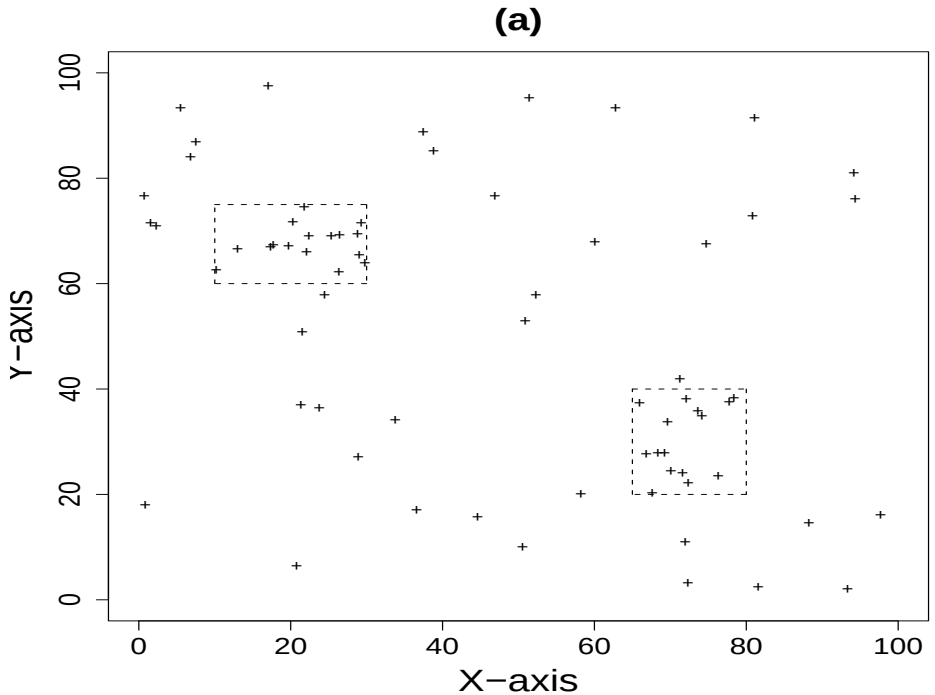
Détection d'agrégats d'événements ponctuels

Nicolas Molinari
Institut Universitaire de Recherche Clinique
Université Montpellier I
CHU de Nîmes

Lundi 26 Novembre 2007

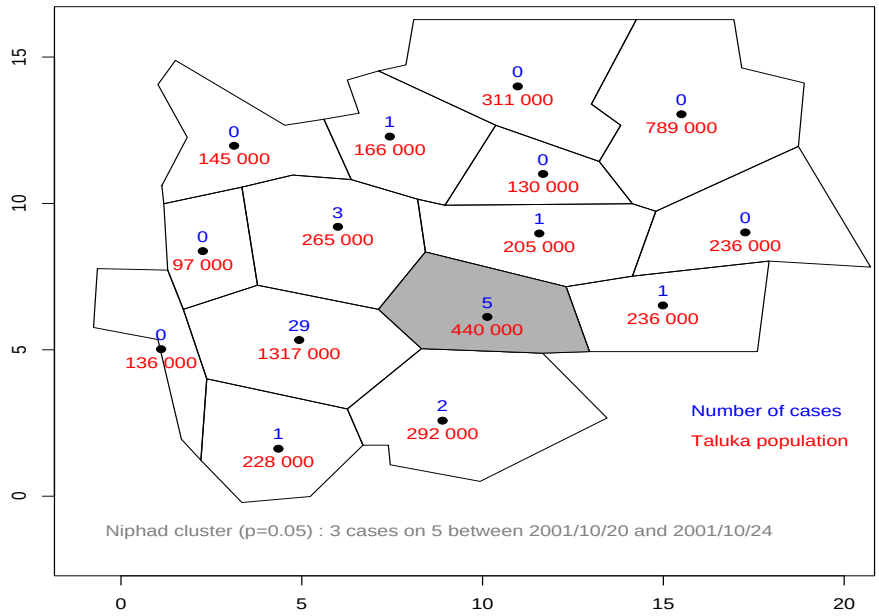
Introduction

Un **cluster** est un agrégat de points de R^p de densité "anormalement" élevée, non due au hasard.



Généralités

- Tests globaux (tendance générale à l'agrégation)
- Tests de détection (ou locaux)
- Tests de concentration (autour d'un foyer pré-défini)



Clusters temporaux, $\boldsymbol{R}^1$

A *Clusterville*, $n = 42$ cas d'une maladie rare se sont produis au cours de l'année. Un cas tous les 10 jours sauf entre les jours 181 et 241 où il y eu un cas tous les 5 jours.

Y a t'il un cluster durant cette période?

Ederer, F., Myers, E. and Mantel, N. (1964)
Naus, J. I. (1966), $p = 0.38$
Tango, T. (1984), $p = 0.2$
Nagarwalla, N. and Kulldorf, M. (1995), $p = 0.09$

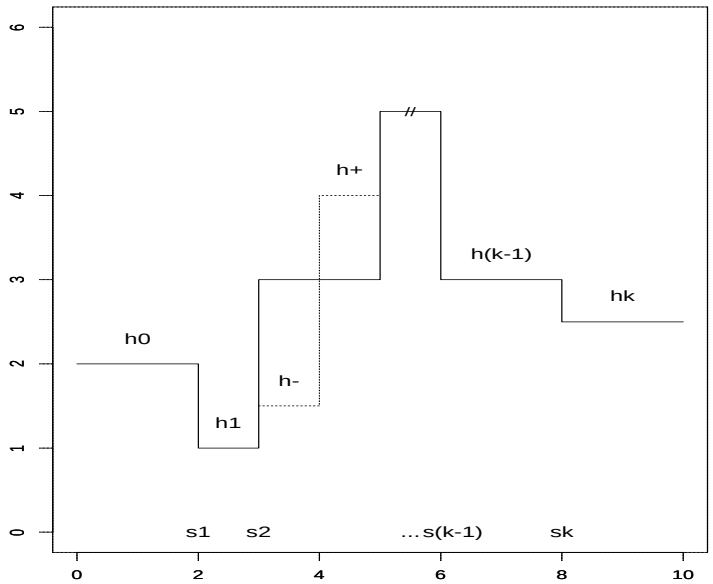
Clusters temporaux, rjMCMC

$$\lambda(t) = \begin{cases} \theta_0 & \text{si } 0 < t < s_1 \\ \vdots & \\ \theta_k & \text{si } s_k < t < T \end{cases}$$

$x = (k, s_1, \ldots, s_k)$

Distribution cible de la chaîne :

$\pi(x) = p(x|y) = p(k|y)p((s_1, \ldots, s_k)|y)$



Clusters temporaux, rjMCMC (cont.)

Quatre types de mouvements :
1) changement de la valeur du taux sur un intervalle,
2) changement de la position d'un point de rupture s_j aléatoirement choisi,
3) naissance d'une marche en choisissant aléatoirement une nouvelle position et estimation des taux,
4) suppression d'une marche et estimation du taux.

Clusters temporaux, rjMCMC (cont.)

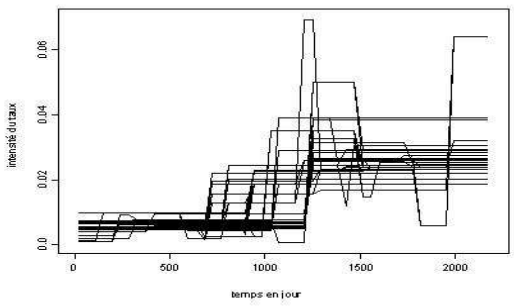
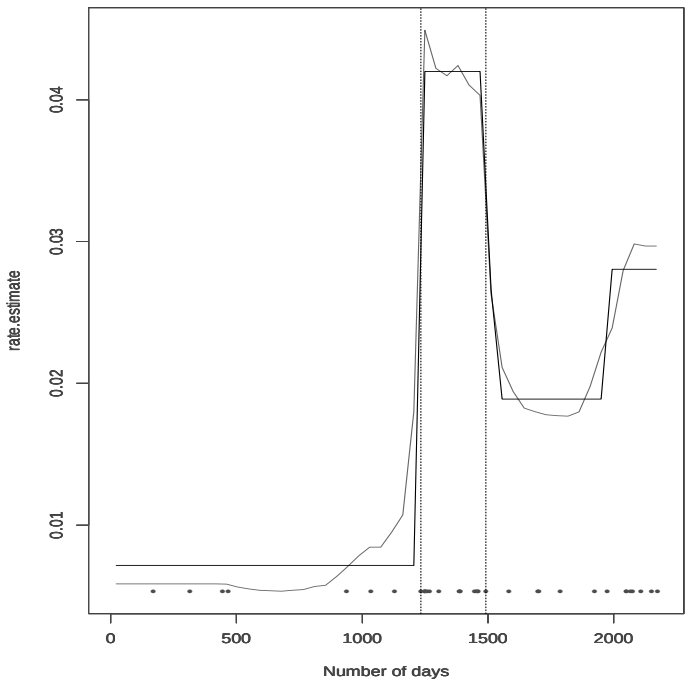


Fig 8.2: Exemple de fonctions taux par morceaux (*a posteriori*) construites par l'algorithme.



Approche par régression

- Transformation des données (trajectoire)
- Régression (détermination du ou des clusters)
- Inférence (significativité des clusters)

Clusters temporaux, le modèle

$X_1, \cdots, X_n$ les dates d'occurrence de n événements.

$Y_i = X_{(i)} - X_{(i-1)}$
Sous H_0, Y_i suit une $\beta(1, n)$.

Considérons les données $(i, y_i)_{i=1,\cdots,n}$, sous l'hypothése de répartition uniforme, on peut proposer comme modéle de régression

$$f(i) = \bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j.$$

En supposant que $x_{n_{j-1}}, \cdots, x_{n_j}$ sont regroupés en cluster,

$$\bar{y}_c = \frac{1}{n_j - n_{j-1}} \sum_{i=n_{j-1}}^{n_j - 1} y_i < \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}.$$

Clusters temporaux, le modèle (cont.)

Ainsi, le modéle

$$f(i) = \sum_{j=1}^{m+1} \bar{y}_{[n_{j-1}+1;n_j]} \times \mathbf{1}_{[n_{j-1}+1;n_j]}(i)$$

est "préférable".

On résoud

$$\min_{0<n_1<\ldots<n_m<n} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(i))^2.$$

Clusters temporaux, algorithme

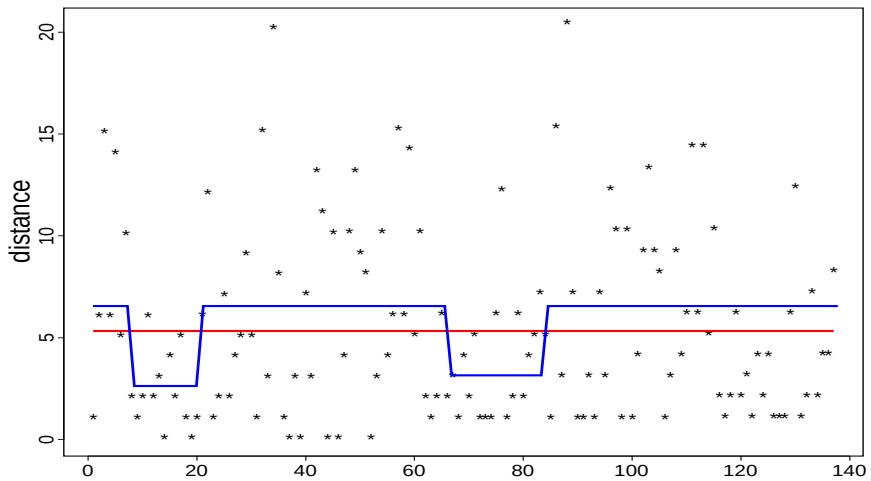
Statistique de test (Bai & Perron, 1998, 2003).
Pour m breaks,

$$F(\hat{n}_1, ...\hat{n}_m) = \frac{n-m-2}{m}\hat{\delta}R'(R\hat{V}(\hat{\delta})R')^{-1}R\hat{\delta}.$$

Statistique du double maximum pour déterminer m et méthode de rééchantillonnage.

Clusters temporaux, illustration

Cas d'hémoptysie au CHU de Nice



Correction de l'évolution de la population "à risque".

Clusters temporaux, inégalité exponentielle

Proposition :
Notons $N = \hat{n}_{k+1} - \hat{n}_k$ et $(Y_i)_{i=1}^{N} = (Y_i)_{i=\hat{n}_k+1}^{\hat{n}_{k+1}}$. Si $(Y_i)_{i=1}^{N}$ iid $\beta(1, n)$ avec $n \geq N$. Pour tout $i \in \{1, \ldots, N\}$, on définit $Z_i = (n+1)Y_i$. Notons $T = \frac{1}{N}\sum_{i=1}^{N} Z_i$ et $t > 0$. Alors,

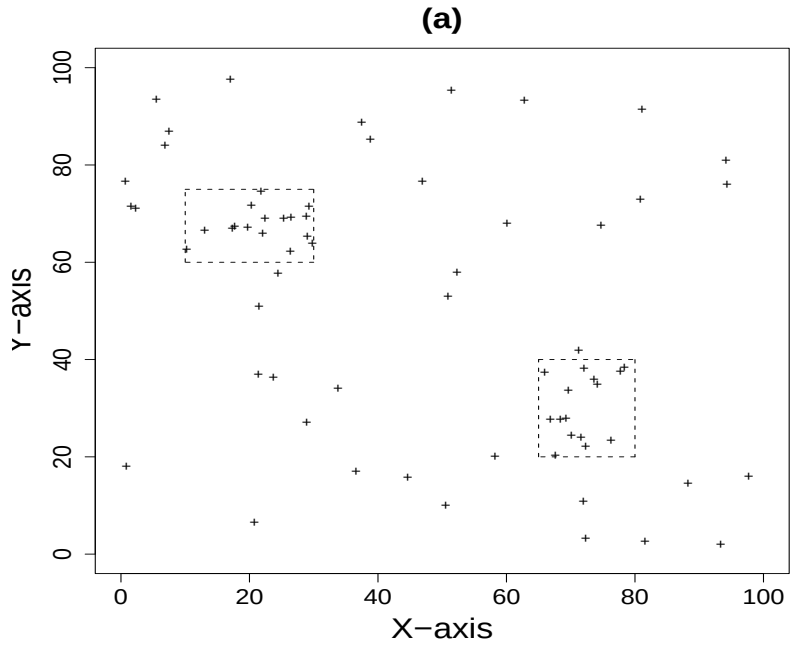
$$\mathbb{P}\left(T \leq 1 - \frac{t}{N}\right) \leq \exp\left(-\frac{t^2}{\frac{2nN}{n+2} + \frac{2t}{3}}\right). \tag{1}$$

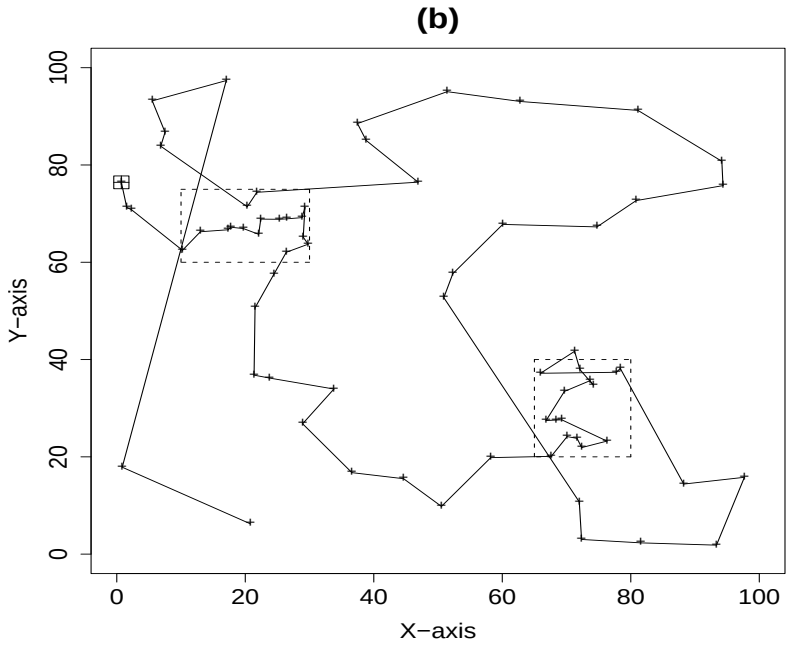
Corollaire :
Pour $\alpha \in (0, 1)$,

$$\mathbb{P}\left(T \leq 1 - \frac{t_\alpha}{N}\right) \leq \alpha \text{ avec } t_\alpha = -\frac{\ln(\alpha)}{3} + \sqrt{\left(\frac{\ln(\alpha)}{3}\right)^2 - \frac{2nN\ln(\alpha)}{n+2}}. \tag{2}$$

Clusters spatiaux, R^2

Pas d'ordre temporel entre les différentes occurences, définition d'une trajectoire

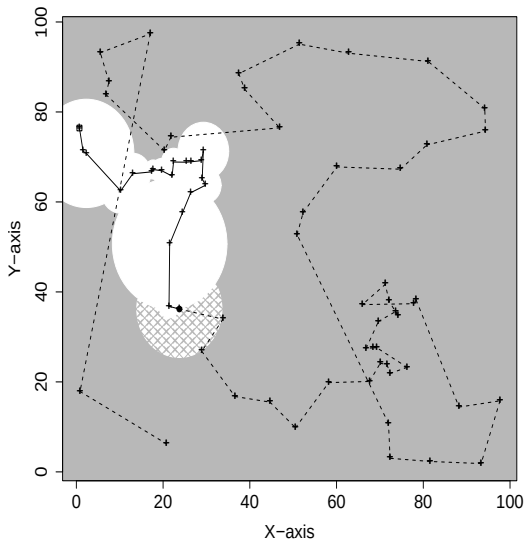


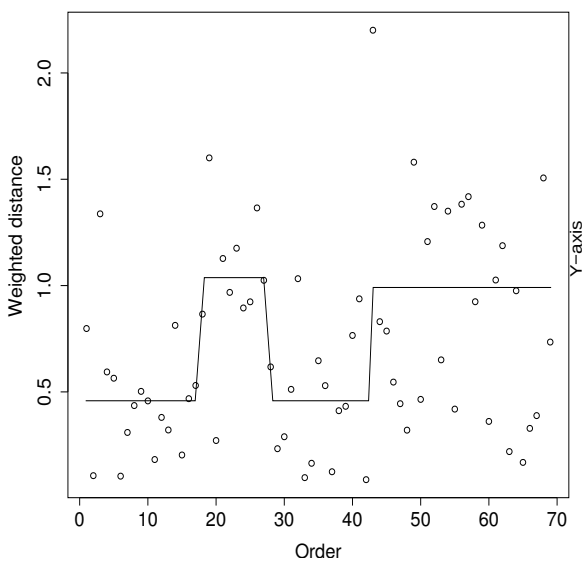


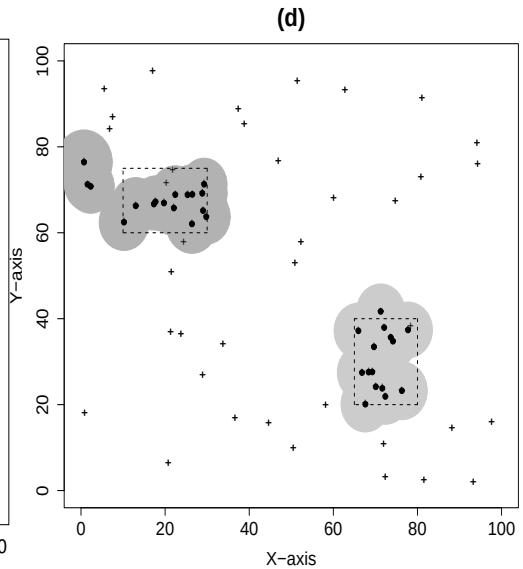
Clusters spatiaux, effets de bord et de trajectoire

$$E[D_k/x_1,...,x_k] = \int P(D_k > r/x_1,...,x_k)dr,$$

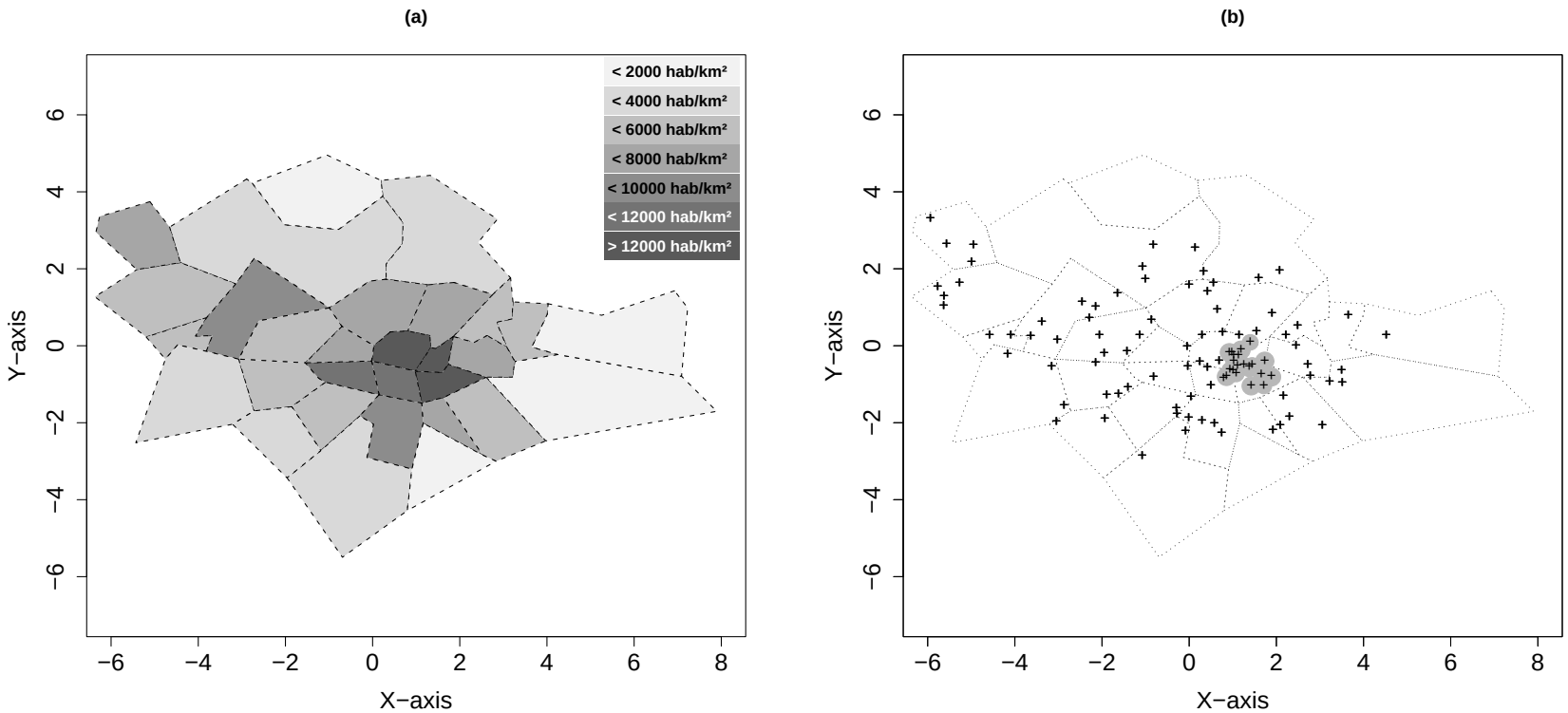
$$P(D_k > r/x_1,...,x_k) = \left[1 - \frac{\int_{grisee(k,r)} f(x)dx}{\int_{gris(k)} f(x)dx}\right]^{n-k}.$$



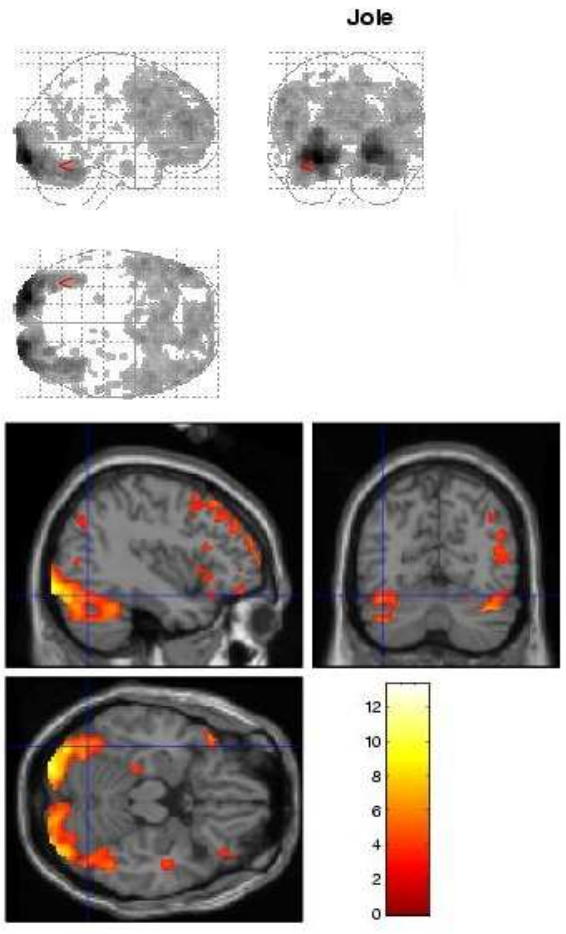


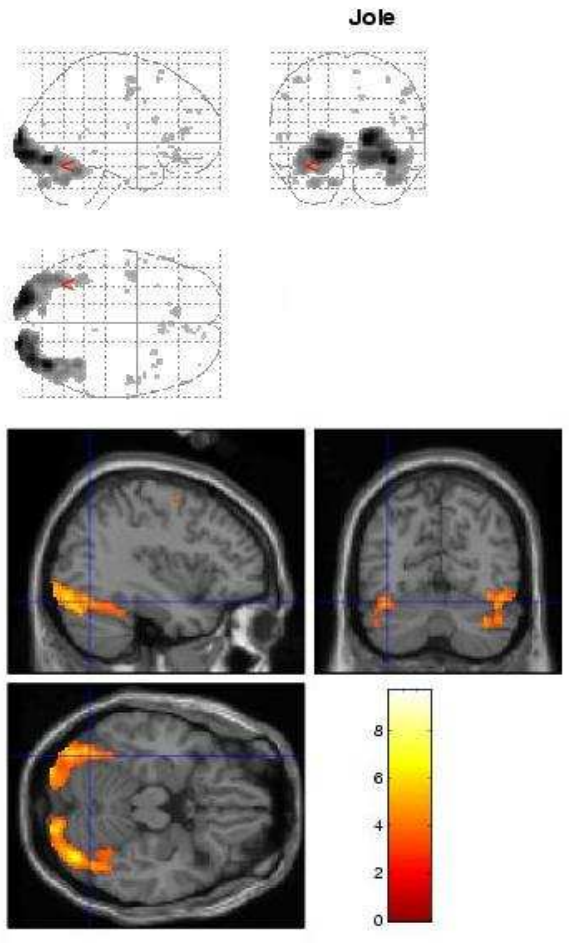


Clusters spatiaux, applications



Dans $\mathbb{R}^p$, $p > 2$





Application R à l'IRMf