Détection d'agrégats pour données ponctuelles

Lionel Cucala

Mardi 27 Novembre 2007

Plan de l'exposé

Plan de l'exposé

Les processus ponctuels

Les processus ponctuels

- Jeux de données : semis de points consistant en la localisation d'événements de même nature : $\{s_1, \cdots, s_n\}, s_i \in D \subset \mathbb{R}^d$.

Les processus ponctuels

- Jeux de données : semis de points consistant en la localisation d'événements de même nature : $\{s_1, \cdots, s_n\}, s_i \in D \subset \mathbb{R}^d$.

- Modélisation par un processus aléatoire : le nombre N et les positions $\{S_1, \cdots, S_N\}$ des événements sont aléatoires.

Les processus ponctuels

- Jeux de données : semis de points consistant en la localisation d'événements de même nature : $\{s_1, \cdots, s_n\}, s_i \in D \subset \mathbb{R}^d$.

- Modélisation par un processus aléatoire : le nombre N et les positions $\{S_1, \cdots, S_N\}$ des événements sont aléatoires.

- Intensité (de premier ordre) :

$$\lambda(s) = \lim_{\nu(ds) \to 0} \frac{\mathbb{E}N(ds)}{\nu(ds)}$$

où ds : volume infinitésimal centré en s.

La détection d'agrégats

La détection d'agrégats

- Définition :
  Agrégat= zone où la concentration en évènements est anormalement élevée, non due au hasard.

La détection d'agrégats

- Définition :
  Agrégat= zone où la concentration en évènements est anormalement élevée, non due au hasard.

- Question :
  Y a-t-il un ou plusieurs agrégats ? Si oui, où ?

La détection d'agrégats

- Définition :
  Agrégat= zone où la concentration en évènements est anormalement élevée, non due au hasard.

- Question :
  Y a-t-il un ou plusieurs agrégats ? Si oui, où ?

- Attention !
  Adaptation à la densité de population $f(s)$.

L'hypothèse nulle

L'hypothèse nulle

- On cherche à réfuter : $(s_1, \cdots, s_N)$ issu d'un processus de Poisson inhomogène d'intensité $\lambda(s) = af(s), a > 0$.

L'hypothèse nulle

- On cherche à réfuter : $(s_1, \cdots, s_N)$ issu d'un processus de Poisson inhomogène d'intensité $\lambda(s) = af(s), a > 0$.

- Sachant $N = n$, on cherchera à réfuter
  $H_0 : (s_1, \cdots, s_n)$ i.i.d. de densité $f(s)$.

L'hypothèse nulle

- On cherche à réfuter : $(s_1, \cdots, s_N)$ issu d'un processus de Poisson inhomogène d'intensité $\lambda(s) = af(s), a > 0$.

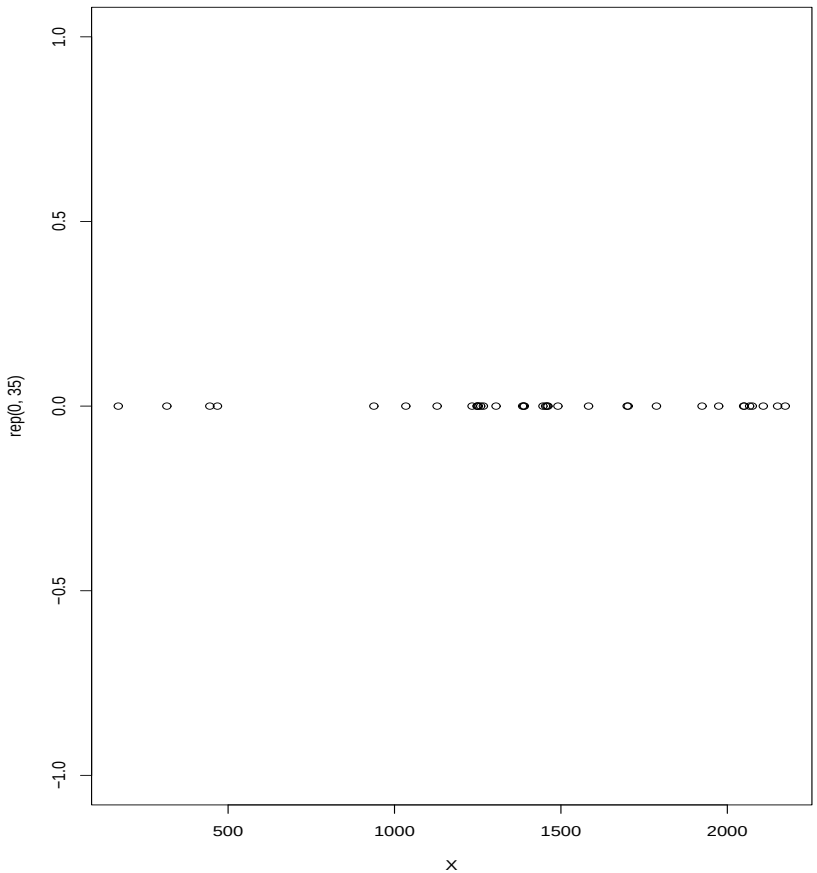
- Sachant $N = n$, on cherchera à réfuter
  $H_0 : (s_1, \cdots, s_n)$ i.i.d. de densité $f(s)$.

- Objectifs : détection du (ou des) agrégat(s) **et** test de significativité par rapport à H_0.

Plan de l'exposé

Plan de l'exposé

Processus ponctuels temporels



Détection d'agrégats temporels

Détection d'agrégats temporels

- Première étape :
  adaptation à la densité de population $f(t)$.

Détection d'agrégats temporels

- Première étape :
  adaptation à la densité de population $f(t)$.
- Evénements :

$$\left\{T_1, \cdots, T_n\right\} \text{ sur } [0, T] \rightarrow \left\{X_1, \cdots, X_n\right\} \text{ sur } [0, 1]$$

où $X_i = \frac{\int_0^{T_i} f(t)\nu(dt)}{\int_0^{T} f(t)\nu(dt)}$.

Détection d'agrégats temporels

- Première étape :
  adaptation à la densité de population $f(t)$.
- Evénements :

$$\left\{T_1, \cdots, T_n\right\} \text{ sur } [0, T] \rightarrow \left\{X_1, \cdots, X_n\right\} \text{ sur } [0, 1]$$

  où $X_i = \frac{\int_0^{T_i} f(t)\nu(dt)}{\int_0^{T} f(t)\nu(dt)}$.
- $H_0 : \left\{X_1, \cdots, X_n\right\}$ i.i.d. $\sim U([0, 1])$.

La statistique de balayage

La statistique de balayage

- $\Lambda(d, n) = \max_{a \in [0,1-d]} \sum_{i=1}^{n} \mathbb{1}(X_i \in [a, a+d])$,

  $a^* = \operatorname{argmax}_{a \in [0,1-d]} \sum_{i=1}^{n} \mathbb{1}(X_i \in [a, a+d])$.

La statistique de balayage

- $\Lambda(d, n) = \max_{a \in [0, 1-d]} \sum_{i=1}^{n} \mathbb{1}(X_i \in [a, a+d])$,

  $a^* = \operatorname{argmax}_{a \in [0, 1-d]} \sum_{i=1}^{n} \mathbb{1}(X_i \in [a, a+d])$.

- Distribution de $\Lambda(d, n)$ sous H_0 exprimée et tabulée (Huntington & Naus, 1975).

La statistique de balayage

- $\Lambda(d, n) = \max_{a \in [0, 1-d]} \sum_{i=1}^{n} \mathbb{1}(X_i \in [a, a+d])$,

  $a^* = \operatorname{argmax}_{a \in [0, 1-d]} \sum_{i=1}^{n} \mathbb{1}(X_i \in [a, a+d])$.

- Distribution de $\Lambda(d, n)$ sous H_0 exprimée et tabulée (Huntington & Naus, 1975).

- Agrégat $[a^*, a^* + d]$ jugé significatif si $P_0\big(\Lambda(d, n) > \sum_{i=1}^{n} \mathbb{1}(x_i \in [a^*, a^* + d])\big) < \alpha$.

La statistique de balayage

- $\Lambda(d, n) = \max_{a \in [0, 1-d]} \sum_{i=1}^{n} \mathbb{1}(X_i \in [a, a+d])$,

  $a^* = \operatorname{argmax}_{a \in [0, 1-d]} \sum_{i=1}^{n} \mathbb{1}(X_i \in [a, a+d])$.

- Distribution de $\Lambda(d, n)$ sous H_0 exprimée et tabulée (Huntington & Naus, 1975).

- Agrégat $[a^*, a^* + d]$ jugé significatif si $P_0\big(\Lambda(d, n) > \sum_{i=1}^{n} \mathbb{1}(x_i \in [a^*, a^* + d])\big) < \alpha$.

- Problème : Longueur de l'agrégat d fixée à priori.

Introduction d'une fenêtre variable

Introduction d'une fenêtre variable

Comment comparer 2 intervalles de longueurs différentes ?

Introduction d'une fenêtre variable

Comment comparer 2 intervalles de longueurs différentes ?

Indice de concentration $I(m, d)$ pour un intervalle de longueur d contenant m événements.

Introduction d'une fenêtre variable

Comment comparer 2 intervalles de longueurs différentes ?

Indice de concentration $I(m, d)$ pour un intervalle de longueur d contenant m événements.

Exemple : $I(m, d) = m/d$ (# d'événements / unité de longueur)
$\Rightarrow$ Sélection du plus petit intervalle contenant 2 événements.

La statistique de balayage (fenêtre variable)

La statistique de balayage (fenêtre variable)

Hypothèse alternative H_1 : $\left\{X_1, \cdots, X_n\right\}$ i.i.d. $\sim g(.)$

où $g(t) = \begin{cases} \frac{m}{nd} & \text{si } t \in [a, a+d], \\ \frac{1-m/n}{1-d} & \text{si } t \in [0, a] \cup [a+d, 1]. \end{cases}$

La statistique de balayage (fenêtre variable)

Hypothèse alternative H_1 : $\left\{X_1, \cdots, X_n\right\}$ i.i.d. $\sim g(.)$

où $g(t) = \begin{cases} \frac{m}{nd} & \text{si } t \in [a, a+d], \\ \frac{1-m/n}{1-d} & \text{si } t \in [0, a] \cup [a+d, 1]. \end{cases}$

Test de H_0 contre H_1 :

La statistique de balayage (fenêtre variable)

Hypothèse alternative H_1 : $\left\{X_1, \cdots, X_n\right\}$ i.i.d. $\sim g(.)$

où $g(t) = \begin{cases} \frac{m}{nd} & \text{si } t \in [a, a+d], \\ \frac{1-m/n}{1-d} & \text{si } t \in [0,a] \cup [a+d,1]. \end{cases}$

Test de H_0 contre H_1 :

$$\begin{aligned} \frac{L_1(X_1, \cdots, X_n)}{L_0(X_1, \cdots, X_n)} &= \left(\frac{m}{nd}\right)^m \left(\frac{1-m/n}{1-d}\right)^{n-m} \\ &= I_{scan}(m,d) \end{aligned}$$

La statistique de balayage (fenêtre variable)

Hypothèse alternative H_1 : $\left\{X_1, \cdots, X_n\right\}$ i.i.d. $\sim g(.)$

où $g(t) = \begin{cases} \frac{m}{nd} & \text{si } t \in [a, a+d], \\ \frac{1-m/n}{1-d} & \text{si } t \in [0,a] \cup [a+d, 1]. \end{cases}$

Test de H_0 contre H_1 :

$$\begin{aligned} \frac{L_1(X_1, \cdots, X_n)}{L_0(X_1, \cdots, X_n)} &= \left(\frac{m}{nd}\right)^m \left(\frac{1-m/n}{1-d}\right)^{n-m} \\ &= I_{scan}(m,d) \end{aligned}$$

$\Lambda_{scan} = \sup_{m \geq n_0} I_{scan}(m,d)$.

La statistique de balayage (fenêtre variable)

Hypothèse alternative H_1 : $\left\{X_1, \cdots, X_n\right\}$ i.i.d. $\sim g(.)$

où $g(t) = \begin{cases} \frac{m}{nd} & \text{si } t \in [a, a+d], \\ \frac{1-m/n}{1-d} & \text{si } t \in [0,a] \cup [a+d, 1]. \end{cases}$

Test de H_0 contre H_1 :

$$\begin{aligned} \frac{L_1(X_1, \cdots, X_n)}{L_0(X_1, \cdots, X_n)} &= \left(\frac{m}{nd}\right)^m \left(\frac{1-m/n}{1-d}\right)^{n-m} \\ &= I_{scan}(m,d) \end{aligned}$$

$\Lambda_{scan} = \sup_{m \geq n_0} I_{scan}(m,d)$.

Distribution de Λ_{scan} sous H_0 inconnue.

La statistique des espacements anormaux

La statistique des espacements anormaux

- $0 = X_{(0)} \leq X_{(1)} \leq \cdots \leq \cdots \leq X_{(n)} \leq X_{(n+1)} = 1.$

La statistique des espacements anormaux

- $0 = X_{(0)} \leq X_{(1)} \leq \cdots \leq \cdots \leq X_{(n)} \leq X_{(n+1)} = 1.$
- $D_i = X_{(i)} - X_{(i-1)}, i = 1, \cdots, n+1.$

La statistique des espacements anormaux

- $0 = X_{(0)} \leq X_{(1)} \leq \cdots \leq \cdots \leq X_{(n)} \leq X_{(n+1)} = 1.$
- $D_i = X_{(i)} - X_{(i-1)}, i = 1, \cdots, n+1.$
- Idée : Observer les espacements successifs.

La statistique des espacements anormaux

- $0 = X_{(0)} \leq X_{(1)} \leq \cdots \leq \cdots \leq X_{(n)} \leq X_{(n+1)} = 1.$
- $D_i = X_{(i)} - X_{(i-1)}, i = 1, \cdots, n+1.$
- Idée : Observer les espacements successifs.
- $S_{j,k} = \sum_{i=j+1}^{k} D_i = X_{(k)} - X_{(j)}, 1 \leq j < k \leq n.$

La statistique des espacements anormaux

- $0 = X_{(0)} \leq X_{(1)} \leq \cdots \leq \cdots \leq X_{(n)} \leq X_{(n+1)} = 1.$
- $D_i = X_{(i)} - X_{(i-1)}, i = 1, \cdots, n+1.$
- Idée : Observer les espacements successifs.
- $S_{j,k} = \sum_{i=j+1}^{k} D_i = X_{(k)} - X_{(j)}, 1 \leq j < k \leq n.$
- $\Lambda_{j,k} = F_0(S_{j,k}) = B_{inc}(S_{j,k}, k-j, n+1-k+j).$

La statistique des espacements anormaux

La statistique des espacements anormaux

- $\forall j = 1, \cdots, n-1, \forall k = j+1, \cdots, n$:

$$\Lambda_{j,k} \sim U([0,1]) \text{ sous } H_0.$$

La statistique des espacements anormaux

- $\forall j = 1, \cdots, n-1, \forall k = j+1, \cdots, n$:

$$\Lambda_{j,k} \sim U([0,1]) \text{ sous } H_0.$$

- $I_{spac}(m,d) = 1/\Lambda_{j,k}$

avec $m = k - j + 1$ et $d = X_{(k)} - X_{(j)}$.

La statistique des espacements anormaux

- $\forall j = 1, \cdots, n-1, \forall k = j+1, \cdots, n$:

$$\Lambda_{j,k} \sim U([0,1]) \text{ sous } H_0.$$

- $I_{spac}(m,d) = 1/\Lambda_{j,k}$

avec $m = k - j + 1$ et $d = X_{(k)} - X_{(j)}$.

- Statistique retenue :

$$\Lambda_{spac} = \sup_{m \geq n_0} I_{spac}(m,d).$$

La statistique des espacements anormaux

- $\forall j = 1, \cdots, n-1, \forall k = j+1, \cdots, n$:

$$\Lambda_{j,k} \sim U([0,1]) \text{ sous } H_0.$$

- $I_{spac}(m,d) = 1/\Lambda_{j,k}$

avec $m = k - j + 1$ et $d = X_{(k)} - X_{(j)}$.

- Statistique retenue :

$$\Lambda_{spac} = \sup_{m \geq n_0} I_{spac}(m,d).$$

- Distribution de Λ_{spac} sous H_0 inconnue.

Utilisation de la statistique

Utilisation de la statistique

- Agrégat le plus probable : $[X_{(j)}, X_{(k)}]$ maximisant $I_{spac}(m, d)$ (resp. $I_{scan}(m, d)$)
  avec $m = k - j + 1$ et $d = X_{(k)} - X_{(j)}$.

Utilisation de la statistique

- Agrégat le plus probable : $[X_{(j)}, X_{(k)}]$ maximisant $I_{spac}(m, d)$ (resp. $I_{scan}(m, d)$)
  avec $m = k - j + 1$ et $d = X_{(k)} - X_{(j)}$.

- Agrégat jugé significatif (rejet de H_0) si Λ_{spac} (resp. Λ_{scan}) supérieur au quantile d'ordre α sous H_0 (obtenu par simulation).

Détection de plusieurs agrégats

Détection de plusieurs agrégats

Ce qui se fait en pratique :

Détection de plusieurs agrégats

Ce qui se fait en pratique :

- Agrégat significatif sur $[X_{(j1)}, X_{(k1)}]$.

Détection de plusieurs agrégats

Ce qui se fait en pratique :

- Agrégat significatif sur $[X_{(j1)}, X_{(k1)}]$.
- On cherche $[X_{(j2)}, X_{(k2)}]$ maximisant $I(m2, d2)$ et tel que $[X_{(j1)}, X_{(k1)}] \cap [X_{(j2)}, X_{(k2)}] = \emptyset$.

Détection de plusieurs agrégats

Ce qui se fait en pratique :

- Agrégat significatif sur $[X_{(j1)}, X_{(k1)}]$.
- On cherche $[X_{(j2)}, X_{(k2)}]$ maximisant $I(m2, d2)$ et tel que $[X_{(j1)}, X_{(k1)}] \cap [X_{(j2)}, X_{(k2)}] = \emptyset$.
- Agrégat significatif sur $[X_{(j2)}, X_{(k2)}]$ si $I(m2, d2)$ supérieur au quantile d'ordre α.

Détection de plusieurs agrégats

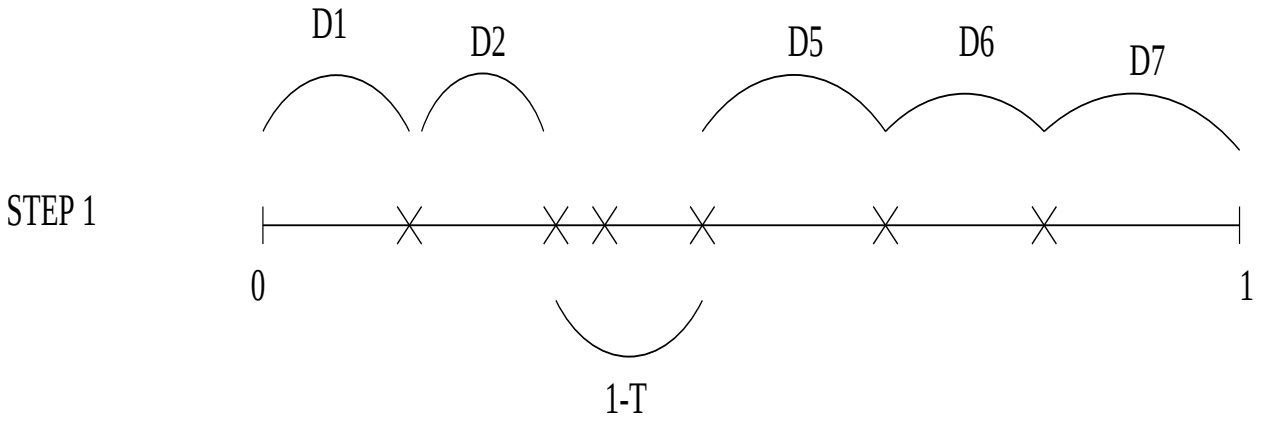
Ce qui se fait en pratique :

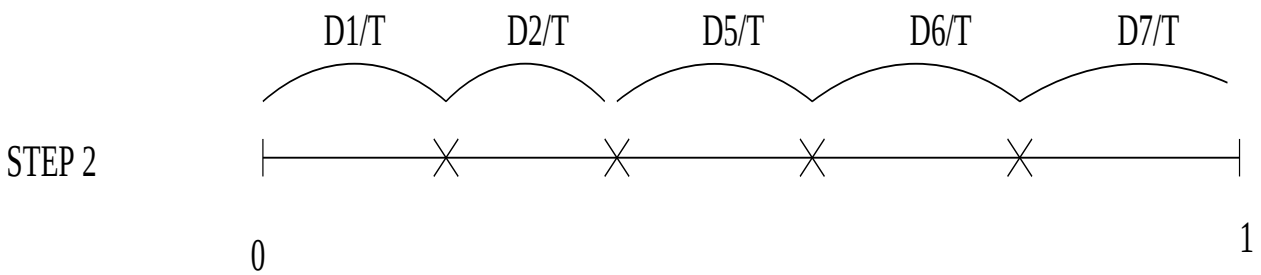
- Agrégat significatif sur $[X_{(j1)}, X_{(k1)}]$.
- On cherche $[X_{(j2)}, X_{(k2)}]$ maximisant $I(m2, d2)$ et tel que $[X_{(j1)}, X_{(k1)}] \cap [X_{(j2)}, X_{(k2)}] = \emptyset$.
- Agrégat significatif sur $[X_{(j2)}, X_{(k2)}]$ si $I(m2, d2)$ supérieur au quantile d'ordre α.

Problème de test multiple $\Rightarrow$ procédure conservative.

Une procédure multiple

Une procédure multiple





Une procédure multiple

Une procédure multiple

- Agrégat significatif : $[X_{i^*}, X_{j^*}]$.

Une procédure multiple

- Agrégat significatif : $[X_{i^*}, X_{j^*}]$.
- On analyse $\{X_k^{(2)}, k = 1, \cdots, n^{(2)}\}$ avec

$$X_k^{(2)} = \begin{cases} \frac{X_{(k)}}{T^*} & \text{if } 1 \leq k \leq i^*, \\ \frac{X_{(k+j^*-i^*)} - X_{j^*} + X_{i^*}}{T^*} & \text{if } i^* + 1 \leq k \leq n - j^* + i^*, \end{cases}$$

où $T^* = 1 - X_{j^*} + X_{i^*}$ et $n^{(2)} = n - j^* + i^*$.

Une procédure multiple

- Agrégat significatif : $[X_{i^*}, X_{j^*}]$.
- On analyse $\{X_k^{(2)}, k = 1, \cdots, n^{(2)}\}$ avec

$$X_k^{(2)} = \begin{cases} \frac{X_{(k)}}{T^*} & \text{if } 1 \leq k \leq i^*, \\ \frac{X_{(k+j^*-i^*)} - X_{j^*} + X_{i^*}}{T^*} & \text{if } i^* + 1 \leq k \leq n - j^* + i^*, \end{cases}$$

où $T^* = 1 - X_{j^*} + X_{i^*}$ et $n^{(2)} = n - j^* + i^*$.

- Itération tant que agrégat significatif.

Une procédure multiple

- Agrégat significatif : $[X_{i^*}, X_{j^*}]$.
- On analyse $\{X_k^{(2)}, k = 1, \cdots, n^{(2)}\}$ avec

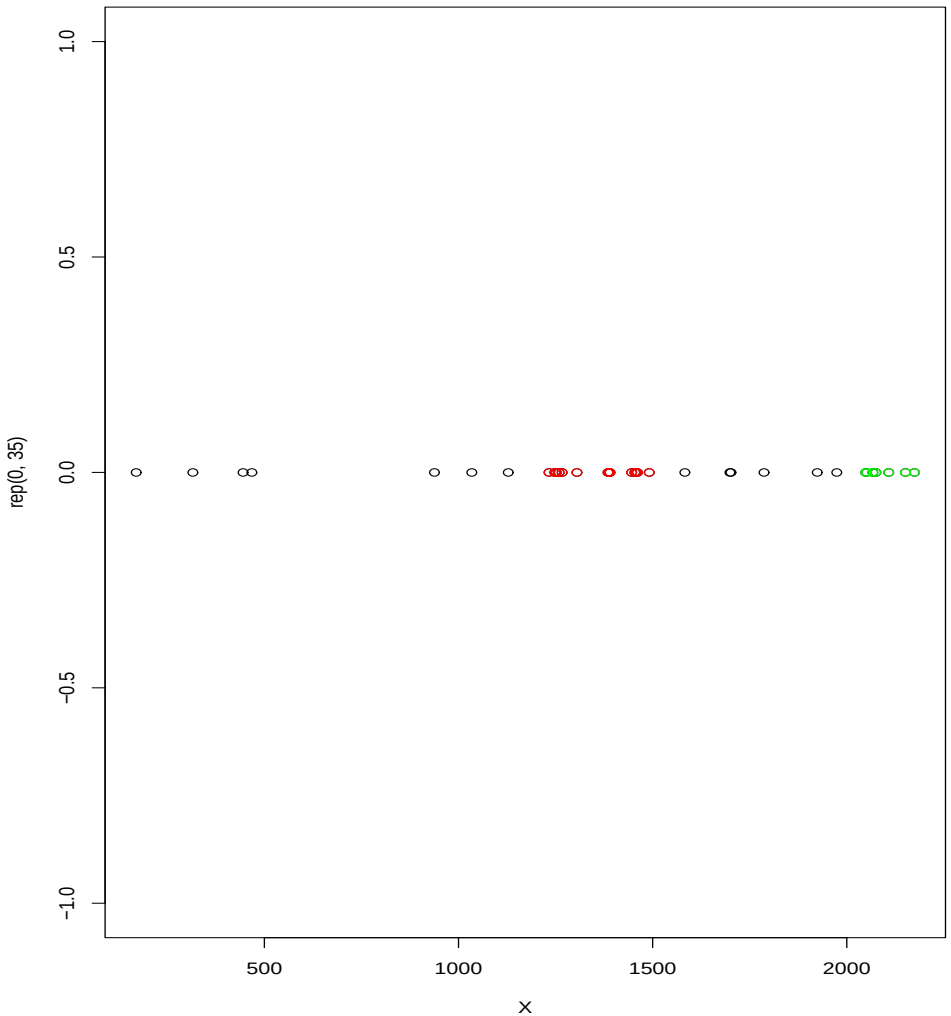
$$X_k^{(2)} = \begin{cases} \frac{X_{(k)}}{T^*} & \text{if } 1 \leq k \leq i^*, \\ \frac{X_{(k+j^*-i^*)} - X_{j^*} + X_{i^*}}{T^*} & \text{if } i^* + 1 \leq k \leq n - j^* + i^*, \end{cases}$$

où $T^* = 1 - X_{j^*} + X_{i^*}$ et $n^{(2)} = n - j^* + i^*$.

- Itération tant que agrégat significatif.
- $C \subseteq [X_{i^*}, X_{j^*}] \Rightarrow X_1^{(2)}, \cdots, X_{n^{(2)}}^{(2)} \sim H_0$.

Agrégats identifiés

Agrégats identifiés



Significativité des agrégats

Erreur 1ère espèce : $\alpha = 0.2$

Tab.: Tests applied to Knox data set

Méthode	$\hat{C}$	p-valeur
$M\Lambda_{spac}$	$[1233, 1491]$	0.0039
	$[1233, 1491] \cup [2049, 2174]$	0.1209
$M\Lambda_{scan}(n_0 = 2)$	$[1233, 1491]$	0.1298
$M\Lambda_{scan}(n_0 = 5)$	$[1233, 1491]$	0.0086
	$[1233, 1491] \cup [2049, 2174]$	0.0806

Données sismiques

Secousses enregistrées en Italie en Janvier 2006.

Tab.: Tests applied to Italy earthquakes data set

Method	$\hat{C}$	p-value
$M\Lambda_{HF}$	$[01/08/2006, 01/08/2006]$	0.0058
$M\Lambda_{scan}(n_0 = 2)$	$[01/08/2006, 01/08/2006]$	0.0263
$M\Lambda_{scan}(n_0 = 5)$	$[01/07/2006, 01/08/2006]$	0.0204
$M\Lambda_{scan}(n_0 = 4)$	$[01/08/2006, 01/08/2006]$	0.0016

Application à des données simulées

Application à des données simulées

Simulation de 1000 jeux de données de 100 évènements (agrégat plat).

Application à des données simulées

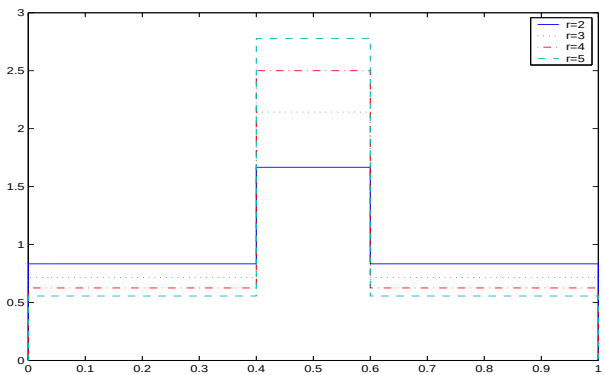
Simulation de 1000 jeux de données de 100 évènements (agrégat plat).

$$f_1(x) = \begin{cases} \frac{r}{0.8+0.2r} & \text{si } 0.4 \leq x \leq 0.6, \\ \frac{1}{0.8+0.2r} & \text{si } x \in [0, 0.4] \cup [0.6, 1], \\ 0 & \text{sinon.} \end{cases}$$

Application à des données simulées

Simulation de 1000 jeux de données de 100 évènements (agrégat plat).

$$f_1(x) = \begin{cases} \frac{r}{0.8+0.2r} & \text{si } 0.4 \leq x \leq 0.6, \\ \frac{1}{0.8+0.2r} & \text{si } x \in [0, 0.4] \cup [0.6, 1], \\ 0 & \text{sinon.} \end{cases}$$



Application à des données simulées

Application à des données simulées

Erreur de première espèce : 5%.

Application à des données simulées

Erreur de première espèce : 5%.

$TP = \nu(C \cap \hat{C})$

Application à des données simulées

Erreur de première espèce : 5%.

$TP = \nu(C \cap \hat{C})$

$TN = \nu(\bar{\hat{C}} \cap \bar{C})$.

Application à des données simulées

Erreur de première espèce : 5%.

$TP = \nu(C \cap \hat{C})$

$TN = \nu(\bar{\hat{C}} \cap \bar{C})$.

$I = TP + TN$.

Application à des données simulées

Tab.: Tests applied to a simulated flat cluster

r		Empirical results of the following :			
		Λ_{spac}	$M\Lambda_{spac}$	Λ_{scan}	$M\Lambda_{scan}$
2	Power	**0.489**		0.425	
	TP	0.07008	**0.07091**	0.05681	0.05731
	TN	0.77711	0.77649	**0.77934**	0.77905
	I	0.84719	**0.84740**	0.83615	0.83636
3	Power	**0.948**		0.933	
	TP	0.16447	**0.16481**	0.15860	0.15943
	TN	0.78121	0.77982	**0.78182**	0.77987
	I	**0.94568**	0.94463	0.94042	0.93930
5	Power	**1**		1	
	TP	0.18636	**0.18642**	0.18621	0.18687
	TN	0.79339	0.79243	**0.79339**	0.79050
	I	**0.97975**	0.97885	0.9796	0.97737

Application à des données simulées

Application à des données simulées

Simulation de 1000 jeux de données de 100 évènements (agrégat en cloche).

Application à des données simulées

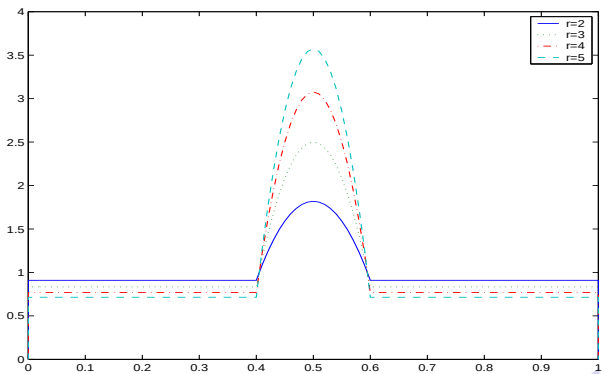
Simulation de 1000 jeux de données de 100 évènements (agrégat en cloche).

$$f_2(x) = \begin{cases} \frac{10}{r+9}\{1 + (r-1) * [1 - 100(x-0.5)^2]\} & \text{si } 0.4 \leq x \leq 0.6, \\ \frac{10}{r+9} & \text{si } x \in [0, 0.4] \cup [0.6, \\ 0 & \text{sinon.} \end{cases}$$

Application à des données simulées

Simulation de 1000 jeux de données de 100 évènements (agrégat en cloche).

$$f_2(x) = \begin{cases} \frac{10}{r+9}\{1 + (r-1) * [1 - 100(x-0.5)^2]\} & \text{si } 0.4 \leq x \leq 0.6, \\ \frac{10}{r+9} & \text{si } x \in [0, 0.4] \cup [0.6, \\ 0 & \text{sinon.} \end{cases}$$



Application à des données simulées

Tab.: Tests applied to a simulated bell cluster

r		Empirical results of the following :			
		Λ_{spac}	$M\Lambda_{spac}$	Λ_{scan}	$M\Lambda_{scan}$
2	Power	**0.318**		0.285	
	TP	0.03542	**0.03558**	0.02701	0.02717
	TN	0.78670	0.78670	**0.78958**	0.78955
	I	0.82212	**0.82228**	0.81659	0.81672
3	Power	**0.833**		0.783	
	TP	0.10780	**0.10838**	0.09797	0.09904
	TN	0.78923	0.78809	**0.79019**	0.78804
	I	**0.89703**	0.89647	0.88816	0.88708
5	Power	**1**		0.999	
	TP	0.14745	**0.14846**	0.14592	0.14732
	TN	0.79499	0.79191	**0.79512**	0.79121
	I	**0.94244**	0.94037	0.94104	0.93853

Application à des données simulées

Tab.: Tests applied to two simulated bell clusters

r		Empirical results of the following :			
		Λ_{HF}	$M\Lambda_{HF}$	Λ_{scan}	$M\Lambda_{scan}$
2	Power	**0.241**		0.214	
	TP	0.03201	**0.03471**	0.02479	0.02623
	TN	0.58689	0.58616	**0.58821**	0.58802
	I	0.6189	**0.62087**	0.61300	0.61425
3	Power	**0.705**		0.652	
	TP	0.10524	**0.14506**	0.09279	0.12277
	TN	**0.57007**	0.56638	0.56764	0.56323
	I	0.67531	**0.71144**	0.66043	0.68600
5	Power	**0.991**		0.981	
	TP	0.18975	**0.27706**	0.19204	0.27279
	TN	**0.53388**	0.52863	0.52504	0.51947
	I	0.72363	**0.80569**	0.71708	0.79226

Plan de l'exposé

Plan de l'exposé

Détection d'agrégats spatiaux

Détection d'agrégats spatiaux

n événements observés en $\{s_1, \cdots, s_n\}$, $s_i \in D \subset \mathbb{R}^2$.

Détection d'agrégats spatiaux

n événements observés en $\{s_1, \cdots, s_n\}$, $s_i \in D \subset \mathbb{R}^2$.

$H_0 : \{s_1, \cdots, s_n\} i.i.d. \sim f(.)$.

Détection d'agrégats spatiaux

n événements observés en $\{s_1, \cdots, s_n\}$, $s_i \in D \subset \mathbb{R}^2$.

$H_0 : \{s_1, \cdots, s_n\} i.i.d. \sim f(.)$.

Principal problème : agrégats potentiels en nombre infini.

Détection d'agrégats spatiaux

n événements observés en $\{s_1, \cdots, s_n\}$, $s_i \in D \subset \mathbb{R}^2$.

$H_0 : \{s_1, \cdots, s_n\} i.i.d. \sim f(.)$.

Principal problème : agrégats potentiels en nombre infini.

Première solution :

Détection d'agrégats spatiaux

n événements observés en $\{s_1, \cdots, s_n\}$, $s_i \in D \subset \mathbb{R}^2$.

$H_0 : \{s_1, \cdots, s_n\} i.i.d. \sim f(.)$.

Principal problème : agrégats potentiels en nombre infini.

Première solution :

- définir une famille finie d'agrégats potentiels.

Détection d'agrégats spatiaux

n événements observés en $\{s_1, \cdots, s_n\}$, $s_i \in D \subset \mathbb{R}^2$.

$H_0 : \{s_1, \cdots, s_n\} i.i.d. \sim f(.)$.

Principal problème : agrégats potentiels en nombre infini.

Première solution :

- définir une famille finie d'agrégats potentiels.
- Mêmes indicateurs de concentration avec :

$$\begin{aligned} I(m,d) &\rightarrow I(m,A) \\ d &\rightarrow \int_A f(s)\nu(ds) = \nu_f(A) \end{aligned}$$

Détection d'agrégats spatiaux

Détection d'agrégats spatiaux

Exemple : la "spatial scan statistic" (Kulldorff, 1997)

Détection d'agrégats spatiaux

Exemple : la "spatial scan statistic" (Kulldorff, 1997)

- Agrégats potentiels : disques dont centre=point d'une grille.

Détection d'agrégats spatiaux

Exemple : la "spatial scan statistic" (Kulldorff, 1997)

- Agrégats potentiels : disques dont centre=point d'une grille.
- $$I_{scan}(m, A) = \left(\frac{m}{n\nu_f(A)}\right)^m \left(\frac{1 - m/n}{1 - \nu_f(A)}\right)^{n-m}.$$

Détection d'agrégats spatiaux

Exemple : la "spatial scan statistic" (Kulldorff, 1997)

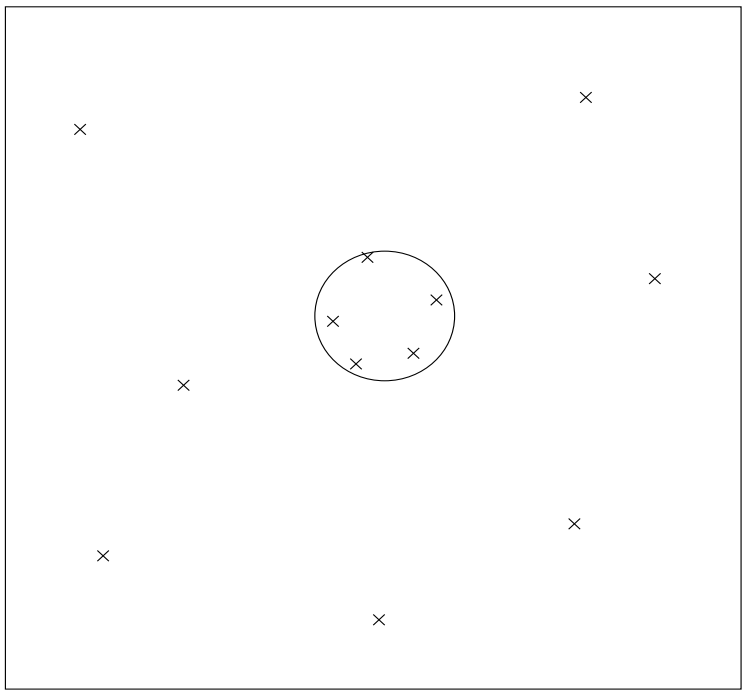
- Agrégats potentiels : disques dont centre=point d'une grille.
- $$I_{scan}(m, A) = \Big(\frac{m}{n\nu_f(A)}\Big)^m \Big(\frac{1 - m/n}{1 - \nu_f(A)}\Big)^{n-m}.$$

Idée : utiliser plutôt

$$I_{spac}(m, A) = 1/B_{inc}(\nu_f(A), m - 1, n + 2 - m).$$

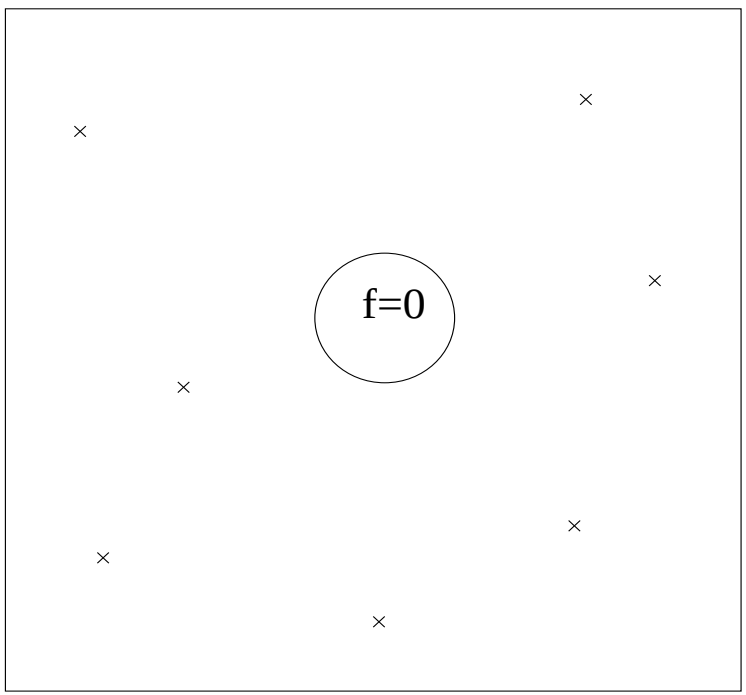
Une procédure multiple

Une procédure multiple



Une procédure multiple

Une procédure multiple



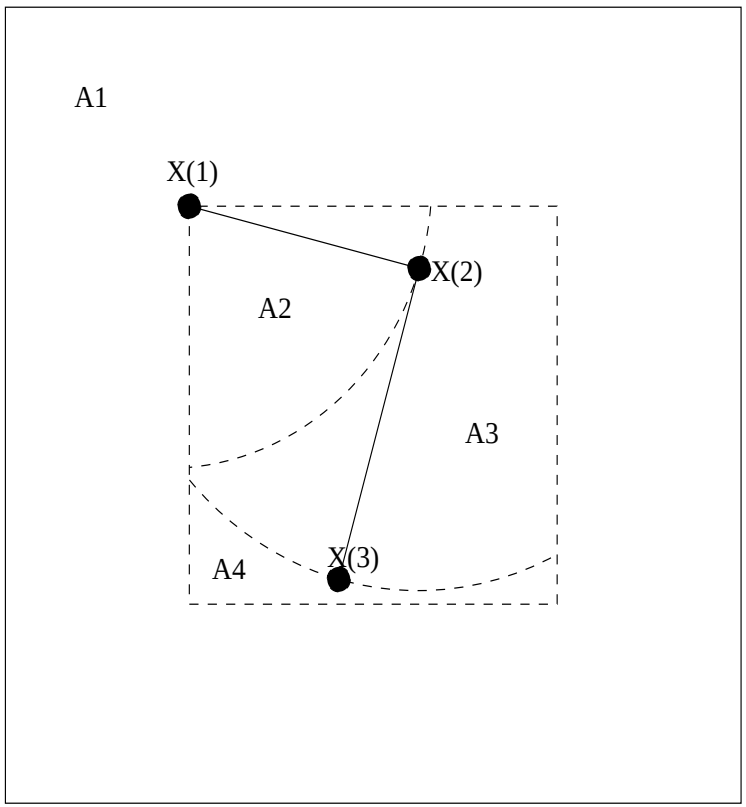
Détection d'agrégats spatiaux

Détection d'agrégats spatiaux

Deuxième solution : transformer les données.

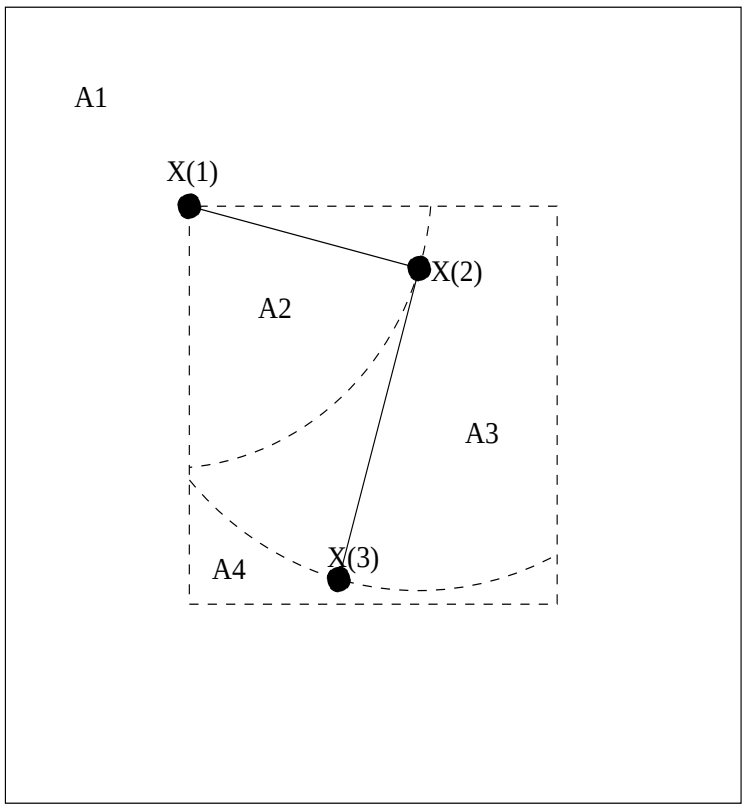
Détection d'agrégats spatiaux

Deuxième solution : transformer les données.



Détection d'agrégats spatiaux

Deuxième solution : transformer les données.



$$A'_i = \int_{A_i} f(s)\nu(ds), i = 1, \cdots, n+1.$$

Détection d'agrégats spatiaux

Détection d'agrégats spatiaux

$B' = \int_B f(s)\nu(ds), \forall B \subset A.$

Détection d'agrégats spatiaux

$B' = \int_B f(s)\nu(ds), \forall B \subset A.$

$S_{1,r} = \{s \in A : d(s, \partial A) \leq r\}.$

Détection d'agrégats spatiaux

$B' = \int_B f(s)\nu(ds)$, $\forall B \subset A$.

$S_{1,r} = \{s \in A : d(s, \partial A) \leq r\}$.

Sous H_0 :

$$\begin{aligned} P(A'_1 \leq S'_{1,r}) &= 1 - P(A'_1 > S'_{1,r}) \\ &= 1 - P(X_i \in \overline{S_{1,r}}, \forall i = 1, \cdots, n) \\ &= 1 - \Big(\int_{\overline{S_{1,r}}} f(s)\nu(ds) \Big)^n = 1 - (1 - S'_{1,r})^n, \end{aligned}$$

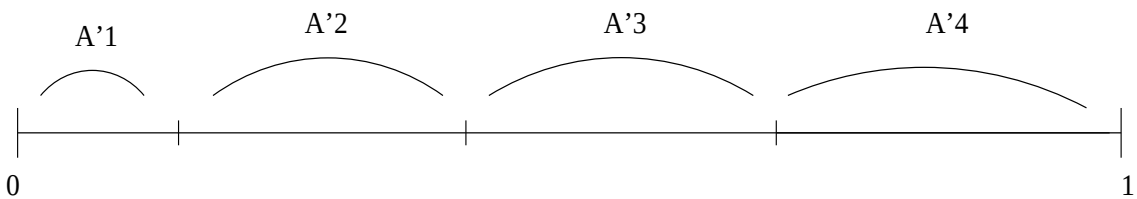
Détection d'agrégats spatiaux

Détection d'agrégats spatiaux

Sous H_0 : $\{A'_1, \cdots, A'_{n+1}\} \sim \{D_1, \cdots, D_{n+1}\}$.

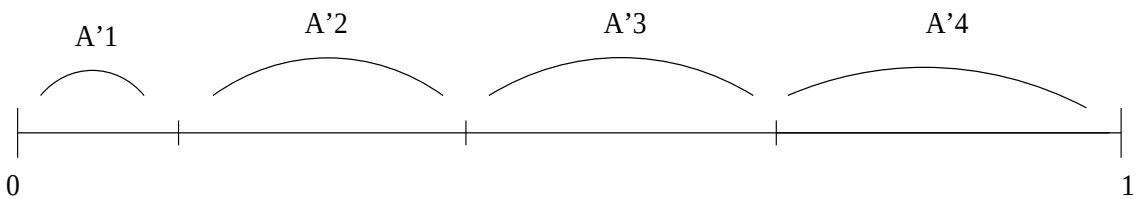
Détection d'agrégats spatiaux

Sous H_0 : $\{A'_1, \cdots, A'_{n+1}\} \sim \{D_1, \cdots, D_{n+1}\}$.



Détection d'agrégats spatiaux

Sous H_0 : $\{A'_1, \cdots, A'_{n+1}\} \sim \{D_1, \cdots, D_{n+1}\}$.



Détection d'agrégats temporels en utilisant $\{A'_1, \cdots, A'_{n+1}\}$ au lieu de $\{D_1, \cdots, D_{n+1}\}$.

Matérialisation des agrégats spatiaux

Matérialisation des agrégats spatiaux

Agrégats temporels : $I_i = [T_{(a_i)}, T_{(b_i)}], 1 \leq i \leq k$.

Matérialisation des agrégats spatiaux

Agrégats temporels : $I_i = [T_{(a_i)}, T_{(b_i)}], 1 \leq i \leq k$.

Agrégats spatiaux : $\hat{C}_i = \bigcup_{j=a_i+1}^{b_i} A_i$.

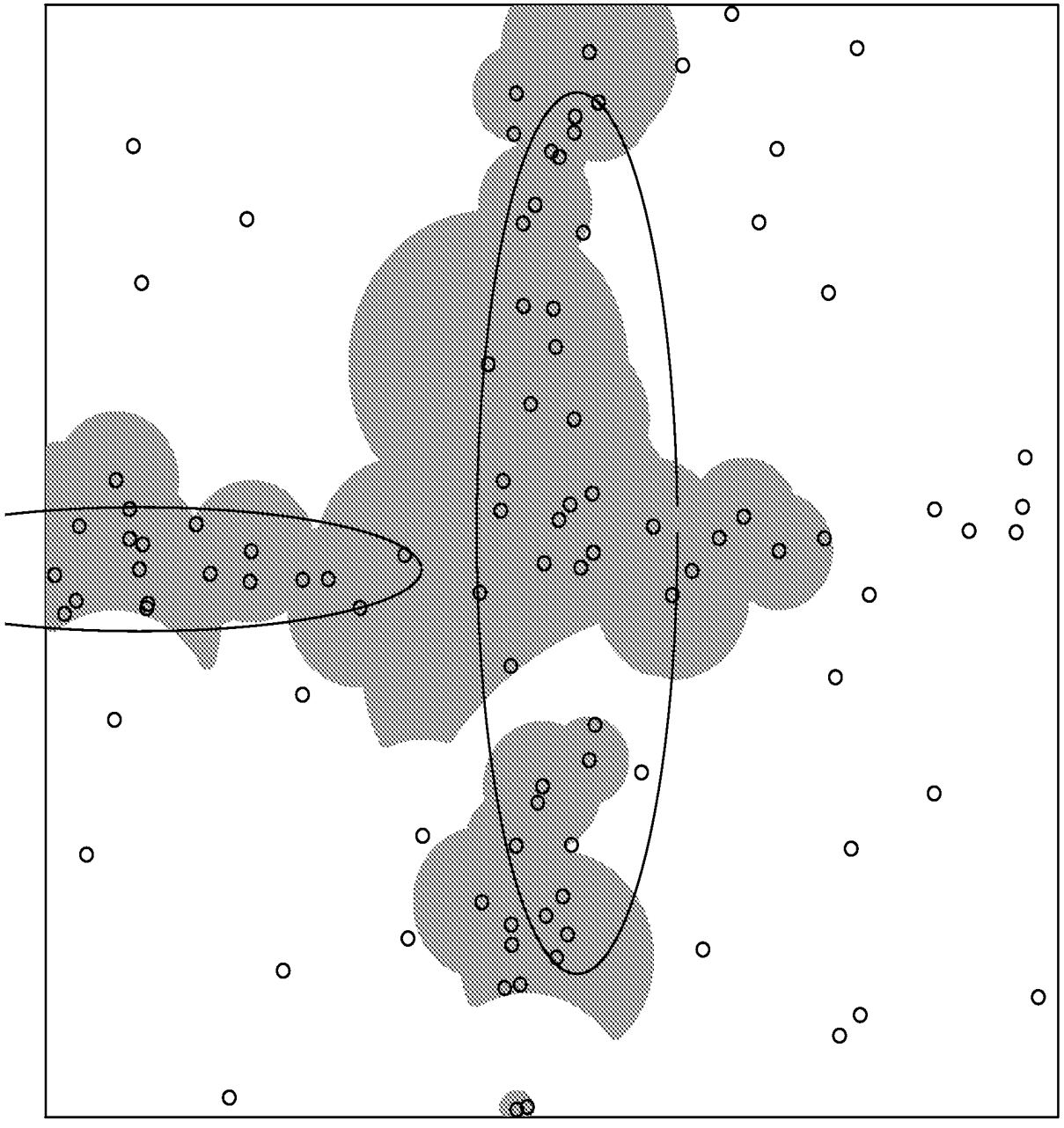
Matérialisation des agrégats spatiaux

Agrégats temporels : $I_i = [T_{(a_i)}, T_{(b_i)}], 1 \leq i \leq k$.

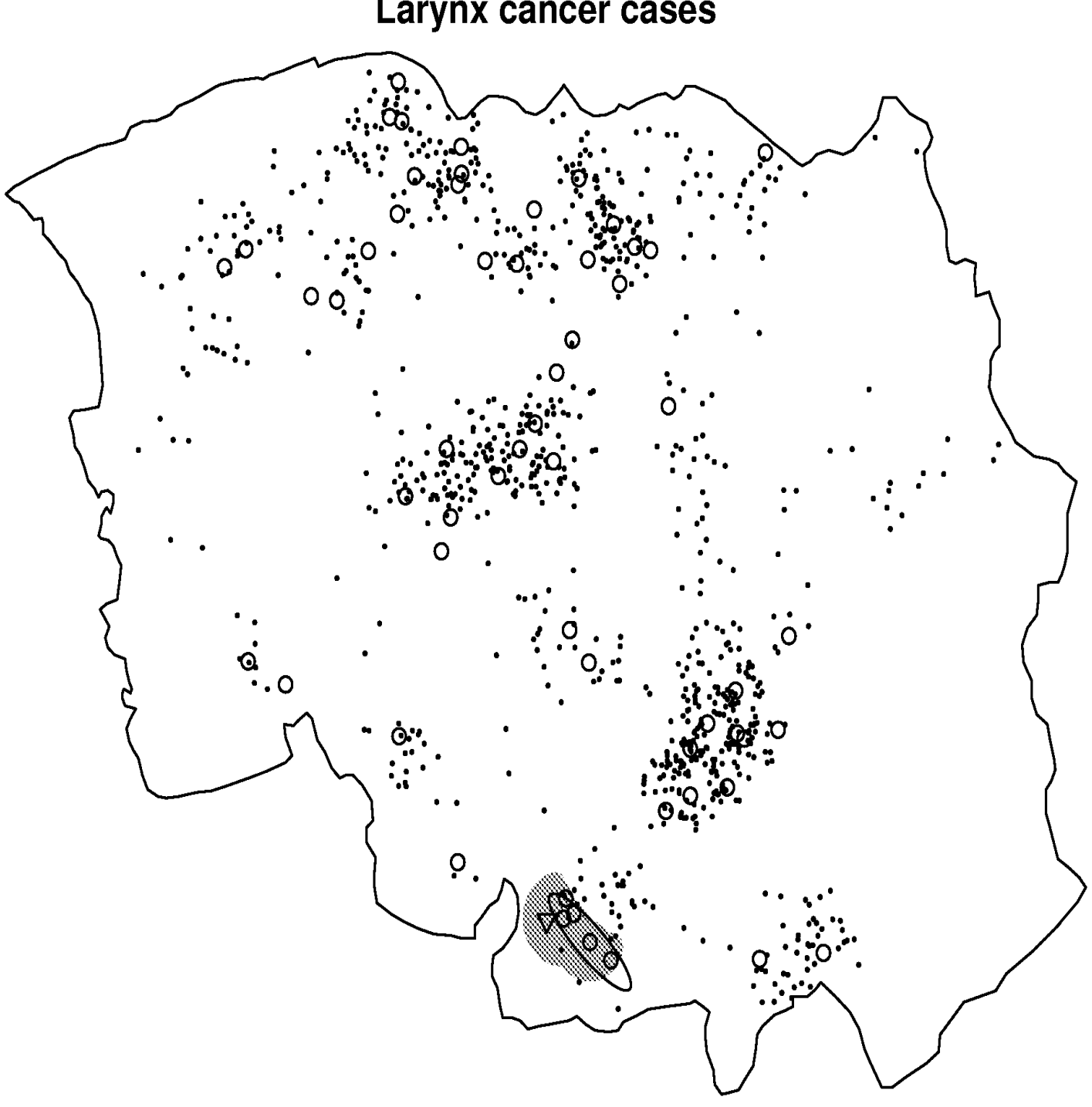
Agrégats spatiaux : $\hat{C}_i = \bigcup_{j=a_i+1}^{b_i} A_i$.

$\hat{C} = \bigcup_{i=1}^{k} \hat{C}_i$.

Application à des données simulées



Application à des données réelles


Larynx cancer cases

Plan de l'exposé

Plan de l'exposé

Conclusions et perspectives

Conclusions et perspectives

- Alternative à la statistique de balayage : indépendante de H_1, plus puissante (temporel).

Conclusions et perspectives

- Alternative à la statistique de balayage : indépendante de H_1, plus puissante (temporel).
- Procédure de détection d'agrégats de toutes formes (spatial).

Conclusions et perspectives

- Alternative à la statistique de balayage : indépendante de H_1, plus puissante (temporel).
- Procédure de détection d'agrégats de toutes formes (spatial).
- Nécessité de pouvoir détecter plusieurs agrégats : mise en place d'une procédure multiple.

Conclusions et perspectives

- Alternative à la statistique de balayage : indépendante de H_1, plus puissante (temporel).
- Procédure de détection d'agrégats de toutes formes (spatial).
- Nécessité de pouvoir détecter plusieurs agrégats : mise en place d'une procédure multiple.
- Travail sur la distribution de Λ_{spac}.

Conclusions et perspectives

- Alternative à la statistique de balayage : indépendante de H_1, plus puissante (temporel).
- Procédure de détection d'agrégats de toutes formes (spatial).
- Nécessité de pouvoir détecter plusieurs agrégats : mise en place d'une procédure multiple.
- Travail sur la distribution de Λ_{spac}.
- Extension au cadre spatio-temporel : travail avec C. Dematteï et N. Molinari (distance spatio-temporelle).