Mesures de la concentration spatiale en espace continu : théorie et applications

Eric Marcon, Florence Puech

Mesures de la concentration spatiale en espace continu : théorie et applications

Éric Marcon [*] **et Florence Puech** [**]

L'agglomération des activités économiques est indéniable (Krugman, 1991) et chacun peut aisément citer des exemples de quartiers spécialisés au sein des villes ou des clusters d'activités par exemple. L'explication des phénomènes d'agglomération semble être à présent bien appréhendée théoriquement (Fujita *et al.*, 1999 ; Fujita et Thisse, 2002) mais les recherches empiriques ne semblent pas avoir atteint un tel stade de maturité (Rosenthal et Strange, 2004, Ellison *et al.*, 2010 ; Gibbons *et al.*, 2014). Durant la dernière décennie, de nombreuses recherches en économie spatiale ont porté sur les mesures de concentration géographique. Les économistes retenaient traditionnellement des mesures reposant sur un zonage du territoire (comme l'indice de Gini) mais des travaux récents ont montré que discrétiser l'espace pouvait engendrer des biais (Briant *et al.*, 2010). L'utilisation de mesures fondées sur les distances (séparant les entités analysées) et non sur un zonage est aujourd'hui recommandée (Combes *et al.*, 2006). Notre contribution méthodologique montre qu'une attention particulière doit encore être portée à la définition de la concentration spatiale pour évaluer l'agglomération des activités économiques.

À partir de la localisation des commerces de détail sur l'aire urbaine de Lyon notamment, nous montrons, en utilisant trois mesures de concentration récemment introduites en économie spatiale (K_d, D et M), que les résultats obtenus ne convergent pas systématiquement. Cette différence dans les estimations provient de la définition de la concentration spatiale retenue qui peut être absolue (présence importante d'activités), topographique (densité élevée d'activités) ou relative (surreprésentation de certaines activités). Nous recommandons alors que le choix de la mesure de concentration soit suffisamment motivé d'un point de vue théorique pour apprécier correctement le phénomène analysé et ainsi apporter une évaluation satisfaisante de la distribution étudiée.

Codes JEL : C10, C60, R12.

Mots-clés : économie géographique, indices de concentration spatiale, mesures fondées sur les distances.

Rappel :

Les jugements et opinions exprimés par les auteurs n'engagent qu'eux mêmes, et non les institutions auxquelles ils appartiennent, ni *a fortiori* l'Insee.

* *AgroParisTech, UMR EcoFoG, BP 709, F-97310 Kourou, Guyane française. Email : eric.marcon@ecofog.gf*
** *Université de Paris-Sud, RITM, 54 Boulevard Desgranges, 92331 Sceaux Cedex, France. Email : Florence.Puech@u-psud.fr*

Nous tenons à remercier les deux rapporteurs anonymes de la revue pour leurs commentaires constructifs, Christophe Baume et la Chambre de commerce et de l'industrie de Lyon sont également remerciés pour nous avoir autorisés à exploiter la base de données des localisations des commerces sur Lyon (France). Cet article a circulé sous le titre « La mesure de la concentration spatiale : un concept polymorphe ».

« *Mesurer adéquatement le mesurable. C'est une nécessité évidente puisque l'information géographique se traite d'une manière scientifique plus rigoureuse et plus fructueuse lorsqu'elle est mesurée correctement. Mais la mesure n'est pas seulement une question de méthode. Au contraire, c'est avant tout une question d'*interprétation *et de jugement géographiques. En effet, il appartient à l'utilisateur de juger, dans le champ de sa discipline, ce qui est mesurable et ce qui ne l'est pas ; d'apprécier à quel degré les diverses mesures qu'il effectue représentent véritablement le phénomène qu'il désire étudier. Cette appréciation est fondamentale car elle conditionne la validité des conclusions tirées d'une analyse quantitative ultérieure, si sophistiquée soit-elle* ».

Béguin (1979), p. 3.

Cette mise en garde d'Hubert Béguin concernant le traitement des données géographiques est loin d'être anodine mais elle n'a peut-être pas attiré suffisamment l'attention des économistes. L'objectif de cet article est de montrer que cet avertissement s'applique notamment à la mesure de la concentration géographique des activités économiques. L'agglomération des activités aux niveaux supranational, infranational et urbain est aujourd'hui un constat unanimement partagé dans notre discipline (Krugman, 1991 ; Fujita et Thisse, 2002). Or, si l'on souhaite comprendre les motivations des agents à se regrouper dans l'espace, une bonne appréhension du phénomène est nécessaire (Rosenthal et Strange, 2004). Les économistes ont pris conscience récemment qu'une réelle réflexion autour de ces mesures devait être entreprise (Combes et Overman, 2004 ; Combes *et al.*, 2006). De nombreuses mesures sont à leur disposition pour évaluer si les activités sont géographiquement concentrées ou dispersées mais malheureusement le choix de l'indice de concentration n'est encore que très peu discuté dans les études. Afin de montrer toute l'importance qu'il serait souhaitable d'accorder à l'outil, nous avons choisi d'illustrer les divergences de résultats obtenus à partir de trois mesures récemment mobilisées dans les études économiques : les fonctions K_d de Duranton et Overman (2005), D de Diggle et Chetwynd (1991) et M de Marcon et Puech (2010). Ces mesures ont la particularité de ne pas reposer sur un zonage géographique (comme les régions par exemple) mais sont définies à partir des distances entre les établissements localisés sur le territoire analysé. A partir de la distribution des commerces de détail sur l'agglomération lyonnaise, nous montrons que les niveaux de concentration obtenus peuvent être sensiblement différents en retenant des indices relativement proches sur le plan statistique. Nous expliquons ces écarts de résultats en soulignant les erreurs d'interprétation possibles et leurs conséquences économiques si le choix de la mesure de la concentration n'a pas reçu une attention suffisante[1].

La première section de l'article rappelle les enjeux économiques de la quantification des niveaux d'agglomération des activités. Dans une deuxième section, nous analysons les trois mesures privilégiées dans les études économiques récentes : K_d, D et M. Puis, après avoir présenté nos données, nous expliquons les écarts de résultats obtenus en appliquant ces trois mesures à deux secteurs d'activité : le commerce de détail de carburants et celui de l'habillement sur Lyon. Nous concluons l'article sur des mises en garde.

Mesurer la concentration spatiale des activités : quels enjeux ?

La démarche méthodologique proposée dans cet article répond à des enjeux économiques importants, à la fois positifs et normatifs.

L'enjeu positif est la nécessité de disposer d'une évaluation rigoureuse des niveaux de concentration spatiale des activités. Il est indispensable de travailler avec une mesure statistique fiable permettant de quantifier les disparités existantes mais aussi d'effectuer des comparaisons robustes des niveaux de concentration obtenus entre secteurs d'activité, entre les territoires et à différents points dans le temps[2]. Or disposer d'une mesure de concentration spatiale épurée de tous les biais statistiques est loin d'être évident. Ce n'est qu'à la fin des années 1990 que les économistes se sont réellement intéressés à cette question (voir notamment Houdebine, 1999). L'évaluation de la concentration était jusque-là trop souvent réduite à l'utilisation d'indices largement employés dans la littérature comme les indices de Gini (Combes *et al.*, 2006) mais qui peuvent se révéler limités pour l'appréhension du phénomène

1. Les termes de concentration spatiale et d'agglomération seront entendus comme des synonymes dans notre article et seront opposés au phénomène de dispersion.
2. Voir par exemple l'étude de Brülhart et Traeger (2005) concernant l'évolution des activités à un niveau international.

de concentration spatiale des activités. Un des problèmes inhérents aux deux indices précités est de ne pas prendre en compte les effets de la structure des industries. Ellison et Glaeser (1997) ont ainsi été les premiers à mettre en évidence que pour un secteur à rendements croissants, une évaluation de la concentration spatiale reposant sur la détection d'écarts entre la distribution de l'emploi dans ce secteur et une distribution théorique fondée sur une équirépartition de son emploi entre les différentes zones était problématique. Un autre problème est le fait que les propriétés de ces indices dépendent du type de découpage retenu pour le territoire. Nous développons ces points plus loin dans le texte et dans l'encadré 1.

Encadré 1

LITTÉRATURE EXISTANTE SUR LES INDICES DE CONCENTRATION

La mesure de la concentration spatiale certainement la plus utilisée en économie spatiale jusque dans les années 1990 est l'indice de Gini (1912) : elle consiste à mesurer la concentration des activités sur une discrétisation prédéfinie du territoire exactement de la même manière qu'on mesure la concentration du revenu au sein d'une population. L'espace analysé est divisé en plusieurs zones distinctes (comme le découpage régional). L'indice de Gini revient à estimer la répartition d'une variable d'intérêt comme l'emploi ou la production au sein de ces zones. Plus précisément, cet indice compare une statistique calculée à partir des données à sa valeur dans le cas d'une répartition égalitaire parfaite (équirépartition) entre les zones géographiques. Concrètement pour un secteur donné, cela peut revenir à estimer l'écart entre la distribution des activités de ce secteur et l'ensemble des activités industrielles par exemple (Krugman, 1991 ; Amiti, 1999 ; Midelfart-Knarvik *et al.,* 2002). Plus l'écart est important, plus les activités du secteur étudié sont géographiquement concentrées (sous-entendu par rapport aux autres activités industrielles).

La pertinence de l'indice de Gini a été remise en cause par l'article d'Ellison et Glaeser (1997). Ces derniers ont montré que cette équirépartition n'est qu'une hypothèse peu satisfaisante car difficilement atteignable. Si une industrie est composée d'établissements identiques, répartis de façon équiprobable dans des zones ayant les mêmes caractéristiques, la distribution observée n'est pas équirépartie : elle comporte une part de hasard qui engendre des fluctuations. Considérons une industrie à rendements d'échelle croissants : une telle industrie serait composée de peu d'établissements car il est coûteux de fragmenter la production. En conséquence, même en l'absence de toute force d'attraction, une concentration spatiale des activités serait détectée avec l'indice de Gini car la distribution des activités de ce secteur ne pourrait suivre la distribution des activités de l'ensemble des industries. L'aspect « grumeleux » (*lumpiness*) de la distribution de ce secteur résultant de la présence d'établissements de grande taille devrait être contrôlé pour éviter toute fausse impression de concentration spatiale. Ellison et Glaeser ont alors proposé un indice de concentration spatiale à partir d'une distribution aléatoire des établissements, selon une probabilité proportionnelle à la taille des zones géographiques. Dans ce cadre d'analyse, l'équirépartition parfaite est un extrême, l'autre extrême étant celui où tous les établissements des industries sont localisés dans une seule zone. Leur mesure de concentration est tirée d'un modèle théorique de choix de localisation dans lequel les avantages naturels et/ou les externalités générées par la présence d'autres établissements expliquent l'agglomération. Ellison et Glaeser ont été les premiers à proposer un cadre théorique dans la construction d'une mesure de concentration spatiale et un test de significativité de résultats. On comprend alors aisément que l'indice d'Ellison et Glaeser ait connu immédiatement un grand succès auprès des économistes (voir par exemple Maurel et Sédillot, 1999 ; Rosenthal et Strange, 2001 ; Holmes et Stevens, 2004). Il a également ouvert la voie à de nouvelles réflexions autour de la définition d'un indice de concentration satisfaisant en économie spatiale. Le lecteur intéressé pourra se reporter à la présentation détaillée et très claire des indices de Gini et d'Ellison et Glaeser donnée par Michel Houdebine dans un article publié dans cette même revue (Houdebine, 1999).

En effet, immédiatement après la publication de leur article, plusieurs économistes ont mis en évidence d'autres limites des indices retenus traditionnellement, notamment le fait que les indices de Gini et d'Ellison et Glaeser sont « aspatiaux » (Arbia, 2001). Ces indices reposent sur un découpage du territoire analysé en zones distinctes mais le positionnement des zones n'est pas pris en compte. Que des zones localisant de forts niveaux d'activité soient contigües ou très éloignées n'a aucune incidence sur les résultats obtenus : ces indices sont donc invariants par permutation des zones. Deux grandes réponses ont alors été apportées. Une première solution proposée par Guimarães *et al.* (2011) vise à compléter l'indice d'Ellison et Glaeser en intégrant une mesure d'autocorrélation spatiale. Une autre solution consiste à supprimer tout zonage et à traiter l'espace en continu. C'est cette seconde approche qui est retenue dans notre article.

Notons enfin qu'il existe d'autres mesures de concentration utilisées en économie spatiale mais leur usage est plus marginal (comme les indices d'entropie par exemple). Le chapitre 10 de l'ouvrage de Combes *et al.* (2006) dresse une présentation complète de tous les outils employés dans notre champ.

L'enjeu est également normatif puisque ces mesures constituent de véritables outils d'aide à la décision. Donnons deux exemples. Tout d'abord, connaître avec précision les niveaux de concentration spatiale d'activités est primordial en matière de politique d'aménagement du territoire. En effet, l'intensité des facteurs expliquant l'agglomération spatiale peut être mise en évidence en régressant les niveaux de concentration obtenus sur les principaux déterminants identifiés par la théorie économique tels que les liens *inputs-outputs*, la présence d'un large marché du travail ou encore l'existence d'externalités de connaissances (Marshall, 1890)[3]. Le niveau de concentration estimé par les indices de concentration doit donc être évalué avec la plus grande précision car l'intensité des facteurs expliquant cette agglomération en dépend. Ainsi, s'il est détecté que les externalités non pécuniaires constituent un facteur explicatif décisif dans l'explication de l'agglomération, cela justifie des financements octroyés pour favoriser l'échange informationnel, source d'innovation et d'augmentation de la productivité comme dans le cas de la politique de pôles de compétitivité menée actuellement en France. Un deuxième exemple illustrant l'importance des mesures de concentration en tant qu'outil d'aide à la décision est donné par l'étude de Barlet *et al.* (2011). Grâce à des résultats obtenus à partir d'indices de concentration, ces auteurs proposent une première identification des services « échangeables » c'est-à-dire des services qui ne sont pas encore échangés internationalement mais qui pourraient le devenir puisqu'aucune entrave technique à l'échange international n'est dans ce cas détectée[4]. L'analyse territoriale proposée dans cette étude permet alors d'estimer les conséquences d'un tel changement pour les différentes zones d'emploi en France métropolitaine. On comprend que les résultats des indices de concentration doivent être irréprochables pour qu'un tel l'exercice de prospective ait un sens. On notera d'ailleurs que les auteurs de cette étude ont procédé à des tests de robustesse de leurs résultats des niveaux de concentration obtenus.

Ces enjeux positifs et normatifs invitent à bien clarifier les critères devant guider le choix de l'indice de concentration à préconiser. Des listes des bonnes propriétés que tout indice de concentration idéal devrait respecter ont été proposées par plusieurs auteurs comme Duranton et Overman (2005), Combes et Overman (2004) ou encore, récemment, par Thomas-Agnan et Bonneu (2014). Cette démarche assez récente de la part des économistes démontre bien que l'évaluation de la concentration spatiale des activités économiques doit répondre à certaines exigences méthodologiques. Le respect de ces critères est maintenant très souvent mentionné dans les études (Duboz *et al.*, 2009). Nous montrons dans notre article à partir d'un cas d'étude sur les activités commerciales sur Lyon que ces critères doivent être encore précisés. En effet, les mesures de la concentration possèdent des propriétés statistiques spécifiques (différences de référentiel retenu pour évaluer les structures spatiales par exemple) et mettent en évidence un type de concentration spatiale bien déterminé. En s'appuyant uniquement sur les critères proposés aujourd'hui dans la littérature économique nous démontrons qu'il est aisé de faire des raccourcis erronés en mobilisant un outil inadapté au regard de la question traitée.

Du maillage territorial à un espace continu

Pourquoi recourir à un espace en continu ?

Les localisations d'établissements sont par nature des données individualisées. La mesure de la concentration géographique de ces établissements devrait par conséquent reposer sur des outils traitant l'espace de manière continue. Or, jusque dans les années 2000, les économistes n'ont pas eu un intérêt marqué pour cette approche (*cf.* encadré 1) : les territoires considérés étaient découpés en plusieurs zones distinctes et l'information à analyser était agrégée à un certain niveau géographique (départements, zones d'emploi…). Retenir un maillage territorial peut être motivé pour la simplicité des calculs à mettre en œuvre et la disponibilité des données. Cependant, discrétiser l'information est problématique car l'agrégation peut gommer des spécificités individuelles… ce que l'on cherche justement à mettre en évidence !

Pour illustrer ce point, considérons l'exemple donné dans le graphique I illustrant une distribution théorique d'établissements (les petits cercles) sur un territoire. Une évaluation de la

3. On pourra se reporter à Rosenthal et Strange (2004) pour une revue complète de la littérature ou encore à l'article d'Ellison et al. *(2010) concernant les phénomènes de co-agglomération.*
4. Leur analyse s'appuie sur une intuition de Krugman (1991) selon laquelle les services « échangeables » peuvent se concentrer géographiquement contrairement à la distribution géographique des services « non-échangeables » qui devrait suivre celle de la demande qui leur est destinée en biens intermédiaires ou en biens finals.

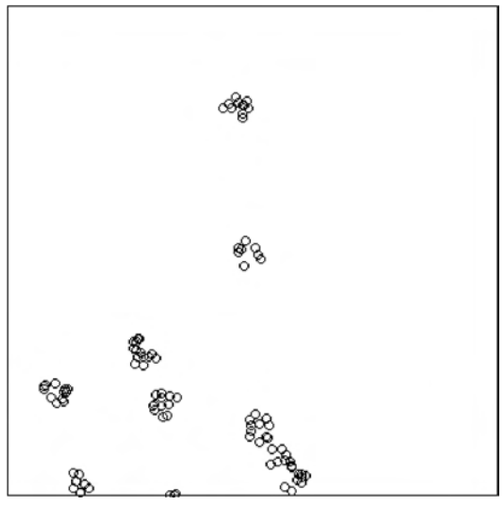
localisation de ces établissements devrait permettre de mettre évidence à la fois :

- Des regroupements d'établissements de type *clusters* puisque l'on peut observer des petits agrégats distincts composés d'une dizaine d'établissements ;

- L'attraction de la région sud-ouest puisqu'un grand nombre d'établissements s'y est localisé.

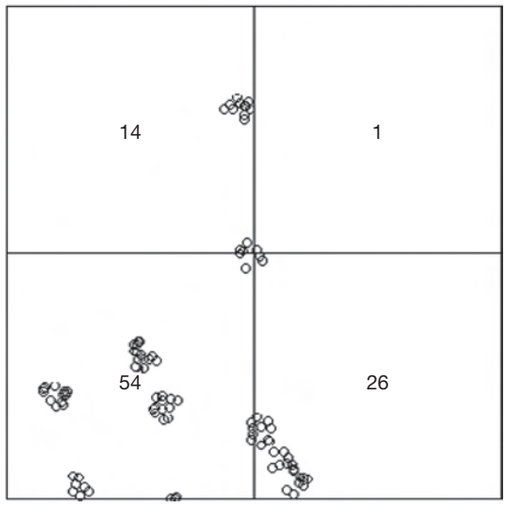
Appliquons à présent un découpage géographique de type carroyage au territoire analysé (*cf.* graphique II). Quatre zones distinctes de même taille sont identifiables : le cadran

Graphique I
Exemple de répartition fictive



Lecture : distribution théorique correspondant à une simulation d'un processus de Matérn (1960). On observe une distribution d'agrégats composés en moyenne de 10 points.
Champ : données simulées à partir d'un processus de Matérn.
Source : programmation des auteurs sous R à l'aide du package spatstat.

Graphique II
Évaluation de la concentration à partir d'un maillage carré



Lecture : même distribution théorique que dans le graphique 1 sur lequel un zonage de maille carrée a été appliqué. Les quatre cellules ont la même forme et la même aire.
Champ : données simulées à partir d'un processus de Matérn.
Source : programmation des auteurs sous R à l'aide du package spatstat.

nord-ouest regroupe ainsi 14 établissements, le cadran nord-est localise 1 établissement, le cadran sud-est en totalise 26 et le cadran sud-ouest en dénombre 54.

Nous constatons tout d'abord qu'un tel zonage discrétise l'espace de manière totalement arbitraire : les *clusters* peuvent ainsi se retrouver divisés géographiquement avec ce découpage. Un exemple est donné pour les établissements du *cluster* localisé au centre du territoire qui sont répartis entre les quatre zones. Puis, nous remarquons qu'une mesure de concentration spatiale appliquée à cette échelle géographique détecterait un fort regroupement d'activités (ce qui est effectivement le cas puisque le cadran sud-ouest localise plus de la moitié des établissements) mais ne mettrait pas en évidence l'attraction manifeste à plus petite échelle (présence de *clusters*).

Toutefois, retenir une échelle géographique plus fine aurait masqué l'attraction du cadran sud-ouest. Un unique niveau géographique semble donc problématique. Le positionnement des frontières tout comme le nombre de zones retenues peut être une source de biais dans les estimations. Ces deux problèmes sont les deux facettes de ce que l'on appelle le Problème des unités spatiales modifiables (*Modifiable Areal Unit Problem* – MAUP) lié à la discrétisation de données qui ne sont pas initialement agrégées (Arbia, 1989)[5].

Quelles sont les spécificités des mesures de concentration en espace continu ?

Jusqu'à la seconde moitié des années 90, les évaluations de la concentration spatiale des activités reposaient sur des mesures bien connues des économistes comme l'indice de Gini, ou l'indice d'Ellison et Glaeser qui se heurtent aux critiques relatives au zonage (*cf.* à nouveau l'encadré 1 pour une présentation succincte). C'est pour préserver la richesse de la distribution des localisations exactes des entités analysées (commerces, établissements industriels) mais aussi afin d'éviter tout biais lié à la MAUP résultant d'un zonage prédéfini du territoire, que de nouveaux outils permettant d'analyser la distribution des activités à toutes les échelles géographiques simultanément ont été progressivement proposés. Ces nouvelles mesures permettent de s'affranchir de tout découpage administratif (comme les départements) ou d'étude (comme les zones d'emploi) et d'analyser les entités de manière individuelle (et non agrégée) à partir de leur localisation exacte. L'évaluation des disparités spatiales repose alors directement sur une analyse effectuée à partir les distances séparant les entités. Aujourd'hui, ces outils dits « fondés sur les distances » ou en « espace continu » sont privilégiés pour évaluer les niveaux de concentration spatiale des activités économiques car ces outils présentent des qualités importantes. Notons que cette approche n'a pas été proposée par des économistes mais par des statisticiens (Ripley, 1981) et elle est depuis largement exploitée dans d'autres sciences comme en foresterie (Moeur 1993 ; Haase, 1995), en épidémiologie (Diggle et Chetwynd, 1991 ; Kingham *et al.*, 1995) ou en écologie par exemple (Harkness et Isham, 1983 ; Gaines *et al.*, 2000).

La mesure la plus connue aujourd'hui est certainement la fonction *K* proposée par B. Ripley (Ripley, 1976, 1977). L'idée de cette mesure est simple. Considérons la distribution des magasins au sein d'une ville. La fonction *K* permet de détecter si autour de chaque magasin il y a en moyenne plus ou moins de magasins qu'il n'y en aurait sous l'hypothèse nulle d'une distribution complètement aléatoire des magasins en ville. Si l'on détecte plus de magasins en moyenne que sous l'hypothèse nulle, on assimilera la structure spatiale observée à de la « concentration géographique » puisque les « magasins s'attirent ». En revanche si l'on détecte moins de magasins en moyenne que sous l'hypothèse nulle, on assimilera ce phénomène à de la « dispersion » puisqu'alors les « magasins se repoussent ». Nous voyons que la fonction *K* a pour but de détecter les « relations de voisinage » existantes entre les points analysés (attraction ou répulsion). Plus formellement, la fonction *K* se définit en référence à la théorie des processus ponctuels qui fournit un cadre statistique rigoureux à l'analyse des structures spatiales de points. Un processus ponctuel est l'équivalent d'une variable aléatoire dont les réalisations sont des semis de points dans un espace connu et délimité (le territoire analysé). Une façon intéressante de décrire un processus ponctuel dont on ne connaît pas la loi consiste à estimer ses propriétés de premier ordre (la densité) et de second ordre (les relations de voisinage). La distribution de référence à laquelle la distribution observée va être comparée pour la fonction *K* est une distribution complètement aléatoire, c'est-à-dire homogène (une densité constante en tout point du territoire, propriété de

5. Voir Briant et al. *(2010) pour une tentative d'estimation des biais statistiques de la MAUP sur données françaises.*

premier ordre) et indépendante (la position d'un point ne dépend pas de la position des autres, propriété de second ordre)[6]. Plusieurs exemples de résultats de la fonction *K* sur des cas théoriques simples sont proposés dans l'encadré 2.

De nombreuses études ont suivi pour proposer des extensions de cette première mesure fondée sur les distances. Citons notamment les fonctions *D* et *M* respectivement proposées par Diggle et Chetwynd (1991) et Marcon et Puech (2010). Ces auteurs ont préféré ne pas retenir l'hypothèse de référence d'une distribution complètement aléatoire (espace homogène) mais plutôt tenir compte d'une possible

6. Il s'agit donc d'une distribution de Poisson homogène.

Encadré 2

TROIS STRUCTURES SPATIALES SIMPLES EN ESPACE CONTINU

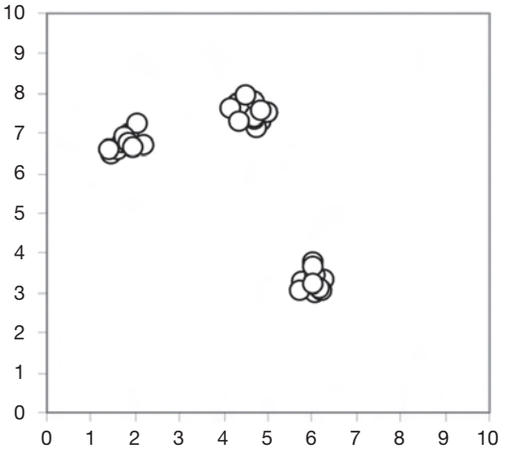
Les mesures développées en espace continu permettent d'identifier les « relations de voisinage » entre les points d'une distribution. Plus exactement, ces mesures permettent de détecter :

- Si les points s'attirent : la distribution est dite concentrée car des agrégats de points se forment, comme cela est observable sur le graphique A1.

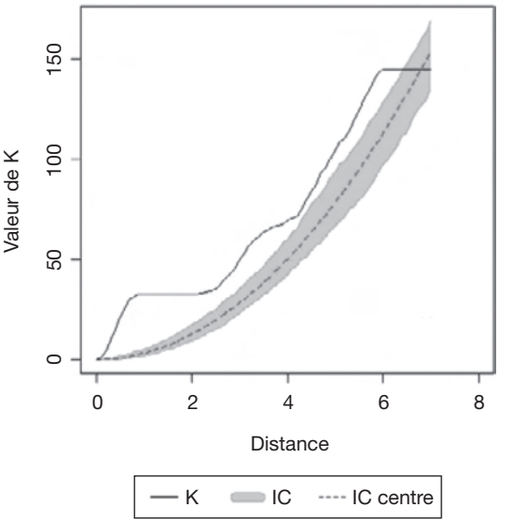
- Si les points se repoussent : la distribution est dite régulière ou dispersée. Le graphique B1 illustre le cas d'une dispersion maximale puisque les points se situent à égale distance les uns des autres (structure dite en nid d'abeilles).

- Si les points se répartissent « au hasard » les uns sans rapport aux autres : les points sont qualifiés d'indépendants puisque la position géographique des points ne dépend alors pas de la position des autres. Sur le graphique C1, les points ont été localisés de manière aléatoire sur l'aire d'étude, il n'y a bien dans ce cas aucune relation d'attraction ou de répulsion entre les points.

Graphique A1
Distribution concentrée



Graphique A2
Fonction *K* associée

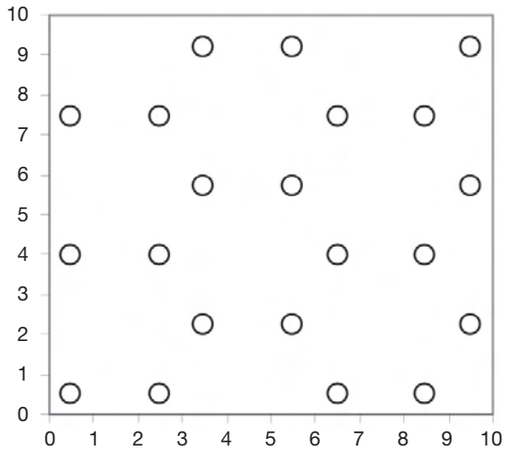


Lecture : à gauche, une distribution théorique de points localisés en un petit nombre d'agrégats est proposée sur un domaine 10 x 10. Elle correspond à la simulation d'un processus de Matérn c'est-à-dire un processus qui permet de générer des points « fils » aléatoirement autour de points « pères » (les centres des agrégats). Les paramètres du processus de Matérn sont déterminés ex ante *et sont égaux dans cet exemple à 0,03 pour la densité du processus de Poisson pour les centres des agrégats, 15 pour le nombre moyen de points par agrégat, à 0,5 pour le rayon des cercles et la fenêtre de simulation. Sur la figure de droite, sont représentés la fonction K associée ainsi que son intervalle de confiance global – IC (seuil de confiance 95 %, 1 000 simulations). Des niveaux de concentration spatiale sont détectés : la distribution étant plus agrégée qu'une distribution aléatoire. Graphiquement, K est au-dessus de son intervalle de confiance. Trois pics de concentration sont détectés : approximativement à la taille des agrégats (soit 0,5) et à plus longue distance (approximativement à 3,75 et à 6) correspondant à la distance moyenne entre agrégats (« agrégat d'agrégats »).*
Champ : données simulées d'un processus de Matérn.
Source : programmation des auteurs sous R avec le package *dbmss. La fonction K a été calculée avec un pas de calcul de 0,1 jusqu'à une distance de 7. Seul l'intervalle de confiance global a été reporté.*

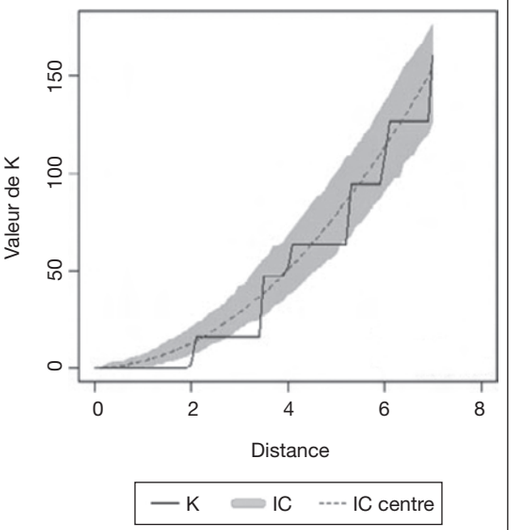
➔

Encadré 2 (suite)

Graphique B1
Distribution dispersée

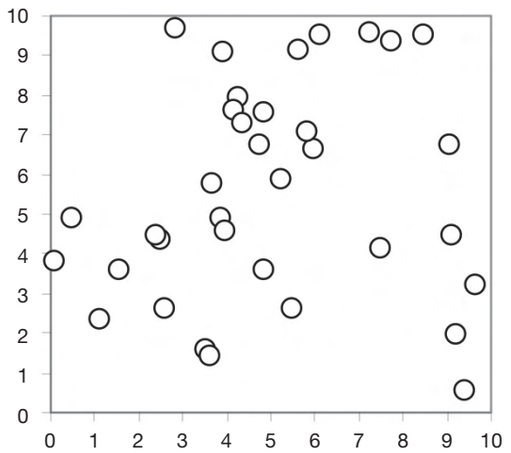


Graphique B2
Fonction *K* associée

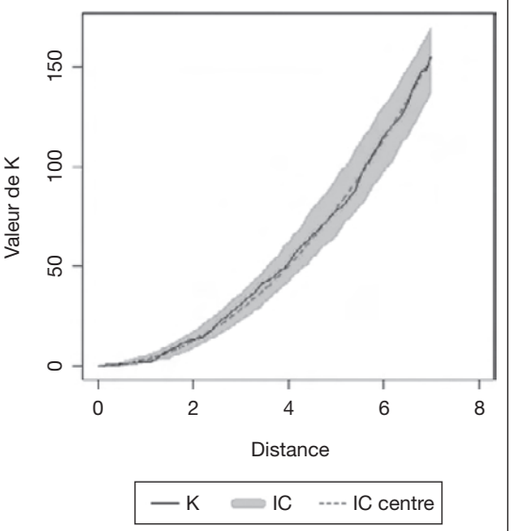


Lecture : à gauche, distribution de points parfaitement régulière (structure en nid d'abeilles) sur un domaine 10 x 10. À droite, la fonction K associée ainsi que son intervalle de confiance global – IC (seuil de confiance 95 %, 1 000 simulations). Lorsque les valeurs de K sont significatives, K est située en dessous de son IC indiquant une distribution dispersée. La distribution est effectivement plus régulière qu'une distribution aléatoire. Le premier pic négatif de dispersion correspond à la distance maximale entre les premiers voisins (côté de l'hexagone, soit une distance juste inférieure à 2). Les pics suivants sont des pics secondaires correspondant aux deuxièmes plus proches voisins etc.
Champ : données simulées d'une structure en nid d'abeilles.
Source : programmation des auteurs sous R avec le package *dbmss. La fonction K a été calculée avec un pas de calcul de 0,1 jusqu'à une distance de 7. Seul l'intervalle de confiance global a été reporté.*

Graphique C1
Distribution aléatoire



Graphique C2
Fonction *K* associée



Lecture : à gauche, distribution théorique de points localisés aléatoirement sur un domaine 10 x 10. Cette distribution est obtenue d'après une simulation d'un processus de Poisson homogène : les points sont tirés aléatoirement et indépendamment les uns des autres sur le domaine (la densité déterminée ex ante est égale à 0,3). À droite, fonction K associée ainsi que son intervalle de confiance – IC (seuil de confiance 95%, 1000 simulations). Quelle que soit la distance du voisinage considérée, aucune concentration spatiale ou dispersion significative n'est détectée (K étant comprise pour toutes les distances dans l'intervalle de confiance).
Champ : données simulées d'un processus de Poisson homogène.
Source : programmation des auteurs sous R avec le package *dbmss. La fonction K a été calculée avec un pas de calcul de 0,1 jusqu'à une distance de 7. Seul l'intervalle de confiance global a été reporté.*

➔

variation de densité au sein des territoires (espace hétérogène). L'espace homogène est en effet un cadre d'analyse peu pertinent pour évaluer la distribution des activités économiques : l'existence de zones non constructibles (montagnes, lacs) ou la tendance naturelle des activités économiques à se regrouper nécessite la prise en compte de l'hétérogénéité de l'espace. Duranton et Overman (2005) ont également proposé une nouvelle mesure spécialement développée pour mesurer la concentration spatiale industrielle. Bien qu'elle se nomme K_d, la mesure introduite par ces auteurs n'a aucun lien avec la fonction originelle *K* de Ripley[7].

Deux choix structurants : définir les notions de voisinage et de concentration

Nous venons de voir que les caractéristiques de l'espace sont importantes. La définition de l'espace considéré doit encore être précisée. L'évaluation les disparités existantes entre les territoires nécessite d'avoir préalablement déterminé d'une part la notion de voisinage et d'autre part le type de concentration (relative, topographique ou absolue) retenus. Évaluer la concentration spatiale en « espace continu » consiste à étudier le voisinage moyen des points d'intérêt (par exemple les établissements appartenant à un secteur donné). Le terme de « voisinage » peut recouvrir deux réalités : selon l'objet de son étude, le praticien peut vouloir analyser le voisinage des établissements à une distance donnée ou dans un rayon donné. Dans le premier cas, le voisinage sera évalué dans une couronne, dans le second cas sur un disque. Le graphique III montre sur un exemple la différence entre ces deux approches. Sur le graphique de gauche, le « voisinage » du point centre dans un rayon *r* est composé de trois voisins (tous les points localisés sur un disque de rayon *r*). Sur le graphique de droite, un seul voisin appartient au « voisinage » du point centre à une distance *r*. Il est localisé dans la couronne à une distance *r* du point centre (si d'autres points avaient été localisés sur cette couronne, ils auraient aussi été considérés comme appartenant au voisinage du point centre).

Concrètement, cette distinction nous permet de définir deux grandes familles de mesures en continu : les fonctions de densité ou les fonctions cumulatives. Des exemples théoriques sur

7. D'autres mesures fondées sur les distances en espace hétérogène existent (voir par exemple Baddeley et al.*, 2000) et sont présentées en détail dans Marcon et Puech (2014).*

Encadré 2 (suite)

La fonction *K* proposée par Ripley (1976, 1977) permet de détecter ces structures spatiales. Techniquement, la fonction *K* peut être estimée pour toute distance *r* en calculant l'espérance du nombre de voisins dans un rayon *r* rapportée à la densité observée sur le territoire. Un intervalle de confiance des résultats est simulé par la méthode de Monte-Carlo en générant un certain nombre de distributions aléatoires et indépendantes (le seuil de confiance généralement retenu étant de 95 %). L'intervalle de confiance de la fonction *K* est centré sur la valeur de référence $\pi.r^2$ puisque le nombre attendu de voisins dans un disque de rayon *r* est égale à la densité.$\pi.r^2$ (numérateur de la fonction *K*). Les fonctions estimées pour les trois cas de cet encadré sont représentées sur le graphique A2 pour la distribution agrégée (concentrée), sur le graphique B2 pour la distribution régulière (dispersée) et sur le graphique C2 pour la distribution aléatoire. Pour les distributions agrégée et régulière, les intervalles de confiance globaux ont été obtenus à partir de 1 000 simulations et des résultats significatifs au niveau de risque de 5 % sont observés : les fonctions *K* estimées sont en dehors de leur intervalle de confiance. Dans le cas de la distribution agrégée, la fonction *K* se situe au-dessus de l'intervalle de confiance de l'hypothèse nulle. Pour de faibles distances, les points ont en moyenne plus de voisins que sous l'hypothèse nulle d'une distribution aléatoire de points, puisque les points sont localisés dans les agrégats. Le plus fort niveau de concentration est observé à une distance égale à la taille des agrégats. Au contraire, pour la distribution dispersée, la fonction *K* détecte qu'il y a en moyenne moins de points dans le voisinage des points qu'il y en aurait sous l'hypothèse nulle d'une distribution entièrement aléatoire de points sur toute l'aire d'étude. La distribution régulière est détectée par des valeurs de *K* négatives et en dehors de l'intervalle de confiance. Graphiquement, un maximum de régularité est observé lorsque la dispersion entre les points est la plus forte. Ainsi, des « pics » négatifs sont nettement visibles juste avant la taille de la maille (côté de l'hexagone), puis juste avant les deuxièmes plus proches voisins etc. Enfin, dans le cas de la distribution aléatoire, aucune concentration ou dispersion n'est détectée, la fonction *K* reste dans l'intervalle de confiance quelle que soit la distance considérée.

les conséquences de cette distinction sont donnés dans Marcon et Puech (2010).

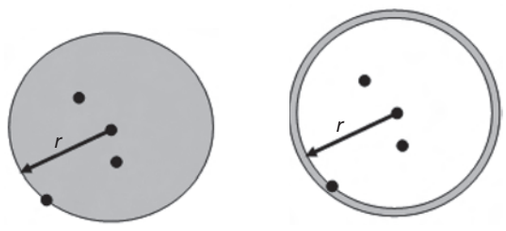
Ces deux notions de voisinage peuvent être combinées à trois appréhensions distinctes de la notion de concentration : « absolue », « topographique » et « relative » (l'apport de l'article de Brülhart et Traeger, 2005, sur ce point est fondamental).

Commençons tout d'abord par les *mesures dites absolues*. Ces dernières reposent sur un simple comptage du nombre de voisins d'intérêt localisés sur la surface d'un disque (pour les fonctions cumulatives) ou d'une couronne (pour les fonctions de densité). Ces mesures absolues sont à préconiser par exemple si la question d'étude porte sur la fréquence d'apparition des voisins d'intérêt (généralement les magasins d'un secteur d'activité donné). Ce calcul n'intègre donc pas de valeur de référence. Si tel était le cas, on parlerait alors de mesures topographiques ou de mesures relatives (même si ces deux notions ne sont pas équivalentes).

Les *mesures topographiques* retiennent comme valeur de référence l'espace physique : le nombre de voisins d'intérêt est divisé par une surface pour obtenir une densité. Pour les fonctions en continu, l'unité d'aire est donc la surface d'un disque ou d'une couronne selon le choix de la définition de voisinage adoptée. La concentration topographique sera à préconiser si l'on s'interroge par exemple sur la densité des magasins d'habillement sur un territoire ; si les magasins d'habillement sont regroupés spatialement, une concentration topographique sera détectée. Ce type de mesure fait implicitement l'hypothèse que le territoire analysé est homogène c'est-à-dire avec la même probabilité de localisation en tout lieu. Cette hypothèse peut être jugée trop forte : il conviendrait alors de prendre en compte l'hétérogénéité de l'espace. Des mesures topographiques en espace hétérogène existent. L'idée est de comparer une densité à une autre densité par différence. Concrètement, on retient des mesures qui permettent d'effectuer la différence entre les densités de voisins du secteur d'intérêt et de la distribution de référence (l'ensemble des secteurs d'activité). Une autre possibilité serait de comparer les densités non par différence mais en faisant leur rapport. Si tel est le cas, la surface de référence disparaît et on obtient *une mesure dite relative*. Concrètement, une surreprésentation ou sous-représentation d'une activité par rapport à une tendance qui est en général l'ensemble des activités peut dès lors être clairement établie.

Notons qu'avant l'article de Brühart et Traeger (2005), généralement les mesures absolues étaient dans la littérature uniquement opposées aux mesures relatives (Haaland *et al.*, 1999). Ces dernières étaient définies comme des mesures visant à comparer la distribution des voisins d'intérêt à une autre variable. Le regain d'intérêt des économistes dans la notion de la concentration spatiale d'une part et l'utilisation de mesures topographiques en économie spatiale d'autre part ont permis de clarifier et de préciser le vocabulaire utilisé. Enfin, toutes les mesures utilisées en économie spatiale pour évaluer la concentration en espace continu peuvent être classifiées en utilisant les notions de concentration topographique (homogène et hétérogène) absolue et relative (*cf.* la typologie proposée par Marcon et Puech, 2014).

Graphique III
Deux notions de voisinage



Lecture : une distribution théorique de points est considérée (points noirs sur les deux graphiques). Pour le point centre, le voisinage peut être assimilé au nombre de points localisés dans un rayon r *c'est-à-dire sur un disque de rayon* r *centré sur le point centre (graphique de gauche). Une autre notion de voisinage consisterait à dénombrer le nombre de points localisés exactement à un rayon* r *du point centre soit dans une couronne de largeur infinitésimale* dr *localisée à une distance* r *du point centre (graphique de droite). Les zones grisées représentent à gauche la surface du disque de rayon* r *et à droite la surface de la couronne à une distance* r *du point centre.*
Source : les auteurs.

Présentation des trois mesures retenues

Nous pouvons à présent justifier notre choix de retenir les trois mesures en espace continu suivantes : les fonctions *D* de Diggle et Chetwynd (1991), K_d de Duranton et Overman (2005) et *M* de Marcon et Puech (2010). Tout d'abord, nous allons voir que ces trois fonctions mettent chacune en évidence un type de concentration différent : topographique (hétérogène) pour *D*, relative pour *M* et absolue pour K_d. Des comparaisons de résultats paraissent dès lors indispensables pour montrer les avantages et les limites de chaque définition de la concentration spatiale retenue. Ces trois fonctions ont été choisies récemment dans la littérature pour évaluer les disparités entre les territoires ; parmi les applications, on notera l'application de K_d par Duranton et Overman (2008), le choix de la fonction *D* par Arbia *et al.* (2008) ou encore la fonction *M* retenue par Jensen et Michel (2011). La mesure K_d de Duranton et Overman semble toutefois plus souvent plébiscitée dans le domaine de l'économie spatiale (voir par exemple les études de Fratesi, 2008 ; Klier et Mc Millen, 2008 ; Nakajima *et al.*, 2012 ; Barlet *et al.*, 2013 ; Koh et Riedel, 2014 ou Behrens et Bugna, 2013).

Ces trois mesures partagent des propriétés mais se différencient également par certaines caractéristiques. Nous allons les présenter et montrer leurs spécificités afin de pouvoir comprendre quelles structures spatiales elles sont finalement en mesure de détecter. Rares sont les études qui justifient précisément le choix d'une mesure pour évaluer la distribution spatiale des activités. Pourtant nous verrons que ce choix devrait être au contraire soigneusement motivé étant donné que ces mesures ne sont pas équivalentes.

Présentation et comparaison des mesures *D*, K_d et *M*

Au sein d'une ville, considérons la distribution des magasins dont les positions géographiques exactes et les secteurs d'activité sont connus. Supposons que nous souhaitons analyser plus précisément les éventuelles interactions (attraction ou répulsion) entre les magasins d'un secteur donné au sein de cette ville, par exemple les magasins du secteur de l'habillement.

Afin d'identifier ces interactions, une première possibilité est de comparer la proportion moyenne de magasins d'habillement observée localement (c'est-à-dire dans leur voisinage à une distance inférieure à une valeur choisie *r*) à leur proportion à l'échelle de la ville. Si les magasins d'habillement s'attirent, leur proportion sera localement plus forte autour des autres magasins d'habillement qu'au niveau de la ville en général : une concentration spatiale relative sera détectée. Au contraire, si les magasins d'habillement se repoussent, leur proportion sera relativement plus faible dans le voisinage des magasins d'habillement : une dispersion relative sera identifiée. C'est l'idée de la fonction cumulative *M* proposée par Marcon et Puech (2010) qui est une mesure relative. Techniquement, cela revient à étudier tout d'abord le « voisinage d'un magasin d'habillement » comme les magasins situés à une distance inférieure ou à égale à *r* de ce magasin. Puis, pour ce rayon *r*, on calcule le rapport entre la proportion relative locale de magasins d'habillement autour d'un magasin d'habillement à la proportion relative observée en ville. On répète cette opération pour tous les magasins d'habillement et on calcule la moyenne des rapports de ces proportions relatives. Un voisinage relatif moyen des magasins d'habillement est ainsi obtenu pour ce rayon. La valeur de référence de la fonction *M* est 1. Une valeur supérieure à 1 indique une concentration relative des magasins d'habillement à l'intérieur de ce rayon *r* et une valeur inférieure à 1 indique une dispersion relative. L'avantage de la mesure en continu est de pouvoir effectuer ces estimations pour tous les rayons possibles par exemple par incrémentation de 50 mètres. On peut dès lors obtenir une caractérisation complète des interactions entre les magasins à toutes les distances possibles et, ainsi, détecter à quelle(s) distance(s) les niveaux de concentration ou de dispersion sont observés.

Une deuxième approche pour identifier les interactions potentielles des magasins de l'habillement consiste à comparer leur structure spatiale, mesurée par la fonction *K* de Ripley, à celle des autres magasins. Cette démarche est celle proposée par la fonction *D* de Diggle et Chetwynd (1991) qui est une mesure cumulative de concentration topographique (en espace hérérogène). Plus précisément, la fonction *D* résulte de la comparaison de la structure spatiale de deux types de magasins uniquement : les magasins d'intérêt (« les cas ») et les autres magasins (« les contrôles »). Si les cas sont plus concentrés que les contrôles, on détectera de la concentration spatiale au sens topographique. Si au contraire les cas sont moins concentrés que les contrôles, une dispersion topographique

sera identifiée. *D* contrôle l'hétérogénéité de l'espace grâce à la prise en compte de la structure spatiale des « contrôles ». Techniquement, la fonction *D* est définie comme la différence de deux fonctions *K* de Ripley pour les cas et pour les contrôles. La valeur de référence est 0 quel que soit le rayon d'étude : des valeurs positives de *D* signaleront une concentration spatiale du secteur, des valeurs négatives de la dispersion.

Enfin, une troisième possibilité pour caractériser la structure spatiale est d'estimer la probabilité de trouver un magasin d'habillement à une distance *r* de chaque magasin d'habillement. Plus précisément, il s'agit d'estimer une densité de probabilité au sens mathématique du terme puisque la distance *r* est une variable continue (la probabilité de trouver un voisin exactement à distance *r* est nulle, mais elle peut être estimée autour de *r*, par intervalles successifs et lissage). Une telle mesure est de type absolu car sans référentiel. C'est la fonction K_d proposée par Duranton et Overman (2005). Elle est calculée à chaque distance *r* et non à une distance inférieure ou égale à *r* comme pour la fonction *M* ou *D* précédemment présentées.

Afin de juger de la significativité des résultats, on peut associer à ces trois indices des intervalles de confiance générés par la méthode de Monte Carlo sous l'hypothèse nulle d'indépendance de la localisation des magasins. Une concentration spatiale significative des activités est détectée si les valeurs observées des fonctions K_d, *D* et *M* sont supérieures à la borne supérieure de leur intervalle de confiance respectif. On détectera au contraire de la dispersion géographique significative des activités si les valeurs observées des fonctions sont inférieures à la borne inférieure de l'intervalle de confiance associé. Enfin, notons que les valeurs de référence (sous l'hypothèse nulle) ne sont pas les mêmes pour les trois fonctions : quel que soit le rayon, elle est de 1 pour *M*, 0 pour *D*, mais dépendante des données pour K_d (on l'estime par le centre de l'intervalle de confiance). Seules les valeurs de *M* sont interprétables : K_d et *D* ne sont mobilisables que pour la détection de la concentration spatiale (ou de la dispersion) mais ne pourront pas la quantifier.

Une présentation formelle de ces mesures est proposée dans l'encadré 3.

Encadré 3

PRÉSENTATION FORMELLE DES TROIS MESURES EN ESPACE CONTINU

Considérons une aire d'étude notée *A* sur laquelle une distribution de *N* points (les magasins) est observée. *A* est donc la surface de l'aire d'étude. On identifie les localisations respectives des points $1,...,i,...N$ par leur position géographique exacte $x_1,...,x_i,...x_N$ sur l'aire d'étude *A*. L'évaluation de la concentration spatiale repose sur l'étude du voisinage moyen des points d'intérêt (par exemple les magasins appartenant à un secteur donné). Pratiquement, pour une distance donnée, tous les points seront pris successivement comme centre du cercle, et tous les autres points testés comme voisins potentiels. Puis, on répétera cette opération pour toutes les distances afin d'obtenir une description détaillée de la distribution analysée. La notion de voisinage dépend de la fonction analysée :

- Dans le cas d'une fonction de densité, le voisinage est représenté par la surface de la couronne de rayon *r*. Les points localisés dans le voisinage seront dans ce cas ceux localisés dans la couronne de rayon *r*.

- Dans le cas d'une fonction cumulative, le voisinage est représenté par la surface du disque de rayon *r*. Les points considérés comme appartenant au voisinage du point centre dans un rayon *r* seront dans ce cas ceux localisés sur la surface du disque de rayon *r*.

La fonction *M* de Marcon et Puech (2010) est une mesure cumulative relative puisqu'elle compare en moyenne la proportion de points d'intérêt dans un rayon *r* à celle que l'on observe sur toute l'aire *A*. La fonction *M* pour le secteur d'intérêt *S* s'écrit pour un rayon *r* :

$$M(r) = \sum_{i,i\in S} \frac{\sum_{j,i\neq j,j\in S} 1\left(\left\|x_i - x_j\right\| \leq r\right)}{\sum_{j,i\neq j} 1\left(\left\|x_i - x_j\right\| \leq r\right)} \Bigg/ \frac{N_S - 1}{N - 1} \qquad (1)$$

où :

- S désigne l'ensemble des points appartenant au secteur d'intérêt *S* ;

- L'indicatrice $1\left(\left\|x_i - x_j\right\| \leq r\right)$ vaut 1 si la distance entre les deux points x_i et x_j est inférieure à *r* ;

- N_s est le nombre total de points appartenant au secteur *S*.

Par souci d'harmonisation avec les autres fonctions, les pondérations des points (par exemple les effectifs pour les magasins) ne sont pas mentionnées dans l'équation (1). Marcon *et al.* (2012b) ont montré que la fonction *M* est une généralisation de la fonction *K* de Ripley (1976, 1977) aux processus ponctuels hétérogènes.

La fonction *D* proposée par Diggle et Chetwynd (1991) est une mesure cumulative topographique en espace non homogène. Plus exactement, elle repose sur une comparaison des structures spatiales de deux types de points uniquement : les points d'intérêt (« les cas »)

➔

Encadré 3 (suite)

et les autres points de l'échantillon (« les contrôles »). Mathématiquement, la fonction *D* se définit comme la différence de deux fonctions *K* de Ripley (1976 ; 1977). Désignons par *S* le secteur d'intérêt et notons K_{cas} et $K_{contrôles}$ respectivement la fonction de Ripley pour les cas et pour les contrôles, il vient :

$$D(r) = K_{cas}(r) - K_{contrôles}(r) \quad (2)$$

avec

$$K_{cas}(r) = \frac{A}{N_S(N_S-1)} \sum_i \sum_{j,i\neq j,j\in S} 1\left(\left\| x_i - x_j \right\| \leq r\right) c(i,j,r)$$

et

$$K_{contrôles}(r) = \frac{A}{N_{\bar{S}}\left(N_{\bar{S}}-1\right)} \sum_i^{N_{\bar{S}}} \sum_{j,i\neq j,j\in \bar{S}}^{N_{\bar{S}}} 1\left(\left\| x_i - x_j \right\| \leq r\right) c(i,j,r)$$

où :

- S désigne l'ensemble des points appartenant au secteur d'intérêt *S* et N_S le nombre total de points appartenant au secteur *S* ;

- $\bar{S}$ désigne le complémentaire de S ;

- $N_{\bar{S}}$ désigne le nombre total de points de l'échantillon n'appartenant pas au secteur d'intérêt *S* ;

- L'indicatrice $1\left(\left\| x_i - x_j \right\| \leq r\right)$ vaut 1 si la distance entre les deux points x_i et x_j est inférieure à *r* ;

- $c(i,j,r)$ est une correction des effets de bord indispensable si les points sont localisés au bord du domaine. Le nombre de voisins serait en effet sous-estimé sans correction d'effet de bord puisqu'une partie du disque se trouverait en dehors du domaine.

La fonction *D* est bien une fonction cumulative puisque les voisins sont décomptés systématiquement jusqu'à une distance *r*. On peut la qualifier de « topographique » puisque le nombre de voisins d'intérêt $\left(\sum_i^{N_{\bar{S}}} \sum_{j,i\neq j,j\in \bar{S}}^{N_{\bar{S}}} 1\left(\left\| x_i - x_j \right\| \leq r\right) c(i,j,r)\right)$ est rapporté à sa densité sur le territoire : $(N_{\bar{S}} - 1) / A$. La fonction D tient enfin compte de l'hétérogénéité de l'espace puisqu'elle est le résultat de la différence entre deux fonctions *K* caractérisant les structures spatiales des cas et des contrôles.

La fonction *D* est une première généralisation de la fonction *K* de Ripley au semis de points hétérogènes, c'est-à-dire permettant de prendre en compte les variations de densité. Dans leur étude originale, Diggle et Chetwynd s'intéressaient à la détection des maladies rares chez l'enfant dans le Nord Humberside au Royaume-Uni. Les « cas » étaient constitués par la sous-population enfantine malade et les « contrôles » représentaient les enfants sains.

Enfin, **la mesure K_d** proposée par Duranton et Overman (2005) estime la probabilité de trouver un voisin à la distance *r* de chaque point. Le nombre de voisins est évalué théoriquement *à* la distance *r*, ce qui nécessite l'utilisation d'une fonction de lissage qui comptabilise les voisins dont la distance est *autour* de *r*. Grâce à une normalisation appropriée, K_d est une fonction de densité de probabilité (de trouver un voisin à la distance *r*) donc K_d est une mesure de densité absolue, sans référentiel. Techniquement, on retient un estimateur de noyau gaussien de bande passante *h* (Silverman, 1986). Soit $k\left(\left\| x_i - x_j \right\|, r\right)$ cet estimateur : il atteint sa valeur maximum si la distance entre les deux points x_i et x_j est exactement *r*, et décroît selon une distribution gaussienne d'écart-type *h* si la distance s'écarte de *r*. Il peut être défini en termes techniques par :

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

Comme Duranton et Overman (2005) dans ce qui suit nous retenons un noyau gaussien de bande passante optimale au sens de Silverman (1986). La fonction K_d se définit de la façon suivante :

$$K_d(r) = \frac{1}{N(N-1)} \sum_i \sum_{j,i\neq j} k\left(\left\| x_i - x_j \right\|, r\right) \quad (3)$$

K_d est une fonction développée apparemment indépendamment des travaux de Ripley.

Duranton et Overman (2005) recommandent de retenir dans les estimations des intervalles de confiance globaux plus restrictifs que les intervalles de confiance locaux (une discussion sur la génération de ces deux types d'intervalles de confiance (local, global) est donnée dans Marcon et Puech, 2010). Pour les trois fonctions, des simulations sont effectuées en attribuant aléatoirement les étiquettes des secteurs sur l'ensemble des localisations existantes (Marcon et Puech, 2014). Les valeurs de référence quel que soit le rayon d'étude sont 1 pour *M*, 0 pour *D* et le centre de l'intervalle de confiance pour K_d.

Enfin, notons que les fonctions dans cette contribution sont présentées comme une recherche de voisins d'intérêt à une distance inférieure égale à *r* pour les fonctions de densité ou sur un disque de rayon *r* pour les fonctions cumulatives. Une présentation alternative serait de travailler directement sur les distances observées entre les points (distances inter-établissements) mais les résultats obtenus sont naturellement identiques quelle que soit la méthode choisie. Ainsi par exemple, décompter les distances inter-établissements pour connaître leur fréquence à un rayon donné est l'approche retenue pour la fonction K_d de Duranton et Overman. Pour les fonctions cumulatives, le nombre de distances inter-établissements serait calculé jusqu'au rayon *r* (et non à un rayon *r* donné).

D, K_d et *M* à l'épreuve des critères

La pertinence de ces trois mesures dans notre champ d'étude peut être jugée en confrontant ces outils aux critères de Duranton et Overman (2005) ou Combes et Overman (2004) définissant une mesure satisfaisante pour évaluer la concentration spatiale des activités économiques.

Tout d'abord, le recours à des zonages doit être évité car il peut être source de biais statistiques (Briant *et al.*, 2010 ; Crozet et Lafourcade, 2011), comme cela a été illustré plus haut par l'exemple théorique du graphique II. Retenir des mesures en espace continu, c'est-à-dire fondées uniquement sur les distances séparant les entités analysées, permet d'éviter ces biais liés à l'agrégation des données. Les trois mesures *D*, K_d et *M* répondent donc favorablement à cette première exigence que doivent respecter les mesures.

Un deuxième critère stipule que les mesures doivent prendre en compte la tendance générale des activités à se concentrer. Plusieurs méthodes peuvent dès lors être mobilisées pour y arriver. Une première solution est de recourir à des mesures topographiques contrôlant l'hétérogénéité de l'espace c'est-à-dire les mesures qui prennent en compte une densité non constante sur le territoire. La fonction *D* repose sur ce cadre d'analyse. En effet cette fonction est construite à partir de la différence de deux fonctions topographiques et identifie une concentration spatiale pour un secteur si la distribution de ses établissements est plus agrégée que celle des autres activités en général. Une deuxième solution est d'opter pour des mesures relatives qui détectent dans le voisinage des établissements du secteur étudié une proportion d'établissements du même secteur supérieure à celle que l'on observe en moyenne sur le territoire. Cette approche est celle retenue par la fonction *M* et généralement en économie spatiale pour les mesures discrètes (indices de Gini ou d'Ellison et Glaeser par exemple). Une dernière possibilité pour respecter ce critère est celle explorée par Duranton et Overman en proposant la fonction K_d. Cette dernière repose sur une comparaison de mesures absolues puisqu'elle permet de comparer la fréquence des distances bilatérales entre les établissements d'un même secteur réellement observée à une fréquence théorique définie en conservant tous les emplacements existants (« sites actifs ») mais en redistribuant les établissements sur ces emplacements. Le contrôle de la tendance générale des activités est par conséquent réalisé avec la comparaison de la distribution observée de celle sous l'hypothèse nulle. Il est donc intéressant de voir que *D*, K_d et *M* respectent ce deuxième critère même si ce concept renvoie à une notion assez large du référentiel optimal et peut ainsi recouvrir différentes réalités. Nous reviendrons sur ce point dans la section suivante.

Un troisième critère respecté par les trois mesures est le fait d'associer un niveau de significativité aux résultats. En effet, un intervalle de confiance de l'hypothèse nulle peut être généré systématiquement pour ces outils comme nous l'avons précisé précédemment.

Un quatrième critère souligne la nécessité de contrôler la concentration industrielle c'est-à-dire la structure industrielle des secteurs qui dépend à la fois du nombre d'établissements au sein des industries et des effectifs associés. C'est ce point qui avait été mis en lumière par Ellison et Glaeser et qui constituait une limite forte à l'encontre de l'indice de Gini (voir encadré 1). *M* peut contrôler la concentration industrielle en pondérant les établissements par leur nombre d'employés, et en conservant cette pondération sous l'hypothèse nulle. La fonction K_{emp} a été proposée par Duranton et Overman (2005) pour pondérer K_d de la même façon. En revanche, la fonction *D* est issue de la théorie des processus ponctuels, dans laquelle chaque établissement est représenté par un point, sans poids. Dans le travail présenté ici, nous nous limiterons à des établissements de poids égaux à 1 pour *M* et à la fonction non pondérée K_d pour permettre des comparaisons entre les trois fonctions.

Un cinquième critère indique que les résultats auxquels aboutissent les mesures doivent être robustes aux comparaisons interindustrielles. La fonction *M* est interprétable comme la proportion des établissements du secteur d'intérêt autour de chaque établissement de référence rapportée à la même proportion dans toute l'aire d'étude. Ce nombre, qui est un quotient de localisation, peut être comparé entre secteurs. La fonction K_d fournit quant à elle une densité de probabilité plus difficilement interprétable, qui reste comparable entre secteurs dans une étude sur le même espace, mais pas entre études portant sur des espaces différents (toutes choses égales par ailleurs, la densité de probabilité diminue si la taille de l'aire d'étude augmente). Enfin, une comparaison de résultats obtenus avec la fonction *D* entre plusieurs secteurs ne serait que peu convaincante puisque *D* est construite comme la différence de deux

fonctions *K* entre celle du secteur d'intérêt et celle des « autres secteurs ». Lors des comparaisons intersectorielles, les « autres secteurs » ne seraient pas les mêmes.

Enfin, notons qu'aucune des trois mesures (ni aucune mesure existante à notre connaissance) ne répond favorablement aux deux derniers critères mentionnés par Combes et Overman (2004)[8] :

- L'indépendance de la mesure au découpage sectoriel : le choix de la classification sectorielle retenue ne doit pas introduire des biais dans les estimations des niveaux de concentration. Le problème sous-jacent est de même nature que le Problème des unités spatiales modifiables (MAUP) précédemment décrit mais renvoie ici au problème de « frontières » intersectorielles. En effet, retenir un niveau sectoriel plus ou moins agrégé par exemple est susceptible de biaiser les évaluations des niveaux de concentration spatiale.

- L'intégration de l'outil à la théorie économique : les mesures sont des statistiques descriptives d'un semis de points, pas des statistiques synthétisant le résultat d'un modèle économique. Seul l'indice d'Ellison et Glaeser respecte cette propriété dans le sens où sa valeur peut être interprétée comme le résultat d'un modèle (même simple) de choix de localisation pour maximiser les profits de chaque établissement mais cet indice repose sur un zonage et non sur les distances inter-établissements. Ce critère ne doit pas être négligé car il permettrait de déterminer le choix de l'indice à retenir selon la question posée. Cette question sera probablement au centre des recherches futures sur les mesures de concentration spatiale.

Que peut-on conclure de cette courte analyse comparative ? Tout d'abord, parmi les trois mesures de concentration géographique considérées, seules deux répondent à un maximum de bonnes propriétés dictées par la littérature : la fonction K_d de Duranton et Overman et la fonction *M* de Marcon et Puech. Cependant, il serait incorrect de considérer K_d et *M* comme équivalentes. Si elles partagent certaines propriétés (elles ne reposent pas sur un zonage pré-défini, elles tiennent compte de la concentration industrielle…) elles sont également sensiblement différentes : K_d est par exemple une fonction de densité de probabilité et *M* une fonction cumulative (Marcon et Puech, 2010). Dans la partie suivante, nous allons mobiliser les trois mesures et nous appuyer sur leurs spécificités individuelles pour obtenir une analyse précise et complète de la concentration spatiale de plusieurs activités commerciales sur Lyon.

Cas 1 : exemple où les résultats des trois mesures convergent

Étudions un premier cas pour lequel les trois mesures aboutissent à la même conclusion. Nous allons considérer pour cela un cas théorique où les magasins du secteur d'intérêt sont plus concentrés topographiquement que ceux des autres secteurs. Nous verrons dans un second temps que ce cas d'étude est observable sur données réelles à partir de la distribution des magasins sur Lyon.

Exemple théorique

Considérons une ville où uniquement deux types de magasins sont implantés : ceux du secteur d'intérêt et ceux des autres secteurs. Les magasins du secteur d'intérêt sont dénommés les « cas » et ceux des autres secteurs les « contrôles ». Les localisations des magasins sont issues de simulations de processus ponctuels connus : processus de Matérn pour la simulation d'agrégats et processus de Poisson pour des distributions aléatoires de magasins au sein de la ville[9]. Le *package spatstat* sous le logiciel R (R Development Core Team, 2012 ; Baddeley et Turner, 2005) permet de simuler de telles réalisations.

Dans notre exemple simulé, on dénombre 168 magasins localisés dans la ville. La distribution de ces commerces est donnée sur le graphique IV. Deux zones denses en commerces sont observables ; chacune compte en moyenne 40 magasins composés de cas (points noirs sur la carte) et de contrôles (représentés par des triangles). Dans notre exemple, 32 cas et 44 contrôles sont au total présents dans ces deux zones. 92 contrôles sont également implantés aléatoirement en ville (les triangles hors des deux agrégats).

Les trois fonctions K_d, *M et D* associées à cette distribution sont données sur le graphique V.

8. Notons que la littérature sur la définition des bonnes propriétés n'est pas encore stabilisée (Thomas-Agnan et Bonneu, 2014).
9. Voir l'encadré 2 pour une présentation des processus de Matérn et Poisson.

Les intervalles de confiance globaux (IC) ont été établis à partir de 1 000 simulations au seuil de confiance de 95 %. Tous les calculs ont été réalisés sous le logiciel R avec le *package* dbmss que nous avons développé (R Development Core Team, 2012 ; Marcon *et al.*, 2012a).

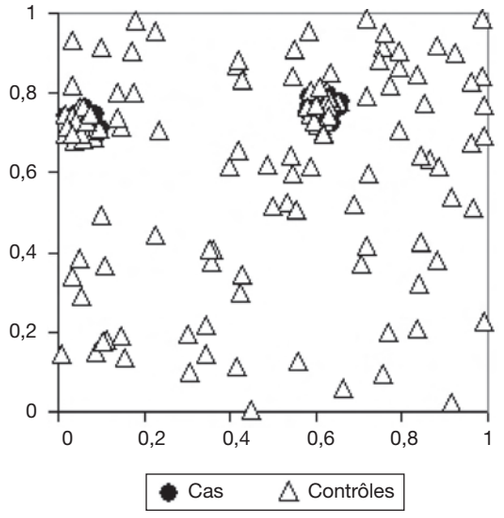
Les trois fonctions détectent une concentration spatiale des cas pour de faibles distances.

La fonction *M* (graphique VA) indique une concentration relative des cas à partir d'un rayon approximativement égal à 0,025 et jusqu'à une distance de 0,25. Elle est maximale à courtes distances : les agrégats simulés de cas sont de petite taille. Puis, plus le rayon d'étude augmente, plus le nombre de contrôles localisés dans le voisinage augmente : la part relative des cas diminue, graphiquement la courbe *M* décroît. À de très faibles distances, l'intervalle de confiance est très large (peu de magasins sont présents à de très faible rayons), les résultats des niveaux de concentration spatiale sont non significatifs.

La fonction *D* quant à elle (graphique VC) compare au sens topographique la distribution des agrégats de cas à la distribution complètement aléatoire des contrôles : une concentration spatiale des cas est à nouveau détectée. Les cas sont en effet plus concentrés spatialement que les contrôles : les cas ne sont localisés que dans les zones commerciales denses alors que les contrôles sont présents également dans ces deux pôles commerciaux mais, par construction, sont majoritairement présents sur toute la ville. Les pics de concentration de *D* et *M* apparaissent approximativement à une distance égale à la taille des agrégats.

K_d détecte également la concentration des cas et le pic de concentration est plus marqué et observable à une distance de 0,05 (graphique VB). Ce pic de concentration correspond exactement au rayon des agrégats de cas. Une fonction de densité comme K_d renvoie des estimations locales plus précisément qu'une fonction cumulative. Nous voyons également que pour de larges distances (au delà de 0,17) les résultats de K_d sont négatifs et significatifs, avec une courbe de K_d est située en dessous de l'intervalle de confiance : la distribution des cas est donc dispersée au sens de K_d. Cette dispersion indique qu'il n'y a pas de cas entre les agrégats. Comment expliquer que les fonctions *M* et *D* ne révèlent pas cette dispersion ? Nous constatons en effet

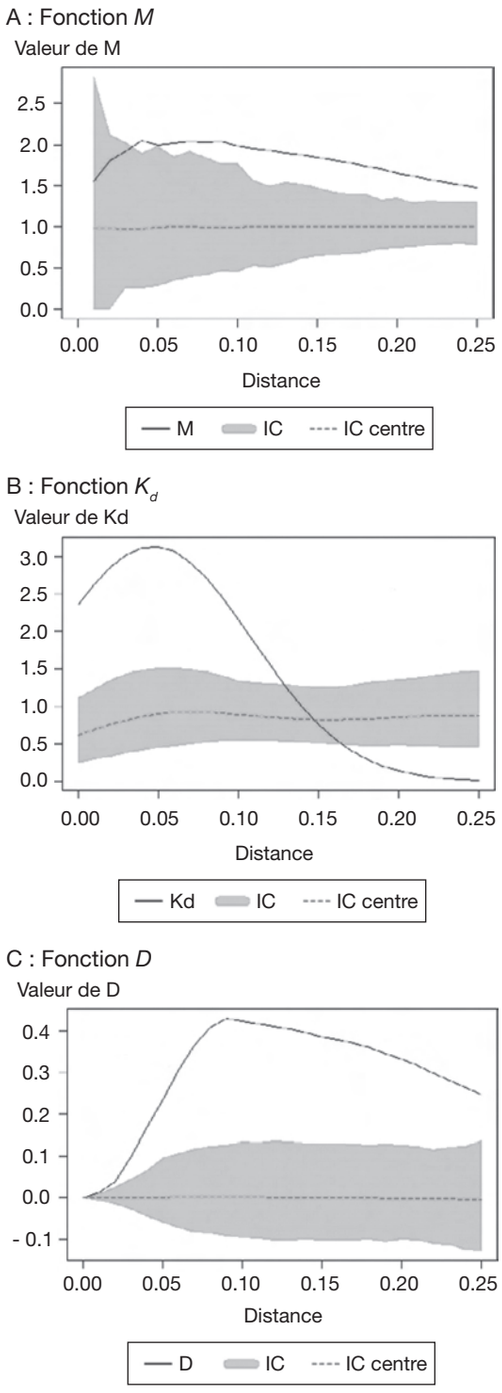
Graphique IV
Distribution théorique de cas et de contrôles du cas 1



Lecture : le domaine est de 1x1 et représente une ville composée uniquement de deux types de magasins (les contrôles et les cas). La distribution théorique de magasins est issue de simulations de deux processus ponctuels. Un processus de Matérn est simulé pour les deux agrégats composés de cas (les cercles) et de contrôles (les triangles). Les paramètres de ce processus sont : 2 pour la densité du processus de Poisson pour les centres des agrégats, 0,05 pour le rayon des agrégats et 40 pour le nombre moyen de points par agrégat. Afin de différencier des cas et des contrôles au sein des deux agrégats de taille n, nous définissons les cas à partir d'un tirage dans une loi binomiale de paramètres n et 0,5. Un processus de Poisson homogène est également simulé pour obtenir une distribution aléatoire de contrôles au sein de la ville (triangles sur la figure, la densité du processus étant égale à 100).
Champ : données simulées à partir de processus de Matérn et de Poisson homogène.
Source : calculs des auteurs sous R à l'aide du package spatstat.

qu'aux mêmes distances, *D* et *M* décroissent doucement mais restent situées au-dessus de leur intervalle de confiance. Cela est dû à leur propriété partagée de fonctions cumulatives : le voisinage à petites distances influence les estimations à plus longues distances (Wiegand et

Graphique V
Résultats des fonctions K_d, *M et D* pour la distribution théorique de magasins du cas 1



Lecture : les courbes D, M et K_d (en trait plein) situées au-dessus de leur intervalle de confiance respectif (zone grisée) détectent un phénomène de concentration spatiale. Seule la fonction K_d présente des niveaux de dispersion significatifs approximativement sur l'intervalle de distances 0,17-0,25, sa courbe étant alors située en dessous de l'intervalle de confiance.
Champ : données simulées à partir de processus de Matérn et de Poisson homogène.
Source : calculs des auteurs sous R avec le package *dbmss. Les fonctions ont été calculées avec un pas de calcul de 0,01 jusqu'à une distance de 0,25. Seuls les intervalles de confiance globaux sont reportés (1 000 simulations).*

Moloney, 2004). Les courbes *M* et *D* retournent donc seulement doucement vers leur intervalle de confiance respectif. Notons que les estimations du graphique V sont effectuées en prenant une valeur maximale de la distance d'étude égale au quart du côté du domaine analysé, comme cela est recommandé pour *D* issue de la fonction *K* de Ripley (Baddeley et Turner, 2005)[10]. Toutefois, en considérant des distances plus importantes, dans cet exemple *M* retourne effectivement dans son intervalle de confiance (aux alentours de 0,5) puis présente à nouveau un pic de concentration (comme K_d) correspondant à la distance entre les deux agrégats (0,6). *M* ne détecte pas de dispersion entre les pics car il n'y a pas de répulsion à proprement parler entre les cas des deux pôles (assimilables aux résultats non significatifs de *M*). La dispersion détectée par K_d à ces distances indique uniquement un « manque de cas ».

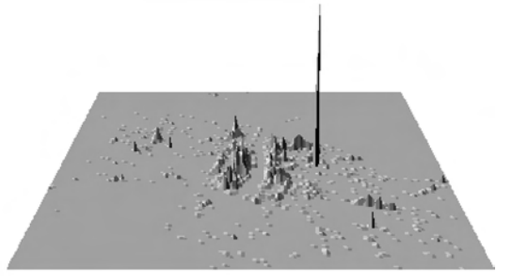
Exemple empirique

L'accès à une base de données de la Chambre de commerce et de l'industrie de Lyon nous a permis de travailler sur les emplacements exacts de 3 124 commerces de détail non-alimentaires identifiés en avril 2012 sur Lyon (France). Ces activités sont ventilées en 26 secteurs et correspondent aux sous-secteurs de 47.30Z à 47.79Z de la Nomenclature d'activités françaises 2008 (révision 2). La carte IA donne un aperçu de la densité de commerces non alimentaires sur Lyon.

10. Afin d'effectuer des comparaisons de résultats entre les trois fonctions, nous avons restreint à un quart du domaine la distance pour les fonctions M et K_d également (graphique VB et graphique VC).

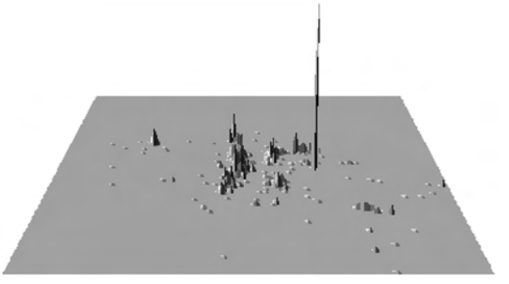
Carte I
Densité des commerces non alimentaires sur Lyon en avril 2012 ainsi que celle des commerces d'habillement en magasin spécialisé (NAF 47.71Z)

A : Densité de commerces de détail non alimentaires



Lecture : la densité des 3 124 implantations de commerces de détail non alimentaires sur Lyon est indiquée. Plus la zone est en relief, plus la densité commerciale est élevée. Au centre de la carte, on distingue nettement la presqu'île de Lyon très commerçante. À l'est de la presqu'île, la rive gauche du Rhône présente également une forte densité de commerces de détail non alimentaires.
Le pic de densité commerciale particulièrement notable à l'est de la rive gauche du Rhône indique un centre commercial.
Champ : ensemble des magasins non alimentaires en avril 2012 sur Lyon.
Source : carte et calculs réalisés sous R par les auteurs (données de la CCI de Lyon).

B : Densité de commerces d'habillement en magasins spécialisés



Lecture : la densité des 931 commerces d'habillement en magasin spécialisé est représentée sur la carte. Nous constatons que cette activité est localisée dans les zones où la densité de commerces de détail non alimentaires est élevée (zones en relief sur la carte 2A).
Champ : ensemble des commerces d'habillement en magasin spécialisé (code NAF 47.71Z) en avril 2012 sur Lyon.
Source : carte et calculs réalisés sous R par les auteurs (données de la CCI de Lyon).

Analysons plus précisément dans cet exemple la distribution des 931 commerces d'habillement en magasin spécialisé (code NAF 47.71Z). Si nous comparons sur Lyon la densité des commerces d'habillement (carte IB) avec la densité des commerces non alimentaires (carte IA) nous constatons que les magasins d'habillement sont localisés dans les zones où la présence d'implantations commerciales non alimentaires est forte notamment sur la presqu'île de Lyon (au centre de la carte IB) et sur la rive gauche du Rhône (zone à l'Est de la presqu'île). Afin de caractériser la structure spatiale des magasins d'habillement à Lyon, les trois fonctions fondées sur les distances K_d, D et M ont été calculées avec un pas de calcul de 100 mètres jusqu'à 3 000 mètres. Les intervalles de confiance globaux (IC) ont été élaborés à partir de 1 000 simulations au niveau de confiance de 95 %. Les distances inter-établissements sont les distances euclidiennes. Afin de pouvoir comparer les résultats entre les trois fonctions (*i*) les effectifs des commerces analysés ne sont pas considérés dans notre analyse empirique et (*ii*) notre comparaison repose sur la détection de niveaux de concentration ou de dispersion mais nous ne cherchons en aucun cas à interpréter les niveaux obtenus (puisque seule M est à même d'apporter de type d'information). Les résultats des trois fonctions M, K_d et D pour le commerce de détail d'habillement sont donnés sur le graphique VI.

Nous constatons que les trois mesures détectent une concentration spatiale correspondant à une concentration géographique au sens relatif (pour M, graphique VIA), absolu (pour K_d, graphique VIB) et topographique (pour D, graphique VIC). M détecte que la proportion des magasins d'habillement relativement aux autres magasins non alimentaires est en moyenne plus élevée autour des emplacements de commerces de détail d'habillement que sur l'ensemble de la ville de Lyon. Il y a donc une concentration relative de ce secteur. K_d détecte également une attraction particulièrement marquée approximativement à des distance de 250 mètres, 1 000 mètres et 1 750 mètres[11], correspondant aux regroupements de ces activités au sein des grandes zones commerciales (premier « pic ») et entre les différentes zones commerciales sur l'aire urbaine de Lyon (deuxième et troisième « pics »). Cela est par exemple le cas entre la presqu'île (cœur de ville) et les grandes rues commerciales de la rive gauche (à l'Est de Lyon)[12]. Enfin, la fonction D compare la concentration spatiale topographique de ce secteur par rapport à l'ensemble des autres secteurs non alimentaires analysés et montre qu'elle est supérieure. Cela souligne que les commerces de l'habillement sont plus concentrés géographiquement que sous l'hypothèse nulle dans laquelle les commerces de l'habillement et les autres commerces non alimentaires ont la même répartition spatiale. En d'autres termes, l'apport de ces trois fonctions nous permet d'identifier des concentrations d'activités commerciales de l'habillement au sens relatif, topographique et absolu soit des regroupements très localisés de ces activités dans des zones à forte densité commerciale. Ce résultat est typiquement le cas de quartiers spécialisés dans le centre-ville.

Cas 2 : exemple où les résultats des mesures divergent

Mais les mesures peuvent également conduire à des résultats divergents. On va l'illustrer en commençant à nouveau par un exemple théorique, caractéristique du cas où les magasins du secteur d'intérêt sont plus dispersés topographiquement que ceux des autres secteurs. Puis, nous vérifieront empiriquement les prédictions de cet exemple théorique en étudiant la localisation du commerce de détail des carburants en magasin spécialisé sur Lyon.

Exemple théorique

Considérons comme dans le cas précédent, une ville dans laquelle uniquement deux types de magasins sont implantés : ceux du secteur d'intérêt (les « cas ») et ceux des autres secteurs (les « contrôles »). 112 magasins sont localisés dans cette ville. Les contrôles sont au nombre de 88 : 41 magasins sont implantés aléatoirement et 47 magasins forment un agrégat pouvant être assimilé à un pôle commercial au sein de la ville. Les cas sont au nombre de 24 et sont aléatoirement distribués au sein de la ville. Les localisations des magasins sont

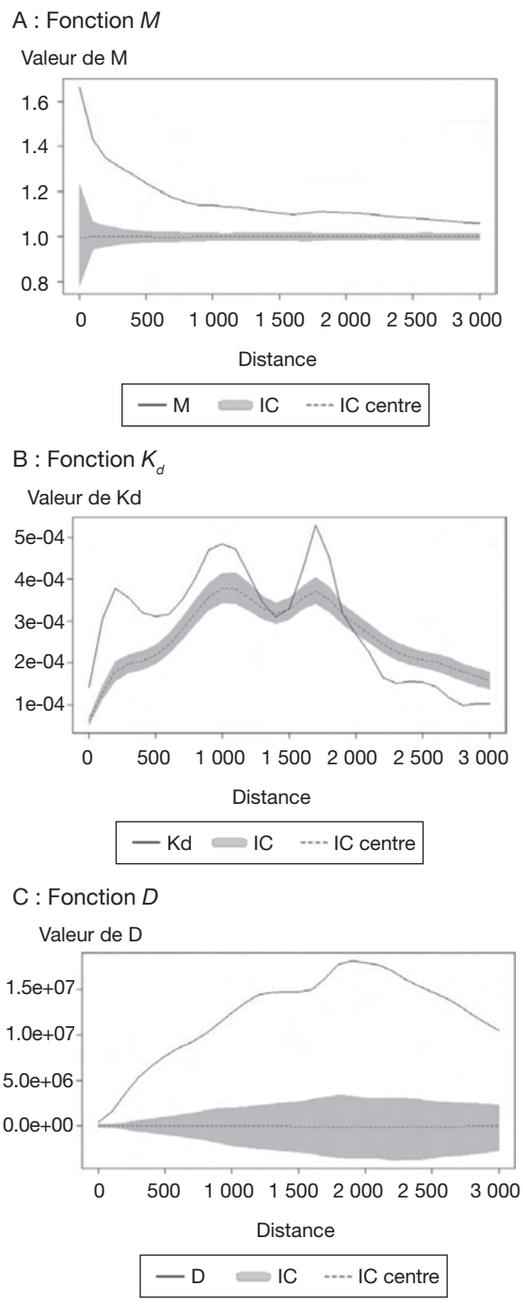
11. Ces trois « pics » de concentration apparaissent aux mêmes distances sur les courbes M et D (graphiques VIA et C) mais sous forme de ruptures de pente puisqu'il s'agit de fonctions cumulatives, d'une manière qui est donc beaucoup discrète que sur une fonction de densité comme K_d.

12. Duranton et Overman (2005 ; 2008) recommandent de ne pas analyser les résultats de K_d au-delà d'une distance économiquement pertinente qu'ils définissent comme la distance médiane de toutes les paires de commerces sur le territoire analysé soit 1 741 mètres dans notre cas. La dispersion détectée sur le graphique VIB au-delà de 2 000 mètres ne sera donc pas étudiée.

issues ici encore de simulations de processus de Matérn pour les agrégats et de processus de Poisson pour les distributions aléatoires à l'aide du *package spatstat*[13]. La distribution des magasins de la ville est représentée sur le graphique VII ; le pôle commercial se distingue aisément. Les trois fonctions K_d, *M et D* sont données par le graphique VIII. L'intervalle de confiance global a de nouveau été simulé à partir de 1 000 simulations au seuil de confiance

13. L'algorithme de simulation de l'exemple théorique peut aboutir à des résultats non significatifs selon les aléas des tirages aléatoires.

Graphique VI
Résultats des fonctions K_d, *M et D* pour les commerces de détail d'habillement (47.71Z) sur Lyon



Lecture : les courbes D, K_d et M (en trait plein) situées au-dessus de leurs intervalles de confiance respectifs (zone grisée) détectent un phénomène de concentration spatiale. Cela est le cas de D et M quelle que soit la distance r considérée (indiquée en abscisse). La courbe de K_d présente à la fois de niveaux de concentration (par exemple jusqu'à 1 250m) et de dispersion géographique (au-delà de 2 km).
Champ : ensemble des magasins non alimentaires dont le commerce de détail d'habillement en magasin spécialisé (Code NAF 47.71Z).
Source : données de la CCI de Lyon, calculs des auteurs sous R avec le package *dbmss.*

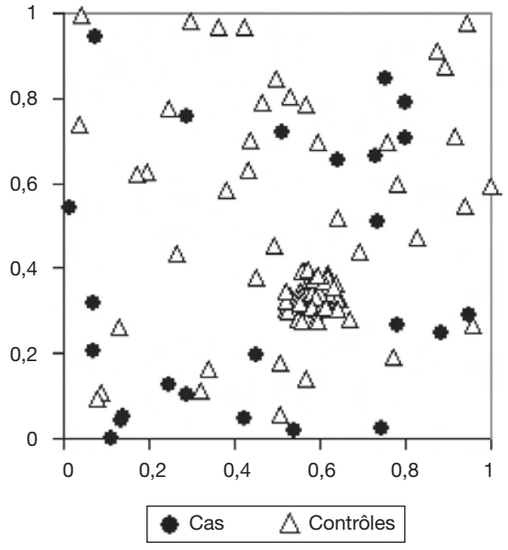
de 95 %. Expliquons à présent les divergences entre les résultats obtenus par ces trois mesures.

Commençons par le graphique VIIIC : la fonction *D* détecte une dispersion spatiale des cas par rapport aux contrôles. Cela est logique car la distribution des contrôles est plus concentrée que celle des cas du fait de l'existence d'un agrégat de contrôles. La mesure topographique capture parfaitement le fait que les cas présentent une distribution plus régulière que celle des contrôles.

La fonction K_d détecte également la dispersion des cas (graphique VIIIB). Ce résultat est là encore compréhensible en se reportant à nos simulations de processus. Tout d'abord, rappelons que les cas sont localisés aléatoirement dans la ville et donc, par construction, la zone à forte densité commerciale (le pôle commercial) ne contient que des contrôles (la présence de cas est, si elle existe, anecdotique). Donc lorsque l'on simule l'hypothèse nulle en redistribuant les cas sur les positions des cas ou des contrôles, les cas sont plus agrégés que pour la distribution observée initialement. Ainsi, la probabilité de trouver un cas à la distance *r* de chaque cas (définition de K_d) est plus faible que sous l'hypothèse nulle de sorte que K_d détecte logiquement de la dispersion pour les cas à toutes les distances. Néanmoins, la fonction K_d retourne plus vite que la fonction *D* dans l'intervalle de confiance pour de larges distances : ceci s'explique par le fait que *D* est une fonction cumulative donc est localement moins sensible à des variations de densités. En d'autres termes, si un niveau de concentration est détecté à petites distances (comme cela s'observe pour les cas), ce résultat influencera à plus grande échelle les estimations de *D*.

Enfin étudions les résultats de *M* (graphique VIIIA). Lorsque les estimations sont significatives, cette mesure relative détecte une concentration spatiale des cas, contrairement aux deux autres fonctions. Pour l'expliquer, gardons à l'esprit que les cas sont localisés aléatoirement et, comme nous l'avons dit, en dehors du pôle commercial. Donc autour d'un cas, la densité relative des cas par rapport aux contrôles est plus forte que sous l'hypothèse nulle : une concentration spatiale au sens de *M* est de ce fait mise en évidence. Au final, seuls les résultats des mesures topographique et absolue sont en accord.

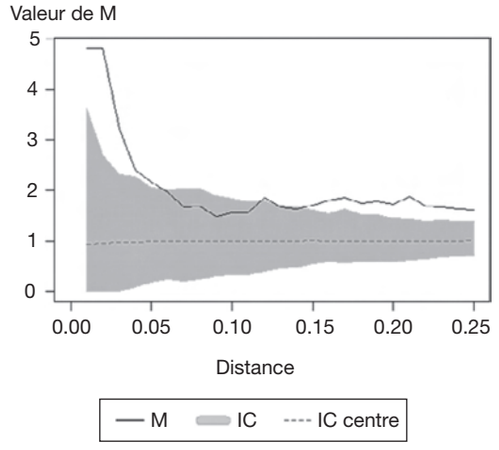
Graphique VII
Distribution théorique de cas et de contrôles du cas 2



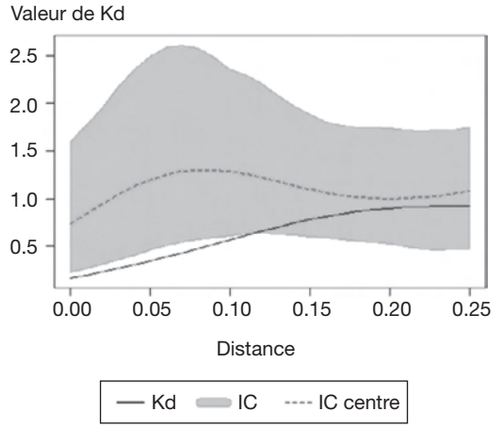
Lecture : distribution théorique de magasins issue de simulations de trois processus ponctuels. Un processus de Matérn est simulé pour l'agrégat de contrôles (agrégat de triangles). Les paramètres de ce processus sont : 2 pour la densité du processus de Poisson pour les centres des agrégats, 0,07 pour le rayon des agrégats et 40 pour le nombre moyen de points par agrégat. Un processus de Poisson homogène est également simulé pour obtenir la distribution aléatoire des contrôles (triangles sur la figure, la densité du processus étant égale à 40), un second processus de Poisson est estimé pour simuler la distribution aléatoire des cas (cercles sur la figure, la densité du processus étant égale à 20). Le domaine est de 1x1 et représente une ville composée uniquement de deux types de magasins (les contrôles et les cas).
Champ : données simulées à partir de processus de Matérn et de Poisson homogène.
Source : calculs des auteurs sous R avec le package spatstat.

Graphique VIII
Résultats des fonctions K_d, *M et D* pour la distribution théorique de magasins du cas 2

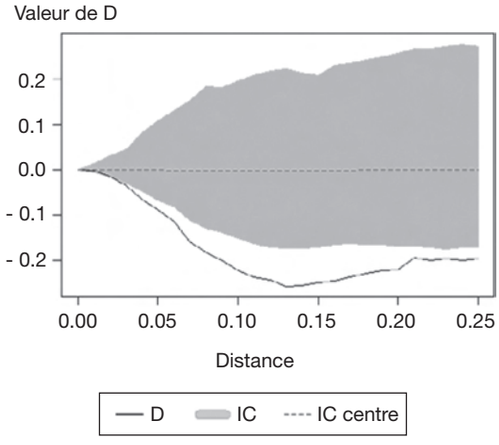
A : Fonction *M*



B : Fonction K_d



C : Fonction *D*



Lecture : les courbes D et K_d (en trait plein) situées en dessous de leur intervalle de confiance respectif (zone grisée) détectent un phénomène de dispersion spatiale. Les valeurs de M significatives (c'est-à-dire les distances où M est au-dessus de son intervalle de confiance) indiquent un phénomène de concentration géographique.
Champ : données simulées à partir de processus de Matérn et de Poisson homogène.
Source : calculs des auteurs sous R avec le package *dbmss. Les fonctions ont été calculées avec un pas de calcul de 0,01 jusqu'à une distance de 0,25. Seuls les intervalles de confiance globaux au niveau de confiance 95 % sont reportés (1 000 simulations).*

Exemple empirique

Ce genre de contradiction peut effectivement se rencontrer dans la pratique. Nous reprenons la distribution des magasins non alimentaires sur Lyon en avril 2012 et nous nous intéressons maintenant au secteur du commerce de détail des carburants en magasins spécialisés (23 commerces, code NAF 47.30Z). La carte II indique que ce secteur ne présente pas les mêmes stratégies d'implantation que les commerces de détail non alimentaires en général. Le faible nombre d'établissements de ce secteur nous permet de nous rendre compte de sa structure spatiale à partir de la distribution observée des commerces. Nous constatons que les stations-services (points noirs sur la carte) ne sont pas localisées au sein des zones où les activités commerciales de détail non alimentaires sont particulièrement présentes (zones en relief).

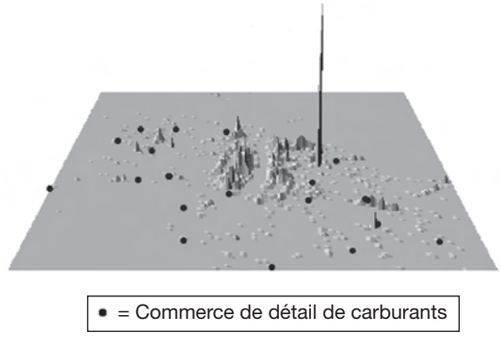
Afin d'identifier les structures spatiales de ce secteur d'activité, les trois fonctions M, K_d et D sont données sur le graphique IXA, B et C. Nous conservons les mêmes choix d'estimations que pour les calculs du secteur de l'habillement précédemment présenté : le pas de calcul est toujours de 100 mètres jusqu'à 3 000 mètres, les intervalles de confiance globaux (IC) ont été simulés à partir de 1 000 simulations au niveau de confiance 95 % et les distances euclidiennes inter-établissements sont retenues.

Nous remarquons que les résultats des courbes M (graphique IXA) et K_d (graphique IXB) sont significatifs jusqu'à une distance de 2 000 mètres, les courbes se situant en dehors de leurs intervalles de confiance respectifs. Toutefois alors que M détecte une concentration spatiale des commerces de carburants (puisque la courbe M est au-dessus de la borne supérieure de l'intervalle de confiance), K_d met en évidence une répulsion de ces établissements aux mêmes distances (K_d étant située en dessous de la borne inférieure de son intervalle de confiance).

Ce résultat apparemment contradictoire traduit le fait que ces deux mesures caractérisent de manière différente la structure spatiale du sous-secteur analysé. Les commerces de détail de carburants ont la particularité d'être localisés plutôt dans des zones où il y a peu d'activités commerciales. Pour juger de la significativité des résultats, les simulations de K_d reposent sur une redistribution des établissements sur les emplacements commerciaux existants. Sous l'hypothèse nulle, les commerces de carburants seront donc situés dans des zones à densité commerciale plus élevée qu'ils ne le sont dans la distribution réellement observée. La probabilité de trouver un voisin à courte distance sera par conséquent plus grande sous l'hypothèse nulle que dans la réalité (comme constaté sur la graphique IXB).

La fonction M mesure quant à elle la densité relative des commerces de carburants c'est-à-dire la densité de cette activité commerciale par rapport aux autres activités commerciales non alimentaires. Dans les zones à faible densité de commerces, les commerces de détail de carburants sont surreprésentés : M détecte cette concentration spatiale relative.

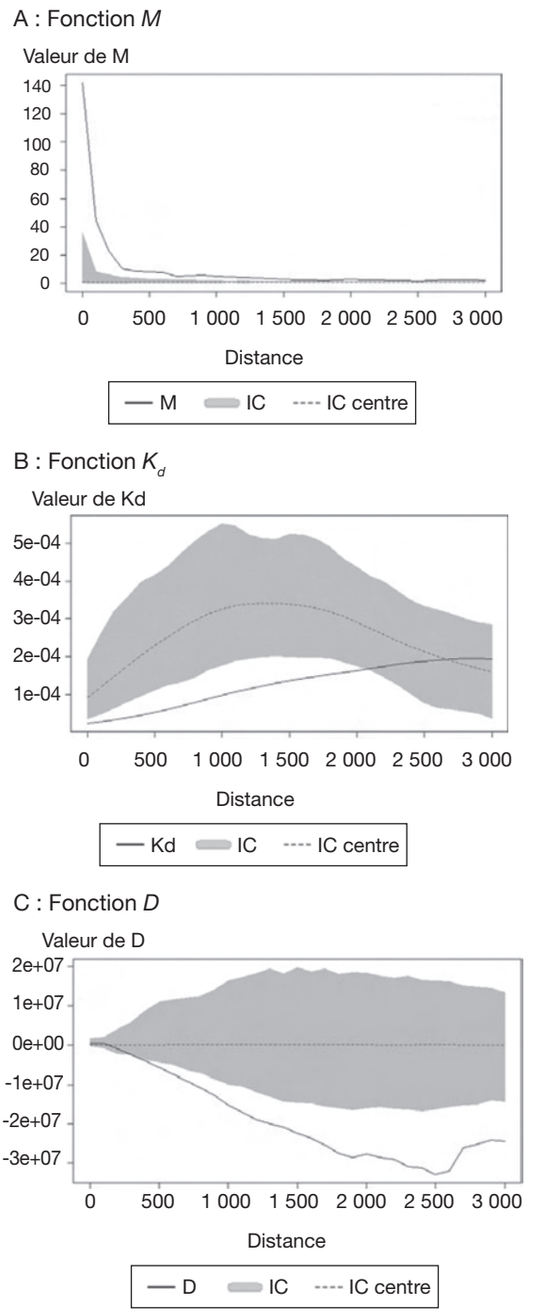
Carte II
Localisation des commerces de détail de carburants (NAF 47.30Z) sur Lyon en avril 2012



Lecture : les 23 implantations de commerces de détail de carburants sont indiquées par des points noirs sur la carte. Nous constatons que cette activité n'est pas localisée dans les zones où la densité de commerces non alimentaires de détail est élevée (zones en relief). Le pic de densité commerciale particulièrement notable est observé pour un centre commercial.
Champ : ensemble des magasins non alimentaires (dont le commerce de détail de carburants en magasin spécialisé, code NAF 47.30Z) en avril 2012 sur Lyon.
Source : carte et calculs réalisés sous R par les auteurs (données de la CCI de Lyon).

Reportons-nous enfin à la dernière fonction analysée, la mesure *D* de Diggle et Chetwynd, qui détecte quant à elle l'écart de concentration spatiale des stations-services par rapport à la tendance générale des activités non alimentaires. La concentration spatiale analysée est dans ce cas topographique, c'est-à-dire par rapport à l'espace physique (au sens de Brülhart et Traeger, 2005). Les commerces de détail de carburants sont localisés dans les zones peu

Graphique IX
Résultats des fonctions *Kd, M et D* pour les Commerces de détail de carburants (47.30Z) à Lyon



Lecture : les courbes D, K_d et M (en trait plein) situées au-dessus de leur intervalle de confiance respectif (zone grisée) détectent un phénomène de concentration spatiale. Cela est le cas de M quelle que soit la distance r considérée (indiquée en abscisse). Les courbes D et K_d situées en dessous de leur intervalle de confiance respectif détectent un phénomène de dispersion géographique. Comme cela est le cas pour D à partir de 500 m et pour le commerce de carburants de K_d jusqu'à 2 km.
Champs : ensemble des magasins non alimentaires (dont le commerce de détail de carburants en magasin spécialisé, code NAF 47.30Z) en avril 2012 sur Lyon.
Source : données de la CCI de Lyon, calculs des auteurs sous R avec le package *dbmss.*

denses sur l'aire de Lyon : ils sont donc plus dispersés topographiquement que l'ensemble des autres commerces non alimentaires. Pour ce secteur, une dispersion de ces établissements commerciaux sera donc détectée au sens de D (graphique IXC).

Ce résultat est comparable à celui de K_d, à la différence importante que D est une fonction cumulative. En effet, K_d compare les probabilités de trouver des voisins à la distance r dans la réalité et sous l'hypothèse nulle. Comme l'espace de référence est le même, compter des magasins (mesure absolue) ou calculer leur densité (mesure topographique) est équivalent. Comparer K_d à sa valeur sous l'hypothèse nulle ne lui donne pas le caractère d'une mesure relative mais d'une mesure topographique. La valeur empirique de K_d et ses valeurs simulées sont calculées sur le même espace : si la probabilité de trouver un voisin à distance r est plus élevée selon les données que sous l'hypothèse nulle, la densité de voisins (par unité de surface) l'est aussi.

Les résultats de M, K_d et D obtenus sont parfaitement en cohérence avec le type de concentration spatiale analysée : M est un indice relatif donc détecte une concentration relative, D mesure une concentration topographique et K_d identifie la concentration absolue (sans autre référentiel que le nombre de voisins observés) mais la comparer à sa valeur sous l'hypothèse nulle la fait se comporter comme une mesure topographique.

* *
*

La conclusion de notre comparaison de mesures souligne l'importance du choix de l'outil et plus précisément du type de concentration qu'il met en évidence. La mesure de la concentration spatiale a fait l'objet d'une grande attention de la part des économistes depuis une dizaine d'années. Duranton et Overman (2005) ont proposé une liste de critères pour mesurer la concentration spatiale. Dans notre article, nous montrons que les résultats peuvent sensiblement diverger en retenant différentes mesures qui respectent les mêmes critères. Par conséquent, cette liste doit donc être complétée : notamment le critère requérant que la mesure prenne en compte la distribution globale de l'activité économique doit être précisé (Duranton et Overman, 2005, p. 1079). Les développements futurs sont indispensables pour améliorer l'intégration de l'outil à la théorie économique comme le soulignent Combes et Overman (2004). Le choix de l'indice de concentration reposera alors sur une démarche méthodologique justifiée théoriquement pour utiliser l'indice approprié correspondant à la question posée. □

BIBLIOGRAPHIE

Amiti M. (1999), « Specialization Patterns in Europe », *Weltwirtschaftliches Archiv*, vol. 135, n° 4, pp. 573-593.

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

Arbia G. (2001), « The role of spatial effects in the empirical analysis of regional concentration », *Journal of Geographical Systems*, vol. 3, n° 3, pp. 271-281.

Arbia G., Espa G. et Quah D. (2008), « A class of spatial econometric methods in the empirical analysis of clusters of firms in the space », *Empirical Economics*, vol. 34, n° 1, pp. 81-103.

Baddeley A. J., Møller J. et Waagepetersen R. (2000), « Non- and semi-parametric estimation of interaction in inhomogeneous point patterns », *Statistica Neerlandica*, vol. 54, n° 3, pp. 329-350.

Baddeley A. et Turner R. (2005), « spatstat: An R Package for Analyzing Spatial Point Patterns », *Journal of Statistical Software*, vol. 12, n°6, pp. 1-42.

Barlet M., Briant A. et Crusson L. (2013), « Location patterns of service industries in France: A distance-based approach », *Regional Science and Urban Economics*, vol. 43, n° 2, pp. 338-551.

Barlet M., Crusson L., Dupuch S. et Puech F. (2011), « Des services échangés aux services échangeables : une application sur données françaises », *Économie et Statistique*, n° 435-436, pp. 105-124.

Béguin H. (1979), *Méthodes d'analyse géographique quantitative*, Librairies Techniques (LITEC), Paris.

Behrens K. et Bougna T. (2013), « An Anatomy of the Geographical Concentration of Canadian Manufacturing Industries », *Cahier de recherche/ Working paper* CIRPEE 13-27.

Briant A., Combes P.-P. et Lafourcade M. (2010), « Dots to boxes: Do the Size and Shape of Spatial Units Jeopardize Economic Geography Estimations? », *Journal of Urban Economics*, vol. 67, n° 3, pp. 287-302.

Brülhart M. et Traeger R. (2005), « An Account of Geographic Concentration Patterns in Europe », *Regional Science and Urban Economics*, vol. 35, n° 6, pp. 597-624.

Combes P.-P. et Overman H. G. (2004), « The spatial distribution of economic activities in the European Union », dans J. V. Henderson et J.-F. Thisse (eds), *Handbook of Urban and Regional Economics*, vol. 4, North Holland, Amsterdam, Elsevier, pp. 2845-2909.

Combes P.-P., Mayer T. et Thisse J.-F. (2006), *Économie géographique. L'intégration des régions et des nations*, Economica.

Crozet M. et Lafourcade M. (2010), *La nouvelle économie géographique*, Collection Repères, La Découverte.

Diggle P. J. et Chetwynd A. G. (1991), « Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations », *Biometrics*, vol. 47, n° 3, pp. 1155-1163.

Duboz M.-L., Guillain R. et Le Gallo J. (2009), « Les schémas de concentration sectorielle au sein de l'Union européenne : l'Est miroir de l'Ouest ? ». *Économie et Statistique*, n° 423, pp. 59-76.

Duranton G. et Overman H. G. (2005), « Testing for localization using micro-geographic data », *Review of Economic Studies*, vol. 72, n° 4, pp. 1077-1106.

Duranton G. et Overman H. G. (2008), « Exploring the Detailed Location Patterns of UK Manufacturing Industries using Microgeographic Data », *Journal of Regional Science*, vol. 48, n° 1, pp. 213-243.

Ellison G. et Glaeser E. L. (1997), « Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach », *Journal of Political Economy*, vol. 105, n° 5, pp. 889-927.

Ellison G., Glaeser E. L. et Kerr W. R. (2010), « What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns », *American Economic Review*, vol. 100, n° 3, pp. 1195-1213.

Fratesi U. (2008), « Issues in the measurement of localization », *Environment and Planning A*, vol. 40, n° 3, pp. 733-758.

Fujita M., Krugman P. et Venables A. J. (1999), *The Spatial Economy*, The MIT Press, Cambridge.

Fujita M. et Thisse J.-F. (2002), *Economics of Agglomeration: Cities, Industrial Location and Regional Growth*, Cambridge University Press, New York.

Gaines K. F., Bryan A. L. Jr. et Dixon P. M. (2000), « The Effects of Drought on Foraging Habitat Selection of Breeding Wood Storks in Coastal Georgia », *Waterbirds: The International Journal of Waterbird Biology*, vol. 23, n°1, pp. 64-73.

Gibbons S., Overman H. G. et Patacchini E. (2014), *Spatial Methods*, dans G. Duranton, J.V. Henderson and W.C. Strange (eds), *Handbook of Regional & Urban Economics*, vol. 5, North Holland, Amsterdam, Elsevier (chapitre à paraître).

Gini C. (1912), *Variabilità e mutabilità*, C. Cuppini, Bologna.

Guimarães P., Figueiredo O. et Woodward D. (2011), « Accounting for neighboring effects in measures of spatial concentration », *Journal of Regional Science*, vol. 51, n° 4, pp. 678-693.

Haaland J. I., Kind H. J., Midelfart-Knarvik K. H. et J. Torstensson (1999), « What Determines the Economic Geography of Europe? », *CEPR discussion paper* n° 2072.

Haase P. (1995), « Spatial pattern analysis in ecology based on Ripley's K function: Introduction and methods of edge correction », *Journal of Vegetation Science*, vol. 6, n° 4, pp. 575-582.

Harkness R. et Isham V. (1983), « A Bivariate Spatial Point Pattern of Ants' Nests », *Applied Statistics*, vol. 32, pp. 293-303.

Holmes T. J. et Stevens J. J. (2004), « Spatial Distribution of Economic Activities in North America », dans J. V. Henderson et J.-F. Thisse (eds), *Handbook of Urban and Regional Economics*, vol. 4, North Holland, Amsterdam, Elsevier, pp. 2797-2843.

Houdebine M. (1999), « Concentration géographique des activités et spécialisation des départements français », *Économie et Statistique*, n° 326-327, pp. 189-204.

Jensen P. et Michel J. (2011), « Measuring spatial dispersion: exact results on the variance of random spatial distributions », *The Annals of Regional Science*, vol. 47, n° 1, pp. 81-110.

Kingham S. P., Gatrell A. C. et Rowlingson B. S. (1995), « Testing for clustering of health events within a geographical information system framework », *Environment and Planning A*, vol. 27, n° 5, pp. 809-821.

Klier T. et McMillen D. P. (2008), « Evolving Agglomeration In The U.S. Auto Supplier Industry », *Journal of Regional Science*, vol. 48, n° 1, pp. 245-267.

Koh H.-J. et Riedel N. (2014), « Assessing the Localization Pattern of German Manufacturing and Service Industries: A distance-based Approach », *Regional Studies*, vol. 48, n° 5, pp. 823-843.

Krugman P. (1991), *Geography and Trade*. MIT Press.

Marcon E., Lang G., Traissac S. et Puech F. (2012a), « dbmss: Distance-based measures of spatial structures », téléchargeable sur : http://cran.r-project.org/web/packages/dbmss/index.html (dernière consultation le 07 janvier 2015).

Marcon E. et Puech F. (2003), « Evaluating the Geographic Concentration of Industries Using Distance-Based Methods », *Journal of Economic Geography*, vol. 3, n° 4, pp. 409-428.

Marcon E. et Puech F. (2010), « Measures of the Geographic Concentration of Industries: Improving Distance-Based Methods », *Journal of Economic Geography*, vol. 10, n° 5, pp. 745-762.

Marcon E. et Puech F. (2014), « A typology of distance-based measures of spatial concentration », *HAL-SHS Working Paper*, n° halshs-00679993, version 3.

Marcon E., Puech F. et Traissac S. (2012b), « Characterizing the relative spatial structure of point patterns », *International Journal of Ecology*, Article ID 619281.

Marshall, A. (1890), *Principle of Economics*. Macmillan, London.

Matérn, B. (1960), « Spatial variation », *Meddelanden från Statens Skogsforskningsinstitut* vol. 49, n° 5, pp 1-144.

Maurel F. et Sédillot B. (1999), « A Measure of the Geographic Concentration in French Manufacturing Industries », *Regional Science and Urban Economics*, vol. 29, n° 5, pp. 575-604.

Midelfart-Knarvik K. H., Overman H. G, Redding S. et Venables A. J. (2002), « Integration and Industrial Specialisation in the European Union », *Revue économique*, vol. 53, n° 3, pp. 469-481.

Moeur M. (1993), « Characterizing spatial patterns of trees using stem-mapped data », *Forest Science*, vol. 39, n° 4, pp. 756-775.

Nakajima K., Saito Y. U. et Uesugi I. (2012), « Measuring economic localization: Evidence from Japanese firm-level data », *Journal of the Japanese and International Economies*, vol. 26, n° 2, pp. 201-220.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, Vienna, Austria, R Foundation for Statistical Computing.

Ripley B. D. (1976), « The Second-Order Analysis of Stationary Point Processes », *Journal of Applied Probability*, vol. 13, n° 2, pp. 255-266.

Ripley B. D. (1977), « Modelling Spatial Patterns », *Journal of the Royal Statistical Society* B, vol. 39, n° 2, pp. 172-212.

Ripley B. D. (1981), *Spatial statistics*, John Wiley & Sons, New York.

Rosenthal S. S. et Strange W. C. (2001), « The determinants of agglomeration », *Journal of Urban Economics*, vol. 50, n° 2, pp. 191–229.

Rosenthal S. S. et Strange W. C. (2004), « Evidence of the nature and sources of agglomeration economies », dans J. V. Henderson et J.-F. Thisse (eds), *Handbook of Urban and Regional Economics*, vol. 4, North Holland, Amsterdam, Elsevier, pp. 2119-2171.

Silverman B. W. (1986), *Density estimation for statistics and data analysis*, Chapman and Hall, London.

Thomas-Agnan C. et Bonneu F. (2014), « Measuring and testing spatial mass concentration of micro-geographic data », *TSE Working Papers*, n° 474, janvier 2014.

Wiegand T. et Moloney K. A. (2004), « Rings, circles, and null-models for point pattern analysis in ecology », *Oikos*, vol. 104, n° 2, pp. 209-229.