Statistiques de scan spatial: Détection et Inférence

Moustapha Sarr

Master 2 ISN
Université de Lille

Plan

1. Introduction
2. Statistique de scan spatial pour des données laticielles (agrégées)
3. Statistique de scan spatial pour des données ordinales

Introduction

Un **cluster spatial** est une zone géographique où l'on observe une concentration d'événements (par exemple, **des cas de maladies**).

Une méthode de détection et d'inférence de clusters spatiaux est un ensemble d'approches et de techniques permettant d'identifier des **clusters spatiaux** et de déterminer leur significativité. Ces méthodes sont utilisées pour repérer des zones présentant un risque anormalement élevé ou faible.

Dans les travaux scientifiques, les méthodes sont divisées en trois catégories :

1- Tests globaux :

- Évaluent si les événements sont répartis de manière aléatoire dans l'espace ou s'ils forment une agrégation globale
- Pas de localisation de clusters
- Utile lorsqu'on cherche à savoir si une maladie est infectieuse ou non.
- Exemple de méthode : **Indice de Moran**, **Diggle et Chetwynd (1991)**,...

2- Tests focalisés :

- Connaissance a priori de la localisation du cluster
- Utile lorsque la région étudiée contient un danger potentiel pour la santé
- Exemple: une centrale à charbon, et que l'on soupçonne un regroupement de cas de cancer du poumon autour de la région.
- Exemple de méthode : **Test de Stone**,...

3- Tests de détection :

- Aucune connaissance a priori sur la localisation des clusters
- Détection et test de significativité
- Exemple de méthode : **Statistiques de scan**,...

Statistiques de scan

Nous nous intéressons à la troisième méthode de détection de cluster.

La **statistique de scan** est une **variable aléatoire** utilisée comme **statistique de test** afin de détecter la **présence d'un cluster**, c'est-à-dire une zone de concentration anormale d'un phénomène, au sein d'un espace étudié S (zone géographique, période de temps...).

- Hypothèse nulle H_0 : Répartition aléatoire des événements dans S.

exemple

Cas de maladie répartis de manière homogène sur une période de temps, zone géographique .

- Hypothèse alternative H_1 : Présence d'un cluster $z \in S$ dans lequel la probabilité d'apparition d'un événement est différente de celle dans le reste de S .

exemple

Zone géographique dans laquelle la probabilité d'être atteint du COVID-19 est supérieure au reste dans la zone géographique étudiée .

Principe générale

Notons $\{X_i\}_{i\geq 1}$ les variables observées, une **statistique de scan** vise à tester **l'hypothèse nulle** H_0 : les X_i sont des variables indépendantes et identiquement distribuées selon une loi de probabilité F avec paramètre p :

$$\forall i, X_i \sim F(p),$$

contre **l'hypothèse alternative** H_1 : les X_i sont indépendantes et il existe un **cluster** z tel que :

$$\forall i \in z, X_i \sim F(p_z), \quad \forall i \in z^c, X_i \sim F(p_{z^c}), \quad p_z > p_{z^c}.$$

Données latticielles (agrégées) :

- Les données sont **agrégées** (regroupées) en une unité spatiale (commune, département...)
- La localisation correspond au centre de l'unité spatiale (son centroïde)
- **Une variable aléatoire** est associée à chaque centre d'unité spatiale
- La surface géographique étudiée est considérée comme **discrète** et modélisée sous forme de **graphe** (lattice).

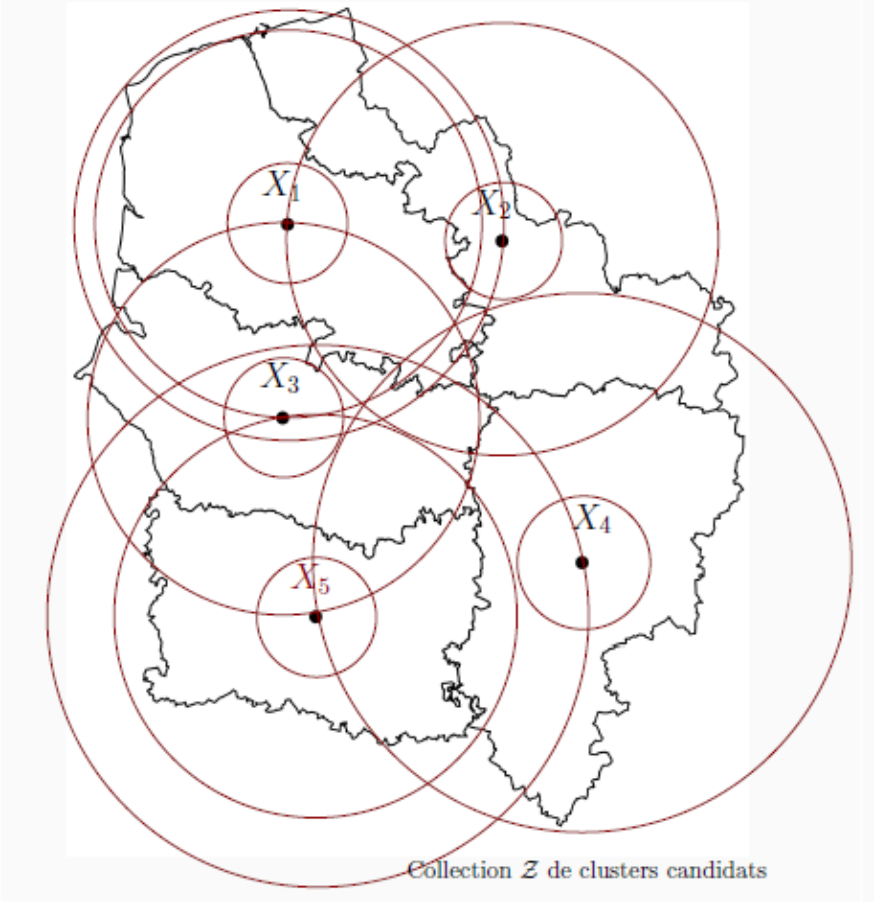
La méthode se décompose en deux phases :

- Détection du cluster le plus probable **(Most Likely Cluster (MLC))**.
- Inférence statistique.

Phase de détection du MLC :

- Processus de scan (de balayage) sur S : On utilise une fenêtre de scan circulaire z centrée en chaque unité spatiale i et de rayon r variant juqu'à un maximum $\mu(z) \leq \frac{\mu(S)}{2}$.

Exemple de collection Z de clusters candidats



Collection $\mathcal{Z}$ de clusters candidats

MLC

$\rightarrow X_i$: mesure au sein de l'unité spatiale i (ex : Nombre de cas de maladie).
$\rightarrow \mu_i$: correspond à la population à risque dans l'unité spatiale i.
$\rightarrow \mu(z)$: correspond à la population à risque du cluster z.
$\rightarrow \mu(S)$: correspond à la population à risque de la zone d'étude R .

A chaque cluster $z \in Z$ est associé indice de concentration $\mathbf{LR(z)} = \frac{L_{H_1^{(z)}}(z)}{L_0}$ (rapport de vraisemblance), où $L(z)$ correspond à la vraisemblance sous H_1^z et L_0 la vraisemblance sous H_0.

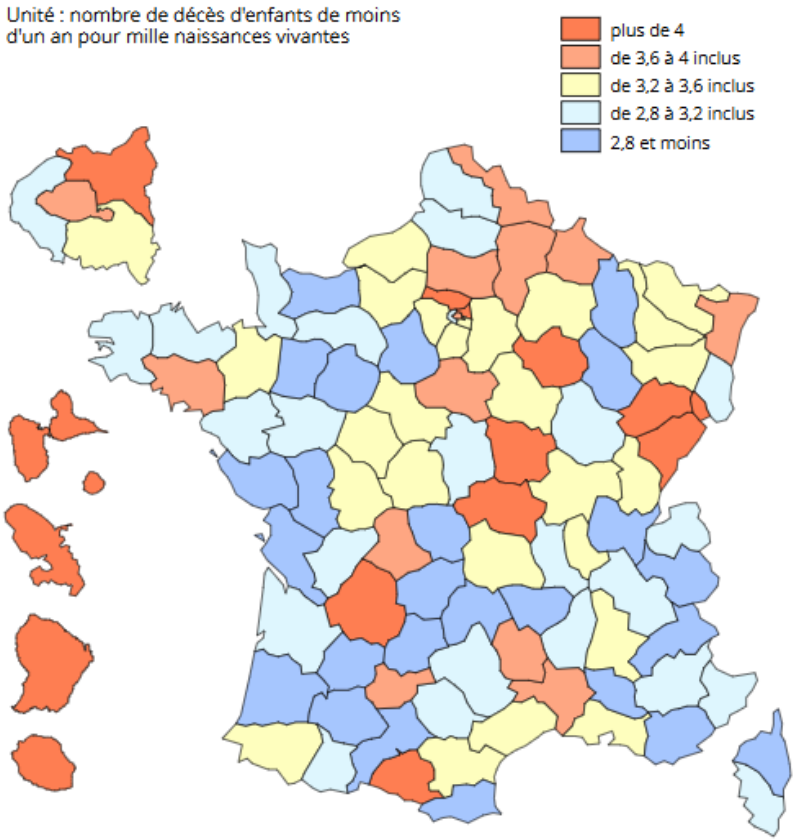
Le cluster le plus probable (MLC) correspond à la zone maximisant le LR :

$$MLC = \arg\max_{z \in Z} LR(z)$$

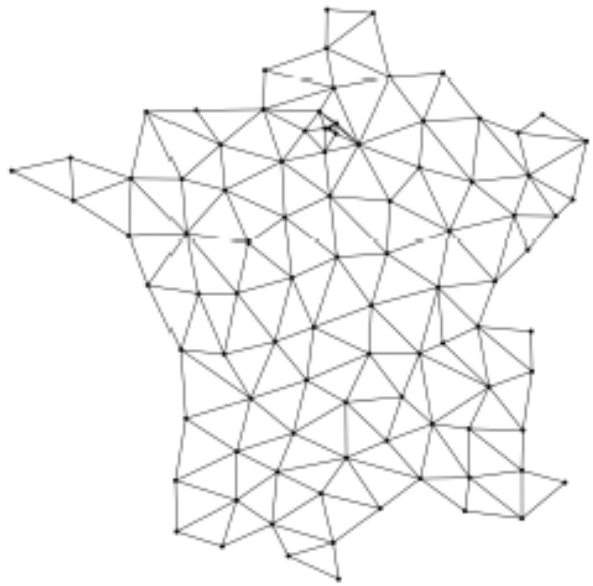
. **La statistique de scan spatiale** :

$$\Lambda = \max_{z \in Z} LR(z) = LR(MLC)$$

Exemples de données latticielles



Taux de mortalité infantile :
France Métropolitaine = 3,37 ‰ ; France = 3,53 ‰
Champ : France, territoire au 31 décembre 2010
Source : Insee, statistiques de l'état civil



(b) Graphe de voisinage $\mathcal{G}$

Phase d'inférence

L'objectif de **la phase d'inférence** est d'associer une **p-value** au MLC.

$\rightarrow$ Problème : La loi de Λ n'a pas de forme analytique sous H_0.
$\rightarrow$ Solution : La loi de Λ est approchée par **simulations de Monte-Carlo**.

- On génère plusieurs jeux de données aléatoires sous l'hypothèse nulle (c'est-à-dire en supposant qu'il n'y a pas de cluster).
- On calcule le Λ pour chaque jeu de données simulé..
- On compare le Λ du cluster observé aux λ des données simulées.

P-Value

Le p-value est défini comme :

$$p = \frac{R}{N+1}$$

où :

- R est le rang de λ observé parmi les λ simulés.
- N est le nombre de simulations Monte Carlo.

exemple

Si on fait 999 simulations et que la valeur observée de λ est parmi les 49 plus grandes valeurs sur 1000, on rejette H_0 avec un seuil de 5%.

Différents modèles de statiques de scan spatial

Il existe différents modèles de statistiques de scan spatiales.

Modèles paramétriques :

- Modèle de Bernoulli : cas / témoin
- Modèle de Poisson : Nombre de cas incidents et population à risque
- Modèle multinomial : Types de cancers...
- Modèle ordinal : stades de cancers...

Modèle de Bernoulli

Ici, on suppose que chaque unité spatiale contient une seule mesure X_i. Dans le contexte des statistiques de scan spatial, l'hypothèse nulle H_0 (absence de cluster) et l'hypothèse alternative associée à un cluster potentiel z ($H_1^{(z)}$) peuvent être définies comme suit :

$$H_0 : \forall i \in S, X_i \sim B(\mu_i, p_S)$$

$$H_1^{(z)} : \forall i \in z, X_i \sim B(\mu_i, p_z) \quad \text{et} \quad \forall i \in z^c, X_i \sim B(\mu_i, p_{z^c}), \quad p_z \neq p_{z^c}.$$

Modèles de Bernoulli et de Poisson

L'indice de concentration proposé par **Kulldorff et Nagarwalla (1995)** est alors donné par :

$$LR_{Ber}(z) = \frac{L_{H_1^{(z)}}(z)}{L_0}$$

où $L_{H_1^{(z)}}(z)$ et L_{H_0} correspondent respectivement aux vraisemblances sous H_1^z et sous H_0.

Ainsi, la statistique de scan spatial Λ_{Ber} est définie par :

$$\Lambda_{Ber} = \max_{z \in Z} \frac{\left(\frac{X(z)}{\mu(z)}\right)^{X(z)} \left(1 - \frac{X(z)}{\mu(z)}\right)^{\mu(z)-X(z)} \left(\frac{X(z^c)}{\mu(z^c)}\right)^{X(z^c)} \left(1 - \frac{X(z^c)}{\mu(z^c)}\right)^{\mu(z^c)-X(z^c)}}{\left(\frac{X(G)}{\mu(G)}\right)^{X(G)} \left(1 - \frac{X(G)}{\mu(G)}\right)^{\mu(G)-X(G)}}$$

Modèles de Bernoulli et de Poisson

- $X(z) = \sum_{i \in z} X_i$
- $X(z^c) = \sum_{i \in z^c} X_i$
- $X(S) = \sum_{i \in S} X_i$
- $\mu(z) = \sum_{i \in z} \mu_i$
- $\mu(z^c) = \sum_{i \in z^c} \mu_i$
- $\mu(S) = \sum_{i \in S} \mu_i$

Modèle de poisson :

Le **modèle de Poisson** proposé par **Kulldorff (1997)** suppose que la variable X_i suit une loi de Poisson :

$$X_i \sim P(p_S \mu_i)$$

où $\mu(i)$ représente la population à risque dans l'unité spatiale i. Ce modèle est particulièrement adapté aux études sur les registres de maladies (ex : registres de cancers) où l'on cherche à détecter des clusters de sur-incidence d'une maladie tout en ajustant sur la population sous-jacente.

Hypothèses

Dans ce cadre, l'hypothèse nulle H_0 est :

$$H_0 : \forall i \in S, X_i \sim P(p_S \mu_i)$$

et l'hypothèse alternative $H_1^{(z)}$ associée au cluster potentiel z est :

$$H_1^{(z)} : \forall i \in z, X_i \sim P(\mu_i p_z), \quad \forall i \in z^c, X_i \sim P(\mu_i p_{z^c}), \quad p_z \neq p_{z^c}.$$

L'indice de concentration proposé par **Kulldorff (1997)** est alors :

$$LR_{Pois}(z) = \frac{L_{H_1^{(z)}}(z)}{L_0}.$$

La statistique de scan spatial Λ_{Pois} est donnée par :

$$\Lambda_{Pois} = \max_{z \in Z} \frac{\left(\frac{X(z)}{\mu(z)}\right)^{X(z)} \left(\frac{X(z^c)}{\mu(z^c)}\right)^{X(z^c)}}{\left(\frac{X(G)}{\mu(G)}\right)^{X(G)}}$$

Cluster secondaire

En plus du cluster le plus probable, il peut être utile d'examiner les clusters secondaires ayant des valeurs élevées du rapport de vraisemblance.

Un cluster secondaire est signalé uniquement s'il n'a pas de chevauchement géographique avec un autre cluster ayant un rapport de vraisemblance plus élevé.
L'interdiction du chevauchement garantit que chaque cluster détecté est réellement distinct et apporte une information nouvelle, évitant ainsi les doublons et assurant une meilleure interprétation épidémiologique.

La significativité statistique d'un cluster secondaire est évaluée indépendamment des autres clusters, en comparant sa valeur du rapport de vraisemblance avec la valeur maximale obtenue à partir des jeux de données simulés.

Ces clusters secondaires permettent d'identifier plusieurs zones à risque plutôt que de se concentrer uniquement sur la plus forte

Modèle multinomial :

Supposons que nous ayons K catégories pour différents types d'une maladie sur une zone d'étude composée de I unités spatiales. Soit c_{ik} le nombre d'observations appartenant à la catégorie k dans l'unité spatiale i $(k = 1, \ldots, K,\ i = 1, \ldots, I)$.
L'hypothèse nulle d'absence de regroupement spatial (clustering) peut être exprimée comme suit : la probabilité d'appartenir à la catégorie k est la même dans toute la zone d'étude pour tous $k = 1, \ldots, K$.

$$H_0 : p_{1z} = p_{1z^c}, p_{2z} = p_{2z^c}, ..., p_{Kz} = p_{Kz^c}$$

où p_{kz} et p_{kz^c} sont les probabilités d'appartenir à la catégorie k à l'intérieur du cluster z et à l'intérieur du cluster z^c , respectivement $(k = 1, \ldots, K)$.
Notez que :

$$\sum_k p_{kz} = \sum_k p_{kz^c} = 1.$$

Modèles multinomial et ordinal

L'hypothèse alternative est qu'il existe au moins une catégorie pour laquelle les probabilités ne sont pas les mêmes.

$$H_1^{(z)} : \exists k \in \{1, \ldots, K\} \quad \text{tel que} \quad p_{kz} \neq p_{kz^c}$$

Soit :

- $c_i = \sum_k c_{ik}$ le nombre total d'observations dans l'unité spatiale i,
- $C_k = \sum_i c_{ik}$ le nombre total d'observations dans la catégorie k,
- $C = \sum_k \sum_i c_{ik}$ le nombre total d'observations dans l'ensemble de la zone d'étude.

L'indice de concentration proposé par **Kulldorff** et **Richard** est alors :

$$LR_{mul}(z) = \frac{L_{H_1^{(z)}}(z)}{L_0}$$

La statistique de scan spatial $\Lambda_{mul} = max$ est définie par :

$$\Lambda_{mul} = \max_{z \in Z} \frac{\prod_k \left(\frac{C_k(z)}{C(z)}\right)^{C_k(z)} \left(\frac{C_k(z^c)}{C(z^c)}\right)^{C_k(z^c)}}{\prod_k \left(\frac{C_k(G)}{C(G)}\right)^{C_k(G)}}$$

Ce modèle est une extension du modèle Bernoulli lorsque $K = 2$.

Modèles multinomial et ordinal

Modèle ordinal:

Supposons que nous ayons une région d'étude composée de I unités spatiales et une variable d'intérêt catégorisée en K niveaux. Les catégories sont de nature ordinale, ce qui signifie que plus k est grand, plus la maladie est grave. Nous définissons ensuite les quantités suivantes :

L'hypothèse nulle H_0 suppose que ces probabilités sont égales à l'intérieur et à l'extérieur de z, comme dans le cas du modèle multinomial :

$$H_0 : p_{1z} = p_{1z^c}, p_{2z} = p_{2z^c}, ..., p_{Kz} = p_{Kz^c}$$

L'hypothèse alternative H_1^z impose une relation ordinale entre ces probabilités :

$$H_1^{(z)} : \frac{p_{1z}}{p_{1z^c}} \leq \frac{p_{2z}}{p_{2z^c}} \leq \ldots \frac{p_{Kz}}{p_{Kz^c}}$$

avec au moins une de ces inégalités étant stricte. Cela signifie que l'on recherche des clusters où les stades les plus graves sont sur-représentés.

Modèles multinomial et ordinal

L'indice de concentration proposé par **Kulldorff** et **Klassen** est alors :

$$LR_{ord}(z) = \frac{L_{H_1^{(z)}}(z)}{L_0}$$

La statistique de scan spatial $\Lambda_{ord} = max$ est définie par :

$$\Lambda_{ord} = \max_{z \in Z} \frac{\prod_k \left(\frac{C_k(z)}{C(z)}\right)^{C_k(z)} \left(\frac{C_k(z^c)}{C(z^c)}\right)^{C_k(z^c)}}{\prod_k \left(\frac{C_k(G)}{C(G)}\right)^{C_k(G)}}$$

Cependant, pour garantir que ces probabilités suivent la contrainte ordinale :

$$\frac{\hat{p}_{1z}}{\hat{p}_{1z^c}} \leq \frac{\hat{p}_{2z}}{\hat{p}_{2z^c}} \leq \dots \frac{\hat{p}_{Kz}}{\hat{p}_{Kz^c}}$$

on applique une procédure de **réarrangement isotone**, en utilisant l'algorithme de pool adjacent violators (PAVA).

Différence entre les deux méthodes

L'algorithme **PAVA (Pool Adjacent Violators Algorithm)** permet de s'assurer que les probabilités estimées respectent bien un ordre croissant. En effet, dans un modèle ordinal, les probabilités doivent vérifier la contrainte :

$$\hat{p}_1 \leq \hat{p}_2 \leq \cdots \leq \hat{p}_K.$$

Cependant, l'estimation brute des probabilités peut ne pas respecter cet ordre. L'algorithme PAVA corrige cette situation en suivant ces étapes :

1. On vérifie si les probabilités sont bien ordonnées.
2. Si une violation est détectée (par exemple, si $\hat{p}_i > \hat{p}_{i+1}$), on ajuste ces valeurs en prenant leur moyenne.
3. On répète ce processus jusqu'à ce que toutes les probabilités soient bien ordonnées.

Cet algorithme permet donc d'obtenir une estimation cohérente des probabilités tout en respectant l'ordre attendu.

le modèle ordinal ajoute une contrainte d'ordre dans l'hypothèse alternative, ce qui permet de détecter des **clusters progressifs (ex: zones où la gravité de la maladie est plus forte**), alors que le modèle multinomial est plus général et détecte toute différence dans la distribution des catégories sans prendre en compte un ordre particulier.