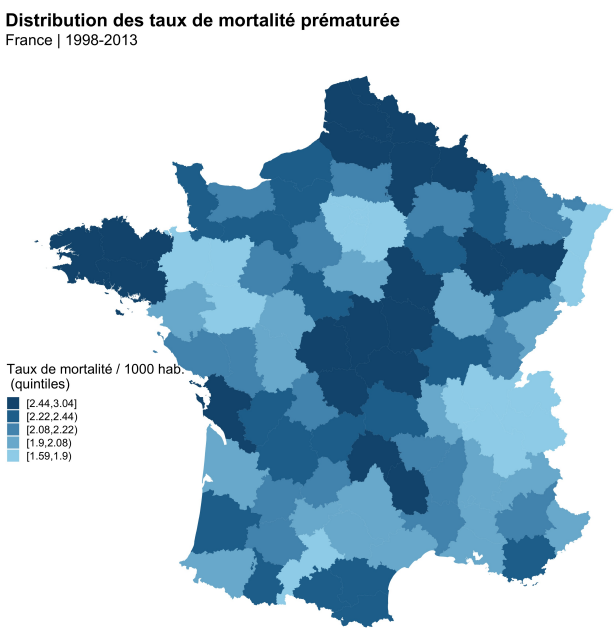
Statistique spatiale

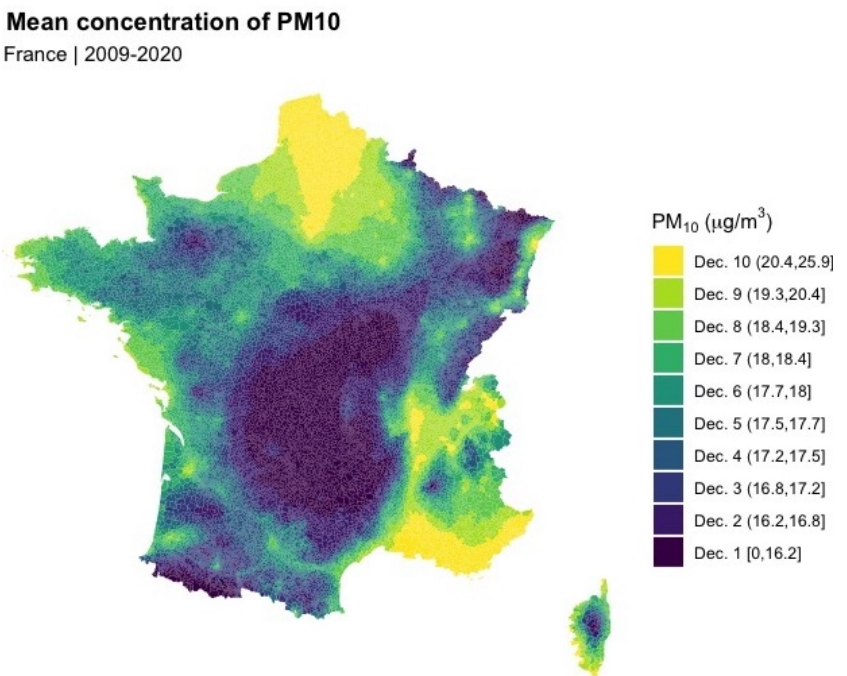
- Modèles bayésiens spatio-temporels pour le *disease mapping* et les études de corrélations écologiques
- Détection de clusters spatiaux et spatio-temporels pour données à haute dimension et de survie

Epidémiologie environnementale

- Analyse des liens entre exposition environnementale et évènements de santé par approches spatiales et spatio-temporelles
- Identification de hotspots/coldspots d'évènement sanitaires / environnementaux

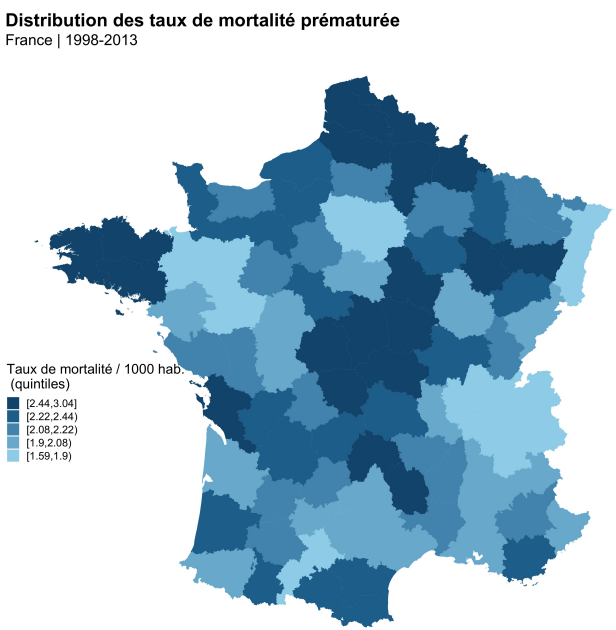
Données spatiales : ensemble de données $\{X_1, \ldots, X_n\}$ mesurées au sein de n sites spatiaux $s_1, \ldots, s_n \in \mathcal{S} \subset \mathbb{R}^2$

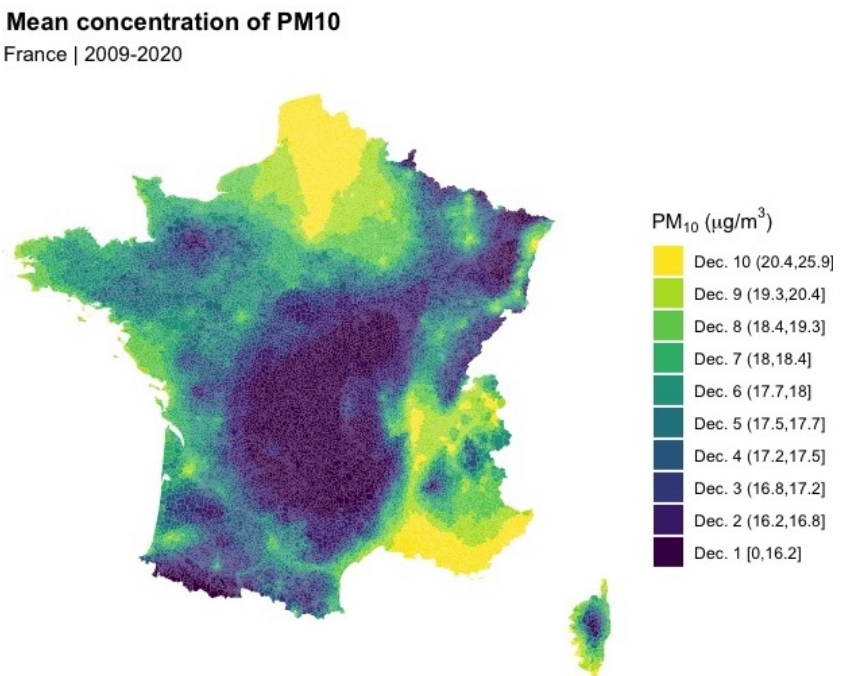




Université de Lille

Données spatiales : ensemble de données $\{X_1, \ldots, X_n\}$ mesurées au sein de n sites spatiaux $s_1, \ldots, s_n \in \mathcal{S} \subset \mathbb{R}^2$



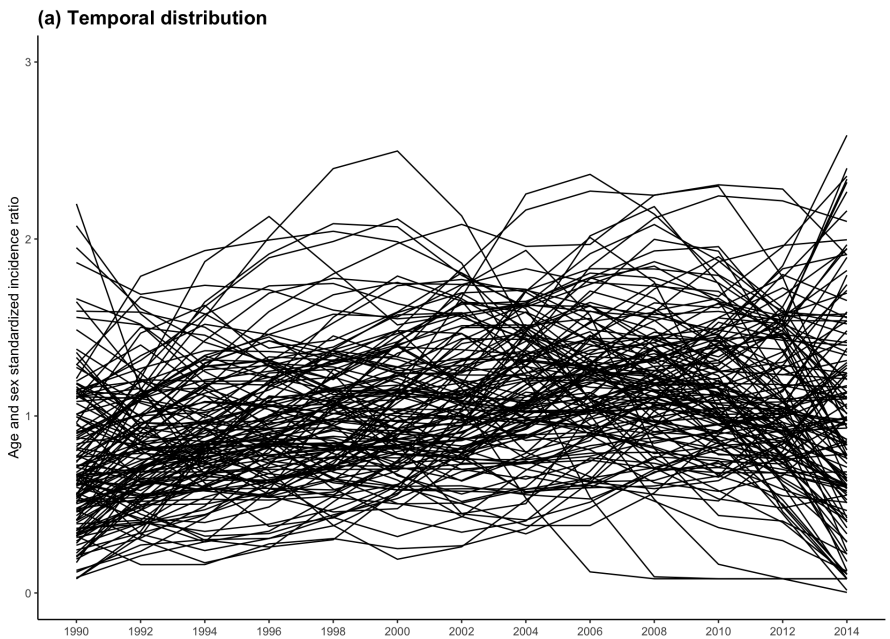


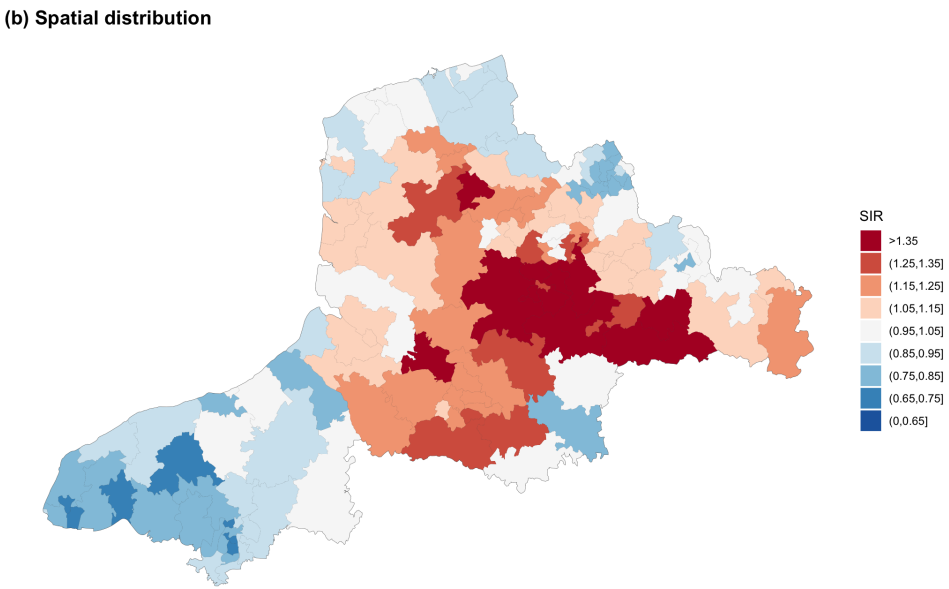
Données spatio-temporelles : ensemble de données $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ mesurées

- au sein de n sites spatiaux $s_1, \ldots, s_n \in \mathcal{S} \subset \mathbb{R}^2$
- au cours du temps : $\mathbf{X}_i = \{X_{i,t_1}, X_{i,t_2}, \ldots, X_{i,m_i}\}$

Données d'incidence d'une pathologie (Registre, système de surveillance. . .)

- Maladie de Crohn (registre EPIMAD)
- Données spatiales : échelle cantonale (140) dans le nord de la France
- Données longitudinales : incidences annuelles (1990-2014)



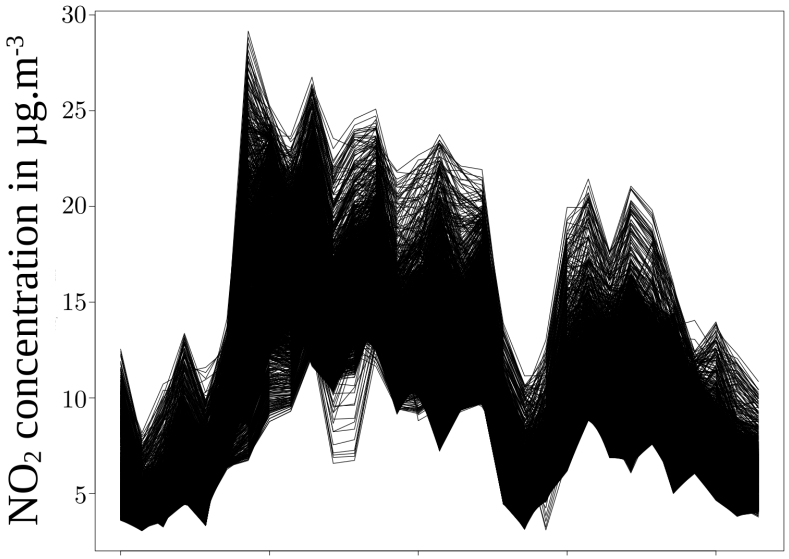


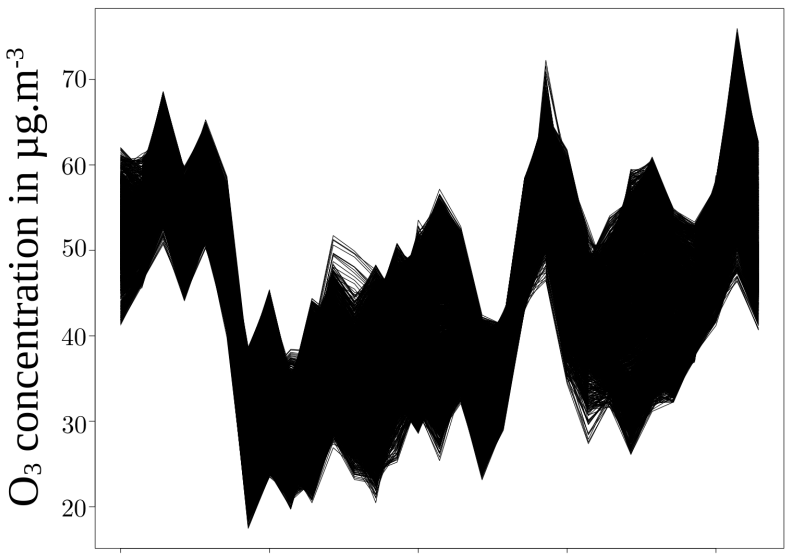
Données environnementales

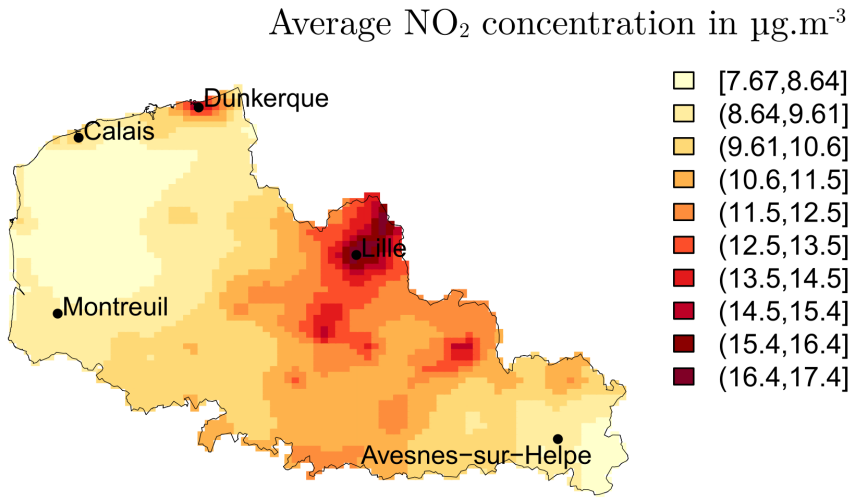
- Données spatiales : capteurs ($\pm$ interpolation spatiale)
- Données longitudinales : mesures horaires, journalières...

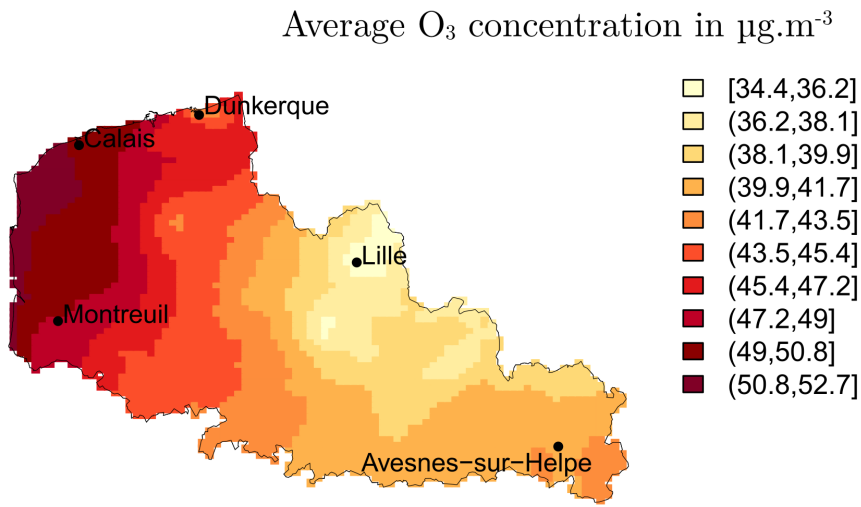
Pollution atmosphérique : données PREV'AIR (Ineris)

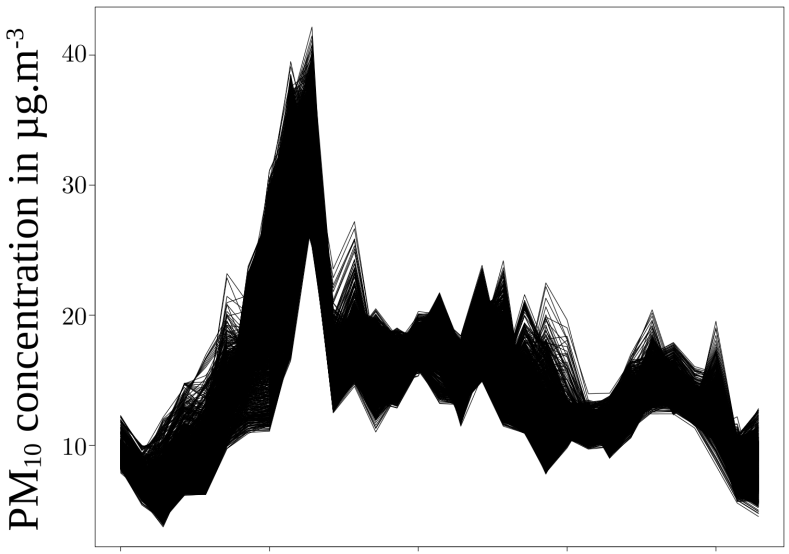
- NO_2, PM_{10}, $PM_{2.5}$, O_3
- Concentrations journalières moyennes ($\mu g.m^{-3}$)
- Période : 1^{er} octobre au 31 octobre 2021
- 3231 cellules de 2 km $\times$ 2 km dans le Nord-Pas-de-Calais

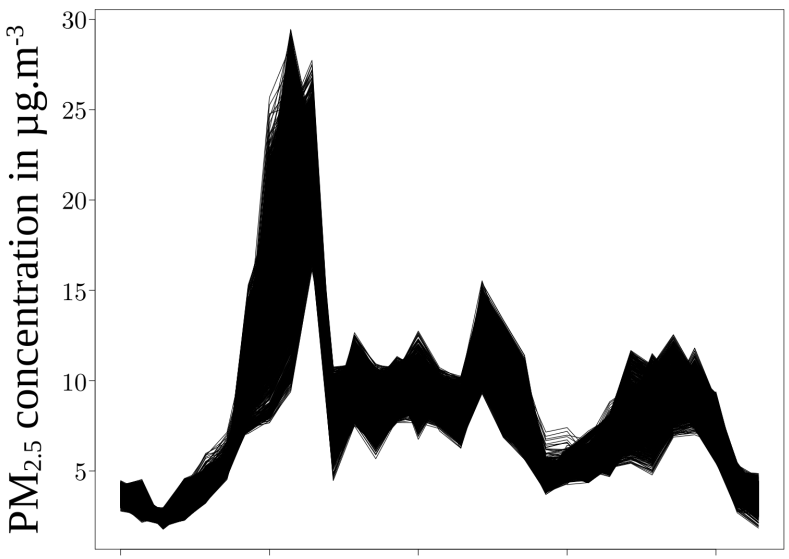

NO_2 concentration in µg.m^{-3}
30
25
20
15
10
5

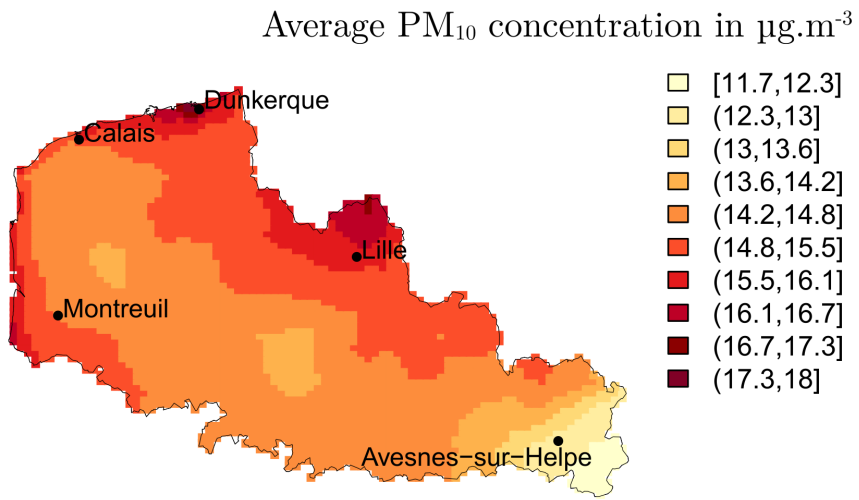

O_3 concentration in µg.m^{-3}
70
60
50
40
30
20

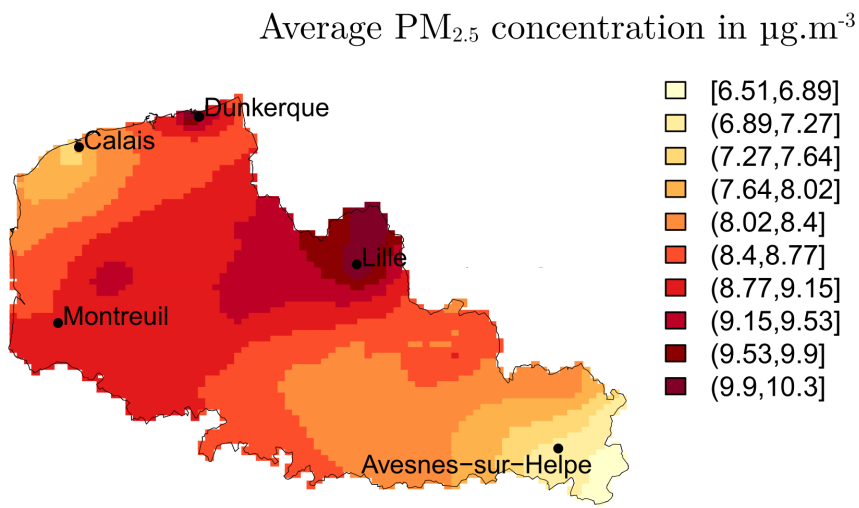

Average NO_2 concentration in µg.m^{-3}
Dunkerque
Calais
Lille
Montreuil
Avesnes−sur−Helpe
[7.67,8.64]
(8.64,9.61]
(9.61,10.6]
(10.6,11.5]
(11.5,12.5]
(12.5,13.5]
(13.5,14.5]
(14.5,15.4]
(15.4,16.4]
(16.4,17.4]


Average O_3 concentration in µg.m^{-3}
Dunkerque
Calais
Lille
Montreuil
Avesnes−sur−Helpe
[34.4,36.2]
(36.2,38.1]
(38.1,39.9]
(39.9,41.7]
(41.7,43.5]
(43.5,45.4]
(45.4,47.2]
(47.2,49]
(49,50.8]
(50.8,52.7]


PM10 concentration in µg.m-3
40
30
20
10


PM2.5 concentration in µg.m-3
30
25
20
15
10
5

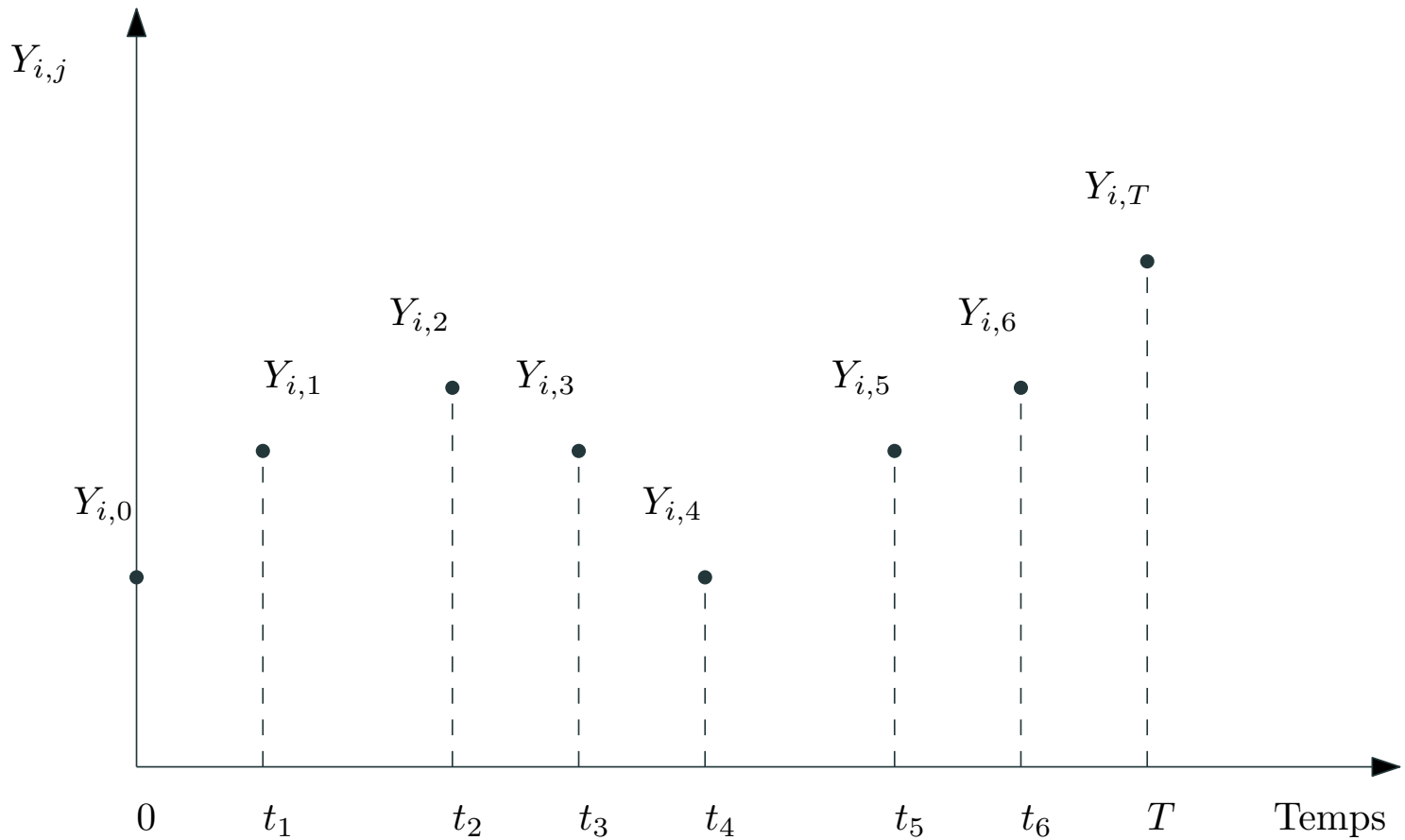

Average PM10 concentration in µg.m-3
Dunkerque
Calais
Lille
Montreuil
Avesnes−sur−Helpe
[11.7,12.3]
(12.3,13]
(13,13.6]
(13.6,14.2]
(14.2,14.8]
(14.8,15.5]
(15.5,16.1]
(16.1,16.7]
(16.7,17.3]
(17.3,18]


Average PM2.5 concentration in µg.m-3
Dunkerque
Calais
Lille
Montreuil
Avesnes−sur−Helpe
[6.51,6.89]
(6.89,7.27]
(7.27,7.64]
(7.64,8.02]
(8.02,8.4]
(8.4,8.77]
(8.77,9.15]
(9.15,9.53]
(9.53,9.9]
(9.9,10.3]

Caractéristiques temporelles de ces données

Pour chaque site i, les données Y_i sont observées en un nombre fini de temps (discret)

$$Y_i = \{Y_{i,0}, \ldots, Y_{i,m_i}\}$$

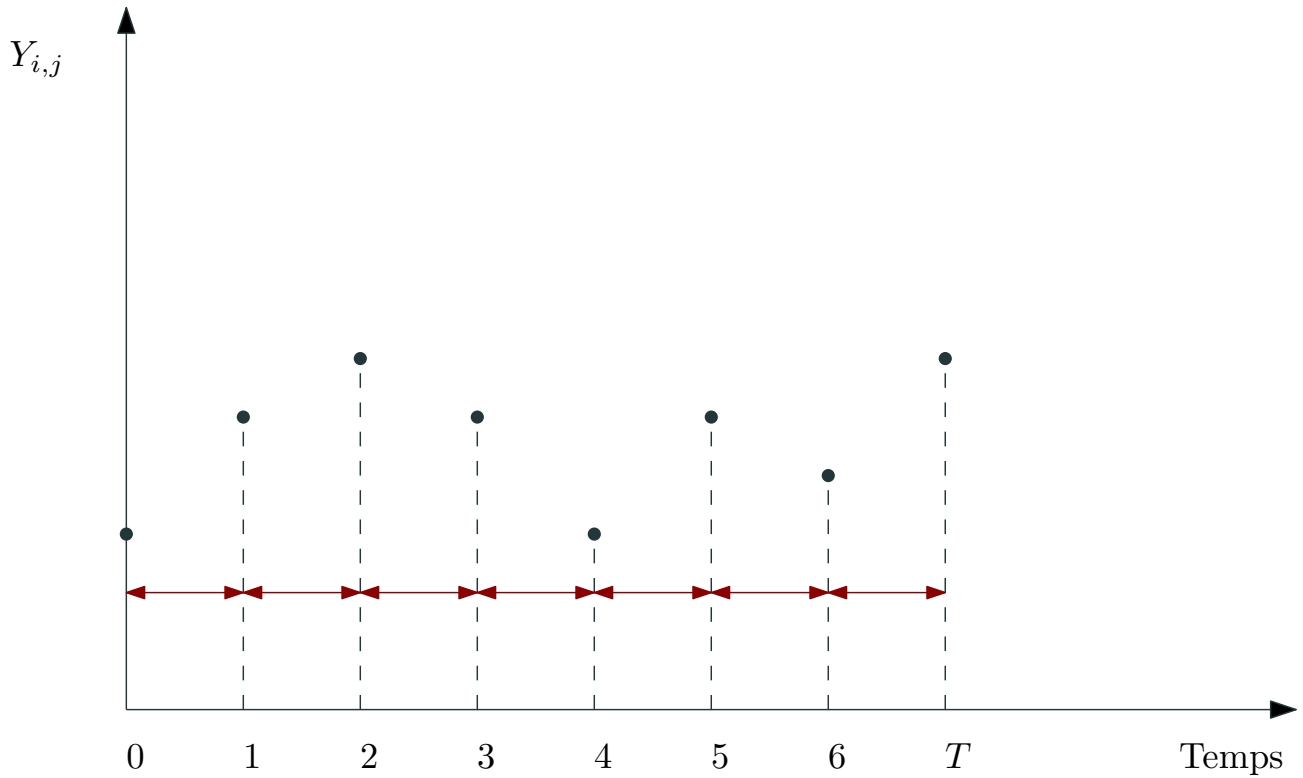


Caractéristiques temporelles de ces données

Pour chaque site i, les données Y_i sont observées en un nombre fini de temps (discret)

$$Y_i = \{Y_{i,0}, \ldots, Y_{i,m_i}\}$$

- Les temps de mesures peuvent être réguliers,

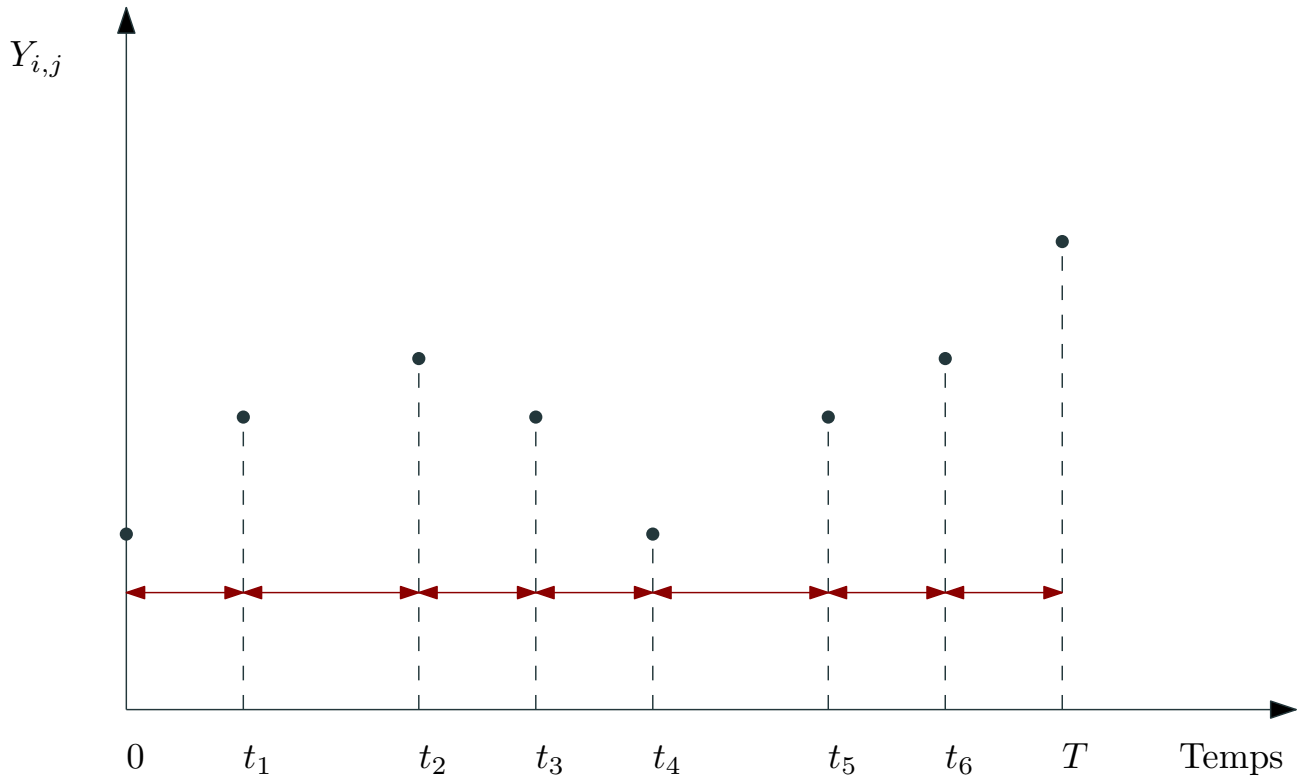


Caractéristiques temporelles de ces données

Pour chaque site i, les données Y_i sont observées en un nombre fini de temps (discret)

$$Y_i = \{Y_{i,0}, \ldots, Y_{i,m_i}\}$$

- Les temps de mesures peuvent être réguliers, irréguliers

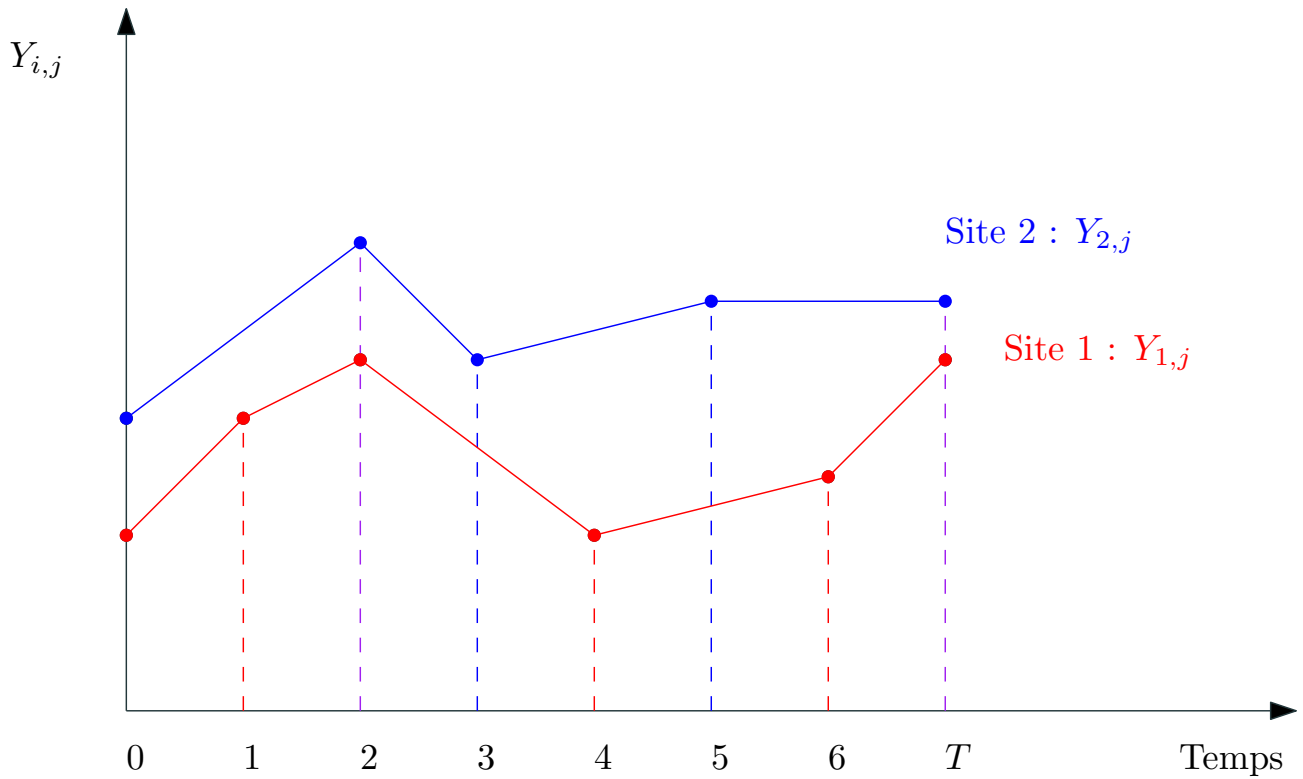


Caractéristiques temporelles de ces données

Pour chaque site i, les données Y_i sont observées en un nombre fini de temps (discret)

$$Y_i = \{Y_{i,0}, \ldots, Y_{i,m_i}\}$$

- Les temps de mesures peuvent être réguliers, irréguliers
- Les temps de mesures peuvent varier d'un site à l'autre

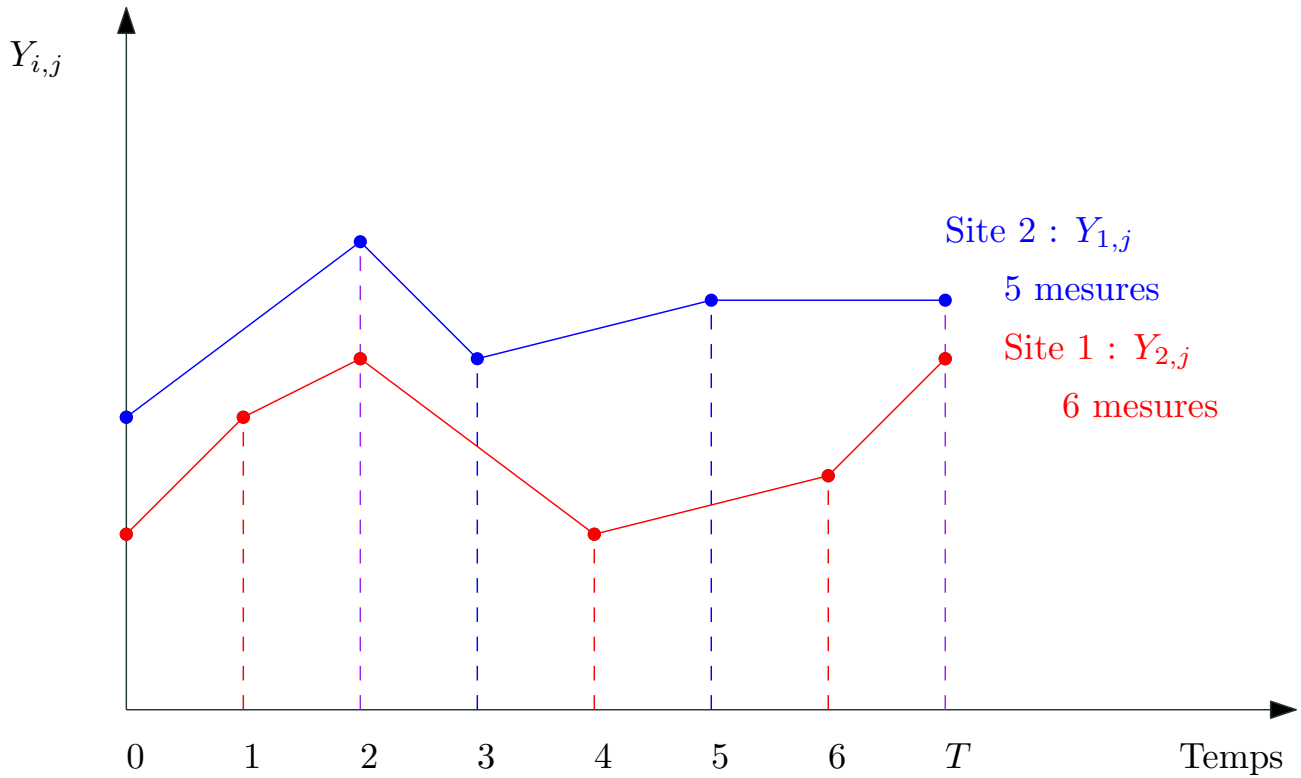


Caractéristiques temporelles de ces données

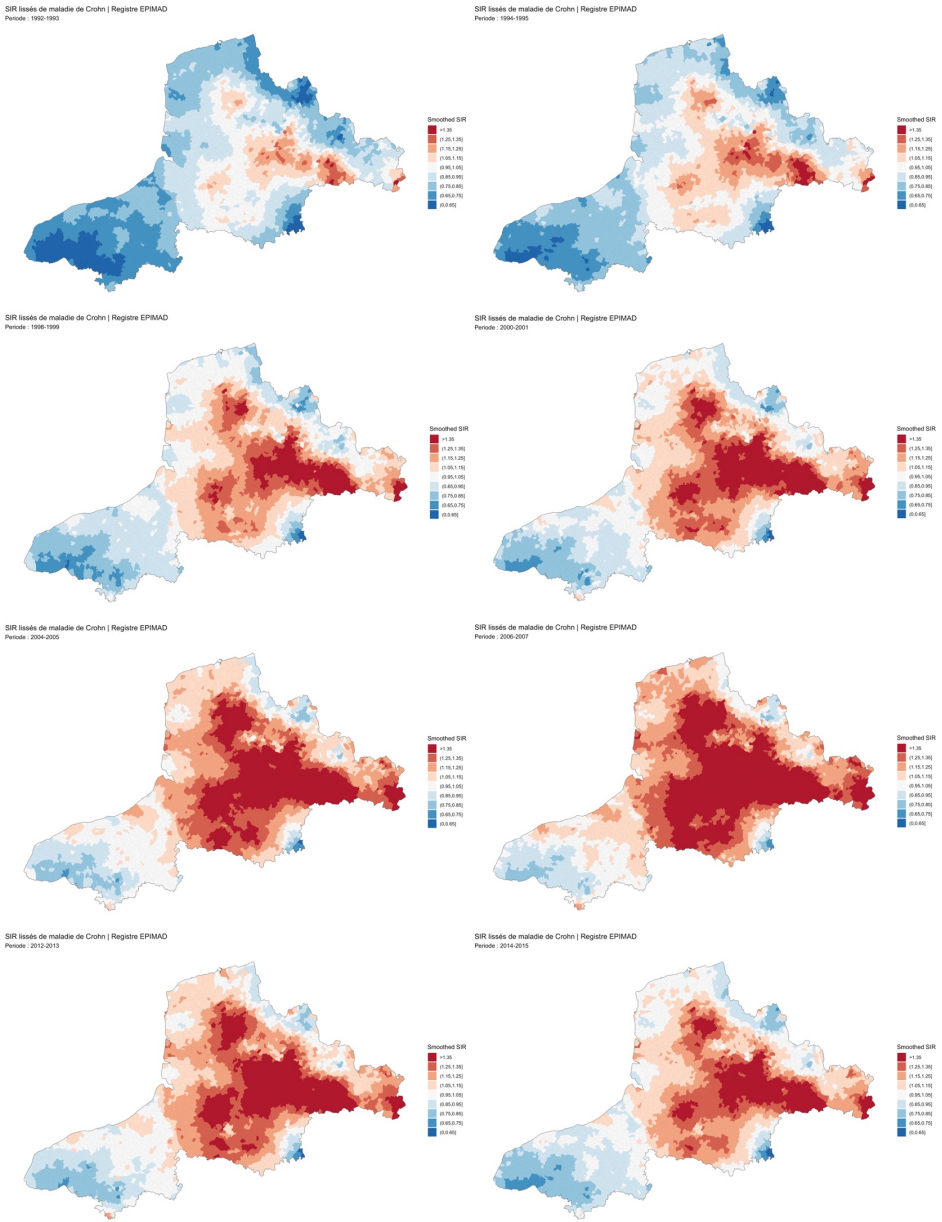
Pour chaque site i, les données Y_i sont observées en un nombre fini de temps (discret)

$$Y_i = \{Y_{i,0}, \ldots, Y_{i,m_i}\}$$

- Les temps de mesures peuvent être réguliers, irréguliers
- Les temps de mesures peuvent varier d'un site à l'autre
- Le nombre de temps de mesures peut varier d'un site à l'autre (missing, ...)



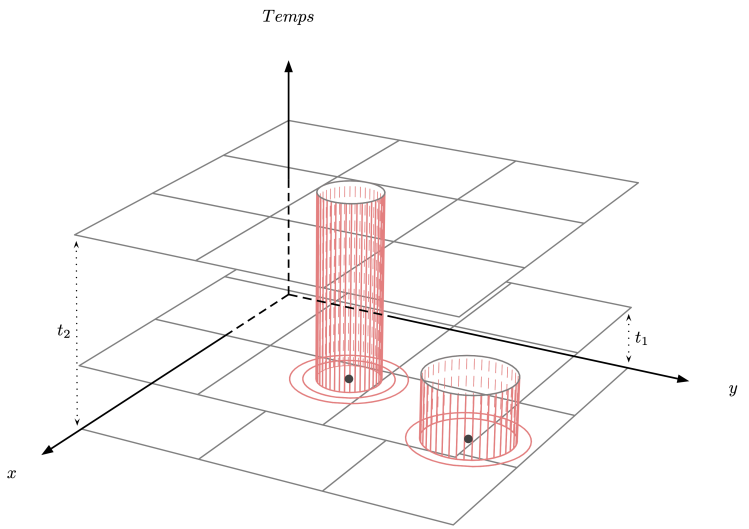
Visualisation



Facteurs associés : régressions

$$Y_{i,t} = f(X^1_{i,t}, X^2_{i,t}, \ldots, X^p_{i,t})$$

Détection de clusters spatio-temporels



Difficultés pour l'analyse statistique spatio-temporelle

Modélisations classiques nécessitent :

- "Au mieux" une donnée observée par site pour l'ensemble des temps de mesures
- Nombre raisonnable de temps de mesure / nombre d'unités spatiales
- Nombre raisonnable d'unités spatiales / nombre de temps de mesure

Difficultés pour l'analyse statistique spatio-temporelle

Modélisations classiques nécessitent :

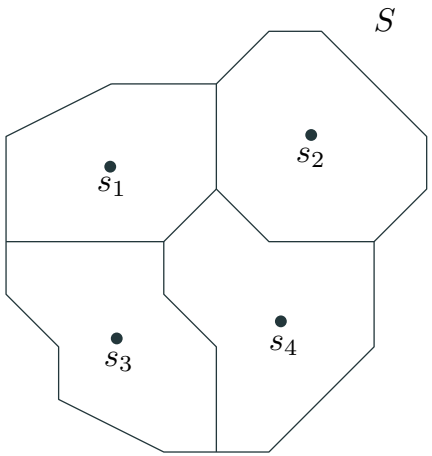
- "Au mieux" une donnée observée par site pour l'ensemble des temps de mesures
- Nombre raisonnable de temps de mesure / nombre d'unités spatiales
- Nombre raisonnable d'unités spatiales / nombre de temps de mesure

Solution

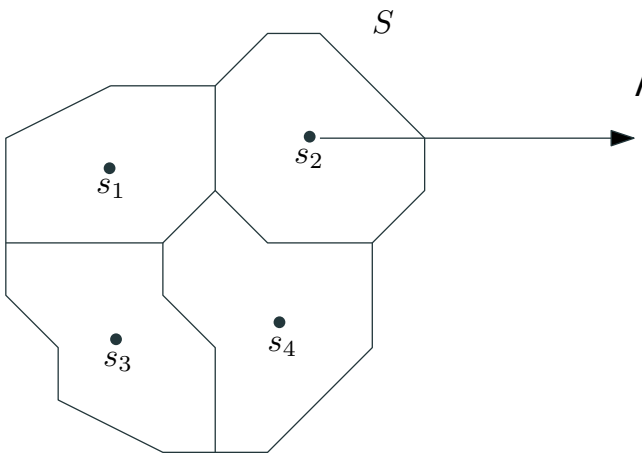
- Modélisation par approche fonctionnelle
- Lissage / interpolation
- Données spatio-fonctionnelles

Données spatio-fonctionnelles

Soient $s_1, \ldots, s_i, \ldots, s_n$ n sites spatiaux d'un domaine $S \subset \mathbb{R}^2$



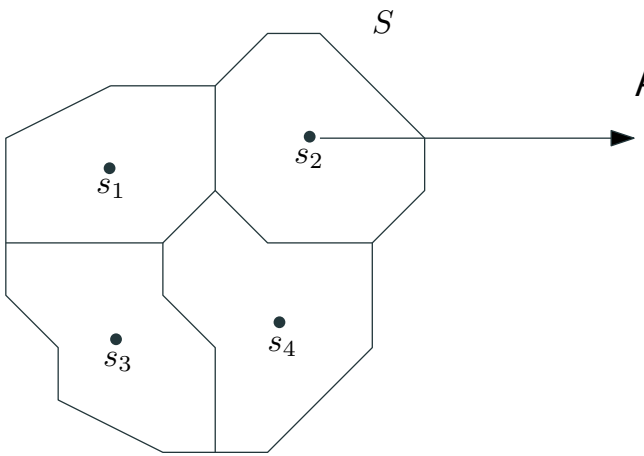
Soient $s_1, \ldots, s_i, \ldots, s_n$ n sites spatiaux d'un domaine $S \subset \mathbb{R}^2$



A chaque site s_i on observe la réalisation d'un p.s. $\{X^{(j)}(t), t \in \mathcal{T}\}$

$$\{X_i^{(1)}(t), X_i^{(2)}(t), \ldots, X_i^{(p)}(t), t \in \mathcal{T}\}, \mathcal{T} = [0, T] \subset \mathbb{R}$$

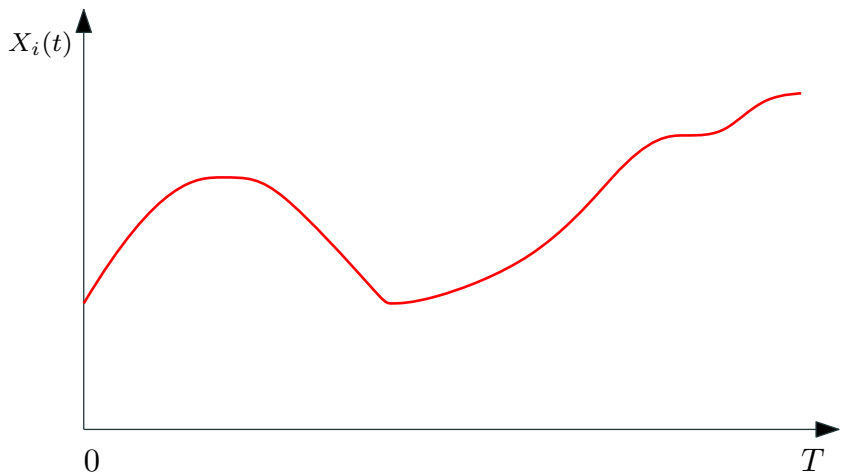
Soient $s_1, \ldots, s_i, \ldots, s_n$ n sites spatiaux d'un domaine $S \subset \mathbb{R}^2$



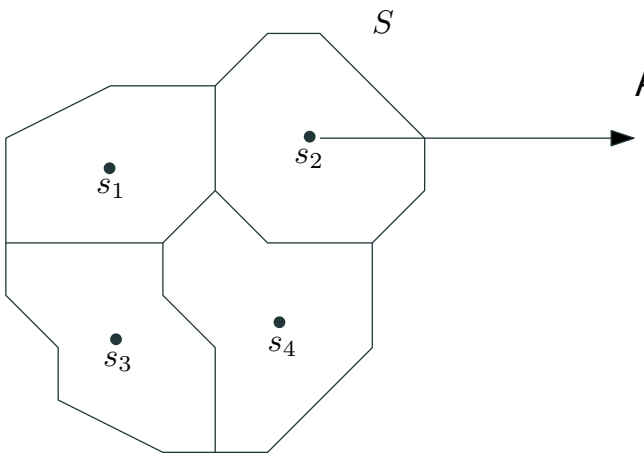
A chaque site s_i on observe la réalisation d'un p.s. $\{X^{(j)}(t), t \in \mathcal{T}\}$

$$\{X_i^{(1)}(t), X_i^{(2)}(t), \ldots, X_i^{(p)}(t), t \in \mathcal{T}\}, \mathcal{T} = [0, T] \subset \mathbb{R}$$

$p = 1 \rightarrow$ Données spatio-fonctionnelles univariées



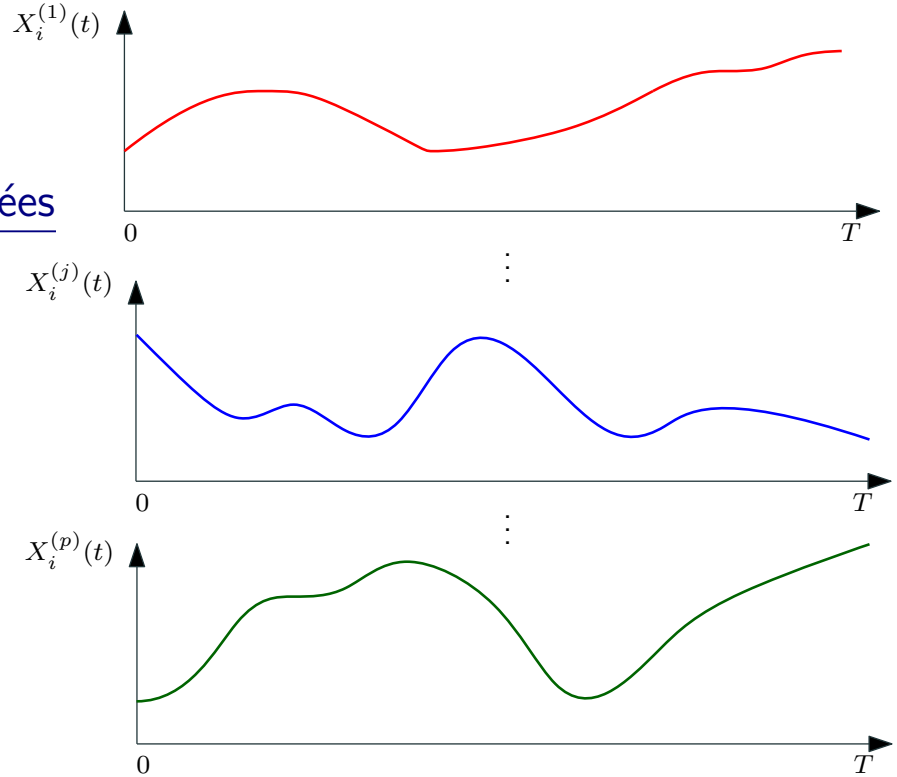
Soient $s_1, \ldots, s_i, \ldots, s_n$ n sites spatiaux d'un domaine $S \subset \mathbb{R}^2$



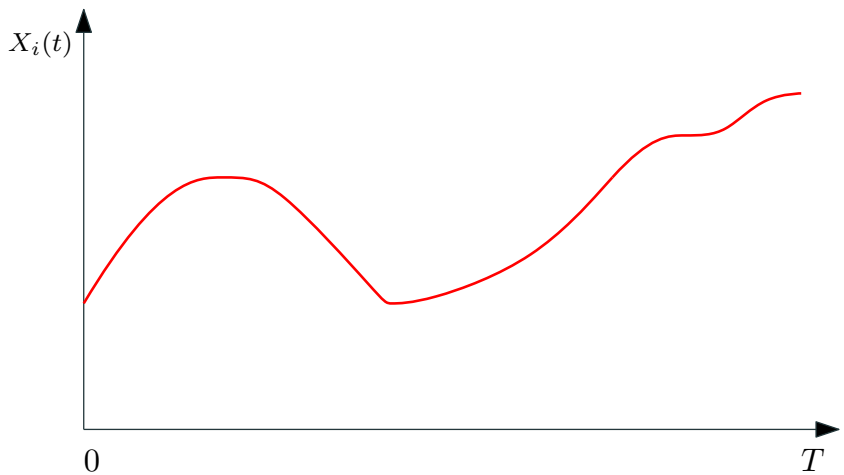
A chaque site s_i on observe la réalisation d'un p.s. $\{X^{(j)}(t), t \in \mathcal{T}\}$

$$\{X_i^{(1)}(t), X_i^{(2)}(t), \ldots, X_i^{(p)}(t), t \in \mathcal{T}\}, \mathcal{T} = [0, T] \subset \mathbb{R}$$

$p \geq 2 \rightarrow$ Données spatio-fonctionnelles multivariées

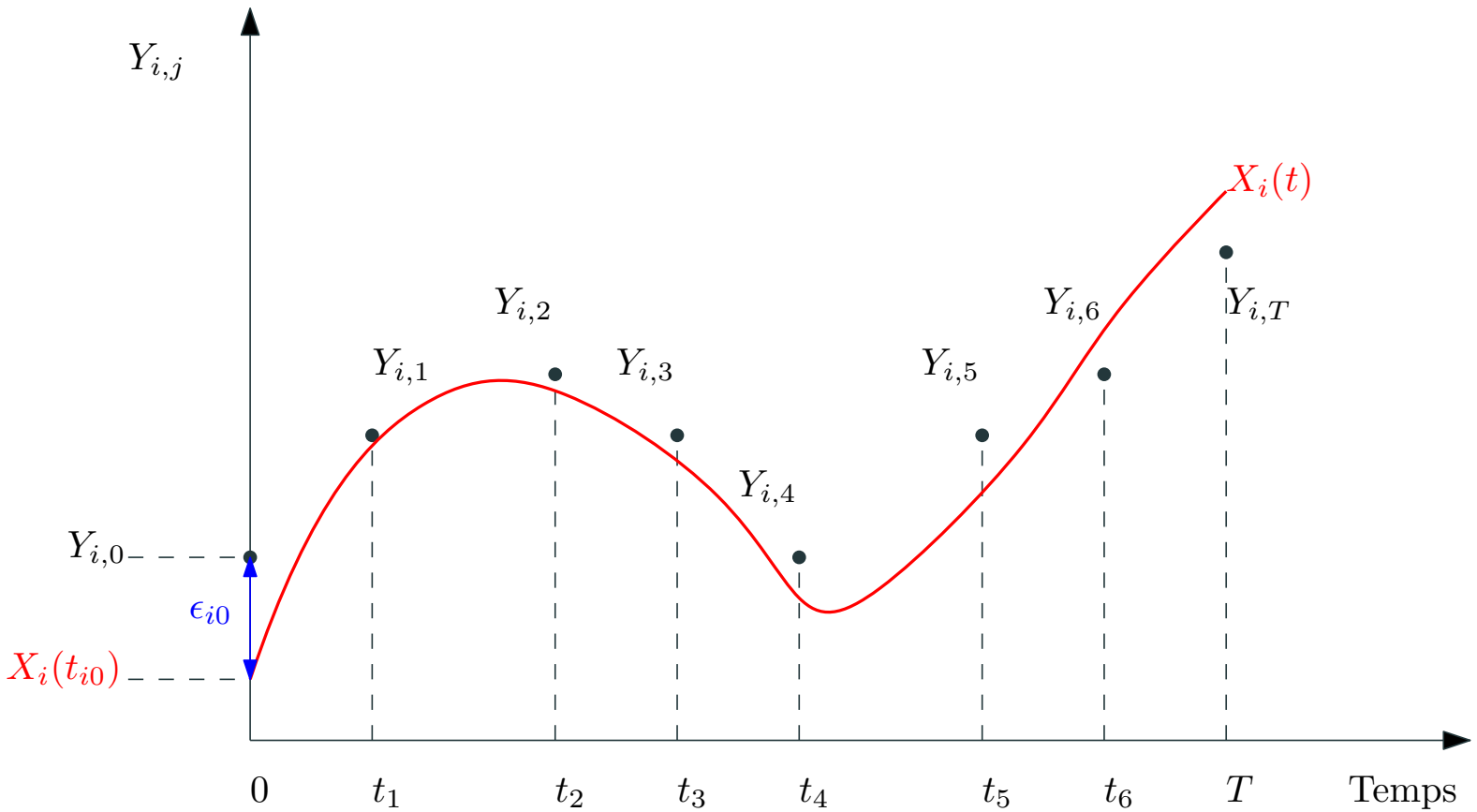


$p = 1 \rightarrow$ Données spatio-fonctionnelles univariées



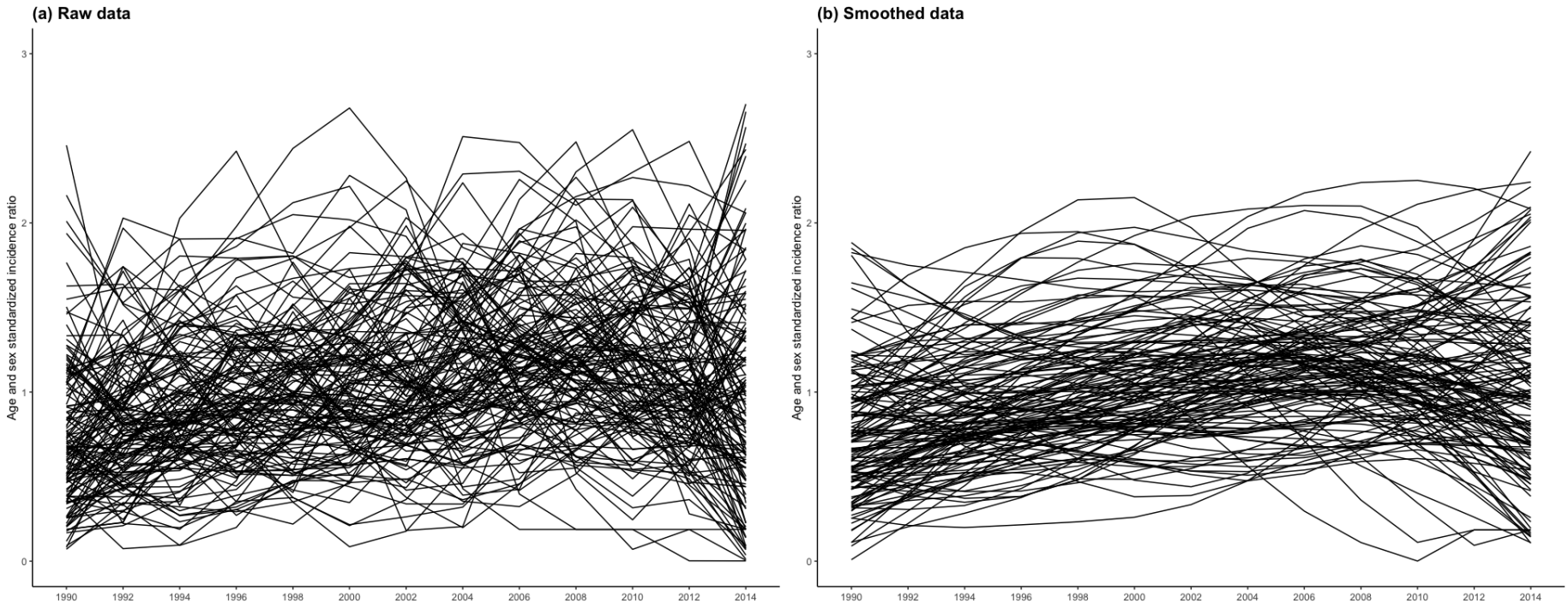
Observations discrètes $Y_{i,j} \rightarrow$ on ne connait pas la fonction $X_i(t)$ sous-jacente...
Mais on peut la modéliser :

$$\underbrace{Y_{i,j}}_{\text{Obs.}} = \underbrace{X_i(t_{ij})}_{\text{Fonction latente}} + \underbrace{\epsilon_{ij}}_{\text{Bruit blanc}}$$



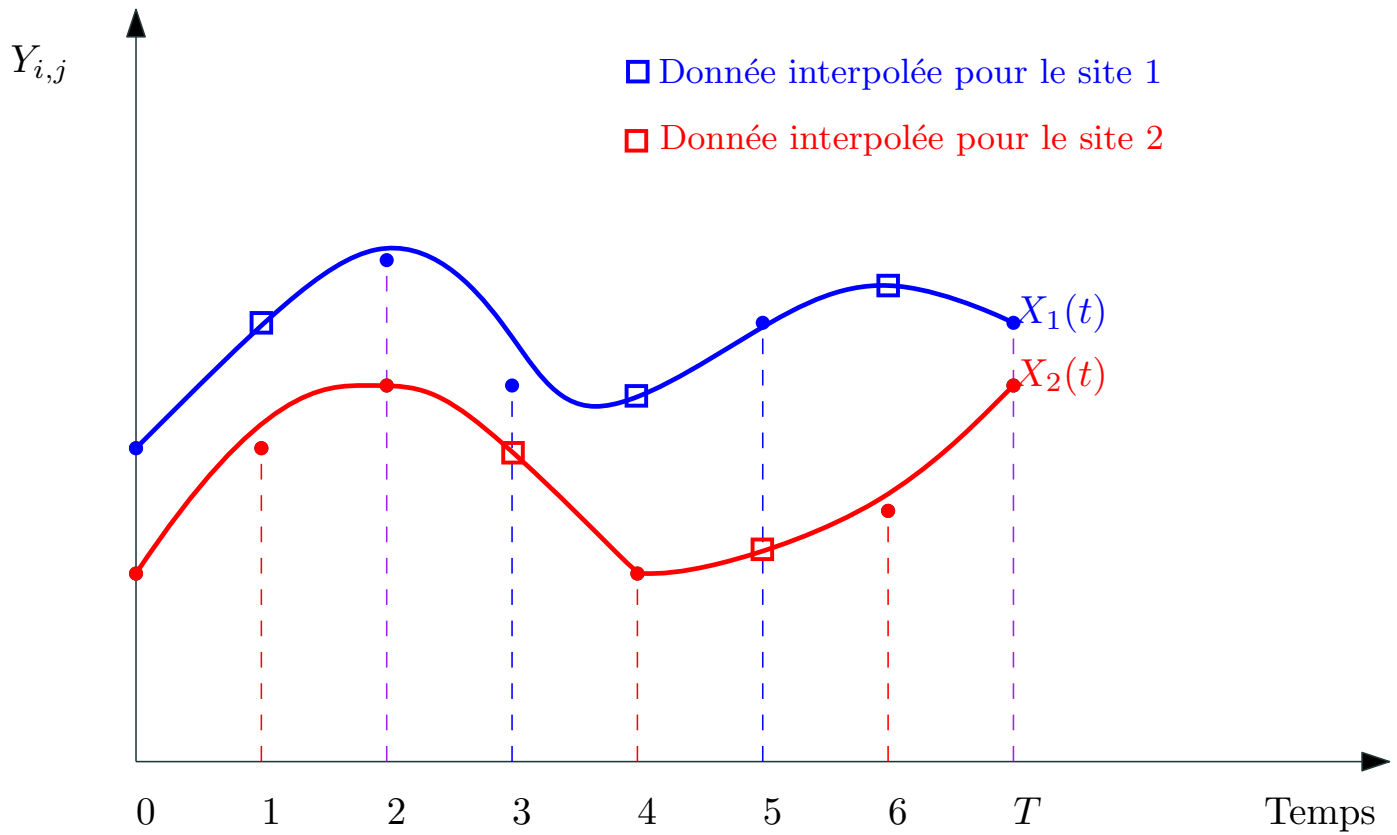
Pour chacun des sites, estimer la fonction $X_i(t)$ permet :

- **de lisser les données, notamment si elles sont observées avec des erreurs**
- de réaliser des interpolations pour pouvoir comparer les observations aux mêmes temps de mesure



Pour chacun des sites, estimer la fonction $X_i(t)$ permet :

- de lisser les données, notamment si elles sont observées avec des erreurs
- **de réaliser des interpolations pour pouvoir comparer les observations aux mêmes temps de mesure**



Pour chacun des sites, estimer la fonction $X_i(t)$ permet :

- de lisser les données, notamment si elles sont observées avec des erreurs
- de réaliser des interpolations pour pouvoir comparer les observations aux mêmes temps de mesure

Problématique : dans la plupart des cas, on ne connaît pas l'expression mathématique de la fonction $\rightarrow$ il faut l'approcher.

Solutions :

- Projection dans des bases de fonctions connues (splines, Fourier, ondelettes...)
- Utilisation de bases de fonctions issues des données (ACP fonctionnelle)
- Approches non-paramétriques (Régression LOESS...)

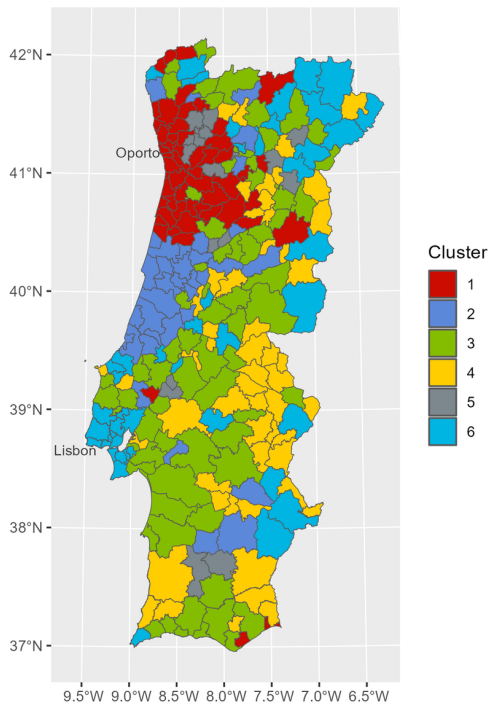
Méthodes d'analyses de données fonctionnelles intégrant la notion de dépendance spatiale :

- Régression [Dabo-Niang *et al.*, 2011 ; Bernardi *et al.*, 2017]
- Réduction de dimension (ACP. . .) [Demsar *et al.*, 2013]
- Clustering [Giraldo *et al.*, 2012 ; Vandewalle *et al.*, 2021]

Méthodes d'analyses de données fonctionnelles intégrant la notion de dépendance spatiale :

- Régression [Dabo-Niang *et al.*, 2011 ; Bernardi *et al.*, 2017]
- Réduction de dimension (ACP. . .) [Demsar *et al.*, 2013]
- Clustering [Giraldo *et al.*, 2012 ; Vandewalle *et al.*, 2021]

⚠ Méthodes de clustering → créer une partition des données spatio-fonctionnelles



Ribeiro *et al.*, 2023

- Les unités spatiales d'un groupe ne sont pas forcément adjacentes
- Typologies de territoires
- Pas de détection de clusters spatiaux
- Pas d'inférence statistique

Statistiques de scan pour données spatio-fonctionnelles

Statistique de scan

▶ Variable aléatoire utilisée comme statistique de test dont l'objectif est de tester la présence d'un cluster d'évènements au sein d'un espace étudié $\mathcal{R}$

Hypothèse nulle $\mathcal{H}_0$: répartition homogène des évènements dans $\mathcal{R}$

Hypothèse alternative $\mathcal{H}_1$: présence d'un cluster $Z \in \mathcal{R}$ dans lequel la probabilité d'apparition de l'évènement est différente de celle dans le reste de $\mathcal{R}$

Statistique de scan

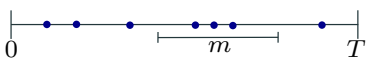
▶ Variable aléatoire utilisée comme statistique de test dont l'objectif est de tester la présence d'un cluster d'évènements au sein d'un espace étudié $\mathcal{R}$

Hypothèse nulle $\mathcal{H}_0$: répartition homogène des évènements dans $\mathcal{R}$

Hypothèse alternative $\mathcal{H}_1$: présence d'un cluster $Z \in \mathcal{R}$ dans lequel la probabilité d'apparition de l'évènement est différente de celle dans le reste de $\mathcal{R}$

Statistiques de scan unidimensionnelles

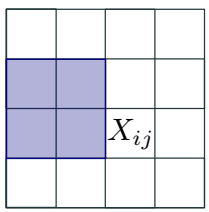
Discrètes



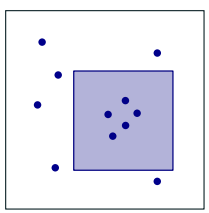
Continues

Statistiques de scan bidimensionnelles

Discrètes



Continues



Statistiques de scan spatiales

▶ Proposées par Kulldorff (1997)
▶ Nombreux développements
▶ Très utilisées en épidémiologie (2639 ref PubMed)

Statistiques de scan spatiales : principe

Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

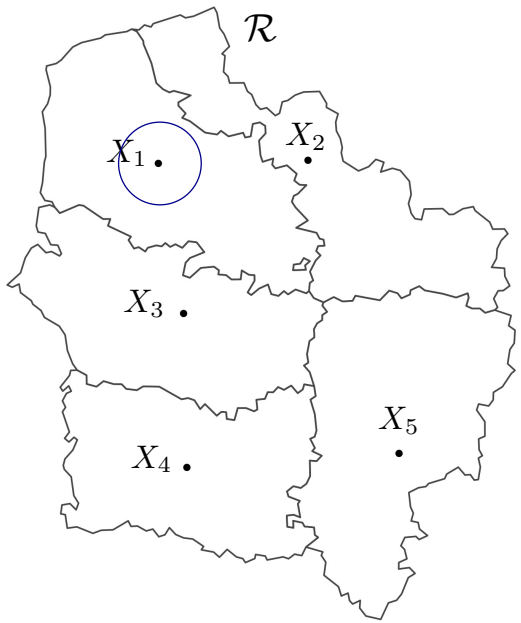
Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

Fenêtre de scan circulaire Z

- Centrée en chaque unité spatiale
- Rayon variable
- Maximum tel que $\mu(Z) \leq \mu(\mathcal{R})/2$

↓

- Cluster potentiel Z



X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

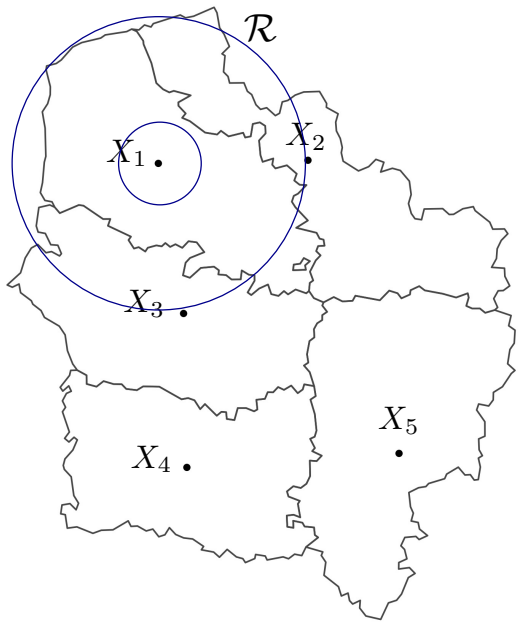
Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

Fenêtre de scan circulaire Z

- Centrée en chaque unité spatiale
- Rayon variable
- Maximum tel que $\mu(Z) \leq \mu(\mathcal{R})/2$

↓

- Cluster potentiel Z



X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

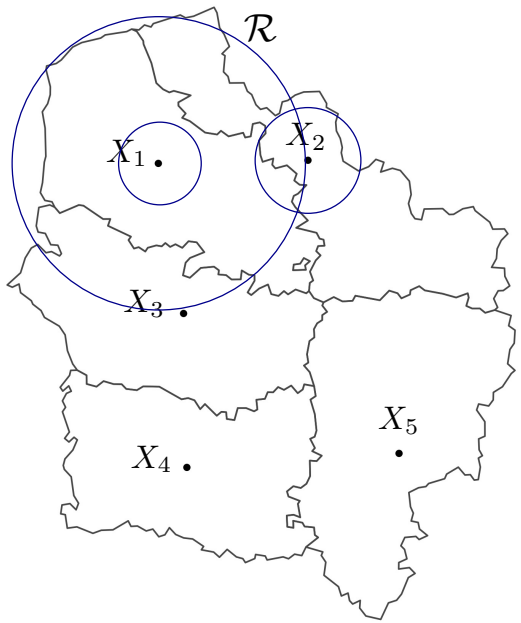
Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

Fenêtre de scan circulaire Z

- Centrée en chaque unité spatiale
- Rayon variable
- Maximum tel que $\mu(Z) \leq \mu(\mathcal{R})/2$

↓

- Cluster potentiel Z



X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

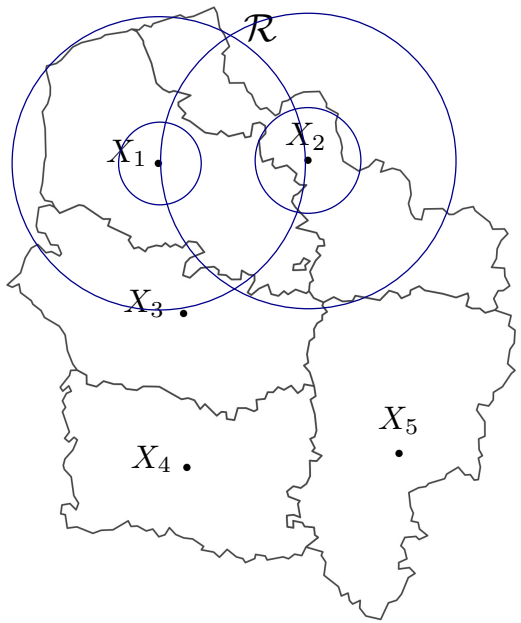
Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

Fenêtre de scan circulaire Z

- Centrée en chaque unité spatiale
- Rayon variable
- Maximum tel que $\mu(Z) \leq \mu(\mathcal{R})/2$

↓

- Cluster potentiel Z



X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

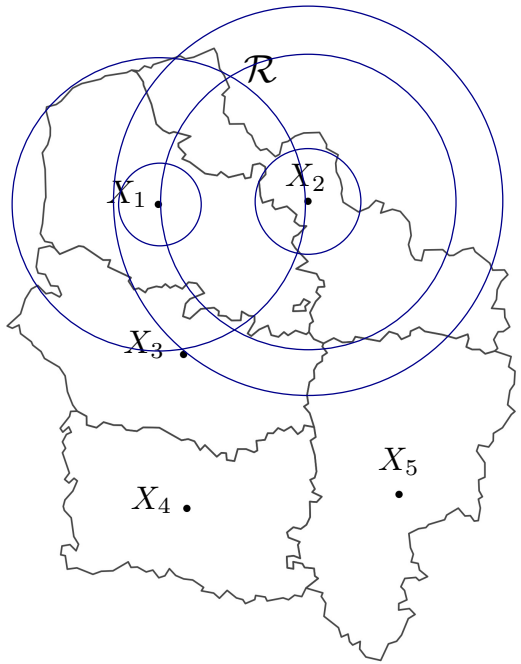
Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

Fenêtre de scan circulaire Z

- Centrée en chaque unité spatiale
- Rayon variable
- Maximum tel que $\mu(Z) \leq \mu(\mathcal{R})/2$

↓

- Cluster potentiel Z



X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

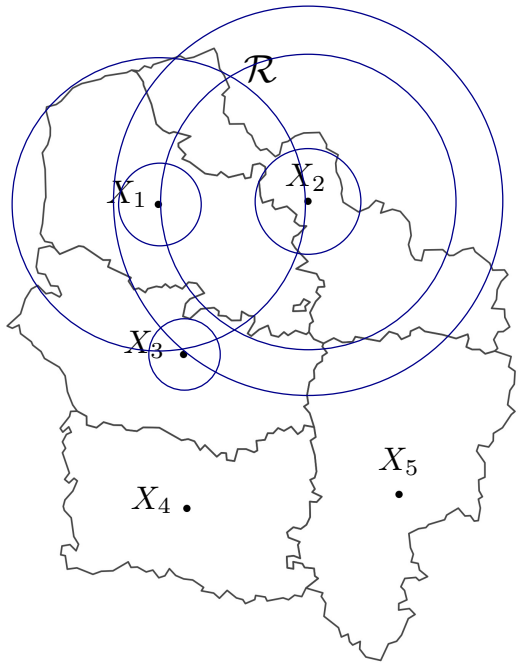
Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

Fenêtre de scan circulaire Z

- Centrée en chaque unité spatiale
- Rayon variable
- Maximum tel que $\mu(Z) \leq \mu(\mathcal{R})/2$

↓

- Cluster potentiel Z



X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

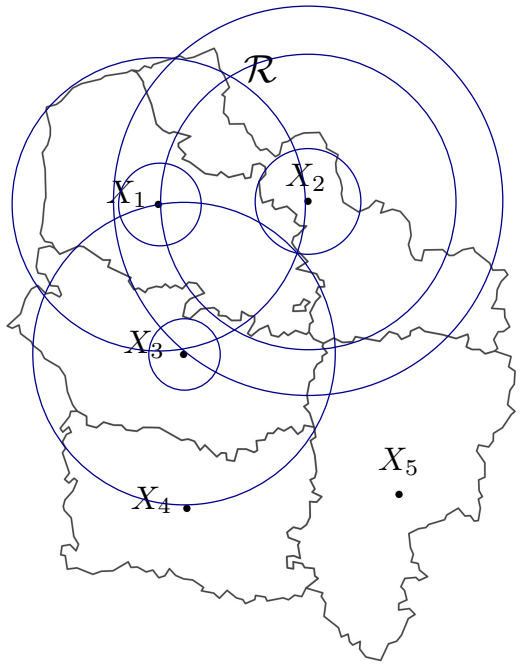
Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

Fenêtre de scan circulaire Z

- Centrée en chaque unité spatiale
- Rayon variable
- Maximum tel que $\mu(Z) \leq \mu(\mathcal{R})/2$

↓

- Cluster potentiel Z



X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

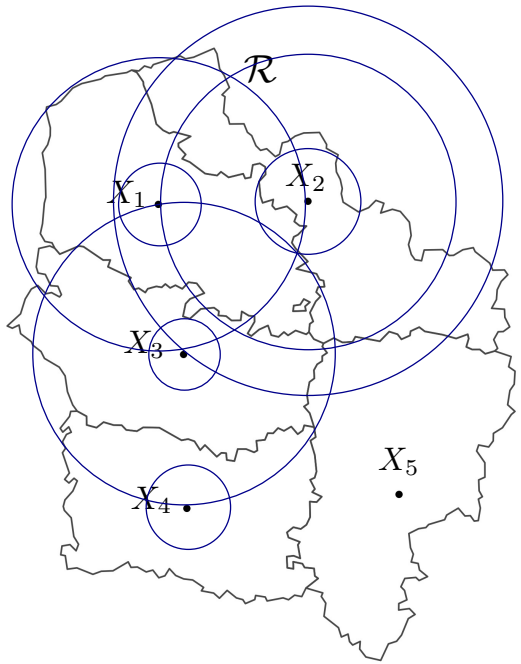
Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

Fenêtre de scan circulaire Z

- Centrée en chaque unité spatiale
- Rayon variable
- Maximum tel que $\mu(Z) \leq \mu(\mathcal{R})/2$

↓

- Cluster potentiel Z



X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

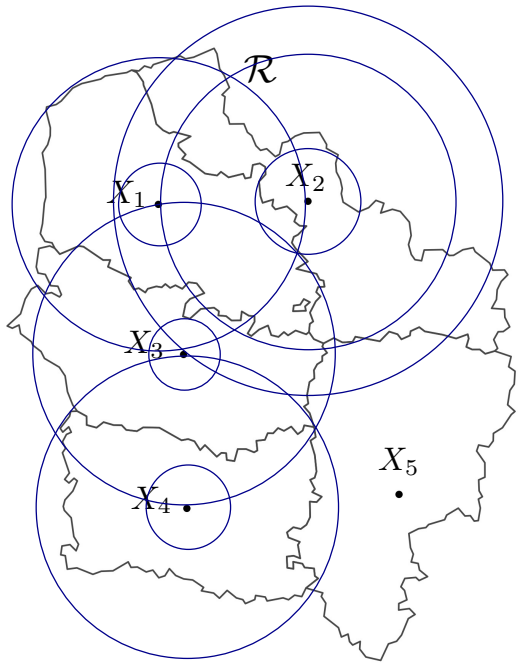
Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

Fenêtre de scan circulaire Z

- Centrée en chaque unité spatiale
- Rayon variable
- Maximum tel que $\mu(Z) \leq \mu(\mathcal{R})/2$

↓

- Cluster potentiel Z



X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

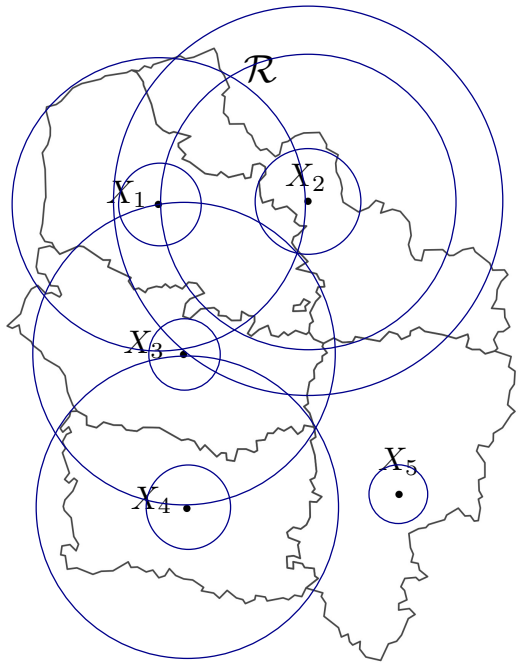
Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

Fenêtre de scan circulaire Z

- Centrée en chaque unité spatiale
- Rayon variable
- Maximum tel que $\mu(Z) \leq \mu(\mathcal{R})/2$

↓

- Cluster potentiel Z



X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

Fenêtre de scan circulaire Z

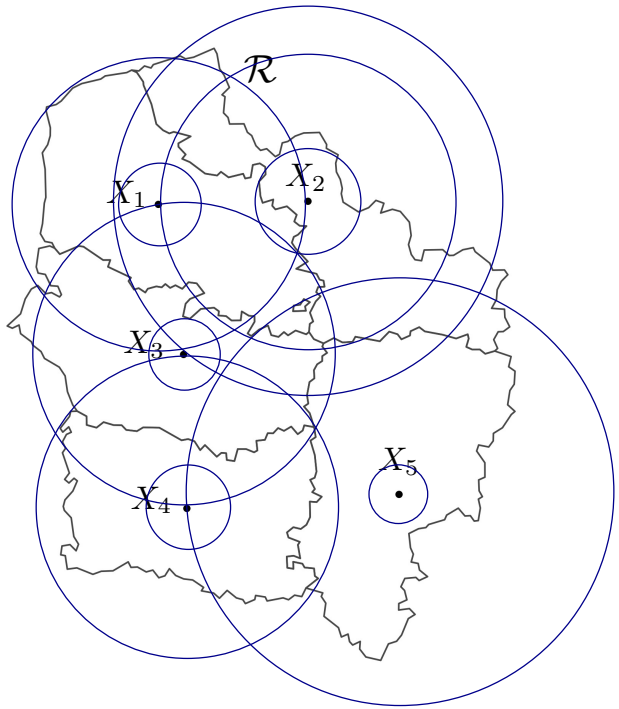
- Centrée en chaque unité spatiale
- Rayon variable
- Maximum tel que $\mu(Z) \leq \mu(\mathcal{R})/2$

↓

- Cluster potentiel Z

↓

- $\mathcal{Z}$: collection de clusters potentiels



X_i : mesure au sein de l'unité spatiale i
- Nb de cas de maladie
- Concentration en pollution

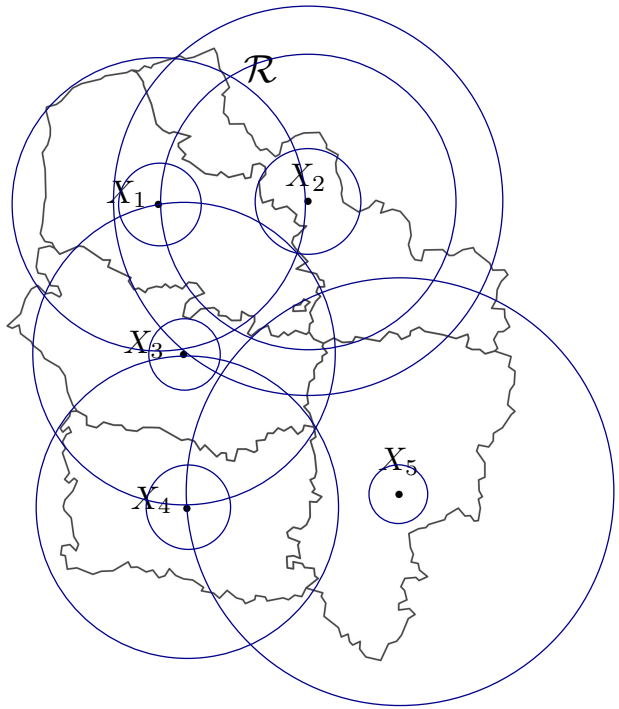
$\mu(X_i)$: mesure d'intensité de l'unité spatiale i
- Population à risque
- Superficie

Statistiques de scan spatiales : principe

Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

A chaque $Z \in \mathcal{Z}$:

- Indice de concentration $I(Z)$



X_i : mesure au sein de l'unité spatiale i

- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i

- Population à risque
- Superficie

Statistiques de scan spatiales : principe

Processus de scan sur $\mathcal{R}$: détection du cluster le plus probable (Most likely cluster (MLC))

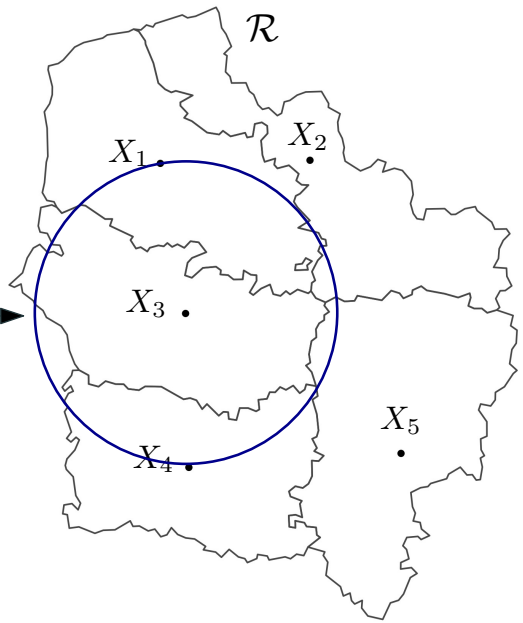
A chaque $Z \in \mathcal{Z}$:

- Indice de concentration $I(Z)$

↓

Most likely cluster $\tilde{Z}$:

$$\tilde{Z} = \underset{z \in \mathcal{Z}}{\operatorname{argmax}}\, \mathsf{I}(Z)$$



X_i : mesure au sein de l'unité spatiale i

- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i

- Population à risque
- Superficie

Statistiques de scan spatiales : principe

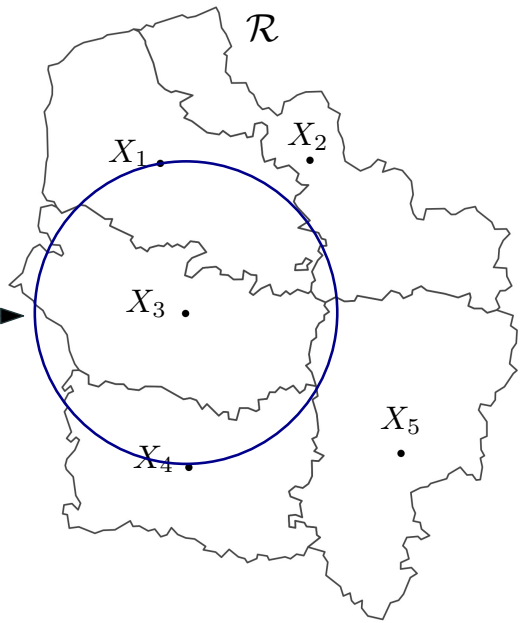
Définition de la statistique de scan spatiale et inférence statistique

A chaque $Z \in \mathcal{Z}$:

- Indice de concentration $I(Z)$

Most likely cluster $\tilde{Z}$:

$$\tilde{Z} = \underset{z \in \mathcal{Z}}{\operatorname{argmax}}\, \mathsf{I}(Z)$$



Statistique de scan spatiale:

$$\Lambda = \max_{z \in \mathcal{Z}} \mathsf{I}(Z) = I(\tilde{Z})$$

Hypothèses de test

$\mathcal{H}_0$: absence de cluster

$\mathcal{H}_1$: présence d'au moins 1 cluster

$$\mathcal{H}_1 = \{\mathcal{H}_1^{(Z_1)}, \mathcal{H}_1^{(Z_2)}, \dots\}_{Z_i \in \mathcal{Z}}$$

Inférence statistique

- Procédure de Monte-Carlo
  - Générations sous $\mathcal{H}_0$
  - Permutations

X_i : mesure au sein de l'unité spatiale i

- Nb de cas de maladie
- Concentration en pollution

$\mu(X_i)$: mesure d'intensité de l'unité spatiale i

- Population à risque
- Superficie

De nombreux modèles

Données univariées : X_i

- Bernoulli, Poisson (Kulldorff and Nagarwalla, 1995 ; Kulldorff, 1997)
- Gaussien (Kulldorff et al., 2009)
- Zero-inflated (de Lima et al., 2015)
- Non-paramétrique (Jung and Cho, 2015 ; Cucala, 2016)
- ...

De nombreux modèles

Données univariées : X_i

- Bernoulli, Poisson (Kulldorff and Nagarwalla, 1995 ; Kulldorff, 1997)
- Gaussien (Kulldorff et al., 2009)
- Zero-inflated (de Lima et al., 2015)
- Non-paramétrique (Jung and Cho, 2015 ; Cucala, 2016)
- ...

Données multivariées : $(X_i^1, \ldots, X_i^p)$

- Gaussien (Cucala et al., 2017)
- Non-paramétrique (Cucala et al., 2019)
- ...

De nombreux modèles

Données univariées : X_i

- Bernoulli, Poisson (Kulldorff and Nagarwalla, 1995 ; Kulldorff, 1997)
- Gaussien (Kulldorff et al., 2009)
- Zero-inflated (de Lima et al., 2015)
- Non-paramétrique (Jung and Cho, 2015 ; Cucala, 2016)
- ...

Données multivariées : $(X_i^1, \ldots, X_i^p)$

- Gaussien (Cucala et al., 2017)
- Non-paramétrique (Cucala et al., 2019)
- ...

Gap : Aucun modèle pour données fonctionnelles

Données univariées : X_i

- Bernoulli, Poisson (Kulldorff and Nagarwalla, 1995 ; Kulldorff, 1997)
- Gaussien (Kulldorff et al., 2009)
- Zero-inflated (de Lima et al., 2015)
- Non-paramétrique (Jung and Cho, 2015 ; Cucala, 2016)
- ...

Données multivariées : $(X_i^1, \ldots, X_i^p)$

- Gaussien (Cucala et al., 2017)
- Non-paramétrique (Cucala et al., 2019)
- ...

Gap : Aucun modèle pour données fonctionnelles

- Scan univarié : résumer l'information sur l'ensemble des temps de mesures
  - ⚠ Importante perte d'information
- Scan multivarié : autant de variables que de temps de mesures
  - ⚠ problème de haute dimension et de fortes corrélations

Qu'est-ce qu'un cluster spatial de données fonctionnelles ?

Qu'est-ce qu'un cluster spatial de données fonctionnelles ?

Univarié : $X \in \mathbb{R} : \mathbb{E}[X_i \mid s_i \in Z] = \mathbb{E}[X_i \mid s_i \notin Z] + \Delta$ avec $\Delta \neq 0$.

Qu'est-ce qu'un cluster spatial de données fonctionnelles ?

Univarié : $X \in \mathbb{R}$: $\mathbb{E}[X_i \mid s_i \in Z] = \mathbb{E}[X_i \mid s_i \notin Z] + \Delta$ avec $\Delta \neq 0$.

Fonctionnel : Dai and Genton (2019) : magnitude and shape outliers

$$\forall t \in \mathcal{T},\ \mathbb{E}[X_i(t) \mid s_i \in Z] = \mathbb{E}[X_i(t) \mid s_i \notin Z] + \Delta(t)$$

Qu'est-ce qu'un cluster spatial de données fonctionnelles ?

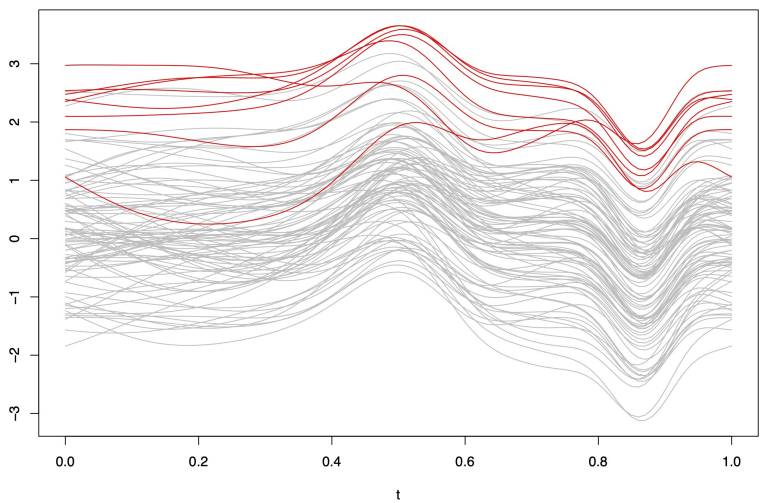
Univarié : $X \in \mathbb{R}$: $\mathbb{E}[X_i \mid s_i \in Z] = \mathbb{E}[X_i \mid s_i \notin Z] + \Delta$ avec $\Delta \neq 0$.

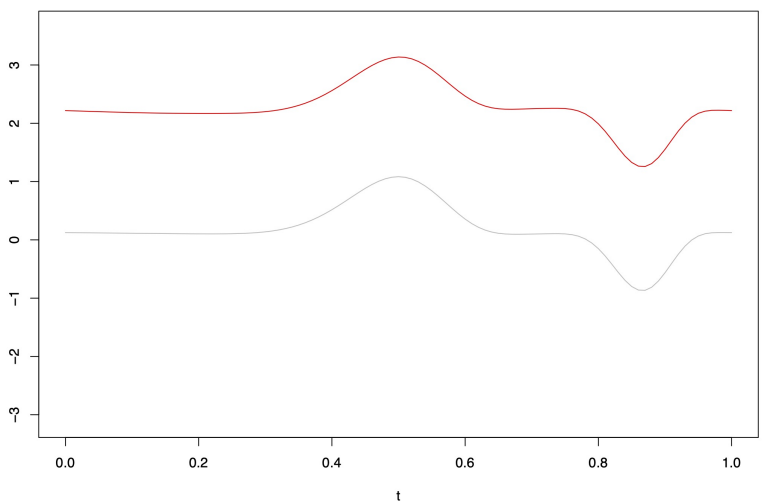
Fonctionnel : Dai and Genton (2019) : magnitude and shape outliers

$$\forall t \in \mathcal{T},\ \mathbb{E}[X_i(t) \mid s_i \in Z] = \mathbb{E}[X_i(t) \mid s_i \notin Z] + \Delta(t)$$

Selon la forme de $\Delta(\cdot)$:

- Cluster en amplitude
- Cluster en forme





Qu'est-ce qu'un cluster spatial de données fonctionnelles ?

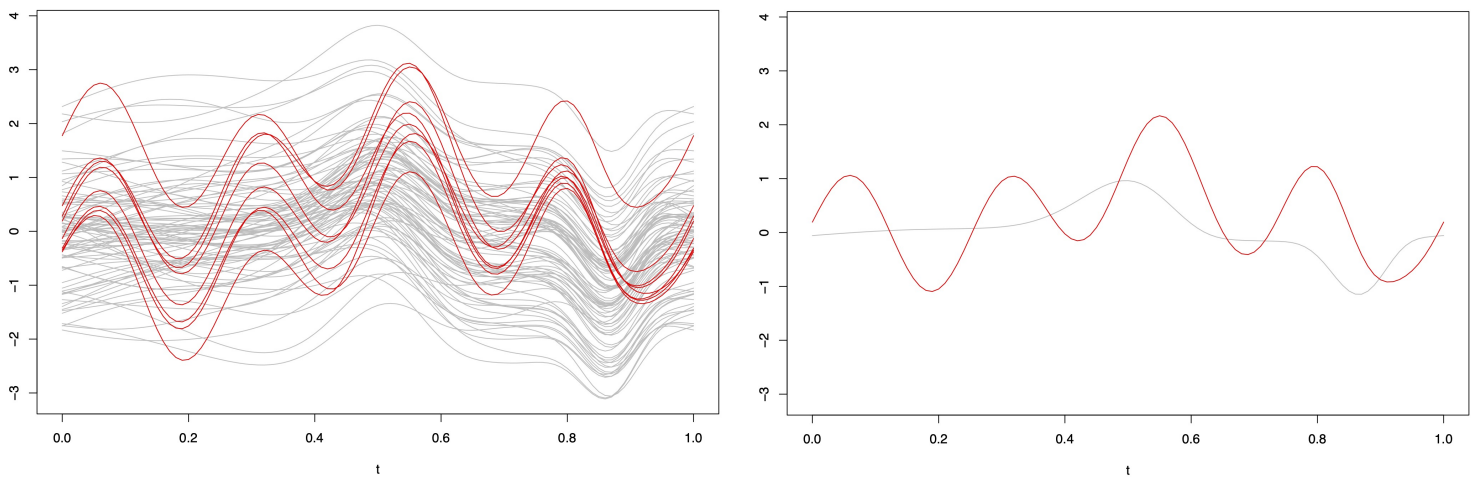
Univarié : $X \in \mathbb{R}$: $\mathbb{E}[X_i \mid s_i \in Z] = \mathbb{E}[X_i \mid s_i \notin Z] + \Delta$ avec $\Delta \neq 0$.

Fonctionnel : Dai and Genton (2019) : magnitude and shape outliers

$$\forall t \in \mathcal{T},\ \mathbb{E}[X_i(t) \mid s_i \in Z] = \mathbb{E}[X_i(t) \mid s_i \notin Z] + \Delta(t)$$

Selon la forme de $\Delta(\cdot)$:

- Cluster en amplitude
- Cluster en forme



Statistiques de scan pour données spatio-fonctionnelles

Statistiques de scan pour données spatio-fonctionnelles

C. Frévent, M.S. Ahmed, M. Marbac, **M. Genin** (2021). Detecting spatial clusters in functional data: new scan statistic approaches. *Spatial Statistics*. (46),100550.

Paramétrique
- Anova fonctionnelle

Non-paramétrique
- Cucala [2014] + Lin [2021]

Statistiques de scan pour données spatio-fonctionnelles

C. Frévent, M.S. Ahmed, M. Marbac, **M. Genin** (2021). Detecting spatial clusters in functional data: new scan statistic approaches. *Spatial Statistics*. (46),100550.

Paramétrique
- Anova fonctionnelle

Non-paramétrique
- Cucala [2014] + Lin [2021]

C. Frévent, M.S. Ahmed, S. Dabo-Niang, **M. Genin**. Investigating spatial scan statistics for multivariate functional data (2023). *Journal of the Royal Statistical Society - Series C* Volume 72, Issue 2, May 2023, Pages 450–475.

Paramétrique
- Manova fonctionnelle

Non-paramétrique
- Extension Frevent et al. 2021
- Multivariate functional Mann-Withney test

Statistiques de scan pour données spatio-fonctionnelles

Univarié

C. Frévent, M.S. Ahmed, M. Marbac, **M. Genin** (2021). Detecting spatial clusters in functional data: new scan statistic approaches. *Spatial Statistics*. (46),100550.

Paramétrique
- Anova fonctionnelle

Non-paramétrique
- Cucala [2014] + Lin [2021]

Multivarié

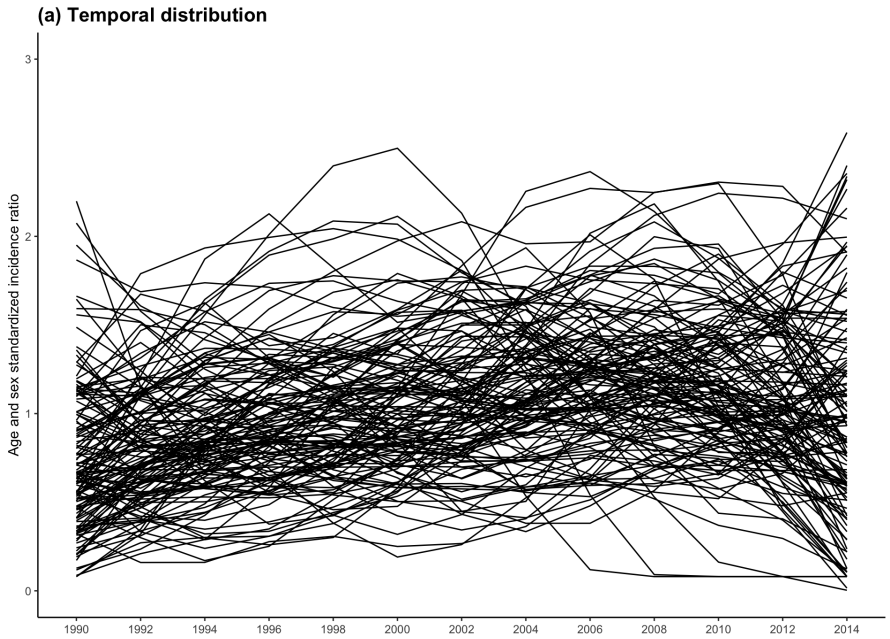
C. Frévent, M.S. Ahmed, S. Dabo-Niang, **M. Genin**. Investigating spatial scan statistics for multivariate functional data (2023). *Journal of the Royal Statistical Society - Series C* Volume 72, Issue 2, May 2023, Pages 450–475.

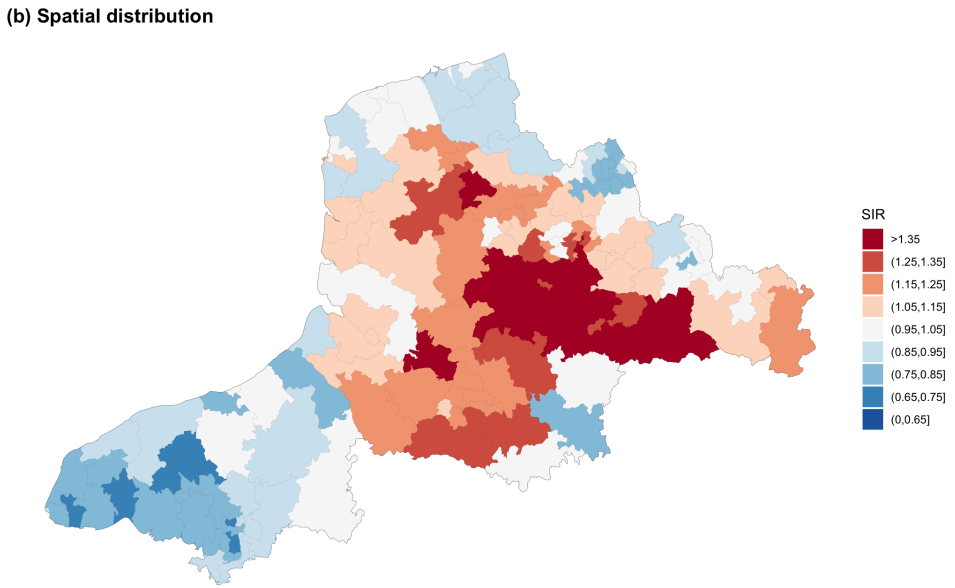
Paramétrique
- Manova fonctionnelle

Non-paramétrique
- Extension Frevent et al. 2021
- Multivariate functional Mann-Whitney test

Ⓡ Package R `HDSpatialScan` CRAN (2021)

C. Frévent, M.S. Ahmed, J. Soula, Z. Smida,L. Cucala, S. Dabo-Niang, **M. Genin** (2022). The R package HDSpatialScan for Multivariate and Functional Spatial Scan Statistics. *R journal*, 14(3),95-120.





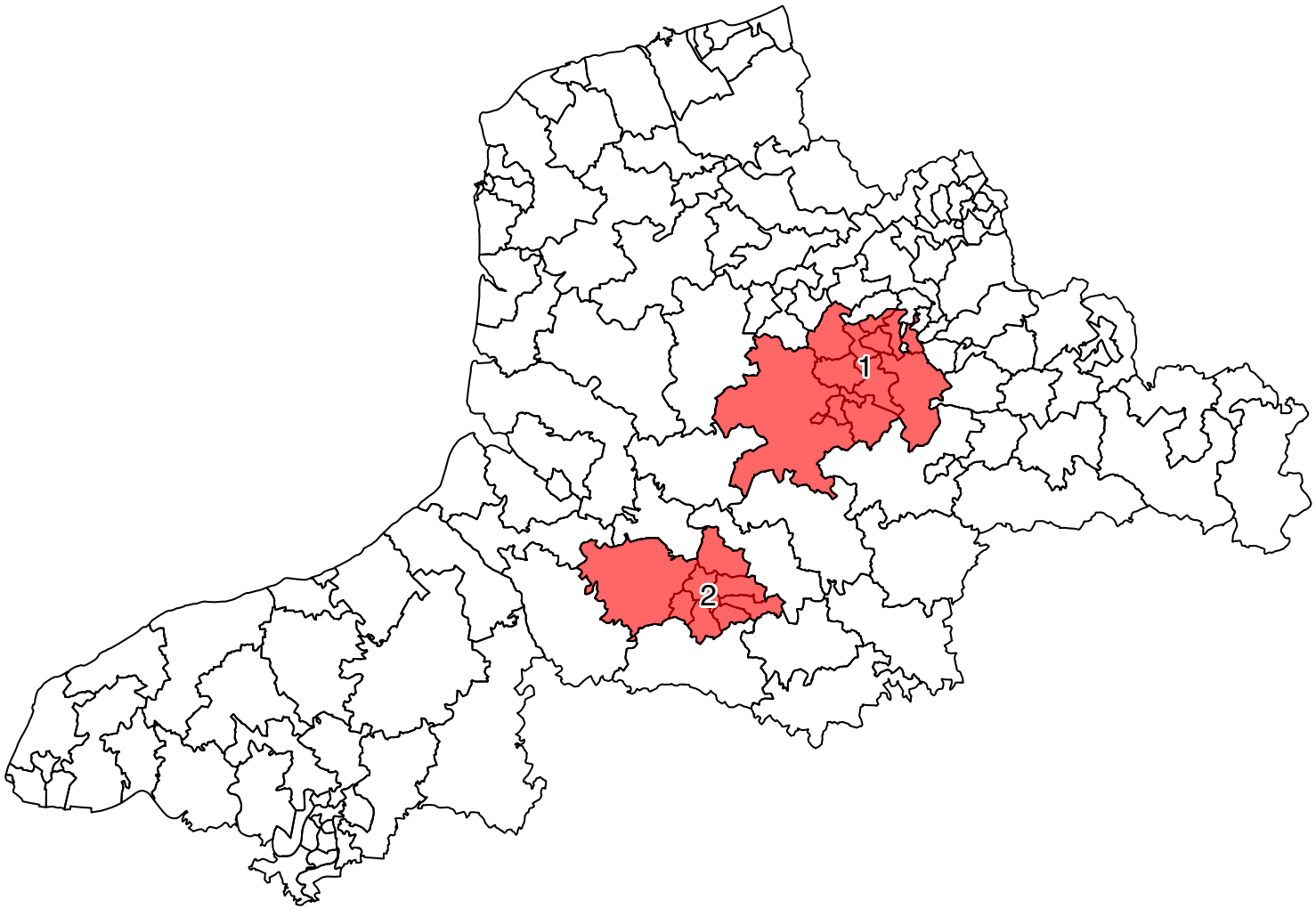
Objectif : détection de clusters spatiaux d'incidence de maladie de Crohn

Méthode :

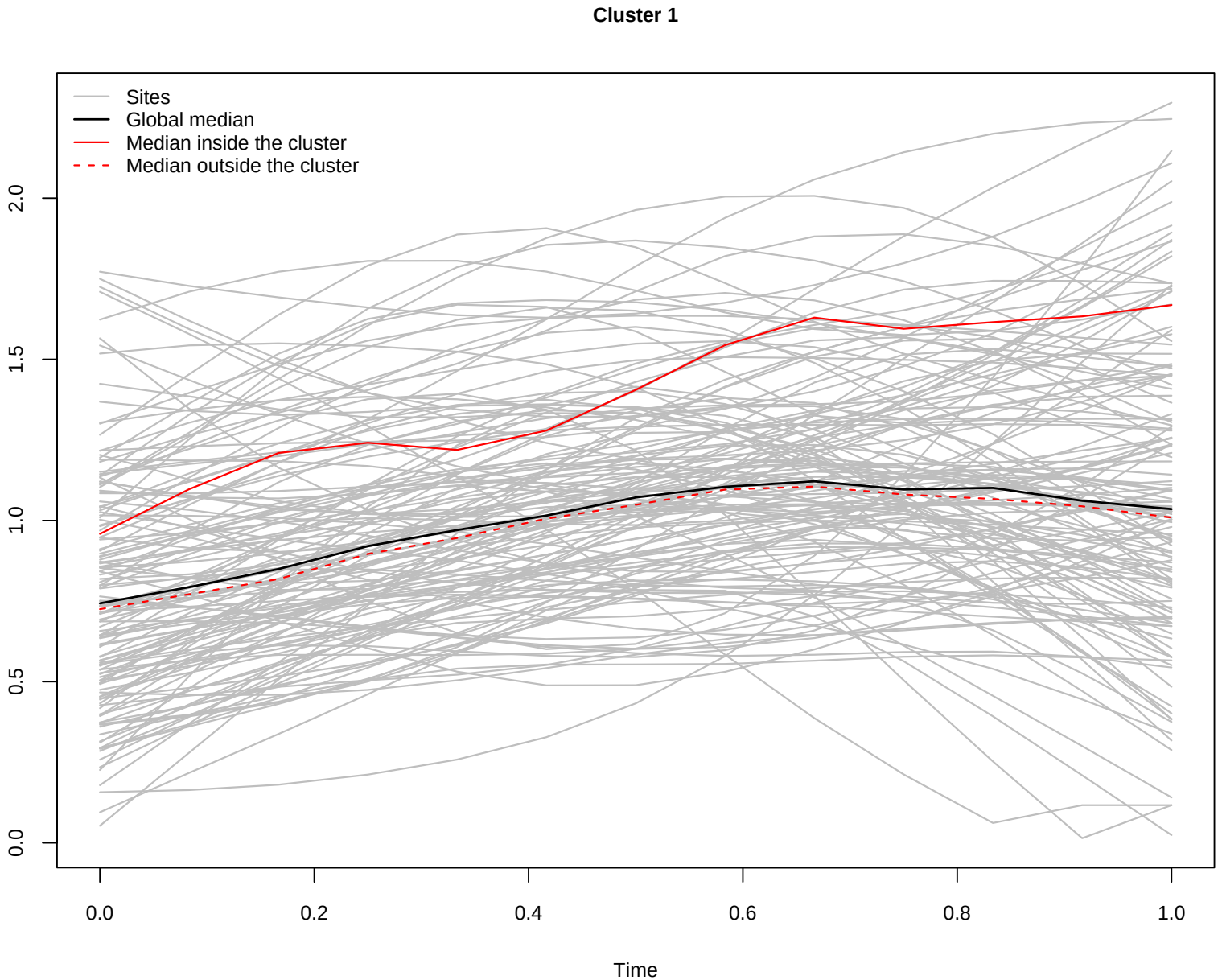
- Lissage par B-splines cubiques
- Scan fonctionnel univarié non-paramétrique

2 clusters spatiaux significatifs :

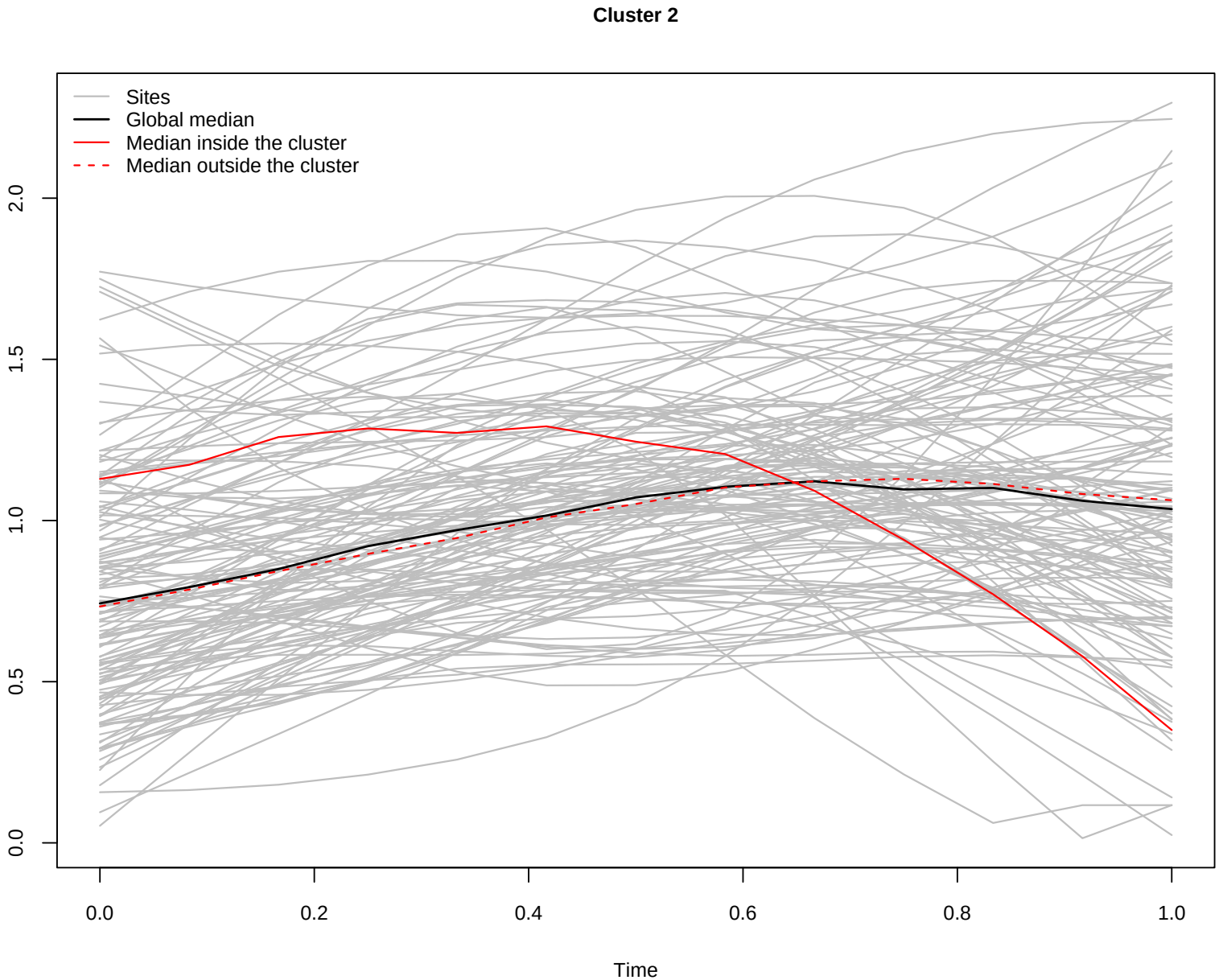
- Cluster 1 : 11 cantons, p=0.005
- Cluster 2 : 8 cantons, p=0.009



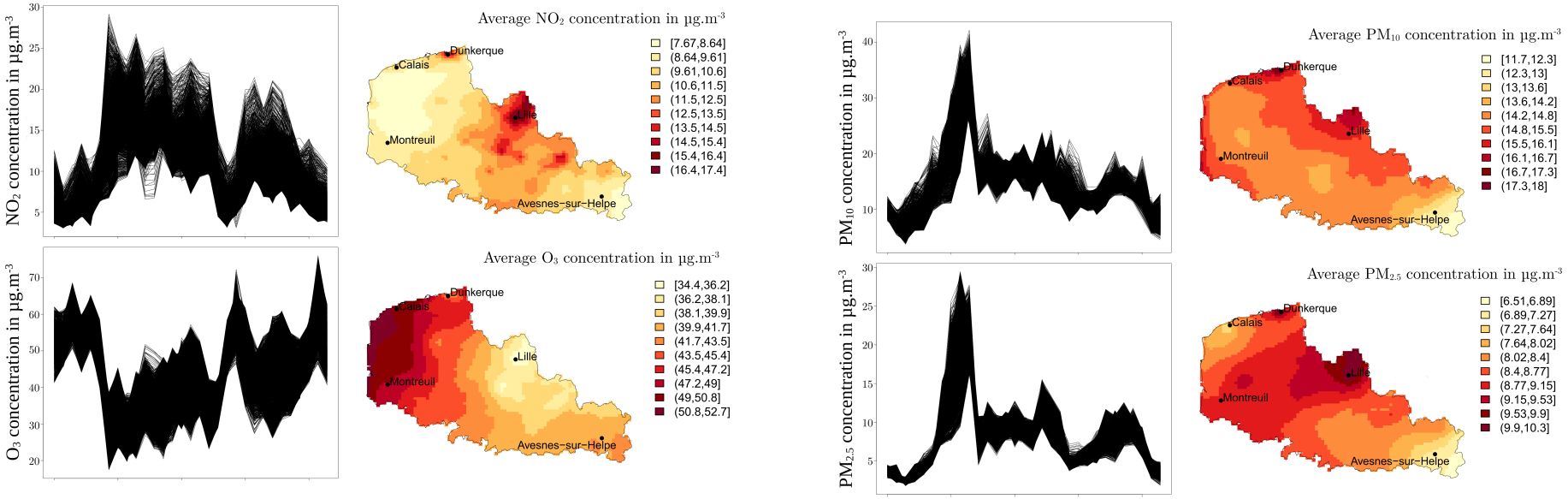
Application dans le cas univarié : Maladie de Crohn



Application dans le cas univarié : Maladie de Crohn



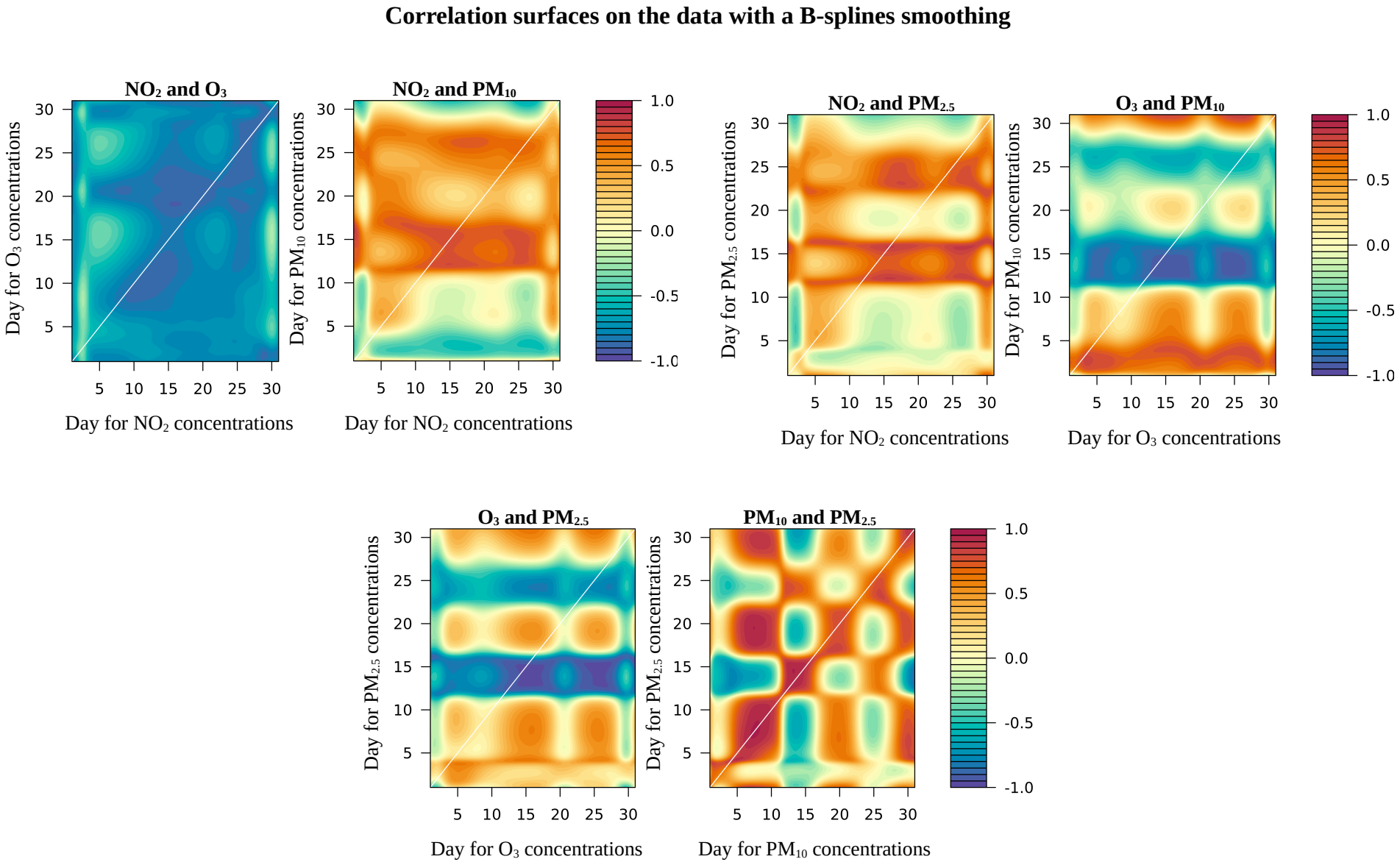
Application dans le cas multivarié : pollution atmosphérique

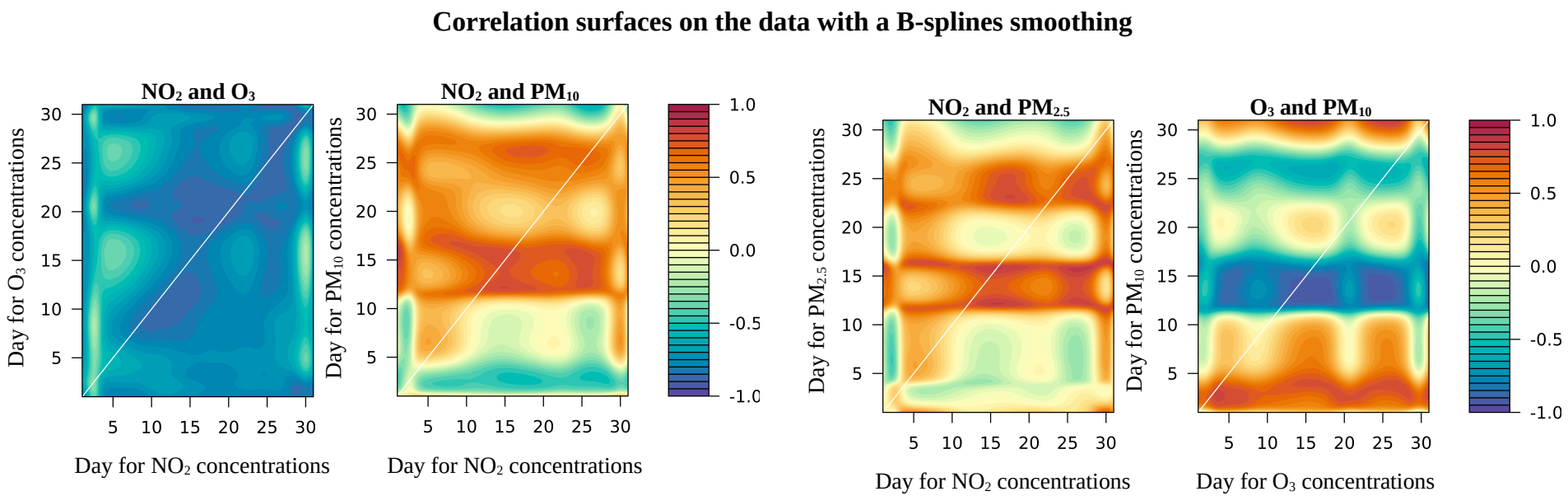


Objectif : détection de points noirs environnementaux

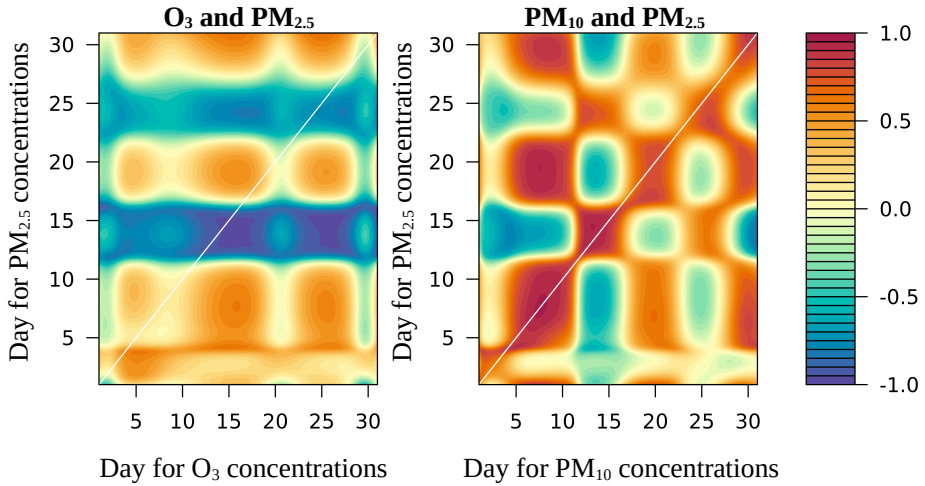
Méthodes :

- Lissage par B-splines cubiques
- Scan fonctionnel univarié sur chacune des composantes ?
  ⇒ Quid des corrélations entre polluants ?

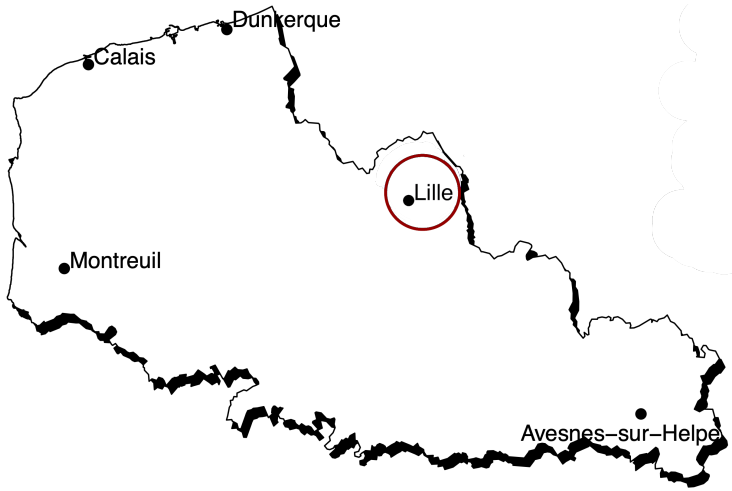

Correlation surfaces on the data with a B-splines smoothing
NO2 and O3
NO2 and PM10
NO2 and PM2.5
O3 and PM10
O3 and PM2.5
PM10 and PM2.5
Day for O3 concentrations
Day for PM10 concentrations
Day for PM2.5 concentrations
Day for NO2 concentrations
Day for O3 concentrations
Day for PM10 concentrations
5 10 15 20 25 30
1.0 0.5 0.0 -0.5 -1.0

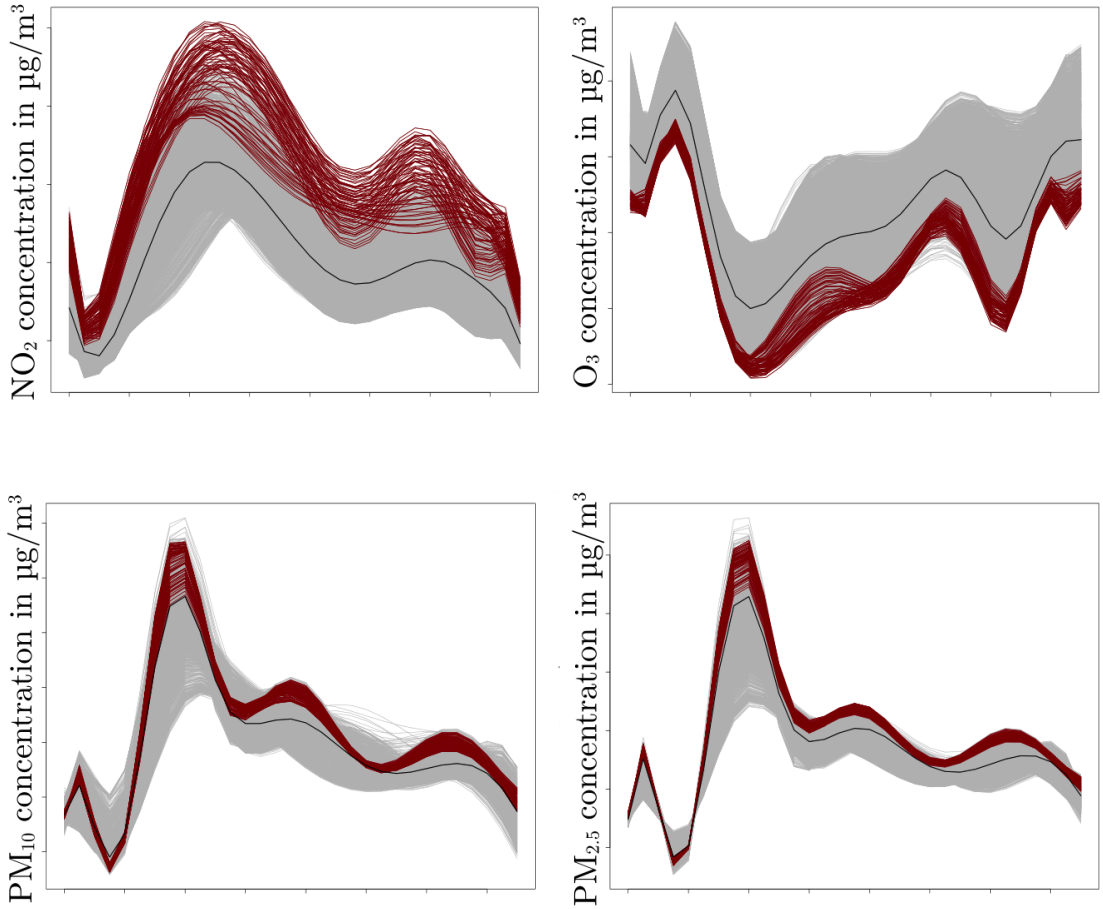


Scan fonctionnel multivarié

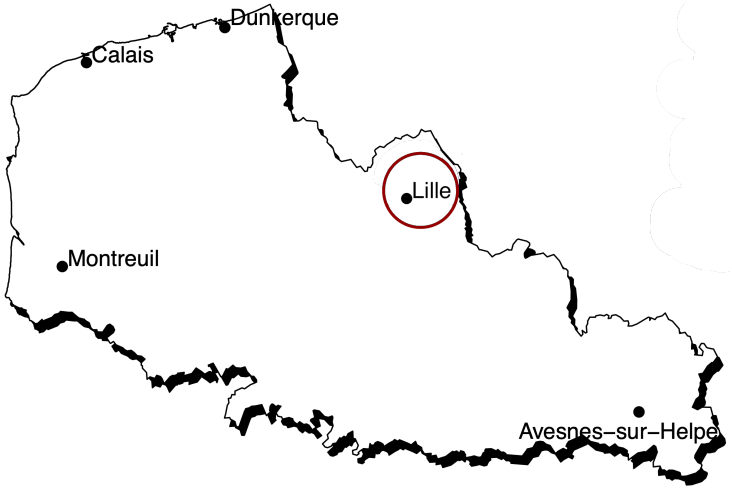


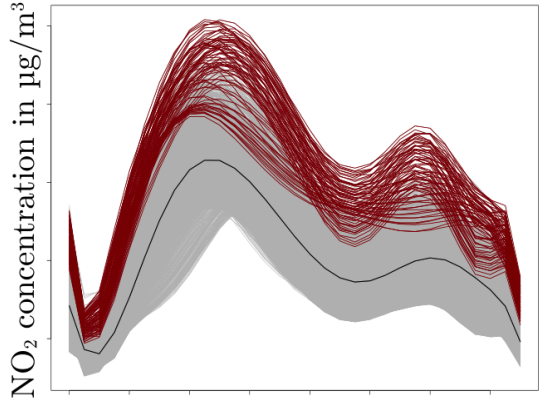
Application dans le cas multivarié : pollution atmosphérique

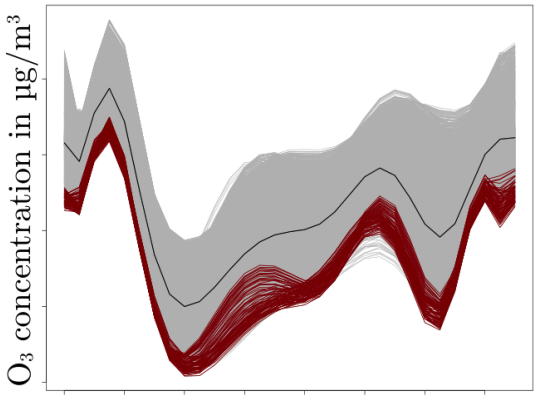


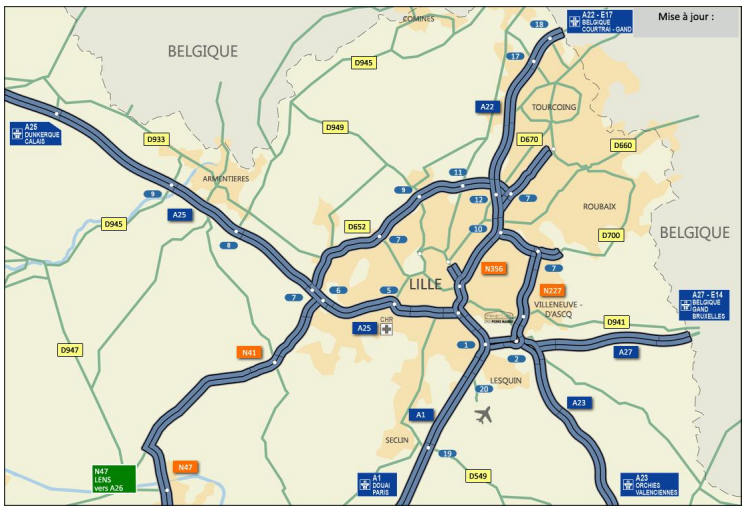


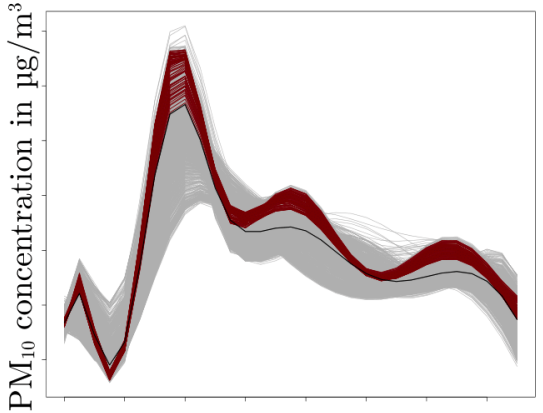
Application dans le cas multivarié : pollution atmosphérique

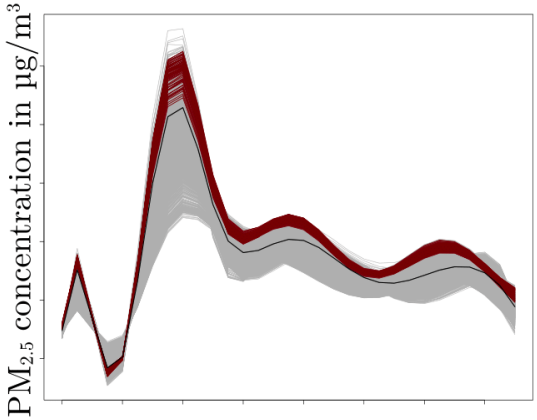












Source : gifex.com

Discussion

Modélisation de données spatio-temporelles par approches fonctionnelles

- Lissage si données observées avec erreurs
- Interpolation si temps de mesures différents selon les sites

Statistiques de scan spatiales pour données fonctionnelles

- Pas de perte d'information / univarié, pas de pb de haute dimension / multivarié
- Range de clusters détectables (en amplitude, en forme...)
- Forme fenêtre adaptable
- Coût computationnel
- Ajustement sur covariables $\rightarrow$ modèles linéaires (généralisés) fonctionnels

$$Y_i(t) = \beta_0(t) + \int_{\mathcal{T}} \beta(s,t)^\top X_i(s)\, \mathrm{d}s + \alpha_w(t)\mathbb{1}_{s_i \in w} + \varepsilon_i(t)$$

- Scan fonctionnel multivarié : co-occurrence de plusieurs pathologies

Merci de votre attention

Functional ANOVA - PFSS

Assumption : $X \in \mathcal{L}^2(\mathcal{T}, \mathbb{R})$. Cuevas et al. (2004) proposed a functional ANOVA.

Step 1 :

$$\mathcal{H}_0 : \mu_w = \mu_{w^c}$$

vs.

$$\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$$

$$F_n^{(w)} = \frac{|w| \ ||\bar{X}_w - \bar{X}||_2^2 + |w^c| \ ||\bar{X}_{w^c} - \bar{X}||_2^2}{\frac{1}{n-2}\left[\sum_{j, s_j \in w} ||X_j - \bar{X}_w||_2^2 + \sum_{j, s_j \in w^c} ||X_j - \bar{X}_{w^c}||_2^2\right]}$$

Functional ANOVA - PFSS

Assumption : $X \in \mathcal{L}^2(\mathcal{T}, \mathbb{R})$. Cuevas et al. (2004) proposed a functional ANOVA.

Step 1 :

$$\mathcal{H}_0 : \mu_w = \mu_{w^c}$$

vs.

$$\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$$

$$F_n^{(w)} = \frac{|w| \ ||\bar{X}_w - \bar{X}||_2^2 + |w^c| \ ||\bar{X}_{w^c} - \bar{X}||_2^2}{\frac{1}{n-2}\left[\sum_{j,s_j \in w} ||X_j - \bar{X}_w||_2^2 + \sum_{j,s_j \in w^c} ||X_j - \bar{X}_{w^c}||_2^2\right]}$$

Step 2 :

$$\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w = \mu_{w^c}$$

vs.

$$\mathcal{H}_1 : \exists w \in \mathcal{W}, \mu_w \neq \mu_{w^c}$$

$F_n^{(w)}$ is considered as a concentration index and maximised over $\mathcal{W}$:

$$\Lambda_{\text{PFSS}} = \max_{w \in \mathcal{W}} F_n^{(w)} \text{ and } \text{MLC} = \arg\max_{w \in \mathcal{W}} F_n^{(w)}$$

Pointwise Student's t-test - DFFSS

Ideas of Cucala et al. (2014) and Lin et al. (2021)
Assumption : $\mathbb{V}[X_i(t)] = \sigma^2(t)$ for $i \in [\![1, n]\!]$

Step 1 : $\mathcal{H}_0 : \mu_w(t) = \mu_{w^c}(t) = \mu_S(t)$

$$I^{(w)}(t) = \frac{|\bar{X}_w(t) - \bar{X}_{w^c}(t)|}{\sqrt{\hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^c}(t)]}}$$

with $\hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^c}(t)] = \widehat{\sigma^2}(t) \left[\frac{1}{|w|} + \frac{1}{|w^c|}\right]$ and

$$\widehat{\sigma^2}(t) = \frac{1}{n-2} \left[\sum_{i, s_i \in w} (X_i(t) - \bar{X}_w(t))^2 + \sum_{i, s_i \in w^c} (X_i(t) - \bar{X}_{w^c}(t))^2\right]$$

Pointwise Student's t-test - DFFSS

Ideas of Cucala et al. (2014) and Lin et al. (2021)
Assumption : $\mathbb{V}[X_i(t)] = \sigma^2(t)$ for $i \in [\![1, n]\!]$

Step 1 : $\mathcal{H}_0 : \mu_w(t) = \mu_{w^c}(t) = \mu_S(t)$

$$I^{(w)}(t) = \frac{|\bar{X}_w(t) - \bar{X}_{w^c}(t)|}{\sqrt{\hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^c}(t)]}}$$

Step 2 : $\mathcal{H}_0 : \mu_w = \mu_{w^c} = \mu_S$
Globalization of the information (Lin et al. 2021) :

$$I^{(w)} = \sup_{t \in \mathcal{T}} I^{(w)}(t)$$

Pointwise Student's t-test - DFFSS

Ideas of Cucala et al. (2014) and Lin et al. (2021)
Assumption : $\mathbb{V}[X_i(t)] = \sigma^2(t)$ for $i \in [\![1, n]\!]$

Step 1 : $\mathcal{H}_0 : \mu_w(t) = \mu_{w^c}(t) = \mu_S(t)$

$$I^{(w)}(t) = \frac{|\bar{X}_w(t) - \bar{X}_{w^c}(t)|}{\sqrt{\hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^c}(t)]}}$$

Step 2 : $\mathcal{H}_0 : \mu_w = \mu_{w^c} = \mu_S$
Globalization of the information (Lin et al. 2021) :

$$I^{(w)} = \sup_{t \in \mathcal{T}} I^{(w)}(t)$$

Step 3 :

$$\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w = \mu_{w^c}$$

vs.

$$\mathcal{H}_1 : \exists w \in \mathcal{W}, \mu_w \neq \mu_{w^c}$$

$$\Lambda_{\text{DFFSS}} = \max_{w \in \mathcal{W}} I^{(w)} \text{ and } \text{MLC} = \arg\max_{w \in \mathcal{W}} I^{(w)}$$

Functional MANOVA - MPFSS

Assumption : $X \in \mathcal{L}^2(\mathcal{T}, \mathbb{R}^p)$

Step 1 : Gorecki et Smaga (2017) proposed a functional MANOVA.

$$\mathcal{H}_0 : \mu_w = \mu_{w^c}$$
$$\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$$

$$\mathrm{LH}^{(w)} = \mathrm{Trace}(H_w E_w^{-1})$$

where

$$H_w = |w| \int_{\mathcal{T}} [\bar{X}_w(t) - \bar{X}(t)][\bar{X}_w(t) - \bar{X}(t)]^\top \, \mathrm{d}t + |w^c| \int_{\mathcal{T}} [\bar{X}_{w^c}(t) - \bar{X}(t)][\bar{X}_{w^c}(t) - \bar{X}(t)]^\top \, \mathrm{d}t$$

and

$$E_w = \sum_{j, s_j \in w} \int_{\mathcal{T}} [X_j(t) - \bar{X}_w(t)][X_j(t) - \bar{X}_w(t)]^\top \mathrm{d}t + \sum_{j, s_j \in w^c} \int_{\mathcal{T}} [X_j(t) - \bar{X}_{w^c}(t)][X_j(t) - \bar{X}_{w^c}(t)]^\top \mathrm{d}t$$

Step 2 :

$$\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w = \mu_{w^c}$$
$$\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$$

$$\Lambda_{\mathrm{MPFSS}} = \max_{w \in \mathcal{W}} \mathrm{LH}^{(w)}$$

Pointwise Hotelling's t-squared test statistic - MDFFSS

Assumption : $\mathbb{V}[X_i(t)] = \Sigma(t,t),\ \forall i \in [\![1, n]\!]$

Step 1 Qui et al. 2021 :

$$\mathcal{H}_0 : \mu_w = \mu_{w^c}$$
$$\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$$

$$T^{(w)} = \sup_{t \in \mathcal{T}} T^{(w)}(t)$$

where

$$T^{(w)}(t) = \frac{|w||w^c|}{n}(\bar{X}_w(t) - \bar{X}_{w^c}(t))^\top \hat{\Sigma}(t,t)^{-1}(\bar{X}_w(t) - \bar{X}_{w^c}(t)),$$

and $\hat{\Sigma}(s,t) =$

$$\frac{1}{n-2}\left[\sum_{i,s_i \in w}(X_i(s) - \bar{X}_w(s))(X_i(t) - \bar{X}_w(t))^\top + \sum_{i,s_i \in w^c}(X_i(s) - \bar{X}_{w^c}(s))(X_i(t) - \bar{X}_{w^c}(t))^\top\right].$$

Step 2 :

$$\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w = \mu_{w^c}$$
$$\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$$

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

Pointwise Wilcoxon-Mann-Whitney test statistic - MRBFSS

Step 1 : $\mathcal{H}_0$: identical distributions in w and w^c

$$W^{(w)} = \sup_{t \in \mathcal{T}} W^{(w)}(t)$$

where $W^{(w)}(t)$ is the Wilcoxon-Mann-Whitney test statistic for multivariate data (Oja and Randles (2004)) adapted to the pointwise case

Step 2 : $\mathcal{H}_0 : \forall w \in \mathcal{W}$, the distributions are identical in w and w^c

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

Pointwise Wilcoxon-Mann-Whitney test statistic - MRBFSS

Step 1 : $\mathcal{H}_0$: identical distributions in w and w^c

$$W^{(w)} = \sup_{t \in \mathcal{T}} W^{(w)}(t)$$

where $W^{(w)}(t)$ is the Wilcoxon-Mann-Whitney test statistic for multivariate data (Oja and Randles (2004)) adapted to the pointwise case

Step 2 : $\mathcal{H}_0 : \forall w \in \mathcal{W}$, the distributions are identical in w and w^c

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

The **statistical significance** of the MLC is estimated as previously.