Méthode d'inventaire et d'analyse

I. Introduction

L'écologie vise à étudier les associations entre êtres vivants ainsi que les associations environnement / êtres vivants. Pour cela, les écologues réalisent très largement des inventaires faunistiques ou floristiques qui dénombrent les espèces répertoriées dans différents sites.

La description des sites par leur composition spécifique (en terme de présence/absence, abondance, biomasse, etc.) ou bien de manière complémentaire, la description des espèces par les sites qu'elles occupent, constitue l'information principale. À cela peuvent s'ajouter les coordonnées des sites qui permettent de réaliser le lien à l'organisation spatiale. Enfin, ces données peuvent être complétées par de l'information supplémentaire caractérisant les sites d'une part, et les espèces d'autre part.

Le but de notre cours a été de proposer, à partir des questions écologiques récurrentes, une méthodologie claire et accessible à tous pour analyser ce type de données. Les questions abordées par la méthodologie proposée sont les suivantes : - Comment sont distribuées les espèces? - Existe-t-il, sur l'aire géographique considérée, des profils de composition spécifique des sites ou bien des profils de répartition d'espèces?

Comprendre et maîtriser des méthodes statistiques qui traitent ces informations individuelles et spatialisées permet de travailler sur des données qui sont aujourd'hui de plus en plus accessibles et recherchées car elles fournissent des analyses très précises sur les comportements des acteurs économiques (ELLISON et al. 2010; BARLET et al. 2013).

Dans ce cadre d'analyse, plusieurs questions méthodologiques importantes se posent alors au statisticien qui dispose de jeux de points à analyser : comment représenter et caractériser spatialement de telles données en utilisant des milliers voire des millions d'observations ? Quels outils statistiques existent et peuvent être mobilisés pour étudier ces observations relatives aux ménages, salariés, firmes, magasins, équipements ou déplacements par exemple ? Comment prendre en compte les caractéristiques qualitatives ou quantitatives des observations étudiées ? Comment mettre en évidence des éventuelles attractions ou répulsions entre les points ou entre différents types de points ? Comment peut-on évaluer la significativité des résultats obtenus ? etc. Pour ce faire, nous nous appuierons sur des méthodes statistiques qui permettent de caractériser des distributions de points et nous expliciterons les enjeux associés.

II. Processus ponctuels

II.1.Le processus de Poisson homogène

Le processus ponctuel permettant de générer des distributions spatiales de points complètement aléatoires (Complete Spatial Randomness - CSR). Pour y arriver, on peut démarrer par un processus particulièrement simple, ***U***, qui génère un unique point pouvant être situé de façon aléatoire sur un domaine d'intérêt ***W***. Si u_1 et u_2 sont les coordonnées du point, il est possible de calculer la probabilité que le point généré par ***U*** se trouve dans un petit espace ***B*** choisi arbitrairement :

$$P(U \in B) = \int_B f(u_1, u_2) du_1 du_2.$$

La répartition est uniforme sur ***W*** si $f(u_1, u_2) = \frac{1}{|W|}$ où $|W$ désigne l'aire de ***W***.

On a donc $P(U \in B) = \int_B f(u_1, u_2) du_1 du_2 = \frac{1}{|W|} \int_B du_1 du_2 = \frac{|B|}{|W|}.$

Ce processus permet d'en définir un autre, le processus binomial. n points sont répartis de façon uniforme sur la région W, de façon indépendante. On peut écrire, de façon classique que :

$$P(n(X \cap B) = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{, avec} \quad p = \frac{|B|}{|W|}.$$

La fonction runifpoint du package spatstat permet de générer des configurations de points à partir d'un processus binomial uniforme. Par exemple, sur la figure suivante, 1 000 points sont attendus sur une fenêtre d'observation 10 x 10.

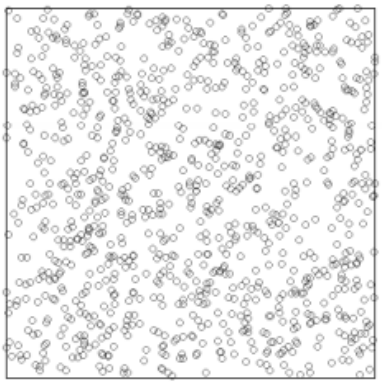

Fig.1 –Tirage de 1 000 points par un processus binomial uniforme
Source : package spatstat, calculs des auteurs.

```
library("spatstat")
plot(runifpoint(1000, win=owin(c(0, 10),c(0, 10))), main="")
```

❖ Pourquoi un tel processus, dans lequel chaque point est placé au hasard de façon uniforme ne convient-il pas pour définir un processus CSR (Complete Spatial Randomness)? On demande dans un premier temps à un tel processus deux propriétés :

— l'**homogénéité** qui correspond à l'absence de "préférence" pour une localisation particulière (c'est bien le cas pour le processus binomial).

— l'**indépendance** traduisant le fait que les réalisations dans une région de l'espace n'ont pas d'influence sur les réalisations dans une autre région.

Processus CSR

Le processus CSR ou processus de Poisson homogène est souvent défini de la façon suivante :

— $P(n(X \cap B) = k) = e^{-\lambda |B|} \frac{\lambda^k |B|^k}{k!}$.
Cela définit le caractère poissonnien de la distribution **(PP1)**;

— $\mathbb{E}[n(X \cap B)] = \lambda |B|$.
Cela définit l'homogénéité **(PP2)**;

— $n(X \cap B_1), \ldots, n(X \cap B_m)$ sont m variables aléatoires indépendantes **(PP3)**;

— une fois fixé le nombre de points, la répartition est uniforme **(PP4)**.

Les propriétés **PP2** et **PP3** sont suffisantes pour définir le processus CSR (DIGGLE 1983), et on peut démontrer que les autres en sont les conséquences.

II.2. L'intensité, propriété d'ordre 1

Les lois des processus sont très complexes (MOLLER et al. 2004), ce qui conduit dans la pratique à utiliser de façon privilégiée des indicateurs qualifiés de propriété d'ordre 1 ou d'ordre 2, comme on utilise les moments d'ordre 1 et 2 (espérance et variance) pour appréhender une variable aléatoire de loi inconnue.

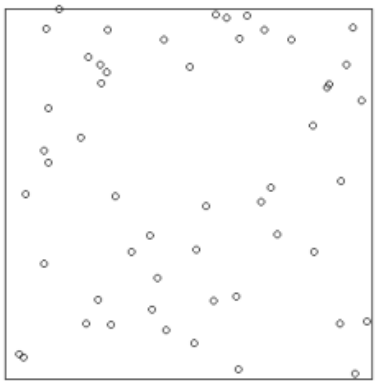

Fig.2 –Tirage de 50 points par un processus binomial uniforme
Source : package spatstat, calculs des auteurs.

```
library("spatstat")
plot(rpoispp(50), main="")
```

- ***Intensité d'un processus***

L'intensité est apparue dans la présentation du processus de Poisson où elle était constante ***(λ).*** Il existe d'autres processus dans lesquels cette hypothèse est relâchée, et où la fonction d'intensité ***λ(x)*** est variable. Elle est définie par :

$$\mathbb{E}[n(X \cap B)] = \mu(B) = \int_B \lambda(x)dx.$$

En appliquant la définition de l'espérance à une petite région centrée en **x** et de surface **dx**, on peut **définir l'intensité** en ce point **x** comme **le nombre de points attendus dans cette petite surface lorsqu'elle tend vers 0**, soit :

$$\lambda(x) = \lim_{|dx| \to 0} \frac{\mathbb{E}[N(dx)]}{|dx|}.$$

II.3. Le processus de Poisson inhomogène

Les processus de Poisson qualifiés d'inhomogènes sont d'intensité variable et leurs points sont distribués indépendamment les uns des autres (la condition **PP3** est conservée). La condition **PP1** sur le caractère poissonnien de la distribution conditionnellement à n est maintenue, le paramètre de loi n'étant plus ***λ |B|,*** mais ***μ(B)*** tel que défini précédemment. La condition **PP4** est modifiée. Conditionnellement à un nombre de points fixé n, les points sont indépendants et identiquement distribués, avec une densité de probabilité $f(x) = \frac{\lambda(x)}{\int_B f(u)du}$.

On trouvera sur la figure 3 deux exemples de processus de Poisson inhomogène, caractérisés par leur fonction d'intensité (où x et y sont les coordonnées).

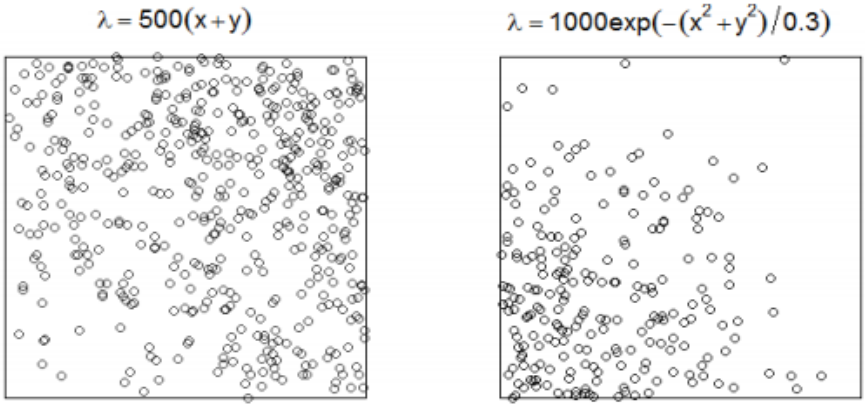

Fig.3 – Exemples de processus inhomogènes
Source : package spatstat, calculs des auteurs.

```
library("spatstat")
par(mfrow=c(1, 2))
plot(rpoispp(function(x, y) {500*(x+y)}), main=expression(lambda==500*(x+y)))
plot(rpoispp(function(x,y) {1000*exp(-(x^2+y^2)/.3)}), main=expression(
    lambda==1000*exp(-(x^2+y^2)/.3)))
par(mfrow=c(1,1))
```

II.4. Les propriétés de second ordre

On va s'intéresser, pour introduire les propriétés du second ordre d'un processus ponctuel, ***à la variance et à la covariance des comptages de points***, que l'on définit ci-dessous :

$$var(n(X \cap B) = \mathbb{E}[n(X \cap B)^2] - \mathbb{E}[n(X \cap B)]^2$$

$$cov\,[n(X \cap B_1), n(X \cap B_2)] = \mathbb{E}[n(X \cap B_1)n(X \cap B_2)] - \mathbb{E}[n(X \cap B_1)]\mathbb{E}[n(X \cap B_2)]$$

- ***Moment d'ordre 2 d'un processus***

Plutôt que d'utiliser ces indicateurs, on définit le moment d'ordre deux de la façon suivante :

$$\nu_{|2|}(A \times B) = \mathbb{E}[n(X \cap A)n(X \cap B)] - \mathbb{E}\,[n(X \cap A \cap B)]\,,$$

Qui vaut pour le processus de Poisson **: λ 2 |A||B|.** Lorsque cette mesure admet une densité, celle-ci, appelée ***intensité d'ordre 2*** et notée ***λ2*** est définie de telle sorte que ***v|2| (C) = R C λ2 (u, v) dudv***. Cette intensité du second ordre peut s'interpréter comme :

$$\lambda_2(x,y) = \lim_{|dx|\to 0|dy|\to 0,} \frac{\mathbb{E}\,[N(dx)N(dy)]}{|dx|\,|dy|}.$$

II.5. Des processus ponctuels aux répartitions observées de points - Répartition au hasard, agrégation, régularité

Lorsque l'on étudie une distribution de points, deux grandes questions se posent : les points observés sont-ils distribués au hasard ou y a-t-il une interaction ? S'il y a une interdépendance, est-elle de nature agrégative ou répulsive ? Selon les réponses à ces questions, ***trois configurations de points*** sont généralement mises en évidence : une distribution dite complètement aléatoire, une agrégée et une régulière. Un exemple de ces trois distributions théoriques est représenté sur la figure 4 et 5.Ces distributions de points sont obtenues à partir de processus ponctuels connus simulés à l'aide du *package spatstat*.

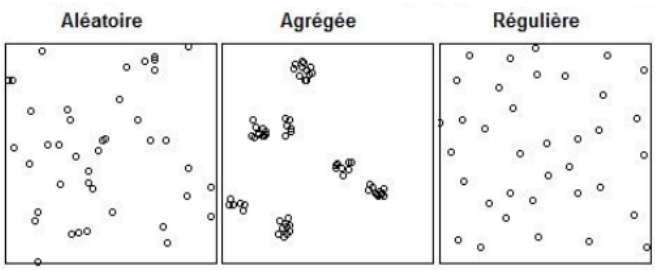

Fig.4 - Les trois configurations classiques de points

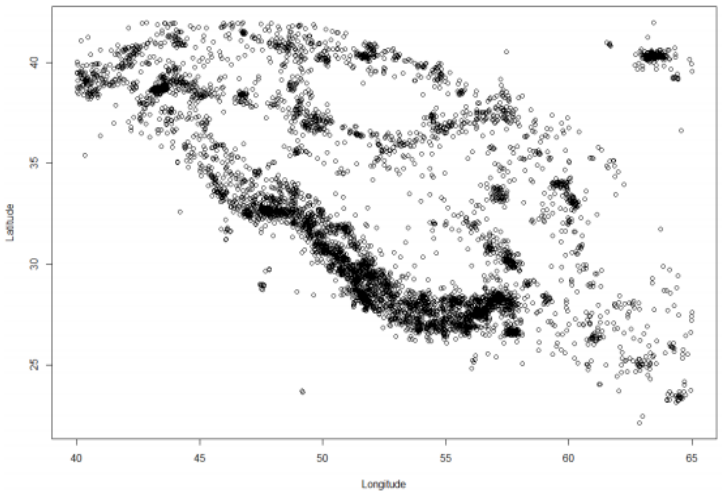

Fig.5 - Localisation de 5 970 épicentres de séismes en Iran survenus de 1976 à 2016

Source : package spatstat, calculs des auteurs.

III. Méthode de Ripley - Fonction K de Ripley et ses variantes

L'indicateur le plus utilisé pour appréhender la corrélation dans les processus ponctuels est la **fonction empirique** $\widehat{K}$, proposée par B.D. Ripley en 1976 (RIPLEY 1976; RIPLEY 1977). Cette fonction nommée couramment ***la fonction de Ripley*** a fait l'objet de nombreux commentaires et développements et de plusieurs variantes. Concrètement, cette fonction va nous permettre d'estimer le nombre moyen de voisins rapporté à l'intensité.

- ***Fonction K de Ripley***

  Son estimateur s'écrit de la façon suivante :

$$\hat{K}(r) = \frac{|W|}{n(n-1)} \sum_i \sum_{j \neq i} \mathbf{1}\{\|x_i - x_j\| \leq r\} c(x_i, x_j; r),$$

Où ***n*** est le nombre total de points sur la fenêtre d'observation, **1 {|| xi –xj| |** } ≤ ***r*** est une indicatrice qui vaut **1** si les points ***i*** et ***j*** sont à une distance au plus égale à ***r*** et ***0*** sinon. **c (xi, xj ; r)** correspond à la correction des effets de bord et ***W*** à l'aire d'étude.

K est une fonction cumulative, donnant le nombre moyen de voisins à distance inférieure à r de chaque point, standardisée par l'intensité du processus **(n/|W|)**, supposé homogène.

Pratiquement, pour étudier le voisinage des points, nous allons balayer toutes les distances r, en calculant la valeur de la fonction ***K*** pour chacune de ces distances. On procède pour cela de la manière suivante :

1. Pour chaque point et distance r, on décompte le nombre de ses voisins (les autres points) localisés sur le disque de rayon r;
2. Puis on calcule le nombre moyen de voisins (en tenant compte d'éventuels effets de bord) pour chaque distance r;
3. Enfin, ces résultats vont être comparés à ceux obtenus sous l'hypothèse d'une distribution homogène (réalisation d'un processus de Poisson homogène), qui sera la valeur de référence attendue

Finalement, on cherchera à détecter s'il existe un écart significatif entre les estimations du nombre de voisins observés et attendus.

Nous avons rapproché sur la figure 6 les trois configurations-types de points vues précédemment et les trois courbes de la fonction ***K*** ainsi obtenues. On représente graphiquement en abscisses la distance r et en ordonnées la valeur de la fonction K estimée à cette distance. Avec le package spatstat, la fonction ***K*** est calculée à l'aide de la fonction ***K*** est. Sur la figure 6, la fonction ***K*** estimée est reportée en noire sur les trois graphiques et la valeur de référence en pointillés rouges. Il vient :

- **lorsque le processus est complètement aléatoire, la courbe s'écarte relativement peu de πr^2.** On peut le constater sur le graphique en bas à gauche de la figure **6** . La courbe de K reste proche de la valeur de référence πr^2, pour tous les rayons r.
- **dans le cas d'un processus régulier**, on obtient : $\hat{K}(r) < K_{pois}(r)$ puisque si les points se repoussent, ils ont moins de voisins en moyenne dans un rayon r que sous l'hypothèse d'une distribution aléatoire de points. Graphiquement, la courbe K détecte cette répulsion : on constate sur le graphique de droite que la courbe K est située sous la valeur de référence (πr^2) pour tous les rayons.
- **dans le cas d'un processus agrégé**, il y a en moyenne plus de points dans un rayon r autour des points que le nombre attendu sous une distribution aléatoire : par conséquent les points s'attirent et $\hat{K}(r) > K_{pois}(r)$. Graphiquement, la courbe K estimée est cette fois-ci située au dessus de la valeur de référence pour tous les rayons d'étude, comme on peut le noter sur le graphique central reportée sur la figure **6** .

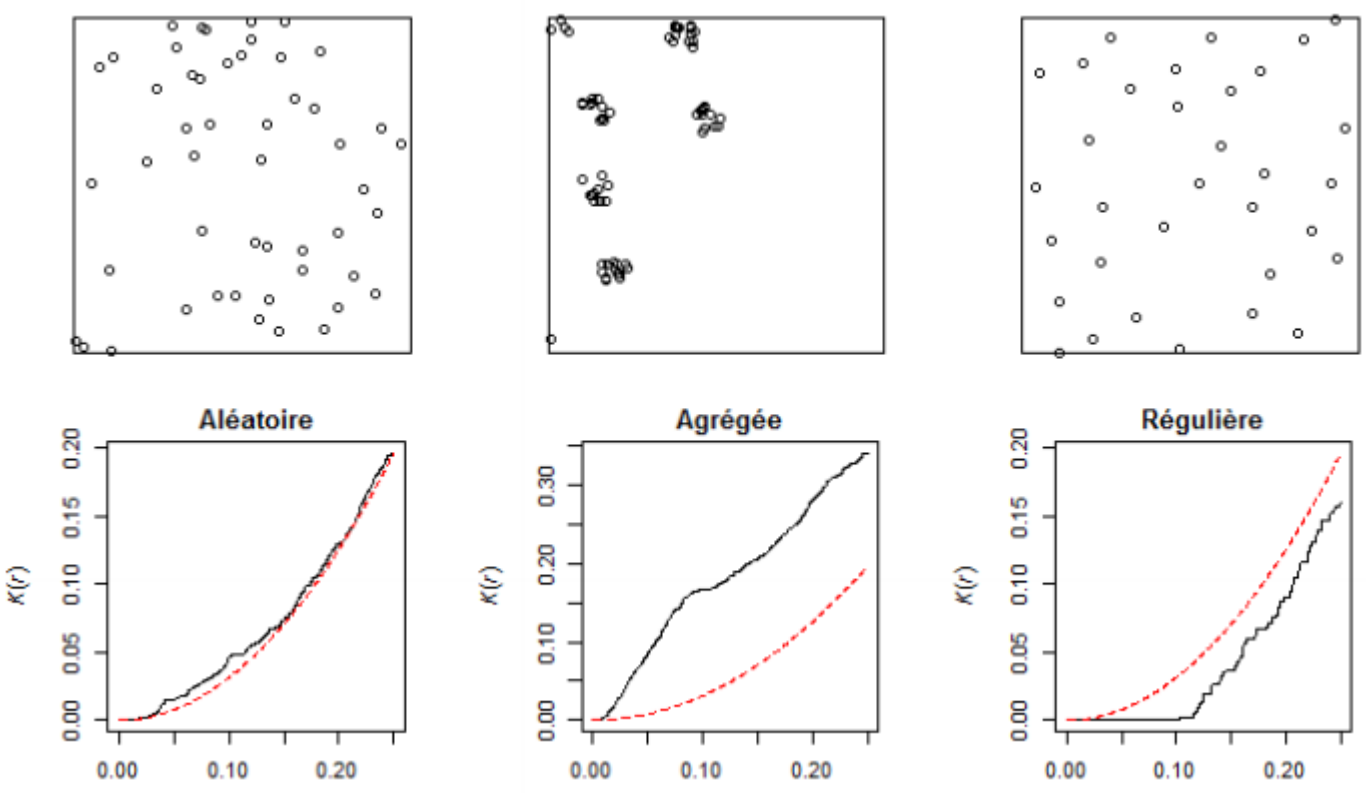

Fig.6– Les fonctions K des trois configurations-types
Source : package spatstat, calculs des auteurs.

```
library("spatstat")
par(mfrow=c(2, 3), mar=c(1, 2, 2, 2))
plot(rpoispp(50), main="")
plot(rMatClust(5, 0.05, 10), main="")
plot(rMaternII(200,0.1), main="")
par(mar=c(4, 4.1, 2, 3))
# Fonction K calculée par spatstat
plot(Kest(rpoispp(50),correction="isotropic"),legend=FALSE,main="Aléatoire"
    )
plot(Kest(rMatClust(5, 0.05, 10),correction="isotropic"),legend=FALSE,main=
    "Agrégée")
plot(Kest(rMaternII(200,0.1),correction="isotropic"),legend=FALSE,main="Ré
    gulière")
par(mfrow=c(1, 1))
```

Bibliographie:

Les configurations de points
JEAN-MICHEL FLOCH
Insee
ÉRIC MARCON
Agro Paris Tech, UMR Eco FoG, BP 709, F-97310 Kourou, Guyane française
FLORENCE PUECH
RITM, Univ. Paris-Sud, Université Paris-Saclay & Crest, 92330 Sceaux, France.78-93 PP.