4. Les configurations de points

JEAN-MICHEL FLOCH
Insee

ÉRIC MARCON
AgroParisTech, UMR EcoFoG, BP 709, F-97310 Kourou, Guyane française

FLORENCE PUECH
RITM, Univ. Paris-Sud, Université Paris-Saclay & Crest, 92330 Sceaux, France

4.1 Cadre d'analyse : les concepts fondamentaux 76
4.1.1 Configurations et processus . . . 77
4.1.2 Processus marqués . . . 77
4.1.3 Fenêtre d'observation . . . 77
4.2 Processus ponctuels : une présentation succincte 78
4.2.1 Le processus de Poisson homogène . . . 78
4.2.2 L'intensité, propriété d'ordre 1 . . . 80
4.2.3 Le processus de Poisson inhomogène . . . 81
4.2.4 Les propriétés de second ordre . . . 82
4.3 Des processus ponctuels aux répartitions observées de points 83
4.3.1 Répartition au hasard, agrégation, régularité . . . 83
4.3.2 Mises en garde . . . 84
4.4 Quels outils statistiques mobiliser pour étudier les configurations de points ? 86
4.4.1 La fonction K de Ripley et ses variantes . . . 86
4.4.2 Comment tester la significativité des résultats ? . . . 91
4.4.3 Point d'étape et mise en évidence de propriétés importantes pour de nouvelles mesures . . . 93
4.5 Mesures fondées sur les distances récemment proposées 98
4.5.1 Indicateur K_d de Duranton et Overman . . . 98
4.5.2 Indicateur M de Marcon et Puech . . . 99
4.5.3 Autres développements . . . 101
4.6 Processus multitypes 101
4.6.1 Fonctions d'intensité . . . 102
4.6.2 Fonctions intertypes . . . 105
4.7 Modélisation des processus 110
4.7.1 Cadre général pour la modélisation . . . 110
4.7.2 Exemples d'application . . . 110

Résumé

Les statisticiens peuvent être amenés à étudier précisément des données spatialisées, par exemple la distribution des revenus des ménages, l'implantation sectorielle d'établissements industriels ou commerciaux, la localisation des établissements scolaires au sein des villes, etc. Des réponses peuvent être apportées grâce à des analyses menées à une ou plusieurs échelles géographiques prédéfinies comme au niveau des quartiers, des arrondissements ou des îlots. Toutefois, il est tentant de vouloir préserver la richesse des données individuelles et travailler en conservant la position exacte des entités étudiées. Si tel est le cas, cela revient pour un statisticien à élaborer des analyses à partir de données géolocalisées sans procéder à une quelconque agrégation géographique. Les observations sont appréhendées comme des points dans l'espace et l'objectif est de caractériser ces distributions de points.

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.

Ce chapitre a pour but d'aider le statisticien à apporter des résultats statistiquement robustes à partir de l'étude de données spatialisées qui ne reposent pas sur un zonage prédéfini. Pour ce faire, nous nous appuierons sur une revue de la littérature des méthodes statistiques qui permettent de caractériser des distributions de points et nous expliciterons les enjeux associés. Nous expliquerons à partir d'exemples simples les avantages et les inconvénients des approches les plus souvent retenues. Le code sous R fourni permettra de reproduire les exemples traités.

Remerciements : Les auteurs remercient Gabriel LANG et Salima BOUAYAD AGHA pour leur relecture attentive d'une première version de ce chapitre ainsi que pour l'ensemble de leurs commentaires constructifs. Marie-Pierre de BELLEFON et Vincent LOONIS qui sont à l'initiative de ce projet sont également remerciés : ce chapitre a indéniablement bénéficié de tous leurs efforts éditoriaux ainsi que de ceux de Vianney COSTEMALLE.

Introduction

L'étude des configurations de points peut paraître plus éloignée des préoccupations des statisticiens publics que d'autres méthodes. Pourquoi leur accorder une place dans ce manuel ? La réponse est simple : la géolocalisation des données permet de disposer d'observations localisées nombreuses sur les entreprises, les équipements, les logements. On est ainsi rapidement amené à s'interroger sur le regroupement possible de ces observations, sur la configuration spatiale de leur implantation aléatoire ou non, sur leur dépendance à d'autres processus (la proximité d'établissements industriels entretenant de forts liens *input-output* peut être recherchée et donc à l'origine d'interactions spatiales entre établissements de différents secteurs). L'objectif de ce chapitre est de présenter une introduction à un corpus de méthodes parfois complexes dans leurs fondements mathématiques, mais qui servent souvent à illustrer des questions assez simples. Les préoccupations des écologues, des forestiers, des épidémiologistes ont été à l'origine du développement de ces méthodes. P.J. Diggle, l'auteur du premier manuel de référence (DIGGLE 1983), est connu pour ses nombreux travaux en épidémiologie (DIGGLE et al. 1991). De ce fait, les exemples pédagogiques permettant d'illustrer les méthodes ponctuelles proviennent souvent de données forestières ou épidémiologiques. Nous nous appuierons dans ce chapitre sur des exemples de ce type fournis dans certains packages de R comme *spatstat* (BADDELEY et al. 2005) ou *dbmss* (MARCON et al. 2015b). Nous retiendrons également des données sur l'implantation d'équipements en France.

Dans l'étude des configurations de points, contrairement aux méthodes fondées sur un zonage ou géostatistiques, on ne mesure pas localement une variable, mais c'est la localisation même des points qui est au cœur du sujet considéré. C'est à partir de ceux-ci que l'on va construire des modèles et faire de l'inférence. Les cartes de la figure 4.1, réalisées à partir d'un extrait des données de la base permanente des équipements (BPE), montrent quatre exemples de localisation d'activités, dans la ville de Rennes (France). [1]

```
library("spatstat")
library("sp")
# Fichier de la BPE sur le site insee.fr :
# Données pour ces exemples :
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))
bpe_eco <- bpe[bpe$TYPEQU=="C104", ]
bpe_pha <- bpe[bpe$TYPEQU=="D301", ]
bpe_vet <- bpe[bpe$TYPEQU=="B302", ]
bpe_med <- bpe[bpe$TYPEQU=="D201", ]
par(mfrow=c(2,2), mar=c(2, 2, 2, 2))
plot(carte, main="Ecoles") ; points(bpe_eco[, 2:3])
plot(carte, main="Pharmacies") ; points(bpe_pha[, 2:3])
plot(carte, main="Magasins de vêtements") ; points(bpe_vet[, 2:3])
plot(carte, main="Médecins") ; points(bpe_med[, 2:3])
par(mfrow=c(1,1))
```

Ces quatre figures simples permettent d'avoir un premier aperçu des grandes différences d'implantation de ces équipements. Ainsi, les magasins de vêtements sont très nombreux, mais extrêmement localisés dans le centre de Rennes. À l'opposé, les écoles primaires semblent réparties de façon plus régulière. Les pharmacies le sont également, mais avec une présence plus importante dans le centre-ville. La localisation des médecins est plus agrégée que celle des pharmacies, mais

1. En cas de localisation imprécise d'équipements, ces derniers sont affectés par défaut au centroïde de l'Iris d'appartenance (zonage de l'Insee en "Îlots Regroupés pour l'Information Statistique", *cf.* https ://www.insee.fr/fr/metadonnees/definition/c1523).

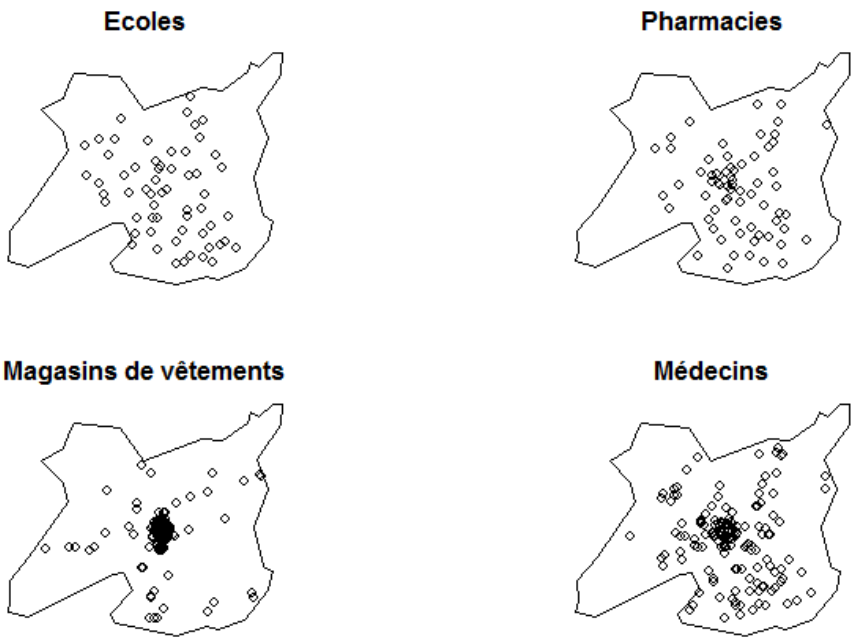

FIGURE 4.1 – Quatre exemples de localisation d'activités dans la commune de Rennes en 2015
Source : *Insee-BPE, calculs des auteurs*

elle est moindre que celle des magasins de vêtements. Ces premières conclusions sur la distribution des activités pourraient être complétées par des analyses spatiales plus avancées, par exemple en rapprochant ces données de la distribution de la population ou de l'accessibilité (proximité plus ou moins importante des grands axes de communication). Les méthodes présentées dans ce manuel permettent justement d'aller au-delà des conclusions apportées par ces premières cartes, certes informatives, mais insuffisantes pour caractériser et expliquer la localisation des entités étudiées.

Dans ce chapitre, nous avons fait le choix de ne pas traiter les méthodes qui discrétisent l'espace, c'est-à-dire des approches reposant sur un zonage d'étude (comme les zones d'emploi en France fondées sur les déplacements domicile-travail) ou un zonage administratif (comme les découpages de la Nomenclature des Unités Territoriales Statistiques - NUTS - d'Eurostat). Des ouvrages dédiés (COMBES et al. 2008) proposent une très bonne introduction en la matière pour tout lecteur intéressé. Ce chapitre se limitera aux méthodes qui tiennent compte de la position géographique exacte des entités étudiées. Notre choix est motivé par au moins deux raisons. La première est liée à l'accès à de telles données, à grande échelle, et au développement de moyens techniques adaptés pour les analyser de manière pertinente. Différents packages sont par exemple accessibles sous le logiciel R. La seconde est qu'en privilégiant des méthodes préservant la nature des données individuelles analysées (position dans l'espace, caractéristiques), le *Problème des Unités Spatiales Modifiables (Modifiable Areal Unit Problem - MAUP)*, bien connu des géographes (OPENSHAW et al. 1979a), sera évité. Le MAUP désigne le fait que la discrétisation de données initialement non agrégées créent potentiellement plusieurs biais statistiques liés à la position des frontières, au niveau d'agrégation, etc. (BRIANT et al. 2010).

4.1 Cadre d'analyse : les concepts fondamentaux

Cette section vise à définir les notions fondamentales sur lesquelles nous nous appuierons dans ce chapitre pour expliquer les méthodes statistiques d'analyse spatiale de données ponctuelles.

4.1.1 Configurations et processus

Pour étudier ces réalisations empiriques que sont les **configurations de points** (ou semis de points) on fait appel à la théorie des **processus ponctuels aléatoires** (*random point process*). Un processus ponctuel peut servir à générer aléatoirement une infinité de **réalisations**, partageant un certain nombre de propriétés. Usuellement, on note X le processus ponctuel et S la réalisation de ce processus. La modélisation des configurations de points fait appel à des méthodes inférentielles qui s'appliquent à des objets dont on n'observe qu'une seule réalisation. Par exemple, pour de nombreuses données, le statisticien ne dispose que d'un seul jeu de points observés à une date donnée. Ainsi, il n'y a qu'une seule répartition des médecins dans la ville de Rennes (voir figure 4.1), des arrêts de bus à Londres, des logements dans la Frise aux Pays-Bas ou des cinémas en Belgique à une date donnée. L'unicité de la réalisation ne doit cependant pas altérer notre analyse : on veillera par conséquent à ce que les données disponibles permettent d'avoir une bonne approximation du processus ponctuel qui l'a générée. Nous reviendrons sur ce point dans ce chapitre.

Définition 4.1.1 — Configuration de points. Dans ce chapitre, une configuration de n points notée $C = \{x_1, ..., x_n\}$ est un ensemble de points de $\mathbb{R}^2$: les objets sont localisés sur une carte. La théorie ne limite pas la dimension de l'espace mais les applications dans des espaces tridimensionnels sont rares, et presque inexistantes dans $\mathbb{R}^d, d > 3$. On note $n(C)$ le nombre de points de la configuration. On considère que les points ne sont pas dupliqués, car cela interdirait l'utilisation de bon nombre de méthodes. **L'ensemble des points contenus dans la région B est noté $C \cap B$, et $n(C \cap B)$ le nombre des points correspondant**.

Le processus X est **défini** si on connaît pour toute région B la loi de la variable aléatoire donnant le nombre des points $n(X \cap B)$, aussi noté $N(B)$ quand aucune confusion n'est possible. En général, on se limite aux processus qualifiés de localement finis, ceux pour lesquels $n(X \cap B) < +\infty, \forall B$.

4.1.2 Processus marqués

Une ou plusieurs caractéristiques peuvent être associées à chaque point. Nous appellerons ces caractéristiques marques du point. Dans ce cas, on parle de **processus ponctuel marqué** (*marked point pattern*). Ce formalisme a beaucoup été utilisé dans les études sur la forêt (voir par exemple MARCON et al. 2012).

Les marques retenues peuvent être qualitatives (différentes espèces d'arbres) ou quantitatives (diamètre du tronc, taille des arbres). Si nous reprenons l'exemple des commerces de vêtements, les marques qualitatives pourraient être le type de magasin (prêt-à-porter ou sur mesure) et les marques quantitatives, la surface du magasin ou le nombre de salariés. Les marques peuvent être plus sophistiquées. Par exemple, Florent Bonneu caractérise la répartition spatiale des sinistres dans la région de Toulouse en 2004 en retenant, pour chaque intervention de pompiers, la charge de travail associée (BONNEU 2007). Cette marque quantitative est obtenue en multipliant la durée de l'intervention et le nombre de pompiers mobilisés.

On se limitera dans un premier temps à des processus non marqués.

4.1.3 Fenêtre d'observation

L'espace pris en compte pour étudier la localisation des points, souvent appelé **fenêtre** (*window*) est dans bien des cas arbitraire. Les auteurs retiennent une aire d'étude carrée (MØLLER et al. 2014), rectangulaire (COLE et al. 1999), circulaire (SZWAGRZYK et al. 1993), une zone administrative (ARBIA et al. 2012) ou un zonage d'étude (LAGACHE et al. 2013).

Les indicateurs utilisés pour détecter les structures spatiales sous-jacentes se fondent sur une analyse du **voisinage des points** : on calcule par exemple pour tous les points étudiés le nombre

moyen de points voisins dans un rayon de 2 km, 4 km etc. La prise en compte des points localisés en bordure de l'espace d'intérêt peut être alors nécessaire. Le risque est en effet de sous-estimer le voisinage des points localisés sur le bord du domaine, une partie de leurs voisins étant localisée hors du domaine. Nous le constatons par exemple sur la figure 4.2. Supposons que le domaine étudié soit une parcelle carrée au sein d'une forêt et que les points représentent des arbres. Le voisinage du point i est décrit comme le disque de rayon r centré sur le point i. Si l'on cherche à connaître le nombre de voisins du point i, ne décompter que les points du disque inclus dans la parcelle sous-estimerait le nombre réel de ses voisins puisque qu'une partie du disque est située hors du domaine d'étude.

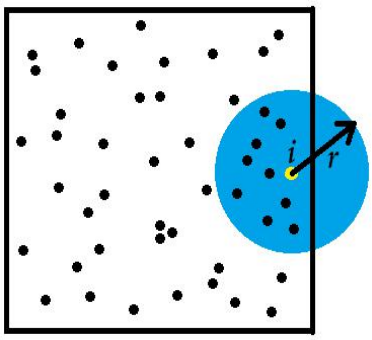

FIGURE 4.2 – Exemple d'effet de bord
Source : *calcul des auteurs.*

L'étude de MARCON et al. 2003 illustre par exemple l'importance d'une non-prise en compte de ce biais sur les estimations de la concentration des activités industrielles en France. Généralement, quel que soit le domaine d'application, ce biais potentiel est jugé suffisamment sévère pour que l'on recoure à **une technique correctrice prenant en compte les "effets de bord"**. Une littérature très importante traite de ces effets de bord et de leur correction (correction globale ou individuelle, création d'une zone-tampon autour du domaine, recours à une correction toroïdale [2], etc.). Le lecteur intéressé pourra se référer aux manuels classiques de statistique spatiale pour de plus amples développements (ILLIAN et al. 2008 ; BADDELEY et al. 2015b). D'un point de vue pratique, les logiciels de calculs (et notamment R) peuvent être utilisés pour traiter ces effets par différentes méthodes de correction. Un exemple sera proposé dans le chapitre 8 : "Lissage spatial".

4.2 Processus ponctuels : une présentation succincte

4.2.1 Le processus de Poisson homogène

Pour débuter, intéressons-nous au 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 :

$$P(U \in B) = \int_B f(u_1, u_2) du_1 du_2. \tag{4.1}$$

2. La correction toroïdale est applicable à une fenêtre rectangulaire. La fenêtre est repliée sur elle-même pour constituer un tore : une continuité est établie entre les limites droite et gauche (respectivement supérieure et inférieure) de la fenêtre qui n'a donc plus de bord.

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 $P(U \in B) = \int_B f(u_1,u_2)du_1du_2 = \frac{1}{|W|}\int_B du_1du_2 = \frac{|B|}{|W|}$.

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 :

$$P(n(X \cap B) = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.2}$$

avec $p = \frac{|B|}{|W|}$.

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 4.3, 1 000 points sont attendus sur une fenêtre d'observation 10 x 10.

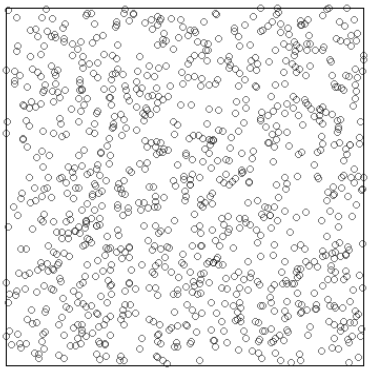

FIGURE 4.3 – Tirage de 1 000 points par un processus binomial uniforme
Source : *package spatstat, calculs des auteurs.*

```
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 ? 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. Ce n'est pas le cas pour le processus binomial (s'il y a k points dans la région B de W, il y en a $n-k$ dans le complémentaire).

L'homogénéité entraîne que le nombre des points attendus dans la région B soit proportionnel à sa surface, soit $\mathbb{E}[n(X \cap B)] = \lambda |B|$. λ est une constante correspondant au nombre moyen de points par unité de surface. La loi de Poisson, qui va servir à caractériser un processus CSR peut être introduite de façon heuristique à partir de la propriété d'indépendance. Celle-ci implique que tous les comptages sur des quadrillages sont indépendants, ceci quelle que soit la taille du carreau. Quand les carreaux, au nombre de m, deviennent extrêmement petits, la plupart d'entre eux ne

contiennent aucun point et quelques uns n'en contiennent qu'un seul. La probabilité qu'une région contienne plus d'un point devient négligeable. En faisant l'hypothèse d'indépendance, $n(X \cap B)$ est le nombre de succès issus d'un grand nombre d'essais indépendants, chaque essai ayant une très faible probabilité de succès. Ce nombre de succès suit une loi binomiale de paramètres m et $\lambda |B| / m$, qui tend vers la loi de Poisson de paramètre $\lambda |B|$ quand m devient grand :

$$P(n(X \cap B) = k) = e^{-\lambda |B|} \frac{\lambda^k |B|^k}{k!}. \tag{4.3}$$

On arrive donc à cette conclusion en partant des hypothèses d'homogénéité et d'indépendance.

Définition 4.2.1 — 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)** ;
- $n(X \cap B_1)$,..., $n(X \cap B_m)$ 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. D'autres propriétés en découlent. Tout d'abord, la superposition de processus de Poisson indépendants de paramètres λ_1 et λ_2 est un processus de Poisson de paramètre $\lambda_1 + \lambda_2$. Si on élimine des points de façon aléatoire avec une probabilité constante p dans un processus de Poisson (*thinned process* que l'on pourrait traduire par processus amaigri), le processus résultant est toujours un processus de Poisson de paramètre $p\lambda$, où p est le paramètre d'amaigrissement.

Le processus de Poisson homogène joue un rôle déterminant dans la modélisation des configurations de points [3]. De très nombreux processus spatiaux ont été définis, nous en donnerons quelques exemples dans ce chapitre. Le package *spatstat* permet de les implémenter. Par exemple, on utilisera la fonction `rpoispp` pour simuler les processus de Poisson homogènes. La figure 4.4 renvoie un tirage d'un processus de Poisson homogène sur une fenêtre d'observation 1 x 1 : 50 points sont attendus et les points sont répartis complètement aléatoirement sur la fenêtre.

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

4.2.2 L'intensité, propriété d'ordre 1

Les lois des processus sont très complexes (MØLLER 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.

3. Un peu comme la loi normale en statistique inférentielle classique (bien que ses propriétés la rapprochent plus de la loi uniforme).

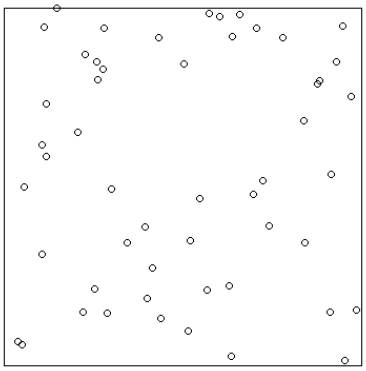

FIGURE 4.4 – Tirage de 50 points par un processus de Poisson homogène
Source : *package spatstat, calculs des auteurs.*

Définition 4.2.2 — 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 $\lambda(x)$ 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 :

$$\lambda(x) = \lim_{|dx| \to 0} \frac{\mathbb{E}[N(dx)]}{|dx|}. \tag{4.4}$$

Dans le cas où elle n'est pas constante, elle peut être **estimée avec les méthodes non paramétriques** utilisées pour l'estimation de la densité. Dans sa version la plus simple, sans correction des effets de bord, l'estimateur de l'intensité s'écrit : $\lambda(u) = \sum_{i=1}^{n} K(u - x_i)$, K désignant le noyau, qui peut être gaussien, ou à support fini (noyau d'Epanechnikov, noyau biweight de Tukey). Il doit vérifier $\int_{\mathbb{R}^2} K(u)du = 1$. Comme dans toutes les méthodes non paramétriques, **le choix du noyau a un impact limité**. En revanche **le choix de la bande passante est extrêmement important** (voir par exemple ILLIAN et al. 2008). On trouvera une présentation de ces méthodes d'estimation dans le chapitre 8 de ce manuel : "Lissage spatial". La fonction utilisée dans le logiciel R est `density` du package *spatstat*, qui permet de fournir des contours, des représentations 3D, des dégradés de couleur. Plusieurs exemples seront donnés dans la section 4.6.1 de ce chapitre.

4.2.3 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 $\lambda |B|$, mais $\mu(B)$ 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 $f(x) = \dfrac{\lambda(x)}{\int_B f(u)du}$.

On trouvera sur la figure 4.5 deux exemples de processus de Poisson inhomogène, caractérisés par leur fonction d'intensité (où x et y sont les coordonnées).

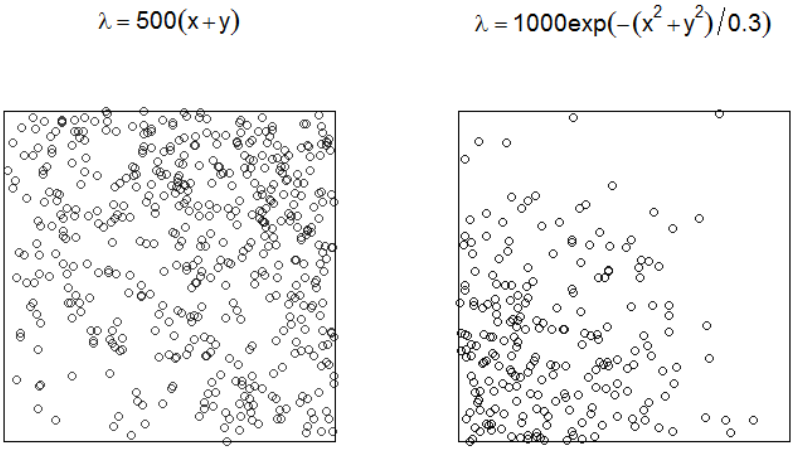

FIGURE 4.5 – Exemples de processus inhomogènes
Source : *package spatstat, calculs des auteurs.*

```
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))
```

4.2.4 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 :

$$var(n(X\cap B) = \mathbb{E}[n(X\cap B)^2] - \mathbb{E}[n(X\cap B)]^2 \tag{4.5}$$

$$cov\,[n(X\cap B_1), n(X\cap B_2)] = \mathbb{E}[n(X\cap B_1)n(X\cap B_2)] - \mathbb{E}[n(X\cap B_1)]\mathbb{E}[n(X\cap B_2)] \tag{4.6}$$

Définition 4.2.3 — Moment d'ordre 2 d'un processus. Plutôt que d'utiliser ces indicateurs, on définit le moment d'ordre deux de la façon suivante :

$$\nu_{|2|}(A\times B) = \mathbb{E}[n(X\cap A)n(X\cap B)] - \mathbb{E}\,[n(X\cap A\cap B)]\,, \tag{4.7}$$

qui vaut pour le processus de Poisson : $\lambda^2\,|A|\,|B|$. Lorsque cette mesure admet une densité, celle-ci, appelée intensité d'ordre 2 et notée λ_2 est définie de telle sorte que $\nu_{|2|}(C) = \int_C \lambda_2(u,v)dudv$.

Cette intensité du second ordre peut s'interpréter comme :

$$\lambda_2(x,y) = \lim_{|dx|\to 0|dy|\to 0,} \frac{\mathbb{E}\,[N(dx)N(dy)]}{|dx|\,|dy|}. \tag{4.8}$$

Les intensités du premier et du second ordres permettent de définir une fonction, appelée fonction de corrélation de paire de points de la façon suivante :

$$g_2(u,v) = \frac{\lambda_2(u,v)}{\lambda(u)\lambda(v)}. \tag{4.9}$$

Dans le cas d'un processus de Poisson homogène, $\lambda_2(u,v) = \lambda^2$, $g_2(u,v) = 1$.

Lorsqu'un processus est **stationnaire (au second ordre)** [4], l'intensité du second ordre n'est pas affectée par la translation et ne dépend que de la différence entre les points : $\lambda_2(x,y) = \lambda_2(x-y)$.

Lorsqu'il est en plus **isotrope**, le processus n'est pas affecté par la rotation et l'intensité de second ordre ne dépend que de la distance entre x et y. Notons que la stationnarité au second ordre et l'isotropie sont indispensables pour de nombreux outils de statistique spatiale.

4.3 Des processus ponctuels aux répartitions observées de points

4.3.1 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 4.6. Ces distributions de points sont obtenues à partir de processus ponctuels connus simulés à l'aide du package *spatstat*.

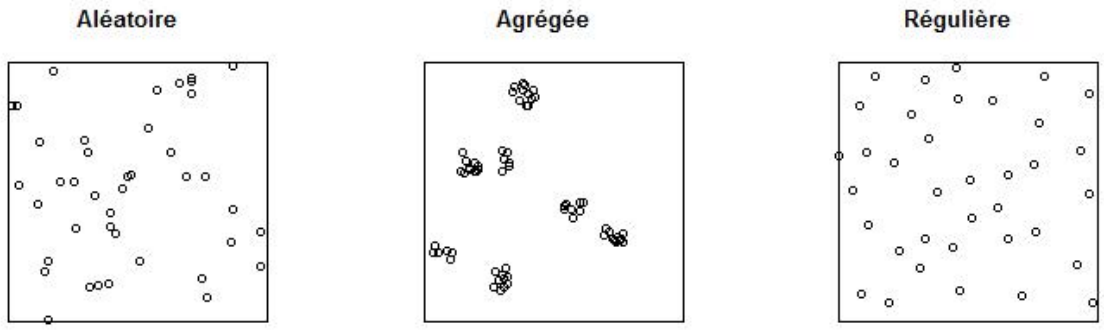

FIGURE 4.6 – Les trois configurations classiques de points
Source : *package spatstat, calculs des auteurs.*

```
library("spatstat")
par(mfrow=c(1, 3))
plot(rpoispp(50), main="Aléatoire")
plot(rMatClust(5, 0.05, 10), main="Agrégée")
plot(rMaternII(200,0.1), main="Régulière")
par(mfrow=c(1,1))
```

La configuration **complètement aléatoire** est centrale pour la théorie. Toutes les configurations de points, en tant que réalisation d'un processus ponctuel, sont aléatoires mais celle-ci correspond

4. Le terme stationnaire sans plus de précision est souvent employé pour les processus d'intensités d'ordres 1 et 2 constantes ; la stationnarité au premier ordre est synonyme d'homogénéité.

à une distribution "complètement au hasard" de points sur une surface : les points sont localisés partout avec la même probabilité et indépendamment les uns des autres. Cette configuration correspond à un tirage d'un processus de Poisson homogène. Il n'y dans ce cas aucune interaction entre les points mais seule l'utilisation d'indicateurs permet de juger si la distribution observée s'écarte *significativement* d'une distribution complètement aléatoire. En effet, il est délicat à l'œil nu d'identifier une telle configuration. Dans cet exemple, nous avons retenu la fonction `rpoispp` dans le package *spatstat* pour simuler les processus de Poisson homogène.

La seconde répartition de points est dite **régulière (ou répulsive)** : on peut penser à la répartition spatiale des arbres dans un verger ou le long des rues en ville, à celle des transats sur un plage etc. Dans une telle configuration, les points sont *plus régulièrement espacés* qu'ils ne le seraient sous une distribution complètement aléatoire. Les points se repoussent et créent une distribution de points dispersée. On peut retrouver un phénomène de dispersion pour certaines activités commerciales, comme les commerces de détail de carburants sur Lyon (MARCON et al. 2015a). Les contraintes de localisations peuvent également créer des dispersions, la distribution géographique des chambres des représentants aux États-Unis en sont un bon exemple (HOLMES et al. 2004). Nous avons représenté sur le graphique de droite de la figure 4.6 une distribution de points dispersée obtenue à partir d'un tirage d'un processus de Matern. Plus précisément, deux exemples simples de processus répulsifs sont fournis par les processus de Matern I et II (voir BADDELEY et al. 2015b). Dans le processus I, tous les couples de points situés à des distances inférieures à un seuil r sont supprimés. Dans le processus II, chaque point est marqué par un temps d'arrivée, variable aléatoire dans $[0,1]$. Les points situés à une distance inférieure à r d'un point arrivé antérieurement sont supprimés. À l'aide du package *spatstat*, les fonctions `rMaternI` et `rMaternII` sont disponibles pour simuler ces deux processus de Matern. Dans l'exemple donné sur la figure 4.6, nous avons retenu une réalisation d'un tirage d'un processus de Matern de type II obtenu grâce à ce package. Il est à noter que d'autres distributions dispersées peuvent être observées : intuitivement par exemple un phénomène de dispersion est observable pour une distribution de points localisés à l'intersection d'une maille en "nid d'abeille" : dans ce cas la distance entre les points est maximale (et elle est plus importante que si la distribution était aléatoire).

Enfin, la dernière configuration possible est qualifiée d'**agrégée**. Dans ce cas, une interaction entre les points est mise en évidence, ils s'attirent, créant des agrégats : une concentration géographique sera alors détectée. En se reportant à la figure 4.1 de l'introduction, il semble que les magasins de vêtements à Rennes sont essentiellement localisés au cœur de la ville. Ce constat pourrait être partagé avec d'autre types de commerces, comme pour l'habillement en magasins spécialisés à Lyon (MARCON et al. 2015a). Une configuration agrégée correspond par exemple au cas théorique central de la figure 4.6 qui est obtenu par un tirage d'un processus de Matern à clusters. L'idée de ce processus pour simuler des agrégats est assez intuitive. Autour de chaque point "parent", dans un disque de rayon r, des points "descendants" sont répartis de façon uniforme. Dans le package *spatstat*, la fonction `rMatClust` permet de simuler des réalisations de processus de Matern à clusters. Nous avons retenu cette fonction pour obtenir la distribution agrégée de la figure 4.6. Nous avons alors notamment spécifié l'intensité du processus de Poisson pour les points parents (égale à 5) et le nombre moyen de points descendants (10) tirés autour des points parents dans un disque de rayon r (égal à 0.05).

4.3.2 Mises en garde

Ces structures spatiales (agrégées, aléatoires ou dispersées) ont une interprétation très intuitive sous l'hypothèse de stationnarité du processus : en comparant les distributions de points observées à une distribution aléatoire, il semble aisé de détecter les interactions de répulsion ou d'attractions à l'origine de phénomènes de dispersion ou de concentration spatiale.

Il ne faut toutefois pas faire de conclusions trop hâtives car il faut bien garder à l'esprit que les mêmes structures agrégées ou dispersées peuvent être obtenues avec un processus de Poisson inhomogène dans lequel l'intensité du processus varie dans l'espace mais les points sont indépendants les uns des autres (voir figure 4.5). Une seule observation de la configuration de points ne permet pas de distinguer les propriétés de premier et de second ordres d'un processus en absence d'informations supplémentaires comme celles apportées par un modèle liant une covariable à l'intensité. ELLISON et al. 1997, ont montré que des avantages naturels (impliquant une plus grande intensité) ont un effet sur la localisation des entreprises non discernable de celui des externalités positives (générant l'agrégation) : la confusion entre les deux propriétés peut concerner aussi les processus.

Une dernière mise en garde concerne l'homogénéité. En effet, dans un premier temps, les méthodes développées en statistiques spatiales ont consisté à tester l'existence d'agrégation ou de répulsion, en assumant l'homogénéité du processus : il s'agissait donc de tester une configuration de points contre l'hypothèse nulle d'une distribution complètement aléatoire (CSR). Pour analyser de tels jeux de données, des mesures comme la fonction originelle K proposée par B.D. Ripley (largement employée dans la littérature statistique) sont adéquates. En revanche, si l'hypothèse nulle d'une distribution de points complètement aléatoire est jugée trop forte, d'autres fonctions doivent être privilégiées. C'est par exemple le cas pour l'étude des tremblements de terre (VEEN et al. 2006). Sur la figure 4.7 sont représentés 5 970 épicentres de séismes en Iran survenus entre 1976 et 2016 (leur magnitude était supérieure à 4.5). Ces données sont issues du package *etas*.

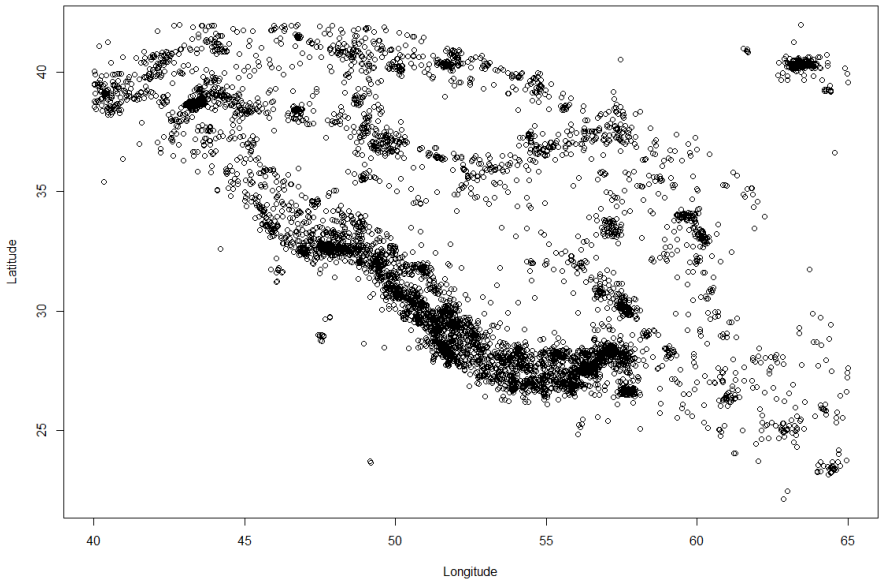

FIGURE 4.7 – Localisation de 5 970 épicentres de séismes en Iran survenus de 1976 à 2016
Source : *package etas, calculs des auteurs.*

```
library("ETAS")
data(iran.quakes,package =  "ETAS")
plot(iran.quakes$lat~iran.quakes$long , xlab="Longitude", ylab="Latitude")
```

On constate alors aisément que retenir une référence à l'homogénéité de l'espace n'est pas optimale car il existe des prédispositions géologiques dans ce cas. La fonction K de B.D. Ripley serait inadaptée pour analyser ce type de données et d'autres outils doivent être retenus comme

la fonction *K-inhomogène* de BADDELEY et al. 2000, que nous présenterons dans ce chapitre. DURANTON et al. 2005 ont par ailleurs également souligné cette limite d'homogénéité de l'espace pour analyser la distribution des activités industrielles et ont proposé une nouvelle fonction K_d.

Une bonne maîtrise des fonctions disponibles est par conséquent indispensable pour caractériser *correctement* une distribution de points. Ce sera l'objet de la section suivante.

4.4 Quels outils statistiques mobiliser pour étudier les configurations de points ?

La réponse à cette question n'est malheureusement pas immédiate. Pour la traiter, il convient d'analyser précisément la question à laquelle on tente de répondre avec des mesures fondées sur les distances (notamment en ce qui concerne la valeur de référence) et d'étudier les propriétés des fonctions. Pour comprendre précisément ce point et donc la difficulté liée au choix de la mesure, cette section débutera par une présentation de la fonction originelle K de Ripley et de développements importants issus de ce travail (sections 4.4.1 et 4.4.2). Puis, nous ferons un point d'étape pour mieux expliciter les déterminants du choix de la mesure (section 4.4.3). Nous verrons alors les avantages et les inconvénients des mesures existantes. Pour une large revue de la littérature ou une comparaison approfondie et plus complète des mesures, on se reportera à l'ouvrage de BADDELEY et al. 2015b ou à la typologie des mesures fondées sur les distances proposée par MARCON et al. 2017.

4.4.1 La 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.

Définition 4.4.1 — Fonction K de Ripley. Son estimateur s'écrit de la façon suivante :

$$\hat{K}(r) = \frac{|W|}{n(n-1)} \sum_i \sum_{j \neq i} \mathbf{1}\left\{\left\|x_i - x_j\right\| \leq r\right\} c(x_i, x_j; r), \tag{4.10}$$

où n est le nombre total de points sur la fenêtre d'observation, $\mathbf{1}\left\{\left\|x_i - x_j\right\| \leq r\right\}$ est une indicatrice qui vaut 1 si les points i et j sont à une distance au plus égale à r et 0 sinon. $c(x_i, x_j; r)$ 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 4.8 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 `Kest`. Sur la figure 4.8, 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 4.8. 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 4.8.

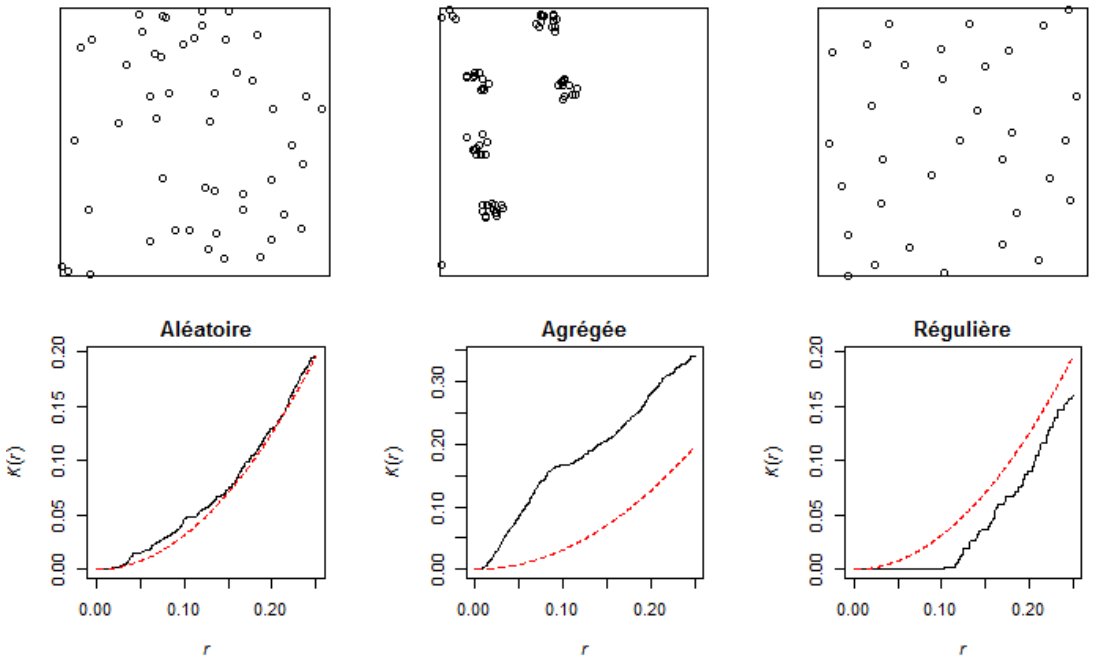

FIGURE 4.8 – Les fonctions K des trois configurations-types
Source : *Package spatstat, calculs des auteurs.*

```
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))
```

Soulignons pour terminer quelques points importants.

Tout d'abord, la fonction K est définie sous l'hypothèse - forte - de stationnarité. Dans le cas de processus de Poisson inhomogènes, l'écart avec la fonction empirique peut être dû à la variation d'intensité plus qu'à un phénomène d'attraction, c'est-à-dire lié à la propriété de second ordre.

De même, l'interprétation est sujette aux mêmes questions qu'en statistique "classique". La corrélation n'entraîne pas la causalité. Une absence de corrélation n'entraîne pas non plus forcément l'indépendance.

De plus, il faut tenir compte du caractère cumulatif de la fonction K. Une grande valeur de K à la distance r_0 peut être due à la conjonction de phénomènes à plus petites distances, alors qu'aucune interaction n'existe entre points distants de r_0.

Notons qu'il existe un **lien entre la fonction K et la fonction de corrélation de paire de points**. On peut l'approcher de la façon suivante : on trace deux cercles concentriques de rayon r et $r+h$, et on compte les points qui se trouvent dans l'anneau ainsi défini. Le nombre attendu est $\lambda K(r+h)-\lambda K(r)$ Si on standardise l'expression par la valeur attendue sur l'anneau pour un processus de Poisson, on obtient :

$$g_h(r) = \frac{\lambda K(r+h)-\lambda K(r)}{\lambda\pi(r+h)^2-\lambda\pi r^2} = \frac{K(r+h)-K(r)}{2\pi rh+\pi h^2}. \tag{4.11}$$

Si on fait tendre h vers 0, $g(r) = \frac{K'(r)}{2\pi r}$ ou $K(r) = \int_0^1 sg(s)ds$, le lien entre la fonction g et la fonction K est donc clair.

Enfin, les valeurs renvoyées par la fonction K permettent de détecter d'éventuelles interactions entre les points pour chacune des distances étudiées, sur *l'ensemble* du territoire analysé. Toutefois, il peut être intéressant d'avoir de l'information localement, comme pour les modèles sur données surfaciques pour lesquels on calcule à côté des indicateurs d'autocorrélation spatiale (comme celui de Moran) des indicateurs locaux appelés LISA (voir chapitre 3 : "Indices d'autocorrélation spatiale"). Dans les modèles ponctuels, **il existe aussi des indicateurs locaux construits sur le principe des indicateurs de Ripley**. On calcule pour chaque point un indicateur $\widehat{K}(r,x_i)$. Les seules paires de points prises en compte sont celles qui contiennent le point x_i. On peut alors représenter graphiquement une des valeurs locales ou l'ensemble des valeurs. Les points qui se distinguent peuvent être repérés de façon graphique ou éventuellement en utilisant des méthodes d'analyse fonctionnelle des données.

La fonction L de Besag 1977.

L'intérêt particulier de la fonction de Ripley et plus généralement des méthodes fondées sur les distances réside dans le fait qu'elles analysent l'espace étudié en parcourant *toutes les distances*

et en ne retenant pas qu'un seul ou quelques niveaux géographiques. Le semis de points est très précisément étudié et aucune distance d'analyse n'est omise. Par conséquent, **seules ces méthodes permettent de détecter *exactement* à quelle(s) distance(s) les phénomènes d'attractions ou de dispersions sont observables, sans biais d'échelle lié à un zonage prédéfini.** S'il y a, par exemple, des agrégats d'agrégats dans les données spatialisées, de telles fonctions peuvent détecter les distances auxquelles se produisent les concentrations spatiales : à la taille de l'agrégat et à la distance entre les agrégats. Les structures spatiales plus complexes pourront aussi être détectées comme des phénomènes multiples d'agglomération pour certaines distances et de répulsion pour d'autres distances (ce sera le cas si plusieurs agrégats sont régulièrement espacés par exemple). Un intérêt additionnel est de pouvoir comparer les valeurs retournées par les fonctions entre plusieurs distances. La fonction K le permet. Dans la version originale de la fonction K, il est peu commode de comparer directement les valeurs estimées pour plusieurs rayons car la valeur de référence, πr^2, nécessite de nouveaux calculs (les comparaisons graphiques hyperboliques n'étant pas immédiates). Comme nous allons le voir, ce point a été l'une des motivations pour apporter des développements à la fonction originelle de Ripley.

Deux transformations de la fonction de Ripley sont fréquemment utilisées. Il n'est pas rare de trouver dans la littérature statistique des applications avec ces variantes plutôt que la fonction K originale (par exemple ARBIA 1989 concernant la distribution des entreprises industrielles, GOREAUD et al. 1999 concernant la distribution des arbres ou encore FEHMI et al. 2001 pour des plantes). La première variante est la fonction $L(r)$ proposée par Besag (BESAG 1977) est définie par : $L(r) = \sqrt{\frac{K(r)}{\pi}}$, qui vaut dans un processus aléatoire $L_{Pois}(r) = r$. Avec le package *spatstat*, la fonction L peut être calculée en utilisant la fonction `Lest`. Une autre version possible est $L(r) - r$, que l'on compare à 0 en cas de répartition complètement aléatoire. Les deux avantages à ces variantes sont d'une part une variance plus stable (GOREAUD 2000) et, d'autre part, des résultats quasiment immédiatement interprétables (MARCON et al. 2003). Ainsi par exemple, en retenant la seconde variante, si la fonction $L(r) - r$ atteint la valeur 2 pour un rayon r égal à 1, cela veut dire qu'en moyenne il y a autant de voisins dans un rayon de 1 autour de chaque point dans cette configuration qu'il n'y en aurait dans un rayon de 3 (=2+1) si la distribution était homogène. Une meilleure normalisation est $\frac{K(r)}{\pi r^2}$ dont la valeur attendue est 1 et la valeur empirique le rapport entre le nombre de voisins observés et attendus (MARCON et al. 2017).

À titre d'illustration, nous avons repris l'exemple d'une distribution agrégée et donné en figure 4.9 les quatre résultats estimés des fonctions K, L, $L - r$ et $K(r)/\pi r^2$ pour cette distribution.

```
library("spatstat")
AGRE <- rMatClust(10, 0.08, 4)
K <- Kest(AGRE,correction="isotropic")
L <- Lest(AGRE,correction="isotropic")
par(mfrow=c(2, 2))
plot(K, legend=FALSE, main="") # K
plot(L, legend=FALSE, main="") # L classique
plot(L, .-r ~ r, legend=FALSE, main="") # L définie comme L(r)-r
plot(K, ./(pi*r^2) ~ r, legend=FALSE, main="") # K(r)/(pi r^2)
par(mfrow=c(1, 1))
```

La fonction D de DIGGLE et al. 1991

Les fonctions K et L peuvent être retenues dans les études si l'hypothèse d'homogénéité de l'espace analysé est vérifiée. Une autre variante de la fonction K permet de prendre en compte la non-homogénéité de l'espace : il s'agit de la fonction D proposée par DIGGLE et al. 1991. Cet

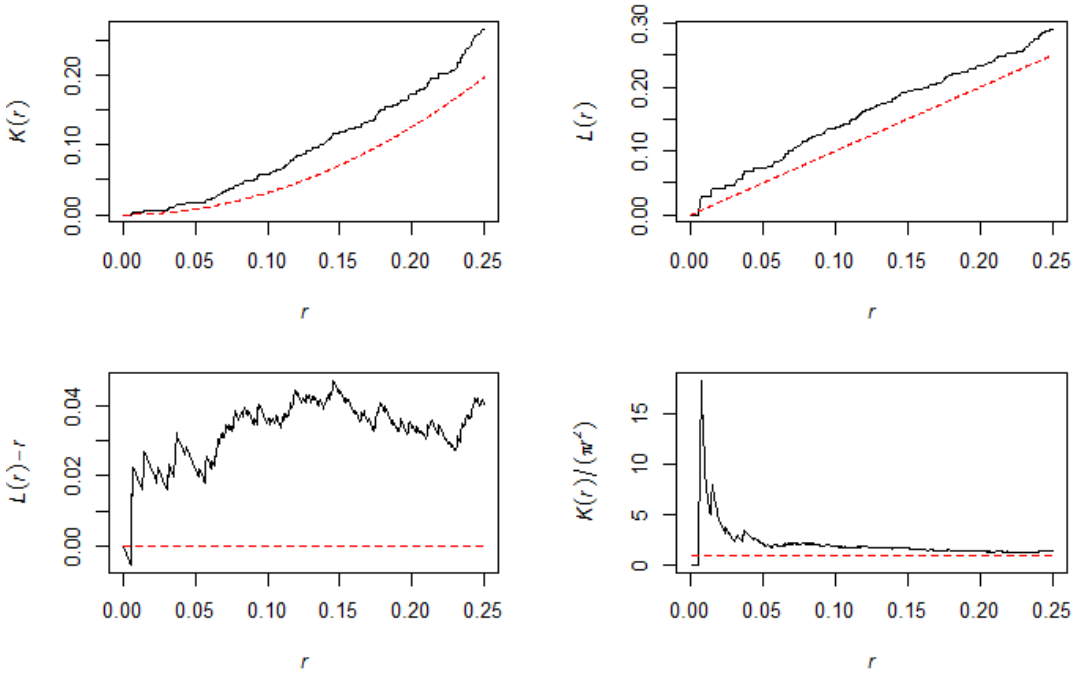

FIGURE 4.9 – Représentation des fonctions K, L, $L-r$ et $K(r)/\pi r^2$ pour l'exemple d'une distribution agrégée
Source : *package spatstat, calculs des auteurs.*

indicateur est directement issu des préoccupations des épidémiologistes, cherchant à comparer la concentration des "cas" (enfants atteints d'une maladie rare dans le Nord de la Grande-Bretagne) et celle des "contrôles" (enfants sains sur la même zone d'étude). Cette fonction se définit très simplement comme la différence entre deux fonctions K de Ripley : celle des cas et celle des contrôles. On obtient :

$$D(r) = K_{cas}(r) - K_{controles}(r) \tag{4.12}$$

La fonction D permet de confronter les distributions de deux sous-populations. Intuitivement, on comprend que si les cas sont plus localisés que les contrôles, une concentration spatiale des cas sera détectée par la fonction D. Inversement, si la distribution des cas est moins concentrée que celle des contrôles, la fonction D détectera que les cas seront spatialement plus dispersés que les contrôles. L'intérêt de recourir à cette fonction est de pouvoir détecter des écarts de la distribution étudiée par rapport à une distribution de référence. Cela peut être par exemple intéressant si l'on souhaite savoir si un certain type d'habitation est géographiquement plus concentré que les autres types d'habitations ou si un type de commerce est plus aggloméré au sein des villes que les autres types de commerces, etc. La différence de deux fonctions K revient à avoir une valeur de comparaison pour D égale à 0, pour tous les rayons d'étude. Toutefois, il est impossible de comparer les valeurs estimées de D du fait du changement de la sous-population de référence. Cette fonction D peut être implémentée dans le logiciel R à l'aide du package *dbmss* : on utilisera alors la fonction nommée `Dhat`. Tout comme la fonction K, il est également possible d'associer un niveau de significativité des résultats en procédant à un étiquetage aléatoire des points (voir infra). On privilégiera alors la fonction nommée `DEnvelope`. Diverses applications sont disponibles dans la littérature concernant la concentration spatiale des activités économiques (comme SWEENEY et al. 1998). Le lecteur intéressé pourra également trouver une variante de la fonction D proposée par ARBIA et al. 2008.

La fonction K_{inhom} de BADDELEY et al. 2000

K_{inhom}, la version de la fonction K de Ripley en espace inhomogène a été proposée par BADDELEY et al. 2000. La valeur estimée de K_{inhom} fait par conséquent intervenir les valeurs estimées de l'intensité (l'hypothèse d'une intensité identique en tout point du territoire étudié doit être relâchée puisque l'espace considéré n'est plus homogène). En notant $\widehat{\lambda}(x_i)$ l'estimation du processus autour du point i et $\widehat{\lambda}(x_j)$ l'estimation du processus autour du point j, la fonction cumulative K_{inhom} peut être définie comme suit :

$$\widehat{K}_{inhom}(r) = \frac{1}{D}\sum_i \sum_{j\neq i} \frac{\mathbf{1}\left\{\left\|x_i - x_j\right\| \leq r\right\}}{\widehat{\lambda}(x_i)\widehat{\lambda}(x_j)} e(x_i, x_j; r) \tag{4.13}$$

avec $D = \frac{1}{|W|}\sum_i \frac{1}{\lambda(x_i)}$.

On montre que dans le cas d'un processus inhomogène : $K_{inhom,pois}(r) = \pi r^2$. Les estimations de K_{inhom} s'interprètent donc de la même façon que dans le cas de la fonction K homogène. D'un point de vue pratique, la fonction nommée `Kinhom` dans le package *spatstat* permet de calculer la fonction K_{inhom}.

Sur le plan théorique, le traitement des processus non stationnaires pourrait être considéré comme résolu, mais la difficulté pratique réside dans l'estimation des densités locales, par la méthode des noyaux. Au delà des difficultés techniques, l'impossibilité théorique de séparer à partir d'une seule observation ce qui tient au phénomène du premier ordre (intensité) et ce qui tient à l'agrégation du phénomène étudié se traduit par des biais importants quand la fenêtre utilisée pour estimer les densités locales est du même ordre de grandeur que la valeur de r considérée. Les applications empiriques de cet indicateur sont encore peu nombreuses (BONNEU 2007 ; ARBIA et al. 2012).

4.4.2 Comment tester la significativité des résultats ?

Plusieurs méthodes statistiques permettent de juger de la significativité des résultats obtenus par les différentes fonctions précédemment présentées. La technique la plus courante étant le recours à la simulation d'un intervalle de confiance par la méthode de Monte Carlo, nous commencerons par l'expliciter.

Méthodes de Monte Carlo

Sans connaissance de la distribution théorique de la fonction K de Ripley sous l'hypothèse nulle d'une distribution complètement aléatoire, **la significativité de la différence entre les valeurs observées et les valeurs théoriques est testée par la méthode de Monte Carlo**. Cette méthode peut être utilisée pour déterminer les intervalles de confiance de toutes les fonctions dérivées de K présentées. On désignera donc de façon générique la fonction d'intérêt par S. Pour cela, on procède de la manière suivante :

1. On génère un nombre q de jeux de données correspondant à l'hypothèse nulle du test. Si l'hypothèse nulle est un processus complètement aléatoire, on génère q processus de Poisson d'intensité correspondant à celle de la configuration de points testée.
2. On définit les courbes $U(r) = max\left\{S^{(1)}(r), ..., S^{(q)}(r)\right\}$ et $L(r) = min\left\{S^{(1)}(r), ..., S^{(q)}(r)\right\}$ qui permettent de définir une enveloppe, représentée en gris dans les graphiques réalisés avec le logiciel R.
3. Pour un test bilatéral, l'enveloppe définie correspond à un risque de première espèce $\alpha = \frac{2}{q+1}$, soit 39 simulations pour un test de niveau 5 %.

Pour chacune des fonctions, on peut construire cette enveloppe qui permet de comparer la statistique construite à partir des données à des statistiques issues de la simulation d'un processus aléatoire correspondant à l'hypothèse nulle testée (un processus de Poisson homogène de même intensité pour la fonction K). Dans le package *spatstat*, c'est la commande générique `envelope` qui permet de faire les simulations de Monte Carlo, et de construire les courbes correspondant aux valeurs supérieures et inférieures de l'enveloppe. L'enveloppe ne doit pas être interprétée comme un intervalle de confiance autour de l'indicateur étudié : elle indique les valeurs critiques du test. Pour donner un exemple simple, utilisons un jeu de données `paracou16` relatif à la localisation des arbres dans un dispositif forestier de Paracou, en Guyane française. Ces données sont disponibles dans le package *dbmss*. Calculons l'intervalle de confiance associée à la fonction K avec 39 simulations. Sur la figure 4.10, la courbe de K obtenue est indiquée (trait plein noir), la courbe en pointillés rouges représente le milieu de l'intervalle de confiance et les deux bornes de l'enveloppe sont données ainsi que l'enveloppe (courbes et enveloppe grises). On constate que, jusqu'à une distance proche de 2 mètres, on ne peut rejeter l'hypothèse nulle de processus CSR à partir de la fonction de Ripley.

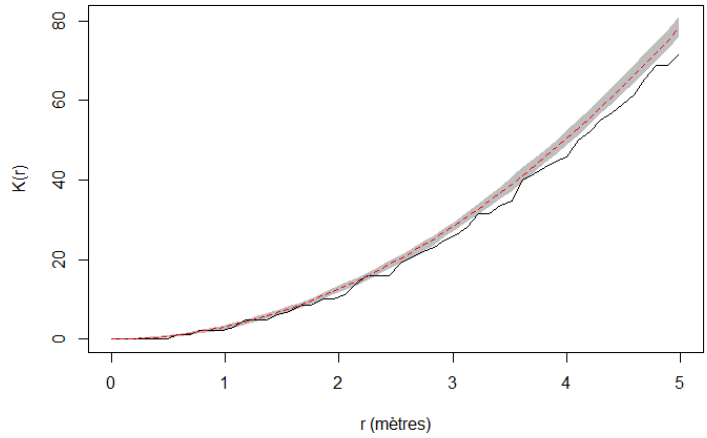

FIGURE 4.10 – Exemple d'enveloppe de confiance pour la fonction K
Source : *package spatstat et dbmss, données "paracou16", calculs des auteurs.*

```
library("dbmss")
# Enveloppe calculée à l'aide du package dbmss, données : 2426 points.
env <- KEnvelope(paracou16, NumberOfSimulations=39)
plot(env,legend =FALSE, main="", xlim=c(0,5), xlab = "r (mètres)", ylab= "K
    (r)")
```

Avec l'augmentation de la puissance de calcul, une pratique répandue consiste à simuler un grand nombre de fois l'hypothèse nulle (1 000 ou 10 000 fois plutôt que 39) et à définir l'enveloppe à partir des quantiles $\alpha/2$ et $1-\alpha/2$ des valeurs de $S(r)$.

Le test est répété à chaque valeur de r : le risque de rejeter par erreur l'hypothèse nulle est donc augmenté au-delà de α. Cette sous-estimation du risque de première espèce n'est pas très grande parce que les valeurs des fonctions cumulatives sont très autocorrélées. Le test est donc couramment utilisé sans précaution particulière. Pourtant, des auteurs comme DURANTON et al. 2005 jugent ce point sérieux et tentent d'y remédier. Une méthode permettant de corriger le problème est présentée dans MARCON et al. 2010 et implémentée dans le package *dbmss* sous le nom d'intervalle de confiance global de l'hypothèse nulle (par opposition aux intervalles de confiance locaux, calculés

à chaque valeur de r). Elle consiste à éliminer itérativement une partie α des simulations dont une valeur au moins contribue à $U(r)$ ou $L(r)$.

Une remarque importante : lorsque l'on calcule une enveloppe sous R, elle est systématiquement associée à une fonction particulière. Dit autrement, les routines de calculs disponibles dans les packages tiennent compte des spécificités des fonctions : les intervalles de confiance sont donc simulés en considérant l'hypothèse nulle correcte. Par exemple, pour simuler l'enveloppe de la fonction K, l'hypothèse nulle est construite à partir d'une distribution de points répartis aléatoirement et indépendamment sur le domaine d'étude. En revanche, pour la fonction D de DIGGLE et al. 1991, élaborer un intervalle de confiance avec les mêmes hypothèses que pour la fonction K serait incorrect. Il faut, pour D, tenir compte des variations d'intensités sur le domaine étudié. Comment procéder ? Rappelons-nous que l'hypothèse nulle correspond pour cette fonction à une situation où la sous-population des cas et la sous-population des contrôles ont la même répartition spatiale. La solution suggérée par DIGGLE et al. 1991 est de procéder à un étiquetage aléatoire (*random labelling*) c'est-à-dire attribuer à chaque simulation, une étiquette "cas" ou "contrôle" pour chaque localisation. Cette permutation aléatoire des étiquettes sur les localisations inchangées est une technique assez intuitive qui sera d'ailleurs reprise pour élaborer des intervalles de confiance d'autres fonctions que nous étudierons dans la section 4.5. Sous R, les packages *spatstat* ou *dbmss* des options pour le calcul des fonctions permettent de simuler cette hypothèse d'étiquetage aléatoire.

Tests analytiques

Les tests analytiques sont peu nombreux dans la littérature et très peu appliqués dans les études, même s'ils présentent l'avantage d'économiser les temps de calculs des intervalles de confiance. Pour K par exemple, des tests analytiques existent sur des domaines d'étude simples (HEINRICH 1991). Dans le cas particulier du test du caractère CSR dans une fenêtre rectangulaire, Gabriel Lang et Éric Marcon ont développé récemment un test statistique classique (LANG et al. 2013) disponible dans la fonction `Ktest` du package *dbmss* (MARCON et al. 2015b). Il retourne la probabilité de rejeter par erreur l'hypothèse nulle d'une distribution complètement aléatoire à partir d'une configuration de points, sans recourir aux simulations : la distribution de la fonction K non corrigée des effets de bord suit en effet une distribution asymptotiquement normale de variance connue. Le test est utilisable à partir de quelques dizaines de points. Il est à noter que tels tests pour des fonctions moins connues sont également proposées dans la littérature (JENSEN et al. 2011).

4.4.3 Point d'étape et mise en évidence de propriétés importantes pour de nouvelles mesures

Les mesures issues de la fonction K de Ripley sont utiles dans de nombreuses configurations pour expliquer les interactions entre les points étudiés. Nous avons d'ailleurs donné de nombreuses références dans des domaines d'application divers. Toutefois, des développements spécifiques peuvent encore être envisagés pour répondre à certaines questions, comme pour la localisation des activités économiques. Pour comprendre ce point, nous allons réfléchir aux atouts et aux limites des mesures issues de la fonction K de Ripley dans ce cadre d'analyse.

Point d'étape : les fonctions dérivées de la fonction K de Ripley sont-elles adaptées pour décrire la concentration spatiale des activités économiques ?

Les outils statistiques présentés dans les sections précédentes sont riches, mais leur utilisation pour appréhender des données comme celles des équipements ou des entreprises ne va pas de soi. Pour s'en convaincre, revenons aux exemples de l'introduction (les quatre équipements) et retenons la fonction K de Ripley pour caractériser les structures spatiales de chacun de ces équipements. Les résultats sont donnés sur la figure 4.11 : la fonction estimée de K de Ripley est représentée en trait plein, les intervalles de confiance obtenus à partir de 99 simulations par la zone grisée, le centre de l'intervalle de confiance est indiqué par la courbe en pointillés et les effets de bord ont été

calculés par la méthode de Ripley. Cette correction des effets de bord repose sur l'idée que, pour un point donné, la partie de la couronne hors du domaine (*cf.* figure 4.2) contient la même densité de voisins que la partie située à l'intérieure du domaine d'étude. Cette hypothèse est acceptable car, rappelons-le, nous considérons dans le cas de fonction K de Ripley une distribution complètement aléatoire de points. [5]

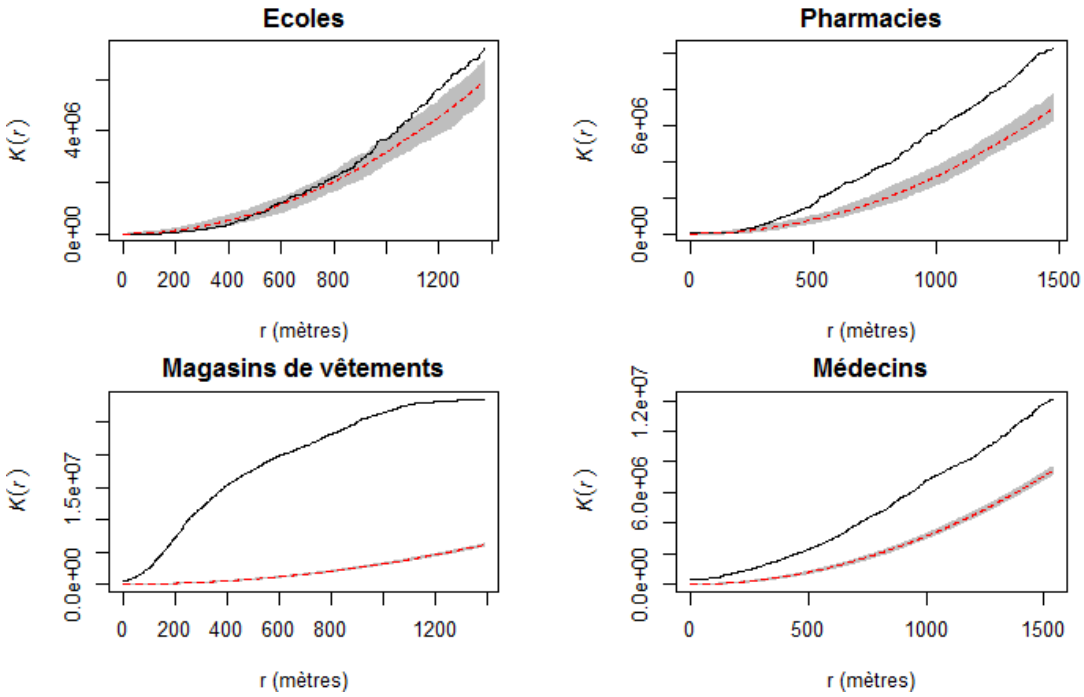

FIGURE 4.11 – Fonctions de Ripley pour les quatre équipements
Source : *Insee-BPE, packages spatstat et dmbss, calculs des auteurs.*

```
library("dbmss")
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))
bpe_eco<- bpe[bpe $TYPEQU=="C104", ]
bpe_pha<- bpe[bpe $TYPEQU=="D301", ]
bpe_vet<- bpe[bpe $TYPEQU=="B302", ]
bpe_med<- bpe[bpe $TYPEQU=="D201", ]

ecole <- as.ppp(bpe_eco[ ,c ("lambert_x", "lambert_y")], owin(c(min(bpe_eco
    [,"lambert_x"]), max(bpe_eco[,"lambert_x"])),c (min(bpe_eco[,"lambert_y
    "]),max (bpe_eco[,"lambert_y"]))))
bpe_ecole_wmppp <- as.wmppp(ecole)
pharma <- as.ppp(bpe_pha[ ,c ("lambert_x", "lambert_y")], owin(c(min(bpe_
    pha[,"lambert_x"]),max (bpe_pha[,"lambert_x"])),c (min(bpe_pha[,"
    lambert_y"]),max (bpe_pha[,"lambert_y"]))))
bpe_pharma_wmppp <- as.wmppp(pharma)
vetem <- as.ppp(bpe_vet[ ,c ("lambert_x", "lambert_y")], owin(c(min(bpe_vet
    [,"lambert_x"]), max(bpe_vet[,"lambert_x"])),c (min(bpe_vet[,"lambert_y
```

5. Techniquement, supposons qu'un voisin d'un point donné est situé sur la couronne de largeur (à l'intérieur du domaine). La correction de Ripley consiste à attribuer à ce voisin un poids égal à l'inverse du rapport du périmètre de la couronne sur le périmètre total de la couronne.

```
    "]), max(bpe_vet[,"lambert_y"])))
bpe_vetem_wmppp <- as.wmppp(vetem)
medecin <- as.ppp(bpe_med[ , c("lambert_x", "lambert_y")], owin(c(min(bpe_
    med[,"lambert_x"]), max(bpe_med[,"lambert_x"])), c(min(bpe_med[,"
    lambert_y"]), max(bpe_med[,"lambert_y"]))))
bpe_medecin_wmppp <- as.wmppp(medecin)

kenv_ecole <- KEnvelope(bpe_ecole_wmppp, NumberOfSimulations=99)
kenv_pharma <- KEnvelope(bpe_pharma_wmppp, NumberOfSimulations=99)
kenv_vetem <- KEnvelope(bpe_vetem_wmppp, NumberOfSimulations=99)
kenv_medecin <- KEnvelope(bpe_medecin_wmppp, NumberOfSimulations=99)
par(mfrow=c(2, 2))

plot(kenv_ecole, legend=FALSE, main="Ecoles", xlab = "r (mètres)")
plot(kenv_pharma, legend=FALSE, main="Pharmacies", xlab = "r (mètres)")
plot(kenv_vetem, legend=FALSE, main="Magasins de vêtements", xlab = "r (mè
    tres)")
plot(kenv_medecin, legend=FALSE, main="Médecins", xlab = "r (mètres)")
par(mfrow=c(1, 1))
```

Les résultats obtenus sur la figure 4.11 confirment les intuitions que nous avions sur la répartition spatiale de chacun des équipements sur Rennes (voir figure 4.1). Pour les médecins, les commerces de vêtements et les pharmacies, des niveaux de concentration spatiale significatifs sont détectés (graphiquement, les courbes K sont situées au dessus de l'intervalle de confiance). S'agissant des écoles, la tendance à la concentration tout comme à la dispersion n'est pas manifeste puisque la courbe K pour ce secteur reste située dans l'intervalle de confiance en dessous d'un rayon d'un kilomètre puis, au-delà ce rayon, la distribution observée des écoles sur Rennes ne semble pas s'écarter de manière importante d'une distribution aléatoire. Enfin, remarquons que la concentration spatiale est particulièrement forte pour les magasins de vêtements (l'écart entre la courbe K et la borne supérieure de l'intervalle de confiance étant le plus important pour ce secteur).

Pouvons-nous considérer ces résultats suffisants pour décrire la structure spatiale de ces équipements ou doivent-ils être complétés ? La réponse est simple : ces conclusions reposent sur des calculs statistiquement corrects, mais, qui peuvent paraître peu pertinents du point de vue économique. Ces résultats se heurtent en effet à plusieurs limites importantes, notamment l'hypothèse d'homogénéité. Tout d'abord, rappelons qu'une concentration spatiale détectée avec la fonction K de Ripley répond ici à une définition particulière : les distributions observées sont plus concentrées que ce qu'elles ne le seraient sous l'hypothèse d'une distribution aléatoire. Cette hypothèse nulle peut paraître bien forte. Prenons le cas de la localisation des pharmacies : on sait qu'elle obéit en France à certaines dispositions réglementaires, liées à la population. La distribution de référence CSR n'apparaît donc pas la plus pertinente dans ce cas. Une solution serait alors de prendre en compte cette non-homogénéité de l'espace par exemple en retenant la fonction D de DIGGLE et al. 1991 pour comparer la distribution des pharmacies à celle des résidents. Sous réserve que les données soient disponibles et accessibles, cela nous permettrait de contrôler l'hétérogénéité du territoire. Cette technique nous permettrait également de régler dans une certaine mesure des contraintes fortes d'implantation (qui empêchent de fait une égale probabilité d'implantation en tout point du territoire analysé) comme l'impossibilité de s'implanter dans des zones non constructibles sur Rennes, dans les parcs urbains etc. : la population comme les commerces ne peuvent s'y localiser. Force est de constater que si cette stratégie est séduisante, elle n'est toutefois pas encore complètement satisfaisante. Par exemple, dans le cas des équipements, et

plus encore des entreprises, on est en présence d'observations ayant des poids généralement très différents (nombre de salariés etc.). Il est donc délicat de considérer que les points analysés ont tous les mêmes caractéristiques. Or, toutes les fonctions présentées jusque-là (K, L, D et K_{inhom}) ne peuvent inclure une pondération des points. Ce constat peut être très problématique d'autant plus que, les travaux sur la concentration industrielle au sens de ELLISON et al. 1997, MAUREL et al. 1999 ont fait converger les préoccupations des économistes et celles des statisticiens spatiaux à la fin des années 1990 vers des indicateurs de concentration spatiale fondés sur un zonage. Des développements ultérieurs en ce sens doivent donc être apportés aux mesures issues du K de Ripley.

Développement des mesures fondées sur les distances pour répondre à des critères économiques

Dans les années 2000, des **listes de critères économiquement pertinents** ont été proposés pour caractériser la concentration spatiale des activités économiques (DURANTON et al. 2005 ; COMBES et al. 2004 ; BONNEU et al. 2015) comme :

— l'insensibilité de la mesure à un changement de définition d'échelles géographiques ;

— l'insensibilité à un changement de définition de niveau sectoriel (suivant la nomenclature sectorielle retenue) ;

— la comparabilité des résultats entre secteurs ;

— la prise en compte de la structure productive des industries (c'est-à-dire la concentration industrielle au sens d'ELLISON et al. 1997 qui dépend à la fois du nombre d'établissements au sein des secteurs et des effectifs) ;

— une référence doit être clairement établie.

Ces questions ont été discutées dans de nombreux travaux notamment pour distinguer les **critères appréciables** comme la comparabilité des résultats entre les secteurs, des **critères indispensables** comme le critère sur l'insensibilité de la mesure suite à un changement de définition d'échelles géographiques (cela renvoie à la MAUP précédemment présentée). L'avantage de toutes les mesures fondées sur les distances présentées dans ce chapitre est d'éviter l'écueil de la MAUP. En revanche, aujourd'hui, aucune mesure encore ne s'est affranchie du découpage sectoriel : le problème soulevé par le second critère de la liste ci-dessus reste donc entier.

Quelles pistes de recherches pour des extensions des mesures présentées ?

Plusieurs développements significatifs ont été proposés dans les années 2000. La poursuite des travaux des spécialistes de statistique spatiale, la prise en compte de l'espace dans les études économiques ont contribué à des innovations importantes en matière d'indicateurs de concentration. On ne présentera pas dans ce cadre l'ensemble des travaux, mais on se limitera à quelques uns des plus utilisés. Nous allons dans un premier temps, introduire une notion peu intuitive qu'est la **valeur de référence**. Lorsque nous essayons de caractériser une distribution de points, nous la confrontons implicitement à une distribution de référence (l'hypothèse nulle du statisticien) et c'est l'écart à cette distribution théorique qui permet d'apprécier la concentration géographique, de la dispersion ou si l'écart n'est pas suffisant pour conclure à des interdépendances entre les points. Pour s'en convaincre, reprenons l'exemple des magasins de vêtements et intéressons-nous à trois types d'indicateurs (MARCON et al. 2015a ; MARCON et al. 2017) pour caractériser leur implantation :

— Les **mesures topographiques** prennent comme valeur de référence l'espace physique (BRÜLHART et al. 2005). Le nombre de voisins des points d'intérêt est rapporté à la surface du voisinage considéré : on se place dans le cadre mathématique des processus ponctuels. Une telle analyse permet de répondre à la question suivante : la densité de magasins de vêtements

est-elle importante autour des magasins de chaussures ? Une réponse positive par exemple permettra de conclure à une concentration topographique des magasins de vêtements (dans le voisinage de ces magasins, la densité des magasins de vêtement est élevée). Les mesures présentées K, L, D et K_{inhom} répondent à cette définition topographique de la valeur de référence (selon les fonctions la densité théorique considérée est constante ou non). Il est intéressant de remarquer que, pour cette valeur de référence, l'hypothèse d'un espace homogène ou inhomogène peut être retenue.

— Les **mesures relatives** prennent comme valeur de référence une distribution qui n'est pas l'espace physique. Le nombre de voisins n'est pas rapporté à la surface, mais au nombre de points de la distribution de référence. On s'écarte clairement de la théorie des processus ponctuels, sauf à considérer la distribution de référence comme une estimation de l'intensité du processus sous l'hypothèse nulle d'indépendance entre les points. Dans notre exemple, cela revient à tester l'existence d'une sur-représentation ou d'une sous-représentation des magasins de vêtements dans le voisinage de magasins de vêtements par rapport à une référence qui peut être l'ensemble des activités commerciales. Attention, la fonction D n'est pas une mesure relative sous ces hypothèses car elle compare une densité à une autre densité, par différence. En revanche une mesure relative répondrait par exemple à la question suivante : autour des magasins de vêtements la fréquence des magasins de vêtements est plus importante qu'en moyenne, sur tout le territoire ? Une réponse positive permet de conclure à l'existence d'une concentration relative des magasins de vêtements.

— Les **mesures absolues** enfin ne font appel à aucune normalisation (par l'espace ou par rapport à une toute autre référence). Dans notre exemple, cela revient à décompter simplement le nombre de magasins de vêtements autour des magasins de vêtements. Le nombre obtenu peut ensuite être comparé à sa valeur sous l'hypothèse nulle choisie, obtenue par la méthode de Monte Carlo.

Suite aux travaux présentés précédemment notamment concernant la fonction K, des indicateurs statistiques ont été proposés dans la littérature statistique pour caractériser ces structures spatiales sous les trois valeurs de référence précédemment énoncées (MARCON et al. 2017). Nous allons développer plusieurs indicateurs dans les sections suivantes et nous verrons qu'une autre différence importante réside dans la notion de voisinage. Par exemple, il est possible d'étudier le voisinage des points analysés *jusqu'à* une certaine distance r. Concrètement, cela revient à caractériser le voisinage des points sur des disques de rayon r et cela définit des fonctions de type cumulative (comme la K de Ripley). Une autre possibilité est d'évaluer le voisinage des points non pas *jusqu'à* une distance r mais *à* une certaine distance r. Le voisinage est évalué dans une couronne (encore appelée anneau) et les fonctions de densité permettent de le caractériser (comme la fonction g que nous avons vue). Une illustration graphique de ces deux définitions est donnée sur la figure 4.12. L'aire grisée correspond sur la figure de gauche à la surface d'un disque de rayon r et, sur la figure de droite, à la surface d'une couronne à un rayon r.

Le choix du voisinage n'est pas anodin. Ainsi, les fonctions de densité sont plus précises autour du rayon d'étude mais ne conservent pas l'information sur les structures spatiales à plus petites distances, contrairement aux fonctions cumulatives. Seule une fonction cumulative pourra par exemple détecter si des agrégats sont localisés aléatoirement ou s'il existe une interaction spatiale entre agrégats (agrégats d'agrégats par exemple). En revanche, comme les fonctions cumulatives accumulent l'information spatiale jusqu'à une certaine distance, l'information locale au rayon r est peu précise, contrairement aux fonctions de densité. Le recours à l'une ou l'autre de ces notions de voisinage présente des avantages et des inconvénients (WIEGAND et al. 2004 ; CONDIT et al. 2000).

MARCON et al. 2017 ont proposé une première classification des fonctions fondées sur les distances selon ces deux critères :

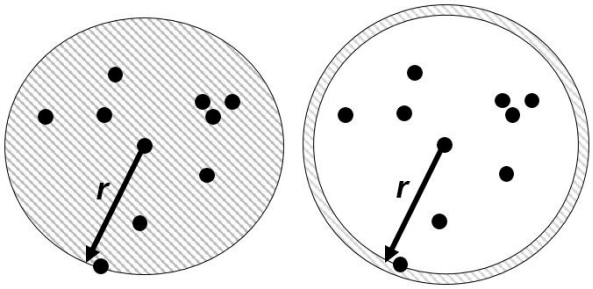

FIGURE 4.12 – Deux notions de voisinages possibles : sur un disque ou sur une couronne
Source : *les auteurs.*

— **le type de la fonction** : densité de probabilité, comme la fonction g ou fonction cumulative comme la fonction K de Ripley ;

— **la valeur de référence** qui peut être topographique (les fonctions de Ripley et leurs variantes directes), relatives à une situation de référence (comme la fonction M que nous allons présenter dans la section suivante) ou absolues (donc sans référence comme la fonction K_d également présentée dans la section suivante).

On comprend mieux pourquoi le choix de la bonne mesure n'est pas immédiat : il convient avant tout d'identifier la question posée pour retenir la mesure la plus adaptée.

4.5 Mesures fondées sur les distances récemment proposées

Nous allons présenter dans cette section deux mesures relatives à deux références non encore traitées : la référence absolue et relative.

4.5.1 Indicateur K_d de Duranton et Overman

Contrairement aux précédentes fonctions présentées, cet indicateur a été développé par des économistes et a été élaboré sans liens directs avec les travaux de Ripley (cités cependant en bibliographie). L'idée de cette fonction est de pouvoir estimer la probabilité de trouver un voisin à la distance r de chaque point.

Définition 4.5.1 — Fonction K_d de Duranton et Overman. Grâce à une normalisation, DURANTON et al. 2005 définissent K_d comme une fonction de densité de probabilité de trouver un voisin à la distance r. On peut par conséquent qualifier cette fonction de densité de mesure absolue car elle n'a pas de référentiel. L'indicateur proposé s'écrit :

$$K_d(r) = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} \kappa \left(\left\| x_i - x_j \right\|, r \right) \tag{4.14}$$

avec n désignant le nombre total de points de l'échantillon et κ, le noyau gaussien tel que

$$\kappa \left(\left\| x_i - x_j \right\|, r \right) = \frac{1}{h\sqrt{2\pi}} exp \left(- \frac{\left(\left\| x_i - x_j \right\| - r \right)^2}{2h^2} \right).$$

On voit ici la difficulté technique de comptabiliser les voisins à une distance r car cela nécessite l'utilisation d'une fonction de lissage (d'où l'utilisation du noyau gaussien dans la fonction). Cette fonction de lissage permet de dénombrer les voisins dont la distance est "autour" de r. La bande passante peut être définie de plusieurs manières mais dans l'article original de DURANTON et al.

2005 celle de SILVERMAN 1986 est mentionnée. Comme pour les autres fonctions fondées sur les distances, un intervalle de confiance de l'hypothèse nulle peut être évalué pour juger de la significativité des résultats obtenus. Les marques (couples poids/type) sont redistribuées sur toutes les localisations existantes (positions occupées par les points) : cette technique permet de contrôler à la fois la concentration industrielle et les tendances générales de localisation de l'ensemble des types de points (deux propriétés listées dans les "bons" critères des indices de concentration applicables pour les activités économiques). L'hypothèse d'une localisation aléatoire des points du type S est rejetée aux distances r, si la fonction K_d est située en dehors de l'enveloppe de confiance de l'hypothèse nulle. Une autre version de K_d prenant en compte la pondération des points existe, elle a été proposée dans l'article original de DURANTON et al. 2005. BEHRENS et al. 2015 ont quant à eux retenu une fonction cumulative K_d. Il est à noter que la fonction K_d a fait l'objet de nombreuses applications empiriques en économie spatiale (par exemple DURANTON 2008, BARLET et al. 2008).

La fonction K_d peut être calculée sous R à l'aide de la fonction `Kdhat` du package *dbmss*. La fonction `KdEnvelope` disponible dans le même package permettra d'associer un intervalle de confiance aux résultats obtenus.

4.5.2 Indicateur M de Marcon et Puech

L'indicateur M de MARCON et al. 2010 est un indicateur cumulatif, comme le K de Ripley puisqu'il est calculé en faisant varier un disque de rayon r autour de chaque point. C'est un indicateur relatif puisqu'il va comparer la proportion de points d'intérêt dans un voisinage à celle que l'on observe sur l'ensemble du territoire analysé. Si l'on considère que les magasins de vêtements s'attirent, leur proportion autour de chaque magasin de vêtements sera plus forte qu'au niveau de la ville. En pratique, pour un rayon r, on va calculer le rapport entre la proportion locale des magasins de vêtements autour des magasins de vêtements à la proportion observée en ville. On réitère ce calcul pour tous les magasins de vêtements et on calcule la moyenne de ces proportions relatives. La valeur de référence de la fonction M est 1, une valeur supérieure traduisant une concentration spatiale relative, une valeur inférieure une tendance à la répulsion (la valeur minimale étant 0). Les valeurs de M sont également interprétables en termes de comparaisons de ratios : par exemple si $M(r) = 3$, cela indique qu'il y a en moyenne une fréquence d'apparition trois fois plus élevée des points d'intérêt autour des points d'intérêt dans un rayon r que celle observée sur toute la fenêtre d'observation. Enfin comme la fonction K_d, M peut intégrer la pondération des points.

Définition 4.5.2 — Fonction M de Marcon et Puech. Formellement, pour les points du type S, on définit la fonction M de Marcon et Puech par :

$$M(r) = \sum_{i \in S} \frac{\sum_{j \neq i, j \in S} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right)}{\sum_{j \neq i} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right)} \Big/ \frac{n_S - 1}{n - 1}. \tag{4.15}$$

où n_S et n désignent respectivement le nombre de points total du type S et de tous types sur la fenêtre d'étude. Cet indicateur doit être lu comme le résultat de deux rapports de fréquences. On compare la moyenne locale de la fréquence des points de type S dans un rayon r autour des points de type S à la fréquence de points de type S sur toute la fenêtre d'observation. Le fait d'enlever un point au dénominateur permet d'éviter un petit biais puisque systématiquement le point centre ne peut être dénombré dans son voisinage.

Comme pour la fonction K_d, une version prenant en compte la pondération des points existe (MARCON et al. 2017). Techniquement cela revient à multiplier l'indicatrice par le poids du point

voisin considéré (par exemple par le nombre de ses employés ou de son chiffre d'affaire si l'on s'intéresse aux établissements industriels). Comme pour les autres indicateurs, on peut générer un intervalle de confiance par des méthodes Monte Carlo. On procède en conservant les spécificités des points (couple poids/secteur). Pour M, comme K_d, le contrôle de la concentration industrielle n'est pas présente dans la définition de la fonction mais dans la définition de l'intervalle de confiance puisque les étiquettes (couples poids/secteur) des points sont redistribuées sur les emplacements existants. Dans leurs derniers travaux, LANG et al. 2015 ont proposé une version non cumulative de l'indicateur M, dénommée m, analogue à la fonction g pour K (voir l'équation (4.11)). Comme dans toutes les situations que nous avons rencontrées, les indicateurs peuvent conduire à des analyses différentes : les valeurs de référence n'étant pas les identiques, ils répondent à des questions différentes. Les analyses apportées sont donc complémentaires (MARCON et al. 2015a ; LANG et al. 2015). Enfin, notons que la fonction M ne nécessite pas la correction d'effets de bord et elle peut être calculée sous R à l'aide de la fonction `Mhat` du package *dbmss*. La fonction `MEnvelope` du même package permet d'associer un intervalle de confiance pour juger de la significativité des résultats obtenus.

Comme exemple d'application, nous proposons d'étudier les structures spatiales des quatre équipements de l'exemple introductif sur la ville de Rennes. La représentation graphique des résultats des fonctions M pour les écoles, les pharmacies, les médecins et les magasins de vêtements est donné sur la figure 4.13.

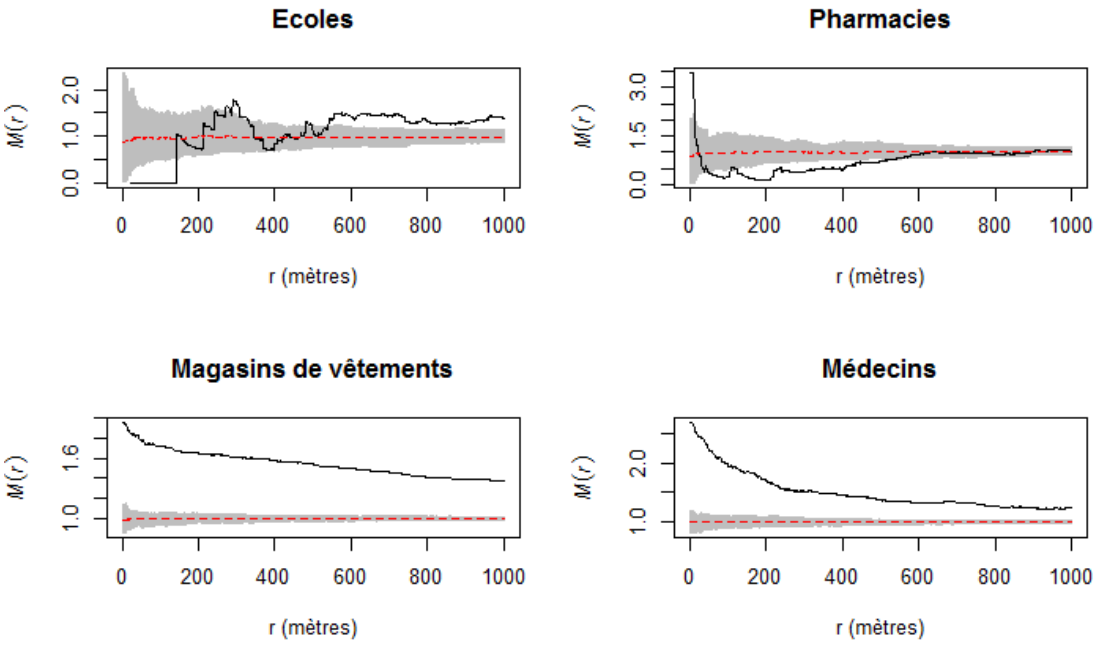

FIGURE 4.13 – Fonctions de Marcon et Puech pour les quatre équipements
Source : *Insee-BPE, packages spatstat et dbmss, calculs des auteurs.*

```
library("dbmss")
# Jeu de points marqués
bpe_equip<- bpe[bpe $TYPEQU %in%c ("C104","D301","B302","D201"),c (2,3,1)]
colnames(bpe_equip) <- c("X", "Y", "PointType")
bpe_equip_wmppp <- wmppp(bpe_equip)
r<- 0:1000
```

```
NumberOfSimulations <- 99
menv_eco <- MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
    ReferenceType="C104")
menv_pha <- MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
    ReferenceType="D301")
menv_vet <- MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
    ReferenceType="B302")
menv_med <- MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
    ReferenceType="D201")
par(mfrow=c(2, 2))
plot(menv_eco, legend=FALSE, main="Ecoles", xlab = "r (mètres)")
plot(menv_pha, legend=FALSE, main="Pharmacies", xlab = "r (mètres)")
plot(menv_vet, legend=FALSE, main="Magasins de vêtements", xlab = "r (mè
    tres)")
plot(menv_med, legend=FALSE, main="Médecins", xlab = "r (mètres)")
par(mfrow=c(1, 1))
```

On constate aisément que des niveaux de concentration spatiale sont observables pour toutes les distances étudiées pour les activités des médecins ou des magasins de vêtements (les deux courbes M associées étant situées au-dessus de leur intervalle de confiance respectif jusqu'à 1km). Comme il est possible de comparer les valeurs obtenues par la fonction M, nous pouvons également conclure que les plus forts niveaux d'aggrégation apparaissent à petites distances. Ainsi, dans les tous premiers rayons d'étude, la proportion de magasins de vêtements autour des magasins de vêtements est approximativement deux fois plus importante que la proportion de magasins de vêtements observée sur la ville de Rennes. Ce résultat est assez proche des conclusions du travail de MARCON et al. 2015a sur la ville de Lyon pour cette activité. Concernant les écoles ou les pharmacies en revanche, il est détecté à la fois des niveaux de concentration ou de dispersion suivant les distances considérées. Les écoles par exemple apparaissent dispersées jusqu'à 150m environ (la courbe M associée est située sous l'intervalle de confiance de l'hypothèse nulle jusqu'à cette distance), puis, au-delà d'une distance de 500m un phénomène de concentration spatiale est détecté. À très courtes distances, les pharmacies apparaissent quant à elles spatialement agrégées alors qu'elles présentent une distribution dispersée dès 50m approximativement. Pour les écoles et les pharmacies, on remarque que les courbes M restent toutefois assez proches de leur intervalle de confiance respectif.

4.5.3 Autres développements

Cette littérature statistique est actuellement bourgeonnante (DURANTON 2008 ; MARCON et al. 2017). Les apports sont variés : les statisticiens définissent le cadre théorique nécessaire et les chercheurs développent des outils applicables aux spécificités de leur domaine. Parmi les travaux menés récemment, BONNEU et al. 2015 proposent une famille d'indicateurs qui a le mérite de montrer les liens entre les indicateurs Bonneu-Thomas (proposé dans l'article), Marcon-Puech et Duranton-Overman. Tous les indicateurs ne sont pas encore implémentés dans les logiciels usuels même si des efforts sont faits en ce sens pour tenir compte des récents développements de la littérature et les rendre disponibles, librement, pour les utilisateurs intéressés.

4.6 Processus multitypes

On a présenté en introduction quatre cartes relatives aux localisations respectives des écoles, des pharmacies, des médecins généralistes et des magasins de vêtements (figure 4.1). On aurait pu rassembler toutes ces informations, chaque activité étant une marque, de nature qualitative,

du processus. Ces marques permettent de constituer des **processus multitypes**, et d'introduire de nouvelles questions à côté de celles qui ont été développées précédemment : entre les types (marques), y a-t-il indépendance dans les localisations ? Si la réponse est négative, observe-t-on des phénomènes de nature attractive ou répulsive ?

Afin d'apporter des réponses à ces questions, nous devons considérer à présent des processus qui ont des caractéristiques propres : il nous est donc possible de définir des indicateurs du premier ordre (l'intensité) et de second ordre (les relations de voisinage), ce que nous ferons successivement dans les deux sous-sections suivantes.

4.6.1 Fonctions d'intensité

L'analyse de la variabilité de l'intensité des processus qui a abouti à la distribution observée des entités analysées est intéressante pour une première analyse.

Dans le domaine de l'écologie, on peut se demander par exemple (*i*) si toutes les espèces d'arbres au sein d'une forêt sont localisées de manière identique, (*ii*) si les arbres morts sont plus agglomérés que les arbres non malades, (*iii*) si la présence des jeunes arbustes suit celle des arbres parents, etc. Pour cela, l'étude de la densité donne une première indication de l'hétérogénéité spatiale observée. Dans l'exemple ci-dessous, nous avons repris les localisations respectives des arbres d'un dispositif expérimental permanent de Paracou en Guyane française, disponibles dans le jeu de données `Paracou16` du package *dbmss*. Trois espèces d'arbres sont répertoriées : les *Vacapoua americana*, les *Qualea rosea* et les espèces mélangées d'arbres regroupées sous le terme *Other*. Le nombre élevé d'arbres présents sur la parcelle Paracou16 (2 426 arbres au total) rend peu identifiables de quelconques tendances de localisation pour chacune des espèces (voir figure 4.14).

FIGURE 4.14 – Localisation des arbres d'espèces *Vacapoua americana*, *Qualea rosea* ou autres (mélange) sur le dispositif forestier Paracou16
Source : *données Paracou16 du package dbmss, calculs des auteurs.*

```
library("dbmss")
data(paracou16)
plot(paracou16, which.marks=2, main = "")
# la 2ème colonne permet de différencier les types de points (espèces)
```

En revanche, une représentation de la densité par espèce est plus informative et permet de mettre en évidence des différences d'implantation selon les espèces d'arbres considérées (voir figure 4.15). Une représentation en 2D de la densité est donnée dans cet exemple et obtenue à partir

de la fonction `density` du package *spatstat*.

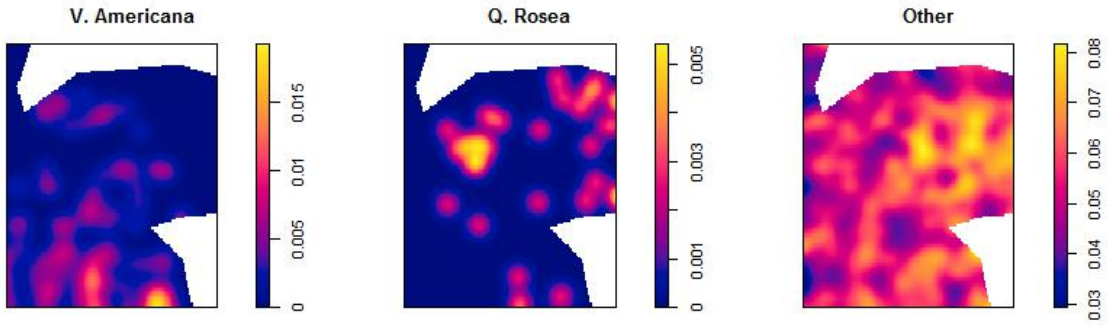

FIGURE 4.15 – Représentation de la densité des arbres d'espèces *Vacapoua americana*, *Qualea rosea* ou autres (mélange) sur le dispositif forestier Paracou16
Source : *données Paracou16 du package dbmss, calculs des auteurs.*

```
library("dbmss")
data(paracou16)
V.Americana<-  paracou16[paracou16$marks$PointType=="V. Americana"]
Q.Rosea<-  paracou16[paracou16$marks$PointType=="Q. Rosea"]
Other<-  paracou16[paracou16$marks$PointType=="Other"]
par(mfrow=c(1,3))
plot(density(V.Americana, 8), main="V. Americana")
plot(density(Q.Rosea, 8), main="Q. Rosea")
plot(density(Other, 8), main="Other")
par(mfrow=c(1,1))
```

Dans le domaine de l'économie spatiale, l'étude de processus multitypes pourrait également être riches d'enseignements. Nous pourrions par exemple nous interroger sur les interactions possibles entre les différents types d'équipements (cabinets de médecins généralistes, écoles, etc.). En reprenant l'extrait issu de la base permanente des équipements sur la ville de Rennes, les quatre sous-configurations de points avaient été représentées sur la figure 4.1. Sur la figure 4.16, nous cartographions les densités de deux équipements : les pharmacies et les médecins. Visuellement, des tendances assez similaires d'implantation semblent être observables, la représentation en 3D sur la figure 4.16 le confirme. La fonction `persp` de *spatstat* est retenue.

```
library("dbmss")
# Fichier de la BPE sur le site insee.fr :
# https://www.insee.fr/fr/statistiques/2387803?sommaire=2410933
# Données pour ces exemples :
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))

bpe_pha<- bpe[bpe $TYPEQU=="D301", ]
bpe_med<- bpe[bpe $TYPEQU=="D201", ]

pharma <- as.ppp(bpe_pha[ ,c ("lambert_x", "lambert_y")], owin(c(min(bpe_
    pha[,"lambert_x"]),max (bpe_pha[,"lambert_x"])),c (min(bpe_pha[,"
    lambert_y"]),max (bpe_pha[,"lambert_y"]))))
bpe_pharma_wmppp <- as.wmppp(pharma)
```

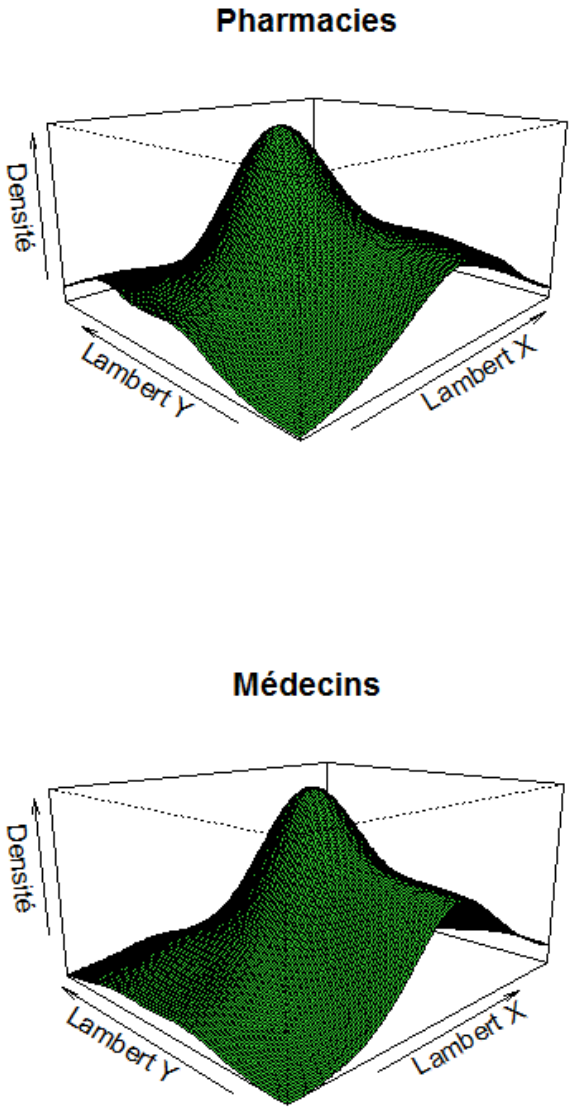

FIGURE 4.16 – Représentation de la densité des pharmacies et des médecins sur Rennes
Source : *Insee-BPE, packages spatstat et dbmss, calculs des auteurs.*

```
medecin <- as.ppp(bpe_med[ , c("lambert_x", "lambert_y")], owin(c(min(bpe_
    med[,"lambert_x"]), max(bpe_med[,"lambert_x"])), c(min(bpe_med[,"
    lambert_y"]), max(bpe_med[,"lambert_y"]))))
bpe_medecin_wmppp <- as.wmppp(medecin)

persp(density(medecin), col="limegreen",
      theta = -45,#angle de visualisation
      xlab = "Lambert X", ylab = "Lambert Y", zlab = "Densité",
      main = "Médecins")
persp(density(pharma), col="limegreen", theta = -45,
      xlab = "Lambert X", ylab = "Lambert Y", zlab = "Densité",
      main = "Pharmacies")
```

Toutefois, seuls les résultats d'une analyse des propriétés de second ordre des processus nous permettra de conclure sur une éventuelle interaction (attraction ou répulsion) entre les espèces d'arbres ou entre les équipements. C'est pour cette raison que l'étude des propriétés de premier ordre n'est qu'une première étape d'analyse pour étudier une configuration de points.

4.6.2 Fonctions intertypes

Différents développements ont été proposés pour étudier les propriétés de second ordre des processus multitypes. Des indicateurs dérivés de la fonction K (univariée) de Ripley ont été proposés pour analyser les localisations relatives des sous-configurations des points liées aux différentes marques. Ces indicateurs sont généralement nommés fonctions intertypes ou fonctions bivariées. Nous en détaillerons deux dans les sous-sections suivantes. D'un point de vue pratique, il est possible d'utiliser des packages R comme *spatstat* ou *dbmss* pour calculer les fonctions et représenter les résultats graphiquement.

La fonction K intertypes

Considérons le cas suivant. Nous souhaitons étudier la structure spatiale entre deux types de points, par exemple les points du type T localisés autour des points du type S. L'appel à une fonction intertype permet alors d'étudier la structure spatiale des points du type T localisés à une distance inférieure ou égale à r de ceux du type S.

Un premier indicateur peut être retenu, la fonction K intertypes. Cette dernière est notée $\widehat{K}_{S,T}$ et se définit comme suit :

$$\widehat{K}_{S,T}(r) = \frac{1}{\widehat{\lambda}_S n_S} \sum_{i \in S} \sum_{j \in T} \mathbf{1}\left\{\left\|x_i - x_j\right\| \leq r\right\}. \tag{4.16}$$

où $\widehat{\lambda}_S$ désigne l'intensité estimée du sous-processus de type S. Sur le domaine d'étude, n_S représente le nombre total de points S.

Dans le cas où S et T sont le même type, on retrouve la définition de la fonction K univariée présentée dans la section 4.4.1. Attention toutefois car la correction des effets de bord n'est ici pas intégrée à la définition de la fonction K intertypes pour alléger l'écriture. La valeur de référence est toujours πr^2, quelque que soit le rayon r considéré, puisque l'on se place sous l'hypothèse nulle d'une distribution complétement aléatoire des points (de type S et T). Si le sous-processus de type S est indépendant du sous-processus de type T, alors le nombre de points de type T se trouvant à une distance inférieure ou égale à r d'un point de type S est le nombre attendu de points de type T localisés dans un disque de rayon r, soit $\lambda_T \pi r^2$. Cette hypothèse nulle correspond à la distribution indépendante de deux types d'établissements industriels par exemple. Une autre hypothèse nulle donnant le même résultat est que les points sont d'abord distribués selon un

processus de Poisson homogène puis reçoivent leur type dans un second temps (par exemple, des emplacements commerciaux sont créés puis occupés par différents types de commerces). Pour toutes les distances r pour lesquelles des valeurs observées de $\widehat{K}_{S,T}(r)$ sont inférieures à πr^2, une tendance à la répulsion des points T autour des points S sera à signaler. Au contraire, des valeurs de $\widehat{K}_{S,T}$ supérieures à πr^2 tendront quant à elles à valider une attraction des points T autour des points S dans un rayon r. La simulation d'un intervalle de confiance par la méthode de Monte Carlo permettra de conclure à une attraction ou une répulsion entre les deux types de points.

Sous le package *spatstat*, la fonction `Kcross` permet d'implémenter la fonction K intertypes. Comme application, nous reprenons ci-dessous l'exemple des données `Paracou16`. En effet, si nous retenons la fonction K intertypes, nous faisons l'hypothèse que l'espace considéré est homogène ; or, cette hypothèse est quasiment systématiquement retenue dans les analyses empiriques en écologie forestière (GOREAUD 2000). Sur la figure 4.17, on a représenté les fonctions K intertypes (ou bivariées) des espèces *Qualea rosea* ou mélangées *Other* avec celle du *Vacapoua americana*. Les courbes noires représentent les fonctions K intertypes observées et celles en pointillés rouges, les fonctions K intertypes de référence. Comme on peut le constater visuellement, il y a un lien de nature répulsive entre les *Qualea rosea* et les *Vacapoua americana* (K intertypes observée est située sous la valeur de référence) alors qu'aucune tendance d'association ne semble exister entre les *Vacapoua americana* et les autres espèces d'arbres (les courbes K intertypes théorique et observée sont confondues pour toutes les distances).

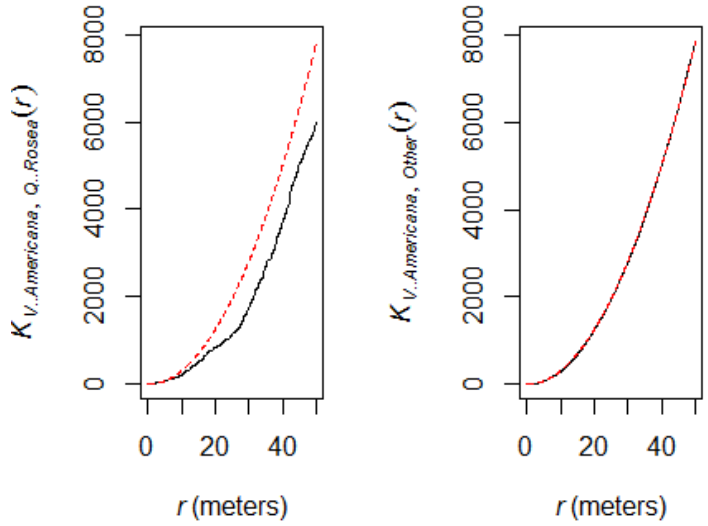

FIGURE 4.17 – Interactions des différentes espèces d'arbres sur le dispositif forestier Paracou16
Source : *données paracou16 du package dbmss, calculs des auteurs.*

```
library("dbmss")
# Simplification des marques
marks(paracou16)<-  paracou16$marks$PointType
par(mfrow=c(1,2))
# calcul de K intertypes pour les arbres de l'espèce "Q.Rosea" autour de
    ceux de l'espèce "Q. Rosea"
plot(Kcross(paracou16, "V. Americana", "Q. Rosea", correction="isotropic"),
```

```
     legend=FALSE, main=NULL)
# calcul de K intertypes pour les arbres de l'espèce "Q.Rosea" autour de
    ceux de l'espèce "Other"
plot(Kcross(paracou16, "V. Americana", "Other", correction="isotropic"),
    legend=FALSE, main=NULL)
par(mfrow=c(1,1))
```

La fonction M intertypes

De la même façon, on peut utiliser la fonction M précédemment présentée comme un outil intertypes. L'idée est toujours de comparer une proportion locale à une proportion globale mais dans le cas de la fonction M intertypes le type de points voisins d'intérêt n'est pas le même type que celui des points centre. Par exemple, si nous suspectons une attraction des points de type T par ceux de type S, nous allons comparer la proportion locale de voisins du type T autour de points du type S à la proportion globale observée sur tout le territoire considéré. Si l'attraction entre les points de type T autour de type S est réelle, la proportion de points de type T autour de ceux du type S devrait être localement plus importante que celle observée sur toute l'aire d'étude. Au contraire, si les points T sont repoussés par ceux du type S, la proportion relative de points de type T autour de ceux du type S sera relativement plus faible que celle observée sur l'ensemble du territoire analysé. L'estimateur empirique non pondéré de M intertypes dans ce cas sera défini par :

$$\widehat{M}_{S,T}(r) = \sum_{i \in S} \frac{\sum_{j \in T} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right)}{\sum_{j \neq i} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right)} \Big/ \frac{n_T}{n-1}. \tag{4.17}$$

où n désigne le nombre total de points sur toute l'aire d'étude, n_S ceux type S. Tout comme la fonction K intertypes, on supposera ici que chaque point n'appartient qu'à un seul type qui peut être S, T ou autre. Pour la fonction M intertypes, la valeur de référence pour toutes les distances r considérées est toujours égale à 1. Pour plus de détails sur cette fonction (prise en compte de la pondération, construction de l'intervalle de confiance associé etc.), on pourra se reporter à l'article de MARCON et al. 2010. Cette fonction intertypes peut être calculée sous R à l'aide de la fonction `Mhat` du package *dbmss*. Pour la construction d'un intervalle de confiance, on utilisera la fonction `MEnvelope` du même package.

Un exemple concret d'application de M intertypes est proposé ci-dessous à partir des équipements rennais considérés dans l'introduction. Si nous soupçonnons des relations d'attraction ou de répulsion entre plusieurs équipements, il est alors possible d'analyser les interactions existantes grâce à la fonction M intertypes. En effet, souvenons-nous que l'utilisation de la fonction M permet de relâcher l'hypothèse d'un espace homogène qui peut être considérée comme une hypothèse forte pour caractériser localisation des activités économiques (voir par exemple DURANTON et al. 2005, p. 1104). Dans ce cas, l'utilisation de M intertypes paraîtrait donc plus appropriée que K intertypes. Sur la figure 4.18, nous avons représenté à partir de l'extrait de données de la base des équipements les liens existants entre les localisations des médecins et des pharmacies sur Rennes. Sur le graphique de droite, sont analysées les localisations des pharmacies dans un voisinage de r mètres des médecins. Une répulsion serait détectée à très petites distances puis de l'agrégation intertypes serait observable jusqu'à 1km. Le graphique de gauche indique que les médecins semblent relativement agglomérés dans un rayon d'un 1km autour des localisations des pharmacies à Rennes (la construction d'un intervalle de confiance avec 100 simulations par exemple nous permettrait de conclure que la tendance à la dispersion à très petites distances est non significative).

```
library("dbmss")
```

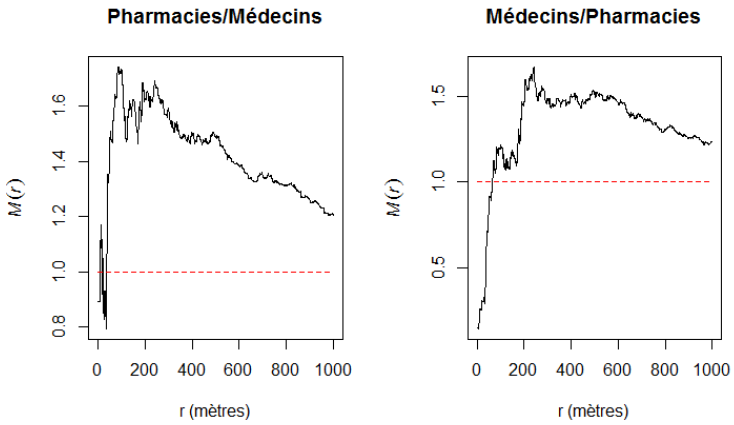

FIGURE 4.18 – Relations de voisinage entre médecins et pharmacies sur Rennes
Source : *packages spatstat et dbmss, calculs des auteurs.*

```
# Fichier de la BPE sur le site insee.fr :
# https://www.insee.fr/fr/statistiques/2387803?sommaire=2410933
# Données pour ces exemples :
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))

# Jeu de points marqués
bpe_equip<- bpe[bpe $TYPEQU %in%c ("C104","D301","B302","D201"),c (2,3,1)]
colnames(bpe_equip) <- c("X", "Y", "PointType")
bpe_equip_wmppp <- wmppp(bpe_equip)

bpe_pha<- bpe[bpe $TYPEQU=="D301", ]
bpe_med<- bpe[bpe $TYPEQU=="D201", ]
pharma <- as.ppp(bpe_pha[ ,c ("lambert_x", "lambert_y")], owin(c(min(bpe_
    pha[,"lambert_x"]),max (bpe_pha[,"lambert_x"])),c (min(bpe_pha[,"
    lambert_y"]),max (bpe_pha[,"lambert_y"]))))
bpe_pharma_wmppp <- as.wmppp(pharma)
medecin <- as.ppp(bpe_med[ ,c ("lambert_x", "lambert_y")], owin(c(min(bpe_
    med[,"lambert_x"]),max (bpe_med[,"lambert_x"])),c (min(bpe_med[,"
    lambert_y"]),max (bpe_med[,"lambert_y"]))))
bpe_medecin_wmppp <- as.wmppp(medecin)

# Jeu de points marqués
r<- 0:1000
#  M intertype : étude des liens entre les localisations des médecins
    autour des pharmacies
M_pha_med<- Mhat(bpe _equip_wmppp, r, ReferenceType="D301", NeighborType="
    D201")
r<- 0:1000
#  M intertype : étude des liens entre les localisations des pharmacies
    autour des médecins
M_med_pha<- Mhat(bpe _equip_wmppp, r, ReferenceType="D201", NeighborType="
    D301")
```

```
par(mfrow=c(1, 2))
plot(M_pha_med, legend=FALSE, main="Pharmacies/Médecins", xlab = "r (mètres
    )")
plot(M_med_pha, legend=FALSE, main="Médecins/Pharmacies", xlab = "r (mètres
    )")
par(mfrow=c(1, 1))
```

L'analyse des relations de voisinage entre les équipements rennais n'est pas la seule qui peut être explorée. Par exemple, nous pourrions suspecter des interactions entre les localisations de certains équipements et la population. Pour examiner cette relation, il convient de rapprocher les données de la figure 4.13 de celles de la population. Le code R pour établir le lien entre population et les quatre types d'équipements considérés à l'aide de la fonction M sont donnés ci-après. On constate aisément sur la figure 4.19 que la distribution des quatre équipements considérés ne semble pas s'écarter significativement de celle de la population (la distance maximum reportée a été limitée à 100 mètres car aucun résultat notable n'est obtenu au-delà de ce rayon).

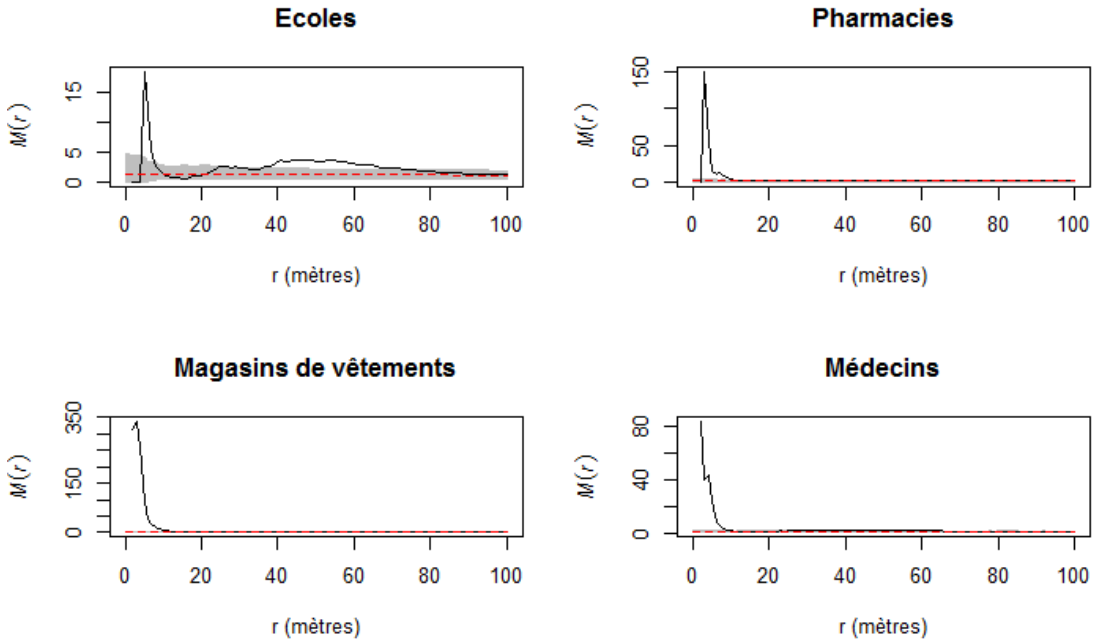

FIGURE 4.19 – Lien équipement/population pour les quatre équipements sur Rennes
Source : *Insee-BPE, package dbmss, calculs des auteurs.*

```
library("dbmss")
colnames(popu) <- c("X", "Y", "PointWeight")
popu$PointType<-  "POPU"
popuwmppp<- wmppp(popu)
# Fusion des jeux de points dans la fenêtre de bpe_equip_dbmms
bpe_equip_popu<-  superimpose(popuwmppp, bpe_equip_wmppp, W=bpe_equip_wmppp
    $window)
# 100 simulations sont retenues par défaut
menv_popu_eco<-  MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="C104", SimulationType="RandomLabeling")
```

```
menv_popu_pha <- MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="D301", SimulationType="RandomLabeling")
menv_popu_vet <- MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="B302", SimulationType="RandomLabeling")
menv_popu_med <- MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="D201", SimulationType="RandomLabeling")
par(mfrow=c(2, 2))
plot(menv_popu_eco, legend=FALSE, main="Ecoles", xlim=c(0,100), xlab = "r (
    mètres)")
plot(menv_popu_pha, legend=FALSE, main="Pharmacies", xlim=c(0,100), xlab =
    "r (mètres)")
plot(menv_popu_vet, legend=FALSE, main="Magasins de vêtements", xlim=c
    (0,100), xlab = "r (mètres)")
plot(menv_popu_med, legend=FALSE, main="Médecins", xlim=c(0,100), xlab = "r
     (mètres)")
par(mfrow=c(1, 1))
```

Notons enfin que M intertypes n'est pas la seule fonction proposée en espace hétérogène. D'autres fonctions univariées ont une version bivariée comme K_d ou K_{inhom} et peuvent être implémentées grâce au package *dbmss* sous R.

4.7 Modélisation des processus

Les processus présentés précédemment, notamment les processus de Poisson, servent également à construire des modèles. On les utilise comme dans les modèles statistiques classiques pour expliquer et prédire. On cherche aussi à trouver parmi les modèles en concurrence celui qui aura le meilleur pouvoir explicatif. Pour construire ces modèles, on utilise des covariables. La souplesse du logiciel R permet d'utiliser des données qui sont associées aux points d'observation, mais aussi des données continues, des images, des grilles.

4.7.1 Cadre général pour la modélisation

Pour ajuster un processus ponctuel de Poisson à un semis de points, on peut spécifier la forme de la fonction d'intensité $\lambda(\cdot)$ et chercher les paramètres qui permettent le meilleur ajustement. Dans le package *spatstat*, la fonction `ppm` est l'instrument essentiel. Si on appelle *trend* le modèle de l'intensité et *monpp* le processus analysé, la commande s'écrit :

```
ppm(monpp~trend)
#où trend désigne de façon générique une tendance et monpp le processus
    analysé
```

La syntaxe de cette commande pour la modélisation de processus ponctuels (dont le sigle est ppm en anglais) est proche de celle de la commande classique `lm` de R, qui sert aux modèles de régression linéaire. Les spécificités de la modélisation peuvent être multiples : les modèles estimés peuvent résulter d'une fonction log-linéaire de la variable explicative, définis à partir de plusieurs variables etc. Le choix et la validation des modèles devront compléter l'analyse pour apporter une réponse concluante. Parmi les solutions, le test du rapport de vraisemblance peut-être mobilisé.

4.7.2 Exemples d'application

Pour traiter une telle question, les jeux de données analysés doivent être suffisamment riches pour répondre de manière satisfaisante aux modèles théoriques. Le lecteur intéressé par cette

approche pourra se reporter notamment aux deux exemples traités en détails dans l'ouvrage de BADDELEY et al. 2005. Le premier repose sur des données (`Bei`) relatif aux arbres de l'espèce *Beischmiedia pendula* disponible dans le package *spatstat*. En effet, en plus de la localisation des arbres de cette espèce dans une forêt humide tropicale de l'île de Barro Colorado, des données sur l'altitude et la pente du terrain sont également fournies. Le second jeu de données, nommé `Murchison` dans le package *spatstat*, est relatif à la localisation des dépôts d'or à Murchison en Australie-Occidentale. Il permet de modéliser l'intensité des dépôts d'or en fonction d'autres données spatiales : la distance à la faille géologique la plus proche (les failles sont décrites par des lignes) et la présence d'un type particulier de roche (décrit par des polygones). La modélisation de l'intensité d'un processus peut donc s'appuyer sur des variables exogènes mesurées ou calculées à partir d'informations géographiques.

Les progrès de la modélisation sont implémentés régulièrement dans la fonction `ppm`. La possibilité de modéliser les interactions entre points (avec l'argument `interaction` de la fonction) en plus de la densité existe actuellement pour un type particulier de processus seulement, ceux de Gibbs, utilisés pour la modélisation de l'agrégation spatiale de l'industrie par SWEENEY et al. 2015. On se reportera à l'aide de la fonction `ppm` pour prendre connaissance de ses mises à jour.

Conclusion

Dans ce chapitre, nous avons tenté de donner un premier aperçu des méthodes statistiques pouvant être retenues pour caractériser les données ponctuelles. Notre objectif était de souligner que la diversité des questions posées nécessite de manier les outils statistiques avec précaution. Avant toute étude, il convient donc de bien définir la question posée et son cadre d'analyse pour retenir la méthode statistique la plus pertinente. Cette mise en garde théorique est importante car les routines de calculs sont aujourd'hui largement accessibles sous le logiciel R notamment et ne posent, en principe, que peu de difficultés pratiques de mises en œuvre. Ces méthodes statistiques peuvent donner lieu à des analyses plus avancées dans ce domaine ou à des études complémentaires, notamment en économétrie spatiale par exemple (voir le chapitre 6 : "Économétrie spatiale : modèles courants").

Références - Chapitre 4

ARBIA, Giuseppe (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht : Kluwer.

ARBIA, Giuseppe, Giuseppe ESPA et Danny QUAH (2008). « A class of spatial econometric methods in the empirical analysis of clusters of firms in the space ». *Empirical Economics* 34.1, p. 81–103.

ARBIA, Giuseppe et al. (2012). « Clusters of firms in an inhomogeneous space : The high-tech industries in Milan ». *Economic Modelling* 29.1, p. 3–11.

BADDELEY, Adrian J., Jesper MØLLER et Rasmus Plenge WAAGEPETERSEN (2000). « Non- and semi-parametric estimation of interaction in inhomogeneous point patterns ». *Statistica Neerlandica* 54.3, p. 329–350.

BADDELEY, Adrian J, Edge RUBAK et Rolf TURNER (2015b). *Spatial Point Patterns : Methodology and Applications with R*. Chapman & Hall/CRC Interdisciplinary Statistics. 810 pages. Chapman et Hall/CRC.

BADDELEY, Adrian J et Rolf TURNER (2005). « Spatstat : an R package for analyzing spatial point patterns ». *Journal of Statistical Software* 12.6, p. 1–42.

BARLET, Muriel, Anthony BRIANT et Laure CRUSSON (2008). *Concentration géographique dans l´industrie manufacturière et dans les services en France : une approche par un indicateur en continu*. Série des documents de travail de la Direction des Études et Synthèses économiques G 2008 / 09. Institut National de la Statistique et des études économiques (Insee).

— (2013). « Location patterns of service industries in France : A distance-based approach ». *Regional Science and Urban Economics* 43.2, p. 338–351.

BEHRENS, Kristian et Théophile BOUGNA (2015). « An anatomy of the geographical concentration of Canadian manufacturing industries ». *Regional Science and Urban Economics* 51, p. 47–69.

BESAG, Julian E. (1977). « Comments on Ripley's paper ». *Journal of the Royal Statistical Society* B 39.2, p. 193–195.

BONNEU, Florent (2007). « Exploring and Modeling Fire Department Emergencies with a Spatio-Temporal Marked Point Process ». *Case Studies in Business, Industry and Government Statistics* 1.2, p. 139–152.

BONNEU, Florent et Christine THOMAS-AGNAN (2015). « Measuring and Testing Spatial Mass Concentration with Micro-geographic Data ». *Spatial Economic Analysis* 10.3, p. 289–316.

BRIANT, Anthony, Pierre-Philippe COMBES et Miren LAFOURCADE (2010). « Dots to boxes : Do the Size and Shape of Spatial Units Jeopardize Economic Geography Estimations ? » *Journal of Urban Economics* 67.3, p. 287–302.

BRÜLHART, Marius et Rolf TRAEGER (2005). « An Account of Geographic Concentration Patterns in Europe ». *Regional Science and Urban Economics* 35.6, p. 597–624.

COLE, Russel G. et Gregg SYMS (1999). « Using spatial pattern analysis to distinguish causes of mortality : an example from kelp in north-eastern New Zealand ». *Journal of Ecology* 87.6, p. 963–972.

COMBES, Pierre-Philippe, Thierry MAYER et Jacques-François THISSE (2008). *Economic Geography, The Integration of Regions and Nations*. Princeton : Princeton University Press.

COMBES, Pierre-Philippe et Henry G OVERMAN (2004). « The spatial distribution of economic activities in the European Union ». *Handbook of Urban and Regional Economics*. Sous la dir. de J Vernon HENDERSON et Jacques-François THISSE. T. 4. Amsterdam : Elsevier. North Holland. Chap. 64, p. 2845–2909.

CONDIT, Richard et al. (2000). « Spatial Patterns in the Distribution of Tropical Tree Species ». *Science* 288.5470, p. 1414–1418.

DIGGLE, Peter J. (1983). *Statistical analysis of spatial point patterns*. London : Academic Press, 148 p.

DIGGLE, Peter J. et A. G. CHETWYND (1991). « Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations ». *Biometrics* 47.3, p. 1155–1163.

DURANTON, Gilles (2008). « Spatial Economics ». *The New Palgrave Dictionary of Economics*. Sous la dir. de Steven N. DURLAUF et Lawrence E. BLUME. Palgrave Macmillan.

DURANTON, Gilles et Henry G. OVERMAN (2005). « Testing for Localization Using Micro-Geographic Data ». *Review of Economic Studies* 72.4, p. 1077–1106.

ELLISON, Glenn et Edward L. GLAESER (1997). « Geographic Concentration in U.S. Manufacturing Industries : A Dartboard Approach ». *Journal of Political Economy* 105.5, p. 889–927.

ELLISON, Glenn, Edward L. GLAESER et William R. KERR (2010). « What Causes Industry Agglomeration ? Evidence from Coagglomeration Patterns ». *The American Economic Review* 100.3, p. 1195–1213.

FEHMI, Jeffrey S. et James W. BARTOLOME (2001). « A grid-based method for sampling and analysing spatially ambiguous plants. » *Journal of Vegetation Science* 12.4, p. 467–472.

GOREAUD, Françcois et Raphaël PÉLISSIER (1999). « On explicit formulas of edge effect correction for Ripley's K-function ». *Journal of Vegetation Science* 10.3, p. 433–438. ISSN : 1654-1103. DOI : 10.2307/3237072. URL : http://dx.doi.org/10.2307/3237072.

GOREAUD, François (2000). « Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude de la modélisation des peuplements complexes ». Thèse de Doctorat. Nancy : ENGREF.

HEINRICH, Lothar (1991). « Goodness-of-fit tests for the second moment funciton of a stationary multidimensional poisson process ». *Statistics : A Journal of Theoretical and Applied Statistics* 22.2, p. 245–268. DOI : 10.1080/02331889108802308.

HOLMES, Thomas J. et John J. STEVENS (2004). « Spatial Distribution of Economic Activities in North America ». *Cities and Geography*. Sous la dir. de J. Vernon HENDERSON et Jacques-François THISSE. T. 4. Handbook of Regional and Urban Economics Chapter 63 - Supplement C. Elsevier, p. 2797–2843.

ILLIAN, Janine et al. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Statistics in Practice. Chichester : Wiley-Interscience, p. 534.

JENSEN, Pablo et Julien MICHEL (2011). « Measuring spatial dispersion : exact results on the variance of random spatial distributions ». *The Annals of Regional Science* 47.1, p. 81–110.

LAGACHE, Thibault et al. (2013). « Analysis of the Spatial Organization of Molecules with Robust Statistics ». *Plos One* 8.12, e80914.

LANG, G., E. MARCON et F. PUECH (2015). « Distance-Based Measures of Spatial Concentration : Introducing a Relative Density Function ». *HAL* hal-01082178.version 2.

LANG, Gabriel et Eric MARCON (2013). « Testing randomness of spatial point patterns with the Ripley statistic ». *ESAIM : Probability and Statistics* 17, p. 767–788.

MARCON, Eric et Florence PUECH (2003). « Evaluating the Geographic Concentration of Industries Using Distance-Based Methods ». *Journal of Economic Geography* 3.4, p. 409–428.

— (2010). « Measures of the Geographic Concentration of Industries : Improving Distance-Based Methods ». *Journal of Economic Geography* 10.5, p. 745–762.

— (2015a). « Mesures de la concentration spatiale en espace continu : théorie et applications ». *Économie et Statistique* 474, p. 105–131.

— (2017). « A typology of distance-based measures of spatial concentration ». *Regional Science and Urban Economics* 62, p. 56–67.

MARCON, Eric, Florence PUECH et Stéphane TRAISSAC (2012). « Characterizing the relative spatial structure of point patterns ». *International Journal of Ecology* 2012.Article ID 619281, p. 11.

MARCON, Eric et al. (2015b). « Tools to Characterize Point Patterns : dbmss for R ». *Journal of Statistical Software* 67.3, p. 1–15.

MAUREL, Françoise et Béatrice SÉDILLOT (1999). « A measure of the geographic concentration in french manufacturing industries ». *Regional Science and Urban Economics* 29.5, p. 575–604.

MØLLER, Jesper et Hakon TOFTAKER (2014). « Geometric Anisotropic Spatial Point Pattern Analysis and Cox Processes ». *Scandinavian Journal of Statistics*. Monographs on Statistics and Applies Probabilities 41.2, p. 414–435.

MØLLER, Jesper et Rasmus Plenge WAAGEPETERSEN (2004). *Statistical Inference and Simulation for Spatial Point Processes*. T. 100. Monographs on Statistics and Applies Probabilities. Chapman et Hall, 300 p.

OPENSHAW, S et P J TAYLOR (1979a). « A million or so correlation coefficients : three experiments on the modifiable areal unit problem ». *Statistical Applications in the Spatial Sciences*. Sous la dir. de N WRIGLEY. London : Pion, p. 127–144.

RIPLEY, Brian D. (1976). « The Second-Order Analysis of Stationary Point Processes ». *Journal of Applied Probability* 13.2, p. 255–266.

— (1977). « Modelling Spatial Patterns ». *Journal of the Royal Statistical Society* B 39.2, p. 172–212.

SILVERMAN, B. W. (1986). *Density estimation for statistics and data analysis*. London : Chapman et Hall, 175 p.

SWEENEY, Stuart H. et Edward J. FESER (1998). « Plant Size and Clustering of Manufacturing Activity ». *Geographical Analysis* 30.1, p. 45–64.

SWEENEY, Stuart H et Miguel GÓMEZ-ANTONIO (2015). « Localization and Industry Clustering Econometrics : an Assessment of Gibbs Models for Spatial Point Processes ». *Journal of Regional Science* 56.2, p. 257–287.

SZWAGRZYK, Jerzy et Marek CZERWCZAK (1993). « Spatial patterns of trees in natural forests of East-Central Europe ». *Journal of Vegetation Science* 4.4, p. 469–476.

VEEN, Alejandro et Frederic Paik SCHOENBERG (2006). « Assessing Spatial Point Process Models Using Weighted K-functions : Analysis of California Earthquakes ». *Case Studies in Spatial Point Process Modeling*. Sous la dir. d'Adrian BADDELEY et al. New York, NY : Springer New York, p. 293–306.

WIEGAND, T. et K. A. MOLONEY (2004). « Rings, circles, and null-models for point pattern analysis in ecology ». *Oikos* 104.2, p. 209–229.