Tests non paramétriques pour des processus de Poisson : Études sur la représentativité spatiale de services

J. Bessac(*), F. Coquet(*)(**), J.-M. Floch(***), M. Fromont(*)(**)

(*) IRMAR, (**) CREST-Ensai, (***) INSEE-DDAR

Paris, 25 janvier 2012

Introduction

- **Étude de données économiques sur la répartition des services dans la ville de Rennes**
  - Comparaison des répartitions de deux services différents
  - Représentativité d'un service par rapport aux logements
- **Le point de vue probabiliste**
  - Modélisation de la distribution d'un service par un processus de Poisson spatial (Himpens, 2008; Bessac, 2009)
  - Modélisation de la distribution des logements par un processus de Poisson spatial ?
- **Le point de vue statistique**
  - Comparaison des lois de deux processus de Poisson spatiaux

Processus de Poisson
Définition

Définition (Møller et Waagepetersen)

Un processus de Poisson N d'intensité s par rapport à la mesure de Lebesgue ν sur $\mathbb{X} \subset \mathbb{R}^2$ est un ensemble de points aléatoires $\{Z_1, \ldots, Z_{|N|}\}$ tels que :

- *Pour tout $B \subset \mathbb{X}$ t. q. $\int_B s(x)d\nu_x < +\infty$, le nombre de points de N tombant dans B, noté $N(B)$, suit une loi de Poisson de paramètre $\int_B s(x)d\nu_x$*
- *Conditionnellement à "$N(B) = n$", $N \cap B$ suit la même loi qu'un n-échantillon de la loi de densité $s/\int_B s(x)d\nu_x$ par rapport à ν*

$\hookrightarrow$ Interprétation "pratique", exemples

Processus de Poisson
Propriétés

Définition

Le processus de Poisson N est dit homogène sur $\mathbb{X}$ si son intensité s par rapport à ν est constante sur $\mathbb{X}$.

Remarque : le processus de Poisson N est homogène si et seulement si conditionnellement à "$|N| = n$", N suit la même loi qu'un n-échantillon de la loi uniforme sur $\mathbb{X}$.

Proposition

Si $B_1, \cdots, B_k$ sont des sous-ensembles disjoints de $\mathbb{X}$, alors $N \cap B_1, \cdots, N \cap B_k$ sont indépendants.

$\hookrightarrow$ Champs d'applications nombreux : fiabilité, étude de trafics, biologie, physique... économie !

Problèmes de type "two-sample"

Deux types de problèmes

Soit $\mathbb{X} \subset \mathbb{R}^2$, N_1 et N_2 deux processus de Poisson observés sur $\mathbb{X}$, d'intensités s_1 et s_2 par rapport à ν, contenant $|N_1|$ et $|N_2|$ points. On note $(X_1, \ldots, X_{|N_1|})$ et $(Y_1, \ldots, Y_{|N_2|})$ les points de N_1 et N_2.

- **Test de proportionnalité :**

$(\mathbf{H_0^p})$ "s_1 et s_2 proportionnelles" contre $(\mathbf{H_1^p})$ "s_1 et s_2 non prop."

$\hookrightarrow$ Tests de (H_0^p) contre "s_1 et s_2 de rapport croissant" par Bovett, Saw (80) et Deshpande, Mukhopadhyay et Naik-Nimbalkar (99)

$\hookrightarrow$ Tests "conditionnels"

Problèmes de type "two-sample"

Deux types de problèmes

- **Test d'égalité :**

$$(\mathbf{H_0})\ "s_1 = s_2"\ \text{contre}\ (\mathbf{H_1})\ "s_1 \neq s_2"$$

$\hookrightarrow$ Tests de proportionnalité... trop conservatifs !

$\hookrightarrow$ Tests de Fromont, Laurent, Reynaud-Bouret (2012)

Problèmes de type "two-sample"

Tests conditionnels

$(\mathbf{H_0^p})$ "s_1 et s_2 proportionnelles" contre $(\mathbf{H_1^p})$ "non prop."

$\Updownarrow$

$(\mathbf{H_0^p})$ "$\tilde{s}_1 = \tilde{s}_2$" contre $(\mathbf{H_1^p})$ "$\tilde{s}_1 = \tilde{s}_2$",
avec $\tilde{s}_1 = s_1 / \int_{\mathbb{X}} s_1(x) d\nu_x$ et $\tilde{s}_2 = s_2 / \int_{\mathbb{X}} s_2(x) d\nu_x$.

Conditionnellement à "$|N_1| = n_1$ et $|N_2| = n_2$", $(X_1, \ldots, X_{|N_1|})$ et $(Y_1, \ldots, Y_{|N_2|})$ suivent les mêmes lois que des n_1 et n_2 échantillons des lois de densités $\tilde{s}_1$ et $\tilde{s}_2$ par rapport à ν indépendants.

$\hookrightarrow$ Tests de type Kolmogorov-Smirnov, Cramer von Mises...

$\hookrightarrow$ Tests de Baringhaus et Franz (2004), Gretton et al. (2008).

Problèmes de type "two-sample"

Tests conditionnels

Test de Baringhaus et Franz (2004) ou test de Cramer

- Statistique de test :

$$T_{Cramer} = \frac{n_1 n_2}{n_1+n_2}\left(\frac{1}{n_1 n_2}\sum_{i,j=1}^{n_1,n_2}\|X_i - Y_j\| - \frac{1}{2n_1^2}\sum_{i,k=1}^{n_1}\|X_i - X_k\| - \frac{1}{2n_2^2}\sum_{j,k=1}^{n_2}\|Y_j - Y_k\|\right)$$

- Valeur critique :

$c^*_{Cramer}(1-\alpha) = (1-\alpha)$ quantile d'une version bootstrap de T_{Cramer} sous (H_0^p) sachant $Z = (X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$

Φ_{Cramer} rejette (H_0^p) si $T_{Cramer} > c^*_{Cramer}(1-\alpha)$

Problèmes de type "two-sample"

Tests conditionnels

Test de Bahr (1996)

- Statistique de test :

$$T_{Bahr} = \frac{n_1 n_2}{n_1+n_2}\left(\frac{2}{n_1 n_2}\sum_{i,j=1}^{n_1,n_2}\varphi\left(\|X_i - Y_j\|^2\right) - \frac{1}{n_1^2}\sum_{i,k=1}^{n_1}\varphi\left(\|X_i - X_k\|^2\right) - \frac{1}{n_2^2}\sum_{j,k=1}^{n_2}\varphi\left(\|Y_j - Y_k\|^2\right)\right)$$

avec $\varphi(t) = 1 - exp(-t/2)$.

- Valeur critique :

$c^*_{Bahr}(1-\alpha) = (1-\alpha)$ quantile d'une version bootstrap de T_{Bahr} sous (H_0^p) sachant Z

Φ_{Bahr} rejette (H_0^p) si $T_{Bahr} > c^*_{Bahr}(1-\alpha)$

Problèmes de type "two-sample"

Tests conditionnels

Test de Gretton et al. (2008)

- Statistique de test :

$$T_{KMMD} = \left(\frac{1}{n_1^2} \sum_{i,k=1}^{n_1} K(X_i, X_k) + \frac{1}{n_2^2} \sum_{j,k=1}^{n_2} K(Y_j, Y_k) - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K(X_i, Y_j) \right)^{1/2}$$

où K est un noyau dit de Mercer ou d'apprentissage, défini positif.

- Valeur critique :

$c^*_{KMMD}(1-\alpha) = (1-\alpha)$ quantile d'une version bootstrap de T_{KKMD} sous (H_0^p) sachant Z, ou majorant du vrai quantile.

$$\Phi_{KMMD} \text{ rejette } (H_0^p) \text{ si } T_{KMMD} > c^*_{KMMD}(1-\alpha)$$

$K(z, z') = exp(-\|z - z'\|^2/\sigma^2)$, $\sigma = \sqrt{2}$: Test de Bahr ?

Problèmes de type "two-sample"

Tests d'égalité

Test "simple" de Fromont, Laurent, Reynaud-Bouret (2012)

On suppose : $s_1, s_2 \in \mathbb{L}^\infty(\mathbb{X}) \cap \mathbb{L}^1(\mathbb{X}, d\nu) \subset \mathbb{L}^2(\mathbb{X}, d\nu)$.

- Statistique de test :

$\hat{T}_K = \sum_{i \neq i'=1}^{|N|} K(Z_i, Z_{i'}) \varepsilon_i^0 \varepsilon_{i'}^0$, où

  - $N = (Z_1, \ldots, Z_{|N|})$ est le processus agrégeant N_1 et N_2,
  - $\varepsilon_i^0 = 1$ si $Z_i \in N_1$ et $\varepsilon_i^0 = -1$ si $Z_i \in N_2$.

$\hookrightarrow$ Si $K(z, z') = \sum_{\lambda \in \Lambda} \varphi_\lambda(z) \varphi_\lambda(z')$, où $\{\varphi_\lambda, \lambda \in \Lambda\}$ est une famille orthonormée de $\mathbb{L}^2(\mathbb{X}, d\nu)$, $\hat{T}_K$ estime sans biais $\|\Pi_{Vect(\varphi_\lambda, \lambda \in \Lambda)}(s_1 - s_2)\|_{2,\nu}^2$.

Problèmes de type "two-sample"
Tests d'égalité

- Choix de noyau possibles :

1. Noyau basé sur une famille orthonormée de $\mathbb{L}^2(\mathbb{X}, d\nu)$: $K(z, z') = \sum_{\lambda \in \Lambda} \varphi_\lambda(z)\varphi_\lambda(z')$,
2. Noyau basé sur un noyau d'approximation k de $\mathbb{L}^2(\mathbb{R}^2)$, t. q. $k(-z) = k(z)$: $K(z, z') = \frac{1}{h^2} k\left(\frac{z-z'}{h}\right)$,
3. Noyau de Mercer (c.f. Gretton et al.) t. q.

$$K(z, z') = \langle \psi(z), \psi(z') \rangle_{\mathcal{H}_K},$$

où ψ et $\mathcal{H}_K$ sont une fonction de représentation et un RKHS associés à K.

Problèmes de type "two-sample"
Tests d'égalité

- Valeur critique :

$q^{(N)}_{K,1-\alpha} = (1-\alpha)$ quantile de $\hat{T}^{\varepsilon}_K$ conditionnellement à N, où $\hat{T}^{\varepsilon}_K$ est une version bootstrap "sauvage" de $\hat{T}_K$ sous (H_0) :
$\hat{T}^{\varepsilon}_K = \sum_{i \neq i'=1}^{|N|} K(Z_i, Z_{i'})\varepsilon_i\varepsilon_{i'}$, où $(\varepsilon_i)_{i\in\mathbb{N}}$ variables de Rademacher i.i.d. indépendantes de N.

On rejette (H_0) si $\hat{T}_K > q^{(N)}_{K,1-\alpha}$

Problèmes de type "two-sample"

Tests d'égalité

Test "agrégé" de Fromont, Laurent, Reynaud-Bouret (2012)

Choix d'un seul noyau K ? Non... $\longrightarrow$ Collection $\{K_m, m \in \mathcal{M}\}$.

$$\Phi_{Agg} \text{ rejette } (H_0) \text{ s'}\exists m \in \mathcal{M}, \text{ t. q. } \hat{T}_{K_m} > q^{(N)}_{K_m, 1-u^{(N)}_\alpha e^{-w_m}},$$

- $\{w_m, m \in \mathcal{M}\}$ réels positifs t. q. $\sum_{m\in\mathcal{M}} e^{-w_m} \leq 1$
- $u^{(N)}_\alpha =$
  $\sup\left\{u > 0, \mathbb{P}\left(\sup_{m\in\mathcal{M}}(\hat{T}^{\varepsilon}_{K_m} - q^{(N)}_{m,1-ue^{-w_m}}) > 0 \;\middle|\; N\right) \leq \alpha\right\}$

Propriétés non asymptotiques :

- Φ_{Agg} est de niveau exactement α,
- Φ_{Agg} vérifie une inégalité de type oracle,
- Φ_{Agg} est adaptatif au sens du minimax.

Problèmes de type "two-sample"

Étude de simulation

Estimation des niveaux et puissances des tests

- Alternatives :

$$\begin{aligned} f_{a,\varepsilon}(z) &= \mathbb{1}_{(0,1)^2}(z) + \varepsilon \mathbb{1}_{(0,a)^2}(z) - \varepsilon \mathbb{1}_{(a,2a)^2}(z), \\ g_{\sigma,\mu}(z) &= (2\pi\sigma^2)^{-1} \exp\left(-\|z-\mu\|^2/(2\sigma^2)\right). \end{aligned}$$

Simulation de deux processus de Poisson d'intensités respectives :

- $200 f_{0,0}$ et $200 f_{a,\varepsilon}$,
- $200 g_{0.15,0.5}$ et $200 g_{0.15,\mu}$.

Niveau choisi : $\alpha = 0.05$.

Logiciels : R pour Φ_{Cramer} et Φ_{Bahr}, Matlab pour Φ_{KMMD} et Φ_{Agg} (noyau gaussien et Epanechnikov).

Problèmes de type "two-sample"

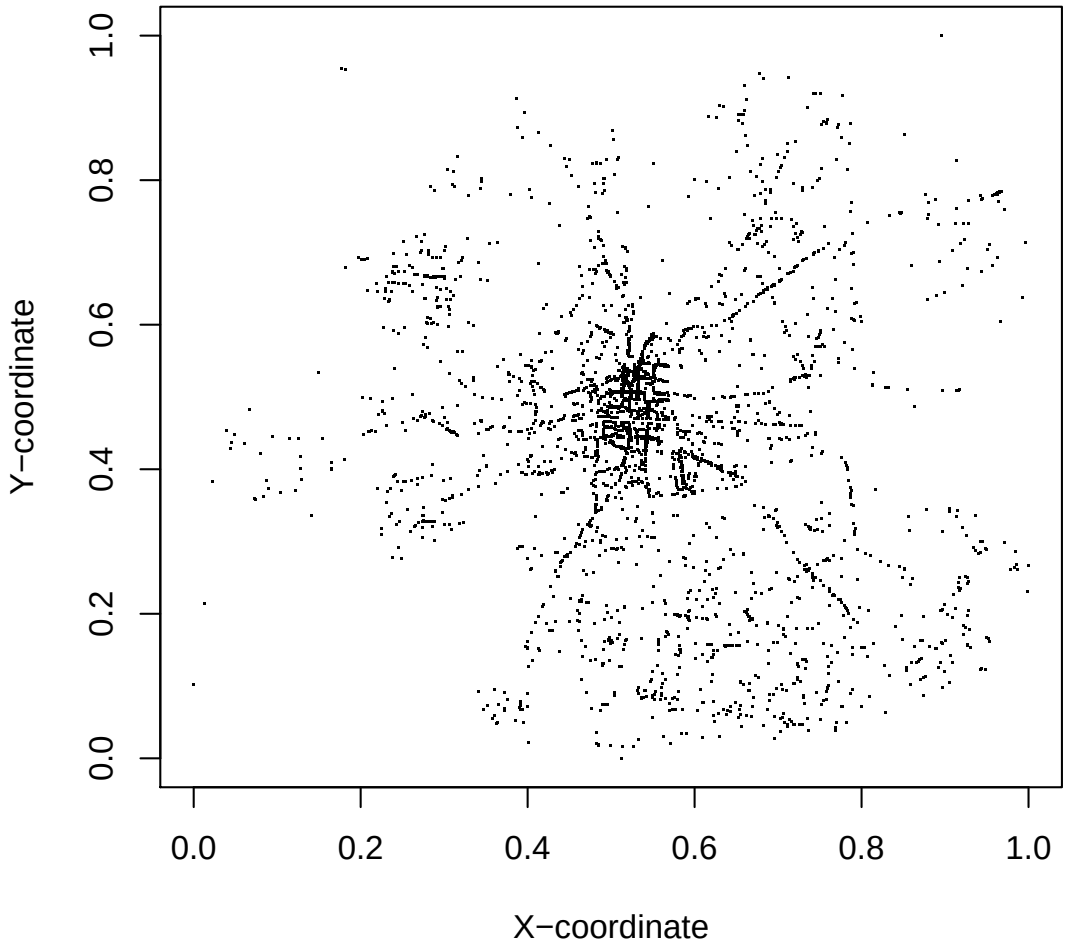
Étude de simulation

- Résultats :

Densités	Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}	$\Phi_{Agg,G}$	$\Phi_{Agg,E}$
$(f_{0,0}, f_{0,0})$	0.052	0.052	0.06	0.0485	0.046
$(f_{0,0}, f_{0.25,0.8})$	0.10	0.09	0.15	0.17	0.18
$(f_{0,0}, f_{0.25,0.9})$	0.10	0.09	0.18	0.23	0.21
$(f_{0,0}, f_{0.25,1})$	0.14	0.11	0.21	0.26	0.25
$(g_{0.15,0.5}, g_{0.15,0.5})$	0.048	0.046	0.043	0.0485	0.046
$(g_{0.15,0.5}, g_{0.15,0.52})$	0.36	0.37	0.26	0.21	0.18
$(g_{0.15,0.5}, g_{0.15,0.54})$	0.90	0.91	0.83	0.69	0.66

Études sur la répartition des services

Représentation des services dans Rennes



Études sur la répartition des services

Comparaison des répartitions de services

On modélise les représentations de deux services différents par des processus de Poisson indépendants.

Pour chaque couple de services, on met en œuvre les tests suivants.

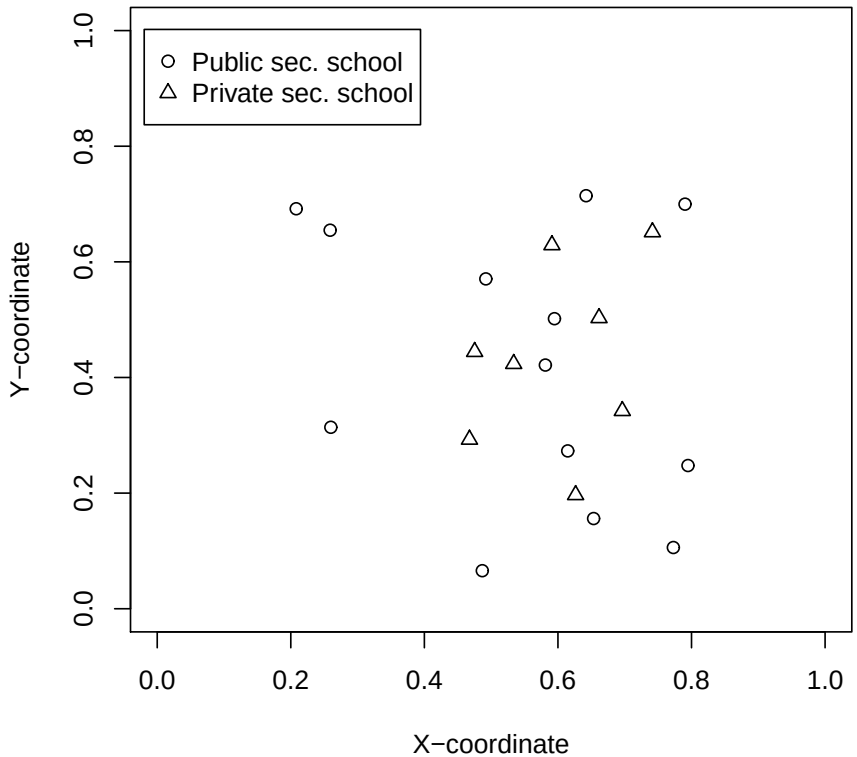
- Tests Φ_{Cramer}, Φ_{Bahr}, Φ_{KMMD} : le rejet de (H_0^p) est codé 1, l'acceptation 0 (p-value).
- Tests $\Phi_{Agg,G}$ et $\Phi_{Agg,E}$: le rejet de (H_0) est codé 1, l'acceptation 0.

Niveau choisi : $\alpha = 0.05$.

Études sur la répartition des services

Comparaison des répartitions de services

Collèges publics et privés

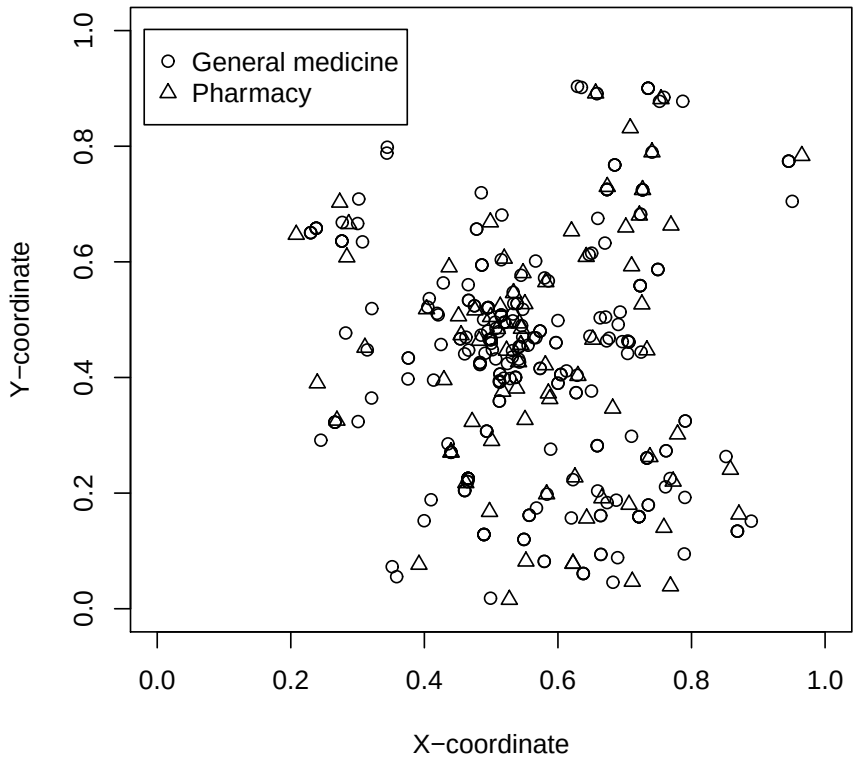


Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}	$\Phi_{Agg,G}$	$\Phi_{Agg,E}$
0(0.60)	0(0.75)	0	0	0

Études sur la répartition des services

Comparaison des répartitions de services

Médecins généralistes et pharmacies

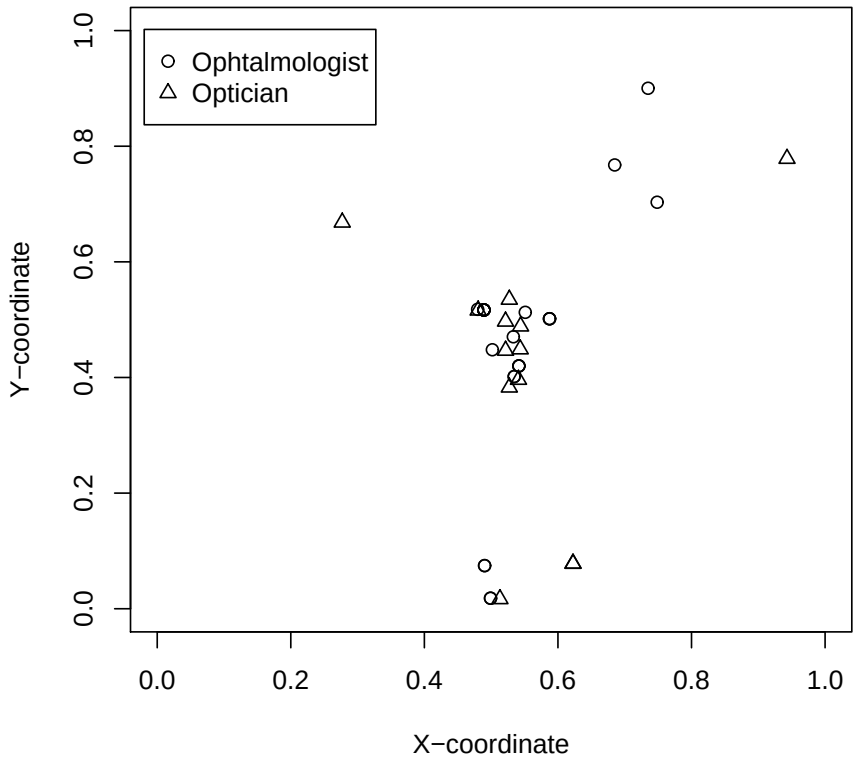


Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}	$\Phi_{Agg,G}$	$\Phi_{Agg,E}$
0(0.99)	0(0.89)	0	1	1

Études sur la répartition des services

Comparaison des répartitions de services

Ophtalmologistes et opticiens

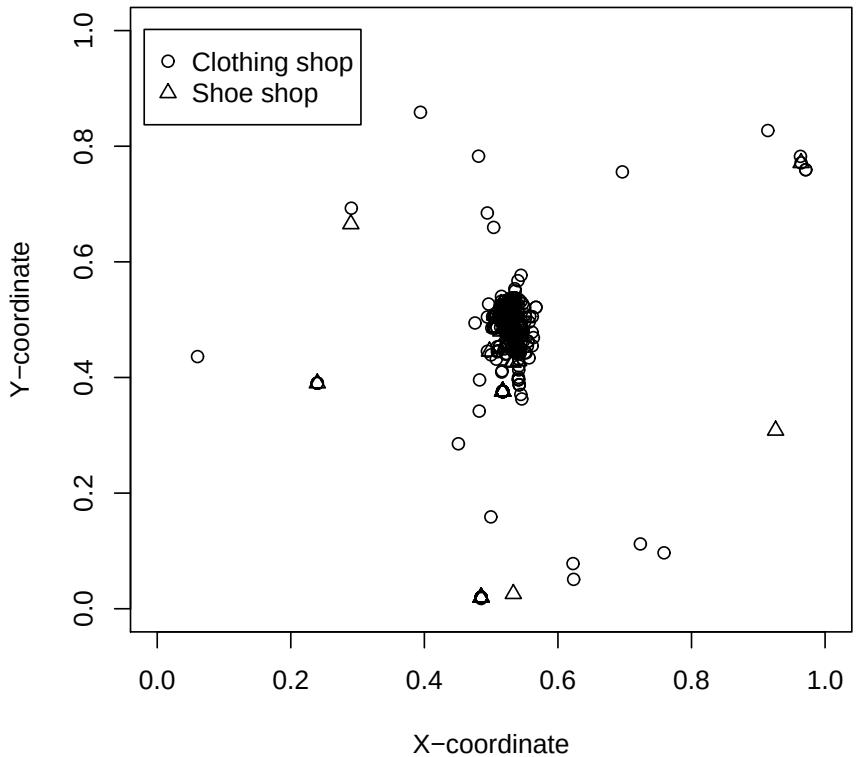


Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}	$\Phi_{Agg,G}$	$\Phi_{Agg,E}$
0(0.60)	0(0.71)	0	0	0

Études sur la répartition des services

Comparaison des répartitions de services

Magasins de vêtements et de chaussures

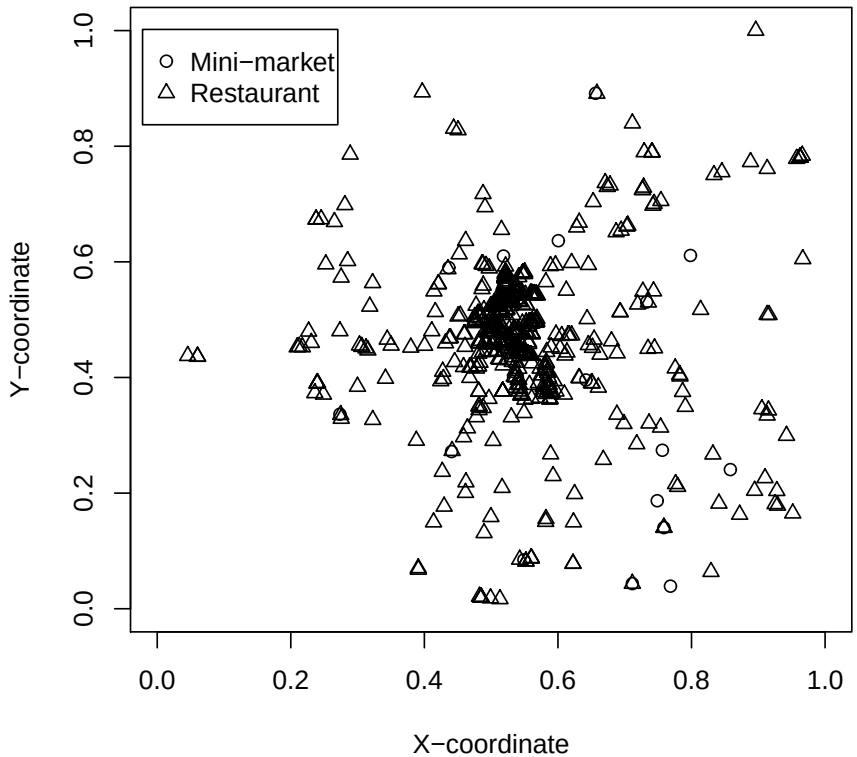


Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}	$\Phi_{Agg,G}$	$\Phi_{Agg,E}$
0(0.39)	0(0.31)	0	1	1

Études sur la répartition des services

Comparaison des répartitions de services

Supérettes et restaurants

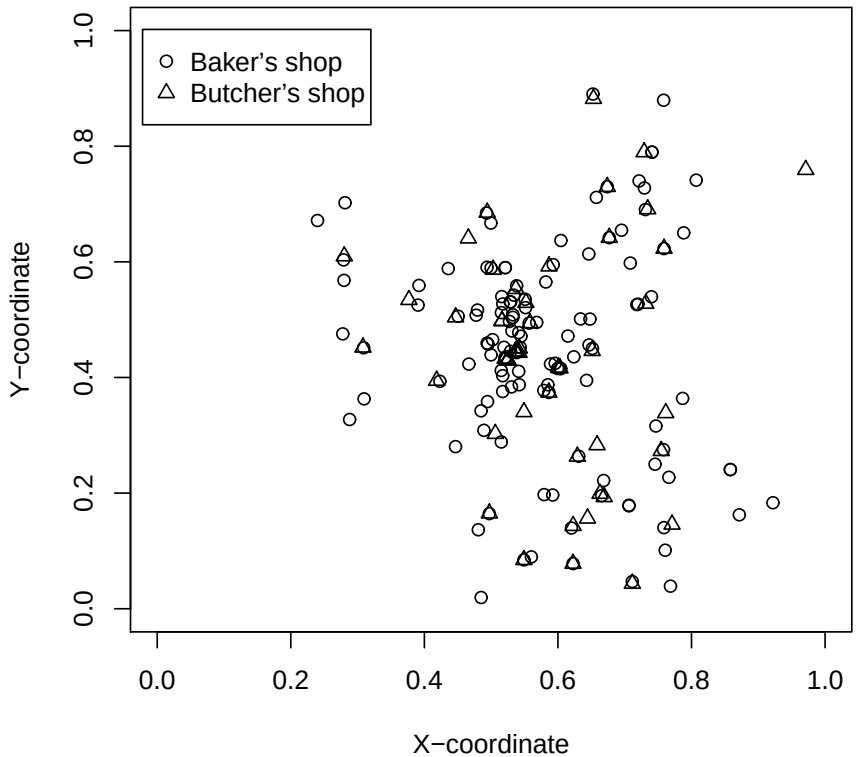


Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}	$\Phi_{Agg,G}$	$\Phi_{Agg,E}$
1(0)	1(0)	1	1	1

Études sur la répartition des services

Comparaison des répartitions de services

Boulangeries et boucheries



Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}	$\Phi_{Agg,G}$	$\Phi_{Agg,E}$
0(0.99)	0(0.95)	0	1	1

Études sur la répartition des services

Représentativité des services / logements

On modélise les représentations d'un service et des logements dans une zone géographique restreinte du centre ville par deux processus de Poisson indépendants.

Pour chaque service, on met en œuvre les trois tests de proportionnalité conditionnels.

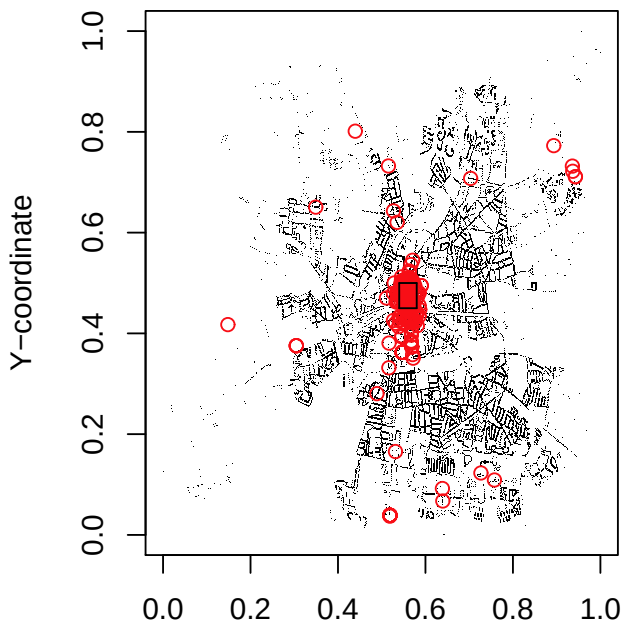
- Tests Φ_{Cramer}, Φ_{Bahr}, Φ_{KMMD} : le rejet de (H_0^p) est codé 1, l'acceptation 0 (p-value).

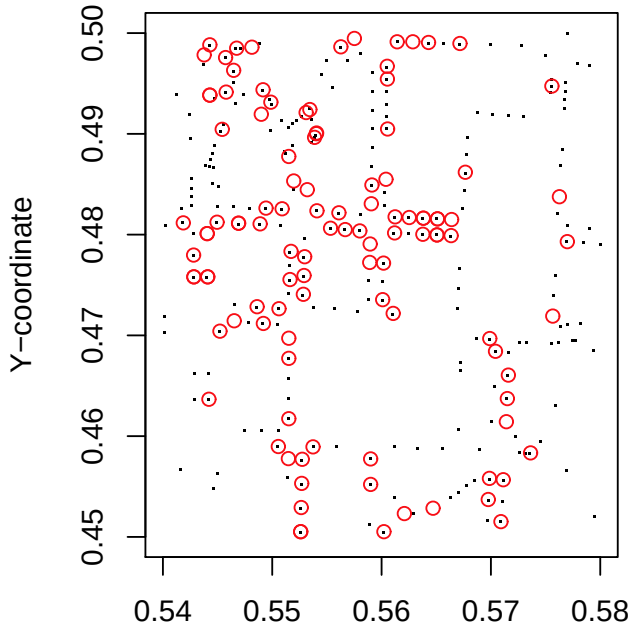
Niveau choisi : $\alpha = 0.05$.

Études sur la répartition des services

Représentativité des services / logements

Magasins de vêtements



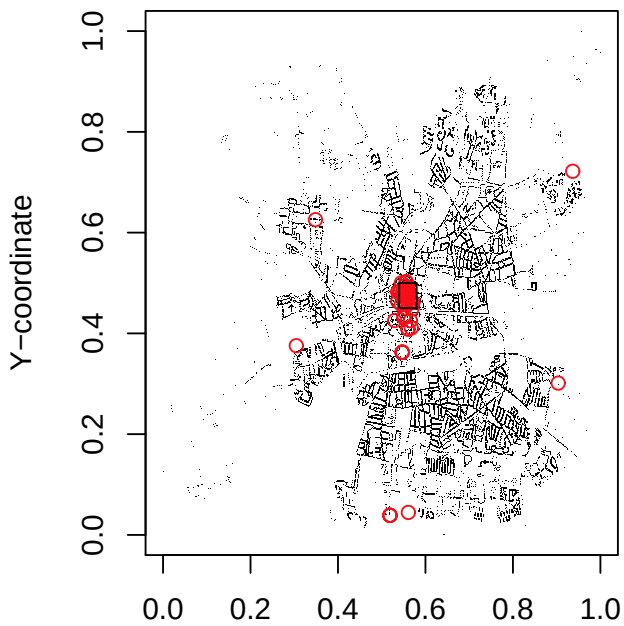


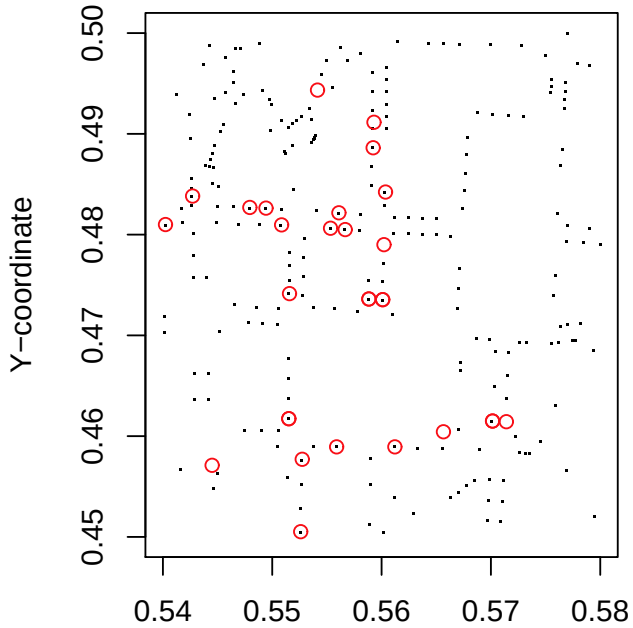
Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}
0(0.16)	0(0.20)	0

Études sur la répartition des services

Représentativité des services / logements

Magasins de chaussures



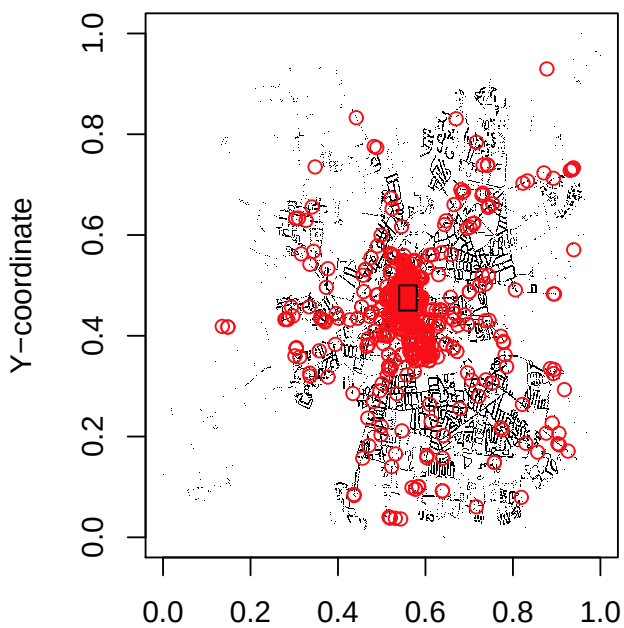


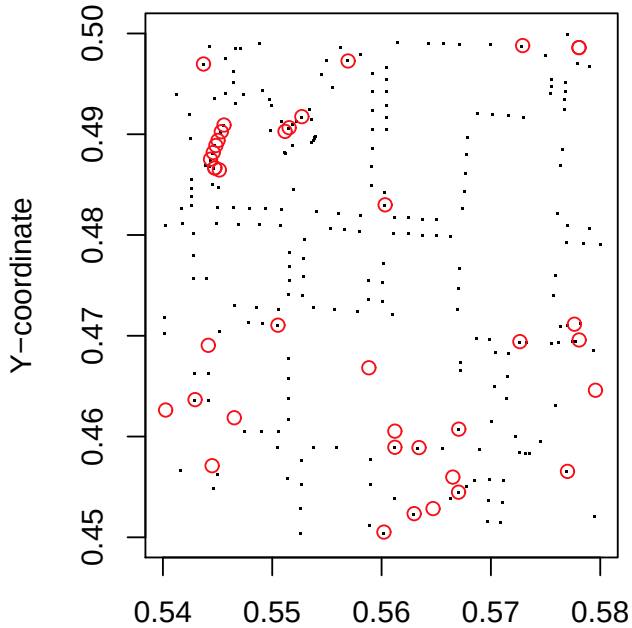
Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}
1(0.018)	1(0.029)	1

Études sur la répartition des services

Représentativité des services / logements

Restaurants



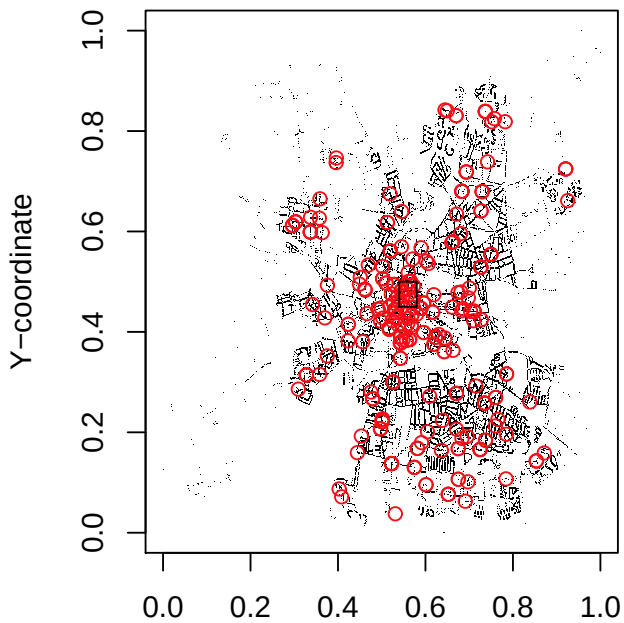


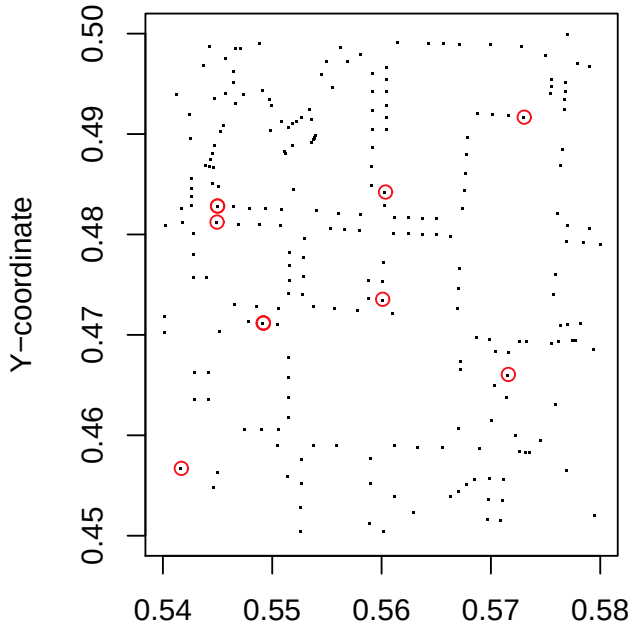
Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}
0(0.09)	0(0.23)	1

Études sur la répartition des services

Représentativité des services / logements

Médecins généralistes



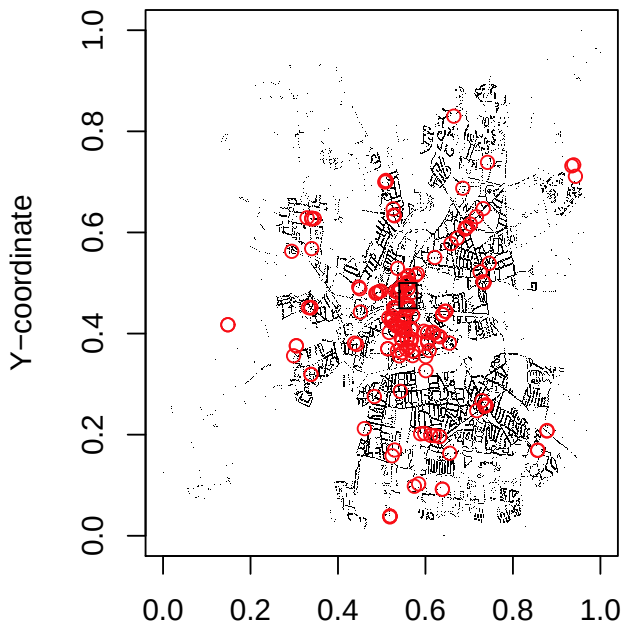


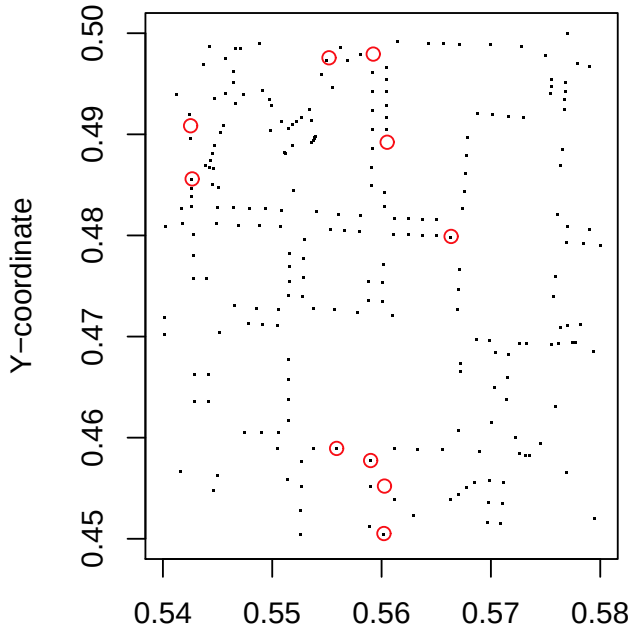
Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}
0(0.18)	0(0.32)	1

Études sur la répartition des services

Représentativité des services / logements

Banques





Φ_{Cramer}	Φ_{Bahr}	Φ_{KMMD}
0(0.51)	0(0.71)	1

Conclusions, perspectives

Conclusions

- Résultats des tests en accord avec les résultats intuitifs dans de bonnes conditions d'application $\longrightarrow$ limites

Perspectives

- Tests de comparaison moins sensibles aux problèmes de taille : prise en compte de la structure en réseau de la représentation des services ou des logements
- Tests d'homogénéité sur un réseau d'adresses ou de rues

Pistes

- Détermination de la structure mathématique la mieux adaptée à la modélisation de ce réseau
- Utilisation de noyaux adaptés à cette structure...

La vie n'est bonne qu'à deux choses : découvrir les mathématiques et enseigner les mathématiques

Siméon-Denis Poisson