Détection d'agrégats temporels, spatiaux et spatio-temporels: contributions aux méthodes de balayage.

Lionel Cucala

Mercredi 8 Juin 2022

Sommaire

Sommaire

Données localisées

- Dans le temps, dans l'espace ou dans les deux.

- Domaine d'observation : $W \subset \mathbb{R}^d$.

- Localisations : $s_1, \cdots, s_n$.

Détection d'agrégat(s) (cas non marqué)

- Agrégat (Naus, 1963) : Ensemble de localisations "anormalement" proches.

- Adaptation à une mesure de population sous-jacente $\mu(.)$.

- Agrégat : $Z \subset W$ tel que $n(Z)$ "anormal" par rapport à $\mu(Z)$.

Détection d'agrégat(s) (cas marqué)

- Variable X mesurée en chaque localisation.

- Observations : $(s_1, x_1), \cdots, (s_n, x_n)$.

- Agrégat : $Z \subset W$ tel que $\{x_i : s_i \in Z\}$ "anormalement différent" de $\{x_i : s_i \in Z^c\}$.

Méthodes de balayage

- Objectif : balayer la fenêtre W et identifier l'agrégat le plus probable.

- Statistique de balayage : indice de concentration maximum sur un ensemble d'agrégats potentiels

$$\lambda = \max_{Z \in \mathcal{C}} I(Z)$$

- Elle dépend de :
  - l'ensemble des agrégats potentiels $\mathcal{C}$.
  - l'indice de concentration $I(Z)$.

Significativité

- Agrégat le plus probable :

$$\hat{C} = \arg\max_{Z \in \mathcal{C}} I(Z).$$

- Significativité estimée par procedure Monte-Carlo (T simulations) :

$$\text{p-val} = \frac{1 + \sum_{j=1}^{T} \mathbb{1}(\lambda^{(j)} > \lambda)}{T+1}.$$

- Simulations :
  - Cas non marqué : $s_1^{(j)}, \cdots, s_n^{(j)}$ i.i.d. $\sim \mu(.)$.
  - Cas marqué : permutation aléatoire des marques.

Sommaire

Sommaire

Choix classiques : agrégats potentiels à géométrie contrainte

- Cadre temporel (Nagarwalla, 1996) :

$$\mathcal{C} = \{[s_i, s_j] : s_i < s_j\}.$$

- Cadre spatial (Kulldorff, 1997) :

$$\mathcal{C} = \{D_{i,j} : 1 \leq i, j \leq n\}$$

où $D_{i,j}$: disque centré en s_i passant par s_j.

- Cadre spatio-temporel (Kulldorff et al., 1998) :

$$\mathcal{C} = \{C_{i,j} : 1 \leq i, j \leq n\}$$

où $C_{i,j}$: cylindre contenant s_i et s_j.

Agrégats potentiels sans contrainte de forme

- Utilisation d'agrégats elliptiques (Kulldorff et al., 2006).

- Utilisation des cellules de Voronoi (Duczmal et al., 2011).

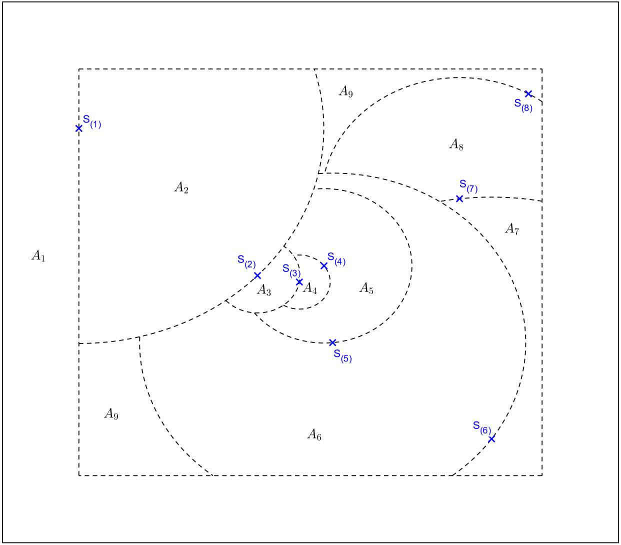
Piste 1 : Parcourir les événements de façon pertinente

- Idée de Dematteï et al. (2007) : parcourir les localisations $s_1, \cdots, s_n$ de proche en proche.

- $$s_{(1)} = \underset{s \in \{s_1, \cdots, s_n\}}{\arg\min} \; d(s, \partial W).$$

$$\forall i \in [\![2, n]\!], \quad s_{(i)} = \underset{s \in \{s_1, \cdots, s_n\}, s \notin \{s_{(1)}, \cdots s_{(i-1)}\}}{\arg\min} d(s, s_{(i)}).$$

Piste 1 : Parcourir les événements de façon pertinente



Piste 1 : Parcourir les événements de façon pertinente

- Définition des aires d'espacement : $A_1, \cdots, A_{n+1}$.

- Propriété de distribution :

$$(A_1, \cdots, A_{n+1}) \sim (E_1, \cdots, E_{n+1})$$

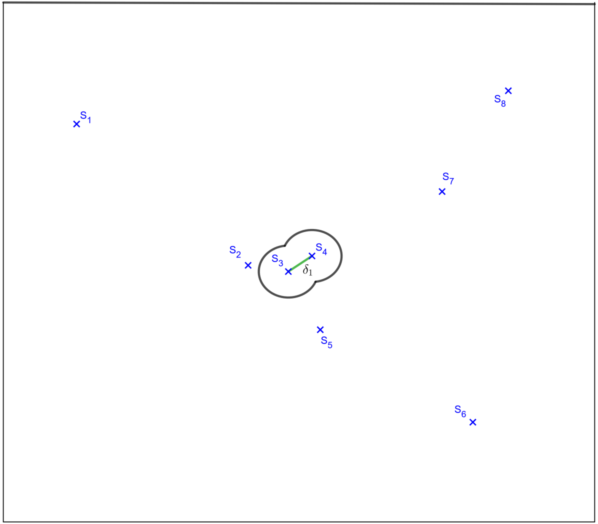
  où $E_1, \cdots, E_{n+1}$: espacements uniformes sur $[0, 1]$.

- On se ramène à de la détection d'agrégat temporel.

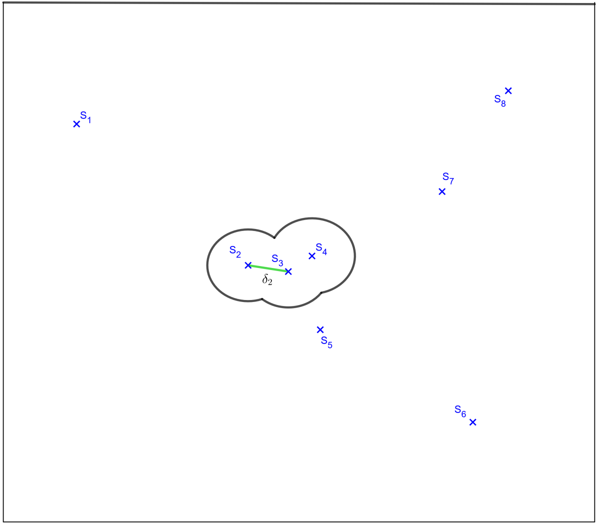
Piste 2 : Des agrégats potentiels basés sur les distances

- Algorithme de Bar-Hen et al. (2007) : création de graphes.
- Soit $\delta > 0$. On définit le graphe $\mathcal{G}(\delta)$:
  - Sommets : $\{1, \cdots, n\}$.
  - Arêtes : $\{(i, j) : d(s_i, s_j) \leq \delta \text{ et } 1 \leq i < j \leq n\}$.
- Agrégats potentiels : composantes connexes de $\mathcal{G}(\delta)$ à mesure que δ augmente.

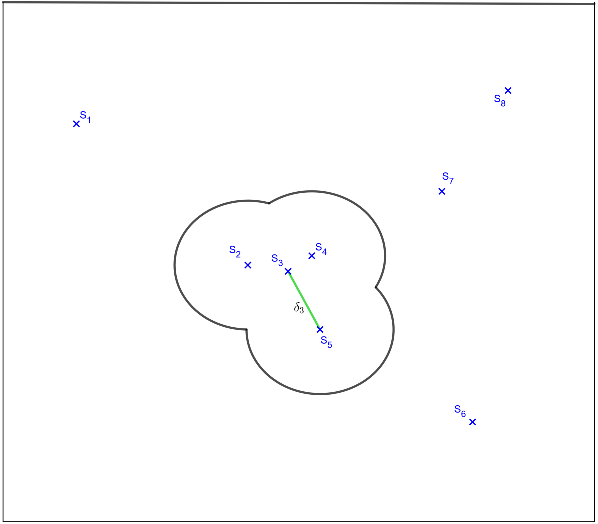
Piste 2 : Des agrégats potentiels basés sur les distances



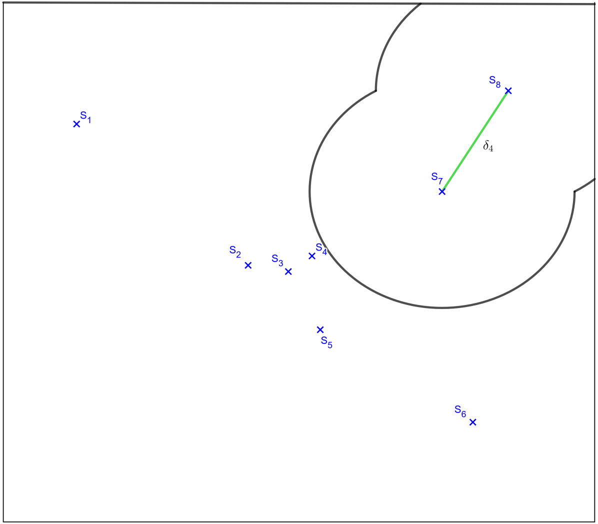
Piste 2 : Des agrégats potentiels basés sur les distances



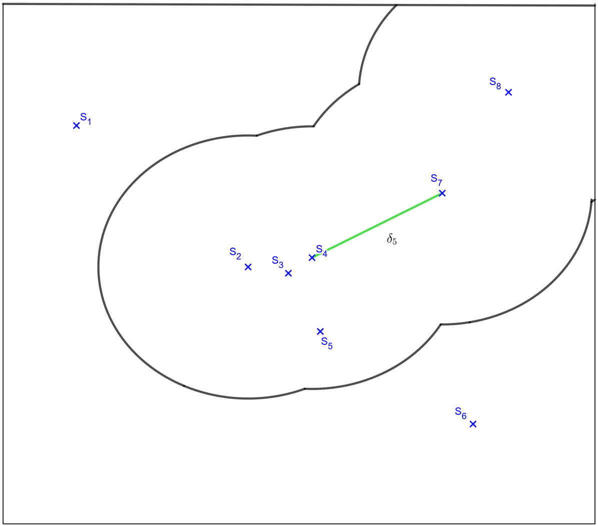
Piste 2 : Des agrégats potentiels basés sur les distances



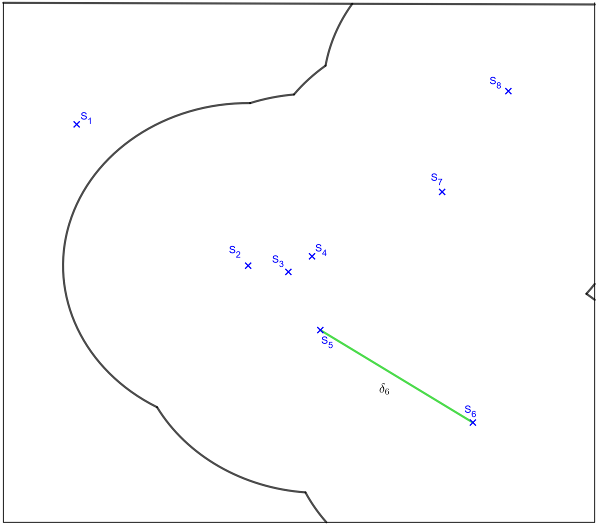
Piste 2 : Des agrégats potentiels basés sur les distances



Piste 2 : Des agrégats potentiels basés sur les distances



Piste 2 : Des agrégats potentiels basés sur les distances



Piste 3 : Une distance spatio-temporelle

- Fenêtre d'observation :

$$W = A \times T.$$

- $D = 2\sqrt{\frac{|A|}{\pi}}$ équivaut à $|T|$.

-

$$d_{ST}\big((y,t),(y',t')\big) = \sqrt{d_S(y,y')^2 + \frac{D^2}{|T|^2} d_T(t,t')^2}.$$

Publications associées

- Cucala, L. (2009). A flexible spatial scan test for case event data. Computational Statistics and Data Analysis, 53, p. 2843-2850.

- Dematteï, C. et Cucala, L. (2011). Multiple spatio-temporal cluster detection for case event data : an ordering-based approach. Communications in Statistics – Theory and Methods, 40, p. 358-372.

- Cucala, L., Dematteï, C., Lopes, P. et Ribeiro, A. (2013). Spatial scan statistics for case event data based on connected components. Computational Statistics, 28, p. 357-369.

Sommaire

Sommaire

Importance de l'indice de concentration

- Statistique de balayage : indice de concentration maximum sur un ensemble d'agrégats potentiels

$$\lambda = \max_{Z \in \mathcal{C}} I(Z)$$

- Quel indice de concentration utiliser ?

Processus non marqué

- Objectif : trouver la zone où la concentration en événements est maximale par rapport à $\mu(.)$.

- Problème : comment comparer Z et Z' si $n(Z) > n(Z')$ et $\mu(Z) > \mu(Z')$?

- Idée de Nagarwalla (1996) et Kulldorff (1997) : utiliser un rapport de vraisemblance généralisé.

Processus non marqué : approche paramétrique

- Modèle paramétrique $\mathcal{M}_0$: absence totale d'agrégat.
- Pour chaque agrégat potentiel $Z \in \mathcal{C}$, un modèle paramétrique $\mathcal{M}_{1,Z}$: présence d'un agrégat dans Z.
- Rapport de vraisemblance entre les deux modèles :

$$RV(Z) = \frac{L^*_{1,Z}}{L^*_0}.$$

- Indice de concentration $I(Z)$ dérivé de $RV(Z)$.
- Statistique de balayage :

$$\lambda = \max_{Z \in \mathcal{C}} I(Z).$$

Processus non marqué : approche paramétrique

- Kulldorff (1997) : modèle Poissonnien.
- Nagarwalla (1996) : approche conditionnelle.
- Rapport de vraisemblance identique :

$$RV(Z) = \frac{L^*_{1,Z}}{L^*_0} = \frac{\left(\frac{n(Z)}{\mu(Z)}\right)^{n(Z)} \left(\frac{n(Z^c)}{\mu(Z^c)}\right)^{n(Z^c)}}{n^n}.$$

- Indice de concentration :

$$I_{RV}(Z) = \log\left(RV(Z)\right)\Big(\mathbb{1}\left(n(Z) > n\mu(Z)\right) - \mathbb{1}\left(n(Z) < n\mu(Z)\right)\Big).$$

Un indice de concentration basé sur les espacements

- Cadre temporel : $W = [0, T]$
- $H_0 : S_1, \cdots, S_n$ i.i.d. $\sim \phi(\cdot)$.
- $T_i = \int_0^{S_i} \phi(s)ds$.
- Statistiques d'ordre associées : $T_{(1)} \leq \cdots \leq T_{(n)}$.
- Sous H_0 :

$$D_{i,j} = T_{(j)} - T_{(i)} \sim \beta(j - i, n + 1 - j + i).$$

- Indice de concentration basé sur les espacements :

$$I_{ES}([T_{(i)}, T_{(j)}]) = 1 - B_{inc}(T_{(j)} - T_{(i)}, j - i, n + 1 - j + i).$$

Un indice de concentration basé sur les espacements

- Extension aux cadres spatial et spatio-temporel.

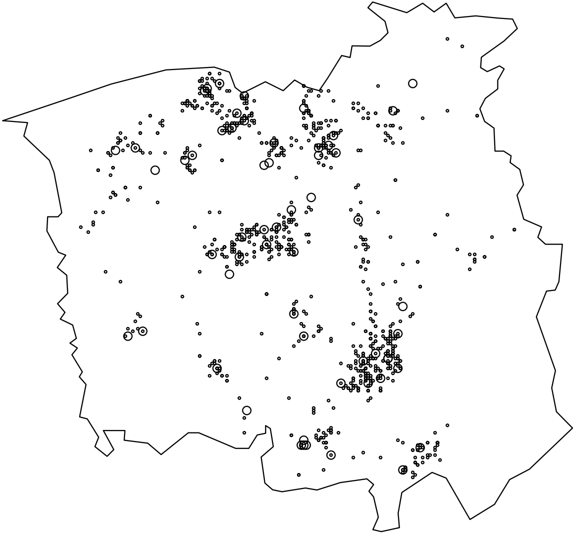
- Indice de concentration basé sur les espacements :

$$I_{ES}(Z) = 1 - B_{inc}\big(\mu(Z), n(Z) - 1, n + 2 - n(Z)\big).$$

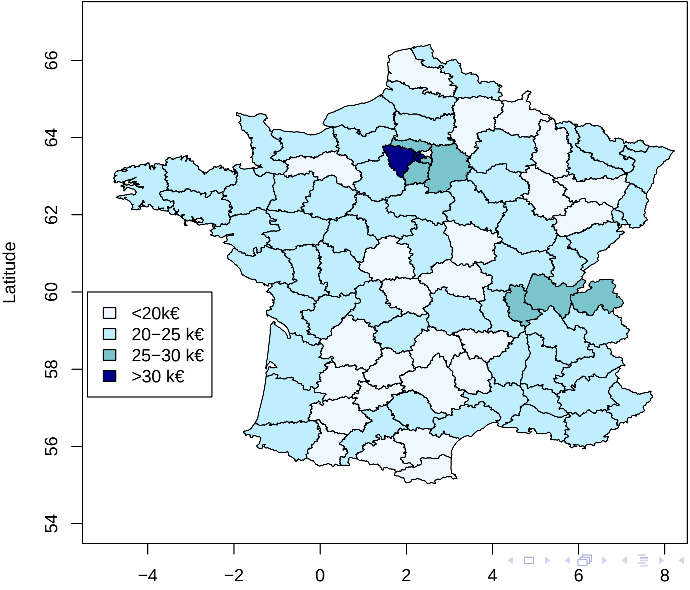
Processus marqué par une variable réelle

- Variable réelle X mesurée en chaque localisation.
- Observations : $(s_1, x_1), \cdots, (s_n, x_n)$.
- Objectif : trouver la zone où la distribution des marques est la plus "anormale".

Processus marqué par une variable réelle



Processus marqué par une variable réelle



Processus marqué : approche paramétrique

- Modèle paramétrique $\mathcal{M}_0$: $X_1, \cdots, X_n$ i.i.d. dans W.
- Pour chaque agrégat potentiel $Z \in \mathcal{C}$, un modèle paramétrique $\mathcal{M}_{1,Z}$: distributions différentes dans Z et Z^c.
- Rapport de vraisemblance entre les deux modèles :

$$RV(Z) = \frac{L^*_{1,Z}}{L^*_0}.$$

- Indice de concentration $I(Z)$ dérivé de $RV(Z)$.
- Statistique de balayage :

$$\lambda = \max_{Z \in \mathcal{C}} I(Z).$$

Processus marqué : approche paramétrique

- Pour une variable binaire : modèle de Bernoulli (Kulldorff, 1997).

- Rapport de vraisemblance proportionnel à

$$\begin{aligned} & x(Z)\log\big(\bar{x}(Z)\big) + \big(n(Z) - x(Z)\big)\log\big(1 - \bar{x}(Z)\big) \\ + \quad & x(Z^c)\log\big(\bar{x}(Z^c)\big) + \big(n(Z^c) - x(Z^c)\big)\log\big(1 - \bar{x}(Z^c)\big). \end{aligned}$$

Processus marqué : approche paramétrique

- Pour une variable continue : modèle Gaussien (Kulldorff et al., 2009).

- Rapport de vraisemblance inversement proportionnel à

$$\frac{n(Z)\left(\overline{x^2}(Z) - (\bar{x}(Z))^2\right) + n(Z^c)\left(\overline{x^2}(Z^c) - (\bar{x}(Z^c))^2\right)}{n}.$$

Processus marqué : approche non-paramétrique

- Approche basée sur les moments.
- Approche basée sur les rangs.
- Hypothèse nulle :

$$H_0 : X_1, \cdots, X_n \text{ i.i.d.}$$

Processus marqué : approche basée sur les moments

- Écart inter-moyennes :

$$D(Z) = \bar{X}(Z) - \bar{X}(Z^c).$$

- Sous H_0, on a :

$$\mathbb{E}_0\big(D(Z)\big) = 0 \text{ et } \mathbb{V}_0\big(D(Z)\big) = \frac{n}{n(Z)n(Z^c)}\sigma^2.$$

- Indice de concentration basé sur les moments :

$$I_M^+(Z) = \frac{\sqrt{n(Z)n(Z^c)}}{\sqrt{n}}\big(\bar{X}(Z) - \bar{X}(Z^c)\big).$$

- Lien avec le test de Student.

Processus marqué : approche basée sur les rangs

- Calcul des rangs (par ordre croissant) des X_i : $R_1, \cdots, R_n$.

- Somme des rangs dans Z : $SR(Z) = \sum_{i=1}^{n} R_i \mathbb{1}_Z(s_i)$.

- Sous H_0, on a :

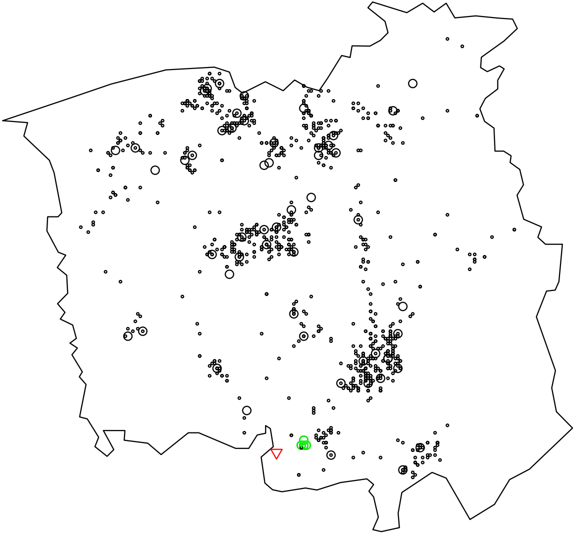
$$\mathbb{E}_0(SR(Z)) = \frac{n(Z)(n+1)}{2} \text{ et } \mathbb{V}_0(SR(Z)) = \frac{n(Z)n(Z^c)(n+1)}{12}.$$

- Indice de concentration basé sur les rangs :

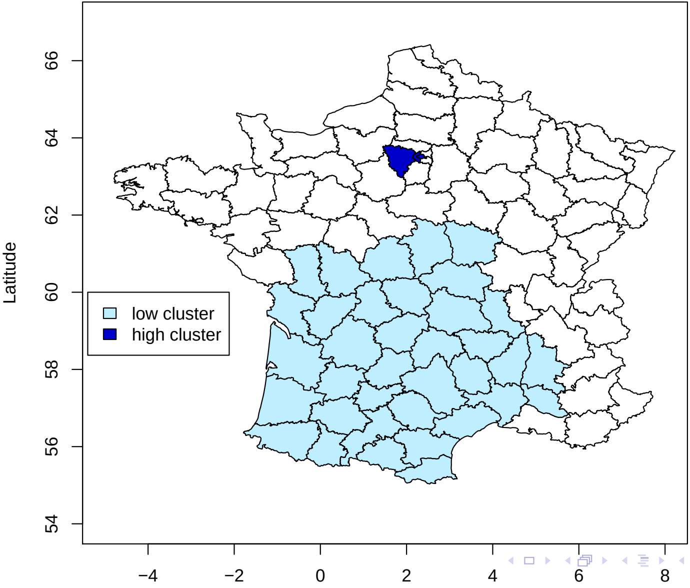
$$I_R^+(Z) = \frac{SR(Z) - \mathbb{E}_0(SR(Z))}{\sqrt{\mathbb{V}_0(SR(Z))}}.$$

- Lien avec le test de Wilcoxon-Mann-Whitney.

Processus marqué par une variable réelle



Processus marqué par une variable réelle



Publications associées

- Cucala, L. (2008). A hypothesis-free multiple scan statistic with variable window. Biometrical Journal, 50, p. 299-310.
- Cucala, L. (2014). A distribution-free spatial scan statistic for marked point processes. Spatial Statistics, 10, p. 117-125.
- Cucala, L. (2016). A Mann-Whitney scan statistic for continuous data. Communications in Statistics – Theory and Methods, 45, p. 321-329.
- Cucala, L. (2016). Scan statistics for detecting high-variance clusters. Journal of Probability and Statistics, Article ID 7591680.
- Cucala, L. et Dematteï, C. (2016). Spatial cluster detection for socio-economic data. CSBIGS, 6, p. 1-9.
- Cucala L. (2017). Variable Window Scan Statistics : Alternatives to Generalized Likelihood Ratio Tests. In : Glaz J., Koutras M. (eds) Handbook of Scan Statistics. Springer.

Sommaire

Sommaire

Processus marqué par plusieurs variables

- Variables réelles $X^1, \cdots, X^p$ mesurées en chaque localisation.

- Observations : $(s_1, x_1), \cdots, (s_n, x_n)$
  où $x_i = (x_i^1, \cdots, x_i^p) \in \mathbb{R}^p$.

- Objectif : trouver la zone où la distribution des marques est la plus "anormale".

Processus marqué par plusieurs variables

- Méthode de Kulldorff et al. (2007) : combinaison des indices de concentration associés à chaque variable.

- Inconvénient : ne prend pas en compte les corrélations entre variables.

- Prise en compte des corrélations :
  - approche paramétrique.
  - approche non-paramétrique.

Marques multivariées : approche paramétrique

- Modèle Gaussien multivarié :

$$\mathcal{M}_0 : X_1, \cdots, X_n \text{ i.i.d. } \sim \mathcal{N}_p(m, \Sigma)$$

$$\mathcal{M}_{1,Z} : X_1, \cdots, X_n \text{ indép. et } \begin{cases} X_i \sim \mathcal{N}_p(m_Z, \Sigma_{Z,Z^c}) & \text{si } s_i \in Z, \\ X_i \sim \mathcal{N}_p(m_{Z^c}, \Sigma_{Z,Z^c}) & \text{si } s_i \in Z^c. \end{cases}$$

- Rapport de vraisemblance inversement proportionnel à

$$|\Sigma^*_{Z,Z^c}|$$

où $\Sigma^*_{Z,Z^c} = \dfrac{n(Z)S(Z) + n(Z^c)S(Z^c)}{n}$.

- Lien avec le test de Hotelling.

Marques multivariées : approche non-paramétrique

- Extension multivariée du test de WMW proposée par Oja and Randles (2004).

- Fonction signe multivariée :

$$\forall x \in \mathbb{R}^p, \quad S(x) = \begin{cases} ||x||^{-1}x & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

- Rang multivarié associé à X_i :

$$R_i = \frac{1}{n}\sum_{k=1}^{n} S_{i,k}$$

où

$$S_{i,k} = S\big(A_x(X_i - X_k)\big)$$

et A_x est la matrice de Tyler.

Marques multivariées : approche non-paramétrique

- Statistique de test d'égalité entre Z et Z^c :

$$U^2_{Z|Z^c} = \frac{p}{c_x^2}[n(Z)||\bar{R}_Z||^2 + n(Z^c)||\bar{R}_{Z^c}||^2$$

où

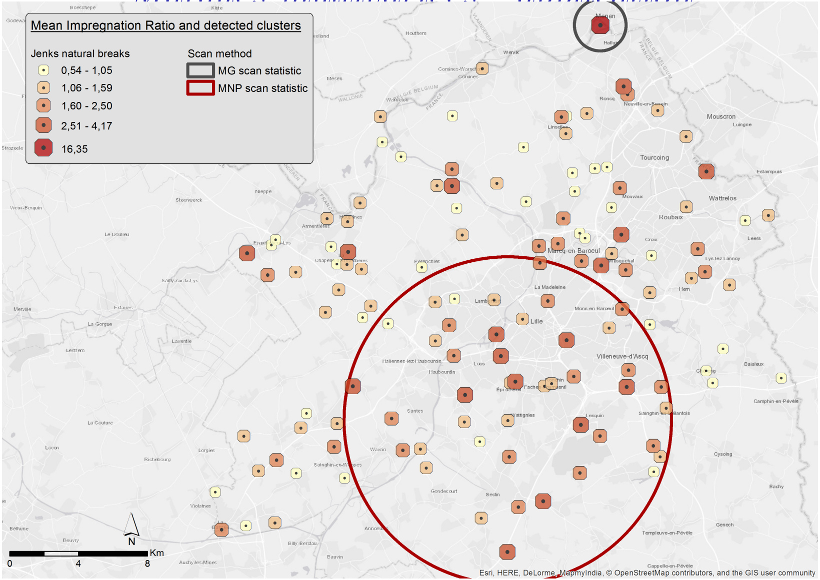
$$\begin{cases} \bar{R}_Z & = & \dfrac{1}{n(Z)} \displaystyle\sum_{i=1}^{n} R_i \mathbb{1}_Z(s_i), \\ c_x^2 & = & \displaystyle\sum_{i=1}^{n} R_i^T R_i. \end{cases}$$

- Si distributions dans Z et Z^c identiques : $U^2_{Z|Z^c} \to \chi^2_p$.

- Indice de concentration multivarié non-paramétrique :

$$I_{MNP}(Z) = U^2_{Z|Z^c}.$$

Marques multivariées : application



Processus marqué par une variable fonctionnelle

- Variable fonctionnelle X mesurée en chaque localisation.
- Observations : $(s_1, x_1), \cdots, (s_n, x_n)$
  où $x_i \in \chi$, espace fonctionnel.
- Objectif : trouver la zone où la distribution des marques est la plus "anormale".

Marques fonctionnelles : approche non-paramétrique

- Extension fonctionnelle du test de WMW proposée par Chakraborty et Chaudhuri (2015).

- Statistique de test :

$$T_{\mathrm{WMW}}(Z) = \frac{1}{n(Z)n(Z^c)} \sum_{\{i:s_i \in Z\}} \sum_{\{j:s_j \in Z^c\}} \frac{X_j - X_i}{\|X_j - X_i\|_\chi}.$$

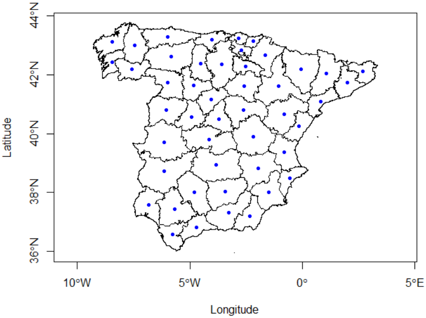
- Si distributions dans Z et Z^c identiques :

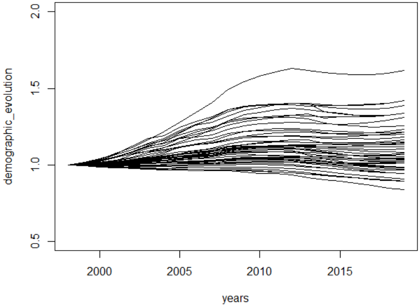
$$\tilde{T}_{\mathrm{WMW}}(Z) = \sqrt{\frac{n(Z)n(Z^c)}{n}}\, T_{\mathrm{WMW}}(Z) \text{ cv faiblement vers } G(0,\Gamma).$$

- Indice de concentration fonctionnel non-paramétrique :

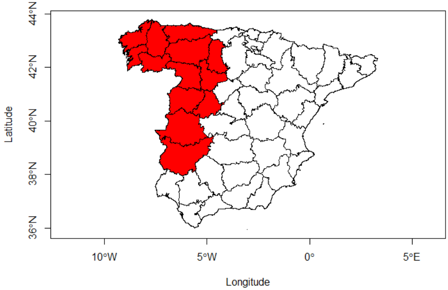
$$I_{FNP}(Z) = \|\tilde{T}_{\mathrm{WMW}}(Z)\|_\chi.$$

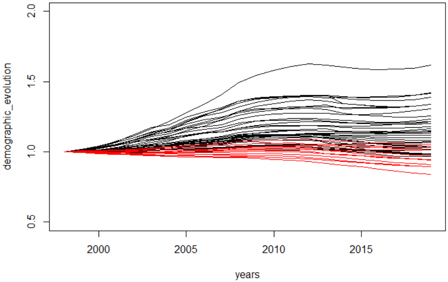
Marques fonctionnelles : application





Marques fonctionnelles : application





Publications associées

- Cucala, L., Genin, M., Lanier, C. et Occelli, F. (2017). A multivariate Gaussian scan statistic for spatial data. Spatial Statistics, 21, p. 66-74.
- Cucala, L., Genin, M., Occelli, F. et Soula, J. (2019). A multivariate nonparametric scan statistic for spatial data. Spatial Statistics, 29, p. 1-14.
- Smida, Z. Cucala, L., Durif, G. et Gannoun, A. (2021). A Wilcoxon-Mann-Whitney spatial scan statistic for functional data. Computational Statistics and Data Analysis, 167, 107378.

Sommaire

Sommaire

Processus marqué par une variable fonctionnelle

➡ Existence de deux statistiques de balayage pour données fonctionnelles :

- une basée sur le test WMW (Smida et al., 2021).
- une basée sur un test de type ANOVA (Frévent et al., 2021).

➡ Possibilités d'indices de concentration basés sur d'autres tests :

- le test de Student fonctionnel de Horvath et al. (2013).
- le test de la médiane de Smida et al. (2022).

➡ Enrichissement du package R HDSpatialScan.

Prise en compte de l'autocorrélation spatiale

- Les méthodes de balayage reposent sur une hypothèse d'indépendance .
- Hypothèse peu réaliste mais simplifie les calculs.
- Prise en compte de l'autocorrélation spatiale pour processus marqué par une variable continue : utilisation du modèle SAR.
- Difficultés pour estimer le paramètre d'autocorrélation spatiale.
- Alternative possible : le modèle SEM (Spatial Error Model).

Méthodes de balayage "global"

- Objectif : comparer les distributions de deux processus non marqués.

- Les statistiques de balayage classiques pointent la plus grande disparité entre les deux processus.

- Idée : utilisation d'un indice de concentration pour introduire une statistique de balayage "global".

- Première tentative : remplacer le maximum de l'indice de concentration par sa variance empirique.

Méthodes de balayage "double"

- Cadre : Processus marqué par une variable continue.

- Objectif : identifier la zone où les événements sont plus nombreux et les marques plus grandes.

- H_0 : localisation issues d'un processus de Poisson et marques i.i.d.

- Deux approches possibles :
  • paramétrique : indice de concentration basé sur un rapport de vraisemblance.
  • non-paramétrique : indice de concentration basé sur un test d'égalité entre Z et Z^c.

Publications associées

➡ Frévent, C., Ahmed, M.S., Soula, J., Smida, Z., Cucala, L., Dabo-Niang, S. et Genin, M. (2021). HDSpatialScan : Multivariate and Functional Spatial Scan Statistics.

➡ Ahmed, M.S. Cucala, L. et Genin, M. (2021). Spatial autoregressive models for scan statistic. Journal of Spatial Econometrics.

➡ Bonneu, F. et Cucala, L. (2021). Global scan methods for comparing two spatial point processes. In : Daouia A., Ruiz-Gazen A. (eds) Advances in Contemporary Statistics and Econometrics. Springer, Cham.

➡ Cucala, L. (2022). A double scan statistic for continuous data.