Processus ponctuels spatiaux pour l'analyse du positionnement optimal et de la concentration

Florent Bonneu

THÈSE

présentée en vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

délivré par l'Université Toulouse I - Sciences Sociales

Discipline : Mathématiques
Spécialité : Statistique

par

Florent Bonneu

intitulée

Processus ponctuels spatiaux pour l'analyse du positionnement optimal et de la concentration.

Directeurs de thèse :
Abdelaati DAOUIA et Christine THOMAS-AGNAN

Soutenue le 19 juin 2009 devant le jury composé de Mesdames et Messieurs :

Avner Bar-Hen (Pr, Université Paris V), Rapporteur.
Liliane Bel (Mcf, AgroParisTech), Examinateur.
Noel Cressie (Pr, The Ohio State University), Rapporteur.
Abdelaati Daouia (Mcf, Université Toulouse I), Directeur.
Michel Goulard (CR, INRA Toulouse), Examinateur.
Marie Lebreton (Mcf, Université Bordeaux IV), Examinateur.
Christine Thomas-Agnan (Pr, Université Toulouse I), Directeur.

A mon épouse Fabienne
Et mes enfants :
Baptiste et Clémence.

Table des matières

Remerciements iii
Résumé vi
Abstract vii

0 Introduction **1**
0.1 Motivations 1
0.2 Processus ponctuels spatiaux 2
0.3 Problèmes de localisation-allocation 11
0.4 Processus empiriques et M-estimation 17
0.5 Indices de concentration et processus ponctuels marqués . . . 24

I Positionnement optimal par modélisation de processus ponctuels marqués **27**

1 Analyse exploratoire et modélisation de sinistres par processus ponctuels spatiaux marqués **29**
1.1 Introduction 30
1.2 Background on summary statistics 33
1.3 Time variation and spatial variation 36
1.4 Analysis of the workload mark 39
1.5 Model 43
1.6 Conclusions 51

2 Modèles de processus ponctuels spatiaux pour des problèmes de localisation-allocation **53**
2.1 Introduction 54
2.2 Literature on location-allocation problems 55
2.3 The case of the fire stations location problem 56
2.4 SPP location-allocation 59
2.5 Conclusions 73

II Propriétés asymptotiques de positions optimales empiriques 75

3 Consistance de positions empiriques conditionnelles 77
3.1 Introduction 77
3.2 Optimal policy specification 79
3.3 Consistency 80
3.4 A numerical illustration 81
3.5 Appendix : Proofs. 85

4 Estimation de positions optimales avec contraintes 89
4.1 Introduction 89
4.2 Main result 92
4.3 Appendix : Lemmas and proof. 93

III Indices de concentration et caractéristiques du second-ordre d'un processus ponctuel spatial marqué 99

5 Indices de concentration et caractéristiques du second-ordre d'un processus ponctuel spatial marqué 101
5.1 Indices de concentration 102
5.2 Caractéristiques du second-ordre pour des processus ponctuels marqués 105
5.3 Indices de concentration basés sur les distances 108
5.4 Comportement asymptotique de notre indice de concentration 111
5.5 Conclusion et perspectives 115

6 Conclusion 117

Bibliographie 118

Remerciements

Je souhaite remercier en premier lieu Christine THOMAS-AGNAN et Abdelaati DAOUIA pour m'avoir encadré durant ma thèse. Je leur suis très reconnaissant pour leur aide et leurs conseils lors de la direction de ce travail de recherche. Sans vouloir réduire leurs qualités à quelques mots, j'ai beaucoup apprécié chez Christine sa confiance et nos rendez-vous scientifiques ainsi que la détermination et la rigueur d'Abdelaati qui m'ont incité à persévérer dans mes recherches.

Je tiens ensuite à remercier Avner BAR-HEN et Noel CRESSIE d'avoir accepté, malgré toutes leurs occupations, d'être les rapporteurs de cette thèse. Je les remercie de l'intérêt qu'ils ont porté à mes travaux ainsi que leurs suggestions qui me permettent de cibler dans mes perspectives de recherche, les plus intéressantes.

Je tiens à exprimer ma reconnaissance à Michel GOULARD pour sa présence dans mon jury. Je considère sa présence comme un "clin d'oeil" sachant qu'avec Christine, ils m'ont fait découvrir la géostatistique pendant qu'Anne RUIZ-GAZEN m'initiait à l'économétrie spatiale lorsque j'étais étudiant.

Je souhaite également remercier Liliane BEL et Marie LEBRETON d'avoir accepté de faire partie de mon jury. J'ai eu l'occasion de les côtoyer plus récemment lors de congrès et les remercie pour l'intérêt qu'elles ont porté à mon mémoire malgré leur emploi du temps chargé.

Je tiens à présent à remercier plus généralement les membres du LSP et du GREMAQ que j'ai eu l'opportunité de côtoyer au cours de ma thèse et auparavant.

Au sein de l'université Paul Sabatier, je pense tout d'abord à Philippe BESSE et Alain BACCINI pour la qualité de leurs enseignements et leur écoute, à Fabrice GAMBOA qui m'a permis de découvrir le monde de la recherche lors d'un stage de maîtrise à l'INSERM, à Michel LEDOUX et Franck BARTHE pour avoir pris le temps de répondre à mes questions mathématiques, à Sébastien DÉJEAN pour des questions un peu plus informatiques, ainsi qu'à Jérémie BIGOT pour sa bonne humeur. Je n'oublie pas Marie-Laure AUSSET, Françoise MICHEL et Agnès REQUIS qui ont toujours été disponibles et efficaces pour les diverses questions administratives ainsi que Jacqueline dont nos discussions mycologiques me permettaient de m'évader de mes problèmes mathématiques.

En tant que moniteur à l'université des sciences sociales, j'ai eu le plaisir de partager mes enseignements avec Sandrine CASANOVA et Anne RUIZ-GAZEN et les remercie de leur encadrement et de leur gentillesse. J'ai eu beaucoup de plaisir à discuter avec mes anciens camarades de DESS, présents maintenant dans les bureaux de la Manufacture : Thibault LAURENT et VALÉRIE OROZCO. Enfin, en quelques mots, je peux dire que j'ai apprécié l'ambiance chaleureuse au GREMAQ.

Ces années de thèse m'ont permis de rencontrer de nombreux doctorants, arrivés par vagues successives. Tout d'abord, ceux qui m'ont accueilli au début de ma thèse : Agnès, Christophe, Delphine, Lionel et Renaud. Je me souviendrai longtemps des quizz à la cafétéria ou des quelques parties de foot sur les terrains de l'université, où Renaud, Lionel et Christophe savaient animer ces moments de détente à leur façon. Ensuite, ceux que j'ai eu le plaisir d'accueillir lors de leur arrivée en thèse : Jean-Paul, Mathieu, Michel, Maxime F. avec qui j'essaie de partager de temps en temps quelques bons moments autour d'une table. Je pense aussi à la composante sud-américaine de mon bureau actuel : Mary-Ana, Luiz et Angelica que je félicite pour leur intégration et que je remercie pour leur simplicité et leur gentillesse.

Pour finir, les derniers que je citerai mais non les moindres sont ceux qui m'ont accompagné depuis le DEA. J'ai eu beaucoup de plaisir à passer ces années de thèse en compagnie de ces trois mousquetaires : Amélie, Laurent et Maxime. Je remercie Amélie pour les discussions que nous avons eu, notamment autour de nos passions communes (plongée, jeux de société), et pour sa technique de recherche des meilleurs parcours SNCF. Je tiens aussi à remercier Maxime pour son sens de l'humour et nos discussions sportives ainsi que son aide pour les questions d'informatique en réseau. Enfin, je terminerai mon tour d'effectif par Laurent, qui a toujours partagé le même bureau que moi, même s'il m'a quitté ces derniers mois. Je tiens à lui dire combien j'apprécie sa simplicité et son extrême gentillesse et, pour répondre à ces propres remerciements de thèse, je suis persuadé que l'on trouvera le temps pour une collaboration scientifique.

Mes remerciements s'adressent aussi à mes amis qui se sont tenus au courant de l'avancée de ma thèse et qui m'ont écouté. Merci à Marlène et Jean, Séb et Béa, Laurie et Nico, Virginie et Mamour, Marie et Guillaume.

Enfin, je veux dire à mes parents et à mon frère combien ils me sont chers et que je les remercie de leur soutien durant ces années très chargées. Je sais

que mes parents sont fiers de me voir soutenir ma thèse, et que mon grand-père l'aurait été aussi, et tout ceci, c'est aussi grâce à eux. Pour finir, je dis un énorme MERCI à mon épouse Fabienne, que je sais soulagée aujourd'hui et qui m'a beaucoup soutenu. Sans le savoir, mes enfants Baptiste et Clémence ont contribué à me rendre les années de thèse beaucoup moins dures (en un certain sens), c'est pourquoi je leur dédie mon mémoire de thèse ainsi qu'à Fabienne pour tout ce qu'ils m'offrent chaque jour.

Résumé

Les processus ponctuels spatiaux forment une branche de la statistique spatiale utilisée dans des domaines d'application variés (foresterie, géo-marketing, sismologie, épidémiologie,...) et développée par de récents travaux théoriques. Nous nous intéressons principalement dans cette thèse à l'apport de la théorie des processus ponctuels spatiaux pour des problèmes de positionnement optimal, ainsi que pour la définition de nouveaux indices de concentration basés sur les distances en économétrie.

Le problème de positionnement optimal s'écrit souvent comme un problème d'optimisation prenant en compte des données geo-référencées auxquelles peuvent être associées des caractéristiques. Pour prendre en compte l'aléa, nous considérons ces données issues d'un processus ponctuel spatial pour résoudre un problème de positionnement stochastique plus réaliste qu'un modèle déterministe. A travers l'étude du positionnement optimal d'une nouvelle caserne de pompiers dans la région toulousaine, nous développons une méthode de résolution stochastique permettant de juger de la variabilité de la solution optimale et de traiter des bases de données volumineuses. L'approche implémentée est validée par des premiers résultats théoriques sur le comportement asymptotique des solutions optimales empiriques. La convergence presque sure des solutions optimales empiriques de l'étude de cas précédente est obtenue dans un cadre i.i.d. en utilisant la théorie de Vapnik-Cervonenkis. Nous obtenons aussi la convergence presque sure des solutions optimales empiriques, dans un cadre plus général, pour un problème de positionnement dérivé du problème de transport de Monge-Kantorovich.

Nous nous intéressons ensuite à des indices de concentration basés sur des distances en économétrie. Ces indices de concentration peuvent s'écrire comme des estimateurs de caractéristiques du second ordre de processus ponctuels marqués. Nous définissons ensuite un estimateur non-paramétrique d'une nouvelle caractéristique d'un processus ponctuel spatial marqué définissant ainsi un nouvel indice de concentration améliorant ceux déjà existants. Dans un cadre asymptotique avec fenêtre d'observation bornée, notre estimateur est asymptotiquement sans biais.

Mots clés : Processus ponctuels spatiaux marqués, problème de localisation-allocation, caractéristiques du second ordre, non et semi-paramétrique, problème de transport, M-estimation, Vapnik-Cervonenkis, indices de concentration, asymptotique sur domaine borné.

Abstract

Spatial point processes form a branch of spatial statistic used in various application areas (forestry, geo-marketing, seismology, epidemiology,...) and developed by recent theoretical results. We are interested primarily in this thesis to the use of spatial point processes theory for solving optimal positioning problems and for defining new concentration indices based on distances in econometrics.

The optimal positioning problem is often written as an optimization problem taking into account geo-referenced data with associated characteristics. We consider that the inputs of the problem are a realization of a spatial point process and solve a stochastic positioning problem more realistic than the deterministic model. Through the study of optimal positioning of a new fire station in the Toulouse area, we introduce a new stochastic approach that gives an indication of the spatial variability of the optimal solution and allows to solve larger problems. The implemented approach is validated by preliminary theoretical results on the asymptotic behavior of empirical solutions. The almost sure convergence of empirical optimal solutions of the case study is obtained in an i.i.d. framework using the Vapnik-Cervonenkis theory. We also obtain the almost sure convergence of empirical optimal solutions, in a more general framework, for a positioning problem derived from the Monge-Kantorovich transport problem.

Then we turn attention to concentration indices based on distances in econometrics. These concentration indices can be written as estimators of second order characteristics of marked point processes. We then define a nonparametric estimator of a new characteristic of a spatial marked point process defining a new concentration index improving existing ones. In an infill asymptotic framework, our estimator is asymptotically unbiased.

Key words : Marked spatial point processes, location-allocation problem, second-order characteristics, non and semi-parametric, transport problem, M-estimation, Vapnik-Cervonenkis, concentration indices, Infill asymptotics.

Chapitre 0

Introduction

0.1 Motivations

Pendant ma thèse, je me suis tout d'abord intéressé à l'apport de la théorie des processus ponctuels spatiaux pour l'étude de la solution optimale d'un problème de positionnement. La détermination d'une localisation optimale est un problème fréquent dans de nombreux domaines, malheureusement cette problématique est souvent traitée d'un point de vue déterministe, alors que la nature des données est en général aléatoire.

La partie I de ce mémoire concerne une étude de cas relative au positionnement optimal d'une nouvelle caserne de pompiers dans la région toulousaine. L'originalité de l'étude est de considérer les données du problème d'optimisation comme aléatoires et issues d'un processus ponctuel spatial. L'étape préliminaire d'analyse exploratoire de la base de données des sinistres et la modélisation par un processus ponctuel marqué est réalisée dans le chapitre 1 (Bonneu, 2007). La méthode "SPP location-allocation" introduite et implémentée dans le chapitre 2 (Bonneu et Thomas-Agnan, 2008) fournit une représentation de la variabilité de la position optimale.

La partie II de ce mémoire étudie le comportement asymptotique des solutions de problèmes de localisation-allocation empiriques vers la solution du problème de positionnement théorique. Ces résultats théoriques sont importants car ils permettent de justifier en pratique l'approximation d'une solution théorique inconnue par la solution d'un problème d'optimisation empirique. Dans un problème de localisation-allocation dérivé du problème de transport de Monge-Kantorovich, le chapitre 3 (Bonneu et Daouia, 2008) établit la convergence forte de l'estimateur de la position optimale dans un

cadre général où les données peuvent être corrélées ou indépendantes. La technique de preuve est basée sur les propriétés des distances de probabilités dans les problèmes de transport optimal. Dans le cadre du problème de localisation d'une nouvelle caserne de pompiers, des résultats asymptotiques sont obtenus pour les solutions de problèmes d'optimisation empiriques dans le chapitre 4. Pour des données indépendantes et identiquement distribuées (i.i.d.), la convergence forte des solutions optimales empiriques est démontrée en s'appuyant sur la théorie de Vapnik-Chervonenkis.

Enfin, la partie III de ce mémoire présente des propriétés asymptotiques de caractéristiques du second ordre pour des processus ponctuels spatiaux marqués. Ces caractéristiques du second ordre sont directement liées à la définition d'un nouvel indice de concentration basé sur les distances améliorant ceux déjà existants en économétrie. Ces indices de concentration permettent par exemple d'analyser les déterminants de la localisation des entreprises. La recherche des facteurs influant sur la concentration des entreprises peut aboutir à la définition des orientations d'une politique générale d'aménagement du territoire. Même si ce n'est pas le cas aujourd'hui, leur utilisation pourrait ainsi s'étendre aux problèmes de positionnement optimal.

Cette introduction rappelle de façon concise quelques notions essentielles à la lecture des chapitres de ce mémoire et présente les contributions de la thèse dans les domaines considérés.

0.2 Processus ponctuels spatiaux

0.2.1 Présentation

Les processus ponctuels spatiaux représentent une branche de la statistique spatiale où l'on étudie des collections d'évènements localisés par leur coordonnées géographiques, appelées semis de points. Grâce aux moyens technologiques actuels, ces jeux de données apparaissent dans de nombreux domaines d'application (foresterie, géo-marketing, sismologie, épidémiologie,...) et sont de plus en plus volumineux. On peut ainsi être amené à étudier la disposition des arbres dans une forêt, la localisation d'entreprises, les épicentres de secousses sismiques, des adresses de patients atteints d'une maladie,...Des variables de différents types (réelles, entières, booléennes,...) peuvent être associées à chaque évènement et seront appelées "marques". A titre d'exemple, la position d'une entreprise est souvent plus intéressante lorsque l'on a des informations sur son secteur d'activité, son nombre d'employés, son chiffre

d'affaire,...Une marque un peu particulière est la marque temporelle dont la connaissance permet d'ajouter une composante dynamique à l'étude du semis de points. Les ouvrages de références traitant de processus ponctuels spatiaux sont les monographies de Moller et Waagepetersen (2004), Cressie (1993), Stoyan et Stoyan (1994) ou Diggle (2003).

Un semis de point est la réalisation d'un processus ponctuel sur un espace polonais (A, d) et se définit comme une collection non ordonnée de points x_i de A, pour tout $i = 1, \ldots, n$; notée $\{x_i; i = 1, \ldots, n\}$. Un processus ponctuel spatial représente donc une configuration aléatoire de positions dans A notée $\{\xi_i; i = 1, \ldots, N\}$ où le nombre total d'évènements N est aussi aléatoire. Bien qu'envisageable, la répétition de points $\xi_i = \xi_j$ pour $i \neq j$ est exclue dans diverses définitions ou propriétés des processus ponctuels spatiaux. Sans répétition de points le processus est dit simple. De plus, toute configuration d'un processus ponctuel spatial doit être localement finie, c'est à dire qu'elle ne doit pas avoir de points d'accumulation.

0.2.2 Caractéristiques du 1er et 2nd ordre

Les processus ponctuels spatiaux possédent des caractéristiques définies à partir des moments de leur mesure de comptage aléatoire Φ (Cressie, 1993). Ces caractéristiques jouent le même rôle que les moments définis pour une variable aléatoire. Ainsi, la connaissance de quelques unes de ces statistiques élémentaires ne permet pas d'identifier complètement un processus ponctuel spatial. Cependant, l'utilisation d'estimateurs de ces statistiques est abondante en analyse exploratoire ou en modélisation lors de l'ajustement et la validation d'un modèle. De manière générale, nous rappelons les mesures de moments d'ordre $p \in \mathbb{N}^*$ et présentons plus précisément les caractéristiques du second ordre pour des processus ponctuels inhomogènes.

Caractéristiques théoriques

Nous commençons par rappeler la définition de la caractéristique du premier ordre.

Définition 0.1. *La mesure d'intensité μ se définit pour tout borélien D de A par*

$$\mu(D) = \mathbb{E} \sum_{\xi \in X} \mathbb{1}[\xi \in D].$$

Si la mesure intensité peut s'écrire sous la forme

$$\mu(D) = \int_D \lambda(\xi) d\xi,$$

avec λ une fonction positive, alors λ est appelée fonction intensité.

Lorsque λ est constante le processus est dit homogène ou stationnaire du premier ordre. Dans ce cas, l'intensité λ représente l'espérance du nombre de points par unité de volume. Si λ n'est pas constante, le processus est dit inhomogène et de façon heuristique, on dit que $\lambda(\xi)d\xi$ est la probabilité d'occurence d'un point dans une boule infinitésimale de centre ξ et de volume $d\xi$.

Nous rappelons maintenant les mesures moments d'ordre p de la mesure aléatoire $\Xi = \sum_{\xi \in X} \frac{1}{\lambda(\xi)} \delta_\xi$, avec $\lambda(\xi) > 0$ presque sûrement pour tout $\xi \in X$. Ces mesures introduites dans Baddeley *et al.* (2000) sont une extension des mesures $\mu^{(p)}$ et $\alpha^{(p)}$ définies à partir de la mesure de comptage Φ (voir par exemple Moller et Waagepetersen, 2004).

Définition 0.2. *Pour un processus ponctuel X d'intensité λ dans A et tout $p \in \mathbb{N}^*$, on définit la mesure de moment d'ordre p de Ξ, $\nu^{(p)}$ sur A^p par*

$$\nu^{(p)}(D) = \mathbb{E} \sum_{\xi_1, \cdots, \xi_p \in X} \frac{\mathbb{1}[(\xi_1, \cdots, \xi_p) \in D]}{\lambda(\xi_1) \cdots \lambda(\xi_p)}$$

où D est un borélien de A^p, et la mesure factorielle de moment d'ordre p de Ξ, $\beta^{(p)}$ sur A^p par

$$\beta^{(p)}(D) = \mathbb{E} \sum_{\xi_1, \cdots, \xi_p \in X}^{\neq} \frac{\mathbb{1}[(\xi_1, \cdots, \xi_p) \in D]}{\lambda(\xi_1) \cdots \lambda(\xi_p)}$$

où D est un borélien de A^p. La notation $\sum^{\neq}_{\xi_1, \cdots, \xi_p \in X}$ indique que les $\xi_1, \cdots, \xi_p$ sont distincts.

Par la suite, on s'intéressera plus particulièrement à la mesure factorielle $\beta^{(2)}$ pour définir des caractéristiques théoriques du second ordre d'un processus ponctuel.

Définition 0.3. *Le processus ponctuel X est stationnaire du second ordre avec pondération par l'intensité ("second-order intensity-reweighted stationary") si $\beta^{(2)}(D_1 \times D_2) = \beta^{(2)}((D_1 + x) \times (D_2 + x))$ pour tout D_1, D_2 boréliens de A et x un veteur de A, où $D_1 + x$ désigne la translation de D_1 par le vecteur x.*

Des exemples de processus ponctuels stationnaires du second ordre avec pondération par l'intensité sont les processus ponctuels de Poisson inhomogènes, certains processus Log Gaussiens inhomogènes et tous les processus ponctuels obtenus par éclaircissement ("thinning") d'un processus stationnaire par un champ aléatoire indépendant.

La fonction K introduite par Ripley (1976) pour des processus ponctuels stationnaires se généralise à des processus ponctuels stationnaires du second ordre avec pondération par l'intensité (Baddeley *et al.*, 2000).

Définition 0.4. *Soit X un processus ponctuel stationnaire du second ordre avec pondération par l'intensité λ. La fonction K_{inhom} de X est définie par*

$$K_{inhom}(r) = \frac{1}{|A|}\mathbb{E}\sum_{\xi_i \in X}\sum_{\xi_j \in X}^{\neq} \frac{\mathbb{1}(\|\xi_i - \xi_j\| \leq r)}{\lambda(\xi_i)\lambda(\xi_j)}, \quad r \geq 0.$$

où $|A|$ désigne l'aire de A (mesure de Lebesgue).

Nous présentons des conditions nécessaires pour assurer l'existence d'une fonction K_{inhom} pour des processus ponctuels inhomogènes et montrons le lien avec la fonction de corrélation des paires. Si les mesures μ et $\beta^{(2)}$ se définissent respectivement à partir de fonctions positives notées λ et $\rho^{(2)}$ alors on a

$$\beta^{(2)}(D_1 \times D_2) = \int_{D_1}\int_{D_2} \frac{\rho^{(2)}(u,v)}{\lambda(u)\lambda(v)} dudv = \int_{D_1}\int_{D_2} g(u,v)dudv$$

où $g(u,v) = \rho^{(2)}(u,v)/(\lambda(u)\lambda(v))$ pour tout $u, v \in A$ est la fonction de corrélation des paires de X.

Si g est invariant par translation, c'est à dire $g(x, x+h) = g_0(h)$ pour une fonction $g_0 : A \to [0, +\infty)$, alors X est stationnaire du second ordre avec pondération par l'intensité et la fonction K_{inhom} est donnée par

$$K_{inhom}(r) = \int_{B(0,r)} g_0(h)dh$$

D'autres fonctions dérivées de K_{inhom} peuvent être définies comme dans le cas stationnaire, comme par exemple la fonction $L_{inhom} = (K_{inhom}/\pi)^{1/2}$. Dans le cas Poissonien, $K_{inhom}(r) = \pi r^2$ et donc $L_{inhom}(r) = r$. Les valeurs de ces caractéristiques obtenues pour des processus ponctuels de Poisson servent de référence pour tester l'hypothèse Poissonienne d'un processus ponctuel. La section suivante présente quelques estimateurs de ces caractéristiques et les difficultés rencontrées.

Estimation

Pour simplifier la présentation on considèrera dans cette section que A est un sous-ensemble borné de $\mathbb{R}^2$.

L'estimation de l'intensité est la première étape dans l'analyse d'un semis de point. Compte tenu de la définition de la mesure intensité, un estimateur naturel et sans biais dans le cas homogène s'écrit $n/|A|$. Dans le cadre inhomogène, les estimateurs de l'intensité dérivent des méthodes d'estimations non paramétriques de la densité multidimensionnelle (Silverman, 1986), où l'on tient compte du fait que l'intégrale de l'intensité sur le domaine est égale à l'espérance du nombre de points. Malheureusement, des problèmes de sous-estimation apparaissent aux bords du domaine, ce qui nous conduit à considérer l'estimateur de l'intensité dans A proposé par Diggle (1985)

$$\hat{\lambda}(x) = \frac{\sum_{i=1}^{n} h^{-2} w(h^{-1}(x - x_i))}{\hat{c}_{A,h}(x)} \tag{0.1}$$

où $\hat{c}_{A,h}(x)$ est un estimateur du facteur de correction de bord $\int_A h^{-2} w(h^{-1}(x-u))du$. Si le choix du noyau w n'est pas essentiel en revanche celui de la fenêtre de lissage h s'avère souvent primordial. Berman *et al.* (1989) proposent de choisir h à partir de la minimisation d'une estimation de la moyenne des erreurs au carré, mais cette méthode contribue à fournir une valeur de h proche de zéro quand les variations de λ sont trop fortes, comme dans le chapitre 1 ou Diggle *et al.* (2007). Les méthodes paramétriques d'estimation sont possibles mais finalement peu répandues à l'inverse de méthodes semi-paramétriques utilisant des covariables connues ou estimées. Il est ainsi fréquent d'estimer λ par

$$\hat{\lambda}(x) = \exp(\hat{\alpha} + \sum_{j=1}^{J} \hat{\beta}_j \hat{C}_j(x)),$$

où les $\hat{C}_j$ sont estimés lors d'une étape préliminaire si elles sont inconnues et où les paramètres $\alpha, \beta_1, \cdots, \beta_J$ peuvent être estimés par maximum de pseudo-vraisemblance. Les covariables estimées $\hat{C}_j$ peuvent être par exemple, la densité d'espèces d'arbres, la densité de population ou tout autre densité d'une variable pouvant expliquer l'inhomogénéité de la réalisation. Dans le chapitre 1, on considère comme covariable une estimation de la densité de population.

D'autres estimations de l'intensité existent dans des cas particuliers. Pour des données bruitées, Cucala (2008) introduit un estimateur prenant en

compte l'erreur de positionnement. Baddeley *et al.* (2000) introduisent un nouvel estimateur $\bar{\lambda}$ pour minimiser le biais de l'estimateur de K_{inhom} par rapport à celui construit avec l'estimateur classique (0.1).

Lorsque λ est connue, un estimateur sans biais de K_{inhom} s'écrit

$$\hat{K}_{inhom}(r) = \frac{1}{|A|} \sum_{x_i \in X} \sum_{x_j \in X}^{\neq} \frac{w_{x_i,x_j,r} \mathbb{1}(\|x_i - x_j\| \leq r)}{\lambda(x_i)\lambda(x_j)}, \quad 0 \leq r < r^*.$$

où $w_{x_i,x_j,r}$ est un terme de correction de bord et $r^* := \max\{d(x, \partial A); x \in A\}$, avec ∂A désignant le bord du domaine A. Un terme de correction de bord proposé par Ripley (1977) s'écrit $w_{x_i,x_j,r} = |(W + x_i) \cap (W + x_j)|$. Dans la grande majorité des cas λ est inconnue et doit être estimée. Avec l'estimateur (0.1), l'estimateur de K_{inhom} qui en résulte est biaisé. Une diminution du biais est obtenue en prenant l'estimateur $\bar{\lambda}$ présenté dans Baddeley *et al.* (2000).

0.2.3 Quelques processus ponctuels

Les processus de Poisson

Le modèle de processus ponctuels le plus élémentaire est le processus ponctuel de Poisson. Dans le cadre homogène, ce processus correspond à l'hypothèse nulle du test de répartition spatiale totalement aléatoire ("Complete Spatial Randomness"). Les processus ponctuels de Poisson inhomogènes sont entièrement caractérisés par leur fonction intensité d'après le théorème de Slivnyak-Mecke (voir par exemple Moller et Waagepetersen, 2004). Ces processus servent à modéliser une répartition inhomogène de points dans l'espace répartis de façon indépendante. Les processus de Poisson sont tels que les variables aléatoires $\Phi(B_1), \ldots, \Phi(B_m)$ pour tout $B_1, \ldots, B_m$ dans A sont indépendantes et de loi de Poisson de paramètre respectif $\int_{B_1} \lambda(x)dx, \ldots, \int_{B_m} \lambda(x)dx$. De plus, conditionnellement au nombre de points dans A les positions $\xi_1, \ldots, \xi_n$ sont indépendamment distribuées selon la densité bidimensionnelle $\frac{\lambda(x)}{\int_A \lambda(x)dx}$. Au chapitre 1, ces modèles de processus sont utilisés pour tester le caractère Poissonien du processus sous-jacent à notre réalisation.

Les processus de Cox

Les processus de Cox sont des modèles pour des réalisations de processus présentant des agrégats causés par une hétérogénéité environnementale aléatoire. Le processus de Cox est une extension naturelle d'un processus de Poisson, obtenu en considérant la fonction intensité d'un processus de Poisson comme la réalisation d'un champ aléatoire Λ. C'est pour cette raison

qu'ils ont été introduits par Cox (1955) sous le nom de "processus de Poisson doublement stochastiques". Certains processus de Cox sont obtenus en regroupant les points autour de ceux d'un autre processus ponctuel comme les processus d'agrégats de Matern ou de Thomas qui appartiennent à la famille des processus de Neyman-Scott (voir par exemple, Moller et Waagepetersen, 2004). Dans les applications, à moins de connaissances a priori, il n'est pas possible de distinguer un processus de Cox X d'intensité Λ du processus de Poisson correspondant $X|\Lambda$ quand seulement une réalisation du processus est disponible. Les processus de Cox Log Gaussiens X (Moller *et al.* (1998)) sont tels que $Y = \log \Lambda$ est un champ Gaussien, c'est à dire que pour tout entier $n > 0$, toutes positions $\xi_1, \ldots, \xi_n \in A$, et tous nombres $a_1, \ldots, a_n \in \mathbb{R}$, $\sum_{i=1}^n a_i Y(\xi_i)$ suit une loi normale. Le caractère agrégatif des processus de Cox nous a conduit à envisager un certain nombre de modèles au chapitre 1.

Les processus de Markov

Cette famille de processus sert le plus souvent à modéliser un comportement répulsif entre les points, cependant des possibilités existent pour modéliser aussi de l'attraction. Ces modèles sont construits en spécifiant pour le processus ponctuel une densité par rapport à un processus de Poisson et en imposant des conditions pour satisfaire une propriété de Markov. Le lecteur pourra se référer à la monographie de Van Lieshout (2000) pour un état de l'art récent sur les processus ponctuels Markoviens. Au chapitre 1, l'aspect agrégatif de notre semis de points nous a conduit à ne pas considérer de modèles de processus ponctuels de Markov.

0.2.4 Processus ponctuels marqués

Nous présentons dans cette section la définition de processus ponctuel marqué de Poisson, une introduction aux tests de corrélation des positions et une présentation succinte de l'étude de la corrélation des marques.

Processus ponctuels marqués de Poisson

Définition 0.5. *Un processus $Y = \{(\xi, \mathfrak{m}_\xi) : \xi \in X\}$ est un processus ponctuel marqué de Poisson si X est un processus ponctuel de Poisson et si les marques $\{m_\xi : \xi \in X\}$ sont indépendantes entre elles, conditionnellement à X.*

Par cette définition, on a que tout processus ponctuel de Poisson Y dans un espace produit $T \times U$ est un processus ponctuel de Poisson marqué avec

positions dans T et marques dans U. L'inverse est faux en général. La Proposition 3.9 dans Moller et Waagepetersen (2004, p.26) fournit une condition sous laquelle l'équivalence est vérifiée. Ainsi, si un processus ponctuel marqué de Poisson est tel que, conditionnellement à X, chaque marque $\mathfrak{m}_\xi$ a une densité ne dépendant pas de $X \backslash \xi$ alors Y est un processus ponctuel de Poisson dans l'espace produit et sous une condition d'intégrabilité $\{\mathfrak{m}_\xi : \xi \in X$ est un processus ponctuel de Poisson dans l'espace des marques.

Corrélation des positions

La nature des marques associées à des positions est souvent de deux sortes : discrète ou continue. Dans le cas de marques discrètes prenant k valeurs distinctes, on considère le processus multivarié $X = (X_1, \ldots, X_k)$. La structure de corrélation entre les points de X_i et X_j, pour $i, j = 1, \cdots, k$ avec $i \neq j$, est étudiée en utilisant des caractéristiques croisées du second ordre ("cross summary statistics"). Ces caractéristiques sont définies à partir des caractéristiques du second ordre comme les fonctions K,L ou g. Tout d'abord introduites dans un cadre stationnaire, la généralisation à certains processus inhomogènes ("cross second order intensity reweighted") se trouve dans Baddeley *et al.* (2000) et se déduit de la section précédente. Le lecteur peut se référer à Moller et Waagepetersen (2004) pour une présentation plus approfondie des caractéristiques croisées. Dans le cas de processus ponctuels bivariés, Bar-Hen et Picard (2006) réalisent une comparaison entre 9 indices de dissimilarité.

Dans le cas de marques continues, nous présentons deux approches possibles pour étudier la corrélation des positions en tenant compte de la marque. La première consiste à discrétiser la marque et à appliquer les caractéristiques présentées dans le paragraphe précédent. La seconde méthode consiste à analyser le caractère séparable de l'intensité du processus ponctuel marqué (X, M) par les méthodes présentées dans Schoenberg (2004). En effet, si l'intensité du processus ponctuel marqué peut s'écrire comme le produit de l'intensité du processus ponctuel des positions par la densité de la variable aléatoire des marques alors les processus marginaux sont indépendants. Nous appliquons cette méthode au chapitre 1 pour tester la dépendance deux à deux entre les positions, les marques et le temps.

Corrélation des marques

L'analyse de la corrélation des marques s'appuie sur les caractéristiques du second ordre de processus ponctuels marqués et s'appuie sur les méthodes

employées en géostatistique. On étudie ainsi l'espérance de fonctions des marques conditionnelement aux positions. Le variogramme empirique conditionnel est ainsi une adaptation au cadre des processus ponctuels du variogramme défini en géostatistique. Le lecteur peut se référer à Stoyan et Stoyan (1994), Schlather (2001), Mateu (2000) ou au chapitre 5 pour une présentation plus détaillée. Nous rappelons maintenant brièvement quelques notions d'indépendance. L'hypothèse dite de "random labelling" consiste à supposer que les marques sont indépendantes entre elles et indépendantes des positions alors que l'hypothèse dite de "geostatistical marking" suppose seulement que les marques sont indépendantes des positions.

0.2.5 Analyse exploratoire et modélisation de sinistres par processus ponctuels spatiaux marqués

Le chapitre 1 réalise l'analyse exploratoire et la modélisation de sinistres contenus dans une base de données fournie par le Service Départemental d'Incendies et de Secours de Haute-Garonne (SDIS31). Cette base contient les positions et les caractéristiques (temps, durée, nombre de pompiers en intervention) de 20820 sinistres survenus au cours de l'année 2004 dans les environs de Toulouse. Chaque sinistre i sera identifié par sa position X_i et une marque D_i mesurant une charge de travail (durée $\times$ nombre de pompiers). L'analyse exploratoire de ce jeu de données et l'ajustement d'un modèle de processus ponctuel spatial marqué constituent une étape préliminaire importante pour la méthodologie employée au chapitre 2.

Une étude des dépendances entre positions et temps, marques et positions ainsi que marques et temps a été réalisée par des méthodes graphiques et/ou des tests de séparabilité présentés par Schoenberg (2004). Cette analyse des dépendances deux à deux a permis de supposer raisonnablement que le modèle de processus ponctuel spatial marqué pouvait être obtenu en modélisant séparément les réalisations marginales.

Sous l'hypothèse de séparabilité, il est possible de modéliser séparément les charges de travail et l'ensemble des positions des sinistres. Un problème majeur dans la modélisation d'un semis de point est toujours de séparer l'inhomogénéité expliquée par l'intensité λ et les intéractions mesurées par la fonction L_{inhom}. La répartition des sinistres étant bien évidemment liée à la répartition hétérogène de la population, nous devons considérer des modèles de processus ponctuels spatiaux d'intensité λ non constante. Le choix du modèle adéquat se fera grâce à un test d'interaction calculé par méthode

Monte-Carlo pour obtenir une enveloppe de la fonction L_{inhom}. Une attention particulière a aussi été prêtée à l'intensité estimée et à la répartition des points d'une réalisation simulée issue du modèle ajusté.

L'inefficacité des méthodes paramétriques, dans notre étude de cas, nous conduit à étudier l'estimation de l'intensité par des méthodes non-paramétriques. Les estimations non-paramétriques $\hat{\lambda}$ et $\bar{\lambda}$ proposées respectivement par Diggle (1985) et Baddeley *et al.* (2000) ne permettent pas d'aboutir à des modèles satisfaisants du point de vue de la représentation de la fonction L_{inhom}. Le fait d'estimer λ et L_{inhom} sur un même jeu de données apparait problématique. C'est pour cette raison que nous introduisons deux méthodes d'estimation de λ semi-paramétriques basées sur la covariable densité de population, estimée à partir du nombre d'habitants par unité administrative (IRIS). Les meilleurs résultats sont obtenus pour une intensité calculée à partir de la densité de population estimée par la méthode des k-plus proches voisins.

$$\hat{C}_2(s) = \frac{1}{N_{Pop} h_s} \sum_{i=1}^{296} N_i k_e \left(\frac{s - s_i}{h_s} \right)$$

où s_i est le centre de l'IRIS i, N_i son nombre d'habitants, $N_{Pop} = \sum_{i=1}^{296} N_i$, $k_e(s) = \frac{3}{4}(1 - \|s\|^2)$ et $h_s = \|s - \xi\|_{(5)}$ est la cinquième statistique d'ordre des distances entre s et le centre des IRIS.

Plusieurs modèles de processus de Cox sont testés (Matern, Thomas, Log Gaussien). L'inhomogénéité est incorporée par la méthode d'éclaircissement comme dans Waagepetersen (2006). Cette méthode permet de connaître l'expression de la fonction K inhomogène théorique en fonction de paramètres κ et ω pour les modèles explorés. Ces paramètres sont ensuite estimés grâce à la minimisation du contraste

$$\int_0^a (\hat{K}_{inhom}(t)^q - K(t; \kappa, \omega)^q)^2 dt$$

où K est une fonction connue pour les modèles considérés et où $q = 1/4$ et $a = 4000$ sont choisies selon les recommandations dans Diggle (2003). Les modèles de processus ponctuels de Poisson et Log Gaussien apparaissent satisfaisants dans notre étude de cas.

0.3 Problèmes de localisation-allocation

Dans cette section, nous introduisons de façon générale le problème de positionnement optimal et quelques méthodes de résolution. L'approche stochas-

tique des problèmes de localisaton-allocation est peu répandue alors que les données sont souvent les réalisations de processus aléatoires. La section 0.3.3 présente la méthodologie développée au chapitre 2 pour le positionnement d'une nouvelle caserne de pompiers dans la région toulousaine. La section 0.3.2 présente le problème probabiliste de transport optimal et établit le lien avec le problème de positionnement optimal. La formulation du problème de transport sera ensuite reprise dans la section 0.4.4 et au chapitre 3 pour démontrer la convergence de solutions optimales empiriques d'un problème de positionnement particulier.

0.3.1 Positionnement optimal

Les problèmes de positionnement optimal sont nombreux et divers : où placer une nouvelle caserne de pompiers, une nouvelle formation d'enseignement, un point de stockage de bois, un silo agricole,. . .En recherche opérationnelle, ce problème est souvent traité d'un point de vue déterministe. En effet, la position optimale résulte souvent d'un problème d'optimisation construit sur un unique semis de points obtenu par exemple au cours d'une seule année. Les problèmes de positionnement sont souvent appelés problèmes de localisation-allocation car la recherche d'une position optimale est indissociable de l'obtention d'une fonction d'affectation. La difficulté de ce problème est que la recherche du couple (position,affectation) optimal se trouve dans un espace de grande dimension et que position et affectation sont corrélées.

Le problème de positionnement optimal d'une nouvelle caserne de pompiers appartient à la famille des problèmes de localisation-allocation conditionnels. Le mot conditionnel indique que le positionnement d'une nouvelle caserne de pompiers doit tenir compte des casernes existantes. Notre problème de positionnement ne consiste pas à redistribuer l'ensemble des casernes sur le domaine d'étude. Le lecteur peut se référer à ReVelle et Eiselt (2005) pour une présentation complète sur les problèmes de localisation-allocation.

Même si une grande partie de la littérature est axée sur l'approche de résolution déterministe, certains papiers introduisent la dimension aléatoire du problème (positions, temps de trajet,. . .). Cependant, l'approche par processus ponctuels spatiaux semble n'avoir jamais été envisagée. Snyder (2006) fournit un état de l'art sur la prise en compte de l'aléa des problèmes de positionnement et sur les modèles probabilistes développés pour la localisation robuste. Cooper (1974) a considéré l'extension du problème de Weber (où la fonction de coût s'écrit comme la distance Euclidienne) au cas où les

positions sont indépendantes et issues de variables aléatoires Gaussiennes. Drezner (1985) analyse la robustesse de la position optimale d'un problème de Weber lors de petites fluctuations des positions des clients. Le caractère aléatoire des demandes de clients est considéré dans Logendran et Terrel (1988).

Le cas du positionnement optimal d'une nouvelle caserne de pompiers est traité dans Serra et Marianov (1998) pour la ville de Barcelone mais sans prendre en compte l'aléa. Les problèmes de positionnement optimal d'un service d'urgence (caserne, hôpital,...) sont parfois traités sous l'aspect d'un problème de couverture où le temps d'accés d'un service d'urgence ne doit pas dépasser un certain seuil et où on prend en compte une éventuelle indisponibilité temporaire (Daskin (1982), Goldberg et Paz (1991), ReVelle (1991)).

Du point de vue de la résolution de ces problèmes, Cooper (1964) introduit les algorithmes heuristiques qui fournissent des solutions approchées au problème en alternant des phases de positionnement et d'affectation optimale. Brimberg *et al.* (1997) fournissent un résumé intéressant sur les algorithmes heuristiques utilisés pour des problèmes de localisation-allocation. Parmi les méthodes récentes, on peut citer la recherche Tabou (Brimberg et Mladenovic, 1996), les algorithmes génétiques (Houck *et al.*, 1996), les colonies de fourmis (Bischoff et Dachert, 2007),...

Nous présentons dans la section suivante le problème de transport de Monge-Kantorovich qui nous permet ensuite de définir un problème de positionnement optimal à partir de ce problème probabiliste.

0.3.2 Transport optimal

Le lecteur pourra se référer aux monographies de Villani (2003), Rachev (1991), Rachev et Rüschendorf (1998) pour davantage de détails.

Le problème de transport optimal a été introduit par Monge (1781) lors de l'étude d'un problème de déblais et de remblais. Le problème consistait à transporter une pile de sable X dans un trou Y de même volume en un coût minimal. Ce problème se modélise par le transfert optimal de la mesure μ de l'espace mesuré X vers la mesure ν de l'espace mesuré Y. On doit nécessairement avoir $\mu(X) = \nu(Y)$. Le transport du sable nécessite un certain effort, modélisé par une fonction de coût mesurable et positive c définie sur $X \times Y$ telle que $c(x, y)$ représente le coût de transport de la particule en x vers la localisation y. Dans le cas de l'affectation de sinistres aux casernes

de pompiers, si les positions des sinistres et des casernes sont connues alors l'allocation optimale est la solution d'un problème de transport. Cependant, dans le cas du positionnement optimal d'une nouvelle caserne de pompiers, on a évidemment la mesure cible ν qui dépend de la position de la nouvelle caserne et n'est donc pas déterminée.

Parce que le problème de Monge est non-linéaire et parfois difficile à résoudre (ou n'admet pas de solution), Kantorovich (1942, 1948) a défini le problème de transport relaxé portant son nom. Dans le problème de Kantorovich on appelle plans de transfert les mesures π sur l'espace produit $X \times Y$. $d\pi(x, y)$ mesure la quantité de masse transférée de x vers y. La possiblité d'avoir une ou plusieurs solutions pour le problème de Kantorovich et aucune pour le problème de Monge réside essentiellement sur une différence majeure. En effet, dans le problème de Kantorovich la masse en un point x peut être découpée et affectée à plusieurs localisations y. Cette possibilité ne semble pas incongru en pratique si l'on considère l'affectation de clients à des magasins où les clients peuvent choisir d'aller à plusieurs magasins qui leur sont proches.

Le problème de Kantorovich consiste à minimiser le coût de transport total

$$I[\pi] = \int_{X \times Y} c(x, y) d\pi(x, y) \quad \text{pour } \pi \in \Pi(\mu, \nu)$$

où $\Pi(\mu, \nu)$ est l'ensemble des mesures sur $X \times Y$ de marginales μ et ν. Le coût de transport optimal entre μ et ν est donc la valeur

$$\mathcal{A}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} I[\pi]$$

Les π correspondant à la minimisation de ce coût de transport sont appelés plans de transfert optimaux. Le lecteur peut se référer à Villani (2003) pour une écriture probabiliste du problème de transport avec des paires de variables aléatoires.

Dans le problème de Monge, parce que chaque position se voit affectée à une seule destination y, les plans de transfert π s'écrivent sous la forme

$$d\pi(x, y) = d\pi_T(x, y) \equiv d\mu(x)\delta[y = T(x)],$$

où T est une application mesurable de $X \to Y$. La condition pour π_T d'appartenir à $\Pi(\mu, \nu)$ se réécrit $\nu = T\sharp\mu$ qui signifie que

$$\nu(B) = \mu(T^{-1}(B)), \quad \text{pour tout ensemble mesurable } B \subset Y.$$

Le problème de Monge s'écrit donc comme le problème de minimisation du coût total

$$I[T] = \int_X c(x, T(x)) d\mu(x),$$

sur l'ensemble des applications mesurables T telles que $T\sharp\mu = \nu$.

La littérature dans le domaine du problème de transport de Monge-Kantorovich est très abondante en ce qui concerne les questions d'existence et d'unicité des solutions (voir par exemple, Rachev 1985, Gangbo and McCann 1996). La convergence des plans de transfert (ou des fonctions d'affectation) occupe aussi une place importante lorsque l'on considère des problèmes de transport où les mesures sources μ et cibles ν convergent (voir par exemple, Villani 2003). Nous présentons ci-dessous un théorème énoncé dans Villani (2003, p.207) sur les distances de Wasserstein et employé au chapitre 3. Pour cela, on considère (X, d) un espace polonais et des fonctions de coût $c(x, y) = d(x, y)^p$, avec $p \geq 0$. On prend pour convention que $d(x, y)^0 = \mathbb{1}_{x \neq y}$. Les mesures μ et ν sont des mesures de probabilité définies sur X.

Théorème 0.1. *(Distances de Wasserstein).*

(i) Pour tout $p \in [1, \infty)$, $W_p = \mathcal{A}_c^{1/p}$ définit une métrique sur $P_p(X)$.

(ii) Pour tout $p \in [0, 1)$, $W_p = \mathcal{A}_c$ définit une métrique sur $P_p(X)$.

où $P_p(X)$ est l'ensemble des mesures de probabilité avec des moments d'ordre p finis.

Le chapitre 3 présente les résultats de convergence des solutions de problèmes de localisation dérivés du problème de transport de Monge-Kantorovich.

0.3.3 Processus ponctuels spatiaux et problèmes de localisation-allocation

Cette section traite l'étude de cas relative au positionnement d'une nouvelle caserne de pompiers. L'approche théorique de ce problème est présentée au chapitre 2.

L'étude du problème de positionnement d'une nouvelle caserne de pompiers est réalisée au chapitre 2. La méthodologie utilisée se généralise aisément à tout problème de localisation-allocation. Dans une approche d'optimisation

multicritère où l'on souhaite à la fois minimiser la distance totale d'intervention et les charges de travail des pompiers, nous considérons le problème empirique mono-objectif suivant

$$(s^*, \alpha^*) = \underset{s,\alpha}{\operatorname{argmin}}\, \lambda \sum_{i=1}^{N} \| X_i - s_{\alpha(i)} \|^2 + (1-\lambda)\Phi_N(\alpha, \{(X_1, D_1), \cdots, (X_N, D_N)\}).$$

où s est la position de la nouvelle caserne, α la fonction d'affectation des sinistres aux casernes et Φ_N une fonction mesurant le déséquilibre dans la répartition des charges de travail des casernes (fonction entropie, indice de Gini ou écart maximum entre les marques). Le choix du paramètre de régularisation λ est lié au compromis acceptable pour les pompiers entre détérioration d'un objectif et amélioration de l'autre. La solution à ce problème n'a pas de formule explicite mathématique et nécessite l'utilisation d'algorithmes pour approximer cette solution. De plus, nous souhaitons disposer d'une représentation de la variation de cette solution optimale.

La méthode proposée au chapitre 2, appelée "SPP location-allocation", se décompose selon les étapes suivantes :

- L'ensemble des localisations des sinistres et leurs marques est considéré comme la réalisation d'un processus ponctuel spatial marqué pour laquelle la recherche du modèle sous-jacent a été réalisée au chapitre 1.
- Plusieurs réalisations simulées issues du modèle ajusté sont générées. A cette étape, il est possible de générer des échantillons plus petits que le jeu de données réel pour des questions de performance de l'algorithme d'optimisation employé à l'étape suivante.
- Sur chaque réalisation simulée, le problème de positionnement optimal est résolu par un algorithme d'optimisation.

Trois algorithmes d'optimisations sont introduits et comparés : naïf, génétique et heuristique. La méthode naïve est une méthode d'évaluation de la fonction objective lorsque la position de la nouvelle caserne peut se trouver sur un noeud d'une grille régulière fixé par l'utilisateur. Suivant les techniques de résolution héritées de Cooper (1964), nous avons développé un algorithme heuristique adapté à notre problématique. Enfin, l'algorithme génétique est une méthode évolutionniste intéressante pour résoudre notre problème de grande dimension qui n'a pas de solution explicite.

Cette technique permet d'obtenir un ensemble de positions optimales selon différents scénarios. On peut ainsi juger de la variabilité de cette position optimale, des foyers de localisation éventuels et permettre aux décisionnaires de proposer une position optimale robuste et envisageable en pratique.

0.4 Processus empiriques et M-estimation

Dans la partie II, la convergence forte de solutions optimales de problèmes de positionnement empiriques vers la solution du problème théorique a été obtenue en s'appuyant sur les méthodes de M-estimation. Nous rappelons quelques définitions et résultats importants de la théorie des processus empiriques en M-estimation. Pour une présentation plus approfondie, le lecteur peut se reporter aux monographies de Pollard (1984), van der Vaart et Wellner (1996), van der Vaart (1998) ou van de Geer (2000).

0.4.1 Convergence de M-estimateurs

Pour illustrer la méthode de M-estimation, nous considérons $(X_i)_{i\geq 1}$ une suite de variables aléatoires indépendantes et identiquemment distribuées issues d'une variable aléatoire X à valeurs dans un espace mesurable $(\mathcal{X}, \mathcal{A})$ et de loi P. Nous verrons par la suite que le théorème classique de M-estimation ne suppose pas que les données sont i.i.d. Supposons qu'un paramètre θ_0, réel ou fonctionnel, appartenant à un espace métrique (Θ, d), est rattaché à la loi P. Alors, nous nous intéressons à l'estimation de ce paramètre $\theta_0 \in \Theta$. La méthode de M-estimation consiste à prendre un estimateur $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ qui maximise une fonction objective du type $\theta \in \Theta \mapsto M_n(\theta)$. Sous certaines hypothèses relatives à la convergence de la fonction M_n vers une fonction M, on obtient la convergence de $\hat{\theta}$ vers θ_0 le maximum de la fonction $\theta \mapsto M(\theta)$. La méthode des moments généralisée est une méthode populaire en économétrie où la fonction objective M_n s'écrit comme un moment empirique. Pour présenter cette méthode, nous définissons d'abord la mesure empirique qui met un poids $1/n$ à chaque observation X_i.

Définition 0.6. *La mesure empirique $\mathbb{P}_n$ associée à $X_1, \ldots, X_n$ est la mesure définie par $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, où δ_a désigne la mesure de Dirac au point a.*

Soit $\mathcal{F}$ l'ensemble des fonctions mesurables $f : \mathcal{X} \to \mathbb{R}$ et P-intégrables, alors la mesure empirique permet de définir une application de $\mathcal{F}$ dans $\mathbb{R}$ donnée par $f \mapsto \mathbb{P}_n f := n^{-1} \sum_{i=1}^n f(X_i)$. Dans le cas de la méthode des moments généralisée, la fonction critère s'écrit donc $M_n(\theta) = \mathbb{P}_n m_\theta$, où $m_\theta : \mathcal{X} \to \bar{\mathbb{R}}$

est une fonction connue. Si l'espérance $M(\theta) = Pm_\theta = \int m_\theta dP$ existe, alors la loi des grands nombres indique que $M_n(\theta) \xrightarrow{p.s.} M(\theta)$ quand n tend vers l'infini. Par conséquent, il semble raisonnable d'espérer la convergence de

$$\hat{\theta}_n := \operatorname*{argmax}_{\theta\in\Theta} P_n m_\theta$$

vers

$$\theta_0 := \operatorname*{argmax}_{\theta\in\Theta} Pm_\theta.$$

Cependant, cette convergence fonctionnelle de M_n vers M est trop faible. Nous rappelons ci-dessous un théorème classique de convergence en probabilité de quasi-maximums $\hat{\theta}_n$ vers θ_0 présenté dans van der Vaart (1998, p.45). Ce théorème requiert la convergence uniforme des fonctions M_n vers M. Il est aussi demandé à θ_0 d'être un maximum "bien séparé" de la fonction M. Il faut noter que ce théorème est général et ne suppose rien sur la façon dont les fonctions M_n et M sont définies, en particulier, ce théorème s'applique aussi à des données dépendantes.

Théorème 0.2. *Soient M_n des fonctions aléatoires réelles et soit M une fonction de θ telles que pour tout $\varepsilon > 0$,*

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{Pr.} 0$$
$$\sup_{\theta: d(\theta,\theta_0)\geq\varepsilon} M(\theta) < M(\theta_0).$$

Alors toute suite d'estimateurs $\hat{\theta}_n$ avec $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$ converge en probabilité vers θ_0.

La condition de convergence uniforme de M_n vers M est parfois trop forte en pratique et difficile à vérifier. Cependant, des résultats utiles existent dans le cadre de la méthode des moments généralisée. En effet, lorsque $M_n = \mathbb{P}_n m_\theta$ et $M = Pm_\theta$ alors cette condition s'écrit comme une loi des grands nombres uniforme.

Dans la partie II, nous nous sommes intéressés à la version presque sûre du théorème 0.2. Au chapitre 3, la condition de convergence uniforme a été obtenue grâce au théorème 0.1. En revanche, dans le chapitre 4, nous avons dû étudier des conditions nécessaires pour satisfaire une loi forte des grands nombres uniforme, même si notre fonction objective ne s'écrit pas comme un moment ou une fonction classique en M-estimation.

0.4.2 Loi forte des grands nombres uniforme

On s'intéresse donc à la convergence uniforme des processus $\mathbb{P}_n f$ sur des classes de fonction. Le théorème de Glivenko-Cantelli étend la loi des grands nombres à la convergence uniforme sur la classe de fonctions $\mathcal{F} := \{f : t \in \mathbb{R} \mapsto \mathbb{1}_{]-\infty,t]}\}$. On note la distance uniforme $\|\mathbb{P}_n f - Pf\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf|$. Lorsque $\mathcal{F}$ est défini comme précédemment, $\mathbb{P}_n f$ représente une valeur prise par la fonction de répartion empirique pour tout $f \in \mathcal{F}$ et $\|\mathbb{P}_n f - Pf\|_{\mathcal{F}} = \sup_{t\in\mathbb{R}} |\mathbb{F}_n(t) - F(t)|$ est connue comme étant la statistique de Kolmogorov-Smirnov (avec $\mathbb{F}_n$ et F désignant respectivement les fonctions de répartition empirique et théorique).

Théorème 0.3. (Glivenko-Cantelli). *Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires indépendantes et identiquement distribuées issues d'une même variable aléatoire X de fonction de répartition F alors*

$$\sup_{t\in\mathbb{R}} |\mathbb{F}_n(t) - F(t)| \xrightarrow{p.s.} 0.$$

Nous rappelons maintenant la définition d'une P-classe de Glivenko-Cantelli.

Définition 0.7. *Une classe $\mathcal{F}$ de fonctions $f : \mathcal{X} \to \mathbb{R}$ est appelée une P-classe de Glivenko-Cantelli si elle vérifie*

$$\|\mathbb{P}_n f - Pf\|_{\mathcal{F}} \xrightarrow{p.s.} 0 \quad \textit{quand } n \to \infty.$$

La propriété d'être une P-classe de Glivenko-Cantelli dépend de la "taille" de la classe. Ainsi, un ensemble fini de fonctions intégrables est toujours Glivenko-Cantelli. Une façon de mesurer la taille d'une classe s'exprime en terme d'entropie. Des théorèmes avec des conditions suffisantes basées sur l'entropie avec crochets ("entropy with bracketing") existent, mais le calcul du nombre minimum $N_{[]}(\varepsilon, \mathcal{F}, L_1(P))$ de ε-crochets dans $L_1(P)$ pour recouvrir $\mathcal{F}$ s'avère compliqué à calculer en général. Ces conditions ne sont pas nécessaires, cependant les exemples couverts sont nombreux. Il est possible d'utiliser d'autres conditions suffisantes plus simples basées sur l'entropie définie par le nombre uniforme de recouvrement ("uniform covering numbers"). On appelle nombre de recouvrement $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ le nombre minimal de boules de la forme $\{g \in \mathcal{F} : \|g - f\| < \varepsilon : f \in \mathcal{F}\}$ de rayon ε nécessaires pour recouvrir l'ensemble $\mathcal{F}$. Le théorème qui suit nécessite une condition sur la fonction envelope F définie par $x \mapsto \sup_{f\in\mathcal{F}} |f(x)|$.

Théorème 0.4. (Glivenko-Cantelli). *Soit $\mathcal{F}$ une classe de fonctions mesurables avec $\sup_Q N(\varepsilon\|F\|_{Q,1}, \mathcal{F}, L_1(Q)) < \infty$ pour tout $\varepsilon > 0$. Si $PF < \infty$, alors $\mathcal{F}$ est une P-classe de Glivenko-Cantelli.*

Une famille importante de classes de fonctions où le nombre de recouvrement est borné correspond aux classes de Vapnik-Chervonenkis (appelées VC-classes). Les VC classes se définissent au travers de propriétés combinatoires et comprennent de nombreux exemples.

0.4.3 Théorie de Vapnik-Chervonenkis

Une classe de fonctions $\mathcal{F} := \{\mathbb{1}_D : D \in \mathcal{D}\}$, où $\mathcal{D}$ est une collection d'ensembles dans $\mathcal{X}$, est définie comme étant une VC-classe de fonctions si $\mathcal{D}$ est une VC-classe d'ensembles. Nous rappelons ici la définition d'une VC-classe d'ensembles (Vapnik et Chervonenkis, 1971).

Définition 0.8. *Soit $\mathcal{D}$ une collection d'ensembles dans $\mathcal{X}$.*
Pour $x_1, \ldots, x_n \in \mathcal{X}$, définissons

$$\Delta^{\mathcal{D}}(x_1, \ldots, x_n) = \text{card}\{D \cap \{x_1, \ldots, x_n\} : D \in \mathcal{D}\},$$

tel que $\Delta^{\mathcal{D}}(x_1, \ldots, x_n)$ est le nombre d'ensembles différents de la forme $D \cap \{x_1, \ldots, x_n\}$, $D \in \mathcal{D}$.
On définit $m^{\mathcal{D}}(n) = \sup\{\Delta^{\mathcal{D}}(x_1, \ldots, x_n) : x_1, \ldots, x_n \in \mathcal{X}\}$ et l'indice $V(\mathcal{D})$ de la classe $\mathcal{D}$ par

$$V(\mathcal{D}) = \inf_{n \geq 1}\{m^{\mathcal{D}}(n) < 2^n\}.$$

Une collection d'ensembles $\mathcal{D}$ est une VC-classe si $V(\mathcal{D}) < \infty$.

En d'autre termes, une VC-classe d'ensembles $\mathcal{D}$ est une collection d'ensembles telle que, si le nombre de points n d'une configuration $\{x_1, \ldots, x_n\}$ est suffisamment grand alors on ne peut pas écrire tous les 2^n sous-ensembles comme $\mathcal{D} \cap \{x_1, \cdots, x_n\}$. On généralise la notion de VC-classe à des ensembles de fonctions $\mathcal{F}$ par la définition suivante.

Définition 0.9. *On appelle sous-graphe d'une fonction $f : \mathcal{X} \to \mathbb{R}$ l'ensemble*

$$subgraph(f) := \{(x, y) \in \mathcal{X} \times \mathbb{R} : f(x) < y\}.$$

Une classe de fonctions $\mathcal{F}$ est une VC-classe si

$$\{\text{subgraph(f)} : f \in \mathcal{F}\}$$

est une VC-classe d'ensembles.

Pour une VC-classe de fonctions van der Vaart et Wellner (1996, Théorème 2.6.7, p.141) démontre que $\sup_Q N(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q))$ est borné pour tout $r \geq 1$ et $0 < \varepsilon < 1$. Par conséquent, la condition suffisante est vérifiée

dans le Théorème 0.4 et si $PF < \infty$ alors $\mathcal{F}$ est P-Glivenko-Cantelli. Ainsi, toute VC-classe de fonctions est P-Glivenko-Cantelli, mais une collection de fonctions P-Glivenko-Cantelli n'est pas nécessairement une VC-classe de fonctions.

La propriété de VC-classe est préservée par de multiples opérations ensemblistes comme l'intersection, l'union, le complémentaire, le passage à l'enveloppe convexe,...Les espaces vectoriels de dimension finie sont aussi des VC-classes. Les classes définies par des fonctions s'écrivant comme la somme ou le produit de fonctions appartenant respectivement à des VC-classes sont aussi des VC-classes. Le lecteur peut se référer aux monographies de van der Vaart et Wellner (1996), van der Vaart (1998) ou van de Geer (2000) pour d'autres exemples de VC-classes. Certaines propriétés de stabilité de la notion de VC-classe se retrouvent aussi lorsque l'on étudie le fait d'être une classe de Donsker (ou une classe de Glivenko-Cantelli).

La monographie de Peskir (2000) traite de façon complète les approches utilisées pour montrer la loi forte des grands nombres uniforme. Le théorème dû à Vapnik et Chervonenkis (1981) présentant une condition nécessaire et suffisante pour satisfaire la loi forte des grands nombres uniforme y est présenté (Theorem 3.11, p.82) dans le cadre de données indépendantes et identiquement distribuées. On définit le nombre aléatoire de recouvrement $N_n(\varepsilon, \mathcal{F})$ associé à $(X_i)_{i\geq 1}$ comme étant le plus petit nombre de boules de rayon $\varepsilon > 0$ (pour la métrique sup de $\mathbb{R}^n$) nécessaire pour recouvrir l'ensemble $\mathcal{F}_n := \{f(X_1), \ldots, f(X_n) | f \in \mathcal{F}\}$ avec $n \geq 1$.

Théorème 0.5. (Vapnik-Chervonenkis, 1981). *Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires indépendantes et identiquement distribuées. Alors la loi forte des grands nombres uniforme est valide :*

$$\sup_{f\in\mathcal{F}} |n^{-1} \sum_{i=1}^{n} f(X_i) - Pf| \xrightarrow{p.s.} 0 \text{ quand } n \to \infty,$$

si et seulement si la condition suivante est satisfaite :

$$\lim_{n\to\infty} \frac{\mathbb{E}(\log N_n(\varepsilon, \mathcal{F}))}{n} = 0$$

pour tout $\varepsilon > 0$.

Cette condition est satisfaite pour des VC-classes de fonctions. Peskir (2000) présente une généralisation de ce théorème au cas de séries stationnaires β-mélangeantes. Des résultats similaires existent pour des structures de dépendance différentes, notamment pour des données α-mélangeantes (Yu, 1994).

0.4.4 Convergence des solutions aux problèmes de localisation-allocation empiriques

Problèmes de positionnement dérivés du problème de transport de Monge-Kantorovich

Dans le chapitre 3, nous étudions le comportement asymptotique des solutions d'un problème de localisation-allocation empirique dérivé du problème probabiliste de transport de Monge-Kantorovich. Nous considérons le problème du positionnement d'un nouvel établissement conditionnellement à des établissements existants, par exemple des magasins, en fonction des positions des clients. Le cadre asymptotique correspond au cas où la taille de l'échantillon des positions des clients, supposées identiquement distribuées, dépendantes ou non, tend vers l'infini. Nous nous intéressons au transfert optimal de la mesure source μ de la population des clients vers la mesure cible ν des établissements. Cependant, le support de la mesure cible dépend de la position du nouvel établissement y_{J+1} c'est pourquoi elle est notée $\nu(y_{J+1})$. Dans ce cadre, la politique d'affectation est spécifiée par le choix d'une distribution jointe Q dans la classe $\mathcal{P}^{\mu,\nu(y_{J+1})}$ des lois sur $U \times U$ de marginales μ et $\nu(y_{J+1})$. Si $c(x,y)$ est une fonction continue positive sur $U \times U$, interprétée comme le coût de transport du client positionnée en x vers l'établissement localisé en y, alors le coût total minimal de l'allocation de la population de clients vers l'ensemble des établissements est donné par

$$\mathcal{A}_c(\mu, \nu(y_{J+1})) = \inf_{Q \in \mathcal{P}^{\mu,\nu(y_{J+1})}} \int_{U \times U} c(x,y) Q(dx, dy) \tag{0.2}$$

qui existe sous certaines conditions sur μ et c (Gangbo et McCann, 1996). La position optimale théorique y^*_{J+1} du nouvel établissement est choisie telle que la fonctionnelle (0.2) soit minimale parmi l'ensemble des positions possibles y_{J+1} de U pour l'implantation du nouvel établissement. La solution au problème théorique s'écrit alors

$$y^*_{J+1} = \operatorname*{argmin}_{y_{J+1} \in U} \mathcal{A}_c(\mu, \nu(y_{J+1}))$$

et la solution au problème empirique

$$\hat{y}^*_{J+1} = \operatorname*{argmin}_{y_{J+1} \in U} \mathcal{A}_c(\mu_n, \nu(y_{J+1}))$$

Remplacer μ par μ_n dans $\mathcal{A}_c(\mu, \nu(y_{J+1}))$ donne la version discrète du problème de Monge-Kantorovich. Sous la condition classique en M-estimation de "minimum bien séparé" de la fonction $y_{J+1} \mapsto \mathcal{A}_c(\mu, \nu(y_{J+1}))$ qui s'écrit

$$\inf\{\mathcal{A}_c(\mu, \nu(y_{J+1})) : \ y_{J+1} \in U, \ d(y_{J+1}, y^*_{J+1}) > \varepsilon\} > \mathcal{A}_c(\mu, \nu(y^*_{J+1})),$$

pour tout $\varepsilon > 0$, et pour une fonction de coût c dans la classe des fonctions $c(\cdot,\cdot) = d(\cdot,\cdot)^p$ telle que $p > 0$ et $\int_U c(x,a)d\mu(x) < \infty$ pour tout point $a \in U$, nous démontrons que $d(\hat{y}^*_{J+1}, y^*_{J+1}) \xrightarrow{p.s.} 0$ quand $n \to \infty$.
Nous avons aussi obtenu la convergence en probabilité pour des quasi-minimums $\hat{y}^*_{J+1}$, c'est à dire vérifiant

$$\mathcal{A}_p(\mu_n, \nu(\hat{y}^*_{J+1})) \leq \inf_{y_{J+1}\in U} \mathcal{A}_p(\mu_n, \nu(y_{J+1})) + o_P(1) \tag{0.3}$$

avec $\mathcal{A}_p(\cdot,\cdot) = [\mathcal{A}_c(\cdot,\cdot)]^{\min(1,1/p)}$ étant la distance de Wasserstein.

Une étude numérique du résultat de convergence précédent est réalisée sous R en utilisant l'algorithme semiLAPJV d'assignement de Volgenant (1996) sur des échantillons simulés.

Problèmes de positionnement avec contraintes

Dans le chapitre 4, on étudie la convergence asymptotique des solutions de problèmes de positionnement optimal empiriques vers la solution du problème théorique dans le cadre de la localisation d'une nouvelle caserne de pompiers. Les problèmes d'optimisation théorique et empirique s'expriment alors de la manière suivante

$$y^*_{J+1} = \operatorname*{argmin}_{y_{J+1}\in\Omega} \inf_{\alpha\in\Lambda} M(y_{J+1}, \alpha) \tag{0.4}$$

$$\hat{y}_{J+1,n} = \operatorname*{argmin}_{y_{J+1}\in\Omega} \inf_{\alpha\in\Lambda} M_n(y_{J+1}, \alpha), \tag{0.5}$$

où, dans le cas de données i.i.d., $M(y_{J+1},\alpha) := \mathbb{E}(\|X - y_{\alpha(X)}\|^2) + \lambda\Phi(\alpha)$ et $M_n(y_{J+1},\alpha) := \frac{1}{n}\sum_{i=1}^n \|X_i - y_{\alpha(X_i)}\|^2 + \lambda\Phi_n(\alpha)$. Les fonctions Φ_n et Φ, définies au chapitre 2, sont des fonctions empirique et théorique mesurant les inégalités de répartition des charges de travail. La fonction objective équilibre ainsi le coût de transport et les charges de travail. Le paramètre de régularisation λ dont le choix est discuté au chapitre 2 est supposé fixé.
Notre approche est différente des procédures de M-estimation standard utilisées en économétrie car la fonction empirique $\inf_{\alpha\in\Lambda} M_n(y_{J+1},\alpha)$ à minimiser ne s'écrit ni sous la forme d'une moyenne empirique, ni sous la forme d'une fonction dépendant d'un paramètre fonctionnel estimé $M_n(y_{J+1},\hat{\alpha})$, comme dans Chen *et al.* (2008) en Z-estimation.
Hors cadre des processus ponctuels, lorsque les couples positions et marques $(X_1, M_1), \ldots, (X_n, M_n)$ sont supposés indépendants et issus d'une même variable aléatoire (X, M) alors la convergence presque sûre de la solution empirique vers la solution théorique est démontrée sous des conditions naturelles en pratique.

(H1) $\inf_{k\in\{1,\dots,J+1\}} \inf_{\alpha\in\Lambda} \mathbb{E}[M \mathbb{1}_{\alpha^{-1}(k)}(X)] > 0.$

(H2) $\inf_{y:d(y,y^*_{J+1})\geq\varepsilon} \inf_{\alpha\in\Lambda} M(y,\alpha) > \inf_{\alpha\in\Lambda} M(y^*_{J+1},\alpha)$, pour tout $\varepsilon > 0$.

(H3) $\{\alpha^{-1}(j); \alpha \in \Lambda\}$ est une classe d'ensembles de Vapnik-Chervonenkis, pour tout $j = 1, \cdots, J+1$.

Une généralisation de la théorie de Vapnik-Chervonenkis à des séries $(X_i)_{i\geq 1}$ β-mélangeantes est présentée dans Peskir (2000). Compte tenu de ces résultats de loi forte des grands nombres uniformes, les perspectives d'extension de notre théorème de convergence au cas dépendant existent.

0.5 Indices de concentration et processus ponctuels marqués

L'étude de la localisation d'activités économiques s'appuie sur la mesure de l'inégalité de leurs répartitions dans l'espace et surtout sur leurs concentrations en certains points. De nombreux indices de concentration ont été introduits en économétrie pour mesurer des inégalités géographiques à partir de données agrégées (masse salariale, taux d'emploi,...) dans des zones fixées. Nous en présentons quelques exemples au chapitre 5.

Cependant, ce découpage en zones présente plusieurs désavantages, notamment celui d'être biaisé par rapport au choix de l'échelle géographique. Cet inconvénient majeur est à l'origine de nouveaux indices de concentration basés sur les distances comme les indices de Duranton et Overman (2005) et Marcon et Puech (2007). Malheureusement, il n'apparait pas rigoureusement quelles sont les quantités théoriques estimées par ces indices de concentration. C'est pour cette raison que nous avons cherché à exprimer ces indices sous la forme d'estimateurs de caractéristiques de processus ponctuels marqués. Nous rappelons au chapitre 5 les notations et définitions nécessaires pour présenter des caractéristiques du second ordre d'un processus ponctuel spatial marqué. Ensuite, nous définissons l'indice I_{DO} de Duranton et Overman (2005) et l'indice I_{MP} de Marcon et Puech (2007), et les écrivons en fonction d'estimateurs de caractéristiques de processus ponctuels marqués.

Compte tenu de l'écriture des indices I_{DO} et I_{MP}, nous introduisons un nouvel indice de concentration basé sur la définition d'une nouvelle caractéristique générale du second ordre d'un processus ponctuel marqué. Le premier

avantage de notre indice de concentration est d'être présenté comme l'estimateur d'une quantité théorique du processus ponctuel sous jacent. Le second avantage est de pouvoir considérer des processus ponctuels des positions inhomogènes, à l'inverse des indices I_{DO} et I_{MP}.

Dans la dernière partie du chapitre 5, nous définissons un cadre asymptotique avec domaine d'observation borné, dans lequel nous étudions le comportement asymptotique de notre indice de concentration.

Première partie

Positionnement optimal par modélisation de processus ponctuels marqués

Chapitre 1

Analyse exploratoire et modélisation de sinistres par processus ponctuels spatiaux marqués

Sommaire

1.1 Introduction **32**

1.2 Background on summary statistics **35**

1.3 Time variation and spatial variation **38**

1.4 Analysis of the workload mark **41**

1.5 Model **45**

Two density estimates 48

Models based on density estimate $\hat{C}_1$ 49

Models based on density estimate $\hat{C}_2$ 52

1.6 Conclusions **53**

Abstract :

We examine a database of firemen emergencies in the surroundings of the city of Toulouse during the year 2004, using methods of statistical analysis for spatial point patterns. Firemen emergencies are characterized by their positions and different features (time, duration, type, ...) that one can model as a spatio-temporal marked point process. For our study, we consider the following characteristics for firemen emergencies : positions, time of occurences and marks which take into account the duration and the number of

firemen involved. We use graphical methods to explore the structure of the underlying spatial point process with a final objective of choosing a suitable model for future work. We first review the basic concepts and methods used in the paper. Considering the marginal spatio-temporal point pattern, we propose to evaluate the importance of the variation of intensity over time in comparison with spatial variation and to test the dependence between positions and time. Afterwards, we conduct an exploratory analysis of the marks to test their dependence with positions as well as their dependence with time. Our resulting framework of independence allows us to explore the dependence between categories in order to test the random labeling hypothesis. Then, under the hypothesis of random labeling and invariance in time which have been established in the first two parts, we fit a spatial point process model to the unmarked spatial point pattern aggregated over the whole year. Finally, we analyze the goodness-of-fit of our models by exploring the first and second order characteristics of simulations from the fitted model. Throughout this article, the exploratory analysis is made using mainly the R package **spatstat**.

1.1 Introduction

The database of firemen emergencies, provided to us by the fire department SDIS 31, contains the locations and characteristics (time, duration, number of firemen, ...) of emergencies which have required an intervention of firemen in the surroundings of the city of Toulouse, the largest town of the Midi-Pyrénées region in France, during the year 2004. Examples of emergencies include fires but also car accidents, assistance to injured people, ... After removing outliers and emergencies with missing values, we have the locations of 20820 emergencies with 5433 distinct points in an area of 620 km^2. The important number of duplications for this spatial point pattern is caused by a positional error. The location of emergencies has not been recorded exactly but approximated by a nearby location which can be the centroid of the street for example. No information is available for this positional error even if we can think that it is closely linked with the level of urbanization.

This problem arises quite frequently in practice, for instance in econometrics or epidemiology (Benes *et al.*, 2005). For each emergency, we have in this dataset the location in the Lambert II extended coordinate system, the time of occurence (in seconds since 1970) and the corresponding month. The mark we consider is the product of the duration of emergency by the number of firemen allocated to each emergency (number of man-hours) which thus re-

presents a measure of total workload. Therefore, for us, an emergency with a low duration time and a high number of firemen will be considered as important as an emergency with a high duration time and a low number of firemen.

The exploratory analysis of this dataset is a preliminary step in the study of the following problem : find the optimal position of a new fire station in this area. The aim of the present paper is to analyze the point pattern in order to guide the choice of a well-founded spatio-temporal marked point process model to be used in Bonneu & Thomas-Agnan (2008).

The exploratory analysis for spatial point patterns often uses nonparametric estimates of various summary statistics based on first and second order properties of point processes. First of all, these characteristics are useful to test the hypothesis of Complete Spatial Randomness (CSR), which consists in determining whether the point pattern derives from a homogeneous Poisson process. Indeed, Poisson point processes model the absence of interaction between points. The intensity function is an important first order property which can be interpreted for homogeneous point processes as the mean number of points per unit area. Various functional summary statistics measure aggregation/clustering or regularity at distances less than different thresholds. In particular, we will introduce the L function derived from the so-called K function introduced by Ripley (1976) for stationary processes and extended to a more general class by Baddeley *et al.* (2000).

The generalization of these statistics to spatio-temporal marked point processes is theoretically feasible but is not yet implemented in software due to the large dimensions which does not allow straightforward graphics. However, for spatial point processes with categorical marks (multitype point processes), there is a generalization of these statistics which allows one to judge whether the point patterns corresponding to the different categories are generated by the same point process model (Stoyan & Stoyan, 1994 and Schlather, 2001). The hypothesis of Complete Spatiotemporal Randomness (CSTR), which corresponds to a spatio-temporal point process where there is an absence of structure in time as well as in space, can be tested by generalizing the summary statistics to the temporal case (Cressie, 1993). In practice one often ignores the variation in time and the dependence between marks and positions in order to analyze the point pattern aggregated over time and to separately fit a spatial point process model for positions.

We suggest two different methods illustrated by graphics to evaluate the importance of the variation in time of the intensity in comparison with the

variation in space. The first one consists in computing the intensity function for the point patterns associated with each month, for example, and comparing the graphs of their estimates. In the second one, we introduce estimates of a measure of the variation in time and variation in space and we compute the resulting ratio. This ratio enables us to understand if the temporal variation can be viewed as negligible compared to the spatial variation. For testing the dependence between positions and time, we present the results of separability tests of the marginal spatio-temporal point process introduced in Schoenberg (2004) which involve a comparison of the intensity and the product of marginal conditional intensities.

We then test the hypothesis of random labeling, i.e. whether the marks are i.i.d. and independent on the positions and time. The dependence between marks and positions can be explained by several aspects : intrinsic heterogeneity of the domain space, concurrence effects, etc. (Schlather *et al.*, 2004). We can use geostatistical methods to test this dependence if the hypothesis that the point pattern is a realization of a stationary and isotropic spatial point process is reasonable. In our case, we have a high heterogeneity in the population density. Consequently, we discretize the marks into different categories and graphically compare the intensity estimates. We also use the method in Schoenberg (2004) for testing the dependence between marks and positions, and settle with the same separability tests the matter of the dependence between marks and time. Our framework of independence between marks and positions, and also between marks and time, allows us to test the dependence between the workload categories by computing a function denoted by L_{cross}. The absence of correlation between the workload categories suggests the random labeling of the marks. Thid leads us to the search for an adequate model for the marginal distribution of the marks.

The results thus obtained suggest that it is reasonable to aggregate the spatial point pattern over time. But, because of the high number of emergencies and duplicated locations, the modelling of the whole point pattern is very difficult. Consequently, we choose to analyze the emergencies of a particular month, for example, June. In this analysis, the major difficulty in choosing a suitable model consists primarily in adjusting the intensity function as well as possible. We present three methods of estimation of the intensity : parametric, nonparametric and semiparametric. For each different estimate of the background intensity, we test the absence of interaction for this point pattern by plotting an estimate of the L summary statistic and the pointwise envelope from simulations of an inhomogeneous Poisson process. Finally, we choose a fitted model which presents approximately the same first and second

order properties as those of the spatial point pattern.
In this paper, we mainly use the R package **spatstat** for analyzing the spatial point process (Baddeley & Turner, 2005 and Baddeley & Turner, 2006).

1.2 Background on summary statistics

By definition, a spatial point process X is a random countable subset of a space S. As in our example, we focus on point processes X whose realizations are finite subsets of a compact set $W \subset S$. A spatio-temporal marked point process $Y = \{(\mathbf{x}, m_{\mathbf{x}}, t_{\mathbf{x}}) : \mathbf{x} \in X\}$ with points $\mathbf{x} \in S$, marks $m_{\mathbf{x}} \in M$ and times $t_{\mathbf{x}} \in T$ is defined to be a spatial point process on the product space $S \times M \times T$. In the sequel, the definitions are given for a spatial point process X in $W \subset \mathbb{R}^2$ but can be generalized to higher dimensions. For convenience, we number the points of a realization $\mathbf{x} = \{x_1, \cdots, x_n\}$ even if we must keep in mind that a point pattern is unordered. For a spatial point pattern with duplicated points, we often plot the distinct points with their number of duplications. However, due to the high proportion of duplicated points in our case, we choose to plot perturbed locations for a better readability. For the perturbation of locations, we use Gaussian noise with zero mean and standard deviation equal to 50 in each coordinate. Our choice for the standard deviation follows from the empirical distribution study of the inter-events distances. Figure 1.1 (Left) plots the perturbed locations of 2007 emergencies in June, suggesting a high inhomogeneity in the distribution of emergencies due to the density of population. In the sequel, we always consider the perturbed locations of emergencies obtained from the same Gaussian noise. This choice is justified later by the difficulty in using methods based on the K function for a point pattern with a high number of duplicated points (envelope, minimum contrast estimation, ...).

Estimation of the intensity λ

The process first-order characteristic is its intensity function λ defined as

$$\lambda(s) = \lim_{d\delta \to 0} \frac{\mathbb{E}\left[N(d\delta)\right]}{d\delta}$$

where $d\delta$ is the elementary area around s and $N(d\delta)$ the number of events in this area.
If λ is constant, then X is said to be homogeneous with intensity λ, otherwise it is inhomogeneous. The estimation of λ is the first step in any exploratory analysis of a point pattern and aims to evaluate the homogeneity of the

process. In our case, it is inappropriate to assume homogeneity because of the spatial correlation of emergencies with the human settlement pattern which is not stationary. Due to the presence of inhomogeneity, we use a kernel method to estimate the intensity function. Our absence of information about the positional error of emergencies does not allow us to use the new kernel estimators introduced in Cucala (2008). Consequently, we choose to estimate the intensity function on the locations perturbed by the Gaussian noise defined before. The chosen estimate is presented in its anisotropic form with a border effects correction (Diggle, 1985) :

$$\hat{\lambda}(s) = \frac{\sum_{i=1}^{n} K_H(s - x_i)}{\hat{c}_{W,H}(s)}$$

where K_H is the kernel with covariance matrix H defined by $K_H(s) = |H|^{-1}k_2(H^{-\frac{1}{2}}s)$, k_2 is the density function of a standard bi-dimensional Gaussian variable and $\hat{c}_{W,H}(s)$ is an estimate of the edge correction factor $c_{W,H}(s) = \int_W K_H(s-u)du$. In its isotropic form K_h is the kernel with standard deviation h defined by $K_h(s) = h^{-2}k(h^{-1}s)$.
The choice of a good bandwidth h is difficult in practice, notably with very wide variations in λ as mentioned in Diggle *et al.* (2007). Indeed, the method proposed in Berman *et al.* (1989) of minimizing an estimation of the mean square error of $\hat{\lambda}$ produces in our case a value of h close to zero. For selecting an optimal diagonal matrix H we can use a plug-in method implemented in the R package **ks** with binned pilot estimation (Wand & Jones, 1994). The diagonal terms of the bandwidth matrix obtained are sufficiently small to capture changes in population density between urban and rural environments and to avoid problems of under-smoothing. For the point pattern of emergencies in June, the smoothing parameter is between 700 and 900 meters for both coordinates. We subsequently use the isotropic form with bandwidth $h = 800$. Figure 1.1 (Right) represents the logarithm transformation of the intensity estimate of emergencies in June. This transformation achieves an enhancement of variations of intensity around cities smaller than Toulouse.

Estimation of the K-function

To study the spatial dependence over a wide range of scales, we can use summary statistics based on a number of known second order properties. Here, we only consider the L function derived from the K function introduced by Ripley (1976) for stationary processes and extended to the class of second order intensity-reweighted stationary processes by Baddeley *et al.* (2000). The theoretical K function for a stationary spatial point process is

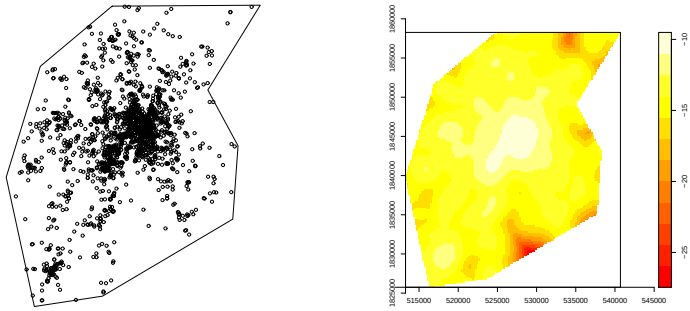

FIGURE 1.1 – Left : Perturbed locations of the 2007 emergencies in June. Right : Logarithm transformation of the intensity of emergencies in June.

the expectation of the number of extra events within distance $r \geq 0$ of a randomly chosen event, divided by the intensity λ. The L function is defined by $L(r) = \sqrt{K(r)/\pi}$ for all $r \geq 0$. At least for small values of r, $L(r) - r > 0$ indicates aggregation/clustering at distances less than r, and $L(r) - r < 0$ indicates regularity. More precisely, because K is a cumulative function, a significant peak of L above 0 shows the maximum range of aggregation and should be interpreted with care beyond this point. For second order intensity-reweighted stationary point processes X, we use the following estimate of the inhomogeneous K function introduced in Baddeley *et al.* (2000) :

$$\hat{K}_{inhom}(r) = \frac{1}{|W|} \sum_{i=1}^{n} \sum_{j \neq i} \frac{\hat{w}_{x_i,x_j,r} \mathbb{1}(\|x_i - x_j\| \leq r)}{\hat{\lambda}(x_i)\hat{\lambda}(x_j)}, \quad r \geq 0$$

where $\hat{w}_{x_i,x_j,r}$ is a boundary correction factor and $|W|$ the area of W. The more common boundary correction factor is the translation correction factor $w_{x_i,x_j,r} = |W \cap W_{x_i - x_j}|^{-1}$, where $W_{x_i - x_j} = \{\xi + x_i - x_j : \xi \in W\}$, but is computationally expensive for large point patterns. So, in our case, we prefer the border correction factor implemented in **spatstat**,

$$\hat{w}_{x_i,x_j,r} = \frac{\mathbb{1}(d(x_i, \partial W) > r)}{\sum_{k=1}^{n} \left(\mathbb{1}(d(x_k, \partial W) > r)/\hat{\lambda}(x_k) \right)}$$

where ∂W is the boundary of the observation window.
Afterwards, in order to study the correlation structure in multivariate point processes such as $X = (Y, Z)$, cross summary statistics $K^{cross}_{inhom}(Y, Z)$ have been introduced for the non-stationary case in the same way as the K function. The definition of $K^{cross}_{inhom}(Y, Z)$ concerns cross second order intensity reweighted stationary processes ; an estimate is given by

$$\hat{K}^{cross}_{inhom}(Y,Z)(r) = \frac{1}{|W|}\sum_{i=1}^{n_{\mathbf{y}}}\sum_{j=1}^{n_{\mathbf{z}}}\frac{\hat{w}_{y_i,z_j,r}\mathbb{1}(\|y_i - z_j\| \leq r)}{\hat{\lambda}_Y(y_i)\hat{\lambda}_Z(z_j)}, \quad r \geq 0 \qquad (1.1)$$

The L^{cross}_{inhom} function is the extension to the multivariate case of the L_{inhom} function for the univariate case. For convenience and where there is no possible confusion, we use the notations K, L and L_{cross} respectively for K_{inhom}, L_{inhom} and L^{cross}_{inhom}.

Envelope

In general, let us consider a statistic $L(r)$ and a given hypothesis H_0. Typically, the null hypothesis can be the absence of interaction, the random labelling hypothesis or the goodness-of-fit of a given model. Critical intervals are necessary to judge the deviance from the null hypothesis of a nonparametric estimate of a summary statistic. Let $\hat{L}(r)$ be the estimate computed from the observed point process X in W, and $\hat{L}_1(r), \cdots, \hat{L}_m(r)$ those obtained from i.i.d. simulations $X_1, \cdots, X_m$ under H_0. For each value of r, we can estimate any quantile for the distribution of $\hat{L}(r)$ under H_0 from the empirical distribution of $\hat{L}_1(r), \cdots, \hat{L}_m(r)$, if m is large enough. The quantiles $L_l(r)$ and $L_u(r)$ used to construct the critical interval are called respectively the lower and the upper envelope. We obtain a pointwise envelope because we have a critical interval for each value of r. Throughout this paper, the envelope is computed from 39 simulations with pointwise minima and maxima in order to have for each r a 5% probability that the estimate of $L(r)$ falls outside the interval.

1.3 Time variation and spatial variation

In this section, we focus on comparing the relative importance of time variation and spatial variation in the intensity of the spatio-temporal point process. In practical situations, unless the importance of time is well-known and predominant (earthquakes,...), the dependence on time is often ignored by aggregating the spatial point process over time. Even so, there does exist some literature on spatio-temporal point processes models (Diggle (2006)). We would like to make sure this aggregation is justified in our case. Because it will be impracticable to simultaneously model space, time and marks, we choose to ignore here the possible dependence between marks and time. We thus perform this investigation by aggregating the marked process into a

single unmarked one. Diggle *et al.* (2005) propose a Monte Carlo test to investigate temporal changes. We propose to decompose the spatio-temporal process into 12 monthly realizations. A first approach is to compare the logarithm transformation of estimates of the intensities for each month (Figure 1.2). At first sight, the estimates of intensities do not show important modifications in shape and seem to present a temporal trend with a low rate of change. Indeed, we note a slight trend in the total number of emergencies through the year.

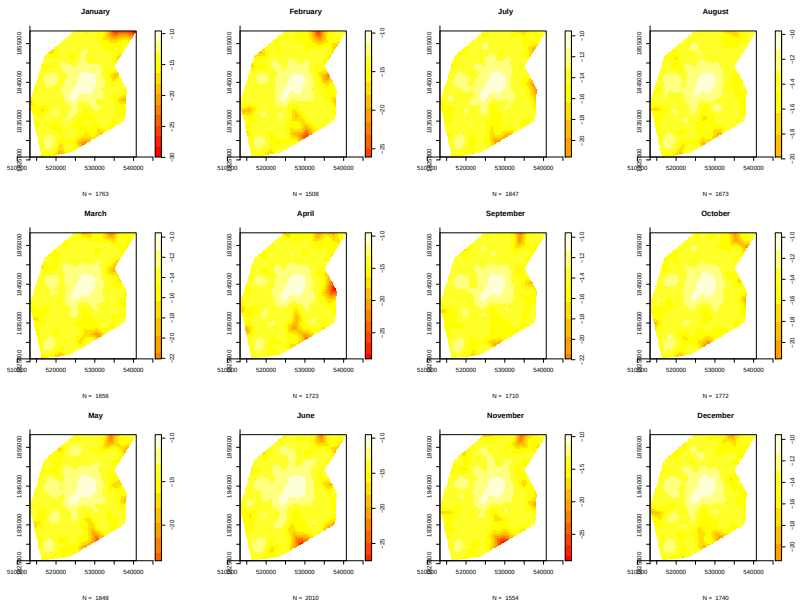

FIGURE 1.2 – Logarithm transformation of intensity by month.

A second approach consists in computing the ratio between an estimate of the time variation and an estimate of the spatial variation of intensity. To measure the time variation at a location s, we introduce

$$TMSE(s) = \frac{1}{12}\sum_{k=1}^{12}(\hat{\lambda}_k(s) - \bar{\lambda}(s))^2$$

where $\hat{\lambda}_k$ is the intensity estimate of the month k and $\bar{\lambda} = \frac{1}{12}\sum_{k=1}^{12}\hat{\lambda}_k$ is the mean intensity. This measure is computed on a regular grid of m points $\mathbf{s} = \{s_1, \cdots, s_m\}$ in the domain space. We denote $TMSE$ the pixel image giving the value at each point of the grid.
To measure the spatial variation, we introduce

$$SMSE = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{\lambda}(s_i) - \frac{n}{|W|}\right)^2$$

The image ratio $TMSE/SMSE$ indicates that the time variation is negligible in comparison with the spatial variation. Indeed, the time variation represents 0.5 percent of the spatial variation at most. Figure 1.3 shows the logarithm of this ratio which reflects well the high spatial inhomogeneity in this domain.

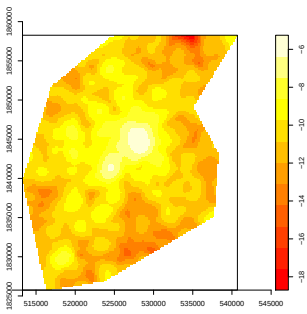

FIGURE 1.3 – Logarithm of the image ratio between the time variation and the spatial variation of intensity.

Finally, we investigate the separability of the intensity function of the spatio-temporal point process (X, D) as in Schoenberg (2004), i.e. we test whether we have

$$\lambda_{X,D}(s,t) = \lambda_X(s) f_D(t), \quad s \in W \text{ and } t \in T.$$

where $\lambda_{X,D}$ and λ_X are respectively the intensities functions of (X, D) and X, and f_D the density function of D. We denote by $\hat{\lambda}$ the estimate of $\lambda_{X,D}$ and by $\tilde{\lambda}$ that of $\lambda_X f_D$. We want to judge the difference between the two. The fact that $s \in W \subset \mathbb{R}^2$ and $t \in T \subset \mathbb{R}$ do not allow us to present straigthforward graphics as in section 1.4 where we discuss dependence between marks and time. To compare the two estimates, we compute four statistics defined in the Schoenberg's article on a regular grid of m points $(\mathbf{s},\mathbf{t}) = \{(s_i, t_j) \in W \times T : i = 1, \cdots, m_{\mathbf{s}}; j = 1, \cdots, m_{\mathbf{t}}\}$.

$$
\begin{aligned}
S_1 &= \sup_{i,j}\{|\hat{\lambda}(s_i,t_j) - \tilde{\lambda}(s_i,t_j)|/\sqrt{\tilde{\lambda}(s_i,t_j)}; (s_i,t_j) \in (\mathbf{s},\mathbf{t})\} \\
S_2 &= \inf_{i,j}\{|\hat{\lambda}(s_i,t_j) - \tilde{\lambda}(s_i,t_j)|/\sqrt{\tilde{\lambda}(s_i,t_j)}; (s_i,t_j) \in (\mathbf{s},\mathbf{t})\} \\
S_5 &= \frac{1}{m} \sum_{(s_i,t_j)\in(\mathbf{s},\mathbf{t})} (\hat{\lambda}(s_i,t_j) - \tilde{\lambda}(s_i,t_j))^2 \\
S_6 &= \sup_{i,j}\{(\hat{\lambda}(s_i,t_j) - \tilde{\lambda}(s_i,t_j))^2; (s_i,t_j) \in (\mathbf{s},\mathbf{t})\}
\end{aligned}
$$

Abnormally large value of these tests statistic indicate a departure from the separability hypothesis. The intensities and the probability density are computed with the *kde* function in the R package **ks**, initially programmed for density estimation. This function allows computing of the density/intensity for three dimensional point pattern. So, in order to obtain intensity estimates, we multiply the result by the number of points. These estimates are not adjusted by a correction factor for border effects. To judge the significance of these statistics, we construct one-sided Monte-Carlo tests from 19 simulations of a Poisson point process under the null hypothesis of separability. If the statistic test S is lower than the maximum value obtained from the simulations then we accept the separability assumption at level 5%. For computational reasons, we limit our study to a subsample of 2000 emergencies randomly chosen. Here, the four tests conclude to the separability assumption. Note that this Monte-Carlo inference is based on simulations from the Poisson model, an assumption that we will discuss later in the paper.
The different approaches show that the variation in time is negligible in comparison with the variation in space and that there is independence between positions and time. We therefore consider that we can aggregate the point pattern over time without loosing important information.

1.4 Analysis of the workload mark

Dependence between marks and positions

Marks and positions are often assumed to be independent but this may not hold in practice. For instance, in forestry, the diameters of trees can be dependent on the nature of the soil and of the presence of others trees nearby. In the case of firemen emergencies, it is possible that the frequency of large workloads emergencies is higher in some areas. Another type of dependence arises from the fact that an occurence may have an influence on the marks of

future emergencies around it. The first type of dependence seems more likely here.
Summary statistics for marked point processes are introduced in Stoyan & Stoyan (1994) and Schlather (2001) to test the dependence between continuous marks and positions. However, these statistics are just defined for stationary and isotropic marked point processses. In these articles, the marks process is modeled as a random field and the authors can apply geostatistical methods. In Schlather (2004), the test of dependence is valid for any random field model where the marks are given by a strictly monotone transformation of a Gaussian random field. This last assumption on the marks is not necessary for the test based on the conditional expectation of marks developed in Guan (2006). Guan's method allows the treatement of examples with bimodal distribution of marks. However, to our knowledge, tests of dependence between continuous marks and positions in the case of inhomogeneous point processes are not available.

We next use two empirical approaches to test the validity of the independence. We first use the same method developed for testing the temporal trend. We discretize the logarithm of workloads, to mitigate the influence of outliers, into three categories : Low, Medium, and High. This discretization is performed by applying the k-means method minimizing the within category variance. Figure 1.4 represents the logarithm transformation of the estimated intensity for the different categories of the multitype point patterns. The patterns of estimates are close together across categories but different in total mass. This suggests that the point patterns could be generated by the same point process model with a different expectation of the number of points.

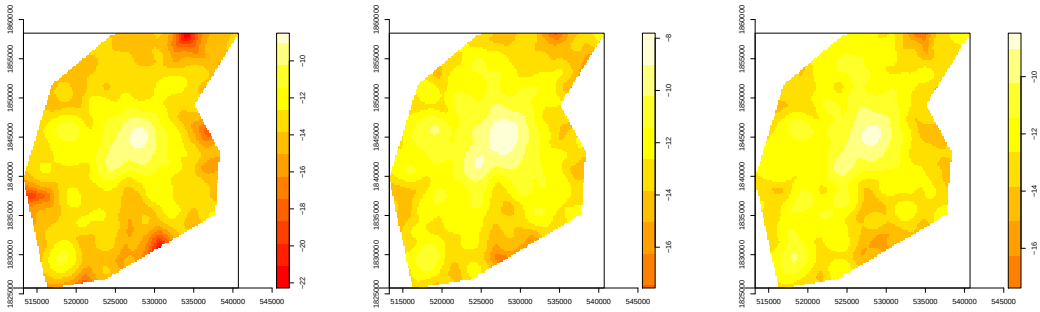

FIGURE 1.4 – Logarithm transformation of intensity by category.Left : Small workloads. Middle : Medium workloads. Right : High workloads.

To confirm the conclusion of independence between marks and positions given by the previous approach we now investigate the Schoenberg's method.

The statistic tests based on S_1 and S_2 accept the separability assumption whereas those based on S_5 and S_6 reject this hypothesis. This situation is not clear-cut and allows us to consider one of the two cases. However, taking into account this dependence implies looking for a more complicated model. One can find some reasons to believe in dependence between marks and positions for some categories of emergencies (fires, car accidents, ...) but we think that this dependence is not very important when considering all types of emergencies simultaneously.

Finally, taking into account the different methods used, the hypothesis of independence between the occurences of emergencies and the workload marks is not as clearly established as in the case of time and location. Nevertheless, we maintain this hypothesis of independence in order to avoid an intractable model.

Dependence between marks and time

We investigate the dependence between marks and time through the separability method. In this case, the bidimensional framework allows to present straightforward plots of the estimates of $\hat{\lambda}$ and $\tilde{\lambda}$ for the marginal marked temporal process (Figure 1.5).

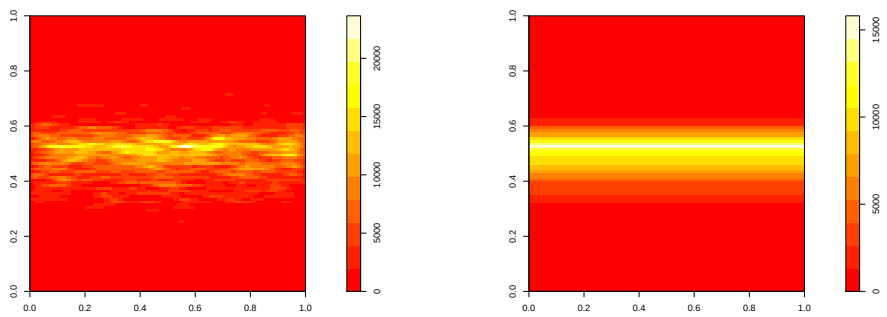

FIGURE 1.5 – Left : Intensity estimate $\hat{\lambda}$. Right : Intensity estimate $\tilde{\lambda}$.

The graph supports the separability assumption. This conclusion is emphasized by the Monte-Carlo separability tests which do not reject the separability assumption.

Dependence between mark categories

Next, we test the dependence between the emergencies of different categories of marks by estimating the functions L_{cross} for all pairs of categories. These

functions measure the dependence between the points of types i and j at distances $r \geq 0$. The calculation of L_{cross} requires a great deal of memory, so, we restrict our study to the June emergencies. We estimate the overall intensity of all the emergencies by a semiparametric method using a model with a single covariate (population) as is done in section 1.5. As in Moller & Waagepetersen (2004), the estimated intensity for each category is chosen to be proportional to the overall intensity estimate in order to have an expectation of the number of points equal to the number of emergencies in each category. The 39 simulations for the envelope calculation are obtained by taking the same positions of the multitype point pattern but with a random permutation of categories. Figure 1.6 presents the estimates and envelopes of L_{cross} corresponding to the three pairs of categories : (Low,Medium), (Low,High) and (Medium,High). For the three pairs of categories the estimated $L_{cross}(r) - r$ lies within the envelope even though it appears to track the upper envelope boundary and sometimes exceed it in a few instances. So, we can consider that the emergencies of different categories of marks are independent. The independence is not a surprising hypothesis in this practical example and is verified under the assumption of independence between marks and locations (proportional intensities).

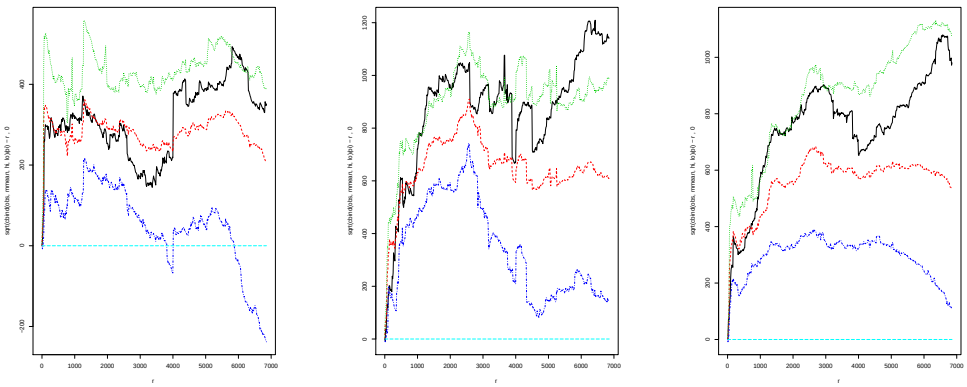

FIGURE 1.6 – Estimated $L_{cross}(r) - r$ for the three pairs of categories on emergencies in June (solid line), average and envelope from 39 multitype point patterns with same locations but categories given by a random permutation (dashed lines).

Marginal distribution

The previous sections have concluded that we can model the marginal point pattern of positions and marks separately in order to avoid a more com-

plicated model. This is the reason why now we analyse now the marginal distribution of marks. Figure 1.7 (Left) presents the histogram of the logarithm of workloads. At first sight, one may think that a log-normal model is acceptable considering the empirical marginal distribution of marks. However, the Normal Q-Q plot of the logarithm of marks in Figure 1.7 (Right) show that it is not a reasonable choice. The kurtosis value is far away from the kurtosis value of the adjusted normal model. Many others transformation with the aim to obtain a normal distribution as well as different models were attempted but none of these were satisfactory. The transformations considered include the Box-Cox transformation with an optimal parameter chosen by *boxcox.fit* (package **geoR**), inverse transformation,. . .. Moreover, we have tried in vain to fit several models from the logarithm of marks (Cauchy, Gaussian Mixture, . . .) or directly from the marks (Exponential, Generalized Pareto, Generalized Extreme Values, . . .). This difficulty in obtaining a satisfactory model for the workload marks suggests that we should consider a bootstrap procedure for generating "simulated" samples.

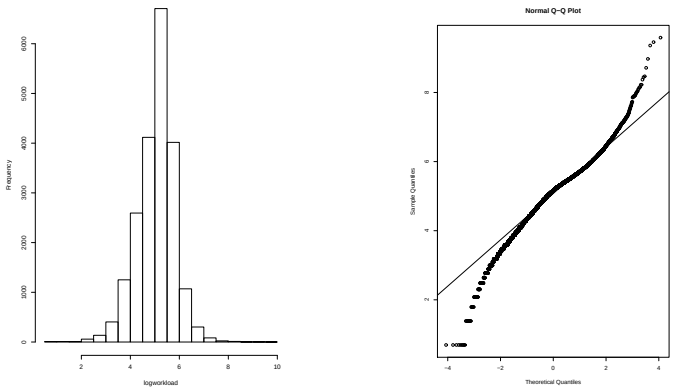

FIGURE 1.7 – Left : Histogram of the logarithm of workloads. Right : Normal Q-Q plot of the logarithm of workloads (xaxis : Theoretical Quantiles, yaxis : Sample Quantiles).

1.5 Model

On the basis of the previous analysis, we decide to consider in this section the marginal spatial point pattern of positions aggregated over the year for fitting a spatial point process model ignoring the marks. But, due to the high number of locations, we take a subsample of this point pattern corresponding to emergencies of a particular month, for example, June. For testing the absence

of interaction, we choose to apply a Monte-Carlo test by computing simulated envelopes of the inhomogeneous L function under an inhomogeneous Poisson process model. The first step in order to estimate the L function and to simulate realizations of an inhomogeneous Poisson process is to estimate the intensity function. We investigate three methods for estimating the intensity : parametric, nonparametric and semiparametric with one covariate.

Parametric and Nonparametric estimation

The parametric method consists in estimating the logarithm of the intensity with a polynomial in the coordinates. We estimate the polynomial coefficients by the method of maximum pseudo-likelihood (Moller & Waagepetersen (2004)). However, parametric models with a reasonable degree (< 5) are often unsatisfactory in the presence of high inhomogeneity of the locations of points in the domain. The resulting intensity is a rough estimate and the coefficients are difficult to compute for higher degrees.
An alternative is to use nonparametric methods that are more adaptable. A major problem is always to separate inhomogeneity explained by the intensity λ and interactions measured by the L function. Figure 1.8 (Left-Middle1) shows that our choice of $h = 800$ yields an intensity estimate and a simulated point pattern close to the point pattern of emergencies in June. So, from the point of view of the first order characteristic, an inhomogeneous Poisson point process seems to be an appropriate model. The choice of the bandwidth for the kernel estimation is of primary importance. As in Diggle (2003), the estimated $L(r)-r$ in Figure 1.8 (Middle2) shows that the selected bandwidth is too small and involves an over-fitting problem. Figure 1.8 (Right) also displays the estimated $L(r) - r$ and envelope when we use the leave-one-out estimate $\bar{\lambda}$ of the intensity function introduced in Baddeley *et al.* (2000) to correct the bias in the estimate of $L(r) - r$. Its formula is given by

$$\bar{\lambda}(s) = \frac{1}{\hat{c}_{W,h}(s)} \sum_{i=1}^{n} K_h(s - x_i) \mathbb{1}(x_i \neq s)$$

If $\hat{\lambda}$ and $\bar{\lambda}$ are approximated by their values evaluated at a fixed grid of points, the two estimators of the intensity surface agree with probability 1. The difference with the usual estimator consists in not taking into account in the summation the point of the pattern at which we estimate the intensity. The use of this estimate in the estimation of K gives a better bias in the simulation example of Baddeley *et al.* than the classical one. The bandwidth $h = 800$ with the leave-one-out estimator gives here an envelope which is

difficult to interpret due to the surprising form of the mean curve under the null hypothesis of a Poisson process (Figure 1.8 (Right)).

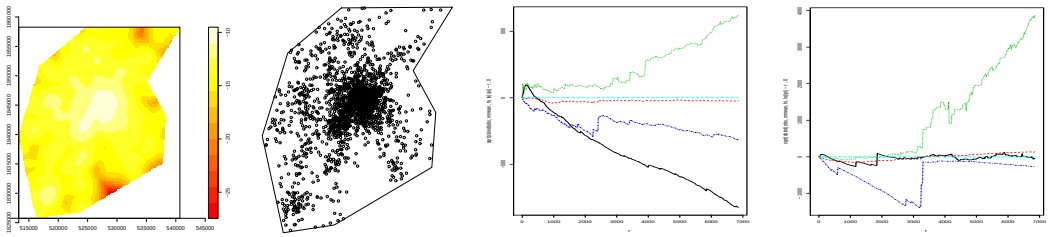

FIGURE 1.8 – Left : Nonparametric density estimation with bandwidth $h = 800$ (June emergencies). Middle1 : A simulation from an inhomogeneous Poisson process model. Middle2 : Estimated $L(r) - r$ on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous Poisson process (dashed lines) with $\hat{\lambda}$. Right : As previously with the leave-one-out estimation $\bar{\lambda}$ of the intensity.

Moreover we think that the use of the same data to estimate nonparametrically both λ and L is problematic. Indeed, we obtain better results by using the emergencies in May for the estimation of λ, which is then used in the estimate of the L function for the point pattern in June. Figure 1.9 shows that this method allows to fit a Poisson process model with a similar first order characteristic, a good simulated process and a better graph for $L(r) - r$. Finally, the envelope of $L(r) - r$ implies that we conclude to neither aggregation nor repulsion in the point pattern ; neither do we conclude to an over-fitted model. However, in this case, the envelope is highly dependent on the choice of the subsample for the estimation of λ. For instance, the choice of the month of July for the estimation of λ would involve "artifact" on the envelope estimate. This is a reason why we did not pursue this direction further.

Semiparametric estimation with one covariate

In many cases, the intensity of the spatial point pattern depends on covariates. For instance, our spatial point pattern is influenced by environmental and economic covariates : population, presence of woods,.... In our study, we have a population covariate which allows us to estimate the intensity of emergencies from an estimate of the population density. Our population covariate is the number of inhabitants in 296 INSEE administrative units named IRIS. We know the total population and the centroid of each IRIS. Figure

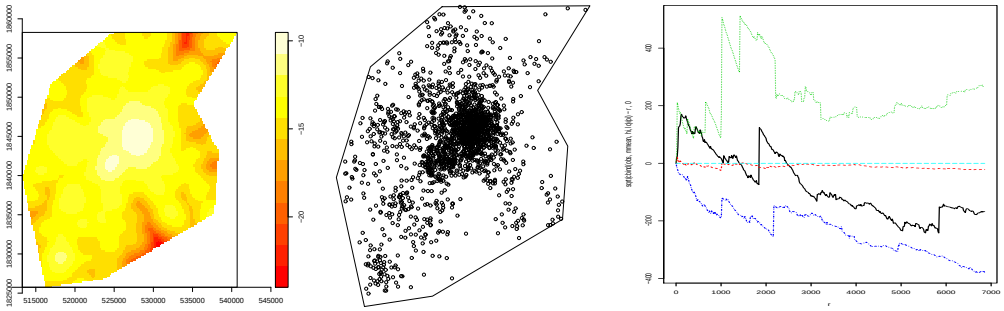

FIGURE 1.9 – Left : Nonparametric intensity estimation with bandwidth $h = 800$ (May emergencies). Middle : A simulation from an inhomogeneous Poisson process model. Right : Estimated $L(r) - r$ on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous Poisson process (dashed lines).

1.10 represents these units with a circle centered at those centroids with radius proportional to the number of inhabitants. We denote by $\xi_1, \cdots, \xi_{296}$ the centroids of the administrative units and by $N_1, \cdots, N_{296}$ their number of inhabitants. It is necessary to know the values of this covariate at every point in the window in order to estimate the background intensity. Consequently, we predict the covariate on a regular grid with a nonparametric predictor and then estimate the coefficient α and β in the expression $\lambda(s) = \exp(\alpha + \beta \log(\hat{C}(s)))$ by maximum pseudo-likelihood, where $\hat{C}(s)$ is the estimate of the covariate.

Two density estimates

We present two nonparametric methods to estimate the population density. The first one $\hat{C}_1(s)$ is a classical nonparametric kernel method with a selected global bandwidth $h = 900$ and a border correction factor. The second kernel method uses an adaptive choice of bandwidth based on the k-nearest neighbors. At each point of a regular grid, we estimate the population density by applying an Epanechnikov kernel k_e with its support adapted in order to take into account only k centroids. We arbitrarly choose $k = 5$. The expression of this estimator is

$$\hat{C}_2(s) = \frac{1}{N_{Pop} h_s} \sum_{i=1}^{296} N_i k_e \left(\frac{s - \xi_i}{h_s} \right)$$

where $N_{Pop} = \sum_{i=1}^{296} N_i$, $k_e(s) = \frac{3}{4}(1 - \|s\|^2)$ and $h_s = \|s - \xi\|_{(5)}$ is the fifth

order statistic of distances between s and the IRIS centroids. We do not use here any correction factor for border effects.
Figure 1.10 displays the logarithm transformation of these two density estimates. We note that the second approach has the advantage of clearly identifying the biggest cities in this region. Infortunately, the intensity is under-estimated near the boundary of the observation region.

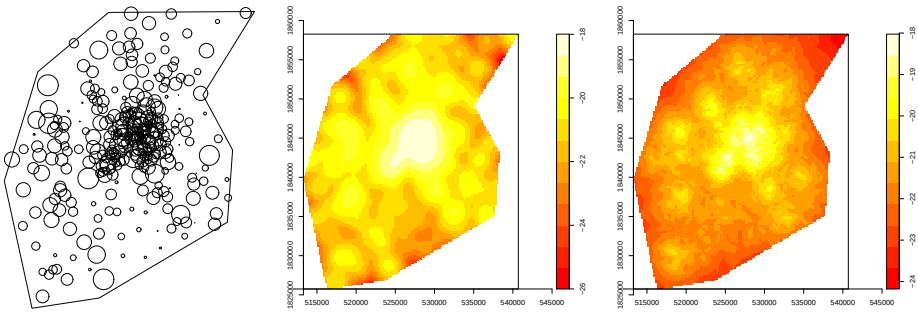

FIGURE 1.10 – Left : Proportionnal symbol map of number of inhabitants per IRIS. Middle & Right : Logarithm transformation of the population density estimated by a nonparametric kernel method with a global bandwidth (Middle) and a local bandwidth obtained by k-nearest neighbors (Right).

Models based on density estimate $\hat{C}_1$

First of all, we focus on models constructed from the density estimate $\hat{C}_1$ obtained by the first method. By maximum pseudo-likelihood, we obtain $\hat{\alpha}$ and $\hat{\beta}$ and write $\hat{\lambda}_1(s) = \exp(\hat{\alpha} + \hat{\beta}\log(\hat{C}_1(s)))$ for all s in the regular grid. Figure 1.11 shows a simulation of a Poisson point process with intensity $\hat{\lambda}_1$ and an envelope which lead us to reject the hypothesis of no interaction, and suggests aggregation for $r \leq 1500$.
We tested three inhomogeneous point processes models of clustering : the Matern cluster process and the Thomas cluster process which belong to the class of Neyman-Scott processes and the Log Gaussian Cox Process (Moller *et al.* (1998)). Neyman-Scott processes and Log Gaussian Cox processes are cluster processes in the class of Cox processes. A Cox process is obtained by considering the intensity function of the Poisson process as a realisation of a random field. Neyman-Scott processes are obtained by clustering points around a homogeneous Poisson point process with intensity κ ("mother" process). A realization of a Neyman-Scott process is obtained by all the realizations of independent Poisson processes at each "mother" point. This daughter point process has an intensity function which depend on a kernel

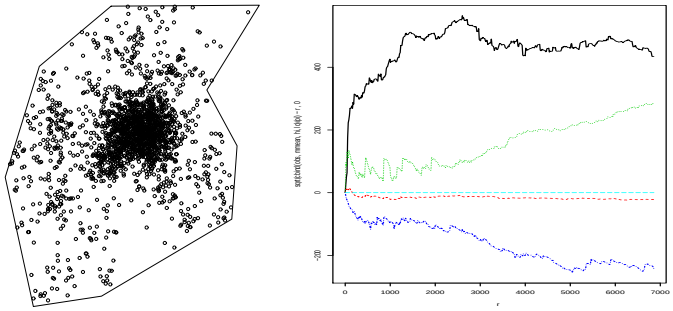

FIGURE 1.11 – Left : A simulation from an inhomogeneous Poisson process model with intensity $\hat{\lambda}_1$. Right : Estimated $L(r) - r$ on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous Poisson process (dashed lines).

function. The two point process models considered here are given by a specific kernel (Moller & Waagepetersen (2004)). For a Log Gaussian Cox process, the intensity function is the exponential transformation of a Gaussian field (Moller et al. (1998)).
The inhomogeneity can be incorporated by different methods (Jonsdottir (2004)). But, for the class of Neyman-Scott processes, it is necessary to incorporate this inhomogeneity by thinning as in Waagepetersen (2006). Indeed, this method is the only one that allows to get an inhomogeneous Neyman-Scott process which is second-order intensity reweighted and for which the inhomogeneous K function is well defined. Thinning is an easy method by which to simulate inhomogeneous point processes : we simulate a realization of a stationary point process X and afterwards apply an independent thinning method by the field defined from $\hat{\lambda}_1$ to obtain a realization of an inhomogeneous point process Y. The advantage is also that the inhomogeneous K function of Y coincides with the K function of X. This fact allows us to estimate the parameters κ and ω of the point process model by minimizing the contrast

$$\int_0^a (\hat{K}_{inhom}(t)^q - K(t;\kappa,\omega)^q)^2 dt$$

where $K(t;\kappa,\omega)$ is known for the class of point process models presented before. For the choice of a and q, Diggle (2003) recommends to choose a considerably smaller than the dimension of the observation plot and $q = 1/4$. We take $a = 4000$ meters.
The thinning method modifies the structure of the point process model. For

example, Figure 1.12 (Left) displays a simulation of the fitted inhomogeneous Thomas cluster process which shows that the expected number of points per cluster is different. If we incorporate the inhomogeneity by considering an inhomogeneous Poisson point process for the "mother" points, then the usual structure is maintained in comparison with the stationary case. Indeed, the expected number of points per cluster is constant in this case. Figure 1.12 presents a simulation of the fitted Thomas point process which looks quite different from the emergencies in June. The minimum contrast estimation yields an expected number of "mother" points and a scale parameter which are too small. By construction, the second order characteristic of this fitted point process model is close to that of the point pattern (Figure 1.12 (Right)).

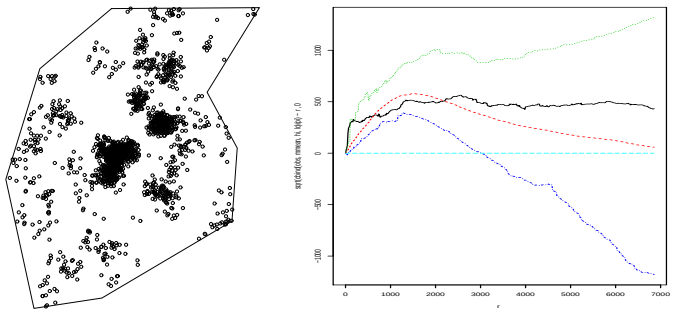

FIGURE 1.12 – Left : A simulation from an inhomogeneous Thomas point process model obtained by thinning. Right : Estimated $L(r) - r$ on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous Thomas point process model obtained by thinning (dashed lines).

We now generalize to the case of Log Gaussian Cox processes (LGCP) the method proposed in Waagepetersen (2006) for Neyman-Scott processes.
The simulation of the fitted LGCP in Figure 1.13 (Left) features aggregation areas as in our point pattern. However, these areas are wider and mainly concentrated around the city of Toulouse. Moreover, several areas exhibit no points or very few points in the simulation whereas they are important areas of emergencies in the point pattern. This is the case of the area in the bottom-left of the region which corresponds to the large city of Muret. We also note that the boundary of the region has generally few points in the simulation. The envelope of $L(r) - r$ is large due to the fact that the variability of the expected number of points in the simulations is relatively important. The envelope suggests that the goodness-of-fit of this model is satisfactory.

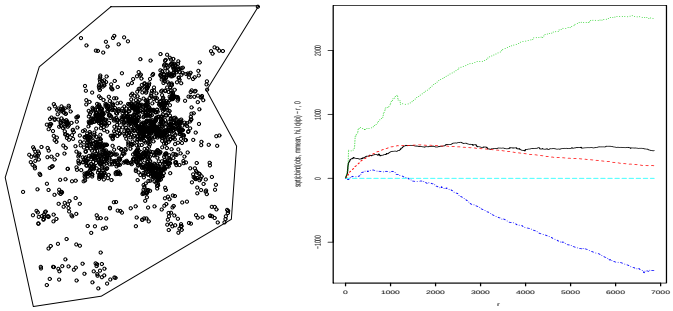

FIGURE 1.13 – Left : A simulation from an inhomogeneous LGCP model obtained by thinning. Right : Estimated $L(r) - r$ on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous LGCP model obtained by thinning (dashed lines).

Models based on density estimate $\hat{C}_2$

We consider the case where the background intensity estimate $\hat{\lambda}_2$ is derived from the density estimate $\hat{C}_2$. Compared to the simulation of a Poisson point process with estimated intensity $\hat{\lambda}_1$, the simulation in Figure 1.14 (Left) now features more aggregated areas and reveals the area of the city of Muret in the bottom-left of the region. From the first order characteristic point of view, this point process is a good model of the emergencies. The Monte-Carlo test of no interaction in Figure 1.14 concludes that our point pattern is more aggregated for approximately $r \leq 1500$ and more regular for large r than the Poisson model. We reject the hypothesis of no interaction and fit a LGCP model next.

The parameters of the LGCP are estimated as previously by the minimum contrast method and the inhomogeneity is incorporated by thinning. The obtained simulation shows that this point process model is interesting because the distribution of the aggregated areas is close to those of our point pattern (Figure 1.15 (Left). There is no void large area except near the boundary of the region. Therefore, the inhomogeneous LGCP model obtained from the thinning by the field $\hat{\lambda}_2$ yields a satisfactory model of the point pattern of emergencies in June. We only add that the estimate $\hat{\lambda}_2$ should be improved by considering an edge correction factor in the density estimate $\hat{C}_2$.

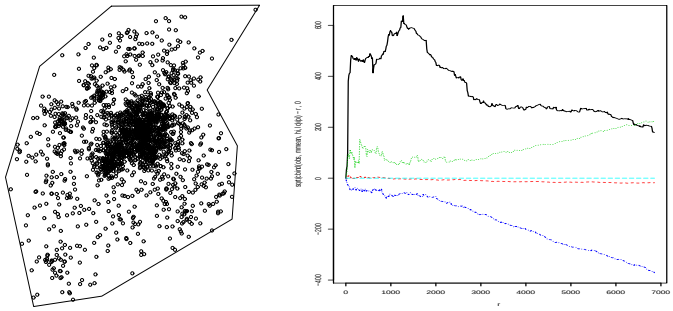

FIGURE 1.14 – Left : A simulation from an inhomogeneous Poisson process model with intensity $\hat{\lambda}_2$. Right : Estimated $L(r) - r$ on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous Poisson process (dashed lines).

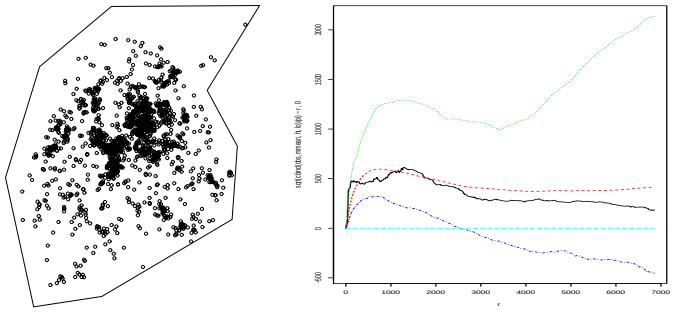

FIGURE 1.15 – Left : A simulation from an inhomogeneous LGCP model obtained by thinning. Right : Estimated $L(r) - r$ on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous LGCP model obtained by thinning (dashed lines).

1.6 Conclusions

The study of this spatio-temporal marked point pattern of emergencies during one year underlines the numerous difficulties faced when analyzing complex and large data sets. First of all, the high inhomogeneity and the many duplicated points are a major problem in the estimation of the background intensity of the emergencies. These difficulties result in problems in finding a good bandwidth h which does not lead to over-fitting.

In the case of inhomogeneity of the positions, the global test of independence between positions, time and continuous marks is intricate. So, we have

tested this dependence two by two on different subsamples for computational reasons. It seems hard to make a definite choice between the different point process models considered here. Indeed, the Poisson point process with intensity estimated nonparametrically yields a simulation with localizations of events close to that of our point pattern, but the estimate of the L function presents some over-fitting. It is difficult to decide whether this phenomenon is due or not to an "artifact" in the estimate of L.

With the semiparametric approach, we observe that the adaptive kernel estimation of the population density yields better point process models than the classical kernel estimation with a global bandwidth. So, we retain as acceptable models for our data set, the Poisson point process with intensity $\hat{\lambda}_2$ and the LGCP with inhomogeneity obtained with the thinning by $\hat{\lambda}_2$. The goodness-of-fit is satisfactory for the first order characteristic for the Poisson model while it is good for the first and second order characteristics for the LGCP model. In spite of the boundary errors generated by the adaptive kernel estimation and the variability of the number of points per simulation, the LGCP appears to us as a good enough model of the June emergencies.

The generalization to the other months is made by considering intensities proportional to $\hat{\lambda}_2$ according to the expected number of points for each month. The marks realizations are obtained by a bootstrap procedure and are affected independently to the point pattern of positions.

Chapitre 2

Modèles de processus ponctuels spatiaux pour des problèmes de localisation-allocation

Sommaire

2.1 Introduction **56**

2.2 Literature on location-allocation problems **57**

2.3 The case of the fire stations location problem . . **58**

2.3.1 The data set 58

2.3.2 The optimization problem 59

2.4 SPP location-allocation **61**

2.4.1 Fitting the point process model 62

2.4.2 Bootstrapping the spatial locations 62

2.4.3 Optimization strategy 63

Selecting the compromise between distance and equilibrium improvement 63

The naive method 64

The genetic algorithm 65

The stochastic heuristic algorithm 66

Toy example 67

2.4.4 Analyzing the results on the firemen data 71

2.5 Conclusions **75**

Abstract :

The problem of finding an optimal location frequently occurs in geomarketing, economics and other fields : positioning a new branch of a bank, a supermarket, a fire station, a plant, designing a traffic network, etc. The optimal location of the source facility is the argument-minimum of an optimization problem parameterized by some characteristics of the clients. The random nature of some of these characteristics has already been recognized, but few stochastic models for location allocation problems address the issue of uncertainty of the locations of the clients, and even then they do it with very naive tools. It is proposed to recognize uncertainty in the spatial positions of the clients, and possible spatial autocorrelation as well, by considering the random inputs of the optimization as one realization of a spatial marked point process. The method, called SPP location-allocation, involves fitting a point process model, simulating from the adjusted process, and solving a family of optimization problems for each simulated set of observations. The advantage of this approach over the deterministic one is twofold : it gives an indication of the spatial variability of the optimal solution, and it allows one to solve larger problems. Finally an application to the optimal positioning of a new fire station in the Toulouse area (France)is presented with some heuristic algorithms.

Key words : Spatial point processes, conditionally multisource location-allocation problem, optimal location.

2.1 Introduction

We consider the general problem of finding the optimal location of a set of a given number of source facilities for clients so as to minimize a given cost function and balance the workload of facilities. Classical examples of such problems include locating a new store so as to minimize the transportation cost of supplies to stores in order to satisfy customers' demand, or locating a new plant so as to minimize the transportation cost of workers to plants. We use a practical case to illustrate our method, which is locating one or possibly several new fire stations in order to minimize the total access time of firemen to emergencies, and reach a relative equilibrium of firemen workload.

As one can tell from the above examples, the different inputs into the problem are random quantities. However this random nature is often ignored by the classical operational research algorithms. Where random approaches to such problems (see random or stochastic or uncertain programming models) have been used, they usually ignore uncertainty about location, or assume

independent random inputs, or even sometimes identically distributed random inputs. The problem with these assumptions is that they ignore the facts that the locations of the clients (customers, workers, emergencies) may be unevenly distributed across space, and there may be some spatial correlations between locations and/or characteristics of the clients. To take this into account, we propose modeling these inputs as realizations of a spatial point process (possibly marked). The purpose of this paper is to illustrate the advantages of this novel perspective through a case study, which is presented in the next section, about locating a new fire station, using standard optimization tools and standard statistical techniques.

2.2 Literature on location-allocation problems

The fire station optimization problem we consider in the next section consists of locating a new fire station, given the locations and characteristics of past emergencies, and given the locations of existing fire stations. This problem belongs to the location-allocation family and is a conditional multisource Weber problem ; this field is very active in the operations research community. The word conditional refers to the fact that only new fire stations have to be located. We refer the reader to ReVelle and Eiselt (2005) for a survey on location-allocation problems and to Brimberg et al. (1997) for a survey on heuristic algorithms for the multisource Weber problem.

Although most of this literature deals with deterministic formulations, some papers consider that the environment may change, and introduce a random dimension to the problem. Snyder (2005) offers a good review of stochastic and robust facility location models. Cooper (1974) considers a stochastic extension of the Weber unconditional problem with independent Gaussian random locations but minimizing the expected cost. Stochastic demand is considered in Logendran and Terrell (1988), who treat the case of an uncapacitated transportation plant location-allocation problem, and in Zhou and Liu (2003), who introduce a hybrid intelligent system for a capacitated location-allocation problem. Drezner (1985) analyses the sensitivity of the optimal location in a Weber problem to small fluctuations in the demand locations.

A number of papers focus more specifically on the siting of fire station or emergency services, but with different viewpoints. Daskin (1982), Goldberg and Paz (1991) and ReVelle (1991) address the problem of adjusting the number of emergency medical service vehicles to obtain a given coverage rate in a given time period. Serra and Marianov (1998) treat the case of locating a fire station in the region of Barcelona but their approach does

not take into account uncertainty on the locations of emergencies. Daskin (1982) develops an expected covering location model for emergency services, accounting for the possibility of a vehicle being busy.
From the algorithmic point of view, these are difficult problems because of their high dimensionality and the existence of many near-maximal solutions. Cooper(1964) introduced heuristic algorithms for these problems, alternating between a location step and an allocation step. Even today, they are useful to get satisfying local solutions in a reasonable computing time, and are often combined with other methods. Among recent methods let us mention D.C. programming (Chen et al., 1998), tabu search (Brimberg et al., 1996), genetic algorithms (Houck et al., 1996), ant colony algorithms (Bischoff and Dächert (2007) and Huang, Liu and Chandramouli (2006)) and swarm optimization. We have restricted attention to techniques that were already or could easily be implemented in the R software in order to be able to couple it easily with the modeling phase, which is most easily done with R.

2.3 The case of the fire stations location problem

2.3.1 The data set

The data set has been provided to us by the SDIS 31, "Service Départemental d'Incendie et de Secours de la Haute-Garonne" (Haute-Garonne is the "département" in which Toulouse is the main city). It consists of the locations and characteristics of about $20,000$ emergencies in and around the city of Toulouse during 2004 (this area will be denoted by Ω). We define an emergency as any event resulting in a call to a fire station : it includes fires but also accidents, and all sorts of incidents. Let N be the total number of emergencies, $X_1, \cdots, X_N$ in $\mathbb{R}^2$ be the locations of the emergencies, and $D_1, \cdots, D_N$ be their associated workload (duration in minutes from first arrival of firemen on site until last departure, times number of firemen involved). N, $X_1, \cdots, X_N$ and $D_1, \cdots, D_N$ will be modelled as random in section 2.4.1. We are given the locations of $J = 6$ existing fire stations $(s_j, j = 1, \cdots, J)$ and their respective number of firemen z_j $(Z = \sum z_j)$: J and the z_j are naturally considered as non random quantities. The median number of firemen per fire station is around 64 and the median workload is around 174 minutes in this base. The data base also contains some other characteristics of the emergencies such as the number of vehicles involved, and the type of emergency (fire, accident, and so on). The two most frequent emergencies are accidents and sicknesses in the street or the road. We have deliberately

ignored these other characteristics of the emergencies in order to first tackle a simpler problem. The SDIS 31 would like to create one (or several) new fire stations in order to reduce the overall travel time to the emergency locations and to relieve the overload of some of the existing ones. The emergencies locations and durations are intrinsically random : they vary over time and the data base is just a particular picture of the situation at a given time upon which we base the long term decision of building a new fire station. We will restrict attention to the case of positioning one new fire station although most of the presentation can be adapted to the case of positioning several new fire stations. An equivalent problem is that of relocating a site (for satisfying administrative constraints) and it is in fact this one that the SDIS 31 was concerned with in the first place. We assume that the number of firemen in the new fire station (numbered $J+1$) is given (equal to 60 in the application).
To evaluate the relevance of a location for the new fire station we need to re-allocate the emergencies to the set of stations including the new one : let $\alpha(i)$ denote the index of the fire station allocated to emergency i. Let Δ_j denote the mean workload in minutes of a fireman in fire station j

$$\Delta_j = \frac{\sum_{i:\alpha(i)=j} D_i}{z_j},$$

and $p_j = \frac{\Delta_j}{D}$ denote the fraction of total workload supported by an average fireman in fire station j, where $D = \sum_{i=1}^{N} D_i$ is the overall workload. Note that given the definition of the durations D_i, this workload does not include travel time.

2.3.2 The optimization problem

The first objective we want to minimize in that problem is related to the total travel time to the emergency sites. Given the location of an emergency, the travel time could also be considered as being random, because conditions of traffic vary according to the time of the day. However, modeling this randomness would increase the degree of complexity of the optimization algorithm and obscure our purpose so we ignore it in a first stage. A good proxy for the travel time would then be the Euclidian distance between an emergency and the closest fire station. Indeed for firemen, it is important to penalize large distances because a large travel time may have disastrous consequences, which is why we decided to use the square of the Euclidian distance to proxy the cost of travel time. We are fully aware that this choice may induce a large influence of outliers if any. However a careful inspection of large distances in the data set should be able to discriminate between outliers due to

reporting errors (which should be eliminated) and true large distances (they represent a small fraction of the total and occur in rural areas). An alternative choice would be to use the maximum distance objective as a proxy for an upper bound on access to a fire but it would induce the same problems with respect to outliers. In addition, the choice of squared distance turns out to simplify the optimization problem further as we will see below.
Our second objective is to achieve a relative equilibrium in the workload of firemen across facilities. To measure this objective, we need to choose a criterion $\Phi_N(\alpha, \{(X_1, D_1), \cdots, (X_N, D_N)\})$ which is minimal when the mean workloads of fire stations are equal. A simple choice is to use the maximum difference of workloads

$$\Phi_N = \max_{j=1,\ldots,J+1} \Delta_j - \min_{j=1,\ldots,J+1} \Delta_j. \tag{2.1}$$

This criterion is minimal and equal to zero when all mean workloads are equal. However it ignores the behaviour of the intermediate workloads. More elaborate choices that take into account the whole of the distribution would be the Gini index or the entropy criterion, but the first one has the advantage that its value is directly interpretable and this is the reason why we use for parameter selection it in the sequel. Recall the definition of the Gini index

$$\Phi_N = \frac{\sum_{j,l} \frac{z_j}{Z}\frac{z_l}{Z}|\Delta_j - \Delta_l|}{2\sum_j \frac{z_j}{Z}\Delta_j} = \frac{\sum_{j=1}^{J+1} z_j \sum_{l=1}^{J+1} z_l|\Delta_j - \Delta_l|}{2ZD}. \tag{2.2}$$

We then have

$$\Phi_N = 0 \Leftrightarrow \Delta_j = D/Z \quad \forall j = 1, \cdots, J+1. \tag{2.3}$$

For the entropy criterion, we have

$$\Phi_N = \sum_{k=1}^{Z} p_{j(k)} \log p_{j(k)} = \sum_{j=1}^{J+1} z_j p_j \log p_j, \tag{2.4}$$

where $j(k)$ is the index of the station to which fireman k belongs. This criterion is minimal when $p_1 = \cdots = p_{J+1} = \frac{1}{Z}$ and in that case $\Phi_N = -\log(Z)$.
As mentioned above, when locating a new facility, we need to re-allocate the entire set of clients to the set of facilities including the new one. Therefore the problem of finding the optimal location cannot be separated from the problem of finding an optimal allocation map $\alpha : \Omega \longrightarrow \{1, \cdots, J+1\}$. Of course, this optimal allocation has limited practical use, since it allocates emergencies of

the past in a situation of the future : it is just a tool in the model. Even if we were considering a dynamic version of the allocation problem, a dynamic optimal allocation would have no sense for the firemen, who consider that an emergency has to be assigned in real time to the closest-available fire station, the equality of workloads being just a long run goal. Finally the position of the new fire station $s_{J+1} = s$ and the allocation α of emergencies to fire stations must be optimized simultaneously, but the optimal allocation is a nuisance parameter in our framework.

To keep the computations manageable, some other constraints have been deliberately ignored in our first pass at the problem, such as bounds on the total workload of a station, but these would be easy to include. Using a multicriteria optimization approach (Ehrgott, 2005), we consider the following mono-objective problem

$$(s^*, \alpha^*) = \arg\min_{s,\alpha} \lambda \sum_{i=1}^{N} \| X_i - s_{\alpha(i)} \|^2 + (1-\lambda)\Phi_N(\alpha, \{(X_1, D_1), \cdots, (X_N, D_N)\}). \tag{2.5}$$

We later discuss the choice of the respective weights assigned to each part of the objective function.

2.4 SPP location-allocation

The method we propose here consists of several steps. In the first step, we consider that the locations of the emergencies $X_1, \cdots, X_N$ in $\mathbb{R}^2$, and their durations $D_1, \cdots, D_N$, constitute a realization of a marked point process (X, D). The statistical theory of marked point processes is well established and they can be readily simulated; see for example Moller and Waagepetersen (2004). For the parameter estimation as well as the simulations, the implementation can be done with the R package "spatstat" of Baddeley and Turner [1](2005 and 2006). Model fitting is discussed in section 2.4.1.

The aim of the second step is to obtain small-size replications of the phenomenon under study. We argue in section 2.4.2 that the smooth bootstrap is a good solution for this. The size of the data set for each of the optimization problems is controlled by the user, so the first advantage of this approach is that, even though one has to repeat the optimization, each of the optimization problems is small.

A common approach in stochastic location is to minimize the expected value of the objective function. This leads to a unique final optimal location

[1]http ://cran.r-project.org/src/contrib/Descriptions/spatstat.html

and this is why we argue that it is preferable to apply an optimization algorithm (see section 2.4.3) to each replication, find a set of optimal locations for the new fire station and then analyze its statistical properties. From the set of optimal locations thus obtained, one can produce contour plots of optimal locations, as we will see in section 2.4.4, associated with levels of confidence. This conveys a sense of the spatial variability of the solution, and is the second advantage of this method. A decision maker can use the contour plots to determine an optimal zone corresponding to a given confidence, and then use other arguments (land availability for example) to select a location in this zone.

2.4.1 Fitting the point process model

In Bonneu (2007), several models are explored for modeling this data set. Difficulties arise from the fact that there is a strong lack of homogeneity, and from the presence of duplicated points (positions of emergencies are approximated by the nearest point of a given network). It is found that variation in space is more important than variation in time, and that temporal stationarity is justified so we can aggregate over time. Separability issues between time and space, space and workload, workload and time are also considered concluding to pairwise independence. Three methods are tested to fit the intensity : a parametric approach with polynomials in the coordinates, a purely non-parametric method by kernel smoothing, and a semi-parametric estimate involving a non-parametric density estimation of the population with adaptive bandwidth. In the parametric and semi-parametric case, the intensity is fitted by maximum pseudo likelihood (Moller and Waagepetersen, 2004). The semi-parametric estimate $\hat{\lambda}$ including population density as a covariate turns out to yield the best results. The second order properties of the model (presence of interaction) is then explored with three models : the Matern cluster process, the Thomas cluster process and the Log Gaussian Cox process (LGCP). Finally, two models are considered as acceptable : a Poisson point process model and an LGCP model. The marginal distribution of the duration proved difficult to approximate in a parametric family so we used simple bootstrap in the simulations of the workloads.

2.4.2 Bootstrapping the spatial locations

A basic bootstrap consists of sampling from the empirical distribution function of the data, whereas a smooth bootstrap consists of sampling from some estimated d.f. obtained by fitting some model to the data, or by smoothing the e.d.f. The difficulty for bootstrapping dependent data such as point pro-

cesses is to find a procedure that retains the dependence structure of the data. For spatial data, Hall (1985) appears to have been the first to use some kind of block resampling procedure. Davison and Hinkley (1997) give a brief overview of methods for spatial bootstrap. In the case of stationary and isotropic point processes, Loh and Stein (2004) introduce a nonparametric method where marks are computed and assigned to observed points before performing bootstrap. Cowling, Hall and Phillips (1996) describe the resampling methods for an inhomogeneous Poisson process, including the smooth bootstrap, for constructing confidence regions for the intensity function. Guan and Loh (2007) use a thinned block bootstrap procedure on a stationary point process derived from a second-order reweighted stationary point process in order to make inferences on the regression parameters. The problems encountered with the resampling methods and the block bootstrap procedure are, for example, that too many events coincide in the bootstrapped samples compared to the original one, or that events are abnormally close or far away from one another. Moreover these bootstrap methods give poor results for the class of non-stationary point processes. Snethlage (1999) criticizes the bootstrap approaches and proposes alternatives for variance estimation of the pair correlation function and confidence regions for the intensity of an inhomogeneous Poisson process. Finally, the smooth bootstrap seems to us to be the best alternative.

2.4.3 Optimization strategy

Although the emphasis in this paper is not on discussing optimization algorithms for this type of problem, we present in this paragraph the particular choices we have made in our case study. We have selected three types of algorithms : a naive algorithm consisting in evaluating on a finite grid, a so called heuristic algorithm, and a genetic algorithm. In order to select an appropriate method for our specific case, we have tested these algorithms on a simulated toy example.

Selecting the compromise between distance and equilibrium improvement

In the objective function (2.5), one has to select the parameter λ that regulates the balance between the distance term and the need to equalize workloads (equilibrium term). It is clear that this decision has to be made by experts (in our case : the firemen). In order to guide them for this choice, we use a slightly different way of writing the objective function. We divide the distance part of this objective function Q_7 by the value of the sum of

squared distances between the emergencies and the closest fire stations in the initial situation Q_6 (with 6 fire stations) and call it the distance improvement rate. Similarly, we divide the equilibrium part C_7 by its value in the initial situation C_6 and call it the equilibrium improvement rate. In this way, the two contributions can be interpreted as rates of change from the initial to the final situation. We found that criterion (2.1) was easier to work with for interpretation purposes.

The objective function (2.5) can then be written $M = \lambda \frac{Q_7}{Q_6} + (1-\lambda)\frac{C_7}{C_6}$.

Two sets of location-allocation outcomes corresponding respectively to Q_7^1, C_7^1 and Q_7^2, C_7^2 yielding the same value of the objective must satisfy

$$\frac{\lambda}{1-\lambda}\frac{C_6}{Q_6} = \frac{C_7^2 - C_7^1}{Q_7^1 - Q_7^2}. \tag{2.6}$$

The parameter $\tau = \frac{\lambda}{1-\lambda}\frac{C_6}{Q_6}$ can then be interpreted as follows. In the initial situation, the mean distance of an emergency to its nearest fire station is $3,600$ meters and the maximum difference of workload (2.1) is 8.57 hours. If the firemen are ready to improve the mean distance by 500 meters with the price of increasing the maximum difference of workload by 1 hour, this choice corresponds to a value of λ, given by (2.6), equal to 0.86. We apply the same principle in the toy example, which results in a different value of the parameter for each draw.

The naive method

Our naive method consists of searching for an optimal location for the new fire station out of a finite set of locations. This set of locations can be the locations of the observed emergencies, or nodes on a rectangular grid. We adopt this second option, which allows us to control the amount of calculation required by selecting the number of evaluation points. For each position in this finite set, we compute the objective function (2.5) by allocating the emergencies automatically to their nearest fire station, and then derive the corresponding optimal location. The accuracy of the optimal location depends on the dimension of the grid chosen by the user. Graphs of each term of the objective at the grid points can bring information about their respective behaviors. The drawback of this method is that in the case of several fire stations the problem dimensionality of the naive algorithm (size of the grid) becomes rapidly intractable. The following method improves on that point.

The genetic algorithm

Genetic algorithms attempt to construct improved solutions from predecessors, in an evolutionary type process (Holland, 1975), like the idea of "natural selection" in biology introduced by Darwin. In location-allocation problems, this type of stochastic algorithm has already been developed for the resolution of a multisource Weber problem by Houck *et al* (1996).

The general scheme is as follows : the first step is to introduce a population of feasible solutions, called individuals. These individual solutions are composed of one or several chromosomes which can also be composed of one or several genes. A fitness function is chosen to evaluate the quality of the solutions. A proportion of the existing population is selected through a fitness-based process, to breed a new generation. There are several tools for obtaining new individuals like the crossover operator, the mutation rate, and so on. We repeat the generation of this new population until a termination occurs using a criterion chosen by the user. Several formulations of genetic algorithms are possible for our problem and we need to make a careful choice in order to converge in a reasonable amount of time.

The dimensionality of the location-allocation variable is very large and genetic algorithms with a large feasible domain of individuals are greedy, so we have two options : optimize on the variable allocation, using the fact that the optimization on the variable location, knowing allocation, is straightforward in some cases, or optimize on location, using allocation to the nearest fire station. These two alternatives involve individuals made of one chromosome with several genes (N or 2 respectively).

The first method consists of considering a chromosome of N genes, where gene i is the index of the fire station to which the emergency x_i is allocated. At the initialization step, this index is randomly chosen among all fire stations $\{1, \ldots, J+1\}$. A faster version is to restrict the set of indices for each emergency to a few stations like the nearest fire station and the new one : $J+1$. This last option allows the problem to converge to a solution in a reasonable time for larger data sets.

In the second alternative, a chromosome is composed of 2 genes. The first gene represents the x-coordinate and the second one the y-coordinate of the position of the new fire station. Given the dimensionality of our problem, we consider this approach in our problem. We describe below the different steps and the parametrization of our genetic algorithm, which we implemented with R.

- Step 1 : Generate uniformly 100 different initial locations.
- Step 2 : For each position, compute the fitness function which consists

of the objective function with allocation to the nearest fire station.

- Step 3 :
  Repeat (1) through (5)

  1. Draw, with replacement, 50 couples of positions from the initial population and retain one position per couple with the tournament method. This method consists of selecting the location that yields the minimum fitness criterion.
  2. Mix pairs of selected locations with a crossing operator of rate 0.8 to create new locations. For a couple of locations (x_1, y_1) and (x_2, y_2), new locations are : $(1/3)(x_1, y_1) + (2/3)(x_2, y_2)$ and $(2/3)(x_1, y_1) + (1/3)(x_2, y_2)$.
  3. Modify the new solutions with a mutation operator of rate 0.01. The mutation of a location consists of changing a randomly chosen coordinate by another value uniformly drawn in the observation window.
  4. Insert the solutions into the initial population.
  5. Retain 100 positions corresponding to the best 100 evaluations of the fitness.

  Until number of generations equals to 15.

At the end of the genetic algorithm, we obtain a population of 100 feasible positions. Our optimal position for the new fire station is the solution that minimizes among these 100 positions the objective function computed with an allocation to nearest fire station.

Compared to the naive method, the genetic algorithm is easily adaptable to the case of several fire stations and can treat larger size problems. However, these two methods present the drawback of searching in the subset of allocations to the nearest fire station. The following method improves on that point.

The stochastic heuristic algorithm

We propose a heuristic algorithm for this problem based on the fact that, given the allocation function, we know the analytic solution to the optimization problem because of the simple form of the cost function : it is simply the centroid of the emergencies allocated to the new station. The procedure is as follows :

- Step 1 : initialize the position and the allocation of emergencies for the new fire station by the optimal solutions obtained by the naive method. Then compute the resulting workloads Δ_j, for $j = 1, \ldots, J+1$.

- Step i $\rightarrow$ Step i+1 : $m_\Delta = \frac{1}{J+1}\sum_{j=1}^{J+1}\Delta_j$ is the mean of the workloads Δ_j. For each fire station, compute a selection probability $sp_j = \frac{|\Delta_j - m_\Delta|}{\sum_{j=1}^{J+1}|\Delta_j - m_\Delta|}$. These selection probabilities reflect the contribution of each fire station to the workload unbalance. Randomly select a fire station d using these selection probabilities. Search for the nearest fire station j to the fire station d. If $\Delta_d < \Delta_j$ then search for the emergency allocated to j which is nearest to fire station d, and re-allocate it to d. Otherwise, search for the emergency allocated to d which is nearest to the fire station j, and re-allocate it to j.

- Compute the new Δ_j for $j = 1, \ldots, J+1$ and the position of the new fire station corresponding to the centroid of the emergencies allocated to it.

- Repeat this procedure until the objective cannot be improved upon during 50 iterations.

Note that this heuristic is only valid for locating a single new fire station but could be adapted to the case of several.

Toy example

In order to test the methods on different situations, we adopt the following two models to simulate the emergencies. The observation window is the unit square and there are initially four stations with 15 firemen each, the new station having 20 of them. We simulate 100 independent realizations of each model and keep the points falling in the observation window. Since the purpose here is just to test the optimization, we restrict attention to simple models without interaction. In the first scenario, the locations of the fire stations are the points $(0.2, 0.2)$, $(0.5, 0.6)$, $(0.7, 0.8)$, and $(0.8, 0.6)$, and the emergencies are obtained by drawing 20 i.i.d. points from 3 Gaussian distributions whose means are located at the points $(0.2, 0.3)$, $(0.4, 0.5)$, $(0.7, 0.6)$, and whose standard deviations are respectively 0.07, 0.08 and 0.12. Figure 2.1 represents the positions of the fire stations (in red) and their workload (size of the corresponding bubble) and for one draw the positions of the emergencies and their duration (size of the bubble).

For the naive algorithm, Figures 2.2 (for the first scenario) and 2.5 (for the second) represent at the grid points, from left to right

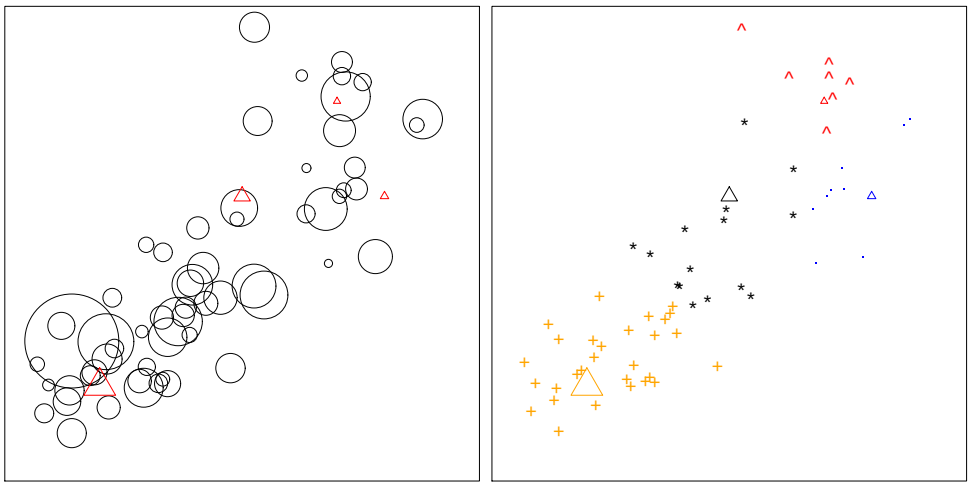

FIGURE 2.1 – Left : Initial locations of fire stations (red triangle) and emergencies (black circle) with workload (proportional size), right : initial allocations to nearest fire station.

- the distance improvement rate ($\frac{Q_7}{Q_6}$),
- the Gini improvement rate ($\frac{C_7}{C_6}$ for C defined by Gini),
- the maximum difference of workload improvement rate ($\frac{C_7}{C_6}$ for C defined by the maximum difference criterion) and
- the objective function based on criterion (2.1).

There is little difference between the graphs of Gini and maximum difference improvement rates on Figure 2.2 which means that one does not loose much by considering only the end points of the distribution of workloads (with a better interpretability). The comparison of the distance criterion and the objective function on Figure 2.3 shows that, for this choice of λ, the objective is mainly driven by the first term.

Table 1 shows the simulation results with, for each of the naive, heuristic and genetic methods, the mean computing time, the mean distance improvement rate, the mean Gini improvement rate, the mean maximum difference of workload improvement rate, and the objective function with standard deviations in parenthesis. The first line of the table shows, for reference, the results of the optimization of the distance alone without the equilibrium part of the objective.

In the second scenario, the locations of the fire stations are the points $(0.6, 0.63)$, $(0.48, 0.08)$, $(0.71, 0.25)$ and $(0.24, 0.62)$, and the emergencies are

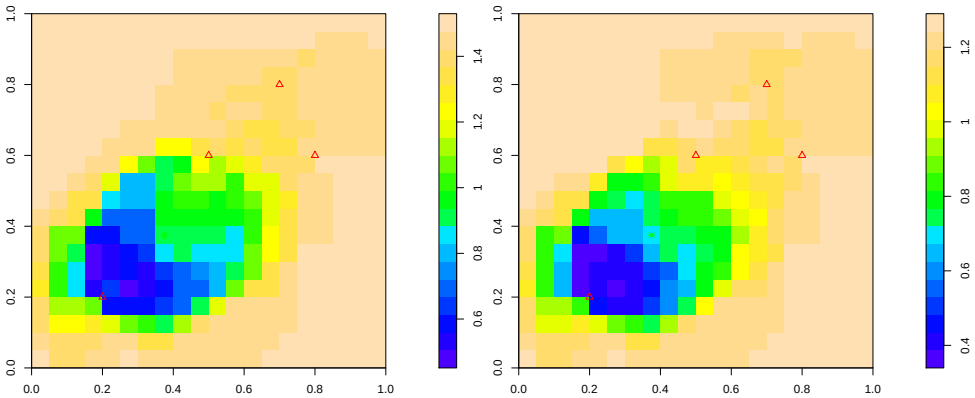

FIGURE 2.2 – Left to right : Gini and maximum difference improvement rates.

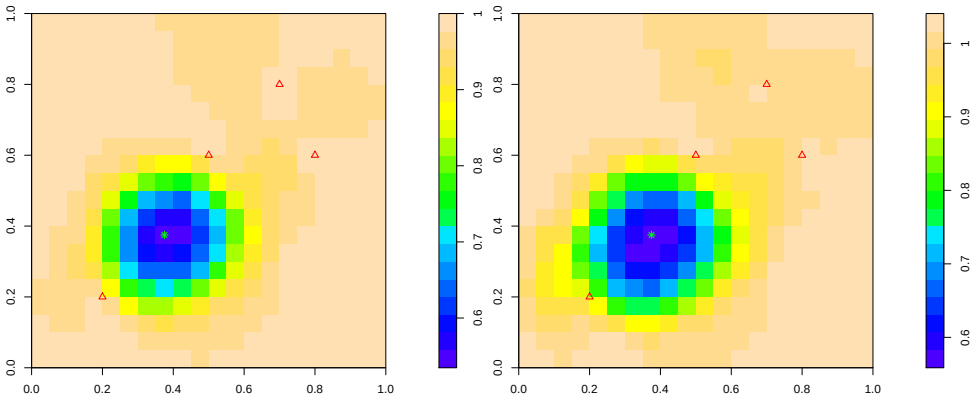

FIGURE 2.3 – Left to right : distance and objective function.

Table 1					
Method	Time (in seconds)	Distance rate	Gini rate	Maxdiff rate	Objective function
Distance alone		0.495	0.923	0.756	
		(0.055)	(0.161)	(0.139)	
Naive	8.569	0.497	0.895	0.732	0.53
	(0.107)	(0.055)	(0.17)	(0.147)	(0.048)
Heuristic	11.012	0.504	0.825	0.661	0.526
	(0.799)	(0.058)	(0.169)	(0.157)	(0.049)
Genetic	15.587	0.493	0.884	0.714	0.524
	(0.349)	(0.055)	(0.164)	(0.149)	(0.049)

obtained by drawing 100 i.i.d. points uniformly in the square. The number of firemen per station is 15 for the old ones and 20 for the new one.

Figure 2.4 represents the positions of the fire stations (in red) and their workload (size of the corresponding bubble) and for one draw the positions of the emergencies and their duration (size of the bubble).

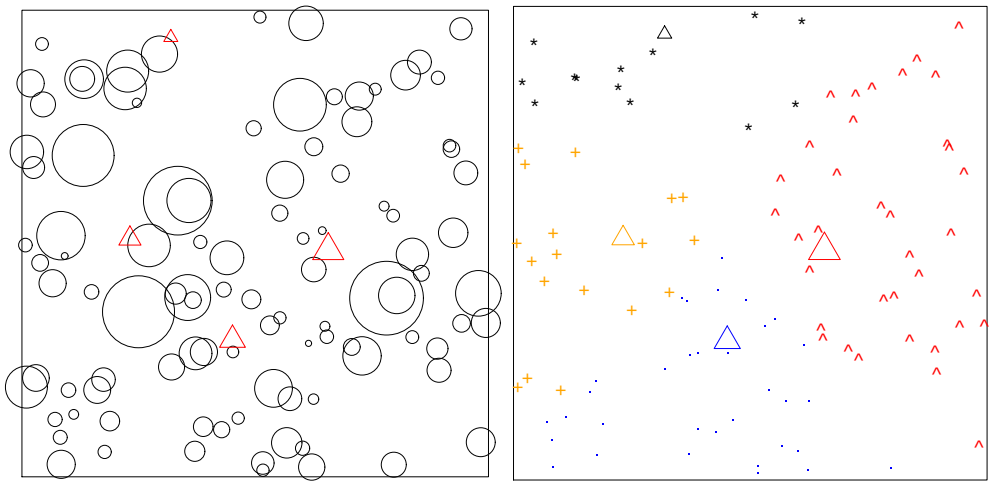

FIGURE 2.4 – Left : Initial locations of fire stations (red triangle) and emergencies (black circle) with workload (proportional size), right : initial allocations to nearest fire station.

Figure 2.5 shows that in the second scenario and for this choice of λ, there is a better equilibrium between the two parts of the objective function.

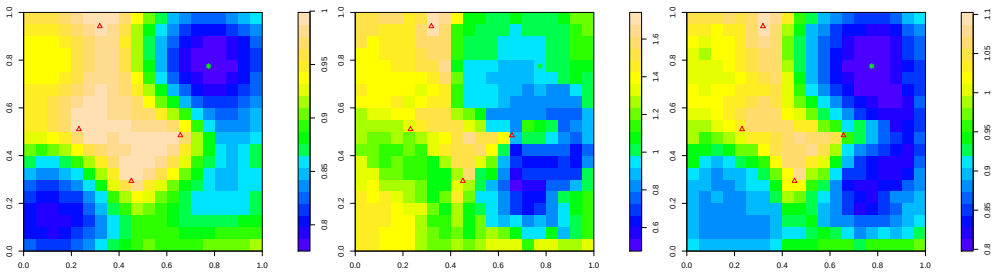

FIGURE 2.5 – Left to right : distance, maximum difference improvement rates and objective function.

Table 2 contains the simulation results for scenario 2 presented as in Table 1.

Overall from Tables 1 and 2, the heuristic method yields the best improvement in the balance of workloads between the facilities, whereas the

Table 2					
Method	Time (in seconds)	Distance rate	Gini rate	Maxdiff rate	Objective function
Distance alone		0.755	1.371	1.113	
		(0.036)	(0.62)	(0.459)	
Naive	8.681	0.775	1.053	0.845	0.785
	(0.113)	(0.053)	(0.489)	(0.361)	(0.065)
Heuristic	11.353	0.796	0.713	0.554	0.762
	(0.628)	(0.061)	(0.435)	(0.326)	(0.056)
Genetic	15.809	0.775	1.019	0.816	0.781
	(0.491)	(0.054)	(0.464)	(0.35)	(0.065)

naive and genetic methods focus on the minimization of the distance rate. Globally, we can say that the heuristic and genetic methods outperform the naive method from the point of view of minimizing the objective function. In the sequel, we test the performance of our methods on our real data set.

2.4.4 Analyzing the results on the firemen data

Figure 2.6 represents the same results as in Figure 2.5 but for the firemen data.

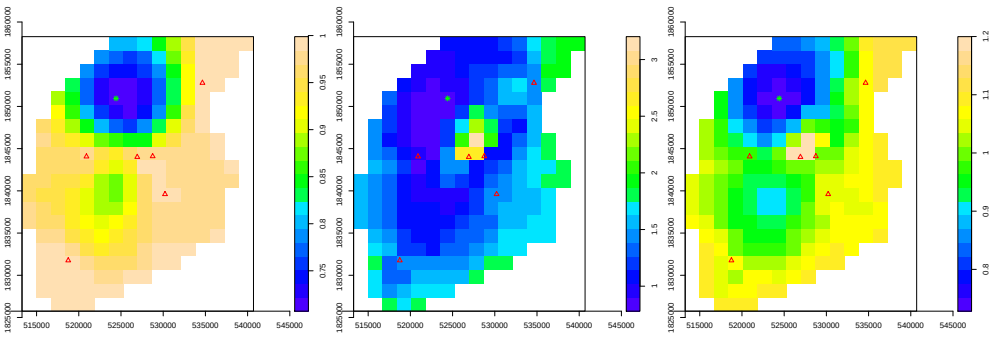

FIGURE 2.6 – Left to right : distance, maximum difference rates and objective function.

Figures 2.7 and 2.8 represent, as in Figure 2.1, the results of the heuristic and genetic methods respectively for the set of emergencies in June. On this example, the optimal locations of the new fire station are nearly identical for the naive and the heuristic method, compared to the location obtained with the genetic method so we omitted the naive method in the figures. On

the other hand, the allocation of emergencies in the heuristic method differs from the other two, particularly near the boundary of the allocation map, notably between the green region and the red one. This discrepancy is also clear in Table 3, and is due to the importance given by the heuristic method to the workloads equilibrium criterion.

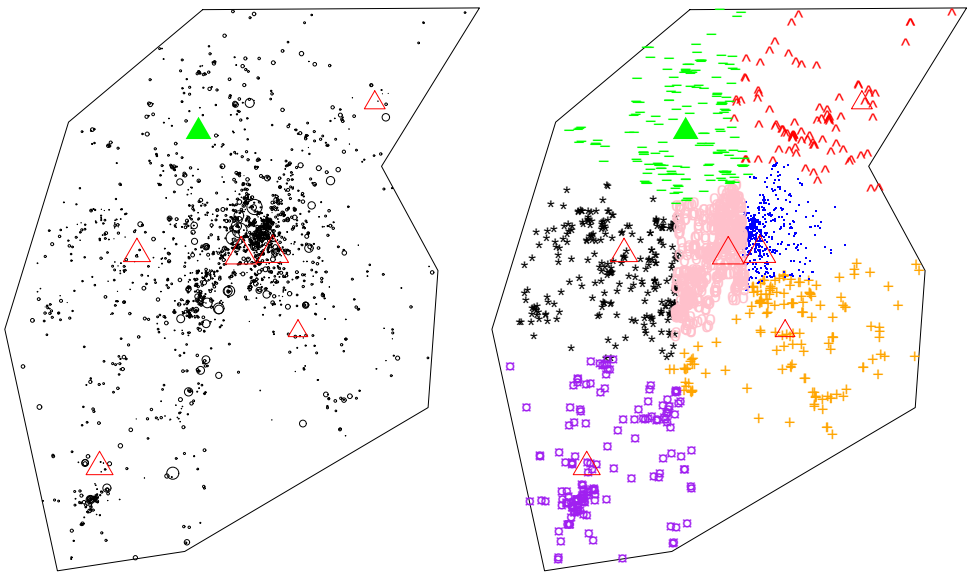

FIGURE 2.7 – Left : Final positions (existing fire station in red triangle, optimal position in green solid triangle), right : final allocations for the heuristic method.

Table 3 shows the results of the optimization methods defined previously on the firemen data set during the month of June. The first line of the table exhibits the results of the optimization of the distance alone without the equilibrium part of the objective. The naive method is the fastest method and achieves the smallest distance improvement rate here. The heuristic method is the only one to search the whole set of allocations without being restricted to the nearest neighbors, which is why it achieves the lowest rate of the difference of workloads criterion (0.583). However, the Gini improvement rates are similar for the heuristic method and the genetic algorithm. Finally, the lowest objective value is obtained with the heuristic method. Consequently, we keep this algorithm for our SPP location-allocation method.

Assuming time stationarity, one could slice the data base into twelve months and consider them as independent repetitions of a same process, thus obtaining 12 repetitions of around 2000 emergencies each. Figure 2.9 shows the density of optimal locations that one can construct by solving the optimization problem for each month separately with the heuristic method.

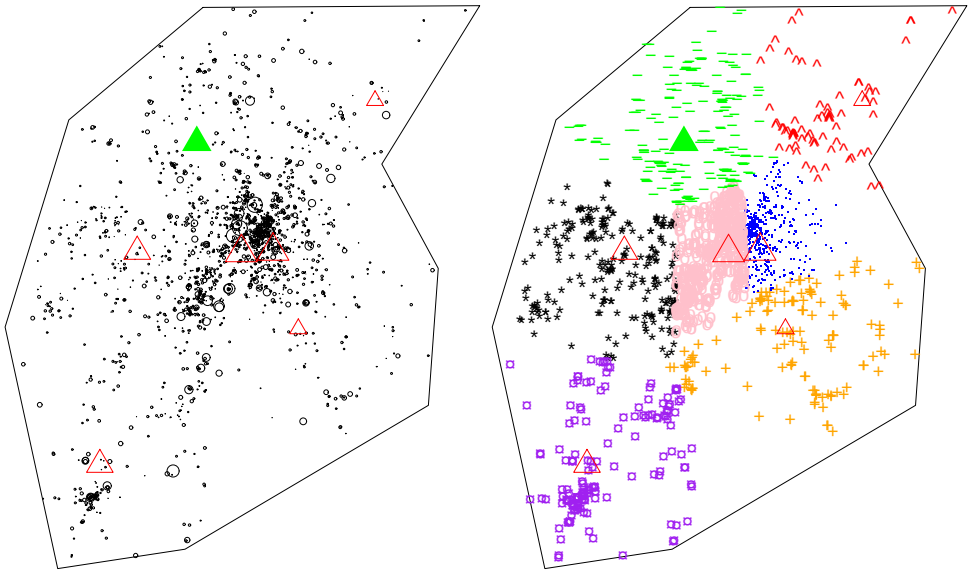

FIGURE 2.8 – Left : Final positions (existing fire station in red triangle, optimal position in green solid triangle), right : final allocations for the genetic method.

Table 3

Method	Time (in seconds)	Distance rate	Gini rate	Maxdiff rate	Objective function
Distance alone		0.707	1.087	0.963	
Naive	222.42	0.708	1.103	0.845	0.727
Heuristic	413.06	0.727	0.808	0.583	0.707
Genetic	505.36	0.712	1.04	0.803	0.725

This number of repetitions is quite low and induces an unreliable density estimation but seems to indicate a unimodal behavior of the distribution of the optimal location.

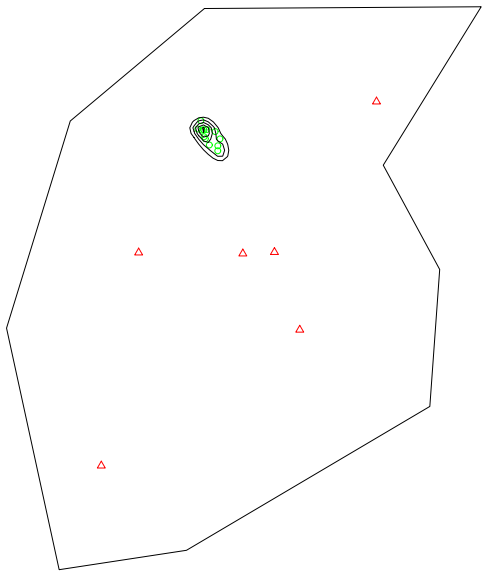

FIGURE 2.9 – Contour plot of the estimated density of optimal locations based on 12 months of data. Existing fire station (red triangle), optimal position (green circle).

Finally, after fitting the two models presented in section 2.4.1, the density of optimal positions obtained with 100 simulations of the fitted inhomogeneous Poisson point process and the fitted LGCP process and optimized with the heuristic method are presented in Figure 2.10. The execution time for the simulation step are respectively 45.95 seconds for the Poisson model and $13,203$ seconds for the LGCP model, whereas the execution time for the optimization step are respectively $11,560$ for the Poisson model and $13,187$ for the LGCP model.

Figure 2.10 clearly shows that the LGCP model implies a high variability in the distribution of the optimal location of the new fire station in comparison with the results obtained with the Poisson model. The variability of the optimal solutions seems to be too large with an LGCP model, in comparison with the results obtained in Figure 10. The representation of the optimal solutions from different simulated realizations from the fitted Poisson model is more informative than the single optimal position in Figure 8. This method allows one to obtain optimal regions, which can be useful when there

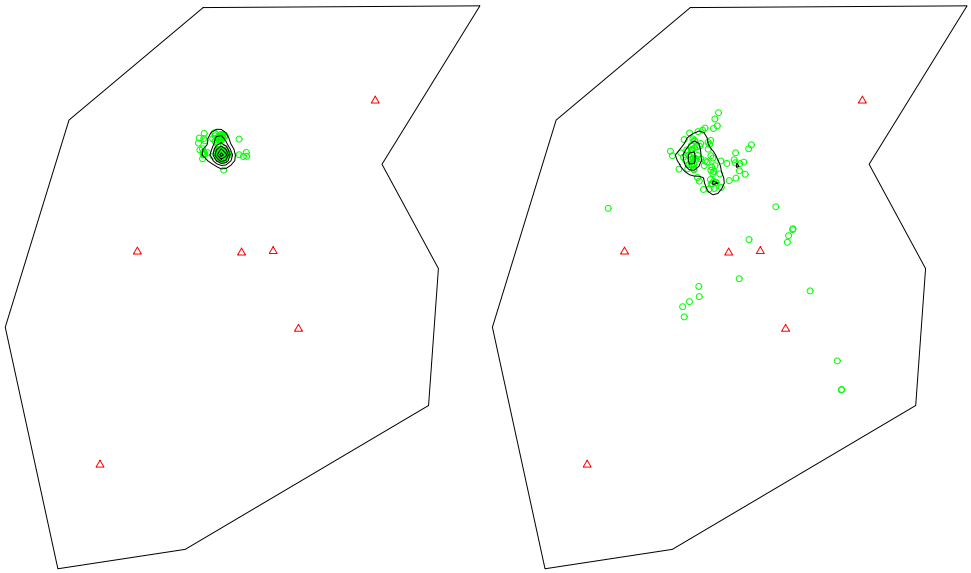

FIGURE 2.10 – Contour plots of the estimated density of optimal locations based on 100 simulations of the Poisson model (left) and LGCP model (right). Existing fire station (red triangle), optimal position (green circle).

are several nearly-optimal solutions. Note that multi-modality may arise as is the case in the right panel of Figure 2.10.

2.5 Conclusions

By fitting a spatial point process to the data we take into account the entire dimension of the natural randomness of this problem. Moreover, this approach allows one to solve larger problems : the statistician summarizes the large data set in a small number of parameters and then controls the sample size when simulating the replications, thus applying the optimization algorithms to smaller-size problems. In addition, the method allows one to capture the spatial variability of the phenomenon and to possibly identify several areas for possible optimal location. The decision maker has to take into account other considerations (quantifiable or not) in the final decision and thus needs the extra information contained in maps like Figure 2.10. From a theoretical point of view, statistical convergence of estimates of the optimal position in such a problem is currently under study. The case of identically distributed positions is treated in Bonneu and Daouia (2008) using a link between optimal position problems and mass transference in the case of

fixed capacity contraints.

Deuxième partie

Propriétés asymptotiques de positions optimales empiriques

Chapitre 3

Consistance de positions empiriques conditionnelles

Sommaire

3.1 Introduction **79**

3.2 Optimal policy specification **81**

3.3 Consistency **82**

3.4 A numerical illustration **83**

3.4.1 Optimization procedure 84

3.4.2 Simulated samples 84

3.5 Appendix : Proofs. **87**

Abstract : We consider the problem of finding the optimal locations of new facilities given the locations of existing facilities and clients. We analyze the general situation where the locations of existing facilities are deterministic while the locations of clients are stochastic with the same unknown marginal distribution. We show how this conditional location-allocation problem can be modeled as a variation of the standard Monge-Kantorovich mass transference problem. We provide a probabilistic formulation of the optimal locations of the new facilities and derive consistent estimators of these theoretical locations from a sample of identically distributed random clients.

3.1 Introduction

In location theory, the standard problem is to find optimal locations of a set of facilities in such a way as to minimize the global cost involved by the allo-

cation of clients to the facilities with prescribed capacity constraints. This is a frequent problem in regional science and economics, which can be modeled as a mass transport problem. In the literature on location research, while the source distribution μ of the population of clients is often assumed to be known, the target distribution ν of facilities is supported on a finite number points to be determined simultaneously with the transport plan so that the cost is minimal (see McAsey and Mou 1998, and the references therein for examples). In probability theory, this problem of finding the location of the support of ν with the optimal allocation map corresponds to optimal coupling of random variables (see *e.g.* Cuesta-Albertos, Matran and Tuero-Diaz 1997, Rachev 1985).

The range of applications where the source measure μ is unknown being wider, we focus in this paper on the estimation of the optimal support of the target measure ν from a sample of random clients drawn from the unknown measure μ. We consider the following general situation : the locations of facilities and clients are supposed to belong to (U, d), a complete separable metric space. We denote by $\mathcal{P}(U)$ the set of all probability measures on U and by $X_1, \ldots, X_n \in U$ a sequence of random locations of n clients which can be dependent or independent. Let $y_1, \ldots, y_J \in U$ be the deterministic locations of J existing facilities and let $q_1, \ldots, q_J \in [0, 1]$ be their respective capacity constraints. Our aim is to handle an optimal policy of the location $y_{J+1} \in U$ of a new facility with known capacity constraint q_{J+1}. For example, the location of a new branch of a management, or public facility, can be searched in a city in order to minimize the transportation of clients under the condition that each branch has a given number of clients. Such a problem of best location policy should be realized in such a way as to minimize the total cost involved by the allocation of clients to the whole set $Y_{J+1} = \{y_1, \ldots, y_{J+1}\}$ of facilities with the prescribed positive masses $q_1, \ldots, q_{J+1}$ such that $\sum_{j=1}^{J+1} q_j = 1$.

In Section 2, we show how this conditional location-allocation problem can be modeled as a variation of the standard Monge-Kantorovich mass transference problem. We introduce empirical versions of the resulting theoretical optimal location based on the sample $\{X_1, \ldots, X_n\}$. Here the optimal location of the new facility is obtained conditionally to the existing ones. When all the facilities positions are unknown, this probabilistic formulation can be found in McAsey and Mou (1998) where the authors show in Theorem 2 the existence of an optimal support for ν. However, in their procedure the source measure μ is assumed to be known, which is not the case in our approach.

In Section 3, we establish the consistency of the constructed estimators

under quite general conditions. To our knowledge, the consistency of empirical optimal locations derived from a Monge-Kantorovich formulation has never been studied. Only the convergence of the total transportation cost has been analyzed in several contexts (see *e.g.* Villani 2003, Rachev and Rüschendorf 1998, Bouchitté, Jimenez and Rajesh 2002). We also extend our approach to the more general setting of finding the optimal locations of $k \geq 1$ new facilities relative to the locations of existing ones. In Section 4, we present a heuristic algorithm to solve the empirical location-allocation problem and we illustrate our procedure through a simulation study to confront theoretical results with empirical behavior. The proofs are reported in the appendix.

3.2 Optimal policy specification

In this section, we deal with the simple case of finding the optimal location y_{J+1} of a new facility ($k = 1$) given the locations of existing facilities and clients. Assuming that y_{J+1} has been found, the problem of optimal allocation of the clients to the fixed set Y_{J+1} of facilities can be modeled by the standard Monge-Kantorovich mass transference problem. Let δ_{y_j} be the point mass concentrated at y_j for $j = 1, \ldots, J+1$. Then the source measure μ and the target measure $\nu(y_{J+1}) = \sum_{j=1}^{J+1} q_j \delta_{y_j}$ describe respectively the mass distribution of the population of clients and that of the set Y_{J+1} of facilities, with equal total weight given by 1. The initial distribution of mass μ is to be transported from the population of clients to the set Y_{J+1} so that the result is the final distribution of mass $\nu(y_{J+1})$. An allocation policy is specified by the choice of a joint distribution Q in the class $\mathcal{P}^{\mu,\nu(y_{J+1})}$ of all laws on $U \times U$ with marginals μ and $\nu(y_{J+1})$. If $c(x, y)$ is a given continuous non-negative function on $U \times U$, interpreted as the cost of transferring the mass from x to y, then the minimal total cost of the allocation of the population of clients to the set Y_{J+1} of facilities is given by the minimum

$$\mathcal{A}_c(\mu, \nu(y_{J+1})) = \inf_{Q \in \mathcal{P}^{\mu,\nu(y_{J+1})}} \int_{U \times U} c(x, y) Q(dx, dy), \tag{3.1}$$

which exists under general conditions on μ and c (see *e.g.* Rachev 1985, Gangbo and McCann 1996). Therefore, the desired optimal conditional location y^*_{J+1} of the new facility can be chosen such that the functional (3.1) is minimal among all possible points y_{J+1} in U , *i.e.*,

$$y^*_{J+1} = \operatorname*{argmin}_{y_{J+1} \in U} \mathcal{A}_c(\mu, \nu(y_{J+1})). \tag{3.2}$$

The existence of (3.2) could be derived under fairly general conditions on μ and c by adapting for example Theorem 2 of McAsey and Mou (1998) to our conditional setup.

The statistical problem is now to estimate the theoretical optimal location y^*_{J+1} from the sample of random locations of clients $X_1, \ldots, X_n$. A natural estimator is given by replacing the unknown source distribution of mass μ in (3.2) with the empirical measure $\mu_n = (1/n)\sum_{i=1}^n \delta_{X_i}$, that is

$$\hat{y}^*_{J+1} = \operatorname*{argmin}_{y_{J+1}\in U} \mathcal{A}_c(\mu_n, \nu(y_{J+1})). \tag{3.3}$$

Replacing μ with μ_n in $\mathcal{A}_c(\mu, \nu(y_{J+1}))$ yields the discrete version of the Monge-Kantorovich problem (see *e.g.* Rachev and Rüschendorf 1998 ; Cuesta-Albertos *et al.* 1996, and the references therein). We will not explicit here the technical conditions which ensure the existence of the estimate (3.3). In what follows, we assume that both values (3.2) and (3.3) exist.

3.3 Consistency

We first prove that the estimator $\hat{y}^*_{J+1}$ is strongly consistent under the standard condition in M-estimation that y^*_{J+1} should be a well-separated point of minimum of the map : $y_{J+1} \mapsto \mathcal{A}_c(\mu, \nu(y_{J+1}))$, that is

$$\inf\{\mathcal{A}_c(\mu, \nu(y_{J+1})) : \ y_{J+1} \in U, \ d(y_{J+1}, y^*_{J+1}) > \varepsilon\} > \mathcal{A}_c(\mu, \nu(y^*_{J+1})) \tag{3.4}$$

for every $\varepsilon > 0$. We also need the cost function c to be in the class of functions :

$$c(\cdot,\cdot) = d(\cdot,\cdot)^p \text{ such that } p > 0 \text{ and } \int_U c(x,a)d\mu(x) < \infty \text{ for any point } a \in U. \tag{3.5}$$

Theorem 3.1. *Let the conditions (3.4) and (3.5) hold. Then for any sequence of estimators $\hat{y}^*_{J+1}$ satisfying (3.3), we have*
$d(\hat{y}^*_{J+1}, y^*_{J+1}) \xrightarrow{a.s.} 0$ *as* $n \to \infty$.

Condition (3.5) is only needed to guaranty the uniform almost sure convergence of $\mathcal{A}_c(\mu_n, \nu(\cdot))$ to $\mathcal{A}_c(\mu, \nu(\cdot))$. Theorem 3.1 can then be extended to any class of cost functions $c(\cdot,\cdot)$ for which this uniform convergence holds.

We also obtain the weak consistency for estimators $\hat{y}^*_{J+1}$ that nearly minimize $\mathcal{A}_p(\mu_n, \nu(\cdot))$, *i.e.*,

$$\mathcal{A}_p(\mu_n, \nu(\hat{y}^*_{J+1})) \le \inf_{y_{J+1}\in U} \mathcal{A}_p(\mu_n, \nu(y_{J+1})) + o_P(1) \tag{3.6}$$

with $\mathcal{A}_p(\cdot,\cdot) = [\mathcal{A}_c(\cdot,\cdot)]^{\min(1,1/p)}$ being the Wasserstein distance.

Theorem 3.2. *Under the conditions of Theorem 3.1, we have* $d(\hat{y}^*_{J+1}, y^*_{J+1}) \xrightarrow{P} 0$ *as* $n \to \infty$, *for any sequence of estimators* $\hat{y}^*_{J+1}$ *satisfying* (3.6).

A general problem can be stated as follows. Given J existing facilities with deterministic locations $y_1, \ldots, y_J$ in U, a population of clients drawn from a probability measure $\mu \in \mathcal{P}(U)$, prescribed positive masses $q_1, \ldots, q_{J+k}$ with $q_1 + \cdots + q_{J+k} = 1$, find locations $y_{J+1}, \ldots, y_{J+k} \in U$ of $k \geq 1$ new facilities such that the minimal total cost of allocation of the population of clients to the set of $J + k$ facilities $\{y_1, \ldots, y_{J+k}\}$, given by

$$\mathcal{A}_c[\mu, \nu(y_{J+1}, \ldots, y_{J+k})] = \inf_{Q \in \mathcal{P}^{\mu, \nu(y_{J+1}, \ldots, y_{J+k})}} \int_{U \times U} c(x, y) Q(dx, dy),$$

is minimal among all locations $y_{J+1}, \ldots, y_{J+k}$ in U, where $\nu(y_{J+1}, \ldots, y_{J+k}) = \sum_{j=1}^{J+k} q_j \delta_{y_j}$ is the target measure and $\mathcal{P}^{\mu, \nu(y_{J+1}, \ldots, y_{J+k})}$ is the class of all probability measures on $U \times U$ with marginals μ and $\nu(y_{J+1}, \ldots, y_{J+k})$. The resulting theoretical optimal locations

$$(y^*_{J+1}, \ldots, y^*_{J+k}) = \operatorname*{argmin}_{(y_{J+1}, \ldots, y_{J+k}) \in U^k} \mathcal{A}_c[\mu, \nu(y_{J+1}, \ldots, y_{J+k})],$$

can be estimated from a random sample of locations of n clients $\{X_1, \ldots, X_n\}$ drawn from the unknown source measure μ, by

$$(\hat{y}^*_{J+1}, \ldots, \hat{y}^*_{J+k}) = \operatorname*{argmin}_{(y_{J+1}, \ldots, y_{J+k}) \in U^k} \mathcal{A}_c[\mu_n, \nu(y_{J+1}, \ldots, y_{J+k})].$$

Then the consistency of each estimate $\hat{y}^*_{J+\ell}$ with $\ell = 1, \ldots, k$, can be easily derived by modifying the condition (3.4) and adapting the proofs of Theorems 3.1-3.2. Indeed, if (3.5) holds with

$$\begin{aligned} \inf\{\mathcal{A}_c[\mu, \nu(y_{J+1}, \ldots, y_{J+k})] : \ & y_{J+1}, \ldots, y_{J+k} \in U, \ d(y_{J+\ell}, y^*_{J+\ell}) > \varepsilon\} \\ & > \mathcal{A}_c[\mu, \nu(y^*_{J+1}, \ldots, y^*_{J+k})] \quad \text{for every } \varepsilon > 0, \end{aligned}$$

then it is easy to see that $d(\hat{y}^*_{J+\ell}, y^*_{J+\ell}) \xrightarrow{a.s.} 0$. Likewise $d(\hat{y}^*_{J+\ell}, y^*_{J+\ell}) \xrightarrow{P} 0$ when $(\hat{y}^*_{J+1}, \ldots, \hat{y}^*_{J+k})$ nearly minimizes $\mathcal{A}_p(\mu_n, \nu(\cdot))$, *i.e.*,

$$\mathcal{A}_p[\mu_n, \nu(\hat{y}^*_{J+1}, \ldots, \hat{y}^*_{J+k})] \leq \inf_{(y_{J+1}, \ldots, y_{J+k}) \in U^k} \mathcal{A}_p[\mu_n, \nu(y_{J+1}, \ldots, y_{J+k})] + o_P(1).$$

3.4 A numerical illustration

Our simulation experiments illustrate how the convergence results work out in practice.

3.4.1 Optimization procedure

The determination of the optimal location (3.2) cannot be separated from the determination of an optimal allocation (3.1). Our optimization method is based on an assignment algorithm introduced by Jonker and Volgenant (1987) named LAPJV, that we have implemented in the R software. This augmenting path algorithm has a good and stable average performance from the point of view of computational time as it is shown in the state of the art of algorithms for assignment problems by Dell'Amico and Toth (2000). In order to use it, we transform our semi-assignment problem into an assignment problem by "duplicating" the columns of the initial cost matrix of size $n \times (J+1)$ into a matrix of size $n \times n$, taking into account the capacity constraints. This transformation induces an important computational time in the resolution step. Many possibilities exist to reduce this computational time by using the semiLAPJV algorithm or a version for sparse matrix named semiLAPMod (Volgenant 1996), which are specifically built for semi-assignment problems, and by making the implementation in the C++ language.

Our 2-step procedure to compute $\hat{y}^*_{J+1}$: In the first step, we introduce a regular grid whose nodes represent a set of feasible locations for the new facility. For each considered node, we search the optimal allocation with the LAPJV algorithm and compute the resulting optimal total cost. At this step, the best location for the new facility, among the evaluated nodes, corresponds to the position associated with the minimal total cost. The accuracy of the optimal location at this stage depends on the size of the grid but we do not know whether this dependence is linear, exponential... As often, the user have to make a compromise between the accuracy of the solution and the computational time. The second step is optional because it depends on the considered cost function. Since we know the clients allocated to the new facility, we can compute for certain cost functions an optimal location for the new facility. For example, in the case of a cost function derived from the Euclidean distance, the optimal location corresponds to the centroïd of the locations of clients which are allocated to it. But in the case of a cost function derived from the ℓ_1 distance, the optimal location of the new facility, when we know the clients allocated to it, is the solution to the Fermat-Weber problem which is not easy to approximate by algorithms.

3.4.2 Simulated samples

Our toy example derives from a set of optimal configurations of 2 to 11 centers of production, in a unit square domain in $\mathbb{R}^2$, with uniformly distributed

consumers, introduced in Bolton & Morgan (2002). In their paper, the aim is to locate simultaneously several facilities. Our framework is different in the sense that we want to locate a new facility conditionally to the existing ones.

We consider in Figure 3.1 (Left) the configuration with 4 centers with the location $y^*_{J+1} = (0.75, 0.75)$ being the one to be estimated. Here, the fixed $J = 3$ existing facilities are represented by the triangles and the known theoretical location of the new facility is given by the green circle. We choose the cost function $c(\cdot,\cdot) := d(\cdot,\cdot)^2$, with d being the Euclidean distance, and we use equal capacity constraints $(0.25, 0.25, 0.25, 0.25)$. For a simulated sample of small size, $n = 40$ clients, the obtained empirical optimal location $\hat{y}^*_{J+1}$ is displayed in Figure 3.2 (blue circle). Here we use the same color to indicate the clients allocated to each facility. Figure 3.1 (Right) shows the value of the objective function at each node when it is chosen as the location of the new facility.

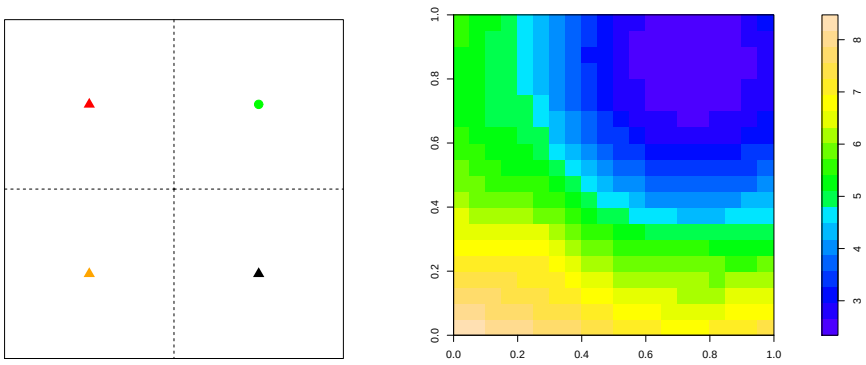

FIGURE 3.1 – Left — existing facilities (triangles) and theoretical new one (green circle), with Voronoï tessellation, under a uniform distribution of clients. Right — values of the optimal total cost with the new facility located at a node of a regular grid of points (20×20).

In Table 3.1, we compute the sample mean and standard deviation of the distance between the estimated optimal locations and the theoretical one, with different values of $n = 40, 100, 200$, for 100 iterations of each configuration. As expected, the empirical optimal location of the new facility is all the more closer to the true theoretical optimal location as the number of clients increases.

Figure 3.3 represents the contour lines of the density estimates of the optimal location, with different numbers of clients in the upper-right square of the initial domain. It is difficult in this particular case to judge whether

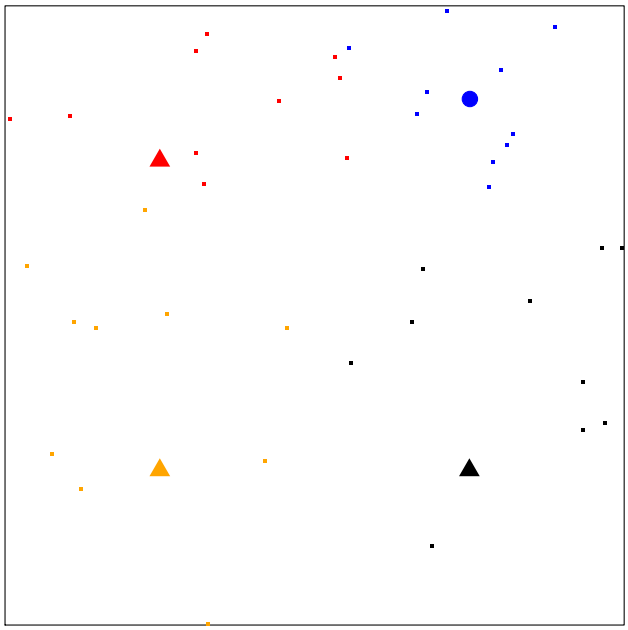

FIGURE 3.2 – Estimated optimal location (blue circle) and its corresponding allocations (blue data points) with 40 i.i.d. simulated clients.

n	40 clients	100 clients	200 clients
mean	0.018	0.004	0.002
std	0.051	0.004	0.002

TABLE 3.1 – Accuracy of the empirical optimal location evaluated with 100 simulations, with different numbers of clients, represented by the sample mean and standard deviation.

the asymptotic law of the empirical optimal location is normal. It would be then interesting to investigate the precise asymptotic distribution of $\hat{y}^*_{J+1}$. Another important topic of interest for future research is the study of the case where the random locations of clients $X_1, \ldots, X_n$ are not identically distributed. The range of applications in this case is wider.

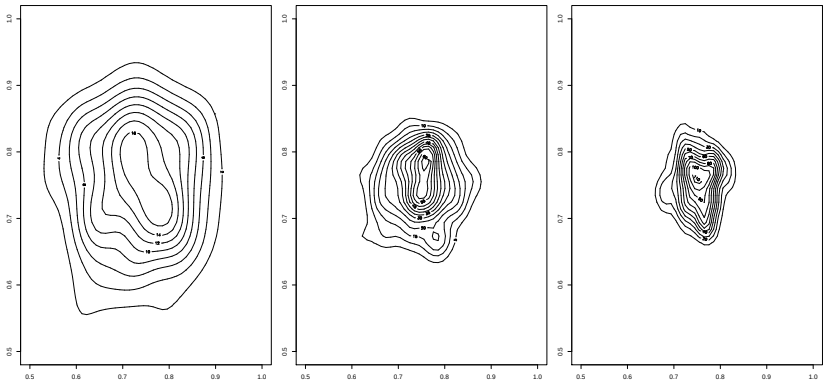

FIGURE 3.3 – Contour lines of the estimated densities of the optimal location for the configurations corresponding respectively to $n = 40, 100, 200$ clients considered in the square $(0.5, 1) \times (0.5, 1)$.

3.5 Appendix : Proofs.

Proof of Theorem 3.1 Let $\mathcal{A}_p(Q_1, Q_2) = \inf_{Q \in \mathcal{P}^{Q_1,Q_2}} [\int_{U \times U} d(x,y)^p Q(dx,dy)]^{p'}$ with $p' = \min(1, 1/p)$. Then we have

$$y^*_{J+1} = \operatorname*{argmin}_{y_{J+1} \in U} \mathcal{A}_p(\mu, \nu(y_{J+1})) \quad \text{and} \quad \hat{y}^*_{J+1} = \operatorname*{argmin}_{y_{J+1} \in U} \mathcal{A}_p(\mu_n, \nu(y_{J+1})).$$

Moreover $\mathcal{A}_p(\cdot, \cdot)$ is a metric on the space $\mathcal{P}_p(U)$ of probability measures on U with finite moment of order p according to Villani (2003, Theorem 7.3, p.207). Since $\mu, \mu_n \in \mathcal{P}_p(U)$ and $\nu(y_{J+1}) \in \mathcal{P}_p(U)$ for all $y_{J+1} \in U$, we obtain $\sup_{y_{J+1} \in U} |\mathcal{A}_p(\mu, \nu(y_{J+1})) - \mathcal{A}_p(\mu_n, \nu(y_{J+1}))| \leq \mathcal{A}_p(\mu_n, \mu)$ by the triangle inequality. On the other hand, by the generalized Glivenko-Cantelli-Varadarajan Theorem (see Rachev 1991, Corollary 11.1.2, p.215), we have

$$\mathcal{A}_p(\mu_n, \mu) = [\mathcal{A}_c(\mu_n, \mu)]^{p'} \xrightarrow{a.s.} 0 \quad \text{as} \quad n \to \infty.$$

Whence

$$\sup_{y_{J+1} \in U} |\mathcal{A}_p(\mu, \nu(y_{J+1})) - \mathcal{A}_p(\mu_n, \nu(y_{J+1}))| \xrightarrow{a.s.} 0 \quad \text{as} \quad n \to \infty. \tag{A.1}$$

It follows that

$$W_n := \mathcal{A}_p(\mu_n, \nu[y^*_{J+1}]) - \mathcal{A}_p(\mu, \nu[y^*_{J+1}]) \xrightarrow{a.s.} 0 \quad \text{as} \quad n \to \infty. \tag{A.2}$$

From now on let $\hat{y}^*_{J+1}(n) := \hat{y}^*_{J+1}$ in (3.3). Since $\mathcal{A}_p(\mu_n, \nu[\hat{y}^*_{J+1}(n)]) \leq \mathcal{A}_p(\mu_n, \nu[y^*_{J+1}])$ by definition (3.3) of $\hat{y}^*_{J+1}(n)$, we obtain $\mathcal{A}_p(\mu_n, \nu[\hat{y}^*_{J+1}(n)]) \leq \mathcal{A}_p(\mu, \nu[y^*_{J+1}]) + W_n$. Whence

$$\begin{aligned} 0 \leq \mathcal{A}_p(\mu, \nu[\hat{y}^*_{J+1}(n)]) - \mathcal{A}_p(\mu, \nu[y^*_{J+1}]) \\ &\leq \mathcal{A}_p(\mu, \nu[\hat{y}^*_{J+1}(n)]) - \mathcal{A}_p(\mu_n, \nu[\hat{y}^*_{J+1}(n)]) + W_n \\ &\leq \sup_{y_{J+1} \in U} |\mathcal{A}_p(\mu, \nu(y_{J+1})) - \mathcal{A}_p(\mu_n, \nu(y_{J+1}))| + W_n. \end{aligned}$$

It follows from (A.1) and (A.2) that $\mathcal{A}_p(\mu, \nu[\hat{y}^*_{J+1}(n)]) - \mathcal{A}_p(\mu, \nu[y^*_{J+1}]) \xrightarrow{a.s.} 0$ as $n \to \infty$, which is equivalent to say that

$$\lim_{n \to \infty} P[|\mathcal{A}_p(\mu, \nu[\hat{y}^*_{J+1}(m)]) - \mathcal{A}_p(\mu, \nu[y^*_{J+1}])| \leq \eta, \ \forall m \geq n] = 1 \tag{A.3}$$

for each $\eta > 0$ (see, *e.g.*, Serfling 1980, p. 6, for this equivalent condition of the almost sure convergence). Now in order to prove $d(\hat{y}^*_{J+1}(n), y^*_{J+1}) \xrightarrow{a.s.} 0$ as $n \to \infty$, it suffices to show

$$\lim_{n \to \infty} P[d(\hat{y}^*_{J+1}(m), y^*_{J+1}) \leq \varepsilon, \ \forall m \geq n] = 1, \tag{A.4}$$

for each $\varepsilon > 0$. Let $\varepsilon > 0$. By (3.4) there exists $\eta > 0$ such that

$$\inf\{\mathcal{A}_p(\mu, \nu(y_{J+1})) : \ y_{J+1} \in U, \ d(y_{J+1}, y^*_{J+1}) > \varepsilon\} - \mathcal{A}_p(\mu, \nu(y^*_{J+1})) > \eta. \tag{A.5}$$

Then, for each $m \geq 1$, the event $\{d(\hat{y}^*_{J+1}(m), y^*_{J+1}) > \varepsilon\}$ implies $\{\mathcal{A}_p(\mu, \nu[\hat{y}^*_{J+1}(m)]) > \mathcal{A}_p(\mu, \nu[y^*_{J+1}]) + \eta\}$. This is equivalent to say that $\{|\mathcal{A}_p(\mu, \nu[\hat{y}^*_{J+1}(m)]) - \mathcal{A}_p(\mu, \nu[y^*_{J+1}])| \leq \eta\}$ is contained in the event $\{d(\hat{y}^*_{J+1}(m), y^*_{J+1}) \leq \varepsilon\}$, for all $m \geq 1$. Therefore, for all n (large enough), the event $\{|\mathcal{A}_p(\mu, \nu[\hat{y}^*_{J+1}(m)]) - \mathcal{A}_p(\mu, \nu[y^*_{J+1}])| \leq \eta, \ \forall m \geq n\}$ is contained in $\{d(\hat{y}^*_{J+1}(m), y^*_{J+1}) \leq \varepsilon, \ \forall m \geq n\}$. Thus (A.4) follows immediately from (A.3). $\square$

Proof of Theorem 3.2 Since $\mathcal{A}_p(\mu_n, \nu[\hat{y}^*_{J+1}]) \leq \mathcal{A}_p(\mu_n, \nu[y^*_{J+1}]) + o_P(1)$ by definition (3.6) of $\hat{y}^*_{J+1}$, we have $\mathcal{A}_p(\mu_n, \nu[\hat{y}^*_{J+1}]) \leq \mathcal{A}_p(\mu, \nu[y^*_{J+1}]) + W_n + o_P(1)$. Whence

$$\begin{aligned} 0 \leq \mathcal{A}_p(\mu, \nu[\hat{y}^*_{J+1}]) - \mathcal{A}_p(\mu, \nu[y^*_{J+1}]) \\ &\leq \sup_{y_{J+1} \in U} |\mathcal{A}_p(\mu, \nu(y_{J+1})) - \mathcal{A}_p(\mu_n, \nu(y_{J+1}))| + W_n + o_P(1). \end{aligned}$$

Then $\mathcal{A}_p(\mu, \nu[\hat{y}^*_{J+1}]) - \mathcal{A}_p(\mu, \nu[y^*_{J+1}]) \xrightarrow{P} 0$ by (A.1) and (A.2). It follows from (A.5) that

$$\lim_{n\to\infty} P[d(\hat{y}^*_{J+1}, y^*_{J+1}) > \varepsilon] \leq \lim_{n\to\infty} P[|\mathcal{A}_p(\mu, \nu[\hat{y}^*_{J+1}]) - \mathcal{A}_p(\mu, \nu[y^*_{J+1}])| > \eta] = 0$$

for every $\varepsilon > 0$, which ends the proof. $\square$

Chapitre 4

Estimation de positions optimales avec contraintes

Sommaire

4.1 Introduction **91**

4.2 Main result **94**

4.3 Appendix : Lemmas and proof. **95**

Abstract : The search of the optimal localization of a new facility often depends on the existing ones and an optimal allocation map of clients to them. We consider this location-allocation problem when characteristics of the population of clients are available. The criterion function to minimize is then the combination of a transportation cost and an equilibrium measure of workloads between the facilities, and it is natural to try to incorporate this information into the estimation procedure. We provide quite general and natural sufficient conditions for the strong consistency of nonparametric estimators of this class of optimal locations where the criterion function to minimize is unsmooth and depends on an infinite dimensional parameter. We investigate in this paper the setting where the pairs of clients and their workload marks are supposed independent and identically distributed.

Key words : Location-allocation, Strong consistency, Vapnik-Cervonenkis classes.

4.1 Introduction

The search of the optimal localization of a new facility often depends on the existing ones and an optimal allocation map of clients to them. The optimal

location represents the parameter to estimate whereas the allocation map of clients to facilities is a tool parameter in our setup. The optimal location of the source facility can be written as the argument-minimum of a criterion function based on an econometric model with an unknown distribution of positions and characteristics of clients. However, because the theoretical optimization problem is intractable or difficult to evaluate, we search optimal location estimators which are solutions of empirical optimization problems.

We can distinguish between two somewhat related estimation procedures to our approach : M- and Z-estimation. In M-estimation problems, the solution parameter is the argument-minimum (or argument-maximum) of a criteron function, whereas in Z-estimation, the solution parameter is the unique solution where the criterion function is null. A particular case of M- and Z-estimation closely related to our approach is the Generalized Method of Moments (GGM), a popular method in econometrics. In GGM the criterion functions have the form of a theoretical moment and a sample moment, whereas the objective function in our formulation is expressed as the optimum over a functional space of a function of moments. The existing theories allow either for unsmooth criterion functions of finite dimensional parameters (e.g., Pakes and Pollard 1989) or smooth objective functions of both finite and infinite dimensional parameters (e.g., Bickel, Klaassen, Ritov and Wellner 1993) or unsmooth criterion functions with simultaneously finite and infinite dimensional parameters (e.g., Chen, Linton and Van Keilegom 2003). To our knowledge, there exists no M-estimation formulation where the objective function of a finite parameter is written as an infinimum function taken over a functional space. We explore this estimation problem through the case study presented in Bonneu and Thomas-Agnan (2008) and follow very closely their notations.

We consider the location-allocation problem which consists in locating a new fire station, conditionally to the existing ones, by minimizing the total access time of firemen emergencies, and reaching a relative equilibrium of firemen workload. The originality in Bonneu and Thomas-Agnan (2008) is to take into account the random nature of the characteristics through the modeling of a spatial point process and to solve a family of optimization problems for several simulated observations from the fitted model. Despite an elegant statistical formulation of their location-allocation problem is available, no attention however was devoted to theoretical bases. In this paper, we investigate the asymptotic properties of this new statistical technique in order to establish the integrity of the optimal location estimators, in the sense that when a large sample is available the method should not be grossly

inefficient.

Let n be the total number of emergencies, $\{X_1, \ldots, X_n\}$ be their locations in a bounded subset[1] $\Omega \subset \mathbb{R}^2$, endowed by a metric d, and $\{W_1, \ldots, W_n\}$ be their associated workload marks. Suppose the locations $y_1, \ldots, y_J$ of J existing fire stations and their number of firemen $z_1, \ldots, z_J$ are available. We denote by y_{J+1} and z_{J+1} the unknown location of the new fire station and its fixed number of firemen. The optimal location of the new fire station is linked to an optimal allocation of emergencies to fire stations. Consequently, the search of a location minimizing the cost function is done over a set Λ of measurable allocations α from Ω to the set of fire stations' indexes $\{1, \ldots, J+1\}$. The theoretical optimal location can be expressed as

$$y^*_{J+1} = \operatorname*{argmin}_{y_{J+1} \in \Omega} \inf_{\alpha \in \Lambda} M(y_{J+1}, \alpha),$$

and its empirical counterpart as

$$\hat{y}_{J+1,n} = \operatorname*{argmin}_{y_{J+1} \in \Omega} \inf_{\alpha \in \Lambda} M_n(y_{J+1}, \alpha), \tag{A.1}$$

where $M(y_{J+1}, \alpha) = \mathbb{E}(\|X - y_{\alpha(X)}\|^2) + \lambda \Phi(\alpha)$ and $M_n(y_{J+1}, \alpha) = \frac{1}{n} \sum_{i=1}^n \|X_i - y_{\alpha(X_i)}\|^2 + \lambda \Phi_n(\alpha)$, with

$$\begin{aligned}
\Phi(\alpha) &= \sum_{j=1}^{J+1} \frac{\mathbb{E}(h_{j,\alpha}(X,W))}{\mathbb{E}(W)} \log \frac{\mathbb{E}(h_{j,\alpha}(X,W))}{z_j \mathbb{E}(W)}, \\
\Phi_n(\alpha) &= \sum_{j=1}^{J+1} \frac{\sum_{i=1}^n h_{j,\alpha}(X_i,W_i))}{\sum_{i=1}^n W_i} \log \frac{\sum_{i=1}^n h_{j,\alpha}(X_i,W_i)}{z_j \sum_{i=1}^n W_i},
\end{aligned}$$

and $h_{j,\alpha}(x,w) = w \mathbb{1}_{\alpha^{-1}(j)}(x)$. The objective function balances a transportation cost and an equilibrium measure of workloads to find an optimal location of the new fire station. The regularization parameter λ is used to give more or less importance at each term of the objective function and is fixed in advance. See Bonneu and Thomas-Agnan (2008) for a discussion on the choice of this regularization parameter. The equilibrium criteria $\Phi_n(\alpha)$ and $\Phi(\alpha)$, based on the entropy function, are minima when all fire stations have identical workloads. In this difficult context, we focus in Section 4.2 on the particular setting where the positions and marks are independent and identicaly distributed, showing that the built estimator $\hat{y}_{J+1,n}$ converges to the optimal location y^*_{J+1} with probability one.

[1] Our convergence results are also valid if we consider a bounded subset Ω in any complete separable metric space $(\mathcal{X}, d)$.

4.2 Main result

We suppose here that the pairs of locations and workload marks $(X_1, W_1), \ldots, (X_n, W_n)$ are independent copies sampled from (X, W) with positive distribution measure π on $\Omega \times \Psi$, where Ψ is a bounded subset of $(0, \infty)$. We denote by μ the marginal distribution of X and by ν the marginal distribution of W. We use the notations μ_n and ν_n respectively for the empirical measures of $(X_1, \ldots, X_n)$ and $(W_1, \ldots, W_n)$. The following conditions will be needed to prove the almost sure convergence of $\hat{y}_{J+1,n}$ to y^*_{J+1}.

(H1) $\inf_{k \in \{1,\ldots,J+1\}} \inf_{\alpha \in \Lambda} \mathbb{E}[W \mathbb{I}_{\alpha^{-1}(k)}(X)] > 0.$

(H2) $\inf_{y: d(y, y^*_{J+1}) \geq \varepsilon} \inf_{\alpha \in \Lambda} M(y, \alpha) > \inf_{\alpha \in \Lambda} M(y^*_{J+1}, \alpha)$, for all $\varepsilon > 0$.

(H3) $\{\alpha^{-1}(j); \alpha \in \Lambda\}$ is a Vapnik-Chervonenkis class of sets, for each $j = 1, \ldots, J+1$.

Assumption (H1) is quite natural since it supposes that in expectancy the minimum workload fire station is positive. (H2) is a standard condition in M-estimation which assumes that y^*_{J+1} is a well-separated point of minimum of the functional $\inf_{\alpha \in \Lambda} M(\cdot, \alpha)$. Assumption (H3) gives a quite natural and general condition to prove the almost sure convergence $\inf_{\alpha \in \Lambda} M_n(y_{J+1}, \alpha) \xrightarrow{a.s.} \inf_{\alpha \in \Lambda} M(y_{J+1}, \alpha)$, uniformly in $y_{J+1} \in \Omega$. In Appendix we assume that the reader is familiar with the theory of empirical processes in M-estimation involving VC classes (see, *e.g.*, van der Vaart (1998, Section 19) or van de Geer (2000, Section 3)). The following examples present two well-known methods in location-allocation literature which satisfy (H3), but there exist many other examples.

Example 1. The set Λ of allocations α such that $\{\alpha^{-1}(1), \ldots, \alpha^{-1}(J+1)\}$ forms a Voronoi partition satisfies the condition (H3).

Example 2. Let $\{A_1, \ldots, A_k\}$ be a partition of the population of clients Ω, where the number k of allocation zones is at least equal to the number $J+1$ of facilities. Consider the mode of allocation for which each zone A_l in Ω should be allocated to a unique facility y_j, and each facility should have at least one allocation, that is the set of allocation maps $\alpha \in \Lambda$ such that

$$\forall l = 1, \ldots, k, \ \exists! j = 1, \ldots, J+1 \quad \text{s.t.} \quad A_l \subset \alpha^{-1}(j)$$

with $\alpha^{-1}(j) \neq \emptyset \ \forall j = 1, \ldots, J+1$.

This zoning system which allows to view the location-allocaton problem as a matter of allocating building blocks, satisfies Assumption (H3). This method is well-known in location theory where the blocks are often delimited by a grid with a finite number of nodes. However, every kind of partition of the domain can be considered and the number of blocks can be as large as desired.

Theorem 4.1. *For any sequence of estimators $\hat{y}_{J+1,n}$ satisfying* (A.1)*, we have*

$$\hat{y}_{J+1,n} \xrightarrow{a.s.} y^*_{J+1} \quad as \quad n \to \infty$$

provided the conditions (H1)-(H3) hold.

4.3 Appendix : Lemmas and proof.

We need the following lemma which extends the weak consistency of M-estimators (see *e.g.* van der Vaart 1998, Theorem 5.7) to the almost sure sense.

Lemma 4.1. *Let Θ be a metric space and $\theta^* = \operatorname{argmin}_{\theta\in\Theta} M(\theta)$, where M is a nonrandom real-valued function defined on Θ. Let $\hat{\theta}_n = \operatorname{argmin}_{\theta\in\Theta} M_n(\theta)$, where M_n is a real-valued function depending on a random sample of size n. If,*

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{a.s.} 0 \quad as \quad n \to \infty, \tag{A.1}$$

$$\inf_{\theta: d(\theta,\theta^*)\geq\varepsilon} M(\theta) > M(\theta^*), \text{ for all } \varepsilon > 0. \tag{A.2}$$

then $\hat{\theta}_n$ converges almost surely to θ^* as $n \to \infty$.

Proof: By Condition (A.1), we have

$$Z_n := M_n(\theta^*) - M(\theta^*) \xrightarrow{a.s.} 0 \quad \text{as} \quad n \to \infty. \tag{A.3}$$

Since $M_n(\hat{\theta}_n) \leq M_n(\theta^*)$ by definition of $\hat{\theta}_n$, we obtain $M_n(\hat{\theta}_n) \leq M(\theta^*) + Z_n$. Whence

$$0 \leq M(\hat{\theta}_n) - M(\theta^*) \leq M(\hat{\theta}_n) - M_n(\hat{\theta}_n) + Z_n \leq \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| + Z_n.$$

It follows from (A.1) and (A.3) that $M(\hat{\theta}_n) - M(\theta^*) \xrightarrow{a.s.} 0$, which is equivalent to say that

$$\lim_{n\to\infty} P[|M(\hat{\theta}_k) - M(\theta^*)| \leq \eta, \forall k \geq n] = 1 \tag{A.4}$$

for any $\eta > 0$ (see, *e.g.*, Serfling 1980, p.6, for this equivalent condition on the almost sure convergence). In order to prove that $d(\hat{\theta}_n, \theta^*) \xrightarrow{a.s.} 0$ as $n \to \infty$, it suffices to show

$$\lim_{n\to\infty} P[d(\hat{\theta}_k, \theta^*) \leq \varepsilon, \forall k \geq n] = 1 \tag{A.5}$$

for any $\varepsilon > 0$. Let $\varepsilon > 0$. By Condition (A.2) there exists $\eta > 0$ such that

$$\inf_{\theta : d(\theta,\theta^*) \geq \varepsilon} M(\theta) - M(\theta^*) > \eta.$$

Then, for each $k \geq 1$, the event $\{d(\hat{\theta}_k, \theta^*) \geq \varepsilon\}$ implies $\{M(\hat{\theta}_k) > M(\theta^*) + \eta\}$. This is equivalent to say that $\{|M(\hat{\theta}_k) - M(\theta^*)| \leq \eta\}$ is contained in the event $\{d(\hat{\theta}_k, \theta^*) < \varepsilon\}$, for all $k \geq 1$. Therefore, for all n (large enough), the event $\{|M(\hat{\theta}_k) - M(\theta^*)| \leq \eta, \forall k \geq n\}$ is contained in $\{d(\hat{\theta}_k, \theta^*) \leq \varepsilon, \forall k \geq n\}$. Thus (A.5) follows immediately from (A.4). $\square$

We also need to prove the almost sure convergence to zero of the following quantities :

$$\begin{aligned}
A_{j,n} &:= \sup_{\alpha\in\Lambda} \left| \frac{1}{n} \sum_{i=1}^{n} \|X_i - y_j\|^2 \mathbb{1}_{\alpha^{-1}(j)}(X_i) - \int_\Omega \|x - y_j\|^2 \mathbb{1}_{\alpha^{-1}(j)}(x) d\mu(x) \right|, \\
&\text{avec } 1 \leq j \leq J+1, \\
B_n &:= \sup_{y_{J+1}\in\Omega} A_{J+1,n} \\
C_{k,n} &:= \sup_{\alpha\in\Lambda} \left| \frac{1}{n} \sum_{i=1}^{n} h_{k,\alpha}(X_i, W_i) - \mathbb{E}[h_{k,\alpha}(X, W)] \right|, \ k = 1, \ldots, J+1.
\end{aligned}$$

Lemma 4.2. *If (H3) holds, then*
(i) $A_{j,n} \xrightarrow{a.s.} 0$ *as* $n \to \infty$, *for each* $j = 1, \ldots, J$.
(ii) $B_n \xrightarrow{a.s.} 0$ *as* $n \to \infty$.
(iii) $C_{k,n} \xrightarrow{a.s.} 0$ *as* $n \to \infty$, *for each* $k = 1, \ldots, J+1$.

Proof: (i) For each $j = 1, \ldots, J$, we denote $\mathcal{F}_j := \{f_{j,\alpha} : \alpha \in \Lambda\}$ the set of functions $f_{j,\alpha} : x \in \Omega \mapsto \|x - y_j\|^2 \mathbb{1}_{\alpha^{-1}(j)}(x)$. Since a collection of sets is a VC class of sets if and only if the collection of corresponding indicator functions is a VC class of functions (see e.g. van der Vaart 1998, p.275), we obtain by Assumption (H3) that $\mathcal{F} := \{f_\alpha : \alpha \in \Lambda\}$, the set of functions $f_\alpha : x \in \Omega \mapsto \mathbb{1}_{\alpha^{-1}(j)}(x)$, is a VC class of functions with envelope 1 (an envelope for a class of functions $\mathcal{F}$ is any function $F \in L^1(\mu)$ such that $|f| \leq F$ for all $f \in \mathcal{F}$). On the other hand, it is obvious that $\{f : x \in \Omega \mapsto \|x - y_j\|^2\}$

is a VC class of functions because it is a set of a single function. Then, by Lemma 2.14 in Pakes and Pollard (1989), $\mathcal{F}_j$ is also a VC class of functions with envelope given by the function f which belongs to $L^1(\mu)$ in view of the boundedness of Ω. Therefore, it follows from Corollary 3.12 in van de Geer (2000) that $\sup_{f\in\mathcal{F}_j} |\int f d\mu_n - \int f d\mu| \xrightarrow{a.s.} 0$, which completes the proof.

(ii) Here also, it suffices to show that $\mathcal{F}_{J+1} := \{f_{y_{J+1},\alpha} : y_{J+1} \in \Omega, \alpha \in \Lambda\}$, the set of functions $f_{y_{J+1},\alpha} : x \in \Omega \mapsto \|x - y_{J+1}\|^2 \mathbb{1}_{\alpha^{-1}(J+1)}(x)$, is a VC class of functions with an envelope in $L^1(\mu)$. To do this, we first need to prove that the set $\mathcal{G}_{J+1} = \{g_{y_{J+1}} : y_{J+1} \in \Omega\}$ of functions $g_{y_{J+1}} : x \in \Omega \mapsto \|x - y_{J+1}\|^2$ is a VC class of functions. Following van der Vaart (1998, p.275), for the collection $\mathcal{G}_{J+1}$ to be a VC class of functions, it suffices to show that the collection of all subgraphs $\{(x,t) \in \Omega \times \mathbb{R} : g_{y_{J+1}}(x) < t\} =: subgraph(g_{y_{J+1}})$ forms a VC class of sets in $\Omega \times \mathbb{R}$. By definition, to show that $\mathcal{D} := \{subgraph(g_{y_{J+1}}) : y_{J+1} \in \Omega\}$ is a VC class of sets it suffices to prove that its corresponding VC index denoted by $V(\mathcal{D})$ is finite (for the description of this index, see e.g. van de Geer 2000, Definition 3.3, p.40). Let $\mathbb{G}$ on $\Omega \times \mathbb{R}$ be the 5-dimensional vector space spanned by the following basis functions linearly independent to each other

$$\Phi_1(x,t) = \|x\|^2;\ \Phi_2(x,t) = x_{(1)};\ \Phi_3(x,t) = x_{(2)};\ \Phi_4(x,t) = t;\ \Phi_5(x,t) = 1,$$

where $x = (x_{(1)}, x_{(2)})$, and let $pos(\mathbb{G})$ be the collection of sets

$$pos(g) := \{(x,t) \in \Omega \times \mathbb{R} : g(x,t) > 0\}, \quad g \in \mathbb{G}.$$

According to Dudley (1978, Theorem 7.2, p.920), we have $V(pos(\mathbb{G})) = 6$. It is easy to see that, for any $y_{J+1} \in \Omega$, we have $subgraph(g_{y_{J+1}}) = \{(x,t) \in \Omega \times \mathbb{R} : t - g_{y_{J+1}}(x) > 0\} = pos(g)$, where $g(x,t) := t - g_{y_{J+1}}(x)$ belongs to the vector space $\mathbb{G}$. Hence, $\mathcal{D}$ is contained in $pos(\mathbb{G})$. It follows that $V(\mathcal{D}) \leq V(pos(\mathbb{G})) = 6$ and thus $\mathcal{D}$ is a VC class of sets. Therefore, $\mathcal{G}_{J+1}$ is a VC class of funcions with envelope given by the constant $diam(\Omega)^2 \in L^1(\mu)$. Finally, since $\mathcal{F} = \{\mathbb{1}_{\alpha^{-1}(j)} : \alpha \in \Lambda\}$ is a VC class of functions with envelope 1, we conclude by Lemma 2.14 in Pakes and Pollard (1989) that $\mathcal{F}_{J+1}$ is a VC class of functions with envelope given by the constant $diam(\Omega)^2$.

(iii) Following the same lines of the proof of (i), for each $j = 1, \ldots, J+1$, it is not hard to verify that the set $\mathcal{H}_j = \{h_{j,\alpha} : \alpha \in \Lambda\}$ of functions $h_{j,\alpha} : (x,w) \in \Omega \times \Psi \mapsto w\mathbb{1}_{\alpha^{-1}(j)}(x)$ is a VC class of functions with an envelope in $L^1(\pi)$, given by the function $(x,w) \in \Omega \times \Psi \mapsto w$. The desired conclusion follows immediately from Corollary 3.12 in van de Geer (2000). $\square$

Finally, we need to show the following result.

Lemma 4.3. *Under Conditions (H1) and (H3),* $\sup_{\alpha\in\Lambda}|\Phi_n(\alpha)-\Phi(\alpha)| \xrightarrow{a.s.} 0$ *as* $n\to\infty$.

Proof: Putting $r_{j,n}(\alpha) := \frac{\sum_{i=1}^n h_{j,\alpha}(X_i,W_i))}{\sum_{i=1}^n W_i}$, $r_j(\alpha) := \frac{\mathbb{E}(h_{j,\alpha}(X,W))}{\mathbb{E}(W)}$ for each $j = 1,\dots,J+1$, and $\bar{W}_n := (1/n)\sum_{i=1}^n W_i$, we have

$$\begin{aligned}\sup_{\alpha\in\Lambda}|r_{j,n}(\alpha)-r_j(\alpha)| &= \sup_{\alpha\in\Lambda}\left|\frac{\sum_{i=1}^n h_{j,\alpha}(X_i,W_i))}{\sum_{i=1}^n W_i} - \frac{\mathbb{E}(h_{j,\alpha}(X,W))}{\mathbb{E}(W)}\right| \\ &\le \frac{1}{\mathbb{E}[W]\bar{W}_n}(\sup_{\alpha\in\Lambda}\mathbb{E}[h_{j,\alpha}(X,W)]|\bar{W}_n-\mathbb{E}[W]|+\mathbb{E}[W]C_{j,n}).\end{aligned}$$

Since, $\sup_{\alpha\in\Lambda}\mathbb{E}[h_{j,\alpha}(X,W)] \le \mathbb{E}[W] < \infty$, $|\bar{W}_n - \mathbb{E}[W]| \xrightarrow{a.s.} 0$ by the strong law of large numbers and $C_{j,n} \xrightarrow{a.s.} 0$ by Lemma 4.2 (iii), we have $\sup_{\alpha\in\Lambda}|r_{j,n}(\alpha) - r_j(\alpha)| \xrightarrow{a.s.} 0$ for each j. Now, putting $v(x) = x\log(x)$ for $x>0$, we get

$$\begin{aligned}\sup_{\alpha\in\Lambda}|\Phi_n(\alpha)-\Phi(\alpha)| &= \sup_{\alpha\in\Lambda}\sum_{j=1}^{J+1}\left|r_{j,n}(\alpha)\log\frac{r_{j,n}(\alpha)}{z_j} - r_j(\alpha)\log\frac{r_j(\alpha)}{z_j}\right| \\ &\le \sum_{j=1}^{J+1}\left(\log(z_j)+\sup_{\alpha\in\Lambda}|v'(\lambda_{j,\alpha}r_{j,n}(\alpha)+(1-\lambda_{j,\alpha})r_j(\alpha))|\right)\sup_{\alpha\in\Lambda}|r_{j,n}(\alpha)-r_j(\alpha)|\end{aligned}$$

where $\lambda_{j,n}(\alpha)\in]0,1[$. To end the proof it is enough to show that $\sup_{\alpha\in\Lambda}|v'(\lambda_{j,\alpha}r_{j,n}(\alpha)+(1-\lambda_{j,\alpha})r_j(\alpha))|$ is bounded, for all n sufficiently large, with probability 1. Since $\inf_{\alpha\in\Lambda} r_j(\alpha) > 0$ by (H1) and

$$|\inf_{\alpha\in\Lambda}[\lambda_{j,\alpha}r_{j,n}(\alpha)+(1-\lambda_{j,\alpha})r_j(\alpha)] - \inf_{\alpha\in\Lambda} r_j(\alpha)| \le \sup_{\alpha\in\Lambda}|r_{j,n}(\alpha)-r_j(\alpha)| \xrightarrow{a.s.} 0,$$

we have for all n large enough

$$\inf_{\alpha\in\Lambda}[\lambda_{j,\alpha}r_{j,n}(\alpha)+(1-\lambda_{j,\alpha})r_j(\alpha)] > \inf_{\alpha\in\Lambda} r_j(\alpha)/2 \tag{A.6}$$

with probability 1. On the other hand, we have

$$\begin{aligned}&\sup_{\alpha\in\Lambda}|v'(\lambda_{j,\alpha}r_{j,n}(\alpha)+(1-\lambda_{j,\alpha})r_j(\alpha))| \\ &\qquad\le 1-\log\{\inf_{\alpha\in\Lambda}[\lambda_{j,\alpha}r_{j,n}(\alpha)+(1-\lambda_{j,\alpha}r_j(\alpha)]\} \\ &\qquad\le 1-\log\{\inf_{\alpha\in\Lambda} r_j(\alpha)/2\}.\end{aligned}$$

The last inequality follows from (A.6) and holds for all n large enough with probability 1. The upper bound is finite and strictly larger than 1 since $0 < \inf_{\alpha\in\Lambda} r_j(\alpha) \leq 1$. This completes the proof. $\square$

Proof of Theorem 4.1 According to Lemma 4.1, it suffices to show that $\sup_{y_{J+1}\in\Omega} |\inf_{\alpha\in\Lambda} M_n(y_{J+1},\alpha) - \inf_{\alpha\in\Lambda} M(y_{J+1},\alpha)| \xrightarrow{a.s.} 0$ as $n \to \infty$. This is immediate by applying the fact that

$$\sup_{y_{J+1}\in\Omega} |\inf_{\alpha\in\Lambda} M_n(y_{J+1},\alpha) - \inf_{\alpha\in\Lambda} M(y_{J+1},\alpha)| \leq \sum_{j=1}^{J} A_{j,n} + B_n + \lambda \sup_{\alpha\in\Lambda} |\Phi_n(\alpha) - \Phi(\alpha)|$$

in conjunction with Lemmas 4.2-4.3. $\square$

Troisième partie

Indices de concentration et caractéristiques du second-ordre d'un processus ponctuel spatial marqué

Chapitre 5

Indices de concentration et caractéristiques du second-ordre d'un processus ponctuel spatial marqué

Sommaire

5.1 Indices de concentration 104

5.2 Caractéristiques du second-ordre pour des processus ponctuels marqués 107

5.1 Processus ponctuels spatiaux marqués stationnaires 107

5.2 Processus ponctuels spatiaux marqués non-stationnaires 108

5.3 Indices de concentration basés sur les distances . 110

5.1 Indice de Duranton et Overman (2005) 110

5.2 Indice de Marcon et Puech (2007) 111

5.3 Nouvel indice de concentration 112

5.4 Comportement asymptotique de notre indice de concentration 113

5.1 Notations 113

5.2 Cadre asymptotique 113

5.3 Biais asymptotique 115

5.5 Conclusion et perspectives 117

Dans ce chapitre, nous présentons de façon non exhaustive une revue bibliographiques des indices de concentration introduits en économétrie. Duranton et Overman (2005) définissent les propriétés fondamentales requises pour être un "bon" indice de concentration. Pour s'affranchir d'un découpage du domaine d'observation en zones, de nouveaux indices basés sur les distances ont été introduits. Ensuite, nous rappelons quelques caractéristiques du second ordre d'un processus ponctuel spatial marqué et définissons de nouvelles caractéristiques lorsque le processus ponctuel des positions est stationnaire du second ordre avec pondération par l'intensité λ. Puis, nous présentons les indices basés sur les distances introduits par Duranton et Overman (2005) puis Marcon et Puech (2007), et les écrivons en fonction d'estimateurs de caractéristiques théoriques de processus ponctuels spatiaux marqués. Nous obtenons ainsi clairement quelle quantité théorique est estimée par ces indices. Enfin, nous présentons un nouvel indice de concentration issu de l'estimation d'une caractéristique du second ordre d'un processus ponctuel marqué.

5.1 Indices de concentration

Pendant des années, l'étude de la concentration géographique s'est limitée à considérer des mesures d'inégalités de données agrégées dans des zones déterminées (départements, cantons, IRIS, ...). Le lecteur pourra se reporter à Curry et George (1983) pour une présentation détaillée des indices de concentration suivant : concentration ratio, Rosenbluth index (1961), Comprehensive concentration index (1970), Pareto slope (1971), Linda index (1976), Hannak-Kay index (1977), U index (1980) et les mesures basées sur l'entropie. Beaucoup d'autres indices ont été définis plus récemment mais en l'absence d'un article de référence établissant un état de l'art il est difficile d'établir une liste exhaustive. Nous rajouterons malgré tout à la liste d'indices précédents l'indice de Theil (1967), l'indice d'Atkinson (1996) et l'indice de Krugman (1991). D'autres références présentent une introduction aux questions de la mesure de la concentration géographique comme les articles de Valeyre (1993) et Houdebine (1999). Les indices d'Herfindahl (1950), de Gini (1991) et d'Ellison-Glaeser (1997) sont d'ailleurs présentés dans Houdebine (1999) comme les plus utilisés en pratique. L'indice d'Ellison-Glaeser est fondé sur une description probabiliste du comportement des entreprises. Cet indice présente l'avantage d'améliorer les indices de Herfindahl et de Gini du point de vue des propriétés fondamentales requises pour être un "bon" indice de concentration. Ces propriétés fondamentales énoncés par Duranton et

Overman (2005) sont présentés par la suite. Pour des unités géographiques, notées $R_1, \cdots, R_M$ et des industries positionnées en $x_1, \cdots, x_N$ avec des effectifs respectifs $m_1, \cdots, m_N$, l'indice d'Ellison-Glaeser s'écrit

$$I_{EG} = \frac{G_{EG} - H}{1 - H} \quad \text{avec } G_{EG} = \frac{\sum_{j=1}^{M}(u_j - v_j)^2}{1 - \sum_{j=1}^{M} v_j^2},$$

où $u_j := \sum_{i=1}^{N} m_i \mathbb{1}(x_i \in R_j) / \sum_{i=1}^{N} m_i$; $v_j := \sum_{i=1}^{N} m_i / N$ et $H = \sum_{j=1}^{M} m_j^2 / (\sum_{j=1}^{M} m_j)^2$ est l'indice d'Herfindahl. Maurel et Sédillot (1999) proposent un indice de concentration construit dans les mêmes lignes que celui d'Ellison-Glaeser mais avec des estimateurs quelque peu différents. L'indice de Maurel et Sédillot (1999) présenté comme plus naturel dans sa construction que celui d'Ellison-Glaeser s'écrit

$$I_{MS} = \frac{G_A - H}{1 - H} \quad \text{avec } G_A = \frac{\sum_{j=1}^{M} u_j^2 - \sum_{j=1}^{M} v_j^2}{1 - \sum_{j=1}^{M} v_j^2}.$$

Cependant, l'agrégation de données relatives à des établissements au niveau de zones spatiales fixées introduit de fausses corrélations entre les variables agrégées. Ce problème est connu sous le nom MAUP ("Modifiable Areal Unit Problem"). De plus, ce découpage souvent arbitraire peut scinder un regroupement d'établissements d'un secteur d'activité dans deux zones géographiques et ainsi perdre la structure réelle de la répartition. Ceci est d'autant plus préjudiciable que la majorité des indices de concentration traitent des unités spatiales proches de la même façon que si elles étaient d'un bout à l'autre du territoire considéré. Pour corriger ce dernier désanvantage, Dawkins (2004) a introduit un indice de Gini spatial pour tenir compte de la notion de régions voisines. Il est à noter que des indices d'autorrélation spatiale comme l'indice de Moran ou de Geary (voir par exemple, Cliff et Ord, 1981) ne sont pas convenables pour mesurer de la concentration spatiale. C'est pour cette raison qu'Arbia et Piras (2009) définissent un test statistique dérivé de ces indices pour mesurer la concentration spatiale.

Ainsi, depuis les années 70, les économistes ont amélioré la mesure des indices de concentration géographique pour évaluer plus précisément l'agglomération de secteurs d'activités, et permettre de répondre aux questions d'aménagement économique du territoire. Cette amélioration s'est faite grâce à la détermination de propriétés fondamentales demandées pour un "bon" indice de concentration. Duranton et Overman (2005) suggèrent qu'un "bon" indice de concentration doit :

(i) être comparable entre les différents secteurs d'activité,

(ii) contrôler la tendance globale d'agrégation,

(iii) contrôler la concentration productive de chaque secteur d'activité au travers de leur taille,

(iv) être non biaisé par rapport au choix de l'échelle géographique,

(v) permettre de tester la significativité des résultats.

La propriété (ii) consiste à prendre en compte les facteurs entrainant nécessairement une agrégation des secteurs d'activité. Par exemple, on peut logiquement s'attendre à ce qu'un secteur d'activité ait un taux d'emploi important dans des zones où la densité de population est forte. Cette propriété naturelle est à distinguer de la propriété (iii) qui consiste à prendre en compte la structure productive de chacun des secteurs d'activité considérés. En effet, des secteurs avec peu d'établissements auront une structure de production concentrée les empêchant de se répartir de façon homogène sur le territoire. Ainsi, la concentration géographique des secteurs dont l'activité est, par nature, très dispersée (par exemple les services de proximité) ne peut pas être directement comparée à celle de secteurs où la concentration productive est plus importante (par exemple l'industrie automobile). Il est ainsi nécessaire d'utiliser la connaissance de la taille des établissements, par exemple en nombre d'employés, pour corriger cette concentration inhérente à chaque secteur. Ces propriétés, classées dans l'ordre chronologique d'apparition, ont permis d'affiner la détermination de "bons" indices de concentration.

Grâce aux moyens actuels, il est de plus en plus fréquent de bénéficier des informations sur le positionnement exact des établissements. Ainsi, il est dommageable de perdre de l'information en agrégeant les données sur les établissements dans des zones administratives souvent économiquement arbitraires. Les développements concernant la mesure d'indices de concentration à partir de données ponctuelles dans un espace continu sont relativement récents en économie. Les articles de Duranton et Overman (2005), puis Marcon et Puech (2007), ouvrent la voie à de nouveaux indices basés sur les distances, qui s'affranchissent ainsi d'un découpage en zones à l'intérieur desquelles des données ponctuelles sont agrégées. Malheureusement, les liens entre ces indices et les quantités théoriques qu'ils estiment ne sont pas clairement établis en l'abscence d'un cadre mathématique adapté. Par la suite, nous uniformisons l'écriture de ces indices dans le cadre de la théorie des processus ponctuels spatiaux marqués. Pour cela, nous rappelons et introduisons des définitions de caractéristiques théoriques du second ordre.

5.2 Caractéristiques du second-ordre pour des processus ponctuels marqués

Dans cette section, nous revenons tout d'abord sur les définitions de caractéristiques du second-ordre dans le cadre stationnaire, pour des processus ponctuels marqués. Ensuite, nous introduisons une généralisation dans le cadre non-stationnaire, mais stationnaire du second-ordre avec pondération par l'intensité ("second-order intensity reweighted").

5.1 Processus ponctuels spatiaux marqués stationnaires

Des caractéristiques du second ordre de processus ponctuels marqués dans le cadre stationnaire ont été définies et employées notamment dans Stoyan et Stoyan (1994), Schlather (2001) ou Mateu (2000).

Soit $(X, M) = \{(\xi, \mathfrak{m}_\xi) : \xi \in X\}$ un processus ponctuel marqué stationnaire simple sur un espace borné $A \subset \mathbb{R}^2$ avec marques positives, et notons la σ-algèbre de A par $\mathcal{A}$. Pour simplifier les notations, on notera le processus ponctuel marqué $\{(\xi_i, \mathfrak{m}_i)_{(i\geq 1)} : \xi_i \in X\}$. Une caractéristique du second ordre d'un processus ponctuel marqué anisotrope (resp. isotrope) est une quantité conditionnelle sachant que les points sont éloignés d'une direction t (resp. distance r). Parce que la probabilité de trouver une paire de points éloignés d'une direction t (resp. distance r) est nulle pour des processus ponctuels simples dans un domaine borné, nous considérons la mesure de moment réduit du second ordre d'une fonction f, notée $\alpha_f^{(2)}$.

Définition 5.1. *Pour toute fonction mesurable positive f sur $\mathbb{R}_+^2$, la mesure $\alpha_f^{(2)}$ sur A^2 est définie par*

$$\alpha_f^{(2)}(B_1 \times B_2) = \mathbb{E}\left[\sum_{\xi_1 \in X} \sum_{\xi_2 \in X}^{\neq} f(\mathfrak{m}_1, \mathfrak{m}_2) \mathbb{1}_{B_1}(\xi_1) \mathbb{1}_{B_2}(\xi_2)\right] \quad \textit{avec } B_1, B_2 \in \mathcal{A}.$$

Si $\alpha_f^{(2)}$ est absolument continue par rapport à la mesure de Lebesgue alors il existe une fonction densité notée $\rho_f^{(2)}$. Dans le but d'analyser les marques, on introduit une version normalisée de $\rho_f^{(2)}$, en divisant par la caractéristique du second ordre du processus non marqué $\rho^{(2)}$. Sous l'hypothèse d'anisotropie (resp. isotropie), on note $\kappa_f(t) = \rho_f^{(2)}(t)/\rho^{(2)}(t)$ (resp. $\kappa_f(r) = \rho_f^{(2)}(r)/\rho^{(2)}(r)$) où $\kappa_f(t)$ (resp. $\kappa_f(r)$) correspond à l'espérance conditionnelle de $f(\mathfrak{m}_i, \mathfrak{m}_j)$ sachant qu'il existe deux points ξ_i et ξ_j de X tels que $\xi_i - \xi_j = t$ (resp. $\|\xi_i - \xi_j\| = t$) de marques respectives $\mathfrak{m}_i$ et $\mathfrak{m}_j$. Le choix de la fonction f

dépend de la nature des marques et du contexte d'étude. On peut citer par exemple :

(ii) $f(m_i, m_j) = \mathbb{1}(m_i = m_j)$ quand les marques sont discrètes,

(ii) $f(m_i, m_j) = m_i m_j$, quand les marques représentent des tailles d'objet,

(iii) $f(m_i, m_j) = \min\{|m_i - m_j|, \pi - |m_i - m_j|\}$, quand les marques sont des angles compris entre 0 et π.

Dans le cadre de marques continues, de nombreuses caractéristiques du second ordre ont été construites à partir de κ_f pour des fonctions f particulières. Une caractéristique très connue est le variogramme des marques (Cressie, 1993) pour laquelle $f(m_1, m_2) = (m_1 - m_2)^2$. Nous renvoyons le lecteur à Schlather (2001) et aux références qu'il contient pour d'autres exemples.

5.2 Processus ponctuels spatiaux marqués non-stationnaires

Dans cette section, nous définissons la mesure aléatoire $\beta_f^{(2)}$ à partir de $\beta^{(2)}$, en suivant le même schéma permettant de définir $\alpha_f^{(2)}$ à partir de $\alpha^{(2)}$. Nous reprenons les notations et hypothèses de la section précédente avec $(X, M) = \{(\xi, \mathfrak{m}_\xi) : \xi \in X\}$ un processus ponctuel marqué simple dont le processus marginal des positions X est stationnaire du second ordre avec pondération par l'intensité λ.

On considère que $f(m_1, m_2)$ s'écrit comme le produit $k(m_1)q(m_2)$. Nous verrons par la suite que les fonctions f prises par Duranton et Overman (2005) et Marcon et Puech (2007) appartiennent à cette famille de fonctions. Nous pouvons définir la caractéristique du premier ordre λ_k à partir de la mesure aléatoire μ_k.

Définition 5.2. *Pour toute fonction mesurable k sur $\mathbb{R}_+$, la mesure μ_k sur A est définie par*

$$\mu_k(B) = \mathbb{E}\left[\sum_{\xi \in X} k(\mathfrak{m}) \mathbb{1}_B(\xi)\right] \quad \textit{avec } B \in \mathcal{A}.$$

Si μ_k et μ_q sont absolument continues par rapport à la mesure de Lebesgue alors on note respectivement λ_k et λ_q leur densité. Nous introduisons maintenant une mesure aléatoire d'ordre 2, notée $\beta_f^{(2)}$.

Définition 5.3. *Pour toute fonction mesurable positive f sur $\mathbb{R}_+^2$, s'écrivant $f(m_1, m_2) = k(m_1)q(m_2)$, la mesure $\beta_f^{(2)}$ sur A^2 est définie par*

$$\beta_f^{(2)}(B_1 \times B_2) = \mathbb{E}\left[\sum_{\xi_1 \in X} \sum_{\xi_2 \in X}^{\neq} \frac{f(\mathfrak{m}_1, \mathfrak{m}_2)}{\lambda_k(\xi_1)\lambda_q(\xi_2)} \mathbb{1}_{B_1}(\xi_1)\mathbb{1}_{B_2}(\xi_2)\right] \quad \text{avec } B_1, B_2 \in \mathcal{A},$$

si $\lambda_k(x) > 0$ et $\lambda_q(x) > 0$ presque sûrement pour tout $x \in A$.

Sous l'hypothèse d'absolue continuité de $\beta_f^{(2)}$ par rapport à la mesure de Lebesgue, on note g_f la densité associée. Nous obtenons sans difficulté que g_f s'écrit sous la forme

$$\begin{aligned} g_f(x_1, x_2) &= \frac{\rho_f^{(2)}(x_1, x_2)}{\lambda_k(x_1)\lambda_q(x_2)} = \frac{\mathbb{E}\left[k(\mathfrak{m}_i)q(\mathfrak{m}_j)|\xi_i = x_1, \xi_j = x_2\right]\rho^{(2)}(x_1, x_2)}{\mathbb{E}[k(\mathfrak{m}_i)|\xi_i = x_1]\mathbb{E}[q(\mathfrak{m}_j)|\xi_j = x_2]\lambda(x_1)\lambda(x_2)} \\ &= \frac{\mathbb{E}\left[k(\mathfrak{m}_i)q(\mathfrak{m}_j)|\xi_i = x_1, \xi_j = x_2\right]}{\mathbb{E}[k(\mathfrak{m}_i)|\xi_i = x_1]\mathbb{E}[q(\mathfrak{m}_j)|\xi_j = x_2]} g(x_1, x_2). \end{aligned}$$

On rappelle que la fonction de corrélation des paires g est égale à 1 pour un processus ponctuel de Poisson d'après le théorème de Slivnyak-Mecke. Si l'on considère un processus ponctuel marqué de Poisson $Y_P = (X_P, M_P)$ pour lequel conditionnellement à X_P, chaque marque $\mathfrak{m}_\xi$ a une densité ne dépendant exclusivement que de ξ alors on obtient

$$\begin{aligned} \rho_f^{(2)}(x_1, x_2) &= \mathbb{E}\left[k(\mathfrak{m}_i)|\xi_i = x_1, \xi_j = x_2\right]\mathbb{E}\left[q(\mathfrak{m}_j)|\xi_i = x_1, \xi_j = x_2\right]\rho^{(2)}(x_1, x_2) \\ &= \mathbb{E}\left[k(\mathfrak{m}_i)|\xi_i = x_1\right]\lambda(x_1)\mathbb{E}\left[q(\mathfrak{m}_j)|\xi_j = x_2\right]\lambda(x_2) \\ &= \lambda_k(x_1)\lambda_q(x_2). \end{aligned}$$

Par conséquent, g_f est égale à 1 pour des processus ponctuels de la même famille que Y_P. On pourrait ainsi définir un test nous permettant d'indiquer si une configuration de points peut être une réalisation d'un processus ponctuel marqué de Poisson où en chaque point ξ la densité de $\mathfrak{m}_\xi$ ne dépend pas de $X \backslash \xi$, bien que ce ne soit certainement pas une condition nécessaire et suffisante.

La caractéristique du second ordre g_f que nous introduisons pour des processus ponctuels marqués, dont le processus ponctuel des positions est stationnaire du second ordre avec pondération par l'intensité, va permettre de définir notre nouvel indice de concentration.

5.3 Indices de concentration basés sur les distances

Nous présentons dans cette section deux indices de concentration basés sur les distances, l'un introduit par Duranton et Overman (2005) et l'autre par Marcon et Puech (2007). Nous utilisons les notations (x_i, m_i) pour définir respectivement la position et la marque associée, correspondant au nombre d'employés, de l'établissement i.

5.1 Indice de Duranton et Overman (2005)

La mesure de la concentration de l'emploi définie par Duranton et Overman (2005) est donnée par l'indice suivant :

$$I_{DO}(r) = \frac{\sum_i \sum_{j>i} h^{-1} w\left(\frac{r-\|x_i-x_j\|}{h}\right) m_i m_j}{\sum_i \sum_{j>i} m_i m_j}$$

où h est une fenêtre de lissage et w est une fonction noyau. Cet estimateur non-paramétrique consiste à calculer la densité des distances entre toutes les paires d'établissements d'un secteur, pondérées par une fonction du nombre d'employés. Cette fonction f est telle que $f(m_i, m_j) = m_i m_j$. Dans le cas d'un processus stationnaire isotrope, un estimateur non paramétrique de $\rho_f^{(2)}(r)$ est défini dans Stoyan et Stoyan (1994) sans correction de bord par

$$\hat{\rho}_f^{(2)}(r) = \frac{1}{2\pi r|A|} \sum\sum_{i \neq j} h^{-1} w\left(\frac{r - \|x_i - x_j\|}{h}\right) m_i m_j, \quad \forall r > 0.$$

Etant donné qu'un estimateur de λ^2 dans la cas stationnaire est donné par $n(n-1)/|A|^2$, on a

$$\begin{aligned} I_{DO}(r) &= \frac{\sum\sum_{i\neq j} h^{-1} w\left(\frac{r-\|x_i-x_j\|}{h}\right) m_i m_j}{\sum\sum_{i\neq j} m_i m_j} \\ &= \frac{2\pi r \hat{\rho}_f^{(2)}(r)}{|A|\widehat{\mathbb{E}[\mathfrak{m}]^2}\widehat{\lambda^2}}. \end{aligned}$$

Un lien est ainsi établi entre la définition de l'indice de concentration I_{DO} et les estimations de caractéristiques de processus ponctuel marqué. On entrevoit ainsi la quantité théorique qui est estimée et pouvons nous demander si cet indice est légitime compte tenu du facteur multiplicatif $2\pi r/|A|$. Il serait

préférable de choisir comme indice $\hat{\rho}_f^{(2)}(r)/\widehat{\mathbb{E}[m]^2}\widehat{\lambda^2}$ qui est un estimateur de g_f dans le cadre d'un processus ponctuel marqué stationnaire isotrope sous l'hypothèse de "random labelling" (c'est à dire, marques indépendantes des positions et indépendantes entre elles). De plus, cet indice ne présente pas de correction de bord ce qui peut certainement aboutir à une sous estimation de la quantité théorique.

5.2 Indice de Marcon et Puech (2007)

Marcon et Puech (2007) soulignent que l'indice de Duranton et Overman (2005) ne permet pas une interprétation des résultats bien qu'il satisfasse les propriétés fondamentales pour être un "bon" indice de concentration. Cet inconvénient les pousse à définir un nouvel indice de concentration que nous noterons I_{MP}. En effet, Marcon et Puech (2007) reprochent à l'indice de concentration I_{DO} de ne pas permettre de quantifier les différences en terme de nombre d'employés ou de nombre d'établissements. Cet inconvénient ne paraît cependant pas rédhibitoire étant donné qu'il apparaît aussi pour des indices largement utilisés comme par exemple l'indice de Gini. L'indice de Marcon et Puech (2007) consiste à regarder pour chaque établissement d'un secteur s, localisé en $x_{i,s}$, le ratio entre le nombre d'employés des établissements voisins appartenant au secteur s et le nombre d'employés de tous les établissements voisins. Tous ces indices locaux sont ensuite additionés, puis une normalisation est effectuée en divisant par le ratio entre le nombre d'employés total du secteur s et le nombre total d'employés. L'indice I_{MP} est ainsi défini par

$$I_{MP}(r) = \sum_{i=1}^{N_s} \frac{\sum_{j=1,j\neq i}^{N_s} m_j \mathbb{1}(\|x_{i,s} - x_{j,s}\| \leq r)}{\sum_{j=1,j\neq i}^{N} m_j \mathbb{1}(\|x_{i,s} - x_j\| \leq r)} / \sum_{i=1}^{N_s} \frac{\sum_{j=1,j\neq i}^{N_s} m_j}{\sum_{j=1,j\neq i}^{N} m_j} \quad \forall r > 0,$$

où N_s est le nombre d'établissements du secteur s, N le nombre total d'établissements et $x_{i,s}$ la position d'un établissement d'indice i appartenant au secteur s. Pour simplifier, on constate que l'indice $I_{MP}(r)$ peut s'écrire sous la forme $id(r)/id(\infty)$ où

$$id(r) = \sum_{i=1}^{N_s} \frac{\sum_{j=1,j\neq i}^{N_s} m_j \mathbb{1}(\|x_{i,s} - x_{j,s}\| \leq r)}{\sum_{j=1,j\neq i}^{N} m_j \mathbb{1}(\|x_{i,s} - x_j\| \leq r)}$$

Ce type de normalisation d'une fonction par sa valeur en l'infini n'est pas surprenante et apparait aussi quand on considère les caractéristiques du second ordre de processus ponctuels marqués (voir par exemple, Mateu (2000)).

Nous établissons maintenant l'écriture de l'indice $id(r)$ en fonction de caractéristiques estimées d'un processus ponctuel marqué.

$$\begin{aligned} id(r) &= \frac{N_s(N_s-1)}{|A|^2(N-1)} \sum_{i=1}^{N_s} \sum_{j=1,j\neq i}^{N_s} \frac{m_j \mathbb{1}(\|x_{i,s}-x_{j,s}\| \leq r)}{\frac{N_s(N_s-1)}{|A|^2} \sum_{j=1,j\neq i}^{N} \frac{m_j \mathbb{1}(\|x_{i,s}-x_j\| \leq r)}{N-1}} \\ &= \frac{N_s(N_s-1)}{|A|^2(N-1)} \sum_{i=1}^{N_s} \sum_{j=1,j\neq i}^{N_s} \frac{m_j \mathbb{1}(\|x_{i,s}-x_{j,s}\| \leq r)}{\widehat{\lambda}^2 \widehat{\mathbb{E}}[\mathfrak{m}_j \mathbb{1}(\|\xi_{i,s}-\xi_j\| \leq r) | \xi_{i,s}=x_{i,s}]} \end{aligned}$$

Cet indice $id(r)$ est proportionnel à l'estimateur d'une version pondérée de la fonction K prenant en compte les marques. On constate que les processus ponctuels marqués considérés sont tels que, pour chacun d'entre eux, le processus ponctuel des positions est stationnaire et isotrope. Cependant, Marcon et Puech (2007) prennent en compte la dépendance des marques avec les positions voisines. La prise en compte de l'ensemble des secteurs pour le calcul de leur indice de concentration sur un secteur ne permet pas de corriger l'inhomogénéité des positions, ni d'apporter une correction de bord mais de prendre en compte la dépendance des marques avec les positions voisines.

5.3 Nouvel indice de concentration

Nous choisissons d'introduire dans cette section un indice de concentration basé sur la même forme que l'estimation de densité dans l'indice I_{DO} et non pas sous une forme cumulative comme l'indice id défini dans I_{MP}. Ce choix est simplement dû à la plus grande facilité d'interprétation de notre indice par rapport à l'indice cumulatif associé que nous aurions pu construire. La formulation des indices précédents qui ne prennent pas en compte la répartition hétérogène des points nous conduit à introduire un estimateur de g_f pour définir un nouvel indice de concentration. Si g_f est invariant par translation l'estimateur proposé s'écrit

$$\hat{g}_f(t) = \sum_{i=1}^{N} \sum_{j=1,j\neq i}^{N} \frac{h^{-2} w\left(\frac{t-\xi_i+\xi_j}{h}\right) k(\mathfrak{m}_i) q(\mathfrak{m}_j)}{|A \cap (A-\xi_i+\xi_j)| \hat{\lambda}_k(\xi_i) \hat{\lambda}_q(\xi_j)} \quad \forall t,$$

où $\hat{\lambda}_k(x)$ peut être choisi comme égal à $\hat{\mathbb{E}}\left[k(\mathfrak{m}_\xi)|\xi=x\right]\hat{\lambda}(x)$, avec $\hat{\lambda}$ l'estimateur de l'intensité défini par Diggle (1985).

Si g_f est aussi invariant par rotation alors on peut définir

$$\hat{g}_f(r) = \frac{1}{2\pi r} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \frac{h^{-1} w\left(\frac{r-\|\xi_i-\xi_j\|}{h}\right) k(\mathfrak{m}_i) q(\mathfrak{m}_j)}{|A \cap (A - \xi_i + \xi_j)| \hat{\lambda}_k(\xi_i) \hat{\lambda}_q(\xi_j)} \quad \forall r > 0,$$

Il est à noter que nous aurions pu aussi définir un indice de concentration cumulatif comme le font Marcon et Puech (2007).

5.4 Comportement asymptotique de notre indice de concentration

Dans cette section, nous nous intéressons aux propriétés asymptotiques de l'estimateur $\hat{g}_{f,n}(t)$ d'un processus ponctuel marqué (X_n, M_n) dans un cadre asymptotique spécifique.

5.1 Notations

Nous notons $A \subset \mathbb{R}^2$, un ouvert borné, la région où les réalisations des processus ponctuels marqués sont observées. Soit $|A|$ l'aire de A, h une constante positive et w un noyau produit symétrique et borné de support $C \times C \subset \mathbb{R}^2$ défini par

$$w(x) = w_0(x_{(1)}) w_0(x_{(2)}), \quad \forall x = (x_{(1)}, x_{(2)}) \in \mathbb{R}^2,$$

où le noyau w_0 satisfait donc

$$\int_C z w_0(z) dz = 0 \text{ et } \int_C z^2 w_0(z) dz < \infty.$$

Pour alléger les notations, nous introduisons $w_{h_n}(\cdot) = h_n^{-2} w(\frac{\cdot}{h_n})$.

5.2 Cadre asymptotique

En estimation de densité dans $\mathbb{R}^d$, le cadre asymptotique consiste à faire tendre la taille de l'échantillon n vers l'infini alors que la fenêtre de lissage h tend vers 0 de telle sorte que nh^d tend vers l'infini.

En théorie des processus ponctuels, plusieurs cadres asymptotiques sont possibles. Tout d'abord, on peut choisir d'augmenter l'espérance du nombre de points du processus par l'accroissement de la région d'observation A (Cressie, 1993). Guan *et al.* (2007) établissent ainsi la consistance et la normalité

asymptotique du variogramme empirique des marques $\hat{\kappa}_f(t)$ sachant les positions, pour un processus ponctuel marqué stationnaire et anisotrope. Dans le cadre géostatistique, García-Soidán *et al.* (2004) introduisent un estimateur de Nadaraya-Watson du variogramme et démontre la consistance pour un champ stationnaire isotrope. Un autre cadre asymptotique adoptée par Diggle et Marron (1988) consiste à accroître l'intensité du processus ponctuel. Nous verrons que notre cadre asymptotique est très proche de ce dernier. Un cadre mixte où intensité et fenêtre d'observation augmentent a aussi était considéré. Enfin, un cadre plus proche de celui de l'estimation de densité serait de considérer plusieurs réalisations d'un même processus et de faire tendre n vers l'infini (Kutoyants, 1998).

En statistique spatiale, le cadre asymptotique avec domaine borné apparait impopulaire, cependant il ne semble pas incongru en pratique. En géostatistique, Stein (1999) souligne l'importance de considérer l'interpolation en des points non observés (krigeage) lorsque le nombre de points à proximité augmente et non lorsque le nombre de points éloignés augmente par le biais de l'agrandissement du domaine. Dans le cadre des processus ponctuels, nous sommes souvent confrontés en pratique à des domaines d'observation fixés pour des raisons techniques ou réelles dont le nombre de points d'une réalisation augmente. Nous avons donc décidé d'étudier la convergence de notre estimateur vers la caractéristique théorique pour des processus hétérogènes lorsque l'intensité du processus augmente dans un domaine fixé.

Nous allons maintenant définir notre cadre asymptotique avec la région A fixée. Considérons une séquence de processus ponctuels (X_n, M_n) dont les caractéristiques associées sont notées λ_n, $\rho_n^{(2)}$, g_n et $g_{f,n}$. On considère une séquence de fenêtre de lissage h_n. On définit par λ, $\rho^{(2)}$, g et g_f les caractéristiques associées à (X_0, M_0). On suppose que g_f est une fonction au moins différentiable jusqu'à l'ordre 2 et de dérivées partielles jusqu'à l'ordre 2 bornées. Notre cadre asymptotique est bâti sur les hypothèses suivantes :

(H1) $\lambda_n = n\lambda$ et $\rho_n^{(2)}(x_1, x_2) = n^2\rho^{(2)}(x_1, x_2)$,

(H2) $\mathbb{E}\left[k(\mathfrak{m}_i)|\xi_i \in X_n\right] = \mathbb{E}\left[k(\mathfrak{m}_i)|\xi_i \in X_0\right]$ $\quad [\lambda_{\mathbf{k},\mathbf{n}} = \lambda_{\mathbf{k}}]$
$\mathbb{E}\left[q(\mathfrak{m}_j)|\xi_j \in X_n\right] = \mathbb{E}\left[q(\mathfrak{m}_j)|\xi_j \in X_0\right]$ $\quad [\lambda_{\mathbf{q},\mathbf{n}} = \lambda_{\mathbf{q}}]$

(H3) $\mathbb{E}\left[k(\mathfrak{m}_i)q(\mathfrak{m}_j)|\xi_i, \xi_j \in X_n\right] = \mathbb{E}\left[k(\mathfrak{m}_i)q(\mathfrak{m}_j)|\xi_i, \xi_j \in X_0\right]$
pour tout $n \in \mathbb{N}$,

(H4) $h_n = O(n^{-\beta})$ avec $\beta \in]0, 1[$.

L'hypothèse $(H1)$ est une hypothèse vérifiée lorsque l'on considère que X_0 est l'"amincissement"(thinning) de X_n avec une probabilité de rétention constante égale à $1/n$ (Proposition 4.2 p.31, Moller et Waagepetersen 2004). Cette hypothèse est naturelle dans notre cadre asymptotique où l'on souhaite que X_n soit une "intensification" de X_0. Les hypothèses $(H2)$ et $(H3)$ consistent à dire que les espérances de fonction des marques, conditionnellement aux positions, ne dépendent pas des paramètres du processus ponctuel des positions mais de la famille à laquelle il appartient. Cependant, lorsque n augmente, l'espérance du nombre de points et l'espérance de la somme des marques augmentent. Ces hypothèses sont naturelles si l'on considère que le nombre d'industries et le nombre d'employés dans toute une région A augmentent mais que la dépendance spatiale du nombre d'employés d'un établissement dépend uniquement de sa position et pas des caractéristiques du processus ponctuel sous jacent. Enfin, les hypothèses $(H1)$, $(H2)$ et $(H3)$ impliquent que $g_{f,n}(x_1, x_2) = g_f(x_1, x_2)$. Quand n tend vers l'infini, l'hypothèse $(H4)$ implique que $h_n \to 0$ et $nh_n \to \infty$.

5.3 Biais asymptotique

Dans cette section, le théorème présenté montre que notre estimateur de g_f est asymptotiquement sans biais. Pour cela, nous avons besoin des hypothèses suivantes :

(H5) $g(x_1, x_2) = g(x_1 - x_2)$,

(H6)

$$\frac{(\mathbb{E}\left[k(\mathfrak{m}_i)q(\mathfrak{m}_j)|\xi_i, \xi_j \in X_0\right])}{(\mathbb{E}\left[k(\mathfrak{m}_i)|\xi_i \in X_0\right])(\mathbb{E}\left[q(\mathfrak{m}_j)|\xi_j \in X_0\right])} = \gamma_{k,q}(x_1 - x_2).$$

Les hypothèses $(H5)$ et $(H6)$ correspondent à des hypothèses d'invariance par translation et permettent d'obtenir l'invariance par translation de g_f, c'est à dire que $g_f(x_1, x_2) = g_f(x_1 - x_2)$.

Théorème 5.1. *Sous les hypothèses* $(H1) - (H6)$, *nous avons*

$$\mathbb{E}[\hat{g}_{f,n}(t)] = g_f(t) + \frac{h_n^2}{2}\left(\frac{\partial^2 g_f}{\partial x_{(1)}^2}(t) + \frac{\partial^2 g_f}{\partial x_{(2)}^2}(t)\right)\int_C z^2 w_0(z)dz + O(h_n^3)$$

Preuve

$$\mathbb{E}[\hat{g}_{f,n}(t)] = \mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\frac{h_n^{-2}w((t-\xi_i+\xi_j)h_n^{-1})k(\mathfrak{m}_i)q(\mathfrak{m}_j)}{|A\cap(A-\xi_i+\xi_j)|\lambda_{k,n}(\xi_i)\lambda_{q,n}(\xi_j)}\right]$$
$$= \mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\frac{h_n^{-2}w((t-\xi_i+\xi_j)h_n^{-1})\mathbb{E}\left[k(\mathfrak{m}_i)q(\mathfrak{m}_j)|\xi_i,\xi_j\in X_n\right]}{|A\cap(A-\xi_i+\xi_j)|\lambda_{k,n}(\xi_i)\lambda_{q,n}(\xi_j)}\right]$$

D'après la Proposition 4.1 dans Moller et Waagepetersen (2004, p.31), cette espérance est égale à

$$\int_A\int_A\frac{w_{h_n}(t-x_1+x_2)}{|A\cap(A-x_1+x_2)|}\frac{\rho_{f,n}^{(2)}(x_1,x_2)}{\lambda_{k,n}(x_1)\lambda_{q,n}(x_2)}dx_1dx_2$$
$$= \int_A\int_A\frac{w_{h_n}(t-x_1+x_2)}{|A\cap(A-x_1+x_2)|}g_{f,n}(x_1-x_2)dx_1dx_2$$

Puis par les changements de variables $u = x_1 - x_2$ et $v = x_2$, on a

$$\int_{A-A}\int_{A\cap(A-u)}\frac{w_{h_n}(t-u)}{|A\cap(A-u)|}g_f(u)dvdu = \int_{A-A}w_{h_n}(t-u)g_f(u)du$$
$$= \int_{\frac{t+A-A}{h_n}}w(v)g_f(t-h_nv)dv$$

D'après le développement de Taylor-Lagrange

$$g_f(t-h_nv) = g_f(t) - h_n\left(v_{(1)}\frac{\partial g_f}{\partial x_{(1)}}(t) + v_{(2)}\frac{\partial g_f}{\partial x_{(2)}}(t)\right)$$
$$+\frac{h_n^2}{2}\left(v_{(1)}^2\frac{\partial^2 g_f}{\partial x_{(1)}^2}(t) + v_{(2)}^2\frac{\partial^2 g_f}{\partial x_{(2)}^2}(t) + v_{(1)}v_{(2)}\frac{\partial^2 g_f}{\partial x_{(1)}\partial x_{(2)}}(t)\right) + O(h_n^3),$$

et en introduisant $I(h_n) := \int_{\frac{t+A-A}{h_n}}w(v)dv$, l'expression de $\mathbb{E}[\hat{g}_{f,n}(t)]$ devient

$$g_f(t)I(h_n) - h_n\frac{\partial g_f}{\partial x_{(1)}}(t)\int_{\frac{t+A-A}{h_n}}v_{(1)}w(v)dv - h_n\frac{\partial g_f}{\partial x_{(2)}}(t)\int_{\frac{t+A-A}{h_n}}v_{(2)}w(v)dv$$
$$+\frac{h_n^2}{2}\frac{\partial^2 g_f}{\partial x_{(1)}^2}(t)\int_{\frac{t+A-A}{h_n}}v_{(1)}^2w(v)dv + \frac{h_n^2}{2}\frac{\partial^2 g_f}{\partial x_{(2)}^2}(t)\int_{\frac{t+A-A}{h_n}}v_{(2)}^2w(v)dv$$
$$+\frac{h_n^2}{2}\frac{\partial^2 g_f}{\partial x_{(1)}\partial x_{(2)}}(t)\int_{\frac{t+A-A}{h_n}}v_{(1)}v_{(2)}w(v)dv + O(h_n^3)I(h_n)$$

Pour n assez grand, $C \times C \subset (t + A - A)h_n^{-1}$, et puisque w est un noyau produit symétrique, on a

$$(1) \quad I(h_n) = \int_C w_0(z)dz = 1,$$

$$(2) \quad \int_{\frac{t+A-A}{h_n}} v_{(1)}w(v)dv = \int_{\frac{t+A-A}{h_n}} v_{(2)}w(v)dv = 0,$$

$$(3) \quad \int_{\frac{t+A-A}{h_n}} v_{(1)}^2 w(v)dv = \int_{\frac{t+A-A}{h_n}} v_{(2)}^2 w(v)dv = \int_C z^2 w_0(z)dz,$$

$$(4) \quad \int_{\frac{t+A-A}{h_n}} v_{(1)}v_{(2)}w(v)dv = \left(\int_C z w_0(z)dz\right)^2 = 0,$$

Par conséquent, on obtient

$$\mathbb{E}[\hat{g}_{f,n}(t)] = g_f(t) + \frac{h_n^2}{2}\left(\frac{\partial^2 g_f}{\partial x_{(1)}^2}(t) + \frac{\partial^2 g_f}{\partial x_{(2)}^2}(t)\right)\int_C z^2 w_0(z)dz + O(h_n^3)$$

□

Si l'on étudie cet estimateur dans le cas où $f(m_1, m_2) = m_1 m_2$, on a une version de l'indice I_{DO} convenablement normalisée qui s'applique à certains processus ponctuels marqués hétérogènes. Le résultat du théorème montre que notre estimateur est asymptotiquement sans biais et que la correction de bord joue un rôle primordial. En effet, sans correction de bord on obtient que l'estimateur noté $\hat{g}_{f,n,scb}$ est asymptotiquement biaisé.

$$\mathbb{E}[\hat{g}_{f,n,scb}(t)] = \int_{A-A}\int_{A\cap(A-u)} \frac{w_{h_n}(t-u)}{|A|} g_f(u)dvdu \to \frac{|A \cap (A-t)|}{|A|} g_f(t).$$

Ce résultat asymptotique est une première étape dans l'étude des propriétés asymptotiques de notre indice de concentration.

5.5 Conclusion et perspectives

Le cadre mathématique de la théorie des processus ponctuels a permis de présenter les indices de concentration basés sur les distances, introduits par Duranton et Overman (2005) et Marcon et Puech (2007) en économétrie, comme des estimateurs de caractéristiques du second ordre de processus ponctuels spatiaux marqués. Nous avons défini un nouvel indice de concentration construit comme l'estimateur d'une nouvelle caractéristique du second

ordre pour certains processus ponctuels marqués inhomogènes, en considérant une fonction des marques de la forme $f(m_1, m_2) = k(m_1)q(m_2)$.

Nous avons prouvé que notre nouvel indice de concentration est asymptotiquement sans biais et une perspective de travail intéressante serait d'étudier le comportement asymptotique de la covariance entre $\hat{g}_{f,n}(t_1)$ et $\hat{g}_{f,n}(t_2)$. Ceci permettrait d'avoir un résultat de convergence en probabilité de notre estimateur. Une autre perspective d'extension serait d'étudier les propriétés asymptotiques de notre estimateur défini lorsque g_f est invariant par rotation. D'un point de vue numérique, nous envisageons une étude basée sur des simulations pour illuster les résultats théoriques de convergence.

Enfin, dans une approche comparative, des exemples simulés de répartition de secteurs d'activités permettraient de juger de la qualité de notre indice de concentration, par rapport aux indices de concentration agrégés et ceux existants basés sur les distances.

Chapitre 6

Conclusion

Dans cette thèse, nous nous sommes intéressés à l'apport de la théorie des processus ponctuels spatiaux pour des problèmes de positionnement optimal, ainsi que pour la définition de nouveaux indices de concentration basés sur les distances en économétrie.

Dans un premier temps, l'étude de cas relative au positionnement optimal d'une nouvelle caserne de pompiers a permis de proposer une nouvelle méthode de résolution prenant en compte la nature aléatoire des positions et de leurs caractéristiques. Découpée en une phase de modélisation et d'optimisation, cette méthode a permis de juger de la variabilité de la solution optimale et de permettre de traiter des bases de données volumineuses. La prise en compte de l'aléa est souvent limitée en recherche opérationnelle et l'utilisation de la théorie des processus ponctuels spatiaux permet d'introduire une méthode d'optimisation stochastique innovante pouvant prendre en compte d'éventuelles intéractions des positions. L'analyse exploratoire de la base de donnée des sinistres demande un travail considérable et on peut encore se poser quelques questions pour améliorer l'étude qui en est faite. Ainsi, il peut être intéressant de chercher à étudier globalement l'indépendance ou à éviter le découpage en classes de la marque pour tester la dépendance entre marques et positions. Ces perspectives d'amélioration montrent le degré de complexité de notre jeu de données et semblent non triviales à mettre en œuvre. De plus, la phase de modélisation est rendue difficile dans notre étude de cas par l'erreur de positionnement et la forte inhomogénéité de la répartition des positions. Dans ce cas, une perspective d'amélioration serait de considérer un modèle hiérarchique où les positions des sinistres seraient issues d'un processus ponctuel latent et ensuite agrégées à leur nœud le plus proche sur la grille prédéfinie par les pompiers. La correction de l'erreur d'agrégation par la prise en compte du modèle de processus ponctuel latent

est une possibilité d'amélioration de la phase d'analyse pour notre étude de cas. A ma connaissance, ce problème d'erreur d'agrégation a été considéré en économétrie spatiale mais pas en théorie des processus ponctuels et représente une perspective de recherche intéressante, notamment pour l'étude de caractéristiques du premier et second-ordre d'un processus ponctuel.

Dans un second temps, des résultats de convergence forte ont été démontrés pour les solutions optimales de problèmes de positionnement optimaux. Dans le cadre de notre étude de cas, la convergence presque sure des solutions optimales empiriques a été prouvée sous l'hypothèse i.i.d. des couples formés par les positions et les marques, ainsi que sous d'autres hypothèses naturelles. Il paraît raisonnable de penser généraliser ce théorème à un processus de Poisson. Limité au cadre i.i.d. par la théorie de Vapnik-Cervonenkis sur la loi des grands nombres uniforme, la perspective de recherche immédiate est de travailler sur la monographie de Peskir (2000) afin de généraliser notre résultat à des données dépendantes. Cette étape pourrait constituer un grand pas vers l'écriture d'un théorème de convergence pour des processus ponctuels spatiaux dans le cadre asymptotique avec domaine d'observation borné. Enfin, dans le cas d'un problème de positionnement optimal dérivé du problème transport de Monge-Kantorovich, nous avons démontré la convergence presque sure des positions optimales empiriques.

La théorie des processus ponctuels spatiaux est utilisée depuis peu en économétrie où de nouveaux indices de concentration basés sur les distances ont été introduits. Nous avons présenté un indice de concentration s'écrivant comme l'estimateur d'une nouvelle caractéristique du second ordre d'un processus ponctuel marqué. L'avantage de cet indice est de prendre en compte l'inhomogénéité du premier ordre, la dépendance entre marques et positions et une correction de bord, de façon à pouvoir comparer les indices de concentration du second ordre entre les différents secteurs d'activité. Dans un cadre asymptotique avec fenêtre d'observation bornée, nous avons prouvé que notre estimateur est asymptotiquement sans biais dans une forme anisotrope. Une perspective de recherche serait d'étudier le comportement asymptotique du terme de covariance afin d'en déduire une éventuelle convergence en probabilité. Ensuite, l'extension au cadre isotrope permettrait d'établir la convergence de notre indice de concentration vers la caractéristique du second ordre du processus ponctuel marqué. Enfin, une étude approfondie sur des données simulées permettrait de souligner la pertinence de cet indice et de l'utiliser ensuite sur des données réelles. L'introduction de ces indices de concentration permet aussi d'envisager leur utilisation pour la détection d'agrégats lorsque l'on considère des réalisations de processus ponctuels marqués.

Bibliographie

[1] Arbia, G. et Piras, G. (2009) Spatial Concentration. *Document de travail.*

[2] Baddeley, A., Moller, J. et Waagepetersen, R. (2000) Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* **54** 329-350.

[3] Baddeley, A. et Turner, R. (2005) spatstat : an *R* package for analysing spatial point patterns. *Journal of Statistical Software* **12** 1-42.

[4] Baddeley, A. et Turner, R. (2006) Modelling spatial point patterns in *R*. In *Case studies in spatial point process modeling* **185** 23-74. Springer, New York.

[5] Bar-Hen, A. and Picard, N. (2006) Simulation study of dissimilarity between point process. *Computational Statistics* **21** 487-507.

[6] Benes, V., Bodlak, K., Moller, J. et Waagepetersen, R. (2005) A case study on point process modelling in disease mapping. *Image Analysis and Stereology* **24** 159-168.

[7] Berman, M. et Diggle, P. (1989) Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society, Series B.* **51** 81-92.

[8] Bickel, P., Klaasen, C., Ritov, Y. and Wellner, J. (1993) Efficient and adaptative estimation for semiparametric models. Baltimore and London : The John Hopkins University Press.

[9] Bischoff, M. et Dächert K. (2007) Allocation Search Methods for a Generalized Class of Location-Allocation Problems. *European Journal of Operational Research*, Article In Press.

[10] Bolton, R. et Morgan, F. (2002). Hexagonal economic regions solve the location problem. The American Mathematical Monthly, 109 (2), 165–172.

[11] Bonneu, F. (2007) Exploring and modelling firemen emergencies with a spatio-temporal marked point process approach. *Case Studies in Business, Industry and Government* **1** (2). http ://www.bentley.edu/csbigs/.

[12] Bonneu, F. and Daouia, A. (2008) The consistency of the empirical optimal conditional locations. *Submitted for publication.*

[13] Bonneu, F. et Thomas-Agnan, C. (2008) Spatial point process models for location-allocation problems. *To appear.*

[14] Bouchitté, G., Jimenez, C. and Rajesh, M. (2002). Asymptotics of an optimal location problem. C. R. Math. Acad. Sci. Paris, 335 (10), 853–858.

[15] Brimberg, J. et Mladenovic N. (1996) Solving the continuous location-allocation problem with Tabu search. *Studies in Locational Analysis* **8** 23-32.

[16] Brimberg, J., Hansen, P., Mladenovic N. et Taillard, E. (1997) Improvements and Comparison of Heuristics for Solving the Multisource Weber Problem. *Les Cahiers du GERAD*, Montreal, Canada.

[17] Chen, P., Hansen, P., Jaumard, B. et Tuy, H. (1998) Solution of the Multisource Weber and Conditional Weber Problems by D.-C. Programming. *Operations Research* **46** (4) 548-562.

[18] Chen, X., Linton, O. and van-Keilegom, I. (2008) Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71** (5) 1591-1608.

[19] Cliff, A. et Ord, J.K. (1981) *Spatial processes. Models and applications.* Pion, Londres.

[20] Cooper, L. (1964) Heuristic Methods for Location-Allocation Problems. *SIAM Review* **6** (1) 37-53.

[21] Cooper, L. (1974) A random locational equilibrium problem. *Journal of Regional Science* **14** 47-54.

[22] Cowling, A., Hall, P. et Phillips, M.J. (1996) Bootstrap confidence regions for the intensity of a Poisson point process. *Journal of the American Statistical Association*, **91** (436) 1516-1524.

[23] Cox, D.R. Some statistical models related with series of events. *Journal of the Royal Statistical Society Series B* **17** 129-164.

[24] Cressie, N. (1993) *Statistics for spatial data.* Wiley, New York. 2nd edition.

[25] Cucala, L. (2008) Intensity estimation for spatial point processes observed with noise. *Scandinavian Journal of Statistics* **35** (2) 322-334.

[26] Cuesta-Albertos, J.A., Matran, C., Rachev, S. T. and Rüschendorf, L. (1996). Mass transportation problems in probability theory. Math. Scientist, 21, 37–72.

[27] Curry, B. et George, K. D. (1983) Industrial Concentration : A Survey. *The Journal of Industrial Economics* **31** (3) 203-255.

[28] Dell'Amico, M. and Toth, P. (2000). Algorithms and codes for dense assignment problems : the state of the art. Discrete applied mathematics, 1000, 17–48.

[29] Daskin, M. (1982) Application of an expected covering model to emergency medical service system design. *Decision Sciences* **13** 416-439.

[30] Davison, A.C. et Hinkley, D.V. (1997) *Bootstrap Methods and Their Applications.* Cambridge University Press, Cambridge.

[31] Dawkins, C. J. (2004) Measuring the Spatial Pattern of Residential Segregation. *Urban Studies* **41** (4) 833-851.

[32] Diggle, P.J. (1985) A kernel method for smoothing point process data. *Applied Statistics* **34** 138-147.

[33] Diggle, P.J. (2003) *Statistical Analysis of spatial point patterns.* Arnold.

[34] Diggle, P.J. (2006) *Statistical analysis of spatio-temporal point process data.* In Finkenstadt, B., Held, L., and Isham V., editors, Semstat2004, 1-45. CRC Press, London.

[35] Diggle, P.J. (1988) Equivalence of smoothing parameter selectors in density and intensity estimation. *Journal of the American Statistical Association* **83** 793-800.

[36] Diggle, P.J., Zheng, P. et Durr, P. (2005) Nonparametric estimation of spatial segregation in a multivariate point process : bovine tuberculosis in Cornwall, UK. *Journal of the Royal Statistical Society, Series C.* **54** 645-658.

[37] Diggle, P.J., Gomez-Rubio V., Brown, P.E., Chetwynd, A.G. et Gooding, S. (2007) Second-order analysis of inhomogeneous spatial point processes using case-control data. *Biometrics* **63** (2) 550-557.

[38] Drezner, Z. (1985) Sensitivity analysis of the optimal location of a facility. *Naval Research Logistics Quarterly* **32** 209-224.

[39] Dudley, R.M. (1978) Central limit theorems for empirical measures. *The annals of Probability* **6** (6) 899-929.

[40] Duranton, G. et Overman, H.G. (2005) Testing for localization using micro-geographic data. *Review of Economic Studies* **72** 1077-1106.

[41] Ehrgott, M. (2005) *Multicriteria optimization* Second edition. Springer-Verlag, Berlin.

[42] Gangbo, W., and McCann, R.J (1996). The geometry of optimal transportation. Acta Mathematica 177, 2, 113–161.

[43] García-Soidán, P.H., Febrero-Bande, M. and González-Manteiga, W. (2004) Nonparametric kernel estimation of an isotropic variogram. *Journal of Statistical Planning and Inference* **121** 65-92.

[44] Goldberg, J et Paz, L. (1991) Locating emergency vehicle bases when service time depends on call location. *Transportation Science* **25** 264-280.

[45] Guan, Y. et Loh, J.M. (2007) A thinned block bootstrap procedure for modeling inhomogeneous spatial point patterns. *Journal of the American Statistical Association.* **102** 1377-1386.

[46] Guan, Y., Sherman, M. and Calvin, J.A. (2007) On asymptotic properties of the mark variogram estimator of a marked point process. *Journal of Statistical Planning and Inference* **137** 148-161.

[47] Hall P. (1985) Resampling a coverage process. *Stochastic processes and their applications* **20** 231-246.

[48] Holland, J.H. (1975) Adaptation in natural artificial systems. University of Michigan Press : Ann Arbor, MI.

[49] Houck, C.R., Joines, J.A. et Kay M.G. (1996) Comparison of genetic algorithms, random restart and two-opt switching for solving large location-allocation problems. *Computers & operations research* **23** (6) 587-596.

[50] Houdebine, M. (1999) Concentration géographique des activités et spécialisation des départements français. *Economie et Statistique* **326-327** 189-204.

[51] Huang B., Liu N. et Chandramouli M. (2006) A GIS supported Ant algorithm for the linear feature covering problem with distance constraints. *Decision Support Systems* **42** 1063-1075.

[52] Jonker, R. and Volgenant, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing, 38, 325–340.

[53] Jonsdottir, K., Hahn U. et Jensen, E. (2004) Inhomogeneous spatial point processes with a view to spatio-temporal modelling. In *Proceedings of the Conference on Spatial Point Process Modelling and its Applications, Castellon, April 2004* (eds. A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan), 131-136.

[54] Kantorovich, L.V. (1942) On the translocation of masses. *C.R. (Dokl.) Acad. Sci. URSS 37* 199-201.

[55] Kantorovich, L.V. (1948) On a problem of Monge. *Uspekhi Mat. Nauk. 3* 225-226.

[56] Kutoyants, Y. (1998) *Statistical inference for spatial Poisson processes.* Lectures Notes in Statistics. 134, Springer.

[57] Logendran, R. et Terrell, M.P. (1988) Uncapacitated plant location-allocation problems wth price sensitive stochastic demands. *Computers & operations research* **15** (2) 189-198.

[58] Loh, J.M. et Stein, M.L. (2004) Bootstrapping a spatial point process. *Statistica Sinica* **14** 69-101.

[59] Marcon, E. et Puech, F. (2007) Measures of the geographic concentration of industries : improving distance-based methods. Preprint.

[60] Mateu, J. (2000) Second-order characteristics of spatial marked processes with applications. *Nonlinear Analysis : Real World Applications* **1** 145-162.

[61] Maurel, F. et Sédillot, B. (1999) A measure of the geographic concentration in french manufacturing industries. *Regional Science and Urban Economics* **29** 575-604.

[62] McAsey, M. and Mou, L. (1998). Optimal locations and the mass transport problem. "Monge Ampere Equation : Applications to Geometry and Optimization," Contemporary Mathematics 226, edited by L. Caffarelli and M. Milman, 131-148.

[63] Moller, J., Syversen, A.R. et Waagepetersen, R.P. (1998) Log Gaussian Cox processes. *Journal of Applied Probability* **25** 451-482.

[64] Moller, J. et Waagepetersen, R.P. (2004) *Statistical inference and simulation for spatial point processes.* vol. 100. Chapman & HallCRC.

[65] Monge, G. (1781) Mémoire sur la théorie des déblais et des remblais. *Histoire de l'académie Royale des sciences de Paris* 666-704.

[66] Pakes, A. and Pollard, D. (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* **57** (5) 1027-1057.

[67] Peskir, G. (2000) From uniform laws of large numbers to uniform ergodic theorems. *Lecture Notes Series, Dept. Math. Univ. Aarhus* **66**.

[68] Pollard, D. (1984) *Convergence of stochastic processes.* Springer-Verlag, New York.

[69] Rachev, S. T. (1985). The Monge-Kantorovich mass transference problem and its stochastic applications. Theory of probability and its applications, 29, 647-676.

[70] Rachev, S. T. (1991). Probability metrics and the stability of stochastic models. John Wiley, New York.

[71] Rachev, S. T. and Rüschendorf, L. (1998). Mass Transportation Problems. Vol. I : Theory, Vol.II : Applications. Probability and its applications. Springer-Verlag, New-York.

[72] ReVelle, C. (1991) Siting ambulance and fire companies : new tools for planners. *Journal of the American Planning Association* **57**(4) 471-484.

[73] ReVelle, C.S. et Eiselt, H.A. (2005) Location analysis : a synthesis and survey. *European Journal Of Operational Research* **165**(1) 1-19.

[74] Ripley, B.D. (1976) The second-order analysis of stationary point processes. *Journal of Applied Probability* **13** 255-266.

[75] Ripley, B.D. (1977) Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society Series B* **39** 172-212.

[76] Schlather, M. (2001) On the second-order characteristics of marked point processes. *Bernoulli* **7**(1) 99-117.

[77] Schlather, M., Ribeiro, P.J. et Diggle, P.J. (2004) Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society, Series B* **66** 79-93.

[78] Schoenberg, F.P. (2004) Testing separability in spatial-temporal marked point processes. *Biometrics* **60** 471-481.

[79] Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics, Wiley, New York.

[80] Serra, D. et Marianov, V. (1998) The p-median problem in a changing network : The case of Barcelona. *Location Science* **6** 383-394.

[81] Silverman, B.W. (1986) Density estimation for statistics and data analysis. Chapman and Hall.

[82] Snethlage, M. (1999) Is bootstrap really helpful in point process statistics ? *Metrika* **49** 245-255.

[83] Snyder, L. (2006) Facility location under uncertainty : A review. *IIE Transactions* **38**(7) 537-554.

[84] Stoyan, D. et Stoyan, H. (1994) *Fractals, random shapes and point fields.* Wiley, Chichester.

[85] Valeyre, A. (1993) Mesures de dissemblances et d'inégalité interrégionales : principes, formes et propriétés. *Revue d'économie régionale et urbaine* **1** 17-53.

[86] van de Geer, S. (2000) Empirical processes in M-estimation, Cambridge university press.

[87] van der Vaart, A.W. (1998). Asymptotic Statistics. Cambridge University Press, Cambridge.

[88] van der Vaart et Wellner (1996) *Weak convergence and empirical processes.* With applications to statistics. Springer Series in Statistics. New York : Springer-Verlag.

[89] van Lieshout, M.N.M. (2000) *Markov point processes and their applications.* Imperial College Press, London.

[90] Vapnik, V.N. et Chervonenkis, A. Ya. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Application* **16** 264-280.

[91] Vapnik, V.N. et Chervonenkis, A. Ya. (1981) Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory Probab. Appl.* **16** 264-280.

[92] Villani, C. (2003). Topics in Optimal Transportation. Graduate Studies in Mathematics, 58. American Mathematical Society, Providence, RI.

[93] Volgenant, A. (1996). Linear and semi-assignment problems : a core oriented approach. Computers and Operations Research, 23 (10), 917–932.

[94] Waagepetersen, R.P. (2006) An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics* **63** (1) 252-258.

[95] Wand, M.P. et Jones, M.C. (1994) Multivariate plug-in bandwidth selection. *Computational Statistics* **9** 97-116.

[96] Zhou, J. et Liu, B. (2003) New stochastic models for capacitated location-allocation problem. *Computers and industrial engineering* **45** (1) 111-125.